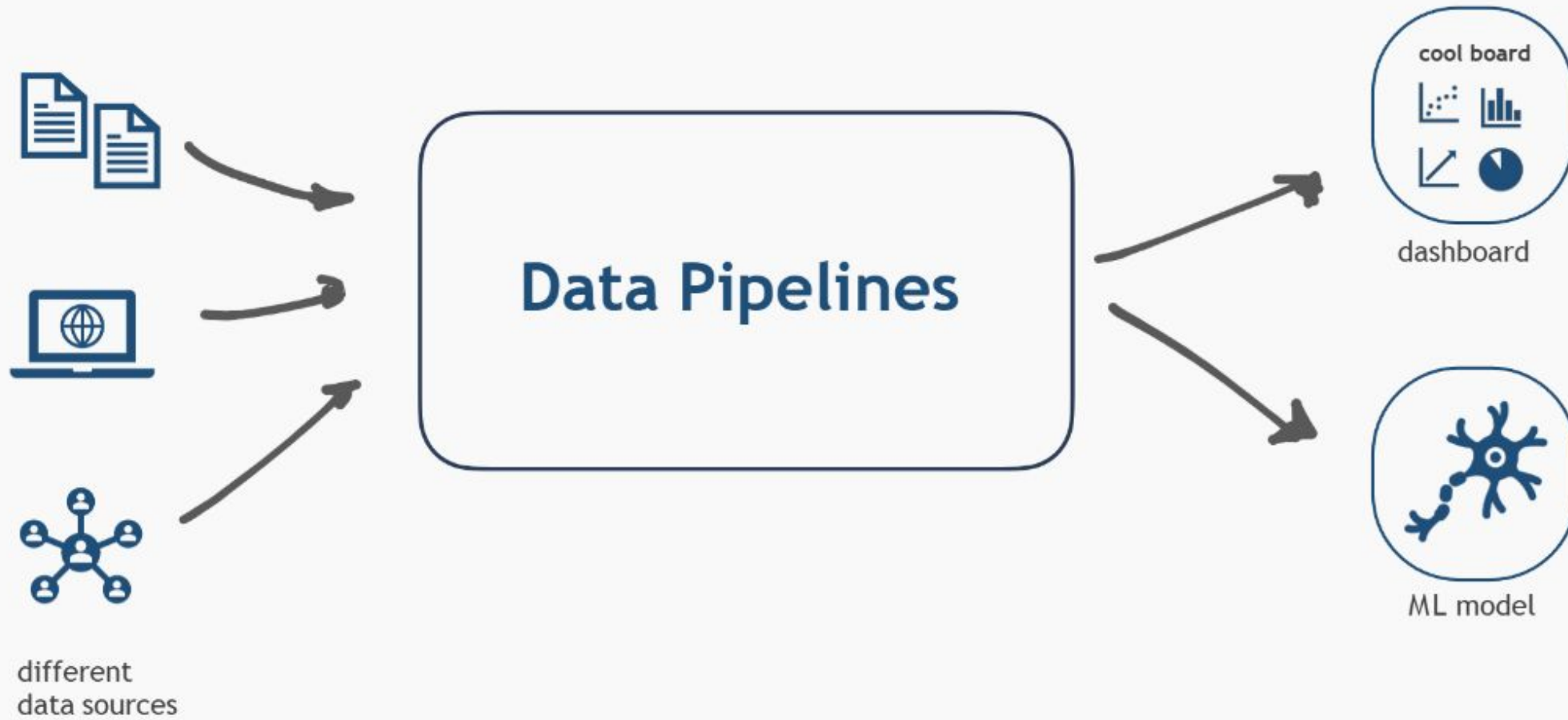


Esoon Ko
IT Höskolan

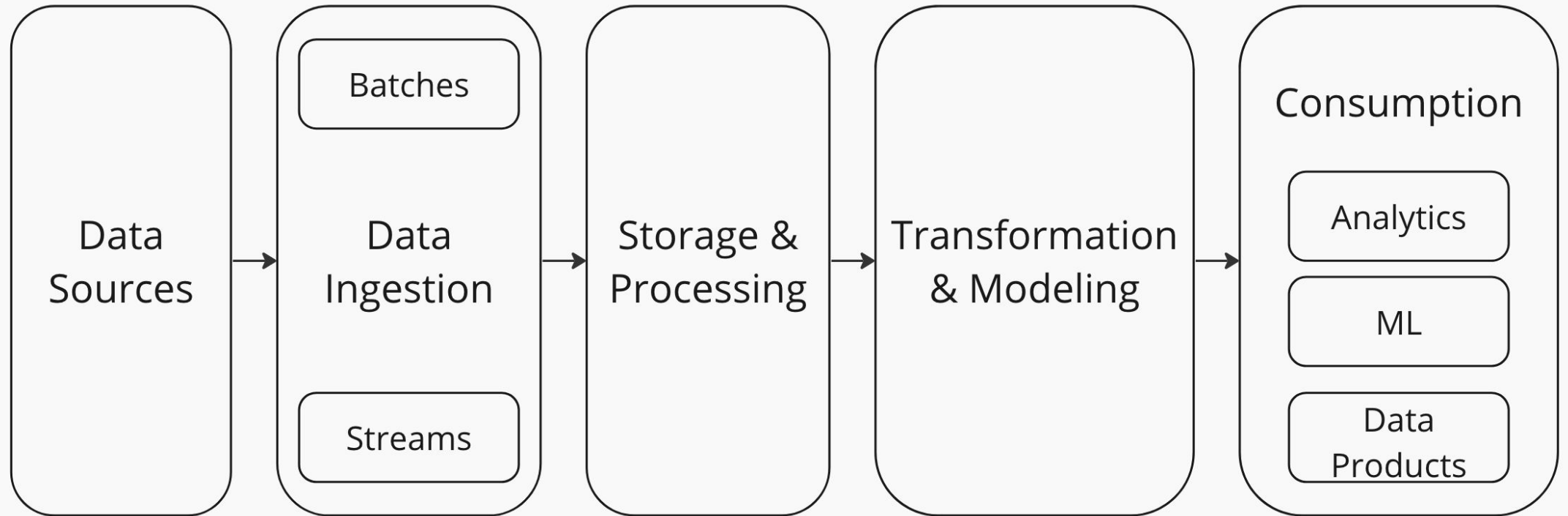
