LINEAMIENTO DE GOBIERNO DEL DATAWAREHOUSE

CONTROL DE CAMBIOS

Fecha	Versión	Observaciones
11/02/2021	1.0	Creación del documento
04/03/2021	1.1	Modificación del documento

UNIVERSIDAD EAFIT

CONTENIDO

Ilustraciones

INTRODUCCIÓN

La gobernanza de datos sobre un DWH es definida como el ejercicio de autoridad y control (planeación, monitoreo y refuerzo) sobre la gestión de los activos de datos que serán migrados al repositorio central. Todas las organizaciones toman decisiones sobre los datos, independientemente de si tienen un programa de gobierno de datos establecido. Aquellas que establecen un programa de gobierno de datos formal ejercen autoridad y control con una mayor intencionalidad; estas organizaciones son más capaces de aumentar el valor que obtienen de sus activos de datos.

El propósito de la gobernanza de datos sobre el proyecto de DWH Moderno es garantizar que los datos se gestionen correctamente, de acuerdo con las políticas y las mejores prácticas. Si bien el motor de la gestión de datos, en general, es garantizar que una organización obtenga valor de sus datos, la gobernanza de datos se centra en cómo se toman las decisiones sobre los datos y cómo se espera que las personas y los procesos se comporten en relación con los datos. El alcance y el enfoque de un programa de gobierno de datos en particular dependerá de las necesidades de la organización; en este sentido, este documento tiene como propósito dar a conocer los elementos necesarios para implementar el programa de gobierno de datos para la UNIVERSIDAD EAFIT dentro de nuestra arquitectura de DWH Moderno implementado en la nube de Azure.

OBJETIVOS

Objetivo General

Construir un marco de trabajo operativo personalizado para EAFIT que se ajuste a sus necesidades y pueda ser plenamente adoptado teniendo en cuenta el rol que desempeñan los datos dentro de la organización, su modelo de negocio, el impacto de las regulaciones que le aplican y su nivel de madurez actual en gobierno de datos.

Objetivos Específicos

- Diseñar el proceso propuesto para la gestión del gobierno de datos sobre el DWH
- Diseñar una propuesta de modelo de formalización de los roles que deben crearse para la implementación del gobierno de datos sobre el DWH
- Proponer las mejores prácticas relacionadas con la seguridad y calidad del dato
- Realizar recomendaciones generales y personalizadas para la implementación del DWH en EAFIT de acuerdo con lo observado durante el proceso de consultoría

1. MARCO DE TRABAJO

Para la implementación de un programa de gobierno sobre nuestro DWH es necesario definir el framework o marco de trabajo con el cual se desea operar, sin embargo, crear un modelo de operación que sea correctamente adoptado por la organización puede ser difícil. La construcción de un modelo operativo para una organización debe considerar los siguientes puntos:

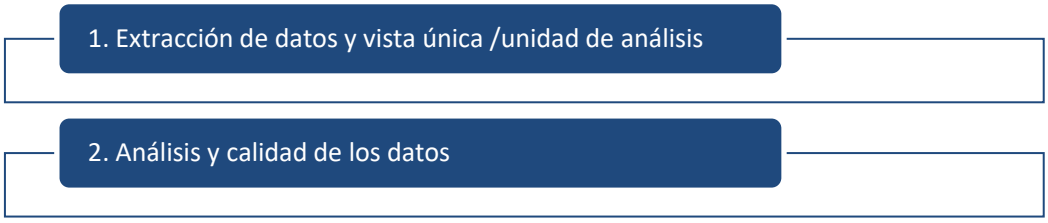
- **El valor de los datos dentro de la organización:** Si una organización vende datos o hace parte de su estrategia comercial por supuesto el gobierno tiene un gran impacto dentro de la organización.
- **Modelo de negocio:** Dependiendo, si se trata de una organización centralizada o descentralizada, local o internacional entre otros factores que influencian el funcionamiento de la organización y por lo tanto el modelo operativo del gobierno de datos. El modelo de negocio influirá en la estrategia de TI, la arquitectura de datos y las labores de integración de los datos lo cual a su vez se verá reflejado en el modelo operativo del gobierno de datos.
- **Impacto de las regulaciones:** Las organizaciones que son altamente reguladas tendrán una mentalidad diferente a aquellas que son menos reguladas. El modelo operativo podría entonces tener relación con el grupo de gestión de riesgos o legal de la organización.
- **Nivel de madurez en gobierno de datos:** De acuerdo con el nivel de madurez de las personas, políticas y procesos se deberán definir roles al nivel técnico que las personas estén en la capacidad de asumir, además de políticas y procesos que la organización esté en capacidad funcional y operativa de implementar.

De esta manera, como primeros pasos dentro del Gobierno de datos se propone la implementación de un proceso de análisis de calidad de datos, el cual permitirá conocer el estado actual de la información y tener una línea base para desplegar estrategias de mejora continua sobre este activo de la institución.

En principio se sugiere realizar un análisis de calidad de referencia sobre nuestro MVP de alertas de deserción ya que aún no se tienen metas establecidas dentro de los indicadores ni rangos de aceptabilidad.

1.1 Metodología para el Análisis calidad de Datos

Parte del conocimiento de los campos que conforman las bases de datos desde las cuales se extraerá la información para el modelo MVP, identificando posibles bloques de información que componen a nivel individual cada unidad de análisis, por ejemplo, bloque demográfico, bloque de información académica, entre otros. Posteriormente se procede con la extracción y transformación de los datos en la herramienta de minería y se realiza un completo análisis descriptivo. Para ello se definen las siguientes actividades claves.



1.1.1 Actividad 1: Extracción de Datos y Vista Única de Unidad de Análisis

Se realiza una extracción de la información mediante las herramientas ETL y de análisis, identificando la unidad de medida a analizar. Una de las bondades de la metodología es que no permite la duplicidad y se genera la visión única por registro. La presente actividad se encuentra cubierta dentro de la implementación del MVP puesto que uno de los scripts entregados por parte de la universidad, correspondiente al Modelo de alertas tempranas, comprende la creación del maestro de estudiantes.

1.1.2 Actividad 2: Análisis y Calidad de Datos

Para garantizar un resultado con altos estándares técnicos se hace necesario la construcción de un indicador de calidad global de las bases de datos, en este caso de aquellas vistas que generen la información para el MVP, que permita dimensionar adecuadamente el alcance del proyecto y garantizar adecuadas estrategias de poblamiento, depuración y construcción de indicadores de conocimiento. La calidad de los datos no solo implica que estos sean buenos y carentes de defectos, sino que sean precisos, consistentes y completos para un correcto desarrollo del proyecto, garantizando en todo momento información correcta y útil.

De esta manera, una vez consolidado el inventario de archivos y registros de las bases de datos de interés, se construyen reglas y ponderaciones que determinarán el nivel de calidad de datos. Para este ejercicio, EAFIT genera ponderaciones y establecerá las causales de incumplimiento sobre campos que se consideran de alto valor, obteniendo como resultado el indicador de calidad que permite dar un diagnóstico claro del estado de la información, donde el valor del indicador dependerá de la calidad que tiene cada variable y del peso asignado.

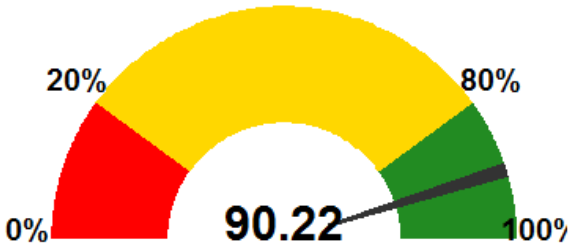
1.1.2.1 Construcción del Indicador de Calidad

El indicador de calidad es un método analítico que cuantifica la calidad de los registros de una base de datos con el propósito de dar un diagnóstico claro del estado de la información y así diseñar estrategias de actualización, procesos de captura y comunicaciones integradas de relacionamiento. Generalmente, los resultados del análisis de calidad de datos son consumidos a través de tableros de control en herramientas de visualización como Power BI.

Beneficios

- ✓ Conocimiento del estado del arte de la información que posee la compañía.
- ✓ Diagnóstico de calidad por unidad de análisis e identificación de su causal de incumplimiento.
- ✓ Impulsar mejores resultados en los proyectos de analítica al tener datos fiables.

Ilustración 1. Ejemplo de tipo de gráfico para medir la calidad de los datos



Así pues, una vez implementado el análisis de calidad de datos en la institución, el proceso propuesto para la gestión del gobierno de datos sobre nuestro DWH se describe en los siguientes pasos:

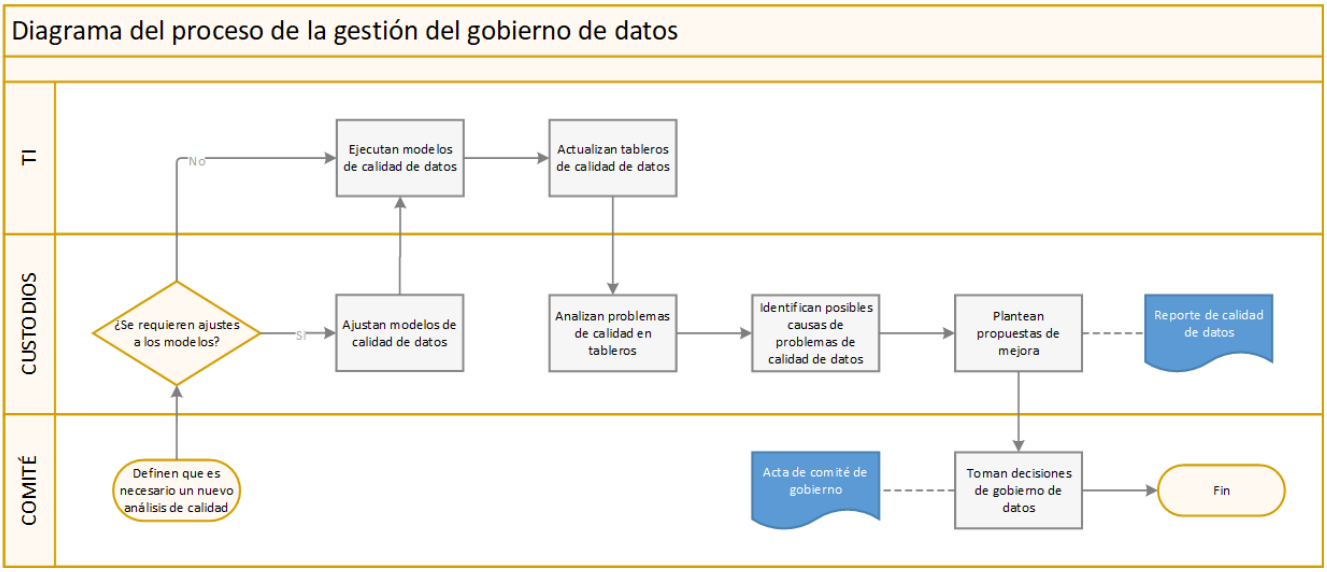


Ilustración 2 Diagrama del proceso de la gestión del gobierno de datos

1.2 Definir la necesidad de realizar un análisis de calidad

Es necesario definir cuándo es necesario realizar un análisis de calidad, bien sea por medio de la ocurrencia de un evento (por ejemplo, un indicador o grupo de indicadores de calidad de datos no se encuentra dentro de su rango de aceptabilidad o fueron añadidos nuevos atributos al análisis) o simplemente definir una periodicidad de análisis.

1.3 Plantear propuestas de mejora

Las propuestas de mejora deben ser del tipo:

1.3.1 Acciones preventivas

La mejor manera de crear datos de alta calidad es previniendo que datos de mala calidad ingresen a la organización. Las acciones preventivas impiden los errores sobre los datos nuevos. Inspeccionar los datos luego de que hayan ingresado a la organización no va a mejorar su calidad. Entre las acciones preventivas están:

i. **Establecer controles en las entradas de datos:** Crear reglas que prevengan que datos inválidos o imprecisos ingresen a los sistemas. Por ejemplo, es posible restringir los valores de entrada de un atributo a través de listas desplegables, restringir los tipos de valores recibidos (únicamente números, únicamente caracteres, etc), limitar la cantidad de caracteres a ingresar, entre otros controles.

ii. **Capacitar a los productores de datos:** Asegurarse que el personal que ingresa información a los sistemas comprenda el impacto de sus datos en los usuarios intermedios. Se debe asegurar que los nuevos usuarios que ingresen a los aplicativos reciban el entrenamiento necesario para el manejo de este, dándole a conocer los estándares y políticas relacionadas con el ingreso y modificación de los datos. Es necesario que existan metadatos que permitan conocer información relacionada con el linaje del dato y con este insumo poder realizar un seguimiento al personal. Entre las dimensiones más útiles a considerar relacionadas con el linaje del dato y las cuales se recomienda incluir dentro de los modelos de datos actuales están:

- Usuario que creó o modificó el dato
- Fecha de creación y modificación del dato
- Aplicativos en donde fue creado o modificado el dato
- Motivo por el cual fue creado o modificado el dato
- Método de ingreso del dato que ha sido creado o modificado
- Fecha en que se inactiva o elimina un dato
- Motivo por el cual fue inactivado o eliminado un dato
- Método de inactivación o eliminación de un dato

Nota: *los anteriores ítems serán una política dura a nivel de roles y accesos en el DWH*

iii. **Definir y aplicar reglas:** Crear un "firewall de datos", que tenga una tabla con todas las reglas de calidad de datos comerciales que se utilizan para verificar si la calidad de datos es buena, antes de utilizarlos en una aplicación como un almacén de datos. Un firewall de datos puede inspeccionar el nivel de calidad de los datos procesados por una aplicación, y si el nivel de calidad está por debajo de los niveles aceptables, los custodios de datos pueden ser informados sobre el problema. Un primer firewall de datos será los tableros de calidad de datos elaborados sobre el MVP para el consumo de los custodios de datos, sin embargo, puede ser necesario que se apliquen controles adicionales por medio de herramientas especializadas en el análisis de calidad de datos como por ejemplo la herramienta DQ Analyzer.

