CPSC 531-01

Advanced Database Management

E-commerce Analysis

Report By
Akash Avinash Butala
Abhishek Nagesh Shinde

Table of content

1: Introduction

- Technology is evolving rapidly, and shopping trends shift daily, e-commerce is in a constant state of adaptation that can leave businesses blundering in the dark. To keep up with e-commerce trends, businesses must anticipate changes in the market using reliable data insights. In short, they need effective e-commerce analysis.
- E-commerce analysis simply refers to a strategy designed to analyze large amounts of data to produce meaningful insights. It helps to measure user behavior, performance trends, and ROI.

E-commerce analysis can provide several benefits for businesses. Some potential benefits of conducting e-commerce analysis include the following:

- Improved understanding of customer behavior: E-commerce analysis can help businesses better understand how customers interact with their website and online store. This can include information about what products are most popular, how long customers are spending on the site, and what pages or products are driving the most sales. This information can be used to make informed decisions about how to improve the customer experience and increase sales.

- Increased sales and revenue: By analyzing data from their e-commerce activities, businesses can identify opportunities to increase sales and revenue. This can include identifying high-performing products and promotions and areas for improvement in the customer journey.

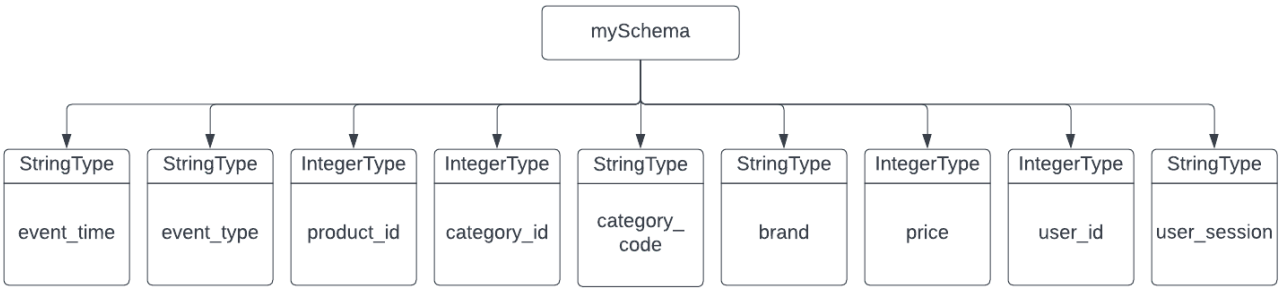
- Improved customer satisfaction: E-commerce analysis can help businesses identify areas where the customer experience can be improved. For example, analysis can reveal areas of the website that are confusing or difficult to navigate, allowing businesses to make changes to improve the customer experience.

- Enhanced market competitiveness: By conducting e-commerce analysis, businesses can gain a better understanding of their competitors and the broader market. This can help them identify opportunities to differentiate their offering and position themselves as a leader in their industry.

- Better decision-making: E-commerce analysis can provide businesses with valuable data and insights that can be used to inform decision-making. This can help businesses make more data-driven decisions rather than relying on guesswork or intuition.

2: Functionalities:

- **Dataset:** Dataset (eCommerce behavior data from multi-category store) is taken from Kaggle.
  Size: 9GB
  Schema:



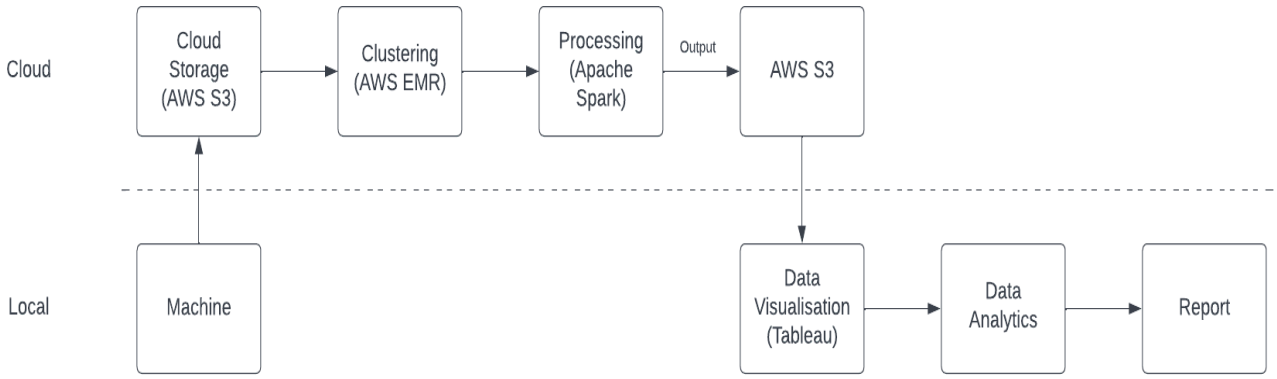
- **Processing Tool:** Apache Spark.
  - In our project, we have PySpark Library for processing our data.
  - PySpark is an interface for Apache Spark in Python.
  - The majority of Spark's functionality, including Spark SQL, DataFrame, Streaming, MLlib (Machine Learning), and Spark Core, are supported by PySpark.

- **Data Storage:** AWS S3 (Simple Storage Service) Bucket
  - AWS S3 is used for storing our large dataset.

- **Clustering:** AWS EMR (Elastic Map Reduce)
  - Amazon EMR is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark , on AWS to process and analyze vast amounts of data.

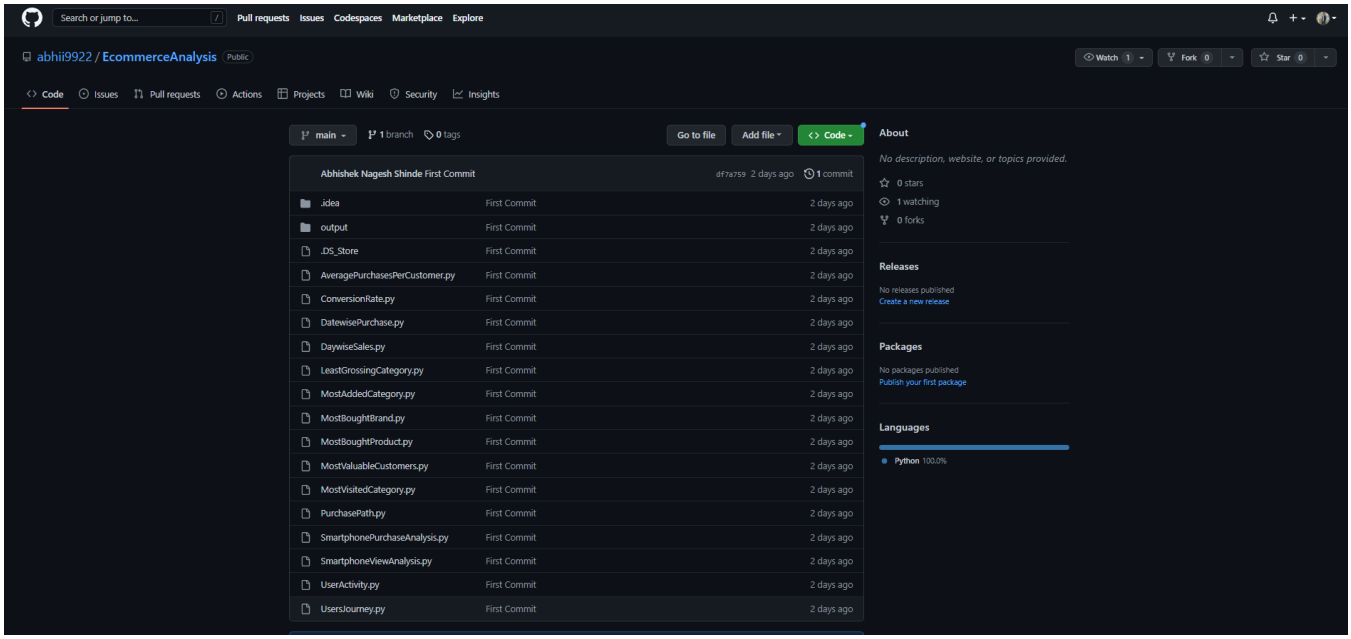
- **Data Analysis and Visualization:** Tableau
  - Tableau is a fantastic business intelligence and data visualization application for reporting and analyzing massive amounts of data.

3: Architecture & Design



4: GitHub Location of Code

https://github.com/abhii9922/EcommerceAnalysis.git



5: Deployment Instructions

Steps to Connect Elastic Map Reduce:

Step 1: **Log in to AWS**

-We have deployed our spark applications on AWS EMR.
-The first step is to login into the AWS services console.
-Create an account and login

Step 2: **Creating Cluster**

-This step includes configuring the cluster.

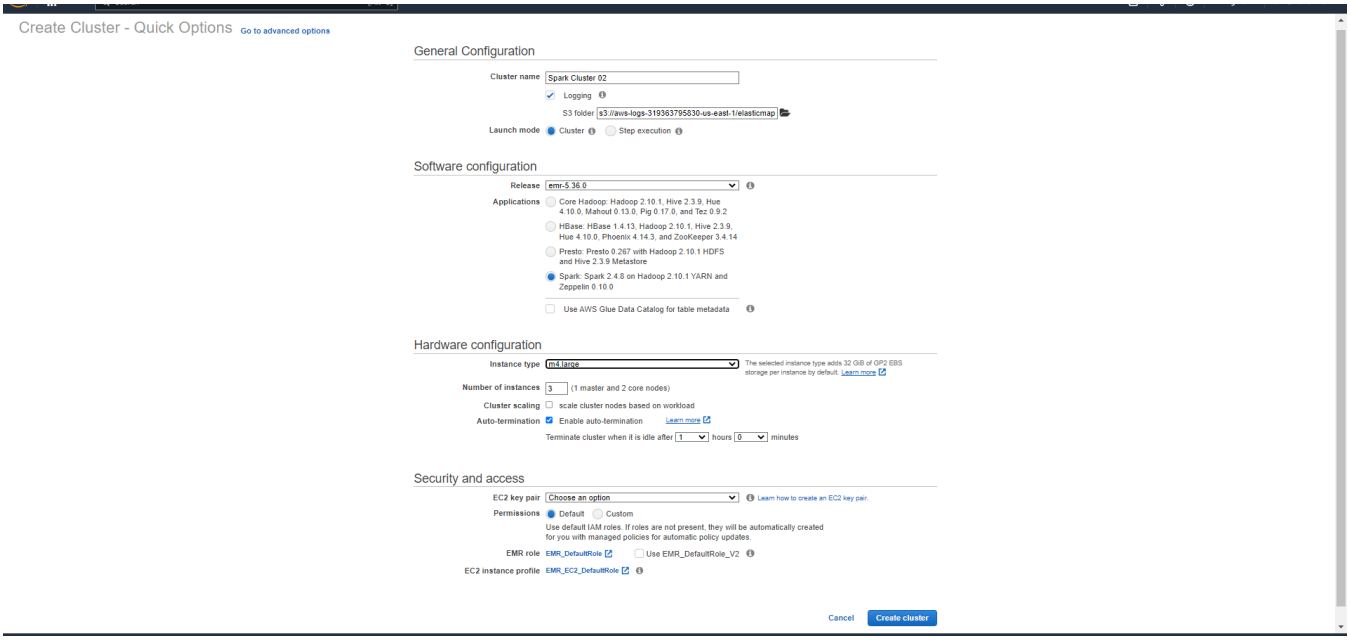
-It has four configurations:

1. **General Configuration**: Naming a cluster.

2. **Software Configuratio**n: Selecting compatible EMR version (we have selected the 5.36.0 version) and all the other software (Spark, Hadoop, Yarn, etc.) that needed to be auto-configured during cluster creation.

3. **Hardware Configuratio**n: Selecting instance type and several instances depending on the requirement. (We have selected an m4.large instance with 2 cores and 1 master instance)

4. **Security and Access**: Need to generate and use key (PPK) for connecting to cluster. The generated key needs to be downloaded and saved for later use. (Authentication)



-After Clicking on create a cluster, the cluster gets started.

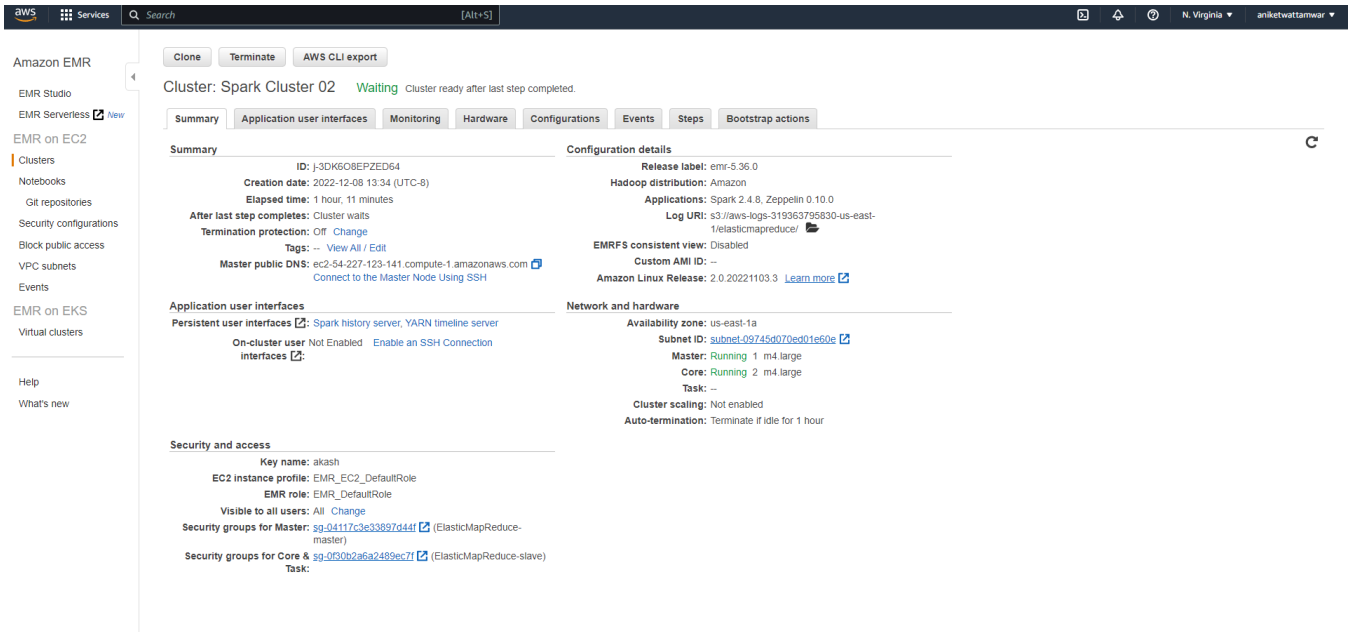
- There are 3 states:

1. **Waiting**: In this state, we have to wait till the whole configuration is completed.

