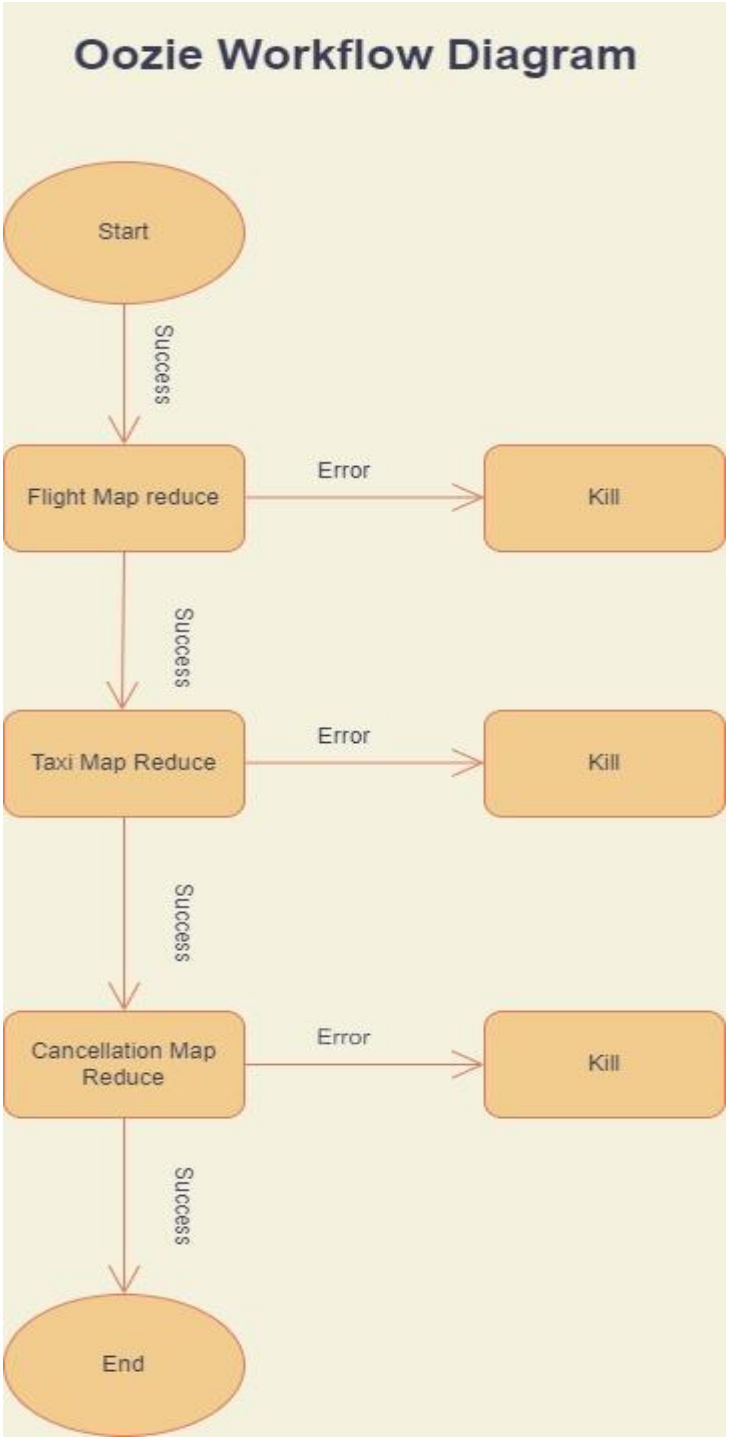
Flight Data Analysis

Team Members:

Marri Hari Prasad Reddy-hm458@njit.edu
Avinash Sunku-as4324@njit.edu

Flight Data Analysis Project Report

A. Oozie Workflow Diagram



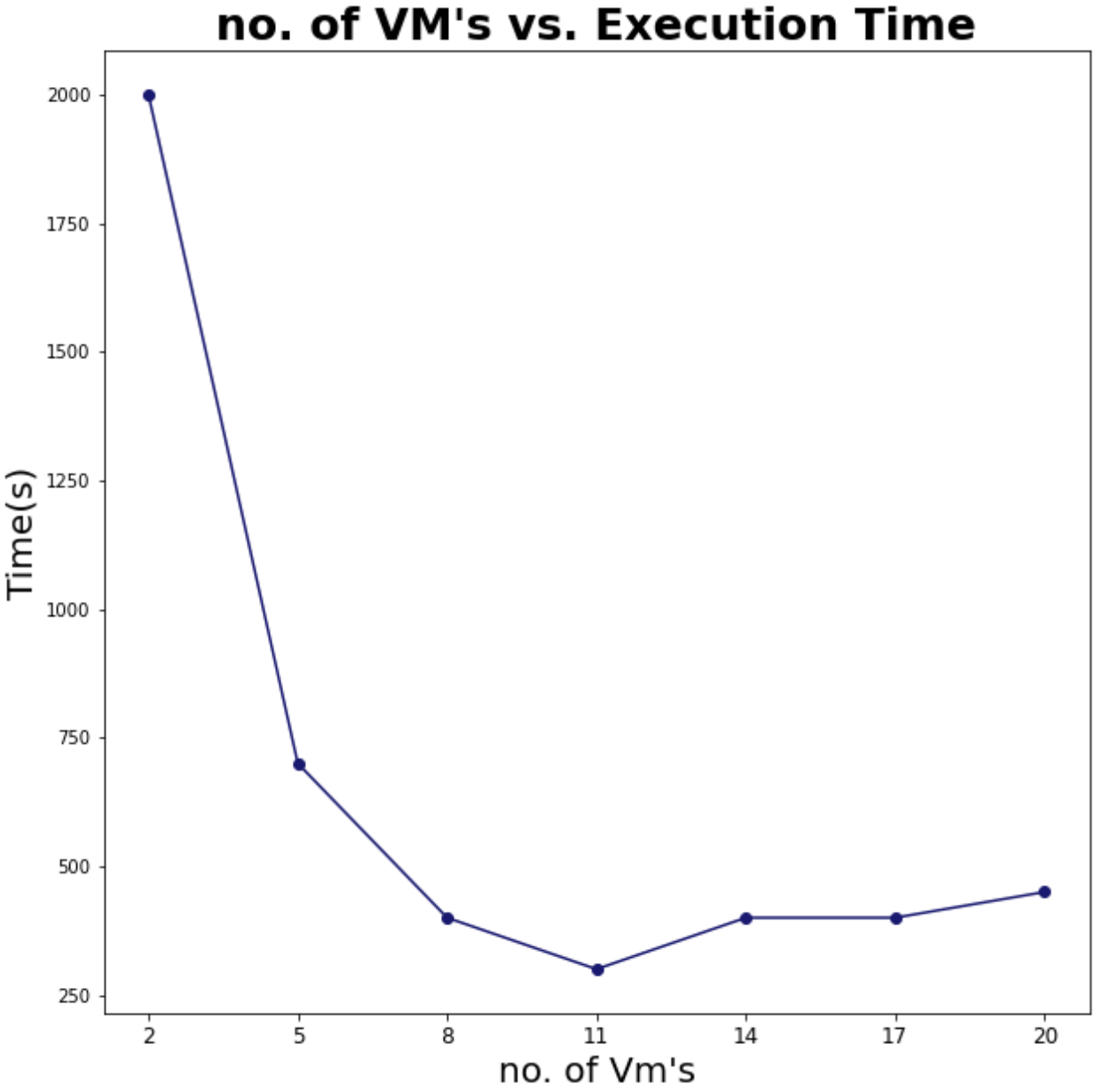
B. Algorithm

First Map-Reduce: On Schedule Airlines

1. Mapper <key,value>:<UniqueCarrier,1or 0>
2. The Mapper reads each line of data one at a time, ignoring the first line and NA data.
Produce UniqueCarrier,1> if the Arr Delay column contains data that is less than or equal to 10 minutes; otherwise, produce UniqueCarrier,0>.
3. Reducer:UniqueCarrier,probability>:key,value>
4. The total of the values from the mapper of the same key will be the number of this airline when it is on schedule. Calculate the total number of 0 and 1 before calculating the probability of this airline arriving on time.
5. After the Reducer, sort using the Comparator function.
After sorting, output the three airlines with the highest and lowest likelihood.
6. No output is available if the data is NULL.

SecondMap-Reduce:Taxi Time at Airports
1. Mapper:: or 2. The Mapper reads the data line by line, ignoring the initial line. Print the following if the TaxiIn or TaxiOut column data is not NA: Reducer: 4. Reducer adds the value from the same key's mapper (normal) and determines the total number of times this key has been detected. IATA airport code, TaxiTime> (all). The average TaxiTime for each key is then calculated using the equation normal/all.
5. After the Reducer, sort using the Comparator function. Identify the three airports with the longest and shortest average cab waits after sorting.
6. No output is available if the data is NULL since no value may be used.
ThirdMap-Reduce:Cancell Reasons
1. CancellationCode, 1> Mapper key, value
2. The Mapper reads the data line by line, ignoring the first line. Print CancellationCode, 1> if the Cancelled value is 1 and the CancellationCode is not NA.
3. Reducer: CancellationCode, sum of the 1s
4. The reducer adds the value from the mapper for the same key.
5. After the Reducer, sort using the Comparator function.
After sorting, output the most common reason for flight cancellations.
6. Print the following if the data is NULL: Several causes contribute to flight cancellations.