1.3.2 Acciones correctivas

Las acciones correctivas se implementan después de que se haya producido y detectado un problema. Los problemas de calidad de los datos deben abordarse sistémicamente y en sus causas fundamentales para minimizar los costos y riesgos de las acciones correctivas. "Resolver el problema donde sucede" es la mejor práctica en la gestión de la calidad de los datos. Esto generalmente significa que las acciones correctivas deben incluir la prevención de la recurrencia de las causas de los problemas de calidad, inicialmente aplicado el MVP.

Se pueden realizar acciones correctivas de 3 maneras diferentes:

i. **Correcciones automatizadas:** Las técnicas de corrección automatizada incluyen estandarización, normalización y corrección basadas en reglas. Los valores modificados se obtienen o generan y se comprometen sin intervención manual. Un ejemplo es la corrección automática de direcciones, que envía las direcciones de entrega a un estandarizador de direcciones que ajusta y corrige las direcciones de entrega mediante reglas, análisis, estandarización y tablas de referencia. La corrección automatizada requiere un entorno con estándares bien definidos, reglas comúnmente aceptadas y patrones de error conocidos. La cantidad de correcciones automatizadas puede reducirse con el tiempo si este entorno está bien administrado y los datos corregidos se comparten con los sistemas ascendentes. Las soluciones de este tipo son implementadas normalmente por medio de la automatización robótica de procesos (RPA’s) o por medio de la configuración de procesos de extracción, carga y transformación entre aplicativos (ETL’s)

ii. **Correcciones dirigidas manualmente:** Se pueden utilizar herramientas automatizadas para corregir datos, pero esto requiere de una revisión manual antes de confirmar que las correcciones se almacenen en las bases de datos de manera definitiva. Normalmente, se usa para aplicar correcciones de nombres, direcciones y datos relacionados con la identidad de las personas y se recomienda utilizar un software especializado el cual de acuerdo con criterios ingresados previamente acordados por los custodios se realicen correcciones basadas en patrones identificados en los datos. Para realizar las correcciones anteriormente mencionadas se utilizan herramientas del tipo MDM, el cual se detallará en una sección posterior.

iii. **Correcciones manuales:** A veces la corrección manual es la única opción en ausencia de herramientas o reglas automatizadas de los sistemas o si se determina que el cambio se maneja mejor a través de la supervisión humana. Las correcciones manuales se realizan mejor a través de una interfaz con controles y ediciones, que proporcionan una pista de auditoría para los cambios. La alternativa de hacer correcciones y comprometer los registros actualizados directamente en entornos de producción es extremadamente arriesgada. Evitar usar este método.

Para plantear las propuestas de mejora se recomienda crear un formato estándar llamado el reporte de calidad de datos en donde se documenten los siguientes elementos:

- Entidades, fuentes y atributos evaluados
- Diagrama de flujo de datos y diagrama del proceso evaluado
- Problemas identificados y sus posibles causas (diagrama causa efecto)
- Propuestas de acciones correctivas y preventivas

1.4 Toma de decisiones de gobierno del Data Warehouse

La toma de decisiones por parte del comité de gobierno es un paso clave para que el gobierno de nuestro DWH pueda desarrollarse dentro de la organización ya que es el comité mismo que, de acuerdo con su visión de negocio, el que determina cuáles de las acciones propuestas por los custodios de datos son las más convenientes de acuerdo con el impacto esperado que estas tengan según la estrategia planteada. En el ítem de propuestas de mejora fueron detalladas las posibles acciones preventivas y correctivas que pueden tomarse sobre los procesos, personas y tecnologías; sin embargo, es necesario acompañar dichas decisiones de políticas de gobierno de datos que especifiquen de manera formal las reglas que deberán aplicarse en cada caso específico.

Una política es un principio o regla para guiar las decisiones y lograr los resultados esperados en un proceso. El término normalmente no se usa para denotar lo que realmente se hace, esto normalmente se conoce como procedimiento o protocolo.

Los esfuerzos en calidad de datos deben estar soportados y a su vez soportar las políticas de gobierno de datos. Por ejemplo, las políticas de gobierno pueden establecer la periodicidad de las auditorías de calidad y exigir el cumplimiento de las normas y las mejores prácticas. Cada política debe incluir:

- Propósito, alcance y aplicabilidad de la política.
- Definición de términos de negocio. Dicha definición debe ser acorde con los diccionarios de datos.
- Responsabilidades de las personas involucradas con el proceso asociado.
- Responsabilidades de otras partes interesadas.

- Vigencia y fecha de inicio.
- Informes que indiquen cómo se ha aplicado la política.
- Implementación de la política, incluyendo enlaces a riesgos, medidas preventivas, cumplimiento de las regulaciones, requerimientos de protección y seguridad de datos.

Todas las políticas de calidad de datos deben ser avaladas por el comité de gobierno, teniendo en cuenta que, de acuerdo a cómo sean establecidas se procederá al momento de resolver conflictos que se presenten en los datos. Las políticas de calidad de datos deben ser el pilar sobre el cual se deben tomar las decisiones en el futuro relacionadas con la calidad de datos.

Para evidenciar las decisiones del comité de gobierno se recomienda crear un acta de gobierno de DWH en donde se documenten los siguientes elementos:

Revisión del acta anterior

- Revisión de pendientes: Tareas realizadas y no realizadas
- Revisión de obstáculos de tareas no realizadas y definición de estrategias para su cumplimiento
- Revisión del impacto en los indicadores de las tareas realizadas

Revisión del reporte de calidad de datos

- Revisión de problemas identificados y sus respectivas causas
- Propuestas de mejora
- Toma de decisiones de gobierno. Deben especificarse los compromisos y tareas a realizar especificando responsables y fechas de ejecución
- Revisión del estado actual de los indicadores involucrados y el impacto esperado con las tareas planteadas

1.5 Seguimiento a las decisiones implementadas

Para el seguimiento a las decisiones implementadas el comité de gobierno puede apoyarse en los indicadores de calidad global de tal manera que, según la evidencia del aumento o disminución en la calidad de los datos, el efecto en la calidad de los datos pueda ser medido y determinar si las decisiones de gobierno tienen el efecto esperado.

Inicialmente, para los tableros de calidad de datos, no se tiene una meta establecida a lograr. A través de la experiencia y luego de completar en repetidas ocasiones el proceso para la gestión del gobierno de datos es posible que entre el comité de gobierno y los custodios de datos acuerden unas metas de mejora en la calidad de datos a las cuales se desea llegar en el corto plazo. Se recomienda para cada indicador de calidad de datos definir:

- Metas al corto y largo plazo

- Rangos de alerta alta, media y rango de aceptabilidad del indicador
- Acciones para realizar en caso de estar en los rangos de alerta

2. FORMALIZACIÓN DE ROLES PARA LA IMPLEMENTACIÓN DEL GOBIERNO DE DATOS SOBRE DWH

Tener definido un marco de trabajo es fundamental en la implementación de gobierno de datos, pero definir su operatividad es de suma importancia, ya que define la interacción entre la organización de gobierno y las personas responsables de proyectos e iniciativas en la Universidad EAFIT, permitiendo una cohesión en toda la organización alrededor del ciclo de vida del dato dentro del DWH.

2.1 Roles de Gobierno de datos en la institución

Los roles recomendados necesarios para dar inicio al gobierno de datos dentro de la institución y su respectiva jerarquía se ilustran en el siguiente diagrama:

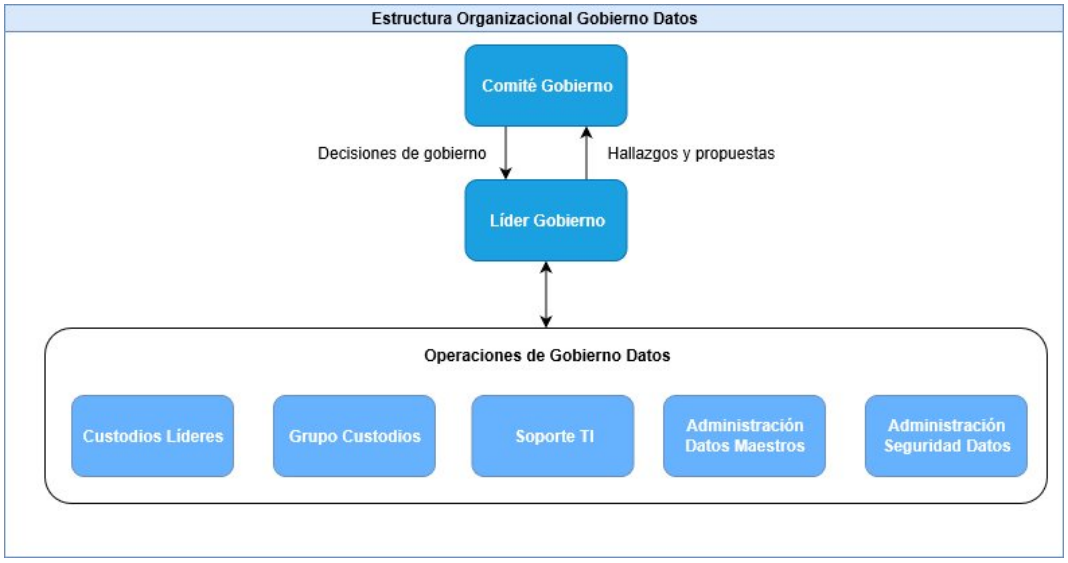


Ilustración 3. Roles en el gobierno de datos de un DWH

Las funciones de cada de uno de los roles anteriormente mencionados son:

2.1.1 Comité de gobierno

Son los patrocinadores del programa de gobierno de datos, donde sus funciones se centran en:

- Entender cómo fluye la información en los procesos de la empresa para aprovechar su uso en la toma de decisiones estratégicas.
- Establecer de acuerdo con su visión de negocio las metas que desean alcanzarse mediante el gobierno de datos.
- Revisar y aprobar los insumos suministrados por la comunidad de custodios para la implementación del gobierno de datos: políticas, estándares, herramientas y mejores prácticas.

- Gestionar las soluciones que ayuden al custodio de datos a resolver las dificultades que pueda tener al ejercer su función.
- Analizar en qué medida dichos proyectos que se proponen en la hoja de ruta afectan al flujo de información, identificar los riesgos, aprobar los planes de contingencia y autorizar, facilitar y monitorear las implementaciones.
- Identificar oportunidades de mejora en los procesos actuales y promover la creación de proyectos que sistematicen la captura, procesamiento y automatización para generar y gestionar información a través de la tecnología
- Validar procesos de estimación de costos y arquitecturas en nube con base en formatos de necesidades y cumplimiento presupuestal

Actualmente dicho comité ya se encuentra constituido en la institución. Liderado por el área de Direccionamiento estratégico y conformado por la dirección de Informática, de Mercadeo Institucional, Administrativa y financiera, Desarrollo Institucional, Desarrollo Humano, la Oficina de Admisiones y Registro, un delegado de cada vicerrectoría y uno de la academia.

2.1.2 Líder de Gobierno

Puede estar integrado por uno o varios funcionarios, los cuales deben poseer conocimientos de negocio y técnicos a nivel de BI y/o Infraestructura en nube de Azure, para que puedan ser la conexión entre todos los roles que comprenden el grupo de operaciones y comité de gobierno, permitiendo que todas las propuestas de mejora sean siempre viables y realizables desde todos los puntos de vista. La función principal será coordinar y apoyar las actividades del grupo de operaciones de gobierno de datos y en toda la organización, adicionalmente debe:

- Velar que todas las decisiones que hayan sido tomadas por el comité de gobierno sean asignadas a las personas correspondientes y que sean ejecutadas de forma exitosa.
- Recopilar todos los hallazgos y propuestas de mejora sugeridas por parte del grupo de operaciones de gobierno para lograr una priorización y llevar al comité una presentación concisa y organizada.
- Estar presente en los nuevos proyectos corporativos para ayudar a definir las reglas relacionadas para los datos y vigilar el cumplimiento de estas.

