Aalto University
School of Science

CS-E4640 Big Data Plaforms

Issues in Time-series Data Ingestion

Minh Tri Nguyen
PhD student at Department of Computer Science
Researcher at AaltoSEA

Content

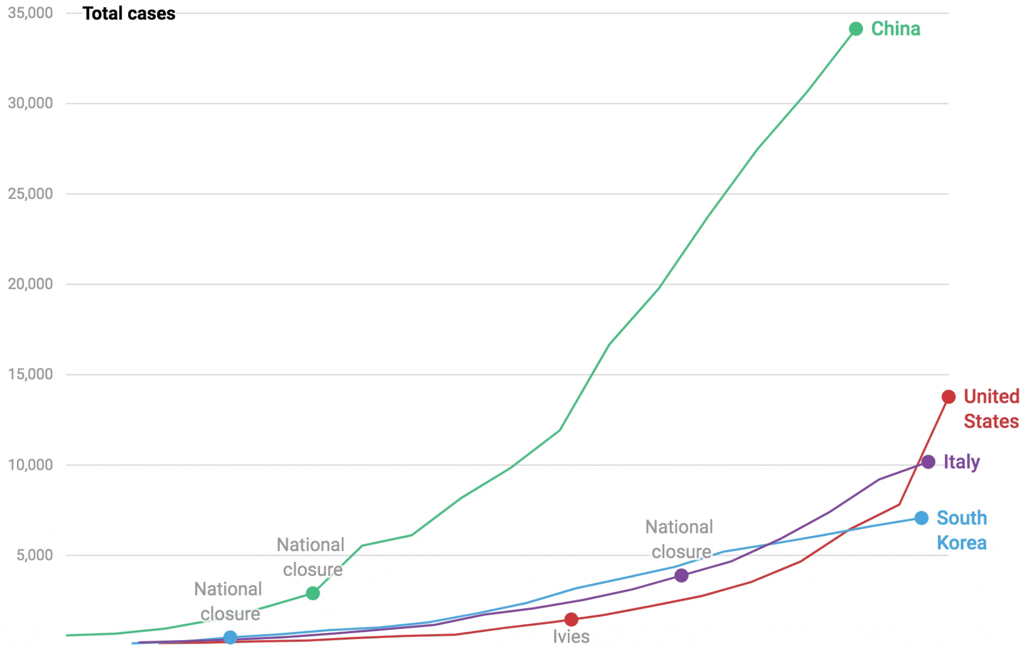
- **What is time series data?**
  - The applications of time series
  - Characteristic
- **Challenges in time series data ingestion**
  - Handling streaming data
  - Database
  - Data Partitioning

My Experiences

- Who was I?
  - A data analyst
- Applications:
  - Predicting stock value:
    - Dataset: NASDAQ 100
  - Predicting the popularity of online contents
    - Dataset: Youtube, MovieLens,...
  - Predicting alarm events
    - Dataset: BTS

What is time series?

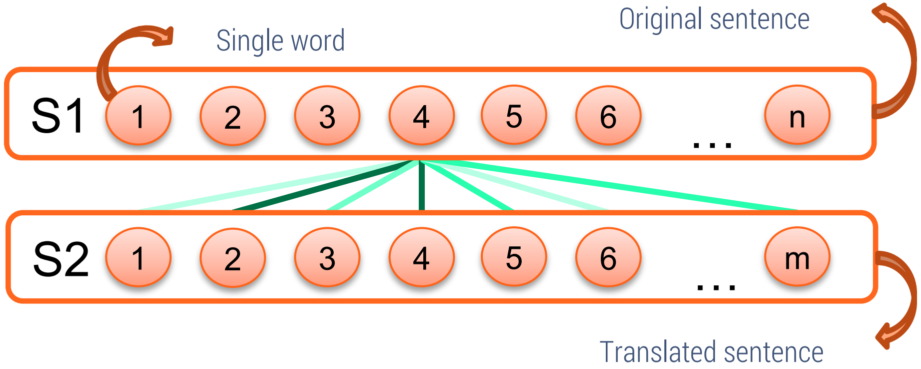
- Time series data is a sequence of data point indexed in time order. The observation is collected by repeating measurements
  - Fixed/dynamic time intervals
  - Triggered event
  - ➢ Tracking changes over time.



