Project Statement for Milestone 2

You team name

(Team members' names)

Overview:

At the end of Milestone 2 of the project, teams should have prepared the data, including reducing, cleansing, and transforming, for storage and processing. Teams should also have chosen a NoSQL database, designed a non-relational schema, and ingested the (reduced) dataset into the database. The team should have validated the data ingestion process using appropriate database queries.

Teams should not use a hosted and managed NoSQL database service for this milestone.

Teams are not required to perform distributed data processing using Hadoop or Spark in Milestone 2. They are, however, expected to choose a NoSQL database that integrates well with Hadoop and/or Spark.

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 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 Preparation:
   a. Describe the data reduction, data cleansing and data transformation steps you have performed so far. Include pseudo-code for each of these steps in your description.
   b. You may have developed a parser to transform the raw data into semi-structured data. Briefly describe the parser algorithm with the help of a pseudo-code or source code snippet.

2. Database System:
   a. What NoSQL database are you using to store the data? Does the database system scale well with the data?
   b. Describe the non-relational schema you have implemented for the data. Why is the schema an appropriate one for your project?

3. Data Ingestion and Query:
    a. Describe how you ingested the dataset into the database. Also, describe how you validated the ingestion step, perhaps through database queries.
    b. Provide performance results of your data ingestion and query operations. How would these results scale with data?

4. Source Code:
    a. Provide source code of your data preparation, data reduction, data transformation, and data ingestion steps 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 made good design decisions and implemented a good solution. The team has provided all relevant information, including descriptions, pseudo-code / code snippets, and diagrams / images.