Responsables del rol: Asistente Direccionamiento Estratégico.

2.1.3 Operaciones de gobierno

Grupo de personas conformado por diferentes núcleos de conocimiento donde en conjunto ejecutan la mayoría de las funciones necesarias para dar cumplimiento al gobierno de datos. Los roles necesarios son:

2.1.3.1 Custodios líderes

Son los principales responsables de vigilar y velar por la calidad de los datos en EAFIT para las entidades priorizadas y esquema de almacenamiento en el DWH. Los resultados de calidad estarían disponibles en reportes de Power BI los cuales serán unos de los principales insumos para la medición de calidad junto con el diccionario de datos y el conocimiento de negocio que posean. Otras funciones asociadas a este rol son:

- Realizar el levantamiento y seguimiento a los flujos de procesos de datos dentro de la organización.
- Realizar mediciones de calidad de datos por medio de los reportes de Power BI, identificando las razones de incumplimiento.
- Sugerir mejoras a nivel de políticas, estándares y procesos en la organización que permitan mejorar la calidad de los datos, que deben ser analizadas en el grupo de operaciones de gobierno para que posteriormente puedan ser elevadas al comité.
- Generar reportes de cumplimiento a las reglas o estándares definidos, por parte de las diferentes áreas de la organización.
- En caso de ser necesario calcular y graficar nuevos indicadores de calidad que permitan análisis más detallados según la necesidad.
- Realizar actualizaciones o correcciones a los diccionarios de datos en casos donde se presenten modificaciones en los aplicativos, actualizaciones de políticas organizacionales, entre otros.
- Velar porque se respeten las políticas de calidad de los datos en toda la organización.

Responsables del rol: Analista de Soluciones – Célula de Bi y Analítica y Analista Direccionamiento estratégico.

2.1.3.2 Grupo de Custodios/Administradores de datos

Grupo de personas con conocimientos de negocio en la organización cuyo papel principal será brindar apoyo a los custodios principales al momento de los hallazgos y propuestas de calidad de datos, garantizando que las posibles mejoras sean transversales en toda la organización. Adicionalmente pueden realizar las documentaciones de las decisiones realizadas por el comité y ayudar con su respectivo seguimiento.

Responsables del rol: Una persona encargada por cada dominio de información identificado en la institución.

2.1.3.3 Soporte TI

Está conformado por uno o más funcionarios de la organización miembros del área de TI y Bases de datos, que tendrán funciones para apoyar el desarrollo e implementación del gobierno de datos como:

- Desarrollar y ejecutar componentes para la integración, búsqueda y modificación masiva de datos cuando sea requerido, dependiendo de los hallazgos de los custodios y las decisiones tomadas por parte del comité.
- Construir políticas y estándares de arquitectura de datos.
- Apoyar gestión de metadatos.
- Realizar actualizaciones de la arquitectura de datos para garantizar que la información siempre esté disponible en todos los aplicativos.

- Apoyar a la actualización y/o corrección del diccionario de datos en los ítems que están relacionados con los almacenamientos en los sistemas de información.
- Velar porque se respeten las políticas de seguridad y custodia de los datos en toda la organización.
- Desarrollar mecanismos de control de crecimiento de presupuesto y disponibilidad de arquitecturas.

Responsables del rol: Responsable Arquitectura desplegada y Coordinador de Bases de datos.

Los siguientes dos roles son necesarios en el gobierno de datos, pero no es de obligatorio cumplimiento que existan personas dedicadas a estas funciones, ya que pueden ser suplidas por parte de los custodios o por soporte de TI. Lo importante es que estas funciones siempre sean desarrolladas.

2.1.3.4 Administración de datos maestros

Conformado por uno o más personas con roles de administrador en los diferentes aplicativos de la organización, que deben contar con un conocimiento transversal de todos los sistemas de información existentes, donde deben saber por medio de que aplicativos o servicios es ingresada la información y en qué base de datos y tablas queda almacenada finalmente la información, permitiendo tener garantía de que los cambios que se llegasen a realizar sean los adecuados. Las funciones asociadas a este rol son:

- Gestionar las solicitudes de creación o modificación de los datos maestros en los sistemas de información, de acuerdo con las decisiones tomadas en el comité para la mejora de gobierno de los datos.
- Realizar procesos de limpieza de datos en los maestros con problemas de calidad de datos.
- Verificar que las propuestas de mejora realizadas por los custodios si sean posibles de ejecutar en el sistema de información y en caso contrario, proponer nuevas opciones que permitan corregir la falencia detectada.
- Apoyar en la actualización del diccionario de datos.
- Participar en los nuevos desarrollos de aplicativos para garantizar que las reglas de negocio y proceso de gestión del dato que se establezcan sean replicadas y mantener una articulación del ecosistema de información.

Responsables del rol: Departamento de Soluciones de Software.

2.1.3.5 Administración de seguridad de datos

Puede estar conformado por una o varias personas, que tienen como objetivo principal planear, desarrollar y ejecutar las políticas y procesos de seguridad, que permitan un adecuado uso de los datos almacenados en los sistemas de información existentes en la organización. Sus funciones son:

- Evaluar el riesgo y definir controles para gestionar riesgos de seguridad actuales.
- Evaluar los requerimientos de seguridad de los datos de acuerdo con los hallazgos de calidad de datos proporcionados por los custodios.
- Implementar procedimientos y controles necesarios para garantizar la seguridad de los datos.
- Ayudar a definir políticas-estándares-procesos de seguridad de datos.
- Realizar y actualizar documentaciones relacionadas con la seguridad de datos.

Responsables del rol: Área de riesgos, área de gestión documental, área de seguridad informática y un delegado del grupo de Gobierno de datos.

Luego de definir e implementar los roles necesarios para el gobierno, es importante definir las interacciones entre todos los niveles del área de gobierno, permitiendo un correcto desarrollo del plan de gobierno de datos, así que las reuniones para la toma de decisiones de gobierno se sugieren deben realizarse con los siguientes intervalos de tiempo:

- Las reuniones entre el grupo de operaciones para compartir los hallazgos de calidad de datos y el levantamiento de propuestas de mejora debe realizarse de forma semanal. Esta periodicidad se puede modificar de acuerdo con los avances que se hayan realizado entre cada sesión.
- Las reuniones entre el grupo de operaciones y el comité de gobierno para analizar y tomar decisiones de gobierno deben realizarse de forma mensual. Adicionalmente, se debe dar seguimiento a las decisiones que se hayan tomado con anterioridad para verificar el nivel de implementación y apropiación en la organización.
- Considerar realizar reuniones entre el grupo de trabajo y el líder de gobierno para no solo tratar temas relacionados a calidad de datos, sino tener en cuenta todos los temas de iniciativas nuevas en la organización que impliquen la creación de nuevos aplicativos o funcionalidades nuevas en los sistemas para mejorar la seguridad de los datos. Todo esto con el fin de que todo el equipo de trabajo esté alineado.
- Considerar como un participante activo de las decisiones que se tomen en el área de gobierno al grupo de gestión del cambio, permitiendo de que la organización esté informada de las nuevas implementaciones de gobierno.

2.2 Roles dentro del DWH

El MVP de alertas tempranas supone una puesta en marcha no solo de los servicios a usar sino de los roles que es necesario mapear y tener en cuenta para la correcta administración del DWH y ejecución del modelo.

Según la arquitectura y los servicios, los roles considerados a tener en cuenta son:

- Arquitecto Cloud
- IT Manager: Administrador de Redes, Comunicaciones y Sistemas Operativos
- Master Data-Data Engineer
- DBA (Administrador de bases de datos)
- Data Mining-Data Scientist
- Business Intelligence

2.2.1 Arquitecto Cloud

Un Arquitecto Cloud es responsable de administrar y coordinar la estructura cloud computing en una organización. Las funciones del Arquitecto Cloud engloban todo lo relacionado con servidores, plataformas de almacenamiento, conectividad y software, que dependen del modelo de cloud elegido: público, privado o híbrido.

- Liderar el cambio cultural y empresarial que supone la adopción de la nube
- Desarrollar y coordinar la arquitectura cloud
- Desarrollar una estrategia de adopción de cloud y coordinar el proceso

Responsables del rol: Responsable de la arquitectura desplegada y Analista de infraestructura.

2.2.2 IT Manager: Administrador de Redes, Comunicaciones y Sistemas Operativos

Es responsable por mantener y monitorear la infraestructura de redes de la Organización. Entre sus principales competencias está mantener el funcionamiento de redes informáticas internas y conexiones a redes externas, de acuerdo con los niveles de servicio operacional y de seguridad que se establezcan.

Nombres de cargos similares

Soporte de infraestructura IT, Soporte de redes, Especialista en redes, Técnico de redes, Ingeniero de redes, Administrador de redes, comunicaciones y sistemas operativos.

Este rol particularmente para el proyecto implementa la VPN para la comunicación entre el sitio On-premise y la nube de Azure.

Responsables del rol: Grupo de trabajo de plataformas y Servidores – Departamento de Infraestructura.

2.2.3 Master Data-Data Engineer

El Master Data o Data Engineer, es el encargado de asegurarse de definir e implementar un flujo de datos desde su origen hasta su explotación de una forma controlada y automatizada.

En primer lugar, debe tener un conocimiento fluido de bases de datos relacionales y del lenguaje de consulta SQL, ETLs, dominio sobre servicios de almacenamiento no estructurado y herramientas como Datafactory, Datalake, Datawarehouse Azure, etc. Esto le da el conocimiento sobre las técnicas de modelado de datos más utilizadas y conocimiento sobre cómo acceder a los datos de origen cuando estos residen en este tipo de almacenamiento.

Entre sus tareas habituales suele encontrarse en definir la configuración del clusters de procesamiento, número de cores, memoria y otros parámetros de bajo nivel, así como también implementar los flujos ETL. Todo enfocado a que la ejecución de los modelos sea lo más eficiente.

Responsables del rol: Analista de Soluciones – Célula de Bi y Analítica y Analista Direccionamiento Estratégico.

2.2.4 DBA (Administrador de bases de datos)

Los DBA son las personas o grupo de personas encargadas de administrar, supervisar y asegurar el adecuado uso de los datos dentro de un DBMS (Database Management System), Estos sistemas permiten manejar grandes montañas de datos de una manera eficiente, permitiendo así disponer de una mejor herramienta para la toma de decisiones de negocios.

Dentro de las funciones más importantes se encuentran las siguientes: Gestión general de Bases de datos, Modelado de datos y diseño de Bases de Datos, Auditoría, Integración con aplicaciones, Resguardo y recuperación de datos, administración de cambios, entre otras.

Responsables del rol: Grupo coordinación de Bases de datos.

2.2.5 DataMining-Data Scientist

Los DataMining-Data Scientist son los profesionales encargados de descubrir patrones en enormes volúmenes de datos. Para ello se valen de herramientas como la inteligencia artificial, el aprendizaje automático, la estadística y sistemas de base de datos.

Requiere saber de matemáticas, estadística y Machine Learning, de lenguajes de programación como R o Python, de uso de notebooks y de ecosistemas Big Data, pero lo que creemos que diferencia al Data Scientist es que es el encargado de sacarle valor a los datos por medio de herramientas como Databricks.

Responsables del rol: Asistente Direccionamiento Estratégico y Asistente Dirección de formación integral.

2.2.6 Analistas de Business Intelligence

Un analista de inteligencia de negocios es responsable de construir tableros y reportes para el análisis de datos en la institución, que son usados para dar soporte a la toma de decisiones. Trabaja con ese tipo de datos para maximizar su utilidad.

Generalmente, un analista de inteligencia de negocios tiene conocimientos de informática y de Business Intelligence. Pero, además, muchos de estos puestos requieren también conocimientos específicos de temas como SQL o herramientas concretas de Business Intelligence como Power BI u otras similares

Responsables del rol: Analistas de Sistema – Departamento Soluciones de Software y Analistas y Asistente Direccionamiento estratégico.

3 HERRAMIENTAS DE GOBIERNO DE DATOS

Al iniciar un proceso de gobierno de datos es de vital importancia definir las herramientas que permitan alcanzar el objetivo de incrementar el nivel de madurez en la organización, así que la primera parte consta de una arquitectura que permita mejorar la calidad de los datos e integración de información, la segunda es la definición de una gestión de datos maestros - MDM.

