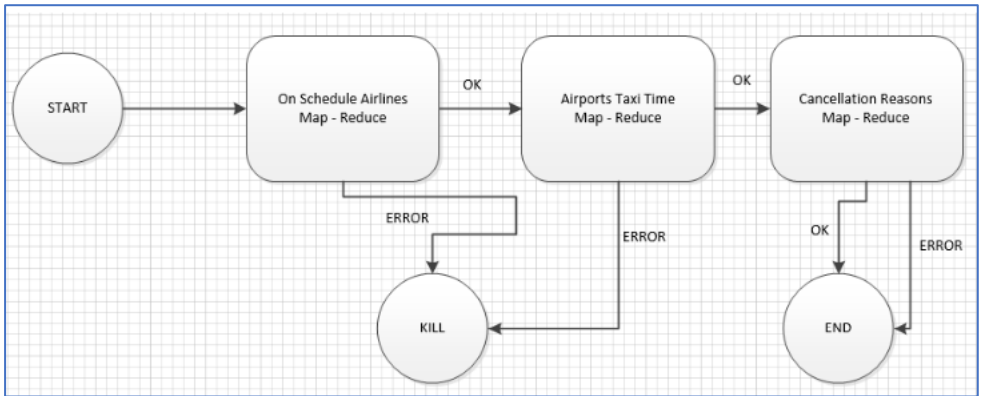
FLIGHT DATA ANALYSIS

Members:

Name: Prem Kumar Maharajan

Course: CS644

- ## Structure of Oozie Workflow



- ## Algorithm Descriptions

FlightOnSchedule:

- Mapper <key, value>:<UniqueCarrier, 1 or 0>
- Mapper reads the data line by line (ignores the first line and the NA data)
- If the value of the ArrDelay column is less than or equal to 10 minutes output: <UniqueCarrier, 1> else output: <UniqueCarrier,0>
- Reducer <key, value>:<UniqueCarrier, probability>
  Probability = (No of 1) / (No of 1 and 0)
- Reducer sums the values from the mapper of the same key and calculates the total number of 0 and 1 and then calculate the on-schedule probability of the airline
- Reducer then uses the Comparator function to perform sorting. After sorting, output the three airlines with the highest and lowest probability
- If the data is Null, then output “No Possible Output”

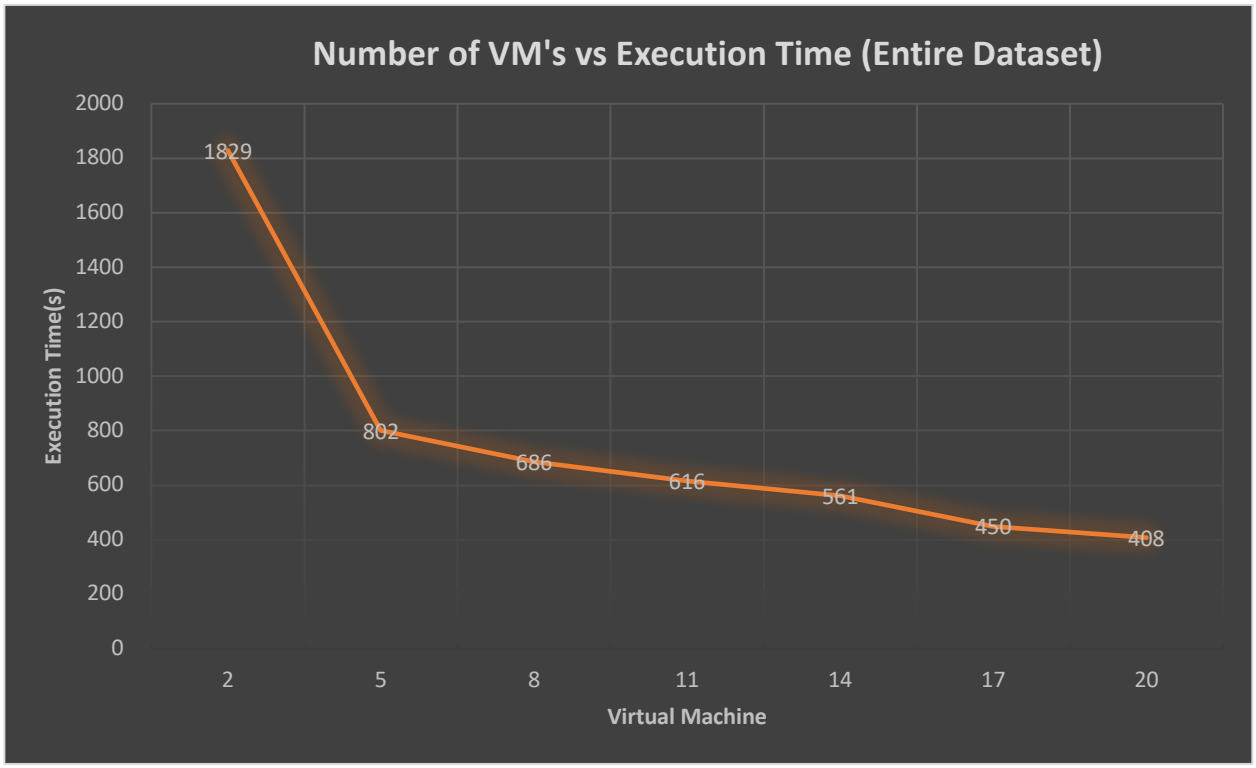
FlightCancellation:

