***E-Business***

Final Project Presentation  
Spring 2020  
Prof. James Curry

Group 4

Amita Akole  
Yongian Cao  
Vishal Goyal  
Shivam Khare  
Sunil Kumar  
Sushil Satya

**I. Introduction:**

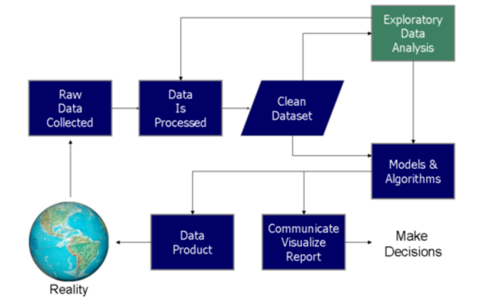
E-commerce is a very dynamically evolving industry, and this is primarily because of its underlying ever-changing technology. Companies like Amazon, EBay are capable of building predictive algorithms being executed in real time in a big data environment.

The main vision of our project is to successfully provide other businesses with knowledgeable insights from Amazon sales to see a remarkable growth in company’s sales and to increase company’s profit. In this project we will be focusing on:

* Analyzing the previous sale’s record to know what items customers purchased previously
* What is in their online shopping cart or on their wish list
* Which products they reviewed and rated, and what items they search for most.

According to Wikipedia, EDA “is an approach to analyzing datasets to summarize their main characteristics, often with visual methods”.

In my own words, it is about gaining important information, future insights and statistically backed predictions based on data that is already available to us through various Mining processes.



**II. Overview of the Dataset:**

1. There are 8 columns in our E-Business dataset.
2. The column names are Event\_time, Event\_type, Product\_Id, Category\_Id, Brand, Price, User\_Id, Session\_Id, and City\_Id respectively.

· Event\_time: Record what time customer is frequently visiting the website.

· Event\_type: What is the behavior of the customer doing? Like add product into cart,

Remove product from the cart or they just take a look at the product.

· Product\_Id: What is the id of Product?

· Category\_Id: Selected product belongs to which category

· Brand: The brand name of product.

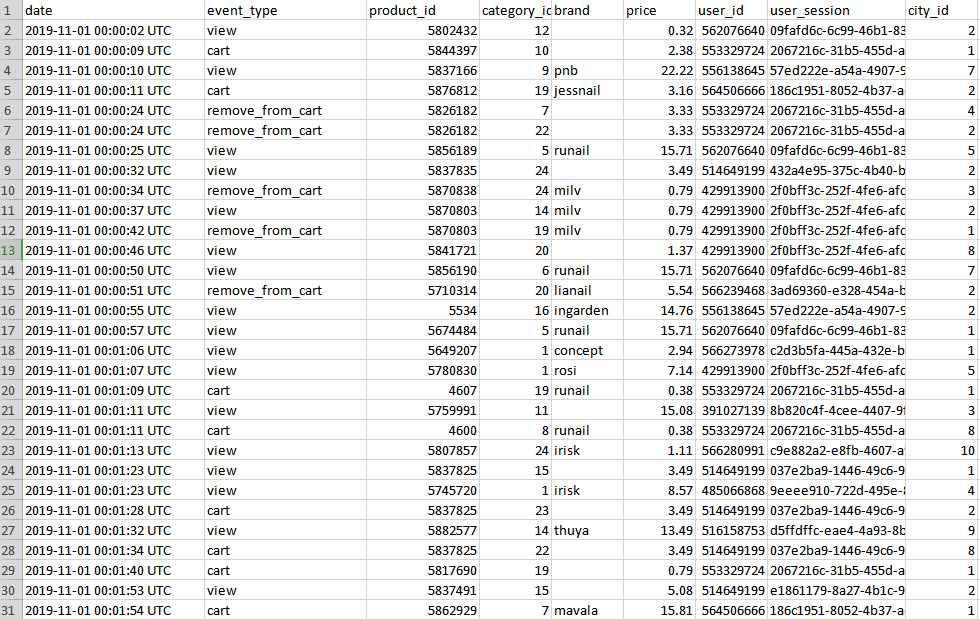
· Price: The single price of product.

