Curso de Especialización de
Inteligencia Artificial y Big Data (IABD)

Big Data Aplicado

UD05. Aplicación práctica de tecnologías Big Data.
Resumen.

JUAN ANTONIO GARCIA MUELAS

Las tecnologías Big Data se empezaron a probar porque los equipos de tecnología vieron la capacidad de resolver casos de uso para el negocio.

El Datawarehouse tiene un modelo de almacenamiento de los datos columnar y suele resolver a consultas analíticas.

Una de las principales ventajas de las tecnologías Big Data en lugar de la tradicionales de Datawarehouse es que permite tomar todos los datos para hacer los análisis.

Las herramientas ETL preparan los datos para el Datawarehouse.

Las tecnologías de Datawarehouse tienen un coste superior al de las tecnologías Big Data.

Data Mining es la disciplina que tradicionalmente ha creado modelos predictivos sobre los datos del Datawarehouse.

Un plan de capacidad requiere conocer qué necesidades de datos tendrá una plataforma.

Una vez instalado un Hadoop on-premise, es necesario hacer optimizarlo o tunearlo para obtener el máximo rendimiento.

Un **Data Lake** es un **repositorio centralizado** que permite **almacenar** todos los datos de una empresa a cualquier escala, **sin modificarlos y sin** tener que **estructurarlos** primero, y también **permite ejecutar** diferentes tipos de **análisis**: desde cuadros de mando y reporting hasta análisis en tiempo real y machine learning.

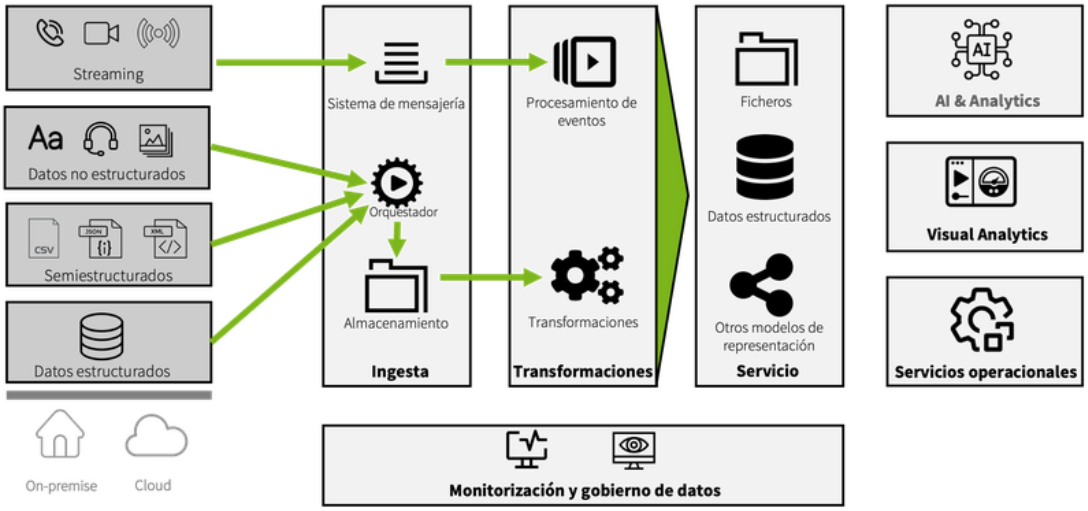
El **objetivo** de un Data Lake es **construir una plataforma de datos** que permita recopilar todos los datos disponible, tanto externos como internos, estructurados y no estructurados, para **gestion**arlos de forma **centralizada** y mejorar los procesos de toma de decisiones.

Un Data Lake **ofrece más funcionalidades que un Datawarehouse**, pero su **gestión** es **más compleja**, al tener mayor variedad de tecnologías, datos y usuarios.

Los **datos**, ya sean estructurados o no estructurados, **se guardan tal cual se reciben** y se les denomina **Raw Data**.

MODELO DE CAPAS DE UN DATA LAKE.

Arquitectura del Data Lake



La **ingesta** puede ser de dos tipos:

Para la ingesta de **datos en streaming,** se utiliza un sistema de mensajería como **Apache Kafka**, para **independizar** la **generación** de los **datos con** su **consumo**, **o** un sistema como Apache **Flume**, para su **volcado directo a HDFS**.