3.2 Arquitectura propuesta sobre Azure

La correcta implementación de un gobierno de datos debe estar soportada en una arquitectura de datos altamente flexible y potente, que permita controlar la seguridad, consistencia y calidad de la información, para este fin pueden ser integrados servicios interconectados que cumplan tares puntuales de forma autónoma, pero ensamblando una gran solución donde el resultado final se convierta en datos de gran valor para la toma de decisiones de la compañía.

A continuación, se diagrama una solución de arquitectura donde se muestra desde una visión general un correcto funcionamiento y flujo de los datos que se debería tener inicialmente para poder potenciar a corto y mediano plazo el desarrollo del mapa de ruta del gobierno de datos dentro del MVP planteado para EAFIT.

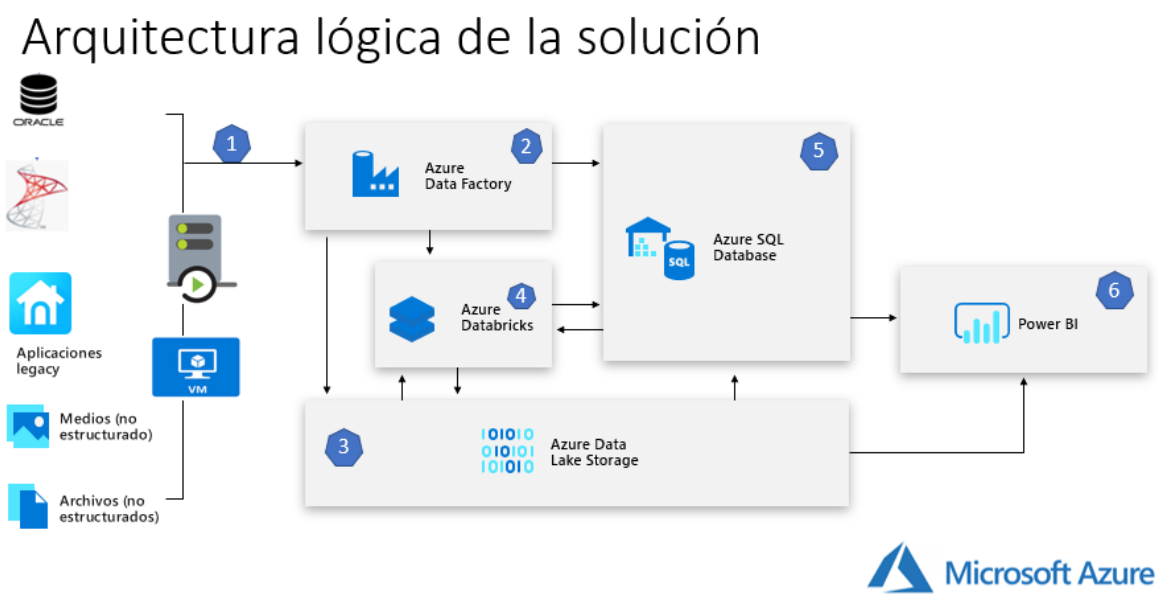


Ilustración 4. Arquitectura propuesta

3.1.2 Descripción de los distintos pasos de la solución.

1. Canal de transferencia de datos mediante VPN a través de máquina virtual.
2. Herramienta data Factory para realizar el proceso ETL de datos
3. Servicio de almacenamiento masivo (Data Lake) y a bajo costo para alojar datos estructurados y no estructurados
4. Databricks para analítica avanzada de datos
5. SQL Database para almacenamiento y modelamiento de datos.
6. Herramienta de BI para visualización de datos mediante tableros y paneles

A continuación una breve descripción de cada servicio por las diferentes capas de la arquitectura:

1. **Capa de orígenes**, identifica todas las posibles fuentes con las distintas sensibilidades, estructuras, contenedores o generadores.

   ✓ *Máquina virtual.*

   Software que permite emular el funcionamiento de un ordenador dentro de otro ordenador gracias a un proceso de encapsulamiento que aísla a ambos.

- ✓ Integration Runtime

Herramienta que permite la conexión entre el sitio On-Premise y la nube, está instalado en la máquina virtual

2. **Capa de ingesta**, será capaz de desplegar distintos Data Transfer Engine, con capacidad para poder adquirir datos estructurados de manera Batch o sincronizando bases de datos espejo en la nube, con ayuda de tecnologías propias de Microsoft Azure o apoyados en herramientas del ecosistema Hadoop, por ejemplo:

- ✓ *Azure DataFactory.*

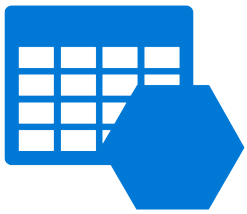
El cual es un pool de herramientas que permiten programar flujos de trabajo programados de copias invocando distintos servicios según las necesidades existentes, incluye soluciones como SSIS y gran cantidad de conexiones a distintas bases de datos como SAP R3, Oracle, SQL Server, Archivos planos.

3. **Capa de almacenamiento**, será capaz de guardar todos los datos que fueron obtenidos a través de los Data Factory en la capa de ingesta.

- ✓ *Azure Sql Database.*

Azure SQL Database es una base de datos relacional de propósito general, proporcionada como un servicio administrado. Con él, puede crear una capa de almacenamiento de datos de alto rendimiento y alta disponibilidad para las aplicaciones y soluciones en Azure.

- ✓ *Data Lake Gen2.*



Almacenará los datos que entren a la solución de forma económica y altamente tolerable a fallos, con capacidad de almacenamiento de grandes volúmenes de datos, integrado con el directorio Activo para poder gestionar permisos de acceso, lectura, escritura de forma centralizada, además brinda los conectores necesarios para que las herramientas de analítica y BI puedan acceder a los datos crudos o finales de forma controlada y eficiente.

4. **Capa de Procesamiento y análisis**, esta capa ejecutará todos los procesos de transformación de los datos a través de los paquetes desarrollados en SSIS y cargados en la nube,

- ✓ *Azure Databricks*

Mediante esta herramienta es posible realizar procesamiento de la información de datos estructurados, no estructurados y semiestructurados, aplicando diferentes técnicas de Big Data y analítica avanzada como algoritmos de Machine Learning y Deep Learning. Aprovecha la computación en memoria y otras optimizaciones. Actualmente mantiene el registro para la clasificación en disco a gran escala.

5. **Capa de utilización y visualización,** Utilizará los resultados de la capa de procesamiento y analítica, permitiendo la consulta de estos datos a través de herramientas BI y reportería como lo es Power BI, office 365 y otras herramientas de visualización de terceros o consumiendo los pronósticos de los modelos entrenados a través de Azure Databricks.

- ✓ *Power BI.*

Servicio de análisis empresarial de Microsoft, su objetivo es proporcionar visualizaciones interactivas y capacidades de inteligencia empresarial con una interfaz lo suficientemente simple como para que los usuarios finales creen sus propios informes.

3.3 Gestión Datos Maestros (MDM)

La gestión de datos maestros provee un conjunto de herramientas para la administración e integración de datos garantizando uniformidad, precisión, administración y coherencia de los datos críticos de una organización, lo cual reduce errores y redundancia en procesos de negocio.

Las soluciones MDM brindan un amplio rango de servicios como:

- **Administración de registros dorados (Golden Record),** realizando la estandarización, eliminación de duplicados y limpieza de datos, permitiendo obtener la información más completa y confiable de registros individuales, lo cual es también denominado *"la versión única de la verdad"* implicando que los usuarios de los datos pueden confiar que es la versión correcta de los datos. Adicionalmente, permite ser totalmente administrable debido a que no requiere creación de códigos de programación.
- **Integración,** datos maestros globales centralizados para toda una organización, debido a que posee conexiones con diversas fuentes, como CRM, bases de datos SQL, ERP y datos en la nube de Microsoft Azure.
- **Flujo empresarial,** refuerza los procesos de negocio ya que permite ver y administrar el rendimiento de las mejoras realizadas.
- **Administración de eventos,** brinda la detección de cambios de los datos en tiempo real.
- **Seguridad,** la administración de permisos sobre los datos están reforzados, como la creación, lectura, actualización y eliminación.

El esquema de implementación de un MDM, parte de conexiones a las diversas fuentes de información que posea la organización que luego es administrada por medio del software MDM, allí se crean las parametrizaciones necesarias para la limpieza, duplicidad, unión de fuentes para la creación de registros dorados, datos maestros entre otros. Finalmente se obtiene una fuente de

información unificada, fiable y coherente que puede ser utilizada para desarrollar las funciones de una organización, proyectos de analítica, informes, etc. La funcionalidad, parametrización, costos de licenciamiento depende del tipo del fabricante del software MDM. A continuación, se ilustra el diagrama de un MDM.

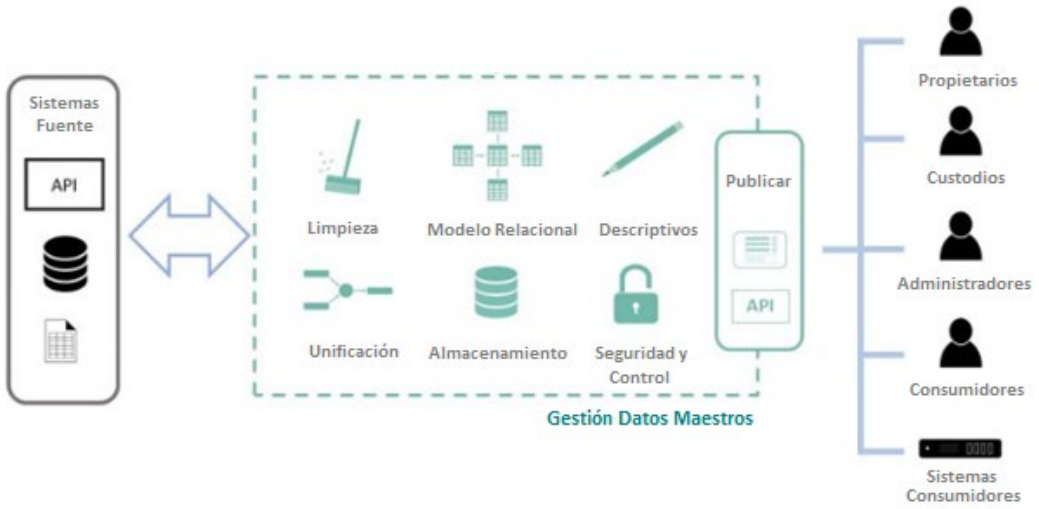


Ilustración 5. Ejemplo de Gestión de Datos Maestros

La implementación de un MDM se realiza para empresas con grandes volúmenes de datos, donde sus datos se encuentran distribuidos en diversas fuentes, que constantemente deben estar realizando cruces de grandes volúmenes de información y la adquisición y/o actualización de datos es de alta frecuencia. Así que el MDM mitiga los tiempos de integración de datos, optimizando la operatividad de las compañías.

Para el caso de EAFIT no se recomienda implementarlo aún, debido al nivel de madurez actual y teniendo en cuenta que el gobierno de datos está en la etapa de definición y aún no se encuentra operativo, además las funciones que realiza un MDM se pueden simular por medio de una correcta implementación de gobierno y uso de herramientas propias de la organización u adquisiciones adicionales como DQ Analyzer, Power BI, etc., que ya han sido descritas en los apartados anteriores del presente documento, ya que el costo de licenciamiento del MDM es bastante elevado.

4 SEGURIDAD DEL DATO

Para EAFIT donde la información de sus estudiantes, productos, servicios entre otros datos se han convertido en su principal activo, es importante que exista un conjunto de buenas prácticas que apliquen transversalmente en el negocio y además contengan procesos claros, estándares de seguridad y trazabilidad. Esto proporcionará las herramientas necesarias que soportarán este pilar fundamental de gobierno de datos sobre nuestro DWH.

Todos estos elementos deberán ser desarrollados de acuerdo con los requerimientos puntuales de algunos actores tales como:

- Los interesados, donde la institución deberá reconocer las necesidades de privacidad y confidencialidad de la información.
- Regulaciones gubernamentales, ya que estas toman lugar en proteger los intereses de algunos actores, asegurando el estricto acceso a la información, transparencia y responsabilidad de custodia.
- Necesidades de acceso del negocio, roles de acceso correctamente identificados que permitan a los usuarios de negocio realizar sus labores de manera adecuada.
- Preocupaciones legítimas del negocio, como obligaciones contractuales, debido a acuerdos que pueden influenciar los requerimientos de la seguridad de los datos.

Ilustración 6. Orígenes de requerimientos de seguridad de datos-Tomado del DAMA

A continuación, se muestra el diagrama contextual general, en referencia al DAMA, donde se puede apreciar el significado de la seguridad de los datos, y los objetivos que se deben tener en cuenta al momento de gestionar la seguridad de los datos.

