Lecture 2: Introduction to the Big Data

Big Data System Design

Table of Contents

Part 1

INTRODUCTION TO BIG DATA

Intro to Big Data

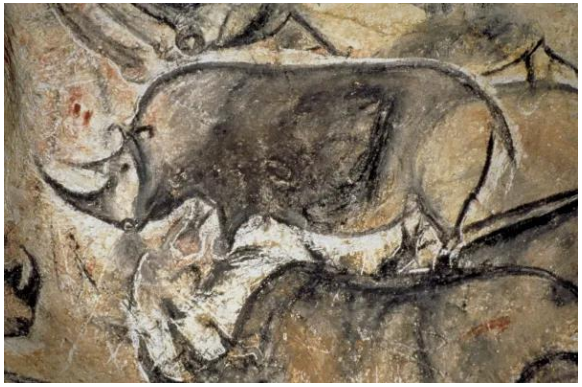
- What is data?
  - Since birth, we are surrounded with data !





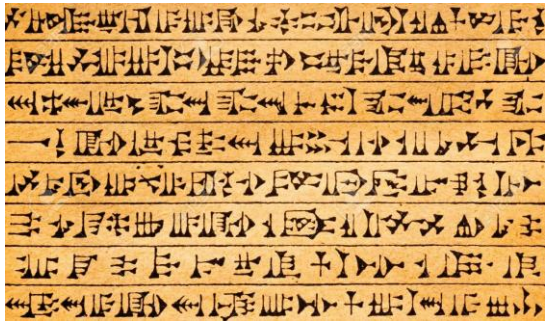

Intro to Big Data

- What is data?



Intro to Big Data

- What is data?
  - From the advent of written language, human observations have been recorded



Intro to Big Data

- What is data?
  - From the advent of written language, human observations have been recorded

Intro to Big Data

- What is data?
  - The advent of computer technologies in 1950s, data most commonly refers to information that is transmitted or stored electronically
  - The electronic sensors has additionally contributed to the volume and richness of recorded data

Intro to Big Data

- What is data?

Data signifies the documented result of quantified or observed phenomena

Intro to Big Data

- Types of data

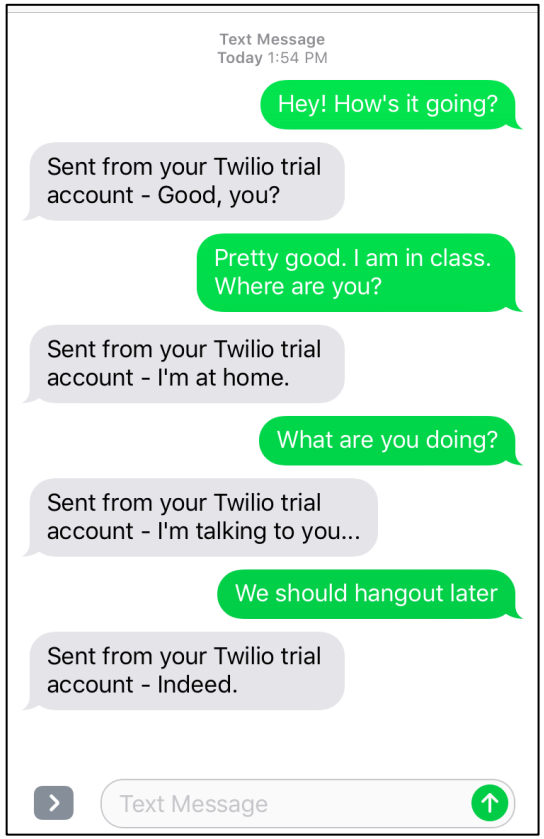
CUSTOMER

CUSTOMER_ID	LAST_NAME	FIRST_NAME	STREET	CITY	ZIP_CODE	COUNTRY
10302	Boucher	Leo	54, rue Royale	Nantes	44000	France
11244	Smith	Laurent	8489 Strong St	Las Vegas	83030	USA
11405	Han	James	636 St Kilda Road	Sydney	3004	Australia
11993	Mueller	Tomas	Berliner Weg 15	Tamm	71732	Germany
12111	Carter	Nataly	5 Tomahawk	Los Angeles	90006	USA
14121	Cortez	Nola	Av. Grande, 86	Madrid	28034	Spain
14400	Brown	Frank	165 S 7th St	Chester	33134	USA
14578	Wilson	Sarah	Seestreet #6101	Emory	1734	USA
14622	Jones	John	71 San Diego Ave	Arlington	69004	USA

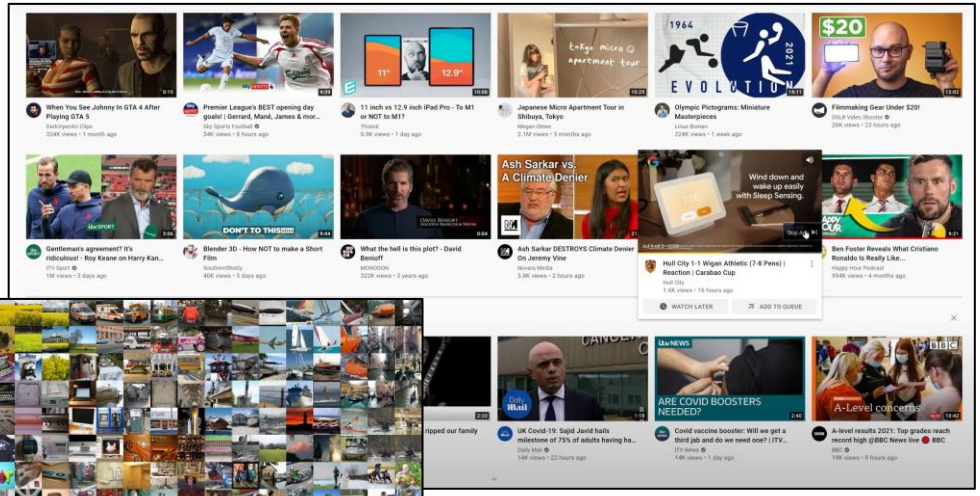
Source: https://www.michael-gramlich.com/what-is-structured-semi-structured-and-unstructured-data/

