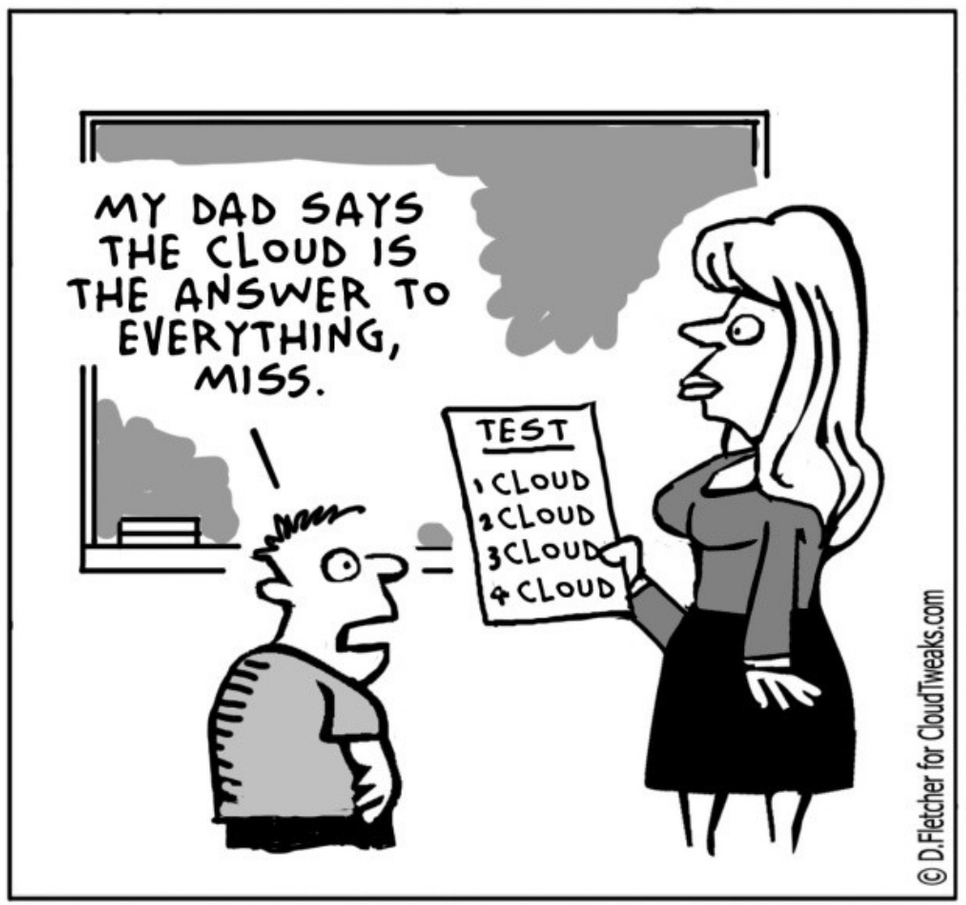
Cloud & Big Data

Mgr. Marcel Bodnár

Agenda

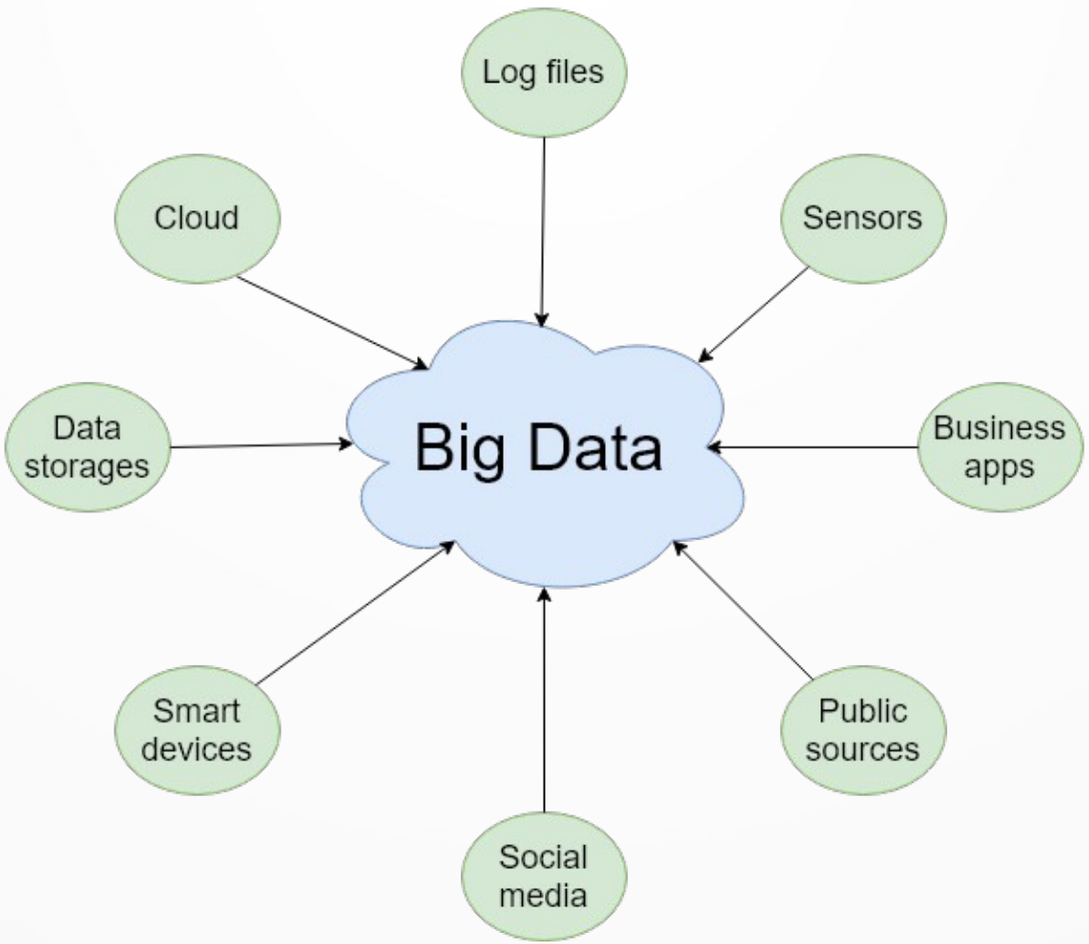
- Big Data challenges
- Data processing
- Cloud concepts



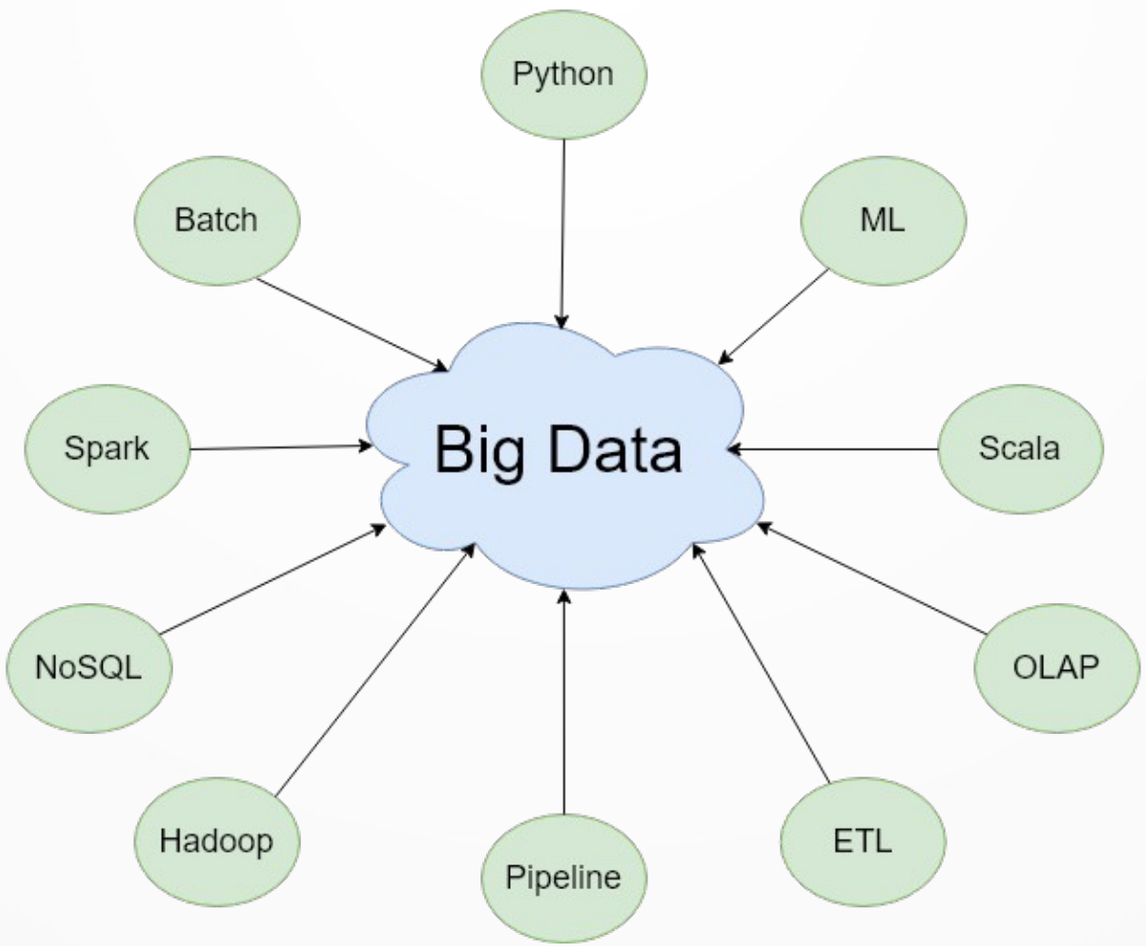
Hands up!

