FLIGHT DATA ANALYSIS

DS644101

Professor: Yajaun Li

Students:

1. Sai Kiran Reddy Chandupatla /sc2565@njit.edu
2. Yashaswini Matam /ym345@njit.edu
3. Geetha Raja Rajeswari Devi Vegesna /gv244@njit.edu
4. Bhaskara Venkata Sai Kumar Rayala /br387@njit.edu

A. Structure of Oozie workflow

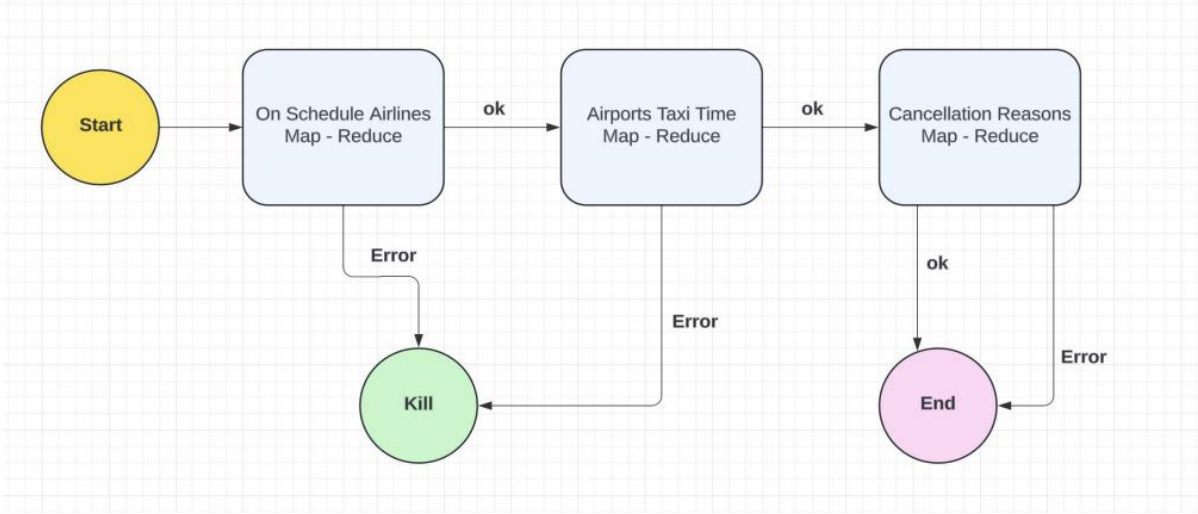

Figure 1

B. Algorithm

On Schedule Airlines was the first map-reduce.

1. Mapper:<UniqueCarrier,1 or 0> <key,value>

2. Ignoring the first line and the NA data, the mapper read the data line by line. Output: <UniqueCarrier,1> if the ArrDelay column's data is less than or equal to ten minutes; else, output: <UniqueCarrier,0>

3. Reducer: UniqueCarrier, probability: <key,value>
    a. Probability is equal to (#of1) / (#of 1 and 0).

4. The total is the airline's number when it operates on schedule. The reducer adds together the values from the mapper for the same key. Next, tally up all the 0s and 1s to find the airline's on-schedule probability.

5. The Reducer uses the Comparator function to sort the data. After sorting, output the three airlines with the highest and lowest probability.

6. If the data is NULL, there will be no output since no value can be utilized.

Second Map-Reduce: Airports Taxi Time

1. <Origin,TaxiOut> or <Dest,TaxiIn> is the TaxiTime for Mapper <key,value>: <IATA airport code.

2. Ignoring the first line, the mapper scanned the data line by line. If the TaxiIn or the output is <IATA airport code, TaxiTime> if the TaxiOut column is not NA.

3. Reducer <key,value>: <Average TaxiTime, IATA airport code>

4. Compute the value from the mapper of the same key (normal) using the reducer sum. The total number of times (all) this key is located. Next, solve the normal/all equation. Determine each key's average taxi time.

5. After that, the Reducer sorts the data using the Comparator function. List the three airports with the longest and shortest average cab times after sorting.

