Storing & Extracting Data

Learning objectives

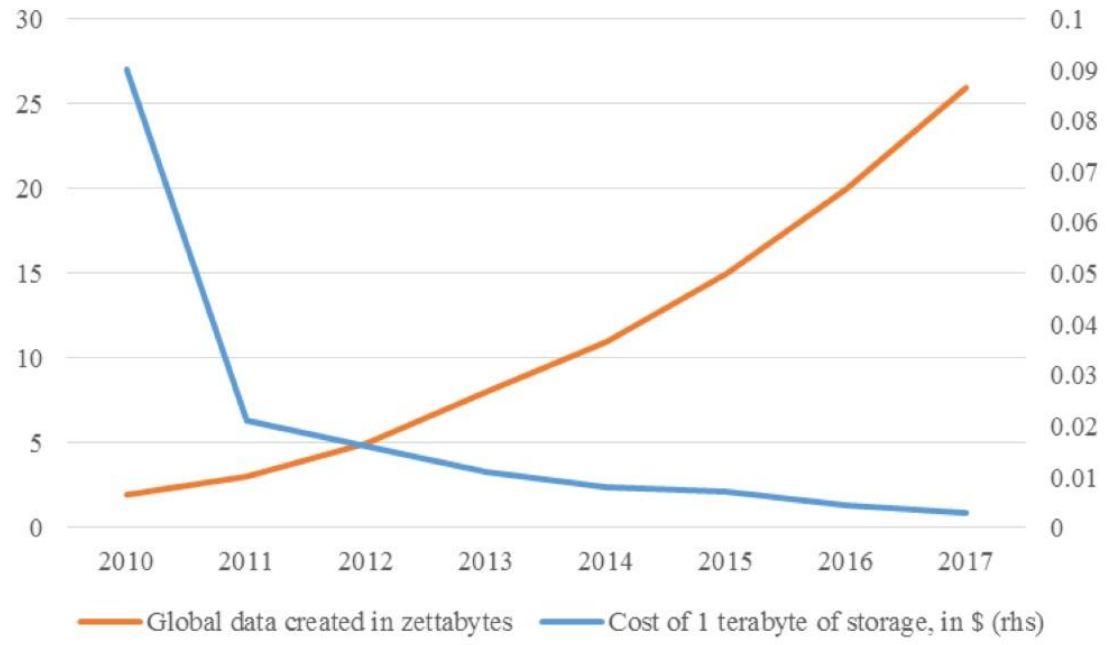
- Exposure to the most common data storage ecosystems.
- Learn the most common challenges around ingestion, storage, and extraction of data.
- Practice loading different types of datasources in Tableau.

Data Trends

Back in 70's, when the first databases were invented, storing data was absurdly expensive.

That has changed and consequently we are capturing exponential amounts of data.

Figure 3: Costs of storage and global data availability, 2009-2017



Source: Reinsel, Gantz and Rydning (2017); Klein (2017). One zettabyte is equal to one billion terabytes.

Spreadsheet Galore

Imagine your organization wants to store operational data in a spreadsheet (e.g. Excel). Do you think this is a good idea or not?

