The Future Of BI

Erik Perjons

DSV, Stockholm University

Concepts in Focus

- Data Warehouse
- Data Lake
- Lakehouse
- Data Mesh
- Analytics and Business Intelligence Platforms (previously called BI tools)

Data Warehouse Definition

- "A data warehouse is a **central repository of information** that can be analyzed to **make more informed decisions**. Data flows into a data warehouse from transactional systems, relational databases, and other sources, typically on a regular cadence. Business analysts, data engineers, data scientists, and decision makers **access the data through business intelligence (BI) tools**, SQL clients, and other analytics applications." (*https://aws.amazon.com/what-is/data-warehouse/*)

Data Lake

- "A data lake is a centralized repository **designed to store, process, and secure large amounts of structured, semistructured, and unstructured data**. It can store data in its native (natural) format and process any variety of it, ignoring size limits."

  (https://cloud.google.com/learn/what-is-a-data-lake)

Architectural Relationship Data Warehouse and Data Lake

- **In modern data architectures, data can be collected and stored in a data lake without being structured upfront**.
- This approach, allows organizations **to store large volumes of raw, semi-structured, or unstructured data** without needing to define how the data will be used in advance.
- **The structuring and transformation of the data typically occur at the point of use**, when the data is accessed for analysis, reporting, or machine learning purposes.

Palmer, M. (2024). Understanding ETL: Data Pipelines for Modern Data Architectures. O'Reilly Media. (can be downloaded from https://www.databricks.com/resources/ebook/understanding-etl)

Architectural Relationship Data Warehouse and Data Lake

- **Data from the data lake can be loaded into a data warehouse**. Unlike data lakes, **data warehouses require a predefined schema** and store **structured, high-quality data that is optimized for fast querying and business analytics**.
- This transformation from raw to structured data is typically handled through **ETL (Extract, Transform, Load)**

Palmer, M. (2024). *Understanding ETL: Data Pipelines for Modern Data Architectures*. O'Reilly Media. (can be downloaded from https://www.databricks.com/resources/ebook/understanding-etl)

Architectural Relationship Data Warehouse and Data Lake

- Once the **data is structured and stored in the warehouse**, it is **accessed and analyzed** through **Analytics and Business Intelligence (ABI) platforms** such as Microsoft Power BI, Tableau, or Looker.
- **Earlier versions of ABI platforms**, were often **called BI tools**.
- Note that it is sometimes possible to use ABI platforms directly towards the data lake – but of the not so optimal since the data is not indexed and aggregated, data might also be incomplete and inconsistent, and some ABI tools requires that data is stored in tables and that there exists meta data, which is missing in data lakes

Lakehouse

- A additional concept that related data warehouse and data lake is **lakehouse**:
- A lakehouse is an **architecture that combines the benefits of data lakes and data warehouses**, providing both the flexibility of data lakes and the management features of data warehouses.
- By **eliminating the need to maintain separate storage and computational systems**, lakehouses reduce complexity, cost, and data duplication.

(https://www.databricks.com/glossary/lakehouse)

Data Mesh

- **Data mesh is a decentralized data architecture** and assigns **ownership to domain-specific teams or departments**, but it **also provide flexible data sharing across an organization**.
- **Traditionally, data has been centralized** in a large data warehouse. **With data mesh, responsibility is distributed**: **each business domain (e.g., HR, sales, production) is accountable for its own data**—how it is stored, documented, and made accessible to others.
- Instead of relying on a centralized team, **a federated model is used**, where **each domain publishes its data as "data products"** that others can consume through a shared self-service platform

(Dehghani, Z. (2020). *Data Mesh: Delivering Data-Driven Value at Scale*. ThoughtWorks. See: https://martinfowler.com/articles/data-monolith-to-mesh.html)

Data Mesh

- A Data mesh architecture means that **each department manages its own data and use it own data structure**, but for **data that needs to be shared across departments for analysis or decision-making, the local structure must be mapped to a common overarching structure or standard.**
- This enables **data from different sources to become compatible, searchable, and usable in a broader organizational context**—without centralizing ownership or responsibility.

(Dehghani, Z. (2020). Data Mesh: Delivering Data-Driven Value at Scale. ThoughtWorks. See: https://martinfowler.com/articles/data-monolith-to-mesh.html)

Trends for Data Warehouse

- Cloud-Based Data Warehouses (DWaaS) that can offer separation of storage and computation
- Multicloud Support – can make use of use multiple cloud providers
- In-Database Analytics – analytics can be performed directly within the data warehouse, without moving data
- Updated in Real Time
- Security and Data Governance
- …

(Based on a number of sources)

Trends for ABI Platforms

- Augmented Analytics
- Large Language Models (LLMs)
- Cloud-Based ABI platforms
- Embedded Analytics in Business Applications
- Self-Service Analytics for Business Users
- ….

(Schlegel, K., et al. (2024) Gartner Group's report "Magic Quadrant for Analytics and Business Intelligence Platforms")