Introduction to Big Data

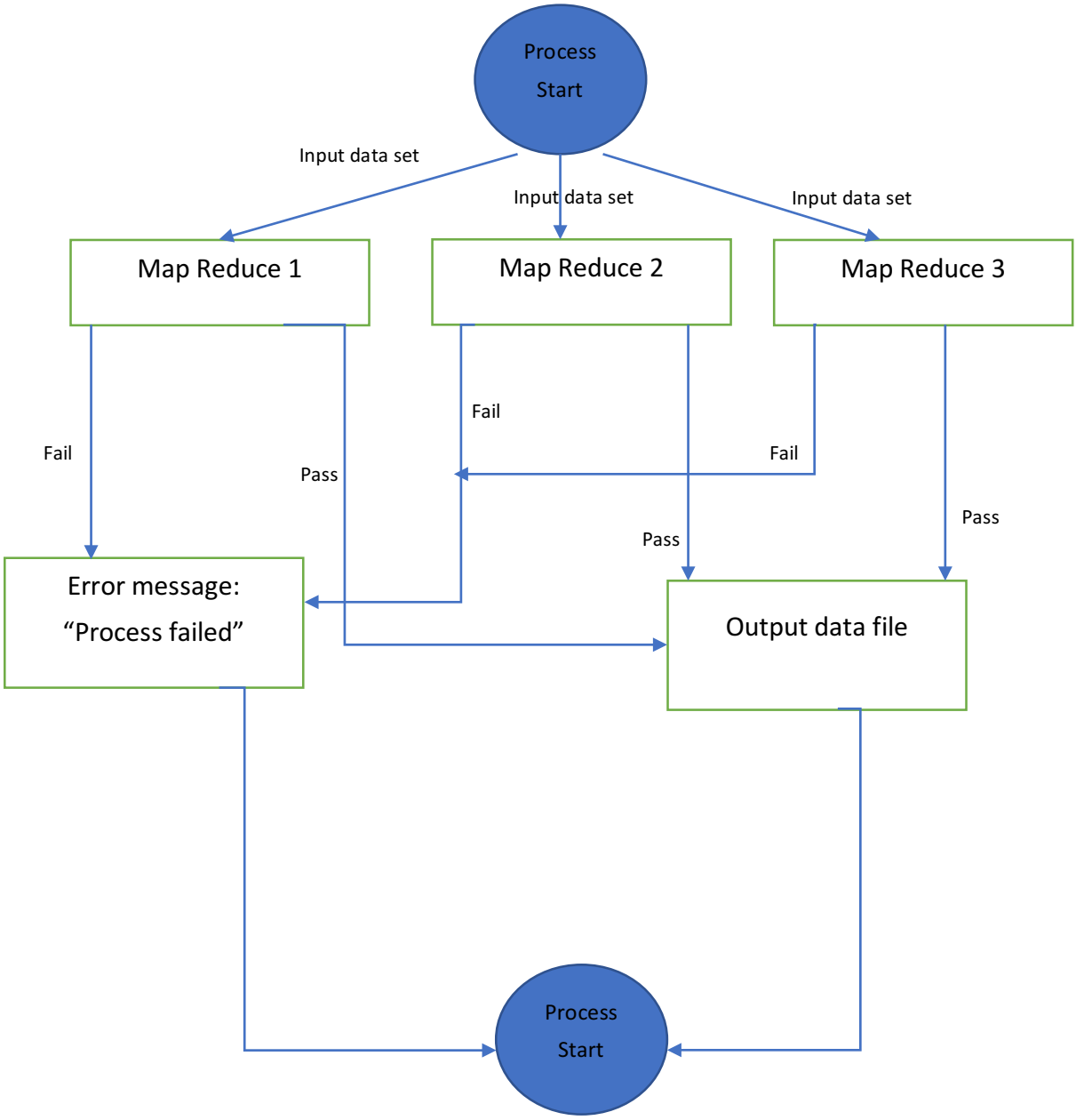
Flight Data Analysis

FINAL REPORT

UNDER THE GUIDANCE OF
Prof. Chase Q. Wu

BY:
Priya Sangale(ps739)

PART A: Structure of the Oozie Workflow:



PART B: Algorithms Used

1. The 3 airlines with the highest and lowest probability, respectively, for being on schedule;

Algorithm:

Function ProbabilityMapper()

Step 1: read the line from the function argument from dataset

Step 2: split each column and store it in token (eg: token[15] is departure time)

Step 3: once the data field is found from the input file using tokens, arrival time and departure time should be checked if it is greater than 10(10 min delay)

Step 4: returns 1 if the time is less than 10 else returns 0

Function ProbabilityReducer()

Step 1: read the line from the function argument from dataset with the help of a separator.

