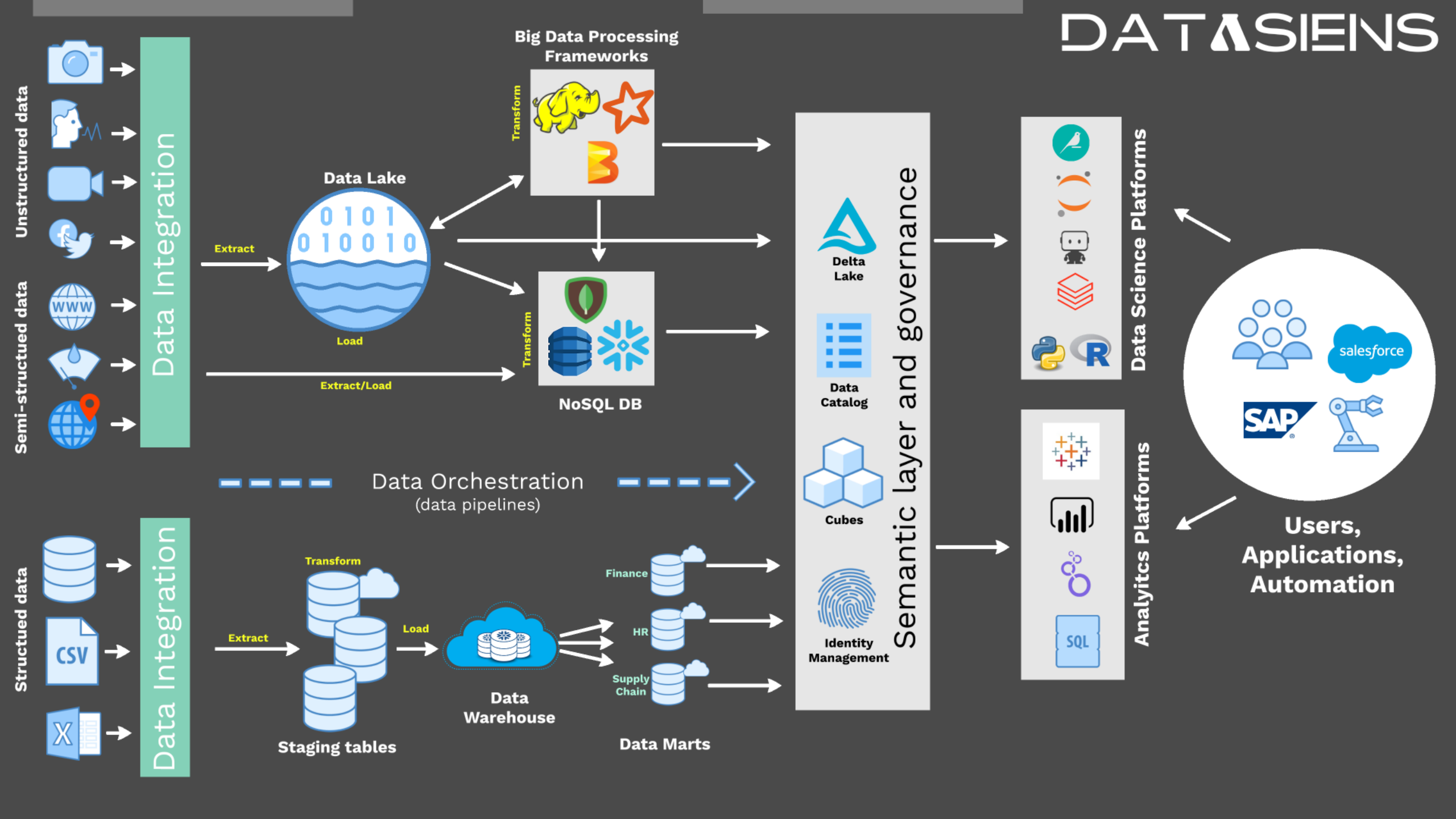
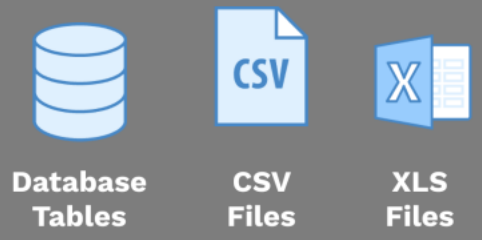

