

# E-commerce Customer Behavior Analysis Report

## 1. Introduction

The dataset used for this analysis can be accessed from the following Kaggle URL: E-commerce Customer Behavior Dataset. It contains a wealth of information that allows us to explore various aspects of customer behavior in the e-commerce domain.

To facilitate the analysis, we set up a Hadoop Docker cluster. The Hadoop cluster played a crucial role in storing and preprocessing the large datasets of user logs and transaction data. Additionally, we utilized Apache Spark, a powerful distributed computing framework, to perform machine learning algorithms for analyzing customer behavior and predicting future buying patterns.

## 2. Methodology

The analysis process involved several steps, as outlined below:

```
norhanswar@norhanswar-HP-Laptop-15s-fq5xxx: ~/Career/Master/BigData/Labs/Lab2Material-20231203/docker-spark-cluster$ sudo docker ps
[sudo] password for norhanswar:
CONTAINER ID   IMAGE                                     COMMAND                  CREATED        STATUS        PORTS
a5a702552e2b   cluster-apache-spark:3.0.2             "/bin/bash /start-sp..." 6 hours ago    Up 5 hours    0.0.0.0:7077->7077/tcp, :::7077->7077/tcp, 7000/tcp, 0.0.0.0:8080->8080/tcp, :::8080->8080/tcp
b775576b3838   bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8 "/entrypoint.sh /run..." 6 hours ago    Up 5 hours    0.0.0.0:9864->9864/tcp, :::9864->9864/tcp
bdbb635916db   bde2020/hadoop-nodemanager:2.0.0-hadoop3.2.1-java8 "/entrypoint.sh /run..." 6 hours ago    Up 5 hours    8042/tcp
a2d82f6b9ee5   bde2020/hadoop-historyserver:2.0.0-hadoop3.2.1-java8 "/entrypoint.sh /run..." 6 hours ago    Up 5 hours    8188/tcp
11d378fb5a32   bde2020/hadoop-resourcemanager:2.0.0-hadoop3.2.1-java8 "/entrypoint.sh /run..." 6 hours ago    Up 5 hours    8088/tcp
ad858515a992   bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8     "/entrypoint.sh /run..." 6 hours ago    Up 5 hours    0.0.0.0:9870->9870/tcp, :::9870->9870/tcp, 0.0.0.0:9010->9010/tcp, :::9010->9010/tcp
3bf1fbc11f35   cluster-apache-spark:3.0.2             "/bin/bash /start-sp..." 4 days ago     Up 5 hours    0.0.0.0:7000->7000/tcp, :::7000->7000/tcp, 7077/tcp, 0.0.0.0:9091->8080/tcp, :::9091->8080/tcp
a93ac807d60a   cluster-apache-spark:3.0.2             "/bin/bash /start-sp..." 4 days ago     Up 5 hours    7077/tcp, 0.0.0.0:9092->8080/tcp, :::9092->8080/tcp
ef41aa42d125   postgres:11.7-alpine                  "docker-entrypoint.s..." 4 days ago     Up 5 hours    0.0.0.0:5432->5432/tcp, :::5432->5432/tcp
norhanswar@norhanswar-HP-Laptop-15s-fq5xxx: ~/Career/Master/BigData/Labs/Lab2Material-20231203/docker-spark-cluster$
```

## 2.1. Setting up the Hadoop Docker Cluster

To begin with, we ran the Docker Compose file to set up the Hadoop Docker cluster.

## 2.2. Storing the Dataset

Once the Hadoop cluster was up and running, we uploaded and stored the dataset in the Hadoop Distributed File System (HDFS). This step ensured that the dataset was readily available for analysis and processing.

---

### Browse Directory

/data/Ass1

Go!

Show

25

entries

Search:

<input type="checkbox"/>		Permission		Owner		Group		Size		Last Modified		Replication		Block Size		Name	
<input type="checkbox"/>		-rw-r--r--		root		supergroup		1.89 KB		Dec 09 16:38		3		128 MB		<a href="#">Analysis.py</a>	
<input type="checkbox"/>		-rw-r--r--		root		supergroup		3.47 KB		Dec 09 16:38		3		128 MB		<a href="#">predict_satisfaction.py</a>	

Showing 1 to 2 of 2 entries

Previous

1

Next

### Browse Directory

/Ass1

Go!

