



BIG DATA SYSTEMS ASSIGNMENT 2

Title:

Selection of suitable Hadoop Eco System platform for sales data analytics

Overview & background:

An e-Commerce company in Europe is analyzing the online retail sales data of 2010 to formulate the sales strategy for 2011. The name of the sales data file is "Assignment2-2024-BDS-DATASET-online_retail_dat.csv". This file contains header and 480232 data records.

The analytics team consists of developers with expertise on different Hadoop frameworks/platforms like MapReduce, Pig Latin, Hive, HBase and Spark. The technical manager is confused about the appropriate framework to be used in the analytics project. The only concern of the manager is that the time taken for execution of the analytics queries should be minimum. In this assignment, you need to implement 2 typical analytics queries in any two of the platforms and do a performance comparison between them and recommend the framework/platform which gives better performance. The scripts/code/queries developed by you on both the platforms, the timing values and the query results are to be submitted for each of the queries mentioned in section 3. Also, you need to submit the UUIDs of computers on which you have done the performance tests.

Input: CSV data with flat schema with multiple records and features.

Description:

1. STORAGE:

The data file should be copied to the local file system of any node in your Hadoop cluster. This data file is to be moved to HDFS of the Hadoop cluster by configuring and running a suitable Flume agent. The block size of the file should be selected for optimum performance. A suitable value for the replication factor of the file should be selected to ensure reliable storage of the data file.

2. METADATA

The data consists of RecordNo, InvoiceNo, StockCode, Description, Quantity, InvoiceDate, Price, CustomerID, and Country. Some of the fields in the data may be blank. If required, you are allowed to remove the first header record containing the schema definition. Or this record may be skipped during reading and or analysis. No other modifications are allowed to the contents of the file.

3. ANALYTIC QUERIES FOR BENCH MARKING:

1. Total revenue (Aggregation of Price) received in the year 2010.
2. List of unique items sold (With same StockCode) and their total sales volume (Aggregation of Quantity) in the year 2010 sorted in ascending order of StockCode.

4. FRAMEWORKS / PLATFORMS TO BE COMPARED:

1. Hadoop group
 - a. Hadoop MapReduce
 - b. Pig Latin Scripts
 - c. Apache HIVE
 - d. Apache HBASE
2. Spark group
 - a. Spark MapReduce
 - b. Spark Dataframes
 - c. Spark Datasets
 - d. SparkSQL

5. GUIDELINES FOR PERFORMANCE COMPARISON:

1. You need to select one framework from Hadoop group and the second framework from the Spark group given in Section 4 above. It is NOT allowed to select two frameworks from the same group. In this assignment, you need to do a query performance comparison between the two frameworks selected by you. Two queries to be used for performance evaluation are given in Section 3 – Analytics queries for benchmarking.
2. If you are using Linux, you can "time" command to time your command. For Windows, you need to find out a method to determine the time taken for execution of each of the queries. Sometimes, the time taken for execution of a query can be less than 1 second and you may not be able to measure time in millisecond range. You have the following 2 options to overcome this problem:
 - a. Repeat the query multiple times, say 10 to 100 and determine the total time taken. Then find out the time taken for executing individual queries.
 - b. Almost all the platforms mentioned above allow you to specify a folder in HDFS as input. You may copy multiple copies of the same data file into the input folder (of course with different file names) and execute the query. Then find out the query time by dividing the total time by the number of copies of the file.

6. CONDITIONS

1. Since this is a group assignment involving comparison of performance on 2 different frameworks, one student should work on 1 platform and other student(s) should work on the second platform. The group leader needs to consolidate the results and submit the assignment.
2. You should use Apache Flume to move data from the local file system to HDFS. If data is moved with the Hadoop put command, marks will be reduced.
3. The Hadoop cluster should be configured on Linux / Windows systems.
4. If only one system is available, you need to configure the cluster in pseudo distributed mode.
5. The Replication factor for the HDFS files should be set as the number of nodes in the cluster.
6. Focus on performance tuning of the framework by selecting proper configuration parameters instead of accuracy of the query results.

