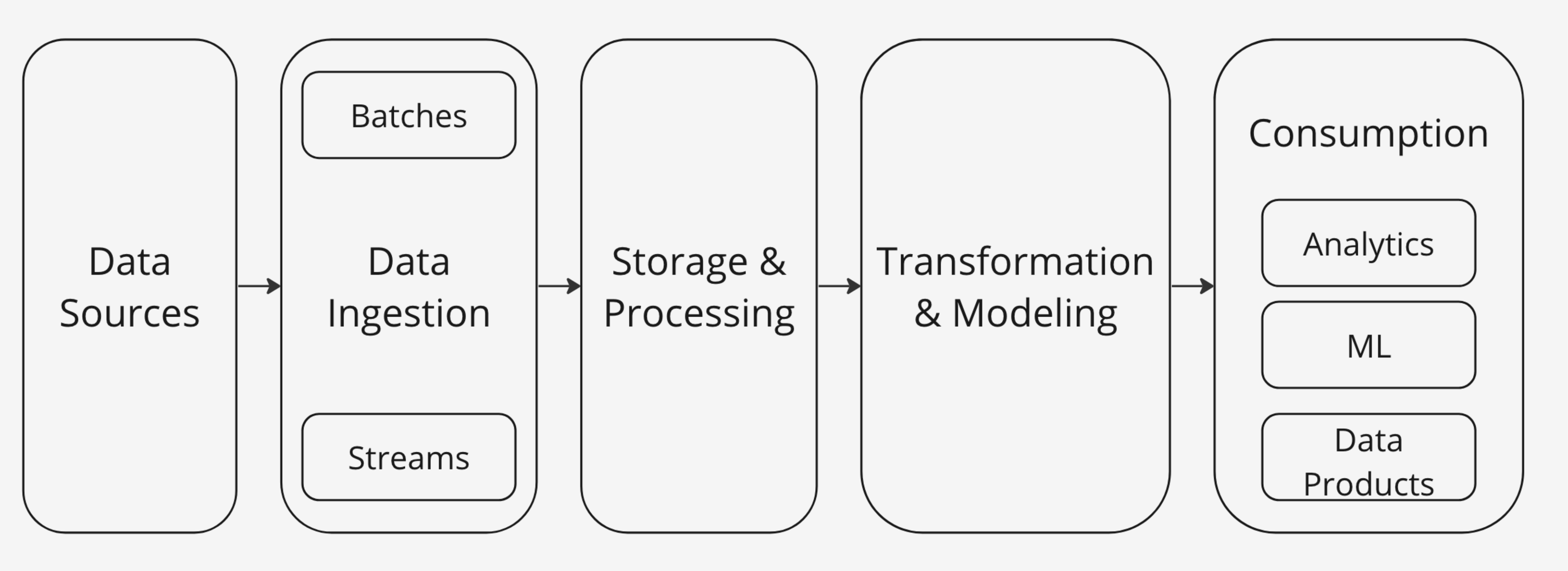
Esoon Ko
IT Högskolan

Getting into **Data Pipelines** and **Data Workflow**

Going back to out **data platform**



Data Sources

Data Ingestion

Batches

Streams

Storage & Processing

