Software and Processes for Data Management

Yapi Donatien Achou

Semcon

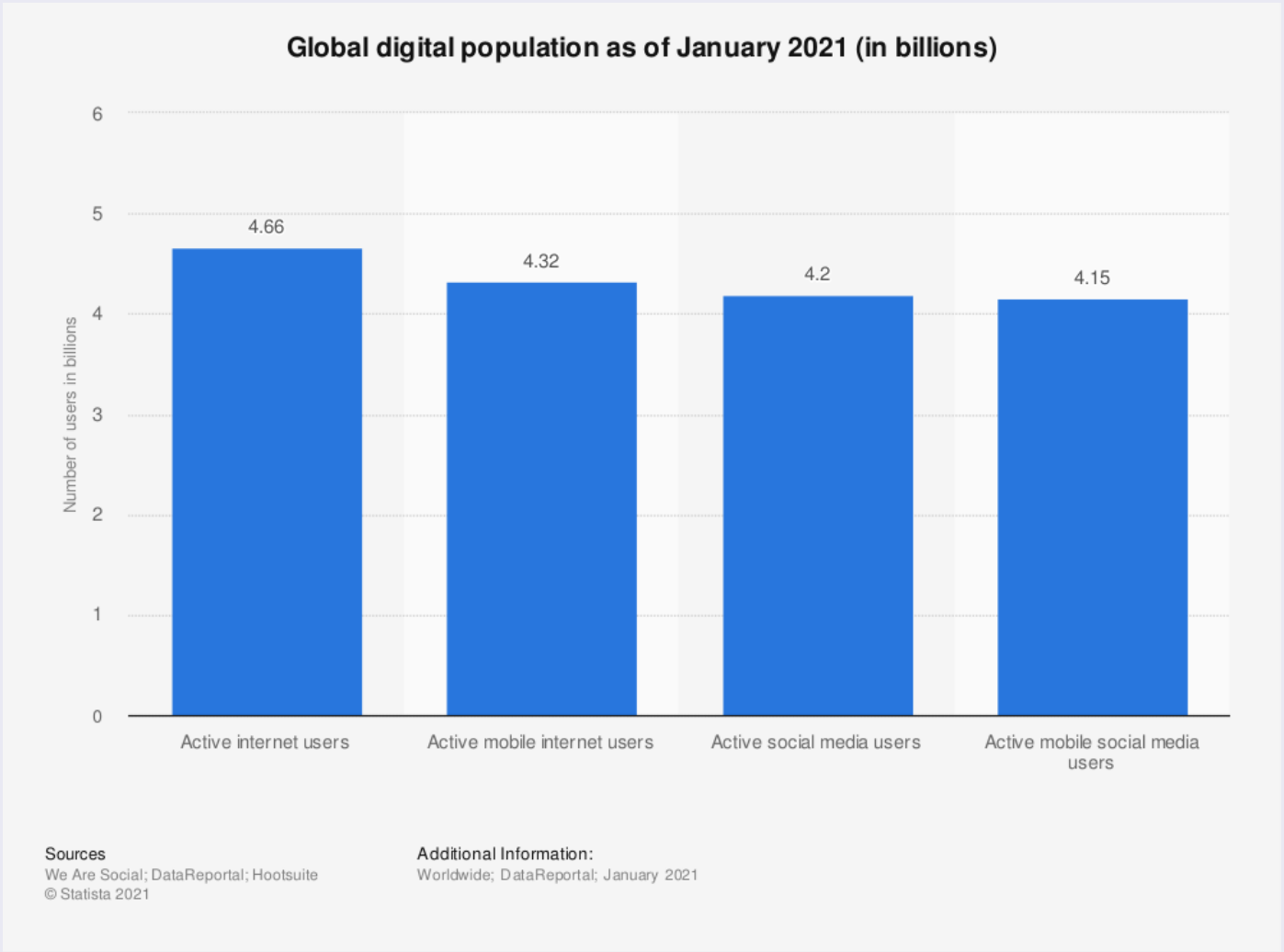
yapi-donatien.achou@semcon.com

August 20, 2021

Overview

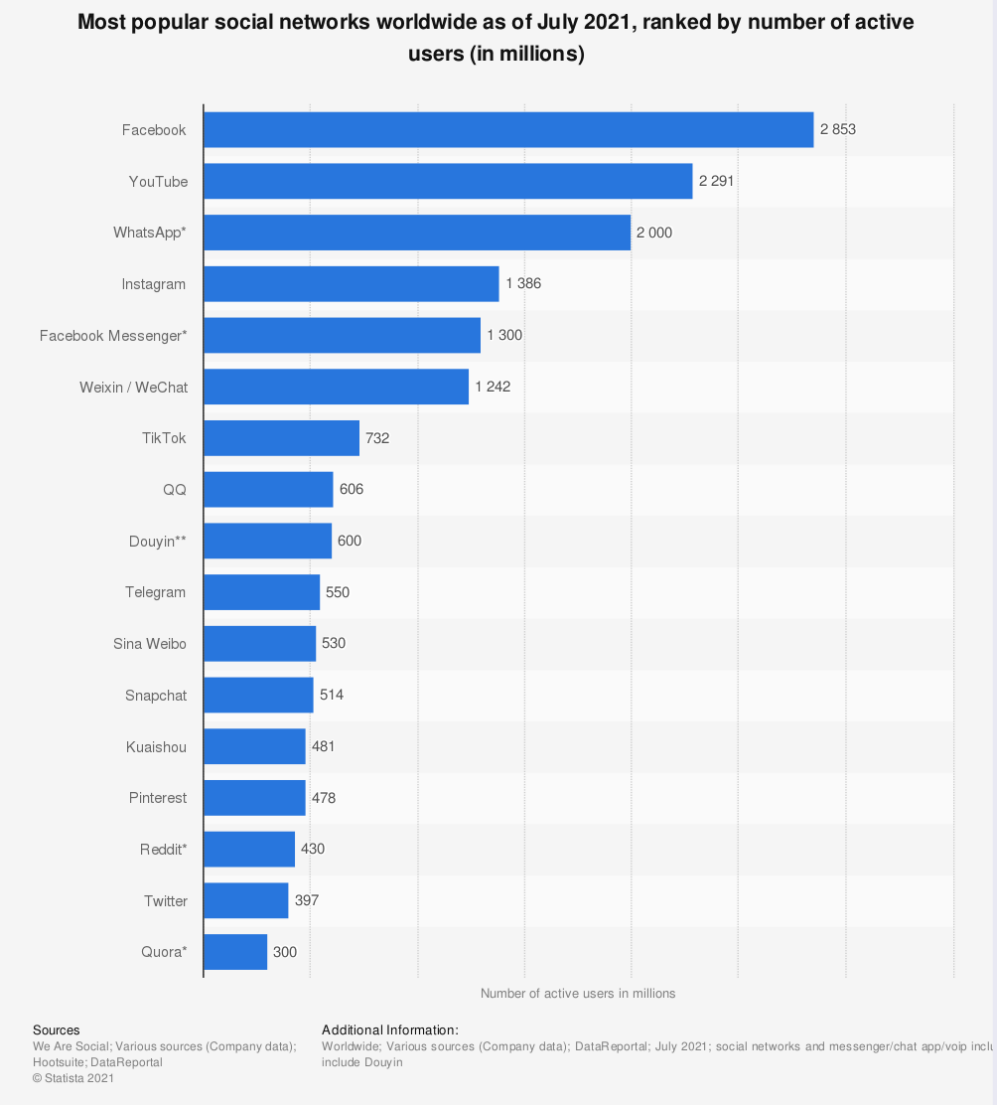
Data source

Internet Users in 2021



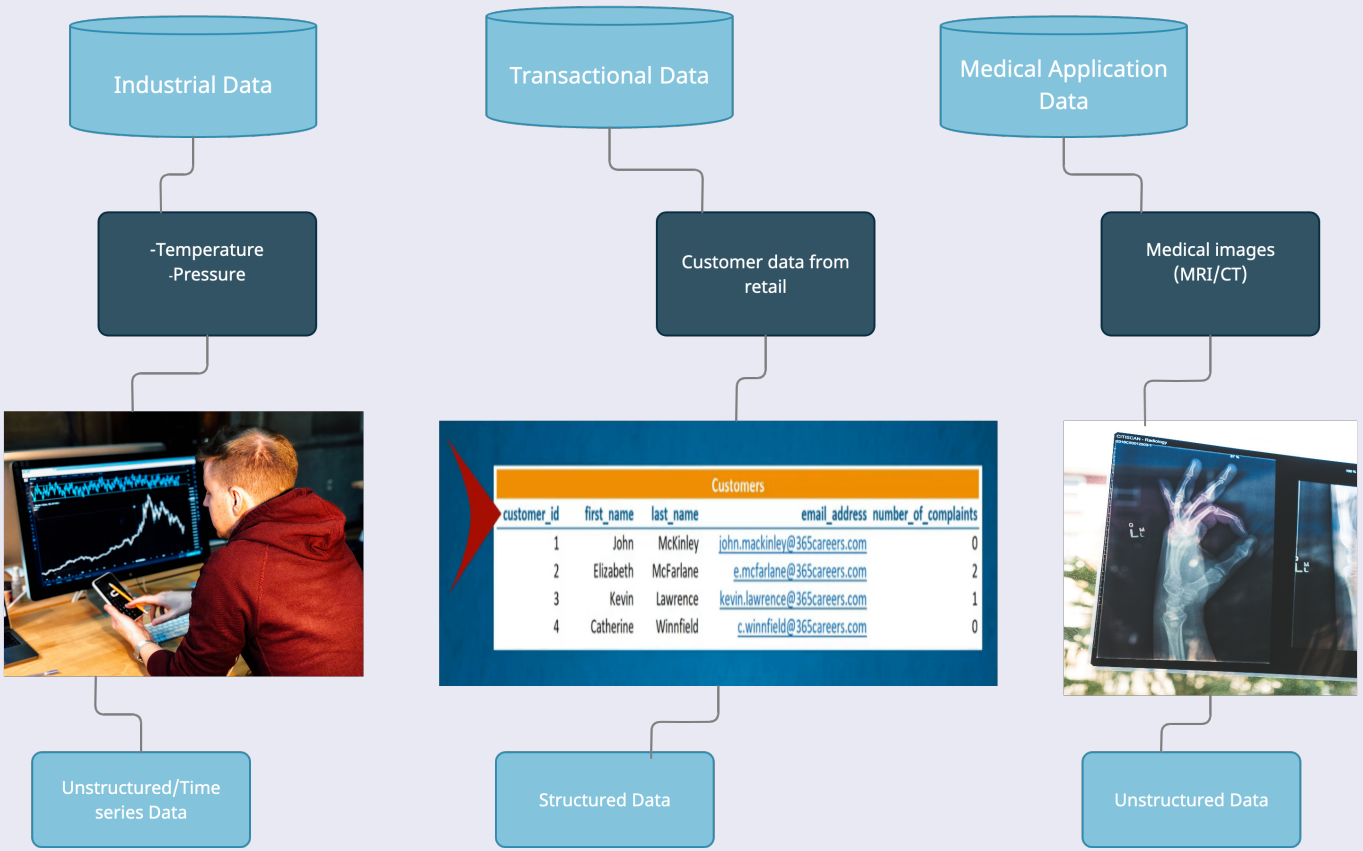
Data source

Social media



Data source

Other data sources



Data Storage

Structured vs unstructured

The type of data dictates how it is stored

Structured data storage

Structured data are typically stored into relational databases

Data Storage

Unstructured data storage

Unstructured data are typically stored in non relational data bases such as time series databases, datalakes

Data Ingestion

Ingestion

Moving data from one or more sources to one or more destinations

Challenges

- From one source/destination to thousand of sources/destinations
- different data types (structures, unstructured)

Example

- Moving Iot
- different data types (structures, unstructured)

