**Final Project**

Felipe Rodriguez

Bellevue University

DSC 650 Big Data

Professor Nasheb Ismaily

March 1, 2024

Product Inventory Management can be a challenge for up-and-coming companies. Identifying trends of transactional data can be beneficial to companies to understand the way their business is running and to understand any potential gaps that may exist. The purpose of this project is to analyze transactional e-commerce data from a UK based company. The data in this study ranges from 01/12/2010 and 09/12/2011. Analyzing this data will give insight to the company’s transactions and will provide stakeholders insight to their sales. The goal is to discover areas that can be improved and create suggestions to improve profits.

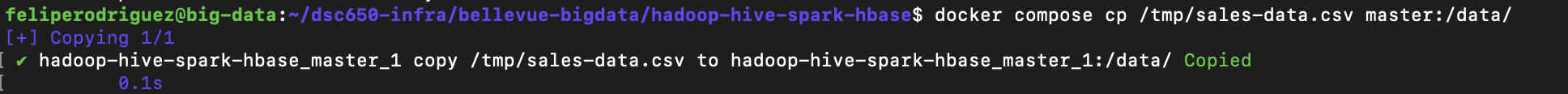
Prior to beginning analytics, the data must be prepared for querying. The tools used in this study are HDFS and Spark. These two are used together in this study because Spark can make use of the data stored in HDFS, spark does not have a system to organize files, while HDFS does (source 1). There are six steps total that are done when preparing the data and querying. The first five steps involve loading the data into HDFS and preparing the data in Spark table for querying. The sixth step is for the queries conducted to reveal insights of the data. Below are the steps laid out and the methods approached.

1. Load data into VM from a local path using scp

A screen shot of a computer

Description automatically generated

1. Copy file into docker container.



1. Add file and confirm it has been loaded into the container

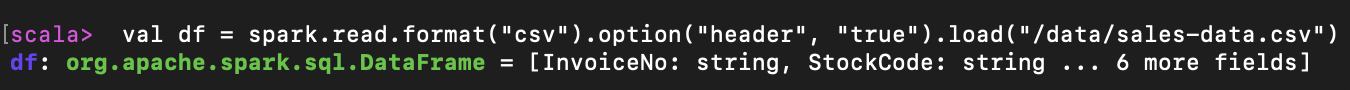
A computer screen with white text

Description automatically generated

1. Create SparkScala Session and load data into Spark from the container.

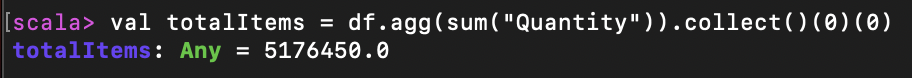
A screen shot of a computer program

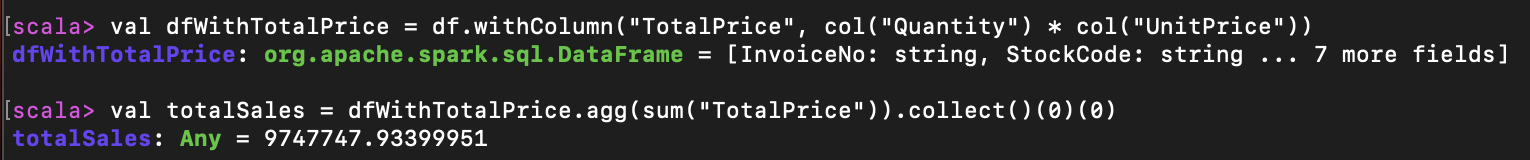
Description automatically generated

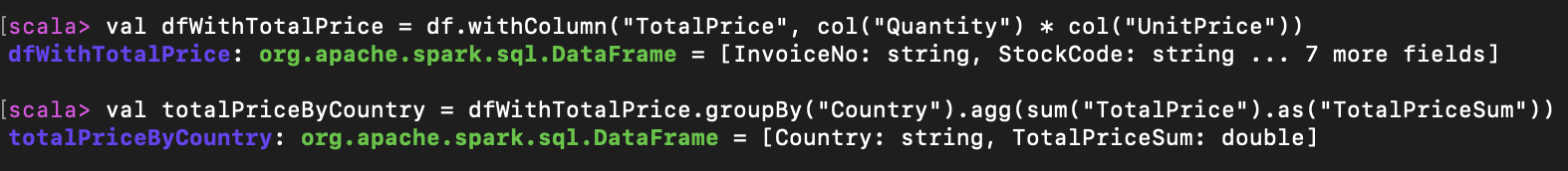
1. A screenshot of a computer screen

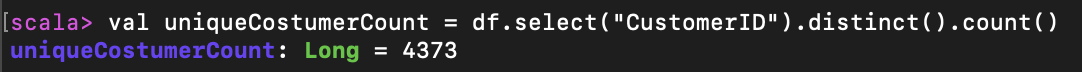
   Description automatically generatedLoad data into table, create temp view and confirm data has loaded.
2. Begin Data Analytics and query the data.