7. SUBMISSION REQUIREMENTS

Your submission should consist of all the following 7 items:

1. Configuration files of Hadoop cluster / Spark and frameworks like Pig, Hive, HBase used in your solution. Include only part of the configuration files which you have modified.
2. The configuration of the Flume agent developed by you to transfer the data file from local filesystem to HDFS folder.
3. The code, scripts, and query developed for any 2 of the selected platform:
 - a. Hadoop MapReduce code
 - b. Pig Latin Script
 - c. Hive Query
 - d. HBase Query
 - e. Spark MapReduce code
 - f. Spark program in scala / python to load the data and manipulate the dataframes.
 - g. SparkSQL queries and associated code in Scala / Python.
4. The output of the 2 bench marking queries obtained from the 2 platforms. For Query 2, include only first 50 rows of the sorted output in your report.
5. Execution time in seconds for each of the 2 queries for the 2 platforms selected by you. Example:
 - a. Query 1 – Platform/Framework 1 = 0.6 secs
 - b. Query 1 - Platform/Framework 2 = 0.65 secs
 - c. Query 2 – Platform/Framework 1 = 0.8 secs
 - d. Query 2 - Platform/Framework 2 = 0.83 secs

6. Screen shot from Hadoop YARN Job History (Sample shown here)

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime
application_1709667623375_0005	jps	Spark SQL Query 2	SPARK		rootjps	0	Wed Mar 6 01:20:59 +0550 2024	Wed Mar 6 01:20:59 +0550 2024	Wed Mar 6 01:21:27 +0550 2024
application_1709667623375_0004	jps	select StockCode_sum(Quantity) from SAL...50 (Stage-2)	MAPREDUCE		rootjps	0	Wed Mar 6 01:18:57 +0550 2024	Wed Mar 6 01:18:58 +0550 2024	Wed Mar 6 01:19:14 +0550 2024
application_1709667623375_0003	jps	select StockCode_sum(Quantity) from SAL...50 (Stage-1)	MAPREDUCE		rootjps	0	Wed Mar 6 01:18:35 +0550 2024	Wed Mar 6 01:18:35 +0550 2024	Wed Mar 6 01:18:55 +0550 2024
application_1709667623375_0002	jps	Spark SQL Query 1	SPARK		rootjps	0	Wed Mar 6 01:14:01 +0550 2024	Wed Mar 6 01:14:01 +0550 2024	Wed Mar 6 01:14:27 +0550 2024
application_1709667623375_0001	jps	select sum(Price) from SALES (Stage-1)	MAPREDUCE		rootjps	0	Wed Mar 6 01:11:21 +0550 2024	Wed Mar 6 01:11:22 +0550 2024	Wed Mar 6 01:11:39 +0550 2024

7. System details of your Hadoop cluster (from all nodes, if you are using more than one node)

- CPU clock speed and number of cores, Memory size in GB.
- UUID of the system – (On Linux - `sudo dmidecode -t system | grep UUID`)
(On Windows - `wmic path win32_computersystemproduct get uuid`)

When you use a virtual machine or Ubuntu Linux as a windows application, you may not be able to get a UUID value. In this case, you have to extract and provide the UUID of the base system.

General Notes:

- Using Canvas, only the first member of the group has to upload the file. No submission over email will be considered.
- Submit the code and a document as a zip file. The document should be a Word or a PDF describing your setup, how to run your code and results as described in "Submission requirements".
- Images / photos of scripts / code captured from screens will not be evaluated.
- The document in the zip file should have full names of the group members along with the BITS Registration no. of each group member.
- Name the zip file in format like "Grp_<your_group_number>.zip" only. Don't add anything into the file names.
- Make sure that you upload the file well ahead of the deadline. At the last moment, we have seen several groups have faced issues while doing the submissions.
- The group leader should give a declaration on the percentage of contributions made by each of the group members. Marks will be reduced for members with low contribution.
- Note - As it's a group assignment, only one submission is expected from each group. Unnecessarily don't upload the solution on individual basis.**
- **Plagiarism will be strictly dealt with and if found will result in cancellation of the Assignment and 0 marks being awarded to all the group members.**
- **The last date of submission will not be extended in any case.**

Academic Integrity

Honesty is primarily the responsibility of each student. The institute considers cheating to be a voluntary act for which there may be a reason, but for which there is no acceptable excuse. It is important to understand that collaborative learning is considered cheating unless specifically allowed for by the professor. The term cheating includes but is not limited to: plagiarism, receiving or knowingly supplying unauthorized information, using unauthorized material or sources, changing an answer after work has been graded and presenting it as improperly graded, illegally accessing confidential information through a computer, doing the assignment for another student or having another student do the assignment for you.