Intro to Big Data

- Types of data



Texts



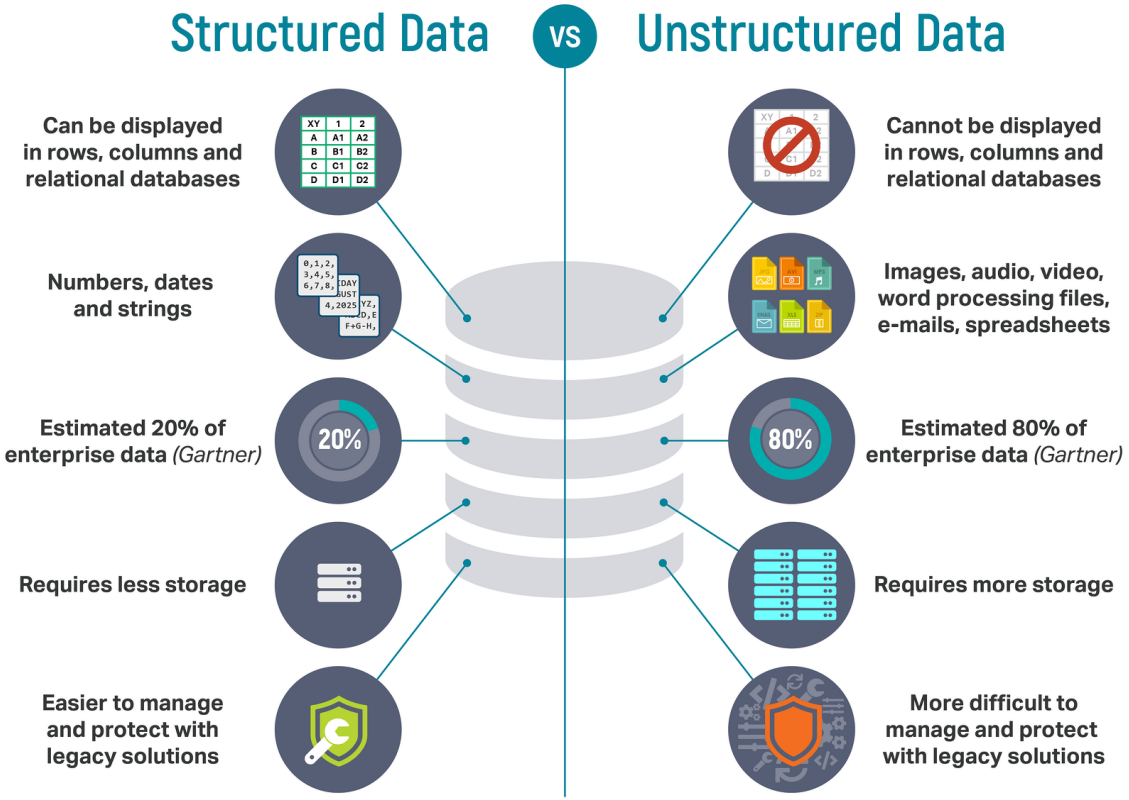
Videos

Images

Sensors

Intro to Big Data

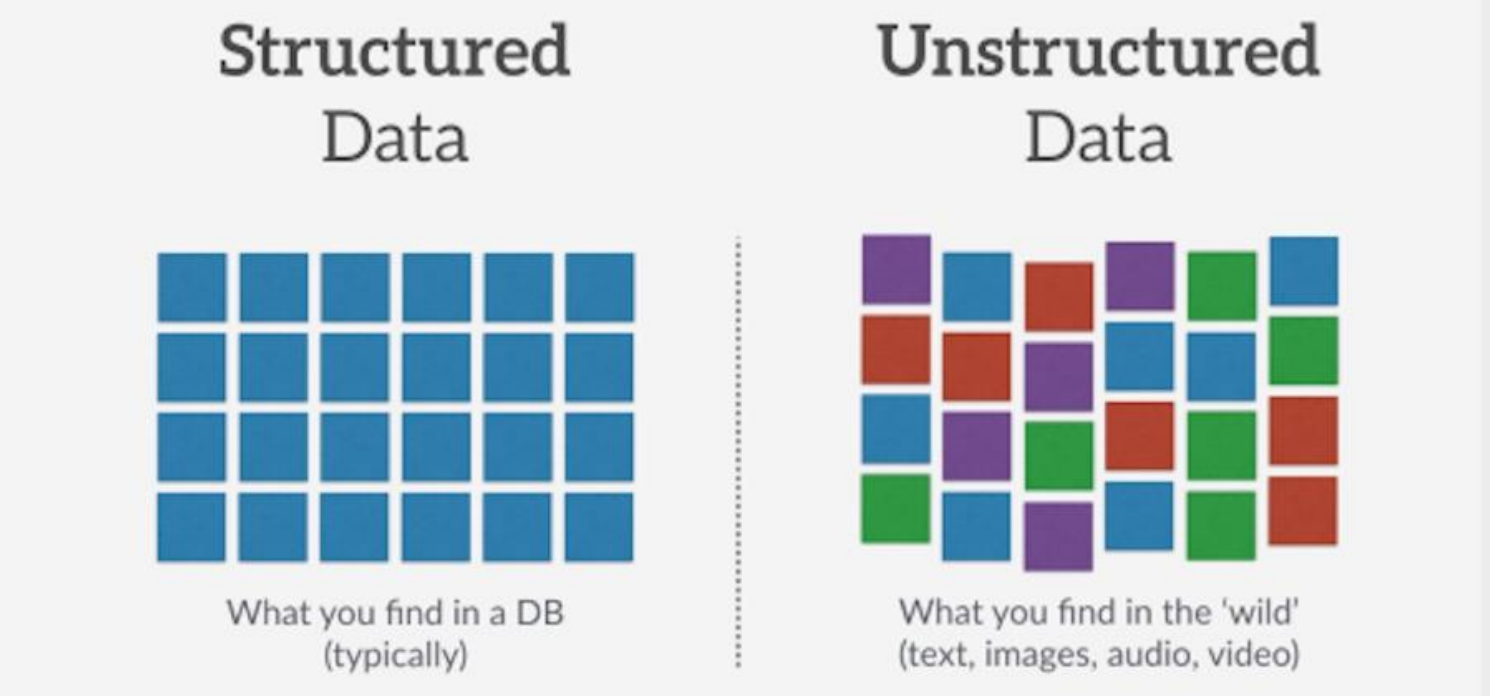
- Types of data
  - Structured data vs. unstructured data



