CCT College Dublin

Assessment Cover Page

Module Title:	Storage Solutions for Big Data, Data Visualization Techniques
Assessment Title:	Integrated Continuous Assessment 2
Lecturer Name:	Muhammad Iqbal, James Garza
Student Full Name:	Anna Georgieva
Student Number:	sbs23039@student.cct.ie
Assessment Due Date:	26/11/2023 @ 11.55pm
Date of Submission:	23/11/2023

Anna Georgieva

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Integrated Continuous Assessment 2

Word count: 2,089

- Github repository: Link
- Dashboard app: Link

Table of contents

List of Figures

Assessment Task for Solutions for Big Data

Question 1:

Word Count: 385

Can you define Big Data? Explain major characteristics of Big Data. Can banks enhance their profits with the support of big data processing and analysis? Research and name top three businesses that have obtained the benefits of big data storage solutions in the recent past.
(20 Marks)

Modern data processing applications and relational database management systems are burdened when faced with the complex tasks of handling, storing, transferring, analysing, predicting, and visualising data, this is where Big Data technologies step in. Organisations must analyse terabytes of data to generate actionable insights and sustain daily business operations effectively. (Lutkevich, 2023) Big data is characterised by the three *Vs:*

- *Volume:* When there is a high demand to store large amount of data;
- *Velocity:* Ensuring that extracted insights keep pace with the continual stream of incoming data;
- *Variety:* Allowing data from different sources like website, wearable technology, mobile devices, computer devices, times and structures to be processed;

In order for business to evolve, it is necessary to examine these data points and make informed decisions based on correlations and patterns seen in the data (Duggal, 2023).

Costley and Lankford (2014) carried out a study centred on the banking sector and they estimated that approximately 50% of banks are developing or have already developed multi-tenant analytics platforms utilising advanced big data technologies and the fundamental solution is Hadoop. Costley and Lankford (2014) state the workloads encountered in this study were focused on card fraud detection, early detection of security frauds, credit risk assessment, real-time exchange feed, social analytics for trading, trade visibility, archival of audit trails to name a few. The shift to big data in banking is due to the regulatory and legal requirements.

American Express analyses customer behaviour with Big data solutions (O' Neill, 2019). By analysing historical transactions with more than 100 variables and employing predictive machine learning models, American express can predict 24% of their accounts, who are more likely to churn within 4 months. It allows them to forecast and evaluate customer loyalty. Amazon leverages big data to strengthen its customer relationships. By utilising their comprehensive data records, they ensure that whenever a user contacts the Support team, the team already has a complete history and understanding of the user's interactions and preferences. This approach enables more personalised and effective support. Netflix is another big corporation, which uses big data and it is primarily for getting insights into the viewing habits of millions of consumers. When Netflix has this information, they purchase the rights of certain films and series, which they know will perform well among certain types of their audiences.

Question 2:

Word Count: 437

What is HDFS, and how does it differ from a traditional file system? Describe and explain different layers of Hadoop framework. Explain five important characteristics of the Hadoop framework. Show the deployment of Hadoop on your virtual machine (VM) by providing the screenshots of (Namenode, Datanode etc.) and your username clearly shows your VM.
(20 Marks)

Hadoop Distributed File System (HDFS) started with Google file systems and Mapreduce (Turkington et al., 2016). These two technologies allowed data processing on a large scale. Traditional file system allows storing the data on a single server and it is suitable for small businesses or personal use (Micronimics, 2022). They are not scalable like distributed file systems and there is a high vulnerability in regards to data loss. Traditional file systems are easier to use, because of a less of a learning curve. They work well on small files or files, which are not stored on the same server. Traditional files are not scalable and they lack redundancy.

While HDFS has five important characteristics of the Hadoop Framework.