Corona virus data [https://www.columbiaspectator.com/contributors/Jun-Yi-Zhang/]

The application of time series

- **Simple applications:**
  - Weather forecast: hourly, daily, weekly,…
  - Health care: heart rate, breathing rate, blood sugar level,…
  - Stock trading value
  - AI system: autonomous, self-driving car, sensor system
- **Complex applications:**
  - NLP
  - Image processing
  - …



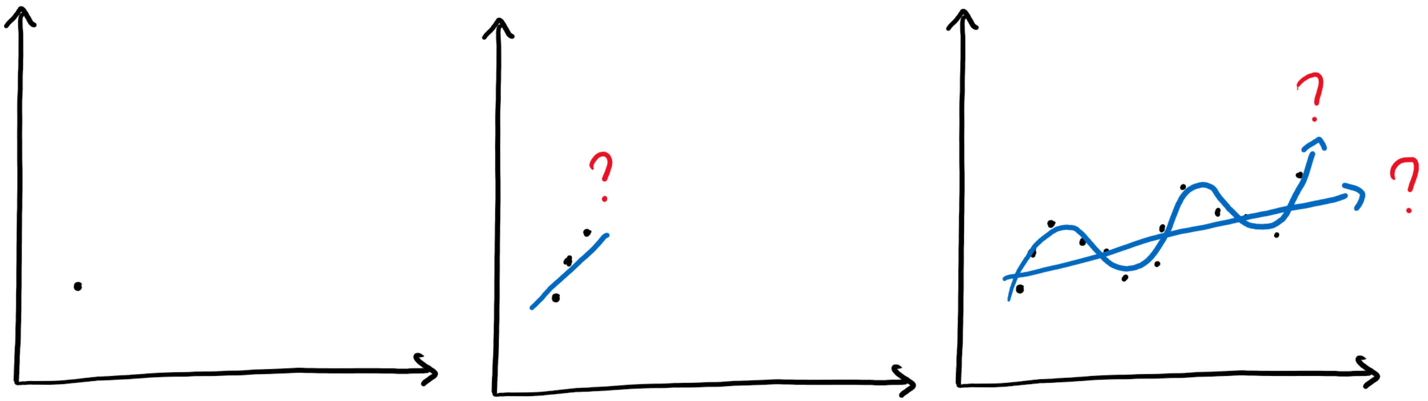
Characteristics of Time series data

- Volume:
  - *Single data point: small (a few KB)*
  - *The whole dataset: big (GB, TB,…)*
- Velocity: every day, hour, minute, second,…
- Variety: Structured, semi structured, unstructured, dynamic
- Veracity: noise, wrong data,…

➢ My initial approaches for each application always based on these characteristics

Time series data in “a whole picture”

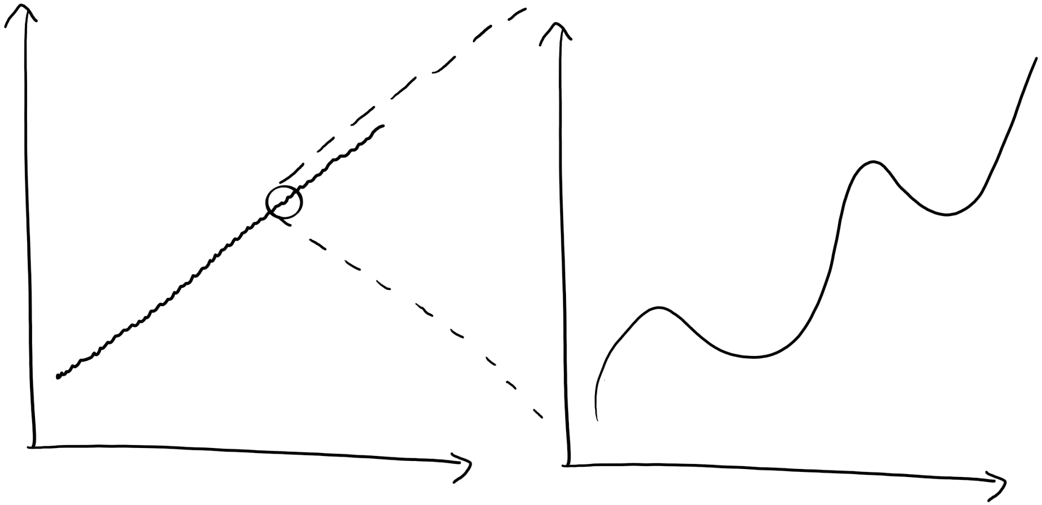
- Trend - the general direction of the changing value within the dataset: upward/downward,…
- Pattern - following a function: linear, cycle, sin(x),…
- Cohesion and correlation with other values,…