The need for a **data engineering**



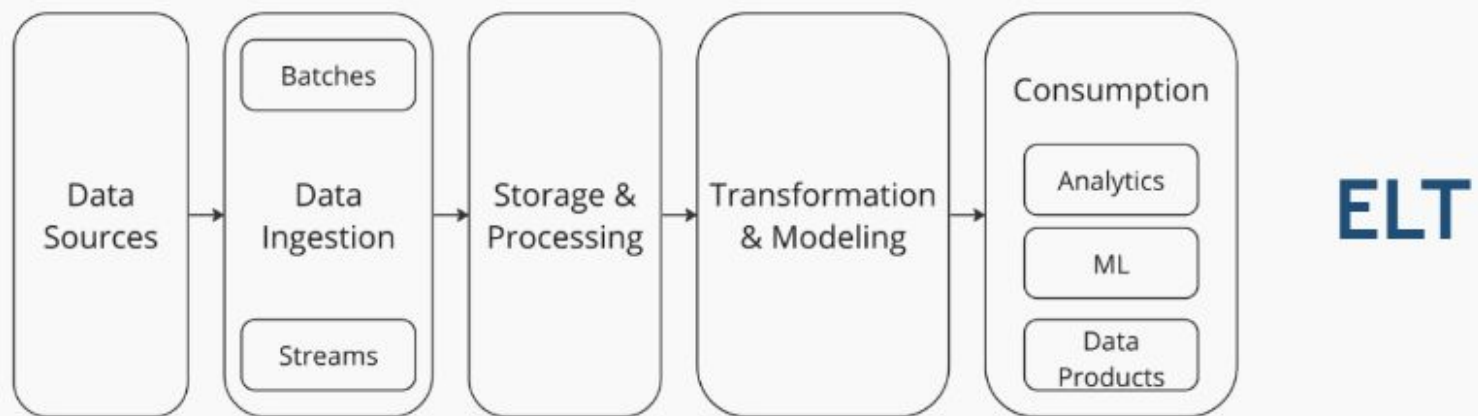
Data engineering



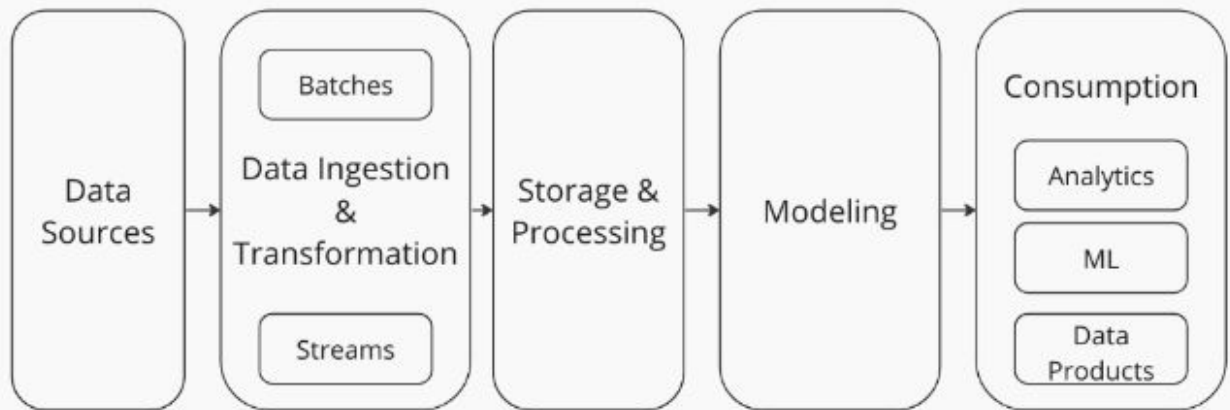