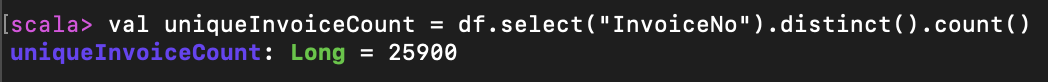


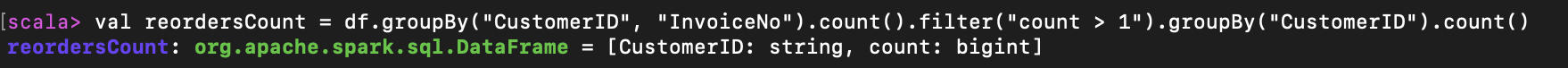
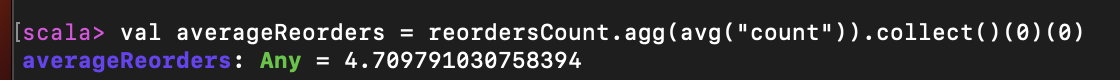




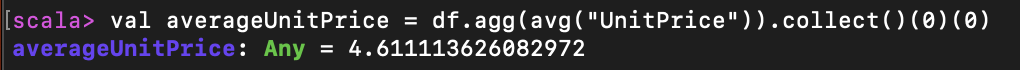


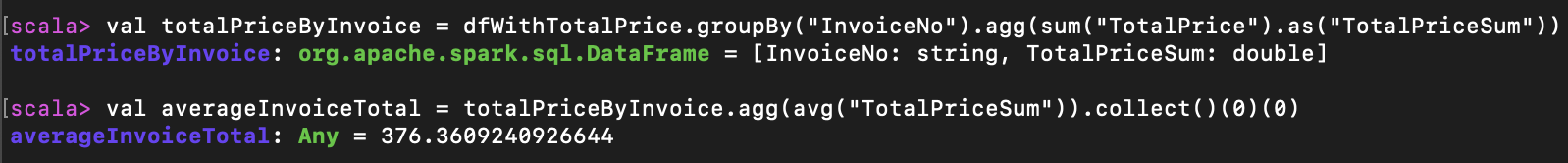
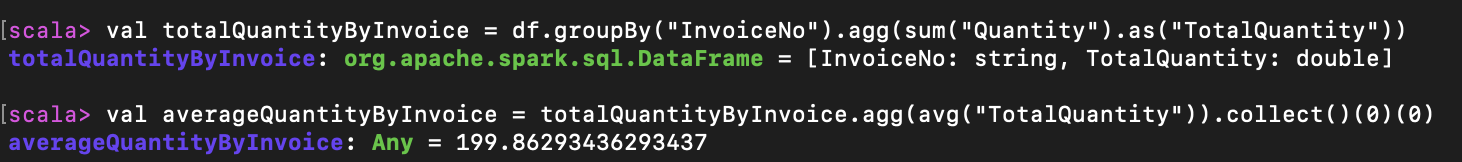




A screenshot of a computer

Description automatically generated





**Findings**

The analysis of this data provided some insights in various areas. It is uncovered that there are a total of 38 countries that products are ordered from. Additionally, there are 4,373 costumer that ordered from 01/12/2010 to 09/12/2011. These customers had a total of 25,900 invoices. The total items sold were 5,176,450 which gave net sales of $9,747,747.94. The average price of the products is $4.61 while the average invoice is $376.36. Each consumer reorders 4.7 times on average. Each invoice had around 200 items. The top ten countries were in the following order: United Kingdom, Netherlands, EIRE, Germany, France, Australia, Switzerland, Spain, Belgium, and Sweden. The Analytical Summary below outlines the data.

**Analytical Summary:** **Top Ten Countries**

|  |  |
| --- | --- |
| Country | TotalPriceSum |
| United Kingdom | 8,187,806.36 |
| Netherlands | 284,661.54 |
| EIRE | 263,276.82 |
| Germany | 221,698.21 |
| France | 197,403.90 |
| Australia | 137,077.27 |
| Switzerland | 56,385.35 |
| Spain | 54,774.58 |
| Belgium | 40,910.96 |
| Sweden | 36,595.91 |

Countries sold to: 38

Unique Customers: 4,373

Total invoices: 25,900

Average Reorders: 4.7

Average Quantity per invoice: 199.86

Average Price per Item: $4.61

Total Items sold: 5,176,450

Total Sales: 9,747,747.94

The findings that stick out the most are the number of items sold and price of the items. When looking at the stats, the price of the items seems low for the quantity being sold. If this was increased by 12 – 25%, the overall margin would improve. Additionally, there seems to be great retention of consumers which is displayed by the number of times they reorder. If the total customer base can be increased, this will create long term business.

In conclusion, the data provides great insight to the company and reveals that there are certain areas for improvement. There are no major areas of concern as the recommendations provide enhancements. This analysis should be refreshed when enhancements have been implemented to analyze the changes and find more opportunities.

Reference:

Carrie. (2017, August 17). *E-commerce data*. Kaggle. https://www.kaggle.com/datasets/carrie1/ecommerce-data

Jeevan, M. (2022, December 14). *Difference between Hadoop and spark: All you need to know*. Simplilearn.com. https://www.simplilearn.com/spark-vs-hadoop-article#:~:text=is%20a%20demand.-,How%20Spark%20and%20Hadoop%20Process%20Data,Distributed%20File%20System(HDFS).