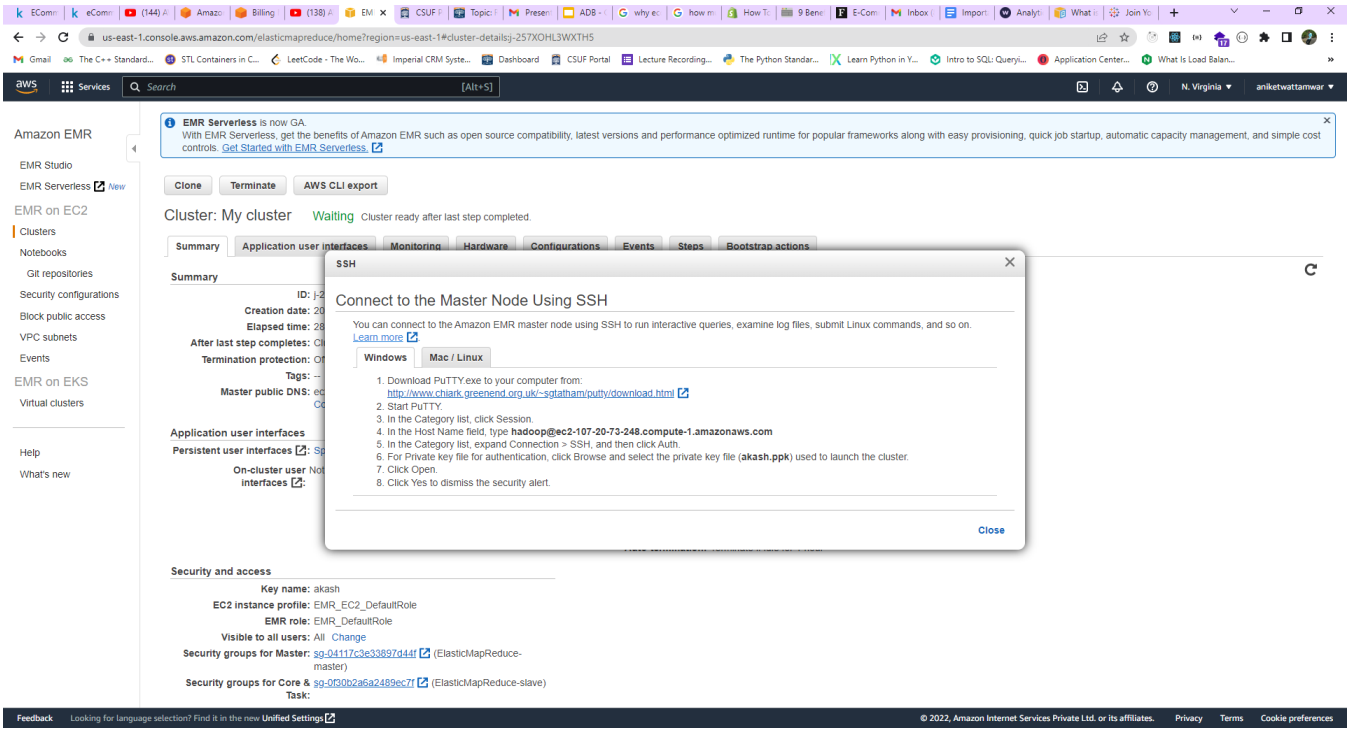
2. **Running**: In this state, we can connect and use the cluster.

3. **Terminating**: The cluster goes into this state when we terminate the cluster.

- After this, we need to update the security group for the master node. Modify the inbound rule and add SSH type to support ipv4 and ipv6 addresses.



Step 3: Connecting to Master Node Using SSH

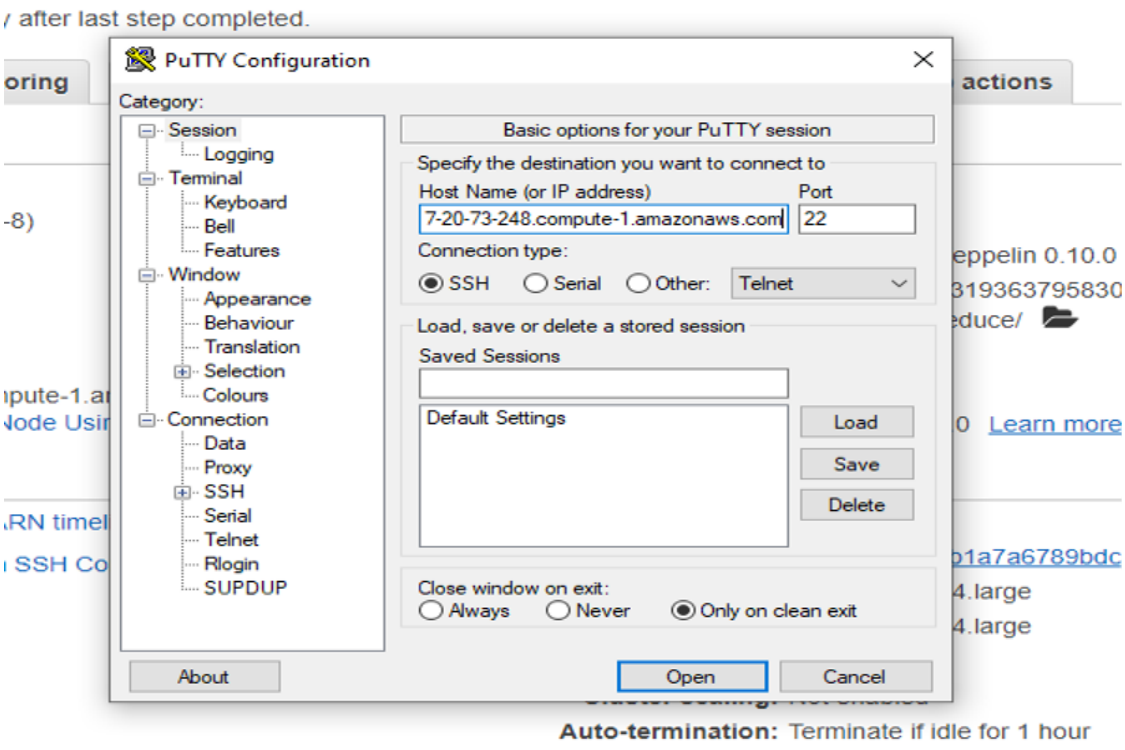


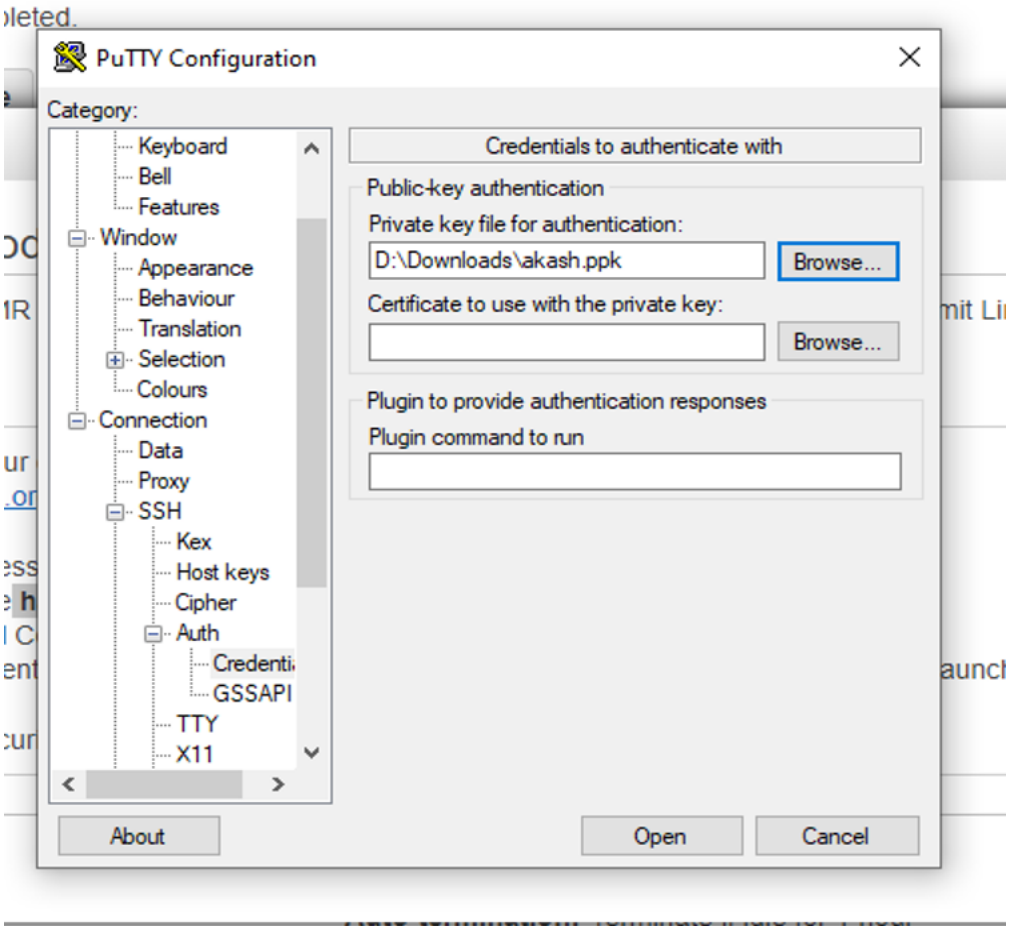
-For connecting the master node, we need to download putty.

-In the hostname field of putty, add Master public DNS.

- Then go to the category section and click on SSH under connection, followed by AUTH and add the key we generated during configuration.

- Finally, click on open, and we will get connected and can use EMR now.


after last step completed.
oring
actions
-8)
eppelin 0.10.0
319363795830
educe/
pute-1.a
Node Usi
0 Learn more
RN timel
SSH Co
b1a7a6789bdc
4.large
4.large
Auto-termination: Terminate if idle for 1 hour
PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
SSH
Serial
Telnet
Rlogin
SUPDUP
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
7-20-73-248.compute-1.amazonaws.com
22
Connection type:
SSH
Serial
Other:
Telnet
Load, save or delete a stored session
Saved Sessions
Default Settings
Load
Save
Delete
Close window on exit:
Always
Never
Only on clean exit
About
Open
Cancel

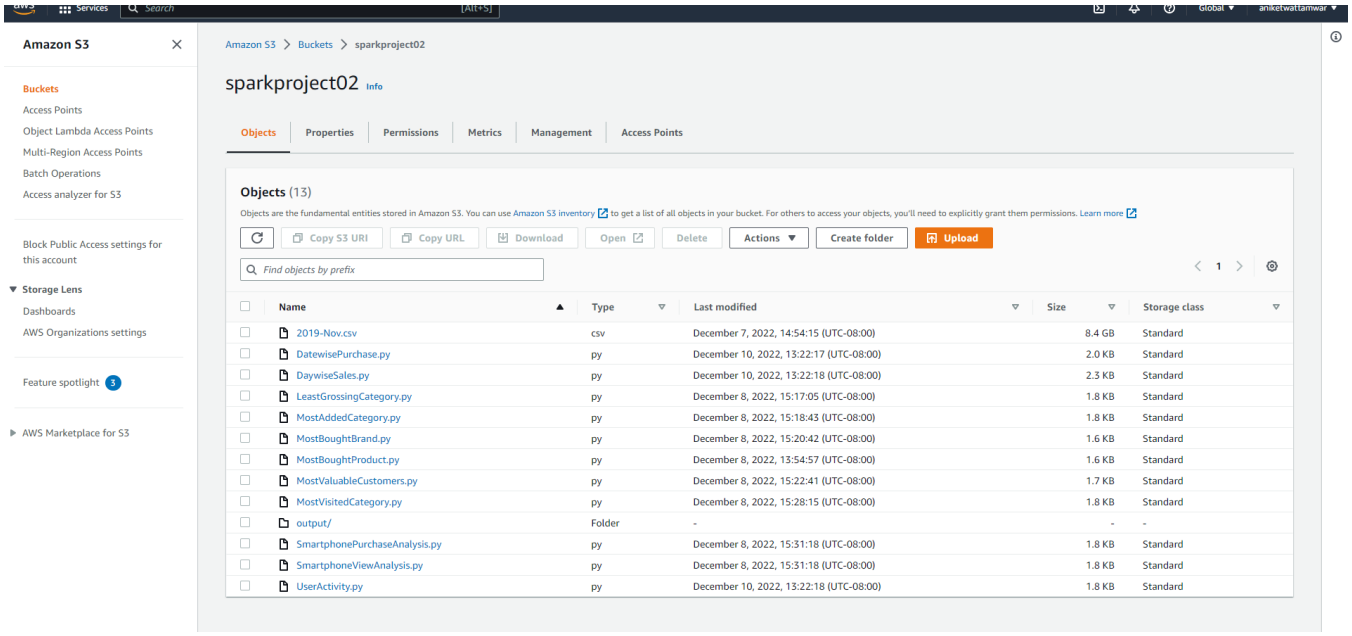

PuTTY Configuration
Category:
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Connection
Data
Proxy
SSH
Kex
Host keys
Cipher
Auth
Credenti
GSSAPI
TTY
X11
Credentials to authenticate with
Public-key authentication
Private key file for authentication:
D:\Downloads\akash.ppk
Browse...
Certificate to use with the private key:
Browse...
Plugin to provide authentication responses
Plugin command to run
About
Open
Cancel

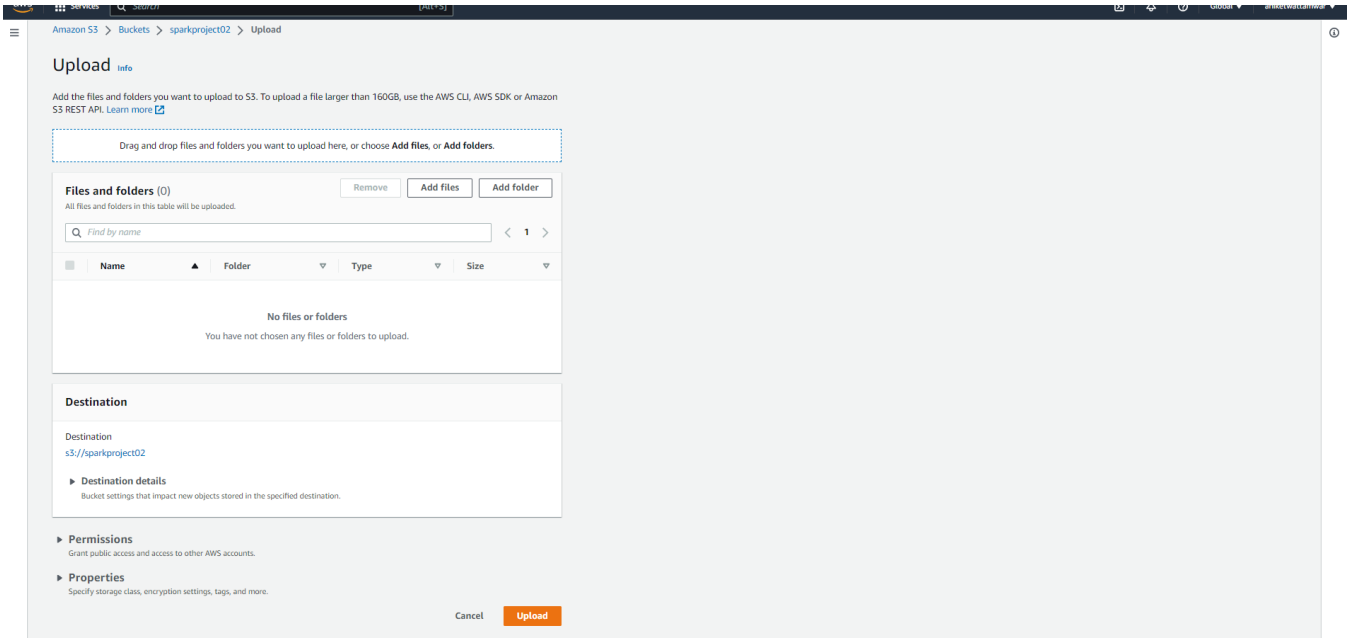

Using username "hadoop".
Authenticating with public key "akash"
Last login: Tue Nov 8 19:47:09 2022
Amazon Linux 2 AMI
https://aws.amazon.com/amazon-linux-2/
22 package(s) needed for security, out of 32 available
Run "sudo yum update" to apply all updates.
[hadoop@ip-172-31-37-116 ~]$

Step 4: AWS S3 bucket:

- We have used an S3 bucket for storing our data and outputs.

-We just have to upload the file we want in the S3 path.