Seguridad del Dato.

Es la definición, planeación, desarrollo y ejecución de políticas y procesos de seguridad que provean una adecuada autenticación, autorización, acceso y auditoria de los datos y la información.

Objetivos

1. Habilitar un correcto acceso a los activos de datos empresariales dentro del tenant de Azure y modelo de DWH.
2. Entender y cumplir con todas las regulaciones y políticas relevantes para la privacidad, protección y confidencialidad de los datos.
3. Se debe asegurar que la privacidad y confidencialidad de todos los interesados estén aplicados y auditados, de acuerdo con las políticas de tratamiento de datos personales que apliquen.

4.1 Seguridad del dato a nivel de plataformas y almacenamientos empresariales

La primera ventana de exposición de los datos se podría concentrar a identificar objetivos claros dispuestos sobre Azure:

- Políticas de seguridad y custodia de los datos.
- Control de acceso centralizado.
- Creación de roles definidos de acuerdo con el nivel de acceso que se desee otorgar.
- Procedimiento de activación, actualización o borrado de accesos a diferentes módulos de aplicativos que permitan controlar la información que se podrá leer o actualizar desde los usuarios funcionales.
- Definición de los grupos de acceso y esquemas de seguridad que tendrán los roles definidos de administración del DWH.
- Procedimientos estandarizados de seguridad del dato, con responsables de autorización en el mapa de proceso de cada aplicativo y/o estimación de proyectos.
- Trazabilidad de todos los procesos que se llevan a cabo, correspondiente al acceso de los datos desde los aplicativos.
- Comunicación constante entre los distintos responsables de los procesos, logrando gestiones unificadas.
- Metodologías, prácticas y herramientas que ayudarán en la consecución de los objetivos.

4.1.1 Diseño de políticas de seguridad de la información.

Al momento de diseñar una política se debe tener definido la falencia que se desea cubrir y el público objetivo, luego se debe publicar para que toda la organización tenga conocimiento de ella. Debido a que durante los procesos del día a día y regulaciones gubernamentales son de constante cambio, se deben realizar revisiones periódicas a las políticas para verificar si se presenta la necesidad de realizar actualizaciones de estas.

En resumen, una redacción de política de seguridad deberá contener mínimo los ítems mencionados en el numeral 1.4 del presente documento.

4.1.2 Logs de registro de cambios

Azure provee un log de actividades para registrar qué operaciones se realizan sobre los recursos, quién realiza las operaciones, cuando se realizan, el estado de las mismas y las propiedades y valores de los cambios realizados, de los últimos 90 días.

Exportar Logs en Data Lake

El análisis de los logs del Data Lake se puede realizar mediante el Storage Analytics el cual registra información detallada sobre solicitudes exitosas y fallidas a un servicio de almacenamiento. Esta información se puede usar para monitorear solicitudes individuales y diagnosticar problemas con un servicio de almacenamiento, así como también se puede indicar a Azure Storage que guarde los registros de diagnóstico para las solicitudes de lectura, escritura y eliminación de los servicios de cola, tabla y blob. La política de retención de datos que establezca también se aplica a estos registros.

Para el caso del Data Lake se debe hacer uso del Log Analytics Workspace, este servicio permite exportar los eventos que ocurren en una cuenta de almacenamiento o en un Event Hubs; se requiere acceder al Azure Monitor donde se podrán exportar los datos requeridos.

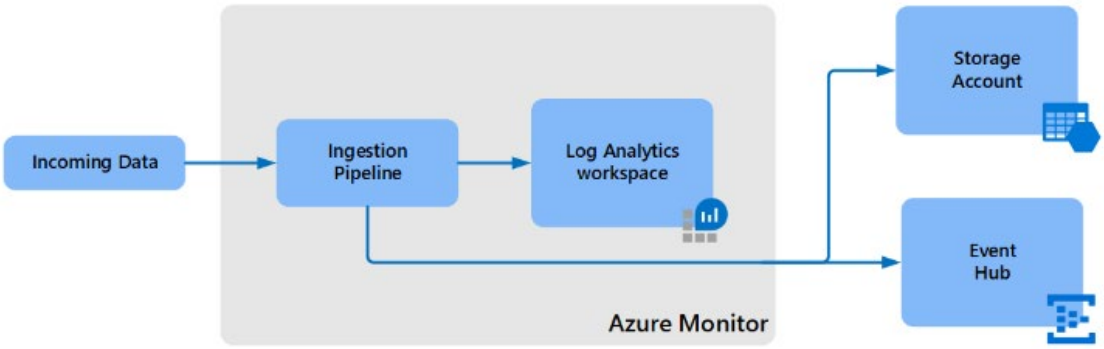
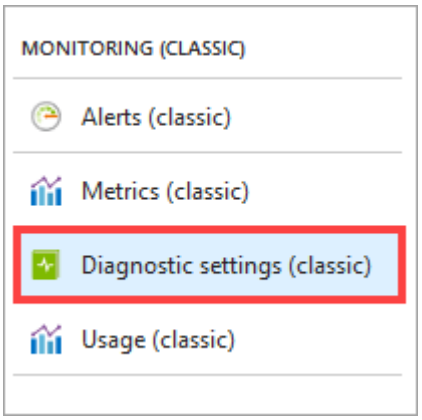


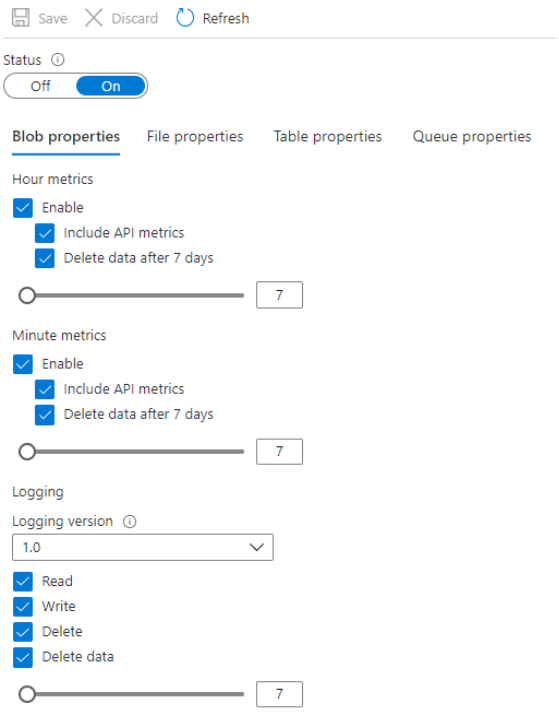
Ilustración 7. Log Analytics Workspace

La imagen muestra como, por ejemplo, un nuevo dato se ingesta y llega al Log Analytics y posteriormente se almacena en el Storage Account o en el Event Hub dependiendo el caso.

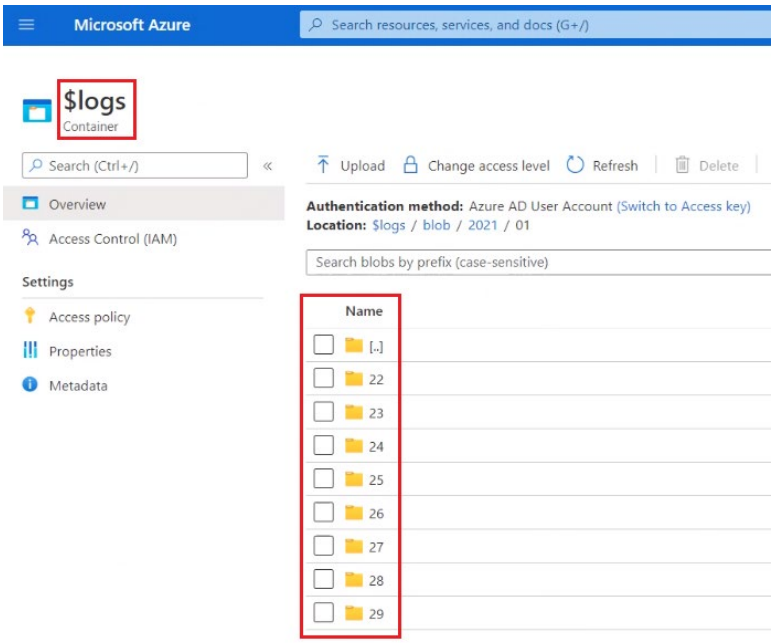
La configuración se debe realizar en la configuración de diagnóstico del Azure Monitor:



Esta opción permite ajustar las siguientes características:



Una vez se configura el servicio, en el blob Storage o Container podrá apreciar el almacenamiento de los logs por jerarquía de fechas así.



En estas carpetas se encuentran los archivos JSON que pueden ser consultados y leídos.

Exportar Logs en Databricks

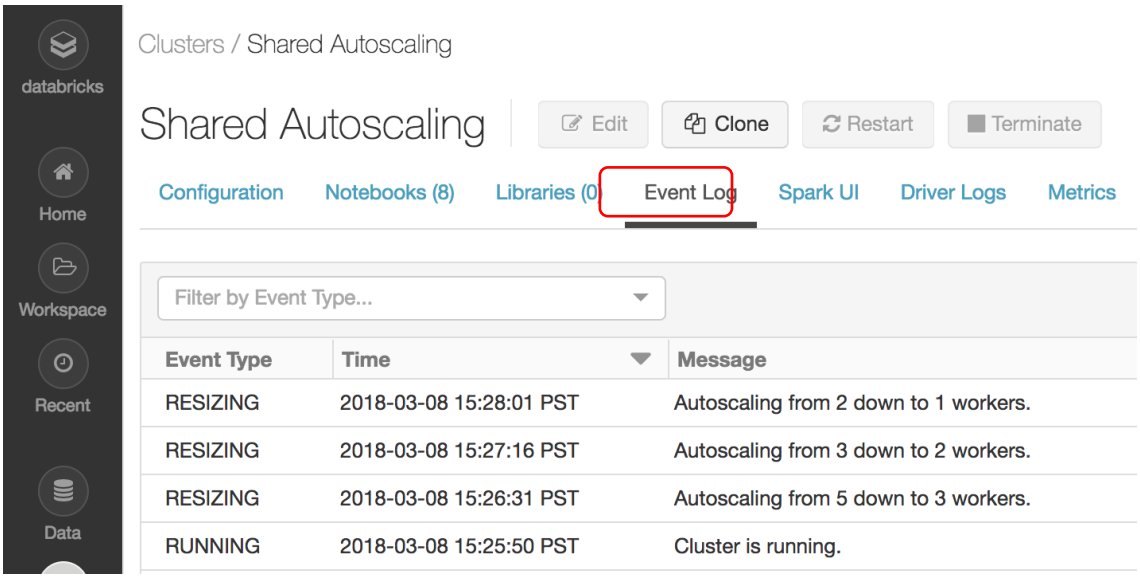
Azure Databricks proporciona tres tipos de registro de actividad relacionada con el clúster:

- Registros de eventos de clúster que capturan eventos del ciclo de vida del clúster, como creación, terminación, ediciones de configuración, etc.
- Registros del controlador y del trabajador de Apache Spark, el cual se puede usar para depurar.
- Registros de secuencias de comandos de inicio de clúster, valiosos para depurar secuencias de comandos de inicio.

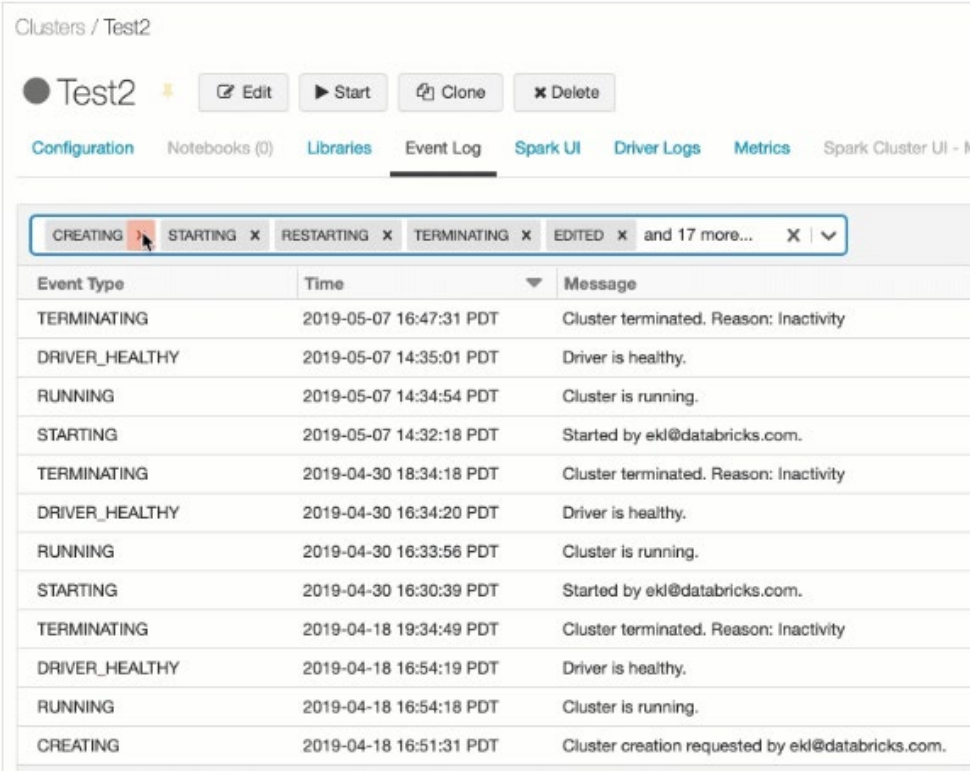
El registro de eventos del clúster muestra eventos importantes del ciclo de vida del clúster que se desencadenan manualmente por acciones del usuario o automáticamente por Azure Databricks. Tales eventos afectan el funcionamiento de un clúster en su conjunto y los trabajos que se ejecutan en el clúster.