Modern Data Stack/**Data platform**

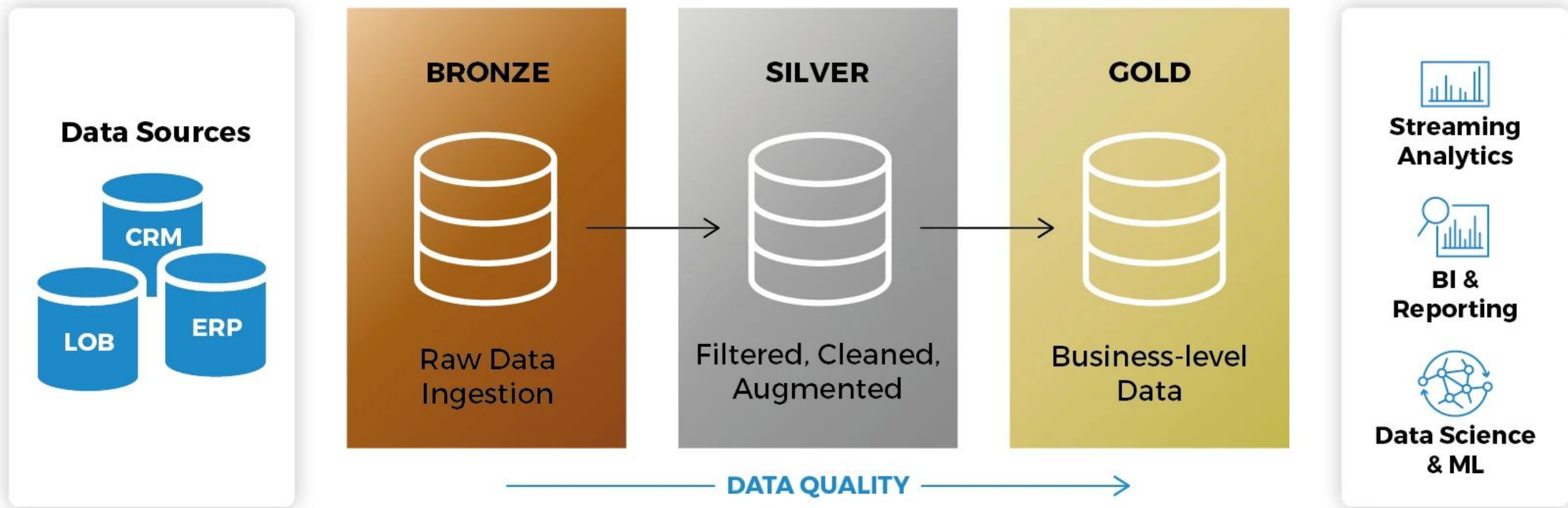


ETL vs ELT



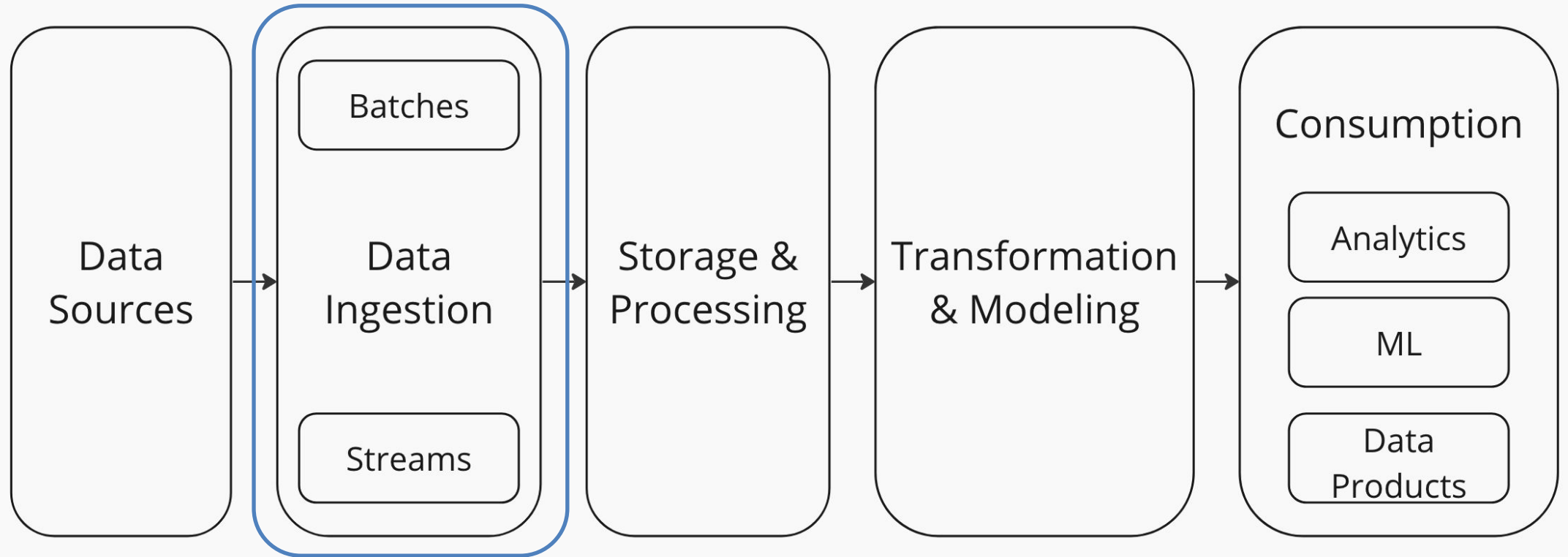
ETL





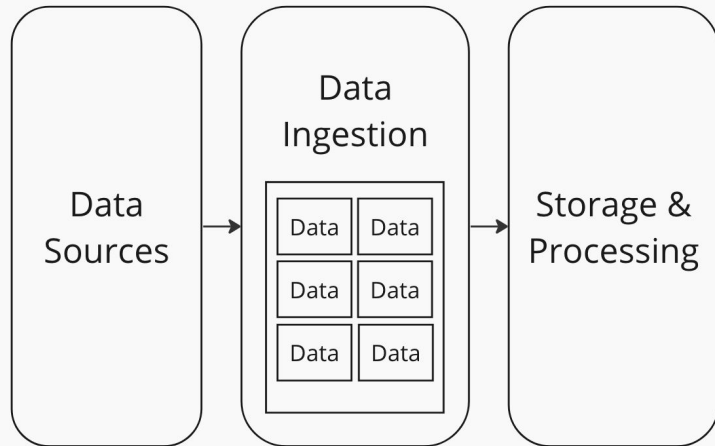
Data Ingestion

Data is ingested from source into storage.