Data and Big Data

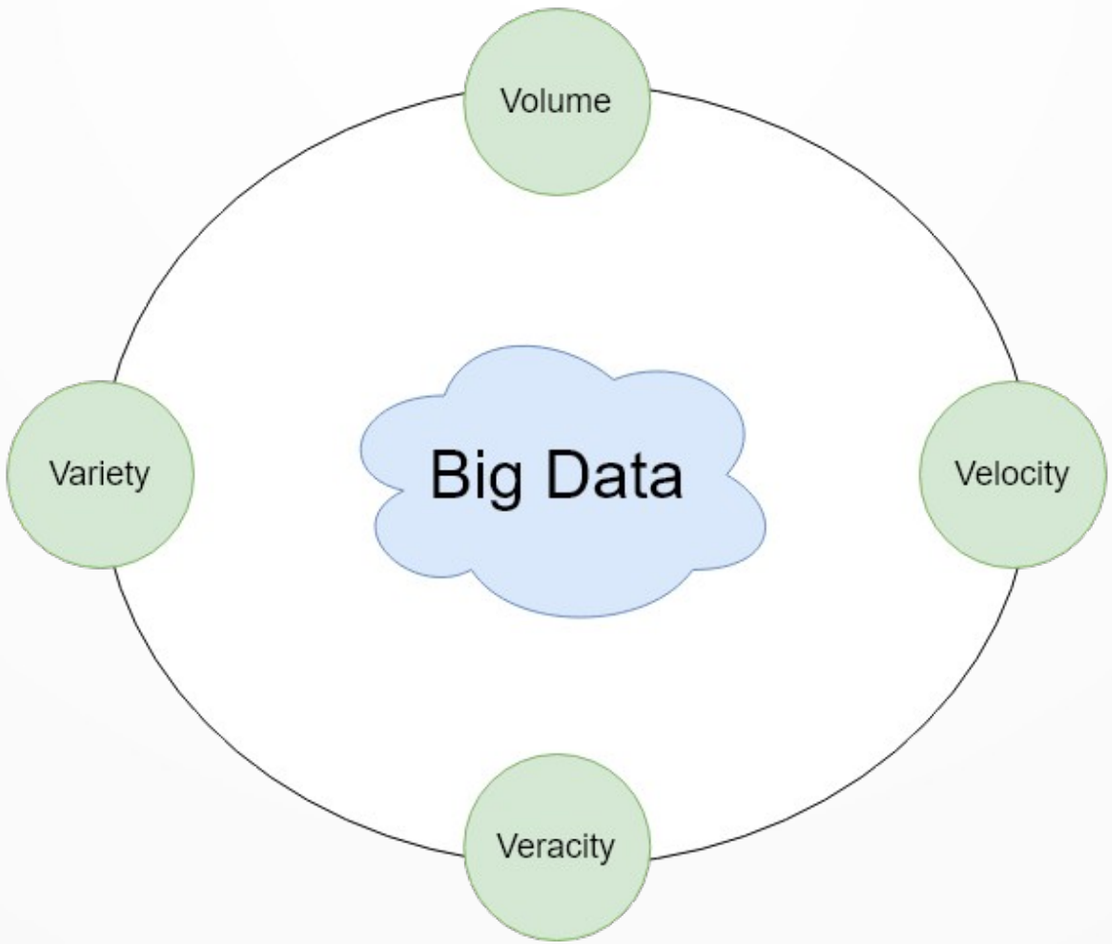
Data sources



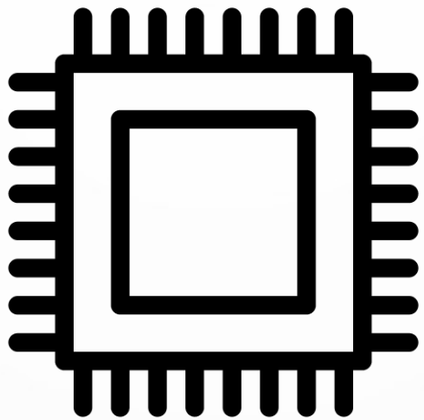
Words like...

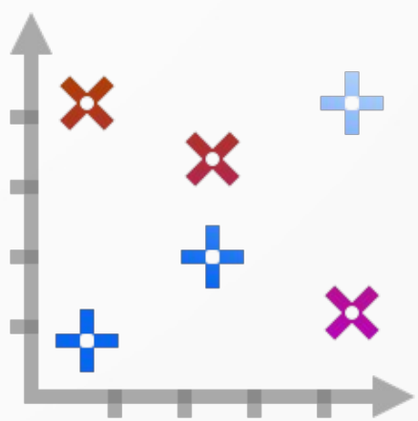

Python
Batch
ML
Spark
Big Data
Scala
NoSQL
OLAP
Hadoop
Pipeline
ETL

Characteristics



Challenges



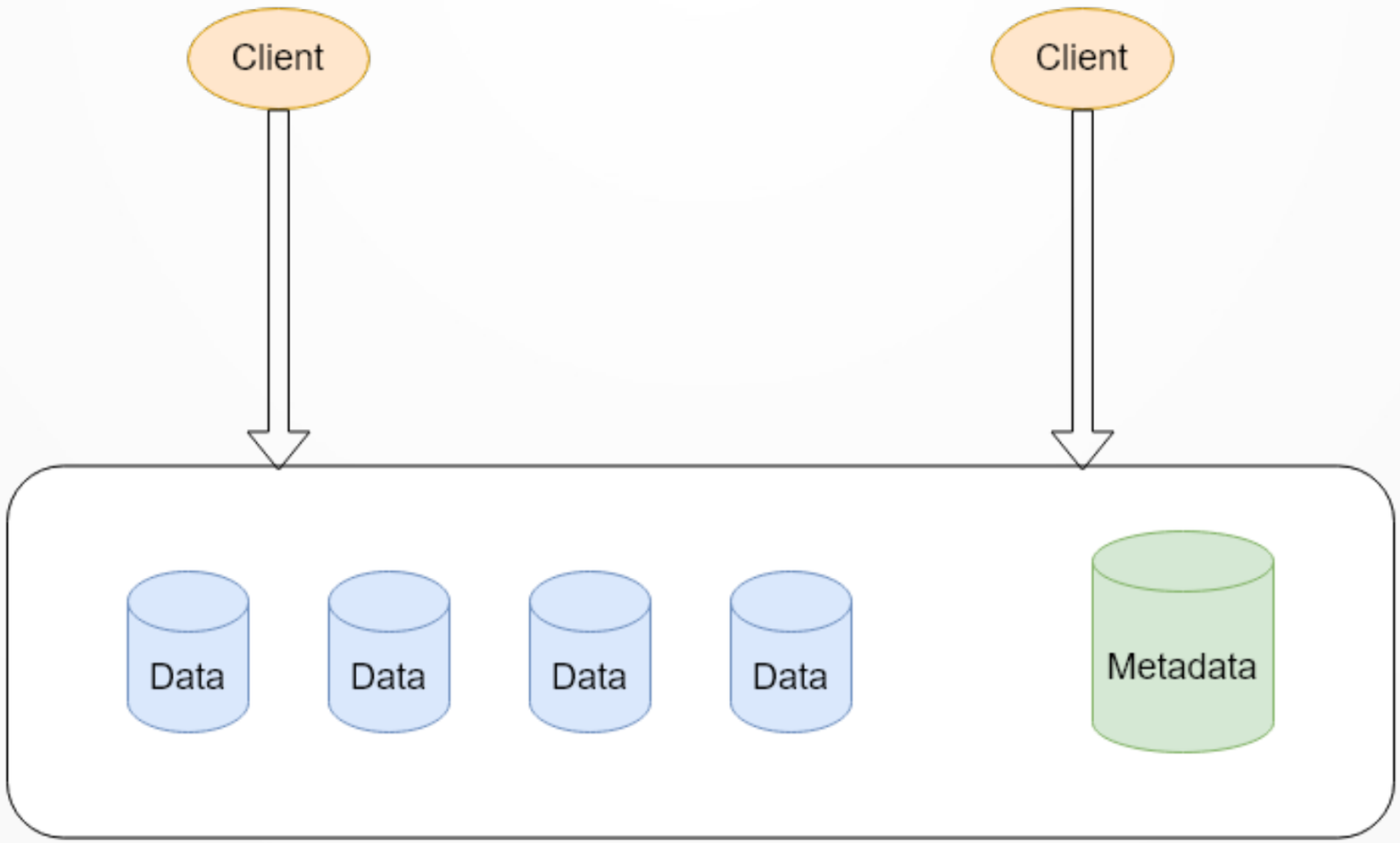


Storage

Storage challenges

- How to store?

