

**ALY 6110 Data Management & Big Data**

**81984**

**Final Project Report**

**Group Epsilon**

**E-commerce Behavior Data**

Keyur Shah ([shah.keyur@husky.neu.edu](mailto:shah.keyur@husky.neu.edu))

Khushboo Oza ([oza.kh@husky.neu.edu](mailto:oza.kh@husky.neu.edu))

Reetika Chaturvedi ([chaturvedi.re@husky.neu.edu](mailto:chaturvedi.re@husky.neu.edu))

Rishhi Kumar ([janakiraman.r@husky.neu.edu](mailto:janakiraman.r@husky.neu.edu))

Shruti Sawale([sawale.s@husky.neu.edu](mailto:sawale.s@husky.neu.edu))

**SUMMARY**

For the final project, we have selected the dataset “eCommerce behavior data from multi category store” taken from Kaggle. The main motive is to perform analysis over this dataset to understand how to apply big data concepts over a large real world associated dataset to know its application in solving real world problems and achieving better business solutions using big data tools like Apache Spark, Spark SQL and Pyspark.

With the selected dataset, we are going to solve certain business questions that will help in achieving effective marketing strategies and identifying which products or categories generate maximum revenue. The percentage of people buying the products and the products that are more preferred come under which category or brand is analyzed using the Spark SQL query. Which category and brand generates maximum revenue will give an insight of customer behaviour on e-commerce.

**CONTENT**

The dataset comprises 4 million rows with unique data and 9 attributes, which are event\_time, event\_type, product\_id, category\_id, category\_code, brand, price, user\_id, user\_session. The attributes that we used are “event\_type” which is used to identify how many people have viewed a product, how many of them have added that product into their cart and how many of them have actually purchased it, then the “category\_code”, that is used as a main category under which further classifications can be done, it is used for easy accessing and sorting of the products. The attribute “brand” further classifies the products according to the various companies that manufactures it. “Price” attribute is used to calculate the revenue that a particular product or category generates.

The business questions that we used to analyze the dataset are:

* Which product and category generates the highest revenue?
* Which brand of particular category is most preferred?
* Customer Behaviour on ecommerce?
* Highest Revenue generating brands under a particular category?

By answering this question, we can know about the customer approach towards the e-commerce sites. The technologies that we used for our project are Apache Spark to handle this big amount of data. We launched Jupyter Notebook and did data cleaning in Python and then launched PySpark in Notebook. Once the data was clean after removing n/a and missing values we used Spark SQL for queries and obtaining the output. Spark SQL is used for structured data processing and can be used as a distributed SQL query engine.

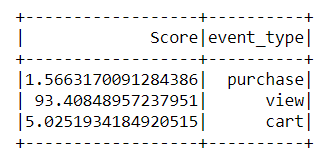
The code here is used to know what percentage of people purchase the product, how many just view the products and how many put the product in their cart.

**Query:**

**df6 = sqlContext.sql("Select (count(event\_type)\* 100 / (select count(\*) from table)) as Score,event\_type From table group by event\_type")**

**df6.show()**

**Output:**



From the obtained result, we came to know that **93.40%** (approx) of the people only view various products whereas just **1.566%** (approx) of the people actually buy the products.

Next code is used to figure out which brand of particular category generates how much revenue.

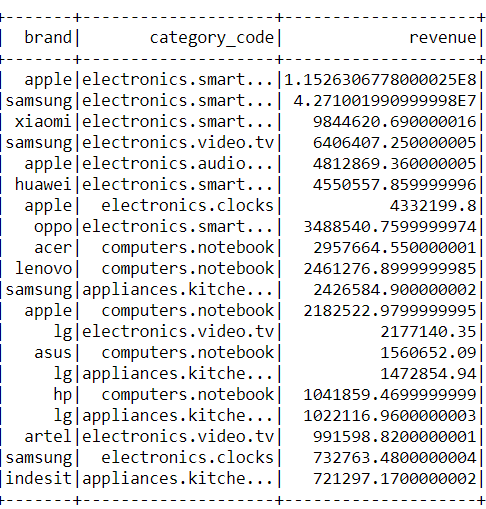
**Query:**

**df6= sqlContext.sql("SELECT brand,category\_code,sum(price) as revenue from table where event\_type='purchase' group by category\_code, brand order by revenue desc")**

**df6.show().format()**

**df6.withColumn("revenue", format\_number("revenue", 2))**

**Output:**



From the above obtained output, we came to know that the brand **“apple”** under **“electronics.smartphone”** and **“electronic.audio.headphones”**, **“Samsung”** under **“electronics.video.tv”** and **“appliances.kitchen”** generate the highest revenue.

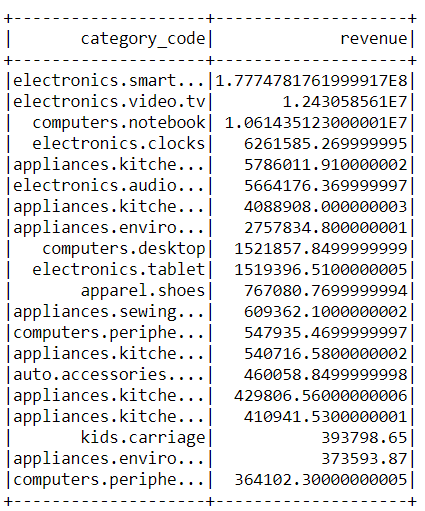
The below mentioned query is to evaluate category generating highest revenue.