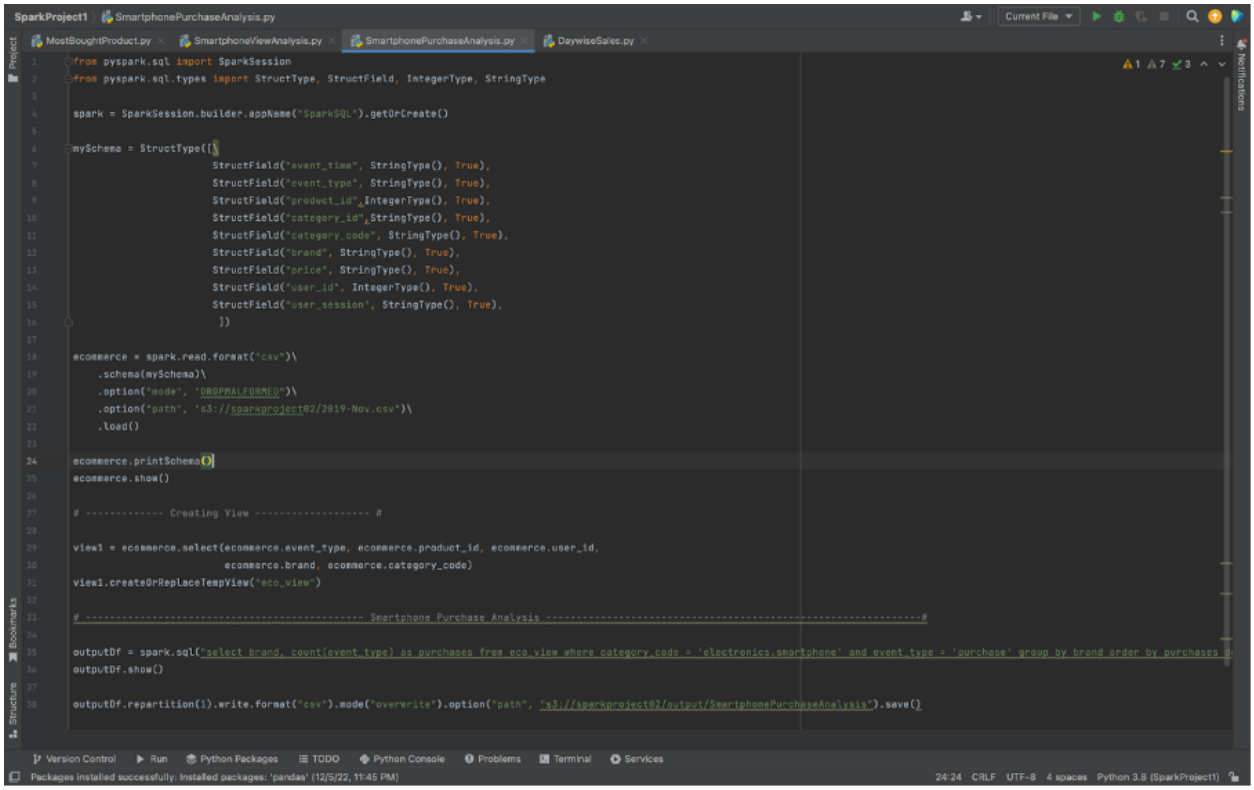
6: Steps to Run the Application

-We have created the Pyspark Application.

- Application includes

    1. Schema Definition
    2. Loading Data
    3. Performing Query
    4. Saving Output
    5. Using Output for Analysis.

- PyCharm IDE is used for locally running the application.



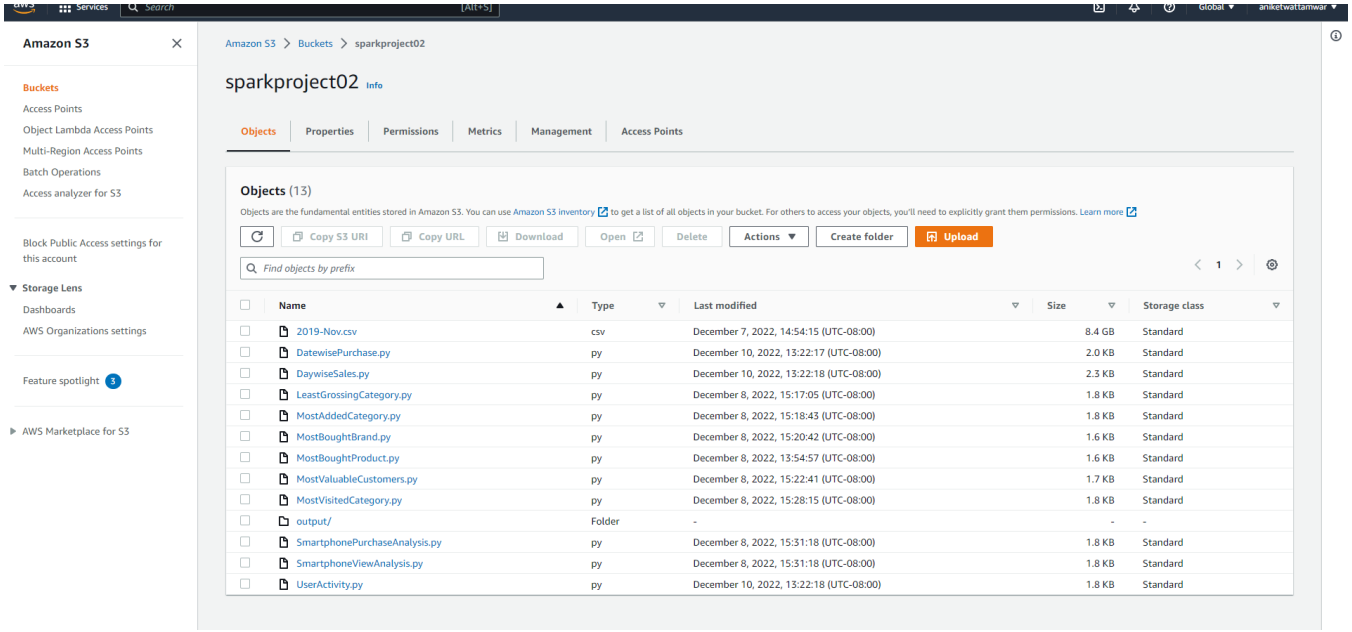
-After testing locally, we moved our application to AWS.

Services used:

1. AWS S3

- We uploaded our dataset (2019-Nov.csv 9 GB) on the S3 bucket.

- Also uploaded all the .py files to the S3 bucket.



2. AWS EMR:

- After completing all the configuration, we connected to EMR through putty.

- For running our .py files, we need to copy the files from s3 bucket to our EMR console.

A command for copying:

$ aws s3 cp s3://sparkproject02/LeastGrossingCategory.py .

-We can check if the .py file is copied successfully using the command:

$ ls

-After copying the .py file, we run this file using the command:

$ spark-submit “file.

-After successfully completing the task, the outputs are saved on the S3 bucket. Then, we downloaded the outputs for analysis and visualization.

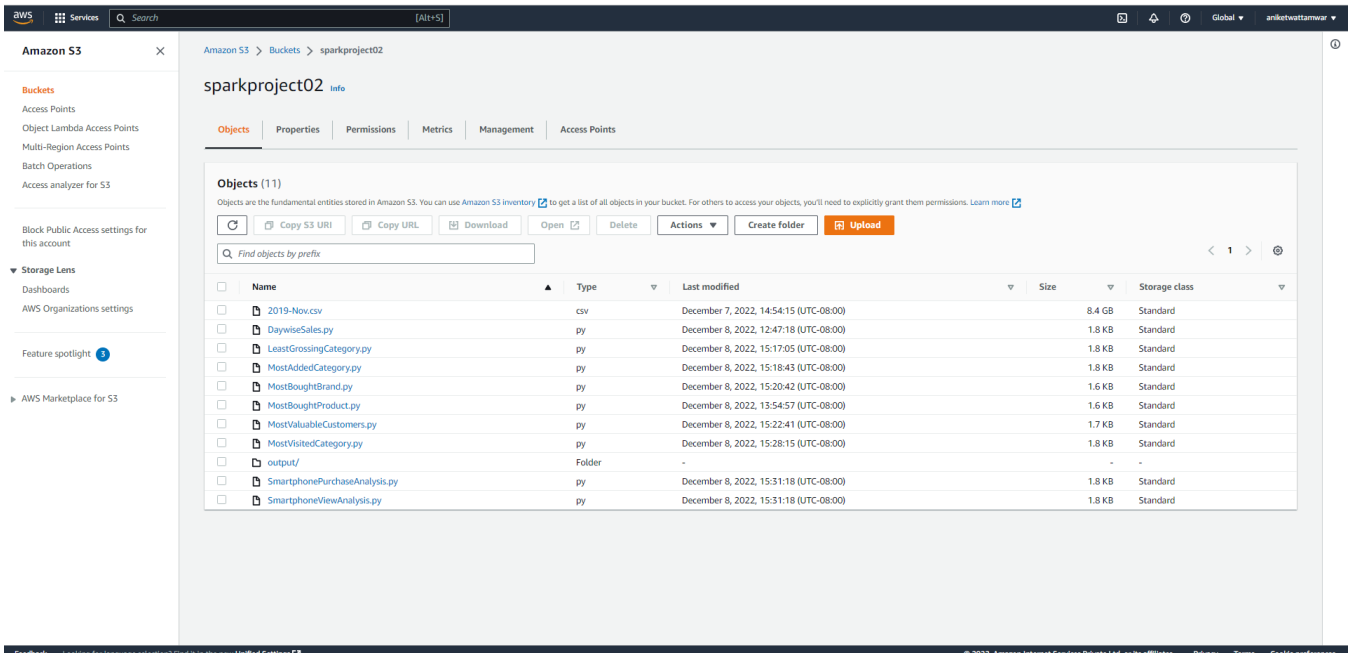
```
[hadoop@ip-172-31-84-178 ~]$ aws s3 cp s3://sparkproject02/MostBoughtProduct.py .
download: s3://sparkproject02/MostBoughtProduct.py to ./MostBoughtProduct.py
[hadoop@ip-172-31-84-178 ~]$ ls
DaywiseSales.py  MostBoughtProduct.py
[hadoop@ip-172-31-84-178 ~]$ clear
[hadoop@ip-172-31-84-178 ~]$ clear
[hadoop@ip-172-31-84-178 ~]$ spark-submit MostBoughtProduct.py
22/12/08 21:56:30 INFO SparkContext: Running Spark version 2.4.8-amzn-2
22/12/08 21:56:30 INFO SparkContext: Submitted application: SparkSQL
22/12/08 21:56:30 INFO SecurityManager: Changing view acls to: hadoop
22/12/08 21:56:30 INFO SecurityManager: Changing modify acls to: hadoop
22/12/08 21:56:30 INFO SecurityManager: Changing view acls groups to:
22/12/08 21:56:30 INFO SecurityManager: Changing modify acls groups to:
22/12/08 21:56:30 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with view permissions: Set(); users  with modify permissions: Set(hadoop); groups with
modify permissions: Set()
22/12/08 21:56:30 INFO Utils: Successfully started service 'sparkDriver' on port 44785.
22/12/08 21:56:30 INFO SparkEnv: Registering MapOutputTracker
22/12/08 21:56:30 INFO SparkEnv: Registering BlockManagerMaster
22/12/08 21:56:30 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
22/12/08 21:56:30 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
22/12/08 21:56:30 INFO DiskBlockManager: Created local directory at /mnt/tmp/blockmgr-3abee17a-ec4f-4808-bae3-2bce5dd2e615
22/12/08 21:56:30 INFO MemoryStore: MemoryStore started with capacity 912.3 MB
22/12/08 21:56:31 INFO SparkEnv: Registering OutputCommitCoordinator
22/12/08 21:56:31 INFO Utils: Successfully started service 'SparkUI' on port 4040.
22/12/08 21:56:31 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://ip-172-31-84-178.ec2.internal:4040
22/12/08 21:56:31 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
22/12/08 21:56:33 INFO RMProxy: Connecting to ResourceManager at ip-172-31-84-178.ec2.internal/172.31.84.178:8032
22/12/08 21:56:33 INFO Client: Requesting a new application from cluster with 2 NodeManagers
22/12/08 21:56:33 INFO Configuration: resource-types.xml not found
22/12/08 21:56:33 INFO ResourceUtils: Unable to find 'resource-types.xml'.
22/12/08 21:56:33 INFO ResourceUtils: Adding resource type - name = memory-mb, units = Mi, type = COUNTABLE
22/12/08 21:56:33 INFO ResourceUtils: Adding resource type - name = vcores, units = , type = COUNTABLE
22/12/08 21:56:33 INFO Client: Verifying our application has not requested more than the maximum memory capability of the cluster (6144 MB per container)
22/12/08 21:56:33 INFO Client: Will allocate AM container, with 896 MB memory including 384 MB overhead
22/12/08 21:56:33 INFO Client: Setting up container launch context for our AM
22/12/08 21:56:33 INFO Client: Setting up the launch environment for our AM container
22/12/08 21:56:33 INFO Client: Preparing resources for our AM container
22/12/08 21:56:33 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to uploading libraries under SPARK_HOME.
22/12/08 21:56:36 INFO Client: Uploading resource file:/mnt/tmp/spark-48ea111a-7810-437a-a4af-1109535d6c2a/__spark_libs__1200444584479479O6.zip -> hdfs://ip-172-31-84-178.ec2.internal:8020/user/hadoop/.sparkStaging/application_1670535573
928_0002/__spark_libs__1200444584479479O6.zip
22/12/08 21:56:36 INFO Client: Uploading resource file:/etc/hudi/conf/hudi-defaults.conf -> hdfs://ip-172-31-84-178.ec2.internal:8020/user/hadoop/.sparkStaging/application_1670535573928_0002/hudi-defaults.conf
22/12/08 21:56:39 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/pyspark.zip -> hdfs://ip-172-31-84-178.ec2.internal:8020/user/hadoop/.sparkStaging/application_1670535573928_0002/pyspark.zip
22/12/08 21:56:39 INFO Client: Uploading resource file:/usr/lib/spark/python/lib/py4j-0.10.7-src.zip -> hdfs://ip-172-31-84-178.ec2.internal:8020/user/hadoop/.sparkStaging/application_1670535573928_0002/py4j-0.10.7-src.zip
22/12/08 21:56:39 INFO Client: Uploading resource file:/mnt/tmp/spark-48ea111a-7810-437a-a4af-1109535d6c2a/__spark_conf__1771894712035028605.zip -> hdfs://ip-172-31-84-178.ec2.internal:8020/user/hadoop/.sparkStaging/application_167053557
3928_0002/__spark_conf__.zip
22/12/08 21:56:40 INFO SecurityManager: Changing view acls to: hadoop
22/12/08 21:56:40 INFO SecurityManager: Changing modify acls to: hadoop
22/12/08 21:56:40 INFO SecurityManager: Changing view acls groups to:
22/12/08 21:56:40 INFO SecurityManager: Changing modify acls groups to:
22/12/08 21:56:40 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(hadoop); groups with view permissions: Set(); users  with modify permissions: Set(hadoop); groups with
modify permissions: Set()
22/12/08 21:56:42 INFO Client: Submitting application application_1670535573928_0002 to ResourceManager
22/12/08 21:56:42 INFO YarnClientImpl: Submitted application application_1670535573928_0002
22/12/08 21:56:42 INFO SchedulerExtensionServices: Starting Yarn extension services with app application_1670535573928_0002 and attemptId None
22/12/08 21:56:43 INFO Client: Application report for application_1670535573928_0002 (state: ACCEPTED)
22/12/08 21:56:43 INFO Client:
        client token: N/A
        diagnostics: AM container is launched, waiting for AM container to Register with RM
        ApplicationMaster host: N/A
        ApplicationMaster RPC port: -1
        queue: default
        start time: 1670536602307
        final status: UNDEFINED
```