Distributed file system

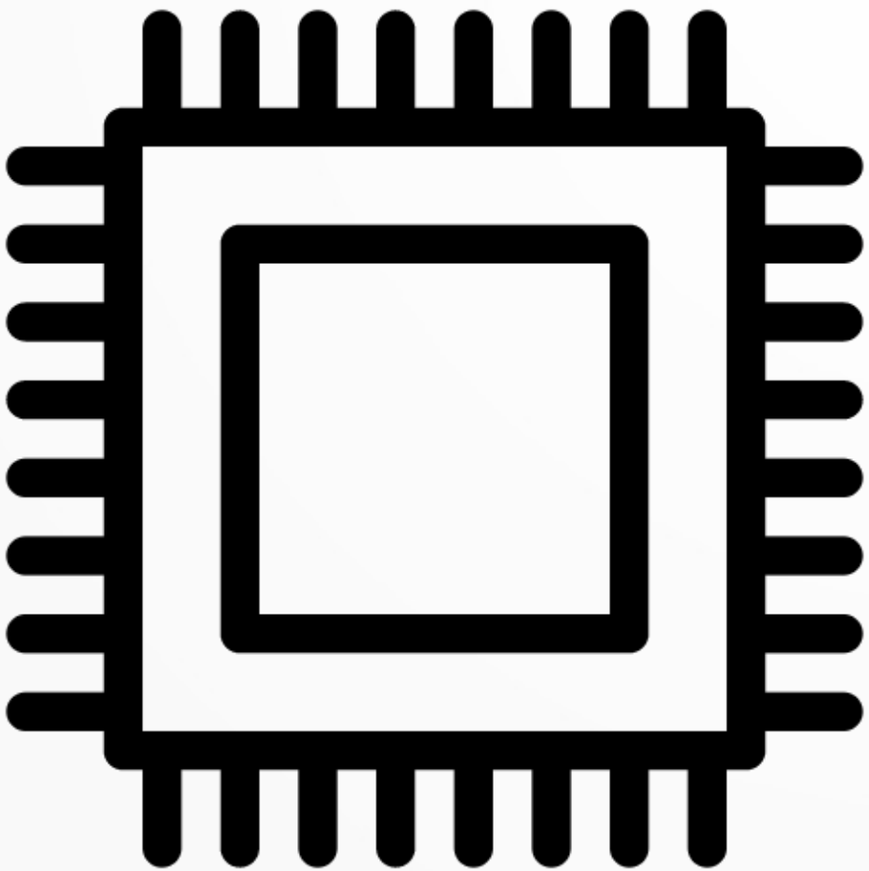


File format matters

- Ideal file format:
  - Fast read
  - Fast write
  - Be split-able
  - Support schema evolution
  - Support advanced compression
- Parquet, AVRO, CSV, JSON, ...

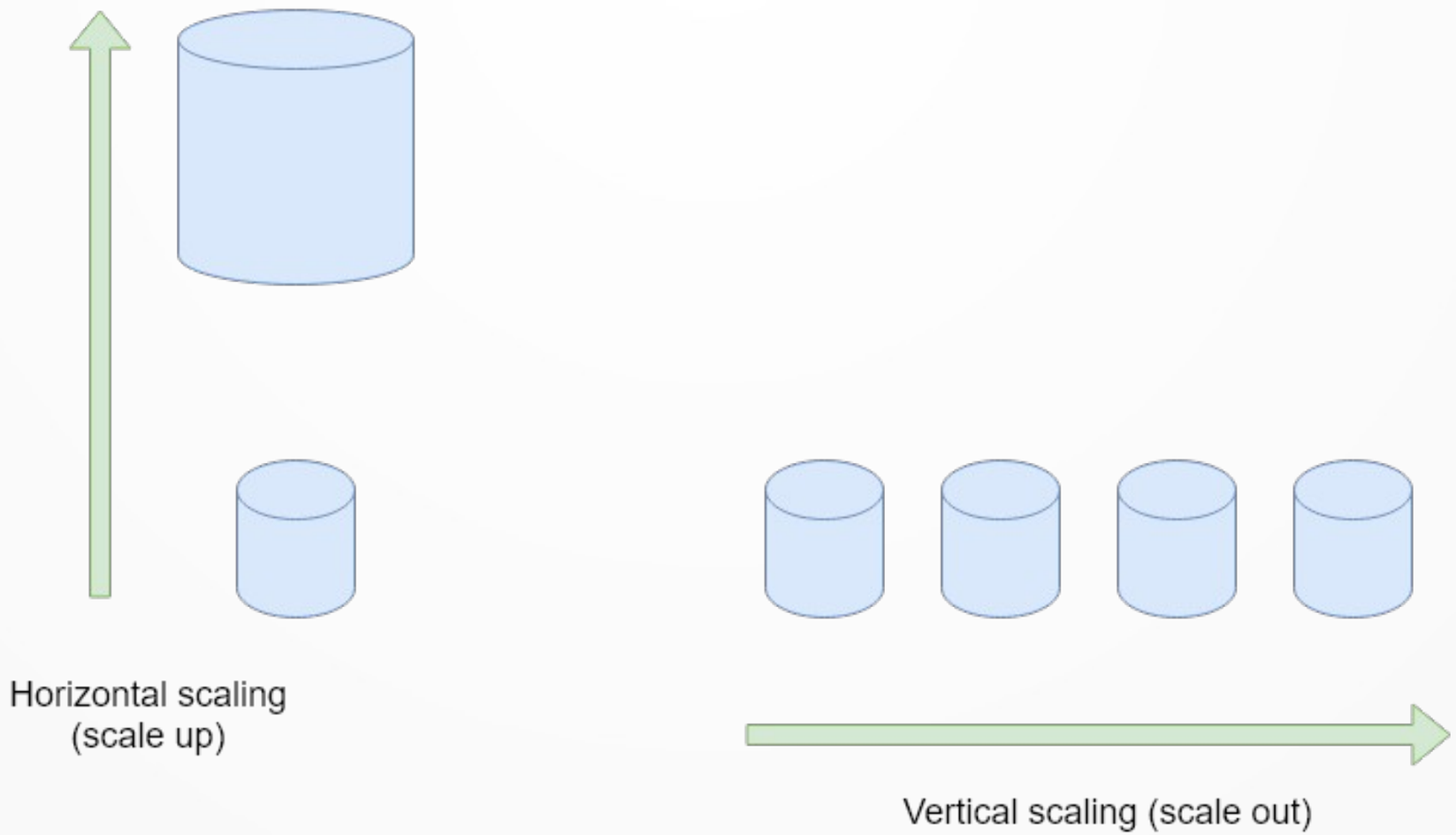
Processing

Processing challenges

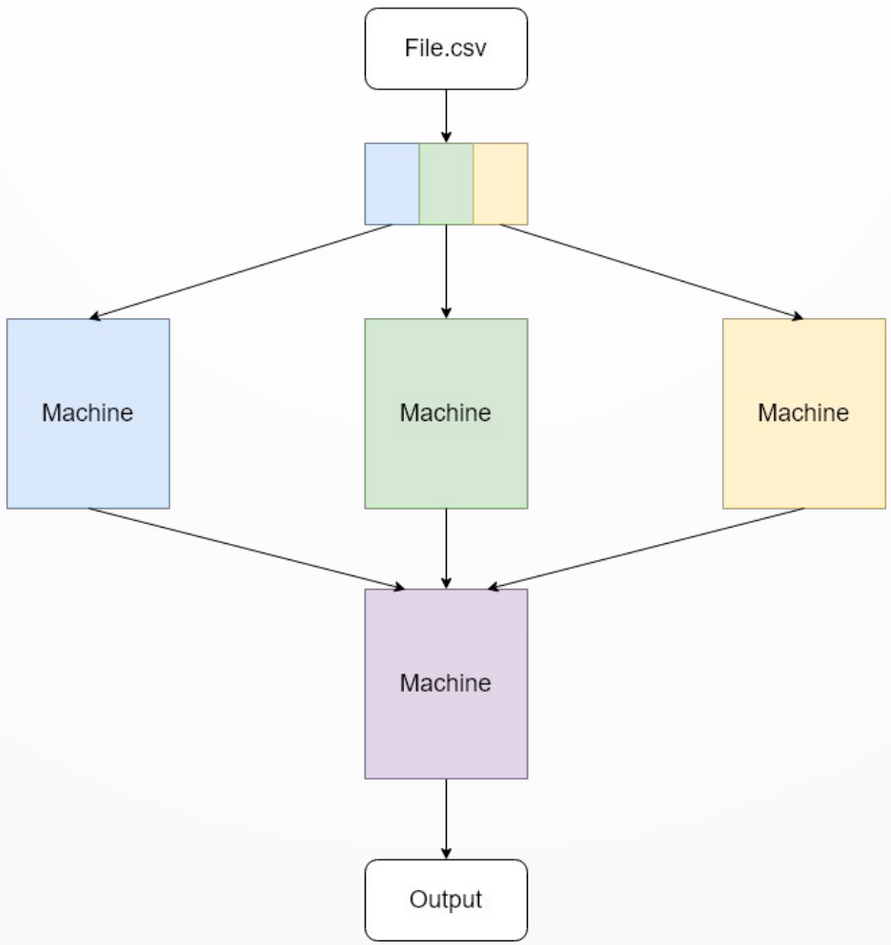


- How to process?