Source: https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

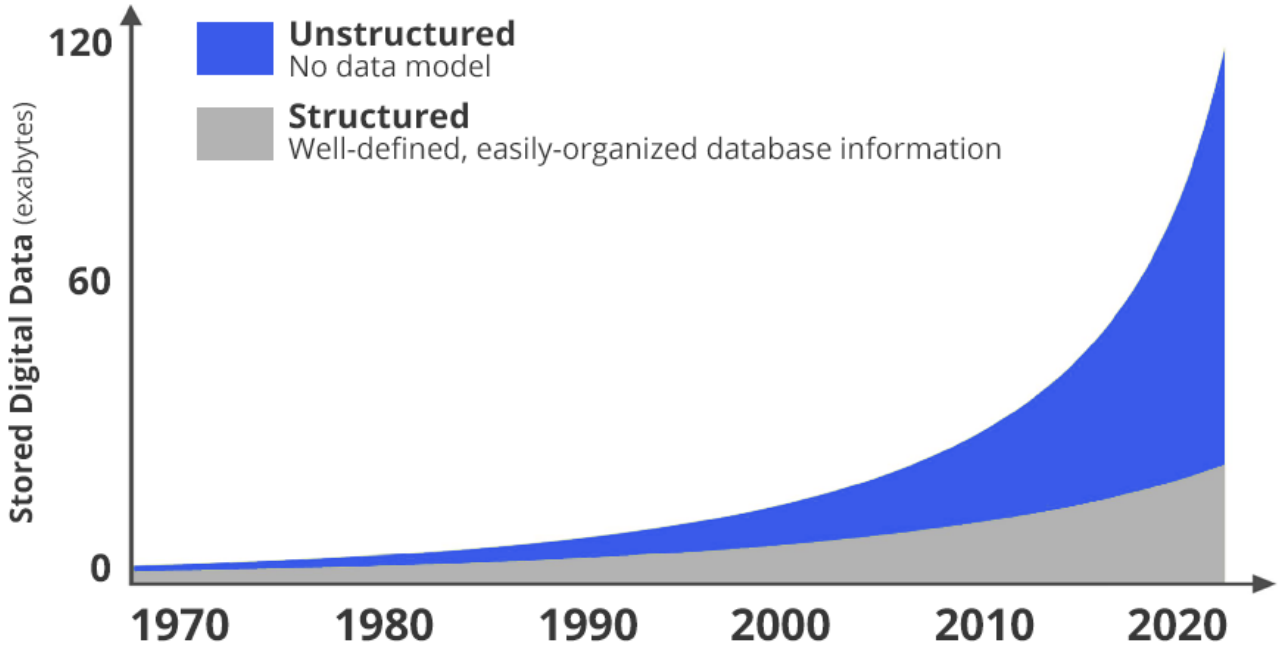
Intro to Big Data

- Types of data
  - Structured data vs. unstructured data



Intro to Big Data

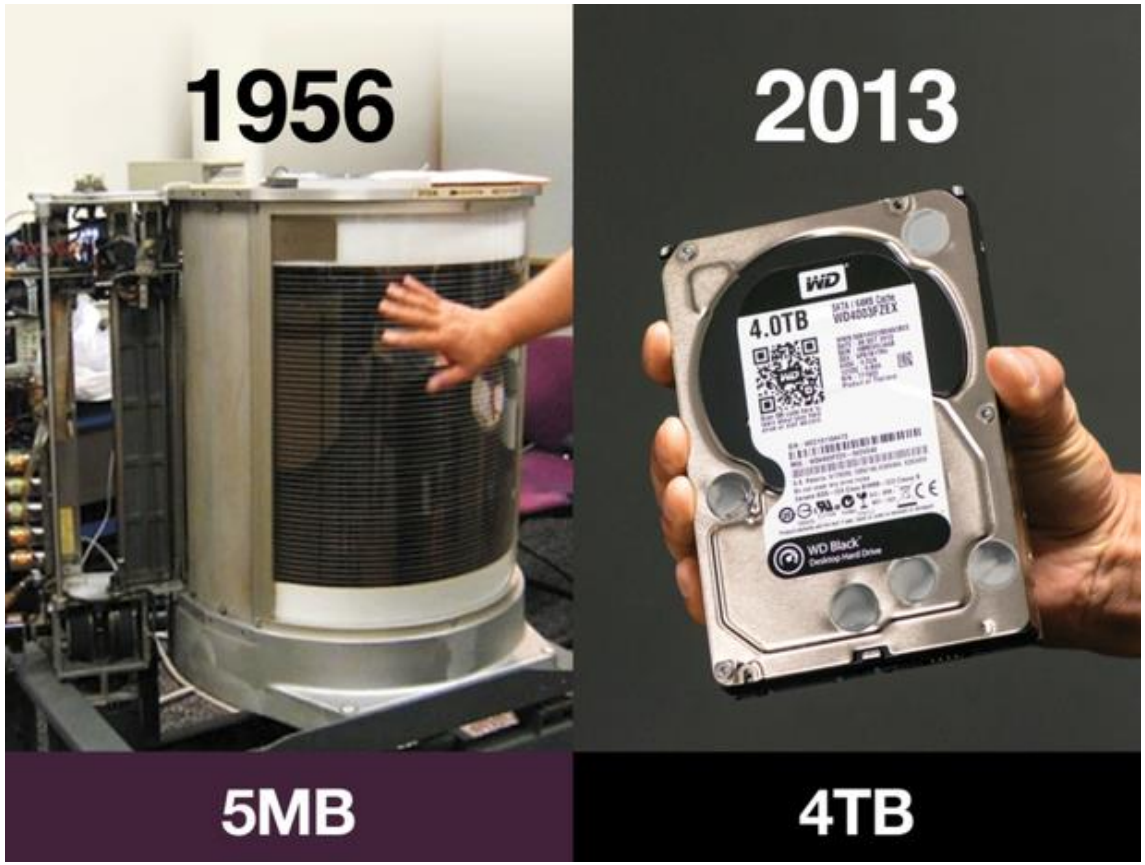
- Types of data
  - Structured data vs. unstructured data



Source: https://www.m-files.com/blog/what-is-structured-data-vs-unstructured-data/

Intro to Big Data

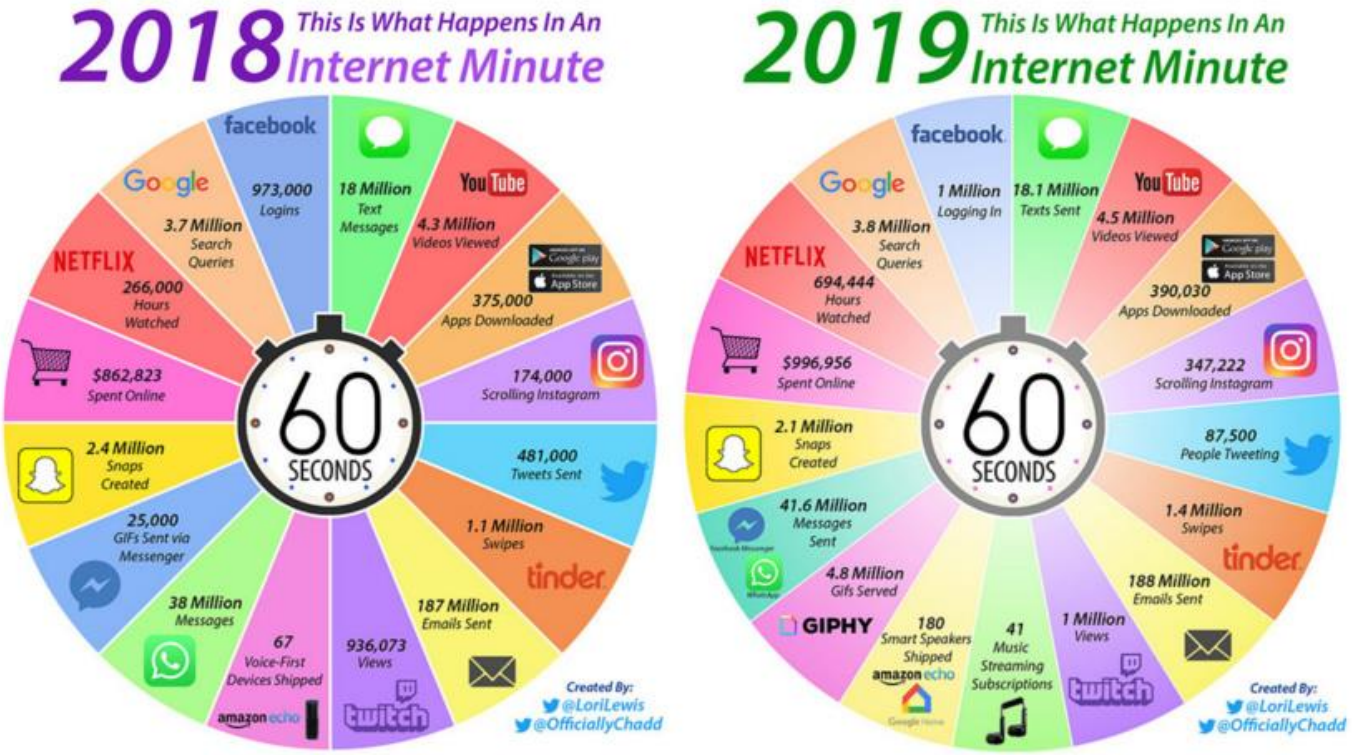
- Reason for growth
  - Rapid advance in computer hardware



Source: https://www.computerworld.com/article/2473980/data-storage-solutions-143723-storage-now-and-then.html