Los eventos se almacenan durante 60 días, lo que es comparable a otros tiempos de retención de datos en Azure Databricks.

Para visualizar los eventos del clúster es necesario acceder a servicio



Dentro de las diferentes actividades es posible validar la terminación, inicio, escalado, edición, nodos perdidos y otras actividades propias del Clúster.



4.1.3 Definición de roles de acceso estándar y responsables de su administración según grupos de accesos

Esto permitirá generar los procedimientos que a su vez se convertirán en políticas de la institución, donde se debe definir claramente, lo siguiente:

- Quién identifica los roles necesarios.
- Criterios de creación y aceptación de un rol.
- Quién los autoriza.
- Quién los crea.
- A cuáles sistemas se pueden aplicar estos roles.
- Matriz de privilegios de roles.

Resumen Privilegio	Roles a nivel gobierno de datos			
	Lector de datos	Editor de datos	Usuario	Administrador
Ingresar al aplicativo	X	X	X	X
Realizar operaciones CRUD (CREAR, LEER, ACTUALIZAR, BORRAR)				X
Realizar operaciones de lectura	X	X		X

Algunos ejemplos de roles comúnmente utilizados en estos casos pueden ser:

- Administrador, el cual podrá tener la potestad de realizar todas las operaciones de tipo CRUD (Crear, Leer, Actualizar, Borrar) sobre los datos u elementos del sistema como usuarios o permisos.
- Lector, puede ver la información puntual sobre la cual tiene acceso, pero no puede modificarla en ningún momento.

4.1.4 Manejo de accesos a la información dentro del DWH.

Se debe estandarizar el acceso a través de procesos y políticas bien definidas, documentadas y centralizadas, además es importante que tanto las políticas y procedimientos de creación, mutación y borrado de los usuarios que acceden a los datos a través de las distintas plataformas empresariales se mantengan controlados, auditados y trazables en el tiempo; adicionalmente, también es conveniente centralizar los accesos **(Single Sign on)** a las distintas plataformas para evitar ambigüedad, duplicidad y falta de control, estos mecanismos se pueden soportar sobre herramientas tecnológicas como integración con directorio activo de la compañía, autenticación multi-factor, sincronizados localmente o desplegados en la nube como se muestra a continuación:

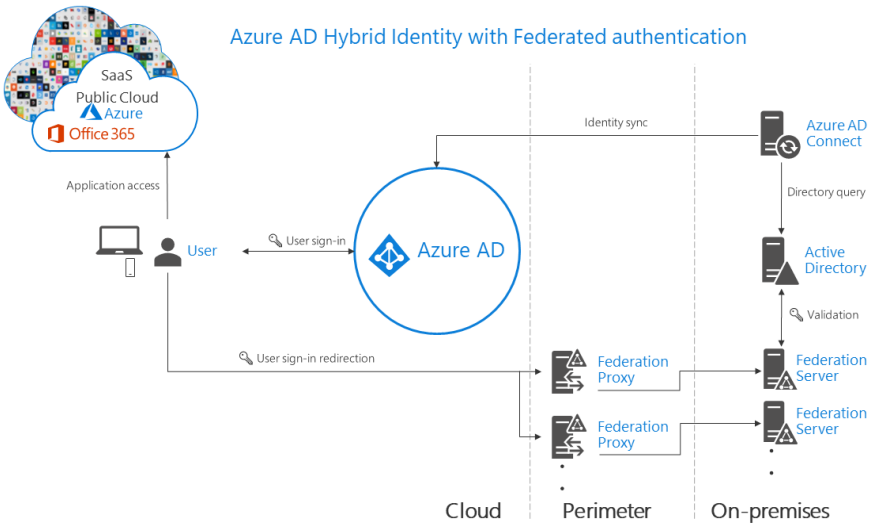


Ilustración 8. Accesos Multi-factor – Microsoft

4.1.5 Roles definidos en Azure

El acceso basado en roles (RBAC) de Azure es un sistema de autorización basado en Azure Resource Manager que proporciona administración de acceso específico a los recursos de Azure como, por ejemplo, proceso y almacenamiento. Existen cuatro roles de Azure fundamentales. Las tres primeras se aplican a todos los tipos de recursos:

Rol de Azure	Permisos	Notas
Propietario	Acceso total a todos los recursos	Al administrador de servicios y a los coadministradores se les asigna el rol de propietario en el ámbito de suscripción.
	Delegar el acceso a otros usuarios	Se aplica a todos los tipos de recursos.
Colaborador	Crear y administrar todos los tipos de recursos de Azure	Se aplica a todos los tipos de recursos.
	Creación de un inquilino en Azure Active Directory	
	No se puede conceder acceso a otros usuarios	
Lector	Ver todos los recursos, pero no le permite realizar cambios.	Se aplica a todos los tipos de recursos.
Administrador de acceso de usuario	Administrar el acceso de usuarios a los recursos de Azure	

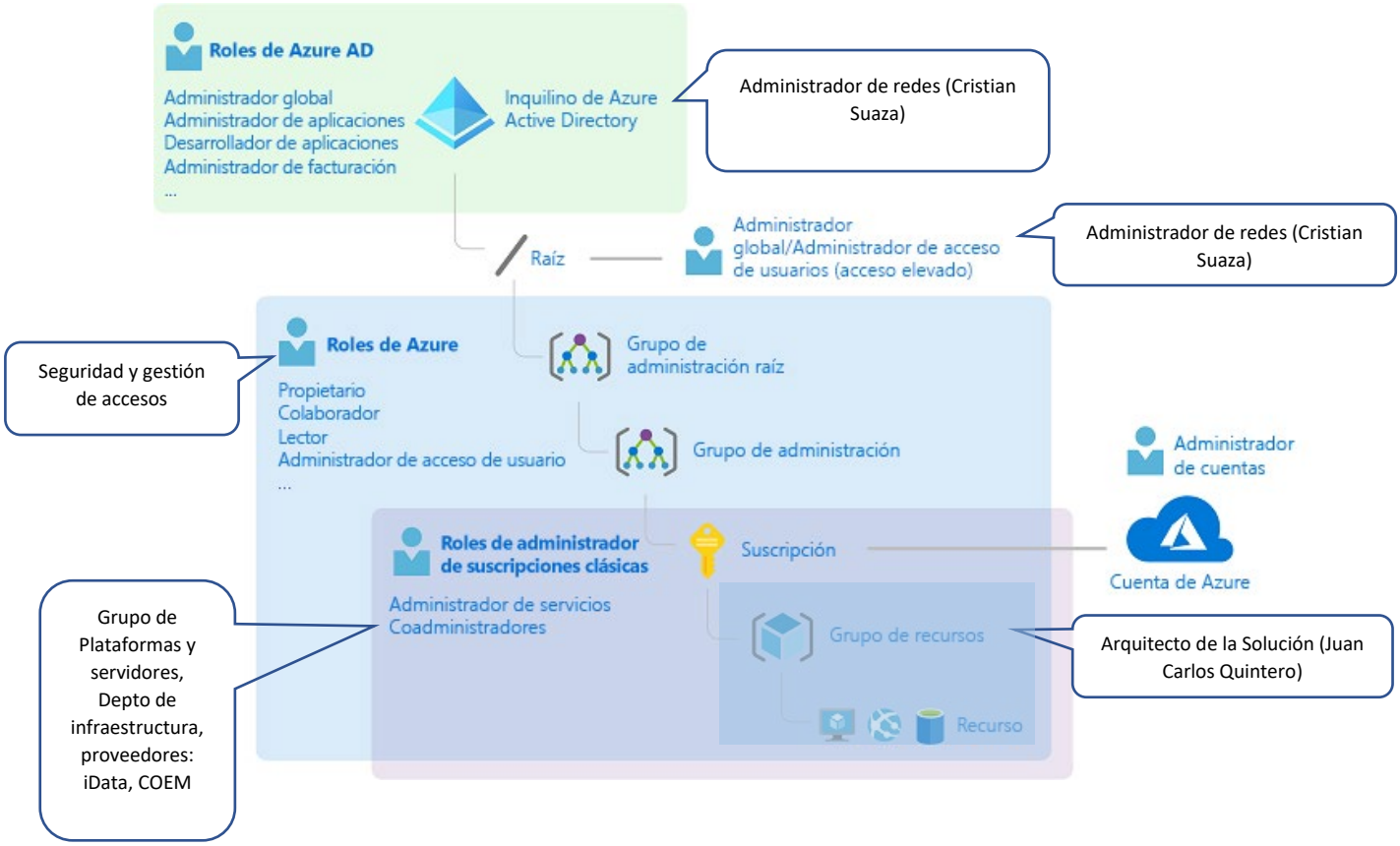


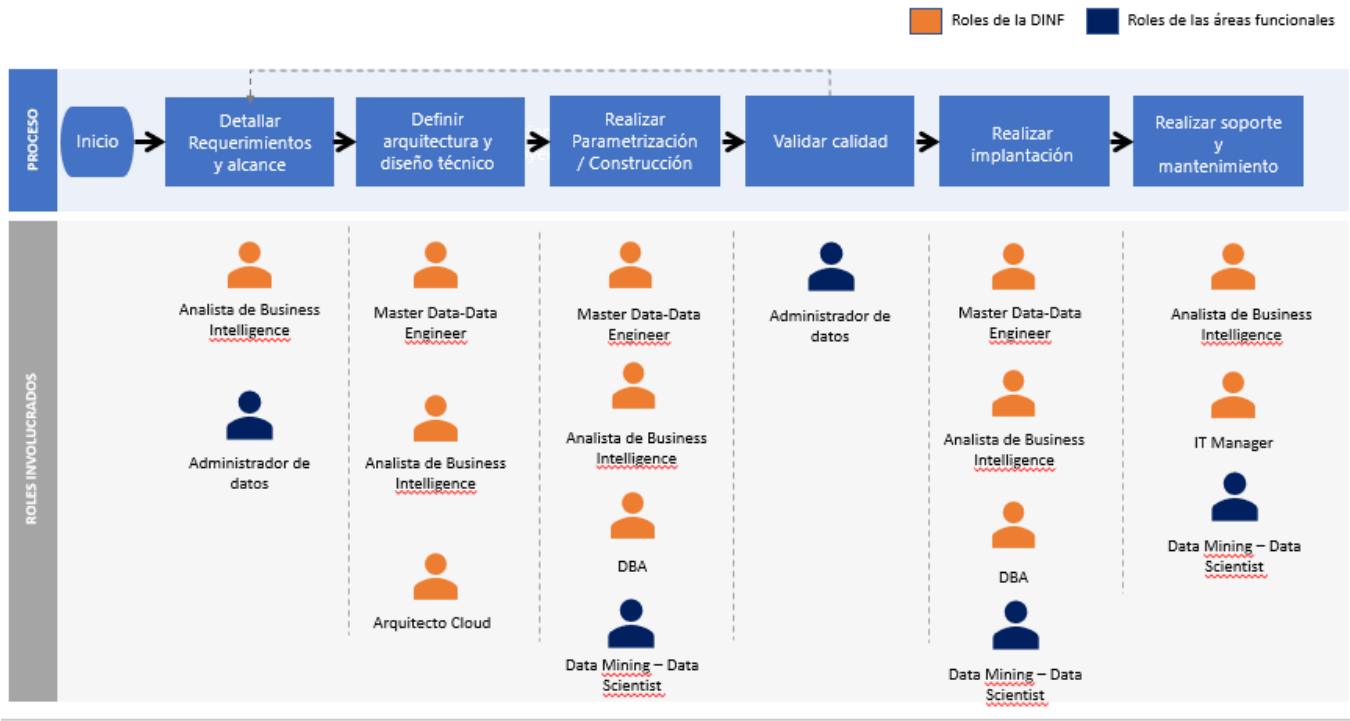
Ilustración 9 Relacionamiento de Roles en Azure