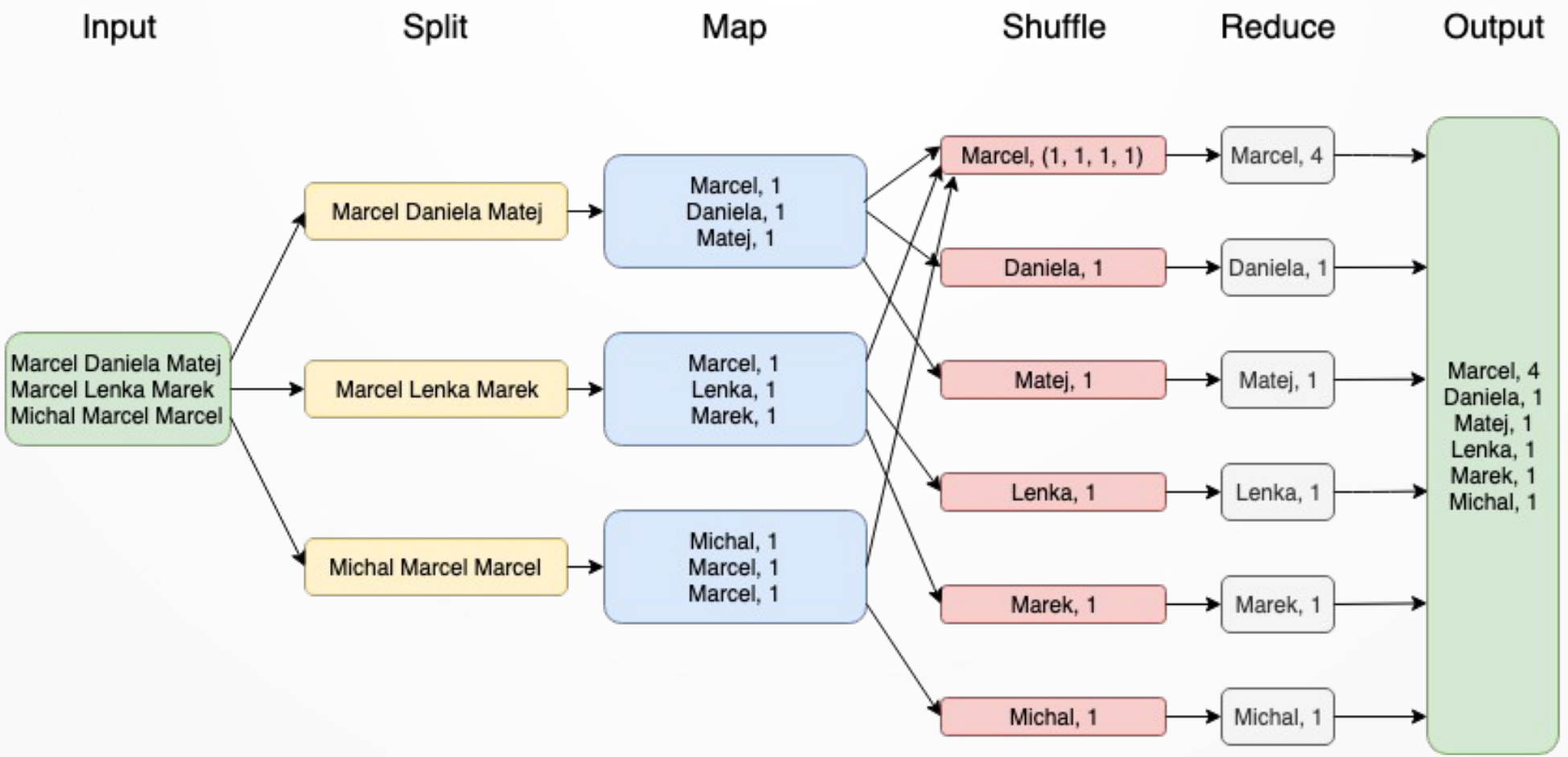
Scaling



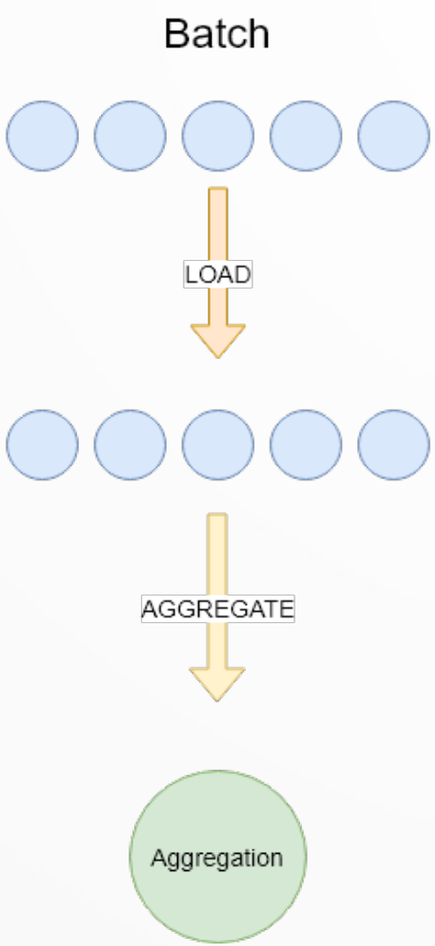
Parallel processing

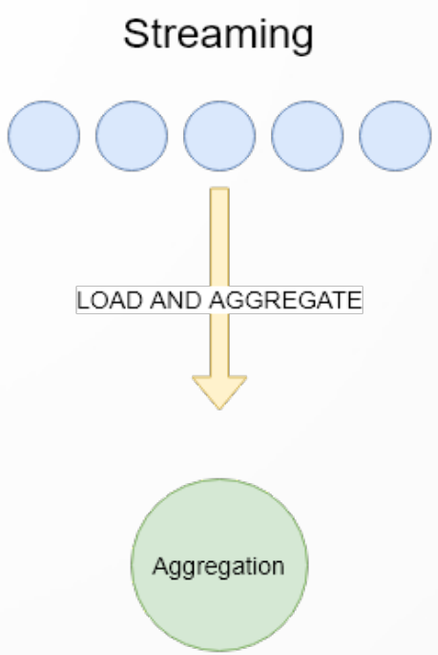


MapReduce as example



Processing models



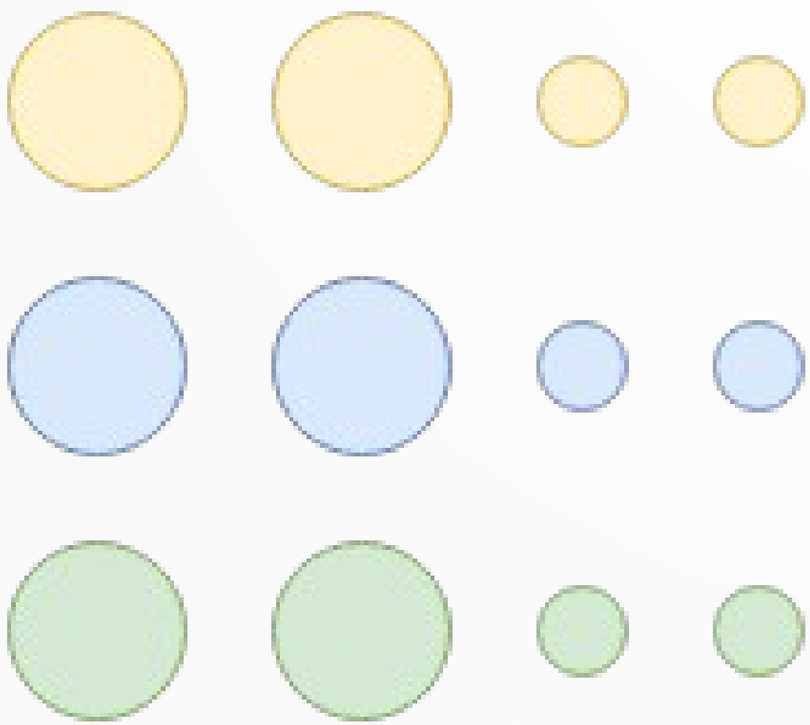


Moving data is hard...

