Project Statement for Milestone 1

DATA KEEPERS

Alhareth Aboud, Logan Kloft, Madee Barnwell, Nam Jun Lee, Stephany Lamas

1. Problem Statement

Give a formal description of the project. What are the inputs and outputs of the problem?

Data Keepers' mission is to provide easy-to-access and intuitive information on route, airline, and airport information. We plan to provide three features: airport and airline search, airline aggregation, and trip recommendation. Each feature is defined by its sets of inputs and outputs.

Airport and airline search feature:

- Given a country X, provide a list of airports operating in country X.
- Given a number of stops X, provide a list of airlines having X stops.
- Given a code share X for airlines, provide a list of airlines operating with code share X.
- Given the option to display active airlines, provide a list of active airlines if selected.

Airline aggregation feature:

- Given the option to display the country or territory with the highest number of airports, provide the country or territory with the highest number of airports if selected.
- Given a number of cities X, provide the top X cities with the most incoming and outgoing airlines.

Trip recommendation feature:

- Given two cities X and Y, provide a list of routes connecting cities X and Y.
- Given two cities X and Y and a number of stops Z, provide a list of routes connecting cities X and Y with fewer than Z stops.
- Given a city X and number of stops Y, provide a list of cities that can be reached from X within Y stops.
- Given all routes, provide the transitive closure of the graph of all routes.

Why is the problem you want to address important? What is its application?

Travel has become an important, and sometimes necessary, part of peoples' lives. Having access to a wealth of information about airports, airlines, plane routes, and countries of operation allow travelers to be self-sufficient when planning travel. A few applications of this program include being able to find a route from point X to Y with the fewest stops, or to plan a multi-city journey that will take the traveler on their bucket list vacation. Users of this system will be able to plan their travel around a desired airline so they can receive miles or points towards future travel. It is also beneficial for companies who regularly book air travel for their employees to be able to determine the most efficient route to a particular destination.

Specify the goal you want to achieve:

The queries needed to support the features of this program are clearly made over a graph data structure. For this reason, the Data Keepers have decided that completing a thorough experimental evaluation of existing graph traversal algorithms will help them understand how they work and when they should be used. Some of the metrics to measure are the total visited nodes, unique visited nodes, memory size, and the time to execute the algorithm. Apache Spark, one of the tools that will be used, has an API called GraphX which contains commonly used graph traversal algorithms. GraphX also claims to allow custom graph algorithms and works with collections such as DataFrames.

2. Team

Team Members:

Alhareth Aboud, Logan Kloft, Madee Barnwell, Nam Jun Lee, Stephany Lamas

Knowledge and Skills from previous courses

The Data Keepers have 22 semesters of combined experience studying computer science, statistics, and management information systems (MIS). Relative programming knowledge includes the completion of 15 different computer science courses in C/C++, C#, Python, Java, JavaScript, CSS, and HTML. The Data Keepers are also experienced in R and SQL; topics covered in various statistics and MIS courses. Through coursework at WSU, the Data Keepers have also gained proficiency in a wide range of programming and analytics tools including GitHub, Git, Visual Studio, Visual Studio Code, IntelliJ, PyCharm, R Studio, pgAdmin, Jupyter Notebook, GitLab, Spyder, Figma, and WSL.

Team Member Roles and Responsibilities:

Alhareth Aboud: Programmer, Database Manager
Logan Kloft: Canvas Group Leader, Researcher, Programmer, Repository Maintainer
Madee Barnwell: Report Designer & Editor, Researcher, Programmer, Document Submitter
Nam Jun Lee: Researcher, Programmer
Stephany Lamas: Discord Communication Manager, Report Designer, Programmer

3. Dataset and Tools

Link and Description of Dataset:

http://openflights.org/data.html

The Airport, Airline, and Route dataset is a publicly available dataset containing the following databases: Airport, Airline, Route, Plane, Country, and Schedule. Among numerous other attributes, the databases contain flight details of various airlines, including airport ID, airport name, the main city served by the airport, the country or territory where the airport is located, airport code, and time zone.

