

# Software and Processes for Data Management

Yapi Donatien Achou

Semcon

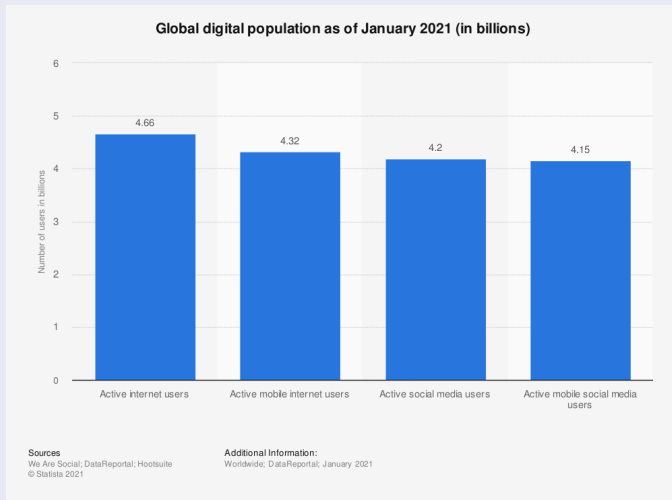
*yapi-donatien.achou@semcon.com*

August 23, 2021

# Overview

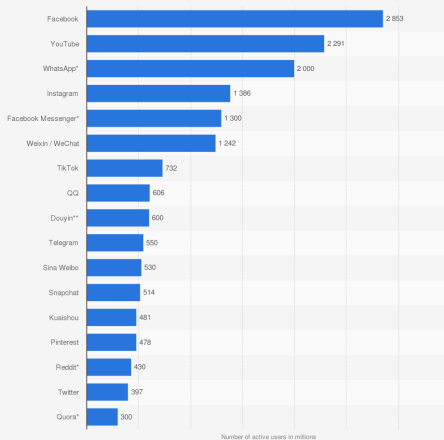
- 1 Data sources
- 2 Data quality
- 3 Data storage
- 4 Data Ingestion

## Internet Users in 2021



## Social media

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



**Sources**

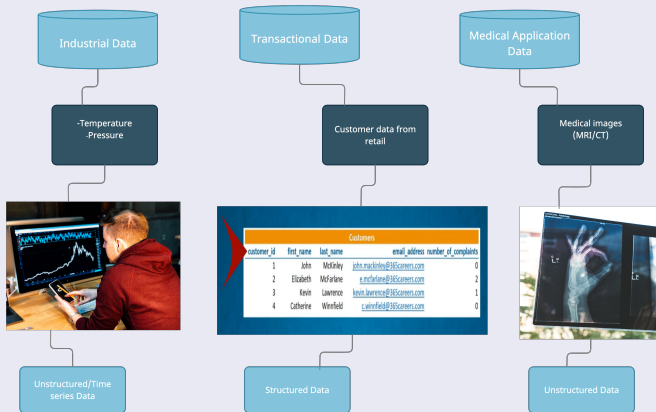
We Are Social; Various sources (Company data);  
Hochschule; DataReportal  
© Statista 2021

**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](#)

# Data quality

# 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



# 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

### **Extract Tranform Load (ETL) pipeline**

- Moving transactional data into a SQL database
- Moving time series data into InfluxDB

## 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

## DATA PIPELINES

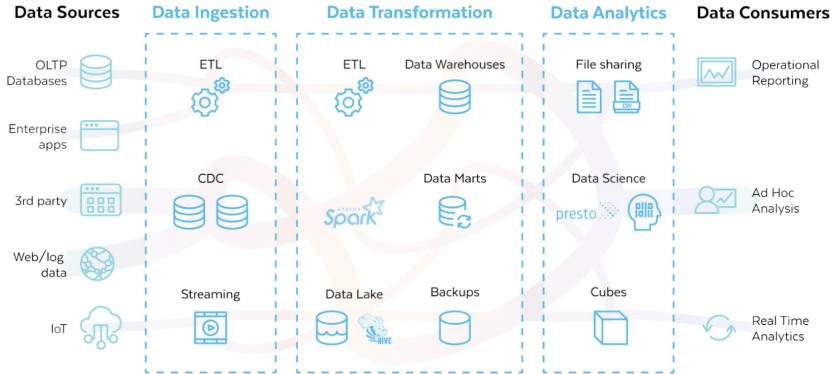


Figure: Cortesie <https://www.snowflake.com/guides/etl-pipeline>

## Machine learning pipeline