Data Ingestion

Data is ingested from source into storage.

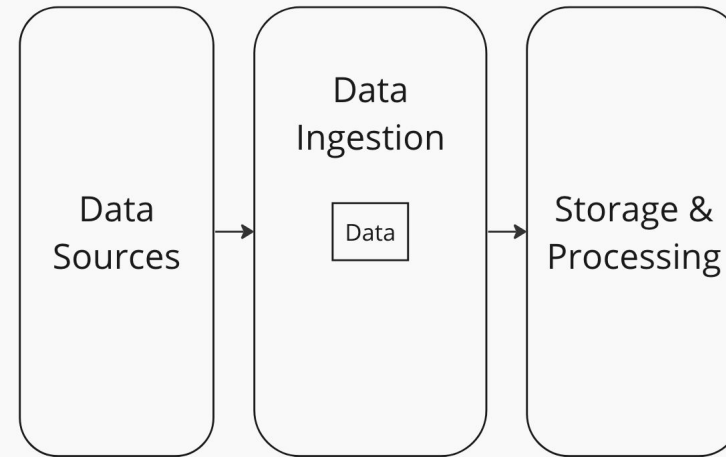


Batch

Data ingestion occurs from a source at a pre-defined time or in pre-defined groupings. Data is often pulled.

Benefits: Suitable for large amount of data that can be ingested in intervals.

Disadvantage: Does not provide data in real time.



