Project Statement for Milestone 3

You team name

(Team members' names)

Overview:

At the end of Milestone 3 of the project, teams should have implemented and tested their algorithms on the stored data set. The algorithms should leverage both a NoSQL database and Hadoop / Spark data processing framework. The algorithms should be designed and implemented with Big Data scale in mind. One or more algorithms should be **advanced** algorithms that leverage Hadoop/Spark's built-in methods (for example, Spark MLlib-based algorithms, Spark GraphX-based algorithms, Spark's map/filter/reduce based algorithms).

Teams should not use a hosted and managed Hadoop/Spark service for this milestone.

Teams are not required to build an end-to-end application with a Graphical user interface in Milestone 3.

All team members are expected to make a significant contribution to the project milestone tasks.

Note: Since some datasets are significantly larger than others, teams are recommended to work with a reduced dataset (>10 MB and <500 MB) during Milestones 1-3. Team would, however, still need to implement the final solution on a large dataset (>1 GB and <20 GB).

Project Report Topics:

The report should cover the following subtopics and answer the questions listed:

1. Data Files:
   a. In addition to the NoSQL database, are you using any distributed data files (e.g., Parquet files, HDFS files) for data storage? If so, describe the data files and data transformation steps.
   b. Include sample data, if applicable.

2. Algorithm Description:
   a. Give a formal description of each of the algorithms you have implemented. It should consist of (1) input, (2) output, and (3) computing operations.
   b. Provide pseudo-code of the algorithms.
   c. Discuss any optimization techniques you have implemented.

3. Algorithm Results:
   a. Provide algorithm results (output) using prepared/stored data (inputs).

b. Provide performance metrics (for example, execution time, accuracy) for the prepared/stored data.
c. Describe how you plan to present the results to the user. Perhaps you plan to run the algorithm on demand and present the results to the user, or you plan to execute the algorithms offline, store their results, and present the results on demand.

4. Algorithm Scalability:
   d. Have you implemented the algorithm with scalability in mind? If so, describe how your algorithm will scale in a distributed storage and processing environment.

5. Source Code:
   a. Provide the source code of your algorithms in a ZIP file.

Peer Evaluation:

Each team member should complete the CATME Peer Evaluation survey.

Grading:

- 50 pts: The team has successfully completed all the tasks described in this milestone template document. The team has successfully designed and implemented the algorithms, and shown good performance on the algorithms. The team has provided all relevant information, including descriptions, pseudo-code / code snippets, and diagrams / images.