Retail grocery industry Data Analyse and Insights

**Name: Mahesh Bodineni**

**Mail Id: mahesh.bodineni@fractal.ai**

**Fractal Id: F07080**

**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

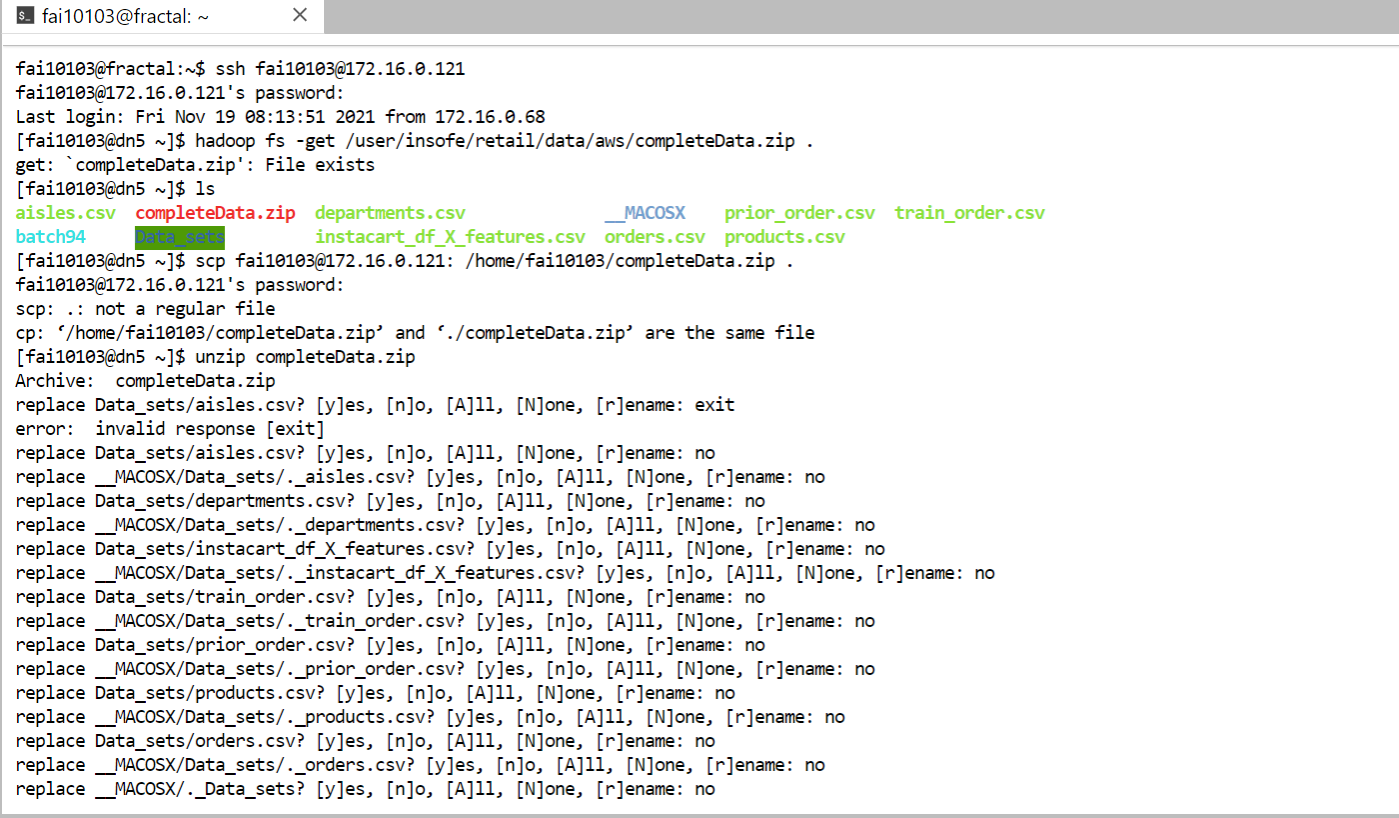
Diagram

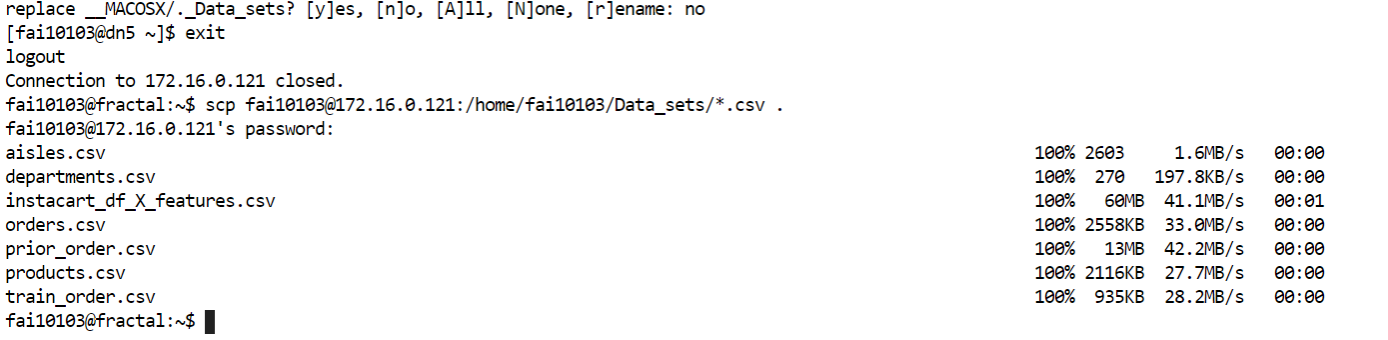
