ETL Pipeline Retail Grocery Store Data Analysis

Table

Description automatically generated with low confidence

Sagar Kalsekar(F06890)

Fractal batch4

sagar.kalsekar@fractal.ai

Data Ingestion Pipeline

**Step No. 1:**

Downloaded \*.csv files from Hadoop cluster locally.

Graphical user interface, text, application

Description automatically generated

**Step No. 2:**

Uploaded all downloaded local files to our local system using SCP command and unzipped the files using UNZIP command.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

**Step No. 3:**

* Started working on transformation using SPARK SQL in Jupyter notebook.

Graphical user interface, text, application, email

Description automatically generated

* Created Spark session and defined schemas.

Text

Description automatically generated

* Extracted data from local and generating output in in output folder.

Graphical user interface, text, application, email

Description automatically generated

* Removing noises from data.

Graphical user interface, text, application, email

Description automatically generated

* Displaying the columns names :

Graphical user interface, text, application, email

Description automatically generated

* Displaying the datatypes of the columns:

Graphical user interface, application

Description automatically generated

Text

Description automatically generated with medium confidence

* Performing Transformation and Displaying records from tables.

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

* Creating aggregation tables.

Text

Description automatically generated

* Output:

Graphical user interface, text, application

Description automatically generated

* Checking null values in the columns:

Graphical user interface, text, application, email

Description automatically generated

* Final output:

Table

Description automatically generated

* Pushing data to Data Lake

Text

Description automatically generated

A picture containing text

Description automatically generated

Text

Description automatically generated with medium confidence

Visualization

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

Chart, pie chart

Description automatically generated

Chart, histogram

Description automatically generated