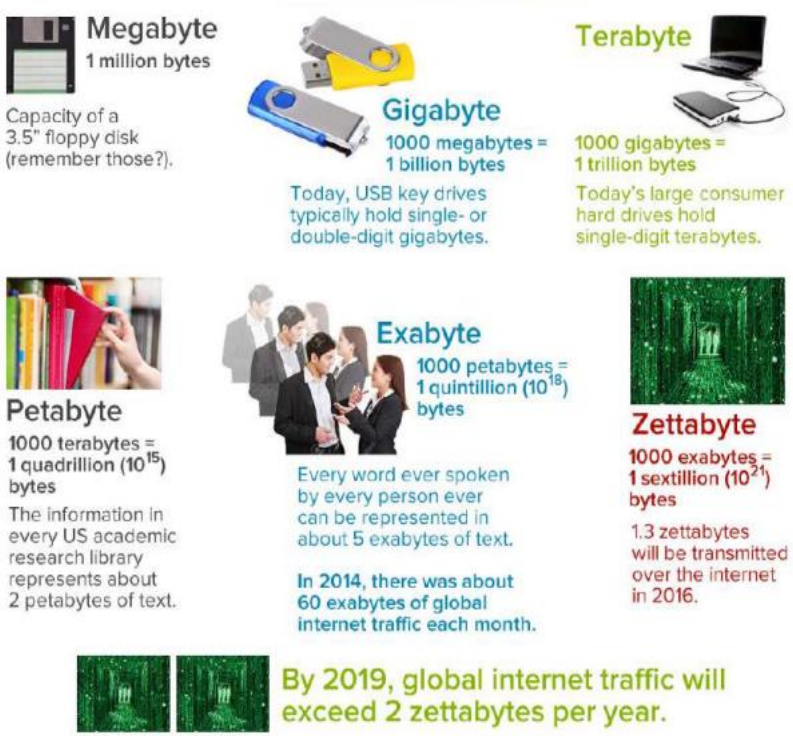
Intro to Big Data

- Reason for growth
  - Rapid advance in social networking



Intro to Big Data

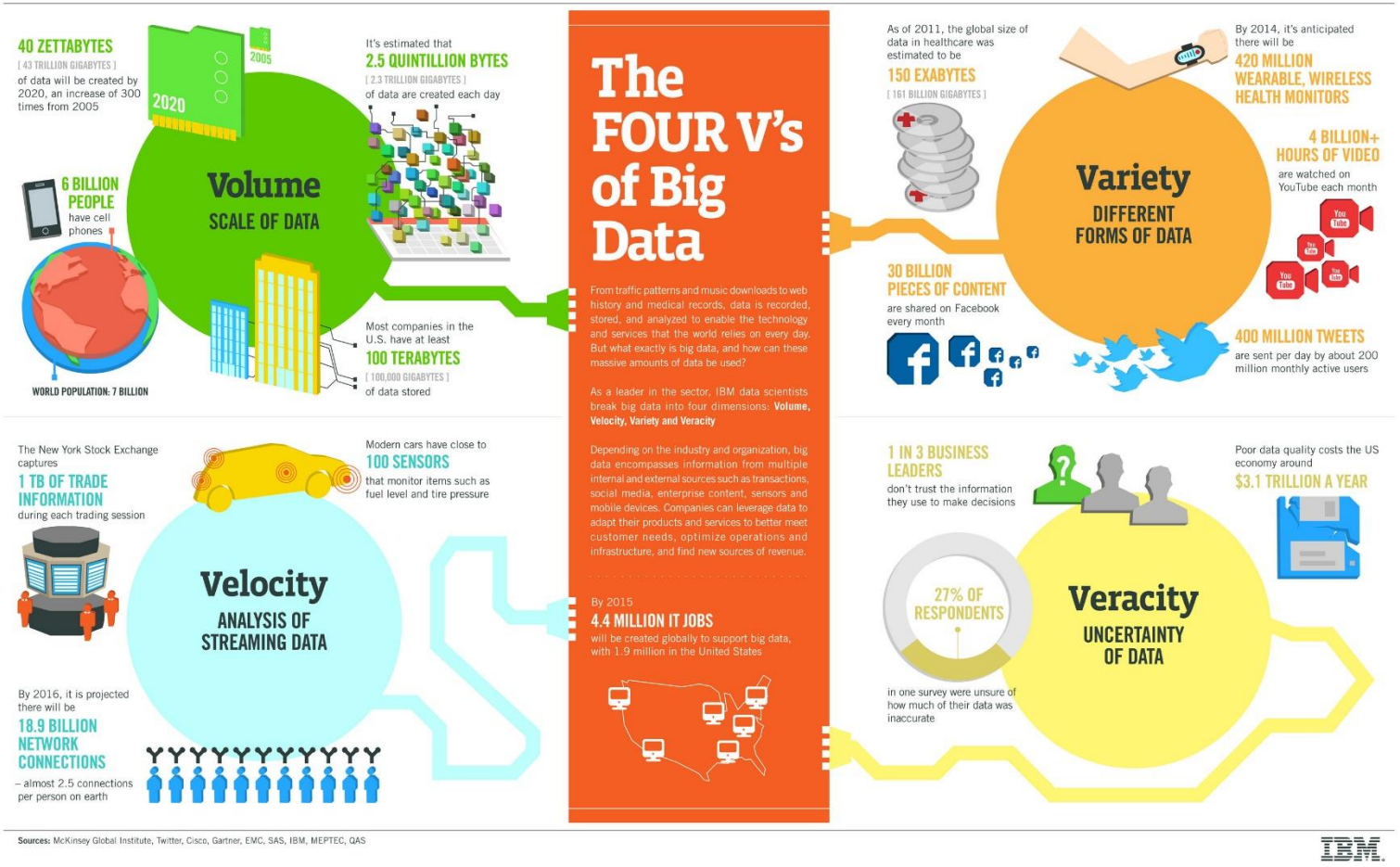
- What is Big Data?
  - According to Seagate, the volume of data generated worldwide will increase from 33 in 2018 to about 175 zettabytes in 2025



1 kilobyte	1,000
1 megabyte	1,000,000
1 gigabyte	1,000,000,000
1 terabyte	1,000,000,000,000
1 petabyte	1,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

Intro to Big Data

- What is Big Data?



Source: https://www.ibmbigdatahub.com/infographic/four-vs-big-data

Intro to Big Data

❖ Hospital Big Data Example

1. Volume

- Hospitals around the world generate a massive amount of data in the form of patient records and test results
- According to IBM, 2.314 Exabytes of medical data collected annually around the world

3. Velocity

- According to IBM, medical data is experiencing a 48 percent annual growth rate

2. Variety

- Hospital can collect medical records in variety form, such as structured and unstructured data
- It can be textual information, excel or images (e.g., X-Ray images)

