Introduction to Big Data

Pooya Jamshidi

pooya.jamshidi@ut.ac.ir

Ilam University

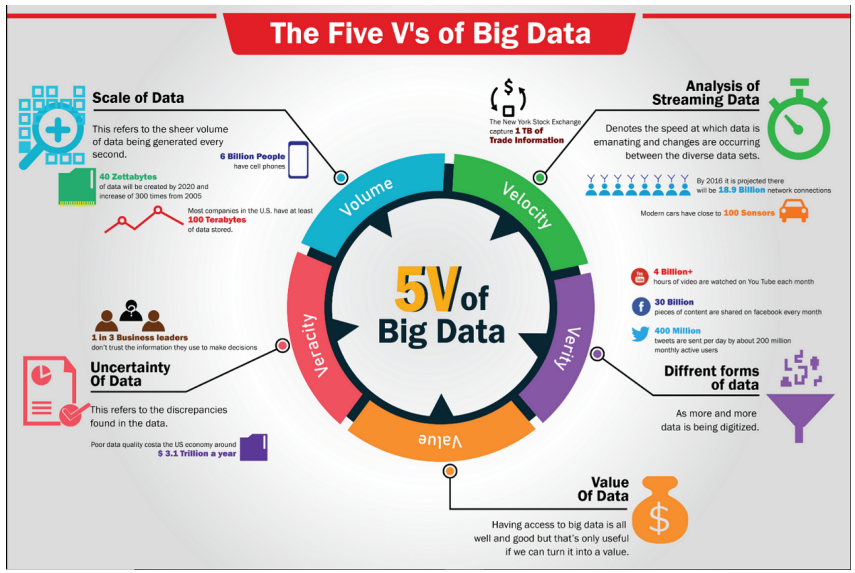
School of Engineering,
Computer Group

February 22, 2025

Big Data Characteristics & Data Mining

Big Data Characteristics

- **3 V's (Laney 2001)**
  - **Volume**
  - **Variety**
  - **Velocity**
- **Plus one**
  - **Value**
- **Another one**
  - **Veracity**
- **Plus many more**
  - Validity
  - Variability
  - Viscosity & Volatility
  - Viability
  - Venue
  - Vocabulary



Source: morioh.com

Volume

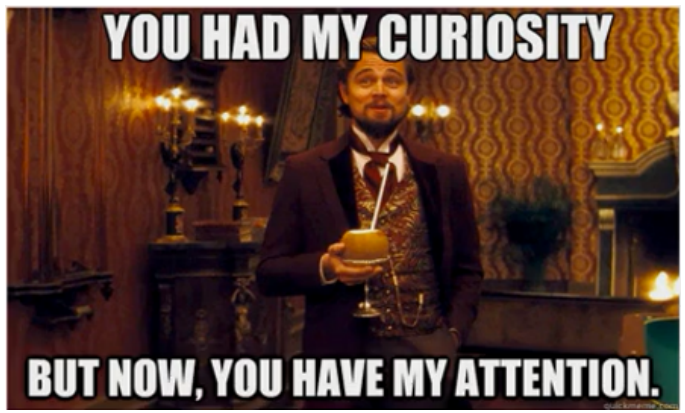
- **How much storage space the data takes up**
  - Driven by exponential growth in **storage capacity**
  - Mediated by technology
    - Parallel processing
    - Better hardware
- **Zetabyte Era:**
  - Cisco Inc. report:
    - The global IP traffic achieved an estimated **1.2 ZB** (or an average of **96 exabytes (EB)** per month) in **2016**.
    - Global IP traffic: All digital data that passes over an IP network which includes, but is not limited to, the public Internet.
    - The largest contributing factor to the growth of IP traffic comes from **video traffic** (including online streaming services like **Netflix** and **YouTube**).

How much is a ZettaByte?

