Project 1

Corso di Sistemi e Architetture per Big Data

A.A. 2021/22

Valeria Cardellini, Matteo Nardelli

Laurea Magistrale in Ingegneria Informatica

Project delivery

- Submission deadline
  - June 9, 2022
  - After the deadline, the maximum achievable score will be decreased by 1 point for each week of delay
- Your presentation
  - June 14, 2022 (to be confirmed)
- What to deliver
  - Link to cloud storage or repository containing project code
  - Project report composed by 3-6 pages in ACM or IEEE proceedings format
  - Presentation slides (max. **15 minutes** per group), to be delivered after your presentation
- Team
  - Target: 2 students per team
  - Also possible 1 student or 3 students per team

Dataset

- You will use a real dataset about taxi trip records in NYC: TLC Trip Record Data
  - Data collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs
  - Available at https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
  - Full dataset in CSV format (from May 13th in Parquet format) and updated monthly
  - We will consider Dec. 2021, Jan. 2022 and Feb. 2022

Dataset

- The yellow and green taxi trip records include fields capturing:
  - pick-up and drop-off dates/times
  - pick-up and drop-off locations
  - trip distances
  - itemized fares
  - rate types
  - payment types
  - driver-reported passenger counts

Dataset: yellow taxi trip records

- Header of CSV file

```
VendorID, tpep_pickup_datetime,
tpep_dropoff_datetime, passenger_count,
trip_distance, RatecodeID,
store_and_fwd_flag, PULocationID,
DOLocationID, payment_type, fare_amount,
extra, mta_tax, tip_amount, tolls_amount,
improvement_surcharge, total_amount,
congestion_surcharge, airport_fee
```

Full description available at
https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

Queries with Hadoop/Spark

- Use Spark framework (or alternatively Hadoop framework and MapReduce programming model) to answer some queries on the dataset
- Include in your report/slides the queries' response time on your reference architecture

Query 1

Average calculation on a monthly basis of the percentage

```
tip_amount/(total_amount – tolls_amount)
```

Queries with Hadoop/Spark

Query 2

Distribution of the number of trips with respect to the departure area (`PULocationID`), average tip and its standard deviation, the most popular payment method, in 1-hour slots

Queries with Hadoop/Spark

Query 3

Identify the top-5 most popular `DOLocationIDs` (TLC Taxi Destination Zones), indicating for each of them the average number of passengers and the mean and standard deviation of `Fare_amount`

Platform and performance evaluation

- Evaluate experimentally the query processing times on the reference platform you used
- Platform can be a standalone node
  - Recommended: use Docker Compose to orchestrate locally the containers
- Alternatively, you can use a Cloud service for Big Data processing (e.g., Amazon EMR) using the available grant

Optional part A

- Compulsory for team composed of 3 students

- Use a higher level framework (Hive or Spark SQL) to address Queries 1, 2 and 3
- Evaluate the performance of all the queries on your reference architecture for both cases

Data acquisition and ingestion

- Which framework to ingest data into HDFS?
  - Flume, NiFi, Kafka, …
- Which format to store data?
  - csv, columnar format (Parquet), row format (Avro), …
- Where to export your results?
  - HBase, Redis, Kafka, …

Optional part B

- Use a visualization framework (e.g., Grafana) to graphically present the query results

Team composition and tasks

- 1 student in the team:
  - Queries 1 and 2
  - Data acquisition and ingestion are optional, HDFS is mandatory
- 2 students in the team:
  - Queries 1, 2 and 3
  - Plus data acquisition and ingestion
- 3 students in the team:
  - Queries 1, 2 and 3
  - Plus data acquisition and ingestion
  - Plus optional part A using a higher level processing framework