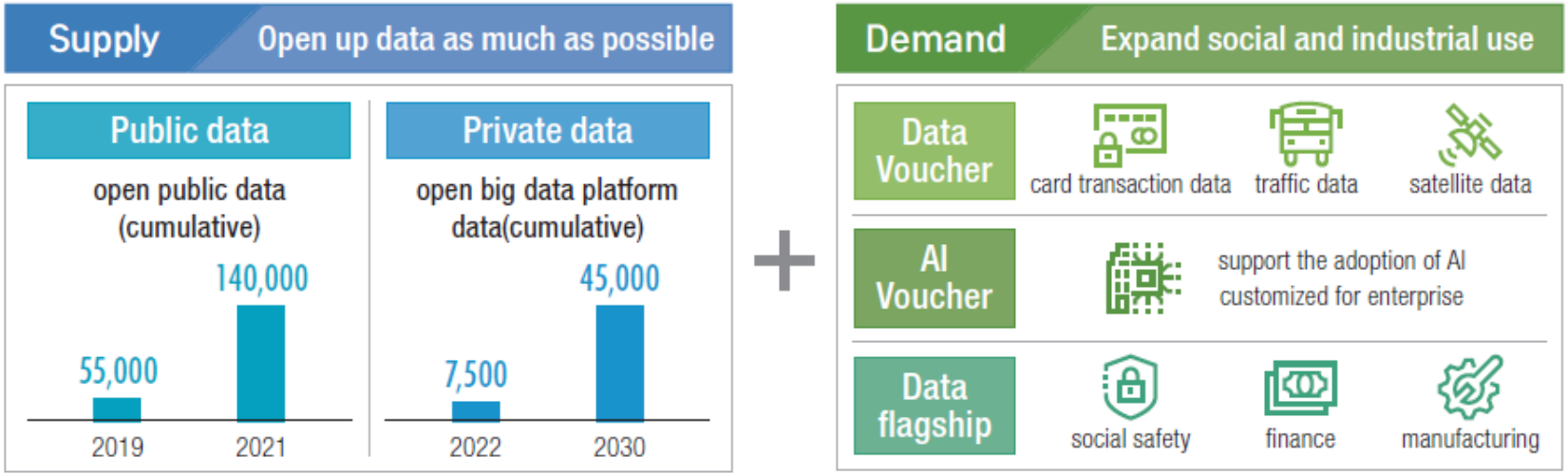
4. Veracity

- Since its healthcare field, the accuracy and trustworthiness of the data must be very high
- High accuracy in medical examination, prediction of disease

Intro to Big Data

❖ Big Data in South Korea

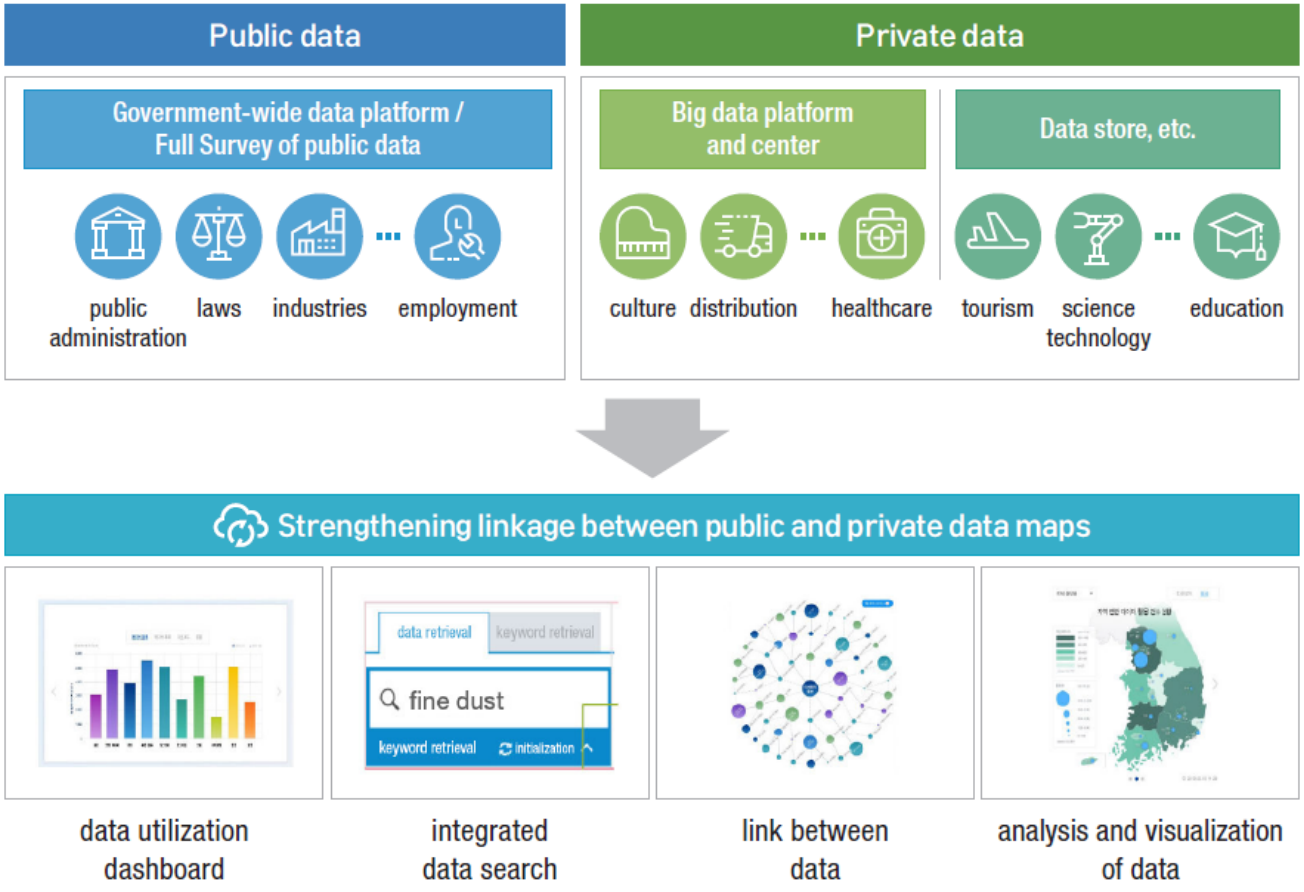
- Promotion of Opening Up Data and Reuse
  - Expanding the construction of AI learning data and securing of AI development infrastructure through the 'AI Hub' platform supply



Intro to Big Data

❖ Big Data in South Korea

- Strengthening Linkage between Public/Private Data Map



Part 2

WHAT IS BIG DATA ANALYTICS

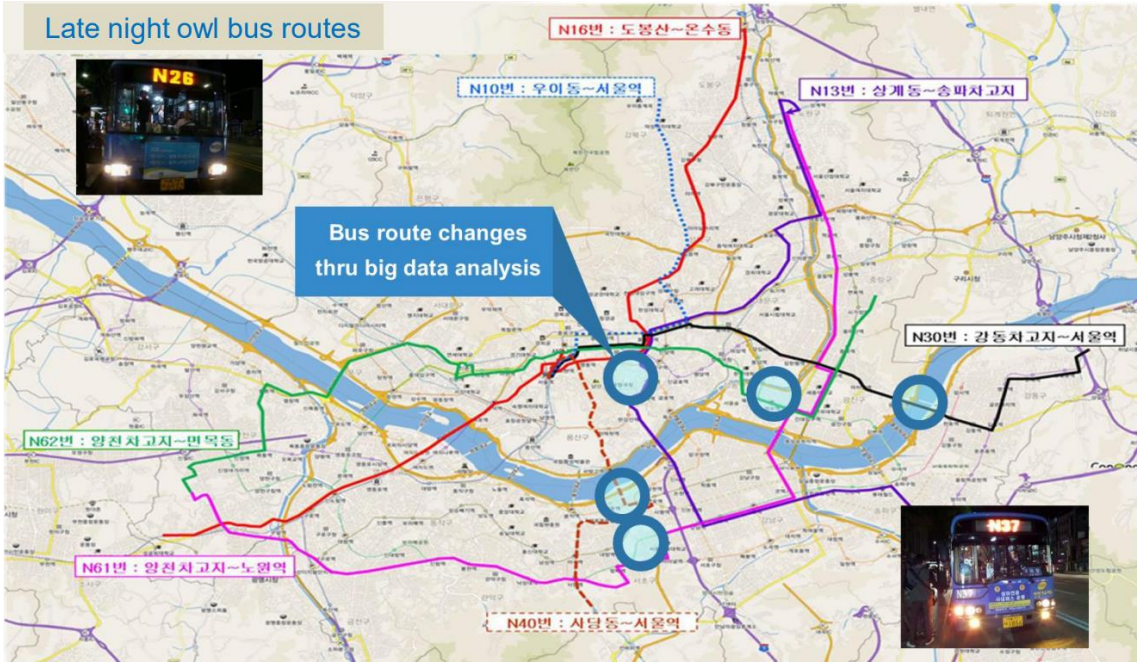
What is Big Data Analytics?