Transformation & Modeling

Consumption

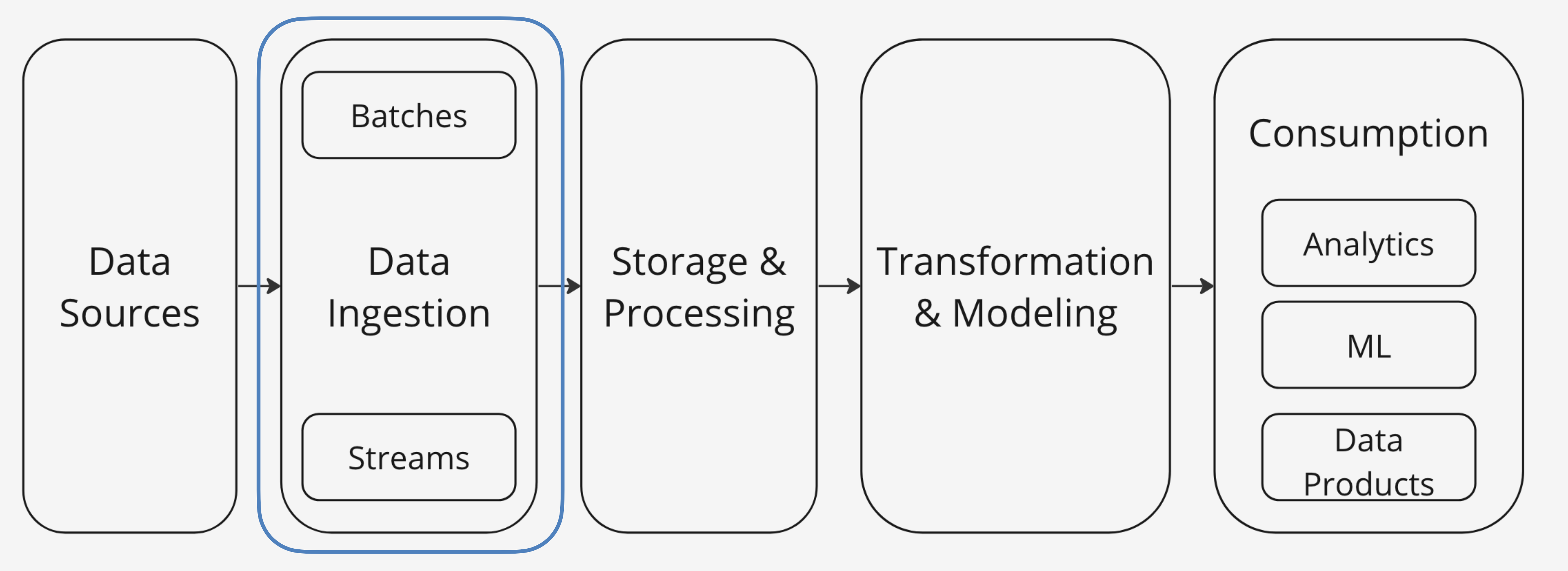
Analytics

ML

Data Products

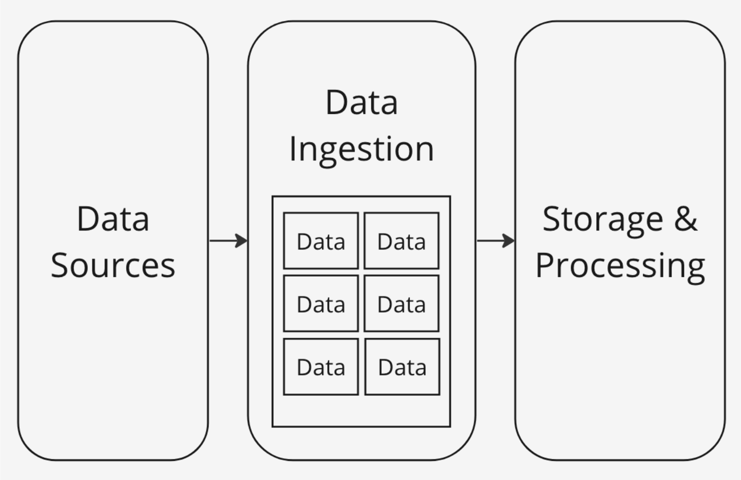
Data Ingestion

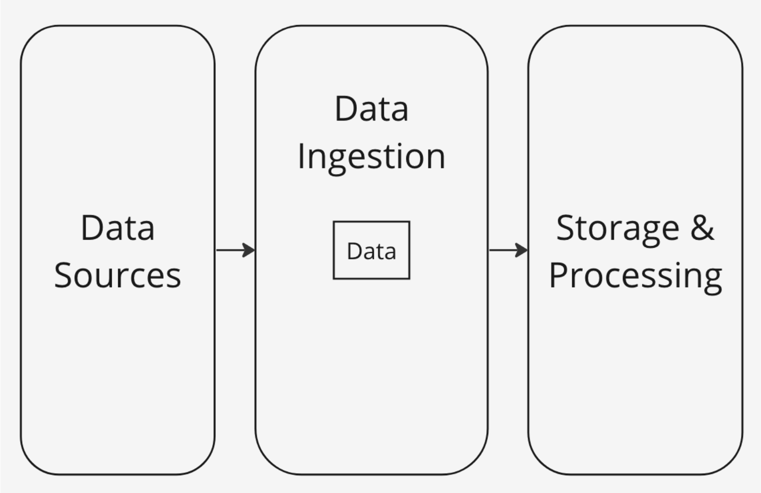
Data is ingested from source into storage.