Show

25

entries

Search:

<input type="checkbox"/>	<div><div></div></div> Permission	<div><div></div></div> Owner	<div><div></div></div> Group	<div><div></div></div> Size	<div><div></div></div> Last Modified	<div><div></div></div> Replication	<div><div></div></div> Block Size	<div><div></div></div> Name	<div><div></div></div>
<input type="checkbox"/>	<a href="#">-rw-r--r--</a>	<a href="#">root</a>	<a href="#">supergroup</a>	21.12 KB	Dec 09 16:39	<a href="#">3</a>	128 MB	<a href="#">E-commerceCustomerBehavior-Sheet1.csv</a>	<div><div></div></div>

Showing 1 to 1 of 1 entries

Previous

1

Next

Hadoop, 2019.

## 2.3. Analysis and Preprocessing with PySpark

Using PySpark, we performed analysis and preprocessing on the dataset. The analysis focused on answering the following key questions:

Segment customers based on their demographic information (Age, Gender, City) and shopping behaviors (Total Spend, Number of Items Purchased, Membership Type)

Membership Type	Age Group	City	sum(Items Purchased)	avg(Total Spend)
Silver	Between 35-50	Los Angeles	288	814.65
Bronze	Between 35-50	Chicago	546	499.8827586206897
Gold	Under 35	New York	901	1165.0355932203393
Silver	Under 35	Los Angeles	401	799.2114285714289
Bronze	Between 35-50	Houston	425	447.6482142857141
Silver	Under 35	Miami	675	690.3896551724141
Gold	Under 35	San Francisco	1141	1460.4561403508774
Gold	Between 35-50	San Francisco	19	1420.8

05/12/09 15:15:15.0 info codegenerator code generated in 11209121 ms

City	sum(Items Purchased)	avg(Total Spend)
Los Angeles	689	805.4915254237288
San Francisco	1160	1459.7724137931039
Chicago	546	499.8827586206897
Houston	425	447.6482142857141
Miami	675	690.3896551724141
New York	901	1165.0355932203393

Which customers are at risk of not making future purchases based on their Days Since Last Purchase and Satisfaction Level

```
23/12/09 13:54:50 INFO CodeGenerator: Code generated in 6.303271 ms
```

Customer ID	Gender	Age	City	Membership Type	Total Spend	Items Purchased	Average Rating	Discount Applied	Days Since Last Purchase	Satisfaction Level
103	Female	43	Chicago	Bronze	510.75	9	3.4	true	42	Unsatisfie
105	Male	27	Miami	Silver	720.4	13	4.0	true	55	Unsatisfie
109	Female	41	Chicago	Bronze	495.25	10	3.6	true	40	Unsatisfie
111	Male	32	Miami	Silver	690.3	11	3.8	true	34	Unsatisfie
115	Female	42	Chicago	Bronze	530.4	9	3.5	true	38	Unsatisfie
117	Male	26	Miami	Silver	700.6	12	3.7	true	48	Unsatisfie
121	Female	43	Chicago	Bronze	505.75	10	3.3	true	41	Unsatisfie
123	Male	27	Miami	Silver	710.4	13	4.1	true	54	Unsatisfie
127	Female	41	Chicago	Bronze	485.25	9	3.6	true	39	Unsatisfie
129	Male	32	Miami	Silver	670.3	10	3.8	true	33	Unsatisfie



Spark Master at spark://a5a702552e2b:7077

URL: spark://a5a702552e2b:7077  
Alive Workers: 2  
Cores in use: 2 Total, 0 Used  
Memory in use: 2.0 GiB Total, 0.0 B Used  
Resources in use:  
Applications: 0 [Running](#), 1 [Completed](#)  
Drivers: 0 Running, 0 Completed  
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
<a href="#">worker-20231209143209-172.19.0.10-7000</a>	172.19.0.10:7000	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	
<a href="#">worker-20231209143209-172.19.0.9-7000</a>	172.19.0.9:7000	ALIVE	1 (0 Used)	1024.0 MiB (0.0 B Used)	

Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

Completed Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
<a href="#">app-20231209144049-0000</a>	PySparkShell	2	1024.0 MiB		2023/12/09 14:40:49	root	FINISHED	10 s