- Big Data analytics is a process used to extract meaningful insights
  - hidden patterns
  - unknown correlations
  - market trends
  - customer preferences

- Big Data analytics provides various advantages
  - It can be used for better decision making, preventing fraudulent activities, reduce cost among other things.

What is Big Data Analytics?

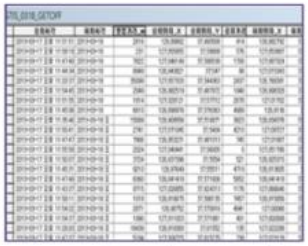
- Example of Big Data Analytics (The OWL Service)
  - Taxi at night is expensive and difficult to catch
  - Through a partnership with Korea Telecom, Seoul Government gained access to anonymized mobile communication data
    - 3 billion mobile call logs, 5 million taxy ride data



What is Big Data Analytics?

❖ Example of Big Data Analytics (The OWL Service)

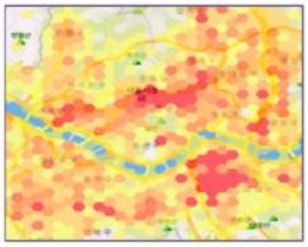
Big Data Analytics for Bus Route Optimization



1. Data collection and analysis

2. Layer modelling based on ride locations

3. Hexagon mapping



4. Floating population density analysis

5. Bus route optimization with floating population

6. Dispatch timetable adjusted accordingly

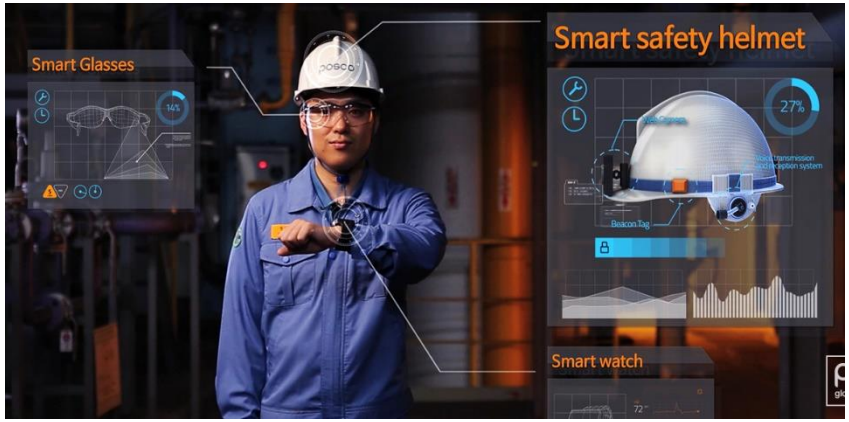
Impact

After three months of operating two routes

- Covers 42% of Seoul residents
- 7,900 passengers per day
- 2.3 million less car trips per year
- $13 million fare savings
- 500 metric tons reduction in greenhouse gas emission per year
- A service satisfaction score of 82 points (74.3 points for standard buses)

What is Big Data Analytics?

- Example of Big Data Analytics (POSCO)
  - POSCO is one of the largest hot rolling plant in the world
  - POSCO reduced energy input by 2% and save 1 billion won annually
    - Collecting and analyzing manufacturing environment data through sensors in factory
    - Maintaining the optimal working conditions through AI



What is Big Data Analytics?

- Example of Big Data Analytics (Siemens Amberg Factory)
  - The most successful case of smart factory using big data
  - Produce prototypes in virtual environment
  - The reasons of defects: materials, man(worker), machine, method
  - Automation 75%, Defects: 0.0009%, Energy: 30%, Time: 50%

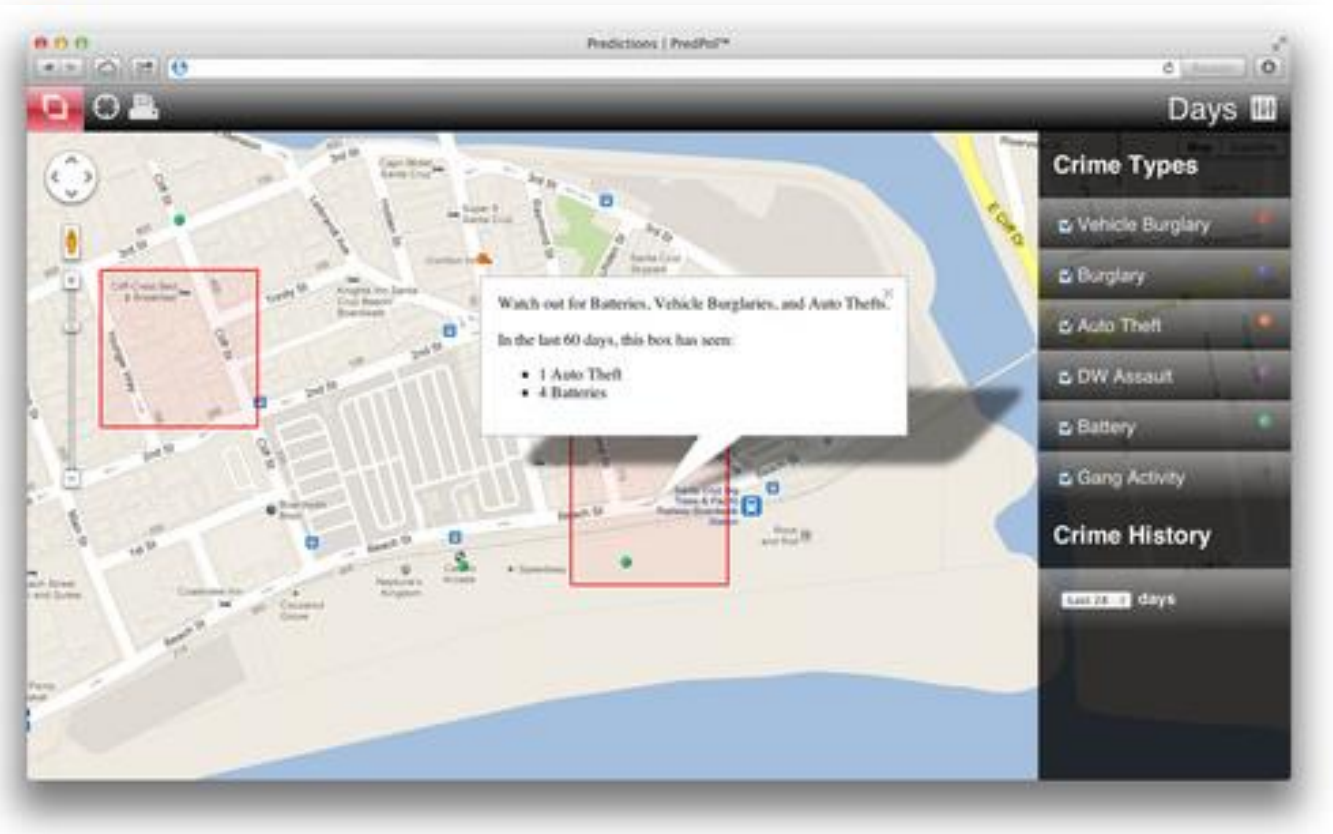

Source: https://www.siemens.com/global/en.html

What is Big Data Analytics?