Description automatically generated

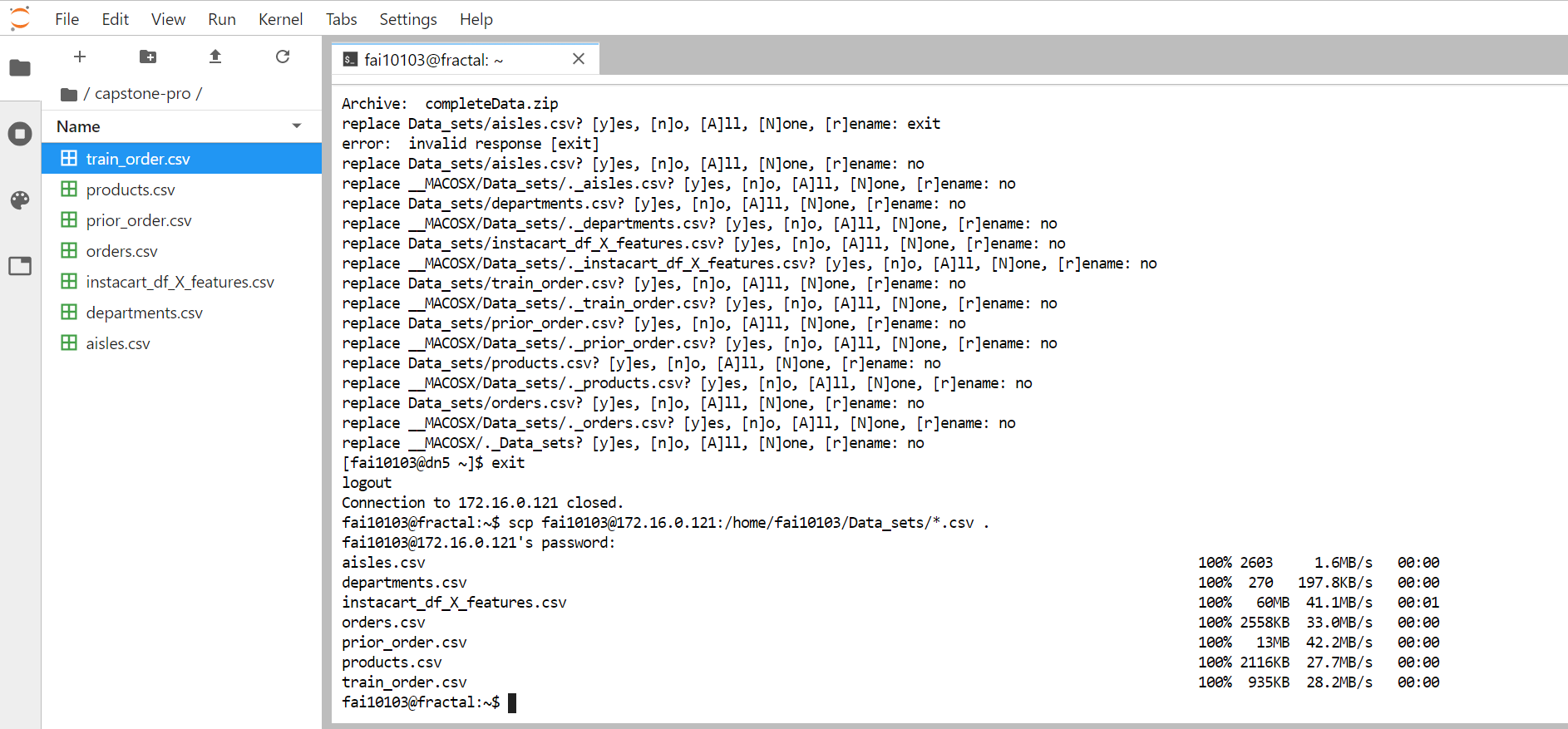
Data Ingestion Pipeline

**Step No. 1:**

Downloaded \*.csv files from Hadoop cluster locally







**Step2:**

Using jupyter notebook to write pySpark code for creating an Ingestion pipeline and Data Aggregation