- ➢ I must be able to see the dataset in a whole picture
  - ➢ *The approach must consider tools/frameworks for data processing and visualization*

Time series data in “a small piece”

- Hidden pattern?



➢ Look at the data in details

Challenges in time series data ingestion

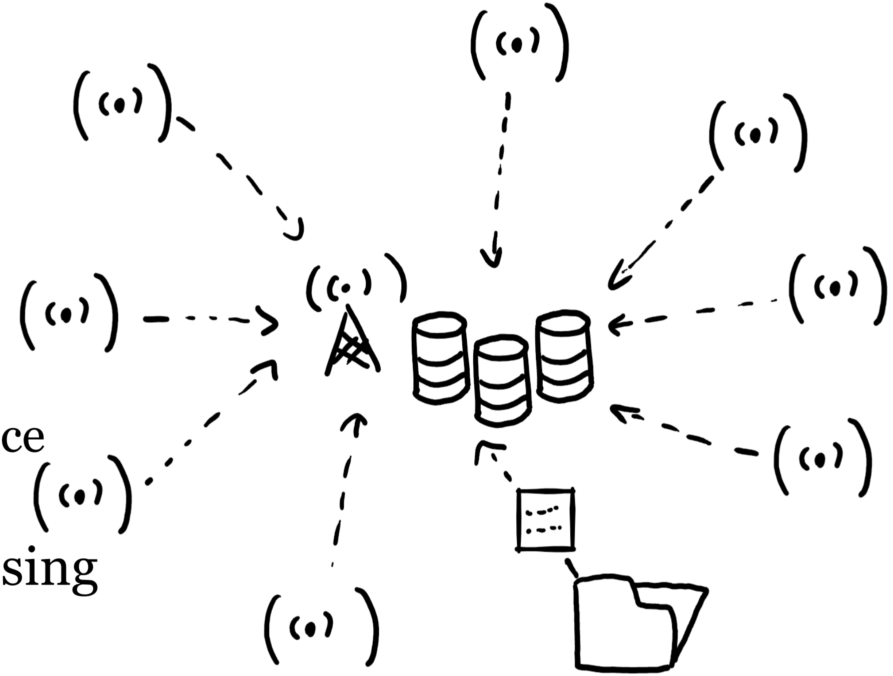
Challenges 1:

- **Data Ingestion**
  - From streaming:
    - *Velocity*
    - *Unstable network connection.*

- **BTS application:**
  - recording alarm event of IoT device failures.

