Retail grocery industry Data Analyse and Insights

**Name: Arghya Nandy**

**Mail Id:** [**arghya.nandy@fractal.id**](mailto:arghya.nandy@fractal.id)

**Fractal Id:** F07002

**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

Diagram

Description automatically generated

Data Ingestion Pipeline

*Step No. 1: Design of the solution*

Check the files in /user/insofe/retail/data/aws directory

Text, letter

Description automatically generated

*Download the file locally and then Unzip it*

Text

Description automatically generated

*Step 2: Analysis of data:*

*Using FractalLab terminal open Pyspark and create an Ingestion pipeline and Data Aggregation code.*

*Open Pyspark in Console*

Text

Description automatically generated

*Extracting Data*

*Creating Schema:*

Text, letter

Description automatically generated

Create pyspark sql dataframe: “aisles\_df”

Text

Description automatically generated

Create pyspark sql dataframe: “department\_df”

Graphical user interface, text, application

Description automatically generated

Create pyspark sql dataframe: “order\_df”

Table

Description automatically generated

Create pyspark sql dataframe: “prior\_order\_df”

Graphical user interface

Description automatically generated

Create pyspark sql dataframe: “products\_df” | Removing noises from products data

Text

Description automatically generated

Create pyspark sql dataframe: “train\_order\_df”

Graphical user interface

Description automatically generated with low confidence

*Showing column names of all data frames*

Text

Description automatically generated

*Showing DataType of all columns*

Text

Description automatically generated

Text, letter

Description automatically generated

*Check null values in columns of all DataFrame*

Text

Description automatically generated with medium confidence

*Transformation*

Text, letter

Description automatically generated

Aggregation: Create Dataframe “combined\_table” after joining all the dataframes



Graphical user interface, text, application

Description automatically generated

Saving the result in csv

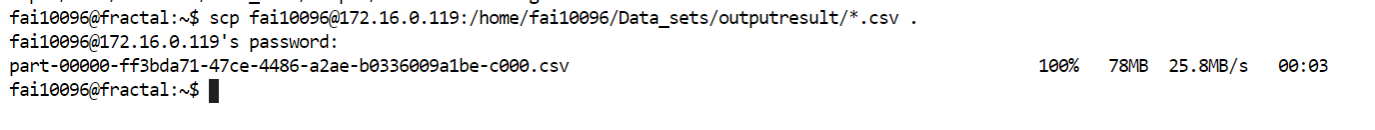
Text

Description automatically generated

Text

Description automatically generated

Download the csv





**

*Step – 3: 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, bar chart

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, histogram

Description automatically generated

*Step 4: GitHub Link:*