CS698: Big Data

Project Report

Ruixin Zhao, rz97@njit.edu

Leyue Wang, lw254@njit.edu

A. Diagram of OOZIE workflow

	Highest 3						*Lowest 3*					
	Item	Value	Item	Value	Item	Value	Item	Value	Item	Value	Item	Value
On-Schedule	HA	86.65%	AQ	79.33%	ML	71.85%	PI	52%	PS	60.4%	AS	61.03%
Ave. Taxi Time	CKB	227	MKK	97.19	VLD	57.29	LAR	8.0	KSM	9.25	VIS	9.36
Flight Cancel.	A(carrier)		317972									

Table 1 Results of Workflow

First of all, the results of the project are listed in Table 1 Results of Workflow.
The airlines with the highest probability for being on schedule are HA, AQ, and ML; and lowest 3 are PI, PS, and AS, respectively.
The airports with the longest average taxi time per flight are CKB, MKK, and VLD, and those with the lowest are LAR, KSM, and VIS.
The most common reason for flight cancellations is carrier, which caused 317972 cancellations.

Figure 1 OOZIE Workflowis the diagram of OOZIE workflow:

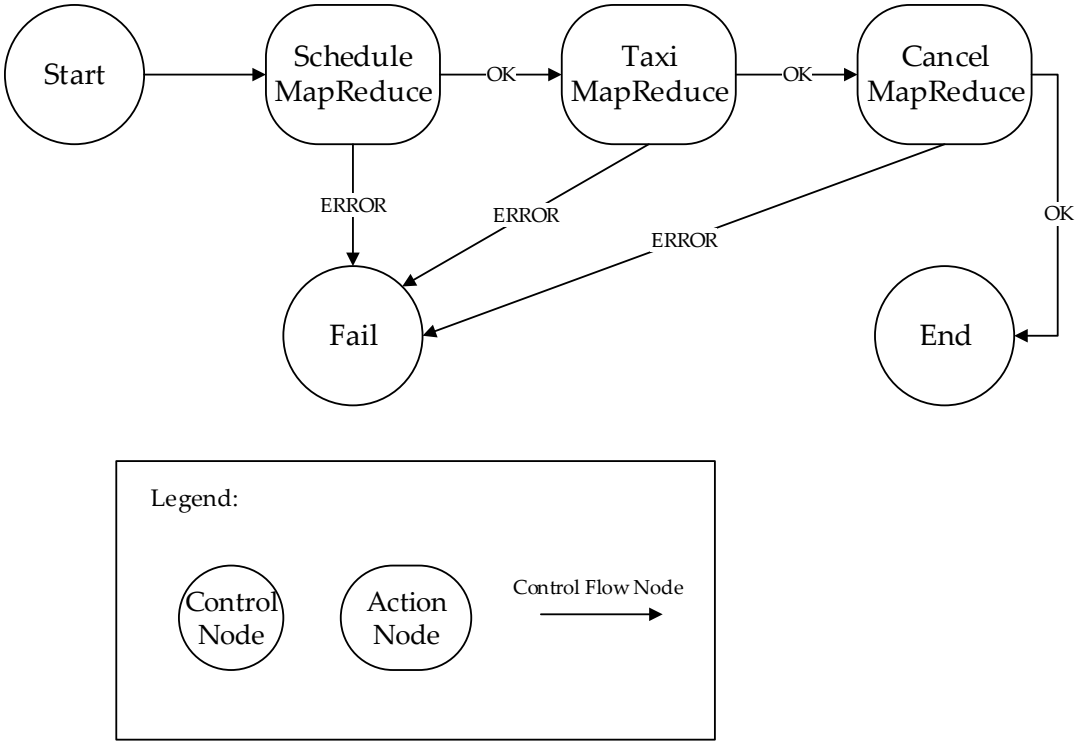

Figure 1 OOZIE Workflow

The OOZIE workflow is shown in Figure 1 OOZIE Workflow. In each action, successful processing leads to the next action node, otherwise to Control Node FAIL.

B. Algorithm Description

- On-Schedule Probability

(1) Read the raw data in rows, excluding the first row;
(2) Choose the data in the 8th column (UNIQUECARRIER) as the key of MAP PHASE, the data that are more than 5 minutes behind schedule in the 14th column (ARRDELAY) as the value;
(3) Shuffle and combine the same UNIQUECARRIER data in REDUCE PHASE;
(4) Calculate the punctuality.

- Average Taxi Time

(1) Read the raw data in rows, excluding the first row;
(2) Choose the 17th column (DEST) as the key of MAP PHASE, the sum of 19th (TEXIIN) and 20th column (TEXIOUT), which are not NA, as the value;
(3) Shuffle and combine the same DEST data in REDUCE PHASE;
(4) Calculate the average waiting time.

- Reason for flight cancellations

(1) Read the raw data in rows, excluding the first row;
(2) Choose the 19th column (CANCELLATIONCODE) that are neither null or NA as the key of MAP PHASE, and output 1 as the value (represents 1 time);
(3) Shuffle and sum the same CANCELLATIONCODE data in REDUCE PHASE;
(4) Calculate the sum of cancellation times of each reason.