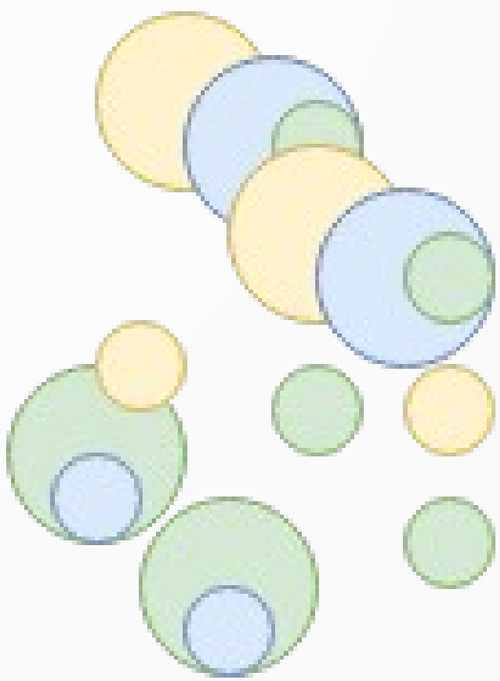
- Exactly once delivery
- Formats
- Protocols
- Security
- Network
- ...

DataWarehouse and Data Lake

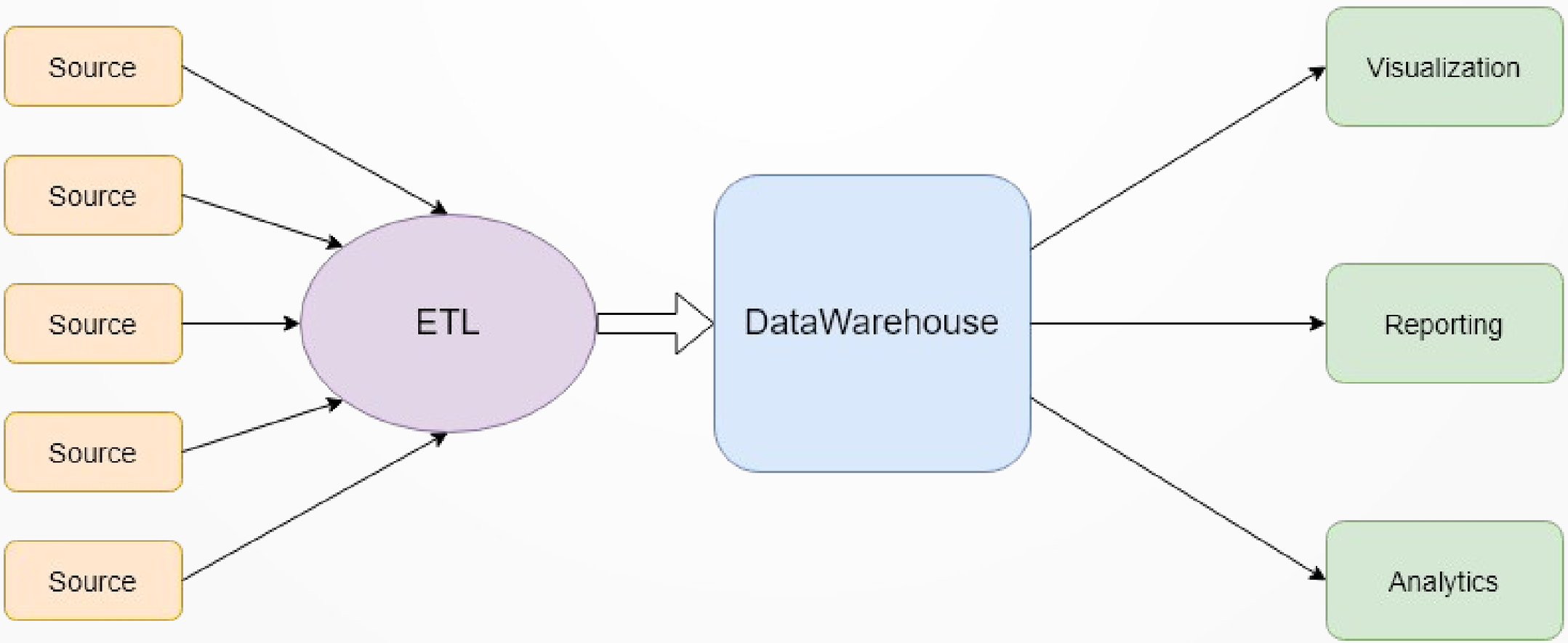
DataWarehouse



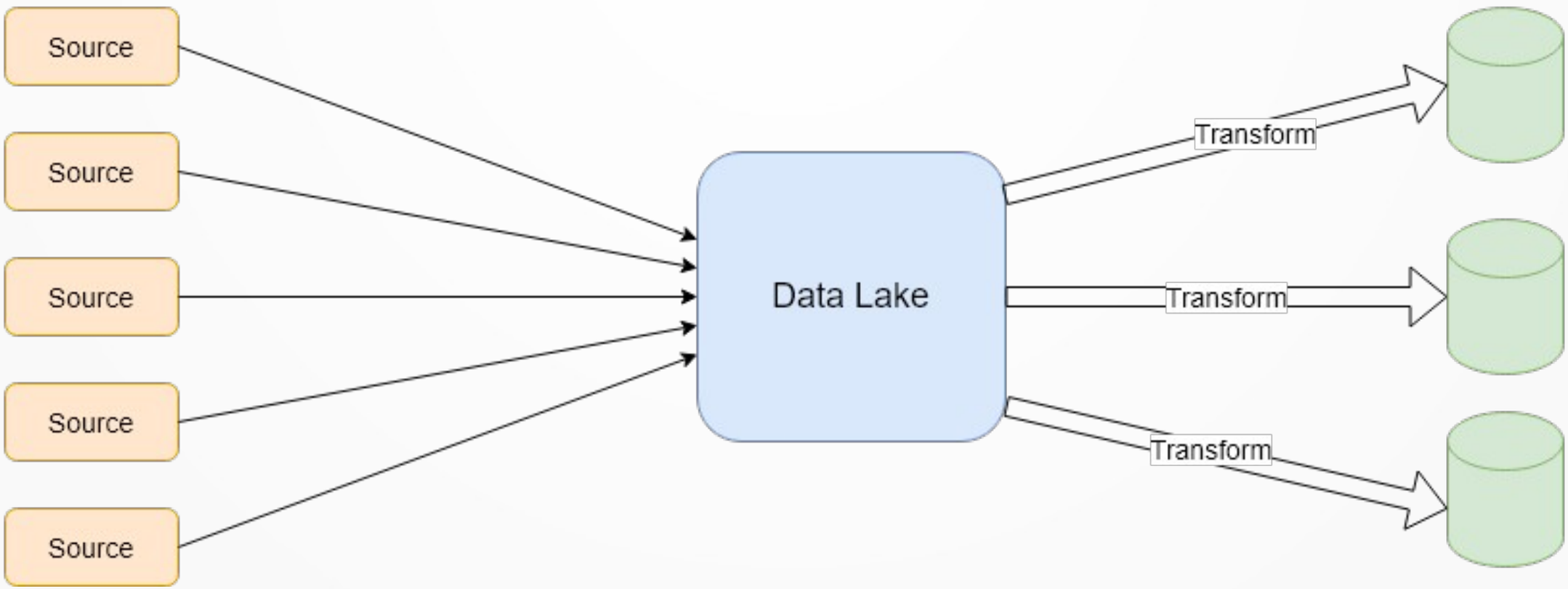
Data Lake



DataWarehouse



Data Lake


Source
Source
Source
Source
Source
Data Lake
Transform
Transform
Transform

Analysis

Analysis challenges

- How to analyze?

Analysis

- Cleaning
- Normalization
- Remove outliers
- …
- Apply ML algorithms

Search

Search challenges

- How to search?

