Uruguay Presidencia | agesic

Marco de trabajo para el análisis de datos

AGESIC

Versión 0.1

2020

Contenido

1. Introducción

La Agencia para el Desarrollo del Gobierno de Gestión Electrónica y la Sociedad de la Información y del Conocimiento (AGESIC) se encuentra impulsando una serie de iniciativas para guiar la transformación digital y la innovación para el fortalecimiento de la relación entre los ciudadanos y el Gobierno.

A partir de este contexto (el cual se profundiza en la sección 2. Antecedentes) y considerando que el análisis de datos de gobierno permite colaborar directa e indirectamente con una gestión gubernamental eficiente y con la definición de políticas públicas basadas en evidencia, se ha definido un **Marco de trabajo para el Análisis de Datos**.

El principal objetivo es proporcionar a organismos estatales de la Administración y a otras entidades vinculadas a la gestión gubernamental, un conjunto de lineamientos y recomendaciones prácticas que guíen y faciliten los distintos tipos de análisis de datos requeridos. Esta primera versión, es disponibilizada y será mantenida y evolucionada por AGESIC, de forma de profundizar y actualizar los lineamientos conforme avance la práctica de analítica gubernamental, así como también en base a la evolución de las disciplinas desarrolladas.

En base a una revisión de las mejores prácticas, antecedentes y situación actual del análisis de datos de gobierno en Uruguay, se han definido las siguientes dimensiones que se desarrollarán en el presente marco de trabajo:

- Gobernanza.
- Tecnología.
- Ciberseguridad.
- Legal.

La dimensión **Gobernanza** [87] define distintos tipos de análisis así como también un proceso de análisis de referencia que será utilizado para desarrollar las buenas prácticas. Si bien existen distintos tipos de análisis de datos de gobierno posible, se entiende el mencionado proceso recoge aquellas actividades comunes a los mismos, facilitando la comprensión y guía en el uso de lineamientos y recomendaciones. Además, se establecen un conjunto de roles y funciones asociadas, categorizando a los mismos como básicos (necesarios para análisis de datos de baja complejidad) y avanzados (requeridos para análisis más complejos, que requieren mayor nivel de madurez y especialización dentro de la organización). Para cada rol se detallan las principales funciones y actividades, con sus respectivas responsabilidades y las capacidades deseables a cumplir por las personas que los desempeñen. Asimismo, se indicará cuál será su participación dentro de cada etapa del proceso de análisis.

La dimensión **Tecnología** establece un conjunto de arquitecturas tecnológicas asociadas a los diferentes tipos de análisis. Para desarrollar las mismas y considerando las mejores prácticas, se utilizó un enfoque de arquitectura empresarial, especificando los distintos dominios (infraestructura, datos y aplicaciones) desarrollando un análisis detallado de las mismas y evaluando

comparativamente distintas soluciones y herramientas en base distintas fuentes de información debidamente referenciadas. Este insumo se entiende clave a los efectos de optimizar los esfuerzos por parte de organismos, facilitando información para la toma de decisiones sobre posibles soluciones tecnológicas que soporten los procesos y actividades del análisis de datos de gobierno.

La dimensión de **Ciberseguridad**, se considera de particular relevancia para la práctica de análisis de datos de gobierno, puesto que desarrolla conceptos y medidas vinculadas a la prevención, detección y control y resiliencia. La preocupación de las organizaciones por la ciberseguridad ha aumentado sistemáticamente conforme al avance en cantidad y sofisticación de los ataques informáticos tanto externos como internos, por tanto, se entiende fundamental su aplicación en el marco de políticas y procedimientos existentes, pero que requieran fortalecimiento y/o complementación para el análisis de datos.

Por último, la dimensión **Legal** describe los principales componentes del marco legal vigente en el Uruguay, describiendo las principales consideraciones y recomendaciones para el cumplimiento de la Ley 18.331 (Protección de Datos Personales).

1.1. Estructura

A continuación, se describen brevemente las principales secciones del marco de trabajo:

- Antecedentes: se describen un conjunto de antecedes e insumos que fueron considerados para el desarrollo del marco de trabajo, considerando: a) iniciativas gubernamentales desarrolladas en Uruguay, b) grupos de trabajo y posicionamiento de Uruguay en evaluaciones internacionales, c) Iniciativas regionales e internacionales y d) estándares, mejores prácticas y marcos de referencia.
- Modelo conceptual: se definen y describen cada uno de las dimensiones del marco de trabajo, con sus correspondientes interrelaciones y referencias.
- Aplicación: en base al proceso de análisis de datos, para cada dimensión del marco de trabajo se detallan un conjunto de recomendaciones y buenas prácticas especificando las actividades del proceso impactadas.
- Glosario: presenta conceptos relevantes para el correcto entendimiento del marco de trabajo.

3. Antecedentes

Existen diversos antecedentes e insumos que revisten interés para el desarrollo y evolución del presente marco de trabajo, los cuales son descritos a continuación en base a la siguiente categorización:

3.1 Iniciativas gubernamentales desarrolladas en Uruguay.
3.2 Grupos de trabajo y posicionamiento de Uruguay en evaluaciones internacionales.
3.3 Iniciativas regionales e internacionales.
3.4 Estándares, mejores prácticas y marcos de referencia.

3.1. Iniciativas gubernamentales desarrolladas en Uruguay

Plan de Gobierno Digital 2020

Una de las principales propuestas para la transformación digital es el **Plan de Gobierno Digital 2020** (PGD) [1], que establece un mapa de ruta dinámico integrando seis áreas de acción complementarias. El Plan busca ser un instrumento acelerador de cambios en estas áreas, impulsando el uso intensivo de tecnologías como Internet, dispositivos móviles, plataformas compartidas y el aprovechamiento de los datos como parte integral de sus políticas de transformación. Su objetivo es crear valor público mediante servicios que satisfagan las necesidades, expectativas y preferencias de los ciudadanos de forma abierta, cercana, inteligente, eficiente, integrada y confiable.

El gobierno digital es de vital importancia para el presente trabajo, dado que aquí radican las distintas iniciativas en las que ha incurrido Uruguay y que lo guían hacia la creación del Marco de trabajo. En relación al PGD, el mismo cuenta con seis áreas de acción como base de este plan:

- Gobierno Cercano.
- Gobierno Abierto.
- Gobierno Inteligente.
- Gobierno Eficiente.
- Gobierno Integrado.
- Gobierno Digital Confiable.

El PGD [1] busca con el **Gobierno Cercano** (Sección 2.1 del PGD) aprovechar al máximo las ventajas que ofrecen las tecnologías digitales para potenciar una atención omnicanal, con servicios de calidad que mejoren la experiencia del ciudadano en su relacionamiento con el Estado. Una experiencia omnicanal implica que un ciudadano no solo pueda hacer trámites o gestiones en Internet, sino que pueda hacerlo por el canal que prefiera –ya sea presencial, telefónico, e-mail, web, móvil u otros– con la certeza de que obtendrá el mismo resultado.

En la Sección 2.2 del PGD, se establece el concepto de **Gobierno Abierto**, en donde mediante la tecnología se busca que los ciudadanos interactúen con el Gobierno de forma directa. Allí se impulsa el concepto de datos abiertos, entendido por aquellos datos que se encuentran disponibles en un formato estándar que

permite su interoperabilidad y que son accesibles al público. El uso de dichos datos, está regido por la normativa vigente, como se explicará pertinentemente en el componente Legal.

Dicha iniciativa provee un conjunto importante de datos que deben ser analizados de forma inteligente para así sacarle el mayor provecho, por lo que es necesario una plataforma que habilite un adecuado análisis de datos.

Otra de las áreas del PGD, estrechamente vinculado a la necesidad de generar un marco para el análisis de datos, es la de **Gobierno Inteligente** (Sección 2.3 del PGD) cuyo objetivo principal es el aprovechamiento de los datos, información y conocimiento como activos para optimizar y mejorar la experiencia de los servicios públicos que brinda el estado. Para esto se busca fortalecer los mecanismos de toma de decisión mediante la tecnología, desarrollando plataformas y modelos analíticos predictivos. Logrando, así, un Estado que actúe como una unidad, intensificando el aprovechamiento de los datos para la toma de decisiones, orientación de políticas públicas y la mejora permanente, anticipándose a las necesidades.

También se plantea el **Gobierno Eficiente**, en la sección 2.4 del PGD, el cual apunta a optimizar el uso de los servicios para reducir costos de operación y modernizar los procesos con un enfoque integral, indispensable para el adecuado desarrollo del Gobierno Digital. Pretende mejorar la gestión transversal e impulsar su adopción en los organismos del Estado, quienes podrán servirse de recursos comunes que contribuyan a la agilidad de los avances.

En la sección 2.5 se explica el concepto de **Gobierno Integrado**, busca la integración tecnológica entre los diferentes organismos del Estado, así como la integración entre el Estado, la ciudadanía, la industria y la academia. Potencia la integración tecnológica y la interoperabilidad de los datos como base del desarrollo y la evolución de los sistemas de gestión. También continuar trabajando en la concepción de un Gobierno Integrado permitirá a las organizaciones públicas intercambiar información en forma oportuna y consistente; además de mejorar la gestión y la creación de nuevos sistemas de análisis que optimicen las políticas públicas y los servicios al ciudadano.

Por último, **el Gobierno Digital Confiable** (Sección 2.6 del PGD), vela por responder a los riesgos, amenazas y desafíos que surgen con el desarrollo de las tecnologías digitales. Se enfoca en generar y hacer disponibles marcos que proporcionen seguridad y confianza en la aplicación y evolución del Gobierno Digital. Propone seguir avanzando en un ecosistema de ciberseguridad, la gestión de riesgos y continuidad operativa, la universalización de la gestión electrónica, y la privacidad y protección de datos personales. Entre sus propuestas se encuentra mejorar las condiciones en el combate del cibercrimen, mediante la adecuación del Marco Normativo de Ciberseguridad, el mismo da soporte a la dimensión Ciberseguridad de presente marco de trabajo.

Política de Datos para la Transformación Digital

La **Política de Datos para la Transformación digital,** fue aprobada en el año 2019 por un grupo de trabajo interinstitucional con el objetivo de impulsar una estrategia nacional de datos que promueva y desarrolle proyectos específicos para la gestión de datos en el Estado. Define el ciclo de vida de los datos de la siguiente manera:

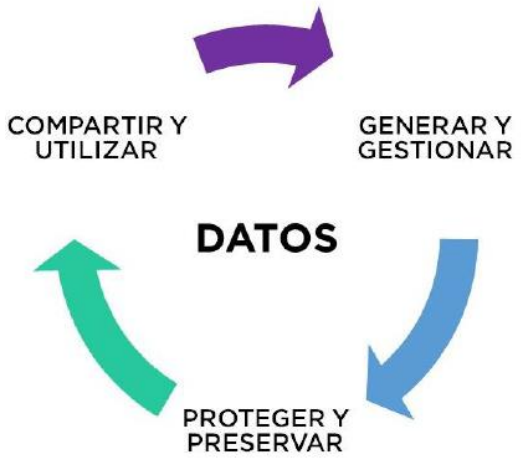

Figura 1 Ciclo de vida de los datos. Fuente: [2]

y sus principios generales son:

- **Generación:** cada organismo recolecta, produce y/o elabora datos. Dichos datos deben poder ser consumidos por otros organismos.
- **Eficiencia:** los organismos deben gestionar de forma eficiente los datos que estén bajo su responsabilidad.
- **Calidad:** los datos deben ser exactos, oportunos y conformes con la realidad.
- **Acceso a los datos:** los datos de los organismos deben estar disponibles para los ciudadanos siempre que no se aplique alguna excepción.
- **Compartir y utilizar:** los organismos deben compartir datos siguiendo los estándares de datos, integración e intercambio establecidos.
- **Datos abiertos:** Los datos del sector público deberán ser abiertos por defecto. Los principios vigentes de datos abiertos del Uruguay se consideran parte integrante de la política de datos para la transformación digital.
- **Protección de datos:** los organismos deben proteger los datos.
- **Seguridad:** los organismos implementarán procedimientos de gestión de los datos para brindar niveles de confiabilidad, integridad, disponibilidad y autenticidad adecuados.
- **Preservación:** Los datos deben mantener su integridad y asegurar su disponibilidad durante el tiempo necesario, según la normativa vigente.

Proyectos estratégicos vinculados a datos impulsados por AGESIC

AGESIC ha desarrollado diversas iniciativas para facilitar la interoperabilidad entre organismos del estado y/o mejorar los servicios de gobierno digital en general. A modo de ejemplo, se describen las siguientes:

- Arquitectura de Gobierno.
- Arquitectura empresarial.
- Arquitectura de datos.
- Plataforma de interoperabilidad (PDI).
- Plataforma de datos.
- Centro Nacional de Respuesta a Incidentes de Seguridad Informática (CERT.uy).

La **Arquitectura de Gobierno** [30] brinda modelos que permitan la prestación de servicios de tecnología consistentes y cohesivos a los ciudadanos y a los organismos del Estado, facilitando la prestación de servicios de TI optimizando costo y eficacia, de manera oportuna, a través de un repositorio de estándares, principios y modelos que asisten en el diseño y entrega de las capacidades de TI, así como, a la vez, servicios del negocio a los ciudadanos.

Mediante la aplicación del marco TOGAF (The Open Group Architecture Framework) para **Arquitectura Empresarial**, se homogeneiza la forma de describir los componentes de la Arquitectura de Gobierno, mediante la estandarización de terminología, componentes, sus relaciones con otros componentes y con componentes externos, así como la definición de principios de arquitectura para los requerimientos, diseño y evolución de las arquitecturas. Por lo tanto, la arquitectura empresarial permite definir de forma rigurosa la estructura de una empresa u organización.

La **Arquitectura de Datos** orienta a los organismos a realizar una gestión de los datos eficiente, proponiendo un esquema de componentes compuesto por herramientas, buenas prácticas y estándares [32]. Sus objetivos son entender las necesidades de la información, para desarrollar y mantener el modelo de datos definido. Definir y mantener la arquitectura de: Base de Datos, integración de datos, Big Data y analítica y metadatos.

A continuación se presentan los principales componentes del modelo de referencia de datos y una breve descripción de los mismos [33]:

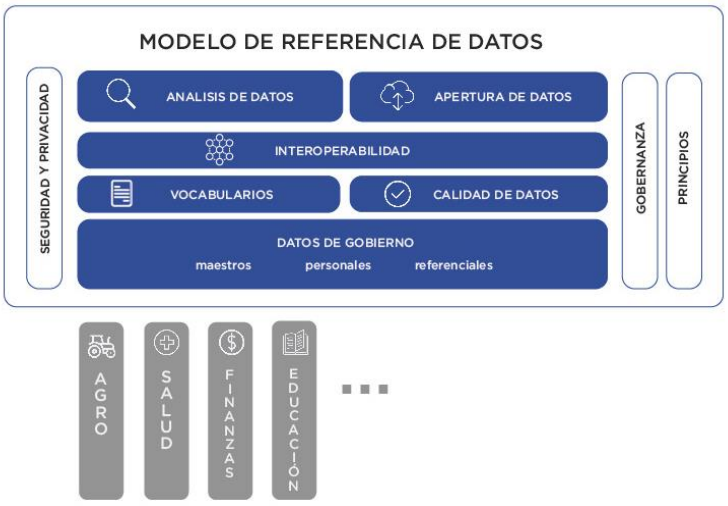

Figura 2 Modelo de referencia de datos. Fuente: [33]

- **Datos de Gobierno:** datos maestros, datos de referencia y datos personales.
- **Calidad de datos:** medición, control y mantenimiento de los procesos de calidad de datos.
- **Vocabularios:** elemento semántico para la interoperabilidad y el análisis de datos.
- **Interoperabilidad:** abarca la interoperabilidad técnica y la semántica. Otorga mecanismos para que los distintos sistemas accedan a los datos de las fuentes correctas y evitar la duplicación de información.
- **Análisis de datos:** obtener información de los datos, mejorando los procesos internos y los servicios brindados a los ciudadanos.
- **Apertura de datos:** cada organismo debe elaborar los procesos necesarios para la apertura y publicación de datos abiertos.
- **Seguridad y privacidad, Gobernanza y Principios:** dichos componentes acompañan el ciclo de vida de los datos y son la base para una correcta gestión de los mismos.

La **PDI** [29] forma parte de la Plataforma de Gobierno Digital de AGESIC y tiene como objetivo general facilitar y promover la implementación de servicios de Gobierno Digital en Uruguay. Para esto, la PDI brinda mecanismos que apuntan a simplificar la integración entre los organismos del Estado y a posibilitar un mejor aprovechamiento de sus activos.

La **Plataforma de Datos** tiene como objetivo permitir la obtención de información relevante, basada en evidencia, oportuna, integrada y de calidad [31]. Dicha plataforma cuenta con los siguientes componentes:

Figura 3 Plataforma de datos. Fuente: [31]

- **Calidad:** brinda modelos de calidad de datos aplicados sobre los vocabularios definidos.
- **Interoperabilidad semántica:** disponibiliza vocabularios definidos para registros federados, conjuntos de datos de referencia y establece un punto único de acceso a los registros del Estado, brindando mecanismos de consulta y administración sobre ellos. Con esto se busca unificar el criterio para el entendimiento de los datos del Estado.
- **Interoperabilidad técnica:** la interoperabilidad técnica permite asegurar y facilitar la comunicación entre los sistemas de información de los organismos del Gobierno, como forma de intercambiar datos y reutilizarlos. Por lo tanto, en la plataforma se amplía dicha interoperabilidad, de forma de agregar protocolos de intercambio, niveles de seguridad, capacidad de integración, orquestación de servicios y transformación de mensajes, entre otras.

- **Seguridad y Privacidad:** garantiza la seguridad de los datos transmitidos y la privacidad de los datos personales. Brinda mecanismos para que los ciudadanos puedan establecer políticas de acceso sobre sus datos personales, así como instrumentos de auditoría sobre los mismos.

El **CERT.uy** fue creado en 2009, con el objetivo de brindar respuestas y análisis continuo a los incidentes de seguridad a nivel gubernamental en Uruguay, proveyendo un seguimiento estadístico de los mismos con frecuencia semestral [89].

Está integrado por el Centro de Respuesta a Incidentes de Ciberseguridad (CERT) y el Centro de Operaciones de Ciberseguridad (SOC).

- **CERT:** asiste en las respuestas de incidentes de ciberseguridad en los organismos estatales y apoya en la implementación de la estrategia de Gestión de Incidentes definida a nivel nacional; además se encarga de desplegar y administrar la infraestructura de ciberseguridad para la gestión de incidentes, colaborando también en la implementación de sistemas seguros en el Estado.

- **SOC:** tiene por principal objetivo detectar en tiempo real eventos e incidentes de ciberseguridad en los activos de información críticos del Estado, así como recolectar y analizar información de ciberseguridad para prevenir y detectar incidentes de la materia.

Con respecto a los últimos datos relevados correspondientes al año 2018, se registró un aumento significativo de incidentes de seguridad informática respecto al mismo período del año anterior. Esto confirma la tendencia detectada en 2017, año en el que el CERT.uy respondió a una cantidad que representó un crecimiento de más del doble con respecto al 2016. Uno de los factores que se cree tuvo una influencia directa sobre este crecimiento es la incorporación del SOC, creado a mediados de 2017. Esto se deriva como se comentó con anterioridad, de que uno de sus principales objetivos es mejorar la capacidad operativa en la detección de incidentes de ciberseguridad, tomando un enfoque 24x7 e incorporando también procesamiento y análisis de grandes volúmenes de datos en el proceso. Estas tendencias validan y reafirman la necesidad de considerar el componente de ciberseguridad como un elemento central en las distintas iniciativas tecnológicas a nivel nacional.

Desde el punto de vista de análisis de datos, es importante destacar que pueden ocurrir incidentes de seguridad en los organismos del Estado Uruguayo, los cuales están obligados por la legislación nacional a reportarlos al CERT.uy, como se indica en la Sección 5.3. Adicionalmente, se establece por ley [93] que toda unidad ejecutora de la Administración Central debe contar con un Responsable de Seguridad de la Información (RSI) designado.

Por último, cabe mencionar que Uruguay cuenta con legislación sobre el manejo de datos y en especial, sobre la protección de los datos personales. Los derechos fundamentales de libertad de información, dar y recibir información, y de intimidad o privacidad, dentro de los cuales está la protección de los datos personales, se encuentran consagrados en la Constitución Nacional y constituyen

verdaderas garantías para el accionar de los individuos. A su vez, el manejo y la protección de los datos se ha vuelto un elemento sustancial debido a los riesgos que esta actividad implica. En particular, la protección de los datos personales se ve cada vez más profundizada o robustecida en la medida que los titulares de los datos toman conciencia de la situación, fundamentalmente de los riesgos que implica la utilización de sus datos por parte de terceros sin consentimiento ni control, y las autoridades públicas cuentan con más facultades para actuar en defensa de los intereses de los titulares, tanto en el ámbito administrativo como judicial.

La propia Ley **No. 18.331** [11] del 11 de agosto de 2008, regulatoria de la materia, contempla en su primer artículo que: "*El derecho a la protección de datos personales es inherente a la persona humana, por lo que está comprendido en el art. 72 de la Constitución de la República*". Y agrega en su **art. 6** que: "*Las bases de datos no pueden tener finalidades violatorias de derechos humanos (...)*".

3.2. Grupos de trabajo y posicionamiento de Uruguay en evaluaciones internacionales.

Uruguay forma parte de distintos grupos de trabajo a nivel internacional y ha sido evaluado mediante diversos rankings globales respecto a temáticas vinculadas directa o indirectamente el análisis de datos de gobierno. A continuación, se describe el posicionamiento de nuestro país respecto a:

- Digital Nations (DN).
- E-Government Development Index (EDGI).
- Índice global de ciberseguridad.

Digital Nations (DN)

Uruguay pertenece al grupo denominado **DN** [4], una red colaborativa de países cuyos Gobiernos son ampliamente digitalizados. El DN busca generar un entorno en donde se puede compartir experiencias relacionadas a la transformación digital para apoyar definir acciones e iniciativas que permitan a los países contar con gestiones gubernamentales cada vez más digitales, ágiles y eficientes. A modo de ejemplo, el DN supone un espacio de trabajo sobre la inteligencia artificial tanto en áreas como la salud, educación, industria, agricultura y energía. En este sentido, cabe destacar que AGESIC promueve una estrategia en el uso de la Inteligencia Artificial en el Gobierno digital [5].

En la quinta cumbre del DN, desarrollada en Israel en 2018, Uruguay planteó generar un grupo temático que aborde diferentes aspectos relacionados con los datos públicos, propuesta que fue acogida por todos los miembros. Uruguay se incorpora al grupo en 2018, siendo junto a México, los primeros países miembro de la región de América Latina y el Caribe, posicionándose así al nivel de los países líderes en el área de gobierno digital. El primero de enero de 2019, Uruguay asume el rol de la presidencia del foro. En la Cumbre ministerial del DN realizada en Uruguay entre el 4 y 6 de noviembre de 2019, se acordó la iniciativa **Datos 360°** (plasmada en la "Declaración de Datos de Montevideo"), con la cual se pretende que

los datos se utilicen de forma inteligente, apoyando en la toma de decisiones y potenciando los servicios ofrecidos por los entes públicos [6].

E-Government Development Index (EDGI).

Las Naciones Unidas elaboran de forma regular éste índice, de forma de evaluar tres dimensiones del Gobierno electrónico: provisión de servicios en línea, conectividad de telecomunicaciones y capacidad humana. En base a la última evaluación realizada en 2020, Uruguay es el país con mejor posicionamiento a nivel de Latinoamérica y segundo en las Américas[7].

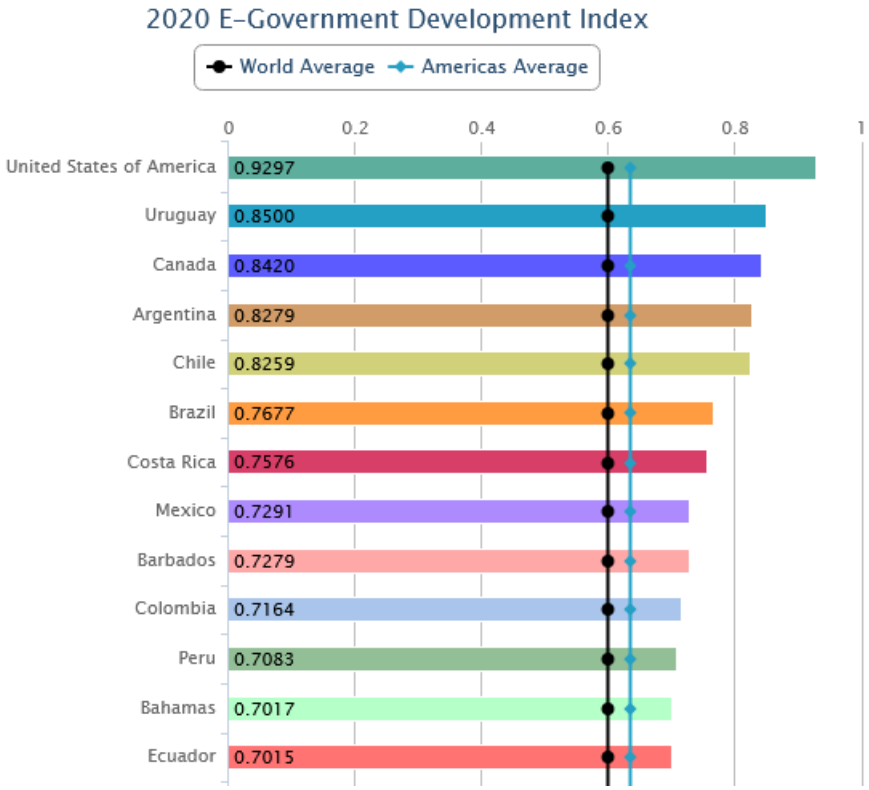

Figura 4 Puntaje EDGI. Fuente: [7]

Índice global de ciberseguridad

El **Índice Global de Cibserseguridad** [9], desarrollado por la International Telecommunication Union (ITU), mide el compromiso de los países con respecto a la ciberseguridad y se evalúa en los siguientes pilares: medidas legales, medidas técnicas, medidas organizativas, desarrollo de capacidades y cooperación. En 2018 Uruguay figura entre los países con nivel de iniciativa alto[1], siendo el único con dicho nivel en América Latina y el Caribe, liderando así la región de las Américas junto a Canadá y Estados Unidos. En particular, se destaca la robustez del pilar organizacional uruguayo, así como la estrategia de ciberseguridad nacional y el marco utilizado para medir el desarrollo del área.

[1] Niveles considerados: bajo, medio y alto. La clasificación surge del puntaje total evaluado en el índice [9].

3.3. Iniciativas regionales e internacionales

A continuación, se presentan algunas iniciativas que revisten interés desde el punto de vista de análisis de datos de gobierno, tanto en la región como a nivel internacional.

3.3.1. A nivel regional

En **América Latina y el Caribe** se han desarrollado diversas iniciativas tendientes a desarrollar una filosofía de datos abiertos, de forma de fomentar la participación de la ciudadanía en la toma de decisiones. Las mismas, plantean mediante el uso de las TICs desarrollar sistemas y soluciones eficientes, escalables y dinámicos, que integren la operación de los organismos del Estado resolviendo necesidades y problemáticas a nivel de: interoperabilidad, compatibilidad, aceleración de procesos de gestión, seguridad y privacidad.

La **CEPAL** (Comisión Económica para América Latina y el Caribe) [34, 35, 36] propuso en 2019 el uso de Big Data para mejorar las capacidades nacionales en la región de América Latina y el Caribe, en particular medir la economía digital utilizando analítica de grandes datos para basar decisiones a nivel regional e internacional.

Colombia a partir del año 2017 ha promovido distintas iniciativas vinculadas al uso de Big Data, Internet de las Cosas, Robótica e Inteligencia Artificial [20, 80]. En particular, el Ministerio TIC creó Centros de Excelencia y Apropiación (CEA) con el fin de capitalizar el análisis de datos en sectores estratégicos. Uno de los objetivos del Ministerio TIC es la creación de valor a partir de Big Data Analytics para: Ciberseguridad, Internet de las cosas y la formulación de política pública. Desde marzo de 2016, se encuentra operando el Centro de Excelencia en Big Data y Data Analytics para desarrollar herramientas y capacidades que permitan manejar y analizar grandes volúmenes de información, para el cual contó con el apoyo del MIT y Data-Pop Alliance tal como se comentó anteriormente [41]. Algunos de los estudios y documentos generados para el mencionado país sobre temáticas vinculadas a Big Data es resumido a continuación [20]:

- Análisis comparativo de estrategias nacionales de Big Data.
- Buenas prácticas que pueden llevar a cabo los Gobiernos para implementar una estrategia de Big Data.
- Diagnóstico de la situación del país en el período 2017-2019.
- Acciones a corto, mediano y largo plazo que le permiten al Gobierno superar los desafíos técnicos, sociales y de capital humano con respecto a datos y Big Data.

En **Argentina** se publicaron gran cantidad de datasets, impulsando la cultura de datos abiertos, entre los cuales se encuentran datos relacionados con actividades de economía, educación, sociedad, seguridad, turismo, cultura y transporte. A su vez, se tiene un plan de estrategia para llevar a cabo los cambios necesarios en la gestión del Gobierno para pasar a un Gobierno electrónico, agilizando los procesos, aplicando buenas prácticas. Este plan posibilitó la interoperabilidad entre las

diversas áreas de Gobierno, teniendo como resultado contar con trámites más eficientes de cara al ciudadano [42, 43].

3.3.2. A nivel internacional

La Comisión Económica Europea de las **Naciones Unidas**, desarrolló una plataforma de Big Data que involucró personal de 20 países [47]. Se realizaron distintos pilotos con información administrativa y redes sociales, entre otros [81].

Hacia finales del 2015, el Gobierno de **Australia** elaboró un informe sobre la gestión de datos en el sector público en el cual llegaron a la conclusión de que existe un suministro insuficiente de datos y análisis a nivel global que limita la capacidad de obtener el máximo valor de los datos disponibles públicamente. Dicho informe sugiere un enfoque de todo el Gobierno para la creación de capacidades de análisis y uso de datos dentro del servicio público australiano. Tal estrategia debería fomentar una "mentalidad de descubrimiento" para permitir un mejor análisis de problemas, desarrollo de soluciones de políticas, prestación de servicios mejorada y eficiencia del sector público. Para ello, elaboraron “Data Skills and Capability Framework”: un marco de habilidades y capacidades que deberían tener las personas que ocupan determinados roles en la gestión y en el análisis de datos. Para lograr que el personal del Gobierno logre alcanzar las habilidades y capacidades necesarias, el estado pone a disposición una serie de cursos que pueden ser destinados tanto para público principiante como para otro altamente calificado [28].

El Gobierno de **Bélgica**, buscando crear un centro de conocimiento para la educación, invirtió en proyectos de análisis de datos, poniéndose énfasis en el intercambio de datos estructurados [47]. En particular, se automatizó el flujo de datos de las escuelas y, trabajando sobre un Data Warehouse, generó reportes de interés sobre dichos datos [82]. Por otro lado, la Agencia de Empleo de dicho país provee servicios sobre empleos, permitiendo que los ciudadanos tengan una herramienta tecnológica rápida y fácil [47]. En concreto, utilizaron Big Data para proporcionarle a los ciudadanos una forma sencilla de recibir información sobre empleos que se ajustan a su perfil [83].

En **Estados Unidos**, más precisamente en la ciudad de Nueva York, en el año 2013, se creó la Oficina de Análisis de Datos del Alcalde (Mayor’s Office of Data and Analytics - MODA), a la cual se la define como: “el centro de inteligencia cívica de la Ciudad de Nueva York”. La creación de dicha oficina permite obtener y analizar datos de todas las agencias de la ciudad, logrando abordar de manera más efectiva los problemas de delincuencia, seguridad pública y calidad de vida. Utiliza herramientas de análisis para priorizar el riesgo de manera más estratégica, brindar servicios de manera más eficiente, hacer cumplir las leyes de manera más efectiva y aumentar la transparencia [21, 22].

El Programa de Análisis de Datos (City Data Analytics Programme) de la Ciudad de Londres, **Inglaterra**, es un centro virtual coordinado por el Equipo de Inteligencia de la Ciudad donde se forman, prueban, ejecutan y comparten proyectos

e ideas de ciencia de datos [23]. En la siguiente imagen se muestra una búsqueda sobre proyectos de la Ciudad de Londres utilizando el sitio web.

Figura 5 Gobernanza: Ejemplo Londres. Fuente: [23]

Otro caso en **Inglaterra**, es la Institución de Transporte de Londres la cual usó una solución tecnológica que le permitió hacer seguimiento y gestión de una flota de más de 8.500 vehículos para proveer información sobre su localización en tiempo real [47, 84].

En **Italia** se está desarrollando el Marco de Análisis de Datos y análisis (Data and Analytics Framework - DAF), que tiene el objetivo de mejorar y simplificar la interoperabilidad y el intercambio de datos entre las Administraciones públicas, promoviendo y mejorando la gestión y el uso de datos abiertos, optimizando las actividades de análisis y generación de conocimiento. Su objetivo es abrir el mundo de la Administración Pública a los beneficios que ofrecen las modernas plataformas de análisis y gestión de Big Data. Utilizando tecnología de Big Data, operarán a lo largo de tres líneas principales [24]:

- Mejorar significativamente el valor de los activos de información a través de la preparación y el uso de herramientas analíticas diseñadas para sintetizar el conocimiento para los tomadores de decisiones, y la difusión de información a los ciudadanos y las empresas.
- Optimizar el intercambio de datos y la implementación de datos abiertos, minimizando los costos de transacción para el acceso y uso de datos.
- Facilitar el análisis de datos y la gestión de datos por parte de equipos de científicos de datos, a fin de mejorar el conocimiento de los fenómenos descritos por los datos y desarrollar aplicaciones "inteligentes", así como tomar iniciativas para promover actividades de investigación científica sobre temas de aplicación de interés para el público.

El Departamento de Criminalística de la Aduana **Lituana**, desarrolló una solución de analítica avanzada que usa modelos de predicción sobre un gran conjunto de datos y perfiles relacionados con las aduanas, para estimar qué tipos de

actividades tienen la mayor probabilidad de corresponderse con operaciones ilegales o fraudulentas [47].

El Gobierno de **Nueva Zelanda** en su sitio data.govt.nz, establece lineamientos para la gestión y el uso de la información. Dentro de los lineamientos se pueden encontrar estándares de contenido de la información, políticas y guía para la gestión de datos, procesos automatizados de análisis de datos y principios para el uso seguro y efectivo de datos y análisis [25, 26, 27].

3.4. Estándares, mejores prácticas y marcos de referencia

A continuación, se desarrollan algunos documentos de interés (estándares, mejores prácticas y marcos de referencia) que fueron insumos clave para el desarrollo del presente marco de trabajo:

- Data Management Body of Knowledge (DAMA-DMBOKII).
- Open Web Application Security Project (OWASP).
- A data-driven public sector: Enabling the strategic use of data for productive (OCDE).
- Marco de Ciberseguridad (AGESIC).

El "**Data Management Body of Knowledge**" (DAMA-DMBOKII) [10], es una de las principales referencias identificadas a nivel mundial relativas a la gestión y el manejo de dato. Propone una serie de principios para guiar la gestión de los datos y pautas generales para su implementación. Aborda distintos ejes relevantes para la práctica de gestión de datos, incluyéndose un capítulo específico para seguridad de datos. Este documento es tenido en cuenta como uno de los insumos principales de referencia para el desarrollo del marco de trabajo, en particular para la dimensión Ciberseguridad.

OWASP [91] es una organización sin fines de lucro cuyo objetivo es proporcionar pautas para el diseño, desarrollo, adquisición, uso y mantenimiento seguro de software. Por lo que las recomendaciones realizadas en 2019 serán referenciadas en la sección de buenas prácticas de ciberseguridad.

La **OCDE** elaboró un documento titulado "A data-driven public sector: Enabling the strategic use of data for productive", donde establece un marco que contiene las prácticas relacionadas con el análisis de los datos existentes, identificando las implicaciones políticas de implementar un enfoque de Gobierno basado en datos y los esfuerzos necesarios para maximizar sus beneficios y proporcionar una referencia a las prácticas emergentes en los países miembros de la OCDE [16].

Uno de los objetivos del Plan de Gobierno Digital (Uruguay) dentro del área de acción Gobierno Confiable es el **Marco de Ciberseguridad** [8]. Éste fue publicado por primera vez en agosto de 2016 por AGESIC, y su principal objetivo es brindar lineamientos y buenas prácticas para un abordaje integral de la ciberseguridad en los organismos del Gobierno. El abordaje del mismo está orientado a reducir el riesgo vinculado a las amenazas cibernéticas que puedan comprometer la seguridad de la información. El Marco de Ciberseguridad integra referencias de distintos

estándares internacionales, brindando un enfoque unificado que contempla además la normativa nacional relativa a seguridad de la información y se encuentra dentro de un conjunto de iniciativas llevadas adelante por AGESIC para mejorar la estrategia a nivel nacional en materia de ciberseguridad. Concluyendo, el Marco de Ciberseguridad tiene una relevancia fundamental para el marco de trabajo ya que es una base de buenas prácticas de ciberseguridad[2].

[2] En este marco se utilizan los términos de ciberseguridad y seguridad de la información en forma indistinta.

4. Modelo Conceptual

El objetivo de esta sección es presentar y desarrollar conceptualmente las dimensiones del marco de trabajo y de cada uno de sus componentes.