Data Ingestion

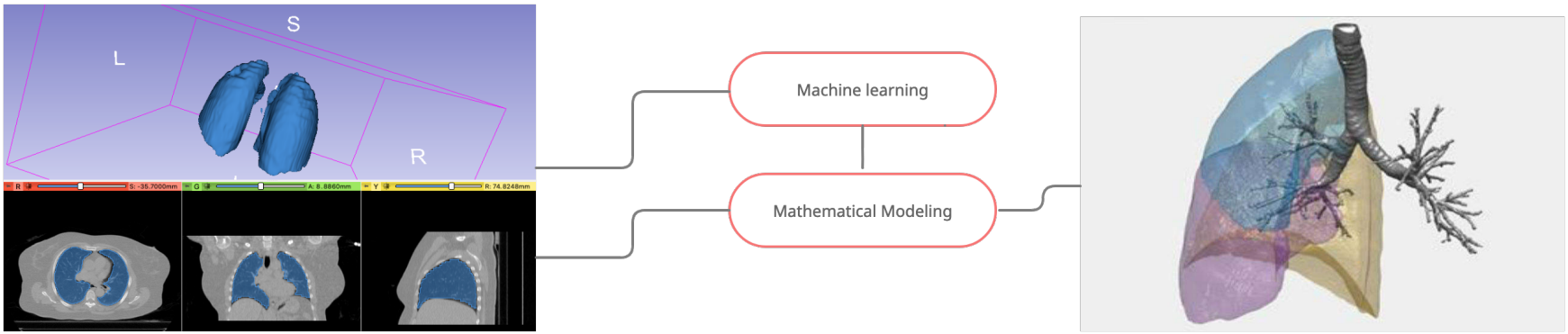
Data pipeline

Data pipeline can be define as a set of processes that allow you to orchestrate and monitore data movements

Software framework

- Apache Airflow

Opportunities in healthcare

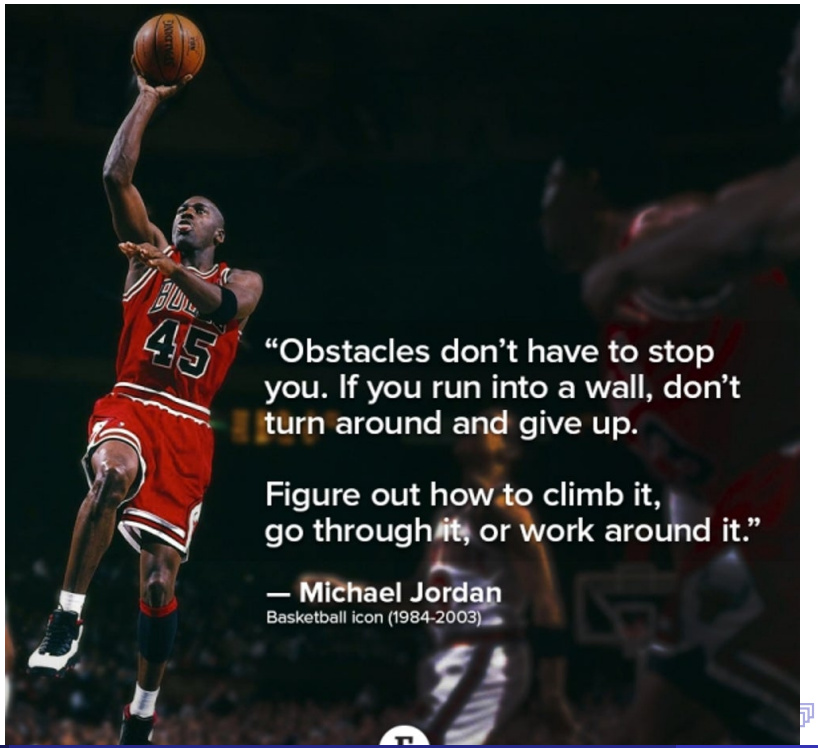


Accurately predict Cancer tumour movement by using

- Medical images (CT, 4DCT, MRI, . . .)
- External sensor data collected from a patient

Challenges

- Thousand of data sources to extract information from
- Data quality issue
- From Gigabits to Terabits of data to process and analyse
- Data privacy concern
- Data quality concern
- A **scalable**, **secure** , **well governed** and **reliable** data management platform



Data pipeline

The data journey from its source to the value that it brings:

- Data ingestion ✔
- Data storing ✔
- Data quality assessment ✔
- data modelling ✔
- Data transformation
- Data analytics

