Project Statement for Milestone 1

Airline Search Engine

CPTS-415_BMW

Brian Joo, Noah Waxman, Mitchell Kolb

1. Problem Statement:
   a. Give a formal description of the project. What's the input and output of the problem?

      The aim of this project is to develop an aviation data search engine using MapReduce, SQL/SPARQL, and graphs using datasets from openflights.org. Values we will obtain from the data set include the airport ID, airport name, location, timezone, and more. Using this information, we aim to provide a solution which can retrieve travel recommendations seamlessly. The project will allow the user to search airports and airlines, countries with the largest number of airports, cities with popular airline travel, and finding the most efficient routes between two cities. Ultimately, the goal of the project is to improve the accessibility of avian information.

   b. Why is the problem you want to address important? What's its application?

      Airlines are the foundation for productivity, not just in America, but worldwide. Not only would the tool improve efficiency in finding and choosing airlines/airports, it would improve operational efficiency of airports, airlines, or even governmental bodies that need access to that information. Additionally, having an updated, efficient database is crucial to maintain safety standards and other regulations that airports/airlines adhere to. Furthermore, analyzing the data could improve fuel efficiency, thus leading to a less-polluted environment. Finally, this tool could also be valuable to researchers who intend to analyze trends and patterns associated with the aviation industry.

2. Project Goals:
   a. Specify the goal you want to achieve (an end-to-end application, and/or a thorough experimental evaluation of existing and new algorithms).

      The goal of this project will be to use the given datasets and to use tools such as MapReduce,SQL to implement a search engine. The user should be able to search for specific arguments such as, name of the airport, city, country, etc…. One of the key components is the trip recommendation feature. There are many subcomponents of this, such as if the user wants to fly to multiple cities. It would be important for the search engine to show the user which airports flying to city a AND to city b is possible.

3. Team Description:

   a. Who are the team members? What knowledge and skills do the team have from previous courses, projects or internships?

      Brian Joo: The knowledge and skills I have that would be more relevant to this project would be python, as we are mostly analyzing data. I have used python in many different forms, such as web development(flask/python), Data mining(315), and introduction to ML(437).

      Noah Waxman: I have experience developing python applications in both personal and educational environments, including web development (Flask, Django, Selenium), games (Pygame), and desktop applications (PyQt). I have also studied algorithms previously.

      Mitchell Kolb: I have experience and skills in data mining, algorithmic analysis, backend development, and relational database design. I think these skills will be useful when assisting my team in the development of this project.

   b. What will be each team member's roles and responsibilities in the current project?

      All members will collaboratively handle data engineering, database design, algorithm development, and user interface design in order to evenly distribute the work for the project.

4. Dataset Description:

   a. Give the description and links of the dataset.

      The description of the dataset for the "Airline Search Engine" is that it was created to collect data from airports, airlines, and the routes that they take. It contains over 10,000 airports, train stations, and ferry terminals that span the globe.

   b. Provide basic statistics on the dataset - number of files, storage size (KB/MB/GB), number of records (rows), number of attributes (columns), etc.

      This dataset contains many collections of data that are dedicated to certain aspects of the travel world. Such as airlines, airports, routes, planes, and countries. Here are the details for each dataset: airlines (400KB) 6162 rows and 8 columns, airports (1.3MB) 7698 rows and 8 columns, routes (2.27MB) 67663 rows and 9 columns, planes (5KB) 173 rows and 3 columns, and country (5KB) 261 rows and 3 columns. Each one of these categories have a single file corresponding to them and the file extension is .dat.

      Here is the link to the dataset: https://openflights.org/data.html