Value	Metric
10^3	**kB** (kilobyte)
10^6	**MB** (megabyte)
10^9	**GB** (gigabyte)
10^{12}	**TB** (terabyte)
10^{15}	**PB** (petabyte)
10^{18}	**EB** (exabyte)
10^{21}	**ZB** (zettabyte)
10^{24}	**YB** (yottabyte)

Volume

- **European Union industry chief Thierry Breton** called on **streaming** platforms to help reduce their load on the continent's infrastructure at the beginning of COVID-19 lockdown.
- **Billion** is the keyword we're looking for...



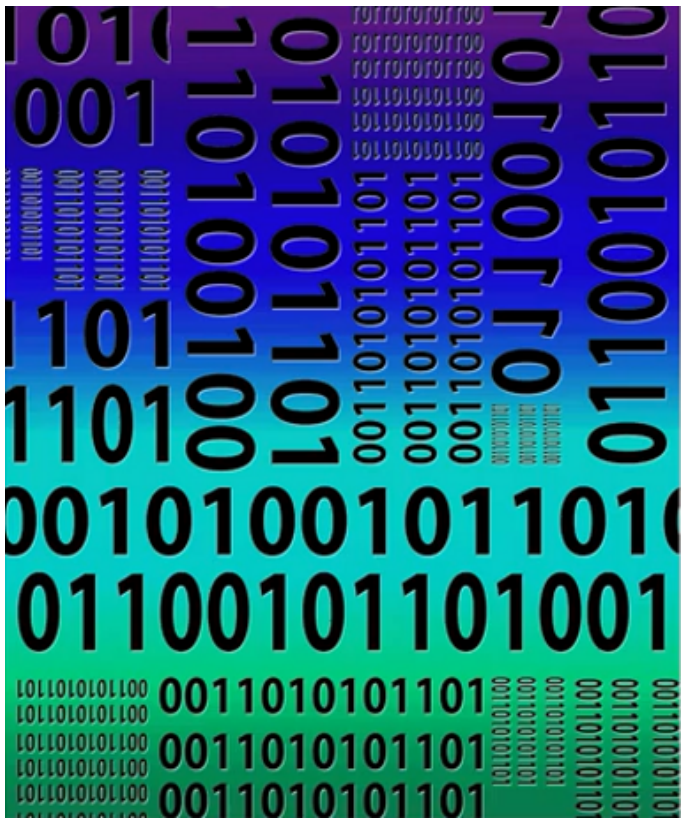
Data Storage Calculation:

$$(10^6 \times 10\text{kB}) = 10\text{GB}$$

$$(10^9 \times 10\text{kB}) = 10\text{TB}$$

Variety

- **How heterogeneous the data is.**
  - Many features per item
  - Irregular structure (as opposed to structured data for **RDBMSes**)
  - Need to store and retrieve different data types quickly, efficiently, **cheaply**
  - Need to align & integrate different representations
  - Dealt with using standards, specs, etc.
- **Big data draws from text, images, audio, and video**
  - It completes missing pieces through **data fusion**.



Dimensions of Variety

- **Content:**
  - **Image**, spectrum, timeseries
- **Form:**
  - **Text**, numeric, relational, graphical, geospatial, sensory
- **Format:**
  - **Plain-text file**, .csv, fixed-width, Excel spreadsheet, HTML table
- **Structure:**
  - Unstructured text, semi-structured email, semantically-marked-up document
- **Source:**
  - **Human-generated**, automated sensor logging, **scientific instruments**, simulations
- **Meaning:**
  - *"This dish is hot."*
- **Representation:**
  - Feb. 20, 2025 vs. **2025/20/02** vs. 2025/02/20

Velocity

- **How quickly** data must be generated and processed
- **Speed** of storage / retrieval / analysis
- **Aspects:**
  - **Real-time** (acted on immediately)
  - **Timeliness** (rate of capture/usage)
  - **Lifespan** (how long it's valuable)
  - **Response time**
- **Strategies:**
  - Simple ingest & access
  - Parallelization
  - Better hardware