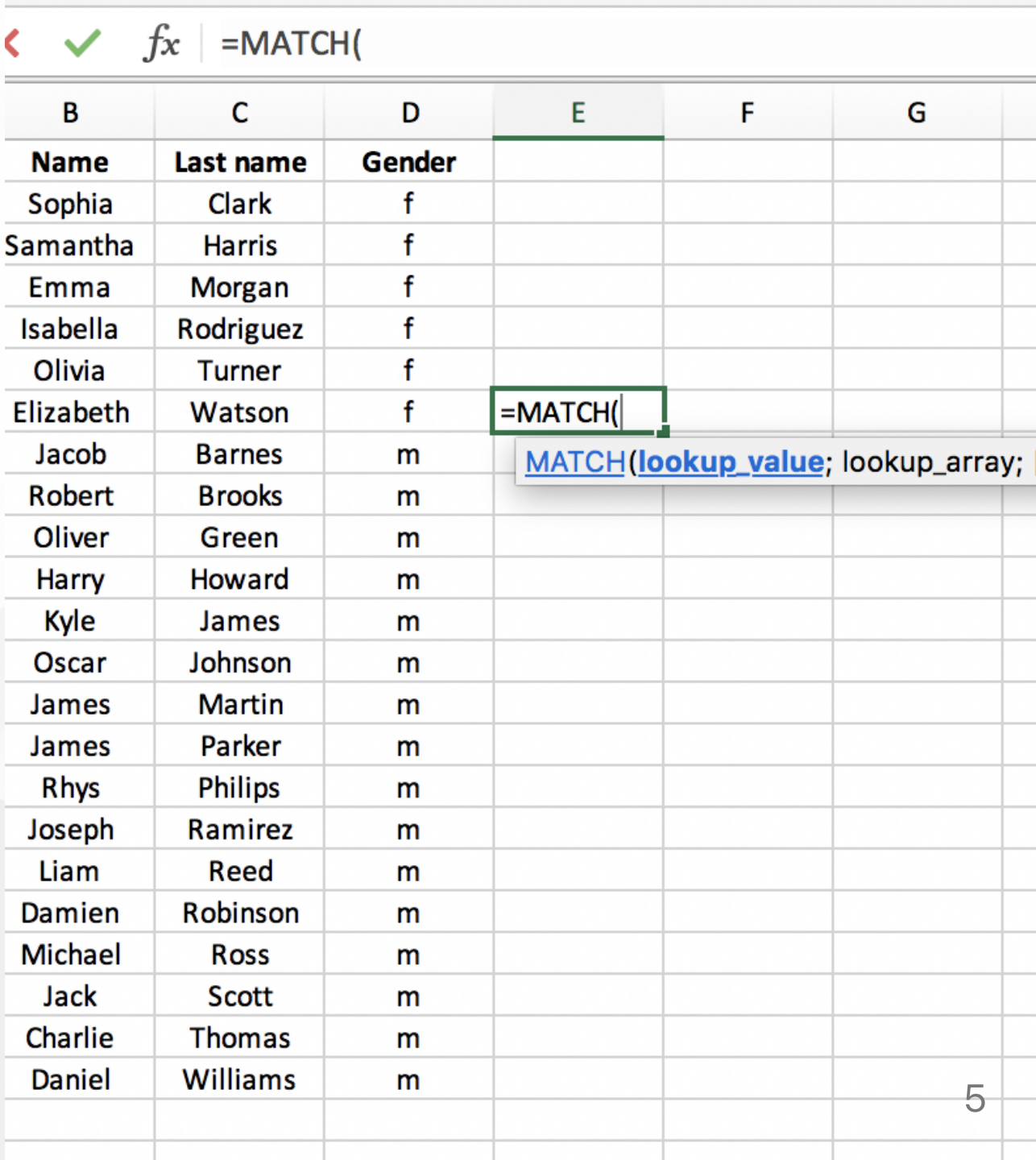
Where is the data?

Data is generally scattered:

- `Flat files`
- `Databases`
- `Data Warehouses`
- `Data Lakes`
- `Source systems`
- `APIs`

Should you store your data in a spreadsheet?

Imagine you work for an organization that wants to start capturing their data. What would be the `pros` and the `cons` of storing their data in a spreadsheet (e.g. Excel)?