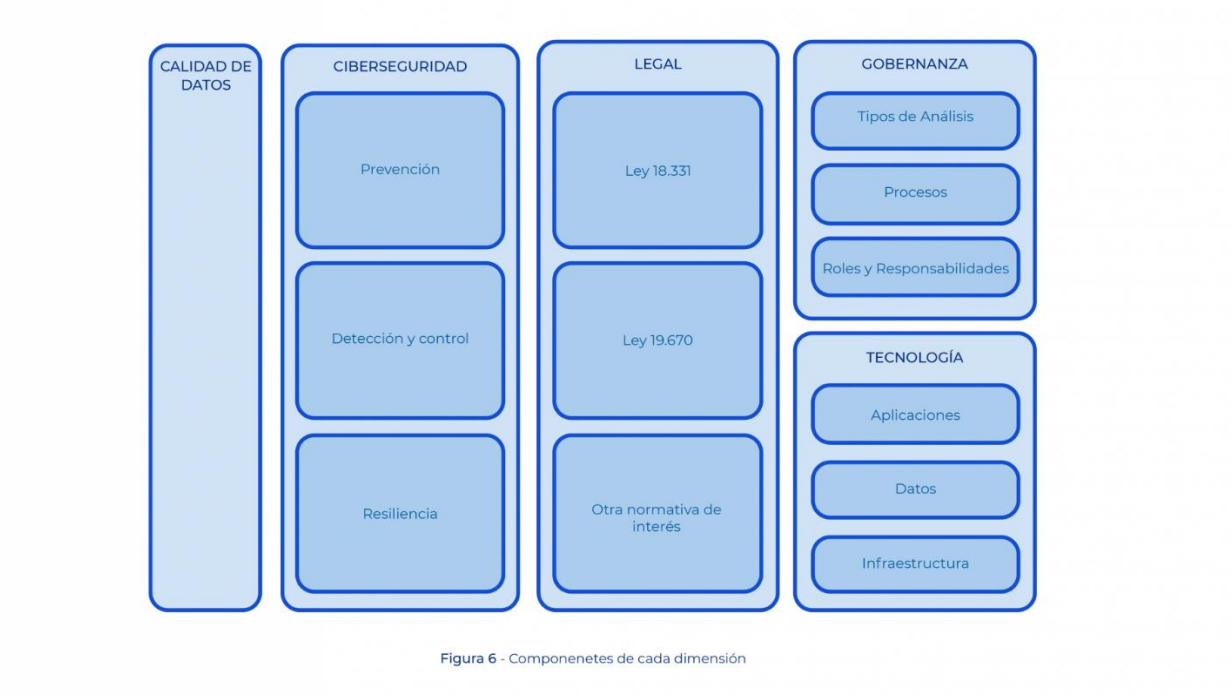

Figura 6 Componentes de cada dimensión

4.1. Gobernanza

En la presente sección se desarrollan los componentes de la dimensión de Gobernanza:

- Tipos de Análisis y Arquitecturas: se presentan dos clasificaciones para el análisis de datos, junto con un conjunto de arquitecturas definidas para soportar tecnológicamente los mismos.
- **Proceso:** se presenta un proceso de referencia para el análisis de datos. Para cada una de las etapas que lo componen se define su objetivo, las entradas o elementos que recibe, las actividades que se realizan y las salidas o resultados que producen.
- Roles y responsabilidades: se definen roles y responsabilidades para enfrentar el análisis de datos. Algunos de estos roles son nuevos en los organismos, mientras que otros son existentes, pero se agregan nuevas responsabilidades referentes a la temática.

4.1.1. Tipos de Análisis, Arquitecturas y Herramientas

En las siguientes subsecciones, se presentan dos tipos de clasificaciones para los tipos de análisis y las arquitecturas y herramientas que los pueden soportar.

4.1.1.1. Tipos de análisis

En la presente subsección se describen dos clasificaciones diferentes de tipos de análisis de datos.

En la primera clasificación que analizaremos se pueden distinguir los diferentes tipos de análisis de datos en función del **nivel jerárquico del destinatario del análisis**:

- **Análisis estratégico:** este tipo de análisis está destinado a la dirección y alta gerencia, parte de la recolección de datos relativos al estado actual y a la evolución de las factores externos e internos claves que afectan al organismo, es decir, del entorno, de los recursos y capacidades de la misma. Su objetivo es que el mismo monitoree los indicadores de desempeño que fueron definidos como claves para el cumplimiento de su estrategia.
- **Análisis de gestión:** aplica a toda la estructura del organismo y de sus componentes para evaluar el grado de eficiencia y eficacia con el cual se están cumpliendo la planificación, organización, la coordinación, la ejecución y control de los objetivos trazados por la entidad, para poder corregir las deficiencias que pudieran existir, optimizando la productividad de la misma. Los destinatarios son la alta gerencia, la gerencia y los jefes.
- **Análisis operativo:** en un análisis operativo los jefes y funcionarios realizan un seguimiento y diagnóstico permanente de las operaciones, detectando de forma temprana oportunidades de mejora. Aplica a todas las áreas y niveles del organismo y su aplicación es clave para conseguir el involucramiento del personal en el cumplimiento de resultados, la detección y consecución de mejoras.

A continuación, se presenta para cada tipo de análisis una sugerencia respecto a su frecuencia y destinatario dentro de la organización.

PROCESO ESTÁNDAR		
Tipo de análisis	Frecuencia sugerida	Destinatario
Análisis estratégico	Trimestral o Semestral	Dirección y alta gerencia
Análisis de gestión	Mensual	Alta gerencia, gerencia y jefes
Análisis operativo	Mensual, semanal o diario	Jefes y funcionarios

Tabla 1 Comparación de tipos de análisis: frecuencia y destinatario

En la segunda clasificación se pueden distinguir los diferentes tipos de análisis de datos en función de **cuál sea el objetivo del mismo**:

- **Descriptivo**: debe responder a la pregunta: ¿qué pasó?, donde se analizan datos completos o una muestra de datos numéricos resumidos.
- **Diagnóstico**: un proceso de análisis diagnóstico debe responder a la pregunta: ¿por qué pasó?, a partir de la información encontrada en el

análisis estadístico. Es útil para identificar patrones de comportamiento de los datos. Si se detecta un nuevo problema, se puede realizar este tipo de análisis para encontrar patrones similares en pasadas y así poder tener la posibilidad de realizar acciones similares para los nuevos problemas.

- **Predictivo**: un proceso de análisis predictivo debe responder a la pregunta: ¿qué va a pasar o qué debió haber pasado?, mediante el uso de datos históricos. Este análisis hace predicciones sobre resultados basados en datos actuales o pasados. El pronóstico es solo una estimación. Su precisión se basa en la cantidad de información detallada que tiene y cuánto se trabaja en ella.
- **Prescriptivo**: un proceso de análisis prescriptivo debe responder a la pregunta: ¿qué se debería hacer?, combina la información de todos los análisis anteriores para determinar qué acción tomar en un problema o decisión actual.

4.1.1.2. Arquitecturas

A continuación, se describe las principales arquitecturas identificadas para dar respuesta a los distintos tipos de análisis desarrollados anteriormente.

- **Análisis de información básico:** proceso en el que se manejan datos de pocas fuentes y se los presenta en grillas, gráficos o tablas dinámicas.
- **Business Intelligence:** es un término general que incluye las aplicaciones, la infraestructura, las herramientas y mejores prácticas que permiten el acceso y análisis de la información para mejorar y optimizar las decisiones y el rendimiento [87].
- **Big Data:** son activos de información de gran volumen, velocidad y / o variedad que exigen formas rentables e innovadoras de procesamiento de información que permitan una mejor comprensión, toma de decisiones y automatización de procesos [87].
- **Streaming (Big Data en tiempo real):** son sistemas de software que realizan cálculos en tiempo real o casi en tiempo real sobre datos de eventos "en movimiento". La entrada es una o más transmisiones de eventos que contienen datos sobre por ejemplo pedidos de clientes, reclamos de seguros, depósitos / retiros bancarios, tweets, publicaciones en Facebook, correos electrónicos, mercados financieros u otros, o datos de sensores de activos físicos como vehículos, dispositivos móviles o máquinas. Las plataformas procesan los datos de entrada a medida que llegan (por lo tanto, "en movimiento"), antes de almacenarlos opcionalmente en algún almacén persistente. Conservan un conjunto de datos de flujo de trabajo relativamente pequeño en la memoria, el tiempo suficiente para realizar cálculos en un conjunto de datos recientes durante un período de tiempo [87].
- **Analítica Avanzada:** es el examen autónomo o semiautónomo de datos o de contenido utilizando técnicas y herramientas sofisticadas, para alcanzar conocimientos más profundos, hacer predicciones o generar recomendaciones. Las técnicas analíticas avanzadas incluyen aquellas

como minería de datos (data mining), machine learning, coincidencia de patrones, forecasting visualización, análisis semántico, análisis de sentimientos, análisis de redes y de clústeres, estadísticas multivariadas, análisis de gráficos, simulación, procesamiento de eventos complejos, redes neuronales [87].

En la siguiente tabla se pueden apreciar ejemplos o escenarios genéricos de análisis de datos. En cada uno de los ejemplos se especifica para las dos clasificaciones de tipos de análisis detalladas anteriormente, las arquitecturas o herramientas que deberían utilizarse.

Ejemplos de aplicación	Tipos de análisis 1	Tipos de análisis 2	Arquitecturas
Indicadores de Gastos del ejercicio anterior con semáforos	Estratégico	Descriptivo	Análisis de información básico o BI Tradicional
Reporte de gastos para rendición de cuentas	Gestión	Descriptivo	BI tradicional
Evaluación de stock de uniformes	Operativo	Descriptivo	Análisis de información básico o BI tradicional
Análisis de comportamiento de proveedores con retrasos	Gestión	Diagnóstico	BI tradicional
Reporte semanal de estados de expedientes	Operativo	Descriptivo	Análisis de información básico o BI tradicional
Estadísticas de Trámites por Canal de entrada	Gestión	Descriptivo	Análisis de información básico o BI tradicional
Proyección de cantidad de trámites (nuevos más actuales) en línea por horario en base a históricos	Operativo	Predictivo	Big Data
Proyección de delincuencia en zonas por horario en base a históricos y datos reales (cámaras, redes sociales)	Estratégico	Predictivo	Streaming
Análisis multas y comportamiento de pago y envío de convenio de pago sugerido	Operativo	Prescriptivo	Analítica avanzada

Tabla 2 Ejemplos de aplicación

4.1.2. Procesos

El objetivo de esta sección es presentar un proceso de análisis de datos de referencia, el cual será utilizado para el desarrollo de las buenas prácticas y recomendaciones. Dicho proceso se compone de etapas y si bien se lo presenta con cierto orden lógico, pueden existir casos donde para un determinado tipo de análisis no se pase por todas las etapas o se necesite realizar varias veces una misma etapa

del proceso (ej: cuando se necesita un dato específico de una única fuente de datos no es necesario realizar la etapa de modelado).

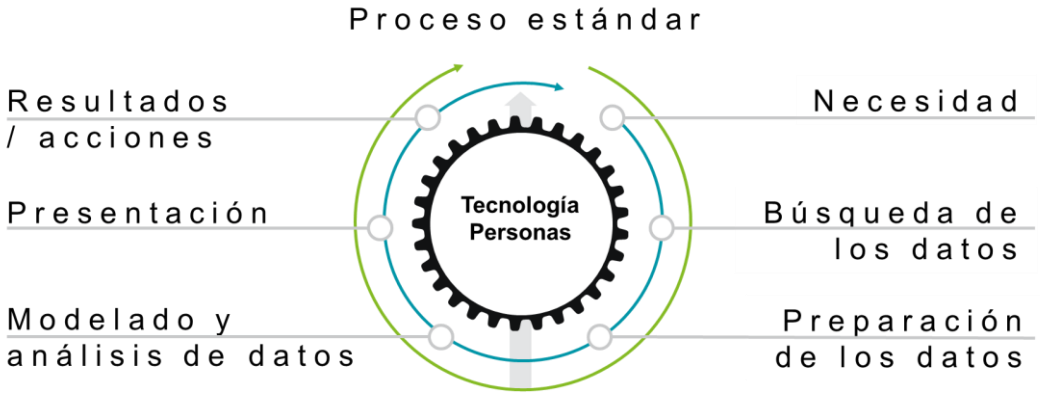

Figura 7 Proceso de referencia

El diagrama presenta las etapas que componen al proceso de análisis de datos. Para cada una de ellas se describe lo siguiente:

- **Objetivo:** es el planteo de una meta o un propósito a alcanzar.
- **Entradas:** son todos aquellos elementos tangibles e intangibles (información, etc.) necesarios para llevar adelante las actividades a realizar.
- **Actividad:** es una unidad de trabajo a realizar. Ésta se puede definir como "una acción sobre un objeto", es decir una actividad se denomina siempre con un verbo (acción) y un sustantivo (objeto).
- **Salidas:** son todos aquellos elementos tangibles e intangibles que cada actividad es responsable de producir.

4.1.2.1. Necesidad

Objetivo

Identificar las necesidades de información del organismo mediante la definición de los objetivos del análisis, los requisitos de la solución, el público objetivo, los criterios de éxito y los requerimientos de los datos a fin de guiar las actividades de recolección de información y posterior análisis.

Entradas

Se inicia con la necesidad de dar respuesta a una pregunta vinculada a la operativa, experimento o estudio que se esté realizando. A modo de ejemplo: ¿Cuántos trámites de pago de sanciones del Ministerio de Transporte y Obras Públicas se realizaron vía web en el 2019?

Actividades

Se comienza por definir claramente la necesidad de información para poder investigar si en la actualidad ya se puede cubrir dicha necesidad con un análisis pre-

existente. En caso de no existir, el equipo deberá revisar si no existe algún antecedente en el que se pueda basar y que mediante la realización de algunos ajustes pueda satisfacer dicha necesidad.

Otros aspectos a definir son los objetivos de análisis, el público objetivo, los criterios de éxito y los requisitos de datos en cuanto a contenidos, formatos, representaciones, etc.

Esta información se va a utilizar para identificar los datos necesarios, el nivel de detalle y la forma de presentación requeridas para el diseño y desarrollo del análisis.

También se utilizarán para identificar los métodos a utilizar, las fuentes de donde se van a obtener los datos (internas y/o externas), evaluando las ventajas y desventajas de cada una y para evaluar los riesgos potenciales que puedan existir. Por ejemplo: tratamiento de los datos personales o sensibles, o no tener acceso a las bases de datos de privados u otros organismos.

El equipo técnico es responsable de cerciorarse si el organismo cuenta con las herramientas necesarias para realizar el análisis que el interesado requiere, en caso de no contar con estas, se podrá sugerir algún camino alternativo o adquirir las herramientas faltantes. En la sección 4.1.1.2. Tipos de Análisis se muestran dos tipos de clasificaciones y algunos ejemplos de aplicación que pueden servir de guía para elegir las herramientas a utilizar.

Un aspecto a tener en cuenta es que luego de avanzar en el modelado y en la presentación de los datos, es posible que surjan nuevos subprocesos con sus propios tipos de análisis y herramientas a utilizar que combinados con el proceso de origen logren satisfacer la necesidad inicial.

Salidas

Al finalizar las actividades, se tienen definidos la necesidad de información, los objetivos del análisis, los requisitos de la solución, el público objetivo, los criterios de éxito y los requerimientos de los datos. Además, se identifican los tipos de análisis a realizar y las herramientas a utilizar.

4.1.2.2. Búsqueda de los datos

Objetivo

Definir las fuentes de datos a utilizar mediante la selección de fuentes confiables para asegurar la validez de los datos obtenidos.

Realizar la recopilación y consolidación (cuando corresponda) de los datos que dan respuesta a las necesidades de información definidas en la etapa anterior, a través de la consulta de las fuentes seleccionadas.

Entradas

Documento con las definiciones acordadas en la etapa anterior.

Actividades

Al comenzar con las actividades de búsqueda de datos se deberá evaluar la viabilidad de la realización del análisis, quien se encargue de realizar la búsqueda deberá tener en cuenta aspectos tales como:

1. ¿Existe personal con disponibilidad para realizar la búsqueda de los datos?
2. ¿Los datos que necesito están en las fuentes de datos disponibles?
3. ¿Cuento con los accesos a las fuentes de datos necesarias?
4. ¿Se cumple con todos los principios de la ética de datos? (OCDE [16])
5. ¿Los datos a utilizar violan algún aspecto legal o confidencial? (Por ejemplo: Ley de Datos Personales)
6. ¿Existen restricciones al uso de los datos que necesito?

Si no existen limitantes a la realización del análisis, es necesario localizar y seleccionar las fuentes disponibles de datos, considerando fuentes internas y/o externas, evaluando la calidad y qué datos son relevantes.

Luego se recolectan los datos disponibles (estructurados, no estructurados y semiestructurados) que sean relevantes para el dominio del problema, tratando de entender cómo permiten responder y atender a los objetivos definidos y analizarlos, teniendo especial consideración en aspectos como la calidad y la posibilidad de integración. En el mismo sentido, cuando queremos integrar nuevos datos debemos evaluar la calidad de los mismos, lo que implica también tener en cuenta cómo se obtuvieron las muestras y su validez, y definir cómo se puede usar en el entrenamiento y la validación de modelos.

De los datos recopilados se tienen que conocer aspectos como: origen, formato, qué representan, vinculación con otros datos, frecuencia de actualización, granularidad, consistencia, confiabilidad, perfilamiento, entre otros.

Una vez recopilados los datos, se debe identificar si existen brechas entre los datos que se tienen y los que se necesitan, esto se debe hacer en forma iterativa, definiendo si se realizan inversiones adicionales para obtener nuevos datos, lo que puede aplazarse hasta tener más claro si existen lagunas en los datos recopilados hasta la fecha. También se debe gestionar el filtrado de datos para evitar sesgos.

Los datos a utilizar podrán ser accedidos directamente a las fuentes originales desde las herramientas de análisis o se puede realizar una copia de los mismos en una base intermedia. Se sugiere que en la ingesta de datos se capturen también metadatos sobre el origen, tamaño y el contenido, entre otros.

Salidas

Al finalizar las actividades previstas en el proceso de Búsqueda de Datos, se obtienen como resultados, las fuentes de datos definidas y el conjunto de datos recolectados. Esto puede implicar, la consolidación de la información en bases intermedias cuando así sea requerido.

4.1.2.3. Preparación de los datos

Objetivo

Procesar, organizar y limpiar los datos para su posterior análisis mediante la aplicación de las técnicas de transformación, normalización y limpieza de datos.

Entradas

Esta etapa inicia con los datos recopilados en la fase anterior, los cuales se encuentran en sus fuentes originales o consolidados en una base intermedia, según corresponda.

Actividades

Esta etapa suele ser la más larga del proceso de análisis debido al trabajo que conlleva la preparación del conjunto de datos que se va a emplear en el resto del proceso.

Para comenzar a preparar los datos debemos comprender lo que hay en ellos, encontrar enlaces entre los datos de las diversas fuentes y alinear los datos comunes para su uso. Entonces, se pueden aplicar técnicas de estadística descriptiva y data science (por ejemplo: análisis de clúster y matrices de correlación, PCA, autoencoders) para explorar los datos y entender cómo estos pueden analizarse, si existen vinculaciones entre varias fuentes de datos, cómo podrían ser los resultados de los análisis y los modelos, encontrar patrones o datos atípicos, etc.

Además, se pueden aplicar técnicas de ingeniería de características (feature engineering) para crear nuevas variables explicativas mediante el conocimiento del tema y las variables disponibles. Dentro de las técnicas de featuring engineering se encuentran: el escalado de datos, binning, binarización, transformación de Box Cox, one-hot encoding o embedding. Para profundizar en algunas de estas técnicas ver [111].

Por la forma que fueron ingresados y almacenados lo datos pueden contener registros duplicados, espacios en blanco o errores, entre otros. Por eso es necesario realizar una limpieza de los mismos. Este es uno de los pasos más importantes del análisis de información porque garantiza una mayor calidad de los datos, que va a redundar en un análisis de mayor precisión. Dentro de las actividades de limpieza de datos más comunes se incluyen: la coincidencia de registros, la identificación de la inexactitud de los datos, eliminación de duplicados y la segmentación de columnas.

También es necesario normalizar y transformar estos datos, que pueden provenir de diversas fuentes de datos, aplicando una serie de reglas de negocio o funciones sobre los mismos, para así llevarlos a un formato único y con la calidad requerida. Por ejemplo, si tengo datos similares de distintos orígenes, pueden requerir ser normalizados para su análisis, sería el caso de datos de direcciones de ciudadanos que en una fuente estén en el formato Nombre de Calle/Número y en otra estén como Número de Calle/Nombre, o que se utilicen abreviaciones como Blvr y en otra fuente diga el nombre completo: Bulevar. Agregar metadatos y en caso que se manejen datos sensibles se deberán aplicar técnicas de sanitización de los mismos.

Finalmente, los datos limpios y prontos para utilizar serán ingestados en un almacén de datos para realizar sobre ellos un análisis de datos.

Salidas

Finalizada la etapa de preparación de los datos, se obtiene el conjunto de datos seleccionados, en un formato adecuado y con la calidad necesaria para llevar adelante el análisis requerido. Además, dichos datos se encuentran ingestados en un almacén de datos para su análisis.

4.1.2.4. Modelado y análisis de datos

Objetivo

Este proceso tiene los siguientes objetivos:

Definir y construir el modelo de análisis de acuerdo al enfoque analítico definido para el estudio de los datos obtenidos.

Analizar los datos mediante la aplicación de las técnicas de análisis definidas para obtener las conclusiones vinculadas a los objetivos de análisis planteados en la primera etapa del proceso.

Entradas

Esta etapa toma como insumo el conjunto de los datos preparados y disponibles en un almacén de datos.

Actividades

El análisis de los datos implica la búsqueda de respuestas a las necesidades planteadas por los interesados, lo cual puede derivar en nuevos hallazgos, o similitudes y diferencias con hallazgos anteriores.

Existen varias técnicas de análisis de datos disponibles para comprender, interpretar y sacar conclusiones basadas en los objetivos definidos. Por ejemplo, la exploración de modelos multidimensionales, el manejo de algoritmos de procesamiento complejos a partir de datos estructurados y no estructurados, el diseño y visualización de datos.

En esta etapa se deben desarrollar las hipótesis y los métodos de testeo, aplicando modelos para encontrar tendencias y correlaciones. Los modelos se realizarán de acuerdo al enfoque analítico elegido y dependen fuertemente de la calidad de los datos y de su robustez en sí.

Un modelo es un conjunto de herramientas conceptuales para describir datos, sus relaciones, su significado y sus restricciones de consistencia.

Si el requerimiento implica la búsqueda de un dato específico y que no requiere el cruce de más de una fuente de datos para su identificación podría no ser necesario ningún tipo de modelado sino solamente la identificación del dato en la fuente disponible y su presentación de forma clara al interesado (ejemplo un semáforo con la cantidad de trámites realizados en el último trimestre).

En el caso que la necesidad sea del tipo diagnóstico se requerirá la construcción de un modelo multidimensional, lo cual implicará definir las

dimensiones, medidas, jerarquías a ser utilizadas para aplicar distintos niveles de análisis (drill down) o para realizar distintos filtros o cortes de los datos.

Si en cambio lo que se requiere es buscar patrones de comportamiento deberán aplicarse técnicas de Machine Learning. Por ejemplo, se requiere la eliminación de outliers, normalización de datos y completar datos faltantes:

- **Detección de outliers:** los outliers son casos que se destacan de los normales. Es importante tomar las correctas acciones para dichos casos, porque dejar la decisión de eliminar o no cada caso tendrá consecuencias importantes sobre los resultados del análisis.
- **Normalización de los datos:** llevar los datos a la misma escala normalmente genera mejores resultados. Por ejemplo, llevar a la escala de 0 a 1, o transformar los valores para representarlos como percentiles en lugar de como valores absolutos.
- **Datos faltantes:** es necesario tomar decisión para completar o eliminar la información faltante. Existen diferentes técnicas para resolver este problema: omisión de fila/columna para la cual falte información, imputación de datos (completar los datos que faltan con posibles valores o valores por defecto).

Básicamente, se puede usar cualquier método, siempre y cuando ayude al analista a examinar la información que se ha recopilado, con el objetivo de darle algún sentido, buscar patrones y relaciones, de forma de ayudar a responder a la necesidad de información original.

Se deben probar diferentes tipos combinaciones de análisis, modelos, algoritmos, variables, parámetros e hiper parámetros sobre el mismo conjunto de datos para responder al caso de uso definido en la etapa inicial (por ejemplo: de tipo matemático, estadístico, econométrico, de machine learning) evaluando cuál tiene el mejor desempeño en la muestra de test.

Con el modelo desarrollado, se debe evaluar en qué medida éste responde la problemática planteada y cuantificar su efectividad antes de comenzar la fase de implementación. Para ello, se debe realizar en primer lugar una evaluación desde el punto de vista funcional para garantizar que aborda el problema de negocio de manera adecuada y completa, y si la solución obtenida no es satisfactoria se debe ajustar el modelado.

En segundo lugar, se tiene que realizar la evaluación cuantitativa del modelo. Existen diferentes medidas para evaluar la efectividad de un modelo dependiendo del tipo de algoritmo implementado y de los objetivos del análisis, y de este proceso surgirá un modelo seleccionado a ser implementado y además uno o más modelos alternativos que servirán para monitorear la efectividad del mismo.

Cuando se utilizan modelos predictivos se suelen utilizar muestras de entrenamiento y de test, y en algunos casos de validación, por lo que se modela con una parte de los datos, se evalúa con los datos de testeo y se realizan ajustes si es necesario.

A medida que los datos son manipulados, es posible que se deba realizar una limpieza o recopilación de datos adicional, por lo que estas actividades se podrían realizar de manera iterativa. Si el problema es muy complejo o novedoso, esta etapa puede suponer el desarrollo o la adaptación de nuevos algoritmos.

Salidas

Como resultado de las actividades vinculadas a esta etapa, el organismo contará con un modelo de análisis documentado.

Además, dispondrá de la información resultante del análisis de los datos seleccionados. Dicha información se encontrará debidamente documentada en el formato de informe que sea acordado en función de las necesidades detectadas en la primera etapa del proceso.

4.1.2.5. Presentación

Objetivo

Comunicar la información resultante del análisis de datos en el formato requerido por usuarios finales mediante la utilización de herramientas de visualización para facilitar la toma de decisiones de los interesados.

Entradas

La información resultante del análisis de datos y los requisitos de representación definidos en la primera etapa del proceso van a ser los insumos para realizar las visualizaciones adecuadas para el público objetivo.

Actividades

La presentación de los resultados del análisis puede variar dependiendo del tipo de interesado que solicite la información. Los encargados de presentar los resultados deben considerar el tipo de destinatario, el canal de información y su periodicidad. Por ejemplo, el interesado puede requerir que le entreguen no solo una “foto” o “vista estática” del resultado sino una “vista interactiva” a partir de la cual pueda navegar para analizar otras dimensiones de la información.

Por ello se deberá definir el tipo de visualización más adecuada para que los interesados puedan entender los resultados que se les están mostrando, teniendo en cuenta que cada una de las visualizaciones generadas debe responder a una pregunta surgida en la primera etapa del proceso (Necesidad) o generar un “insight”.

Las herramientas de visualización ofrecen diferentes funcionalidades de entrega y diseño. Es posible utilizar las capacidades de las herramientas para analizar los datos o simplemente como una herramienta para organizar y presentar la información resultante del análisis. Por ejemplo: se pueden utilizar tablas o gráficos para ayudar a comunicar los mensajes clave contenidos en los datos. Las tablas, por ejemplo, son útiles para un usuario que puede buscar números específicos, mientras que los gráficos (por ejemplo, gráficos de barras o gráficos de líneas) pueden ayudar a explicar los mensajes cuantitativos contenidos en los datos. También se podrán utilizar semáforos u otras señalizaciones para monitorear

indicadores de forma clara y amigable. Incluso los datos muy complicados pueden ser simplificados y entendidos por la mayoría de las personas cuando se representan visualmente.

Una vez definidas las visualizaciones a realizar, los encargados de la presentación de los resultados deben realizarse preguntas del estilo: ¿los resultados que se muestran tienen sentido?, ¿se puede contar una historia con las visualizaciones elegidas? En caso de ser afirmativas, los datos pueden ser mostrados a los interesados.

Tener en cuenta que a medida que se muestran los resultados, los interesados pueden realizar comentarios que resulten en un análisis adicional, por lo que el ciclo de análisis puede volverse iterativo.

Salidas

Como resultado de la Presentación, la información del análisis se encuentra comunicada y sociabilizada entre las partes involucradas, encontrándose disponible para su consulta y utilización para la toma de decisiones y ejecución de acciones.

4.1.2.6. Resultados y acciones

Objetivo

Evaluar los resultados obtenidos y tomar decisiones utilizando el conocimiento generado.

Entradas

Esta etapa toma como insumo, el informe elaborado en el marco del análisis de datos, las visualizaciones realizadas, así como todas las definiciones acordadas durante la ejecución de las etapas del proceso.

Actividades

Una vez que los datos han sido presentados, pueden ser interpretados. Se podrá verificar si lo que se ha recopilado es útil para satisfacer las necesidades de información originales.

¿El análisis realizado le ayuda a contestar alguna de las preguntas surgidas inicialmente?, ¿alguno de los resultados obtenidos es limitante o no es concluyente?, ¿han surgido nuevas necesidades que antes no se identificaron? Si todas las preguntas se tratan con los datos disponibles actualmente, entonces su investigación puede considerarse completa y los datos finales. Ahora se puede utilizar para el propósito para el que se realizó el análisis, ayudar a los interesados a tomar buenas decisiones.

Los interesados van a comunicar si el análisis realizado satisface sus expectativas y los ayuda a tomar mejores decisiones. Si el resultado es positivo, resta definir si el análisis realizado va a ser utilizado por única vez o va a ser de uso habitual convirtiéndose en un modelo estándar de análisis de información.

En caso que se convierta en un modelo estándar, se sugiere implementar el modelo en el entorno de producción o en un entorno de pruebas comparable, en muchos casos de forma limitada hasta que su rendimiento se haya evaluado completamente. Además, se tienen que designar a los responsables de revisarlo, mantenerlo y ajustarlo de ser necesario, ya que el modelo va a ser utilizado por otras personas y con nuevos datos, pudiendo surgir nuevas variables y necesidades para el mismo modelo.

Salidas

Los resultados de estas actividades son las acciones, definiciones y decisiones que se toman a partir del análisis realizado. Por ejemplo: elaboración de políticas, procedimientos, presupuestos o decisiones como ingreso de nuevo personal o recortes de gastos.

4.1.3. Roles y Responsabilidades

En la presente sección se definen una serie de roles que se identificaron como participantes en un proceso de análisis, algunos de los mismos estarán presentes en cualquier tipo de análisis de datos (Ingeniero de datos), otros se harán presentes en procesos de análisis más avanzados (Custodio de datos).

Los roles a definir se dividen en dos grupos:

- Nuevos Roles inherentes al análisis de datos.
- Roles existentes a los que se le agregan actividades relacionadas al análisis de datos.

En la primera tabla, que se muestra a continuación, se detallan los nuevos roles con sus respectivas responsabilidades. Su estructura es la siguiente: la primera columna, “Roles”, incluye la denominación del mismo, las siguientes tres columnas detallan la dedicación al análisis de datos sugerida para el rol, el propósito/misión que debe cumplir y las principales actividades que va a desarrollar. Continúa con las experiencias y las competencias básicas y deseables que deberían tener las personas que cumplan el rol y por último su formación (también básica y deseable).

Por su parte, la segunda tabla detalla los roles y responsabilidades que deben ser agregados a cargos ya existentes. La estructura de esta tabla es la misma que la anterior.

Nuevos Roles y responsabilidades							
Roles	Dedicación al análisis de datos	Propósito/Misión	Principales Actividades	Experiencia y competencias		Formación	
				Básicas	Deseables	Básicas	Deseables
Científico de Datos	Full time	Extrae datos internos o externos y los analiza en función de la demanda o por iniciativa propia para generar valor al organismo	• Recibe requerimientos por parte de los interesados y del Analista de información. • Sugiere herramientas, modelos de análisis y datos a considerar • Aplica diversas técnicas de análisis para analizar datos internos o externos	• Programación (Python, R, etc.) • Funciones del organismo • Herramientas estadísticas y matemáticas • Herramientas de análisis • Estadística y Machine Learning • Manejo de las plataformas de datos y de análisis de datos de AGESIC • Modelos predictivos • Catálogo de datos abiertos	• Minería de Datos • Big Data	• Profesional de Ciencias Económicas o Administración, Ingeniería o similares • Estudios de estadística / Matemática	• Especialización o estudios en ciencia de datos
Arquitecto Análisis de Datos	Full time	Ayuda al organismo a entender sus objetivos estratégicos con respecto a la gestión de datos.	• Diseña la arquitectura de análisis de datos, integrando las fuentes necesarias	• RDBMs y bases de datos no relacionales • Uso de herramientas de análisis de datos	• Manejo de distintos SO • Habilidades de liderazgo • Persuasión	• Analista en Computación	• Ingeniero en Computación • Especialización o estudios en ciencia de datos

		Conoce e identifica necesidades de incorporar herramientas de análisis de datos optimizando costos y esfuerzos en el uso de los recursos	• Evalúa las herramientas de análisis y sistemas fuentes existentes •	• Arquitectura de Software • Funciones del organismo • Manejo de la plataforma de datos y análisis de datos de AGESIC • Catálogo de datos abiertos • Plataforma de interoperabilidad • Arquitectura de datos • Arquitectura del organismo y de la Arquitectura Integrada de gobierno • Detallista • Innovación / Iniciativa • Soluciones de data warehousing • Manejo de activos de gobierno digital			
Especialista / Ingeniero de datos	Full time	Diseñar, desarrollar, testear y mantener sistemas de análisis de datos	• Programa ETL • Investiga sobre herramientas de ETL • Diseña los sistemas de	• Herramientas de ETL • Lenguajes de Scripting • Programación • RDBMs	• Optimización en consultas a bases de datos (relacionales y no relacionales)	• Programador	• Analista en Computación • Especialización o estudios en ciencia de datos

			información junto con el Arquitecto de Análisis de Datos • Gestiona los accesos a las fuentes de datos para análisis • Gestión de la Calidad de datos	• Bases de datos no relacionales • Manejo de las plataformas de datos y análisis de datos de AGESIC • Catálogo de datos abiertos • Plataforma de interoperabilidad • Soluciones de data warehousing • Framework de Calidad de Datos	• Big data		
Analista de información	Full time	Crea dashboards, tableros, reportes y modelos de análisis de acuerdo a los requerimientos de los Tomadores de Decisión	• Trabaja con los interesados para tener conocimiento de lo requerido • Conoce las distintas funciones del organismo pública para entender qué se podría requerir • Diseña reportes, dashboards y visualizaciones interactivas • Explora y encuentra	• Funciones del organismo • Programación básica • Herramientas de Analytics • Consultas SQL • Manejo de la plataforma de datos y análisis de datos de AGESIC • Catálogo de datos abiertos • Orientación al cliente • Innovación / Iniciativa • Calidad del trabajo	• Programación avanzada • Manejo de bases de datos relaciones y no relacionales • Conocimientos estadísticos y de machine learning	• Profesional de Ciencias Económicas o Administración, Ingeniería o similares	• Especialización o estudios en ciencia de datos

			interrogantes y las resuelve mediante el análisis de los datos • Gestiona los impactos en los análisis actuales de posibles cambios del origen de datos o de los requerimientos • Nexo entre el rol Especialista / Ingeniero de Datos y los interesados • Gestiona la Calidad de datos	• Facilidad de comunicación • Habilidad y pensamiento analítico			
Custodio del análisis de datos	Part time	Responsable de la gestión de los datos de las entidades. Por lo que maneja los procesos de datos, políticas de datos y realiza guías sobre el uso de los datos	• Garantizar la disponibilidad y seguridad de los datos • Optimizar el almacenamiento y visualización de los datos • Diseñar, gestionar y controlar los procesos que manipulan datos	• Manejo de bases de datos relaciones y no relacionales • Funciones del organismo • Plataformas de datos y análisis de datos de AGESIC • Catálogo de datos abiertos • Arquitectura de datos	• Modelos de datos físicos y lógicos • Programación	• No Aplica	• Especialización o estudios en análisis y ciencia de datos • Especialista en gestión de proyectos

			• Establecer métricas y requisitos para la calidad de datos. Coordina con otros roles para su cumplimiento en todas las etapas del ciclo de vida de los datos	• Manejo de activos de gobierno digital • Calidad de datos • Framework de Calidad de Datos • Habilidad y pensamiento analítico			

Tabla 3: Nuevos roles y responsabilidades

Roles y responsabilidades que se agregan a cargos existentes							
Roles	Dedicación al análisis de datos	Propósito/Misión	Principales Actividades	Experiencia y competencias		Formación	
				Básicas	Deseables	Básicas	Deseables
Responsable de Seguridad de la Información	Part time	Gestiona de forma integrada las medidas de ciberseguridad con perspectivas de riesgo en la entidad	• Realiza tareas de prevención, detección y control de riesgos • Gestiona proactiva y reactivamente los riesgos • Encargado de diseñar políticas de acceso a los datos para salvaguardar la	• Comunicar conceptos de seguridad a roles técnicos y no técnicos • Auditoria y gestión de riesgos • Capacidades para gestión de incidentes • Plataformas de datos y análisis de datos de AGESIC • Redes y hardware	• Comunicación y persuasión • Integración con el negocio para crear una cultura donde los riesgos de seguridad sean asumidos por toda la organización	• Ingeniero / Licenciado en Computación	• Especialización o estudios sobre seguridad de los datos

			información contenida en la organización • Realiza planes de recuperación y contingencia e investiga brechas de seguridad • Planifica, supervisa y gestiona hackeos éticos a los sistemas del organismo para identificar y reparar posibles vulnerabilidades • Supervisa y gestiona la implementación de controles vinculados a identificación y gestión de accesos • Implementa técnicas de autenticación y autorización en los sistemas • Determina las medidas de seguridad	• Plataforma de interoperabilidad • Arquitectura de datos • Arquitectura del organismo y de la Arquitectura Integrada de gobierno • Marco de Ciberseguridad de AGESIC	• Capacidad para establecer un plan estratégico en seguridad de la información • Asegura que los recursos y presupuesto de seguridad son suficientes para cumplir los objetivos propuestos		

			adecuadas según la clasificación de los distintos tipos de información				
Control de cumplimiento normativo / Delegado de protección de datos	A demanda	Valida que el análisis de datos respete el marco regulatorio y legal vigente	• Monitorea que se cumplan los reglamentos y los estándares • Resuelve cuestiones legales referidos al análisis de datos	• Marco regulatorio y legal sobre análisis de datos • Competencias del organismo		• Abogado o similar	• Estudios en sistemas de información
Tomador de Decisiones / interesados	Part time	Nexo entre el área usuaria, el interesado y los Analistas de información ya que es quien tiene el conocimiento experto	• Recomendación de variables y formas de presentación de los modelos de análisis • Evaluar, junto con el Custodio de datos y el Especialista / Ingeniero de datos, la calidad de los datos • Recomendar las fuentes de datos necesarias para el análisis • Validar los datos resultantes del	• Experto en las funciones del organismo • Datos e información disponible y en que sistemas	• Habilidad y pensamiento analítico • Framework de calidad y análisis de datos	• Estudios que competan a la entidad donde se desempeña	• Capacitación en Analytics/BI

			análisis, verificando que sean coherentes con la realidad • Apoyar en la interpretación de los datos y su análisis • En ocasiones descubrir necesidades y definirlas				
Encargado de Dominio	Part time	Nexo entre el área usuaria, el interesado y quien tiene el conocimiento experto (Analistas de información, Científico de datos, etc.)	• Recomendación de variables y formas de presentación de los modelos de análisis • Evaluar, junto con el Custodio de datos y el Especialista / Ingeniero de datos, la calidad de los datos • Recomendar las fuentes de datos necesarias para el análisis • Validar los datos resultantes del análisis, verificando que	• Experto en las funciones del organismo (dependiendo del área puede contar con conocimientos en Marketing, Operaciones, Finanzas, etc.) • Datos e información disponible y en que sistemas	• Habilidad y pensamiento analítico • Framework de calidad y análisis de datos	• Estudios que competan a la entidad donde se desempeña	• Capacitación en Analytics/BI

			sean coherentes con la realidad • Apoyar en la interpretación de los datos y su análisis • En ocasiones descubrir necesidades y definirlas				

Tabla 4: Roles y responsabilidades que se agregan a cargos existentes

También se definieron los **roles necesarios en un equipo de análisis de datos,** dependiendo del nivel de la estructura organizativa (básica, avanzada) y la tecnología que se va a utilizar.