Value

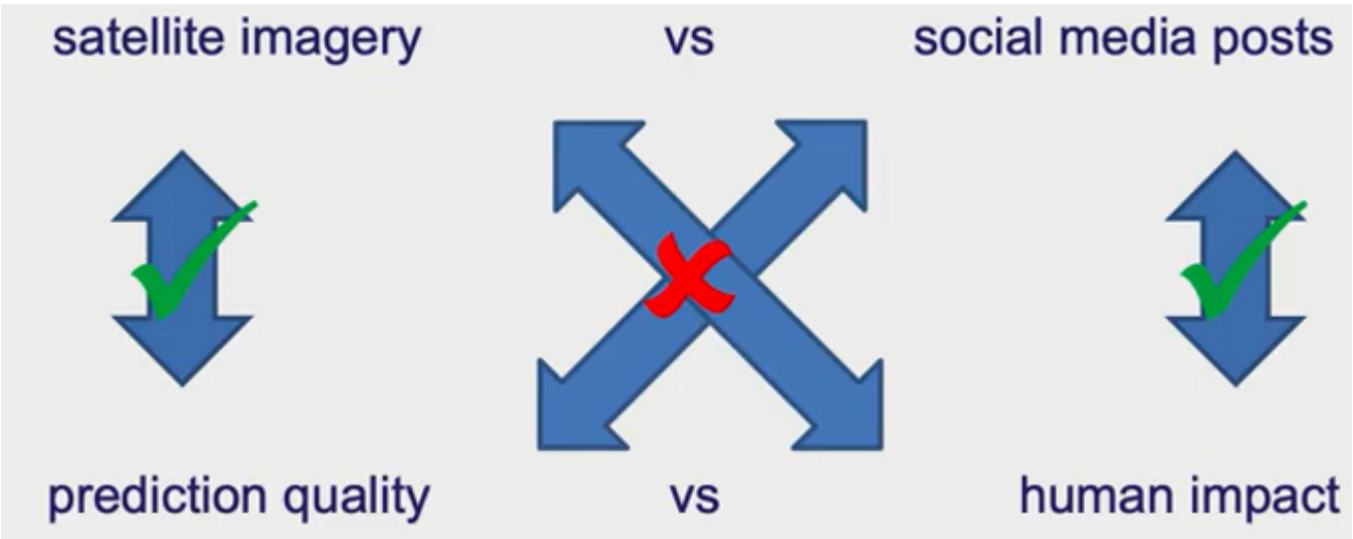
- **Business value** or ROI
- Data value can be achieved by the **processing** and analysis of large datasets.
- Value also can be measured by an assessment of the **other qualities of big data**.
- Value may also represent the **profitability** of information that is retrieved from the analysis of big data.

Veracity

- **Is the data trustworthy?**
  - **Provenance**, reliability, accuracy, **completeness**, ambiguity.
  - Importance of Veracity depends on what the *Value* of the data is.
- **Strategies:**
  - Transparent QC
  - **Provenance tracking**
  - Data management best practice
  - Good governance practices
- **Note:** Provenance and other **veracity metadata** can itself become **Big Data**.

Validity

- **Accuracy and correctness** of the data relative to a **particular use**.
  - Example: Gauging storm intensity



How much data does Facebook have?

- It contains an extremely heterogeneous set of data:
  - **Binary blobs:** e.g., photos & videos
  - **Textual data:** e.g., post contents
  - **Metadata:** e.g., impressions & metadata
- Facebook stores several **exabytes of data**, and the size grows exponentially.

Source 1 — Source 2

Variability

- **How the meaning of the data changes over time**
  - Language evolution
  - Data availability
  - Sampling processes
  - change in the souece of data

Viscosity & Volatility

- Both related to **velocity**.
- **Viscosity**: *data velocity relative to timescale of event being studied*.
- **Volatility**: *rate of data loss and stable lifetime of data*.
  - Scientific data often has practically unlimited lifespan, but social/business data may evaporate in finite time.