- It offers **fault tolerance**, which means that it can continue its operation even if Namenode and Datanode fails, because the data is replicated among different data blocks.
- HDFS **can store** tremendous amounts of data and it is designed for batch processing of large datasets. It is capable of storing diverse data types, ranging from megabytes to petabytes, excels in data streaming.
- It can **scale** its capacity by adding more servers, Hadoop clusters can be easily expanded by adding more nodes. This allows it to handle increasing data volumes efficiently.
- It is **flexible** in regards to data processing, allowing it to work with semi-structures, structured and unstructured data.
- Hadoop runs on **commodity hardware,** making it affordable to store and process large datasets and **incurs no licensing costs**.

Layers of Hadoop Framework

- (Reddy, 2020) (DataFlair Team, 2019) **HDFS** segments files into blocks for storage across a distributed network. It is a data storage of Hadoop. It segregates the data into smaller block units and stores them in a distributed way. It has two demons - NameNode and DataNode. NameNode is responsible for the management of Namespace and regulation of file access, while DataNode is responsible for the storage of the data.
- **YARN** (Yet Another Resource Negotiator): The main principle of YARN is to handle job scheduling and cluster management. It has two demons - ResourceManager acts as a mediator for resource allocation across all applications competing for resources within the system and NodeManager's role involves tracking the resource consumption of containers and relaying this information back to the ResourceManager.

- **MapReduce:** The Map function converts input data into a format suitable for mapping, while the Reduce function takes this data to generate the end result.
- **Hadoop Common:** A collection of Java libraries crucial for starting Hadoop and employed throughout its different modules.

When jps script is run, it lists all java processes, and it shows that Hadoop daemons are running successfully on the system.

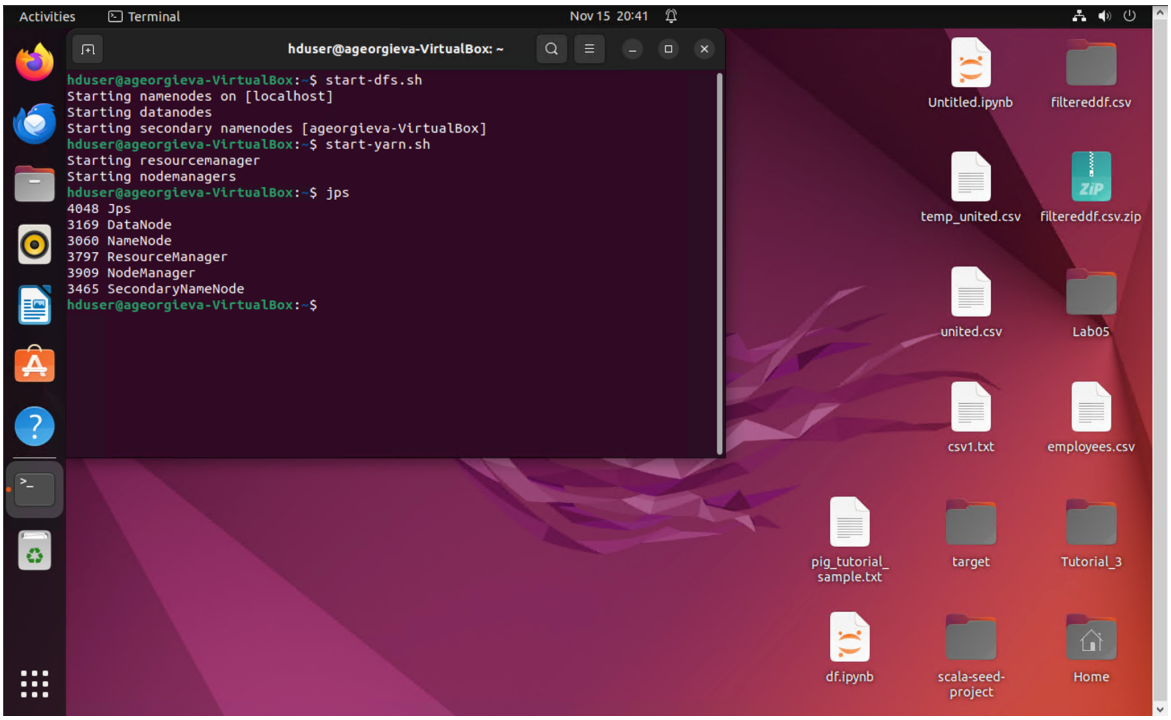


Figure 1: Starting DataNode, NameNode, ResourceManager, NodeManager, SecondaryNameNode