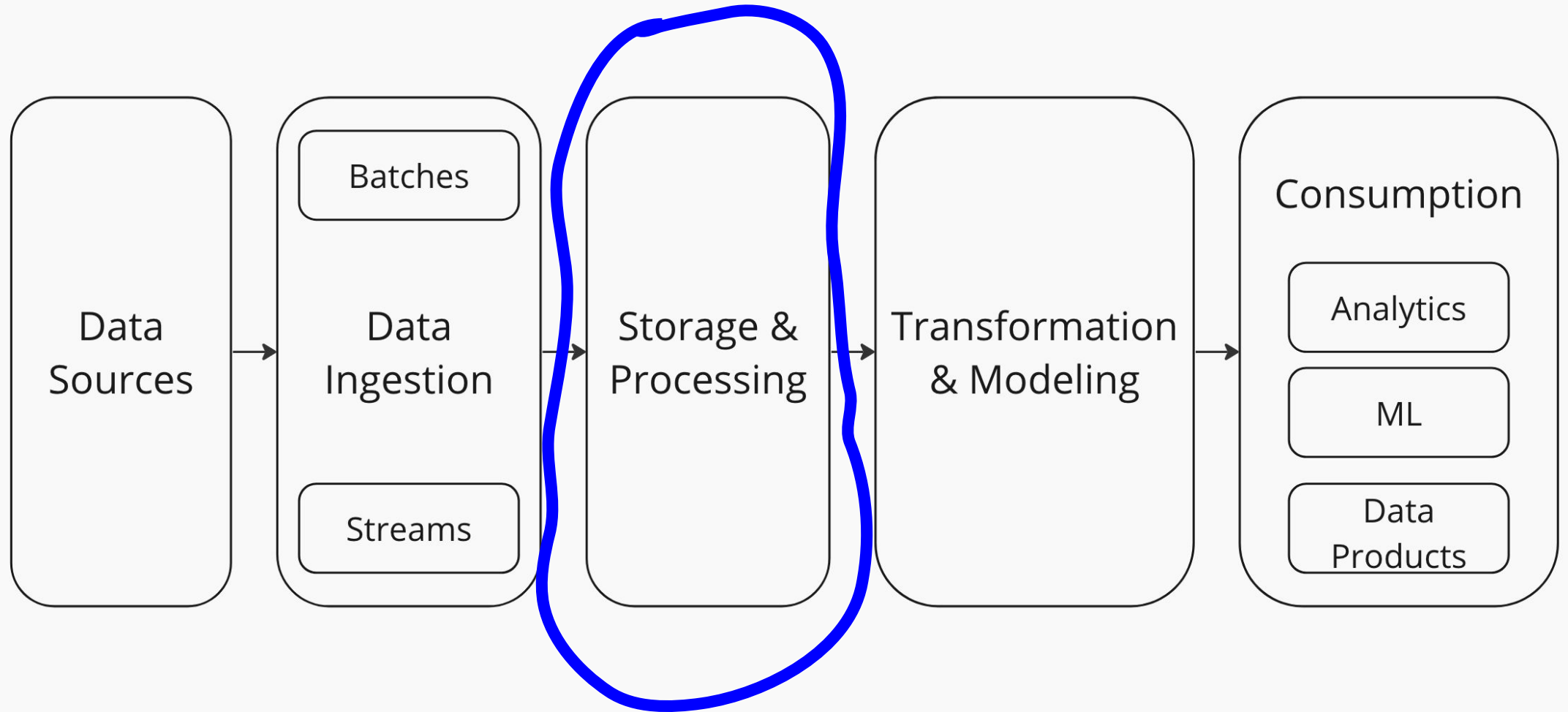
Stream

Data is ingested as soon as it is available in the source. Data is often pushed.

Benefits: Receive data in near real time.

Disadvantage: More difficult to process large amount of data that batch ingestion is capable of.

Storage and processing



Data Storage & Processing

Where data is stored and processed. We have many possibilities here...

On premise? Cloud?

Data warehouse? Data lake?



Data Warehouse

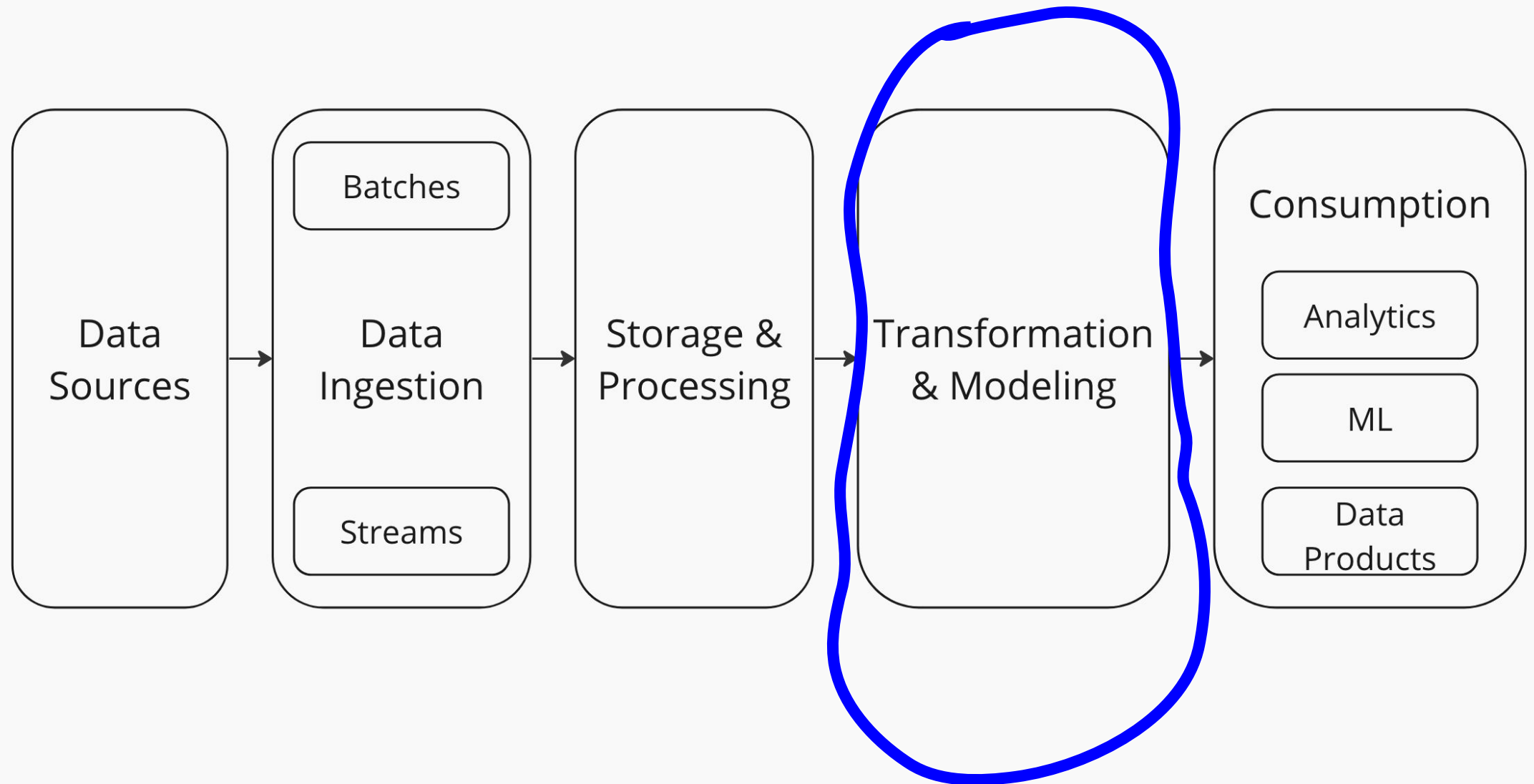
Data repository that provides data storage and compute, usually leveraging SQL queries for data analytics use cases.