C. Execution Time-VM Discussion

VM	2	3	4	5	6	7
Execution Time	40:35	30:06	21:23	18:01	16:11	14:32
Time Delta	\	10:19	8:43	3:22	1:50	1:39

Table 2 Execution time-VM

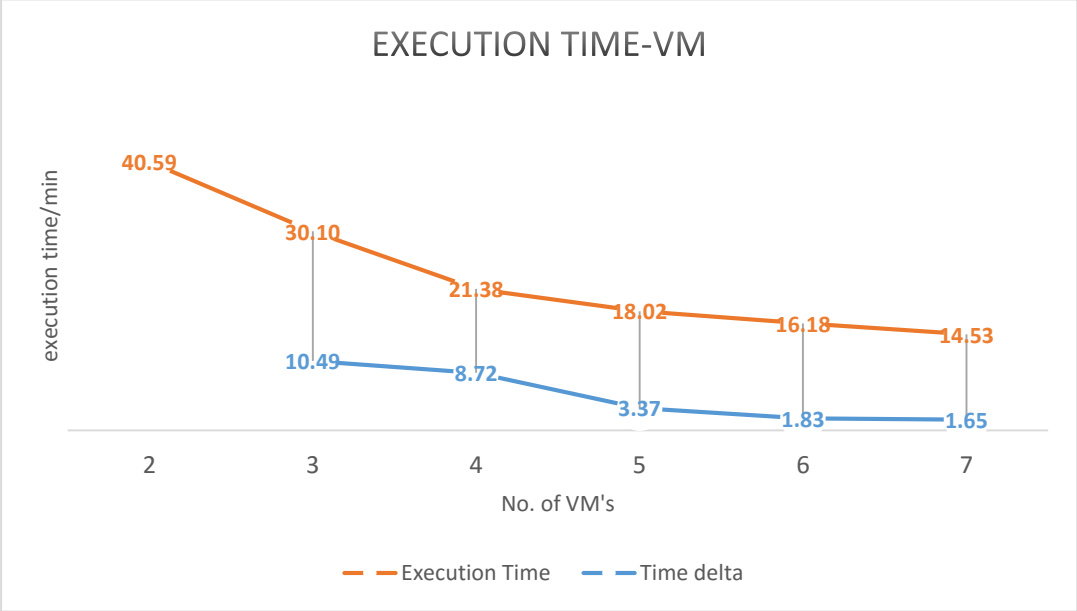

Figure 2 Plot of Execution time-VM Relationship

This is the performance graph that illustrates how the number of VMs (i.e. HADOOP nodes) affects processing time. The orange line shows the execution time as the number of VMs is increased. For 2 nodes the system performs the query in about 40 minutes, for 7 nodes it takes less than 15 minutes.

The blue line (time delta) exhibits the enhancement we get in going from N nodes to N+1 nodes. When the number of VMs is small, the enhancement is significant – going from 2 to 3 nodes reduces the total time by more than 10 minutes, yet becomes less as the nodes increases. In the last two steps, the improvements are both less than 2 minutes. At some point, it will approach a limit where adding more nodes does not decrease processing time significantly.

D. Execution Time-Data Size Discussion

The result of runtime in regarding to data size (by year progression) is:

Year	1	2	3	4	5	6	7	8	9	10	11
Exe.Time	00:23.7	01:13.4	02:00.8	02:54.8	03:43.0	04:48.7	05:17.6	06:11.9	07:06.2	07:54.7	08:57.4

Year	12	13	14	15	16	17	18	19	20	21	22
Exe.Time	12:14.5	13:30.1	14:13.3	13:36.1	13:47.2	19:36.5	17:12.9	17:54.5	19:05.0	21:25.1	21:22.9

Figure 3 Execution Time-Data Size Plotshows how the execution time goes with the progressing of data size. We can see that despite some deviation points, the relationship between them complies with a linear relationship. The data size increasing, considering the so-far immensity of data, usually gets stuck in the bottleneck and leads to a high dimensional increasing in time cost. The linear distribution suggests HDFS' perfect distribution and expansion ability, allowing the increasing of massive data irrelevant to system's time complexity.

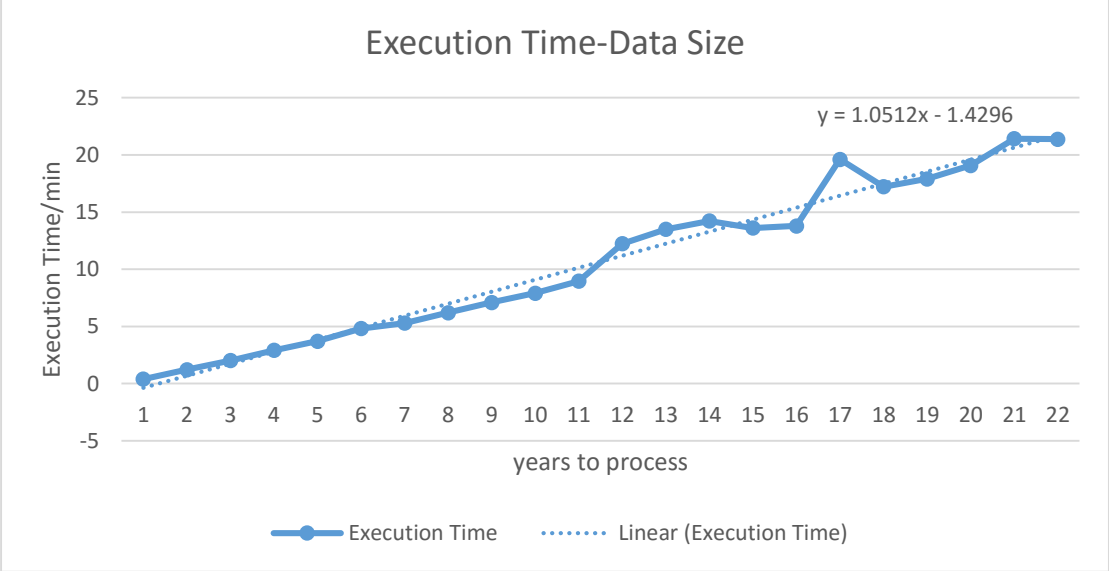

Figure 3 Execution Time-Data Size Plot