9/9/2018

Statistical Analysis of Job run time and number of records

SPARKPROJECT-1

Name: RUKSANA BHANU SHEIK

STATISTICS OF JOB RUN TIME & NUMBER OF RECORDS.

JOB RUN TIME:

Below is the template for time taken for execution of pig and spark programs for single record lookup, filter with some condition and group by accompanied with order by.

	Single Look Up Record	Filter	Group By With Order
Pig	00:11:12 HH:mm:ss	00:10:06 HH:mm:ss	00:15:49 HH:mm:ss
Spark	00:09:38 HH:mm:ss	00:04:23 HH:mm:ss	00:01:50 HH:mm:ss

ANALYSIS:

As per the above statistics of time taken on the job run time, it is evident that spark has taken considerably less time than pig. The reason behind this is that, spark is in-memory execution. Due to this, the seek time, rotational delay and transfer time gets minimized.

Whereas Pig reads and writes from hard disk of data nodes in HDFS. Hence the seek time, transfer time would be very high compared to Spark

CONCLUSION:

Due to in memory execution, spark has taken less time for execution.

NUMBER OF RECORDS:

Below is the template showing number of records obtained as output upon execution of pig and spark programs for single record lookup, filter with some condition and group by accompanied with order by.

	Single Look Up Record	Filter	Group By With Order
Pig	1	31029	4
Spark	1	31029	4

ANALYSIS:

As per the above statistics of number of records for each case, it is evident that both pig and spark programs provided same number of records as output. Spark reads the files as RDD's and executes the program. Pig internally converts the script into map reduce program and executes.

CONCLUSION:

Number of records obtained as outputs for Pig and Spark for each scenario is one and the same. Therefore, Spark is the suggestible approach to solve this problem statement.