**Retail grocery industry Data Analysis and Insights**

**Calendar

Description automatically generated with low confidence**

**Name: Dipanwita Bhaumik**

**Mail Id:** [**Dipanwita.bhaumik@fractal.id**](mailto:Dipanwita.bhaumik@fractal.id)

**DE Crossover Training Batch**: 4

**Fractal Id:** F07068

***Problem Statement***

The retail grocery industry in the United States faces a precarious economic environment. Due primarily to competition from warehouse clubs, supercentres, and e-commerce, retail grocery sales have underperformed the U.S. retail sector and the overall U.S. economy, and employment growth in the industry has been stagnant. Yet, a large proportion of consumers maintain a strong preference for shopping at retail grocery stores, and total grocery industry sales and employment still exceed sales and employment at warehouse clubs/super-centres and e-commerce retailers. To compete in this setting, many retail grocers are turning to third-party online grocery delivery services offering online shopping and same-day grocery delivery, the largest of which is the current retail store.

One of the retail companies and its team came up with a business problem in which after solving, can help the online grocery stores in managing their business to gain an edge over the market. The specific business problem is to drive higher sales volume and customer retention. The solution involved building a ETL pipeline by the data engineering team and perform analysis by the ML team.

As part of this capstone project, build a ETL Pipeline as part of data engineering solution to create a foundation for other applications that are dependent on the engineering solution. Applications like data analytics and modelling may be applied to provide summary reports for decision makers.

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 analyse it and then load the cleaned-up data back-to-back to a data lake.

DATA PROCESSING PIPELINE

HDFS DATA LAKE

Input Files:

Structured Data

Aisles.csv

Departments.csv

Orders.csv

prior\_order.csv

products.csv

train\_order.csv

Analysis

Using

PySpark, SparkSQL

Visualization

A yellow sign with black text

Description automatically generated with medium confidence

Output File:

Retail\_aggregated\_file.csv

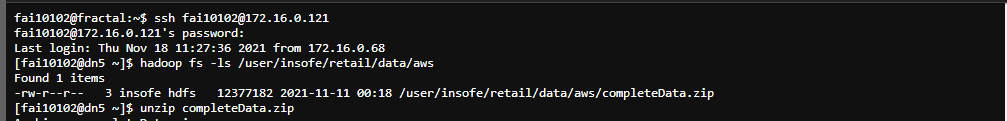
Output File:

Retail\_aggregated\_file.csv

Data Ingestion Pipeline

Step No. 1**:**

*Downloaded \*.csv files from Hadoop cluster locally*



Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

*Step 2:*

*Using Jupyter notebook to write PySpark code for creating an Ingestion pipeline and Data Aggregation*

*setting up* environment

A screenshot of a computer

Description automatically generated

*Specifying File Schemas*

Text

Description automatically generated

*Extracting Data*

Text

Description automatically generated

Text

Description automatically generated

*Removing noises from products data*

Text

Description automatically generated

*Showing columns of all data frames*

A picture containing text

Description automatically generated

*Output*

Text

Description automatically generated

*Showing Data types of all data frames*

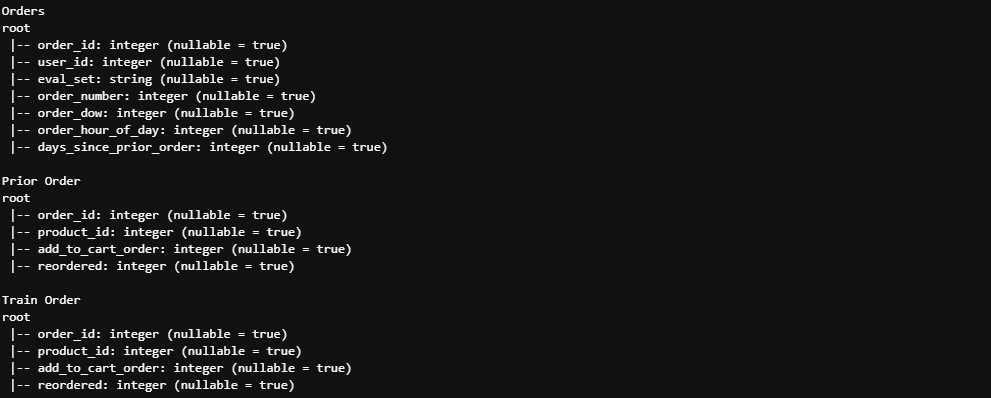
A picture containing rectangle

Description automatically generated

*Output*

A picture containing shape

Description automatically generated



*Transformation: - Data Processing Part*

Text

Description automatically generated

*Displaying the records from tables*

Shape, rectangle

Description automatically generated

*Output*

Shape

Description automatically generated

Text

Description automatically generated

Shape

Description automatically generated

*Creating Aggregated table*

Text

Description automatically generated

*Output*



Graphical user interface, text, application

Description automatically generated

*Checking Null Values*

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

*Final Output Table*

Text

Description automatically generated with medium confidence

*Step 3: Data pushing to Data Lake*

*From Insofe local path: /home/fai10102/Capstone\_Project/Data\_sets/output/Retail\_aggregated\_data.csv*

*output file is copied to HDFS cluster local path: fai10102@172.16.0.121:/home/fai10102/*

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

*/user/insofe/fai10102/Retail\_Capstone\_Project/Output/*

Text

Description automatically generated

Text

Description automatically generated with low confidence

*Step 4:* Data Visualization

Graphical user interface, application

Description automatically generated

Graphical user interface

Description automatically generated Graphical user interface, table

Description automatically generated

Table : Day\_of\_the\_week Table: Hour\_table

Chart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, bar chart

Description automatically generated

*Step 5:* GitHub Link

[*https://github.com/Dipanwita1997/Retail\_Grocery\_DE\_Capstone\_Project.git*](https://github.com/Dipanwita1997/Retail_Grocery_DE_Capstone_Project.git)

Graphical user interface, text, application, email

Description automatically generated