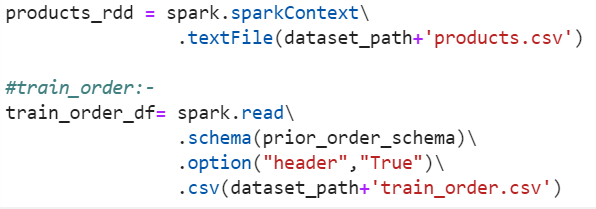
**Environment Setup:**



**Specifing File Schemas:**



**Extracting Data:**

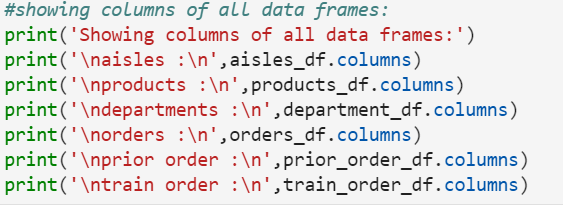


**Removing Noises from products data:**

Text

Description automatically generated

**Showing columns of all data frames:**



**Output:**

Text

Description automatically generated

**Showing datatypes for all data frames:**



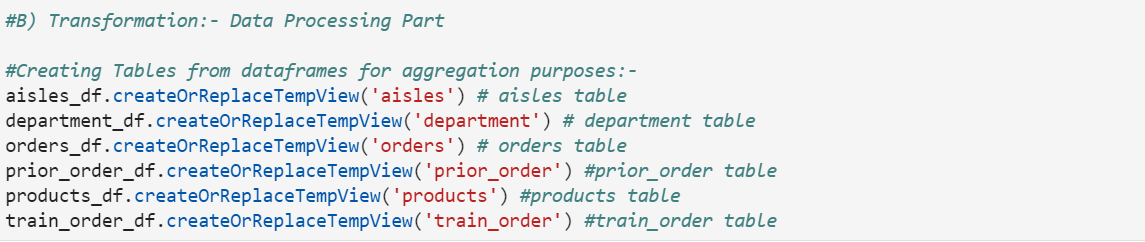
**Output:**

Text

Description automatically generatedText, letter

Description automatically generated

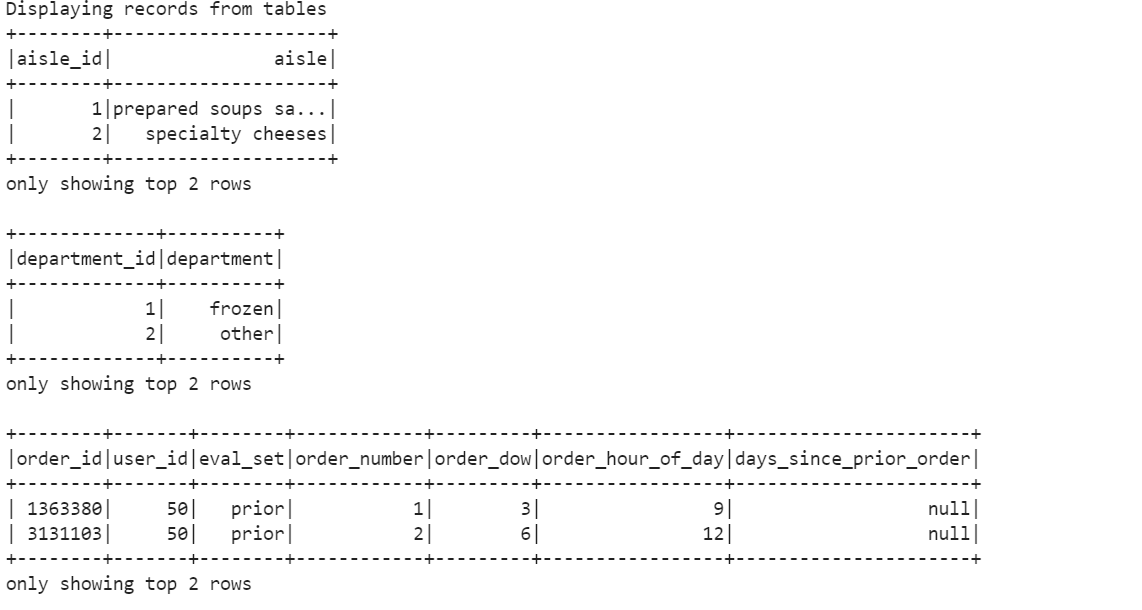
**Transformation:-Data Processing Part:**

