## COMP7305 Term Project Proposal

## Project Title: The Prediction of Delay Flights

## Group No: 21

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Name** | **UID** | **Email** |
| Group Leader | Cai Jiajun | 3035561431 | jjcai@connect.hku.hk |
| Member | Chen Yiming | 3035562801 | u3556280@connect.hku.hk |
| Member | Liang Zhihao | 3035561651 | kelliang@connect.hku.hk |
| Member | Luo Difan | 3035562851 | u3556285@connect.hku.hk |

## Project Description

The project will use a machine learning model to predict whether the flight is delayed. The dataset we used is an open source dataset which contained the flight records of US from 1989 to 2008 and the size of entire dataset is up to 12GB. The whole project will be built on an 8 VMs cluster using a large dataset of size more than 5GB. For better user interaction, this project will also design a web UI to display details and result.

## Software to be installed

* Apache Spark (with Hadoop): <http://spark.apache.org/>
* Apache Spark SQL (with Hadoop): <https://spark.apache.org/sql/>
* OpenRefine: <http://openrefine.org/>
* Apache Maven: <https://maven.apache.org/>

## Open-source programs to be ported and integrated:

We have found some open-source repositories, which may help us to accomplish our design in this project.

* MLlib, Apache Spark's scalable machine learning library.:

<https://spark.apache.org/mllib/>

* D3.js, a JavaScript library for manipulating documents based on data which help us to build the front-end website interface

<https://d3js.org/>

## Data Collection

We may use the flight record of multiple years of the following flight record datasets according our need so that the total size of the dataset used will be larger than 6GB:

**Data Expo DB: Airline on time data (More than 12GB):**

The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed.

■ http://stat-computing.org/dataexpo/2009/

## Job Division

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tasks** | | **Group leader** | **Member** | **Member** | **Member** |
| **Jiajun** | **Yiming** | **Zhihao** | **Difan** |
| Data | Data collection |  | ✔ | ✔ |  |
| Data pre-processing |  | ✔ |  | ✔ |
| Database organization | ✔ |  |  | ✔ |
| Spark  Cluster | Deploy the project to Spark cluster and accomplish the  training process | ✔ | ✔ | ✔ | ✔ |
| Cluster maintenance | ✔ | ✔ | ✔ | ✔ |
| Performance  Evaluation | ✔ | ✔ | ✔ | ✔ |
| Others | Web UI Design | ✔ |  |  | ✔ |
| Prepare presentation PPT | ✔ | ✔ |  |  |
| Live demo | ✔ | ✔ | ✔ | ✔ |
| Demo video preparation |  |  | ✔ |  |
| Write the final report | ✔ | ✔ | ✔ | ✔ |

\*\* All members will give the presentation