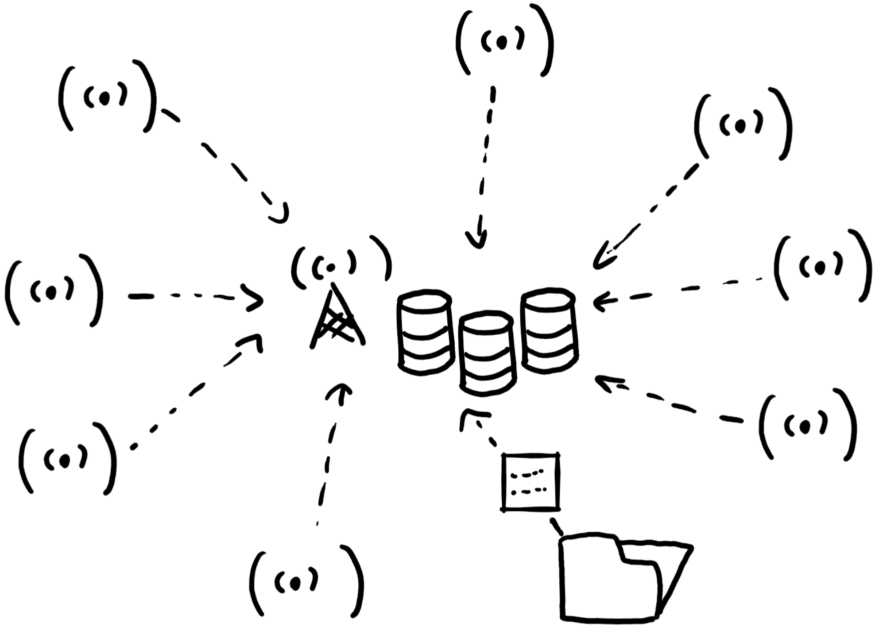
➢ I simulate the streaming data using MQTT

➢ Techniques: buffering, queueing,...



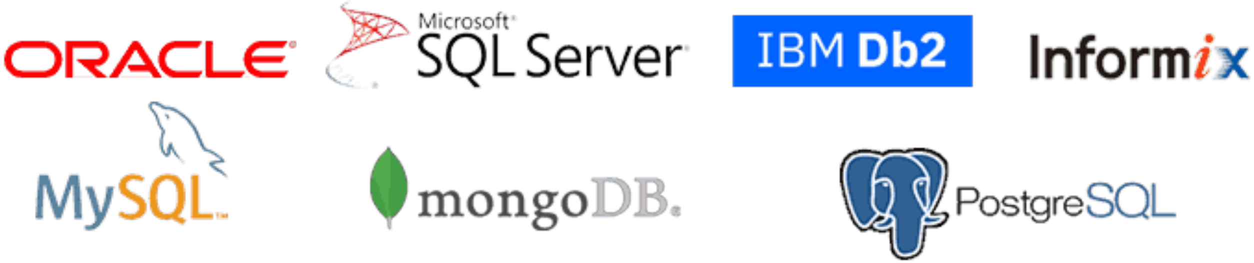
Challenges 1:

- **Data Ingestion**
  - From big files:
    - *Transferring speed*
    - *Secure transmission/privacy*
    - *Data availability*
      - Replication
      - Sharding



Challenge 2:

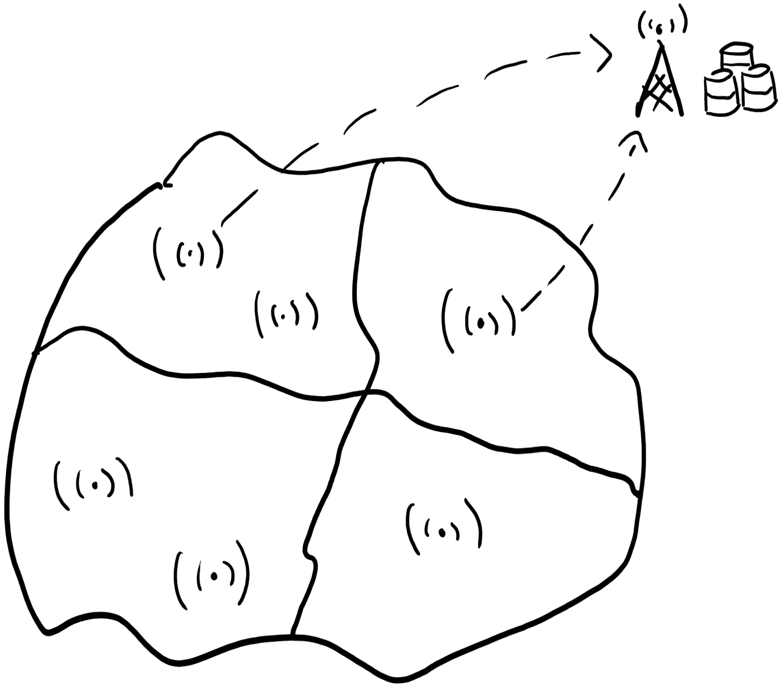
- **Choosing Database:**
  - Data nature: data types, data schema
  - Ingestion method: API, …
  - Operating speed: Move, copy, insert,…
  - Supporting tools/frameworks



➢ For Youtube data, I choose MongoDB - Flexible schema, map-reduce, connector to spark, and other ML tools and frameworks.