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.

Data Storage & Processing

Where data is stored and processed. We have many possibilities here…
On premise? Cloud?
Data warehouse? Data lake?

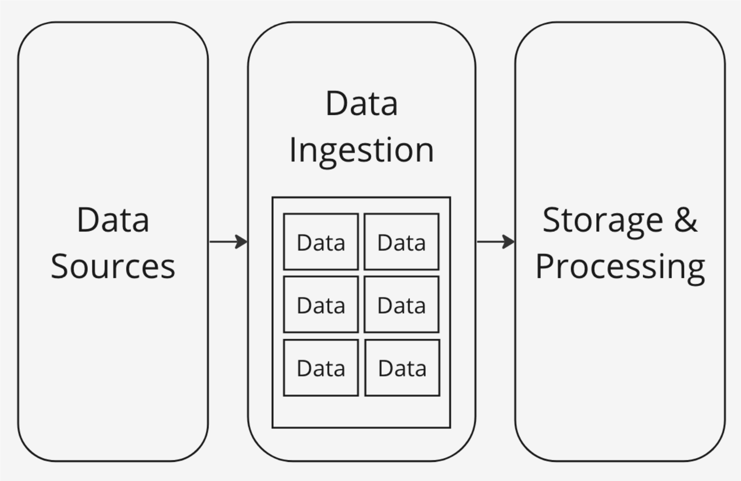
Data Warehouse

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.

Data Ingestion



Lets create a batch ingestion and store it into our warehouse

Batch

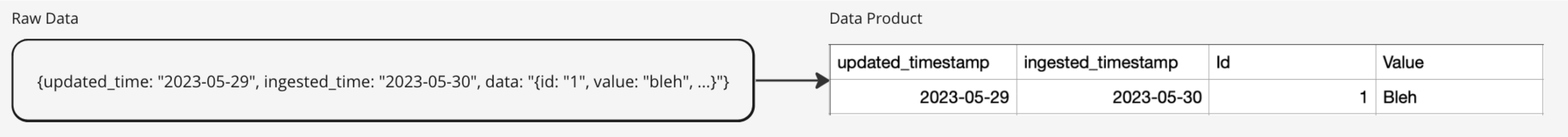
Data ingestion occurs from a source at a pre-defined time or in pre-defined groupings. Data is often pulled.

Benefits: Suitable for large amount of data that can be ingested in intervals.
Disadvantage: Does not provide data in real time.

Data Transformation/Modeling

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

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.



Raw Data

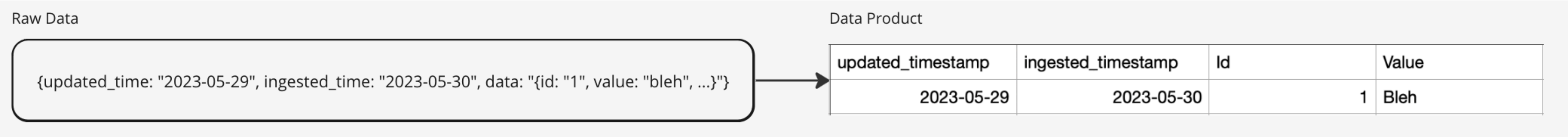
```
{updated_time: "2023-05-29", ingested_time: "2023-05-30", data: "{id: "1", value: "bleh", ...}"}
```

Data Product

updated_timestamp	ingested_timestamp	Id	Value
2023-05-29	2023-05-30	1	Bleh

Data Transformation/Modeling

Lets transform the data so we can use it



Raw Data

```
{updated_time: "2023-05-29", ingested_time: "2023-05-30", data: "{id: "1", value: "bleh", ...}"}
```

Data Product

updated_timestamp	ingested_timestamp	Id	Value
2023-05-29	2023-05-30	1	Bleh