En la siguiente tabla se puede ver para cada nivel de la estructura organizativa, cuáles de los roles definidos anteriormente, son necesarios para la realización de un análisis de datos tradicional.

	Estructura organizativa	
Rol	**Básico**	**Avanzado**
Arquitecto Análisis de Datos		*
Especialista / Ingeniero de datos	*	*
Científico de Datos		*
Analista de información	*	*
Custodio de datos		*

Tabla 5: Roles necesarios por nivel de estructura organizativa

Además, se indican los roles para la realización de un análisis de datos utilizando Big Data o Machine Learning.

	Estructura organizativa	
Rol	**Básico**	**Avanzado**
Arquitecto Análisis de Datos		*
Especialista / Ingeniero de datos	*	*
Científico de Datos	*	*
Analista de información	*	*
Custodio de datos		*

Tabla 6: Roles necesarios por nivel de estructura organizativa

Por último, se indica cuál debería ser la **participación de los roles** en las diferentes etapas del proceso de análisis. Para ello, se utiliza una matriz de asignación de responsabilidades (RACI). Los posibles valores de las celdas de la tabla son:

- **R:** indica que se trata de un responsable de la actividad.
- **A:** rol de asistencia.
- **C:** rol de consulta.
- **I:** rol de informado.

Roles	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
Arquitecto Análisis de Datos	C	A	R	R	R	I
Especialista / Ingeniero de datos	C	R	R	A		I
Científico de Datos	C	A	R	R	R	I
Analista de información	R	A	R / A	R / A	R	A
Custodio análisis de datos	C	C	C	C	C	C
Responsable de Seguridad de la Información		C	C	C	C	
Control Regulatorio y Legal	C	C	C	C	C	C
Tomador de Decisiones / interesados	R			R	R	R
Encargado de Dominio	A	A	A	A	A	A

Tabla 7 Participación de los distintos roles en el proceso

4.2. Tecnología

La presente sección tiene como objetivo definir los distintos componentes tecnológicos para el análisis de datos. Para esto se definen cuatro arquitecturas de distintas complejidades, y para cada una de ellas se explican los componentes.

Entonces, ante la necesidad de analizar datos se debe seleccionar qué arquitectura se ajusta a la necesidad y si se tiene la infraestructura que le da soporte. Así como si los tipos de datos a analizar, pueden ser analizados por la arquitectura escogida.

Se identifican los componentes: Aplicaciones, Datos e Infraestructura, enmarcados en las dimensiones tecnológicas de la Arquitectura Empresarial.

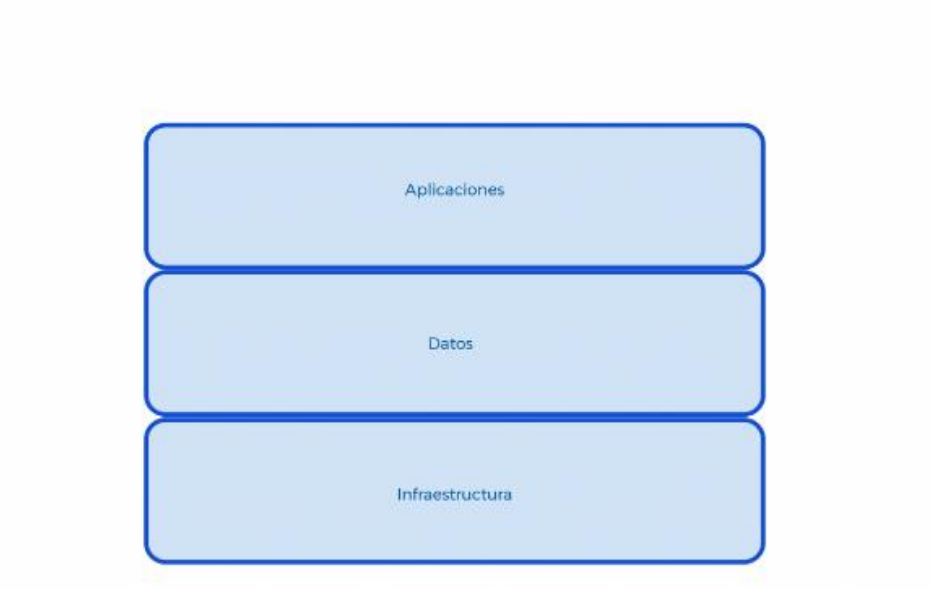

Figura 8 Arquitectura genérica

La dimensión de **Aplicaciones** incluye las aplicaciones que proveen las funcionalidades necesarias para llevar a cabo el análisis de datos. Por lo que se tomarán algunas aplicaciones existentes en el mercado y se las comparará según los siguientes criterios:

- **Grandes volúmenes de datos:** indica si está orientado a grandes volúmenes de datos.
- **Autenticación Web:** explica la forma de autenticarse en la web.
- **Características:** lista de características relevantes de la herramienta.
- **Compatibilidad / Integración:** indica contra qué herramientas se puede integrar o es compatible.
- **Conexión con fuentes de datos:** lista las fuentes a las que se puede conectar la herramienta.
- **Configuración:** indica si se necesita configurar el ambiente antes de que quede operativa la herramienta.
- **Dashboards analíticos:** indica si se puede crear dashboards analíticos.
- **Enfoque:** describe la herramienta a alto nivel.

- **Entrenamiento de modelos:** indica si puede realizar entrenamiento de modelos.
- **Escalabilidad:** indica si la herramienta es escalable.
- **Escritura de datos:** lista los conectores a utilizar para escribir datos.
- **Exploración interactiva:** indica si se permite una interacción interactiva por parte del usuario.
- **Gestión de usuarios:** indica si existe la posibilidad de gestionar el acceso a la información por parte de los usuarios del sistema. Así como la creación de usuarios, grupos de usuarios, permisos, etc.
- **Lectura de datos:** lista los conectores a utilizar para leer datos de las fuentes.
- **Lenguajes soportados:** lista los lenguajes de programación que son soportados por la herramienta.
- **Licenciamiento:** explica el tipo de licenciamiento: Libre o Paga, en caso del segundo detalla si existe una versión gratuita.
- **Machine Learning:** indica si permite técnicas de Machine Learning.
- **Monitoreo de actividades:** describe, si existen, los mecanismos para monitorear las actividades de los usuarios.
- **Open Source:** especifica si la herramienta es Open Source.
- **Paralelismo:** indica si la herramienta permite la ejecución en paralelo.
- **Perfil de usuario:** establece qué tipos de usuarios pueden utilizar la herramienta. Los posibles valores son Funcional y Técnico. Además, se detalla el tipo de conocimiento que debe tener el usuario para un uso exitoso de la herramienta.
- **Performance:** describe si la herramienta tiene buena performance, así como si es escalable y el uso que hace de los recursos.
- **Plataforma:** lista las plataformas en las que se puede ejecutar la herramienta.
- **Requerimientos extra para alta disponibilidad:** indica si se necesita de Hardware extra para tener alta disponibilidad, en caso de que se requiera, detalla cuáles.
- **Tiempo real:** indica si los datos se actualizan en tiempo real.
- **Tipo de fuente:** lista los tipos de fuentes que se pueden manipular.
- **Versión:** indica la versión de la herramienta analizada.

Con respecto a la dimensión **Datos,** detalla cómo los datos pueden ser sujeto de análisis en la arquitectura. Para esto, se utilizan los tipos de datos definidos en la sección Conceptos generales.

Finalmente, la dimensión de **Infraestructura** define los componentes necesarios para el despliegue de la arquitectura.

Tal como se mencionó anteriormente, se definen cuatro arquitecturas base que dan soporte al análisis de datos, es importante notar que éste no es un listado exhaustivo:

- **BI:** conjunto de herramientas que dan soporte al análisis de datos.

- **Big Data en modo batch:** procesamiento de Big Data en modo batch, por lo que se sincroniza la ingesta.
- **Big Data en tiempo real:** procesamiento de Big Data en tiempo real, por lo que la ingesta consiste en la constante recepción de datos.
- **Analítica avanzada:** mediante la aplicación de algoritmos sobre los datos, se permite predecir y encontrar patrones en los mismos.

Entonces, para las cuatro arquitecturas se realizará el mapeo contra la arquitectura genérica, detallando: aplicaciones, datos e infraestructura. Para las aplicaciones se presentan las distintas herramientas necesarias para cumplir con los requerimientos inherentes a cada una, para esto se comparan distintas herramientas según los siguientes criterios:

Es importante notar que no todos los criterios no aplican a todas las arquitecturas.

Por otro lado, en la sección del modelo conceptual de Gobernanza se define BI, Big Data y Analítica avanzada. En esta dimensión se instancian dichos conceptos al alcance de cada arquitectura.

4.2.1. Conceptos generales

Antes de hondar en las arquitecturas, se presentan conceptos claves a utilizar en las mismas.

Los **procesos de ETL** (Extraction, Transformation & Load) implican la extracción, transformación y carga de los datos requeridos para la ejecución de casos de uso [48]. Para esto se pueden utilizar *scripts* que accedan a la base de datos y ejecuten la transformación y filtrado necesario sobre los datos. Se presenta el proceso en la Figura 9[3].

La extracción es el primer paso del proceso de ETL, refiere a la obtención de los datos desde las distintas fuentes de datos. Mediante la transformación de datos, se logra que los mismos puedan ser interpretados de forma correcta para que así, en el último paso se puedan cargar en el almacén de datos.

Dependiendo del orden en que se ejecuten las etapas, se denomina de forma distinta. Por ejemplo, el proceso ELT refiere a ejecutar primero la extracción, luego la carga y por último la transformación de los datos.

Figura 9 Proceso ETL

[3] Link: https://drive.google.com/file/d/1Lvk36mh0dPgK_rvyTQltvj7etPE72_Va/view?usp=sharing

Un **Data Warehouse** (DW) [86] es una colección de datos en donde dos o más fuentes de datos se pueden unir. Generalmente albergan datos bien conocidos y estructurados. Son adecuados para consultas complejas y alta concurrencia.

Un **Data Lake** [86] es una colección de datos que no sufrieron ninguna transformación y que fueron capturados de una gran cantidad de fuentes. Usualmente soporta preparación de datos, análisis exploratorio y actividades de ciencia de datos.

Se definen los siguientes **tipos de datos** [110]:

- **Datos estructurados:** datos que residen en una base de datos relacional, cuenta con una estructura basada en tablas y columnas. Cuenta con claves relacionales que permiten identificar qué columnas predefinidas permiten vincular distintas entidades de información.
- **Datos semi-estructurados:** la estructura de la información sigue determinadas propiedades que permiten un sencillo análisis sin embargo no residen en una base de datos relacional. Para alojar este tipo de datos en una base de datos relacional se requiere la ejecución de procesos de transformación. Algunos ejemplos son: XML, CSV, CLSV y XSL.
- **Datos no estructurados:** los datos no siguen una estructura predeterminada, por lo que pueden estar en cualquier formato como por ejemplo video o imágenes.

Para cada tipo de arquitectura, se especifican qué datos se pueden analizar.

4.2.2. BI

En esta sección, se diseña una arquitectura denominada **BI** que describe los componentes necesarios para realizar un análisis de datos. BI se define en el Modelo Conceptual de Gobernanza.

En la Figura 10[4] se presenta un diagrama en alto nivel que describe el proceso típico que se lleva a cabo en esta arquitectura.

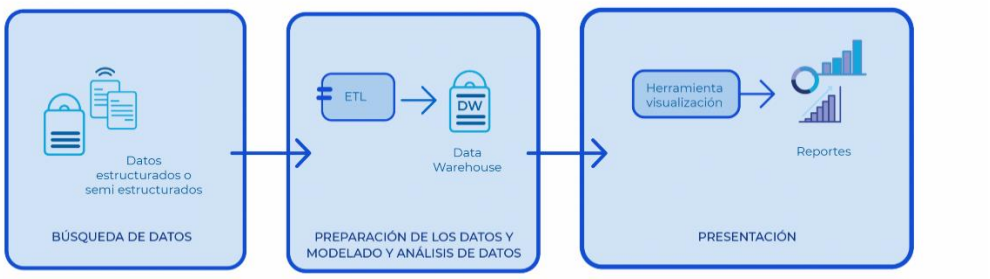

Figura 10 Proceso de BI

[4] Link: https://drive.google.com/file/d/1dsTFAcOPJSJgH0q-Sn14Tjnq3KAodW-F/view?usp=sharing

Como primer paso, en el proceso de análisis de datos, se tiene la “Búsqueda de datos” en donde se identifican las fuentes de datos relacionales y locales del organismo. En la “Preparación de datos”, se realiza el proceso ETL en donde se toman dichos datos y, luego de transformarlos, se alojan en un DW que es diseñado en la etapa de modelado y análisis de datos de forma de contar con un repositorio que conteste las preguntas de la organización. Para el modelado del DW existen dos grandes enfoques [98, 99]: Kimball utiliza una metodología Bottom-Up e Inmon utiliza una **Top-down**.

Finalmente, en la “Presentación” mediante las herramientas de visualización se accede al DW para presentar la información en formatos amigables con el usuario, tanto reportes, como gráficos y tableros.

A continuación, se muestra el mapeo de la arquitectura genérica al caso de la arquitectura de BI.

Figura 11 Mapeo BI

En la dimensión **Aplicaciones** se identifican dos componentes que brindarán las funcionalidades necesarias para un adecuado análisis de datos:

- **ETL:** componente para realizar los procesos de extracción, transformación y carga. Ver definición del proceso ETL.
- **Visualizador**: brinda las funcionalidades para generar tableros y reportes a partir de los datos transformados por el componente ETL.

La dimensión **Datos** se define la estructura de los datos utilizados por el organismo. Los tipos de datos que pueden ser consumidos por la presente arquitectura: estructurados y semi-estructurados.

La dimensión **Infraestructura** describe los componentes de la plataforma tecnológica que dan soporte a la implantación de la arquitectura. Así como también indica las fuentes de información desde donde se toman los datos para su análisis y un almacén de datos que contiene los datos transformados por el ETL, a ser consumidos por el Visualizador.

4.2.2.1. Aplicaciones

Los componentes de la dimensión Aplicaciones en BI son: ETL y Herramientas de visualización. A continuación, se presenta un detalle de cada una.

4.2.2.1.1. ETL

Debido a que existen diversas herramientas para llevar a cabo el proceso ETL, se decidió agruparlos según funcionalidades y detallar cuáles de estas herramientas lo permiten. Además, para la comparación se seleccionaron las herramientas sin querer elaborar una lista taxativa: Excel, Microsoft SQL Server, Pentaho Data Integration, Python, R yTalend, es importante notar que en el mercado existen otras alternativas.

En la Tabla 8 se puede ver la comparación entre las distintas herramientas [49, 50, 51, 52], los criterios a considerar son presentados al principio de la sección.

	Excel	Microsoft SQL Server	Pentaho Data Integration	Python	R	Talend
Enfoque	Planilla de datos	Lenguaje de consulta estructurada	Interfaz Drag & Drop Implementado en Java	Lenguaje de programación, cuya curva de aprendizaje es sencilla	Lenguaje de programación Buenas librerías para estadística y modelado	Interfaz Drag & Drop Generador de código Java
Open Source	No	Sí	Híbrido Tiene versión *Community* y *Enterprise*	Sí	Sí	Híbrido Tiene versión Community y Enterprise
Grandes volúmenes de datos	No es para grandes volúmenes de datos	No es recomendado para grandes volúmenes de datos	Sí	Sí, tiene librerías para Big Data	La performance empeora cuando se tienen grandes cantidades de datos	Sí
Configuración	No	Sí	Sí	Sí	Sí	Sí
Conexión con fuentes de datos	Archivos y bases de datos relacionales por ODBC	Bases de datos relacionales por ODBC	Archivos, base de datos relacionales y no relacionales	Archivos, base de datos relacionales y no relacionales	Archivos, base de datos relacionales y no relacionales	Archivos, base de datos relacionales y no relacionales
Plataforma	Windows, MacOS	Windows	Windows, Linux, MacOS	Windows, Linux, MacOS, AIX, Solaris, entre otras.	Windows, Linux, MacOS	Windows (64 bit), MacOS

	Excel	Microsoft SQL Server	Pentaho Data Integration	Python	R	Talend
Paralelismo	No	Sí	Todos los pasos se ejecutan en paralelo, no se puede hacer secuencialmente	Sí	Sí	Sí en versión Enterprise
Gestión de usuarios	El acceso a los datos puede ser mediante una contraseña, tanto para el acceso como para la modificación	No, existen herramientas adicionales	Gestiona usuarios y sus roles con un rol administrador	No, existen herramientas adicionales	Sí, se utiliza RStudio	Se tiene grupos de usuarios y se les asigna roles
Monitoreo de actividades	No	No	No	Depende de dónde se ejecute. Existen herramientas adicionales para monitoreo de CPU y memoria	Depende de dónde se ejecute. Existen herramientas adicionales para monitoreo de CPU y memoria	Creación de archivos de log sobre conexiones, contenido del servidor e interacción de los usuarios
Versión	Office 2019[5]	SQL Server 2019[5]	Data Integration 8.2[5]	Python 3.8.1[5]	3.6.1 para Windows[5]	Talend Open Studio 7.2.1[5]

Tabla 8 Comparativa de ETLs en BI

[5] Última versión a Diciembre de 2019.

4.2.2.1.2. Herramientas de visualización

Las herramientas de visualización permiten a los usuarios analizar los datos de una forma amigable. Estas herramientas generan dos tipos de resultados, reportes que agrupan la información en base a dimensiones e indicadores y tableros que agrupan reportes y asignan formatos más atractivos para los distintos análisis, mediante el uso de diversos tipos de gráficos.

Existen varias herramientas para conseguir una solución de calidad. Para esta arquitectura, se seleccionaron alternativas que sean fáciles de usar y permitan obtener buenos resultados sin necesidad de tener un profundo conocimiento técnico: JupyterHub, Pentaho Community, Pentaho Enterprise, QlikView, QlikSense, Tableau y PowerBI, es importante notar que en el mercado existen otras alternativas disponibles. Estas herramientas buscan ser de auto-servicio de forma que los perfiles más funcionales puedan generar sus propios análisis de información.

Tal como se explicó al principio de la sección, en la Tabla 9 se presenta un comparativo, a nivel de funcionalidades, entre las herramientas elegidas [53, 54].

	JupyterHub	Pentaho Community	Pentaho Enterprise	QlikView	QlikSense	Tableau	PowerBI
Open Source	Sí	Sí	No	No	No	No	No
Licenciamiento	Libre	Libre	Modelo pago por soporte y software complementario que facilita la instalación y su uso	Paga	Si bien es paga, existe una versión Desktop para Windows gratuita	Versión gratuita que permite navegar en tableros ya elaborados y paga que permite conectar a fuentes y generar tableros	Versión gratuita que permite conectar a fuentes y generar tableros Versión con licencia paga que permite distribuir a otros formatos
Plataforma	El servidor únicamente soporta Unix/Linux	Todos los SO	Todos los SO (Windows 64 bit)	Windows (64 bit), MAC, iPhone, iPad, Web	Windows (64 bit), MAC, iPhone, iPad, Web, Android	Web, Mobile (iPhone, Android) La versión Desktop solamente en Windows (64 bit) y MacOS	Windows
Dashboards analíticos	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Exploración interactiva	No	No	Sí	Sí	Sí	Sí	Sí

	JupyterHub	Pentaho Community	Pentaho Enterprise	QlikView	QlikSense	Tableau	PowerBI
Perfil de usuario	Técnico Requiere conocimientos de programación	Técnico Requiere conocimiento de consultas MDX o SQL para potenciar los dashboards	Técnico Si bien la interfaz es amigable, requiere conocimientos de consultas MDX para potenciar los dashboards	Técnico Requiere conocimientos técnicos de análisis de datos	Funcional Intuitiva y fácil de usar	Funcional Drag and Drop de dimensiones e indicadores y cuenta con una sintaxis propia para cálculos	Funcional No requiere conocimientos técnicos
Gestión de usuarios	El recurso whitelist tiene la lista de los usuarios que tienen permitido acceder al sitio, la misma es gestionada por un usuario administrador	Existe una consola de usuario que permite definir usuarios y roles	Existe una consola de usuario que permite definir usuarios y roles	QMC permite realizar tareas administrativas. Tiene distintas vistas para agregar permisos a los usuarios: a nivel de documentos, objetos en el servidor, grupos, entre otros	Usa QMC para la gestión de usuarios	El usuario administrador gestiona a los usuarios: agrega nuevos, crea grupos y los agrega a estos, gestiona los roles en el sitio, entre otros	En el portal del usuario administrador se puede gestionar los usuarios

	JupyterHub	Pentaho Community	Pentaho Enterprise	QlikView	QlikSense	Tableau	PowerBI
Monitoreo de actividades	Genera logs	Envía mails cuando un *job* falla	Permite ver los logs mediante una interfaz web y gráficas de performance	La herramienta QlikView System Monitor permite capturar los logs	Mediante QMC se puede monitorear el uso de la herramienta	Genera logs sobre conexión de datos, contenido en el servidor e identificación de usuarios. Se tiene herramientas open source (LogShark y TabMon) para analizarlos	En el portal del administrador se puede ver la actividad de los usuarios y sus grupos
Versión	1.0.0[6]	8.2[6]	8.3[6]	12[6]	2.2[6]	Desktop 2019.2[6]	Desktop 2.74.5619.621[6]

Tabla 9 Comparativa de Visualizadores en BI

[6] Última versión a Diciembre de 2019.

4.2.2.3. Infraestructura

La dimensión Infraestructura permite visualizar cómo se despliegan los componentes de la arquitectura y cómo interaccionan ente ellos.

Tal como se mencionó en la introducción a la arquitectura BI, la infraestructura debe contar con un proceso ETL que tome los datos de las bases de datos relaciones y locales de cada organismo, y que almacene dichos datos, luego de procesarlos, en un almacén de datos. Además, se requiere de herramientas de visualización para mostrar los datos y generar valor.

Los distintos componentes pueden estar en cada uno de los organismos. Se entiende como recomendable, mínimamente, contar con un servidor que aloje el almacén de datos (denominado DW en la Figura) y los procesos ETL (Figura 12[7]). Las herramientas de visualización, en su mayoría, cuentan con versiones instalables en equipos de escritorio para una fácil administración directamente desde los usuarios.

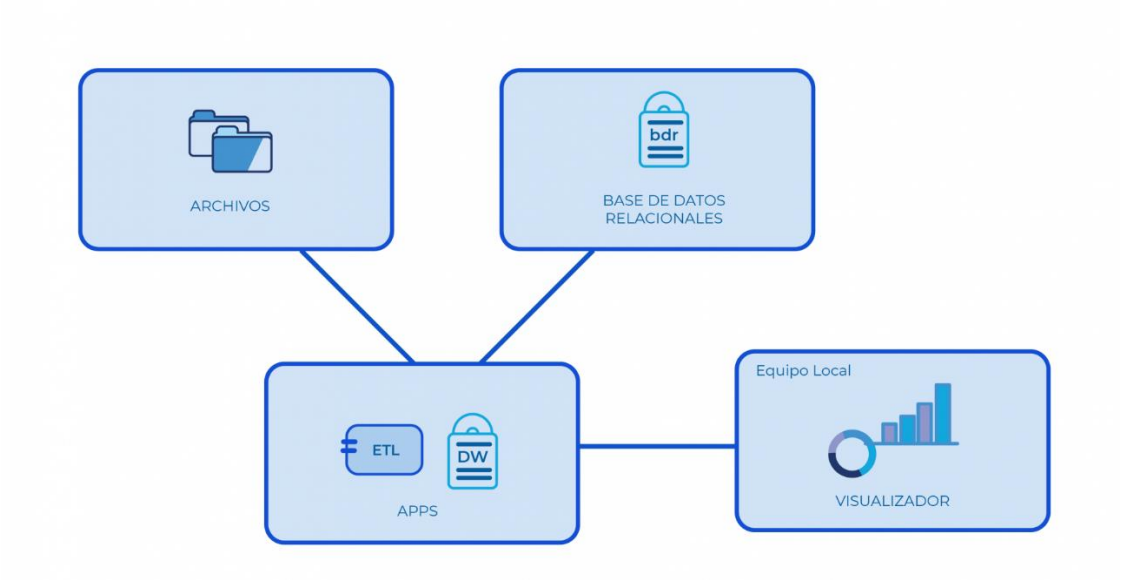

Figura 12 Despliegue básico de BI

Por otro lado, en la Figura 13[8] se presenta otro despliegue de la infraestructura, en donde el componente de ETL y el DW se separan en distintos servidores, brindando mayor robustez a la solución.

[7] Link: https://drive.google.com/file/d/1pHEDZ0EdqIdTSk9e9F1B6mLqmqnaHynL/view?usp=sharing
[8] Link: https://drive.google.com/file/d/1pHEDZ0EdqIdTSk9e9F1B6mLqmqnaHynL/view?usp=sharing

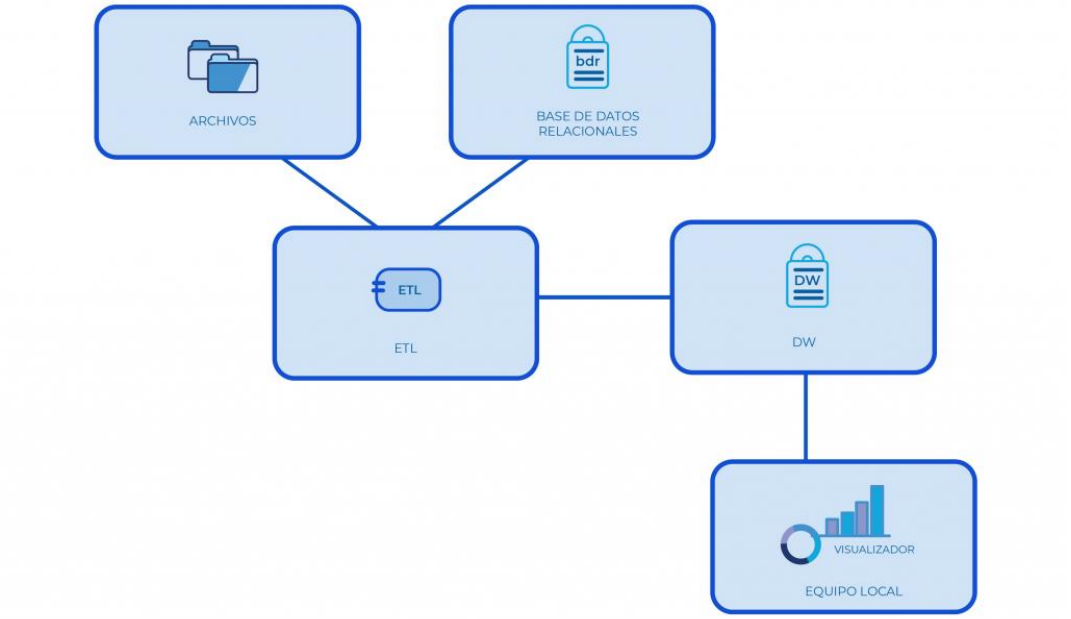

Figura 13 Despliegue robusto de BI

4.2.3. Big Data

Tal como se definió en la Sección de Modelo Conceptual de Gobernanza, Big Data permite generar soluciones de procesamiento innovadoras para una gran variedad de datos, trayendo así beneficios en los organismos. Los datos tratados en este paradigma, cumplen con las siguientes propiedades, denominadas “Las cinco V” (Figura 14):

- **Volumen:** Grandes volúmenes de datos.
- **Variedad:** Los datos provienen de múltiples repositorios, dominios o tipos, por lo que son altamente heterogéneos.
- **Velocidad:** Grandes flujos de datos que deben ser procesados rápidamente.
- **Veracidad:** Como los datos son heterogéneos, pueden tener inconsistencias.
- **Valor:** Los datos por sí solos no generan valor, pero al ser combinados y procesados se logra generar valor para el negocio.

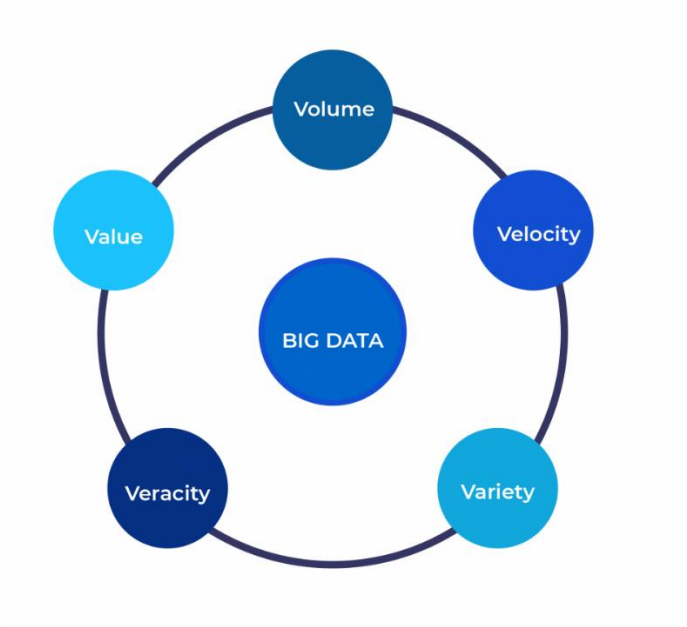

Figura 14 Propiedades de Big Data. Fuente [63]

Al tener en cuenta el proceso estándar, definido por el componente de Gobernanza, se puede detectar qué pasos debe llevar a cabo una arquitectura de Big Data para un exitoso análisis de datos. La Figura 15[9], especifica estos pasos.

Figura 15 Proceso de Big Data

En la etapa de "Búsqueda de datos" se realiza la captura de datos, ésta puede realizarse en modo pasivo o activo. En donde, el modo activo es aquel en que los usuarios son conscientes de que se están obteniendo sus datos y, entonces, los comparten de forma activa. Por otro lado, en la captura de datos en modo pasivo, tal como lo indica su nombre, es cuando los usuarios no son conscientes de que están compartiendo sus datos.

Luego, en la etapa de "Preparación de datos" se tiene el proceso de Ingesta, el cual puede tener dos modalidades:

[9] Link: https://drive.google.com/file/d/1P337UKg1JfuE1HSCySgG4jaSPVx2E8ZS/view?usp=sharing

- **Batch:** Debe existir una herramienta que planifique cuándo se va a ejecutar la ingesta de datos.
- **Streaming:** Da soporte a una ingesta en tiempo real de los datos.

Luego de ingestar los datos, ya sea en tiempo real o de forma planificada, los datos pueden ser transformados, limpiados y fusionados. En la etapa de "Modelado y análisis de datos", como lo dice su nombre, se realiza el análisis de los datos. Finalmente, mediante reportes y visualizaciones se realiza la etapa "Presentación" gracias a la cual se podrán tomar acciones. A lo largo del marco de trabajo, se explicarán con mayor nivel de detalle, junto con las herramientas a utilizarse.

En la Figura 16 se puede ver el mapeo de las arquitecturas definidas a la arquitectura genérica. La única diferencia entre las arquitecturas, es que para el caso de Big Data en tiempo real no se tiene la aplicación Scheduler.

Figura 16 Mapeo Big Data en modo batch

En la dimensión **Aplicaciones** se identifican los siguientes componentes que brindan las funcionalidades necesarias para un adecuado análisis de datos:

- **Procesamiento de datos:** limpieza, normalización, transformación y fusión de los datos.
- **Ingesta de datos:** guardar los datos, ya normalizados, necesarios para el análisis.
- **Scheduler:** permite la planificación de la ingesta (solamente para la arquitectura Big Data en modo batch).
- **Visualizador**: brinda las funcionalidades para generar tableros y reportes a partir de los datos transformados por el componente ETL.

La dimensión **Datos** se define la estructura de los datos utilizados por el organismo. Los tipos de datos que pueden ser consumidos por la presente arquitectura son: datos estructurados, semi-estructurados y no estructurados.

La dimensión **Infraestructura** describe los componentes de la plataforma tecnológica que dan soporte a la implantación de la arquitectura. Así como también las fuentes de información desde donde se toman los datos para su análisis.

4.2.3.1. Aplicaciones

Los componentes de la dimensión Aplicaciones para Big Data en modo batch son: Ingesta de datos, Procesamiento de datos, Scheduler y Herramientas de visualización. Por otro lado, para la arquitectura Big Data en tiempo real son: Ingesta de datos, Procesamiento de datos y Herramientas de visualización. A continuación, se presenta en detalle cada una.

4.2.3.1.1. Ingesta de datos

Los organismos pueden poseer datos en diferentes formatos y fuentes, por lo que es necesario disponerlos en una plataforma de Big Data para que sean procesados. La tecnología más adecuada, varía según el tipo de fuente de datos a ingestar, así como si se necesita hacerlo en tiempo real (arquitectura de Big Data en tiempo real) o no (arquitectura Big Data en modo batch), en este marco se tomó un conjunto acotado de éstas:

- **Base de datos externas:** Sqoop, Attunity Replicate (Qlik).
- **Cola de mensajes:** Apache Kafka, Apache Flume y RabbitMQ.
- **Web Service:** Apache Flume, Apache Nifi.
- **Archivos:** Apache Nifi, Apache Flume.
- **Logs:** Apache Flume, LogStash y Fluentd.

La Tabla 10 muestra la comparación entre las herramientas seleccionadas [55, 56, 57, 58, 59] mencionadas anteriormente, esta lista no pretender ser una lista exhaustiva de herramientas. Se comparan los criterios presentados al principio de la sección.

Es importante notar que, si en la fila “Tiempo real” se tiene una celda con valor “No” para la columna X, quiere decir que la herramienta en la posición X no puede ser utilizada para la arquitectura de Big Data en tiempo real.

	Apache Sqoop	Attunity Replicate (Qlik)	Apache Kafka	Apache Flume	RabbitMQ	Apache Nifi	LogStash	Fluentd
Enfoque	Es una herramienta para transferir bloques de datos forma eficiente entre Hadoop y bases de datos relacionales	Mueve los datos de forma segura, fácil y eficiente con mínimo impacto en el sistema	Es un bus de mensajería optimizado para flujos de datos de alta entrada y la repetición	Servicio distribuido que maneja eficazmente grandes volúmenes de datos	Sistema de colas de mensajes MQ, permite comunicación segura y rápida entre varios actores	Mediante una interfaz Drag & Drop permite automatizar el fujo de datos entre distintos sitemas	Usado para recopilar, parsear y almacenar logs. Tiene gran cantidad de plugins	Usado para recopilar y unificar datos
Open Source	Sí	No	Open Source: Licencia Apache 2.0	Open Source: Licencia Apache 2.0	Open Source: Licencia pública Mozilla	Sí	Sí	Sí

	Apache Sqoop	Attunity Replicate (Qlik)	Apache Kafka	Apache Flume	RabbitMQ	Apache Nifi	LogStash	Fluentd
Tipo de fuente	Base de datos relacionales externas. Import / Export de base de datos relacionales externas y HDFS/Hive	Bases de datos relacionales y no relacionales	Cola de mensajes	Cola de mensajes, Web service, archivos y logs	Cola de mensajes	Archivos y Web service	Log	Log
Tiempo real	No	Sí	Sí	Alto rendimiento	Sí	Sí	No	No
Performance	Escalabilidad limitada debido a su uso de MapReduce y consume mucha CPU y E/S	Escalable	Escalable y tolerante a fallos	Escalable y tolerante a fallos	Con grandes cantidades de datos no tiene buena performance	Escalable	Consume mucha memoria y recursos	Consume poca memoria

	Apache Sqoop	Attunity Replicate (Qlik)	Apache Kafka	Apache Flume	RabbitMQ	Apache Nifi	LogStash	Fluentd
Lectura de datos	Conector JDBC	Conectores JDBC, ODBC, OLE DB, ADO.NET	Variedad de conectores entre ellos: ActiveMQ, IBM MQ, JDBC, JMS, Replicator	Multitud de formatos y aplicaciones: JMS, HTTP, Kafka, IRC, Solr, Kite y Custom Para logs: Avro, Thrift, Syslog y Netca	Conectores para: Avro, Binary, CSV, MySQL, Postgres, JSON, Log, Protobuf, SDC Record, Text, XML	Conectores y transformers para: JDBC, EWS, MongoDB, Cassandra, HDFS, Hive, Amazon S3, HBase, Elasticsearch, HTTP, WebSocket, POP3, TCP, UDP, Syslog, Avro, ORC, Parquet, Base64, JSON, XML, CSV, TEXT, JMS, Kafka, AMQP Customizados: se puede desarrollar	STDIN, Syslog, Files, TCP/UDP, Microsoft windows, Eventlogs, Websocket, Zeromq y extensions customizadas	Logs

	Apache Sqoop	Attunity Replicate (Qlik)	Apache Kafka	Apache Flume	RabbitMQ	Apache Nifi	LogStash	Fluentd
Escritura de datos	Conector HDFS	Azure Data Lake	Variedad de conectores, entre ellos: HDFS, Amazon S3, ElasticSearch y JDBC	Conectores para: HDFS, HBase, Logger, File Roll, Null y ElasticSearch	Conector para HDFS	Ídem lectura	Azure,Elastic, CloudWathc ,Files,github, ganglia,google_pubsub, graphite,HTTP,JMX, RabbitMQ,S3 ,TCP, syslog,WebSocket	Archivos, MongoDB, MySql, Amazon S3
Lenguajes soportados	Bash y Java	Java y .Net	Ruby, Python, Java y Node.js	Java	Erlang	Java	Se pueden exportar métricas utilizando Python, Java o la API HTTP	Se pueden crear custom plugin con Ruby
Gestión de usuarios	Sí, tiene integración con Kerberos. Pueden gestionarse grupos y privilegios		Sí, se cuenta con: SSL, SASL y ACL	Sí, se cuenta con: Kerberos y KDC	Sí, usuarios para acceso a interfaz de monitoreo	Sí	Sí, mediante credenciales	No

	Apache Sqoop	Attunity Replicate (Qlik)	Apache Kafka	Apache Flume	RabbitMQ	Apache Nifi	LogStash	Fluentd
Monitoreo de actividades	No		Sí, exporta métricas en JXM	Sí, exporta métricas en JXM, JSON y Ganglia	Sí, a través de línea de comandos o interfaz gráfica. Integración con Prometheus, Grafana y ELK	Sí, provee una interfaz para consultar estado de procesos, uso de memoria y disco	Sí, provee información de hilos y procesos, estado de la memoria JVM y del sistema operativo	Sí, expone métricas mediante una API REST
Versión	Sqoop 2 1.99.7[10]	6.2[10]	2.3[10]	1.9.0[10]	3.7.22[10]	1.10.0[10]	7.5.0[10]	1.0[10]

Tabla 10 Comparativa de herramientas de ingesta de datos en Big Data

[10] Última versión a Diciembre de 2019.