Question 3:

Word Count: 247

Demonstrate a comparison of MySQL and Apache Hive based on the architecture and performance. Consider a dataset and perform a query on both systems with at least 5,000 rows and at least 5 features. Show the duration of query execution by displaying screenshots obtained from a virtual machine (VM).
(20 Marks)

Comparison of MySQL and HIVE

	MySQL*(seconds)*	**HIVE**(seconds)
Create Table	0.08	6.476
Reading csv into new table	local path, not HDFS 0.79	HDFS 0.219
Select from table Limit 10	0	18.752
Describe table	0.02	0.767
Select Transaction ID 1000000008	0.04	6.25
Select Transaction ID 1000004008	0.01	1.596
Select from table where total items > 5 order by total cost desc	0.01	6.539

Figure 2: Comparison of MySQL and HIVE

I ran queries on both MySQL and HIVE and here are the findings:

Architecture:

- MySQL is a traditional relational database management system that doesn't natively support HDFS. It is required to use SQOOP if I read data from HDFS with MySQL (Community Cloudera, 2014). It operates on a single server and is designed for Online Transaction Processing systems (MySQL, 2023). I got an error while running the import csv into retail_table, so I used a local path instead. I also updated the MySQL configurations with *local-infile,* which allows importing data from *txt* or *csv* files.

- I could import the csv from HDFS into a HIVE table easily and the import took a shorter time than MySQL import. Hive is a data warehouse system built on top of Hadoop. It provides SQL-like language called HiveQL, which translates SQL-like queries into MapReduce jobs executed on Hadoop. Hive is designed to handle large datasets typical in big data applications. HIVE consists of 3 core parts (Guru99, 2023): *Hive Clients*, *Hive Services*, *Hive Storage and compute* .

Performance:

- MySQL is suitable for real-time query processing and handles complex queries on structured data. It's highly performant on smaller datasets and is commonly used for online transaction systems.
- HIVE takes a longer time to query the data, as seen from the comparison table; It is designed for batch processing on large datasets and is not intended for real-time queries. It is traditionally slower than MySQL due to the overhead of MapReduce jobs (Quora, 2019).