Tools for implementing project:

Apache Spark

Apache Spark is a big data distributed processing platform that is fast and provides many analysis libraries for tasks that require repetitive processing to operate on in-memory. Using the Spark SQL provided by this tool, the Data Keepers will derive useful results on the path of airport and airline searches, airline aggregation, and travel recommendations, which are the objectives of the project.

Pandas

Python will be used to load the data into Apache Spark using functions from the Pandas module.

GitHub

The team will use GitHub to collaborate using a single code base and to store the .csv files for the data.

Jupyter Notebook

The team is considering the use of Jupyter Notebooks because of its ease and how it shows the execution of Python code in a more visual manner. This might take place of hosting a website that requires a web server ($$) to run since the project will require a backend for its features.

4. Project Progress and Contributions:

Dataset Preparation:

The first step taken was to upload the data sets to GitHub so they would be accessible to all team members who clone the shared code repository. To prepare the datasets, the team needs to clean and format them so that they can fit the storage schema. Thus far, the Data Keepers have discussed considerations to keep in mind while preparing the data sets. Those considerations are as follows:

Transformations to fit database:

Apache Spark can operate on a wide variety of databases, so the team needs to consider those as options. Once a database has been selected, the team will then have to consider what needs to be done to transform the data to fit the database.

Remove unnecessary information

Not all the information is necessary for the queries of the features. Because of this, the team will trim some attributes of the data sets.

Cleaning data

The .csv files contain rows that have missing information. The team has not considered what will be done to address this problem but will make a decision and take action prior to the completion of Milestone 2.

Team Member Contributions:

Alhareth Aboud will be helping the Data Keepers with programming, taking the lead in managing the database of the project, and assisting the team with other things as well.

Logan Kloft created the canvas group for the team so that one person can submit the assignment for the group. He has started working on setting up the GitHub repository to make it easy to download the needed tools for the project without any hiccups. With respect to Milestone, 1 he filled out the formal description of the project, filled out the goal of the project, and listed his skills for Madee to turn into a cohesive paragraph along with everyone else's skills. He provided the team with a list of possible team member roles and responsibilities and listed his roles. He also added a few tools to the list that Nam Jun started for implementing the project, expanded on what the team has done and what needs to be done for preparing the data sets, and filled out the plans for Milestone 2.

Madee Barnwell's Milestone 1 contributions included detailing the application of the program and why it is important. She provided the description of, and link for, the dataset to be used for this project. She consolidated the group's experience, skills, and coursework into a narrative for the report. She listed the roles she will be taking on throughout the project. She designed, formatted, edited, and submitted the final Milestone 1 report.

Nam Jun Lee's Milestone 1 contributions included narrowing down the choice of which project to proceed with in Discord and helping the other team members understand which tools would be the best to use. He listed the tools to be used, as well as their descriptions and explanations as to why they will be used to start the project. He listed his skills and roles, and actively participated in the team members' Discord discussions.

Stephany Lamas's Milestone 1 contributions included organizing the creation of the team by opening lines of communication via the Microsoft Teams chat. She also created the Data Keepers Discord Server. Additionally, she listed the roles she will take on during the project.

Plan for milestone 2

Milestone 2 focuses on analyzing and extracting the airport, airline, and route datasets. The Data Keepers must consider the tools that have been selected to extract the relevant data. They are heavily considering using Python to parse and extract the data files which are in .csv format, in which case they would use a Python module called Pandas which can read .csv files.

Once the files are extracted using Pandas, the team plans to use Apache Spark's DataFrames to perform the analysis and manipulation of data since it is built for large data sets. All of this can be done in a load.py file where pyspark will be used to call the Apache Spark APIs.

The Data Keepers have also considered how written code will be shared amongst each other and have decided to use a private GitHub repository. Using GitHub will allow the team to work together on the same code base and maintain a backup at the same time with little interruption to their separate workflows.