More V's

- **Viability**
  - Which **data** has meaningful relations to questions of interest?
  - Another take on **value**.
- **Venue**
  - Where does the **data** live and how do you get it?
- **Vocabulary**
  - **Metadata** describing structure, content, & provenance
  - **Schemas**, **semantics**, ontologies, taxonomies, vocabularies

Critiques of Big V's Model

- Big V's model concerns mostly about **scalability** than **understandability**.
- An alternative is *cognitive big data* which concerns around:
  - **Data completeness:** Understanding of the *non-obvious* from data.
  - **Data correlation, causation, and predictability:** Causality as not essential requirement to achieve predictability.
  - **Explainability and interpretability:** Humans desire to understand and accept what they understand, where algorithms do not cope with this.
  - **Level of automated decision making:** Algorithms that support automated decision making and algorithmic self-learning.

Source: A. Lugmayr, et al. *A comprehensive survey on big-data research and its implications - What is really 'new' in big data? It's cognitive big data!* Pacific Asia Conference on Information Systems, 2016.

Data Mining

Data Mining

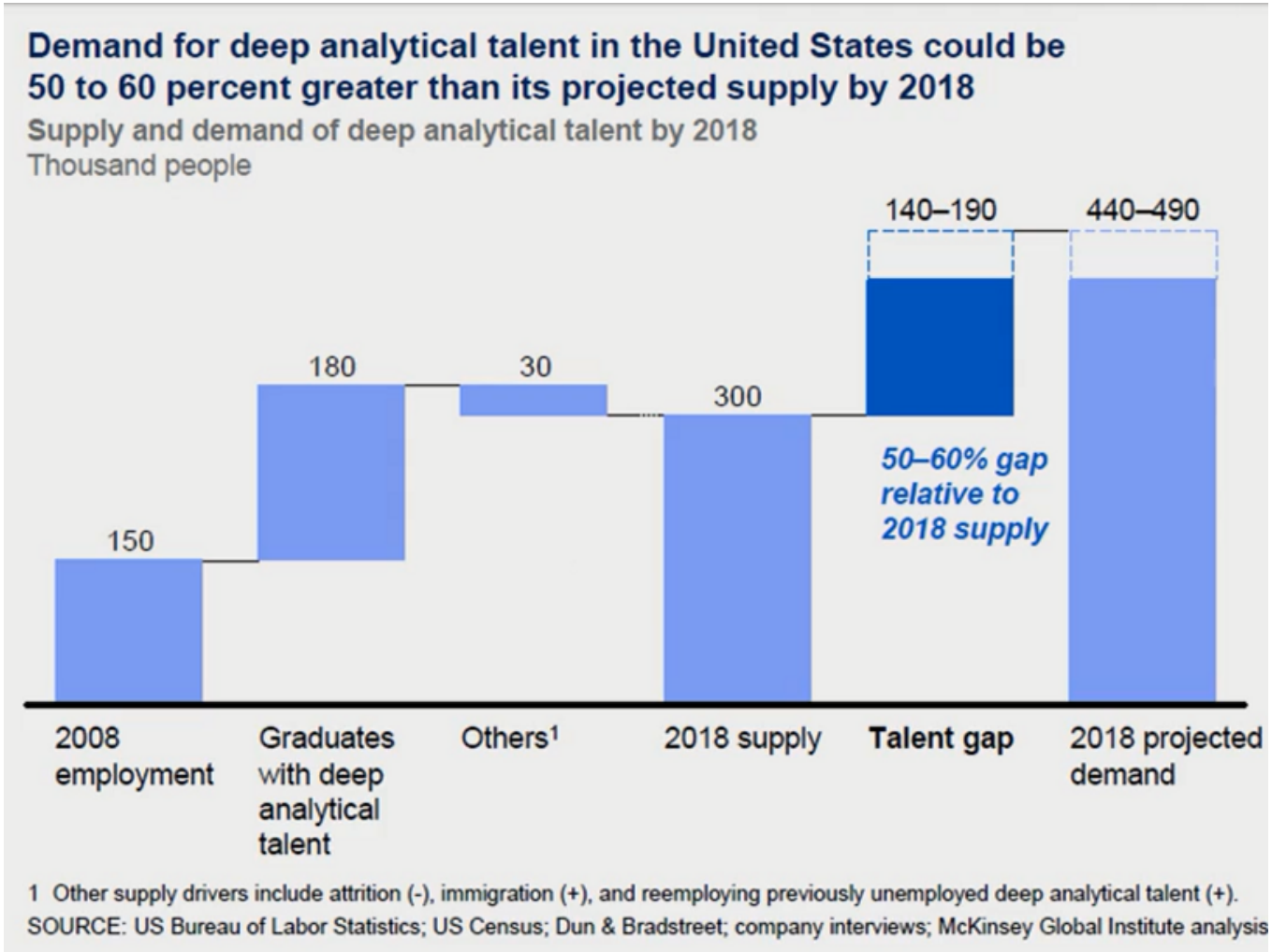
- To extract **knowledge** from data, it needs to be:
  - Stored
  - Managed
  - Analyzed