· User\_Id: The id number of this user.

· Session\_Id: The Session number of one website access behavior.

· City\_Id: The location of visiting action.

1. The following picture shows an overview of how the dataset is created in our software.



**Strength:**

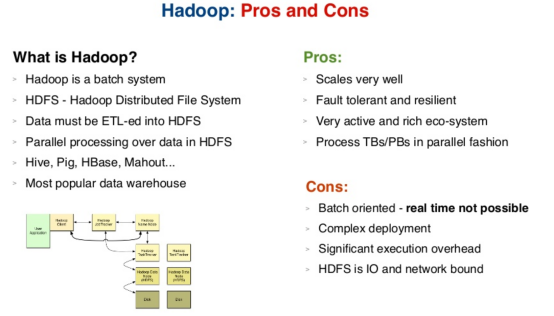
1. This dataset is easier to understand and format friendly rather than json format.
2. This dataset is simple to maintain. Sometimes, we delete some data occasionally. We can restore by using SQL sentences in our information database.

**Projected Outcomes:**

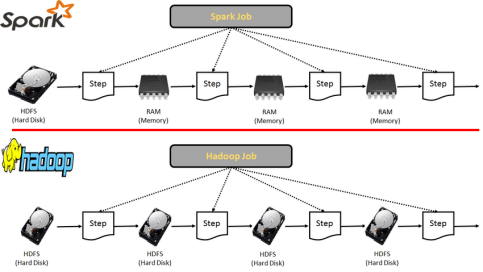
1. Our project will show the need to know how many items are bought by each client and discover the foremost prevalent one.
2. Depending on that result, the information will be used to recommend additional products that other customers purchased when buying those same items. Additionally, this will show results of which product is the most popular, what season/ occasion people prefer for online shopping so that we can advertise more during that phase.

**Differences between Hadoop and Spark:**

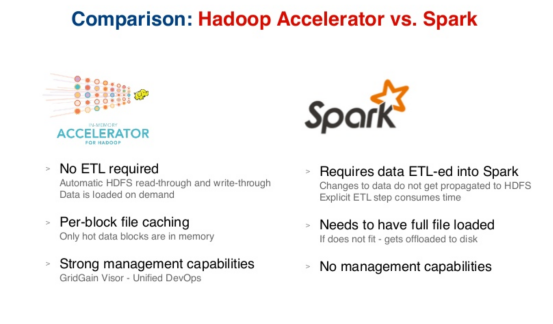
Both Hadoop vs. Spark are popular choices in the market; let us discuss some of the major difference between Hadoop and Spark:



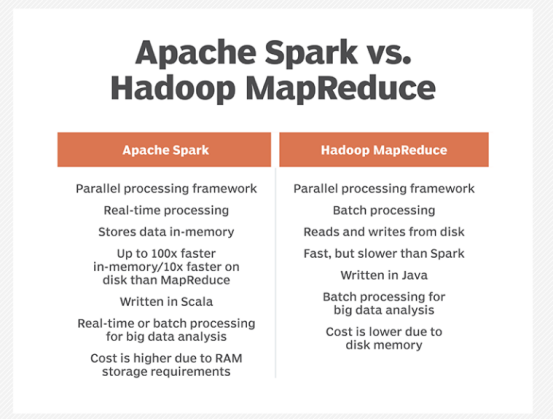
1. Hadoop is an open source framework which uses a MapReduce algorithm whereas Spark is lightning fast cluster computing technology, which extends the MapReduce model to efficiently use more types of computations.



1. Hadoop requires developers to hand code each and every operation whereas [Spark is easy to program with RDD](https://www.educba.com/rdd-in-spark/) –Resilient Distributed Dataset.



1. Hadoop’s MapReduce model reads and writes from a disk, thus slowing down the processing speed whereas Spark reduces the number of read/write cycles to disk and stores intermediate data in-memory, hence faster-processing speed.