Data Lake

...also a data repository that provides storage and compute. But is able to do it for structured and unstructured data. In most implementation data lakes are cloud storage that works similarly to your local file storage.

Storage and processing



Data Transformation

Data transformation usually means cleaning raw data and enriching it in order to make it consumable for analysis or reporting.

Common techniques used in data transformation include:

- Cleaning: This involves removing or correcting errors, inconsistencies, and missing values in the data.
- Normalization: Normalizing data involves scaling it to a common range to eliminate variations in scale that can affect analysis.
- Aggregation: Aggregating data involves summarizing detailed data into a more compact form, often by grouping it based on certain criteria and calculating summary statistics.
- Joining and Merging: Combining data from multiple sources by matching records based on common attributes.
- Derivation: Creating new variables or features from existing ones through calculations or transformations.
- Filtering: Selecting a subset of data based on specific criteria.

Data Modeling

Data modeling involves creating a conceptual representation of the structure and relationships within a dataset. It provides a blueprint for organizing and understanding data, enabling efficient storage, retrieval, and analysis. Data modeling helps ensure that data is organized efficiently, supports accurate analysis, and facilitates communication between stakeholders involved in data management and analysis processes.

Key components of data modeling include:

- Entity-Relationship (ER) Modeling: Identifying the entities (such as objects, people, or events) within a dataset and defining the relationships between them.
- Schema Design: Designing the structure of a database or dataset, including tables, fields, and constraints, based on the data model.
- Normalization: Ensuring that the database design follows normalization principles to minimize redundancy and dependency.
- Data Integrity: Enforcing rules and constraints to maintain the accuracy, consistency, and validity of data.
- Data Flow Diagrams: Visual representations of how data moves through a system or process, depicting inputs, outputs, and transformations.

SO - Basically continuation of the database course!

Data Modeling

Data modeling involves creating a conceptual representation of the structure and relationships within a dataset. It provides a blueprint for organizing and understanding data, enabling efficient storage, retrieval, and analysis. Data modeling helps ensure that data is organized efficiently, supports accurate analysis, and facilitates communication between stakeholders involved in data management and analysis processes.