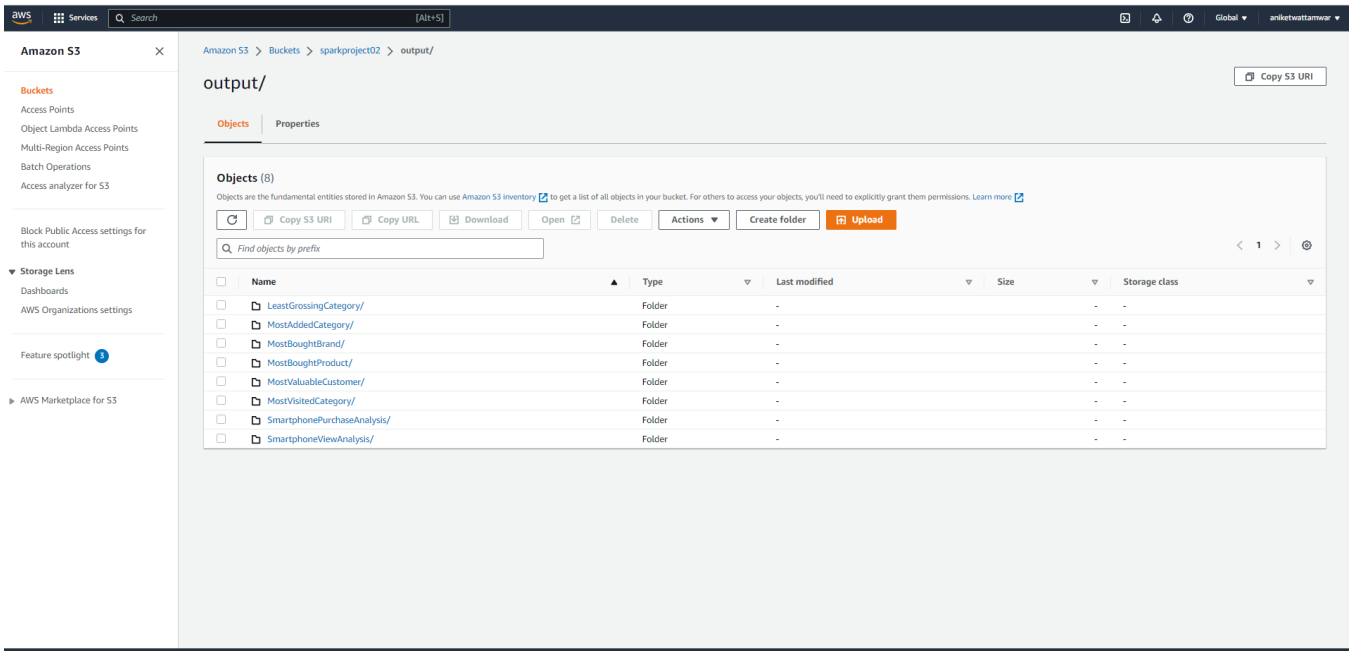
```
22/12/08 21:56:49 INFO SingleEventLogFileWriter: Logging events to hdfs:/var/log/spark/apps/application_1670535573928_0002.inprogress
22/12/08 21:56:49 INFO Utils: Using 50 preallocated executors (minExecutors: 0). Set spark.dynamicAllocation.preallocateExecutors to `false` disable executor preallocation.
22/12/08 21:56:49 INFO YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0
22/12/08 21:56:50 INFO SharedState: loading hive config file: file:/etc/spark/conf.dist/hive-site.xml
22/12/08 21:56:50 INFO SharedState: Setting hive.metastore.warehouse.dir ('null') to the value of spark.sql.warehouse.dir ('hdfs:///user/spark/warehouse').
22/12/08 21:56:50 INFO SharedState: Warehouse path is 'hdfs:///user/spark/warehouse'.
22/12/08 21:56:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL.
22/12/08 21:56:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/json.
22/12/08 21:56:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution.
22/12/08 21:56:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /SQL/execution/json.
22/12/08 21:56:50 INFO JettyUtils: Adding filter org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter to /static/sql.
22/12/08 21:56:50 INFO StateStoreCoordinatorRef: Registered StateStoreCoordinator endpoint
22/12/08 21:56:52 INFO ClientConfigurationFactory: Set initial getObject socket timeout to 2000 ms.
22/12/08 21:56:54 INFO InMemoryFileIndex: It took 66 ms to list leaf files for 1 paths.
22/12/08 21:56:54 INFO YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (172.31.94.230:35828) with ID 1
22/12/08 21:56:55 INFO ExecutorAllocationManager: New executor 1 has registered (new total is 1)
root
 |-- event_time: string (nullable = true)
 |-- event_type: string (nullable = true)
 |-- product_id: integer (nullable = true)
 |-- category_id: integer (nullable = true)
 |-- category_code: string (nullable = true)
 |-- brand: string (nullable = true)
 |-- price: integer (nullable = true)
 |-- user_id: integer (nullable = true)
 |-- user_session: string (nullable = true)

22/12/08 21:56:55 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-94-230.ec2.internal:39123 with 2.1 GB RAM, BlockManagerId(1, ip-172-31-94-230.ec2.internal, 39123, None)
22/12/08 21:56:55 INFO YarnSchedulerBackend$YarnDriverEndpoint: Registered executor NettyRpcEndpointRef(spark-client://Executor) (172.31.81.16:36516) with ID 2
22/12/08 21:56:55 INFO ExecutorAllocationManager: New executor 2 has registered (new total is 2)
22/12/08 21:56:55 INFO FileSourceStrategy: Pruning directories with:
22/12/08 21:56:55 INFO FileSourceStrategy: Post-Scan Filters:
22/12/08 21:56:55 INFO FileSourceStrategy: Output Data Schema: struct<event_time: string, event_type: string, product_id: int, category_id: int, category_code: string ... 7 more fields>
22/12/08 21:56:55 INFO FileSourceScanExec: Pushed Filters:
22/12/08 21:56:56 INFO BlockManagerMasterEndpoint: Registering block manager ip-172-31-81-16.ec2.internal:37721 with 2.1 GB RAM, BlockManagerId(2, ip-172-31-81-16.ec2.internal, 37721, None)
22/12/08 21:56:57 INFO CodeGenerator: Code generated in 812.845795 ms
22/12/08 21:56:57 INFO CodeGenerator: Code generated in 136.41667 ms
22/12/08 21:56:57 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 361.3 KB, free 911.9 MB)
22/12/08 21:56:57 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 32.5 KB, free 911.9 MB)
22/12/08 21:56:57 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on ip-172-31-84-178.ec2.internal:40833 (size: 32.5 KB, free: 912.3 MB)
22/12/08 21:56:57 INFO SparkContext: Created broadcast 0 from showString at NativeMethodAccessorImpl.java:0
22/12/08 21:56:57 INFO GPLNativeCodeLoader: Loaded native gpl library
22/12/08 21:56:57 INFO LzoCodec: Successfully loaded & initialized native-lzo library [hadoop-lzo rev 049362b7cf53ff5f739d6b1532457f2c6cd495e8]
22/12/08 21:56:57 INFO FileSourceScanExec: Planning scan with bin packing, max size: 134217728 bytes, open cost is considered as scanning 4194304 bytes, number of split files: 68, prefetch: false
22/12/08 21:56:57 INFO FileSourceScanExec: relation: None, fileSplitsInPartitionHistogram: ArrayBuffer((1 fileSplits,68))
22/12/08 21:56:58 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
22/12/08 21:56:58 INFO DAGScheduler: Got job 0 (showString at NativeMethodAccessorImpl.java:0) with 1 output partitions
22/12/08 21:56:58 INFO DAGScheduler: Final stage: ResultStage 0 (showString at NativeMethodAccessorImpl.java:0)
22/12/08 21:56:58 INFO DAGScheduler: Parents of final stage: List()
22/12/08 21:56:58 INFO DAGScheduler: Missing parents: List()
22/12/08 21:56:58 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[3] at showString at NativeMethodAccessorImpl.java:0), which has no missing parents
22/12/08 21:56:58 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 14.3 KB, free 911.9 MB)
22/12/08 21:56:58 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 6.9 KB, free 911.9 MB)
22/12/08 21:56:58 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on ip-172-31-84-178.ec2.internal:40833 (size: 6.9 KB, free: 912.3 MB)
22/12/08 21:56:58 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1297
22/12/08 21:56:58 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[3] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))
22/12/08 21:56:58 INFO YarnScheduler: Adding task set 0.0 with 1 tasks
22/12/08 21:56:58 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, ip-172-31-81-16.ec2.internal, executor 2, partition 0, RACK_LOCAL, 8547 bytes)
22/12/08 21:56:58 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on ip-172-31-81-16.ec2.internal:37721 (size: 6.9 KB, free: 2.1 GB)
22/12/08 21:57:00 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on ip-172-31-81-16.ec2.internal:37721 (size: 32.5 KB, free: 2.1 GB)
22/12/08 21:57:16 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 17800 ms on ip-172-31-81-16.ec2.internal (executor 2) (1/1)
22/12/08 21:57:16 INFO DAGScheduler: ResultStage 0 (showString at NativeMethodAccessorImpl.java:0) finished in 17.984 s
22/12/08 21:57:16 INFO YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool
```

