Evolución de la Ingeniería de Datos

Juamini 3 Pro Preview

November 23, 2025

Esta es la historia central del **Capítulo 1** de *Fundamentals of Data Engineering*. Reis y Housley describen la evolución de la ingeniería de datos como un movimiento pendular: de la simplicidad rígida a la complejidad caótica, y finalmente hacia una simplicidad modular.

Aquí tienes la explicación de cada etapa según los autores:

1. Data Warehouse Monolítico (1980s - 2000s)

La era de "Todo en una caja".

- **Contexto:** Las empresas necesitaban reportes (Business Intelligence).
- **Tecnología:** Bases de datos relacionales gigantes y costosas (Oracle, IBM DB2, Teradata) instaladas en servidores físicos propios (*on-premise*).
- **Características:**
  - **Monolítico:** El almacenamiento y el procesamiento estaban atados en una sola máquina o appliance.
  - **Rígido:** Todo debía ser SQL y datos estructurados.
  - **Etiqueta del Rol:** "Ingeniero de BI" o "Desarrollador ETL".
- **El Problema:** Cuando llegó internet (Yahoo, Google, Amazon), el volumen de datos explotó. Estos sistemas monolíticos eran demasiado caros y no podían escalar lo suficiente.

2. Big Data Complejo (2000s - 2015)

La era de "Ingeniería Extrema".

- **El "Big Bang":** Google publica los papers de *MapReduce* y *GFS*, y nace **Hadoop**.
- **Filosofía:** Usar muchas computadoras baratas (*commodity hardware*) en lugar de una supercomputadora cara.
- **Tecnología:** El ecosistema Hadoop (HDFS, MapReduce, Hive, Pig, ZooKeeper).
- **Características:**
  - **Complejidad:** Era extremadamente difícil de mantener. Requería ingenieros que supieran Java/Scala, redes, hardware y sistemas distribuidos de bajo nivel.
  - **El "Big Data Engineer":** Era un rol muy técnico, casi un científico de la computación. Se pasaban el día configurando clústers y arreglando fallos de nodos.
- **El Problema:** Se volvió una moda (*hype*). Las empresas usaban herramientas complejas para problemas simples. El mantenimiento era una pesadilla y se perdía el foco en el valor de negocio.

3. Modern Data Stack Simplificado (2020s - Presente)

La era del "Ciclo de Vida" y la Abstracción.

- **El Cambio:** La nube pública (AWS, Azure, GCP) y las herramientas SaaS maduraron.
- **Filosofía: Abstracción**. Dejar de preocuparse por "instalar" el servidor y enfocarse en usarlo.
- **Tecnología:** Herramientas modulares, "llave en mano" y *plug-and-play.*
  - Almacenamiento: Snowflake, BigQuery, Databricks.
  - Ingesta: Fivetran, Airbyte.
  - Transformación: dbt.

- **Características:**
  - **Descentralización:** Pasamos de monolitos a piezas de LEGO conectables.
  - **Facilidad:** Ya no necesitas saber gestionar un cluster de Hadoop. Puedes empezar con una tarjeta de crédito y SQL.
  - **El Nuevo Ingeniero de Datos:** Ya no es un "cuidador de hardware". Es un **Ingeniero del Ciclo de Vida** que conecta herramientas, gestiona costos, seguridad y DataOps para entregar valor rápido.

Resumen para el Examen

La evolución según Reis es un viaje desde la **rigidez** (Warehouse) pasando por la **complejidad técnica extrema** (Big Data/Hadoop) hasta llegar a la **simplicidad modular y abstracta** (Modern Data Stack) de hoy, donde el foco vuelve a estar en el valor del dato y no en la infraestructura.