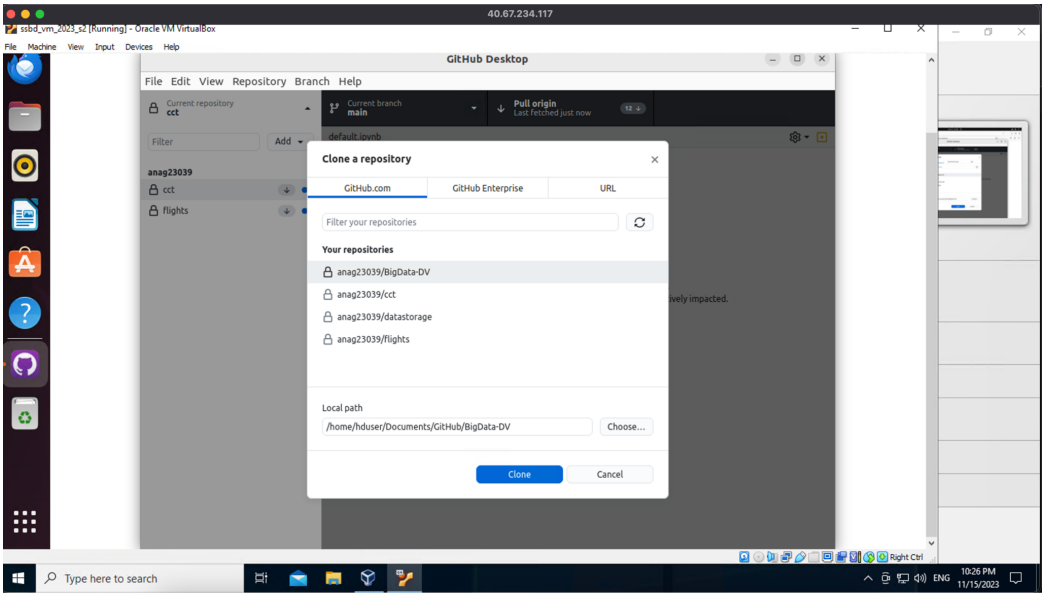


Figure 3: Clone repository on Oracle VM

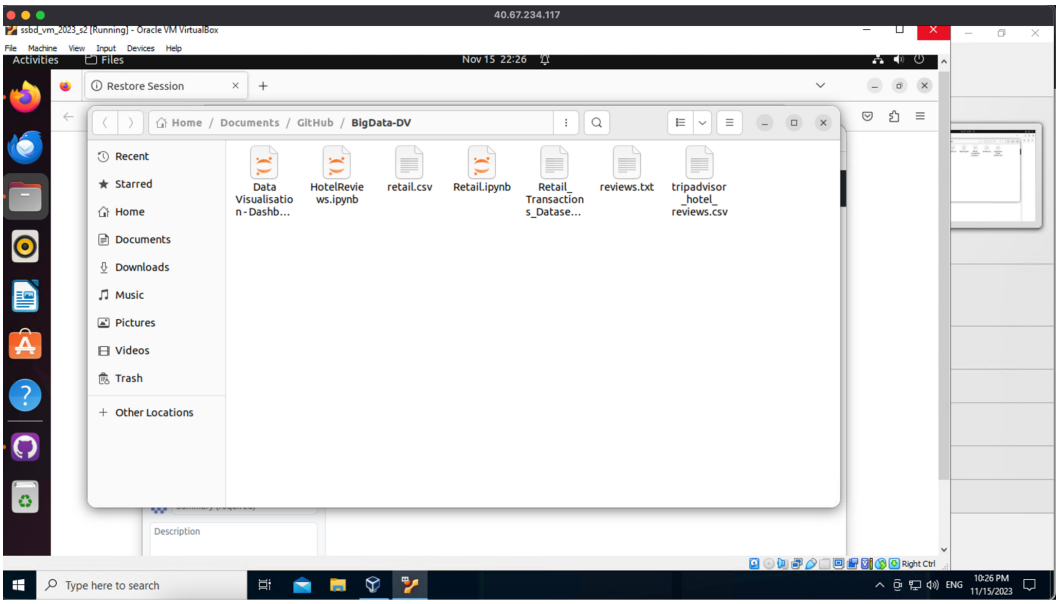


