Project Statement for Milestone 2

You team name

Team members' names

In this report you should focus on the datasets description, dataset preparation and formatting, description of data collection tools you use. Typically, the process involves writing a parser to extract the information you need, using data structures to store the data in memory and uploading / ingesting the date to a database of choice. Your report should cover the following subtopics and answer the questions listed:

1. Data Model and Database tools:
    - Describe the data model you will be use to represent the dataset? Justify why the data model is an appropriate one for the dataset.
    - What database will you be using to store the data?

2. Dataset Statistics:
    - How large is the dataset you will be dealing with? Report the following statistics for your datasets:
        i. If you are using a key-value data model, how may key-value pairs? How many unique keys? What are the data types for keys and values? Are these basic data types or data structures? What's the physical storage size (in KB/MB/GB).

        ii. If you are using a graph data set: how many nodes and edges? How many attributes are there for the nodes/edges? Is it labelled? Directed? What's the average degree of the nodes? What's the density of the graph/network data? What's the physical storage size (in KB/MB/GB).

        iii. If you are using a document data model, how may documents does your model contain? How many elements / sub-elements does each document has? What are the attributes? What's the physical storage size (in KB/MB/GB).

iv. If you are using another non-relational data model, describe your dataset statistics based on this non-relational data model. What’s the physical storage size (in KB/MB/GB).

Note: You would need to present and demonstrate a ‘Big Data’ solution. This means that you should be choosing distributed database / datastore that are readily (freely) available. For this reason, I recommend against choosing a Relational database.

- In data processing, you may need to develop a parser to transform the raw data into the format/tools you are using. Briefly describe the functions of the parser you have implemented so far.
- You should have successfully loaded the data into the database management software of your choice. What’s the estimated loading time, if you have the result?

3. Data structure and auxiliary structure.
   - What data structure have you developed of your own to represent the data, if any? For example, if you are using graphs, put up pseudo code that represent the data. If you are using a collection to represent the data, provide pseudo code on collection definition. Do you use any types of indexes? If so, briefly describe the indexes you developed for fast access the data.

4. Project Plan and Contributions:
   - List each team member’s contribution in Milestone 2.
   - What’s your plan for Milestone 3?