4.2.3.1.2. Procesamiento de datos

Los datos pueden ser tomados de diversas fuentes de datos:

- Bases de datos relacionales.
- Bases de datos no relacionales.
- Archivos con formatos estructurados, semi-estructurados y no estructurados.

Por ende, es necesario **normalizar** y **transformar** estos datos, es decir: llevarlos a un formato legible para facilitar su procesamiento. A su vez, al ser datos de distintos orígenes, pueden requerir ser **fusionados** para generar valor y realizar tareas de **limpieza**. Esto es muy similar al proceso de ETL, pero en vez de en este caso se realiza primero la carga y luego la transformación y es por esto que se referencia como proceso ELT.

Para las tareas de normalización, fusionado y limpieza se utilizan diversas herramientas. Con el fin de realizar una tabla de comparación entre las herramientas, se decidió tomar el siguiente conjunto acotado de las mismas: Apache Nifi, Apache Sqoop, Apache Spark y Apache Hive, Apache Pig. En la Tabla 11 se puede apreciar la comparación [60]. Los criterios relevantes para el procesamiento de datos se presentan al principio de la sección.

	Apache Nifi	Apache Sqoop	Apache Spark	Apache Hive	Apache Pig	Python
Enfoque	- Interfaz de usuario para el manejo y monitoreo de flujo de datos	- Herramienta para transferir datos entre Apache Hadoop y bases de datos relacionales de Apache	- Framework para procesamiento de grandes volúmenes de datos - Multilenguaje (Python, R, Scala)	- Herramienta distribuida que opera con HDFS	- Plataforma para analizar grandes volúmenes de datos sobre Hadoop - Permite funciones de usuarios escritas en Java o lenguajes de scripting que puedan compilarse en Java	- Lenguaje de programación - Técnicas avanzadas de Machine Learning
Open Source	Sí	*Sí*	Sí	Sí	Sí	Sí
Grandes volúmenes de datos	Sí	Sí	Sí	Sí	Sí	Sí, tiene librerías para Big Data
Conexión con fuentes de datos	Archivos, base de datos relacionales y no relacionales	Bases de datos relacionales y Hadoop	Archivos, base de datos relacionales y no relacionales	Base de datos relacionales y no relacionales	Archivos, bases de datos relacionales, no relacionales y otras fuentes usando funciones definidas por el usuario	Archivos, base de datos relacionales y no relacionales
Paralelismo	Sí	Sí	Sí	Sí	Sí	Sí
Machine Learning	Sí	No	Sí	No	Sí	Sí

	Apache Nifi	Apache Sqoop	Apache Spark	Apache Hive	Apache Pig	Python
Perfil de usuario	Funcional Interfaz Drag & Drop	Funcional Conocimientos sobre consultas a bases de datos	Técnico Conocimientos en Programación, curva de aprendizaje sencilla al tener distintos lenguajes a disposición	Funcional Conocimientos sobre consultas a bases de datos, fácil de aprender	Funcional Lenguaje de scripting similar a SQL, curva de aprendizaje sencilla	Técnico Conocimientos de programación, curva de aprendizaje es sencilla
Gestión de usuarios	Sí	Sí, tiene integración con Kerbreros. Pueden gestionarse grupos y privilegios	Provee autenticación y autorización de usuarios en la aplicación web de Spark	Sí, con Apache Knox, Kebreros, LDAP, Ambari y Ranger	No	No, existen herramientas adicionales
Monitoreo de actividades	Sí, provee para consultar estado de procesos, uso de memoria y disco	No	Se puede consultar los logs o la aplicación web de Spark	Sí, se pueden obtener métricas de performance mediante PerfLogger	Se pueden exportar métricas	Depende de dónde se ejecute. Existen herramientas adicionales para monitoreo de CPU y memoria
Versión	1.10.0[11]	Sqoop 2 1.99.7[11]	2.4.2[11]	1.0.0[11]	0.17.0[11]	Python 3.8.1[11]

Tabla 11 Comparación de herramientas de procesamiento de datos en Big Data

[11] Última versión a Diciembre de 2019.

4.2.3.1.3. Scheduler

Las ingestas pueden ser configuradas de forma que se ejecuten cada cierto tiempo, mediante una herramienta de Scheduler [87], dicha herramienta se encarga de la gestión y ejecución de flujos de trabajo que contienen secuencias de acciones para la carga de datos.

En la Tabla 12 se hace una comparación de algunas de las aplicaciones que sirven para planificar las ingestas, para esto se seleccionaron las siguientes herramientas: Apach Oozie, Azkaban y Apache Airflow [61], es importante notar que existen alternativas en el mercado.

Con el fin de comparar las herramientas, se usaron los criterios presentados al principio de la sección.

	Apache Oozie	Azkaban	Apache Airflow
Enfoque	Planificador de flujos de trabajo de Hadoop	Es un planificador de flujos de trabajo de Hadoop	Procesamiento por lotes de uso general
Open Source	Sí	Sí	Sí
Autenticación Web	Kerberos	Contraseña XML	LDAP o con contraseña
Escalabilidad	Es escalable	Es escalable, aunque en menor medida que Apache Oozie	Puede llegar a ser escalable pero debe ser configurado
Monitoreo	Sí	Limitado	Sí
Requerimientos extras para alta disponibilidad	Balanceador de carga para los nodos web, base de datos y Zookeeper	Base de datos	Celery/Dask/Mesos, balanceador de carga y base de datos
Gestión de usuarios	Sí, se pueden definir permisos a nivel de ejecución, escritura y lectura de Jobs	En el archivo de properties se puede: agregar usuarios, definir grupos y definir roles	Usa archivos de configuración, utiliza el mismo parser de Python

	Apache Oozie	Azkaban	Apache Airflow
Monitoreo de actividades	Sí, se puede monitorear memoria y logging. Colas y conexiones	Monitorea las tareas y trackea a los usuarios	Para el monitoreo se debe usar una herramienta externa, como por ejemplo StatsD
Versión	5.1.10[12]	Azkaban 3.80.0[12]	1.10.6[12]

Tabla 12 Comparación de Scheduler en Big Data

4.2.3.1.4. Herramientas de visualización

Las herramientas de visualización pueden clasificarse según cómo muestran los datos:

- **Estática:** no permite navegación en la visualización.
- **Interactiva:** permite navegar en la visualización, interactuando con los datos.
- **Tiempo real:** los datos que se observan están actualizados.

Existen varias herramientas para conseguir una solución de calidad. Para las dos arquitecturas planteadas se escogieron las herramientas: JupyterHub, Apache Zepellin, QlikView, QlikSense, Tableau, PowerBI, Kibana siendo que esta lista no pretende ser exhaustiva, sino con fines comparativos.

En la Tabla 13 se realiza un comparativo, a nivel de funcionalidades, entre las herramientas elegidas [53, 54]. Los criterios utilizados se describen al principio de la sección.

[12] Última versión a Diciembre de 2019.

	JupyterHub	Apache Zeppelin	QlikView	QlikSense	Tableau	PowerBI	Kibana
Open Source	Sí	Sí	No	No	No	No	Sí
Licenciamiento	Libre	Libre	Paga	Si bien es paga, existe una versión Desktop para Windows gratuita	Versión gratuita que permite navegar en tableros ya elaborados y paga que permite conectar a fuentes y generar tableros	Versión gratuita que permite conectar a fuentes y generar tableros Versión con licencia paga que permite distribuir a otros formatos	Paga. Tiene una versión gratuita con la mayoría de las funcionalidades, no posee Alertas, Gestión de Ingesta, ente otras
Plataforma	El servidor únicamente soporta Unix/Linux	Mac OSX, Ubuntu 14.4, CentOS 6.X and Windows 7 pro SP1 (64 bit)	Windows (64 bit), MAC, iPhone, iPad, Web	Windows (64 bit), MAC, iPhone, iPad, Web, Andorid	Web, Mobile (iPhone, Android) La versión Desktop solamente en Windows (64 bit) y MacOS	Windows	Windows (64 bit), MAC, Linux, Docker, Kubernetes
Dashboards analíticos	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Exploración interactiva	No	Sí	Sí	Sí	Sí	Sí	Sí
Tiempo real	Sí	Sí	No	Sí	Sí	Sí	Sí

	JupyterHub	Apache Zeppelin	QlikView	QlikSense	Tableau	PowerBI	Kibana
Perfil de usuario	Técnico Requiere conocimientos de programación	Técnico Requiere conocimientos de programación	Técnico Requiere conocimientos técnicos de análisis de datos	Funcional Intuitiva y fácil de usar	Funcional Drag and Drop de dimensiones e indicadores y cuenta con una sintaxis propia para cálculos	Funcional No requiere conocimientos técnicos	Funcional. Intuitiva y fácil de usar. Requiere conceptos básicos de SQL y expresiones regulares
Gestión de usuarios	El recurso whitelist tiene la lista de los usuarios que tienen permitido acceder al sitio, la misma es gestionada por un usuario administrador	Mediante el Framework Apache Zepellin, se consiguen métodos de autenticación, autorización, criptografía y gestión de la sesión de los usuarios	QMC permite realizar tareas administrativas. Tiene distintas vistas para agregar permisos a los usuarios: a nivel de documentos, objetos en el servidor, grupos, entre otros	Usa QMC para la gestión de usuarios	El usuario administrador gestiona a los usuarios: agrega nuevos, crea grupos y los agrega a estos, gestiona los roles en el sitio, entre otros	En el portal del usuario administrador se puede gestionar los usuarios	Se tiene una app para gestionar usuarios

	JupyterHub	Apache Zeppelin	QlikView	QlikSense	Tableau	PowerBI	Kibana
Monitoreo de actividades	Genera logs	Exporta métricas referentes a los jobs en ejecución, drivers y la memoria de disco usada	La herramienta QlikView System Monitor permite capturar los logs	Mediante QMC se puede monitorear el uso de la herramienta	Genera logs sobre conexión de datos, contenido en el servidor e identificación de usuarios. Se tiene herramientas open source (LogShark y TabMon) para analizarlos	En el portal del administrador se puede ver la actividad de los usuarios y sus grupos	En caso de habilitarlo, se pueden tener métricas sobre su uso
Versión	1.0.0[13]	0.8.2[13]	12[13]	2.2[13]	Desktop 2019.4[13]	Desktop 2.74.5619.621[13]	7.4.2[13]

Tabla 13 Comparación de Visualizadores en Big Data

[13] Última versión a Diciembre de 2019.

4.2.3.3. Infraestructura

La infraestructura que de soporte a estas dos arquitecturas debe estar preparada para manejar con grandes volúmenes de datos de diversas fuentes, y para el caso de Big Data en tiempo real dar soporte en tiempo real para la ingesta.

La Figura 17 muestra los típicos componentes en una arquitectura de Big Data. Cada componente es distribuido en el clúster para proveer un procesamiento y almacenamiento distribuido.

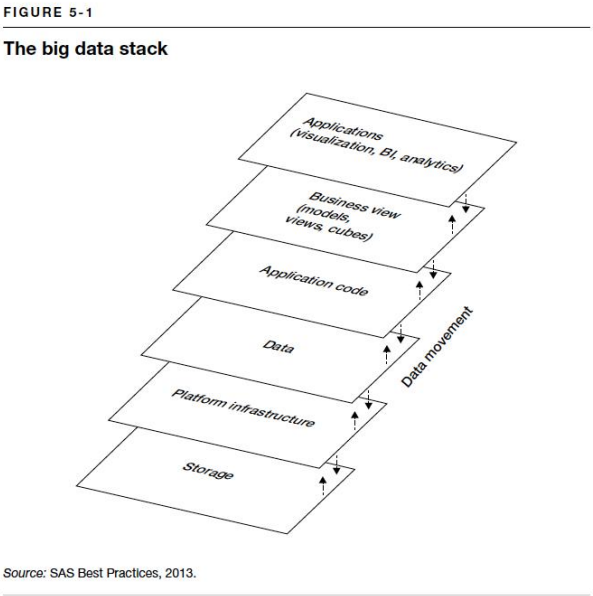

Figura 17 Big Data at Work: Dispelling the Myths, Uncovering the Opportunities

Los componentes son:

- **Almacenamiento:** es necesario almacenar grandes cantidades de datos a bajo costo.
- **Plataforma de Infraestructura:** es necesario contar con una plataforma que disponga de herramientas para el procesamiento de muchos datos, y la capacidad de integrar, manejar y transformar estos datos.
- **Datos:** los datos provienen de fuentes variadas, los organismos deben tomar en cuenta todos los datos que generan y encontrar el potencial valor que los mismos pueden dar.
- **Procesamiento en paralelo:** para el procesamiento de los datos es necesario contar con herramientas que brinden la posibilidad de la paralización del trabajo.
- **Visualización:** es importante contar con herramientas para generar la visualización y reportes a partir de los datos.
- **Aplicaciones:** se dispone de programas y algoritmos que utilizan estas herramientas para finalmente generar el valor deseado.

Tomando en cuenta lo antes mencionado en la Figura 18[14] se presenta el diagrama de la arquitectura, tanto Big Data en modo batch como Big Data en tiempo real, a nivel de componentes:

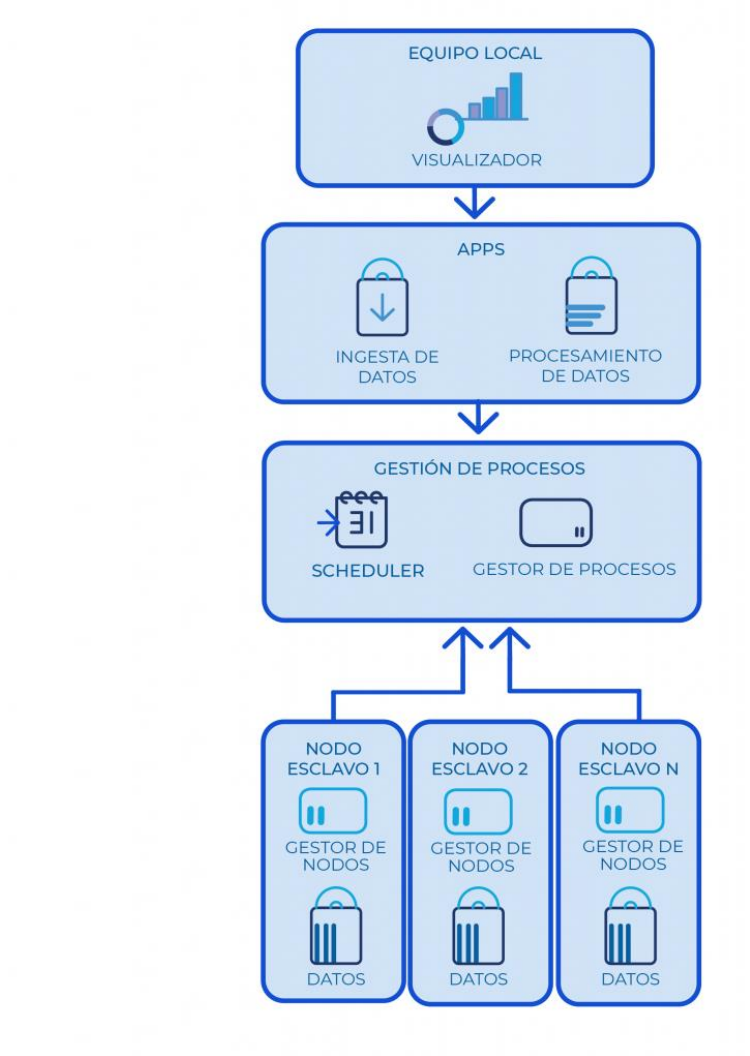

Figura 18 Diagrama de componentes de las Arquitecturas

A continuación, se desarrollan los distintos componentes:

- **Nodos esclavos:** cada nodo almacena datos, lo hacen de forma replicada para así ser tolerante a posibles fallos y no sufrir cuellos de botella. Típicamente se almacenan los datos en: HDFS, Ceph, Apache HBase, S3, Kudo, ElasticSearch, MongoDB, MariaDB, Redis o Apache CouchDB.
- **Gestor de procesos:** encargado de la planificación, únicamente para la arquitectura de Big Data en modo batch, y asignación de recursos para los datos. Algunas de las herramientas utilizadas por la industria son: Yarn, Mesos y Kubernetes.

[14] Link: https://drive.google.com/file/d/1Mi_8KOKqIIfWVx4gAVWGe1nzY4cFl5ie/view?usp=sharing

- **Apps y Visualizador:** se ejecutan las aplicaciones mencionadas en la Sección de Herramientas de visualización.

En el Anexo I: Caso de estudio Big Data, se pueden ver arquitecturas típicamente utilizadas en el paradigma Big Data.

4.2.4. Analítica avanzada

La analítica avanzada consiste en la aplicación de algoritmos que aprenden de los datos, en la Sección de Modelo Conceptual de Gobernanza se da una definición detallada de la misma.

Se identifican dos modalidades del aprendizaje automático: supervisado y no supervisado. La modalidad de aprendizaje **supervisado** se basa en el entrenamiento de un algoritmo que aplica una función sobre variables de entrada que generan una salida. Posee un conjunto de datos que se compone de entradas y salidas originales, a partir del cual el algoritmo aprende. Dentro de este marco se encuentran dos principales aplicaciones:

- **Regresión:** se aplican modelos para predecir los valores que tomará cierta variable, a partir del histórico de datos anteriores o una muestra de datos específica.
- **Clasificación:** de forma similar, a partir del histórico de datos o muestra de datos, el algoritmo tiene la capacidad de determinar la categorización de nuevos datos.

Por otro lado, en la modalidad **no supervisado** no se conoce el output, el algoritmo aprende de los datos de entrada, sin disponer la información sobre las salidas. El algoritmo es capaz de reconocer patrones que refieren a las salidas. Dentro de este campo se encuentra la siguiente categorización general:

- **Clustering:** se agrupan los datos en grupos, donde elementos de un grupo están relacionados entre sí.
- **Asociación de reglas:** identificar patrones en base a reglas en los datos.
- **Detección de anomalías:** se utilizan diferentes técnicas para poder detectar errores o patrones atípicos en los datos.
- **Reducción de dimensiones:** se utilizan técnicas para poder reducir la dimensión de los datos. Esto se puede utilizar para evitar sobreajustes o para poder interpretar los datos, ya sea mediante factores o gráficamente.

Ejemplos de aplicaciones en la industria donde puede aplicarse este tipo de análisis pueden ser para detectar patrones en la sociedad a través de la movilidad de los ciudadanos, encontrar tendencias, agrupar en clúster que categoricen a diferentes tipos de personas, realizar predicciones en función de histórico de datos generados, ya sea por diferentes organizaciones o datos externos como ser redes sociales o el clima. Abre un abanico de posibilidades flexibles que brindan la posibilidad de poder mejorar la calidad de vida de las personas, el servicio que se

les brinda y satisfacer las necesidades actuales. Es importante tomar en cuenta los aspectos sobre la política y privacidad de los datos.

Al tener en cuenta el proceso de análisis de datos, definido por la dimensión de Gobernanza, se puede detectar qué pasos debe llevar a cabo una arquitectura de Analítica avanzada para un exitoso análisis de datos. En la Figura 19[15] se puede apreciar esto.

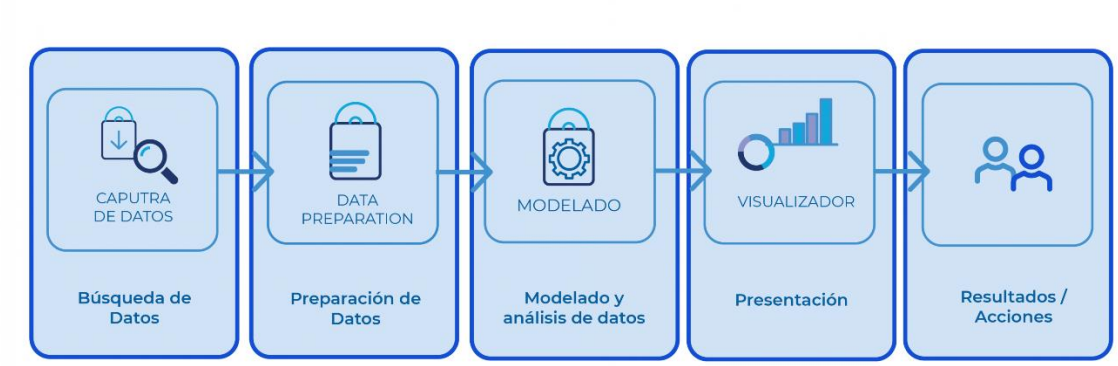

Figura 19 Proceso de Analítica avanzada

En la etapa de "Búsqueda de datos" se realiza la captura de datos, ésta puede realizarse en modo pasivo o activo. En donde, el modo activo es aquel en que los usuarios son conscientes de que se están obteniendo sus datos y, entonces, los comparten de forma activa. Por otro lado, en la captura de datos en modo pasivo, tal como lo indica su nombre, es cuando los usuarios no son conscientes de que están compartiendo sus datos.

Luego, se procede a "Preparación de datos" en la que se transforman los datos para que los algoritmos puedan encontrar patrones en los datos. Esto generalmente requiere la detección de outliers, la normalización de datos y completar los valores faltantes. Por más detalle ver el Modelo Conceptual de Gobernanza.

Además, se incluye el feature engineering, que es el proceso de utilizar el conocimiento en el dominio de los datos para crear features de los datos que permitan aumentar el poder predictivo de los algoritmos.

En la etapa de "Modelado y análisis de datos" se debe diseñar e implementar los algoritmos de aprendizaje automático. Para validar el funcionamiento de los mismos, se generan gráficas para visualizar el aprendizaje y se calculan métricas que miden el nivel de aprendizaje. Estas métricas pueden ser utilizadas no solamente para evaluar el algoritmo por sí solo sino para establecer comparaciones con otros algoritmos que realizan la misma tarea y poder seleccionar el que obtenga mejores resultados, algunos ejemplos son:

- **Para modelos de regresión:** error medio absoluto, error cuadrático medio, raíz del error cuadrático medio.

[15] Link: https://drive.google.com/file/d/1oPP-YVJZ1d_k8TD4ZQsLzko_dxSqE_pf/view?usp=sharing

- **Para modelos de clasificación:** matriz de confusión, accuracy, precision, pérdida logarítmica.
- **Para modelos de clustering:** Silhouette coefficient, Dunn index, Rand index.

Finalmente, en la etapa de "Presentación" es un paso importante del proceso, ya que es el resultado final que ve el usuario. Expresar los resultados de forma amigable y clara es fundamental para poder obtener el valor a partir de los resultados y lograr una correcta interpretación.

Entonces, la arquitectura de Analítica avanzada se basa en el procesamiento de datos variados para generar análisis descriptivos, predictivos y prescriptivos, generando valor a partir de los datos. En la Figura 20 se puede ver el mapeo de la arquitectura genérica al caso de Analítica avanzada.

Figura 20 Mapeo Analítica avanzada

En la dimensión **Aplicaciones** se identifican tres componentes que brindarán las funcionalidades necesarias para un adecuado análisis de datos predictivo:

- **Data Preparation:** tomar decisiones en cuanto a los outliers encontrados y los valores faltantes, normalizar los datos, dejar los datos pre-procesados disponibles para ser utilizados por los modelos de ML.
- **Ejecución de algoritmos de ML:** implica el diseño e implementación de modelos de Machine Learning sobre los datos pre-procesados.
- **Visualizador:** brinda las funcionalidades para generar tableros y reportes a partir de los resultados de los modelos.

La dimensión **Datos** se define la estructura de los datos utilizados por el organismo. Los tipos de datos a ser consumidos por la presente arquitectura son: datos estructurados, semi-estructurados y no estructurados.

Típicamente, antes de la ejecución de los algoritmos los datos deben estar normalizados y se deben solucionar los problemas presentados sobre los mismos (datos que son outliers o que faltan).

Para el entrenamiento de los algoritmos se utiliza un conjunto de los datos, con el objetivo de poder luego con el resto de los datos disponibles validar que el algoritmo aprendió correctamente. Este conjunto de datos se denomina “training set”.

La dimensión **Infraestructura** describe los componentes de la plataforma tecnológica que dan soporte a la implantación de la arquitectura. Así como también las fuentes de información desde donde se toman los datos para su análisis.

4.2.4.1. Aplicaciones

Los componentes de la dimensión Aplicaciones para la arquitectura Analítica avanzada son: Data Preparation, Modelado ML y Herramientas de visualización.

4.2.4.1.1. Data Preparation

Data Preparation representa a un conjunto de metodologías para analizar datos desde distintos puntos de vista, encontrando patrones en los mismos, resumiéndolos y clasificándolos en grupos, así como identificar relaciones entre ellos. Enfocándose en el entendimiento de los datos y su procesamiento.

Se requieren de dos etapas para llevarlo a cabo:

- **Ingeniería de datos:** se convierten los datos en crudo en datos preparados para utilizarse en los algoritmos de analítica avanzada
- **Feature engineering:** se seleccionan las features más relevantes y se crean nuevas a partir de los datos.

Entre las herramientas disponibles para el procedimiento de Data Preparation se realizó una selección, para la cual se establece un cuadro comparativo en la Tabla 14 [66, 67, 68, 69, 70]. Las herramientas a analizar son: Apache Mahout, Jupyter Notebook, H20, R Studio, KNIME, RapidMiner, SAS Enterprise Miner y Weka, no trata de ser una lista exhaustiva sino comparativa. Los criterios utilizados se describen al principio de la sección.

	Apache Mahout	Jupyter Notebook	H2O	R Studio	KNIME	RapidMiner	SAS Enterprise Miner	Weka
Enfoque	Se centra en *data clustering*, clasificación, colaboración y filtrado de datos	Es una plataforma web interactiva. Combina código, ecuaciones, visualizaciones y dashboard, entre otros	Plataforma para ejecutar algoritmos de ML en memoria de forma distribuida. AutoML: *training* y *tunning* automático con un límite de tiempo. Generación de *ensembles*. *deploy* de servicios en Spark	Lenguaje de programación tiene buenas librerías para estadística y modelado	Crear flujos de ML end-to-end mediante una interfaz. Los flujos pueden compartirse y ser manejados y hacer el deploy desde la herramienta	Es una herramienta para el proceso *end-to-end* de Data Science, incluyendo preparación de datos, diseño de modelos, *deploy* y manejo de los mismos	Plataforma con interfaz de drag-and-drop. El proceso de minería de datos de SAS abarca el muestreo, exploración modificación modelado y evaluación (SEMMA)	Interfaz UI que contiene una colección de algoritmos que aplicados directamente a los datos o utilizarse desde código.
Open Source	Sí, Apache license	Sí	Sí, Sparkling water	Sí, AGPL v3 License	Sí, KNIME Analytics Platform	Sí, RapidMiner Studio Core	No	Sí, GNU General Public License

	Apache Mahout	Jupyter Notebook	H2O	R Studio	KNIME	RapidMiner	SAS Enterprise Miner	Weka
Performance	Posee algoritmos escalables, aunque su cantidad es limitada	No afecta mucho a la performance. La misma dependerá de cómo esté implementado el algoritmo y la cantidad de nodos/hilos que se utilicen	Tiene buena performance gracias al procesamiento en memoria y la rápida serialización ente nodos y clúster	La performance empora cuando se tienen grandes cantidades de datos	Ofrece extensiones para escalabilidad y performance. Está diseñado para pequeños o medianos negocios	Tiene buena performance con grandes cantidades de datos	Incluye nodos de alto rendimiento de data mining que potencian su performance.	Para lograr buena performance con gran cantidad de datos debe ejecutarse sobre un clúster
Lenguajes soportados	Scala	Python, R, Scala, Julia, Ruby, JavaScript, Perl y PHP	R, Python, Scala y Java	R y Python	Python, R y Java	R y Python	Python y R	Java, Python y R

	Apache Mahout	Jupyter Notebook	H2O	R Studio	KNIME	RapidMiner	SAS Enterprise Miner	Weka
Plataforma	Linux	Cualquier ambiente que tenga instalado python o Anaconda.	Windows, Mac OS X, Docker, Linux, Hadoop, AWS, Azure y Google Cloud	Windows, Linux y MacOS	Windows, Mac OS X y Linux	Windows, Mac, Linux, Cloud, Hadoop, Amazon EMR, Apache, Microsoft's Azure HDInsight, HDP y Cloudera	HP-UX IFS, Linux x64, 64bit enabled AIX, 64bit enabled Solaris, Solaris x64, Windows For Desktop: Windows	Windows, Mac OS X y Linux
Compatibili-dad / Integración	Puede ser utilizada como librería Java	Docker y Kubernetes	Spark, Kafka, Storm, Hive Jupyter, Tableau, Spotfire, HDFS, S3, NFS y SQL	Dataiku Data Science Studio, Pentaho Data Integration, Docker, Kubernetes y Amazon	AWS y Azure	Hadoop y Spark	Spark	Hadoop y Spark

	Apache Mahout	Jupyter Notebook	H2O	R Studio	KNIME	RapidMiner	SAS Enterprise Miner	Weka
Perfil de usuario	Técnico Requiere conocimientos de programación	Técnico Requiere conocimientos de programación	Funcional Configuración de workflows mediante interfaz amigable con Drag & Drop	Técnico Se necesitan conocimientos de programación	Funcional Configuración de workflows mediante interfaz amigable con Drag & Drop	Funcional Configuración de workflows mediante interfaz amigable con Drag & Drop	Funcional Interfaz amigable	Técnico Requiere conocimientos de programación
Gestión de usuarios	No	Con JupyterHub pueden compartirse notebooks entre varios usuarios	Tiene soporte de autenticación, pero no de autorización o ACL	Sí, soporta OAuth2, PAM y LDAP/AD	Se pueden usar los métodos de autenticación de Tomcat, por default se configura una base de datos (H2) basada en autenticación. Integración con LDAP/AD	Sí	Sí, se pueden gestionar grupos y roles	Sí

	Apache Mahout	Jupyter Notebook	H2O	R Studio	KNIME	RapidMiner	SAS Enterprise Miner	Weka
Monitoreo de actividades	No	JupyterHub provee una API REST para exportar métricas	Se pueden monitorear cada uno de los modelos mediante una API REST, obteniendo valores correspondientes a determinadas métricas	Sí, cuenta con una API HTTP para healthchecks	No	Permite: *logging*, *debugging*, *breakpoints* y macros	Sí, se puede monitorear la actividad de los servidores SAS, obteniendo por ejemplo los clientes conectados y logs	Sí, muestra la cantidad de tareas encoladas y en ejecución
Versión	0.13.0[16]	1.0.0[16]	3.26.0.10[16]	1.2.5019[16]	Server 4.8[16]	9.5[16]	15.1[16]	3.8[16]

Tabla 14 Comparativa de Data Preparation en Analítica avanzada

[16] Última versión a Diciembre de 2019.

4.2.4.1.2. Modelado ML

Luego de comprender los datos, es requerido diseñar e implementar modelos de Machine Learning sobre los datos pre-procesados. Por ende, se deben generar algoritmos de inteligencia artificial que provean al sistema de la habilidad de aprender sin la necesidad de un humano.

El modelado se puede realizar utilizando las herramientas vistas en el punto anterior. Aunque existen algunas librerías específicas que suelen ser utilizadas [71, 72, 73, 74]: Scikit-Learn, TensorFlow, Torch / Pytorch, RapidMiner y MLLIB, aunque es importante notar que existen otras alternativas en el mercado. Las funcionalidades estudiadas son las presentadas al principio de la sección.

	Scikit-Learn	TensorFlow	Torch / Pytorch	MLLIB
Enfoque	Librería de Python, es útil para data mining y análisis de datos	Librería para resolver problemas de computación numérica y Machine Learning de gran escala	Librería Python que provee flexibilidad en desarrollos de deep learning	Librería para Apache Spark escalable para Machine Learning
Open Source	Sí	Sí	Sí	Sí
Plataforma	Linux, MacOS y Windows	Linux, MacOS y Windows	Linux, MacOS y Windows	Spark
Características	Clasificación, regresión, clustering, pre-procesamiento de datos, selección de modelos y reducción de dimensionalidad	Provee librerías para manejo de flujos de datos	Módulos: autograd, optim y nn	Permite aplicar algoritmos de ML, Featurization (extraer, seleccionar y transformar variables), pipelines, persistencia y utilities (estadística, manejo de datos, etc.)
Entrenamiento de modelos	Sí	Sí	Sí	Sí
Lenguajes soportados	Python	C++ y Python	Python	Java, Scala, Python y R

	Scikit-Learn	TensorFlow	Torch / Pytorch	MLLIB
Perfil de usuario	Técnico Conocimientos de programación. Curva de aprendizaje sencilla	Técnico Sintaxis fácil de usar. Variedad de tutoriales	Técnico Conocimientos de programación. Curva de aprendizaje sencilla	Técnico Conocimientos de programación
Gestión de usuarios	No	No	No	No
Monitoreo de actividades	No	No	No	No
Versión	0.21.3[17]	2.0[17]	1.3[17]	2.3[17]

Tabla 15 Comparación de modelado ML en Analítica avanzada

[17] Última versión a Diciembre de 2019.

4.2.4.1.3. Herramientas de visualización

Se encuentran dos tipos de visualizaciones principales en cuanto al fin de uso de las mismas,

- **Exploratoria:** sirve para comprender los datos comprende actividades como ser detección de outliers, identificación de límites y umbrales.
- **Explicatoria:** sirve para mostrar los resultados a modo de confirmación, reportes, facilitando la interpretación.

En cuanto a cómo se muestran los datos, visualización puede ser:

- **Estática:** no permite navegación en la visualización.
- **Interactiva:** permite navegar en la visualización, interactuando con los datos.
- **Tiempo real:** los datos que se observan están actualizados.

El objetivo de los visualizadores en ML es presentar los resultados de los algoritmos de ML. Además, no necesariamente trabajan sobre grandes cantidades de datos por lo que no es necesario que las herramientas, como es el caso de Big Data, manejen grandes cantidades de datos.

Existen varias herramientas para conseguir una solución de calidad. Para las dos arquitecturas planteadas se escogieron las herramientas: D3, ECharts, HighCharts, Leaflet, QlikSense, Tableau y PowerBI siendo que esta lista no pretende ser exhaustiva, sino con fines comparativos.

En la Tabla 16 se realiza un comparativo, a nivel de funcionalidades, entre las herramientas elegidas [75, 76, 77]. Los criterios utilizados se describen al principio de la sección.

	D3	ECharts	HighCharts	Leaflet	QlikSense	Tableau	PowerBI
Enfoque	Librería de JavaScript. Altamente configurable a través de CSS	Librería para gráficos en JavaScript	Librería para gráficos en JavaScript	Librería interactiva de mapas para celulares	Plataforma de analítica	Software de gestión empresarial para el análisis de datos	Servicio de análisis empresarial de Microsoft, para visualizaciones interactivas e inteligencia empresarial
Open Source	Sí	Sí	No	Sí	No	No	No
Licenciamiento	Es gratuita	Es gratuita	Es paga	Es gratuita	Si bien es paga, existe una versión Desktop para Windows gratuita	Versión gratuita que permite navegar en tableros ya elaborados y paga que permite conectar a fuentes y generar tableros	Versión gratuita que permite conectar a fuentes y generar tableros Versión con licencia paga que permite distribuir a otros formatos

	D3	ECharts	HighCharts	Leaflet	QlikSense	Tableau	PowerBI
Plataforma	Windows, Linux, MacOS y Docker	Windows, Linux, MacOS y Docker	Browser y Mobile (iOS y Android)	Brower y Mobile (Safari iOS 7+, Android browser 2.2+, 3.1+, 4+, Chrome, Firefox, IE10+ para Win8 devices)	Windows, MAC, iPhone, iPad, Web, Andorid	Web, Mobile (iPhone, Android) La versión Desktop solamente en Windows (64 bit) y MacOS	Windows
Dashboards analíticos	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Exploración interactiva	Sí	Sí	Sí	Sí	Sí	Sí	Sí
Tiempo real		Sí	Sí	Sí	Sí	Sí	Sí
Perfil de usuario	Técnico Requiere conocimientos de programación Curva de aprendizaje alta	Técnico Requiere conocimientos de programación	Técnico Requiere conocimientos de programación	Técnico Requiere conocimientos de programación	Funcional Intuitiva y fácil de usar	Funcional Drag and Drop de dimensiones e indicadores y cuenta con una sintaxis propia para cálculos	Funcional No requiere conocimientos técnicos

	D3	ECharts	HighCharts	Leaflet	QlikSense	Tableau	PowerBI
Gestión de usuarios	Para tener métodos de autenticación se los debe agregar en el header del JSON	No	La versión Cloud permite dos tipos de grupos de acceso: Categorías y gestión de permisos	No	Usa QMC para la gestión de usuarios	El usuario administrador gestiona a los usuarios: agrega nuevos, crea grupos y los agrega a estos, gestiona los roles en el sitio, entre otros	En el portal del usuario administrador se puede gestionar los usuarios
Monitoreo de actividades	No	No	No	No	Mediante QMC se puede monitorear el uso de la herramienta	Genera logs sobre conexión de datos, contenido en el servidor e identificación de usuarios. Se tiene herramientas open source (LogShark y TabMon) para analizarlos	En el portal del administrador se puede ver la actividad de los usuarios y sus grupos

	D3	ECharts	HighCharts	Leaflet	QlikSense	Tableau	PowerBI
Versión	5.14.2[18]	4.5.0[18]	7.2.1[18]	1.6.0[18]	2.2[18]	Desktop 2019.4[18]	Desktop 2.74.5619.621[18]

Tabla 16 Comparación de Visualizadores en Analítica avanzada

[18] Última versión a Diciembre de 2019.

4.2.4.2. Infraestructura

Las características de los componentes de la arquitectura dependen de la cantidad de datos que se desee procesar y el tipo de reportes a realizar. Dependiendo de estas características puede variar la cantidad de nodos, CPU, almacenamiento y si se requiere de un clúster o no. Los componentes destacados son:

- **Almacenamiento:** podría resultar necesario almacenar los datos pre-procesados a ser consumidos por el algoritmo.
- **Plataforma de Infraestructura:** es necesario contar con una plataforma que disponga de herramientas para el procesamiento de datos, y la capacidad de integrar, manejar y transformar estos datos.
- **Datos:** los datos pueden provenir de fuentes variadas, será necesario transformar los mismos y llevarlos al mismo formato, normalizarlos y completar datos faltantes.
- **Procesamiento en paralelo:** puede ser necesaria la paralización del trabajo. Esto dependerá del tipo de análisis que se realice y la cantidad de datos.
- **Visualización:** es importante contar con herramientas para generar la visualización y reportes a partir de los resultados de los algoritmos de Machine Learning. La correcta interpretación de los datos es fundamental para poder generar el valor deseado.
- **Aplicaciones:** se implementan algoritmos, tanto de aprendizaje supervisado como no supervisado, los cuales utilizan estas herramientas para finalmente generar el valor deseado.