- Example of Big Data Analytics (Netflix)
  - With 115 million subscribers, Netflix collect a huge amount of data
    - Ratings, watch history, searchers and others
  - Recommend the next movie you should watch or smart advertising

What is Big Data Analytics?

- Example of Big Data Analytics (Predpol)
  - Projection of areas where criminal activity is most likely
  - Reduced crime rates in Los Angeles, US



What is Big Data Analytics?

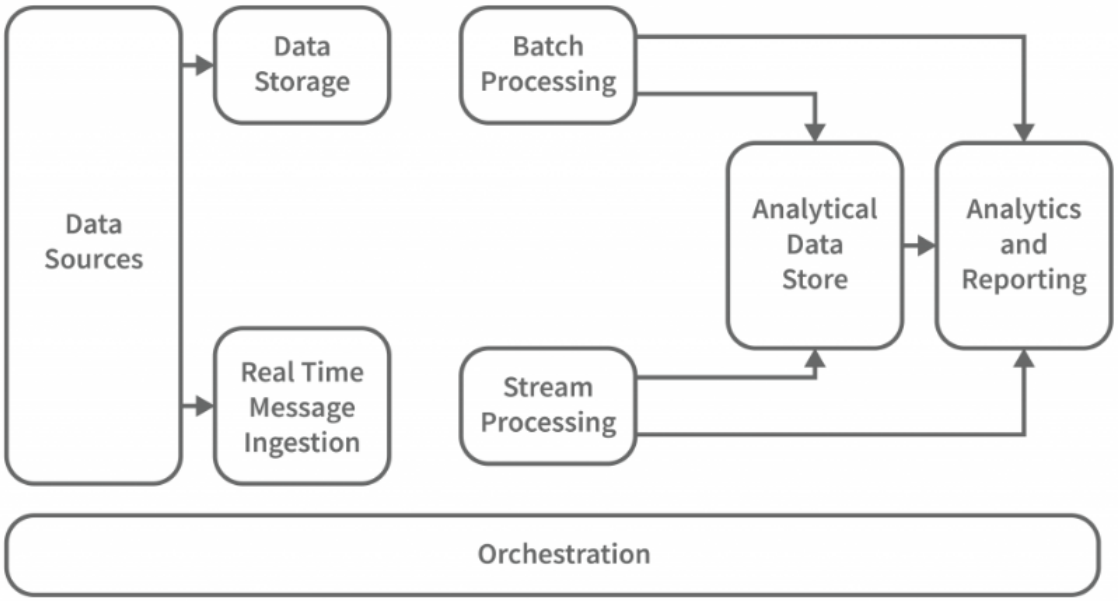
- Example of Big Data Analytics
  - Drug data reveal sneaky side effects

What is Big Data Analytics?

- Example of Big Data Analytics
  - Shoplifting detection using artificial intelligence (AI)

How to design Big Data System?

- Big Data Frameworks (required)
  - Processing tremendous data
  - Real-time data processing
  - Low cost
  - Fault Tolerance



Source: https://www.interviewbit.com/blog/big-data-architecture/

How to design Big Data System?

- Framework
  - Software environment
  - Reuse of the designs and implementations of software functions
  - Support the development of new applications or solutions

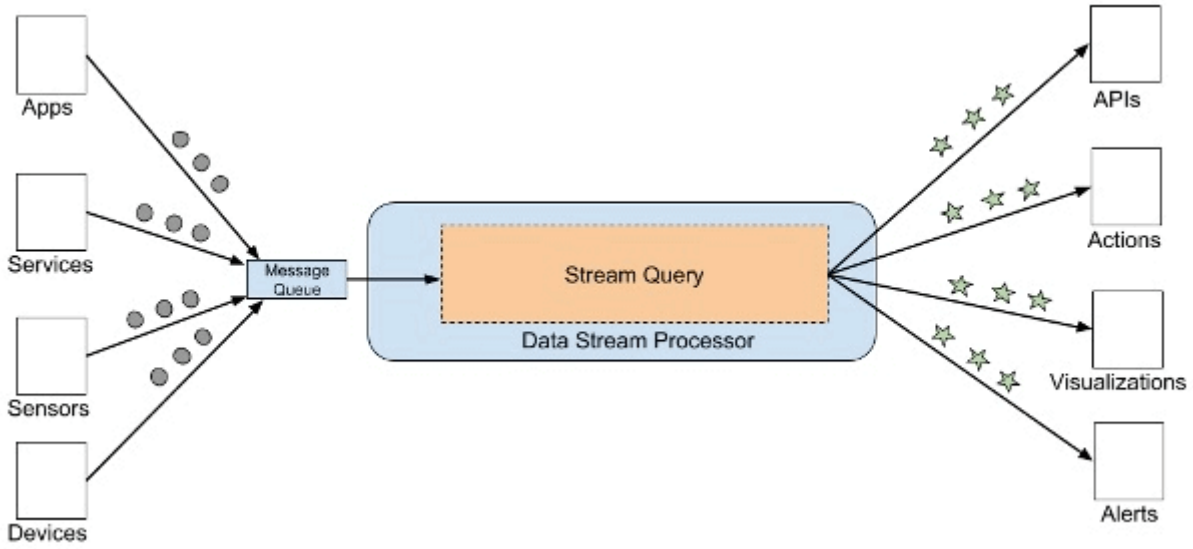
Source: https://www.analyticsvidhya.com/blog/2022/09/the-origin-of-big-data-analytics/

How to design Big Data System?

- Real-time data processing (Why?)
  - Numerous data sources: SNS, IoT, Smartphone, Sensors, RFID, Log, etc
  - Real-time analysis → Pattern identification → Decision or New BM
  - Data integration and analysis techniques (required)
    - Data filtering and cleaning
    - Data regularization and normalization
    - Outlier detection
    - Data interpolation
    - Data integration
    - Data analysis
    - Data visualization, etc
  - Data security & privacy (required)

How to design Big Data System?

- Streaming data
  - Data generated continuously from data sources
  - Typically, data records (in kilobytes) are transferred simultaneously
  - E.g., Customer logs, transactions, SNS, Stock market, Video, GPS, etc

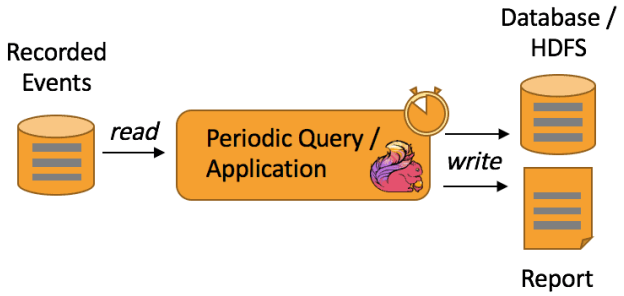


Source: https://www.upsolver.com/blog/streaming-data-architecture-key-components

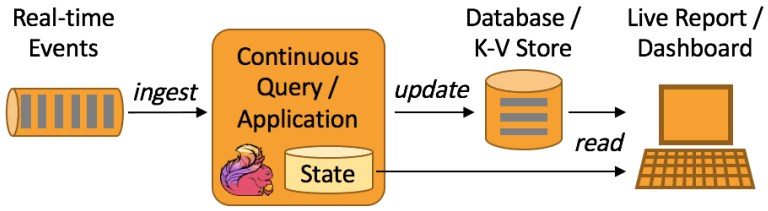
How to design Big Data System?