Data Mining $\approx$ Big Data

$\approx$

Predictive Analytics $\approx$ Data Science

Good News


Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018
Supply and demand of deep analytical talent by 2018
Thousand people
140–190
440–490
180
30
300
150
50–60% gap relative to 2018 supply
2008 employment
Graduates with deep analytical talent
Others[1]
2018 supply
Talent gap
2018 projected demand
1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).
SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

What is Data Mining?

- Given lots of data
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
  - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict *unknown or future values* of other variables
  - **Example:** Recommender systems

Meaningfulness of Analytic Answers

- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless.**
- Statisticians call it **Bonferroni's principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap.

Total Information Awareness (TIA)

- Following the terrorist attack on **Sep. 11, 2001**, it turned out that:
  - Four people enrolled in different flight schools for **commercial aircrafts**.
  - They were not affiliated with any airline.
  - **Conclusion:** There was **enough data** to prevent the attack.
- **Total Information Awareness (TIA)** was created under **DARPA** to mine all the data it could find to track terrorist activity. These data include:
  - **credit-card receipts**
  - **hotel records**
  - **travel data**
- TIA caused great concern among **privacy** advocates, and **the project was eventually killed by Congress**.

Meaningfulness of Analytic Answers

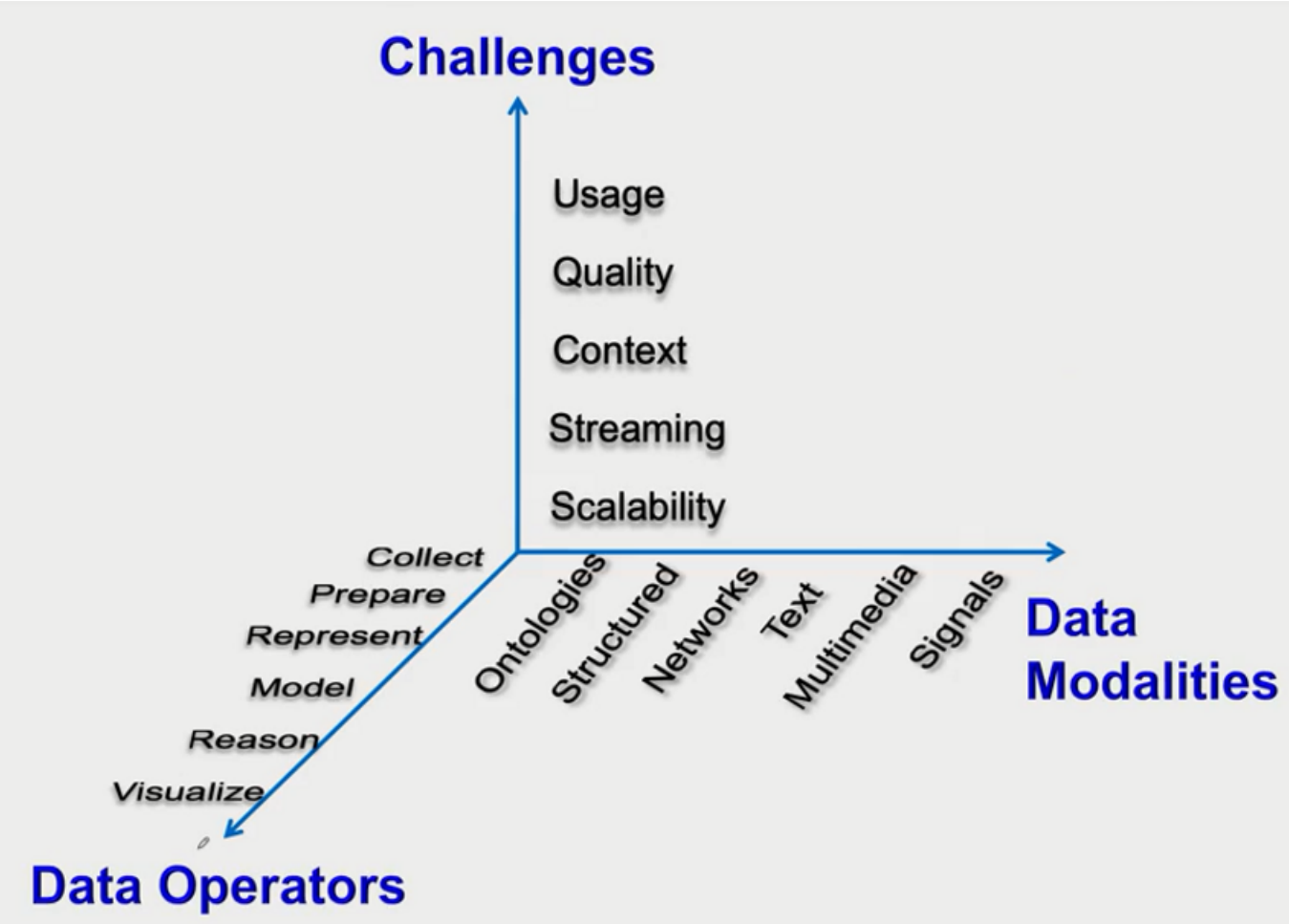
Example:

- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**.
  - 1 billion (10^9) people being tracked
  - 1,000 days
  - Each person stays in a hotel **1%** of time (**1 day out of 100**)
  - Hotels hold 100 people (so 10^5 hotels)
- **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**

Meaningfulness of Analytic Answers (cont'd)

- Probability of any two people both deciding to visit a hotel on any given day is:
$$(10^{-2})^2 = 10^{-4}$$
- Visiting the same hotel:
$$10^{-4} \times 10^{-5} = 10^{-9}$$
- Two people visiting on two different given days is:
$$(10^{-9})^2 = 10^{-18}$$
  - Hotels can be different on the two days.
- **Expected number of "suspicious" pairs of people:**
$$\binom{10^9}{2} \times \binom{1000}{2} \times 10^{-18} \approx 250,000$$
- . . . too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in a more efficient way.

What matters when dealing with data?



Data Mining: Cultures

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms
- **Different cultures:**
  - To a **DB** person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - *Result is the query answer*
  - To an **ML** person, data-mining is the **inference of models**
    - *Result is the parameters of the model*
- **In this class we will do both!**

This Class

- This class overlaps with machine learning, statistics, artificial intelligence, databases but more stress on:
  - **Scalability** (big data)
  - **Algorithms**
  - **Computing architectures**
  - Automation for handling **large data**