La infraestructura de la arquitectura Analítica avanzada es variada, y puede ir desde la ejecución en la máquina local, una máquina virtual o hasta una plataforma de Big Data. Esta decisión depende de las características del análisis que se desee realzar.

Finalmente, si se desea realizar un análisis de datos estático y/o con poco volumen, se sugiere utilizar una arquitectura BI o directamente hacer el análisis sobre planillas. Por otro lado, si se debe trabajar con grandes volúmenes de datos, se recomienda utilizar una arquitectura de Big Data, tomando en cuenta que si debe realizarse trabajos en tiempo real lo más conveniente sería la versión de Big Data en tiempo real.

4.3. Ciberseguridad

La Seguridad de Datos abarca la definición, planificación, desarrollo y ejecución de políticas y procedimientos de seguridad que proveen autenticación, autorización, acceso y auditoría apropiados relativos a datos y activos de información [10]. De la definición anterior, queda claro que el proceso de aseguramiento de datos involucra tareas de distinto tipo. A continuación, se describe el enfoque elegido para llevar a cabo medidas que permitan realizar dicho proceso de forma efectiva, definiéndose a su vez el objetivo del componente y el vínculo que tiene este con el enfoque elegido.

En esta sub sección se describe la estructura del componente de Ciberseguridad del Modelo desarrollado, que clasifica los lineamientos definidos en tres niveles: **prevención**, **detección y control**, y **resiliencia**. Se describen las características generales de cada uno de estos niveles y se presentan los beneficios que aporta la clasificación en el abordaje de la Ciberseguridad y la seguridad de los datos.

En el resto de la sección, se motiva la elección de la categorización propuesta, profundizando los objetivos de cada categoría y del enfoque de riesgos sobre el que se apoya la misma.

Como se explicó en la Sección de Antecedentes, los últimos años se han visto caracterizados por la interconexión de sistemas y el uso generalizado de tecnología, demandando esto que la ciberseguridad sea abordada con un enfoque de riesgo, considerando factores tanto internos como externos de las organizaciones.

A nivel de la estructura organizacional, este cambio de paradigma se está viendo reflejado en la incorporación de la figura del CISO (Chief Information Security Officer, por sus siglas en inglés) o RSI (Responsable de Seguridad de la Información), que gestiona de forma integrada las medidas de ciberseguridad con perspectivas de riesgo en la organización, concientizando a los distintos sectores sobre la relevancia del correcto abordaje de la temática.

Las medidas adoptadas por cada organismo para manejar los riesgos de ciberseguridad deben ser continuamente adaptadas en función de las características y necesidades particulares de cada una, para que los riesgos se gestionen de forma proactiva, así como reactiva. Por este motivo se presentan las recomendaciones de este componente en los tres niveles antes mencionados.

La Figura 21 ilustra las distintas categorías junto con sus objetivos, que aporta al abordaje integral de los riesgos en la organización.

Figura 21 Categorías de recomendaciones con sus objetivos

Las recomendaciones contenidas en el componente de **prevención** están orientadas a aumentar el nivel de protección del sistema, particularmente de las partes que involucran un riesgo elevado[19]. A mayor nivel de seguridad, más tiempo y recursos necesitará un atacante para realizar un ataque de forma exitosa.

Las medidas de **detección y control**, buscan disminuir el tiempo entre que un atacante compromete un sistema y la víctima toma conciencia de ello. Finalmente, luego de que se tomó conocimiento de que el sistema ha sido atacado con éxito, es necesario adoptar acciones para contener el incidente y recuperarse del mismo, cubiertas por los lineamientos de **resiliencia**.

Esta definición del modelo por componentes permite concentrarse en cada una de las etapas fundamentales del proceso de aseguramiento de activos, datos e información con un enfoque integral de riesgos, logrando que el mismo sea robusto.

4.4. Legal

Desde el punto de vista normativo, Uruguay ha sancionado una serie de leyes que regulan la generación y el tratamiento de los datos, entre las que se pueden destacar como más relevantes las siguientes:

- **Ley 17.930:** Presupuesto nacional de sueldos gastos e inversiones. Ejercicio 2005 – 2009. Esta norma, en lo específico, creó AGESIC.
- **Ley 18.331:** Ley de protección de datos personales. Se establece un sistema general para el manejo de los datos personales y de las bases de datos con datos personales.
- **Ley 18.381:** Ley sobre el derecho de acceso a la información pública. Normativa que promueve la transparencia de la función administrativa de todo organismo público, sea o no estatal, y garantiza el derecho fundamental de las personas al acceso a la información pública.
- **Ley 18.719:** Presupuesto nacional de gastos e inversiones. Ejercicio 2010 – 2014. Se creó la Dirección de Seguridad de la Información en AGESIC, que alberga el Centro Nacional de Respuesta a Incidentes de Seguridad Informática (CERTuy) y se dispuso la obligación de las entidades públicas de adoptar medidas y cumplir determinados principios para promover el intercambio de información pública o privada autorizada por el titular, disponible en medios electrónicos.
- **Ley 19.172:** Regulación y control de cannabis. En esta normativa se creó un registro de usuarios, disponiéndose la protección de la identidad de los mismos, la cual es considerada como dato sensible.
- **Ley 19.355:** Presupuesto nacional de gastos e inversiones. Ejercicio 2015 – 2019. Se dispuso la obligación de las entidades públicas de publicar en formato abierto cierta información, según las normas técnicas que determine AGESIC y del Registro Nacional de Antecedentes Judiciales de comunicar sus datos a las autoridades

[19] Un riesgo puede definirse como la probabilidad de que un incidente ocurra y las consecuencias que este pueda tener sobre un activo. El nivel del riesgo se suele establecer en función de su probabilidad de ocurrencia y su impacto. Para profundizar en el tema consultar [90].

judiciales en materia penal exclusivamente con el fin de comprobar la reincidencia. Se facultó al Poder Ejecutivo a determinar los mecanismos de intercambio de información clínica con fines asistenciales (Sistema de Historia Clínica Electrónica Nacional).

5. Aplicación

En la presente sección se brindan recomendaciones prácticas para cada una de las etapas del proceso de análisis antes definido: Necesidad – Búsqueda de datos – Preparación de los datos – Modelado y Análisis de datos – Presentación – Resultados / Acciones. Dichas prácticas se describen para cada uno de las dimensiones.

Para esto, se desarrolló un checklist en formato de tabla con las siguientes columnas, desarrollándose luego cada uno de los puntos:

- **REF:** indica el código de la buena práctica, está formado por la concatenación del prefijo y un número, el prefijo respeta lo siguiente:
  - Gobernanza: se utiliza la letra 'G'.
  - Tecnología: se utiliza la letra 'T'.
  - Ciberseguridad: el prefijo dependerá de la clasificación de las medidas a recomendar,
    - Prevención: se utiliza la letra 'P'.
    - Detección y control: se utilizan las letras 'DC'.
    - Resiliencia: se utiliza la letra 'R'.
- **Nombre:** nombre de la buena práctica a definir.
- **Descripción:** detalle de la buena práctica a definir.
- **Necesidad, Búsqueda de datos, Preparación de los datos, Modelado y análisis de datos, Presentación y Resultados/Acciones:** representa cada una de las etapas del proceso de análisis de datos. Si en la celda se tiene un asterisco, significa que la buena práctica aplica en la etapa del proceso.

Luego, se desarrollan cada uno de las buenas prácticas, especificando las etapas del análisis de datos (AD) a tomar en cuenta.

5.1. Gobernanza

Las siguiente sub secciones presentan las buenas prácticas del componente de Gobernanza.

5.1.1. Checklist

En la Tabla 17 se presentan las recomendaciones y buenas prácticas del componente de Gobernanza el formato de la tabla se explica al principio de la sección.

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
G1	Control de cambios	Generar una metodología de control de cambios y usarla cuando lo crean necesario.		*	*	*	*	*
G2	Metodología ágil	Aplicar una metodología ágil para así poder validar en etapas tempranas lo desarrollado.		*	*	*	*	
G3	Criterios de éxito	Definir los criterios de éxito como SMART.	*					
G4	Solicitud de Análisis de Datos	Claridad en la Solicitud de Análisis de Datos.	*					
G5	Fuentes de origen de los datos	Las fuentes deben estar validadas		*				
G6	Selección de datos	Seleccionar los datos correctos para el análisis actual y pensar en posibles datos necesarios para futuros análisis.		*				
G7	Plan de calidad de datos	Implementar un plan de calidad de datos.			*			
G8	Precisión de los datos	Verificar la precisión de los datos.			*			
G9	Valores atípicos (outliers)	Identificar valores atípicos.			*			
G10	Valores duplicados	Identificar valores duplicados.			*			

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
G11	Validar los datos	Validar la precisión de los datos.			*			
G12	Identificar datos faltantes	Realizar un tratamiento a los datos faltantes.			*			
G13	División de muestras train – test	20% test, 80% train			*			
G14	Datos de referencia	Identificar catálogos ya existentes para datos maestros				*		
G15	Estándares de industria	Utilizar indicadores estándar de la industria.				*		
G16	Modelos multidimensionales	Utilizar metodologías y buenas prácticas para el diseño de modelos multidimensionales.				*		
G17	Datos históricos (modelos predictivos)	En modelos predictivos utilizar varias variables de diferentes tipos.				*		
G18	Pruebas de Significación	Realizar pruebas de significación para realizar una mejor evaluación del modelo.				*		
G19	Entrenamiento de los modelos	Tener en cuenta el riesgo de overfitting.				*		
G20	Monitoreo constante	Realizar un monitoreo constante de los modelos para corregir sesgos.				*		

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
G21	Machine Learning	Utilizar métodos de reducción de dimensiones y técnicas para evitar el sobreajuste.				*		
G22	Audiencia	Conocer a la audiencia.					*	
G23	Contar una historia	Guiar la atención de la audición a lo que se desea resaltar.					*	
G24	Colores en los tableros	Usar colores y fuentes consistentes.					*	
G25	Diseño de tableros	Mostrar sólo los detalles necesarios.					*	
G26	Jerarquía visual	Respetar una jerarquía visual.					*	
G27	Leyendas	Colocar las leyendas estratégicamente.					*	
G28	Orden lógico y estructuras conceptuales	Generar un orden lógico y respetar estructuras conceptuales.					*	
G29	Informar disponibilidad del análisis	Informar a otros posibles interesados sobre la disponibilidad del análisis.						*
G30	Capacitar a usuarios finales	Capacitar a los usuarios finales en el uso de las herramientas a disposición.					*	*
G31	Gestión del conocimiento	El conocimiento adquirido de poder ser transferido a otras personas.						*

Tabla 17 Checklist de buenas prácticas y recomendaciones de Gobernanza

5.1.2. Buenas prácticas

En las siguientes subsecciones se desarrollan los puntos vistos en la sección anterior.

G1. Control de cambios

Se sugiere generar una metodología de control de cambios para supervisar las solicitudes de cambio, aprobar aquellos cambios que se consideren convenientes y gestionar la implementación de esos cambios.

Etapas del AD a tomar en cuenta: Búsqueda de datos, Preparación de los datos, Modelado y análisis de datos, Presentación, Resultados / Acciones.

G2. Metodología ágil

Se recomienda aplicar metodologías ágiles a la hora de realizar el análisis de los datos con el cometido de validar los avances realizados en etapas tempranas. En todo momento se debe contar con la participación de los interesados para que aporten su experiencia y revisen los hallazgos intermedios.

Un ejemplo de es utilizar un proceso Scrum, en el que se aplican un conjunto de buenas prácticas para trabajar de forma colaborativa y obtener, lo antes posibles, resultados positivos. Por más información sobre Scrum ver [103].

Etapas del AD a tomar en cuenta: Búsqueda de datos, Preparación de los datos, modelado y Análisis de datos, Presentación.

G3. Criterios de éxito

Se deben definir los criterios de éxito como SMART, esto significa que deben cumplir con las siguientes características:

- Específicos,
- Medibles,
- Alcanzables,
- Relevantes y
- Realizarse en un tiempo límite

Etapas del AD a tomar en cuenta: Necesidad.

G4. Solicitud de análisis de datos

Utilizar un template de Solicitud de análisis de datos es una buena práctica, ya que en el mismo se detallan objetivos, características de los datos necesarios, fuentes de datos, etc.

Si bien es una práctica recomendada, no es imprescindible, quizás es más realizable en los organismos que cuenten con una madurez importante en el análisis de datos.

A modo de ejemplo se presenta un template genérico de una solicitud de pedido de información:

Template solicitud de pedido de inforn

Etapas del AD a tomar en cuenta: Necesidad.

G5. Fuentes de origen de los datos.

Utilizar la o las fuentes de origen de los datos lo más validadas posibles. Por ejemplo: las bases de datos de DNIC para los nombres completos de las personas.

Etapas del AD a tomar en cuenta: Búsqueda de datos.

G6. Selección de datos

Se trata de encontrar los datos más adecuados para el propósito del análisis. Es importante tener en cuenta que esto no solo debe tratarse de encontrar los datos que necesita en este momento, se debe tener una actitud proactiva y pensar los datos que se pueden necesitar a futuro, cuando surjan necesidades similares.

Etapas del AD a tomar en cuenta: Búsqueda de datos.

G7. Plan de calidad de datos

Se sugiere crear un plan y una estrategia de limpieza de datos definiendo expectativas y estándares mínimos para los datos. Crear indicadores clave de rendimiento (KPI) de calidad de datos. Algunos ejemplos de indicadores clave de rendimiento de calidad de datos son: ratio de datos erróneos, ratio de datos duplicados o número de valores vacíos.

Tener en cuenta dónde ocurren la mayoría de los errores de calidad de datos. Identificar datos incorrectos y comprender la causa raíz del problema de datos.

Etapas del AD a tomar en cuenta: Preparación de los datos.

G8. Precisión de los datos

Debe asegurarse que los datos sean precisos y tengan sentido. La realización de esta tarea es importante y requiere cierto conocimiento del área temática con la

que está relacionado el conjunto de datos por lo cual es deseable consultar al encargado de dominio correspondiente.

Si bien no existe un enfoque específico para verificar la precisión de los datos, algunas acciones que se realizan habitualmente son el perfilamiento de los datos o la comparación de distintas fuentes. La idea básica es formular algunas propiedades que cree que deberían mostrar los datos, y probar los datos para ver si esas propiedades están satisfechas.

Etapas del AD a tomar en cuenta: Preparación de los datos.

G9. Valores atípicos (outliers)

Los outliers son puntos de datos que están distantes del resto de la distribución. Son valores dispares en comparación con el resto del conjunto de datos.

Es necesario estudiar los outliers en cada caso en particular, sobre los detectados se debe determinar:

- Si es realmente un outilier o si es un error en los datos.
- Decidir si eliminarlo o no. Si no se sabe si se lo debe eliminar, se puede entrenar el modelo con y sin outliers y comparar los resultados, y a partir de esto tomar la decisión.

Etapas del AD a tomar en cuenta: Preparación de los datos.

G10. Valores duplicados

Los datos duplicados causan informes inexactos. Se debe asegurar la entrada de datos en buen estado, valide y elimine cualquier duplicado.

Etapas del AD a tomar en cuenta: Preparación de los datos.

G11. Validar los datos

Se deben validar los datos para asegurarnos de que cumplan con los estándares requeridos. Si no es así, debemos alertar al equipo técnico o incluso arreglarlo en el acto.

Uno de los propósitos de la validación de datos es evaluar la precisión y consistencia de los datos capturados. La precisión y la consistencia solo se pueden medir comparando los datos con otra fuente precisa. Esta fuente debe ser correcta, de lo contrario, no tenemos forma de saber que los nuevos datos también son precisos.

Si se trata de un gran conjunto de datos, desarrolle un script o enfoque que pueda validar un pequeño conjunto de datos a la vez. Esto es mucho más fácil de escalar que intentar arreglar un conjunto de datos completo al mismo tiempo.

Se debe llevar un registro de trazabilidad sobre los cambios realizados para poder identificar cuando han sido realizados.

Etapas del AD a tomar en cuenta: Preparación de los datos.

G12. Identificar datos faltantes

Los datos faltantes o los valores faltantes se producen cuando no se almacena ningún valor de datos para la variable en una observación. Por eso pueden causar un riesgo potencial para análisis que se va a realizar. Probablemente sean uno de los problemas de datos más habituales.

Para poder identificarlos es necesario tener conocimiento del negocio y tener la habilidad de inferir cuando se trata de un faltante o no.

Una vez identificados se deben resolver estos faltantes. El método a utilizar va depender del análisis a realizar, pero se puede:

- Reemplazar los valores faltantes con un valor apropiado, o
- Eliminar la fila / registro.

Etapas del AD a tomar en cuenta: Preparación de los datos.

G13. División de muestras train – test

Por lo general, se dividen los datos entre un 20% y un 80% entre las etapas de test y las de train.

Etapas del AD a tomar en cuenta: Preparación de los datos.

G14. Datos de referencia

Cuando el modelo de análisis incluye datos de referencia (catálogos conocidos) se debe consultar y acceder a la fuente de información oficial, ya sea que esté administrada por un sistema o un área.

Por ejemplo: si una de las dimensiones de análisis son las oficinas del organismo, tomar la fuente oficial (consultar a recursos humanos o a una fuente externa de referencia).

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

G15. Estándares de industria

Si existen indicadores con estándares de la industria se pueden usar para la transformación de datos y también considerarlos como una opción para evaluar su uso.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

G16. Modelos multidimensionales

Previo a la construcción del modelo para análisis de información se recomienda traducir el pedido de información en términos de elementos de un modelo multidimensional como ser Dimensiones, Medidas o Indicadores y Jerarquías de modo de poder solicitar los datos con mayor claridad considerando los posibles requerimientos de cruces, granularidad, filtros y presentación de datos.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

G17. Datos históricos (modelos predictivos)

Cuando se realicen modelos predictivos se deben usar datos históricos de varias variables de diferentes tipos: provenientes de sistemas transaccionales, variables demográficas y hasta climáticas.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

G18. Pruebas de Significación

En algunos casos se aplican pruebas de significación además de los indicadores globales de los modelos, para tener una mejor evaluación y justificar la implementación o adopción de medidas de gran impacto.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

G19. Entrenamiento de los modelos

Entrenar los modelos teniendo en cuenta el riesgo de overfitting y validando los mismos antes de ponerlos en producción.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

G20. Monitoreo constante

Es necesario hacer un monitoreo constante ya que los sesgos en los datos generados por el desbalanceo en las poblaciones que impactan en los modelos pueden ajustarse a medida que se va aplicando en producción y se aproxima a los valores poblacionales de las variables.

La construcción y evaluación de los modelos se enriquece por algoritmos específicos para grandes volúmenes de datos y por modelos alternativos a los clásicos que aportan mejoras, como por ejemplo SVM o random forest.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

G21. Machine Learning

En Machine Learning se tienen problemas comunes: multicolinealidad (multicollinearity) y sobreajuste (overfitting).

El problema de **multicolinealidad** se da cuando las variables tienen correlación alta entre sí. Esto puede causar problemas a la hora de interpretar los resultados. Para resolver esto, se aplican métodos de reducción de dimensiones, para así reducir la cantidad de variables y quedarse con las que más influyen en la variable objetivo. También simplifica el modelo, lo que hace más fácil de interpretar y más rápido el entrenamiento.

El **sobreajuste** [101] se da cuando se sobre entrena al modelo con casos particulares, por lo que no será capaz de reconocer nuevos datos de entrada que no tengan las mismas características.

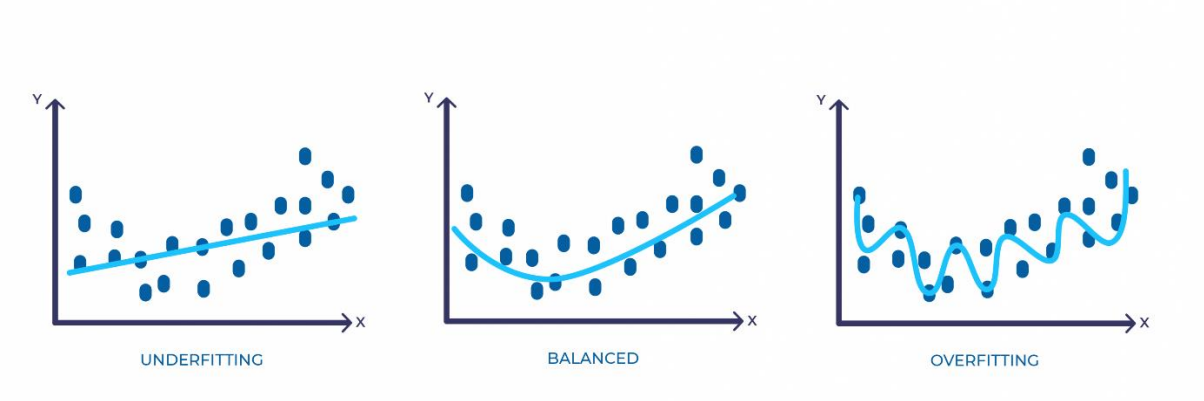

Figura 22 Ajustes del modelo [101]

Como se pude ver en los gráficos cuando un modelo está subajustado no puede capturar la relación entre los ejemplos de entrada y los valores objetivo. Por el contrario, cuando está sobreajustado funciona bien en los datos de entrenamiento, pero no en los de evaluación. Esto se debe a que el modelo está memorizando los datos que ha visto y no puede generalizar a ejemplos no vistos.

En general, se espera que el error del training set y el validation estén cerca, si bien el correspondiente al validation será más grande, la diferencia entre ambos debería ser chica a medida que el algoritmo aprende. Dicho de otro modo, en la Figura 22 se espera que la línea roja (validation) y la azul (training set) estén cerca y a medida que aumente la cantidad de iteraciones, se acerquen.

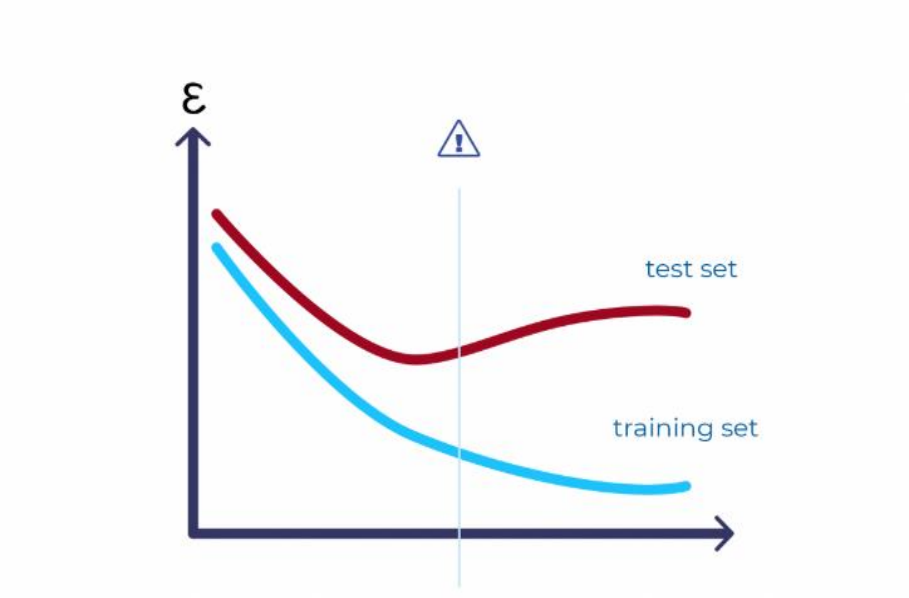

Figura 23 Overfitting. Fuente: [102]

Existen varias técnicas para evitar el sobreajuste:

- **Oversampling:** se usa cuando los datos no poseen muchos casos de clase 1. Cuando se hace le muestra se lo fuerza a que tenga datos de clase para asegurarse de que el modelo se va a entrenar con esa clase.
- **Stratified sampling:** hacer la muestra de manera que contenga la misma proporción sobre los datos para determinadas columnas.
- **Early stopping:** para que el modelo no se entrene con los detalles de los datos del training set, lo detiene antes.
- **Train-validation-test sets:** en lugar de utilizar solamente training y test, se utiliza el validation set para corregir errores en las predicciones producidos durante el training. El test set se utiliza para testear el modelo al final.
- **Cross-validation:** utiliza train-validation-test, alternando los datos con los que se entrena.
- **Regularization:** penaliza los parámetros para reducir "la importancia".
- **Pruning:** se "poda" el árbol en el entrenamiento. Solamente es aplicable a CART (Classification And Regression Trees).

Etapas del AD a tomar en cuenta: Necesidad, Búsqueda de datos, Preparación de los datos, Modelado y análisis de datos.

G22. Audiencia

Una visualización de datos es inútil si no está diseñada para comunicarse claramente con el público objetivo. Debe ser compatible con la experiencia del público y permitir a los espectadores ver y procesar datos de manera fácil y rápida. Tenga en cuenta qué tan familiarizado está el público con los principios básicos que presentan los datos.

Se recomienda la realización de un bosquejo del modelo para cada público objetivo y mostrarlo a grupos representativos para su validación.

Las diferentes características visuales funcionan mejor con diferentes tipos de datos. Por ejemplo, los gráficos de dispersión funcionan bien con dos datos cuantitativos, mientras que los gráficos de líneas funcionan mejor para datos ordinales de fecha, por el contrario, los gráficos de líneas son una mala elección para los datos categóricos (no ordinales) ya que los gráficos de líneas implican continuidad. Asegúrese de saber qué datos utilizará su imagen.

Se presenta una imagen que sirve de guía para la elección de la forma de mostrar los datos.


guía.PNG

Etapas del AD a tomar en cuenta: Presentación.

G23. Contar una historia

Storytelling es una metodología para comunicar información, adaptada a una audiencia específica, con una narrativa convincente.

Algunas de las prácticas más comunes son:

- La historia debe tener un camino narrativo claro que incluya exposición, acción ascendente, clímax, acciones descendentes y denuncia. Sin datos que respalden su historia, se desmoronará.
- Colocar la información más importante en la parte superior del tablero en general, o en la parte superior izquierda. Cuando se mira un tablero, el ojo típicamente va a la parte superior izquierda primero.
- Si está utilizando un gráfico de barras para mostrar la diferencia, generalmente está bien comenzar el eje en cero. Colocar el eje en cero le permite mostrar el tamaño completo de la barra. En el caso de un gráfico lineal, el objetivo es mostrar el aumento o la disminución con el tiempo y no el tamaño absoluto, es una práctica aceptada para no colocar el eje en cero. Si está mostrando múltiples tendencias usando un gráfico lineal, entonces el eje debe ser consistente.

Etapas del AD a tomar en cuenta: Presentación.

G24. Colores en los tableros

Evitar el uso de múltiples diseños de color en un tablero a menos que haya diseños naturales e independientes en los datos. Basarse en colores del organismo.

Se recomienda la utilización de hojas de estilo para poder definir un mismo aspecto en todas las visualizaciones que se desean realizar.

Etapas del AD a tomar en cuenta: Presentación.

G25. Diseño de tableros

Realizar un balance entre un diseño simple y la complejidad de las métricas y análisis. Limitar la cantidad de indicadores en un tablero a 3 o 4. Si se agregan demasiados indicadores, se pierde el panorama general de todos los detalles.

Etapas del AD a tomar en cuenta: Presentación.

G26. Jerarquía visual

Respetar una jerarquía visual que permita ver primero los principales bloques de información e ir guiando a lo que interesa mostrar.

Etapas del AD a tomar en cuenta: Presentación.

G27. Leyendas

Si una leyenda aplica a todas las hojas, se deben colocar juntas con todos los filtros. Si una leyenda aplica a una o varias secciones, se debe colocar cerca de las mismas.

Se pueden indicar aspectos como las fechas de corte, datos de la muestra y de los indicadores.

Etapas del AD a tomar en cuenta: Presentación.

G28. Orden lógico y estructuras conceptuales

Crear filtros de acción que presenten datos en un orden lógico y respetando las estructuras conceptuales, por ejemplo: organismo y unidad ejecutora.

Etapas del AD a tomar en cuenta: Presentación.

G29. Informar disponibilidad del análisis

Es importante comunicar al resto de los interesados o posibles interesados de la disponibilidad de la información y posibles análisis para que en caso corresponda puedan reutilizar este activo de información generado.

Las políticas de gobierno alientan la tendencia a disponibilizar todos los datos.

Etapas del AD a tomar en cuenta: Resultados / Acciones.

G30. Capacitar usuarios finales

Es importante realizar capacitaciones a los usuarios finales para asegurarse de que sepan utilizar las herramientas e interpretar los resultados para hacer un uso correcto de la información brindada.

Las herramientas deben estar disponibles al público objetivo para poder entender los informes de mejor manera. Es deseable además que exista documentación disponible para los usuarios.

Etapas del AD a tomar en cuenta: Presentación, Resultados / Acciones.

G31. Gestión del conocimiento

Es necesario que el conocimiento obtenido pueda ser transferido a otras personas, por eso es indispensable que por ejemplo se guarde la documentación de los análisis realizados y tratar de conseguir que todos los análisis sean reproducibles.

Etapas del AD a tomar en cuenta: Resultados / Acciones.

5.2. Tecnología

La presente sección tiene las buenas prácticas del componente de Tecnología.

5.2.1. Checklist

En la Tabla 18 se presentan las recomendaciones y buenas prácticas del componente de Tecnología [49, 50], el formato de la tabla se explica al principio de la sección.

Cabe destacar que las dos últimas buenas prácticas (T11 y T12), refieren a casos particulares. Los mismos se detallan porque se considera que se agrega valor en el análisis.

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
T1	Elección de la arquitectura	Para la elección de la arquitectura tomar en cuenta: volumen de datos, necesidad de tiempo real en ingesta y visualizador, disponibilidad de hardware, frecuencia de los análisis	*					
T2	Almacenamiento de datos	Diseñar e implementar el almacén de datos según la arquitectura elegida	*					
T3	Hardware	Analizar el hardware a utilizar, por ejemplo: si es Hadoop se debe analizar la distribución de los recursos para los servicios instalados	*					
T4	Crear metadata	Generar metadata del DW para así explicar los datos			*	*		
T5	Repositorio de código	Utilizar un repositorio de datos (ej.: git) para almacenar el código generado	*	*	*	*	*	*
T6	Compatibilidad de versiones	Se debe tener en cuenta que las versiones instaladas o a instalar, deben ser compatibles	*	*	*	*	*	
T7	Herramientas para gestión de usuarios	Herramientas para gestión de usuarios	*	*	*	*	*	*

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
T8	Herramientas para Monitoreo de actividades	Investigar herramientas para dar soporte al monitoreo de las actividades de los usuarios en el sistema	*	*	*	*	*	*
T9	Calidad en las ingestas	Las ingestas deben tomar en cuenta criticidad, privacidad y durabilidad de los datos		*	*			
T10	Optimizar consultas	Particionar las datos con el objetivo de optimizar las consultas y obtener una mejor performance			*	*		
T11	Big Data – Hadoop	En caso de utilizar Hadoop, se debe tener en cuenta consideraciones especiales Ver Anexo	*	*	*	*		
T12	Analítica avanzada	En caso de una arquitectura de Analítica avanzada, se debe tomar en cuenta consideraciones especiales			*	*		

Tabla 18 Checklist de buenas prácticas y recomendaciones

5.2.2. Buenas prácticas

En las siguientes subsecciones se desarrollan los puntos vistos en la sección anterior., según la dimensión de Tecnología.

T1. Elección de la arquitectura

Un análisis exitoso depende de una correcta elección de la arquitectura a elegir. Para seleccionar una de las cuatro arquitecturas presentadas en la sección anterior, se deben responder distintas preguntas.

A continuación se presentan las preguntas mínimas:

- ¿Se manejan grandes volúmenes de datos?: si se debe tratar un gran volumen de datos se recomendaría optar por la arquitectura de Big Data. La modalidad dependerá de la respuesta a la pregunta: ¿es necesaria la ingesta y visualización en tiempo real?
- ¿Es necesaria la ingesta y visualización en tiempo real?: si es una respuesta afirmativa y se tiene gran volumen de datos, se debería considerar optar por Big Data en tiempo real.
- ¿Se desea encontrar patrones en los datos?: en caso de que así sea, una arquitectura de analítica avanzada podría ser la mejor opción.
- ¿Se utilizan tipos de datos estructurados?: si es así, se debería optar por una arquitectura de Big Data o Analítica avanzada.

Etapas del AD a tomar en cuenta: Necesidad.

T2. Almacenamiento de datos

Se debe diseñar e implementar el almacén de datos según la arquitectura elegida. Descubrir un error en etapas posteriores, puede influir de forma negativa.

Para el modelado del data warehouse existen dos grandes enfoques [98, 99]: Kimball utiliza una metodología Bottom-Up e Inmon utiliza una **Top-down**.

Etapas del AD a tomar en cuenta: Necesidad.

T3. Hardware

Analizar el Hardware a utilizar, ya sea si se tiene o si se tiene que conseguir. Esto pude limitar la arquitectura a utilizarse.

Por ejemplo, para el caso de arquitectura Big Data se debe analizar, entre otras cuestiones, la distribución de los recursos para los servicios instalados.

Etapas del AD a tomar en cuenta: Necesidad.

T4. Crear metadata

Para el éxito de un DW, es fundamental la capacidad de explicar los datos. Para esto, se debe generar metadata como parte de ciclo de desarrollo y administrarse como parte de las operaciones en marcha.

Un ejemplo de esto es la creación de metadata para las personas [100], en donde se busca normalización la información de las personas, permitiendo su identificación de forma inequívoca.

Etapas del AD a tomar en cuenta: Transversal.

T5. Repositorio de código

Es una buena práctica utilizar un repositorio de datos, como por ejemplo GIT, para almacenar el código generado y así se pueda gestionar el control de las versiones.

Etapas del AD a tomar en cuenta: Transversal.

T6. Compatibilidad de versiones

Se debe tener en cuenta que las versiones instaladas de los distintos componentes y librerías sean compatibles con el caso de uso a implementar y las nuevas librerías a instalar.

Etapas del AD a tomar en cuenta: Transversal.

T7. Herramientas para Gestión de usuarios

Es requerido tomar medidas de seguridad contra el acceso de los usuarios a los sistemas. Para esto se debe gestionar el acceso de los usuarios a las herramientas, generando permisos y grupos según sus funcionalidades. Por lo tanto, es necesario analizar herramientas que permitan hacerlo.

Por más información ver la buena práctica DC de la dimensión de Ciberseguridad.

Etapas del AD a tomar en cuenta: Necesidad.

T8. Herramientas para el Monitoreo de actividades

Con el fin de tener consciencia sobre el uso que los usuarios les dan a las herramientas, se debe tener un monitoreo sobre sus actividades, pudiendo identificar qué usuario realizó qué actividad. Por lo que es vital, analizar herramientas que permitan dicho monitoreo.

Etapas del AD a tomar en cuenta: Necesidad.

T9. Calidad en las ingestas

Las ingestas de datos deben tomar en cuenta la criticidad, privacidad y durabilidad de los datos.

Etapas del AD a tomar en cuenta: Búsqueda de datos, Preparación de los datos.

T10. Optimizar consultas

Las consultas sobre los datos se deben particionar de forma inteligente, para así obtener una mejor performance. En caso contrario el sistema puede degradarse y ser afectado considerablemente.

Por ejemplo, en bases de datos relacionales las consultas se pueden mejorar con la gestión de índices [108]. Sino, en caso de utilizar Hadoop se pueden realizar particiones guardando los datos en subdirectorios categorizados por los valores de una columna, para profundizar sobre el manejo de particiones ver [109].

Etapas del AD a tomar en cuenta: Preparación de los datos, Modelado y análisis de datos.

5.2.2.1. Casos particulares

En la presente sección se presentan buenas prácticas para arquitecturas en específico.

T11. Big Data - Hadoop

En caso de utilizar Hadoop en arquitecturas de Big Data, se debe validar con un experto la arquitectura de la plataforma de Hadoop, analizando la distribución de los recursos para los servicios instalados.

En cuanto al ambiente y ejecución, se debe tener en cuenta:

- Disponer de un nodo edge para el desarrollo del caso de uso. Y un usuario con permisos necesarios para el desarrollo del mismo.
- El nodo edge deberá contar con clientes HDFS, Hive, Spark, etc.
- El nodo deberá contar con acceso a un repositorio GIT donde se almacenará el código del caso de uso y sus versiones.
- Gestión de colas de YARN: utilizar colas de trabajo para la asignación de recursos.

Etapas del AD a tomar en cuenta: Necesidad, Modelado y análisis de datos.

T12. Analítica avanzada

Se pueden imputar los campos con posibles valores según los siguientes métodos:

- **Cálculos:** los posibles valores pueden generarse de varias formas a partir de los datos disponibles. Tomando el promedio, moda, random.
- **Valores por defecto:** pueden completar con un valor por defecto.
- **Simulación:** pueden simularse utilizando alguna distribución. Por ejemplo, triangular determinados valores correspondientes a mínimo, máximo y moda. O una distribución normal especificando el promedio.
- **Join con otro dataset:** buscar otro dataset que pueda utilizarse para completar los datos faltantes.

La decisión de usar un método u otro, depende del significado de los datos. El objetivo es intentar reproducir los datos reales de la forma más exacta posible.

Etapas del AD a tomar en cuenta: Preparación de los datos.

Por otro lado, se debe tomar en cuenta:

- **Comparación de modelos:** probar varios modelos y comparar su performance. La comparación se establece en base a métricas estadísticas, entre las más comunes se encuentran:
  - RMSE (Root Mean Square Error) [104].
  - R-squared (R2) [105].
  - Accuracy [106].
- **Ensemble:** depende de la complejidad del estudio que se esté llevando a cabo, en ciertos casos es necesario aplicar técnicas de ensemble para conseguir mejores resultados. Dichas técnicas, combinan varios modelos para generar el resultado final. Se tienen los siguientes tipos [107]:
  - Voting (clasificación) / Promedio (Predicción).
  - Bagging: entrena en paralelo el mismo modelo sobre diferentes trainings sets y después se aplica voting o promedio.
  - Boosting: entrena en secuencia el mismo modelo haciendo resampling y agregando los registros que fueron mal clasificados en el paso anterior.
  - Stacking: entrena sobre los mismos datos diferentes modelos.
  - Monitorear el algoritmo: es necesario monitorear los resultados del algoritmo.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos.

5.3. Ciberseguridad

En la presente sección se detallan las buenas prácticas del componente de Ciberseguridad.

Como base general para el abordaje de la ciberseguridad en el proceso de Análisis de Datos, se recomienda tener en cuenta las distintas buenas practicas propuestas en el "**Marco de Ciberseguridad**" elaborado por AGESIC. Como se mencionó en la Sección de Antecedentes, las recomendaciones expuestas toman como base la segunda edición del "Data Management Body of Knowledge" [10], con particular énfasis en el capítulo "Data Security" (Seguridad de Datos), así como recomendaciones dadas por OWASP en 2019 [91].

