**Portafolio 8: Integración de Datos**

**M8 AE 1: Diagnóstico y propuesta de solución ante un problema de integración de datos**

**Contexto:** Una organización presenta demoras en la generación de reportes, errores por duplicación de información y decisiones basadas en datos inconsistentes. El equipo identifica que las áreas de ventas, marketing y logística trabajan con bases de datos aisladas y sin conexión entre sí. Tu rol es analizar la situación y proponer una solución basada en buenas prácticas de integración de datos.

**Consigna:** A partir del caso planteado, describe qué tipo de integración sería más adecuada, qué tareas y herramientas aplicarías, y qué beneficios esperas obtener. Fundamenta tu propuesta con conceptos clave vistos en clase.

**Paso a paso:**

1. **Analiza el caso:**
   1. **¿Qué tipo de problema de datos presenta la organización?**

La organización tiene problemas de silos de información, donde cada área trabaja con información de manera aislada e inconsistente, lo que no permite una buena toma de decisiones.

* 1. **¿Qué riesgos genera para la operación y la toma de decisiones?**

Dentro de los riesgos está la pérdida de información o duplicidad de información, lo que no permite la toma decisiones con datos certeros y actualizados.

* 1. **Elige el tipo de integración más adecuado: ¿Batch, en tiempo real, virtualización, federación?**

El tipo de integración sugerida es la consolidación de datos, en un repositorio central, lo que permitirá el análisis de datos históricos, facilitando la calidad y acceso de los datos. Se sugiere el método Batch para la lectura/carga de datos por lotes.

**Justifica tu elección según la necesidad de sincronización y volumen de datos.**

La consolidación de datos, si bien puede ser costoso, y requiere mayor espacio de almacenamiento, permitirá la mejor emisión de reportes, información fidedigna y sincronización de los datos.

1. **Identifica las tareas del proceso de integración:**
   1. **Menciona cómo llevarías a cabo la extracción, transformación y carga (ETL).**

En primera instancia se debe obtener la información desde los diferentes orígenes y estandarizarla en algunas bases de datos que permitan el acceso de las diferentes áreas y que permitan el almacenamiento de manera centralizada. Para ello se ha mencionado que las herramientas existentes en el mercado permiten realizar el proceso ETL de una forma más dinámica y sencilla.

CÓDIGO ABIERTO: menor costo, comunidad que facilita el uso.

Talend open Studio

Apache NiFi

APLICACIONES DE PAGO: más robustas, con mejor soporte y respaldo, líderes ene l mercado

AWS Glue

Google Cloud DataFlow (GCD)

* 1. **¿Qué validaciones aplicarías para garantizar calidad de datos?**

Buscaría la normalización de los datos, evitando valores duplicados, determinando políticas para el tratamiento de valores nulos o vacíos y estandarizaría las estructuras y realizaría una integración semántica para mejor comprensión y procesamiento.

1. **Seleccione herramientas posibles: Nombra al menos una herramienta o tipo de software (por ejemplo, Talend, Pentaho, Apache Nifi, etc.).**

Dentro de las Herramientas disponibles existen de código abierto y de pago, la elección de ellas dependerá de la decisión de implementación, la proyección del negocio, el conocimiento técnico del equipo y del tratamiento de los datos. Apache NiFi, una plataforma de integración de datos de código abierto que permite crear flujos de datos visuales (dataflows). Es ideal para mover datos entre sistemas, aplicar transformaciones ligeras y realizar enrutamientos dinámicos. Su interfaz intuitiva y su capacidad de procesar datos en tiempo real o por lotes la convierten en una opción popular para equipos técnicos.

Otra tecnología muy utilizada es Talend, una suite completa de integración de datos que incluye conectores a cientos de fuentes, herramientas de transformación visual, limpieza de datos y capacidades de gobierno. Talend ofrece versiones open source y enterprise, e integra funcionalidades de ETL,

calidad, preparación y orquestación de datos.

1. **Redacta una breve propuesta escrita: Incluir el diagnóstico, propuesta de solución y beneficios esperados (como mejora en la agilidad analítica, visión unificada, calidad de datos).**

**Diagnóstico**: La organización sufre de **silos de datos**, donde cada departamento gestiona su información de forma aislada. Esto lleva a **inconsistencias** y a la **duplicación** de información, lo que ralentiza la generación de reportes y afecta negativamente la toma de decisiones estratégicas.