While tracking Data Lineage: Set of transformation that data goes through

Data pipeline

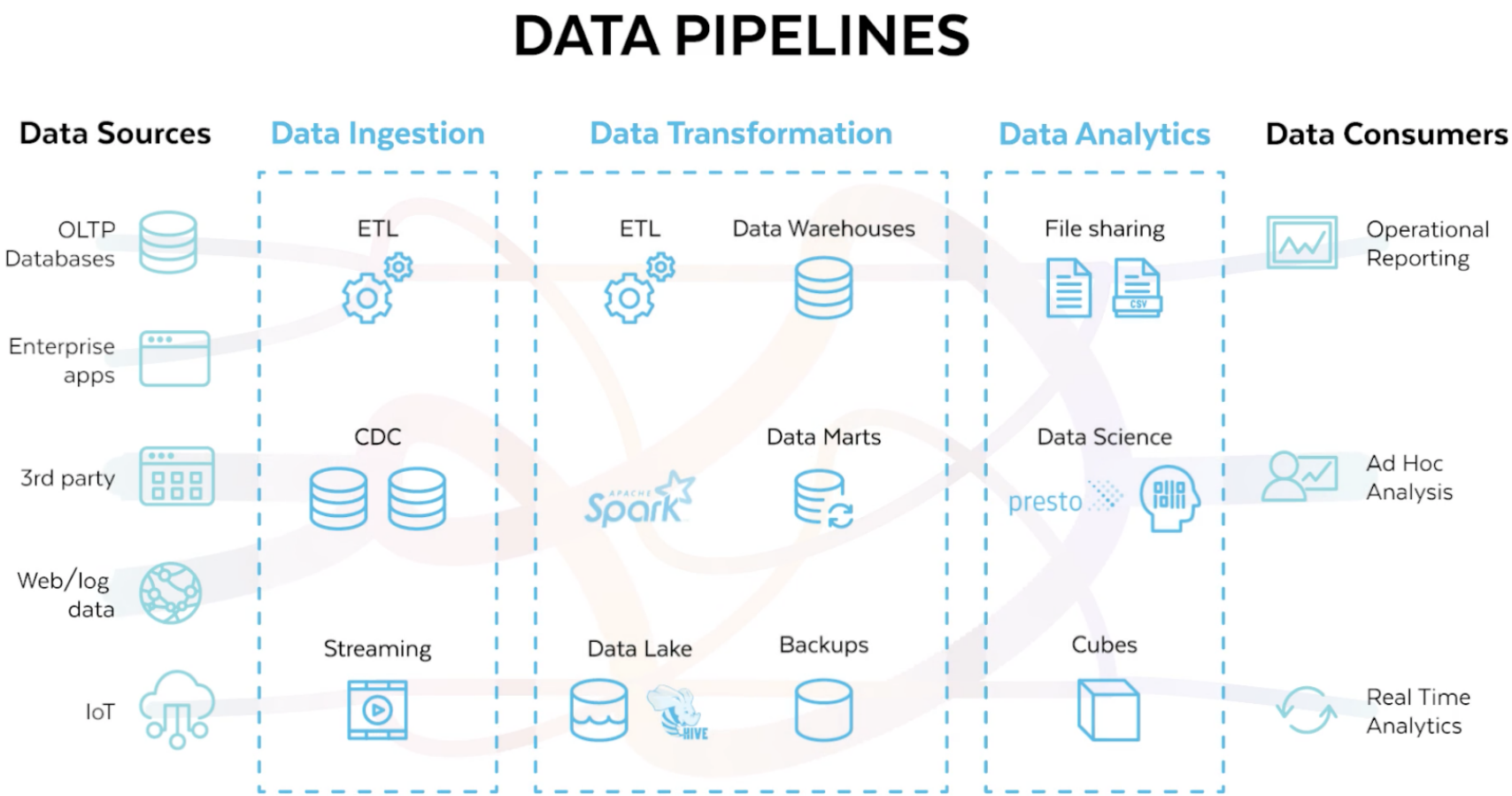


Figure: Cortesie https://www.snowflake.com/guides/etl-pipeline

Tools

Data ingestion

Data storage

Data science

Machine learning pipeline