

Software and Processes for Data Management

Yapi Donatien Achou

Semcon

yapi-donatien.achou@semcon.com

August 21, 2021

Overview

1 Data sources

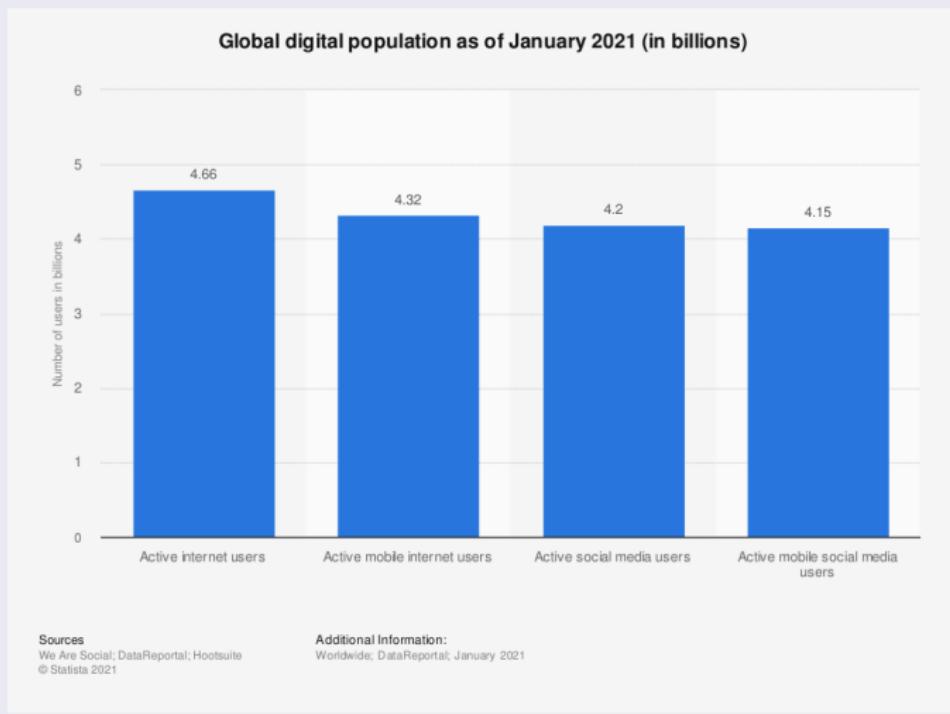
2 Data storage

3 Challenges

4 Workflow and tool

Data source

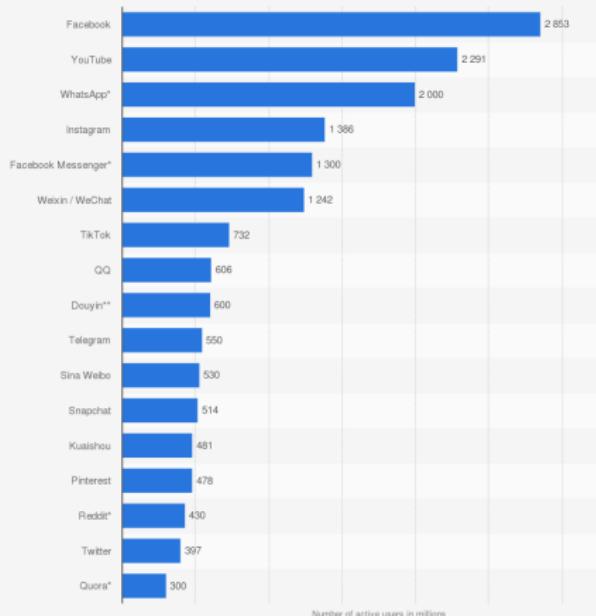
Internet Users in 2021



Data source

Social media

Most popular social networks worldwide as of July 2021, ranked by number of active users (in millions)



Sources:

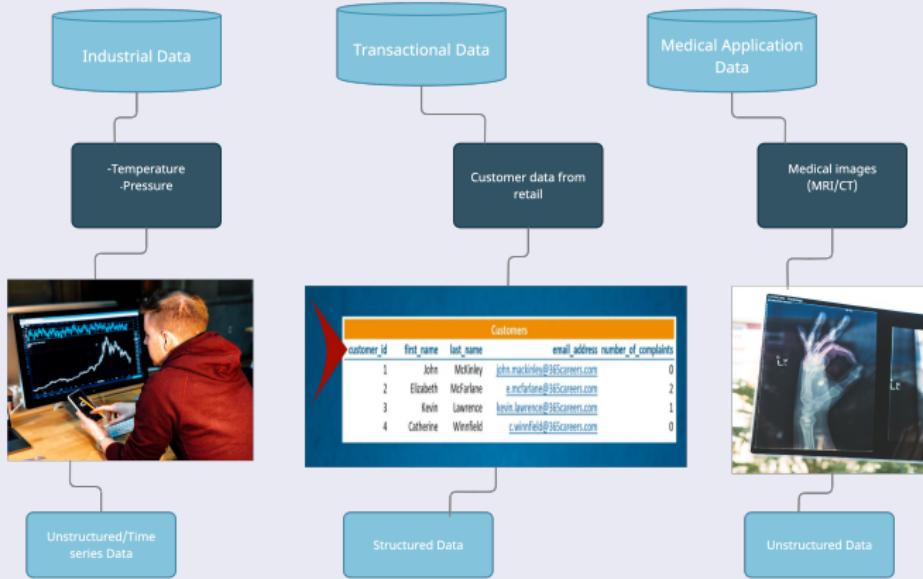
We Are Social; Various sources (Company data);
Hostsite: DataReportal
© Statista 2021