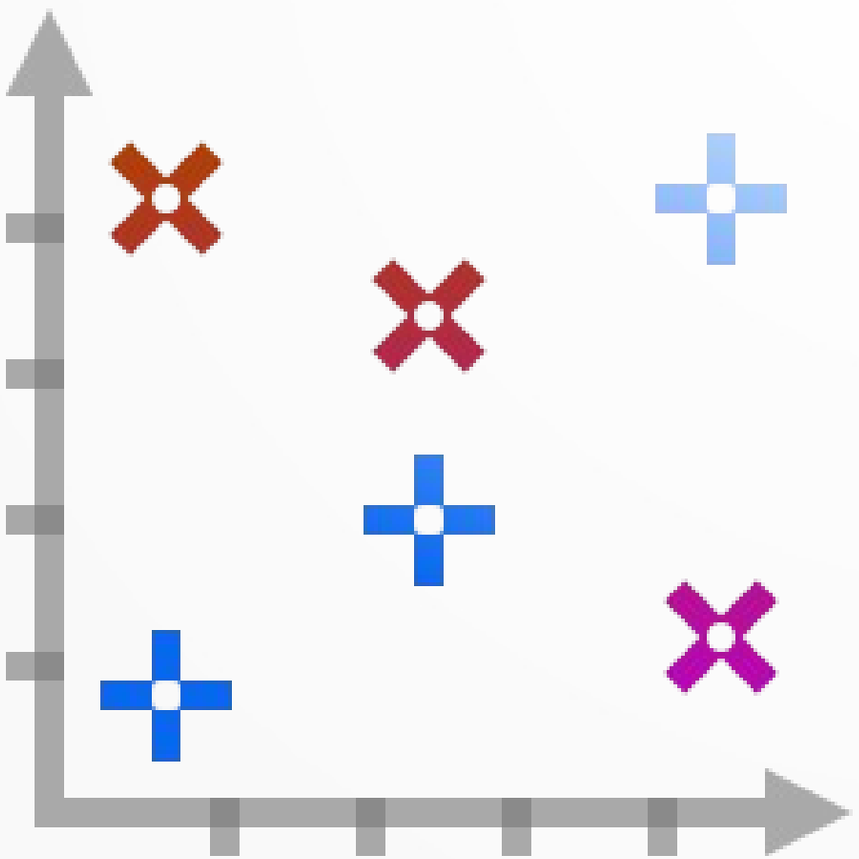
Security

Security challenges

- How to secure?

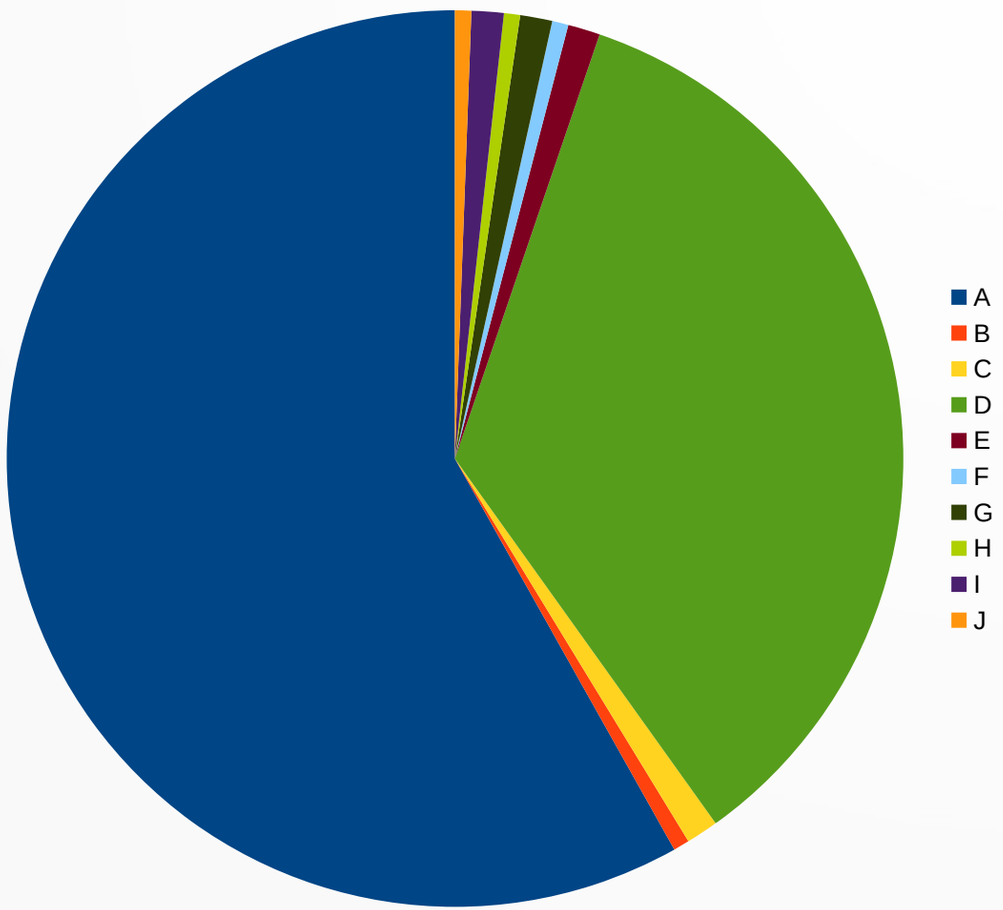
Visualization

Visualization challenges



- How to visualize?

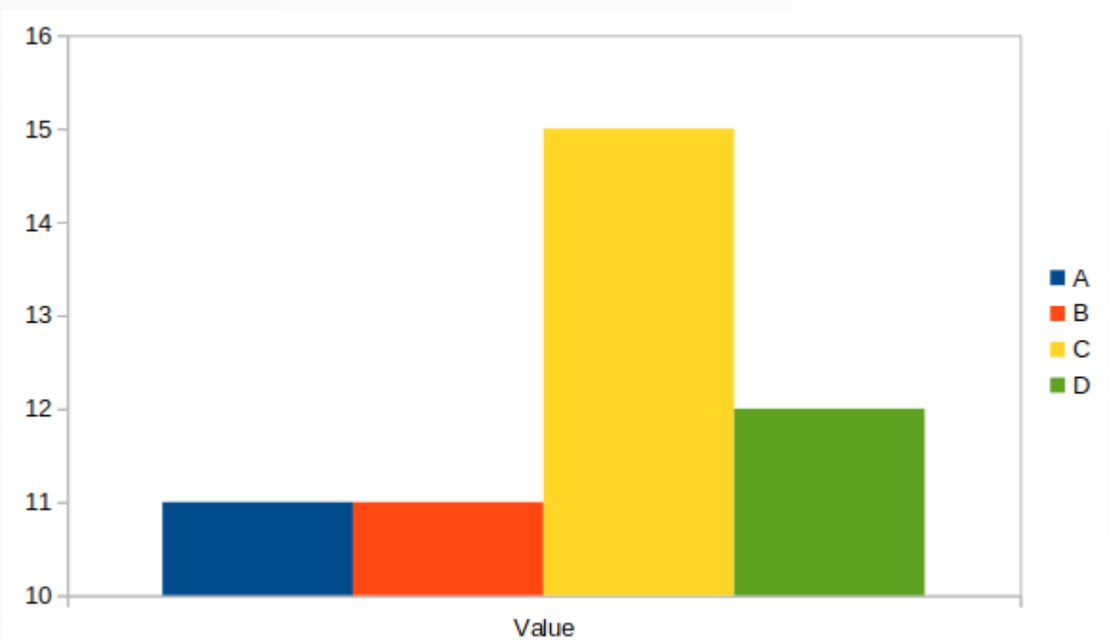
Common mistakes

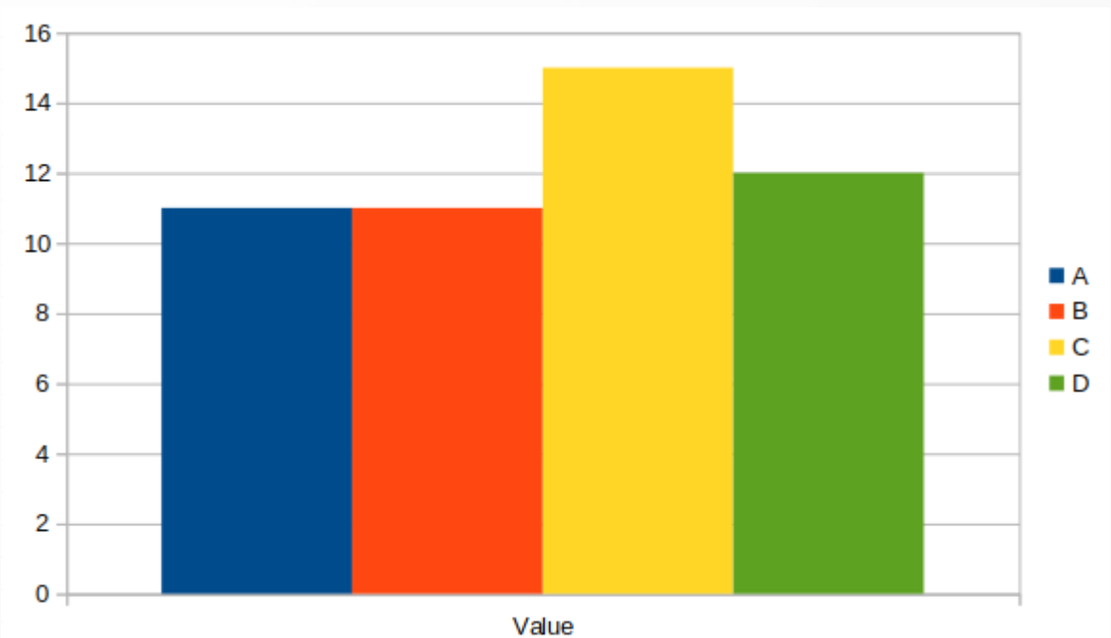


- Too many variables

Common mistakes

- Incorrect scale (truncated graph)