**Query:**

**df6= sqlContext.sql("SELECT category\_code,sum(price) as revenue from table where event\_type = 'purchase' group by category\_Code order by revenue desc")**

**df6.show()**

**Output:**



From the above table, it is observed that maximum revenue is generated with the sales of **“electronics.smartphone”** which is followed by **“electronics.video.tv”.**

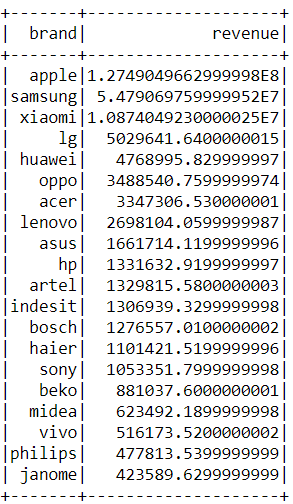
The next code is to figure out which brand generates how much revenue.

**Query:**

**df6= sqlContext.sql("SELECT brand,sum(price) as revenue from table where event\_type = 'purchase' group by brand order by revenue desc")**

**df6.show()**

**Output:**



The above table shows that **“apple”** is the brand generating maximum revenue followed by **“Samsung”** and **“Xiaomi”.**

The code now is used to know which brand under a particular category is most purchased among all.

**Query:**

**df6= sqlContext.sql("SELECT count(brand) as counts,brand,category\_Code from table where event\_type='purchase' group by category\_Code,brand order by counts DESC")**

**df6.show()**

**Output:**



From the obtained output, we came to know that the brand **“samsung”** under category code **“electronics\_smartphones”** is the most purchased one.

From analyzing the dataset, we came to know that **“Apple”** smartphones generate the highest revenue but the most preferred smartphone by customers is **“Samsung”**.

**COMMENTS**

We performed our analysis on Spark SQL. We faced problems while installing and configuring the application of Apache Spark like we had to set several environment variables as a result, we followed the straight forward approach of standalone deployment. The next issue was while loading the data on Apache Spark server which was regarding the availability of the space as per the requirement because of the huge dataset. To resolve this issue, we had to install certain libraries and packages which are com.databricks.spark.csv, etc.

After loading the data on the server, we used sparkSQL context for creating data frame. We had converted this data frame in to data tables. After which, we interacted with the data using SQL queries. The queries are executed by the executing engine of Spark. Formulating correct queries that will fetch accurate results was one of the major challenges that we faced while working on the data. Since the data was huge, we were thoughtful about queries but running them turned out to be a smooth process and we obtained results faster than expectations.

Apache Spark is highly useful when it comes to managing a big amount of data. The simplicity of the tool is amazing. It helps in quick and easy data interaction via application interface. PySpark works superfast and Spark SQL provides a quick output. Handling this data is quite impossible without big data tools. Big data tools like Spark comes as a support and easy way to access this data and at the same time helps in decision making. Spark can prove to be a boon for the organizations as it provides them with a lot of information in minimal time.

**CONCLUSION**

Big data plays a great role in the eCommerce industry. It helps in predicting the trends and finding the best-selling products so that they are able to provide customer satisfaction and stand competition in the market.

The business questions that we analyzed were much relevant for the retail stores to improve upon their product availability and ensure which products would provide them with continuous revenues leading to high profits.

After analyzing the number of purchases taking place, the brand that contributed maximum to the revenue and the product that generated the most revenue is “Apple” under the “electronic.smartphone” category. This helped in establishing a connection with the demands that exist and enable them to provide requisite supply.

Big data also helps a lot in predicting the estimated demand, trends and patterns that exist in the market. Not only retailers are able to be ready with the customer’s needs but also able to plan their stock accordingly.

Also, it helps them to decide which products do not require their investment in terms of revenue generation. Along with this, we also came to know that the “Samsung” smartphone is the most preferred product amongst customers. This comparison between brands can help manufacturers to differentiate and come up the points where they lag behind due to which their products are sold in lesser number, as a result they can improve their products for obtaining better stand in the market.

**REFERENCES:**

1. Kechinov, M. (2019, December 9). eCommerce behavior data from multi category store. Retrieved from<https://www.kaggle.com/mkechinov/ecommerce-behavior-data-from-multi-category-store>
2. Bochet, C. (n.d.). Get Started with PySpark and Jupyter Notebook in 3 Minutes. Retrieved May 14, 2020, from <https://www.sicara.ai/blog/2017-05-02-get-started-pyspark-jupyter-notebook-3-minutes>
3. Pyspark.sql module¶. (n.d.). Retrieved May 15, 2020, from <https://spark.apache.org/docs/1.4.0/api/python/pyspark.sql.html>
4. Penchikala, S. (2015, January 30). Big Data Processing with Apache Spark – Part 1: Introduction. Retrieved from [https://www.infoq.com/articles/apache-spark-introduction /](https://www.infoq.com/articles/apache-spark-introduction%20/)

**DATASET**



**PROJECT MODEL**

****