6. If there is nothing in the data, output: There is no output because no value can be used.

Third Map-Reduce: Cancellation Reasons

1. Mapper: <CancellationCode, 1> <key,value>

2. Ignoring the first line, the mapper scanned the data line by line. Output: <CancellationCode, 1> if the CancellationCode is not NA and the Cancelled value is 1.

3. Lowering <key,value>: < CancellationCode, total of the 1s>

4. The value from the mapper of the same key is summed by the reducer.

5. After reducing, sorting is done using the Comparator function. Determine the most frequent cause of flight cancellations after sorting.

6. If the data is NULL, the following output will appear: The most frequent cause of flight cancellations

C. Increasing number of VMs (entire data set)

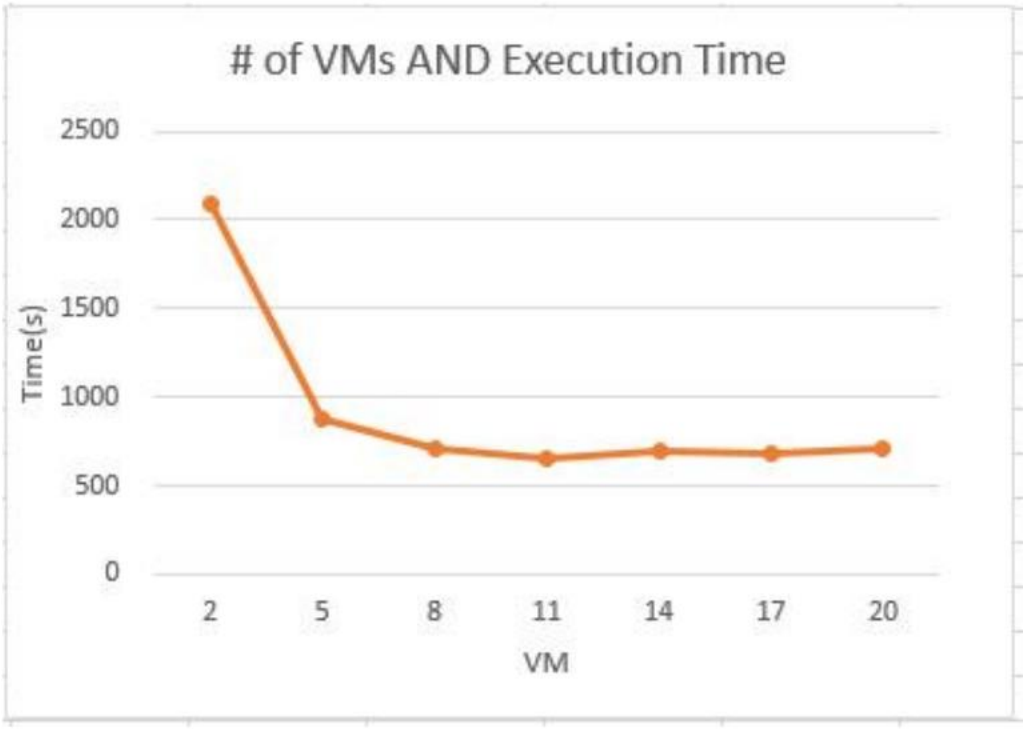

Figure 2

Figure 2 indicates that the process execution time will reduce as the number of virtual machines (VMs) increases. The processing capacity of the Hadoop cluster will improve as the number of virtual machines (VMs) increases since more data nodes can handle the data in parallel. Every map-reduce job will thus be executed faster than it did previously, which will also result in a quicker oozie workflow execution time. Nevertheless, adding more virtual machines (VMs) won't always result in a faster execution time for dealing with the same amount of data. Even with an attempt to expand the number of virtual machines (VMs), the execution time will stop lowering once it reaches a certain point. This is because a Hadoop cluster with more virtual machines (VMs) has more time for information exchange across its datanodes. A Hadoop cluster's information interaction time grows as the number of virtual machines (VMs) rises.