Common mistakes

- Incorrect scaling

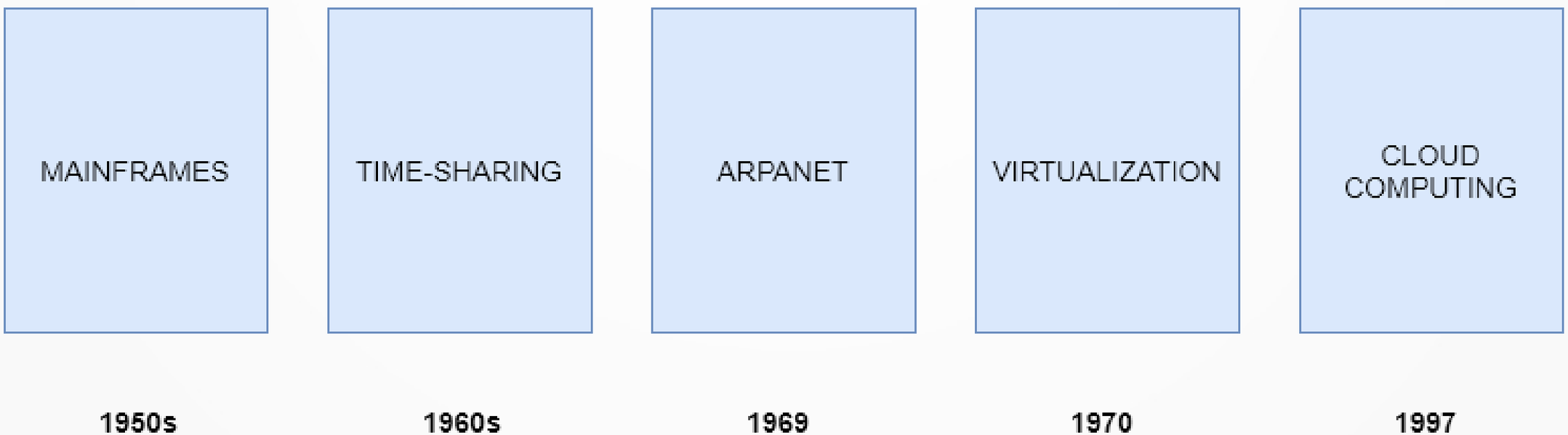
Challenges

Cloud

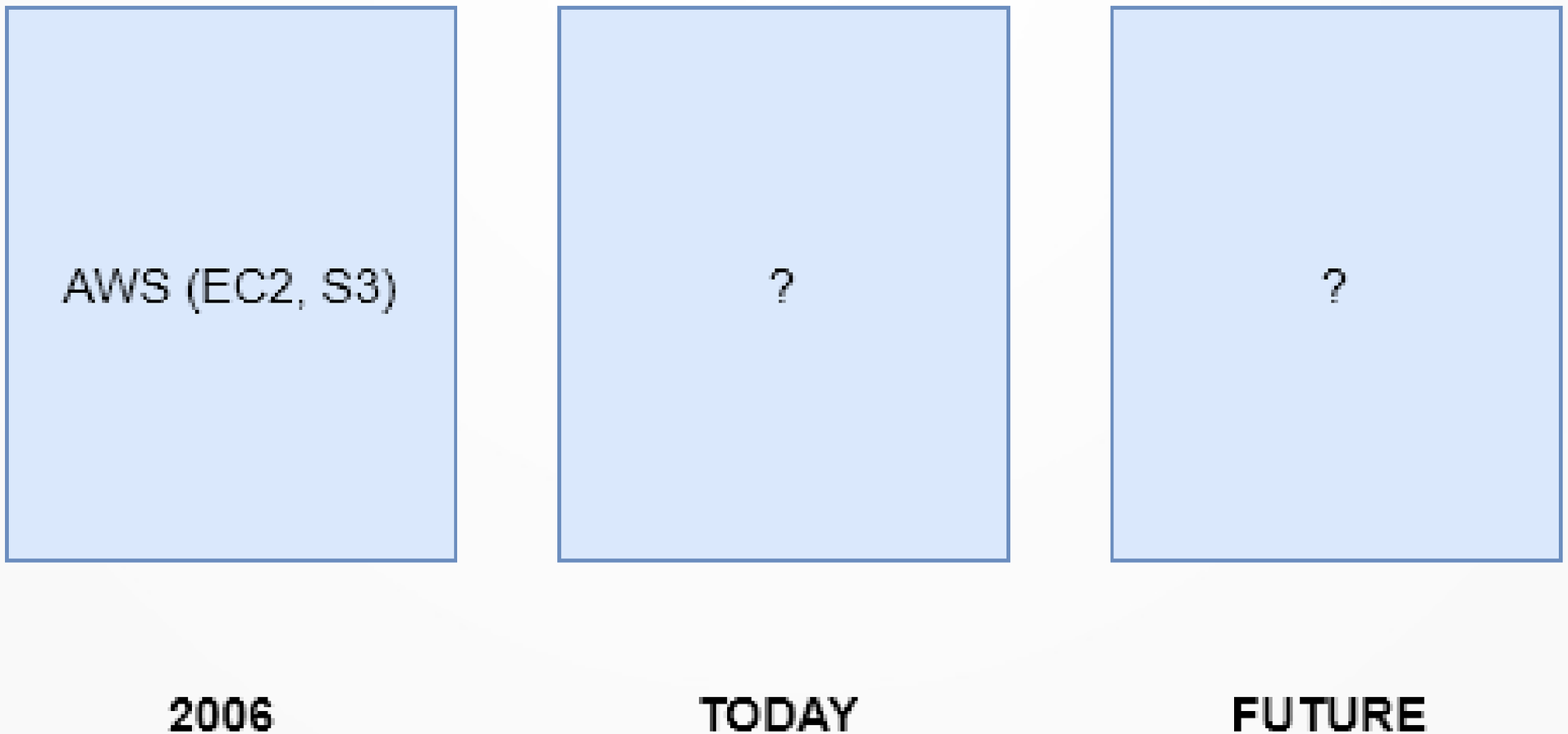
Cloud providers

- AWS (https://aws.amazon.com/)
- Microsoft Azure (https://azure.microsoft.com/)
- Google Cloud (https://cloud.google.com/)
- Others

