Software and Processes for Data Management

Yapi Donatien Achou

Semcon

yapi-donatien.achou@semcon.com

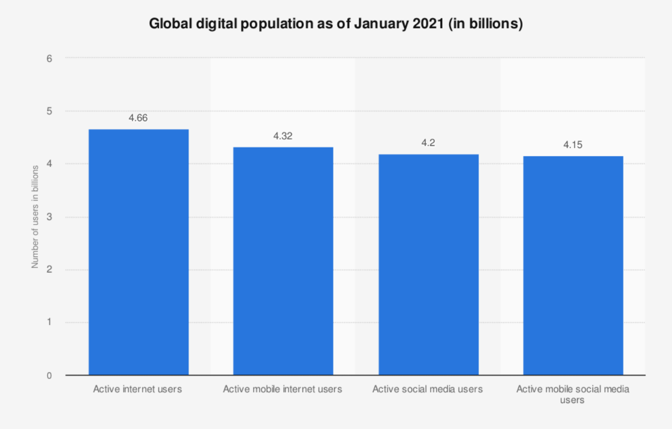
August 25, 2021

Overview

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

Data source

Internet Users in 2021



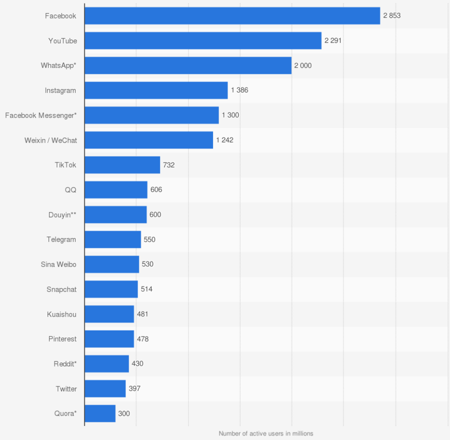
Sources
We Are Social; DataReportal; Hootsuite
© Statista 2021

Additional Information:
Worldwide; DataReportal; January 2021

Data source

Social media

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

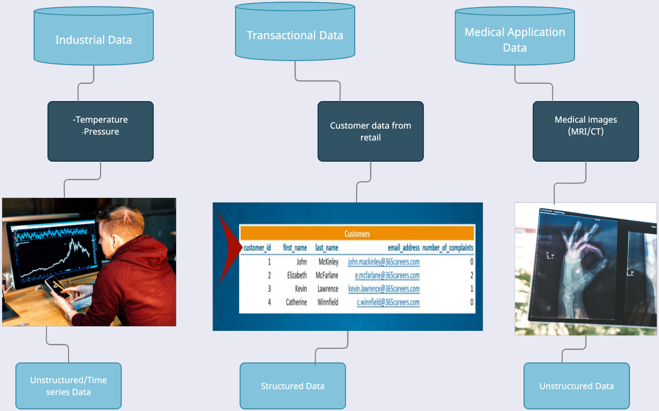


Sources
We Are Social; Various sources (Company data);
Hootsuite; DataReportal
© Statista 2021

Additional Information:
Worldwide; Various sources (Company data); DataReportal; July 2021; social networks and messenger/chat app/voip inclu
include Douyin

Data source

Other data sources



Customers

customer_id	first_name	last_name	email_address	number_of_complaints
1	John	McKinley	john.mackinley@365careers.com	0
2	Elizabeth	McFarlane	e.mcfarlane@365careers.com	2
3	Kevin	Lawrence	kevin.lawrence@365careers.com	1
4	Catherine	Winnfield	c.winnfield@365careers.com	0

Data source

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 quality dimensions	
— Accessibility — Accuracy — Actuality — Auditability — Authorization — Completeness — Conformance to domain models and conceptual model/ Semantic consistency — Conformance to metadata /schema — Consistency — Duplicates	— Format/structural consistency — Integrity — Interoperability — Identifiers — Precision — Processability — Resolution — Timeliness — Traceability — Uncertainty

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)

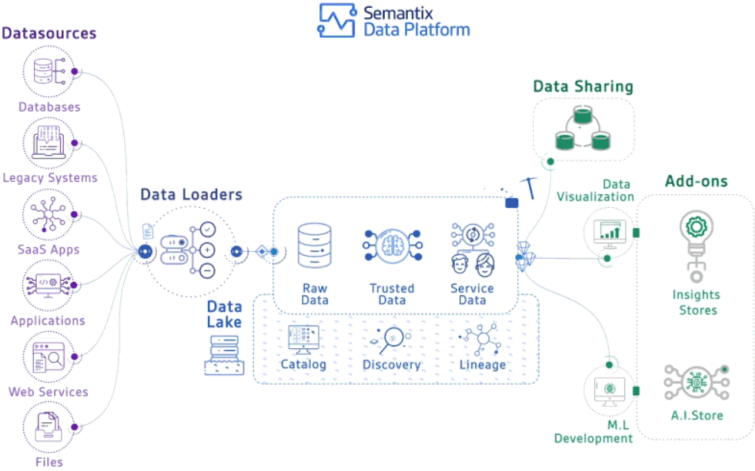
Example

Extract Tranform Load (ETL) pipeline

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

Data Ingestion

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



Data pipeline

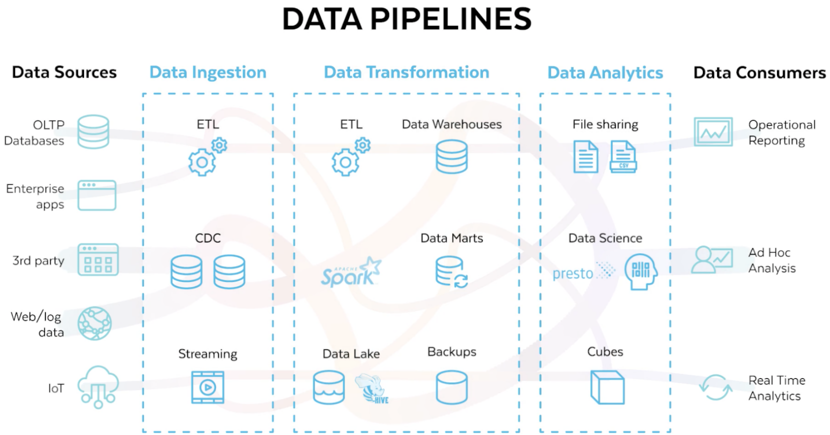
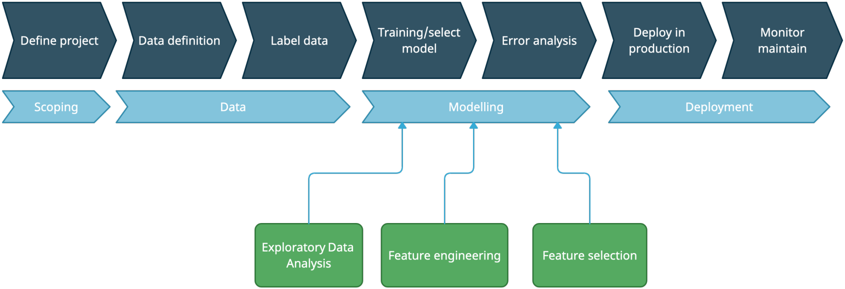


Figure: Cortesie https://www.snowflake.com/guides/etl-pipeline

Data science

Machine learning pipeline


Define project
Data definition
Label data
Training/select model
Error analysis
Deploy in production
Monitor maintain
Scoping
Data
Modelling
Deployment
Exploratory Data Analysis
Feature engineering
Feature selection