D. Increasing data size (20 VMs)

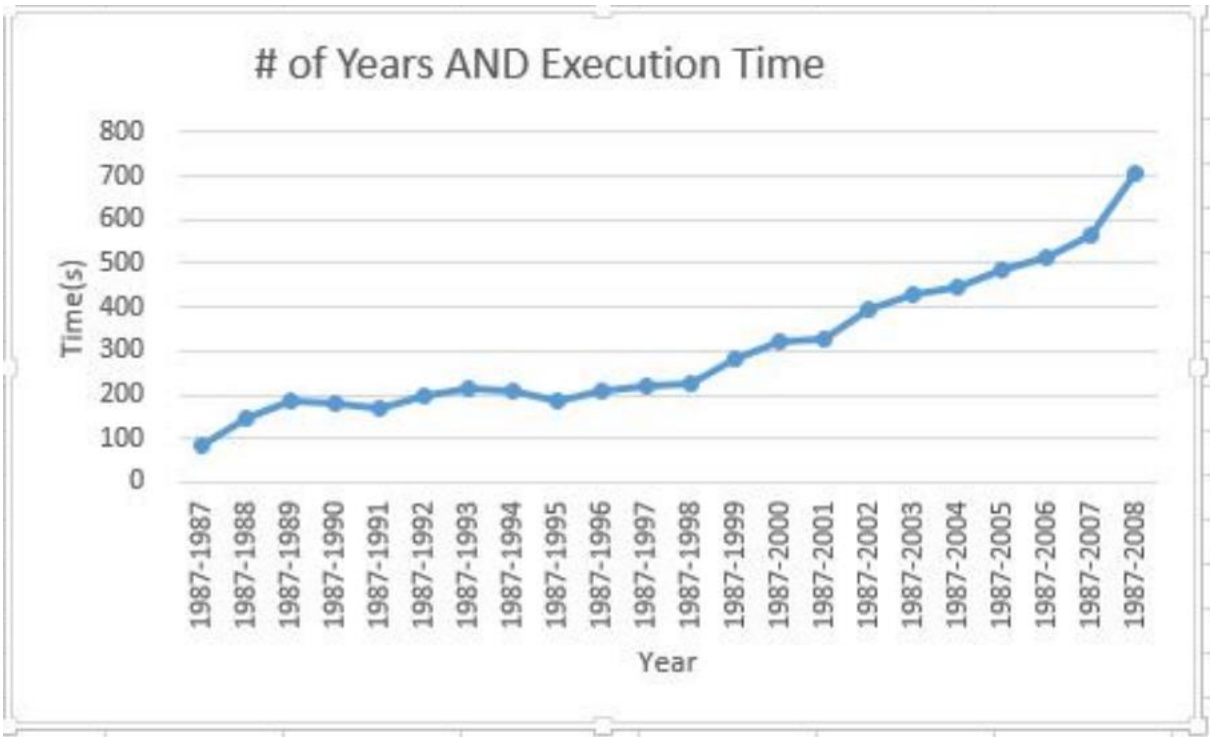

Figure 3 indicates that the oozie workflow's execution time will always increase in tandem with the size of the data. Because there hasn't been much data growth in the first few years, the time-consuming increase initially increases along with the amount of data, but this increase is gradual. In contrast, the time-consuming increased quickly after 1998, and the slope became considerably steeper than it had been in the earlier years. The cause is that, compared to prior years, flight data increased more quickly between 1998 and 2008. It also demonstrates the growing trend of consumers choosing air travel.

Some environment setting of the Hadoop cluster

Instance Information: Ubuntu Server 16.04 LTS (HVM), SSD Volume Type - amif4cc1de2 Family: General purpose
Type: t2.medium
vCPUs: 2
Memory (GiB): 4
Instance Storage (GB): 24GB

Environment of master instance and slave instances:
The master and slave instances are set up with the following environment configurations:

```
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export PATH=${JAVA_HOME}/bin:${PATH}
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
export PATH=$PATH:/usr/local/hadoop/bin
```

export OOZIE_HOME=/usr/local/oozie/distro/target/oozie-4.3.0-distro/oozie-4.3.0
export PATH=$PATH:$OOZIE_HOME/bin

Local Operating System:
The local operating system is Windows, and the user utilizes MobaXterm for managing instances and files.
Note: To modify the input file, it is necessary to update the inputFilePath in the job.properties file provided. Users should adhere to the specified format within the job.properties file.
Hadoop and Oozie Versions:
The Hadoop version employed is 2.8.0, while the Oozie version is 4.3.0