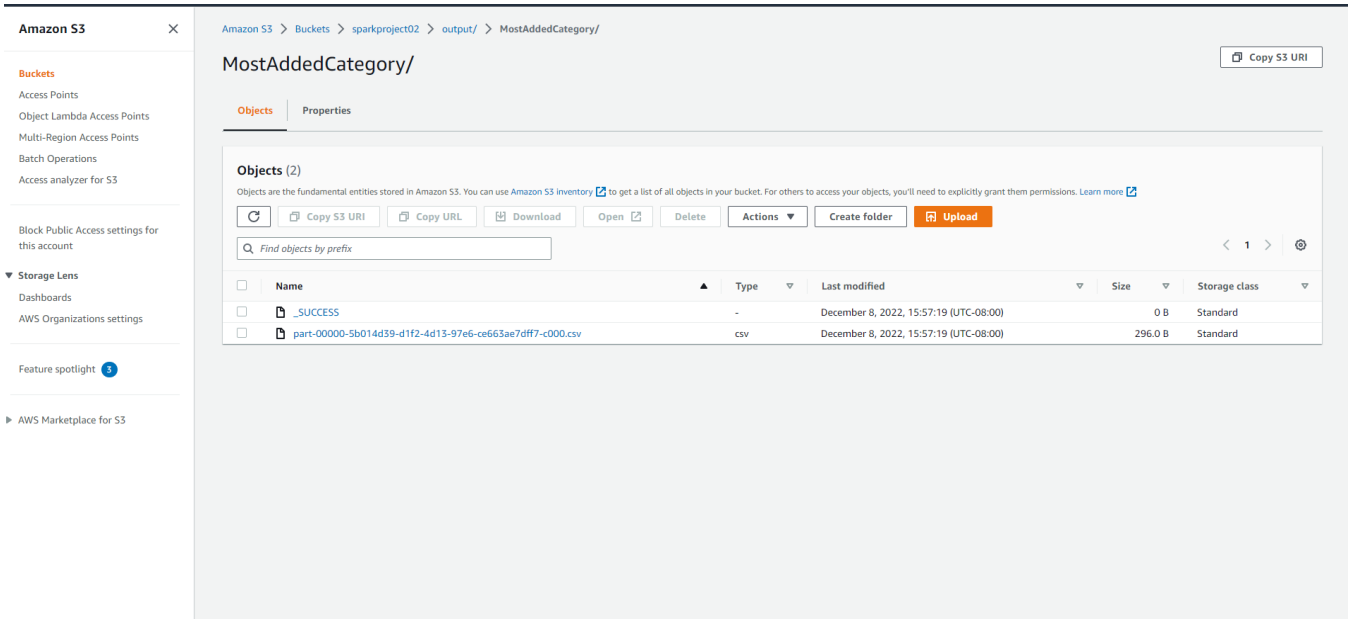
```
22/12/08 21:57:16 INFO YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool
22/12/08 21:57:16 INFO DAGScheduler: Job 0 finished: showString at NativeMethodAccessorImpl.java:0, took 18.167508 s
22/12/08 21:57:16 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
22/12/08 21:57:16 INFO DAGScheduler: Got job 1 (showString at NativeMethodAccessorImpl.java:0) with 4 output partitions
22/12/08 21:57:16 INFO DAGScheduler: Final stage: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0)
22/12/08 21:57:16 INFO DAGScheduler: Parents of final stage: List()
22/12/08 21:57:16 INFO DAGScheduler: Missing parents: List()
22/12/08 21:57:16 INFO DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[3] at showString at NativeMethodAccessorImpl.java:0), which has no missing parents
22/12/08 21:57:16 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 14.3 KB, free 911.9 MB)
22/12/08 21:57:16 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 6.9 KB, free 911.9 MB)
22/12/08 21:57:16 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on ip-172-31-84-178.ec2.internal:40833 (size: 6.9 KB, free: 912.3 MB)
22/12/08 21:57:16 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1297
22/12/08 21:57:16 INFO DAGScheduler: Submitting 4 missing tasks from ResultStage 1 (MapPartitionsRDD[3] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(1, 2, 3, 4))
22/12/08 21:57:16 INFO YarnScheduler: Adding task set 1.0 with 4 tasks
22/12/08 21:57:16 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, ip-172-31-81-16.ec2.internal, executor 2, partition 1, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:16 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 2, ip-172-31-94-230.ec2.internal, executor 1, partition 2, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:16 INFO TaskSetManager: Starting task 2.0 in stage 1.0 (TID 3, ip-172-31-81-16.ec2.internal, executor 2, partition 3, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:16 INFO TaskSetManager: Starting task 3.0 in stage 1.0 (TID 4, ip-172-31-94-230.ec2.internal, executor 1, partition 4, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:16 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on ip-172-31-81-16.ec2.internal:37721 (size: 6.9 KB, free: 2.1 GB)
22/12/08 21:57:16 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on ip-172-31-94-230.ec2.internal:39123 (size: 6.9 KB, free: 2.1 GB)
22/12/08 21:57:17 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on ip-172-31-94-230.ec2.internal:39123 (size: 32.5 KB, free: 2.1 GB)
22/12/08 21:57:35 INFO TaskSetManager: Finished task 2.0 in stage 1.0 (TID 3) in 19682 ms on ip-172-31-81-16.ec2.internal (executor 2) (1/4)
22/12/08 21:57:36 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 19799 ms on ip-172-31-81-16.ec2.internal (executor 2) (2/4)
22/12/08 21:57:42 INFO TaskSetManager: Finished task 3.0 in stage 1.0 (TID 4) in 25800 ms on ip-172-31-94-230.ec2.internal (executor 1) (3/4)
22/12/08 21:57:42 INFO TaskSetManager: Finished task 1.0 in stage 1.0 (TID 2) in 26034 ms on ip-172-31-94-230.ec2.internal (executor 1) (4/4)
22/12/08 21:57:42 INFO YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool
22/12/08 21:57:42 INFO DAGScheduler: ResultStage 1 (showString at NativeMethodAccessorImpl.java:0) finished in 26.063 s
22/12/08 21:57:42 INFO DAGScheduler: Job 1 finished: showString at NativeMethodAccessorImpl.java:0, took 26.088888 s
22/12/08 21:57:42 INFO SparkContext: Starting job: showString at NativeMethodAccessorImpl.java:0
22/12/08 21:57:42 INFO DAGScheduler: Got job 2 (showString at NativeMethodAccessorImpl.java:0) with 20 output partitions
22/12/08 21:57:42 INFO DAGScheduler: Final stage: ResultStage 2 (showString at NativeMethodAccessorImpl.java:0)
22/12/08 21:57:42 INFO DAGScheduler: Parents of final stage: List()
22/12/08 21:57:42 INFO DAGScheduler: Missing parents: List()
22/12/08 21:57:42 INFO DAGScheduler: Submitting ResultStage 2 (MapPartitionsRDD[3] at showString at NativeMethodAccessorImpl.java:0), which has no missing parents
22/12/08 21:57:42 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 14.3 KB, free 911.9 MB)
22/12/08 21:57:42 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 6.9 KB, free 911.9 MB)
22/12/08 21:57:42 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on ip-172-31-84-178.ec2.internal:40833 (size: 6.9 KB, free: 912.2 MB)
22/12/08 21:57:42 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:1297
22/12/08 21:57:42 INFO DAGScheduler: Submitting 20 missing tasks from ResultStage 2 (MapPartitionsRDD[3] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19))
22/12/08 21:57:42 INFO YarnScheduler: Adding task set 2.0 with 20 tasks
22/12/08 21:57:42 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 5, ip-172-31-81-16.ec2.internal, executor 2, partition 5, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO TaskSetManager: Starting task 1.0 in stage 2.0 (TID 6, ip-172-31-94-230.ec2.internal, executor 1, partition 6, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO TaskSetManager: Starting task 2.0 in stage 2.0 (TID 7, ip-172-31-81-16.ec2.internal, executor 2, partition 7, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO TaskSetManager: Starting task 3.0 in stage 2.0 (TID 8, ip-172-31-94-230.ec2.internal, executor 1, partition 8, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO TaskSetManager: Starting task 4.0 in stage 2.0 (TID 9, ip-172-31-81-16.ec2.internal, executor 2, partition 9, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO TaskSetManager: Starting task 5.0 in stage 2.0 (TID 10, ip-172-31-94-230.ec2.internal, executor 1, partition 10, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO TaskSetManager: Starting task 6.0 in stage 2.0 (TID 11, ip-172-31-81-16.ec2.internal, executor 2, partition 11, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO TaskSetManager: Starting task 7.0 in stage 2.0 (TID 12, ip-172-31-94-230.ec2.internal, executor 1, partition 12, RACK_LOCAL, 8547 bytes)
22/12/08 21:57:42 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on ip-172-31-81-16.ec2.internal:37721 (size: 6.9 KB, free: 2.1 GB)
22/12/08 21:57:42 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on ip-172-31-94-230.ec2.internal:39123 (size: 6.9 KB, free: 2.1 GB)
22/12/08 21:57:43 INFO ExecutorAllocationManager: Requesting 3 new executors because tasks are backlogged (new desired total will be 3)
22/12/08 21:57:44 INFO ExecutorAllocationManager: Requesting 1 new executor because tasks are backlogged (new desired total will be 4)
22/12/08 21:57:45 INFO ExecutorAllocationManager: Requesting 1 new executor because tasks are backlogged (new desired total will be 5)
22/12/08 21:58:19 INFO TaskSetManager: Starting task 8.0 in stage 2.0 (TID 13, ip-172-31-81-16.ec2.internal, executor 2, partition 13, RACK_LOCAL, 8547 bytes)
22/12/08 21:58:19 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 5) in 36838 ms on ip-172-31-81-16.ec2.internal (executor 2) (1/20)
22/12/08 21:58:19 INFO TaskSetManager: Starting task 9.0 in stage 2.0 (TID 14, ip-172-31-94-230.ec2.internal, executor 1, partition 14, RACK_LOCAL, 8547 bytes)
22/12/08 21:58:19 INFO TaskSetManager: Finished task 3.0 in stage 2.0 (TID 8) in 37586 ms on ip-172-31-94-230.ec2.internal (executor 1) (2/20)
22/12/08 21:58:20 INFO TaskSetManager: Starting task 10.0 in stage 2.0 (TID 15, ip-172-31-81-16.ec2.internal, executor 2, partition 15, RACK_LOCAL, 8547 bytes)
22/12/08 21:58:20 INFO TaskSetManager: Finished task 4.0 in stage 2.0 (TID 9) in 37889 ms on ip-172-31-81-16.ec2.internal (executor 2) (3/20)
22/12/08 21:58:20 INFO TaskSetManager: Starting task 11.0 in stage 2.0 (TID 16, ip-172-31-81-16.ec2.internal, executor 2, partition 16, RACK_LOCAL, 8547 bytes)
22/12/08 21:58:20 INFO TaskSetManager: Finished task 6.0 in stage 2.0 (TID 11) in 38093 ms on ip-172-31-81-16.ec2.internal (executor 2) (4/20)
22/12/08 21:58:20 INFO TaskSetManager: Starting task 12.0 in stage 2.0 (TID 17, ip-172-31-94-230.ec2.internal, executor 1, partition 17, RACK_LOCAL, 8547 bytes)
```

```
22/12/08 22:04:23 INFO TaskSetManager: Finished task 17.0 in stage 6.0 (TID 153) in 249 ms on ip-172-31-94-230.ec2.internal (executor 1) (17/26)
22/12/08 22:04:23 INFO TaskSetManager: Starting task 25.0 in stage 6.0 (TID 161, ip-172-31-81-16.ec2.internal, executor 2, partition 25, PROCESS_LOCAL, 7767 bytes)
22/12/08 22:04:23 INFO TaskSetManager: Finished task 15.0 in stage 6.0 (TID 151) in 346 ms on ip-172-31-81-16.ec2.internal (executor 2) (18/26)
22/12/08 22:04:23 INFO TaskSetManager: Finished task 18.0 in stage 6.0 (TID 154) in 221 ms on ip-172-31-94-230.ec2.internal (executor 1) (19/26)
22/12/08 22:04:23 INFO TaskSetManager: Finished task 20.0 in stage 6.0 (TID 156) in 218 ms on ip-172-31-81-16.ec2.internal (executor 2) (20/26)
22/12/08 22:04:23 INFO TaskSetManager: Finished task 19.0 in stage 6.0 (TID 155) in 250 ms on ip-172-31-81-16.ec2.internal (executor 2) (21/26)
22/12/08 22:04:24 INFO TaskSetManager: Finished task 21.0 in stage 6.0 (TID 157) in 254 ms on ip-172-31-94-230.ec2.internal (executor 1) (22/26)
22/12/08 22:04:24 INFO TaskSetManager: Finished task 24.0 in stage 6.0 (TID 160) in 150 ms on ip-172-31-94-230.ec2.internal (executor 1) (23/26)
22/12/08 22:04:24 INFO TaskSetManager: Finished task 23.0 in stage 6.0 (TID 159) in 201 ms on ip-172-31-94-230.ec2.internal (executor 1) (24/26)
22/12/08 22:04:24 INFO TaskSetManager: Finished task 22.0 in stage 6.0 (TID 158) in 215 ms on ip-172-31-81-16.ec2.internal (executor 2) (25/26)
22/12/08 22:04:24 INFO TaskSetManager: Finished task 25.0 in stage 6.0 (TID 161) in 137 ms on ip-172-31-81-16.ec2.internal (executor 2) (26/26)
22/12/08 22:04:24 INFO YarnScheduler: Removed TaskSet 6.0, whose tasks have all completed, from pool
22/12/08 22:04:24 INFO DAGScheduler: ShuffleMapStage 6 (showString at NativeMethodAccessorImpl.java:0) finished in 1.337 s
22/12/08 22:04:24 INFO DAGScheduler: looking for newly runnable stages
22/12/08 22:04:24 INFO DAGScheduler: running: Set()
22/12/08 22:04:24 INFO DAGScheduler: waiting: Set(ResultStage 7)
22/12/08 22:04:24 INFO DAGScheduler: failed: Set()
22/12/08 22:04:24 INFO DAGScheduler: Submitting ResultStage 7 (MapPartitionsRDD[16] at showString at NativeMethodAccessorImpl.java:0), which has no missing parents
22/12/08 22:04:24 INFO MemoryStore: Block broadcast_8 stored as values in memory (estimated size 24.5 KB, free 911.4 MB)
22/12/08 22:04:24 INFO MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 12.0 KB, free 911.4 MB)
22/12/08 22:04:24 INFO BlockManagerInfo: Added broadcast_8_piece0 in memory on ip-172-31-84-178.ec2.internal:40833 (size: 12.0 KB, free: 912.2 MB)
22/12/08 22:04:24 INFO SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:1297
22/12/08 22:04:24 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 7 (MapPartitionsRDD[16] at showString at NativeMethodAccessorImpl.java:0) (first 15 tasks are for partitions Vector(0))
22/12/08 22:04:24 INFO YarnScheduler: Adding task set 7.0 with 1 tasks
22/12/08 22:04:24 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 162, ip-172-31-94-230.ec2.internal, executor 1, partition 0, NODE_LOCAL, 8080 bytes)
22/12/08 22:04:24 INFO BlockManagerInfo: Added broadcast_8_piece0 in memory on ip-172-31-94-230.ec2.internal:39123 (size: 12.0 KB, free: 2.1 GB)
22/12/08 22:04:24 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 1 to 172.31.94.230:35828
22/12/08 22:04:24 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 162) in 89 ms on ip-172-31-94-230.ec2.internal (executor 1) (1/1)
22/12/08 22:04:24 INFO YarnScheduler: Removed TaskSet 7.0, whose tasks have all completed, from pool
22/12/08 22:04:24 INFO DAGScheduler: ResultStage 7 (showString at NativeMethodAccessorImpl.java:0) finished in 0.102 s
22/12/08 22:04:24 INFO DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 1.467163 s
+---------+--------------+
|product_id|purchase_count|
+---------+--------------+
|   1004856|         32321|
|   1004767|         22613|
|   1005115|         22244|
|   4804056|         17800|
|   1004833|         13486|
|   1002544|         11678|
|   1004870|         10673|
|   1005100|         10132|
|   1004249|          8881|
|   1005105|          8483|
+---------+--------------+

22/12/08 22:04:24 INFO FileSourceStrategy: Pruning directories with:
22/12/08 22:04:24 INFO FileSourceStrategy: Post-Scan Filters: isnotnull(event_type#1),(event_type#1 = purchase)
22/12/08 22:04:24 INFO FileSourceStrategy: Output Data Schema: struct<event_type: string, product_id: int>
22/12/08 22:04:24 INFO FileSourceScanExec: Pushed Filters: IsNotNull(event_type),EqualTo(event_type,purchase)
22/12/08 22:04:24 INFO FileSourceScanExec: Pushed Filters: EqualTo(none,purchase),IsNotNull(none)
22/12/08 22:04:24 INFO FileOutputCommitter: File Output Committer Algorithm version is 2
22/12/08 22:04:24 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: true
22/12/08 22:04:24 INFO DirectFileOutputCommitter: Direct Write: ENABLED
22/12/08 22:04:24 INFO SQLConfCommitterProvider: Using output committer class org.apache.hadoop.mapreduce.lib.output.DirectFileOutputCommitter
22/12/08 22:04:24 INFO DirectFileOutputCommitter: Nothing to setup since the outputs are written directly.
22/12/08 22:04:24 INFO MemoryStore: Block broadcast_9 stored as values in memory (estimated size 361.3 KB, free 911.1 MB)
22/12/08 22:04:24 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 32.5 KB, free 911.0 MB)
22/12/08 22:04:24 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on ip-172-31-84-178.ec2.internal:40833 (size: 32.5 KB, free: 912.2 MB)
22/12/08 22:04:24 INFO SparkContext: Created broadcast 9 from save at NativeMethodAccessorImpl.java:0
22/12/08 22:04:24 INFO FileSourceScanExec: Planning scan with bin packing, max size: 134217728 bytes, open cost is considered as scanning 4194304 bytes, number of split files: 68, prefetch: false
22/12/08 22:04:24 INFO FileSourceScanExec: relation: None, fileSplitsInPartitionHistogram: ArrayBuffer((1 fileSplits,68))
```

3. Output:

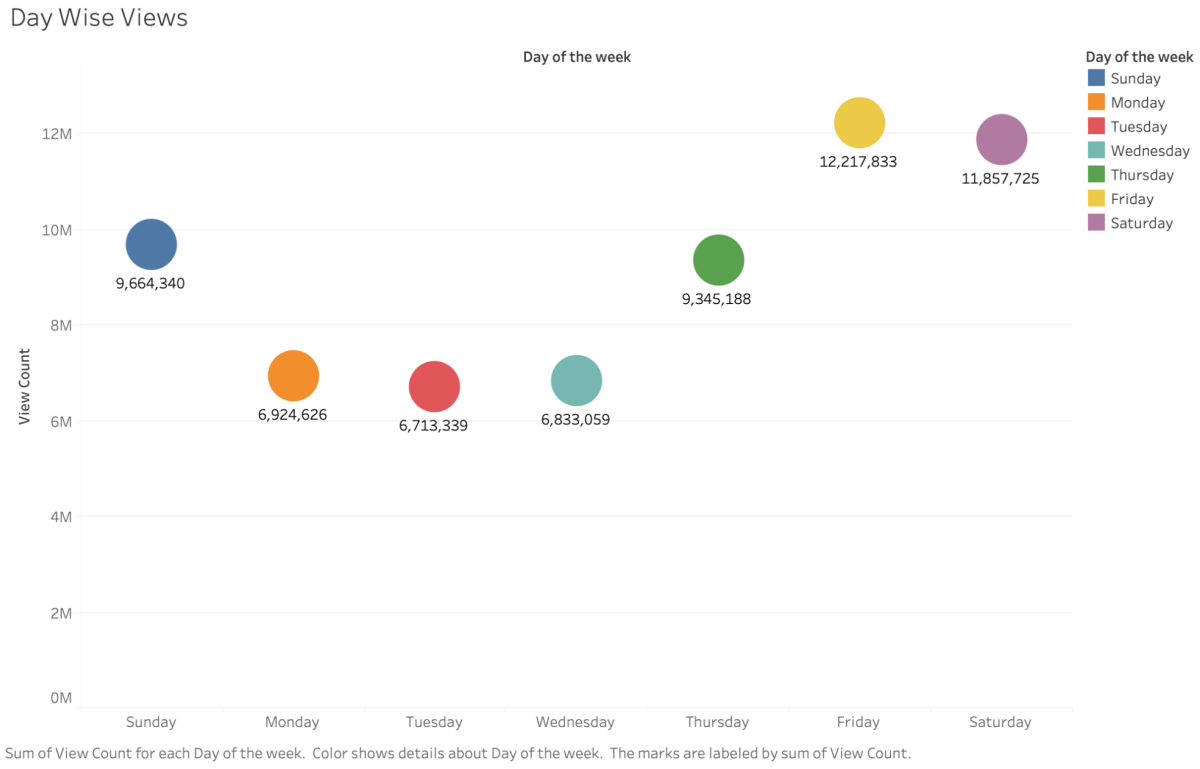

Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > sparkproject02
sparkproject02 Info
Objects
Properties
Permissions
Metrics
Management
Access Points
Objects (11)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
2019-Nov.csv csv December 7, 2022, 14:54:15 (UTC-08:00) 8.4 GB Standard
DaywiseSales.py py December 8, 2022, 12:47:18 (UTC-08:00) 1.8 KB Standard
LeastGrossingCategory.py py December 8, 2022, 15:17:05 (UTC-08:00) 1.8 KB Standard
MostAddedCategory.py py December 8, 2022, 15:18:43 (UTC-08:00) 1.8 KB Standard
MostBoughtBrand.py py December 8, 2022, 15:20:42 (UTC-08:00) 1.6 KB Standard
MostBoughtProduct.py py December 8, 2022, 13:54:57 (UTC-08:00) 1.6 KB Standard
MostValuableCustomers.py py December 8, 2022, 15:22:41 (UTC-08:00) 1.7 KB Standard
MostVisitedCategory.py py December 8, 2022, 15:28:15 (UTC-08:00) 1.8 KB Standard
output/ Folder - - -
SmartphonePurchaseAnalysis.py py December 8, 2022, 15:31:18 (UTC-08:00) 1.8 KB Standard
SmartphoneViewAnalysis.py py December 8, 2022, 15:31:18 (UTC-08:00) 1.8 KB Standard


Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight
AWS Marketplace for S3
Amazon S3 > Buckets > sparkproject02 > output/
output/
Copy S3 URI
Objects
Properties
Objects (8)
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
LeastGrossingCategory/ Folder - - -
MostAddedCategory/ Folder - - -
MostBoughtBrand/ Folder - - -
MostBoughtProduct/ Folder - - -
MostValuableCustomer/ Folder - - -
MostVisitedCategory/ Folder - - -
SmartphonePurchaseAnalysis/ Folder - - -
SmartphoneViewAnalysis/ Folder - - -