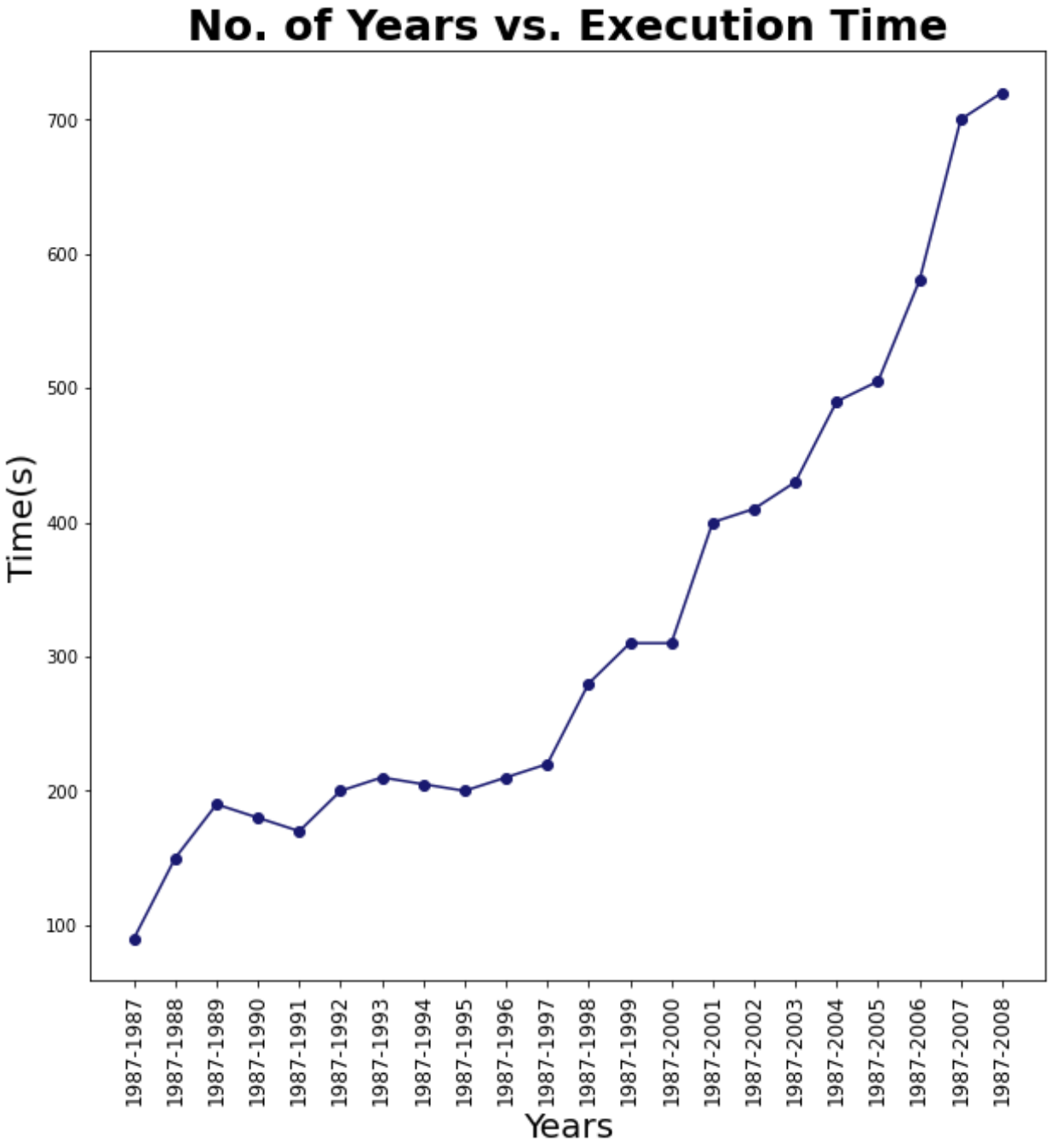
C. A performance measurement plot that compares the workflow execution time in response to an increasing number of VMs



As per the graph above, as the number of VMs grows, the workflow execution time decreases. Because data may be handled in parallel on more datanodes, the Hadoop cluster's processing capacity will increase as the number of VMs grows. Then every map-reduce operation will take less time to complete than previously, and the oozie workflow will also take less time. However, increasing the number of VMs does not always result in a faster trade execution time when the data size remains the same. When the execution time falls within a certain range, adding

more virtual machines will no longer reduce the execution time. The reason for this is that having more VMs means having more time for data to communicate across hadoop datanodes.

D. Increasing data Size



In this experiment, we want to investigate how performance changes as the input data changes.
We'll utilize two virtual machines for this experiment. We'll start by running the process on just one data file (1987.csv). The execution time is really short, as we can see. We are currently increasing the data by one year for each run and logging the execution time. The execution time increases as the amount of input data increases. As a result, performance and input data magnitude are inversely connected.