- Mapper <key, value>: < CancellationCode, 1>
- Mapper reads the data line by line (ignores the first line and the NA data)
- If the value of Cancelled is 1 and CancellationCode is not NA, output:
  < CancellationCode, 1>
- Reducer <key, value>: < CancellationCode, sum of 1>
- Reducer sums the value from the mapper of the same key
- Reducer then uses the Comparator function to perform sorting. After sorting, output the most common reason for flight cancellations
- If the data is Null, then output “No common reason for flight cancellations”

TaxiTime:

- Mapper <key, value>: <IATA airport code, TaxiTime>: <Origin, TaxiOut> or <Dest, TaxiIn>
- Mapper reads the data line by line (ignores the first line and the NA data)
- If the data of the TaxiIn or the TaxiOut column is not NA, output: <IATA airport code, TaxiTime>
- Reducer <key, value>: <IATA airport code, Average TaxiTime>
- Reducer sums the value from the mapper of the same key and then calculates the total time the key is found and then divides the two to find the average TaxiTime of each key
- Reducer then uses the Comparator function to perform sorting. After sorting, output the three airports with the longest and shortest average taxi time
- If the data is Null, then output "No Possible Output"

• Performance Measurement Plot: No of VM vs Entire Dataset Execution Time

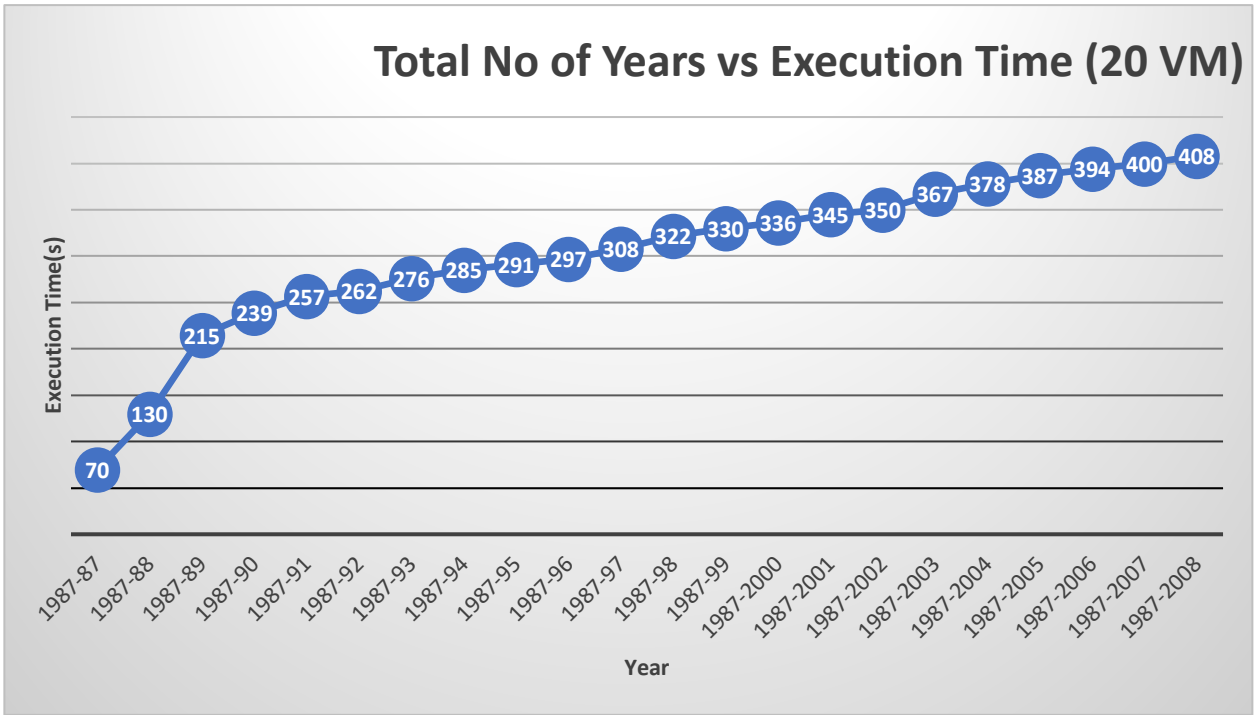


From the above plot, we can confirm that as the number of VMs increases, the workflow execution time significantly decreases to a certain point.

This is because by increasing the number of VMs we are increasing the processing capability of the Hadoop cluster and the data can be handled in-parallel on more data nodes. This, in-turn reduces the execution time of the map-reduce job and thus the execution time of the oozie workflow.

However, the execution time will not always decrease significantly by increasing the number of VMs (evident when no of VM is increased from 17 to 20). When the execution time decreases to a certain threshold, increasing the number of VM will not have a huge impact as the interaction time between the data nodes of a Hadoop cluster increases. This sometimes will have a negative impact on the execution time (execution time increases after a certain no of VM)

- ## Performance Measurement Plot: 20 VM vs Increasing Dataset Execution Time



As per the plot, we can confirm that with the increase in dataset size, the execution time also increases. This shows that more and more people started to travel by Flight from 1987 to 2008.