**III. Simulation:**

The following diagrams are coding sections for product count, city count, and visiting time count of the customers with their respective outputs.

1. Product count result hypothesis: We assume customers prefer to buy some daily necessaries, books, and clothes.
2. City count result hypothesis: We assume some higher population cities present a higher sales count like in New York City, California.
3. Time count result hypothesis: We assume customers would like to visit our website after working time or before bedtime.

**IV. Production Data Set & Analytical Data Set:**

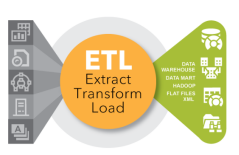
**Cleaning Dirty Data:**

a) **What is ETL?**

ETL is a type of data integration that refers to the three steps (extract, transform, and load) used to blend data from multiple sources. It is often used to build a [data warehouse](https://www.sas.com/en_us/insights/data-management/data-warehouse.html). During this process, data is extracted from a source system, transformed into a format that can be analyzed, and loaded into a data warehouse or other system. Extract, load, transform (ELT) is an alternate but related approach designed to push processing down to the database for improved performance.

b) **Why is ETL important?**

* When used with an enterprise data warehouse (data at rest), ETL provides deep historical context for the business.
* By providing a consolidated view, ETL makes it easier for business users to analyze and report on data relevant to their initiatives.
* ETL can improve data professionals’ productivity because it codifies and reuses processes that move data without requiring technical skills to write code or scripts.
* ETL has evolved over time to support emerging integration requirements for things like streaming data.
* Organizations need both ETL and ETL to bring data together, maintain accuracy and provide the auditing typically required for data warehousing, reporting and [analytics](https://www.sas.com/en_us/insights/analytics/what-is-analytics.html).

****

c) **Why do we need to ETL our dataset?**

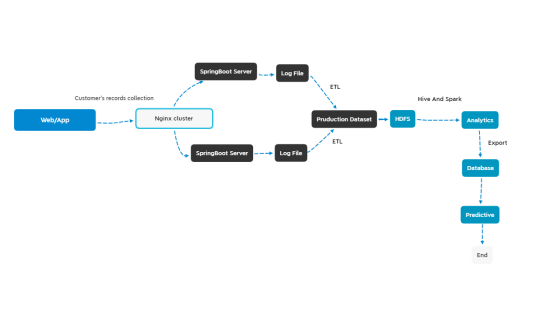
Sometimes, a Web server cannot collect some of the customer’s information and the value is missing. These missing values might affect our results if we did not filter them. In this case, we filter the user-id, product-category-id these columns value is null.

****

**d) Difference between Analytical Data Set and Production Data Set:**

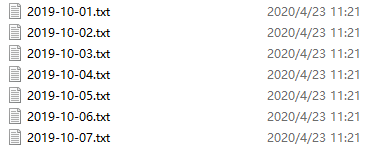
* An analytical dataset is a dataset that is generated by manipulating data through merges, the creation of new fields, application of filters, and so on.
* A production database contains the data used for production tasks such as creating and updating features. Depending on the data model we will be using, data in a production database can be used to create a digital or hard-copy map or chart or a specific type of data.

**e) Data Process Representation:**



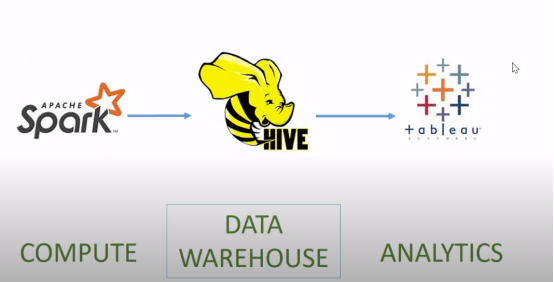
**f) Data mining:**

We want to know every day’s customer’s count and calculate every day’s new customer, the active customer. This business request helps us to find some problems. For example: Our company average customer’s number is 20000 a day but there was 10000 yesterday. So we can test some problems based on this result. Firstly, we need to put customer’s data into different files by their visiting time and we cannot change the data. The output is a text format file.