Es importante tener en cuenta que las medidas presentadas aplican a todos los ambientes utilizados en el proceso de Análisis de datos, incluyendo los de prueba.

5.3.1. Checklist

En la Tabla 19 se presentan las recomendaciones y buenas prácticas del componente de Ciberseguridad, el formato de la tabla se explica al principio de la sección.

BUENAS PRÁCTICAS			PROCESO DE ANÁLISIS DE DATOS					
REF	Nombre	Descripción	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / Acciones
P1	**Identificar tipo de información**	Identificar Modelo de Datos Organizacional y si se está trabajando con información crítica en el análisis.			*			
P2	**Fuentes de Requerimientos**	Considerar partes interesadas, regulaciones gubernamentales, información crítica y reglas de acceso, así como obligaciones contractuales.	*	*	*	*	*	
P3	**Clasificación de información y preprocesamiento**	Verificar que los datos a ser utilizados en el proceso de análisis se encuentran categorizados.			*			
P4	**Enmascaramiento de Datos**	Considerar el uso de técnicas de enmascaramiento u ofuscación de datos.			*		*	
P5	**Métodos y canales de transferencia**	Evaluar el método y canales utilizados para realizar la transferencia de datos.	*	*	*	*	*	*
P6	**Procedencia de los datos**	Evaluar la procedencia de las fuentes de datos utilizadas. Verificar que se encuentren permitidas.	*	*				

BUENAS PRÁCTICAS			PROCESO DE ANÁLISIS DE DATOS					
REF	Nombre	Descripción	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / Acciones
P7	Conexiones y necesidades de actualización	Para las fuentes que requieran conexión, conectarse por el menor tiempo necesario y su periodicidad..		*	*	*	*	*
P8	Definir roles y privilegios	Definir roles, privilegios y responsabilidades asociados a las herramientas utilizadas.	*	*	*	*	*	*
P9	Gestión centralizada de identidades	Gestionar las identidades de los usuarios del sistema y la pertenencia a los distintos grupos de acceso de manera centralizada .	*	*	*	*	*	*
P10	Políticas de acceso a datos	Pautas para definición de las políticas de control de acceso a reportes y exposición de datos en la capa de presentación del sistema.	*	*	*	*	*	*
P11	Política de control de accesos	Implementar en los dispositivos de acceso una política basada en la hora de los accesos, la ubicación y la cantidad de información que está siendo accedida o descargada.	*	*	*	*	*	*

BUENAS PRÁCTICAS			PROCESO DE ANÁLISIS DE DATOS					
REF	Nombre	Descripción	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / Acciones
P12	Cambio información crítica	Requerir re-autenticación cuando se vaya a realizar una actualización de la información crítica de una cuenta.	*	*	*	*	*	*
DC1	Revisión de Resultados de Análisis de Datos	Establecer una política para revisión de los resultados (incluye correlaciones) generadas a partir del proceso de Análisis de Datos.				*		*
DC2	Registro de logs en monitoreo	Mantener y revisar logs de las acciones correspondientes.	*	*	*	*	*	*
DC3	Métricas utilizadas para el análisis de seguridad de datos	Métricas sugeridas para realizar mediciones.	*	*	*	*	*	*
DC4	Registrar y monitorear logs en autenticación	Mantener y monitorizar un log de los intentos de autenticación del sistema.	*	*	*	*	*	*
DC5	Controles de integridad de datos	Considerar hacer uso de un algoritmo de hash criptográfico como una medida posible para asegurar la integridad de los datos almacenados.	*	*	*	*	*	*

BUENAS PRÁCTICAS			PROCESO DE ANÁLISIS DE DATOS					
REF	Nombre	Descripción	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / Acciones
R1	**Procesos de respuesta a revisión de resultados de Análisis de Datos**	Establecer procesos para recibir, analizar y responder a los reportes de los tipos de resultados o correlaciones.						*
R2	**Cifrar respaldos con datos críticos**	Verificar que los respaldos que contengan datos críticos sean cifrados de forma apropiada.	*	*	*	*	*	*

Tabla 19 Checklist de buenas prácticas y recomendaciones en Ciberseguridad

5.3.2. Buenas prácticas

Las buenas prácticas se agrupan de la siguiente forma, donde se subdividen según si son medidas de prevención, detección y control, y resiliencia:

1. **Consideraciones generales asociadas al proceso de Análisis de Datos (Sección 5.3.2.1):** contiene las Buenas Prácticas asociadas directamente al proceso de Análisis de Datos.
2. **Consideraciones de gestión de ambiente asociadas al proceso de Análisis de Datos (Sección 5.3.2.2):** se detallan consideraciones relativas a la correcta gestión del ambiente de la plataforma desde un punto de vista de ciberseguridad con medidas particulares para el contexto de Análisis de Datos.

5.3.2.1. Consideraciones generales asociadas al proceso de Análisis de Datos

Se debe tener en cuenta que las medidas presentadas aplican a todos los ambientes utilizados en el proceso de análisis de datos, incluyendo los de las pruebas.

Prevención (P)

Las buenas prácticas se presentan según la siguiente clasificación:

- Estructura organizacional.
- Requerimientos.
- Clasificación y manejo de Datos e Información.
- Conexiones y necesidad de actualización.

Estructura organizacional

Con respecto a la estructura de la organización, se recomienda:

P1. Identificar tipo de información

Identificar el Modelo de Datos Organizacional[20].

Adicionalmente, se deben identificar las fuentes que se están utilizando para el análisis y si alguna contiene información crítica[21]. Se debe tener presente que, si se está utilizando información crítica[21], se podrían tener restricciones para divulgar la información. Estas restricciones pueden, por ejemplo, incluir tanto la no divulgación a secciones determinadas del organismo, personal externo al mismo o la necesidad de aplicar enmascaramiento u otro procesamiento para poder hacerlo. Para saber si

[20] Una referencia para la definición y construcción de modelos de datos puede encontrarse en [10].

[21] En este marco se considera información crítica, a la clasificada como información reservada o confidencial según la legislación nacional. Para ver esta clasificación, consultar el punto "P5.1" en la sección "Anexo II: Profundización Buenas Prácticas".

la información que está siendo utilizada en el proceso es crítica, consultar el punto "P3.1" del Anexo III: Profundización de Buenas Prácticas en Ciberseguridad.

Roles involucrados: Arq. Análisis de Datos, Especialista / Ing. de Datos, Analista de Información, Científico de Datos.

Etapas del AD a tomar en cuenta: Preparación de los datos.

Requerimientos

Con respecto a los requerimientos del sistema (Fase Necesidad definida en el proceso de la Sección 4.1.2) de Análisis de Datos, se recomienda:

P2. Fuentes de Requerimientos

Considerar las siguientes fuentes de requerimientos en la etapa de análisis: partes interesadas, regulaciones gubernamentales, información crítica[21] y reglas de acceso, así como obligaciones contractuales.

Verificar, por ejemplo, si alguno de los proveedores de los datos utilizados impone restricciones sobre los mismos a nivel de privacidad o confidencialidad.

Verificar en caso de que existan, que estas restricciones están implementadas en el sistema y de lo contrario contactar al encargado de seguridad de la información o responsable del conjunto de datos involucrado. Por ejemplo, si se utiliza un conjunto de datos con información de empleados de una organización y no se debe poder acceder al nombre y género de los mismos, debe verificarse que una vez cargados los datos en el sistema, estos no pueden verse o ser accedidos por los usuarios del mismo.

Por más información, referirse al Anexo III: Profundización de Buenas Prácticas en Ciberseguridad.

Roles involucrados: Arq. Análisis de Datos, Especialista / Ing. de Datos, Analista de Información, Científico de Datos.

Etapas del AD a tomar en cuenta: Necesidad, Búsqueda de datos, Preparación de los datos, Modelado y análisis de datos, Presentación.

Clasificación y manejo de Datos e Información

Con respecto a la clasificación y manejo de datos e información, se recomienda:

P3. Clasificación de información y preprocesamiento

Verificar que los datos a ser utilizados en el proceso de análisis se encuentran categorizados. Mínimamente se debe saber si se está trabajando con información

crítica[21]. Consultar al RSI o al encargado de los datos en caso de que no encuentre la categoría a la que pertenece la información. Se debe tener presente que, si está utilizando información crítica[21], se podría tener restricciones para divulgar la información. Estas restricciones pueden por ejemplo incluir tanto la no divulgación a secciones determinadas del organismo, personal externo al mismo o la necesidad de aplicar enmascaramiento u otro procesamiento para poder hacerlo. Es decir, que en caso de estar creando reportes para personal con distintos permisos puede ser necesario preprocesar los datos antes de mostrarlos en la visualización para algunos usuarios.

Asegurar que en caso de que los datos requieran preprocesamiento, éste sea realizado. En el ejemplo anterior, se pueden definir un conjunto de datos de prueba con algunos registros del personal y validar que una vez expuestos al usuario, la información de nombres y género no está siendo mostrada y no puede ser accedida.

Por más información, referirse al Anexo III: Profundización de Buenas Prácticas en Ciberseguridad.

Roles involucrados: Arq. Análisis Datos, Especialista / Ing. de Datos, Analista de Información, Científico de Datos.

Etapas del AD a tomar en cuenta: Preparación de los datos.

P4. Enmascaramiento de Datos

Considerar el uso de técnicas de enmascaramiento u ofuscación[22] de datos cuando se quiere evitar mostrar información crítica o generar un conjunto de datos para un ambiente de pruebas. Ver tipos y ejemplos de enmascaramiento en Sección Anexo III: Profundización Buenas Prácticas en Ciberseguridad. Por ejemplo, en caso de que no se quiera exponer un subconjunto de los datos utilizados en el análisis a un proveedor externo al organismo, se puede aplicar enmascaramiento a estos datos previo al envío.

Roles involucrados: Arq. Análisis de Datos, Especialista / Ing. de Datos, Analista de Información, Científico de Datos.

Etapas del AD a tomar en cuenta: Preparación de los datos, Presentación.

P5. Métodos y canales de transferencia

Se sugiere evaluar el método y canales utilizados para realizar la transferencia de los datos.

Se recomienda realizar la transferencia de información crítica[21] por canales considerados seguros. En general, se recomienda hacer uso de un canal cifrado punto a punto, es decir que los datos estén cifrados en todo punto de la comunicación. Esto puede por ejemplo ser implementado haciendo uso de una VPN

[22] Técnicas que tienen como objetivo ocultar parcial o totalmente la información.

(red privada virtual por sus siglas en inglés) y TLS (protocolo seguridad de la capa de transporte, por sus siglas en inglés). Para este caso, puede considerarse la transferencia a través de REDuy [78], que implementa mecanismos robustos de autenticación y cifrado de datos. Una guía para implementar TLS correctamente puede encontrarse en [95] y para la implementación de VPNs en [96, 97].

Roles involucrados: Arq. Análisis de Datos, Especialista / Ing. de Datos, Analista de Información, Científico de Datos.

Etapas del AD a tomar en cuenta: Transversal.

P6. Procedencia de los datos

Se sugiere evaluar la procedencia de los datos que se incorporan al sistema de Análisis de Datos. Verificar si existe una lista de entidades u organizaciones permitidas para incorporar datos.

Roles involucrados: Especialista / Ing. de Datos, RSI, Control Regulatorio y Legal.

Etapas del AD a tomar en cuenta: Necesidad, Búsqueda de datos.

Conexiones y necesidad de actualización

Con respecto a las conexiones y necesidad de actualización, se recomienda:

P7. Conexiones y necesidades de actualización

Para las fuentes utilizadas que requieran conexión (como las bases de datos), minimizar la duración de las conexiones, es decir mantenerse conectado por el menor tiempo posible. Evaluar también la periodicidad necesaria con la que se realizan y asegurar que solo se establece una conexión cuando se necesita realizar una tarea específica y que cuando ésta termina, la conexión se cierra lo más rápido posible.

Por ejemplo, en el caso de las actualizaciones de datos del sistema de Análisis de Datos evaluar:

- Si es necesario actualizar los datos en tiempo real.
- Si es necesario realizar la transferencia de todos los datos cada vez que se realiza una carga. Siempre que sea posible preferir realizar actualizaciones incrementales de los mismos. Esto evita que los datos que ya se encuentran en el sistema sean transferidos nuevamente por la red, mitigando el riesgo de que sean interceptados.

Roles involucrados: Especialista / Ingeniería de datos, Arq. Analista de datos.

Etapas del AD a tomar en cuenta: Búsqueda de datos, Preparación de los datos, Modelado y análisis de datos, Presentación, Resultados / Acciones.

Detección y control (DC)

Las buenas prácticas se presentan según la siguiente clasificación:

- Reportes de fallas o anomalías.

Reportes de fallas o anomalías

Con respecto de fallas o anomalías, se recomienda:

DC1. Revisión de Resultados de Análisis de Datos

Establecer una política para revisión de los resultados (incluye correlaciones) generadas a partir del proceso de Análisis de Datos. En particular, se recomienda que los analistas al finalizar un tablero y eventualmente haberlo utilizado para extraer conclusiones sobre los mismos, evalúen si se ha generado información crítica en el proceso. En caso de que se verifique que este tipo de información puede ser extraída del tablero, podría restringirse el acceso al mismo para ciertos roles específicos de usuarios.

Roles involucrados: RSI, Analista de Información, Científico de Datos.

Etapas del AD a tomar en cuenta: Modelado y análisis de datos, Resultados / Acciones.

Resiliencia (R)

Las buenas prácticas se presentan según la siguiente clasificación:

- Respuestas a Incidentes.

Respuesta a incidentes

Con respecto a la respuesta a fallas o incidentes de seguridad en el sistema de Análisis de Datos, se recomienda:

R1. Procesos de respuesta a revisión de resultados de Análisis de Datos

Establecer procesos para recibir, analizar y responder a los reportes de los tipos de resultados o correlaciones definidos en el punto DC1. También se sugiere definir un proceso de respuesta en caso de detectar análisis que se consideren indebidos usando el sistema de Análisis de Datos.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Resultados / Acciones.

5.3.2.2. Consideraciones de gestión de ambiente

Tener en cuenta que las medidas presentadas aplican a todos los ambientes utilizados en el proceso de análisis de datos, incluyendo los de las pruebas.

Prevención (P)

Las buenas prácticas se presentan según la siguiente clasificación:

- Guía de usuarios y accesos.
- Autenticación.

Guía de usuarios y accesos

Con respecto a la gestión de usuarios y accesos, se recomienda:

P8. Definir roles y privilegios

Definir roles, privilegios y responsabilidades asociados a las herramientas utilizadas para el proceso de Análisis de Datos. Esta asignación debe seguir el Principio de Mínimos Privilegios. En particular, definir para los distintos roles las acciones que se pueden realizar en la capa de Presentación del sistema de Análisis de Datos, que pueden incluir la capacidad de compartir datos o introducirlos al sistema.

Considerar que incluso los ciudadanos u otros actores externos podrían realizar análisis sobre los datos expuestos por la organización, sin formar parte de la misma. Por tanto, podría ser necesario enmascarar ciertos datos propios del organismo que estén accesibles a terceros.

Por más información, referirse al Anexo III: Profundización de Buenas Prácticas en Ciberseguridad.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

P9. Gestión centralizada de identidades

Gestionar las identidades de los usuarios del sistema y la pertenencia a los distintos grupos de acceso de manera centralizada. Esto busca evitar problemas de integridad de datos y de seguridad en los accesos. Por ejemplo, que por desincronización de la gestión de grupos de acceso un usuario pueda acceder a un recurso de un grupo del que ya no forma parte porque la desvinculación del grupo no se propagó de forma adecuada a todo el sistema de análisis de datos.

Los cambios en los grupos de acceso deben contar con las autorizaciones correspondientes a nivel de la organización y deben quedar registrados en un sistema de gestión de versiones.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

P10. Políticas de accesos de datos

Para la definición de las políticas de control de acceso a reportes y exposición de datos en la capa de presentación del sistema, considerar si las plataformas o herramientas utilizadas son internas o externas (por ejemplo, servicios en la nube) a la organización. En la definición de los controles y procesos de tratamiento de los datos también debe considerarse si estos son gubernamentales, internos o externos a la organización.

Verificar en caso que se quiera enviar datos a un servicio en la nube o a un proveedor externo de la organización que no existan restricciones para el envío, por ejemplo, sobre el método de trasferencia a utilizar, si existen datos que no pueden ser transferidos o deben ser enmascarados para esto (Ver P4. Enmascaramiento de Datos).

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

P11. Política de control de accesos

Como parte del control de acceso a datos se sugiere implementar en los dispositivos de acceso una política basada en la hora de los accesos, la ubicación y la cantidad de información que está siendo accedida o descargada. Mínimamente, implementar dicha política en los dispositivos que tienen acceso a la capa de presentación del sistema de análisis de datos, eventualmente distinguiendo si son internos o externos a la organización.

Se sugiere considerar los volúmenes usuales de datos descargados del sistema por los usuarios que realizan análisis para definir cuándo se generarán alertas por descarga excesiva de datos. Adicionalmente, se sugiere indicar a los usuarios que si descargan un conjunto de datos que exceda dicho volumen, se generará una alerta de seguridad en el sistema. Si un usuario determina que debe descargar necesariamente una cantidad que generará una alerta, puede considerar avisar al responsable de la seguridad del conjunto de datos.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

Autenticación

Con respecto al proceso de autenticación, se recomienda:

P12. Cambio información crítica

Requerir re-autenticación cuando se vaya a realizar una actualización de la información crítica[21] de una cuenta, como la contraseña o el email asociado.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

Detección y control (DC)

Las buenas prácticas se presentan según la siguiente clasificación:

- Monitoreo y sistemas de detección de intrusión.
- Métricas.
- Autenticación.
- Integridad de datos.

Monitoreo y sistemas de detección de intrusión

Con respecto al monitoreo del sistema de Análisis de Datos y a los sistemas de detección de intrusión, se recomienda:

DC2. Registro de Logs en monitoreo

Para las partes del sistema de Análisis de Datos que se considere requieren monitoreo (Ver puntos DC1 y P10), mantener y revisar logs de las acciones correspondientes. Por ejemplo, al realizar la revisión de los logs (de forma automatizada o manual) considerar: buscar accesos en horarios poco frecuentes, múltiples intentos de autenticación fallidos, así como volúmenes inusuales de descargas. Puede considerarse la utilización de un centralizador de logs para incrementar la efectividad de estas tareas.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

Métricas

Con respecto a las métricas a ser usadas en la evaluación de la seguridad del sistema de Análisis de Datos, se recomienda:

DC3. Métricas utilizadas para el análisis de seguridad de datos

Como parte de los procesos de detección de incidentes y amenazas, así como de control de la seguridad del sistema considerar la evaluación de las siguientes métricas:

- Volumen de datos descargado por usuario y/o por área de negocio.
- Volumen de datos transferidos por redes inseguras.
- Volumen de datos críticos por sector del organismo.
- Intentos infructuosos de login.
- Accesos fuera de horarios operativos.
- Bloqueo de usuarios.
- Usuarios inactivos, por ejemplo, lista de usuarios no utilizados durante los últimos 90 días.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

Autenticación

Con respecto al proceso de autenticación, se recomienda:

DC4. Registrar y monitorear logs en autenticación

Mantener y monitorizar un log de los intentos de autenticación del sistema (Ver punto HC.DC2 en Anexo III: Buenas Prácticas de Ciberseguridad), en particular asegurar que se registran y revisan: los intentos de autenticación fallidos, las cuentas que se inactivan y las contraseñas involucradas en los intentos fallidos.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

Integridad de datos

Con respecto a la integridad de los datos almacenados, se recomienda:

DC5. Controles de integridad de datos

Considerar hacer uso de un algoritmo de hash criptográfico como una medida posible para asegurar la integridad de los datos almacenados. Para esto, al momento de almacenar un dato en el sistema de Análisis de Datos, se debe también computar y almacenar el hash del mismo. Para verificar la integridad del mensaje, se recalcula el hash del dato y se verifica que se obtenga el mismo valor que el del hash previamente almacenado.

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

Resiliencia (R)

Las buenas prácticas se presentan según la siguiente clasificación:

- Respaldo de datos.

Respaldo de datos

Con respecto al respaldo de datos se recomienda:

R2. Cifrar respaldos con datos críticos

La existencia de respaldos de los datos del sistema también genera un riesgo a nivel de seguridad. Por tanto, deben tomarse medidas para protegerlos.

Verificar que los respaldos que contengan datos críticos sean cifrados de forma apropiada y que las claves correspondientes se almacenen de forma segura. Se sugiere que las claves estén disponibles en lugares físicos externos a los del sistema, para poder ser usadas en un plan de recuperación ante desastres (estrategia de recuperación).

Roles involucrados: RSI.

Etapas del AD a tomar en cuenta: Transversal.

5.4. Legal

En la presente sección se desarrollan un conjunto de buenas prácticas y la regulación existente y aplicable en nuestro país para el análisis y tratamiento de datos personales.

5.4.1. Checklist

En la Tabla 20 se presentan las recomendaciones, buenas prácticas y principales normales del componente Legal, el formato de la tabla se explica al principio de la sección.

Estas buenas prácticas fueron pensadas para ser cumplidas por una organización perteneciente a la Administración Central (de ahora en adelante Entidad) que se propone crear una base de datos con datos personales (de ahora en adelante “datos”), tratarlos, conservarlos y, eventualmente, comunicarlos a terceros.

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
L1	**Obtención de los datos**	Todos los datos a tratar deben ser clasificados para determinar si son: datos personales o no y en caso afirmativo, si se trata de datos especialmente protegidos. Además, se debe establecer el origen de los datos	*	*				
L2	**Creación de la base de datos**	Para crear una base de datos, deberá tenerse en cuenta una serie de requisitos en cuanto a su finalidad y al consentimiento de los titulares	*		*			
L3	**Base de datos con datos sensibles**	Tratándose de datos sensibles, la legislación establece una serie de protecciones especiales que deberán cumplirse necesariamente			*	*	*	
L4	**Registro de la base de datos**	Todas las bases de datos deben registrarse para ser consideradas licitas			*			

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
L5	**Exclusiones**	Deberá cumplirse con todo lo previsto anteriormente, siempre que la base de datos no se encuentre excluida en razón de su objeto o regulación			*	*	*	
L6	**Tratamiento de los datos**	Los organismos deben adoptar las medidas necesarias para la seguridad, confidencialidad de los datos y el cumplimiento de los principios	*	*	*	*		
L7	**Encargado del tratamiento**	Si existe un Encargado de tratamiento, se debe documentar contractualmente el alance de sus servicios	*	*	*	*	*	*
L8	**Comunicación de datos**	Para poder comunicar datos de la base de datos a terceras personas, deberán cumplirse una serie de requisitos					*	*
L9	**Transferencia de datos**	Para poder transferir datos de la base de datos internacionalmente, deberán cumplirse una serie de requisitos			*	*	*	*

BUENAS PRÁCTICAS			PROCESO ESTÁNDAR					
REF	Título de buenas prácticas	Descripción de buenas prácticas	Necesidad	Búsqueda de datos	Preparación de los datos	Modelado y análisis de datos	Presentación	Resultados / acciones
L10	Conservación de datos	Una vez cumplida con finalidad para la cual la base fue creada, los datos deberán destruirse, o si así lo desea el titular de los mismos		*	*	*	*	*
L11	Derechos de los titulares	Los titulares de los datos incorporados a la base, en cualquier momento podrán solicitar acceso, rectificación o eliminación de los mismos	*	*	*	*	*	*
L12	Incumplimiento y sanciones	El desconocimiento por parte de las personas involucradas en el proceso, podrá acarrear la adopción de sanciones por parte del órgano de control	*	*	*	*	*	*

Tabla 20 Checklist de buenas prácticas y recomendaciones en Legal

5.4.2. Normativa y buenas prácticas

A continuación, se desarrolla el marco normativo y las buenas prácticas definidas en la subsección anterior.

El manejo de los datos en general, y de los datos personales en particular, representa un desafío para la Administración, la que deberá extremar sus cuidados respecto a su uso, tratamiento, almacenamiento y eventual comunicación de los mismos.

Es por ello que se establecen en este capítulo ciertas prácticas recomendables respecto a su manejo y la normativa vigente más relevante para la recolección de datos y creación de bases de datos con datos personales.

En lo que tiene que ver con las buenas prácticas, se recomienda que el funcionario encargado de la toma de decisión analice pormenorizadamente la necesidad, o no, del manejo de datos personales. Es ésta la primera recomendación, la cual se puede resumir en la siguiente pregunta: ¿necesito trabajar con datos personales o tengo caminos alternativos para conseguir el resultado que me propongo obtener?

Como se dijo anteriormente, dado la complejidad y sensibilidad del manejo de estos datos, es recomendable evitar su utilización en caso de resultar posible.

Una vez formulada la pregunta anterior y en caso de obtener una respuesta afirmativa en el sentido que sí debe trabajarse con datos personales, resulta procedente profundizar el análisis y examinar si el uso que se hará de los datos es el mínimo imprescindible.

No parece lógico tener los datos duplicados en más de un soporte (digital y papel, por ejemplo) o en más de un dispositivo. Tampoco pueden exponerse los datos si no se cuenta con las medidas de seguridad físicas y lógicas adecuadas. Por último, también será relevante el acceso: ¿qué personas estarán en condiciones de acceder a los datos, sólo las que les resulta imprescindible para cumplir sus tareas?

Como puede apreciarse, antes de abocarse a la creación de una base de datos con datos personales, el responsable de la misma deberá analizar todos estos supuestos a los efectos de minimizar los riesgos del tratamiento, así como sus responsabilidades funcionales.

Si definitivamente la persona encargada, luego de realizar este análisis preventivo, llega a la conclusión que tiene la necesidad de recolectar y tratar datos personales, deberá hacerlo en un todo ajustado a derecho.

Para ello, se detalla a continuación la forma correcta de trabajar con datos personales y con bases de datos con datos personales, se encuentren informatizadas o no, según la normativa vigente en el Uruguay. Por lo que primero se explica el glosario utilizado y luego las buenas prácticas presentadas en la subsección anterior.

L0. Conceptos generales

Resulta conveniente definir los conceptos que se utilizarán en este capítulo, de modo de uniformizar su alcance.

- **Base de datos:** indistintamente, designan al conjunto organizado de datos personales que sean objeto de tratamiento o procesamiento, electrónico o no, cualquiera que fuere la modalidad de su formación, almacenamiento, organización o acceso.
- **Comunicación de datos:** toda revelación de datos realizada a una persona distinta del titular de los datos.
- **Consentimiento del titular:** toda manifestación de voluntad, libre, inequívoca, específica e informada, mediante la cual el titular consienta el tratamiento de datos personales que le concierne.
- **Dato personal:** información de cualquier tipo referida a personas físicas o jurídicas determinadas o determinables.
- **Dato sensible:** datos personales que revelen origen racial y étnico, preferencias políticas, convicciones religiosas o morales, afiliación sindical e informaciones referentes a la salud o a la vida sexual.
- **Destinatario:** persona física o jurídica, pública o privada, que recibiere comunicación de datos, se trate o no de un tercero.
- **Disociación de datos:** todo tratamiento de datos personales de manera que la información obtenida no pueda vincularse a persona determinada o determinable.
- **Encargado del tratamiento:** persona física o jurídica, pública o privada, que sola o en conjunto con otros trate datos personales por cuenta del responsable de la base de datos o del tratamiento.
- **Fuentes accesibles al público:** aquellas bases de datos cuya consulta puede ser realizada por cualquier persona, no impedida por una norma limitativa o sin más exigencia que, en su caso, el abono de una contraprestación.
- **Tercero:** la persona física o jurídica, pública o privada, distinta del titular del dato, del responsable de la base de datos o tratamiento, del encargado y de las personas autorizadas para tratar los datos bajo la autoridad directa del responsable o del encargado del tratamiento.
- **Responsable de la base de datos o del tratamiento:** persona física o jurídica, pública o privada, propietaria de la base de datos o que decida sobre la finalidad, contenido y uso del tratamiento.
- **Titular de los datos:** persona cuyos datos sean objeto de un tratamiento incluido dentro del ámbito de acción de la presente ley.
- **Tratamiento de datos:** operaciones y procedimientos sistemáticos, de carácter automatizado o no, que permitan el procesamiento de datos personales, así como también su cesión a terceros a través de comunicaciones, consultas, interconexiones o transferencias.
- **Usuario de datos:** toda persona, pública o privada, que realice a su arbitrio el tratamiento de datos, ya sea en una base de datos propia o a través de conexión con los mismos.

L1. La obtención de los datos

Para la obtención de los datos deben tenerse en cuenta los siguientes puntos:

- Deberá definirse en primer lugar si se trata de datos personales (de personas físicas o jurídicas) de acuerdo a la definición legal (Ley No. 18.331, art. 4, lit. D) y en caso afirmativo si los datos conformarán una base de datos con datos personales (Ley No. 18.331, art. 4, lit. A).
- Tratándose de datos especialmente protegidos, deberá tomarse en cuenta lo expresado en L3.
- Los datos personales que requieran el previo consentimiento informado y se encuentren incluidos en información pública en poder de un Organismo, tienen el carácter de información confidencial (Ley No. 18.381, art. 10).
- Los datos del Sistema de Información Integrada del Área Social (SIIAS) del Ministerio de Desarrollo Social, se encuentran regulados por la Ley No. 18.331. (Ley No. 18.719, art. 621)
- Luego debe establecerse el origen de los datos, los cuales pueden provenir de las siguientes fuentes:
    - Datos propios de dependientes de la Entidad.
    - Datos propios de particulares en poder de la Entidad.
    - Datos de terceros ajenos a la Entidad, y en este caso pueden ser:
        - Obtenidos de fuentes públicas (Ley No. 18.331, art. 9 bis).
        - Entregados voluntariamente por los titulares.
        - Entregados por los titulares en cumplimiento de una obligación legal.

Etapas del AD a tomar en cuenta: Necesidad, Búsqueda de datos.

L2. La creación de la base de datos

Para la creación de una base de datos, informatizada o no, el organismo debe tener en cuenta lo siguiente:

- Sólo podrá crearse una base de datos cuando:
    - Cumpla con los principios generales de la normativa.
    - No tenga una finalidad violatoria de los derechos humanos.
    - No sea contraria a las leyes.
    - No sea contraria a la moral pública (Ley No. 18.331, art. 6).
- Los datos que contenga deberán ser:
    - Veraces.
    - Adecuados.
    - Ecuánimes.
    - Exactos.
    - Actualizados.
    - No excesivos en relación con la finalidad con que se recogieron (Ley No. 18.331, art. 7).

- Deberá contarse con el consentimiento del titular del dato cuando resulte necesario (Ley No. 18.331, art. 9).
- No se requiere consentimiento del titular cuando:
    - Los datos provengan de fuentes públicas de información (registros, publicaciones en medios masivos de comunicación).
    - Se recaben para el ejercicio de funciones propias de los poderes del Estado.
    - Se recaben en virtud de una obligación legal.
    - Se trate de listados (nombres y apellidos, documento de identidad, nacionalidad, domicilio y fecha de nacimiento de las personas físicas; y razón social, nombre de fantasía, registro único de contribuyentes, domicilio, teléfono e identidad de las personas a cargo de las personas físicas).
    - Deriven de una relación contractual, científica o profesional y sean necesarios para su cumplimiento (Ley No. 18.331, art. 9).
    - Se trate de registros y documentos destinados a la protección y contralor del trabajo establecidos por la normativa (Ley No. 19.355, art. 84).
- Cuando se requiera consentimiento éste debe ser:
    - Libre.
    - Previo.
    - Expreso.
    - Informado (Ley No. 18.331, art. 9).
    - Documentado.
    - Gratuito.

Si el consentimiento es prestado juntamente con otras declaraciones, debe estar en forma expresa y destacada. Deberá informársele la finalidad, la existencia de la base, el carácter obligatorio o no de las respuestas y sus consecuencias, así como la posibilidad del titular de ejercer sus derechos (Ley No. 18.331, art. 13).

- El responsable de la base deberá guardar la prueba del consentimiento. (Decreto 414/009 de 31 de agosto de 2009, art. 6)
- No podrán crearse bases de datos para la adopción de decisiones con efectos jurídicos que afecten de manera significativa a los titulares (Ley No. 18.331, art. 16).
- Las bases de datos deberán estar alojadas en centros de datos seguros situados en territorio nacional, excepto que se consideren que no constituye un riesgo para el organismo (Decreto No. 92/014 de 7 de abril de 2014, art. 3).

Etapas del AD a tomar en cuenta: Necesidad, Preparación de los datos.

L3. Base de datos con datos sensibles

Para las bases de datos que contienen datos sensibles debe considerarse (Ley No. 18.331, art. 18):

- Los datos sensibles están especialmente protegidos, razón por la cual está prohibida la formación de bases de datos con datos sensibles o que indirectamente los revelen.
- Las bases de datos que contengan datos sensibles sólo podrán crearse o tratarse cuando existan razones de interés general autorizadas por ley o cuando la Entidad tenga mandato legal para hacerlo.
- Se podrán tratar estos datos con finalidades estadísticas o científicas si se disocian de sus titulares.
- Los establecimientos sanitarios públicos y los profesionales de la salud -física o mental- sólo pueden recolectar y tratar datos relativos a la salud de sus pacientes (Ley No. 18.331, art. 19).

La información relativa a la identidad de los titulares de los actos del registro de usuarios de cannabis, es considerada un dato sensible (Ley No. 19.172, art. 28).

Etapas del AD a tomar en cuenta: Preparación de los datos, Modelado y análisis de datos, Presentación.

L4. Registro de la base de datos

Todas las bases de datos deberán inscribirse en el Registro que a tales efectos lleve la Unidad Reguladora y de Control de Datos Personales - URCDP (Ley No. 18.331, art. 24).

Sólo son consideradas lícitas las bases de datos inscriptas y que cumplan con los principios establecidos en la normativa (Ley No. 18.331, art. 6).

El plazo para la inscripción es de 90 días desde el inicio de la actividad y la información debe mantenerse actualizada, comunicándolo trimestralmente a la URCDP. (Decreto 664/008 de 22 de diciembre de 2008, art. 3 y Decreto 414/009 de 31 de agosto de 2009, arts. 17 y 20).

Etapas del AD a tomar en cuenta: Preparación de los datos.

L5. Exclusiones

No será de aplicación la metodología referida en los anteriores puntos, cuando se trate de bases de datos (Ley No. 18.331, art. 3):

- Creadas y reguladas por leyes especiales.

Tengan por objeto la seguridad pública, la defensa, la seguridad del Estado y sus actividades en materia penal, investigación y represión del delito.

Etapas del AD a tomar en cuenta: Preparación de los datos, Modelado y análisis de datos, Presentación.

L6. Tratamiento de datos

Para el tratamiento de los datos se debe considerar:

- La Entidad debe adoptar las medidas necesarias para asegurar la seguridad y confidencialidad de los datos, evitando la adulteración, pérdida o consulta no autorizada (Ley No. 18.331, art. 10).
- Los datos deberán almacenarse de forma que se pueda cumplir con el derecho de acceso de los titulares (Ley No. 18.331, art. 10).
- Se deberá documentar debidamente con las personas que por su situación laboral o profesional accedan a los datos, su responsabilidad de actuación bajo secreto profesional y con responsabilidad penal por ello (Ley No. 18.331, art. 11).
- Aplicar el principio de responsabilidad proactiva, adoptando las medidas técnicas y organizativas para garantizar un tratamiento adecuado (privacidad por diseño, privacidad por defecto, evaluación de impacto), (Ley No. 18.331, art. 12).
- Una vez que los datos no sean necesarios o pertinentes para cumplir con el fin para el que fueron recolectados, deberán eliminarse (Ley No. 18.331, art. 8).
- Podrán conservarse datos personales con fines históricos, estadísticos o científicos, siempre que se autorice expresamente por la URCDP (Decreto 414/009 de 31 de agosto de 2009, arts. 37 y 38).

En ningún caso podrá comunicarse datos entre bases si no existe ley o previo consentimiento del titular (Ley No. 18.331, art. 8).

Etapas del AD a tomar en cuenta: Necesidad, Búsqueda de datos, Preparación de los datos, Modelado y análisis de datos.

L7. El Encargado del tratamiento

En caso de que exista un Encargado de tratamiento que preste servicios informáticos, o sea, una persona física o jurídica que intervenga en el tratamiento de los datos por cuenta del Responsable, deberá documentarse contractualmente el alcance de los servicios (Ley No. 18.331, art. 30).

No se considera comunicación o cesión de datos el acceso por parte de un encargado de tratamiento que resulte necesario para la prestación de un servicio al responsable. (Decreto 414/009 de 31 de agosto de 2009, art. 14 in fine).

Etapas del AD a tomar en cuenta: Transversal.

L8. Comunicación de datos

Con el fin de comunicar datos, se debe tomar en cuenta:

- Cualquier persona física o jurídica que brinda tratamiento a una base de datos, está obligada a utilizar la información que conozca en forma reservada y destinarla exclusivamente a su actividad, quedándole prohibida la difusión a terceros. Estas personas están obligadas a guardar secreto profesional, aún después de finaliza su relación con el responsable de la base (Ley No. 18.331, art. 11).
- Los mecanismos de intercambio de información clínica con fines asistenciales, mediante el Sistema de Historia Clínica Electrónica

Nacional, deberán asegurar la confidencialidad de la información (Ley No. 19.355, art. 466).

- Sólo se puede comunicar datos en los siguientes casos:
  - Para el cumplimiento de los fines relacionados al interés del emisor y del destinatario.
  - Con el previo consentimiento informado del titular.
- En caso de investigaciones médicas, los datos serán considerados confidenciales y sólo se puede revelar la identidad de los titulares en caso de autorización expresa, garantizando el anonimato (Ley No. 18.286, art. 70).
- Los funcionarios policiales podrán solicitar a la Dirección Nacional de Asistencia y Seguridad Social Policial que transfiera electrónicamente su historia laboral a instituciones de intermediación financiera o de crédito (Ley No. 19.405, art. 52).
- La Dirección General Impositiva y el Banco de Previsión Social pueden dar a publicidad datos de los sujetos que realicen defraudación tributaria por un monto elevado o, cuando la naturaleza del acto realizado afecte el interés general (Ley No. 17.930, arts. 460 y 461).
- Podrá comunicarse a la Dirección General Impositiva y el Banco de Previsión Social, cuando así lo soliciten por escrito, los datos necesarios para el control de los tributos (Ley No. 17.930, arts. 469).
- En cualquier momento el titular puede revocar el consentimiento otorgado (Ley No. 18.331, art. 17).
- No se requiere consentimiento para informar datos, cuando (Ley No. 18.331, art. 17):
  - Exista una ley que lo disponga.
  - Sean datos de salud necesarios por razones sanitarias.
  - De emergencia.
  - Para la realización de estudios epidemiológicos, debiendo disociarse en caso de resultar posible.
  - Se encuentren disociados de sus titulares.
- Las entidades financieras deberán comunicar a la Dirección General Impositiva los datos personales establecidos a texto expreso en el art. 18 de la Ley No. 19.484, de 5 de enero de 2017.
- No es posible comunicar datos ante una solicitud de acceso a la información pública, cuando ésta ha sido declarada confidencial por contener datos personales que requieran el previo consentimiento informado (art. 10 de la Ley No. 18.381, de 17 de 0ctubre de 2008).
- Si los datos personales se encuentran incluidos en información pública con carácter de datos abiertos, los datos deberán estar disociados (Decreto 54/017, de 20 de febrero de 2017, art. 4).