Flat Files

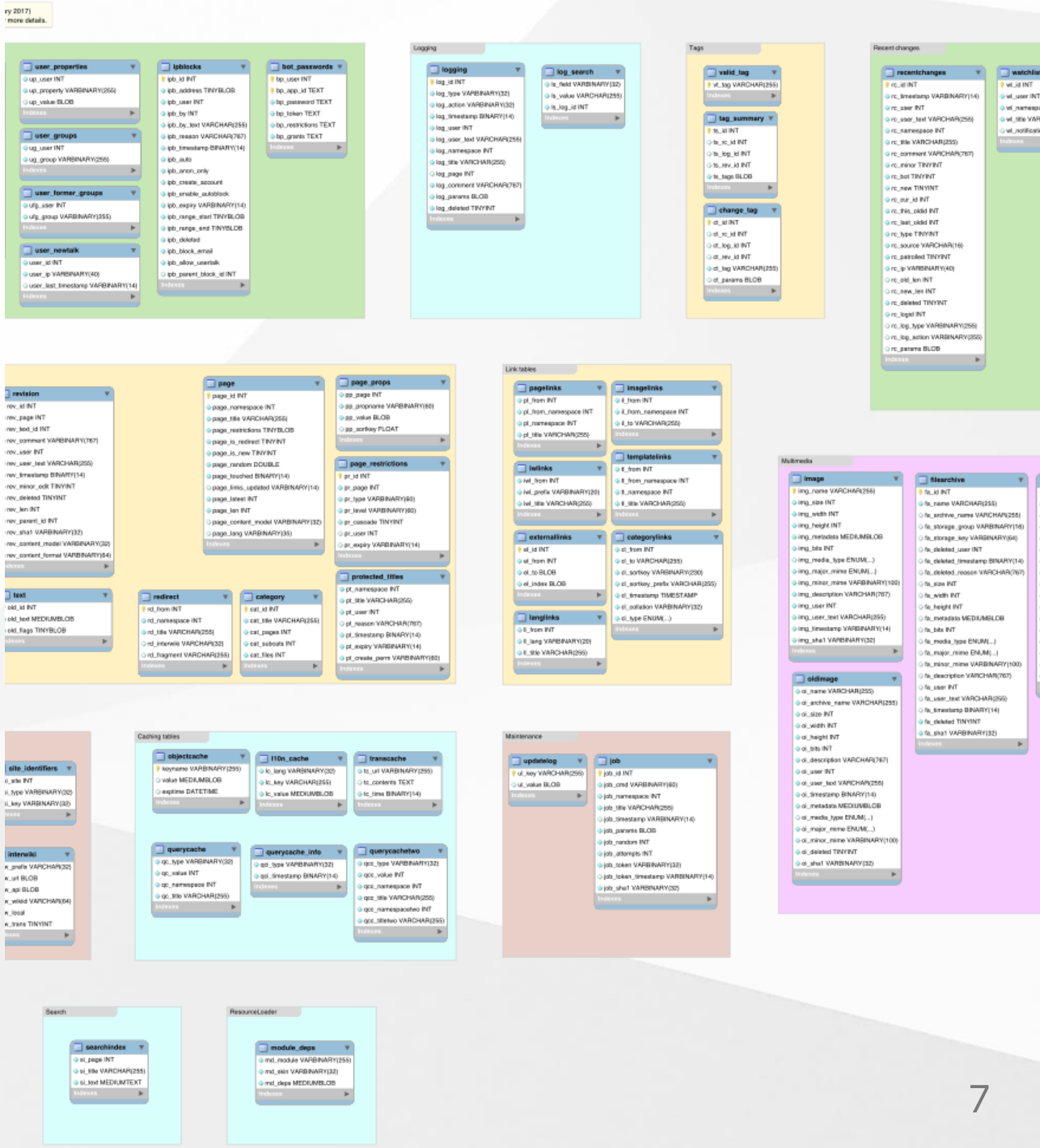
Common formats include `.csv`, `.txt`, `.xlsx`, `.json`, `xml`, `.avro`, `.parquet`.

+ Flexibility.
- Collaboration.
- Security.
- Scalability.

Databases

Databases are an attempt to professionalize data storage. They are composed of spreadsheet-like tables that are related to each other.