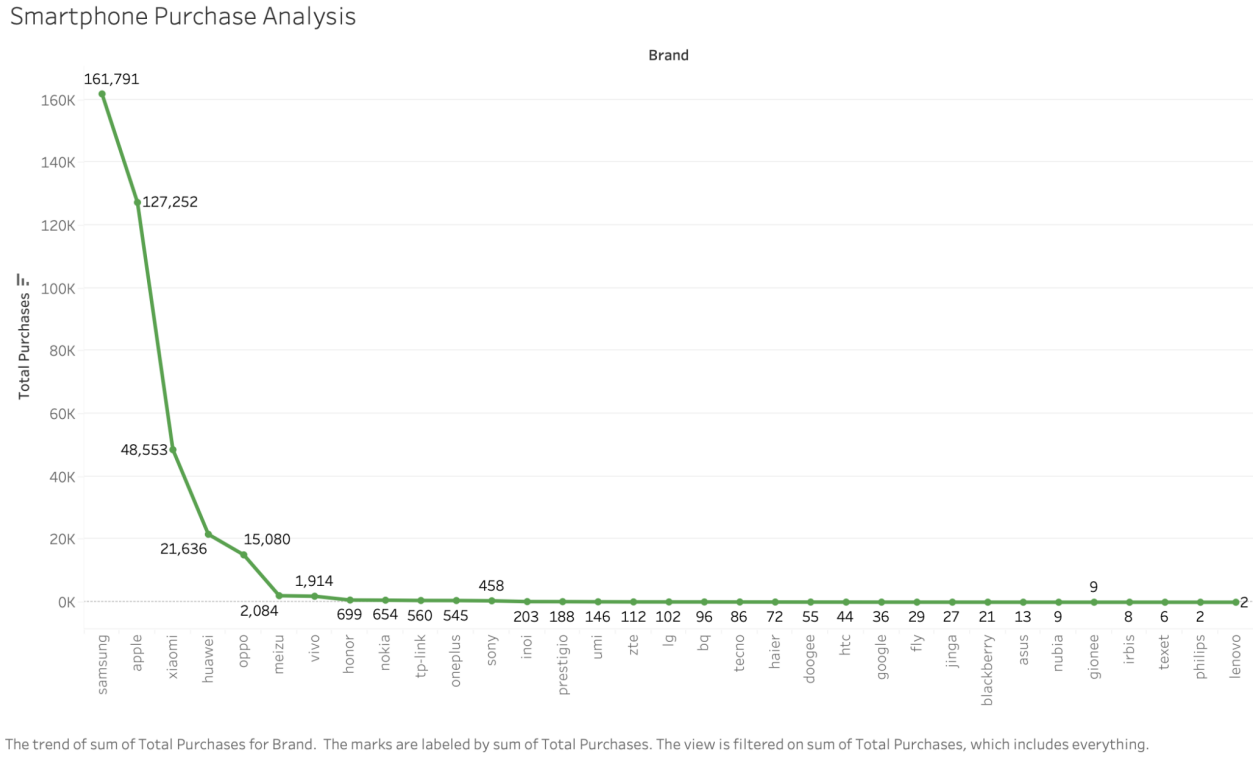

Amazon S3
Buckets
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
Access analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
AWS Organizations settings
Feature spotlight 3
AWS Marketplace for S3
Amazon S3 > Buckets > sparkproject02 > output/ > MostAddedCategory/
MostAddedCategory/
Copy S3 URI
Objects
Properties
Objects (2)
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
_SUCCESS
-
December 8, 2022, 15:57:19 (UTC-08:00)
0 B
Standard
part-00000-5b014d39-d1f2-4d13-97e6-ce663ae7dff7-c000.csv
csv
December 8, 2022, 15:57:19 (UTC-08:00)
296.0 B
Standard

7: Test Results

- **Day Wise Sales:**
  By analyzing day-wise sales, businesses can identify trends in customer behavior and understand which days of the week are most famous for making purchases. This information can be used to make critical decisions about when to run promotions, offer discounts, or launch new products to maximize sales and improve the overall performance of the e-commerce business.
  In addition to tracking day-wise sales, businesses can also analyze sales data by hour of the day, month, or year to gain a more detailed understanding of customer behavior and make more targeted decisions about optimizing their e-commerce operations.



- **Smartphone Purchase Analysis:**
  Key metrics that could be considered in an e-commerce analysis on phone purchases is the popularity of different phone models and brands. This information can help businesses understand which phones are in high demand and should be prioritized in terms of stock and advertising. E-commerce analysis on phone purchases can provide valuable insights to help businesses improve their online sales and better serve their customers.



The trend of sum of Total Purchases for Brand. The marks are labeled by sum of Total Purchases. The view is filtered on sum of Total Purchases, which includes everything.

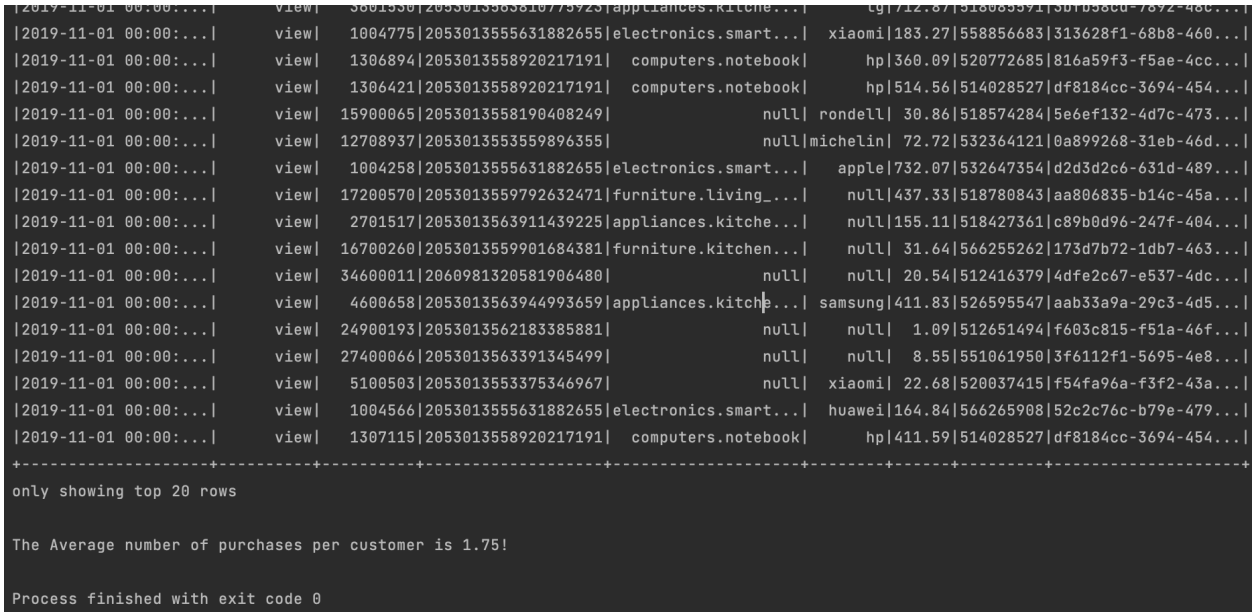
- **Conversion Rate:**
  The conversion rate is the percentage of website visitors who added items to the cart from view, who purchased directly from view, and who made purchases after adding to the cart. This metric can provide insight into how effective a business's website and sales process are at convincing consumers to make a purchase.

```
Rate of conversion between view and purchase events 1.8100796048728667%
Rate of conversion between view and add to cart events 1.522031767054738%
Rate of conversion between add to cart and purchase events 118.92521851731952%

Process finished with exit code 0
```

- **Average Purchases Per Customer:**

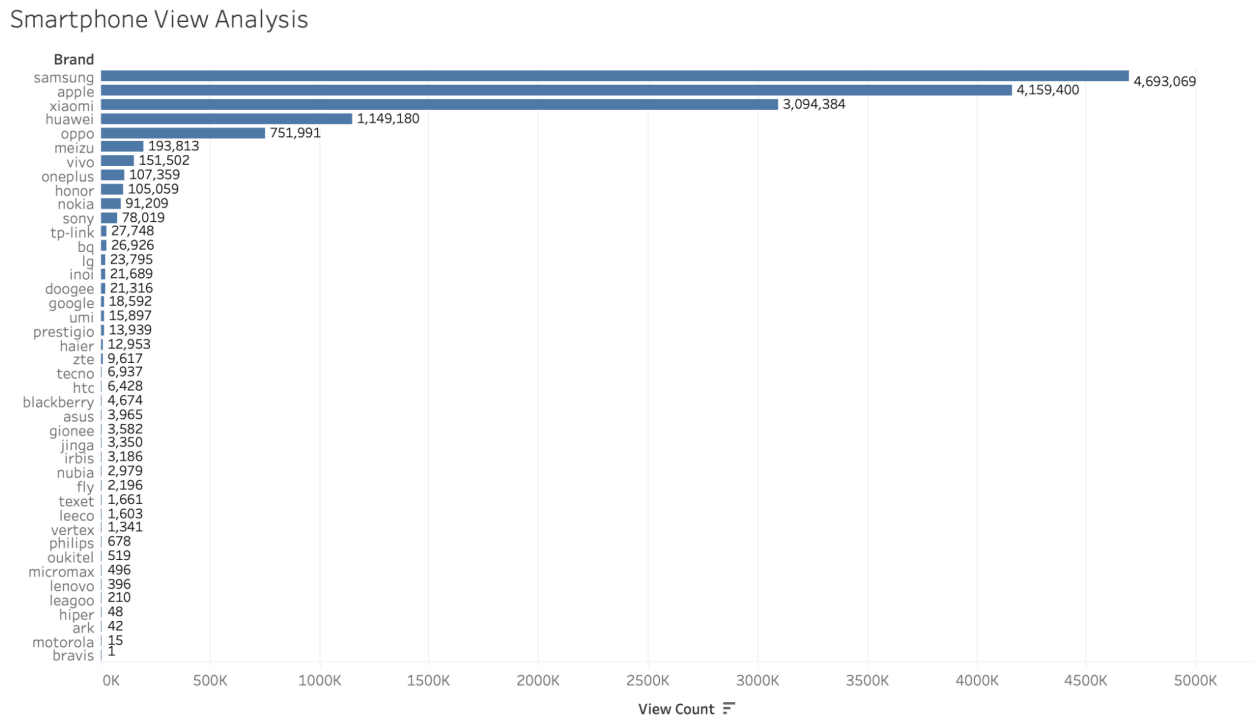
  To calculate the average purchase per customer, businesses can take the total revenue generated from customer purchases and divide it by the number of customers who made a purchase. This will give the average amount that each customer spent on the site.

  By analyzing the average purchase per customer, businesses can gain insight into the overall spending habits of their customers and identify trends in customer behavior. This information can be used to make informed decisions about pricing, product selection, and marketing strategies in order to maximize revenue and improve the overall performance of the e-commerce business

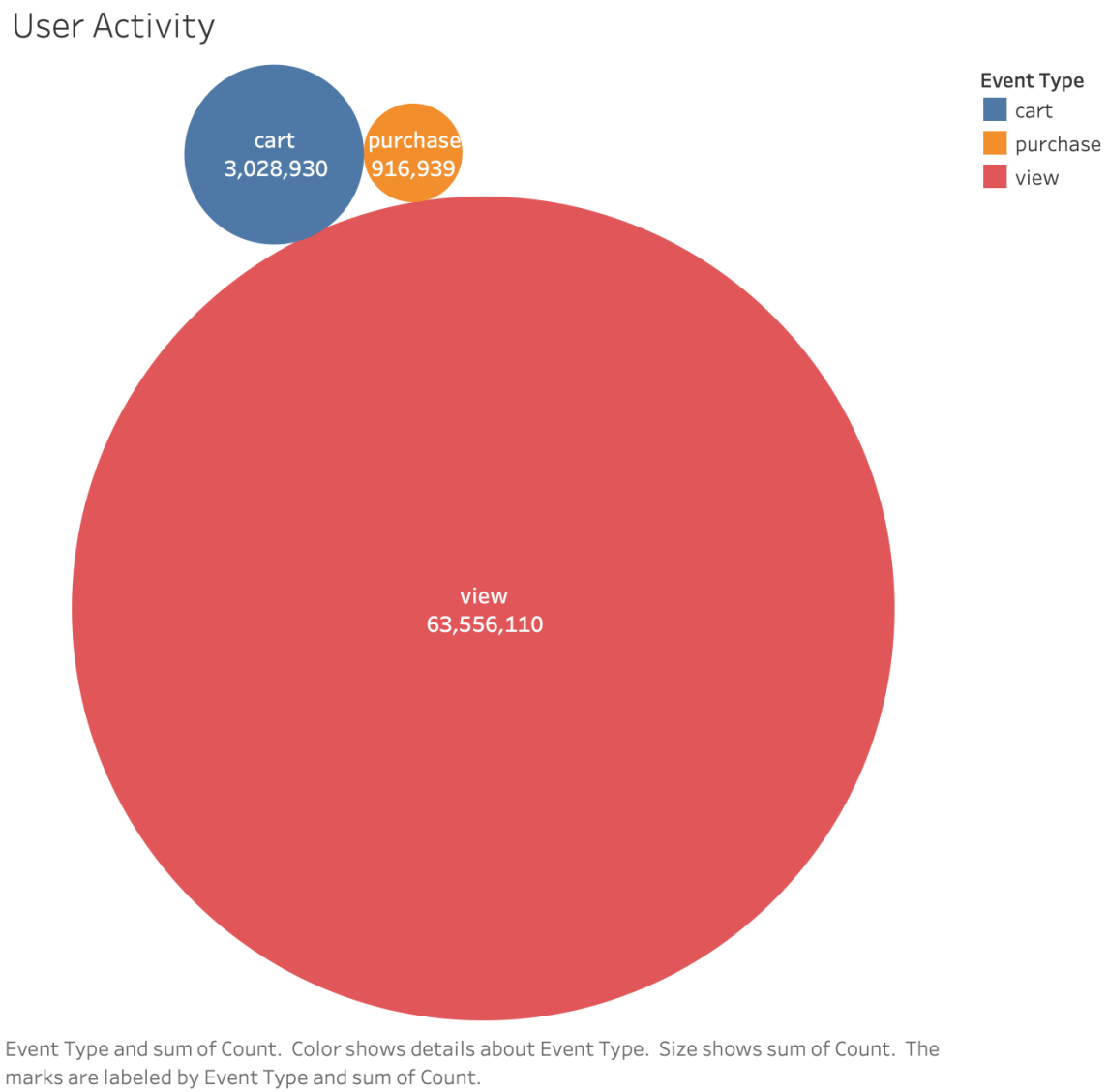


```
|2019-11-01 00:00:...|      view|   3801330|2053013563810775923|appliances.kitche...|      lg|712.87|518083591|30fb38cd-7892-48c...|
|2019-11-01 00:00:...|      view|   1004775|2053013555631882655|electronics.smart...|  xiaomi|183.27|558856683|313628f1-68b8-460...|
|2019-11-01 00:00:...|      view|   1306894|2053013558920217191|  computers.notebook|      hp|360.09|520772685|816a59f3-f5ae-4cc...|
|2019-11-01 00:00:...|      view|   1306421|2053013558920217191|  computers.notebook|      hp|514.56|514028527|df8184cc-3694-454...|
|2019-11-01 00:00:...|      view|  15900065|2053013558190408249|                null| rondell| 30.86|518574284|5e6ef132-4d7c-473...|
|2019-11-01 00:00:...|      view|  12708937|2053013553559896355|                null|michelin| 72.72|532364121|0a899268-31eb-46d...|
|2019-11-01 00:00:...|      view|   1004258|2053013555631882655|electronics.smart...|   apple|732.07|532647354|d2d3d2c6-631d-489...|
|2019-11-01 00:00:...|      view|  17200570|2053013559792632471|furniture.living_...|    null|437.33|518780843|aa806835-b14c-45a...|
|2019-11-01 00:00:...|      view|   2701517|2053013563911439225|appliances.kitche...|    null|155.11|518427361|c89b0d96-247f-404...|
|2019-11-01 00:00:...|      view|  16700260|2053013559901684381|furniture.kitchen...|    null| 31.64|566255262|173d7b72-1db7-463...|
|2019-11-01 00:00:...|      view|  34600011|2060981320581906480|                null|    null| 20.54|512416379|4dfe2c67-e537-4dc...|
|2019-11-01 00:00:...|      view|   4600658|2053013563944993659|appliances.kitche...| samsung|411.83|526595547|aab33a9a-29c3-4d5...|
|2019-11-01 00:00:...|      view|  24900193|2053013562183385881|                null|    null|  1.09|512651494|f603c815-f51a-46f...|
|2019-11-01 00:00:...|      view|  27400066|2053013563391345499|                null|    null|  8.55|551061950|3f6112f1-5695-4e8...|
|2019-11-01 00:00:...|      view|   5100503|2053013553375346967|                null|  xiaomi| 22.68|520037415|f54fa96a-f3f2-43a...|
|2019-11-01 00:00:...|      view|   1004566|2053013555631882655|electronics.smart...|  huawei|164.84|566265908|52c2c76c-b79e-479...|
|2019-11-01 00:00:...|      view|   1307115|2053013558920217191|  computers.notebook|      hp|411.59|514028527|df8184cc-3694-454...|
+-------------------+----------+----------+-------------------+--------------------+--------+------+---------+--------------------+
only showing top 20 rows

The Average number of purchases per customer is 1.75!

Process finished with exit code 0
```