De acuerdo a los perfiles de administración de DWH recomendados en la sección 2.4, se sugiere manejar el siguiente esquema de asignación de permisos:

Rol/Perfil	Servicio	Rol Azure (RBAC) – QA & DEV	Rol Azure (RBAC) - Prod	Colaborador	Notas
Arquitecto Cloud GS_ArquitectoCloud	Datafactory	Owner	Owner	Arquitecto de la Solución	El arquitecto Cloud deberá tener acceso a todos los servicios con rol de colaborador mientras se hace el despliegue de los mismos; después de ello, se recomienda eliminar los accesos o
	Datalake	Owner	Owner	Arquitecto de la Solución	
	Databricks	Owner	Owner	Arquitecto de la Solución	
	SQL Database	Owner	Owner	Arquitecto de la Solución	

					cambiar el rol a lector.
IT Manager GS_itManager	Local Gateway	Colaborador	Colaborador	Administrador de redes	
	Virtual Gateway	Colaborador	Colaborador	Administrador de redes	
	Virtual Network	Colaborador	Colaborador	Administrador de redes	
Master Data-Data Engineer GS_MasterData	Datafactory	Colaborador	N/A	Analista de Sistemas – Célula de Bi y Analítica y Analista de Direccionamiento Estratégico	Se requiere conocimientos en ETL´s
	Datalake	Colaborador	N/A	Analista de Sistemas – Célula de Bi y Analítica y Analista de Direccionamiento Estratégico	
DBA (Administrador de bases de datos) GS_DBMS	Datafactory	Owner	Owner	Grupo Coordinación DBA	El DBA puede apoyar actividades dentro de este servicio
	SQL Database	Owner	Owner	Grupo Coordinación DBA	
Data Mining-Data Scientist GS_DataMining GS_DataScientist	Databricks	Colaborador	N/A	Asistente Direccionamiento Estratégico y Asistente Dirección de formación integral.	
	SQL Database	Colaborador	N/A	Asistente Direccionamiento Estratégico y Asistente Dirección de formación integral.	Deberá realizar procesos de validación de escritura de los datos en la SQL DB
Business Intelligence GS_Business_Intelligence	SQL Database	Colaborador	N/A	Analistas de Sistemas – Departamento Soluciones de Software y Analistas y Asistentes Direccionamiento estratégico.	

Proceso ciclo de vida de la solución de Software



Detallar Requerimientos y alcance: Actividad que realizan entre el analista de BI y el administrador de datos con el objetivo de recibir y entender desde la necesidad funcional aquello que se quiere construir, y co-crear a partir del acompañamiento del analista de BI, un producto entregable viable técnicamente y que satisfaga las expectativas.

Entregable: Formato de documentación de requerimientos y alcance.

Responsable: Analista de BI

Definir arquitectura y diseño técnico: Con base en la información surgida de la actividad previa, el equipo conformado por el analista de BI, el Master Data – Data Engineer y el arquitecto Cloud definirán a la luz de las necesidades del área funcional, las fuentes de datos y la disponibilidad de

tecnología (Componentes del DWH), cual debería ser la arquitectura optima para llevar a cabo el requerimiento.

Entregable: Diagrama de la arquitectura del reporte/analítica.

Responsable: Arquitecto Cloud

Realizar Parametrización / Construcción: Con base en las definiciones realizadas, el equipo compuesto por el Master Data – Data Engineer, el DBA, el analista de BI y el Data Scientist (en caso de tratarse de una iniciativa de analítica) llevaran a cabo las actividades de parametrización y construcción del reporte/analítica, mediante la ingesta de las bases de datos identificadas, la realización de los scripts a través de las vistas materializadas o procedimientos correspondientes y finalmente la visualización definida.

Entregable: Reporte/Analítica en versión de pruebas.

Responsable: Analista de BI

Validar calidad: El administrador de datos junto con los interesados en el área usuaria llevaran a cabo pruebas de validación que determinen el correcto funcionamiento del reporte/analítica o los ajustes pertinentes de cara a la satisfacción del alcance.

Entregable: Visto bueno de la solución o formato de documentación de requerimientos y alcance, incluyendo los ajustes sobre el alcance inicial.

Responsable:

Realizar implantación: Acción de despliegue a producción del reporte/analítica.

Entregable: Reporte/Analítica en versión de producción.

Realizar soporte y mantenimiento: Acciones de soporte, mantenimiento y actualización del reporte/analítica implementado, remitidas bajo demanda por parte del administrador de datos correspondiente.

Entregable: Formato de documentación de requerimientos y alcance.

Accesos Datalake

Con el fin de garantizar un gobierno integral en el componente y sus datos se hará uso del control de accesos basados en listas ACL permitiendo asignar cada archivo y directorio de la cuenta de almacenamiento una lista de control de acceso, haciendo uso de las entidades en Azure Active Directory.

Así mismo se hará integración con el acceso basado en roles (RBAC) donde se tiene varios roles integrados de Azure que puede asignar a usuarios, grupos, entidades de servicio e identidades administradas. Las asignaciones de roles son la forma de controlar el acceso a los recursos de

Azure. Si los roles integrados no satisfacen las necesidades específicas de su organización, puede crear sus propios roles personalizados de Azure.

Accesos Databricks

En Azure Databricks, puede utilizar las listas de control de acceso (ACL) para configurar el permiso para acceder a objetos del área de trabajo (carpetas, cuadernos, experimentos y modelos), clústeres, grupos, trabajos y tablas de datos. Todos los usuarios administradores pueden administrar las listas de control de acceso, al igual que los usuarios que tienen permisos delegados para administrar las listas de control de acceso.

Dentro del propio Databricks se pueden configurar permisos de acceso a objetos del área de trabajo. De forma predeterminada, todos los usuarios pueden crear y modificar objetos de área de trabajo, incluidas carpetas, cuadernos, experimentos y modelos, a menos que un administrador habilite el control de acceso al área de trabajo. Con el control de acceso a objetos de área de trabajo, los permisos individuales determinan las capacidades de un usuario. En este artículo se describen los permisos individuales y cómo configurar el control de acceso a objetos del área de trabajo.

Aptitud	Sin permisos	Lectura	Ejecutar	Editar	Administrar
Enumerar elementos en la carpeta	x	x	x	x	x
Ver elementos en la carpeta		x	x	x	x
Clonar y exportar elementos		x	x	x	x
Crear, importar y eliminar elementos					x
Movimiento y cambio de nombre de elementos					x
permisos, cambiar					x

Para visualizar el control de acceso de todo el servicio ver el siguiente acceso:
https://docs.microsoft.com/es-es/azure/databricks/security/access-control/

Accesos SQL Databse

SQL Database permite administrar centralmente las identidades de usuario de base de datos y otros servicios de Microsoft con la integración de *Azure Active Directory*. Esta funcionalidad simplifica la administración de permisos y mejora la seguridad. *Azure Active Directory admite autenticación*

multifactor para aumentar la seguridad de los datos y de la aplicación, al tiempo que admite un proceso de inicio de sesión único.

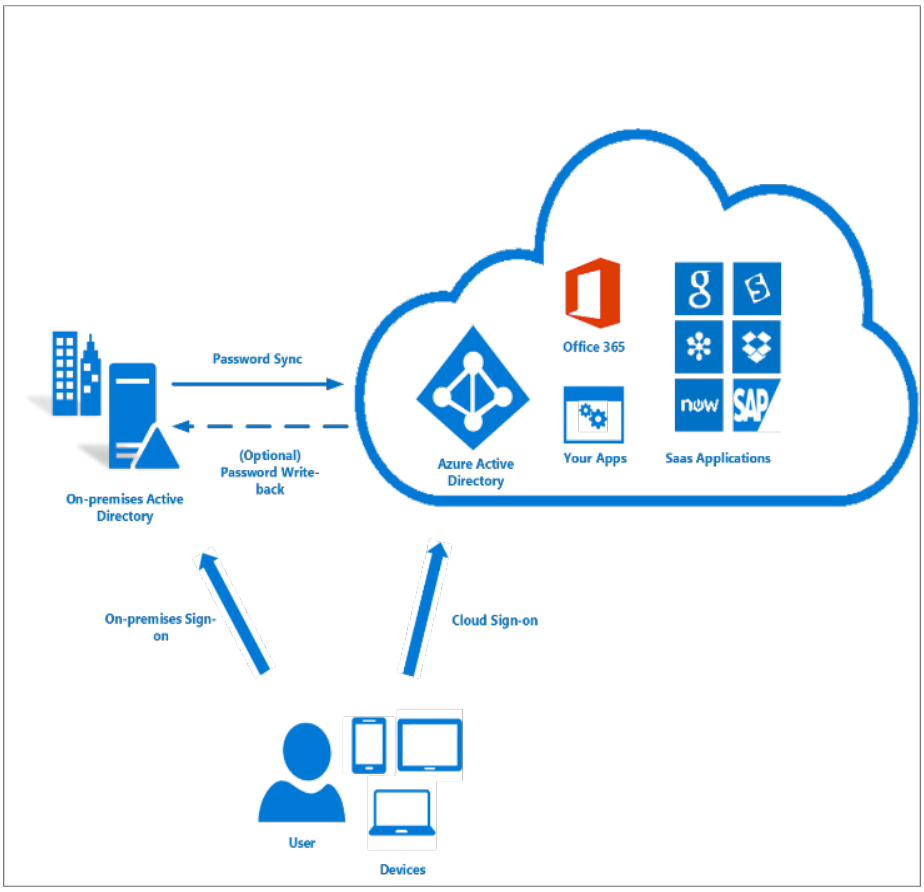


Ilustración 10 Integración Directorio Activo Azure

La autenticación con el directorio activo de Azure ofrece entre muchos beneficios los siguientes:

- Ofrece una alternativa a la autenticación de SQL Server.
- Ayuda a detener la proliferación de identidades de usuario en los servidores.
- Permite la rotación de contraseñas en un solo lugar.
- Los clientes pueden administrar los permisos de la base de datos con grupos externos (Azure AD).
- Azure AD admite conexiones de SQL Server Management Studio que usan la autenticación universal de Active Directory, lo que incluye Multi-Factor Authentication. Multi-Factor Authentication incluye una sólida autenticación con una gama de opciones sencillas de comprobación: llamada de teléfono, mensaje de texto, tarjetas inteligentes con PIN o notificación de aplicación móvil.
- Azure AD admite conexiones similares desde SQL Server Data Tools (SSDT) que usan la autenticación interactiva de Active Directory.

Otra de las características de seguridad ofrecidas por Azure es el servicio de Threat Protection, Advanced Threat Protection para Azure SQL Database, Azure SQL Managed Instance y Azure

Synapse Analytics detecta actividades anómalas que indican intentos inusuales y potencialmente peligrosos de acceder a las bases de datos o de vulnerar su seguridad.

Este servicio permite la crear alertas ya detecta actividades anómalas que indican intentos inusuales y potencialmente dañinos de acceso o ataque a las bases de datos como:

- Una posible vulnerabilidad a la inyección de código SQL
- Intento de inicio de sesión desde una aplicación potencialmente dañina
- Inicio de sesión desde un centro de datos de Azure inusual
- Inicio de sesión desde una ubicación inusual
- Inicio de sesión de un usuario principal que no se ha visto en 60 días
- Inicio de sesión desde una dirección IP sospechosa
- Posible intento de ataque de fuerza bruta de SQL

Dentro de los servicios de seguridad para las bases de datos SQL se encuentra también la encripción de los datos que es un elemento clave dentro del proyecto MVP; Azure ofrece lo denominado como Always Encrypted que es una característica diseñada para proteger la información confidencial, como números de tarjeta de crédito o números de identificación nacional (por ejemplo, los números de seguridad social de EE. UU.), almacenada en bases de datos Azure SQL Database o SQL Server. Always Encrypted permite a los clientes cifrar información confidencial en aplicaciones cliente y nunca revelar las claves de cifrado en Motor de base de datos (SQL Database o SQL Server).