Additional Information:

Worldwide; Various sources (Company data); DataReportal, July 2021; social networks and messenger/chat app/voip incl.
include Douyin

Data source

Other data sources



Example

- ① Read time series data with [pandas](#) and plot with [matplotlib](#)
- ② Read transactional data from a series of json files and set them in a tabular format with [pandas](#)
- ③ read medical images with [matplotlib](#)/ [pydicom](#)

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



Unstructured data storage

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



Microsoft Azure
Data Lake

*



InfluxDB

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 lot
- different data types (structures, unstructured)

Data Ingestion

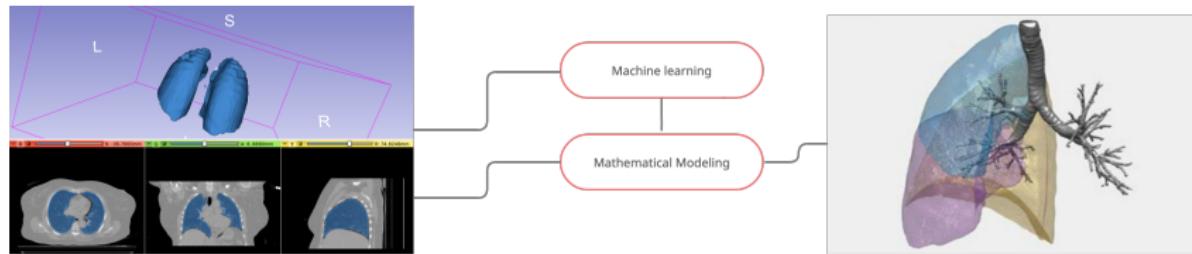
Data pipeline

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

Software framework

- Apache Airflow

Opportunities in healthcare

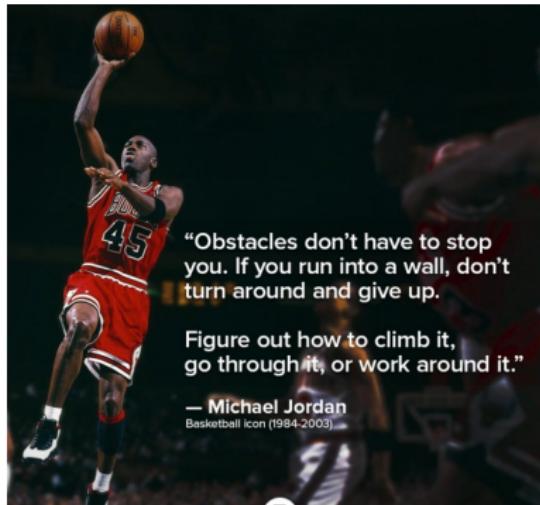


Accurately predict Cancer tumour movement by using

- 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

DATA PIPELINES

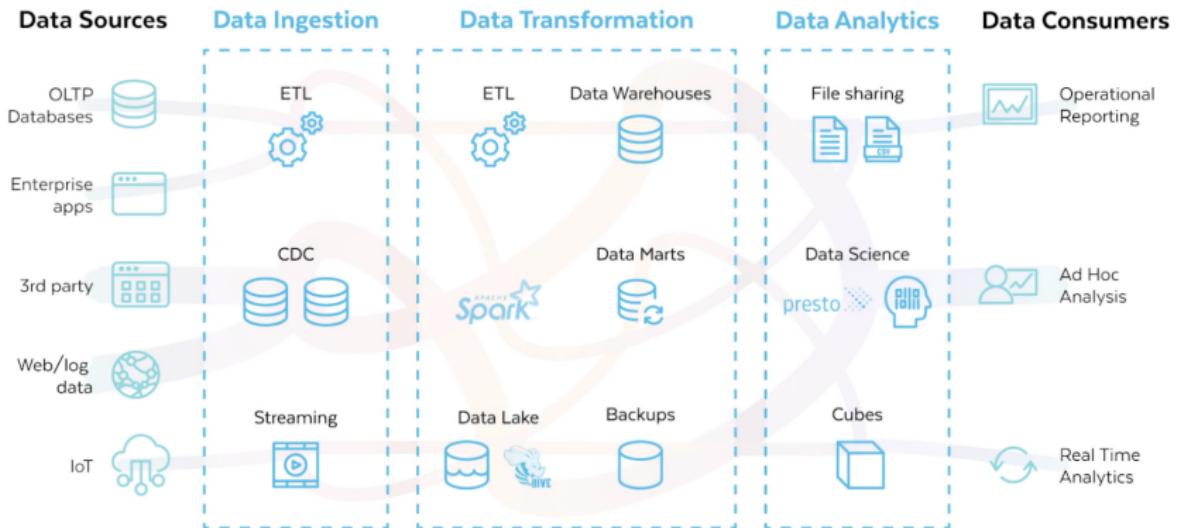


Figure: Cortesie <https://www.snowflake.com/guides/etl-pipeline>

Tools

Data ingestion



Data storage



Machine learning pipeline