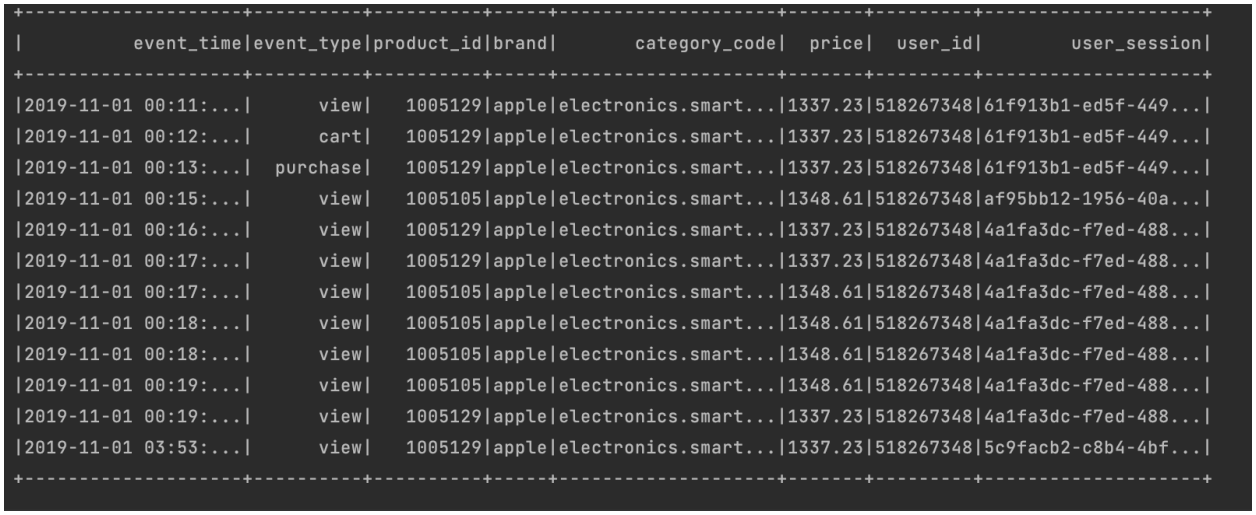
- **Smartphone View Analysis:**
  By analyzing smartphone views, businesses can identify common patterns and trends in user behavior on mobile devices, such as which brands are most popular and which models have high drop-off rates. This information can be used to make changes to the device in order to improve the user experience and increase conversions to the particular brand. For example, if a particular brand has a high drop-off rate on mobile devices, the business can investigate the reason for this and make changes to its model in order to improve sales and conversion on smartphones.



- **User Activity:**
  E-commerce user activity analysis is the process of examining user behavior on an e-commerce website in order to understand and improve the user experience. This can be done by analyzing the count of views, the number of items in the cart, and successful purchases. The goal of e-commerce user activity analysis is to identify trends and patterns in user behavior in order to make informed decisions about how to optimize the website for better user engagement and conversion. By analyzing user activity, e-commerce businesses can better understand their customers and make more informed decisions about improving their online shopping experience.



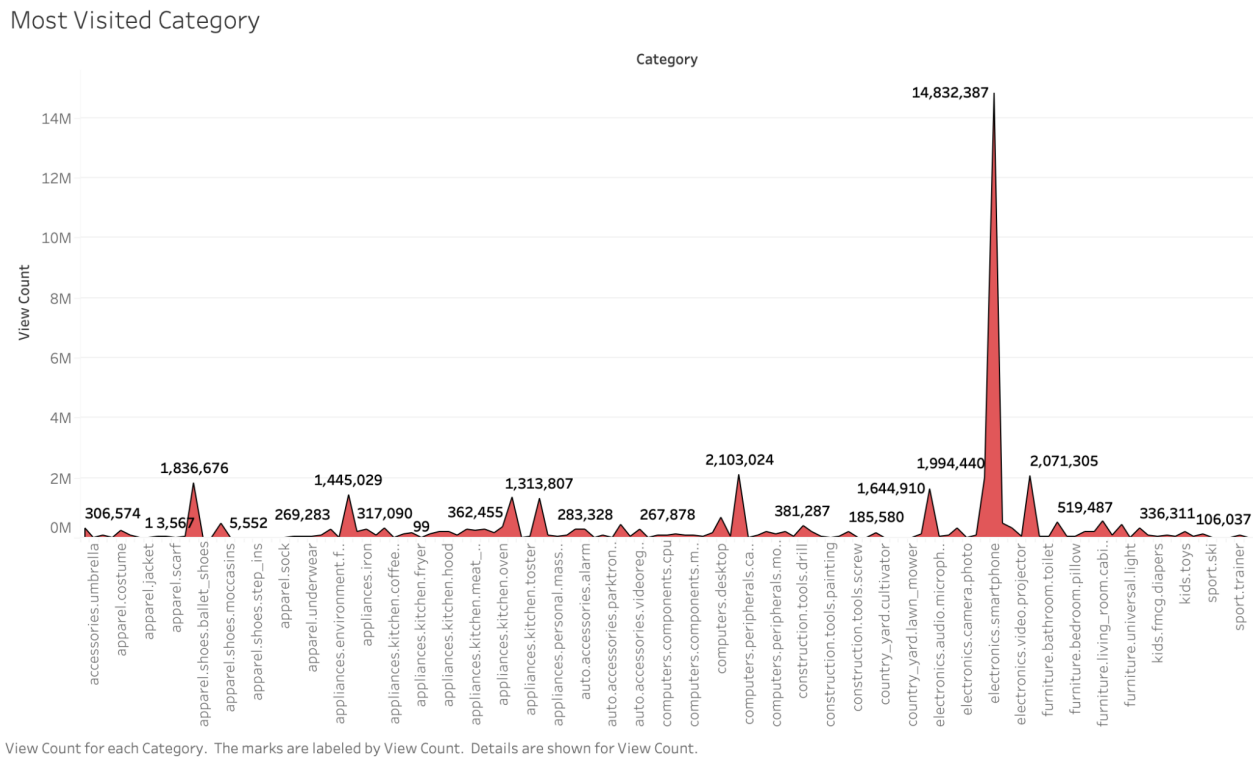
- **Purchase Path:**
  E-commerce analysis on purchase path typically involves analyzing the steps a customer takes from the time they first visit a website or app to the time they complete a purchase. This type of analysis can help businesses understand the customer journey and identify potential barriers or points of friction that may be hindering conversions. For example, analyzing purchase path data may reveal that many customers are dropping off at the checkout page, indicating a problem with the payment process. This information can then be used to make improvements to the checkout experience, such as streamlining the payment process or offering additional payment options. Additionally, purchase path analysis can be used to identify opportunities for upselling or cross-selling, such as by highlighting related products during the checkout process. Overall, purchase path analysis can provide valuable insights into customer behavior and help businesses optimize their e-commerce experience.



```
+-------------------+----------+----------+-----+--------------------+-------+---------+--------------------+
|          event_time|event_type|product_id|brand|       category_code|  price|  user_id|        user_session|
+-------------------+----------+----------+-----+--------------------+-------+---------+--------------------+
|2019-11-01 00:11:...|      view|   1005129|apple|electronics.smart...|1337.23|518267348|61f913b1-ed5f-449...|
|2019-11-01 00:12:...|      cart|   1005129|apple|electronics.smart...|1337.23|518267348|61f913b1-ed5f-449...|
|2019-11-01 00:13:...|  purchase|   1005129|apple|electronics.smart...|1337.23|518267348|61f913b1-ed5f-449...|
|2019-11-01 00:15:...|      view|   1005105|apple|electronics.smart...|1348.61|518267348|af95bb12-1956-40a...|
|2019-11-01 00:16:...|      view|   1005129|apple|electronics.smart...|1337.23|518267348|4a1fa3dc-f7ed-488...|
|2019-11-01 00:17:...|      view|   1005129|apple|electronics.smart...|1337.23|518267348|4a1fa3dc-f7ed-488...|
|2019-11-01 00:17:...|      view|   1005105|apple|electronics.smart...|1348.61|518267348|4a1fa3dc-f7ed-488...|
|2019-11-01 00:18:...|      view|   1005105|apple|electronics.smart...|1348.61|518267348|4a1fa3dc-f7ed-488...|
|2019-11-01 00:18:...|      view|   1005105|apple|electronics.smart...|1348.61|518267348|4a1fa3dc-f7ed-488...|
|2019-11-01 00:19:...|      view|   1005105|apple|electronics.smart...|1348.61|518267348|4a1fa3dc-f7ed-488...|
|2019-11-01 00:19:...|      view|   1005129|apple|electronics.smart...|1337.23|518267348|4a1fa3dc-f7ed-488...|
|2019-11-01 03:53:...|      view|   1005129|apple|electronics.smart...|1337.23|518267348|5c9facb2-c8b4-4bf...|
+-------------------+----------+----------+-----+--------------------+-------+---------+--------------------+
```

- **Most Visited Category:**
  By analyzing the most visited category, businesses can gain insight into the interests and preferences of their customers and make informed decisions about how to optimize their product selection and marketing strategies. For example, if a particular category is consistently the most visited, the business may want to invest more resources in promoting and expanding that category in order to drive more sales and improve the overall performance of the e-commerce business.

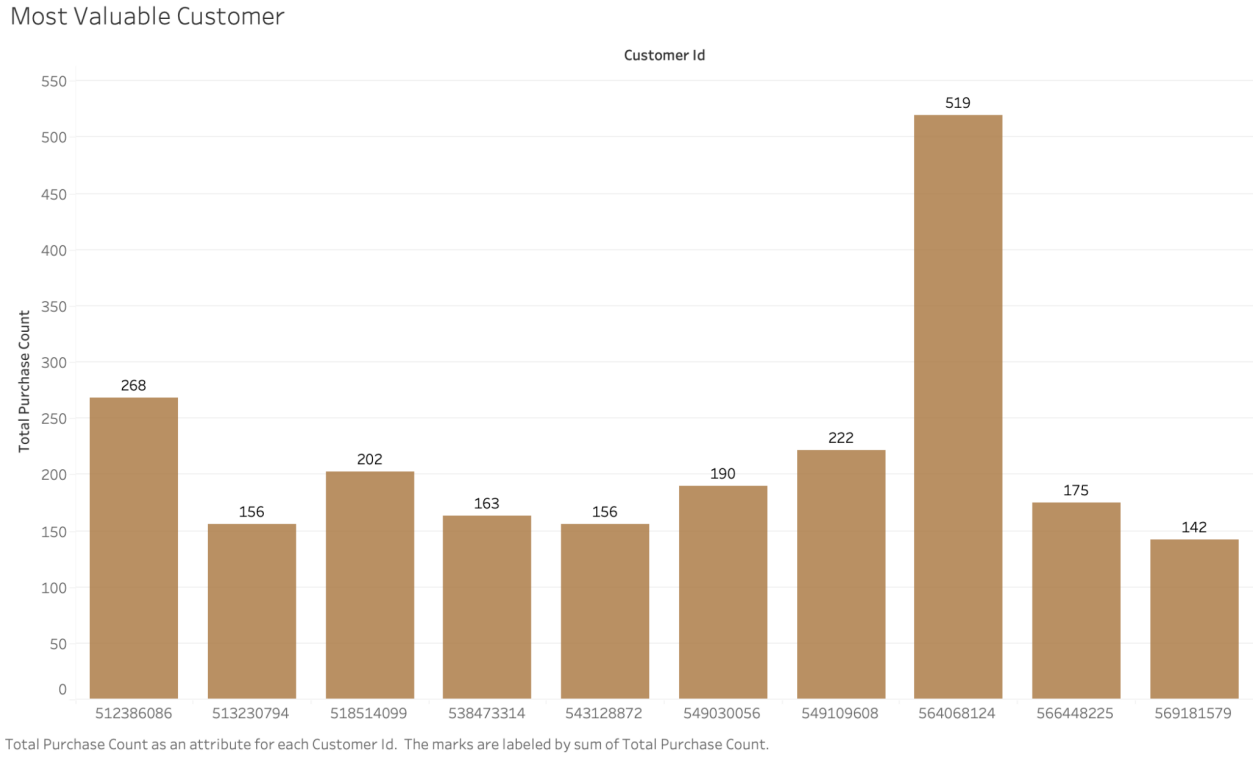
  In addition to analyzing the most visited category, businesses can also study other metrics, such as the average time spent on each category page and the conversion rate for each category, in order to gain a complete understanding of customer behavior and make more informed decisions about how to optimize the e-commerce website.



View Count for each Category. The marks are labeled by View Count. Details are shown for View Count.

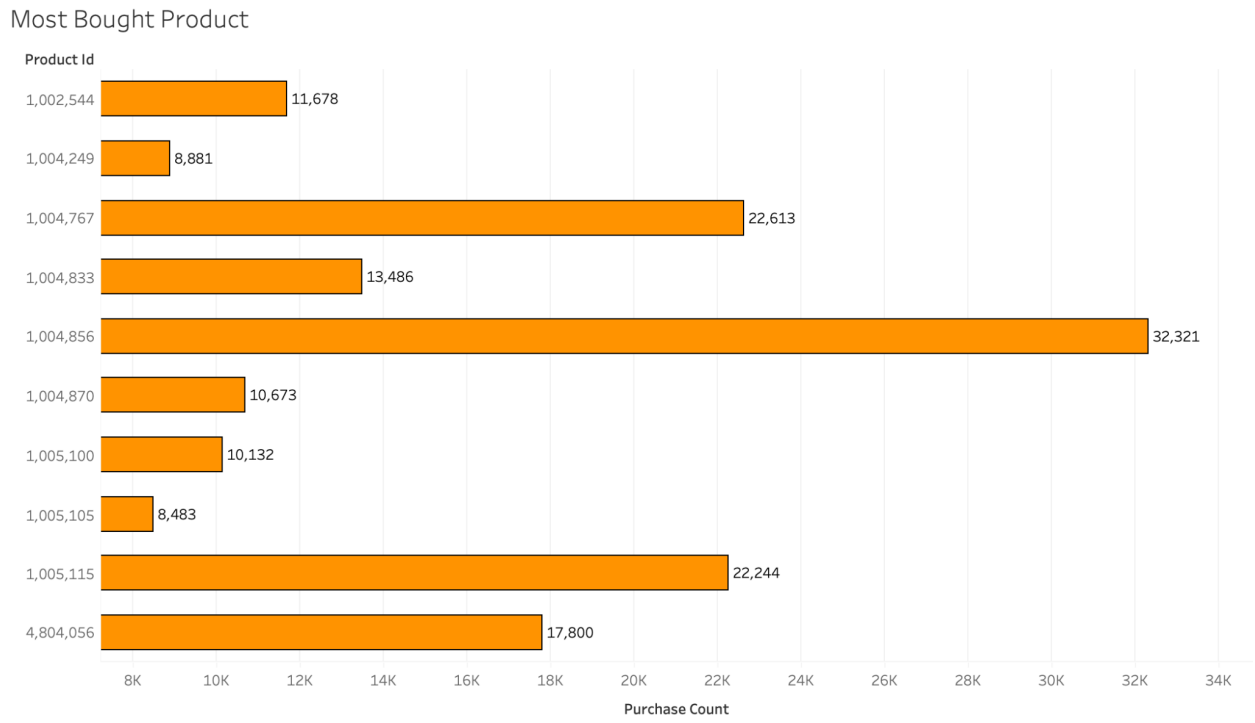
- **Most Valuable Customers:**
  By analyzing the most valuable customers, businesses can gain insight into the spending habits of their customers and identify opportunities to improve customer retention and increase revenue. This information can be used to make informed decisions about pricing, product selection, and marketing strategies in order to maximize the value of each customer and improve the overall performance of the e-commerce business.