**Propuesta de Solución**: Implementar un **proceso ETL de tipo Batch** para centralizar los datos de las áreas de ventas, marketing y logística en un **Data Warehouse**. Este proceso garantizará que los datos se extraigan, limpien, transformen y unifiquen antes de ser cargados en un único repositorio y compartido por todas las áreas.

**Beneficios esperados**:

* **Visión unificada**: Todos los equipos trabajarán con una única fuente de verdad.
* **Mejora en la calidad de los datos**: Se eliminarán inconsistencias y duplicados.
* **Aumento de la agilidad analítica**: La generación de reportes será más rápida y precisa.
* **Toma de decisiones más informada**: Las decisiones estratégicas se basarán en datos fiables y consistentes, lo que mitigará los riesgos operativos.

**M8 AE 2: Diseño de Flujo e Ingesta batch para consolidación de ventas regionales.**

**Contexto: Una organización con múltiples sucursales en todo el país necesita consolidar, de forma diaria, la información de ventas que recibe en archivos .CSV. Estos archivos provienen de distintos sistemas y presentan formatos variables. El área de Business Intelligence solicita implementar un flujo de ingesta batch que procese, transforme y unifique estos datos cada noche, dejándolos listos para análisis.**

**Consigna: Diseña un flujo de ingesta de datos en modalidad batch que resuelva esta problemática utilizando herramientas ETL. El flujo debe contemplar extracción, transformación y carga, especificando tareas técnicas y buenas prácticas.**

**1.Objetivo de la ingesta:** El objetivo de la ingesta de datos es, tras hacer el proceso de transformación y unificación diario de los datos, poder consolidar la información de las diferentes sucursales en un data Warehouse con el objetivo de poder hacer análisis y poder realizar una mejor toma de decisiones.

**2. Describe la Estructura del flujo ETL:** El proceso de obtención y consolidación de los datos se realizará de manera diaria, en batch recibiendo archivos en formato .CSV que provienen de las diferentes sucursales y con formatos variables, la posterior normalización y estandarización de los datos, para finalmente consolidar la información en un repositorio unificado que permite el acceso controlado y con datos unificados para toda la organización.

**Extracción:** El proceso de extracción consistirá en recibir desde cada una de las sucursales lo archivos .CSV mediante una programación diaria que permita contar con toda la información antes de realizar la transformación, por ejemplo: la información debe estar disponible cada día a las 20:30 horas, validando que se recibió todos los archivos de cada una de las sucursales. Se deben considerar procedimientos adicionales y alternativos ante posibles errores o retrasos.

**Transformación:** La transformación consiste en unificar los datos estandarizando sus formatos, normalizando su estructura, eliminando información duplicada, imputando datos en blanco, etc. de acuerdo con definiciones previas corrigiendo información que se considere como errónea o deficiente.

**Carga:** Después de la transformación y cuando los datos ya están validados, se realizará un proceso de carga en un data Warehouse donde los quedarán disponibles para hacer análisis y posteriormente tomar las decisiones que corresponda.

**3. Elige la Herramienta adecuada:** Se sugiere usar Apache NIFI,su interfaz visual basada en flujos (dataflows), su capacidad para configurar transformaciones, control de errores, priorización y manejo de formatos diversos la convierten en una de las opciones más completas y flexibles. Es eficiente por su manejo en memoria y integración con diferentes herramientas. Además, tiene bajo costo en relación con otras herramientas.

**4. Incluye al Menos 2 Buenas Prácticas del proceso Batch:**

**Validación de datos:** Se debe verificar que lleguen todos los datos y que los datos sean válidos, que contengan información consistente. Después de la transformación se debe validar que los datos han sido transformados de manera correcta y que son consistentes y fidedignos para su interpretación.

**Trazabilidad:** Se debe monitorear cada proceso, buscando posibles errores, cuellos de botella o inconvenientes técnicos, entre otros, con la finalidad de poder corregirlos, realizar mejoras y documentar cada proceso en la implementación.

**Programación fuera de horarios de carga del sistema:** Con la finalidad de dar disponibilidad a los procesos y asegurar la incorporación de cada sucursal se programará la recepción de datos después del cierre diario y se pueda contar con la mayor capacidad de recursos de los sistemas.

**Nodo de Manejo de errores:** Permite determinar en paso se producen los errores del proceso y poder detener el proceso o continuar dejando logs del error para corregir.