DATASIENS
Unstructured data
Semi-structued data
Structued data
Data Integration
Data Integration
Big Data Processing Frameworks
Transform
Data Lake
Extract
Load
Extract/Load
Transform
NoSQL DB
Data Orchestration
(data pipelines)
Transform
Extract
Load
Staging tables
Data Warehouse
Finance
HR
Supply Chain
Data Marts
Delta Lake
Data Catalog
Cubes
Identity Management
Semantic layer and governance
Data Science Platforms
Analyitcs Platforms
SQL
salesforce
SAP
Users, Applications, Automation

Structured data

DATASIENS

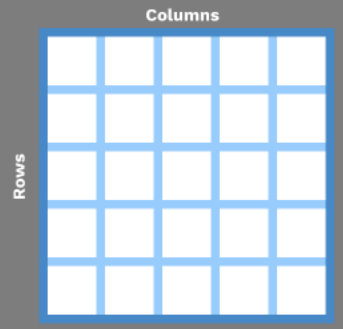
- **Only ~20% of the data in an organisation is structured data.**
- **However constitutes ~80% of the data science use cases.**

Common sources/types of structured data



- Staging Tables
  - Data cleansing
  - Join multiple sources
  - Enrichment
- Data warehouse
  - Hosts the single source of truth for the entire org.
- Data marts
  - Data extracted and curated for specific use cases.