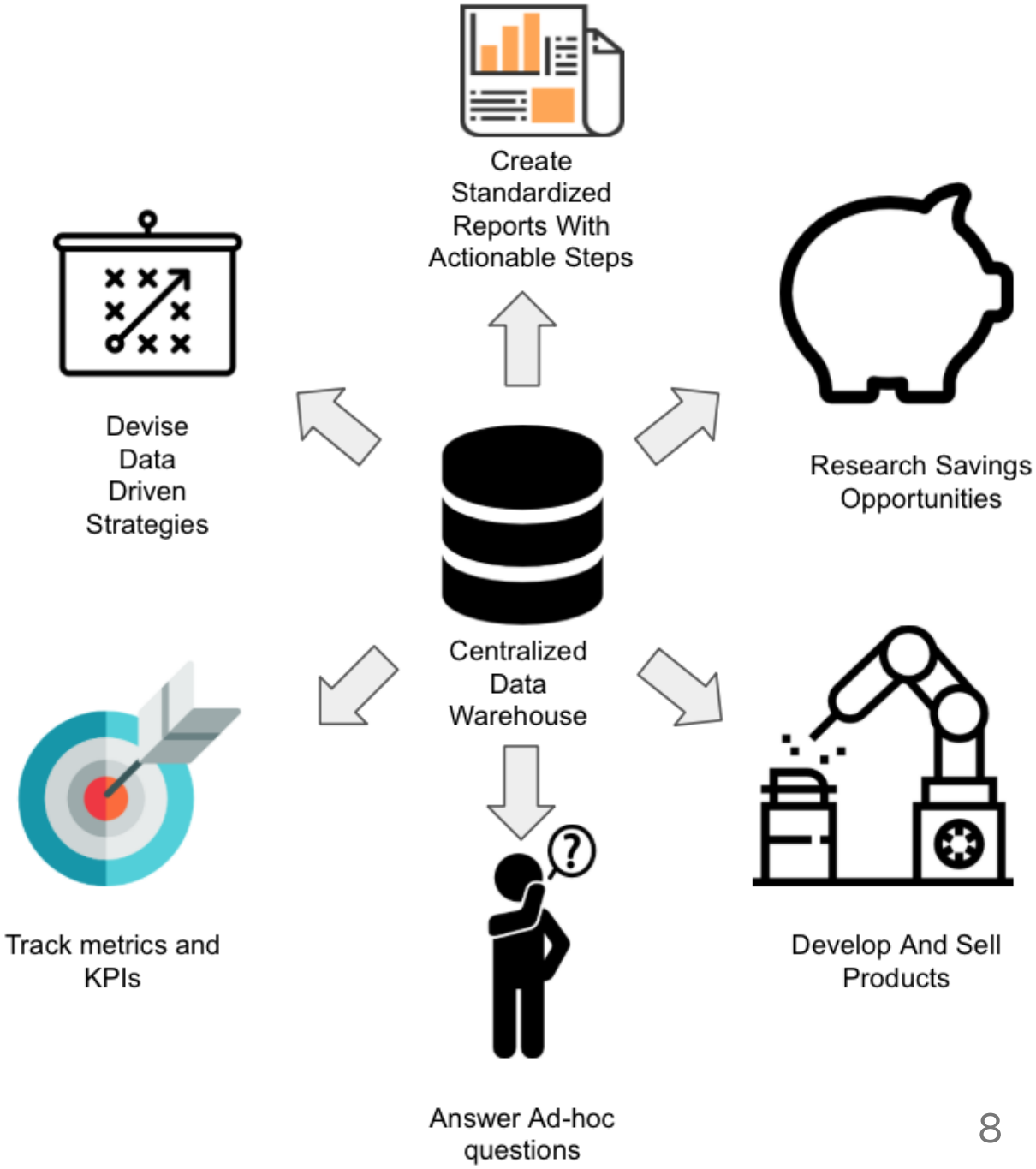
+ Collaboration.
+ Security.
- Flexiblity.
- Scalability.
- Tabular.



Data Warehouse

Data warehouse look and feel like a database. But they are optimized for analytics (instead of powering an application).