- Big Data System
  - Distributed and Parallelized System for managing large amounts of data
  - Real-time data processing and batch data processing
  - Functions: Data collection, management, transmission and analysis

Batch processing



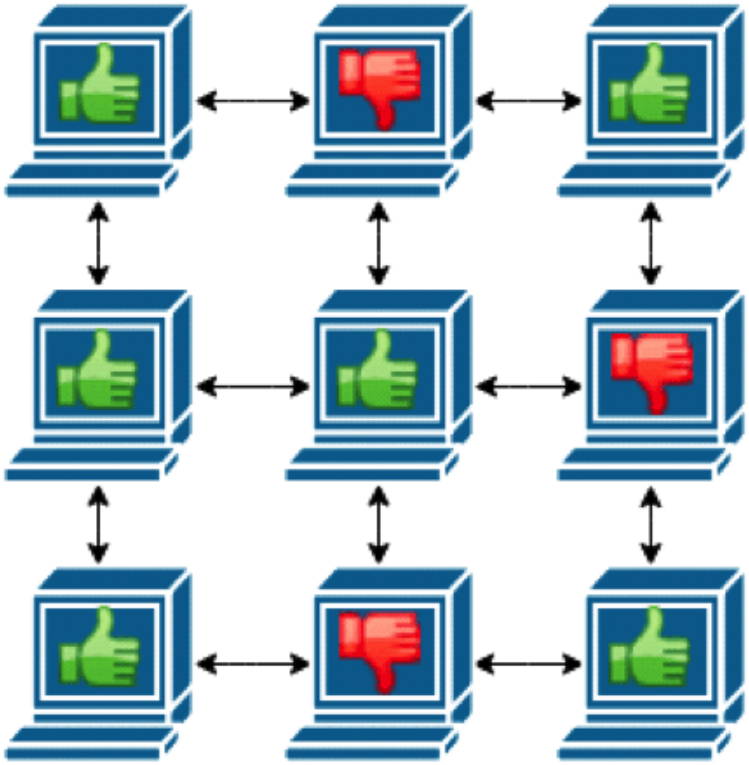
Real-time processing



Source: https://k21academy.com/microsoft-azure/data-engineer/batch-processing-vs-stream-processing/

How to design Big Data System?

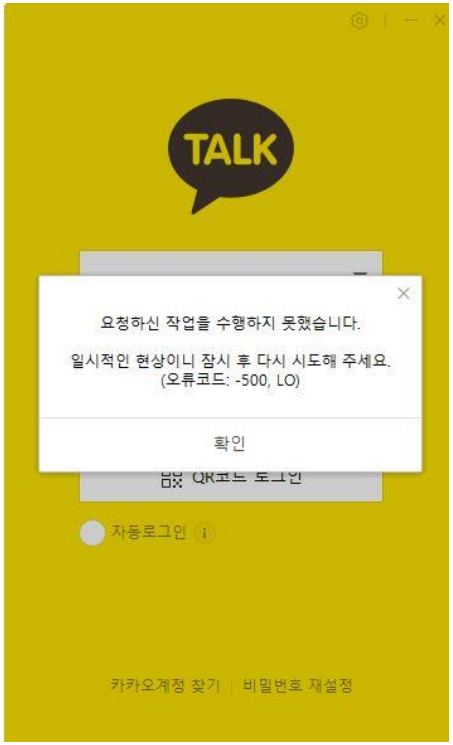
- Fault tolerance system
  - System that performs tasks even when faults occur



Source: https://kariera.future-processing.pl/blog/how-to-build-a-fault-tolerant-system/

How to design Big Data System?

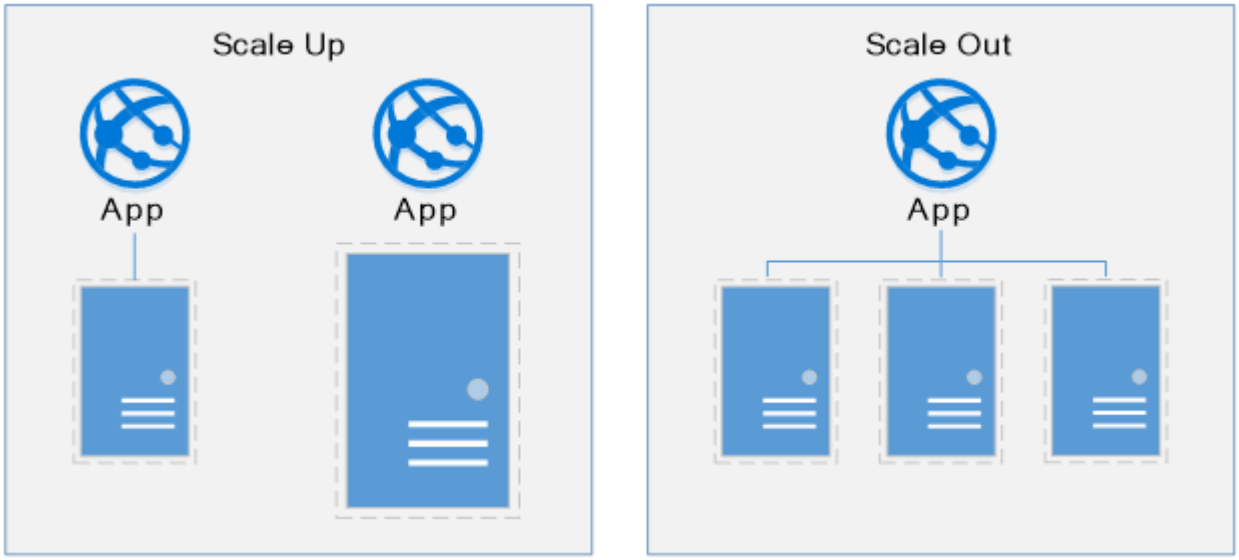
- Fault tolerance system
  - System that performs tasks even when faults occur
  - Faults: Hardware malfunction, Software errors, Computing resources, Data corruption, etc
  - Problem



Source: https://www.kakaocorp.com/page/service/service/KakaoTalk

How to design Big Data System?

- Cost effective system
  - Low cost system



Source: https://stackoverflow.com/questions/63260033/difference-between-scale-up-and-scale-out-app-service-plan

How to design Big Data System?

- Cost effective system
  - Big data systems are not necessarily expensive
  - Reduce cost by configuring the sufficient system for tasks
  - Considerations: data type, processing time, data throughput, and other system requirements
  - Advantage
    - Low initial cost → Competitive price

How to design Big Data System?

- Compatibility with existing systems
  - Compatibility with existing systems for big data collection
  - E.g., existing RDBMS, Hadoop system, applications, etc

How to design Big Data System?

- Opensource
  - Available to anyone (without any restrictions)
  - This is compatible with most existing systems
  - Open source license: GPL, LGPL, MPL, BSD, MIT, Apache

	GPL	LGPL	MPL	BSD	MIT	Apache
Download	O	O	O	O	O	O
Deploy	O	O	O	O	O	O
Modify	X	X	X	O	O	O
Noncommercial	X	X	X	X	X	X
No Derives	X	X	X	X	X	X

Summary and Discussions

- Intro to Big Data
  - Data most commonly refers to information that is transmitted or stored electronically
  - 4Vs of Big Data

- What is Big Data Analytics
  - Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, customer preferences

Questions?

SEE YOU NEXT TIME!