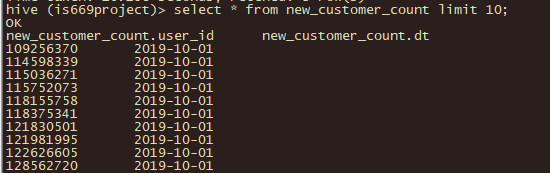
**g) Tool:**

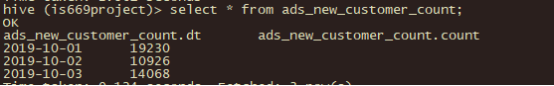
Apache Hive is a data warehouse system for data summarization and analysis and for querying of large data systems in the open-source Hadoop platform. It converts SQL-like queries into Map Reduce jobs for easy execution and processing of extremely large volumes of data.



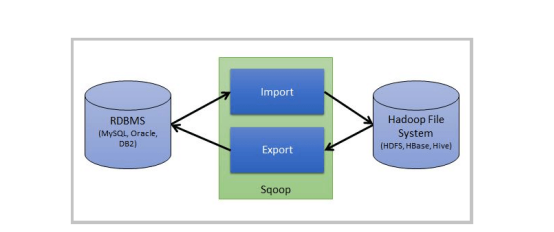
In this case, we need to upload our results into Hadoop and take some HIVE analytics. We use scripts to load data into the data warehouse. The code is in the photo 2.2.

We want to calculate how many new customers every day. We use SQL to join the ‘active customer table’ and ‘new customer table’. For example, if some customers appear in today’s file but not in our database. It shows these customers are new. We calculate every day’s new customer count and active customer count respectively. The data split code is in the 2.2-2.

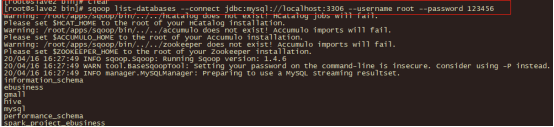


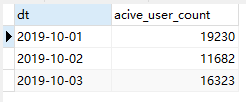


Sqoop is an automated set of volume data transfer tools which allows simple import, export of data from structured based data which stores NoSql systems, relational databases and enterprise data warehouses to Hadoop ecosystems.



We prepare to export the result from our cluster into our database by Sqoop. The reason we take this step is because the manager is easier to observe the result every day and this way helps us to take other business requests. Like, calculate the ratio between new customers and active customers every day. We set the directory of our result and import the destination of our database.



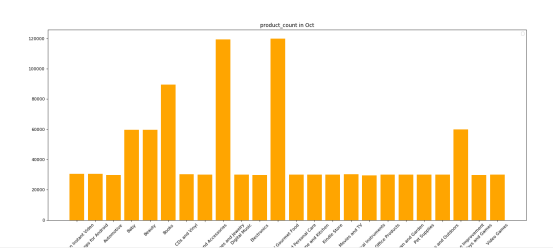


**V. Outcomes of Visualization and Recommendation:**

1. **Visualization:**

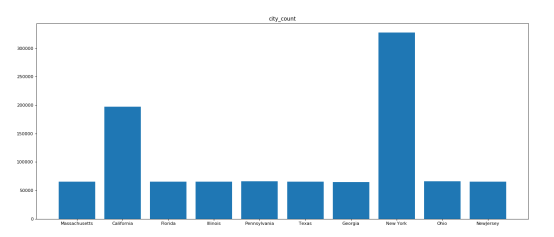
For visualization, we took some business data, then visualized the result and proved our assumption.

1. **Product number count:-**

To know some products that are currently popular in the market and product manager based on our result to extend more size, color, and feature in them we need to visualize every database’s outcome. Firstly, we depict the product count’s result by Python. We need to connect to our database and get the result in this table. Secondly, we convert the data into the table format because we need to set the x-axis and y-axis. We choose the bar diagram because we want to know some more. The last step is to set the picture size and show it. The code is in the photo 3.1.  


