Project 1

Corso di Sistemi e Architetture per Big Data

A.A. 2018/19

Valeria Cardellini, Fabiana Rossi

Laurea Magistrale in Ingegneria Informatica

Project delivery

- Submission deadline
  - May 24, 2019
  - After the deadline, the maximum achievable score will be decreased by 2 points for each week of delay
- Your presentation
  - May 30, 2019
- What to deliver
  - Link to cloud storage or repository containing the project code
  - *Optional*: project report composed by 4-6 pages in ACM or IEEE proceedings format
  - Slides of your presentation (max. **15 minutes** per group), to be delivered after the presentation
- Team
  - Target: 2 students per team
  - Also possible 1 student or 3 students per team

Dataset

- You will use a real dataset
- Weather conditions on a hourly basis
  - Goal: batch analytics of weather conditions over two different countries
  - Weather information: temperature, humidity, pressure, textual description
- Reduced data set is available at http://www.ce.uniroma2.it/courses/sabd1819/projects/prj1_dataset.tgz

Dataset: schema

- Input in CSV format
- city_attributes.csv
  - For each city: latitude and longitude
- humidity.csv
  - For each city: humidity (percentage) on an hourly basis
- pressure.csv
  - For each city: pressure (measured in hPa) on an hourly basis
- temperature.csv
  - For each city: temperature (measured in Kelvin degrees) on an hourly basis
- weather_description.csv
  - For each city: description (expressed in String format) of the weather conditions on an hourly basis
  - E.g., a clear day is described with the string "sky is clear"

Queries with Hadoop/Spark

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

Query 1

For each year of the dataset, identify the cities that have at least 15 days a month in March, April and May

Query 2

Identify, for each country, the average, standard deviation, minimum, maximum temperature, pressure and humidity recorded in each month of each year

Queries with Hadoop/Spark

Query 3

Identify, for each country, the 3 cities that registered in 2017 the maximum difference in average temperature in the local time slot between 12.00 and 15.00 in June, July, August and September compared to January, February, March and April

Compare the city position with the previous year ranking

Optional part

- Compulsory for team composed of 3 students
- Use either Hive (or Pig) or Spark SQL to address the same three queries
- Include in the report the query times using a higher level framework on your reference architecture and compare them to those achieved by your pure Hadoop/Spark-based solution

Queries for the team

- 1 student in the team: queries 1 and 3
- 2 students in the team: all the three queries
- 3 students in the team: all the three queries plus optional part

Data ingestion

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