BIGDATA

El término "big data" se refiere a los datos que son tan grandes, rápidos o complejos que es difícil o imposible procesarlos con los métodos tradicionales.

TERMINOLOGÍA

DATOS ESTRUCTURADOS
Datos cuantitativos, muy organizados que encajan perfectamente en campos y columnas fijos en bases de datos relacionales y hojas de cálculo.

DATOS NO ESTRUCTURADOS
Datos cualitativos, que no pueden procesarse y analizarse utilizando herramientas y métodos convencionales porque no tienen un modelo predefinido, lo que significa que no se pueden organizar en bases de datos relacionales. En cambio, las bases de datos no relacionales o NoSQL son las más adecuadas para administrar datos no estructurados.

DATA WAREHOUSE
Sistema que agrega y combina información de diferentes fuentes en un almacén de datos único y centralizado; consistente para respaldar el análisis empresarial, la minería de datos, inteligencia artificial (IA) y Machine Learning.

MACHINE LEARNING
Es una disciplina del campo de la Inteligencia Artificial que, a través de algoritmos, dota a los ordenadores de la capacidad de identificar patrones en datos masivos y elaborar predicciones (análisis predictivo).

CLOUD COMPUTING
Es la entrega de diferentes servicios a través de Internet. Estos recursos incluyen herramientas y aplicaciones como almacenamiento de datos, servidores, bases de datos, redes y software. Se le denomina así porque la información a la que se accede se encuentra de forma remota en un espacio virtual.

LAS 5V'S

VOLUMEN
El big data implica un volumen enorme de datos. Los datos se presentan en todo tipo de formatos: desde datos numéricos estructurados en bases de datos tradicionales hasta documentos de texto no estructurados, correos electrónicos, vídeos, audios, datos de teletipo y transacciones financieras.

VARIABILIDAD
Los flujos de datos son impredecibles, cambian a menudo y varían mucho. Es un reto, pero las empresas necesitan saber cuándo algo está de moda en los medios sociales, y cómo gestionar los picos de carga de datos diarios, estacionales y desencadenados por eventos.

VELOCIDAD
Con el crecimiento del Internet de las Cosas, los datos llegan a las empresas a una velocidad sin precedentes y deben ser manejados de manera oportuna.

VERACIDAD
La calidad de los datos. Debido a que los datos provienen de tantas fuentes diferentes, es difícil vincular, comparar, limpiar y transformar los datos a través de los sistemas.

VALOR
El valor se obtiene de datos que se transforman en información; esta a su vez se convierte en conocimiento, y este en acción o en decisión.

CICLO DE VIDA DE LOS DATOS

- CAPTURA DE DATOS
- ALMACENAMIENTO
- PROCESAMIENTO Y ANALISIS
- EXPLORACIÓN Y VISUALIZACIÓN

BIGDATA ETL

EXTRACT
Los datos en bruto deben extraerse de una variedad de fuentes, por ejemplo:

- Bases de datos existentes
- Registros de actividad como el tráfico de red, informes de errores, etc.
- Rendimiento y anomalías de aplicaciones
- Incidencias de seguridad
- Otras actividades transaccionales que deben comunicarse para dar cumplimiento normativo

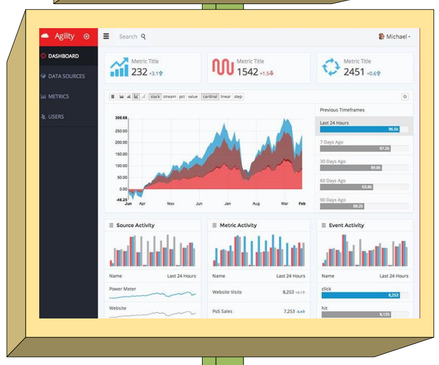
Los datos extraídos en ocasiones se transfieren a otro destino como por ejemplo un data lake o un almacén de datos.

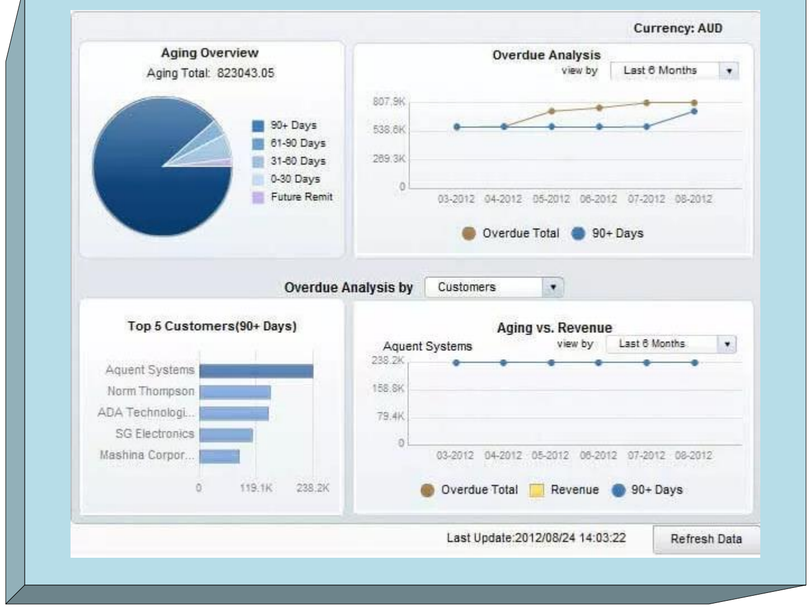
TRANSFORM
La transformación modifica los datos en bruto para que presenten los formatos de notificación correctos.
Normalización: definir qué datos entrarán en juego, cómo se formatearán y almacenarán, y otras consideraciones básicas que definirán las etapas sucesivas.
Eliminación de duplicados
Verificación: permiten seguir cribando los datos no utilizables y pueden alertar sobre anomalías en sus sistemas, aplicaciones o datos.
Clasificación: maximizar la eficiencia de los almacenes de datos agrupando y clasificando elementos como los datos en bruto, audios, archivos multimedia y otros objetos en categorías

LOAD
consiste en almacenar los datos ya transformados en un sistema destino del que se puedan nutrir todas las áreas de la organización. Estos sistemas de almacenamiento reciben el nombre de Data Warehouse y son el origen de datos para distintas herramientas de analítica descriptiva, diagnóstica, predictiva y prescriptiva.

DASHBOARDS

Los dashboards se nutren de datos, mostrándolos visualmente de una manera u otra para una posterior interpretación. La idea principal se basa en extraer conclusiones e información de valor para la empresa, y aquí es donde confluyen los dashboards con el big data.





SAP HANA

SAP HANA es la implementación de SAP SE de la tecnología de base de datos en memoria. Hay cuatro componentes dentro del grupo de software:

- **SAP HANA DB (o HANA DB)** se refiere a la tecnología de base de datos en sí.
- **SAP HANA Studio** se refiere al conjunto de herramientas que proporciona SAP para modelar.
- **SAP HANA Appliance** se refiere a HANA DB como socio de Hardware presentadas en el certificado (véase más adelante) como un dispositivo. También incluye las herramientas de modelado de HANA Studio, así como herramientas de replicación y transformación de datos para mover datos a HANA DB
- **SAP HANA Aplicación en nube** se refiere a la infraestructura basada en la Computación en la nube para la entrega de aplicaciones (típicamente las aplicaciones existentes de SAP reescritas para ejecutarse en HANA).

ZUBIRI VALDEZ HEDSON LEONARDO 7°G