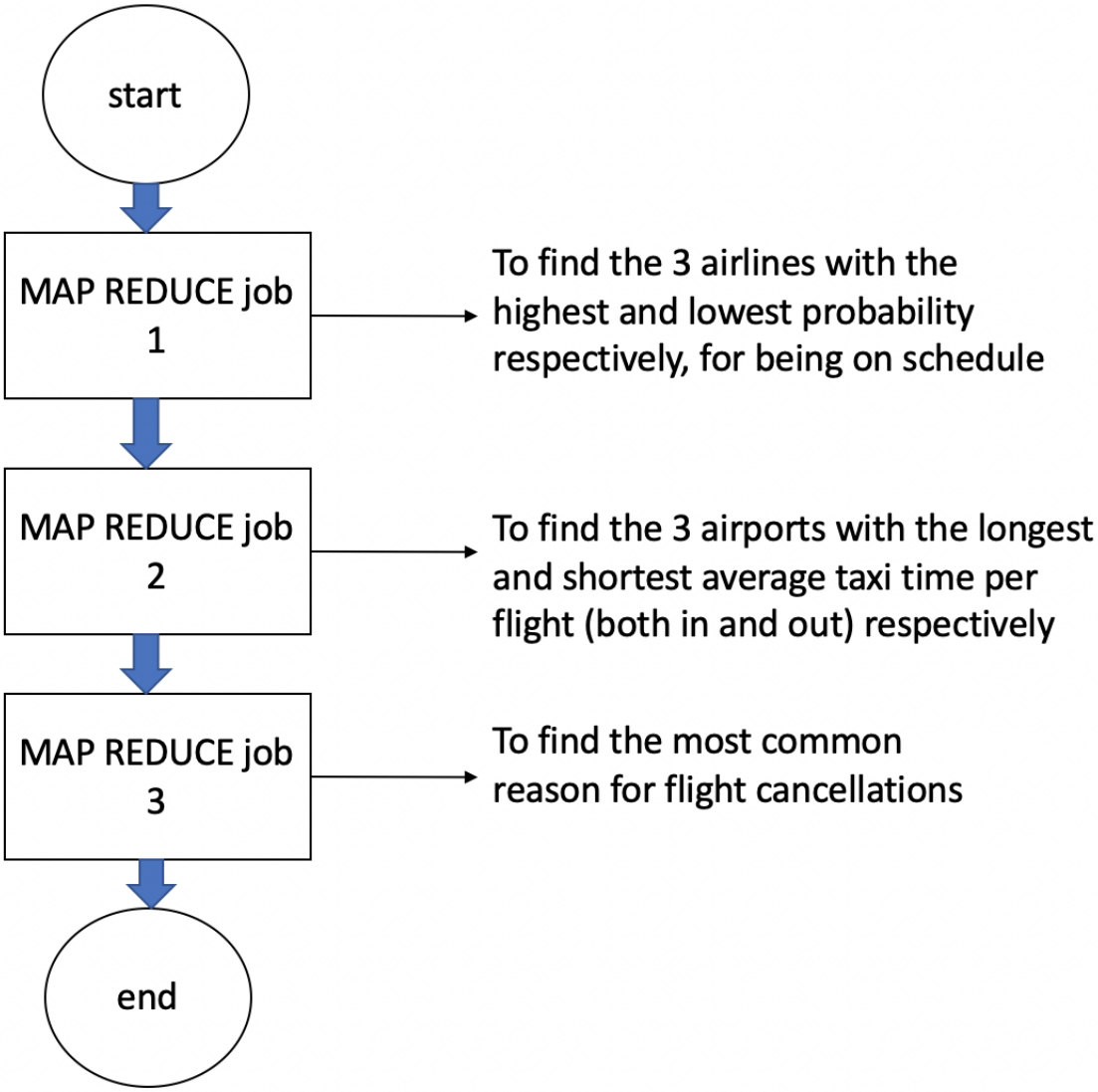
# PROJECT REPORT
# ON
# FLIGHT DATA ANALYSIS

**Subject-** CS644 (Introduction to Bigdata)

**Professor -** Daqing Yun

**TEAM:**
Sirisha Bojjireddy - sb2423
Swetha Mahesh - sm277
Divya Guduru – dg499

**STRUCTURE OF OOZIE WORKFLOW:**



**ALGORITHM:**

1. Create an input folder with all the csv files from the given flight data set.
2. Created a scala class with all the attribute names.
3. Created an object in scala with main function.
4. Initialized variables
5. Built a scala application using local machine
6. Load the flight input raw dataset into spark SQL database
7. Convert the data to a dataframe
8. Convert the dataframe to a dataset
9. Created a relation in Scala SQL
10. Query the table as per the requirements
    a) Select all the data from the flight data (large data- for validation purpose)
    b) Mapreduce job1 -> To find the 3 airline with the highest and lowest probability
    c) Mapreduce job2 -> To find the 3 airports with the longest and shortest average taxi time per flight
    d) Mapreduce job3 -> to find the most common reason for flight cancellations