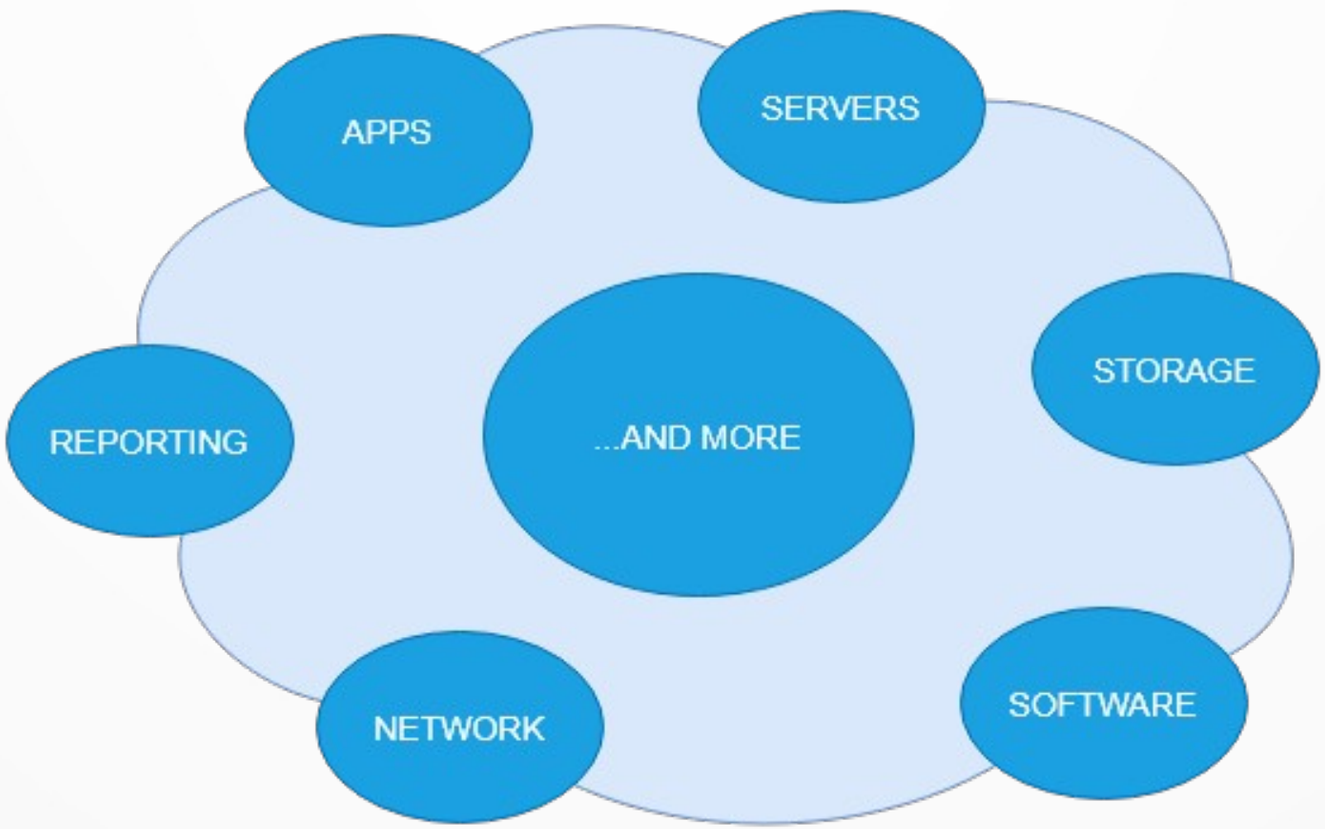
Cloud computing prehistory



Cloud computing history



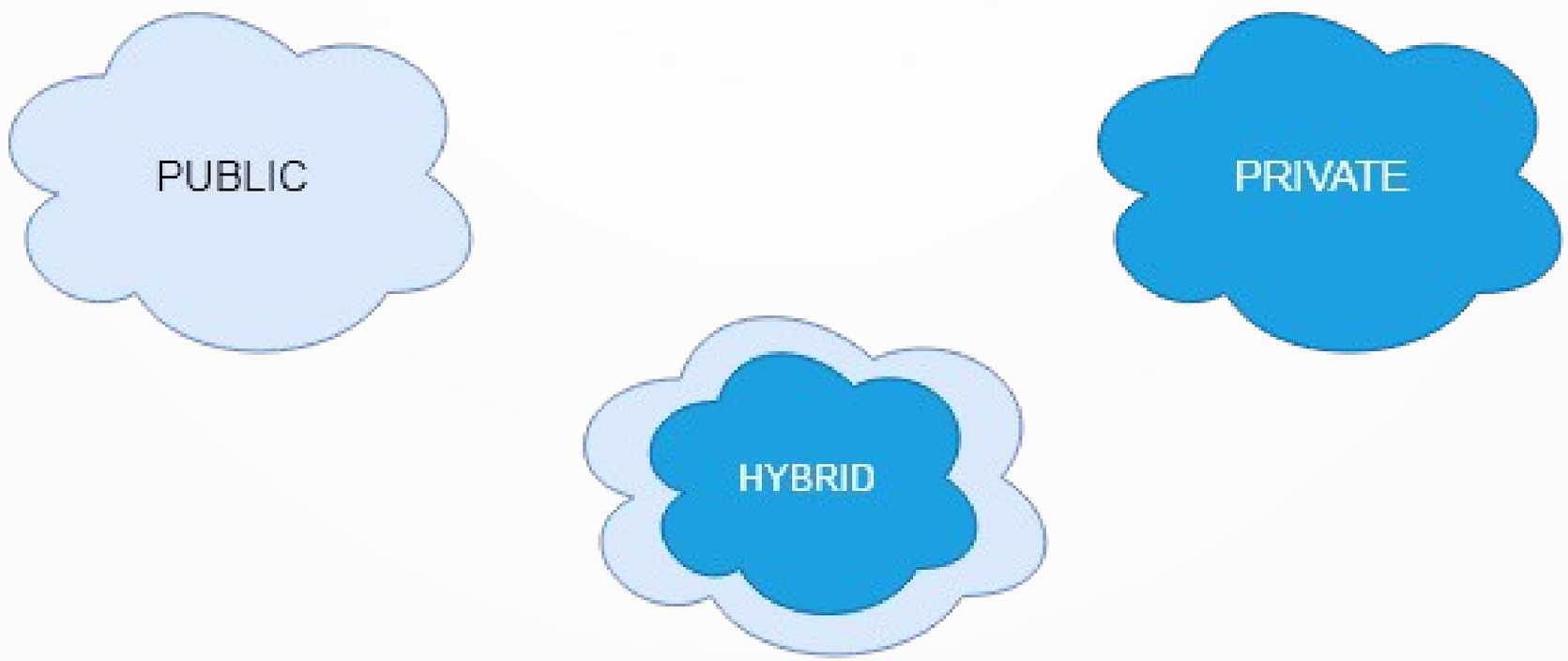
Cloud computing



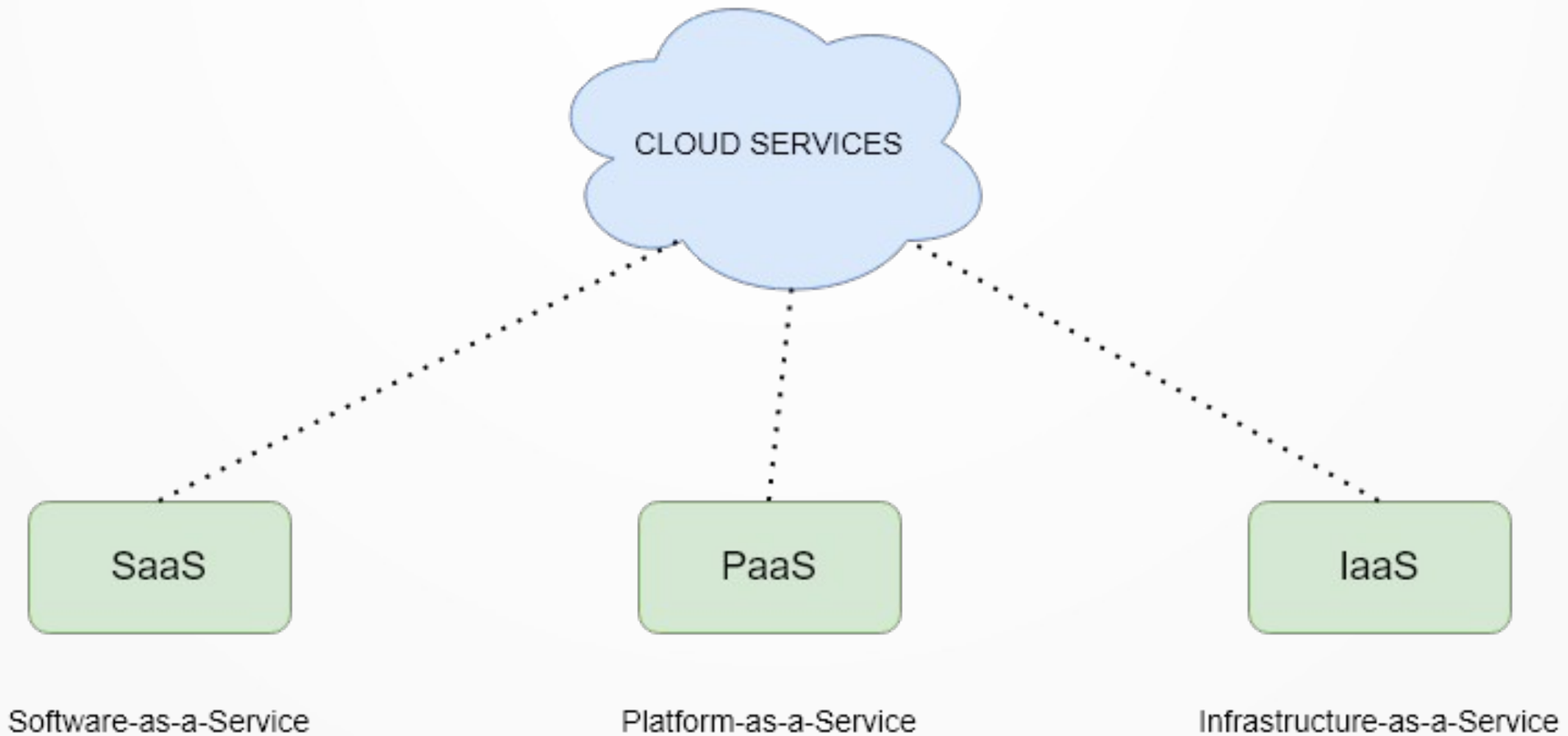
Cloud concepts

- High availability
- Elasticity
- Scalability
- Agility
- Fault tolerance
- Disaster recovery

Cloud models



Cloud service models



Cloud service models

- https://www.youtube.com/watch?v=36zducUX16w

Key benefits of Cloud Computing

- Quick deployment
- Flexibility
- Automatic software updates
- Efficiency and cost reduction
- Data security
- Mobility
- Collaboration

Q & A

Thank you!