Traffic Data Pipeline

→ [3472 gg] allowed
→ [2920 gg] necessary

① Gathering Data

Batch: 1 day → History: Every 5 min. collect historical data from 2013-01-01 to nowadays → 10 days → a server always-on is not necessary (may go beyond the free tier) → do it manually

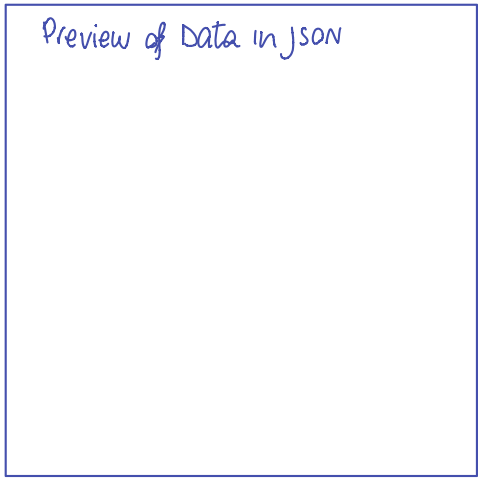
Stream: 10 min → Real-time: Every 10 min. collect real-time data for every Station → always-on

Computing

Note: the typical pattern is every 10 min, so we don't need Pub/Sub

necessity of a server

Preview of Data in JSON



storage

② Store Data

→ Possible solutions:
- DynamoDB → for AWS Cloudwatch
- MySQL
- SQLite
- Spark

(MySQL, SQLite, Spark: if we don't use AWS, Pub/Sub is necessary)

Note: number of requests may go beyond the free tier

→ Discussion

Collect data and divide them into 2 tables:
1. Info about Bluetooth Stations
2. Timestamp, Count, idStation

③ Query the DB

- AWS → boto3 module
- MySQL → mysql
- SQLite → sqlite
- Spark → Spark

communication

4) Build Prediction Model

↳ state of the art
- Neural Networks

5) Store intermediate result of the model

↳ evaluate if necessary

6) Predict from place-hour

Computation

storage

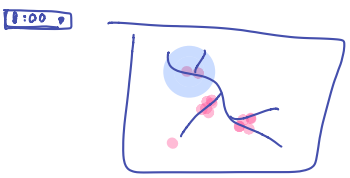
Computation

Communication

7) Result

↳ interactive map of traffic

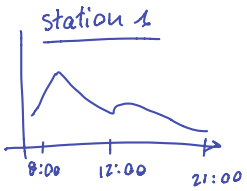
↳ choose time range in terms of hour



Note: if we can compute distances and paths, we can also communicate future traffic

↳ number of vehicles in a given place at a given time → basic result

↳ graph with traffic prediction for a place



Note: for the report, compute accuracy of our model

+ always keep the map up-to-date → more important most recent data than oldest ones.