Step 2: data from the mapper will be sent to reduces, mapper will send the flights which has delay time more than 10, count operation is performed to count the number of delayed flights.

Step: count the total number of flights

Step 3: each delayed record is saved in the form of a key value

Step 4: Probability is counted using the formula Delayed flights/total flights by putting the value of minimum and maximum head size 3.

Step 5: Print the heap value.

Function ProbabilityStatus()

Step 1: Probability is calculated using the formula Flight Delay /Total Flights

Step 2: Probability is counted using the formula Delayed flights/total flights by putting the value of minimum and maximum heap size 3.

Step 3: Print the heap value.

2. The 3 airports with the longest and shortest average taxi time per flight (both in and out), respectively

Algorithm:

Function TaxiMapper()

Step 1: read the line from the function argument from dataset with the help of a separator.

Step 2: retrieve the taxiIn and taxiout using the token separators on input line.

Step 3: get the values of taxiIn using the destination and taxiOut using Origin.

Step 4: Generate a Key Pair using the above values to proceed for Reducer.

Function TaxiReducer()

Step 1: read the line from the function argument from dataset with the help of a separator.

Step 2: retrieve the value of total flights and count of delayed flights.

Function TaxiStatus()

Step 1: Probability is calculated using the formula Flight Delay /Total Flights

Step 2: Probability is counted using the formula Delayed flights/total flights by putting the value of minimum and maximum heap size 3.

Step 3: Print the heap value.

3. The most common reason for flight cancellations.

Function CancellationMapper()

Step 1: read the line from the function argument from dataset with the help of a separator.

Step 2: since we must find the reason for flight cancellations, hence key is 'Reason'

Step 3: Retrive the Flight origin value

Step 4: The above data will make the key value-pair

Function CancellationReducer()

Step 1: read the line from the function argument from dataset with the help of a separator.

Step 2: since we must find the reason for flight cancellations, hence key is 'Reason'

Step3: find the sum of the number of values

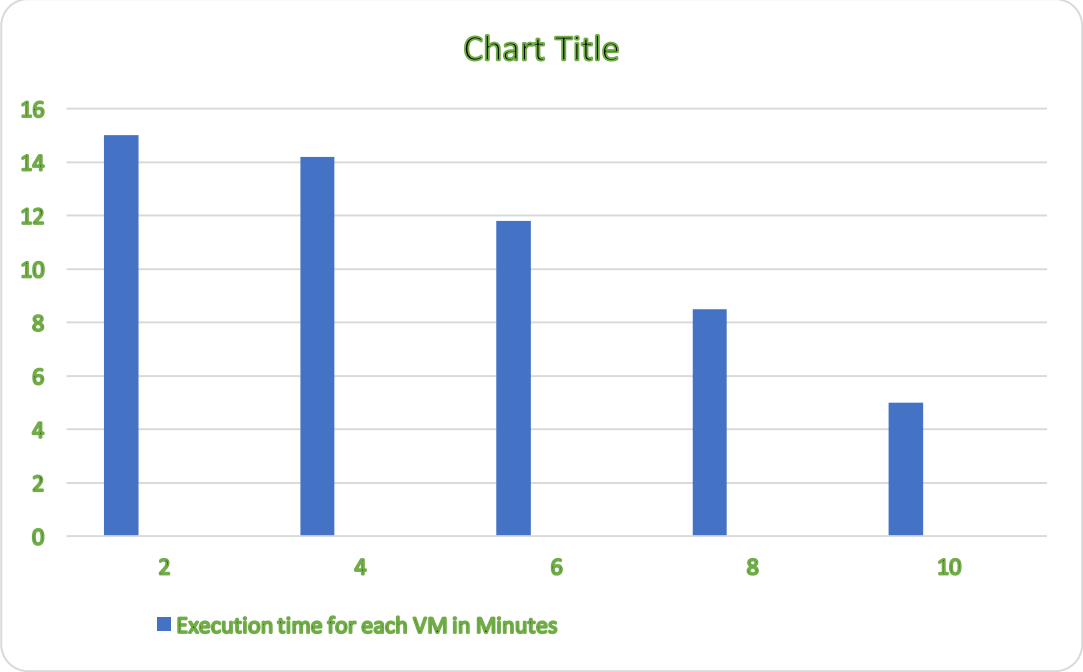
Function Cancellationstatus()

Step 1: Key value pair which is the 'Reason' and the 'Total count' is retrived

Step 2: set the maximum heap status to 1, as only one output is required

Step 3: Print the heap value.