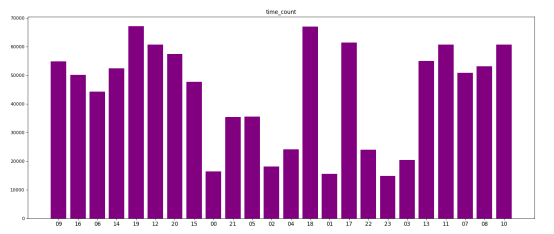
1. **City number count:-**

The second part is showing some popular cities. The advertisement colleagues want to know which city the most is worthy to put their ads. This result helps the advertisement department colleagues to calculate every city's budget and convert their ROI. The detailed code is in the 3.2.



1. **Visiting period count:-**

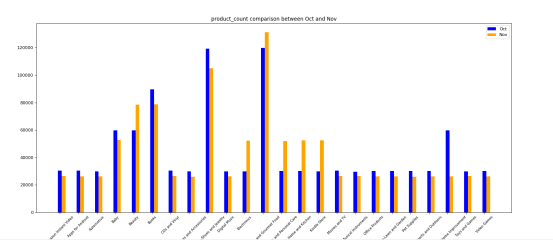
This result means we understand which time customers would like to visit our website and buy our product. It helps us to allocate our web resources. The code is in the photo 3.3.



**VI. Result comparison between Production Data Set & Analytical Data Set:**

**1. Product count result hypothesis:** We assume customers prefer to buy some daily products like electronics, books, and clothes, etc.

Analytical dataset Production dataset  

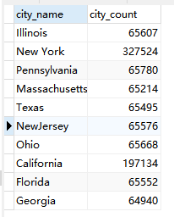
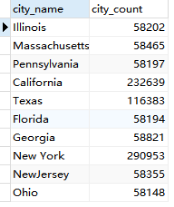


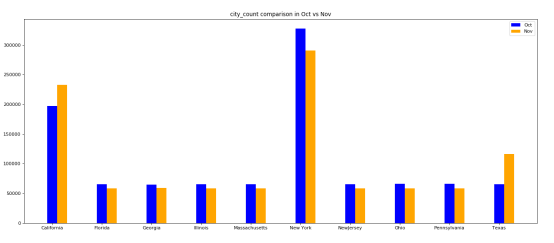
**a) Hypothesis 1 Conclusion:**

Based on the last picture, we see the product “books” is the main source during October and November. Beauty and clothes products are also popular in the market. We should provide this result to the product departments. It helps them to explore more size, color, and features in these kinds of things.

**2. City count result hypothesis:** We assume some higher population cities present a higher sales count like in New York City, California.

Analytical dataset Production dataset

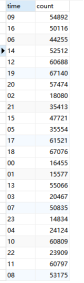
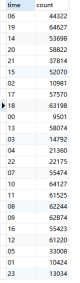
 

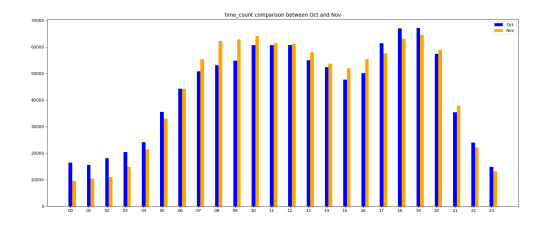


**a) Hypothesis 2 Conclusion:** Base on the comparison between Oct and Nov. Our assumption about some big cities might have higher sales than others are correct. New York City and California present a higher customer’s number. We should put more advertisements in these areas to find more potential customers.

**3. Time count result hypothesis:** We assume customers would like to visit our website after their working hours or before bedtime.

Analytical Data Set Production Data Set



**a) Hypothesis 3 conclusions:** We guess customers would like to visit our website during the evening period than at night. It seems our hypothesis is right. We need to allocate our advertisement in the evening time to improve the ROI.

**VII. Predictive Analysis:**

1) The main aim of our project is to analyze the previous sales record to know what items customers purchased previously, what is in their online shopping cart or on their wish list, which products they reviewed and rated, and what items they search for most. These analytics help us to figure out every customer’s behaviour. These analyses help us to make decisions to increase the company’s profit. For that purpose, we will be using Predictive Analytics in our project.