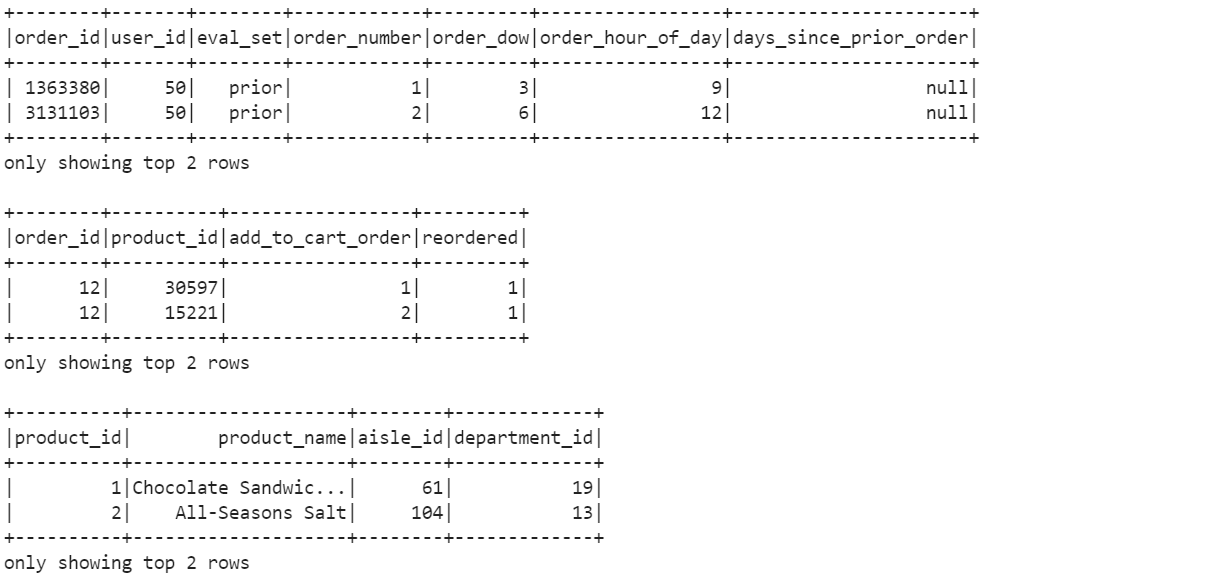


**Display the records from tables:**



**Output:**

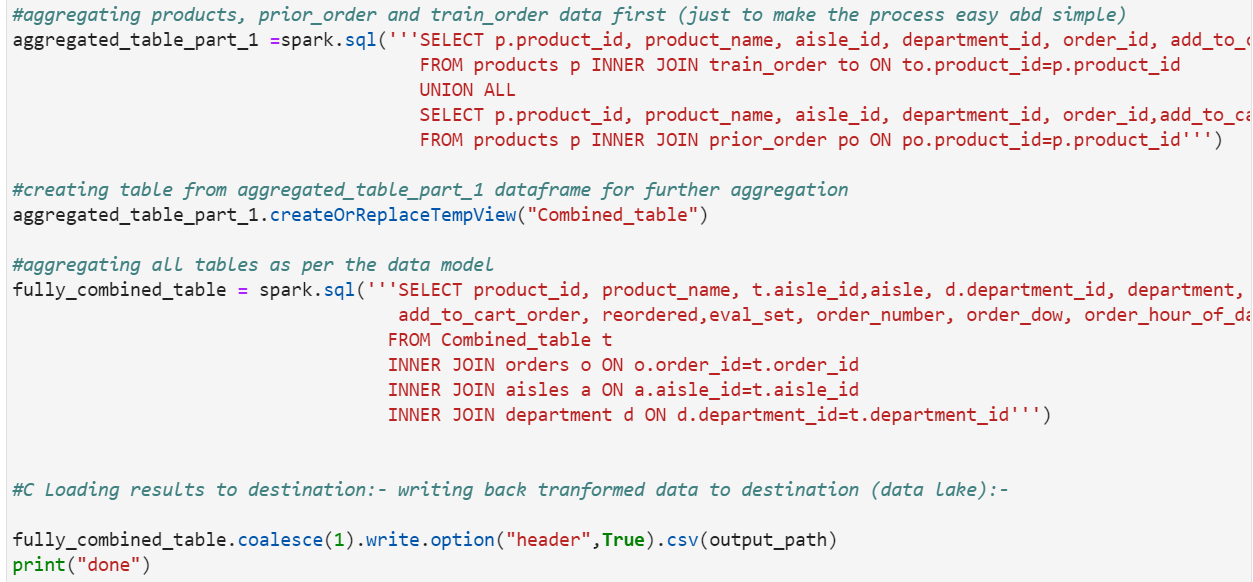