  In addition to analyzing the most valuable customers, businesses can also study other metrics, such as customer lifetime value, customer acquisition cost, and customer retention rate, in order to gain a complete understanding of customer behavior and make more informed decisions about how to optimize the e-commerce website.



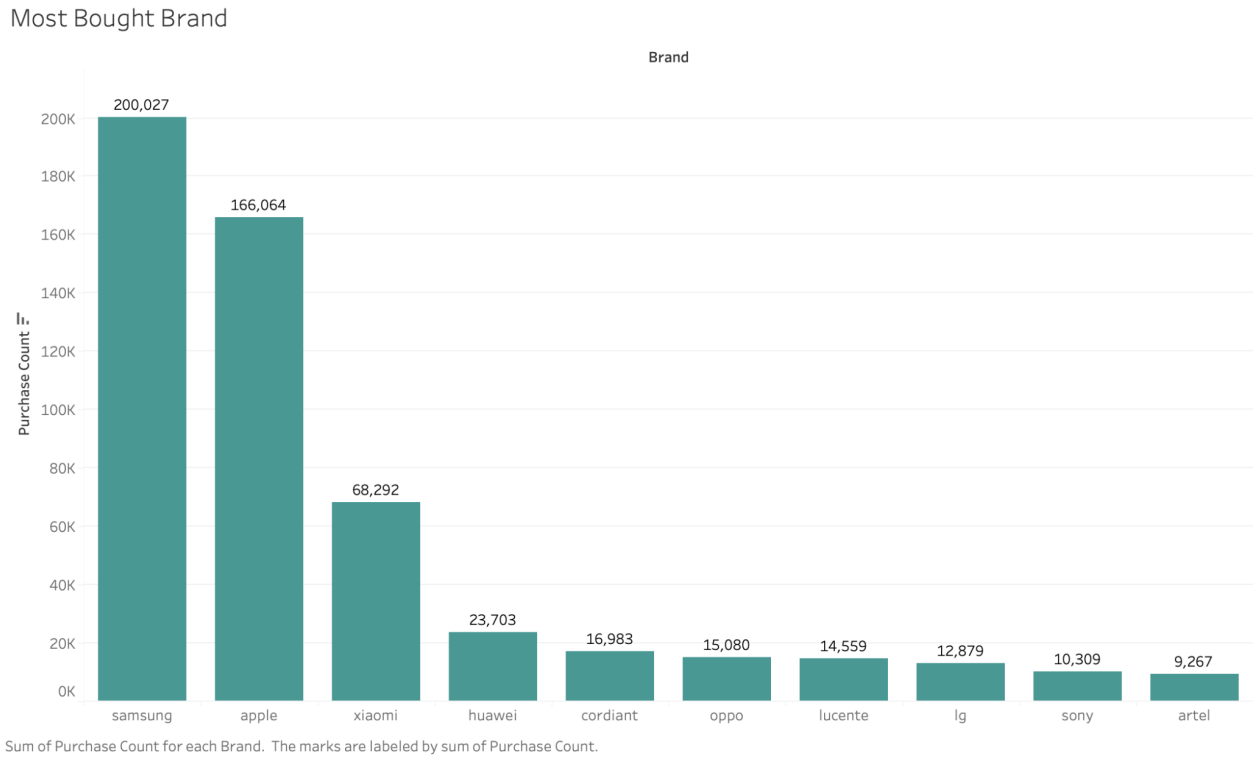
- **Most Bought Product:**
  By analyzing the most bought product, businesses can gain insight into the interests and preferences of their customers and make informed decisions about how to optimize their product selection and marketing strategies. For example, if a particular product is consistently the most bought, the business may want to invest more resources in promoting and expanding that product in order to drive more sales and improve the overall performance of the e-commerce business.

  In addition to analyzing the most bought product, businesses can also study other metrics, such as the average order value and the conversion rate for each product, in order to gain a more complete understanding of customer behavior and make more informed decisions about how to optimize the e-commerce website.



- **Most Bought Brand:**

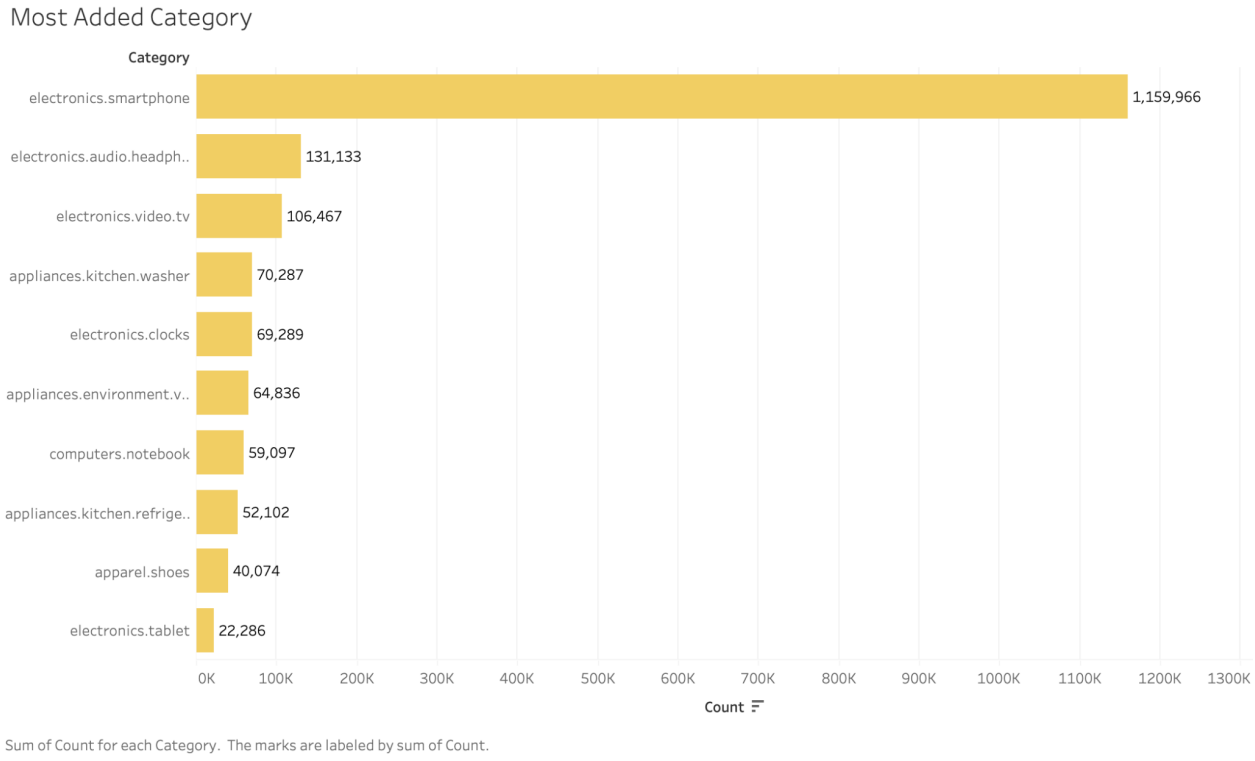
  E-commerce analysis on most bought brand typically include studying consumers purchasing habits, the products they buy, and the brands they prefer. In the case of a "Most Bought Brand," e-commerce analysis involve identifying the brand that is most frequently purchased by consumers on an online platform. This information can be used by companies to better understand consumer preferences and make decisions about their product offerings and marketing strategies.



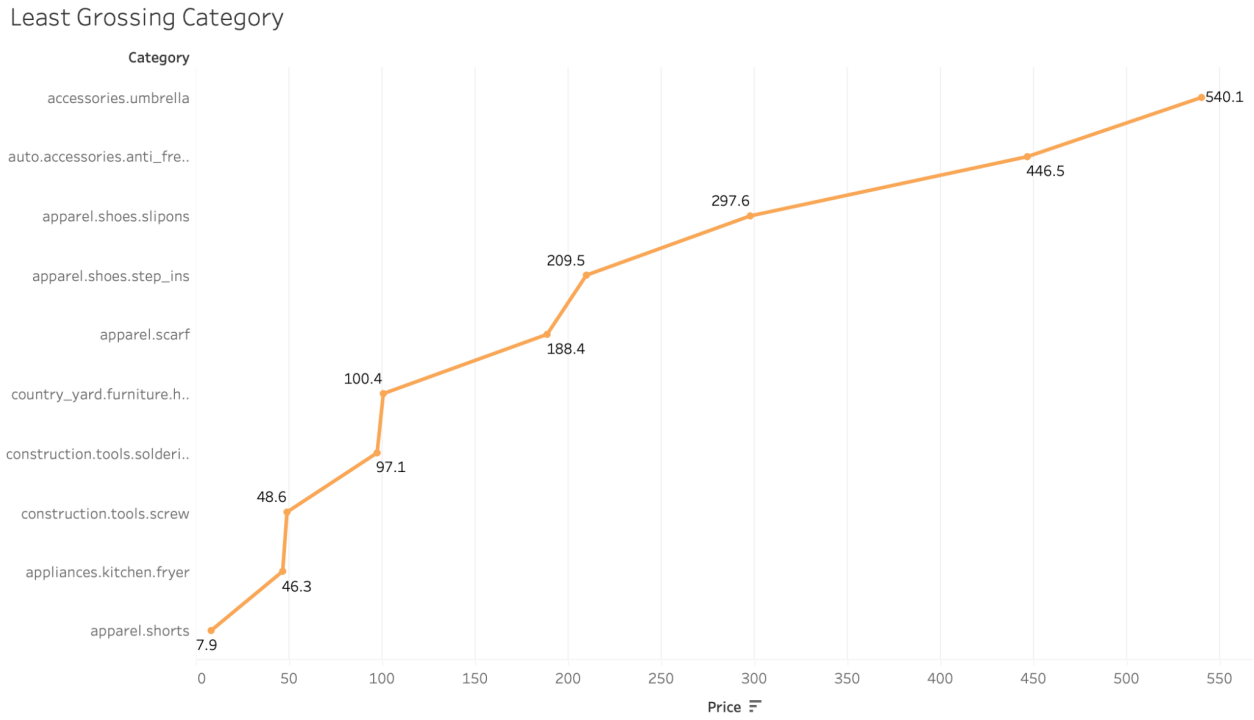
- **Most Added Category:**

  To perform an e-commerce analysis on the Most Added Category, you would need to collect data on the items that customers are adding to their online shopping carts. Once you have collected this data, you can use it to identify trends and patterns in customer behavior. You could also use this data to identify any potential problems or areas for improvement, such as if there are certain products or services that are not being added to shopping carts as often as you would expect.

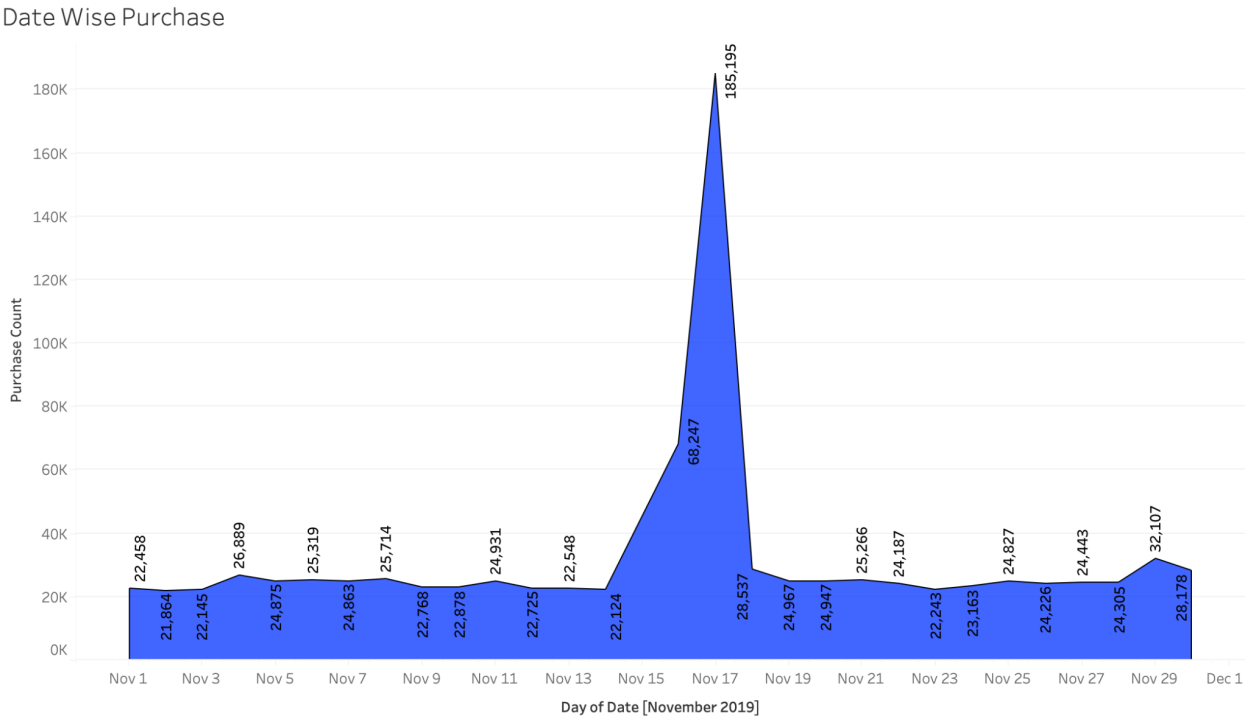
  Overall, an e-commerce analysis of the Most Added Category can provide valuable insights into customer behavior and preferences, which can help businesses to improve their online shopping experiences and increase sales.



- **Least Grossing Category:**
  By analyzing the least grossing category, businesses can gain insight into the areas of the business that may be underperforming and make informed decisions about how to optimize their product selection and marketing strategies. For example, if a particular category consistently generates low revenue, the business may want to review the products in that category and make changes to improve its appeal to customers in order to drive more sales and improve the overall performance of the e-commerce business.



- **Date Wise Purchase:**
  E-commerce analysis on date wise purchase typically involves analyzing data related to the number and value of purchases made on different dates. This type of analysis can help businesses understand trends in consumer behavior and identify opportunities for growth. For example, analyzing date wise purchase data may reveal that sales typically spike during the holiday season(i.e. Black Friday) or on specific days of the week. This information can then be used to inform marketing and sales strategies, such as offering promotions on slow-selling days or increasing inventory levels during peak periods. Additionally, date wise purchase analysis can be used to identify potential issues, such as a sudden decrease in sales on a particular date, which may indicate a problem with the website or payment system. Overall, date wise purchase analysis can provide valuable insights into consumer behavior and help businesses make data-driven decisions.



The plot of sum of Purchase Count for Date Day. The marks are labeled by sum of Purchase Count.