Etapas del AD a tomar en cuenta: Presentación, Resultados / Acciones.

L9. Transferencia de datos

La transferencia internacional de datos es posible cuando (Ley No. 18.331, art. 23):

- Se cuente con el consentimiento inequívoco del titular.
- Si es necesaria para cumplir un contrato entre el interesado y el responsable del tratamiento. Se incluye también las medidas precontractuales solicitadas por el interesado.
- Si es necesaria para celebrar o cumplir un contrato entre responsable y un tercero, siendo en interés del interesado.
- Sea necesaria o exigida para salvaguardar un interés público importante.
- Sea para el reconocimiento, ejercicio o defensa de un procedimiento judicial.
- Sea necesaria para salvaguardia del interés vital del interesado.
- Se haga desde un registro que esté concebido para facilitar información al público, esté abierto a la consulta o se realice por parte de quien demuestre interés legítimo en la consulta.

No está permitido transferir datos internacionalmente a países u organismos internacionales que no cuenten con un nivel de protección adecuado, excepto que el Organismo de Control autorice la transferencia, requiriendo al responsable las garantías suficientes para los titulares de los datos. Esta prohibición no rige en caso de:

- Cooperación judicial internacional.
- Intercambio de datos médicos cuando lo requiera el tratamiento del afectado.
- Transferencias bancarias o bursátiles.
- Tratados internacionales en que Uruguay sea parte.
- Entre organismos de inteligencia en la lucha contra el crimen organizado, el terrorismo y el narcotráfico.

Para la transferencia internacional de datos deberá solicitarse la autorización previa de la URCDP (Decreto 414/009 de 31 de agosto de 2009, arts. 34 y 35)

Etapas del AD a tomar en cuenta: Preparación de los datos, Modelado y análisis de datos, Presentación, Resultados / Acciones.

L10. Conservación de los datos

En cuanto a la conservación de los datos:

- En caso de servicios informatizados de datos, éstos deberán ser destruidos una vez cumplido el contrato, excepto que se tenga autorización del comitente si se presume la existencia de futuros encargos (Ley No. 18.331, art. 30).
- Los datos registrados con fines policiales se cancelarán cuando no sean necesarios para las averiguaciones que los motivaron (Ley No. 18.331, art. 25).

- Si el titular lo requiere, deberán eliminarse los datos de la base cuando (Ley No. 18.331, art. 15):
    - Causen perjuicio a los derechos e intereses legítimos de terceros.
    - Exista notorio error.
    - Se contravenga una obligación legal.

Etapas del AD a tomar en cuenta: Búsqueda de datos, Preparación de los datos, Modelado y análisis de datos, Presentación, Resultados / Acciones.

L11. Derechos de los titulares de los datos

Sobre los derechos de los titulares de los datos se debe tomar en cuenta:

- Los titulares de los datos deberán ser informados previamente a la recolección, de forma expresa, precisa e inequívoca de lo siguiente (Ley No. 18.331, art. 13):
    - La finalidad para la que se recolectan.
    - Quienes pueden ser los destinatarios.
    - La existencia de la base de datos, identidad y domicilio del responsable.
    - El carácter obligatorio o no de las respuestas.
    - Las consecuencias de proporcionar o no los datos.
    - La posibilidad de ejercer sus derechos de acceso, rectificación y supresión.
- El titular de los datos, que acredite su identidad, tendrá derecho a conocer la información que exista en la base de datos sobre su persona (Ley No. 18.331, art. 14).
- Si el titular es una persona fallecida, sus sucesores universales podrán ejercer este derecho (Ley No. 18.331, art. 14).
- El responsable de la base debe brindar la información en el plazo máximo de 5 días.
- La información debe ser suministrada de forma clara, amplia sobre la totalidad de los datos y estar presentada de manera no codificada y con la explicación que pueda corresponder.
- El solicitante podrá optar por el formato en el que deberá suministrarse la información, sea escrito, informático, telefónico, etc.
- Toda persona tiene derecho a solicitar rectificación, inclusión o supresión de la información existente sobre sí en una base de datos (Ley No. 18.331, art. 15).
- En caso de proceder la modificación, el responsable debe corregir la situación en el plazo máximo de 5 días desde la solicitud.
- Si se solicitara acceso a los datos cuando los mismos se encuentran en proceso de revisión, el responsable deberá dejar constancia de esa situación.
- Las modificaciones o supresiones se realizarán sin cargo alguno para el titular solicitante.

Etapas del AD a tomar en cuenta: Transversal.

L12. Incumplimiento y sanciones

El incumplimiento a la normativa por parte del responsable de la Entidad, del encargado del tratamiento o de cualquier persona comprendidas en la misma, podrá aparejar sanciones por parte del Órgano de Control (Ley No. 18.331, art. 35).

Las sanciones se graduarán en atención a la gravedad o reiteración, y podrán ir desde la simple observación hasta la clausura de la base, pasando por el apercibimiento, multa económica y suspensión de hasta 5 días (Ley No. 18.331, art. 35).

Etapas del AD a tomar en cuenta: Transversal.

6. Glosario

Access Control List: lista que especifica los permisos de los usuarios de un sistema. Define cuáles usuarios y grupos pueden acceder, así como qué operaciones pueden realizar. Se abrevia ACL.

Algoritmo de hash: función que recibe un bloque de datos de tamaño variable y da una salida de tamaño fijo.

Algoritmo de hash criptográfico20: función de hash con la propiedad de que los resultados de aplicar la función a un gran número de entradas, produce salidas distribuidas uniformemente y que parecen aleatorias.

Algoritmo: Conjunto de instrucciones ordenadas que ofrecen una solución a un problema.

Almacén de datos: conjunto de datos. Incluye el concepto de Data Warehouse, Data Marts, entre otros.

Autenticación: mecanismo que permite asegurar que un usuario es quien dice ser.

Autorización: mecanismo que permite establecer si un usuario tiene permisos sobre un recurso.

Big Data: son activos de información de gran volumen, de alta velocidad y/o de gran variedad, que exigen soluciones rentables e innovadoras de procesamiento de información, con plataformas especialmente diseñadas en términos de hardware y software, que permitan una mejor comprensión, toma de decisiones y automatización de procesos.

Business Intelligence: conjunto de herramientas que dan soporte al análisis de datos. Se abrevia BI.

Clasificación: determinar a qué clase pertenece un dato.

Clúster: conjunto de servidores.

Clustering: agrupación de los datos en grupos, a partir de las características de los mismos.

Controlador de dominio20: servidor responsable por manejar la información de dominio, como la información de acceso, incluyendo contraseñas.

Data Mining: metodologías para analizar datos desde distintos puntos de vista, encontrando patrones en los mismos, resumiéndolos y clasificándolos en grupos, así como identificando relaciones entre dichos datos.

Data Science: aplicación de algoritmos sobre los datos para generar valor a partir de los mismos.

Data Warehouse: es una colección de datos, integrado, no volátil y dinámico. Ayuda a la toma de decisiones en el organismo que lo utilice. Se abrevia DW.

Datos abiertos: son aquellos datos que se encuentran disponibles en un formato estándar, permitiendo así la interoperabilidad, y que su accesibilidad es abierta al público

Deep learning: conjunto de algoritmos de aprendizaje automático para modelar abstracciones de alto nivel.

DMZ (Zona desmilitarizada) 20: segmento de red que se encuentra entre la red privada de una organización e Internet.

Drill down: técnica que permite obtener un mayor nivel de granularidad en los datos.

Enmascaramiento de Datos12: técnicas cuyo objetivo es ocultar parcial o totalmente la información.

Extraction, transformation & load: procesos que implican extracción, transformación y carga de los datos requeridos para la ejecución de casos de uso. Se abrevia ETL.

Firewall20: un software que monitorea las comunicaciones entre el PC y otras computadoras, bloqueando cierto tráfico no deseado para aumentar la seguridad del equipo.

Framework: es un conjunto estandarizado de conceptos, prácticas y criterios sobre una problemática. Sirve como referencia, para enfrentar y resolver problemas de índole similar.

Gobierno electrónico: es el uso de las TIC por parte de los organismos del Gobierno, con el fin de mejorar los servicios e información que ofrecen a los ciudadanos. Así, como aumentar la eficiencia y eficacia de la gestión pública.

LDAP: protocolo de la capa de aplicación que controla el acceso de la información que está almacenada de forma centralizada en una red. Sus siglas significan Lightweight Directory Access Protocol.

Log: un registro de los eventos que ocurren en los sistemas o redes de una organización.

Machine Learning: mediante la aplicación de algoritmos sobre los datos, se permite predecir y encontrar patrones en los mismos. Se abrevia ML.

Metadatos20: Información utilizada para describir ciertas características, restricciones, parámetros y modalidades de uso de un conjunto de datos.

Modelo: diseño del algoritmo de Machine Learning, se visualiza como una función que se aplica a los datos nuevos de entrada, donde la salida es el resultado final.

Multicolinealidad (Multicollinearity): problema estadístico que consiste en la existencia de correlación entre las distintas variables de un modelo.

Open Source: es un término que se utiliza para denominar los tipos de software que se distribuyen mediante una licencia que le permite al usuario final utilizar el código del programa para analizarlo y modificarlo.

Outlier: es un valor atípico en una observación.

Predicción: hecho o situación que se anuncia que sucederá en el futuro.

Proxy20: es un dispositivo o programa intermediario que provee comunicación y otro tipo de servicios entre un cliente y un servidor. El proxy acepta cierto tipo de tráfico, ya sea entrante o saliente de la red, procesándolo y reenviándolo. Evita el trafico directo entre redes internas y externas.

QMC: consola de gestión de QlikSense. Las siglas significan Qlik Management Console.

Sample: conjunto de datos que es parte del conjunto de datos totales.

Scheduler: planificador de tareas.

Secure Sockets Layer (SSL) y Transport Layer Security (TLS)20: protocolos de autenticación y seguridad ampliamente implementados en navegadores y servidores web. SSL ha sido reemplazado por TLS. Estos protocolos son utilizados para proteger la información en las transmisiones por Internet.

Sistema de Nombre de Dominio (DNS) 21: es un servicio de búsqueda que provee un mapeo entre el nombre de red de un dispositivo en Internet y una dirección numérica.

Sobreajuste de un modelo (Overfitting): se da cuando se sobre-entrena un modelo con casos particulares, como consecuencia no reconoce nuevas entradas que no tengan las mismas características.

Test set: conjunto de datos utilizado para verificar el funcionamiento del algoritmo de Machine Learning una vez finalizado el entrenamiento y de esta forma validar el modelo para el posterior uso con nuevos datos.

Tests de penetración20: Pruebas que verifican el nivel de un sistema, dispositivo o proceso para resistir intentos activos para comprometer su seguridad.

Training set: conjunto de datos que se utilizan para el entrenamiento del algoritmo de Machine Learning.

Validation set: conjunto de datos que se utilizan para tunear los parámetros del algoritmo de Machine Learning para ir corrigiendo los errores generados en el entrenamiento.

VPN (Red privada virtual): Red de datos que permite que dos o más participantes se comuniquen de forma segura a través de una red pública creando una conexión privada o “túnel” entre ellos.

WAF: un Firewall de Aplicaciones Web es un dispositivo de hardware o software que permite proteger los servidores de aplicaciones Web de determinados ataques específicos en Internet.

Referencias

[1] AGESIC, Presidencia de la República Oriental del Uruguay (2019). Plan de Gobierno Digital 2020: Transformación con equidad. *Sitio web Presidencia de la República Oriental del Uruguay*

https://www.gub.uy/agencia-Gobierno-electronico-sociedad-informacion-conocimiento/politicas-y-gestion/plan-de-Gobierno-digital-uruguay-2020. Último acceso 12/11/2019.

[2] AGESIC (2019). Uruguay: Política de Datos para la Transformación Digital*. Sitio web Presidencia de la República Oriental del Uruguay*.

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/noticias/uruguay-politica-datos-para-transformacion-digital. Último acceso 12/11/2019.

[3] AGESIC (2019). Qué son los datos abiertos. *Sitio web Presidencia de la República Oriental del Uruguay.*

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/son-datos-abiertos. Último acceso 12/11/2019.

[4] AGESIC (2019). Uruguay: Gobierno Digital y D9*. Sitio web Presidencia de la República Oriental del Uruguay*.

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/uruguay-gobierno-digital-d9. Último acceso 12/11/2019.

[5] Uruguay preside el D9, el grupo de gobiernos digitalmente más avanzados del mundo. *Sitio web Uruguay sustentable*.

http://www.uruguaysustentable.uy/uruguay-d9/. Último acceso 12/11/2019.

[6] AGESIC (2019). Comenzó “360°. Una visión integral para la gestión de los datos en el Estado”. *Sitio web Presidencia de la República Oriental del Uruguay.*

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/noticias/comenzo-datos-360deg-vision-integral-para-gestion-datos-estado. Último acceso 12/11/2019.

[7] United Nations (2018). E-Government Survey 2018. *Sitio web United Nations.*

https://publicadministration.un.org/egovkb/en-us/Reports/UN-E-Government-Survey-2018. Último acceso 12/11/2019.

[8] AGESIC (2018). Marco de Ciberseguridad. *Sitio web Presidencia de la República Oriental del Uruguay.*

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/marco-de-ciberseguridad. Último acceso 12/11/2019.

[9] ITU (2018). Global Cybersecurity Index (GCI) 2018. *ITUPublications*.

https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf. Último acceso 22/11/2019.

[10] DAMA International (2019). DAMA – DMBOK: Data Management Body of Knowledge. *Technics Publications*.

[11] Ley 18.331. *Sitio IMPO*.

https://www.impo.com.uy/bases/leyes/18331-2008. Último acceso 21/11/2019.

[12] Ley 25.326. *Sitio OAS*.

https://www.oas.org/juridico/pdfs/arg_ley25326.pdf. Último acceso 21/11/2019.

[13] Ley 1581. *Sitio web Ministerio de Comercio y Turismo*.

https://www.mintic.gov.co/portal/604/articles-4274_documento.pdf. Último acceso 21/11/2019.

[14] Comisión Europea. Reforma de 2018 de las normas de protección de datos de la UE. *Sitio web Unión Europea*.

https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_es. Último acceso 21/11/2019.

[15] Pérez V. Impacto en Uruguay del nuevo Reglamento de la Unión Europea sobre protección de datos personales. *Sitio web Presidencia de la República Oriental del Uruguay*.

https://www.gub.uy/unidad-reguladora-control-datos-personales/comunicacion/publicaciones/impacto-en-uruguay-del-nuevo-reglamento-de-la-union-europea-sobre. Último acceso 21/11/2019.

[16] *Sitio web OECDiLibrary*.

https://www.oecd-ilibrary.org/docserver/09ab162c-en.pdf?expires=1571246564&id=id&accname=guest&checksum=FE8D5B21408BDE9A1185217E769A71C7. Último acceso 21/11/2019.

[17] Ebrahim S, Murphy T (2019). Think slower: How behavioral science can improve decision making in the workplace. *Sitio web Deloitte*.

[18] Deloitte. Government & public services. *Sitio web Deloitte*.

[19] *Sitio web Gartner*.

https://www.gartner.com/imagesrv/summits/docs/na/business-intelligence/gartners_business_analytics__219420.pdf. Último acceso 21/11/2019.

[20] Colombia's Big Data Strategy. *Sitio web Data-pop Alliance*.

https://datapopalliance.org/colombia-big-data-strategy/. Último acceso 21/11/2019.

[21] *Sitio web NYC Analytics*.

https://www1.nyc.gov/site/analytics/index.page. Último acceso 21/11/2019.

[22] NYC Analtics. Mayor's Office of Data Analytics (MODA). *Sitio web NYC Analitics*.

https://www1.nyc.gov/assets/analytics/downloads/pdf/MODA-project-process.pdf. Último acceso 21/11/2019.

[23] *Sitio web London Database*.

https://data.london.gov.uk/city-data-analytics-programme/. Último acceso 21/11/2019.

[24] Carilni A, Ercolani A, et al. Data & Analytics Framework (DAF). *Sitio web Digital Transformation Team*.

https://teamdigitale.governo.it/en/projects/daf.htm. Último acceso 21/11/2019.

[25] Gobierno de Nueva Zelenda (2018). Principles for the safe and effective use of data and analytics. *Sitio web Stats NZ*.

https://www.stats.govt.nz/assets/Uploads/Data-leadership-fact-sheets/Principles-safe-and-effective-data-and-analytics-May-2018.pdf. Último acceso 21/11/2019.

[26] *Sitio web data.govt.nz – Manage data*.

https://www.data.govt.nz/manage-data. Último acceso 21/11/2019.

[27] *Sitio web data.govt.nz – Use data*.

https://www.data.govt.nz/use-data. Último acceso 21/11/2019.

[28] Australian Government. Data: skills and capability in the Australian public service. *Sitio web Australian Government*.

https://www.pmc.gov.au/sites/default/files/publications/data-skills-capability.pdf. Último acceso 21/11/2019.

[29] AGESIC. Plataforma de Interoperabilidad: ¿qué es? *Sitio web centro de conocimientos de AGESIC.*

https://centrodeconocimiento.agesic.gub.uy/web/ccio/plataforma-de-interoperabilidad. Último acceso 13/11/2019.

[30] AGESIC. Arquitectura de Gobierno. *Sitio web centro de conocimientos de AGESIC*.

https://centroderecursos.agesic.gub.uy/web/arquitectura-de-gobierno/arquitectura-integrada-de-gobierno/-/wiki/Arquitectura+de+Gobierno/Arquitectura+de+Gobierno+y+Arquitectura+Empresarial. Último acceso 13/11/2019.

[31] AGESIC (2019). Plataforma de Datos. *Sitio web Presidencia de la República Oriental del Uruguay.*

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/plataforma-datos. Último acceso 13/11/2019.

[32] AGESIC (2019). Qué es la Arquitectura de Datos. *Sitio web Presidencia de la República Oriental del Uruguay.*

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/politicas-y-gestion/arquitectura-datos. Último acceso 13/11/2019.

[33] AGESIC. Modelo de Referencia de Datos. *Sitio web centro de conocimientos de AGESIC*.

https://centroderecursos.agesic.gub.uy/web/arquitectura-de-gobierno/arquitectura-integrada-de-gobierno/-/wiki/Arquitectura+de+Gobierno/Modelo+de+Referencia+de+Datos. Último acceso 13/11/2019.

[34] *Sitio web CEPAL*.

https://www.cepal.org/es/proyectos/big-data. Último acceso 13/11/2019.

[35] CEPAL (2019). Cumbre de Inteligencia Artificial en América Latina. *Sitio web CEPAL*.

https://www.cepal.org/es/notas/cumbre-inteligencia-artificial-america-latina. Último acceso 13/11/2019.

[36] *Sitio web AI Latin America SumMIT*.

http://ailatinsum.mit.edu/. Último acceso 13/11/2019.

[41] Elespectador.com (2018). Colombia adopta políticas de Big Data como pionero en A. Latina. *Sitio web AETecno*.

https://tecno.americaeconomia.com/articulos/colombia-adopta-politica-de-big-data-como-pionero-en-latina. Último acceso 13/11/2019.

[42] Clusellas P, Martelli E, et al. (2019). Un gobierno inteligente: el cambio de la Administración Pública de la Nación Argentina 2016-2019.

https://www.boletinoficial.gob.ar/pdfs/gobierno_inteligente.pdf. Último acceso 13/11/2019.

[43] Straface F (2019). Big Data y datos abiertos: cómo crean valor gobiernos y empresas. *Sitio web Cronista*.

https://www.cronista.com/columnistas/Big-data-y-datos-abiertos-como-crean-valor-gobiernos-y-empresas-20190310-0026.html. Último acceso 13/11/2019.

[47] Barbero M, Coutuer J, et al. (2016). *Big data analytics for policy making*.

https://joinup.ec.europa.eu/sites/default/files/document/2016-07/dg_digit_study_big_data_analytics_for_policy_making.pdf. Último acceso 13/11/2019.

[48] Dhanda P, Sharma N (2016). Extract Transform Load Data with ETL Tools. *International Journal of Advanced Research in Computer Science* 7(3).

https://www.ijarcs.info/index.php/Ijarcs/article/download/2662/2650. Último acceso 31/10/2019.

[49] Chou L (2019). Comparison of Data Analysis Tools: Excel, R, Python and BI Tools. *Sitio web Towards Data Science.*

https://towardsdatascience.com/comparison-of-data-analysis-tools-excel-r-python-and-bi-tools-6c4685a8ea6f. Último acceso 1/11/2019.

[50] Bruce D. Pentaho vs. Talend: How the Two Data Integration Tools Compare?. *Sitio web knowledgenile*.

https://www.knowledgenile.com/blogs/pentaho-vs-talend/. Último acceso 1/11/2019.

[51] Uher P (2009). Comparision CloverETL vs. competitors. *Sitio web CloverETL*.

https://is.muni.cz/th/jzajc/Comparison_CloverETL_vs_Talend_Pentaho.pdf. Último acceso 1/11/2019.

[52] Alooma Team (2018). What is CloverETL?. *Sitio web Alooma*.

https://www.alooma.com/answers/what-is-cloveretl. Último acceso 1/11/2019.

[53] GetApp. Tableau vs Pentaho vs Pentaho Comparision Chart. *Sitioweb GetApp.*

https://www.getapp.com/business-intelligence-analytics-software/a/tableau-software/compare/qlikview-9-vs-pentaho/#features. Último acceso 1/11/2019.

[54] Choudhury A (2019). Jupyter vs Zeppelin: A comprehensive comparision of notebooks. *Sitio web Analyticsindamag.*

https://analyticsindiamag.com/jupyter-vs-zeppelin-a-comprehensive-comparison-of-notebooks/. Último acceso 1/11/2019.

[55] Calvo D (2018). Comparativa Kafka, Flume y RabbitMQ. *Sitio web Diego Calvo*.

http://www.diegocalvo.es/comparativa-kafka-flume-rabbitmq/. Último acceso 1/11/2019.

[56] Bearman C (2018). To Sqoop, or Not to Sqoop? That is the Question. *Sitio web Qlik Blog*.

https://blog.qlik.com/to-sqoop-or-not-to-sqoop-that-is-the-question. Último acceso 14/11/2019.

[57] Siciiani T (2017). Big Data Ingestion: Flume, Kafka, and NiFi. *Sitio web DZone*.

https://dzone.com/articles/big-data-ingestion-flume-kafka-and-nifi. Último acceso 14/11/2019.

[58] EDUCBA. 5 Most Important Difference Between Apache Kafka vs Flume. *Sitio web EDUCBA*.

https://www.educba.com/apache-kafka-vs-flume/. Último acceso 14/11/2019.

[59] Abhimanyu (2019). Logstash vs Fluentd – Which one is better! *Sitio web TechManyu*.

https://www.techmanyu.com/logstash-fluentd-which-one-is-better/. Último acceso 14/11/2019.

[60] EDUCBA. Apache Nifi vs Apache Spark – 9 Useful Comparision To Learn. *Sitio web EDUCBA*.

https://www.educba.com/apache-nifi-vs-apache-spark/. Último acceso 14/11/2019.

[61] Xu S (2018). Workflow Processing Engine Overview 2018: Airflow vs Azkaban vs Conductor vs Oozie vs Amazon Step Functions. *Sitio web Medium*.

https://medium.com/@xunnan.xu/workflow-processing-engine-overview-2018-airflow-vs-azkaban-vs-conductor-vs-oozie-vs-amazon-step-90affc54d53b. Último acceso 14/11/2019.

[62] Aller M (2016). Diseño e implementación de una infraestructura Big Data para análisis de mercados financieros. *Trabajo de Fin de Grado, Universidad Carlos III de Madrid*.

https://e-archivo.uc3m.es/bitstream/handle/10016/27031/TFG_Miguel_Aller_Camino.pdf. Último acceso 18/11/2019.

[63] *Sitio web coservit*.

https://coservit.com/servicenav/wp-content/uploads/sites/3/2018/10/BIG-DATA.png. Último acceso 15/11/2019.

[65] Peterson B (2018). 6 essential steps to the data mining process. *Sitio web BarnRaisers*.

https://barnraisersllc.com/2018/10/data-mining-process-essential-steps/. Último acceso 29/11/2019.

[66] Shah D (2017). Data mining. *Sitio web Towards Data Science*.

https://towardsdatascience.com/data-mining-tools-f701645e0f4c. Último acceso 29/11/2019.

[67] DataFlair Team (2018). 19 Best Data Mining tool – Open Source Tools & Techniques. *Sitio web Data Flair*.

https://data-flair.training/blogs/data-mining-tools-techniques/. Último acceso 29/11/2019.

[68] Digital Guide (2017). Data mining tools for better data analysis. *Sitio web Digital Guide*.

https://www.ionos.com/digitalguide/online-marketing/web-analytics/a-comparison-of-data-mining-tools/. Último acceso 29/11/2019.

[69] StackShare. Jupyter vs RStudio. *Sitio web StackShare*.

https://stackshare.io/stackups/jupyter-vs-rstudio. Último acceso 29/11/2019.

[70] *Sitio web H20*.

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html. Último acceso 29/11/2019.

[71] *Sitio web TrustRadius*.

https://www.trustradius.com/machine-learning. Último acceso 29/11/2019.

[72] Equipo Software Testing Help (2019). 11 Most Popular Machine Learning Software Tools in 2019. *Sitio web Software Tools in 2019*.

https://www.softwaretestinghelp.com/machine-learning-tools/. Último acceso 29/11/2019.

[73] Steeves J. A Plethora of Tools for Machine Learning. *Sitio web Knowm*.

https://knowm.org/machine-learning-tools-an-overview/. Último acceso 29/11/2019.

[74] Equipo RapidMiner. Logging and Monitoring. *Sitio web RapidMiner*.

https://docs.rapidminer.com/9.4/radoop/troubleshooting/logging-and-monitoring.html. Último acceso 2/12/2019.

[75] Chou L (2019). 9 Data Visualization Tools That You Cannot Miss in 2019. *Sitio web Towards Data Science*.

https://towardsdatascience.com/9-data-visualization-tools-that-you-cannot-miss-in-2019-3ff23222a927. Último acceso 29/11/2019.

[76] *Sitio web Leaflet*.

https://leafletjs.com/. Último acceso 2/12/2019.

[77] Equipo de HighCharts. How to learn HighCharts. *Sitio web HighCharts*.

https://www.highcharts.com/blog/post/how-to-learn-highcharts/. Último acceso 2/12/2019.

[78] AGESIC (2019). Qué es REDuy. *Sitio web Presidencia de la República Oriental del Uruguay*.

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/que-es-reduy. Último acceso 2/12/2019.

[79] IMPO (2008). Ley N° 18.381. *Sitio web IMPO*.

https://www.impo.com.uy/bases/leyes/18381-2008. Último acceso 2/11/2019.

[80] AGESIC, Presidencia de la República Oriental del Uruguay (2019). Colombia conexión: entrevista con la Viceministra de tecnología. *Sitio web Presidencia de la República Oriental del Uruguay*.

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/noticias/colombia-conexion-entrevista-con-la-viceministra-de-tecnologia. Último acceso 15/01/2020.

[81] *Sitio web de la UNECE*.

https://statswiki.unece.org/display/bigdata. Último acceso 21/02/2020.

[82] *Sitio web Transport for London*.

https://onderwijs.vlaanderen.be/. Último acceso 21/02/2020.

[83] *Sitio web Transport for London*.

https://www.vdab.be/english/vdab.shtml. Último acceso 21/02/2020.

[84] *Sitio web Transport for London*.

https://tfl.gov.uk/. Último acceso 21/02/2020.

[85] *Sitio web Republic of Estonia.*

https://www.emta.ee/eng/contacts-and-about-us/structure-tasks-strategy-board/introduction-and-structure. Último acceso 21/02/2020.

[86] Friedman T, Heudecker N (2020). Data Hubs, Data Lakes and Data Warehouse: How They Are Different and Why They Are Better Together. *Sitio web Gartner.*

https://www.gartner.com/document/3980938?ref=solrAll&refval=242010046. Último acceso 26/02/2020.

[87] *Sitio web Gartner.*

https://www.gartner.com/en/glossary. Último acceso 26/02/2020.

[88] "Ciberseguridad ¿Estamos preparados en América Latina y el Caribe?" Informe Ciberseguridad 2016, Banco Interamericano de Desarrollo y Organización de los Estados Americanos, 2016.

[89] Estadísticas CERTuy.

https://www.gub.uy/centro-nacional-respuesta-incidentes-seguridad-informatica/datos-y-estadisticas/estadisticas. Último acceso 01/11/2019.

[90] A. Refsdal, B. Solhaug, and K. Stølen, "Cyber-risk management". Springer, 2015.

[91] Open Web Application Security Project (OWASP). *Sitio web OWASP*.

https://owasp.org/#. Último acceso 3/2/2020.

[92] CERT.uy, página oficial.

https://www.gub.uy/centro-nacional-respuesta-incidentes-seguridad-informatica/. Último acceso 11/2019.

[93] Decreto N° 452/009.

https://www.impo.com.uy/bases/decretos/452-2009. Último acceso 4/2/2020.

[94] Gartner. *Market Guide for Data Masking. Noviembre 2019.*

[95] OWASP, Transport Layer Protection Cheat Sheet.

https://cheatsheetseries.owasp.org/cheatsheets/Transport_Layer_Protection_Cheat_Sheet.html. Último acceso 10/2/2020.

[96] S.Frankel, P.Hoffman, A.Orebaugh, R. Park. NIST: Guide to SSL VPNs.

https://csrc.nist.gov/publications/detail/sp/800-113/final. Último acceso 10/2/2020.

[97] E.Barker, Q.Dang, S.Frankel, K.Scarfone, P.Wouters. NIST: Guide to IPsec VPNs.

https://csrc.nist.gov/publications/detail/sp/800-77/rev-1/draft. Último acceso 10/2/2020.

[98] Kimball Group. Dimensional Modeling Techniques. *Sitio web Kimball Group*.

https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/dimensional-modeling-techniques/. Último acceso 05/03/2020.

[99] Dertiano V (2016). Arquitectura BI (Parte IV): Comparativa entre Inmon y Kimball. *Sitio web bigeek*.

https://blog.bi-geek.com/arquitectura-comparativa-inmon-y-kimball/. Último acceso 05/03/2020.

[100] AGESIC (2016). Modelo de Referencia de Metadatos de Personas. *Sitio web Presidencia de la República Oriental del Uruguay.*

https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/modelo-referencia-metadatos-personas. Último acceso 06/03/2020.

[101] Amazon. Model Fit: Underfitting vs Overfitting. *Sitio web Amazon.*

https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html. Último acceso 06/03/2020.

[102] *Sitio web ML Wiki*.

http://mlwiki.org/index.php/Overfitting. Último acceso 2/12/2019.

[103] *Sitio web Scrum.*

https://www.scrum.org/resources/what-is-scrum. Último acceso 06/03/2020.

[104] What is Root Mean Square Error (RSME). *Sitio web Statics how to.*

https://www.statisticshowto.datasciencecentral.com/rmse/. Último acceso 05/03/2020.

[105] Hayes A (2020). R-Squared Definition. *Sitio web Investopedia.*

https://www.investopedia.com/terms/r/r-squared.asp. Último acceso 05/03/2020.

[106] Understanding Precision, Accuracy and Basic Statics.

https://blog.pocd.com.au/scientific/understanding-precision-accuracy-and-basic-statistics/. Último acceso 05/03/2020.

[107] Ensemble Methods in Machine Learning. *Sitio web Educba.*

https://www.educba.com/ensemble-methods-in-machine-learning/. Último acceso 05/03/2020.

[108] Database Optimization Techinques #1: Indexing. *Sitio web OptimizDBA.*

https://optimizdba.com/database-optimization-techniques-1-indexing/. Último acceso 06/03/2020.

[109] Mashiach A (2019). Partition Managment in Hadoop. *Sitio web Cloduera Blog.*

https://blog.cloudera.com/partition-management-in-hadoop/. Último acceso 06/03/2020.

[110] Calvo D (2017). Tipos de datos: estructurados, semiestructurados y no estructurados. *Sitio web Diego Clavo.*

http://www.diegocalvo.es/tipos-de-datos-estructurados-semiestructurados-y-no-estructurados/. Último acceso 06/03/2020.

[111] Rençberoğlu, Emre (2019). Fundamental Techniques of Feature Engineering for Machine Learning. Sitio web *Towards data science*.

https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114. Último acceso 13/03/2020.

Anexo I: Caso de estudio de Big Data

Típicamente, las arquitecturas de Big Data se diseñan sobre Hadoop, por lo que es importante explicar cómo se implementa su infraestructura.

Antes de ahondar en dicha arquitectura, cabe destacar que Apache Hadoop es un Framework Open Source que permite el procesamiento distribuido de grandes volúmenes de datos a través de un clúster. Su arquitectura permite escalabilidad y detección de errores en la capa de aplicación. Además, es tolerante a fallos. También existen en el mercado otras alternativas, emergiendo al día de hoy, a Hadoop como plataforma de Big Data.

Una de las alternativas es OpenDataHub (proyecto Open Source) que posee como componentes principales OpenShift y Ceph. En este caso se cuenta con el soporte de licenciamiento de RedHat. La desventaja es que el proyecto aún está en proceso de crecimiento.

Otras alternativas consisten en la combinación de otras herramientas substitutas para los componentes de Hadoop. Por ejemplo, cambiar el almacenamiento HDFS por otra herramienta como ser Cassandra o Ceph, combinando con herramientas como Apache o Apache Storm para el procesamiento. El problema de estas alternativas es que no se cuenta con una licencia donde se incluya soporte a la plataforma.

Retomando Hadoop, su naturaleza distribuida brinda la posibilidad de modificar los datos de acuerdo a lo que se necesite, pudiendo tener almacenado el mismo dato en distintos formatos. Lo cual implica un cambio de paradigma para los organismos, ya que importa más la rapidez de procesamiento ante la consistencia.

Consta de cuatro módulos principales [15]:

- **Hadoop Common:** tiene las librerías de java que dan soporte al resto de los módulos.
- **Hadoop Distributed File System (HDFS):** sistema de archivos distribuido, permite que los datos sean accedidos en alto rendimiento. Tiene una alta tolerancia a errores y funciona en un Hardware de bajo costo.
- **Hadoop Yarn:** framework encargado de la planificación de tareas y la gestión de los recursos del clúster.
- **Hadoop MapReduce:** permite el procesamiento paralelo de datos.

En la Figura 24 se muestra el diagrama de la arquitectura Hadoop. En dicha arquitectura se define un conjunto de componentes necesarios para el correcto funcionamiento de Hadoop:

- **Namenode:** gestiona el HDFS, sabe en dónde se encuentran los ficheros, pero no los almacena.
- **Secondary Namenode:** gestiona al nodo Namenode, está en una máquina diferente.
- **Datanode:** almacena los datos del HDFS. Típicamente se varios de estos nodos que tienen los datos replicados.

- **Resource Manager:** gestiona los recursos del clúster y las aplicaciones que se ejecutan en el módulo Yarn.
- **Node Manager:** mediante la comunicación con el componente Resource Manager, gestiona los recursos de los nodos.

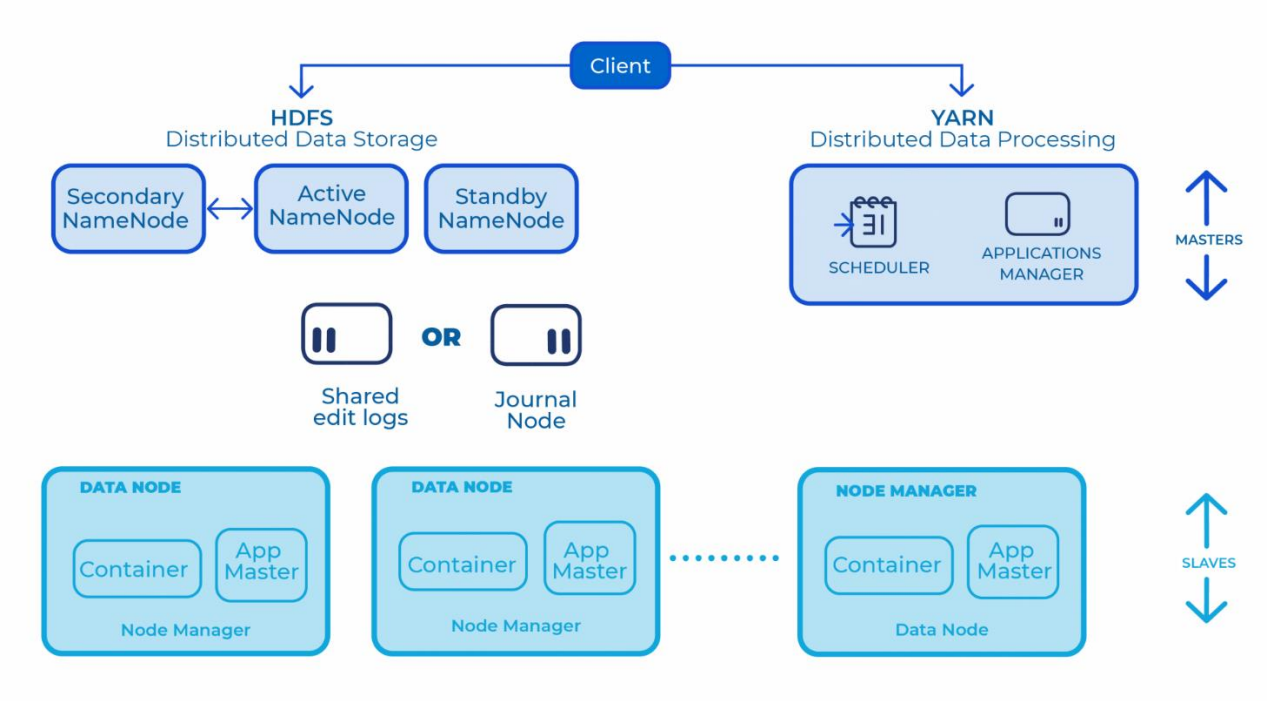

Figura 24 Arquitectura Hadoop. Fuente: [15]

Anexo II: Análisis de posibles escenarios de aplicación

En esta sección se instancian distintos escenarios posibles para el proceso de Análisis de Datos y se enumeran para los considerados más relevantes, el conjunto de requerimientos que se sugiere tener en cuenta para su correcta implementación.