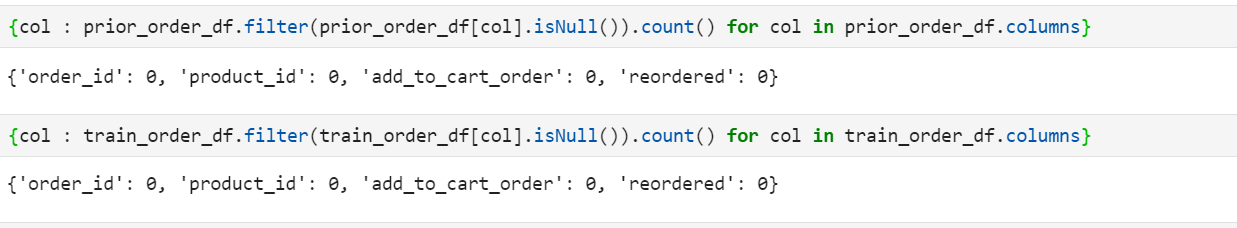
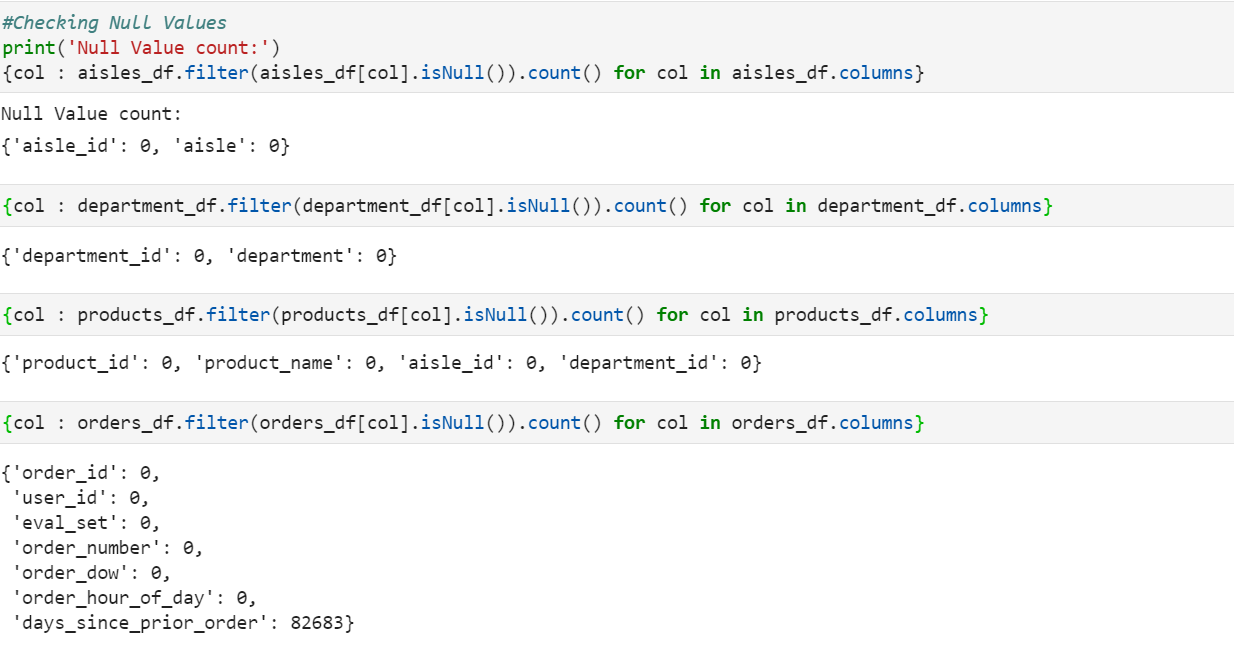
Graphical user interface, application