Challenges 3:

- **Storing Data - Data partitioning**
  - *Geographical location*
  - *Data attributes*
  - …
- ➢ Querying, Visualizing data
  - *Look at data in details*
  - *Quick access*
  - *Lower communication cost*

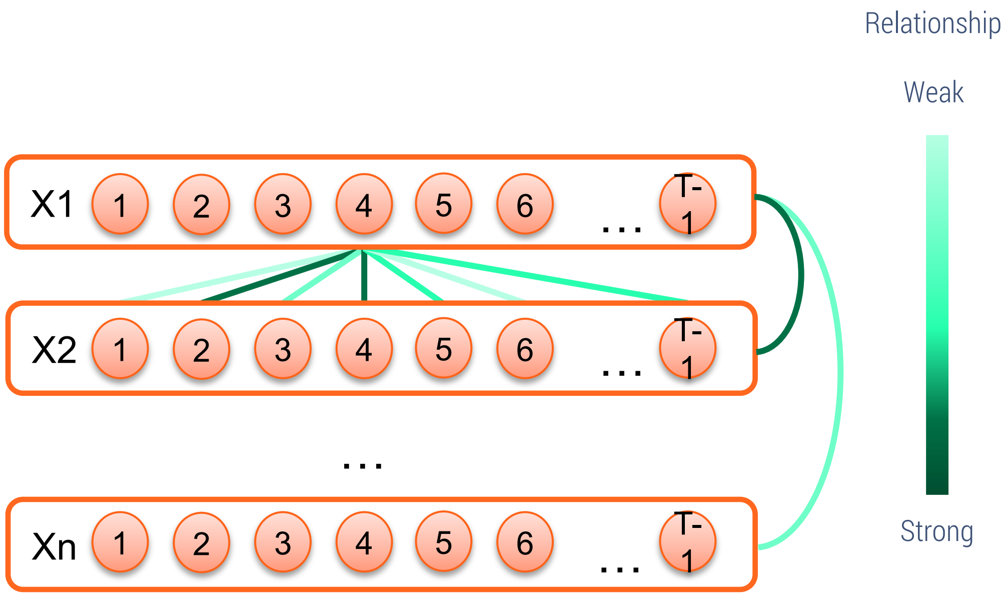


- ➢ Choosing the methods for data partitioning based on how I manipulate the data.

Data partitioning

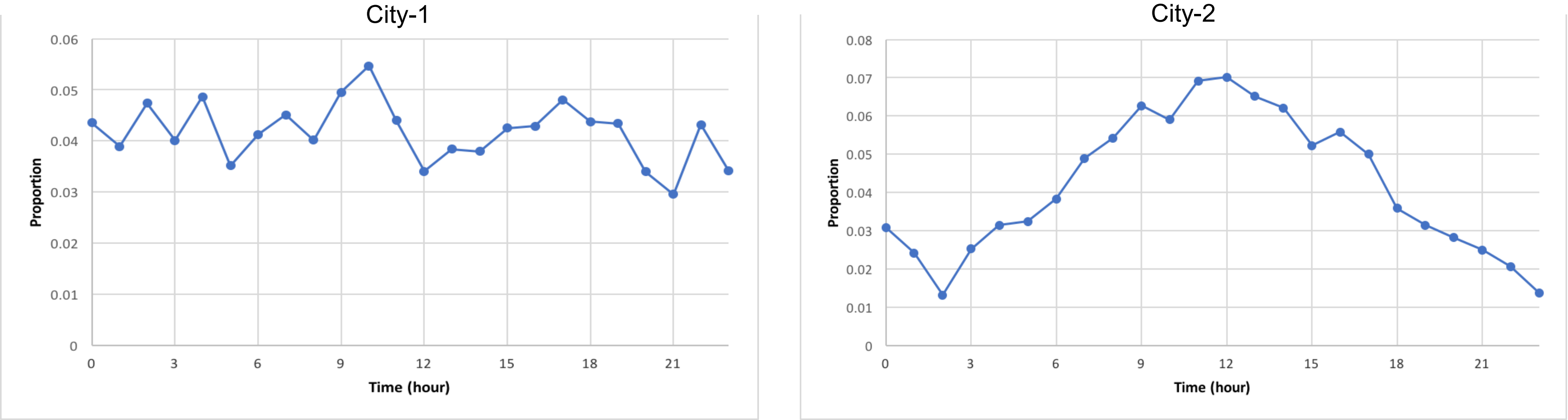
- Cohesion and correlation with other values
  - How do we know?
  - ➢ Visualization, experiments,…