**5.Representa el flujo en forma de esquema Textual**

* Establecer procedimientos para el envío, recepción y procesamiento de datos.
* Automatización de los procesos.
* Recepción de datos y validación
* ETL de los datos, considerando la eliminación de datos duplicados, y normalización.
* Validar los datos procesados, asegurando la consistencia.
* Carga de los datos al Datawarehouse
* Generación de informes, reportes y Dashboards para el análisis de información.

**Diseño de Flujo ETL Batch para consolidación de clientes multi origen:**

**Contexto: Una empresa de retail posee sistemas independientes de registro de clientes en sus tiendas físicas, sitio web y aplicación móvil. Actualmente, la información está fragmentada, duplicada y sin formato unificado. El área de analítica requiere contar con una base única, consolidada y actualizada diariamente para analizar comportamiento y fidelización. Se te solicita proponer un flujo de ingesta batch que resuelva esta necesidad.**

**Consigna: Diseña un flujo de ingesta de datos en modalidad batch, utilizando herramientas ETL, que permita consolidar registros de clientes desde múltiples fuentes, transformarlos y cargarlos a un repositorio unificado.**

**Paso a paso:**

**Define los objetivos del flujo:** Se deben establecer procesos de carga desde las diferentes fuentes, tiendas físicas, sitio web y aplicación móvil. Realizar las transformaciones necesarias para asegurar la calidad de la información y cargar los datos consolidados en un datawarehouse, con procesos de carga en batch que se actualizará diariamente en los horarios de menor carga de los servicios.

**¿Qué problema concreto resuelve? ¿Qué datos deben consolidarse?**

Resuelve el problema de silos de información, consolida los datos y permite mejorar la toma de decisiones. Se debe consolidar los datos de los productos, los clientes, las compras, actividad en la página web y en la APP.

**Describe el proceso ETL:**

**Extracción:** La información proviene de diferentes fuentes en el caso de la tienda física proviene de archivos .CSV consolidados del proceso de venta diario desde la página web se puede ver la interacción de los clientes con la página y los procesos de compra realizados en línea desde la APP se puede monitorear las consultas, de compra y pagos. Toda esta información debe ser consolidada en una base de datos única que se actualizará diariamente, determinando un horario de extracción de datos, tras el cierre de tiendas, por ejemplo, a las 01:00 hrs.

**Transformación:** En el proceso de transformación se debe considerar la unificación o normalización de la información obtenida desde las 3 fuentes se debe verificar que no exista duplicidad en los datos que no que no haya campos en blanco o que los formatos de cada campo sea el correcto de acuerdo con lo definido para ello. Por ejemplo: definir los formatos de RUT, Nombre, validar correos, mayúsculas/minúsculas, etc.

**Carga:** El proceso de carga considera el almacenamiento de información **en la nube** para evitar tener que implementar equipamiento propio y asegurar el funcionamiento permanente de la nueva base centralizada, disponible para los equipos de análisis y con gran disponibilidad.

**Selecciona una herramienta ETL:**

Se sugiere utilizar AWS Glue. Su elección se justifica por ser un servicio en la nube, y a la gran cantidad de herramientas que ofrece para la administración de datos:

Integración nativa con la nube: Al ser un servicio de Amazon Web Services, se integra perfectamente con otros servicios como S3 DynamoDB y Redshift, facilitando la construcción de un ecosistema de datos en la nube.

Procesamiento sin servidor (Serverless): No se necesita gestionar servidores, lo que simplifica la operación. AWS Glue escala automáticamente para manejar la carga de trabajo.

Funcionalidad de deduplicación (FindMatches): AWS Glue tiene una característica que facilita la deduplicación de registros, incluso cuando no hay identificadores únicos exactos.

**Incorpora buenas prácticas:**

**Ventana de ejecución:** La ejecución del flujo se planifica para que ocurra en una ventana de tiempo fuera del horario operativo (por ejemplo, entre la 1 a.m. y 4 a.m.). Esto evita afectar el rendimiento de los sistemas transaccionales y de los usuarios.

**Nodos de validación:** Se incluyen pasos específicos en el flujo para validar los datos. Por ejemplo, si un archivo de una tienda tiene un formato incorrecto o faltan campos obligatorios, el flujo lo detecta y lo desvía a un nodo de manejo de errores, evitando detener los procesos de carga y procesamiento.

**Logs y alertas**: Se configuran logs y alertas en caso de cualquier falla (ej. error de conexión a la base de datos o un archivo corrupto). Especificando las acciones ante las fallas. Además, se registran los detalles del error en un log, lo que permite una revisión posterior para depurar y resolver el problema.