Description automatically generated

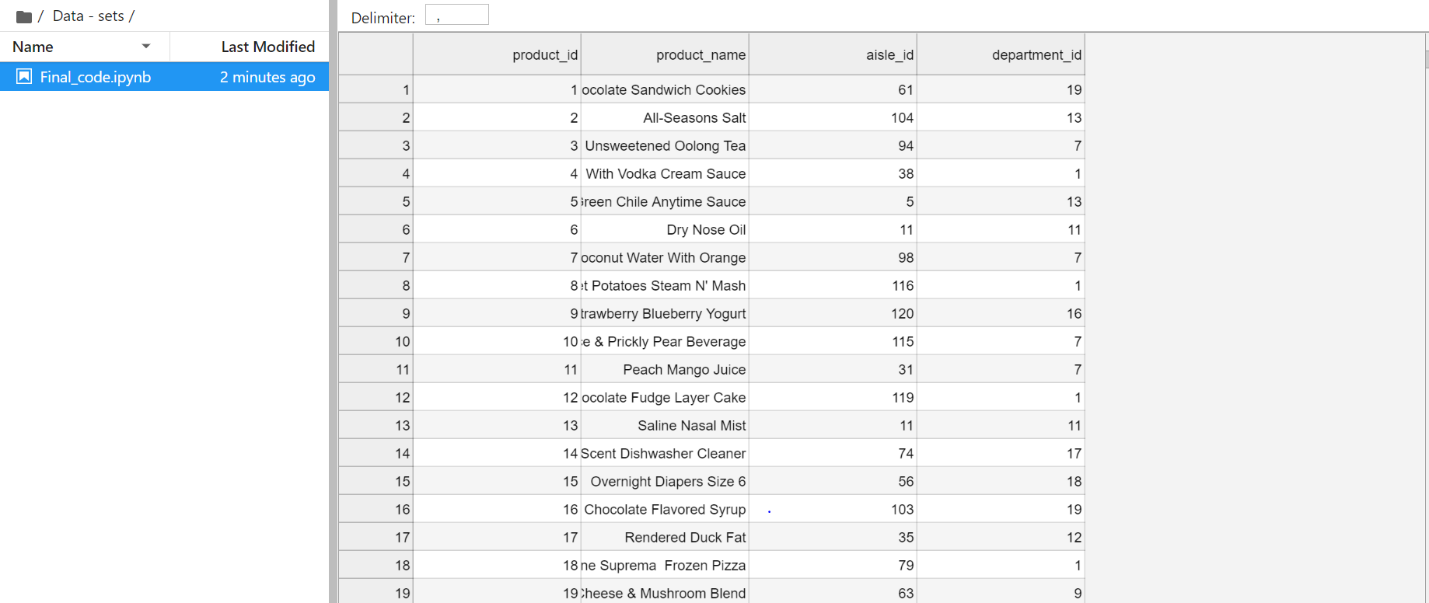
**Creating Aggregated table:**



**Checking Null values:**



**Final out Table****:**



**Step 3:**

Data Pushing to Data Lake

**From Insofe Local Path:**

* /home/fai10103/Capstone\_Project/Data\_sets/output/Retail\_aggregated\_data.csv

