Flight Data Analysis

CS 644: Introduction to Big Data

Professor: Chase Wu

Team Member:
Saumya Jain - SJ634

a. Structure of Oozie workflow

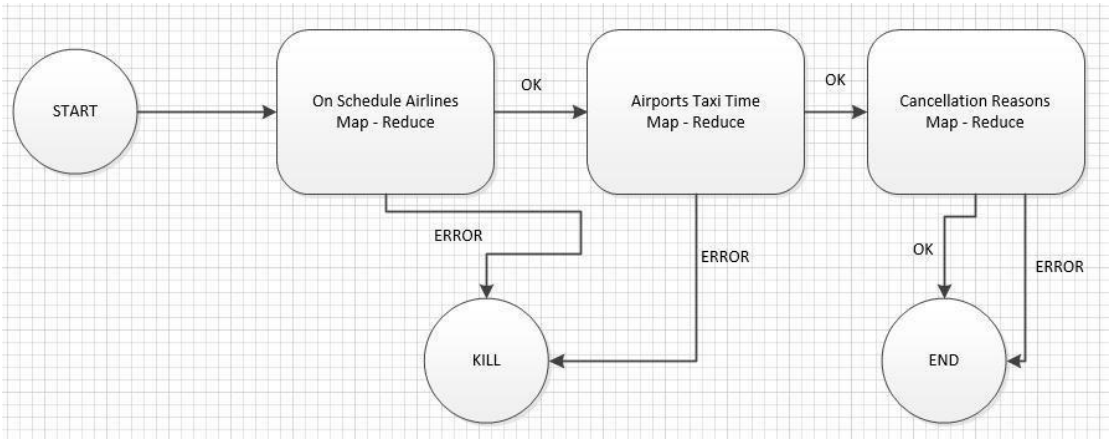

Figure 1

b. Algorithm

First Map-Reduce: On Schedule Airlines

1. Mapper <key,value>:<UniqueCarrier,1 or 0>
2. The Mapper read the data line by line, ignore the first line and the NA data. If the data of the ArrDelay column which is less than or equal to 10 minutes, output: <UniqueCarrier,1>, otherwise output: <UniqueCarrier,0>
3. Reducer <key,value>:<UniqueCarrier,probability>
   Probability = (# of 1) / (# of 1 and 0)
4. Reducer sum the values from the mapper of the same key, the sum will be the number of this airline when it is on schedule. And calculate the total number of 0 and 1, then calculate the on-schedule probability of this airline.
5. Reducer then use the Comparator function do the sorting. After sorting, output the 3 airlines with the highest and lowest probability.
6. If the data is NULL, then output: There is no value can be used, so no output.

Second Map-Reduce: Airports Taxi Time

1. Mapper <key,value>: <IATA airport code, TaxiTime>: <Origin,TaxiOut> or <Dest,TaxiIn>
2. The Mapper read the data line by line, ignore the first line. If the data of the TaxiIn or the TaxiOut column is not NA, output: <IATA airport code, TaxiTime>
3. Reducer <key,value>: <IATA airport code, Average TaxiTime>
4. Reducer sum the value from the mapper of the same key (normal) and calculate the total times this key is found (all). Then do the equation: normal/all to calculate the average TaxiTime of each key.

5. Reducer then use the Comparator function do the sorting. After sorting, output the 3 airports with the longest and shortest average taxi time.
6. If the data is NULL, then output: There is no value can be used, so no output.

Third Map-Reduce: Cancellation Reasons

1. Mapper <key,value>: < CancellationCode, 1>
2. The Mapper read the data line by line, ignore the first line. If the value of the Cancelled is 1 and the CancellationCode is not NA, output: < CancellationCode, 1>
3. Reducer <key,value>: < CancellationCode, sum of the 1s>
4. Reducer sum the value from the mapper of the same key.
5. Reducer then use the Comparator function do the sorting. After sorting, output the most common reason for flight cancellations.
6. If the data is NULL, then output: There is no the most common reason for flight cancellations.

c. Increasing number of VMs (entire data set)