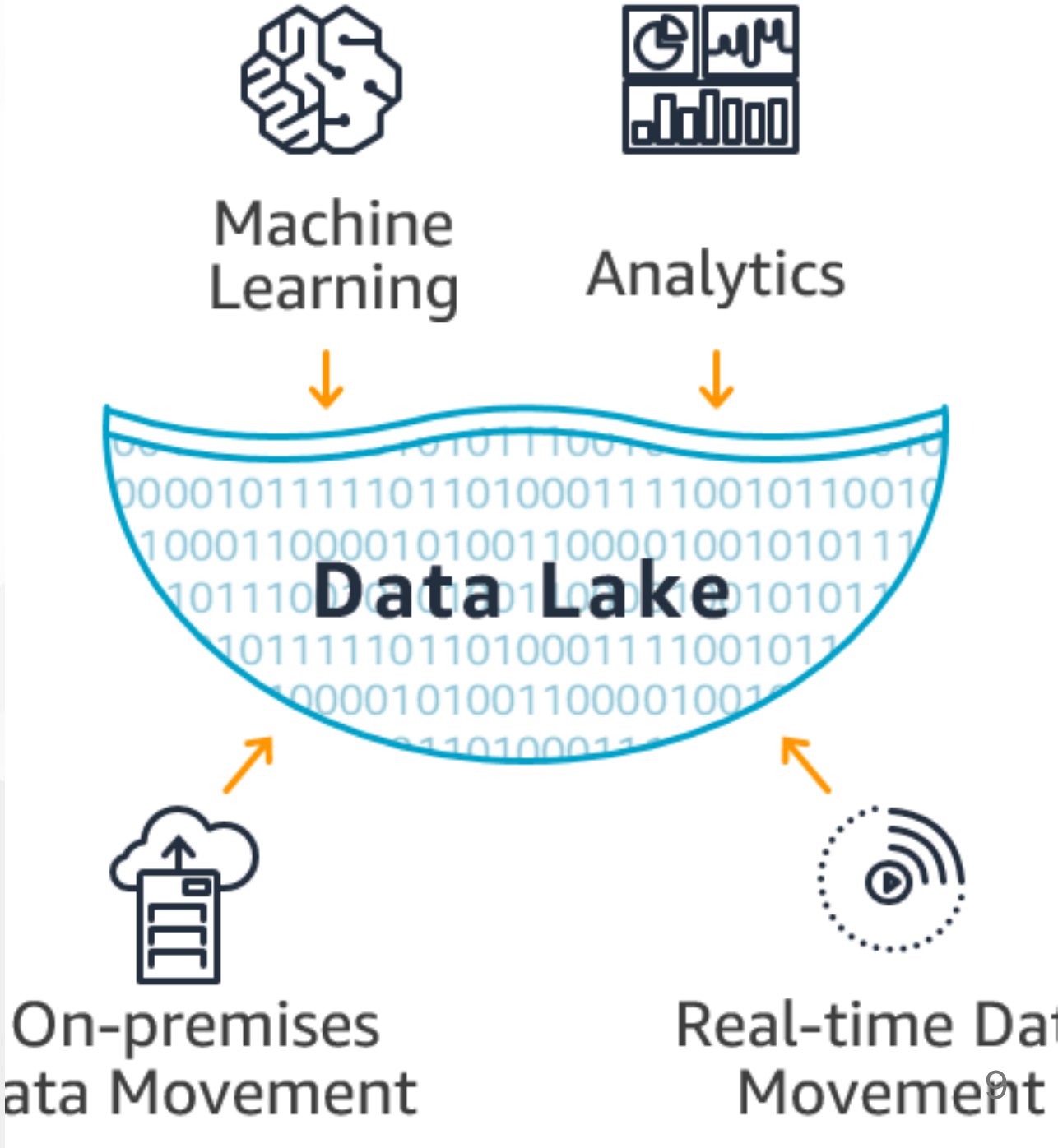
+ Collaboration.
+ Security.
+ Scalability.
- Flexibility.
- Tabular.



Data Lake

Data lakes are a cheaper, more flexible of data warehouses. Data doesn't need to be tabular or relational anymore. But it can get messy.

+ Collaboration.
+ Scalability.
+ Flexibility.
- Order.



Source Systems

Source system is any system that captures data. You generally don't want it to live here.

- Flexibility.
- Centralization.

APIs

APIs can be used as a secure interface to allow anybody to query data.

+ Automation.
+ Security.