Data is organized in rows and columns

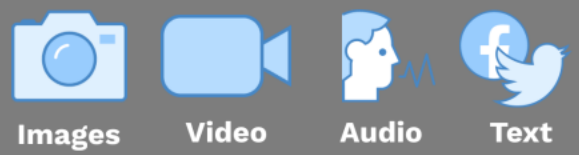


Unstructured data

DATASIENS

~80% - 90% of the data in an organization is unstructured

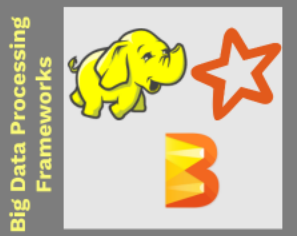
Common Types of unstructured data



Best and most economical storage

Data Lake

Pre-processing and processing



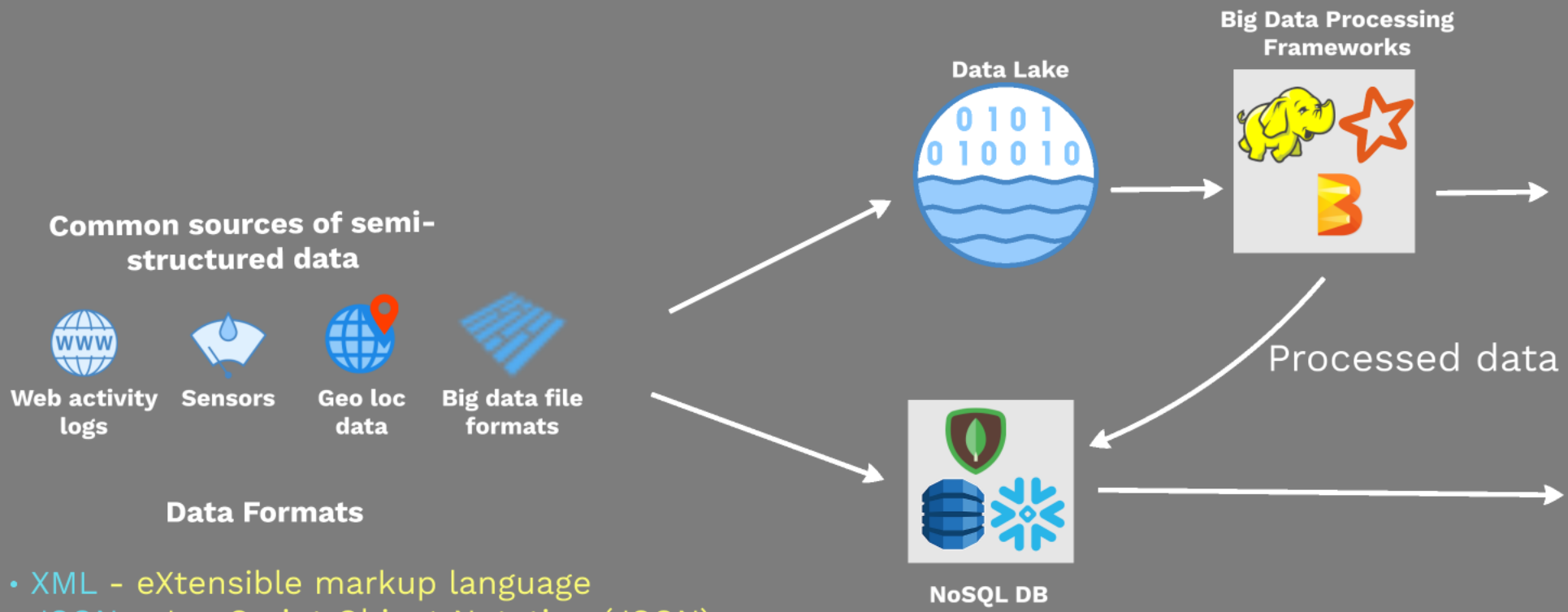
- Scaling images
- Text from audio
- Video annotation
- Tokenization for text

- Storing metadata and location of thumbnails and image in data lake

Common data science use cases

- Computer vision
- Speech to text
- Natural language processing (NLP)

Semi structured data



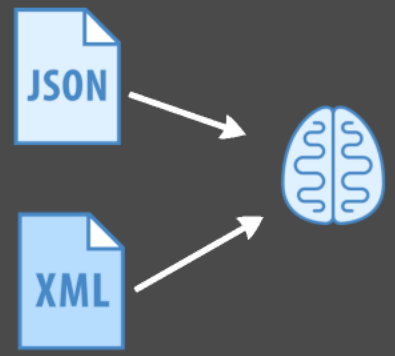
Data Formats

- XML - eXtensible markup language
- JSON - JavaScript Object Notation (JSON)
- Big data file formats like ORC, Parquet and Avro.

- Semi-structured data can be loaded either into a data lake or a NoSQL database.
- The choice depends on
  - data volume
  - amount of processing required
  - existing data engineering skill set
- In some cases data is processed by big data frameworks and the NoSQL database is used as a serving layer.

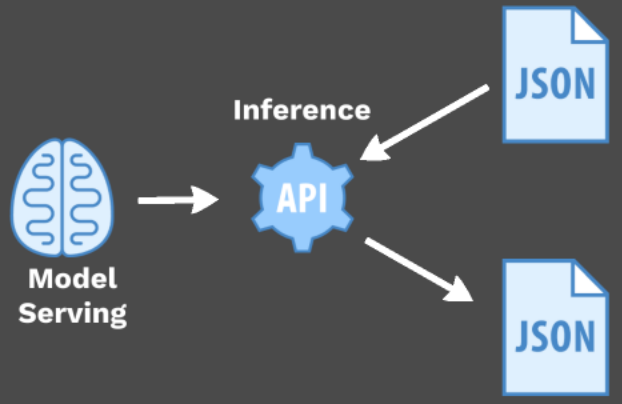
Machine learning with semi-structured data?

DATASIENS



Semi-structured data is never used to train models directly.

Semi-structured data is first flattened and then used for model training.



Model serving API's generally accept semi-structured data and also output data in the same format