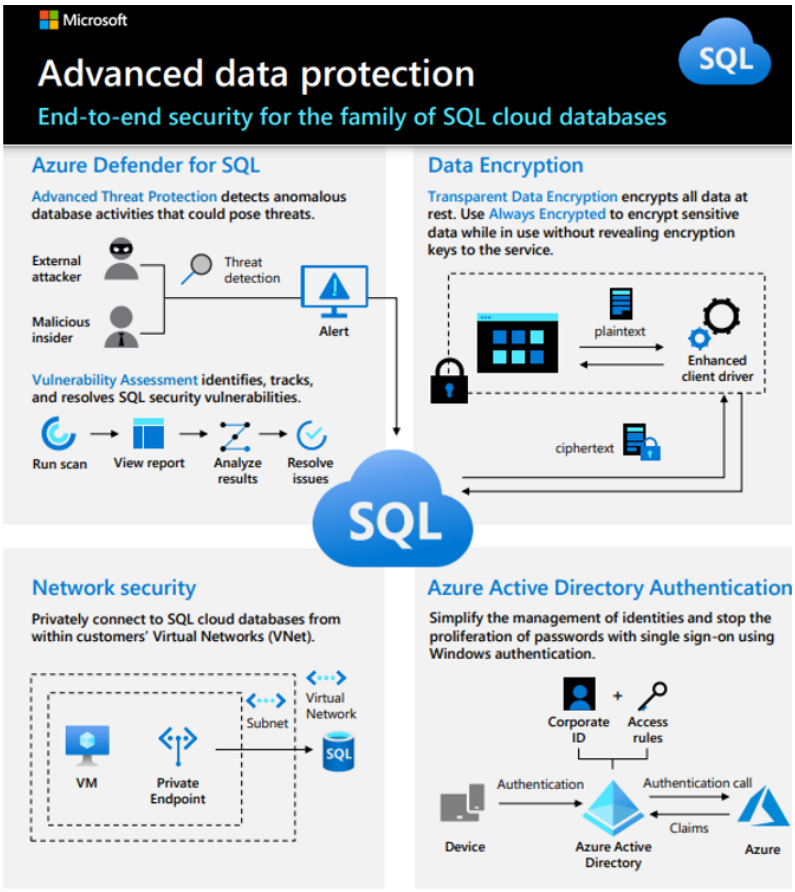


Ilustración 11 Seguridad SQL Database Azure

4.2 Aseguramiento de la continuidad del negocio a través de los datos.

La continuidad del negocio es un elemento crucial para las compañías actuales, los datos hacen parte fundamental de este objetivo, ya que les permite seguir soportando sus procesos misionales de prestación de servicios, sin pausa que afecte las ganancias. Dado lo anterior, toda compañía necesita un plan para la continuidad del negocio en caso de desastre o un evento adverso que afecte sus sistemas y la disponibilidad de uso de sus datos; los administradores de bases de datos (DBA) deben desarrollar un plan de recuperación para todas las bases de datos y servidores de estas, cubriendo escenarios que podrían resultar en la pérdida o corrupción de los datos, tales como:

- Daño de un servidor físico.
- Daño de uno o más discos del dispositivo de almacenamiento.
- Error o pérdida de una base de datos.
- Corrupción de los índices de las bases de datos.

Cada base de datos debería ser evaluada de acuerdo con su nivel de criticidad para ser priorizada. Algunas pueden ser esenciales para el negocio y necesitarían estar disponibles inmediatamente, las menos críticas podrían ser restauradas cuando los sistemas primarios estén en línea de nuevo.

Para este fin, es necesario tener un plan de copias regular regido por un SLA (Acuerdo de Nivel de Servicio) donde se especifique la frecuencia y tipo, como diario diferencial, semanal completa; junto a una política clara que identifique acciones y controles para tener en cuenta, como la cadena de custodia de esas copias. El plan anterior debe estar acompañado de un plan completo de pruebas de recuperación ante desastres periódico apoyado en un proceso establecido, donde se guarden los resultados obtenidos de éstos, que ayudarán a realizar trazabilidad y posteriormente, mejora de los procesos.

Recuperación de desastres en nube con Azure Site Recovery.

En cualquier entorno se está en riesgo de pérdida de la información, por ello es necesario contar con servicios que brinden las opciones necesarias para mitigar estos riesgos; uno de estos servicios es Site Recovery. Se sugiere a EAFIT adoptar una estrategia de continuidad empresarial y de recuperación ante desastres (BCDR) que mantenga sus datos seguros, y sus aplicaciones y cargas de trabajo en línea cuando se produzcan interrupciones planeadas o imprevistas.

Azure Recovery Services ayuda con su estrategia de BCDR:

- **Servicio Site Recovery**: Site Recovery ayuda a garantizar la continuidad empresarial manteniendo las aplicaciones y cargas de trabajo empresariales en funcionamiento durante las interrupciones. Site Recovery replica las cargas de trabajo que se ejecutan en máquinas físicas y virtuales desde un sitio principal a una ubicación secundaria. Cuando se produce una interrupción en el sitio principal, se conmuta por error a la ubicación secundaria y se accede desde allí a las aplicaciones. Cuando la ubicación principal vuelva a estar en ejecución, puede realizar la conmutación por recuperación en ella.
- **Servicio Backup**: El servicio Azure Backup mantiene los datos seguros y recuperables.

Site Recovery puede administrar la replicación de:

- Máquinas virtuales de Azure que se replican entre regiones de Azure.
- Máquinas virtuales locales, máquinas virtuales de Azure Stack y servidores físicos.

Los servicios ofrecidos por Site Recovery son:

- Solución de BCDR simple
- Replicación de máquinas virtuales de Azure
- Replicación de máquinas virtuales local
- Replicación de la carga de trabajo
- Resistencia de datos
- Destinos RTO y el RPO
- Mantener la coherencia de aplicaciones a través de la conmutación por error
- Pruebas sin interrupciones
- Conmutaciones por error flexibles
- Planes de recuperación personalizados
- Integración de BCDR

- Integración de Azure Automation
- Integración de red

El flujo de datos se esquematiza en la siguiente imagen:

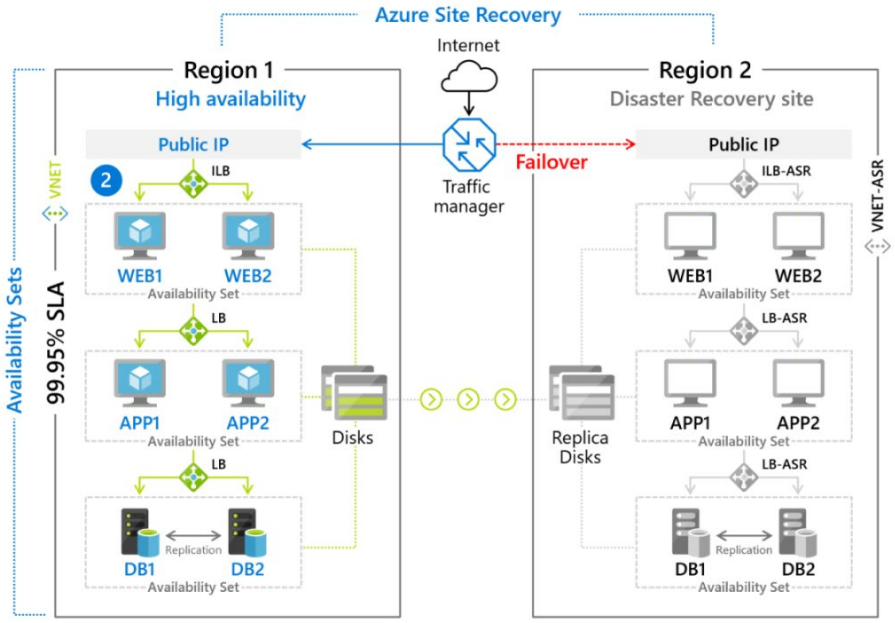


Ilustración 12 Azure Site Recovery

La implementación del Site Recovery permitirá a EAFIT contar con una herramienta mediante la cual se podrá recuperar la información en caso de desastres.

5 RECOMENDACIONES

Las siguientes acciones se recomiendan para iniciar la implementación del programa de gobierno de datos de manera que se puedan lograr los objetivos propuestos:

- **Empoderamiento del área gobierno de datos dentro de LA UNIVERSIDAD EAFIT:** Como primer medida cualquier modificación que afecte directamente los datos en los sistemas de información debe llevar el visto bueno del equipo ampliado de gobierno de datos; de la misma manera los proyectos deben contar con la participación del área de gobierno de datos para ayudar a definir las reglas relacionadas para los datos y vigilar el cumplimiento de estas; finalmente el área de gobierno de datos se encarga, o en su defecto autoriza, quiénes pueden y en qué forma van a actualizar los datos.

- **Revisión inicial de políticas:** Realizar una revisión inicial de las políticas y estándares existentes y realizar su comparación según la calidad del dato, para verificar si realmente se está cumpliendo. De lo contrario, tomar las acciones necesarias ya sean correctivas o preventivas**:** Como primera medida cualquier modificación que afecte directamente a los datos en los sistemas de información debe llevar el visto bueno del área de gobierno de datos.

- **Definición proyectos:** Los proyectos deben contar con la participación del área de gobierno de datos para ayudar a definir las reglas relacionadas para los datos y vigilar el cumplimiento de estas; finalmente el área de gobierno de datos se encarga, o en su defecto autoriza, quiénes pueden y en qué forma van a actualizar los datos.

- **Formalización de políticas de accesos:** Definir las políticas, personal y sistematizar el control de accesos a las fuentes de información, permitiendo tener una trazabilidad y administración de los accesos en toda la organización.

- **Definición de los líderes de administración del DWH:** El comité de gobierno deberá definir los líderes de administración del DWH que se sugiere hagan parte del mismo comité, teniendo en cuenta el rol que desempeñe dentro de la universidad y los skills que posean en administración de recursos en la nube de Azure.

- **Definición de metas:** El comité de gobierno debe implementar la definición de metas de forma periódica, dado que las tecnologías y las metas /visiones organizacionales están en constante evolución.

- **Capacitación continua:** Realizar capacitaciones en el grupo de operaciones de gobierno de datos, que les permita contar con las herramientas y conocimientos necesarios para ejecutar de forma adecuada sus funciones.

- **Revisión y actualización del diccionario de datos:** Buscar en los tableros de Power BI problemas de calidad de datos que según el diccionario de datos no deberían ocurrir de acuerdo con la naturaleza del atributo. Por ejemplo, atributos con listas desplegables y que se evidencie valores diferentes a los habilitados por la lista. Finalmente concluir si está mal diligenciado el diccionario o tomar acciones para la mejora de calidad.

- **Correcciones automáticas de calidad datos:** En caso de haber identificado las falencias de calidad y que éstas se presenten de formas repetitivas como la inserción de espacios al final/inicial de palabras, caracteres especiales, entre otros, crear procesos automáticos en las bases de datos que de forma constante corrijan los errores.

- **Inclusión de metadatos relacionados con el linaje del dato:** Para realizar auditorías en los datos se recomienda incluir metadatos que relacionen el usuario, fecha, aplicativos, motivos y métodos con los cuales un dato ha sido creado o modificado.

- **Enriquecimiento de datos:** Para casos donde existen problemas de calidad en atributos relacionados con contacto como direcciones, correos, nombres entre otros, se puede acudir a

una estandarización de forma paga, ya que existen diversas organizaciones que proveen estos tipos de información con garantía de calidad como lo son las centrales de riesgo.

6 CONCLUSIONES

- Un gobierno de datos exitoso necesita la orquestación formal de personas, procesos y tecnologías que permita a una organización aprovechar los datos como un activo empresarial.
- Gobierno de datos es una iniciativa donde sus resultados se obtiene de mediano a largo plazo, evidenciado en la reducción de costos, riesgos y operatividad.
- La implementación de un programa de gobierno de datos logra que la información sea confiable, ayudando a tomar decisiones acertadas de acuerdo con la realidad y evitar sanciones.
- El empoderamiento del área de gobierno va ligado al apoyo de los altos ejecutivos como también a un oportuno apoyo del plan de gestión y apropiación del cambio.
- Las decisiones de gobierno deben estar enfocadas en acciones tanto correctivas como preventivas.

- Se deben definir siempre las metas que se quieren alcanzar, para tener una ruta a seguir clara.
- Fortalecer las políticas de seguridad de los datos existentes a nivel de control y administración de accesos, restauración y copia de información garantiza la protección de datos como un activo de gran valor para la organización.
- Implementar y definir de forma clara en el equipo ampliado de gobierno de datos el proceso de limpieza de datos, si estos se van a hacer de forma manual verificar que los colaboradores definidos como responsables tienen los permisos, conocimientos necesarios y garantizar que se tiene un plan de acción en caso de que las correcciones no den el resultado esperado y se necesite de un plan de restauración de datos.
- Existen herramientas que facilitan la mejora de calidad de datos y crean una versión única de la verdad como lo es un sistema para la gestión de datos maestros (MDM), pero debido al nivel actual de madurez de EAFIT y los costos que conlleva adquirir esta herramienta, se recomienda hacer uso de las capacidades actuales de los sistemas de información complementado con las entregadas este acompañamiento como Power BI, Databricks y el Datalake. Además, si se ejecuta a cabalidad el marco de trabajo se pueden alcanzar los mismos resultados de un MDM.
- Se debe priorizar la centralización y administración de roles de acceso a los sistemas de información existentes basados en grupos de seguridad.