Los **datos en reposo** suelen ingestarse utilizando un **orquestador que coordina y ejecuta planificadamente procesos** de ingesta. El **orquestador** puede ser **Apache Oozie**, mientras que los **procesos** de ingesta pueden estar **implementados con Sqoop** (para ingesta de datos de **bases de datos relacionales**), **o** simplemente **tareas** ejecutadas como script que recogen ficheros de sistemas **SFTP**.

El **procesamiento** de los **datos raw** suele realizarse mediante herramientas de procesamiento masivo de datos como Apache **Spark o** Apache **Hive**.

El modelo de **arquitectura** del Data Lake es, un **modelo centralizado** con estas **ventajas**:

- ✓ La **especialización** de los **perfiles** de **ingeniería de datos**, que permite tener una mayor productividad en las tareas de ingesta, transformación y ofrecimiento de los datos.
- ✓ La **industrialización** de los **procesos** de **ingeniería de datos**, que ofrece una mayor homogeneidad en las todas labores sobre los datos. **Pretende estandarizar y automatizar todo lo posible las tareas para ganar eficiencia.**
- ✓ La **economía de escala**, tanto en infraestructura como en equipos.

Sin embargo, tiene como principales **problemas**:

- ✓ La **generación de dependencias** sobre un único equipo de ingeniería de datos.
- ✓ La **dificultad para escalar** el modelo de operaciones sobre los datos.

Por estos dos principales motivos, surgió a principios de esta década el concepto de Data Mesh.

Uno de los principales problemas de Data Lake frente a data Mesh es que es más difícil de escalar.

Una **arquitectura Data Mesh** es un **enfoque descentralizado** en el que cada dominio es responsable de preparar los datos y ofrecer al resto de equipos estos datos como producto para que puedan ser utilizados de la forma más sencilla posible.

Data Mesh es principalmente un **enfoque organizacional.** Sigue cuatro **principios básicos**:

- ✓ **Propiedad y arquitectura descentralizada de datos orientada al dominio**: el principio de **propiedad del dominio exige** que los **equipo**s de dominio **asuma**n la **responsabilidad** de sus datos.
- ✓ **Datos como producto**: el principio de datos como producto proyecta una filosofía de **pensamiento de producto sobre datos** analíticos. Este principio significa que hay consumidores para los datos más allá del dominio.
- ✓ **Infraestructura de datos de autoservicio como plataforma**: la plataforma **debe simplificarse** para que los dominios puedan **publicar o consumir** los datos de otros dominios **de** una **forma sencilla**.
- ✓ **Gobierno computacional federado**: Un **único equipo de Gobierno de datos**, esta disciplina es compartida por todos los dominios.

Todos estos principios, seguidos **sin control pueden generar** un problema que se conoce como **Data Swamp**. Para evitarlo, se cuenta con equipos de **Gobierno de Datos o Data Governance** (**conjunto de procesos, roles, políticas, estándares y métricas que garantizan el uso eficiente y efectivo de los datos, alineado con los objetivos de las empresas.).**

Uno de los principales beneficios de Data Lake frente a Data Mesh es permite homogeneizar más las actividades de preparación de datos.

En un entorno multitenancy es importante definir una estructura de directorios clara.

Amazon EMR, son las siglas de Amazon Elastic MapReduce, es un servicio de Amazon Web Services que **permite crear clusters Hadoop a demanda**.

HDInsight es la solución de **Microsoft Azure de Hadoop como servicio**, que permite **arrancar clústers Hadoop a demanda en** una modalidad de **pago por uso**. Un servicio muy parecido a Amazon EMR, aunque tiene algunas **diferencias**, entre las que destacan:

- ✓ HDInsight **utiliza la distribución HDP de Hortonworks** en lugar de una propia en el caso de EMR.
- ✓ HDInsight **proporciona Ambari** como herramienta de administración y monitorización, mientras que **EMR proporciona una herramienta propia**, mucho menos potente, **y Ganglia** para monitorización.