**Propuesta de forma esquemática (paso a paso textual)**

**Activación Diaria:** El flujo se inicia automáticamente a la hora programada.

**Extracción de Orígenes:**

Leer los archivos .CSV de tiendas.

Conectar a la base de datos del sitio web y extraer datos.

Conectar a la API de la APP móvil y extraer datos.

**Validación:** Se aplican reglas de validación. Los datos que no cumplen los requisitos se envían a un repositorio de errores y se genera una alerta.

**Transformación:** Los datos válidos se limpian, normalizan y, se deduplican.

**Unificación:** Todos los datos extraídos se consolidan en un único conjunto de datos.

**Carga:** El conjunto de datos final, limpio y consolidado se carga de forma incremental en la tabla de registro de clientes en la nube AWS.

**Finalización:** Se notifica al equipo de analítica que los datos están listos para su uso y se archivan los archivos de origen.

**M8 AE 3:** **Monitoreo en tiempo real de transacciones con Kafka**

**Contexto: Una empresa fintech desea implementar un sistema de monitoreo en tiempo real para detectar patrones sospechosos en las transacciones de sus clientes. Actualmente, la información se procesa en batch al final del día, lo que genera demoras críticas en la detección de fraudes. Se necesita implementar un flujo de ingesta en streaming que permita capturar y procesar eventos a medida que ocurren utilizando una plataforma de mensajería como Apache Kafka.**

**Consigna: Diseña un flujo de ingesta de datos en modalidad streaming utilizando una plataforma de mensajería asíncrona (Kafka, Flink o Spark Streaming), orientado a resolver la problemática de monitoreo en tiempo real.**

**Paso a paso:**

**Define el objetivo del flujo:** el objetivo es hacer la ingesta de datos en tiempo real para dar respuestas inmediatas a la detección de fraudes, y tomar las acciones necesarias, de esta forma evitamos tener que espera los procesamientos en batch, lo que significaría tomar acciones muy tarde ante posibles riesgos.

**¿Qué tipo de eventos serán capturados?:** Se capturarán todos los datos de las transacciones en tiempo real, como giros, transferencias, pagos, con la finalidad de detectar posibles fraudes

**¿Qué se busca detectar o procesar en tiempo real?** Busca detectar fraudes y operaciones atípicas, como movimientos muy frecuentes, de altos montos o desde dispositivos o lugares sospechosos.

**Elige la herramienta principal: ¿Kafka, Flink o Spark Streaming? Justifica tu elección brevemente:**

Se implementará la solución en Kafka, considerando su procesamiento por eventos y la velocidad de procesamiento, la posibilidad de recibir información de variadas fuentes, integrarse con diferentes plataformas, etc.

**Esquematiza el flujo general: Origen de datos:** Los datos pueden venir desde diferentes fuentes, APP, aplicación de escritorio, directamente desde la institución, las cuales deben ser analizadas.

**Plataforma de mensajería:** Se debe implementar mensajería en ambos mecanismos (pull y push), con la finalidad de rescatar la información desde los productores y recibir la información que se genere.

**Procesamiento:** Se debe analizar la información, validar que es correcto y gatillar alertas que permitan tomar las acciones necesarias para detener los posibles fraudes.

**Destino final:** Se envía los datos a los departamentos correspondientes, con la finalidad de tomar acciones para detener el posible fraude. De acuerdo con lo determine la organización se podrían tomar otras acciones, como bloqueo temporal de cuentas, confirmación del cliente, etc. Finalmente, toda la información es almacenada para posterior análisis y posible entrenamiento de procesos automáticos.

**Indica al menos dos beneficios de usar streaming frente al procesamiento batch.**

El procesamiento en streaming se realiza en el momento, en tiempo real, permitiendo procesar los datos de manera inmediata. Batch, por otro lado, debe esperar de acuerdo con lo programado para enviar la información.

Permite realizar acciones o tomar decisiones en tiempo real, mejorando los resultados inmediatos y la experiencia del cliente.