2) Predictive Analysis helps us with a deeper understanding of customer habits and preferences. Predictive search capabilities have been built-in to an analytics solution that allows us to analyze customers past click-through behaviours, shopping history, and product preferences, in real-time.

3) We will be implementing Predictive Analysis for the same purpose as it allows continuous analysis of customer data and will provide the most relevant results and recommendations to users.

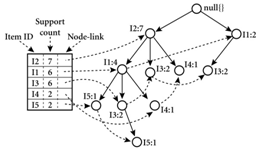
**VIII. FP-Growth Algorithm:**

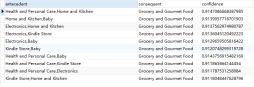
1) In our project, we will use the FP-Growth algorithm to find the connection between products and customers as it has proved to be the most advanced and efficient implementation of frequent pattern mining.

2) The FP-Growth Algorithm is an alternative way to find frequent products without using candidate generations, thus improving performance. For so much, it uses a divide-and-conquer strategy. The core of this method is the usage of a special data structure named frequent-pattern tree, which retains the item-set association information.

3) The mining data is decomposed into sub-datasets according to the frequent patterns identified which leads to the more focused search of smaller databases.

4) Due to the divide and conquer strategy, FP-Growth Algorithm is the most efficient method for data mining. Additionally, there are no repeated scans of the whole database.



Product relations in October Product relations in November  

**IX. Conclusion:**

It’s clear that baby and beauty stuff presents a stronger relation with grocery things in October records. Thus, we can display some beauty and baby products when customers search for baby products. And health products connect with the grocery foods in November records. As the same, we need to show more health related products when people visit grocery departments. The code is in the 4.1.

# **Current State:**

While at the beginning of recommendation systems it was important to find explicit similarity in people and products, a more effective method has been used to look at the similarity of latent attributes. This is done by using matrix factorization. To oversimplify, all of the attributes for an item or a customer are combined in a way that reveals relationships that have not yet been realized.

# **Future State — Deep Learning**

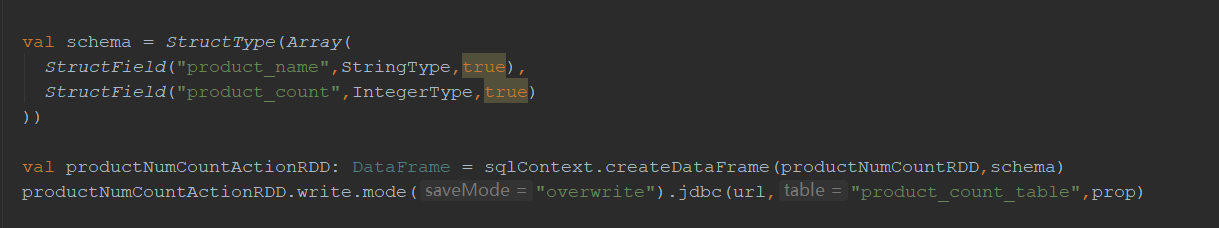
* Recommendation systems can be a very powerful tool in a company’s arsenal, and future developments are going to increase the business value even further. Some of the applications include being able to anticipate seasonal purchases based on recommendations, determine important purchases, and give better recommendations to customers which can increase retention and brand loyalty.
* Most businesses will have some use for recommendation systems, and I encourage everyone to learn more about this fascinating area.

1. **Limitations:**

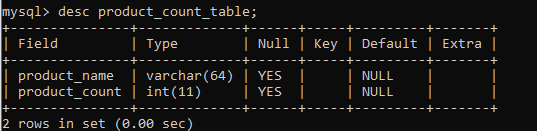
* **Missing Values:** There are some missing values in the data set and we cannot take a few data to make some prediction.
* **Spreader Dataset:** The city name and category name in the other document. We need to take lots of “join sentence” in SQL. It takes more time and has adverse effects on code efficiency.
* **Gender Data Missing:** The data set doesn’t show the user’s gender. It causes a little inconvenience in some analysis that we will be performing.

