Large Scale and Multi-Structured Data Bases
University of Pisa
Academic Year 2019-2010
Workgroup Task 2

Title: Designing and Implementing an application which interacts with a Document Database

Task 1.1: Design of the application

The workgroup ***must first design*** a complete application which manages a "big dataset" stored in a Document Database.

Constrains:

- CRUD operation must be implemented
- Some Analytics and Statistics on the dataset must be included as main functionalities of the application. Thus, at least two aggregation pipelines must be designed
- At least three indexes must be defined
- Replica set must be used (define appropriately functional and non-functional requirements)
- Some administration use cases must be defined

Dataset

The dataset must be a ***real dataset*** characterized by a ***large volume*** (at least 200mb).
It will be appreciated if the dataset will have at least one of the following features:

- Variety: Multi-sources and/or multi-format. For example, if an application will handle comments on hotels, it will be appreciated if the database will be built using different sources for comments such as TripAdvisor and Booking.
- Velocity: data may lose importance after a certain time interval since new data quickly arrives.

Web scraping and crawling algorithms can be used for retrieving real dataset from web sites and services. Their usage will be appreciated.
It is suggested to ***informally discuss*** with the instructor about the selected dataset before starting the design process.

Design

The following stages must be carried out and described appropriately in the ***Design Document*** (to be approved by the instructor):

1) To brief ***describe*** the proposed application by words, like a sort of storytelling towards non experts in the field of computer engineering.
2) To identify the ***main actors*** of the application and to define the main functional and non-functional ***requirements***.
3) To draw the UML diagram of the main ***Use Cases***
4) To draw the UML diagram of the ***Analysis Classes*** (specify at least the main fields describing each class)

5) To define the ***data model*** to be used in the document database (namely, the main entities and the structure of the collections).
6) To identify the ***architecture of the overall platform*** and the ***frameworks*** to use for the implementation of the application and of the data base.

Task 1.2: Implementation and test of the application

Once the Design document will be approved by the instructor, the workgroup will be allowed to start with the implementation.

A final document must be produced including:

- A description of the main modules of the software
- A description of the performed tests
- A short manual of usage of the application

It will be appreciated if the database will be deployed in a remote cluster (it will be provided by the instructor) and if some tests for evaluating the features of the application in relation with the CAP theorem. Some statistic and performance tests when evaluating read and write operations against the database will be appreciated.

At the end of the task, all the artifacts (reports, code, database dump and executable files) must be uploaded *only by the reference person* of the group. Avoid multiple uploads.
Groups *may use git archive*. In this case, the reference person must specify the address to download the repository. Please upload documentation always on the e-learning platform.

Deadline: December 10, 2019.