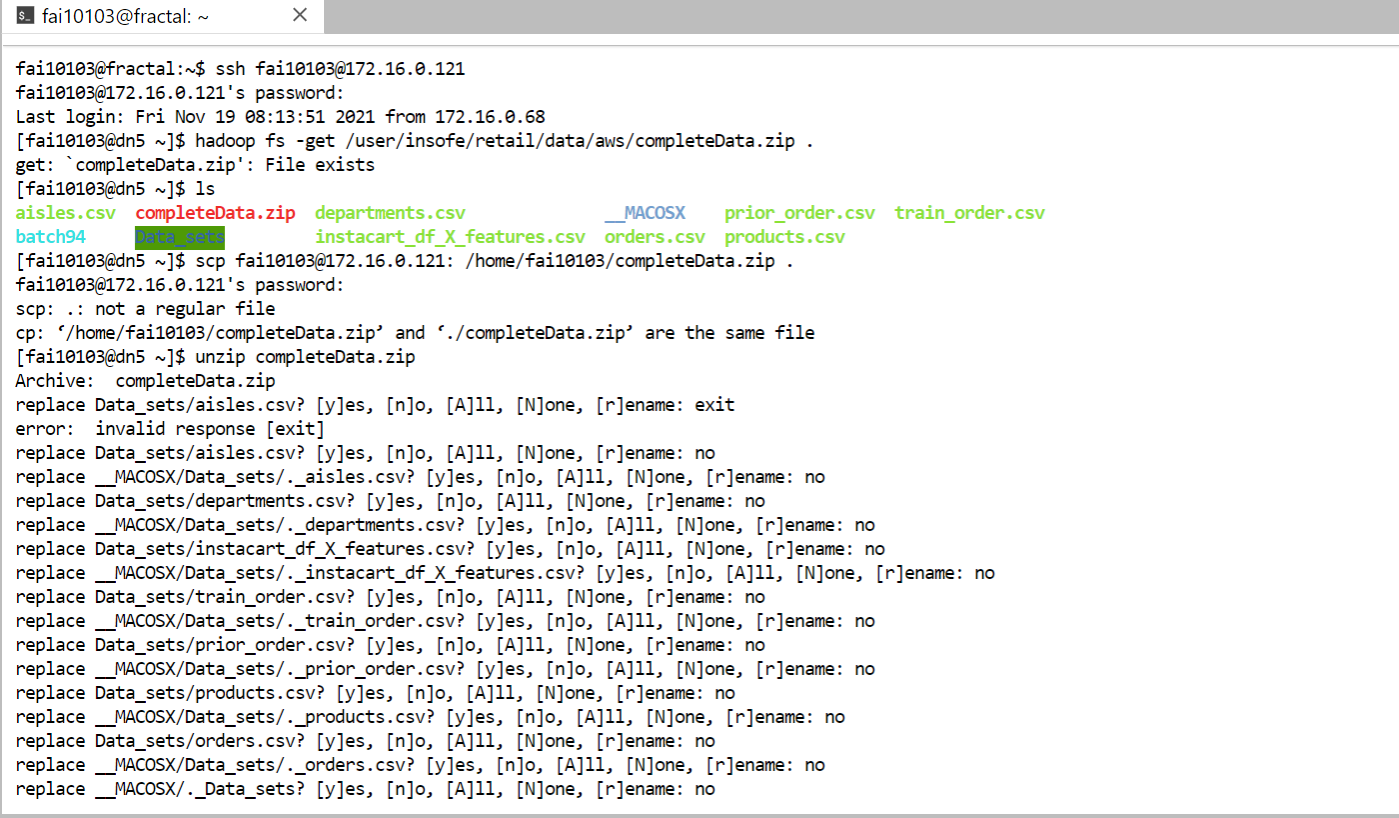
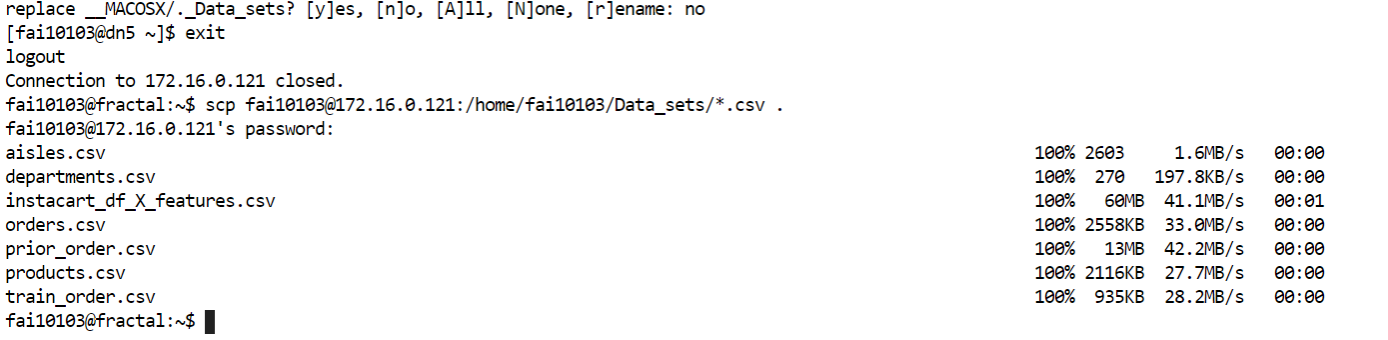
**Output file is copied to HDFS cluster local Path:**

* [fai10103@172.16.0.121:/home/fai10103/](mailto:fai10103@172.16.0.121:/home/fai10103/)