**X. ARTIFACTS:**

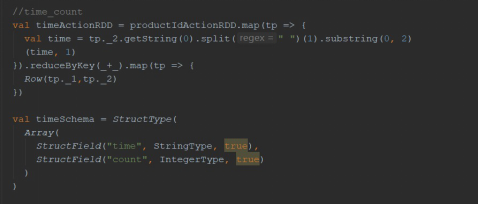
Code Screenshots:



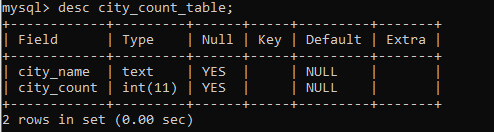
1.1 - Product-Count Table Creation



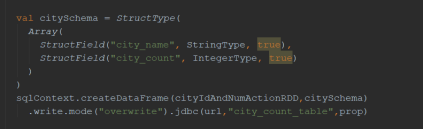
1.1-2 Product count table description



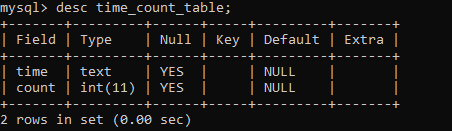
1.2 - Time-Count Table Creation



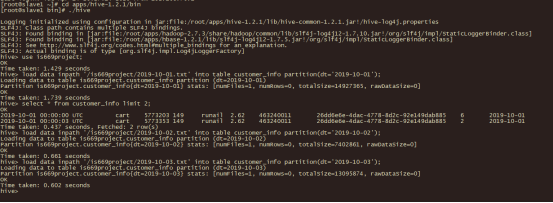
1.2-2 City count table description



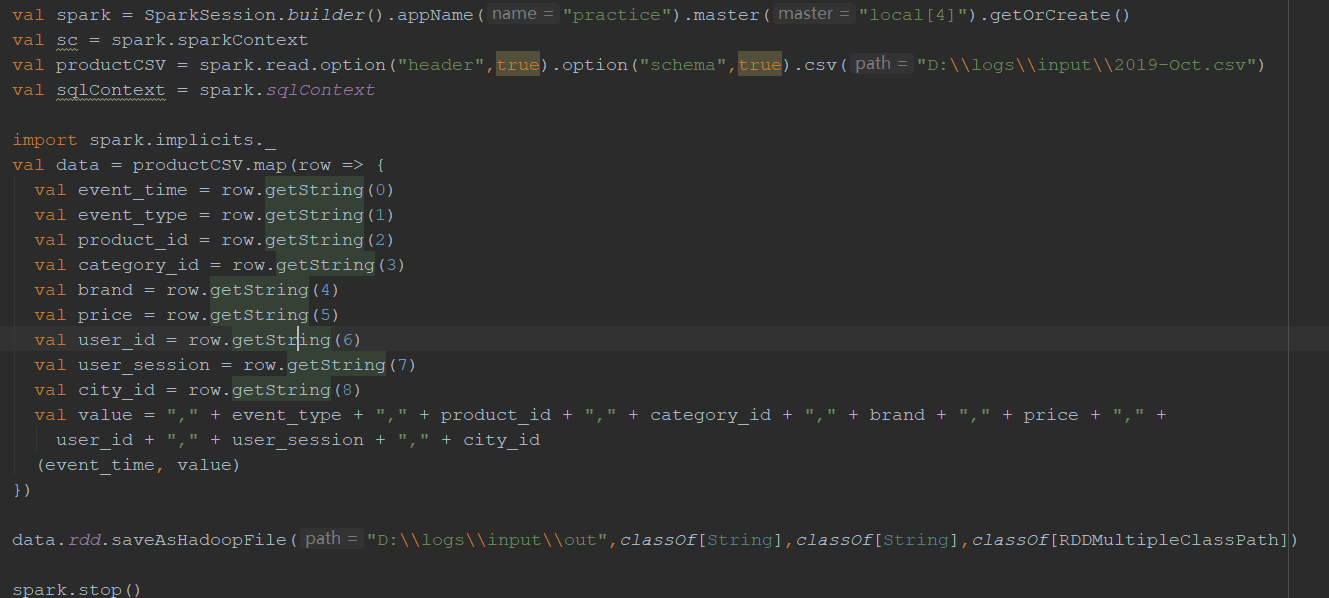
1.3 City-Count Table Creation



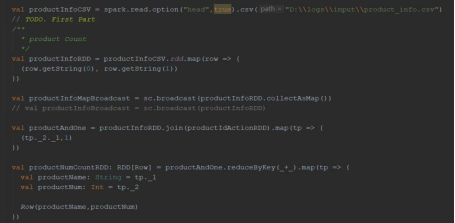
1.3-2 City-Count table description



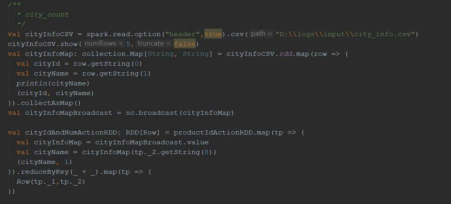
2.2 - Dataset load in Hadoop



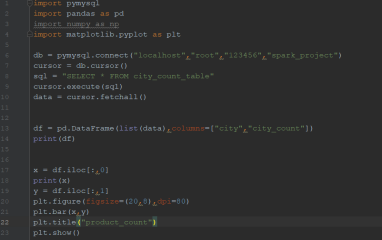
2.2-2 Data split



