

## BIG DATA SYSTEMS – ASSIGNMENT 2

### Title:

### **Implementation of a system for sales data analytics using Hadoop Eco System**

### 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 "Assignment-2 2023 BDS DATA SET online\_retail\_data.csv". This file contains 525461 records. The schema of this data set is (Record No, Invoice No, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, Country). Some of the fields in some records may be blank. The analysis should be carried out using Hadoop eco system products after storing the data on HDFS. The results of the analysis along with the code and queries developed should be submitted.

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

### Description:

#### 1. STORAGE:

The data file should be moved to HDFS of the Hadoop cluster. 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 Record No, Invoice No, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country. Some of the fields in the data may be blank. If required, you are allowed to remove the first record containing the schema definition. Or this record may be skipped during reading and or analysis. No other modifications are allowed on the contents of the file.

### 3. ANALYTICS:

You need to perform analysis of the data to find out the following 7 parameters for finalizing the sales strategy of the company for year 2011:

1. *Total number of unique customers in the "given country".*
2. *Country from which the maximum revenue was collected from sales in the month of March 2010.*
3. *Month of 2010 in which maximum number of items were sold.*
4. *In the month of January 2010, find the country in which maximum number of items were sold*
5. *The StockCode of the item with the highest number of sales in the "given country" in the year 2010*
6. *StockCode of the item for which the maximum revenue was received by sales in the month of December 2010.*
7. *The country in which minimum number of sales happened in 2010.*

#### Conditions and suggestions:

1. At least one of the solutions for the problems should be implemented using MapReduce code in a language of your choice.
2. The solution for at least one of the problems should be developed using Hadoop Eco system products like Apache Pig, Hive, and Hbase.
3. The solutions for the remaining problems can be implemented using a method/product of your choice. ie. MapReduce,Pig, Hive or HBase.
4. You are allowed to use Apache Spark to process the data stored on HDFS.
5. The Hadoop cluster should be configured on Linux / Windows systems.
6. If only one system is available, you need to configure the cluster in pseudo distributed mode.
7. The Replication factor for the HDFS files should be set as the number of nodes in the cluster.
8. When the term "given country" is used in a question, you need to execute the query / code with a country of your choice. Most of the European countries are present in the data. You may examine the data in Excel or any other viewing tool to find out the countries present in the data.

## Submission Requirements:

1. Configuration files of Hadoop cluster / Spark and products like Pig, Hive, HBase if used in your solution. Include only the configuration files which you have modified.
2. The MapReduce code developed by you. Please include a screen shot of the MapReduce job history after running the job.
3. The output of the MapReduce job. The location of the output file stored on HDFS and the name of the output file should be mentioned.
4. If Spark is used, code written in Scala, Python, Java. And outputs generated.
5. If used, the Pig Latin Script and the output generated
6. If used, the Hive query and the output generated
7. If used, HBase query and the output generated
8. A Video recording of the screen on which the queries are fired / programs are run. The video should start when the queries are entered on the screen and stop only when the outputs appear on the screen. In case, the outputs go to HDFS / local files, they are to be copied to the screen and video graphed.
9. You should provide clear instructions on how to reproduce the submission on the Evaluator's setup.
10. Your code and results should be reproducible.
11. The implementation should be general purpose for any other CSV input file.

## 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".
  - ☐ 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.
- **Note - As it's a group assignment, only one submission is expected from each group. Unnecessarily don't upload the solution on individual basis. If it's observed, then the penalty (25%)**
  - **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.**