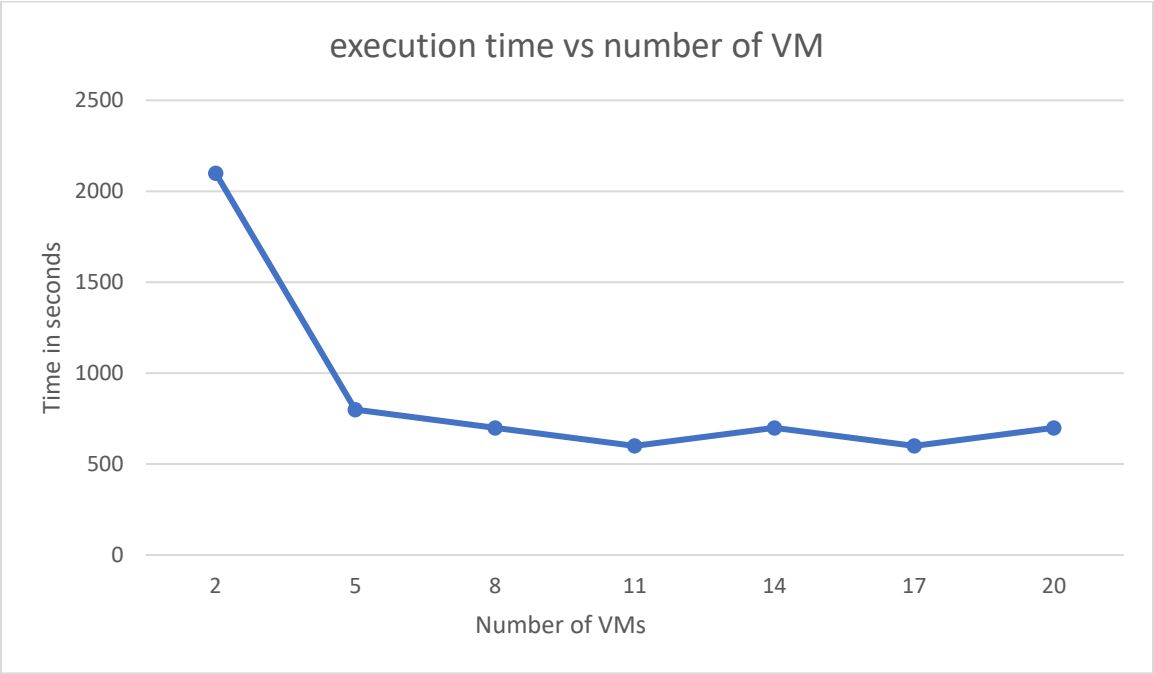
11. The jar file is generated, which is used to run the oozie workflow
12. On error, kill

**To run a oozie application**

```
oozie job -oozie http://localhost:11000/oozie -config finalproject/spark-FlightAnalysis/map-reduce/job.properties -run
```

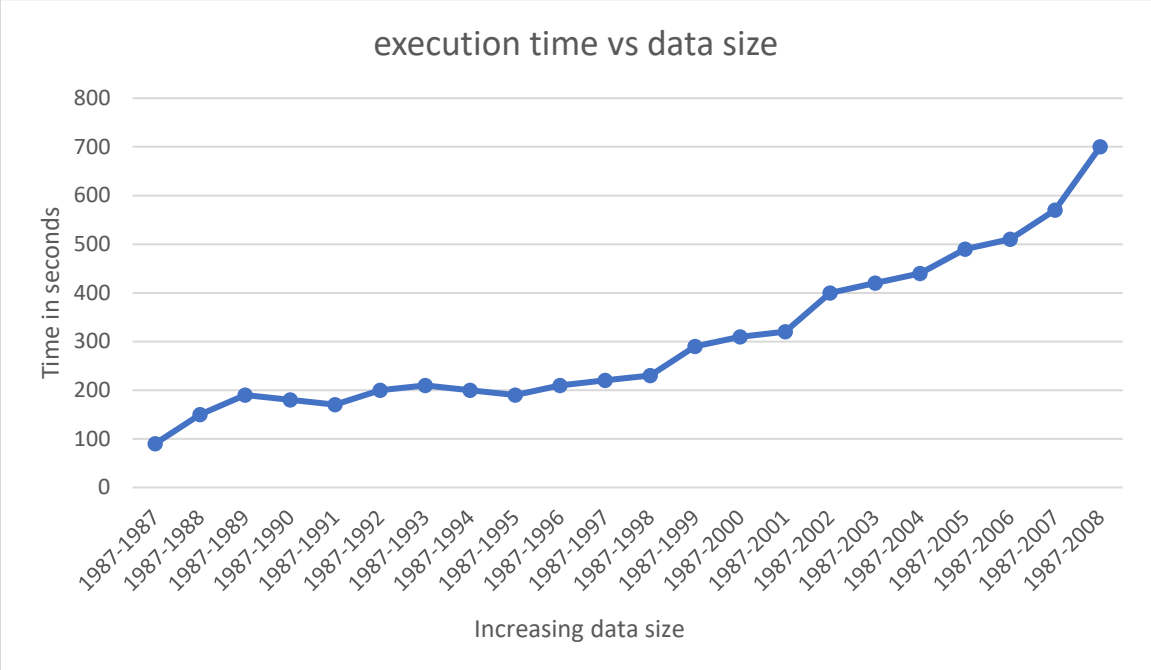
## PERFORMANCE MEASURE PLOTS

### A) WORKFLOW EXECUTION TIME VS NUMBER OF VMS USED




From the plot, it is illustrated that with increase in the number of VM, the execution time decreases. When we split the task paralelly with more number of nodes, the ability of Hadoop cluster increases thus reducing the mapreduce jobs and the oozie workflow.

## B) WORKFLOW EXECUTION TIME VS DATA SIZE USED




From the plot, it is illustrated that with increase in the data size, the execution time increases. When we keep adding more data every year, the time is increased which implies more people chose to travel by flight.