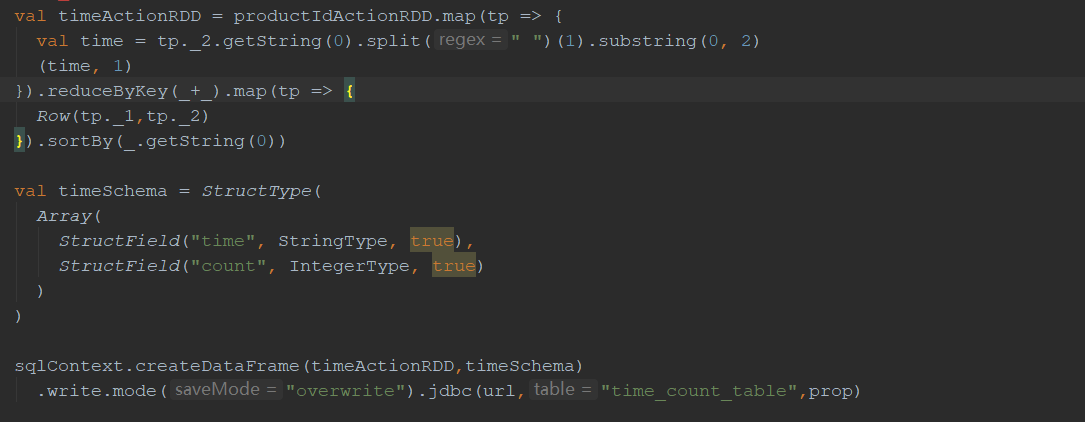
3.1 - Product count code



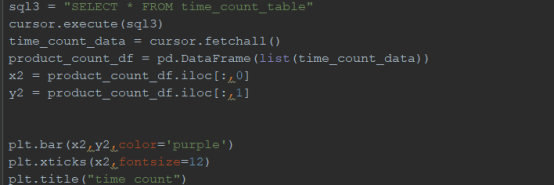
3.2 - City count code



3.2.2 - City count result visualization



3.3.1 - Time count code



3.3.2 - Time count result visualization



4.1 Product-relations code

References:

<https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/>

<https://www.bigdataframework.org/short-history-of-big-data/>

<https://www.scnsoft.com/blog/spark-vs-hadoop-mapreduce>

<https://intellipaat.com/blog/what-is-apache-spark/>

<https://towardsdatascience.com/recommender-systems-in-practice-cef9033bb23a>

<https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_FP-Growth_Algorithm>