**Redacta el flujo en forma de lista o paso a paso claro.**

1. **Captura de datos:** La aplicación registra una nueva transacción y la envía como un evento a un **productor de Kafka**.
2. **Ingesta:** El productor escribe el evento de la transacción en el clúster de Kafka.
3. **Procesamiento:** Un trabajo de **Apache Flink** consume los eventos del clúster y clasifica las transacciones.
4. **Detección de fraude:** Flink procesa cada evento, aplica las reglas de detección y ejecuta los modelos de *machine learning* para asignar un puntaje de riesgo a la transacción.
5. **Generación de alertas:** Si el puntaje de riesgo supera un umbral predefinido, Flink envía una alerta con los detalles de la transacción a un segundo tópico de alertas en Kafka,
6. **Visualización y acción:** Un **consumidor de Kafka** lee el tópico de alertas y desencadena una acción, como notificar al equipo de seguridad, enviar una alerta automática a la aplicación móvil del cliente o incluso bloquear la transacción. Al mismo tiempo, los datos se envían a un **dashboard** para su visualización y análisis.
7. **Finalización:** La información se clasifica y se almacena para futuros análisis y entrenamiento de nuevos modelos de machine learning.

**M8 AE 4: Plataforma de automatización de Flujo de datos:**

**Contexto: Una empresa con operaciones online y físicas recibe diariamente grandes volúmenes de información desde diferentes canales de venta: e-commerce, puntos físicos, aplicaciones móviles y marketplaces. Actualmente, el equipo de datos integra manualmente los reportes, lo que genera cuellos de botella y errores humanos. Se requiere automatizar este flujo para que los datos lleguen estandarizados y en tiempo real al data warehouse de la compañía.**

**Consigna: Diseña un flujo de datos para procesar altos volúmenes de información de ventas multicanal utilizando una plataforma de automatización. Deberás contemplar las fuentes de entrada, etapas de procesamiento, reglas de decisión y destino de los datos, aplicando buenas prácticas.**

**Paso a paso: Define los componentes del flujo.**

**Fuente de datos: ¿de dónde proviene la información?**

La información proviene de diversas fuentes: e-commerce, puntos físicos, aplicaciones móviles y Marketplace.

**Procesador: ¿qué transformaciones deben realizarse?**

Se debe limpiar la información, eliminando duplicados, imputar en los datos en blanco, etc. Además, se puede realizar cálculos que sean necesarios. Se debe automatizar los procesos de carga y transformación, evitando pérdidas de tiempo y ejecutando los procesos en períodos de baja carga de procesamiento.

**Enrutador o decisor: ¿se aplican reglas para enviar los datos a distintos destinos?**

Se pueden definir cargas de datos temporales que permitan visualizar datos en Dashboard o APP para consultas o decisiones inmediatas. Además, se debe almacenar la información de salida para análisis, consultas y entrenamiento de Machine Learning. También se puede segmentar la información, enviando de manera directa resúmenes o cargando Dashboards para algún área específica.

**Destino: ¿dónde deben almacenarse los datos? (data lake, DWH, API, etc.)**

Se almacenarán los datos en un data warehouse, con acceso de toda la organización.

**Selecciona o menciona una plataforma de automatización posible:**

Se sugiere el uso de Apache NiFi, con ingesta de datos en batch. Apache NIFI considera interfaz visual y baja complejidad. Permite procesamiento en tiempo real y escalabilidad horizontal.

**Detalla los beneficios esperados:** Se espera la reducción de tiempo en los procesos, reducción de errores y disponibilidad de la información en tiempo real, permitiendo el monitoreo de errores o detección de cuellos de botella mediante el seguimiento de logs y triggers de alertas.

**Redacta tu propuesta de flujo paso a paso:**

**Ingesta:** Se realiza la carga de información de forma automatizada, desde las diversas fuentes:

* **E-commerce:** Se realiza lectura desde la base de datos de la página web, considerando períodos determinados de tiempo, ejemplo cada 8 horas.
* **Puntos físicos:** Se realiza lectura desde los puntos de venta, consolidando la información por determinados períodos de tiempo, por ejemplo, posterior a la lectura de e-commerce.
* **Aplicaciones móviles:** Se realiza lectura desde la información almacenada de las aplicaciones móviles
* **Marketplace:** Se consolida la información de los marketplace y se carga de manera coordinada con los procesos anteriores.

**Transformación:** Se normalizan los datos, buscando estandarizar la información de los diferentes orígenes y se eliminan datos en blanco se corrigen valores nulos y se dispone de la data para la validación.

**Validación:** Se validan los datos, determinando que cumpla con las reglas esperadas,como rangos de valores, datos correctos, etc. Si se encuentran incongruencias se debe correr procesos tendientes a la corrección o revisión de los datos, para lo cual, previamente se deben haber considerado las acciones, como alertas, informes, registros en log, etc.

**Enriquecimiento:** Se mejora la data con procesos de transformación, cálculos o nueva data que pueda servir a la toma de decisiones.