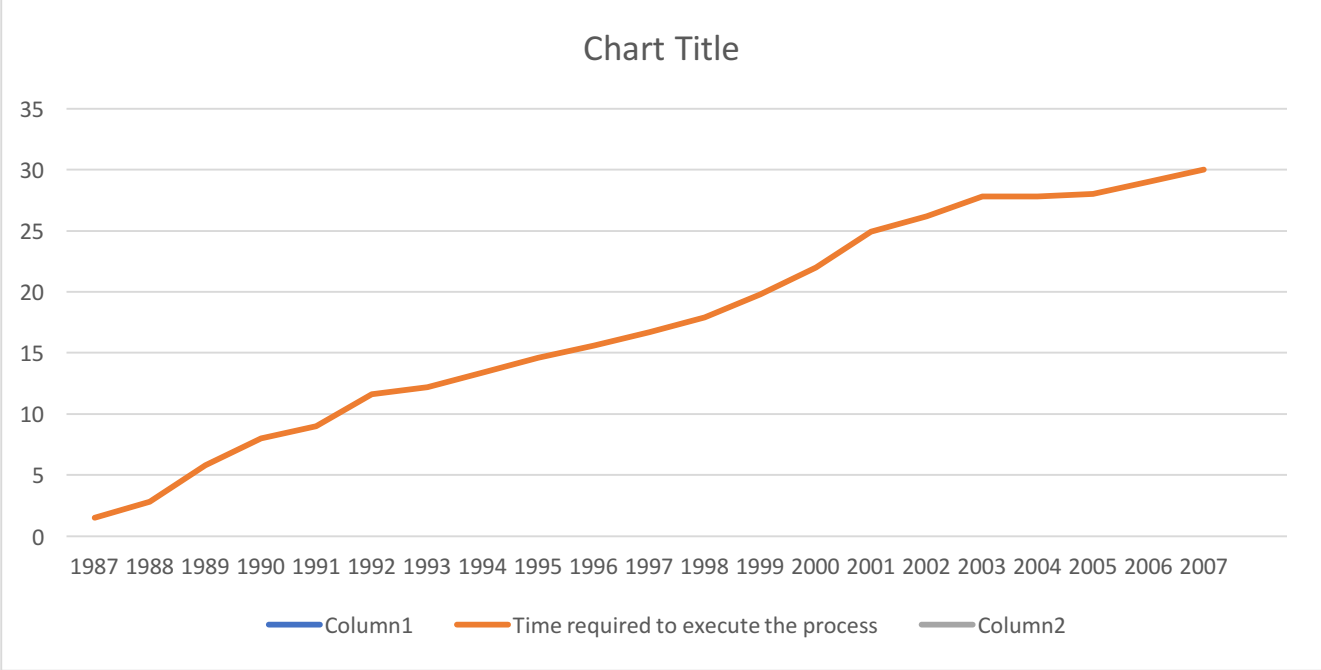
PART 3: Performance measurement plot that compares the workflow execution time in response to an increasing number of VMs used for processing the entire data set (22 years) and an in-depth discussion on the observed performance comparison results



Number of slave nodes	Execution time for each VM in Minutes
2	15
4	14.2
6	11.8
8	8.5
10	5

- As the number of virtual machines increases the time required for process decreases

<u>PART 4:</u> A performance measurement plot that compares the workflow execution time in response to an increasing data size (from 1 year to 22 years) and an in-depth discussion on the observed performance comparison results



Data Set	Time required to execute the process
1987	1.5
1988	2.8
1989	5.8
1990	8
1991	9
1992	11.6
1993	12.2
1994	13.4
1995	14.6
1996	15.6
1997	16.7
1998	17.9
1999	19.8
2000	22
2001	24.9

2002	26.2
2003	27.8
2004	27.8
2005	28
2006	29
2007	30

- As the data size of the input file increase, time required for the process in oozie increase.