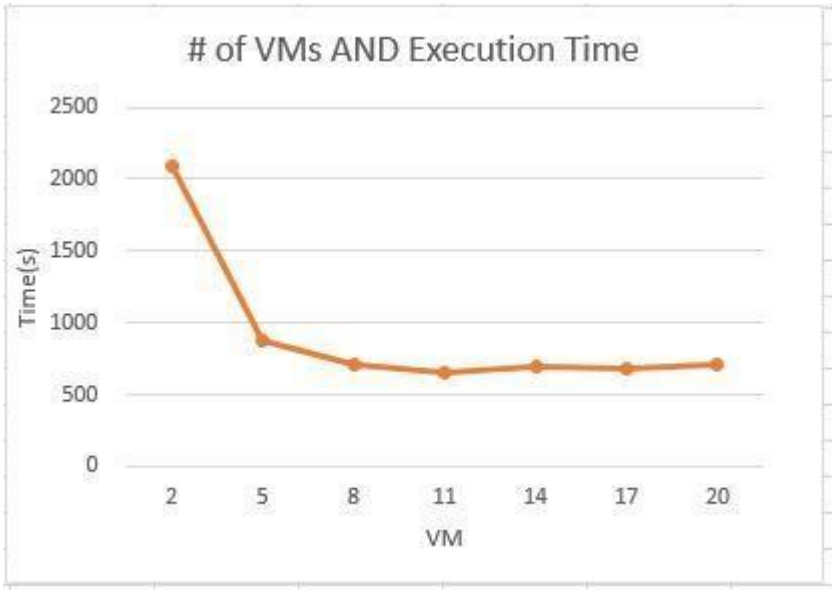

Figure 2

According to the Figure 2, along with the increasing the number of the VMs, the workflow execution time will decrease. By increasing the number of the VMs, the processing ability of the Hadoop cluster will also increase, because the data can be dealt with in parallel on more data nodes. Then the execution time of every map-reduce job will be shorter than before, thus the execution time of the oozie workflow will be shorter than before too. However, the execution time of deal with the same data size will not always decrease by increasing the number of VMs. When the execution time decrease to a certain range, although trying to increase the number of VM, the execution time will no longer decreasing anymore. The reason is more VMs means more information interaction time between the

data nodes of a Hadoop cluster. Information interaction time of a Hadoop cluster increases when the number of VMs increases.

d. Increasing data size (20 VMs)

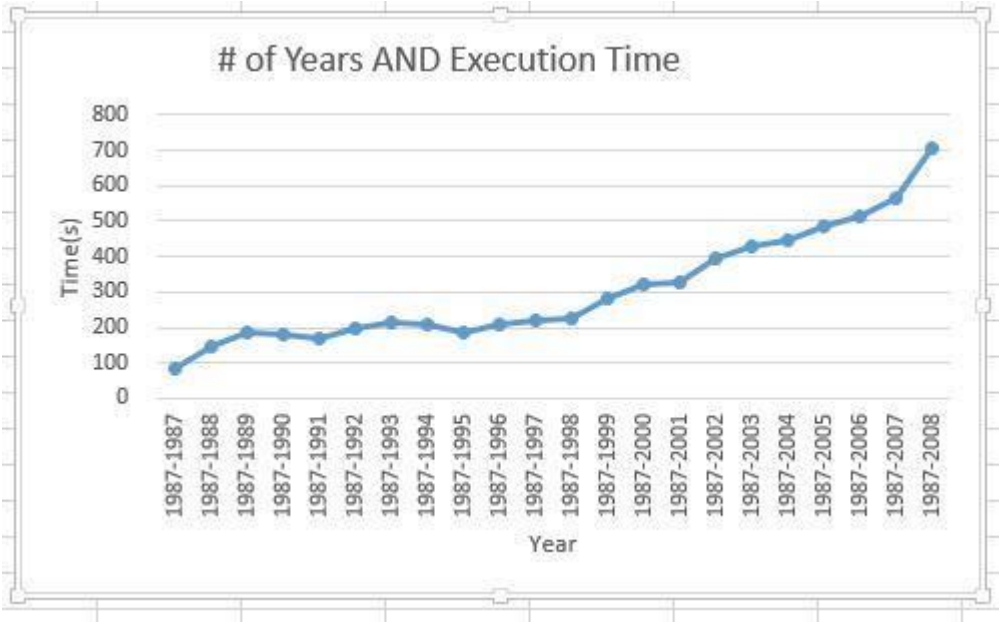

Figure 3

According to the Figure 3, along with the increasing data size, the execution time of the oozie workflow will always increase too. In the beginning, the time-consuming increase with the increase in the amount of data, but the time-consuming increasing is slow, this is because the data increasing of first few years is not that much. On the contrary, after year 1998, the time-consuming increase very fast, the slope is become much steep compare to the first few years. The reason is the flight data between year 1998 to year 2008 is increase faster than the previous years. It also shows more and more people choose traveling by plane.

Some environment setting of our Hadoop cluster

Instance Information: Ubuntu Server 16.04 LTS (HVM), SSD Volume Type - ami-f4cc1de2
Family: General purpose
Type: t2.medium
vCPUs: 2
Memory(GiB): 4
Instance Storage(GB): 24GB

Environment of master instance and slave instances:

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: Windows Operation System
User use MobaXterm to manage instances and files

NOTE：
To change the input file, please change the inputFilePath in the job.properties file which we provided. Please refer the format in the job.properties file version of hadoop: 2.8.0

version of oozie: 4.3.0