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

* /user/insofe/fai10103/Retail\_Capstone\_Project/output/

Final Structured data pushing to Data Lake



**Step 4:**

Data Visualization

Graphical user interface, application

Description automatically generated

Graphical user interface, application

Description automatically generated Graphical user interface

Description automatically generated

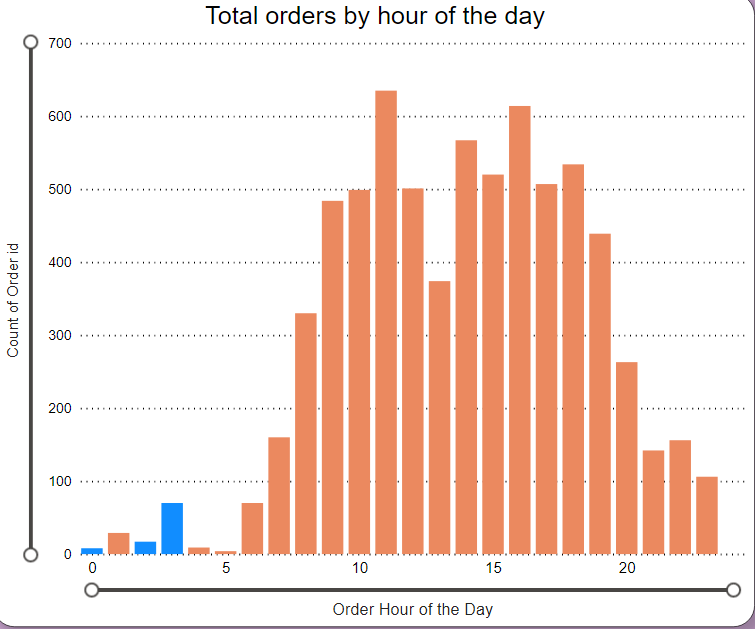
Days Table Hourly Table

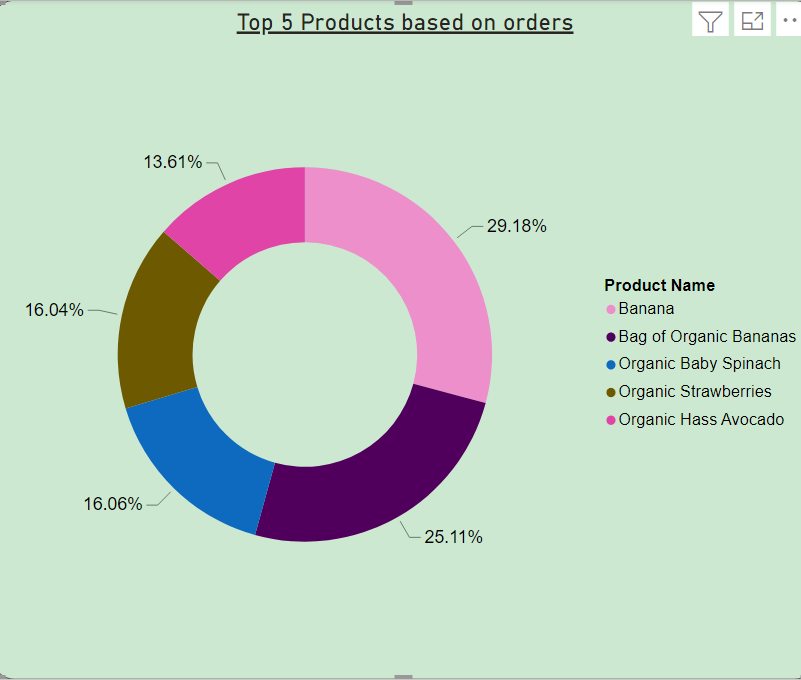
Chart, pie chart

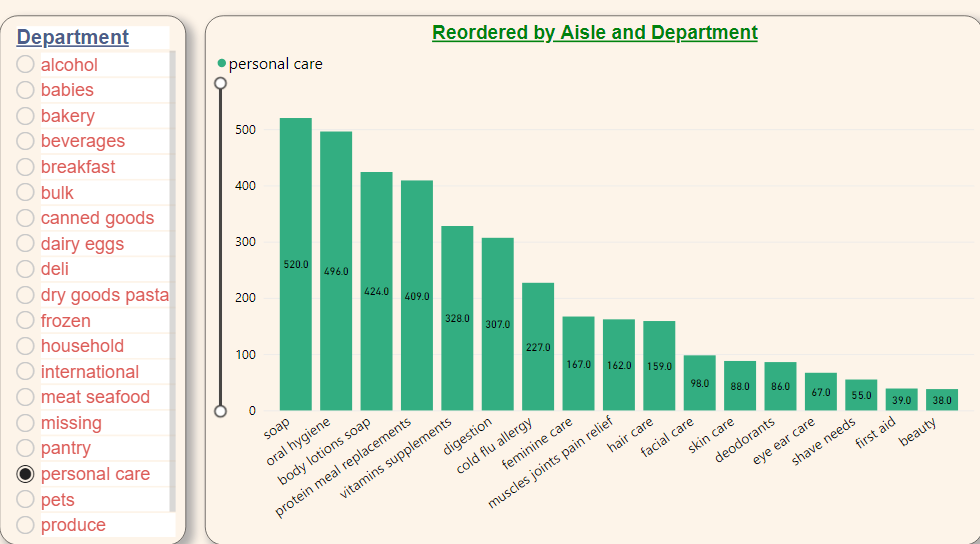
Description automatically generated

Text

Description automatically generated with medium confidence







**Step 5:** GitHub Link

<https://github.com/maheshfractal/Capstone_Project.git>