Report on Big Data project: (short subtitle)

Name Surname - Mat. 004815
Name Surname - Mat. 162342

May 14, 2020

Contents

1 Teachers' notes

The goal of the project is to assess the students' skills in writing jobs of low/medium complexity and to correctly reason about the jobs' performances. The projects must be agreed with the teachers (do not start without explicit consent) and it consists of:

- finding a complex-enough dataset: about 1GB, possibly consisting of more tables;
- loading the dataset on HDFS/Hive;
- implement an analytical job in both MapReduce and Spark;
- writing a short report to describe it.

To deliver the project, each group must **send by email to both teacher the link to their repository**. The repository must:

- be forked from the assignment available on the Github classroom: ;
- contain a report folder with the PDF of the final report (based on this template). The report can be written in either Italian/English and Latex/Word at your discretion. Be concise and go straight to the point: an excessively verbose report is a waste of time for you and for the teachers;
- contain a README.md file with the instruction to run the jobs. Indeed, the teachers must be able to clone the repository and run the jobs from their accounts on the cluster. This means that the dataset must be accessible on HDFS/Hive and the code must compile and run correctly. Please make the jobs repeatable (i.e., the job checks and possibly deletes old data to avoid errors when re-running the code).

This guide is based on the "MapReduce+Spark" kind of project. However, we remind that a different kind of project may be agreed upon.

The evaluation will be based on the following.

- Compliance of the jobs with the agreed upon specifications.
- Compliance of the report with this guide.
- Job correctness.
- Correct reasoning about optimizations.

Appreciated aspects.

- Code cleanliness and comments.
- Further considerations in terms of job scalability and extensibility.

2 Introduction

2.1 Dataset description

Please provide:

- A brief description of the dataset.
- The link to the website publishing the dataset (e.g., `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`).
- Direct links to the downloaded files, especially if more than one files are available in the previous link (e.g., `https://s3.amazonaws.com/nyc-tlc/trip+data/yellow_tripdata_2017-01.csv`).

2.1.1 File description

For each file, briefly indicated the available data and the fields used for the analyses; examples are welcome.

3 Data preparation

Please provide:

- The paths to each file on HDFS and/or its corresponding location in Hive (database and table); consider relying on the structured data lake organization.
- A subsection with details on the pre-processing of the data (only necessary if the data is dirty and/or it contains a significant amount of useless information).

4 Jobs

One subsection for each job.

4.1 Job #1: short description

Provide a brief, general description of the job. Then, one subsubsection for each implementation.

4.1.1 MapReduce/Spark implementation

Please provide:

- Input and output files/tables.
- Execution time and amount of resources.

- Direct links to the application's history on YARN (e.g., `http://isi-vclust0.csr.unibo.it:18088/history/application_15...`).
- Description of the implementation. A schematic and concise discussion is preferrable to a verbose narrative. Focus on how the data is manipulated in the job (e.g., what do keys and values represent across the different stages, what operations are carried out).
- Performance considerations with respect the (potentially) carried out optimizations, e.g., in terms of:
  - allocated resources and tasks;
  - enforced partitioning;
  - data caching;
  - combiner usage;
  - broadcast variables usage;
  - any other kind of optimization.
- Short extract of the output and discussion (i.e., whether there is any relevant insight obtained).

5 Miscellaneous

If necessary, feel free to add sections to explain any other relevant information.