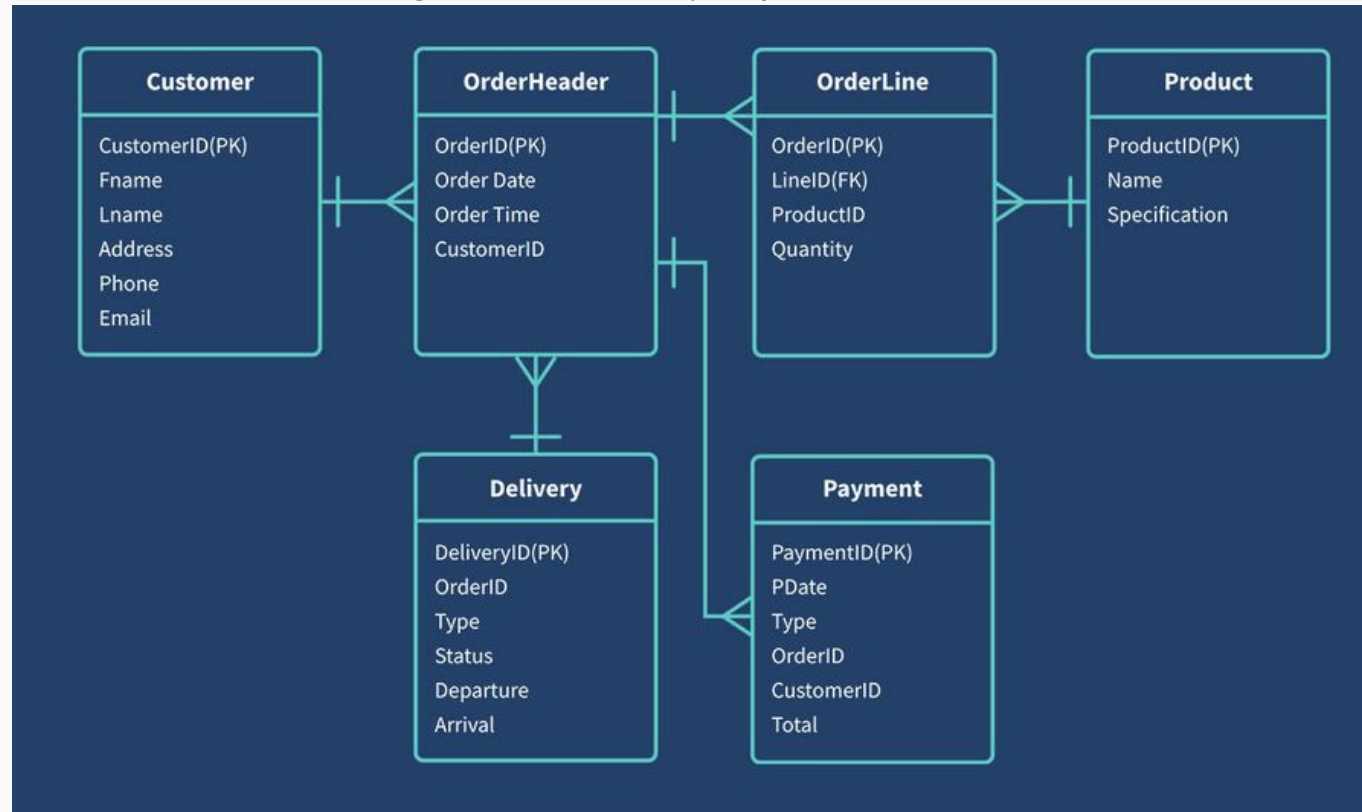


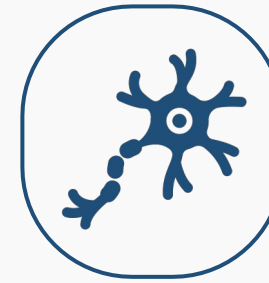
Image taken from “What is data modeling” - Qlik

Data Consumption Layer

The data consumption layer, also known as the presentation layer or the access layer, is a crucial component is usually the final layer. Its primary function is to facilitate the access, visualization, and interpretation of data by end-users. Essentially, it serves as the interface between the underlying data sources and the users who need to interact with the data for decision-making, analysis, or reporting purposes.



dashboard



ML model