Gestión de Datos

Gestión y Arquitectura de Datos, Universidad de San Andrés

Si encuentran algún error en el documento o hay alguna duda, mandenmé un mail a rodriguezf@udesa.edu.ar y lo revisamos.

1. Introducción

La gestión de datos es el proceso de recolectar, almacenar, organizar, proteger, recuperar y mantener datos dentro de una organización. Es fundamental para garantizar que los datos sean precisos, accesibles y utilizables para la toma de decisiones.

2. Ciclo de Vida de los Datos

2.1. Creación y Captura

La creación y captura de datos es el proceso de obtener datos de fuentes internas o externas y almacenarlos en el sistema de información. Estos pueden venir por distintos medios:

- Ingreso manual de datos
- Captura automática
- Integración de sistemas
- Sensores y dispositivos IoT

2.2. Almacenamiento

El almacenamiento de datos es la forma en que se guardan los datos para su posterior uso. Puede ser guardados en distintos contenedores, todos ellos los veremos más adelante en la materia:

- Bases de datos relacionales
- Bases de datos NoSQL

- Data Lakes
- Sistemas de archivos distribuidos

2.3. Procesamiento

El procesamiento de datos es el proceso de transformar los datos para que sean más útiles. Puede ser procesados de distintas maneras:

- ETL (Extract, Transform, Load)
- Limpieza de datos
- Validación
- Enriquecimiento

2.4. Análisis

El análisis de los datos es importantísimo para poder tomar decisiones informadas. Esto se puede hacer de disintas maneras y formas, pero podemos mencionar algunas:

- Business Intelligence
- Analytics
- Machine Learning
- Visualización

2.5. Archivado y Eliminación

El archivado y eliminación de datos es el proceso de guardar los datos para su posterior uso o eliminarlos cuando ya no son necesarios. Esto se puede hacer de distintas maneras:

- Políticas de retención: La empresa debe tener políticas claras sobre cuánto tiempo se deben mantener los datos.
- Cumplimiento regulatorio: La empresa debe cumplir con las leyes y regulaciones aplicables.

- Backup y recuperación: La empresa debe tener un plan de backup y recuperación de datos en caso de un desastre.
- Eliminación segura: La empresa debe tener un proceso de eliminación de datos que sea seguro y eficiente.

3. Calidad de Datos

3.1. Dimensiones de Calidad

Las dimensiones de calidad de los datos son las características que deben tener los datos para ser considerados de calidad. Estas dimensiones son:

- Precisión: La precisión de los datos es la medida de qué tan cercanos son los datos a la realidad.
- Completitud: La completitud de los datos es la medida de qué tan completos son los datos.
- Consistencia: La consistencia de los datos es la medida de qué tan coherentes son los datos.
- Actualidad: La actualidad de los datos es la medida de qué tan recientes son los datos.
- Unicidad: La unicidad de los datos es la medida de qué tan únicos son los datos.
- Accesibilidad: La accesibilidad de los datos es la medida de qué tan fácil es acceder a los datos.
- Seguridad: La seguridad de los datos es la medida de qué tan seguros son los datos.

3.2. Métricas de Calidad

Las métricas de calidad son las medidas que se utilizan para evaluar la calidad de los datos. Estas métricas son:

- Tasa de error: La tasa de error es la medida de qué tan frecuentes son los errores en los datos.

- Tasa de duplicación: La tasa de duplicación es la medida de qué tan frecuentes son los duplicados en los datos.
- Tasa de actualización: La tasa de actualización es la medida de qué tan frecuentes son las actualizaciones en los datos.

4. Gobierno de Datos

El gobierno de datos es el proceso de gestionar los datos dentro de una organización. Existen distintos roles y responsabilidades dentro de lo que es el gobierno de datos, cada uno con sus responsabilidades y objetivos.

4.1. Roles y Responsabilidades

- Chief Data Officer (CDO): Es el encargado de la dirección y gestión de los datos dentro de la organización.
- Data Owner: Responsable de un conjunto de datos, generalmente un gerente de área.
- Data Steward: Es el responsable operativo de los datos, manteniendolos actualizados y consistentes.

4.2. Políticas y Procedimientos

Las políticas y procedimientos son los documentos que regulan el uso y la gestión de los datos dentro de la organización. Estos documentos son:

- Políticas de seguridad
- Políticas de privacidad
- Políticas de retención
- Políticas de uso
- Políticas de acceso

5. Arquitectura de Datos

La arquitectura de datos moderna se compone de cuatro capas principales que trabajan en conjunto para manejar el ciclo de vida completo de los datos.

5.1. Fuentes de Datos

La primera capa son las **fuentes de Datos**, que incluyen sistemas operacionales como ERPs y CRMs, APIs externas que nos permiten obtener datos de terceros, y fuentes de datos no estructurados como documentos, imágenes y videos.

5.2. Almacenamiento

Para el **almacenamiento**, las organizaciones utilizan diferentes tipos de repositorios según sus necesidades. El Data Warehouse es ideal para datos estructurados y análisis histórico, mientras que el Data Lake permite almacenar datos en su formato original. Los Data Marts son subconjuntos especializados del Data Warehouse para áreas específicas del negocio.

5.3. Procesamiento

El **procesamiento** de datos puede realizarse de diferentes maneras. El procesamiento por lotes (Batch) es adecuado para grandes volúmenes de datos que no requieren tiempo real. El procesamiento en streaming es ideal para datos que necesitan ser procesados inmediatamente. La arquitectura Lambda combina ambos enfoques para ofrecer flexibilidad.

5.4. Consumo

Finalmente, la capa de **consumo** permite que los datos sean accesibles para los usuarios finales a través de reportes estructurados, dashboards interactivos y APIs que permiten la integración con otros sistemas.