Report on Big Data project: (short subtitle)

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

April 12, 2019

Contents

1 Teachers' notes

Each group should designate a reference user. In the local home directory of such user there must be an `exam` folder exclusively containing the jobs to run (e.g., MapReduce jar, Spark scala file, Spark jar). Also, please send (either by email or by sharing a Git project) the following files.

- The source code of the jobs; if more versions have been developed, only send the most efficient one.
- A text file with the commands to run the job.
- The PDF file of the report; use Italian/English and Latex/Word at your discretion. Be concise and go straight to the point; do not waste time and space on writing a verbose report.

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 name of the reference user (i.e., the one in whose home directory is the `exam` folder).
- The machine name (or IP address) of the reference user.
- 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(SQL) implementation

Please provide:

- The command to run the job from the reference user's home directory; explain possibly different parameter configurations.
- Direct link to the application's history on YARN (e.g., `http://isi-vclust0.csr.unibo.it:18088/history/application_15...`).
- Input files/tables.
- Output files/tables.
- 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.