**ETL Pipeline Retail Grocery Store**

**Data Analysis**

**Diagram

Description automatically generated**

Mihir Shah (F06881)

Fractal Batch 4

[mihir.shah@fractal.ai](mailto:mihir.shah@fractal.ai)

**Data Ingestion Pipeline**

Step 1:

Download unstructured data from Hadoop cluster locally.

**Text

Description automatically generated**

**Text

Description automatically generated**

**Text

Description automatically generated**

Step 2:

Downloading files to our local system using **SCP** command.

**Text

Description automatically generated**

Step 3:

Unzipping the zip files in our local system.

**Text

Description automatically generated**

Step 4:

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

* Setting up environment:

Text

Description automatically generated

* Specifying file schemas.

Text

Description automatically generated

* Extracting Data

Text

Description automatically generated

* Removing noises from products data

A screenshot of a computer

Description automatically generated with medium confidence

* Showing columns of all data frames.

Text

Description automatically generated

* Showing data types of all Data Frames.

Text

Description automatically generated

Text

Description automatically generated

* Transformation and Displaying records from tables.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

* Creating Aggregated table.

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

* Checking Null Values

A screenshot of a computer

Description automatically generated

* Final Output Table

Text

Description automatically generated with medium confidence

Step 5:

Final Structured data pushing to Data Lake

A computer screen capture

Description automatically generated with medium confidence

Text

Description automatically generated

Step 6: Data Visualization:

Chart, bar chart

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