**Entrega:** Se entrega la información como reportes, dashboards, APP o bases de consulta, que permitan su análisis y toma de decisiones de acuerdo con la estrategia del negocio.

**Auditoría y Monitoreo:** Se registran logs, métricas y eventos, se envían informes de alertas que permitan controlar, mantener y mejorar la calidad de los datos, la adaptación de los procesos y su funcionamiento constante.

**Orquestación Automatizada para consolidación masiva de registros clínicos:**

**Contexto: Una red de clínicas privadas necesita consolidar, diariamente, grandes volúmenes de información médica proveniente de diferentes sedes y sistemas. Los registros incluyen datos de pacientes, estudios realizados, recetas y turnos. Actualmente, la carga se realiza de forma manual y asíncrona, generando errores, duplicaciones y demoras críticas. El área de datos debe diseñar un flujo automatizado para orquestar la recolección, transformación y carga de datos, utilizando una plataforma que permita gestionar esta complejidad de manera eficiente y escalable.**

**Consigna: Diseña un flujo automatizado para orquestar el procesamiento de grandes volúmenes de datos médicos, identificando la plataforma más adecuada (Airflow, NiFi o Talend) según el caso. Detalla las etapas del flujo, sus características técnicas, los beneficios esperados y las buenas prácticas que aplicarías.**

**Paso a paso:**

**Define la necesidad principal del flujo:** Se requiere de la implementación de un proceso de automatización para la consolidación de la información proveniente de las diferentes fuentes. Se debe automatizar también la transformación de los datos, optimizando los tiempos de procesamiento, disminuyendo los costos y mejorando la calidad de los datos.

**¿Qué tipos de datos serán procesados y qué objetivo persigue la automatización?**

Se requiere procesar información de diferentes fuentes, con información variada (no indica si es información estructurada o no). Con la automatización, se busca coordinar los tiempos de recolección de datos y reducir los tiempos de ejecución, además se busca reducir la posibilidad de errores generadas por los usuarios y transformar la información de acuerdo con los parámetros establecidos.

**Selecciona la plataforma adecuada entre NiFi, Airflow o Talend:**

Se sugiere el uso de Apache NiFi, pues es una herramienta que permite cargas y la entrega de datos de diferentes fuentes y destinos, respectivamente. Además, la documentación indica que es un muy buen orquestador de las herramientas necesarias.

**Justifica tu elección considerando el tipo de procesamiento (batch/streaming), el perfil técnico del equipo y la complejidad del flujo.**

Se realizará ingesta de datos en Batch, se requiere de personal técnico capacitado en la implementación y mantenimiento de Apache NiFi, se debe considerar las posibles fallas para segmentar la información con errores y corregir cuando corresponda. También se debe considerar la posibilidad de ampliar los procesos de recopilación o entrega de información.

**Describe el flujo paso a paso en función de estas tres etapas clave:**

**Ingesta: ¿cómo y desde dónde se capturan los datos?**

Los datos provienen de diversas sedes y sistemas, por lo que se deben rescatar desde las diferentes fuentes con agendamiento diario que permita consolidar la información mediante procesos batch. Se debe considerar el monitoreo de los procesos de captura de datos, con la finalidad de activar alarmas o acciones ante posibles errores, se debe monitorear los metadatos.

**Procesamiento: ¿qué transformaciones, validaciones o reglas se aplican?**

Se debe validar la información, asegurándose de contar con toda la información necesaria y consolidar la información para la transformación, donde se deben normalizar los datos, eliminar datos duplicados, con errores, fuera de rangos, etc. Las reglas de transformación deben ser evaluadas y definidas previamente para aplicarlas en este proceso, definiendo, por ejemplo, cuando se eliminan datos registros con datos en blanco, cuando se imputan datos, como media, mediana, moda, etc.

**Entrega: ¿dónde se almacenan los datos procesados? (data lake, base de datos, APIs, etc.)**

Posterior a la transformación, los datos se almacenan en un data warehouse, donde los sistemas tendrán acceso a datos normalizados e históricos. Además, se pueden generar vistas o dashboards para mejorar la toma de decisiones.

**Agrega al menos 3 buenas prácticas que aplicarías.**

Se debe considerar:

* Configurar alertas para evitar cuellos de botella y mover a una base datos temporal los posibles datos erróneos.
* Validar datos y flujos antes de ponerlos en producción.
* Documentar cada proceso y cambio para posteriores consultas y definira los responsables.
* Monitorear el sistema en operación para actualizarlo o detectar posibilidades de mejora.