➢ Correlated data should be in the same data partitions for quicker access and analysis.



Data partitioning

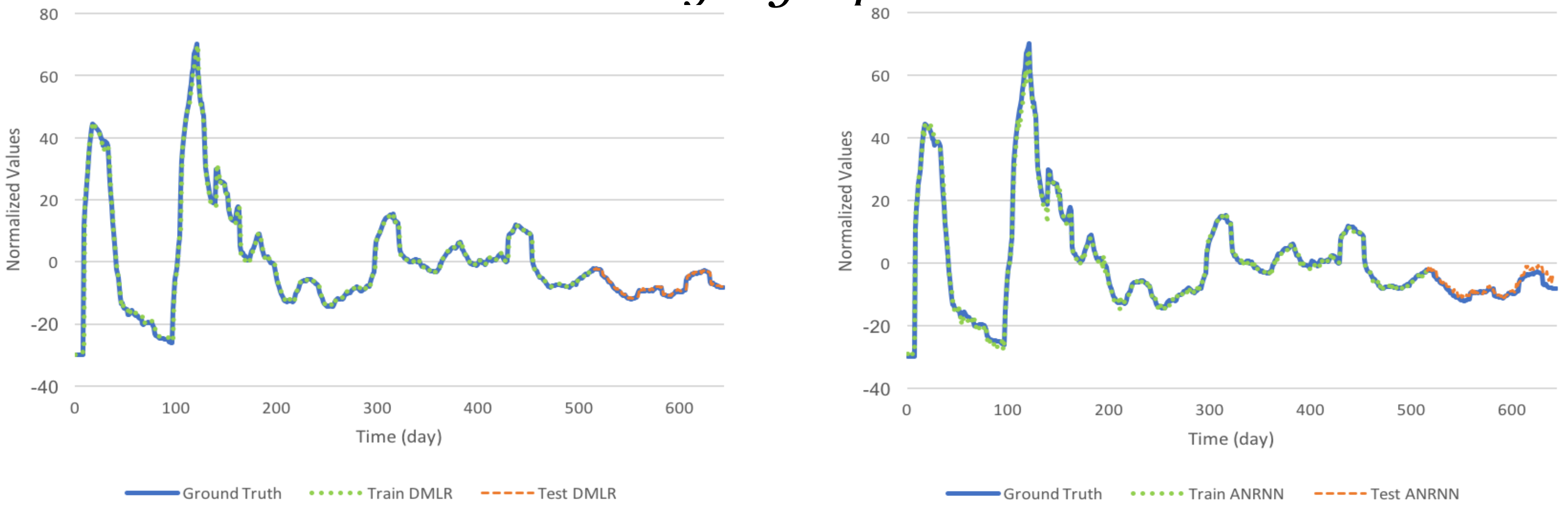
- MovieLens dataset: recording the movie's views with location.
  - *Partition data based on geographical location.*



View-count distribution within a day in 2 different places (MovieLens dataset)

