**Retail Industry Sales Data Insights and Analysis**

Name: Saurabh Joshi

Mail ID: Saurabh.joshi@fractal.ai

Fractal ID: F06891

Fractal Batch 5

Problem Statement: In this project, a series of applications need to be built using Python, SQL, Spark that can download data from a data lake, process and analyze it and then load the cleaned-up data back-to-back to a data lake.

Diagram

Description automatically generated

**Data Ingestion Pipeline**

Step No. 1

Connecting to Hadoop cluster.

Text, letter

Description automatically generated

Step No 2:

Download \*.csv files from Hadoop cluster to local.



Step No 3:

Unzipped/Extracted the completeData.zip folder

Text, letter

Description automatically generated

Step No 4:

Uploaded all downloaded local files to our local system using **SCP** command.

A picture containing text

Description automatically generatedA picture containing table

Description automatically generated

Step No 5:

Setting up Environment and Importing packages.

Graphical user interface, text, application, email

Description automatically generated

Step No 6:

Specifying the file schema as per requirement.

Text

Description automatically generated

Step No 7:

Loading/Reading the files into data frames.

Text

Description automatically generated with medium confidence

Graphical user interface, text, application

Description automatically generated

Step 8:

Removing Noise from the products data.

Text

Description automatically generated

Step 9:

Showing Columns of all data frames.

Graphical user interface, text, application

Description automatically generated

Output:

Text

Description automatically generated

Step 10:

Showing Data types of all data frames.

Graphical user interface, application

Description automatically generated with medium confidence

Output:

Text

Description automatically generated

Step 11:

Creating Tables using data frames.

Text

Description automatically generated

Step 12:

Displaying Records from the tables created.

Graphical user interface, text, application

Description automatically generated

Output:

Table

Description automatically generated

Step 13:

Doing Aggregation part of the tables.

Text

Description automatically generated with medium confidence

Step 14:

Uploading the final output csv file to INSOFE cluster.

From Insofe local path: /home/fai10090/Capstone\_Project/Output/Final\_output.csv

output file is copied to HDFS cluster local path: fai10102@172.16.0.118:/home/fai10090/

From HDFS cluster local path output file is copied to Insofe path:

/user/insofe/fai10090/Capstone\_Project/Output/

Graphical user interface, text, application

Description automatically generated

**Visualization**

Graphical user interface, application

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generated

GitHub:

[sj-saurabh11/Capstone\_Project\_Saurabh (github.com)](https://github.com/sj-saurabh11/Capstone_Project_Saurabh)