Se organizan los distintos escenarios en el diagrama de flujo mostrado en la Figura 25 Posibles escenarios, según las siguientes variables de decisión:

1. Ubicación almacenamiento: Lugar donde están almacenados físicamente los datos.
2. Ubicación Procesamiento: Lugar donde se realiza el procesamiento de los mismos, incluyendo método y canal de transferencia (A dónde se envían, cómo y por qué canal).
3. Análisis: Quién realiza el análisis de los datos.

A continuación, se explican los valores posibles para las variables definidas anteriormente:

- Ubicación de almacenamiento:
  - **Organismo:** Los datos se encuentran almacenados en el organismo.
  - **Proveedor:** Los datos se encuentran almacenados en un proveedor externo al organismo.
  - **Nube de Gobierno:** Los datos se encuentran almacenados en el servicio de Nube de Presidencia de la República brindado por Agesic.
- Ubicación procesamiento:
  - **Organismo:** Los datos son procesados por el organismo.
  - **Proveedor:** Los datos son procesados por un proveedor externo al organismo.
- Análisis:
  - **Organismo:** El análisis de los datos es llevado a cabo por personal del organismo.
  - **Proveedor:** El análisis de los datos es llevado a cabo por un proveedor externo al organismo.

Nota: Si existen fuentes provenientes de más de una ubicación (según la variable 1), se deben tener en cuenta los conjuntos de requerimientos de todos los escenarios correspondientes.

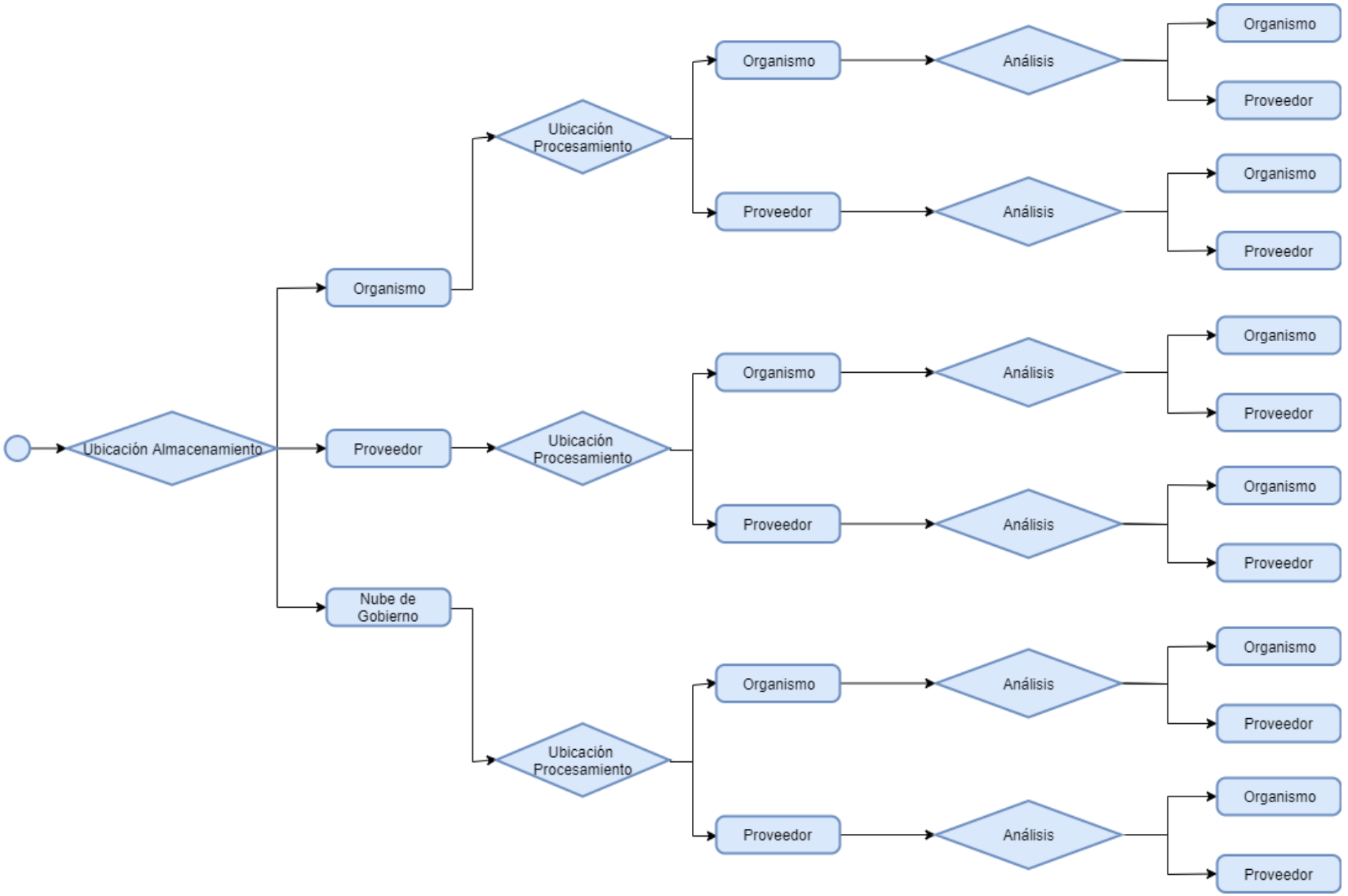

Figura 25 Posibles escenarios

Anexo III: Profundización Buenas Prácticas de Infraestructura y Ciberseguridad

En el presente anexo se describen buenas prácticas a nivel de Infraestructura y Ciberseguridad que pueden ser clasificadas de la siguiente forma:

1. Extensión de buenas prácticas del Checklist (Ciberseguridad)
2. Buenas prácticas complementarias al análisis de datos (Infraestructura)

III.1. Extensión de buenas prácticas del checklist (Ciberseguridad)

A continuación, se extienden algunas buenas prácticas presentadas en el checklist, se agrupan según la clasificación Prevención (P), Detección y Control (DC) y Resiliencia (R). Además, cada buena práctica se presenta con la numeración utilizada en el checklist.

Prevención (P)

Las buenas prácticas se presentan según la siguiente clasificación:

- Requerimientos.
- Clasificación y manejo de Datos e Información.

Requerimientos

P2. Fuentes de Requerimientos

En la etapa de análisis, es necesario tener en cuenta desde el punto de vista de la seguridad de datos:

- **Partes interesadas:** Identificar las necesidades de privacidad y confidencialidad de las organizaciones o grupos de personas, que están siendo consideradas como partes interesadas en el sistema.
- **Información crítica[21] y reglas de acceso según normativa:** identificar y tener en consideración la información que es considerada como crítica de cada organización que integrará el sistema, así como las reglas de acceso definidas para la información considerada.
- **Necesidades de acceso:** relevar las necesidades de acceso a los datos de cada organización o grupos de personas participantes y en base a esto definir los accesos legítimos que podrán ser realizados en el sistema. Evaluar si para acceder a datos específicos es necesario obtener el consentimiento de alguna de las partes involucradas.

- **Obligaciones contractuales:** las obligaciones contractuales y acuerdos de no divulgación a los que están sometidos las distintas organizaciones participantes. En particular, considerar este punto cuando se establezcan contratos con otras organizaciones.

Asegurar que los requisitos relevados son tomados como insumo en la etapa de diseño e implementación del sistema. Esta información también puede ser tenida en cuenta en la selección de proveedores o paquetes de software.

Clasificación y manejo de Datos e Información

P3. Clasificación de información y preprocesamiento

Una vez definidos los datos a ser incluidos en el sistema, estos deben ser clasificados con el fin de definir los niveles de protección. Este proceso debería contener las siguientes etapas:

- Identificar y clasificar los activos que contienen información crítica[23] del sistema.
- Localizar los datos críticos, estableciendo sus lugares de almacenamiento. Observar que los requerimientos de seguridad pueden diferir según el lugar de almacenamiento de los datos. Por ejemplo, si una cantidad significativa de información crítica[24] se encuentra almacenada en un mismo lugar, el riesgo asociado a una fuga de información puede ser elevado.
- Determinar cómo proteger los distintos activos, las medidas de protección dependen del tipo de datos almacenados, así como del tipo de tecnología utilizada.

Los desarrolladores tanto de partes del sistema, así como de dashboards o reportes deben saber cómo se clasifica para los distintos esquemas la agregación de distintos tipos de datos según la clasificación individual de los mismos. Según la clasificación de la agregación se pueden tomar medidas para tratar la información resultante.

P3.1 Clasificación de información según legislación uruguaya

A continuación, se detalla la clasificación de información definida en la legislación nacional (Ley N° 18381 [79]). Únicamente se incluyen las

[23] En este marco se considera información crítica, a la clasificada como información reservada o confidencial según la legislación nacional. Para ver esta clasificación, consultar el punto "P5.1" en la sección "Anexo II: Profundización Buenas Prácticas".

[24] En este marco se considera información crítica, a la clasificada como información reservada o confidencial según la legislación nacional. Para ver esta clasificación, consultar el punto "P5.1" en la sección "Anexo II: Profundización Buenas Prácticas".

características que debe cumplir la información por cada categoría, para profundizar sobre quién debe realizar la clasificación y otros detalles vinculados a la legislación referirse a la Ley.

Información pública (Artículo 2)

Se considera información pública toda la que emane o esté en posesión de cualquier organismo público, sea o no estatal, salvo las excepciones o secretos establecidos por ley, así como las informaciones reservadas o confidenciales.

Información reservada (Artículo 9)

Como información reservada podrá clasificarse aquella cuya difusión pueda:

1. Comprometer la seguridad pública o la defensa nacional.
2. Menoscabar la conducción de las negociaciones o bien, de las relaciones internacionales, incluida aquella información que otros estados organismos internacionales entreguen con carácter de reservado al Estado uruguayo.
3. Dañar la estabilidad financiera, económica o monetaria del país.
4. Poner en riesgo la vida, la dignidad humana, la seguridad o la salud de cualquier persona.
5. Suponer una pérdida de ventajas competitivas para el sujeto obligado o pueda dañar su proceso de producción.
6. Desproteger descubrimientos científicos, tecnológicos o culturales desarrollados o en poder de los sujetos obligados.
7. Afectar la provisión libre y franca de asesoramientos, opiniones o recomendaciones que formen parte del proceso deliberativo de los sujetos obligados hasta que sea adoptada la decisión respectiva, la cual deberá estar documentada.

Información confidencial (Artículo 10)

Se considera información confidencial:

1. Aquella entregada en tal carácter a los sujetos obligados, siempre que:
    a. Refiera al patrimonio de la persona.
    b. Comprenda hechos o actos de carácter económico, contable, jurídico o administrativo, relativos a una persona física o jurídica, que pudiera ser útil para un competidor.
    c. Esté amparada por una cláusula contractual de confidencialidad.

2. Los datos personales que requieran previo consentimiento informado.

Tendrán el mismo carácter los documentos o secciones de documentos que contengan estos datos.

P3.2 Usando Metadatos para clasificación de información

Es posible hacer uso de metadatos para clasificar la información del sistema. Esto puede realizarse tanto para datos individuales, así como para conjuntos de datos.

Es deseable almacenar los criterios y niveles de clasificación utilizados en un repositorio accesible por todos los usuarios del sistema de Análisis de Datos. Adicionalmente, al momento de generar reportes, una herramienta podría posibilitar el etiquetado de los archivos correspondientes según los distintos niveles de confidencialidad. También se puede realizar el etiquetado manual de los archivos en caso de que no se cuente con la funcionalidad anterior.

P4. Enmascaramiento de Datos

Las técnicas de enmascaramiento u ofuscación de datos tienen como objetivo ocultar parcial o totalmente la información, haciendo uso de técnicas de mezclado, remoción, o cambios en la apariencia de los datos sin perder de forma permanente su significado o las relaciones entre ellos. Estas técnicas suelen ser de utilidad cuando el objetivo es evitar mostrar datos críticos en un sistema o generar un conjunto de datos para un ambiente de pruebas.

Existen dos tipos de técnicas de enmascaramiento:

- **Enmascaramiento Persistente:** los datos son alterados de manera permanente, es decir que el proceso es irreversible.
- **Enmascaramiento Dinámico:** altera la apariencia de los datos sin cambiar de forma permanente los datos subyacentes.

Las técnicas pueden ser implementadas mediante los siguientes métodos:

- **Sustitución:** remplazar uno o más caracteres de un dato por otro valor, que puede ser fijo o aleatorio.
- **Mezclado:** intercambiar el dato de un registro por otro del mismo tipo.
- **Varianza temporal:** realizar un corrimiento temporal en un rango acotado.
- **Varianza de valor:** combinar el dato con un valor en un rango acotado.

- **Borrado:** eliminar un dato determinado.
- **Aleatorización:** reemplazar parte o todo un dato con valores aleatorios.
- **Cifrado:** cifrar los datos.
- **Enmascarado con expresión:** cambiar los valores de un dato por el resultado de una expresión, que puede ser fija. Por ejemplo, sustituir todos los valores de un campo determinado por la expresión “Información crítica”.

Ejemplos

A continuación, se muestran ejemplos de los distintos tipos de técnicas tomando sobre un esquema relacional de base de datos.

Supongamos que el esquema contiene las tablas que se muestran a continuación con los siguientes datos:

Id cliente	CI	Nombre	Dirección	Ciudad	País	Código postal
1	5233423	Juan Pérez	…	Montevideo	UY	…
2	4326743	Nicolas Gómez	…	Montevideo	UY	…
3	4299031	Alberto González	…	San Pablo	BR	…

Tabla 21 Tabla clientes

Id producto	Nombre	Precio unitario
1	P1	200
2	P2	150
3	P3	175

Tabla 22 Tabla productos

Id vendedor	C.I	Nombre
1	4192201	Felipe Rodríguez
2	4892330	Alicia Fernández

Tabla 23 Tabla vendedores

Id venta	Id cliente	Id vendedor	Id producto	Cantidad	Fecha venta
1	1	2	2	2	21/7/2020
2	2	1	1	5	12/5/2020
3	1	1	2	6	3/9/2020

Tabla 24 Tabla ventas

Sustitución

Por ejemplo, se puede sustituir la C.I de las tablas por valores aleatorios y sustituir los nombres por valores en un conjunto de valores posibles. Para la tabla vendedores se podría obtener:

Id vendedor	C.I	Nombre
1	21093720393	Tom Smith
2	12098420934	John Ford

Tabla 25 Tabla vendedores sustituyendo datos

- Se mantiene la asociación entre los atributos sustituidos y el resto de los campos del registro, pero podría no ser posible establecer tendencias entre los valores de los campos sustituidos. En este caso, la sustitución de nombres en los vendedores no mantiene el género y por tanto no podría establecerse una relación entre el género del vendedor y la cantidad de ventas.

Mezclado

Se intercambian datos entre los valores de los distintos registros.

Id venta	Id cliente	Id vendedor	Id producto	Cantidad	Fecha venta
1	1	1	1	6	21/7/2020
2	2	1	2	5	3/5/2020
3	1	2	2	2	12/9/2020

Tabla 26 Tabla ventas con datos mezclados

- Pueden conservarse tendencias, en este caso se conserva la cantidad total de ventas de cada producto.

- Se pierden relaciones con otros atributos del mismo registro, en este caso la cantidad de ventas por vendedor.

Varianza temporal

Se realiza un corrimiento temporal (hacia adelante o atrás) mensual en el rango {2,3}.

Id venta	Id cliente	Id vendedor	Id producto	Cantidad	Fecha venta
1	1	2	2	2	21/10/2020
2	2	1	1	5	12/8/2020
3	1	1	2	6	3/3/2020

Tabla 27 Tabla ventas con varianza temporal

- Puede no mantener las tendencias temporales, pero sí entre las magnitudes de los datos. En este caso si se evalúa cantidad de unidades vendidas de un producto determinado y se aplica varianza temporal, no se podrá analizar el vínculo entre el mes del año y la cantidad de ventas, pero sí la tendencia general del volumen de las mismas.

Varianza de valor

Se suma a las cantidades vendidas de los productos un valor en el rango {2,5}.

Id venta	Id cliente	Id vendedor	Id producto	Cantidad	Fecha venta
1	1	2	2	7	21/7/2020
2	2	1	1	7	12/5/2020
3	1	1	2	11	3/9/2020

Tabla 28 Tabla ventas con varianza de valor

- La varianza de valor puede afectar los valores individuales de los datos. Por ejemplo, si como en este caso a la cantidad de ventas de productos se le suma un valor fijo, se mantiene la tendencia de evolución, pero se pierde el valor máximo y mínimo del atributo.

Borrado

Se eliminan las columnas distintas de “Id cliente” y “C.I” de la tabla de clientes.

Id cliente	C.I
1	5233423
2	4326743
3	4299031

Tabla 29 Tabla clientes con columnas eliminadas

- Se pierden los datos eliminados, no pudiendo estos ser considerados en el análisis.

Aleatorización

Análogo al caso de sustitución de los valores en el campo “C.I”.

Cifrado

Se cifran los valores del campo “Ciudad” en la tabla clientes.

Id cliente	C.I	Nombre	Dirección	Ciudad	País	Código postal
1	5233423	Juan Pérez	…	WOb��-m|| UY	…
2	4326743	Nicolas Gómez	…	WOb��-m|| UY	…
3	4299031	Alberto González	…	w�n�^�XH'c��Sa	BR	…

Tabla 30 Tabla clientes con cifrado

- Mientras los datos se encuentran cifrados, estos son inaccesibles y en particular no pueden ser usados en el análisis.
- Se agrega la tarea de gestión de claves y de cifrado y descifrado en el almacenamiento y acceso a los datos.

Enmascarado con expresión

Se sustituye todos los valores de un campo determinado por la expresión “Información crítica”.

Id cliente	C.I	Nombre	Dirección	Ciudad	País	Código postal

1	Información crítica	Juan Pérez	…	Montevideo	UY	…
2	Información crítica	Nicolas Gómez	…	Montevideo	UY	…
3	Información crítica	Alberto González	…	San Pablo	BR	…

Tabla 31 Tabla clientes con enascaramiento

- En el caso de una sustitución genérica para todos los valores de un campo o de un subgrupo de estos valores (por ejemplo, los que sean considerados información confidencial), no será posible utilizarlos en el análisis.

P8. Definir roles y privilegios

La asignación debe seguir el Principio de Mínimos Privilegios, es decir que un usuario, proceso o programa solo debe poder acceder a la información que necesita para cumplir con su propósito legítimo.

Detección y control (DC)

DC3. Métricas utilizadas para el análisis de seguridad de datos

Para cada métrica se sugieren definir adicionalmente: umbrales deseables, procesos de monitoreo y de corrección de desviaciones.

III.2. 2. Buenas prácticas complementarias al análisis de datos (Infraestructura)

Se presentan las siguientes buenas prácticas de la dimensión de Ciberseguridad, las cuales no fueron presentadas en el checklist de la sección de Buenas Prácticas porque o bien hacen referencia a herramientas y componentes concretos que pueden ser adicionados al análisis o son pautas para la gestión de tecnología utilizada en el proceso:

3. **Consideraciones asociadas a gestión de Tecnología:** presenta medidas vinculadas al uso de Tecnología en el proceso de Análisis de Datos.
4. **Consideraciones asociadas a herramientas o componentes para Análisis de Datos:** detalla pautas específicas vinculadas a herramientas o componentes a incorporar para la realización de Análisis de Datos.

Las medidas se presentan según si son categorizadas como una medida para “Personas” y “Tecnologías”. La primera hace referencia a puntos que involucren la identificación de personas o la realización de procesos que involucren a las mismas, mientras que la segunda refiere a medidas vinculadas a la tecnología utilizada en el proceso (configuración o buenas prácticas para su utilización).

III.2.1. Consideraciones asociadas a Gestión de Tecnología

En esta sección se presentan medidas vinculadas al uso y gestión de Tecnología en el proceso de Análisis de Datos. Dichas medidas se agrupan según si son de Prevención (P), Detección y Control (DC) o Resiliencia (R), a cada una de las medidas se las referencia por el prefijo “GT.” concatenado con el tipo de agrupación (“P”, “DC”, “R”).

Prevención (GT.P)

Las buenas prácticas se presentan según la siguiente clasificación:

- Estructura organizacional.
- Guía de usuarios y accesos.
- Configuración de herramientas.
- Autenticación.
- Manejo de sesión.
- Segmentación de redes.

Estructura organizacional

Con respecto a la estructura de la organización, se recomienda:

GT.P1. Identificar al Responsable de Seguridad de la Información

Identificar al Responsable de Seguridad de la Información (RSI), así como a los responsables de los datos e involucrarlos en los esfuerzos de aseguramiento de los mismos. Ante un incidente de Seguridad de la Información o necesidad de asesoramiento en el área, tenga presente que estas personas deben ser contactadas. Los objetivos principales de esta medida son:

- En caso de que exista un incidente de ciberseguridad que involucre los datos, se puedan obtener instrucciones de cómo gestionarlo o eventualmente de quiénes deben hacerse cargo del mismo.
- Ante dudas sobre medidas concretas de ciberseguridad, se pueda tener un referente que brinde orientación en la materia o que en su

defecto conozca los riesgos que pueden generarse en base a los datos considerados.

En el Decreto N° 452/009 [93] se establece que toda unidad ejecutora de la Administración Central debe contar con un RSI designado. En general, las personas designadas como responsables de un conjunto de datos son las que conocen su estructura, las necesidades organizacionales a las que responden, dónde están almacenados y eventualmente el vínculo con otros conjuntos de datos.

Según lo que establezca la normativa, la función de seguridad de la información puede ser realizada por un grupo dedicado específicamente a la misma. En medianas y grandes empresas, es habitual que exista la figura de Responsable de Seguridad de la Información (RSI), que puede reportar al Responsable de Información en caso de que el cargo exista, o a la Gerencia General. Notamos que en general, el equipo dedicado a seguridad de la información tiende a estar más orientada a aspectos técnicos del área.

Si la organización no cuenta con un grupo dedicado específicamente a seguridad de la información, esta responsabilidad debe ser asumida por los responsables de los distintos conjuntos de datos. En cualquier caso, es necesario que los responsables de los datos sean involucrados en los esfuerzos realizados para asegurarlos. En particular, los responsables deben ser involucrados activamente con el personal de tecnología, desarrolladores y profesionales de ciberseguridad de modo que sea posible:

- Identificar los datos que deben que se encentran regulados.
- Identificar y proteger apropiadamente los sistemas críticos.
- Diseñar controles para asegurar confidencialidad, integridad y cumplimiento regulatorio relativo a los datos.

Roles involucrados: Arq. de Datos, Ing. de Datos, Analista de Información, Científico de Datos.

Categoría: Personas.

GT.P2. Coordinación áreas – Información y actualizaciones de seguridad

Verificar que el equipo de Seguridad de la Información revisa las nuevas herramientas tecnológicas que se incorporan en el proceso y que controla que estas dispongan de las últimas actualizaciones y parches de seguridad disponibles.

En cuanto a la coordinación entre las áreas que incorporan fuentes de datos y la de Seguridad de la Información considerar que: Cuando se agregan nuevas fuentes de datos, se verifique que estás sean adecuadamente clasificadas y tratadas en consecuencia (incluyendo los controles de Seguridad de la Información).

Roles involucrados: Realización de actividad: RSI; Verificación de actividad: Arq. de Datos, Ing. de Datos, Analista de Información y Científico de Datos.

Categoría: Personas.

Guía de usuarios y accesos

Con respecto a la gestión de usuarios y accesos, se recomienda:

GT.P3. Asegurar credenciales de administración

Asegurar las credenciales de las cuentas de administración y que estas solo pueden ser accedidas en caso de emergencia con la documentación y autorizaciones apropiadas. La actividad de estas cuentas debería ser monitoreada (ver puntos Detección y Control) y una vez utilizadas su validez expirar en un periodo de tiempo reducido.

Roles involucrados: RSI.

Categoría: Personas.

GT.P4. Estandarizar asignación de permisos

En caso de utilizar asignación de permisos de acceso por grupo, estos pueden ser organizados de dos formas: de forma tabular o jerárquica como se explica a continuación.

La forma tabular consiste en asignar en una tabla para cada junto de datos D, los roles que pueden acceder a los datos de ese conjunto (R) según su nivel de confidencialidad (N). A continuación, se muestra un ejemplo:

		Nivel de confidencialidad		
		N1	N2	N3
Conjunto de datos	D1	R11	R12	R13
	D2	R21	R22	R23

Tabla 32 Niveles de confidencialidad

En este ejemplo podría ocurrir, D1=Datos Historias Clínicas con nivel N3=Confidencial que pueden ser accedidos únicamente por los usuarios con roles R13=Ministerio de Salud Pública (MSP).

En la forma jerárquica se organizan los roles en un árbol donde los roles hijos tienen menos permisos de acceso que su rol padre en el árbol.

Roles involucrados: RSI.

Categoría: Personas.

Configuración de herramientas

Con respecto a la configuración de herramientas, se recomienda:

GT.P5. Configuración adecuada de herramientas

Asegurar que se eliminan las cuentas de administración por defecto de las herramientas utilizadas, que las credenciales establecidas por defecto son cambiadas, las funcionalidades o servicios no utilizados son deshabilitaos y que los mensajes de error no revelan información interna del sistema.

Roles involucrados: RSI.

Categoría: Tecnología.

Autenticación

Con respecto al proceso de autenticación, se recomienda:

GT.P6. Política segura Contraseñas

Hacer uso de una política segura para definir contraseñas.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P7. Nombres de usuarios únicos y case insensitive

Utilizar nombres de usuarios que no distingan entre mayúsculas y minúsculas (case insensitive). Los nombres de usuarios deberían ser únicos.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P8. Validación de campos

Realizar validación de campos cuando corresponda, tanto sintáctica como semánticamente.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P9. Mecanismo seguro de cambio de contraseña

Implementar un mecanismo seguro para recuperación de contraseña, en particular deshabilitar todas las sesiones previas cuando un cambio de contraseña se realiza de forma exitosa.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P10. Almacenamiento seguro de contraseñas

Almacenar las contraseñas de forma segura.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P11. Transmisión segura de contraseñas

Transmitir contraseñas usando TLS, SSL u otro protocolo seguro. Esto también aplica para la transmisión del Id de sesión, como se explica luego en esta sección o cualquier otro dato crítico que se transmita desde la aplicación.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P12. 2FA para aumentar seguridad

Hacer uso de autenticación de dos factores (2FA, por sus siglas en inglés) cuando se quiera aumentar el nivel de seguridad.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P13. Mensajes de errores genéricos

En el caso de que el usuario o la contraseña proporcionadas por un usuario sean incorrectas, la aplicación debe responder con un mensaje de error

genérico, sin importar si es el nombre de usuario o la contraseña el valor incorrecto. Tampoco debe indicar si existe una cuenta asociada al nombre de usuario proporcionado.

Roles involucrados: RSI.

Categoría: Tecnología.

Manejo de sesión

Con respecto al manejo de sesiones de usuarios, se recomienda:

GT.P14. Campo id de sesión

Hacer uso de un nombre que no sea demasiado descriptivo para el campo que contiene el Id de sesión. Se debe evitar que sea posible inferir del nombre del campo de Id de sesión alguna herramienta utilizada para el desarrollo (Por ejemplo, el Framework de alguna de las herramientas utilizadas). Se recomienda cambiar el nombre por defecto del campo a un nombre genérico como “id”.

El Id de sesión debe ser suficientemente largo con el fin de evitar un ataque por fuerza bruta que permita a un atacante verificar la existencia de sesiones válidas.

El Id de sesión no debe contener información relevante sobre el usuario correspondiente.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P15. Transmisión segura id de sesión

Transmitir el Id de sesión usando TLS, SSL u otro protocolo seguro.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P16. Tiempo para inactivar id de sesión

Definir un período de tiempo a partir del que se inhabilita una sesión en el caso de inactividad.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.P17. Terminación de sesión

Cuando un usuario se “desloguea” del sistema, se debería terminar la sesión existente asociada.

Roles involucrados: RSI.

Categoría: Tecnología.

Segmentación de redes

Con respecto a la segmentación de redes, se recomienda:

GT.P18. Segmentación de red

Segmentar la red en subredes según los requerimientos específicos de seguridad de cada parte del sistema. Esto busca prevenir que, si un atacante logra vulnerar las defensas perimetrales de la red, gane acceso al resto de los dispositivos de la misma, quedando la amenaza contenida en un segmento determinado. En general, se definen segmentos llamados “Zonas Desmilitarizadas” (DMZ por sus siglas en inglés) que contienen dispositivos expuestos a internet que no deberían estar en el mismo segmento que la red interna de la organización.

A continuación, se detalla una posible segmentación:

1. Un segmento para los servidores Web, Reverse Proxies u otros que reciban conexiones directas desde internet (DMZ).
2. Un segmento con los servidores de aplicaciones.
3. Un segmento con los servidores de base de datos.
4. Un segmento con los servidores internos a la organización (controladores de dominio, DNS, proxy, etc.).
5. Un segmento para los dispositivos de trabajo de la organización

Con respecto a la comunicación entre los segmentos anteriores se tiene:

- Tráfico bidireccional entre el segmento 1 e Internet.
- Tráfico unidireccional del segmento 1 al segmento 2.
- Tráfico unidireccional del segmento 2 al segmento 3.
- Tráfico unidireccional del segmento 4 hacia Internet.
- Tráfico unidireccional del segmento 5 al segmento 4.

La separación de segmentos puede implementarse mediante Firewalls.

Roles involucrados: RSI.

Categoría: Tecnología.

Detección y control (DC)

Las buenas prácticas se presentan según la siguiente clasificación:

- Reportes de fallas o anomalías.
- Gestión de usuarios y accesos.

Reportes de fallas o anomalías

Con respecto de fallas o anomalías, se recomienda:

GT.DC1. Política reporte de fallas

Establecer una política para reportar fallas o vulnerabilidades de seguridad del sistema de Análisis de Datos.

Roles involucrados: RSI.

Categoría: Personas.

Gestión de usuarios y accesos

Con respecto a la gestión de usuarios y accesos, se recomienda:

GT.DC2. Preferir acceso basado en roles/grupos

Preferir el acceso basado en roles o grupos sobre el basado en permisos a nivel individual usuario (Ver HC.P2).

Preferir el acceso basado en roles o grupos sobre el basado en permisos a nivel individual de cada usuario. Esto permite a un administrador de seguridad definir los privilegios para cada grupo de acuerdo a los distintos roles en la organización y luego asignar a los distintos usuarios a los grupos que les correspondan.

Si bien es posible asignar a un usuario a más de un grupo, esta práctica hace que se dificulte comprender de manera unificada los privilegios de acceso concedidos a un usuario específico. Por tanto, se recomienda que de ser posible se asigne a cada usuario a un único grupo de acceso.

Roles involucrados: RSI.

Categoría: Tecnología.

GT.DC3. Control niveles de privilegio

Un proceso para implementar la política de control de niveles de privilegio puede consistir en:

- Validar asignación de permisos, verificando que para cada permiso asignado existe una solicitud del usuario en el sistema de gestión de cambios para dicho permiso.
- Requerir un proceso de aprobación de permisos para documentar cada solicitud de cambio.
- Asegurar la existencia de un proceso para eliminar los permisos de las personas que ya no pertenecen a la organización o que cambiaron de departamento o sección dentro de la misma con otros permisos de acceso.

Roles involucrados: RSI.

Categoría: Personas.

Resiliencia (R)

Las buenas prácticas se presentan según la siguiente clasificación:

- Respaldo de datos.
- Planes de recuperación.
- Respuesta a incidentes.

Respaldo de datos

Con respecto al respaldo de datos se recomienda:

GT.R1. Plan periódico de respaldo (Proyectos)

Definir un plan periódico de respaldo del sistema de Análisis de Datos, contemplando la plataforma y los datos almacenados. Verificar si los datos sobre los que se está realizando el análisis deben ser respaldados y si el respaldo se está llevando a cabo de forma correcta. En caso de que el mismo no se esté realizando correctamente, contactar al responsable de seguridad de los datos.

Roles involucrados: RSI, Analista de Información, Científico de Datos.

Categoría: Personas.

Planes de recuperación

Con respecto a los planes de recuperación, se recomienda:

GT.R2. Estrategia de recuperación

Verificar si existe una estrategia de recuperación y de manejo de fugas de información. En particular, deberían existir estrategias de recuperación para centros de datos, salas de servidores y aplicaciones.

Roles involucrados: RSI, Analista de Información, Científico de Datos.

GT.R3. Guías de restauración

Que los planes y procedimientos de recuperación incluyan guías y procedimientos detallados para restaurar un centro de datos, sistemas, redes y/o aplicaciones. Familiarizarse con los planes y procedimientos que aplican a las herramientas que utiliza para el análisis.

Roles involucrados: RSI, Analista de Información, Científico de Datos.

Categoría: Personas.

GT.R4. Pruebas estrategias y planes de recuperación

Realizar pruebas periódicas de las estrategias y los planes de recuperación.

Roles involucrados: RSI.

Categoría: Personas.

Respuesta a incidentes

Con respecto a la respuesta a fallas o incidentes de seguridad en el sistema de Análisis de Datos, se recomienda:

GT.R5. Procesos de respuesta a fallas

Establecer procesos para recibir, analizar y responder a fallas o vulnerabilidades de seguridad (Ver GT.DC1) reportadas. Recordar la obligatoriedad de todos los organismos públicos de reportar los incidentes de Seguridad de la Información al CERTuy (decreto 451/009).

Roles involucrados: RSI.

Categoría: Personas.

III.2.2. Consideraciones asociadas a herramientas o componentes para Análisis de Datos

En esta sección se presentan medidas vinculadas a herramientas o componentes a incorporar para la realización de Análisis de Datos. Dichas medidas se agrupan según si son de Prevención (P), Detección y Control (DC) o Resiliencia (R), a cada una de las medidas se las referencia por el prefijo "HC." concatenado con el tipo de agrupación ("P", "DC", "R").

Prevención (P)

Las buenas prácticas se presentan según la siguiente clasificación:

- Clasificación y manejo de Datos e Información.
- Autenticación.
- Tests de penetración.

Clasificación y manejo de Datos e Información

Con respecto a la clasificación y manejo de datos e información, se recomienda:

HC.P1. Gestión de riesgos

Identificar los riesgos[25] externos e internos a los que está expuesto el sistema de Análisis de Datos. En particular, asegurarse de que toda la información crítica[21] utilizada en el análisis es necesaria y evitar su uso en ambientes no productivos, dado que estos suelen ser más vulnerables. Considerar generar conjuntos de datos de prueba que no contengan información crítica. Si se considera que un subconjunto de los datos es innecesario para el análisis, estos no deben ser incluidos.

Además de realizar la clasificación de la información que se incorpora al sistema, es necesario evaluar los riesgos a los que está expuesto el sistema, como se mencionó. Esto incluye tanto los riegos externos (generados por agentes externos al sistema), así como los internos (generados por agentes internos al sistema como el personal y los procesos de la organización).

Análisis de riesgos: para cada riesgo identificado, se debe definir su probabilidad e impacto de ocurrencia, junto con las escalas correspondientes que serán utilizadas. De la probabilidad y el impacto, se debe definir una regla para calcular la severidad del riesgo. Además, puede definirse una estimación para el costo aproximado para remediar el eventual daño si se incurre en el

[25] Consultar las categorías "Evaluación de riesgos" y "Estrategia para gestión de riesgos" del Marco de Ciberseguridad [8].

riesgo y el costo para remediarlo o mitigarlo. Se debe definir un proceso formal de priorización de riesgos que involucre a las partes interesadas del sistema.

Roles involucrados: Arq. Análisis de Datos, Especialista / Ing. de Datos, Analista de Información, Científico de Datos.

Categoría: Personas.

HC.P2. Monitoreo de cuentas genéricas y de servicios

Asegurar que se monitorean el uso y actividades de las cuentas genéricas y de servicios, que la asignación de privilegios se da según el Principio de Mínimos Privilegios y que en caso de las cuentas sean asignadas a un usuario este cuenta con las autorizaciones necesarias. Al evaluar herramientas para incorporar en el sistema de Análisis de Datos buscar que estás tengan las funcionalidades para cumplir este punto. En particular, definir la actividad de qué cuentas debe ser monitoreada y las herramientas con las que se realizará el control.

El uso de cuentas genéricas y de servicios aumenta el riesgo de fugas de datos y dificulta la trazabilidad de la persona responsable en caso de un incidente. Se sugiere que las cuentas genéricas no sean usadas por defecto.

Roles involucrados: RSI.

Categoría: Personas.

Autenticación

Con respecto al proceso de autenticación, se recomienda:

HC.P3. Deshabilitar acceso, múltiples intentos incorrectos

Prevenir ataques por fuerza bruta deshabilitando temporalmente el acceso en el caso de múltiples intentos de autenticación incorrectos. Considerar el tiempo por el que las cuentas permanecerán bloqueadas para evitar el bloqueo de grandes bloques de usuarios, que puede darse si un atacante accede a sus nombres de usuario. La 2FA también limita los ataques por fuerza bruta, dado que uno de los componentes involucrados varía.

Roles involucrados: RSI.

Categoría: Tecnología.

Tests de penetración

Con respecto al proceso de verificación del sistema desde el punto de vista de la seguridad de la información, se recomienda:

HC.P4. Tests de penetración

Además de realizar auditorías de seguridad, llevar a cabo tests de penetración del sistema de Análisis de Datos.

Roles involucrados: RSI.

Categoría: Tecnología.

Detección y control (DC)

Las buenas prácticas se presentan según la siguiente clasificación:

- Monitoreo y sistemas de detección de intrusión.

Monitoreo y sistemas de detección de intrusión

Con respecto al monitoreo del sistema de Análisis de Datos y a los sistemas de detección de intrusión, se recomienda:

HC.DC1. Incorporación de un IDS

Evaluar incorporar un sistema de detección de intrusión (IDS o HIDS por sus siglas en inglés) en el sistema de Análisis de Datos. El objetivo de estos sistemas es detectar posibles ataques mediante monitoreo pasivo del sistema (sobre la red o los servidores de la misma), brindando los servicios de notificación correspondientes a los administradores del mismo. La falta de este tipo de controles permite que un atacante realice distintas estrategias de ataque sin restricciones, hasta encontrar una exitosa. La incorporación de un sistema IDS aumenta la probabilidad de detectar al atacante antes de que tenga éxito.

Roles involucrados: RSI.

Categoría: Tecnología.

HC.DC2. Monitoreos del sistema

Determinar las partes del sistema de Análisis de Datos que requieren monitoreos (manuales o automáticos), durante qué periodos se realizarán los monitores y qué acciones serán llevadas a cabo en caso de generarse una alerta.

Por ejemplo, se sugiere determinar un proceso de respuesta a las alertas que indican que un usuario está descargando una cantidad inusual de datos del sistema o accediendo en horarios poco frecuentes.

Roles involucrados: RSI.

Categoría: Personas.