Data partitioning

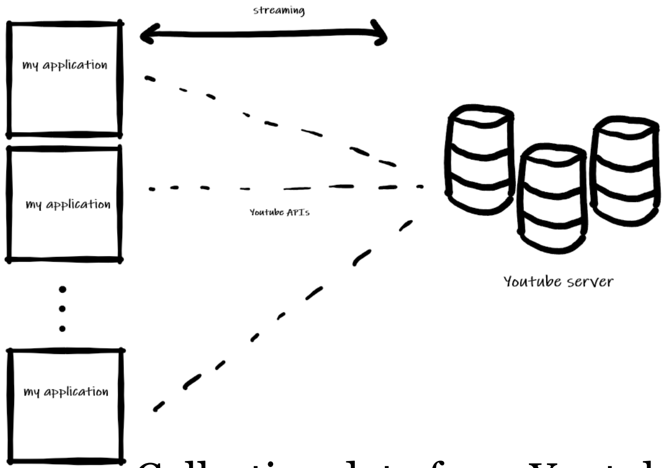
- MovieLens dataset: recording the movie's views with location.
  - *Partition data based on geographical location.*



Predicting Movie's views (MovieLens dataset)

Data partitioning

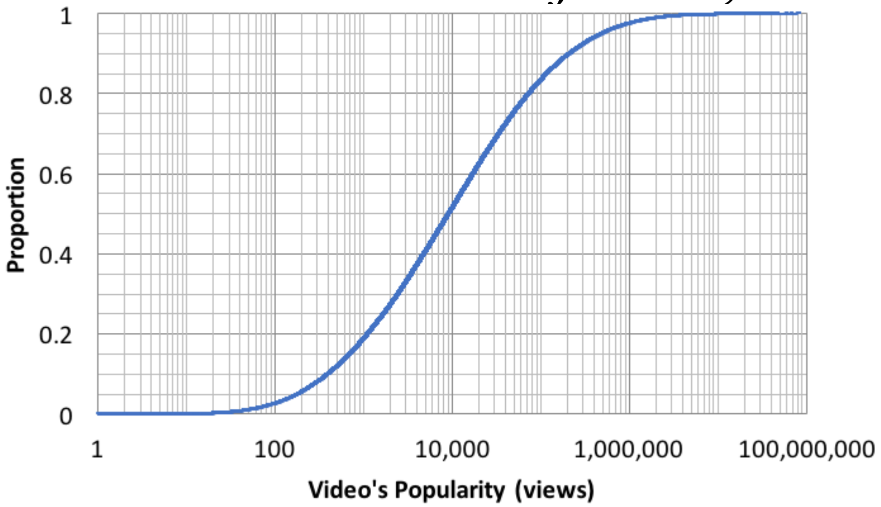
- Youtube dataset: recording the views number of 50 most popular videos in 50 countries.
  - *Partition data based on number of views, author, genre,…*



Collecting data from Youtube (https://developers.google.com/youtube/v3)

Data partitioning

- Youtube dataset: recording the views number of 50 most popular videos in 50 countries.
  - *Partition data based on number of views, author, genre,…*



View distribution on Youtube dataset

Aalto University
School of Science

Sum up

- **Different applications will come up with different approaches**
  - Streaming/files ingestion
  - Volume
  - Velocity
  - Data nature
  - Supporting tools/frameworks
  - …
  - ➢ Database, techniques, …
- **Always look at the dataset with different views**
  - Within different views, I may want to partition the dataset in different ways.
  - Visualizing, performing a lot experiments to find the optimal solutions.

References and further information

- https://version.aalto.fi/gitlab/bigdataplatforms/cs-e4640
- https://grouplens.org/datasets/movielens/
- https://cseweb.ucsd.edu/~yaq007/NASDAQ100_stock_data.html#:~:text=Description,2017%2C%20in%20total%20191%20days.
- https://developers.google.com/youtube/v3
- https://version.aalto.fi/gitlab/bigdataplatforms/cs-e4640/-/tree/master/data%2Fbts
- https://ieeexplore.ieee.org/abstract/document/8855675

Thank you!

Any Question?