Figure 4: Path to files

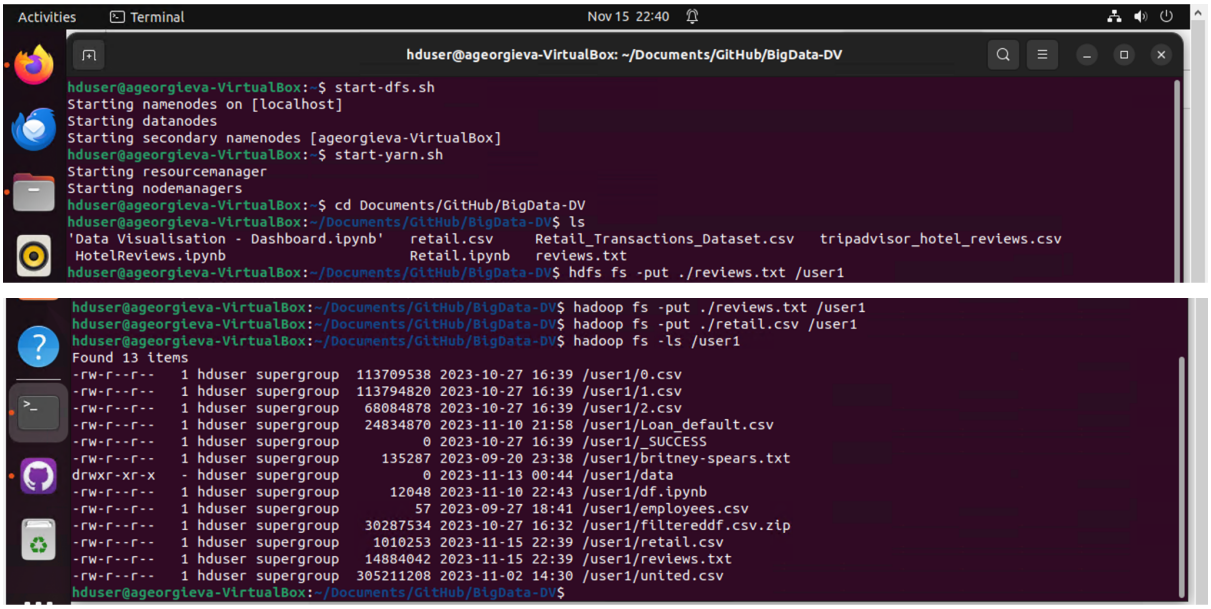
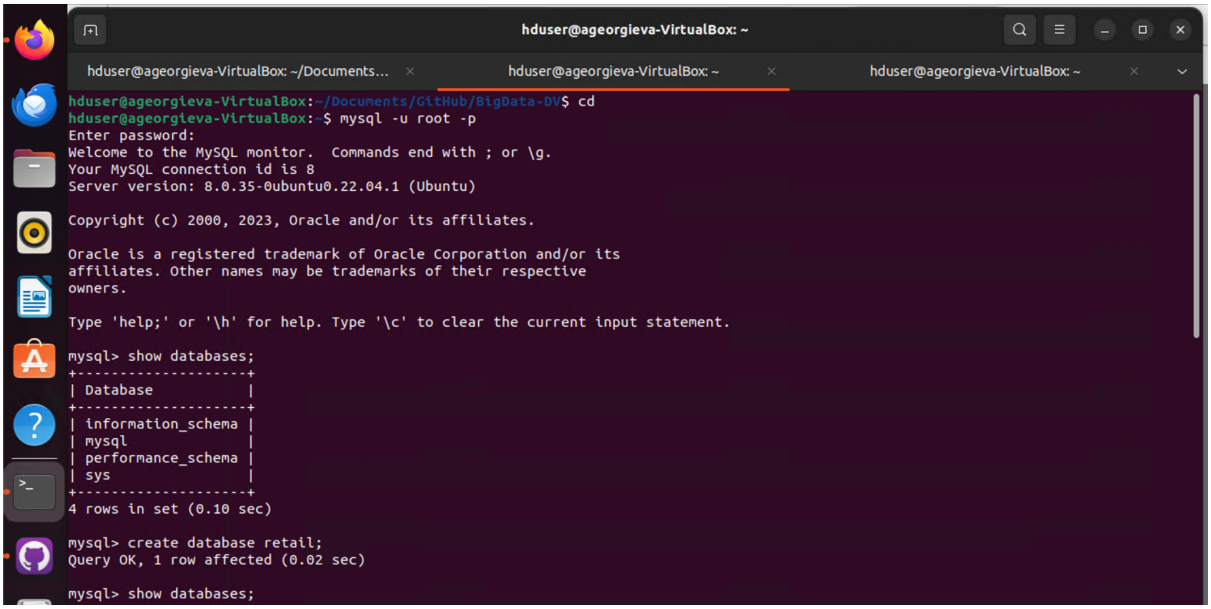
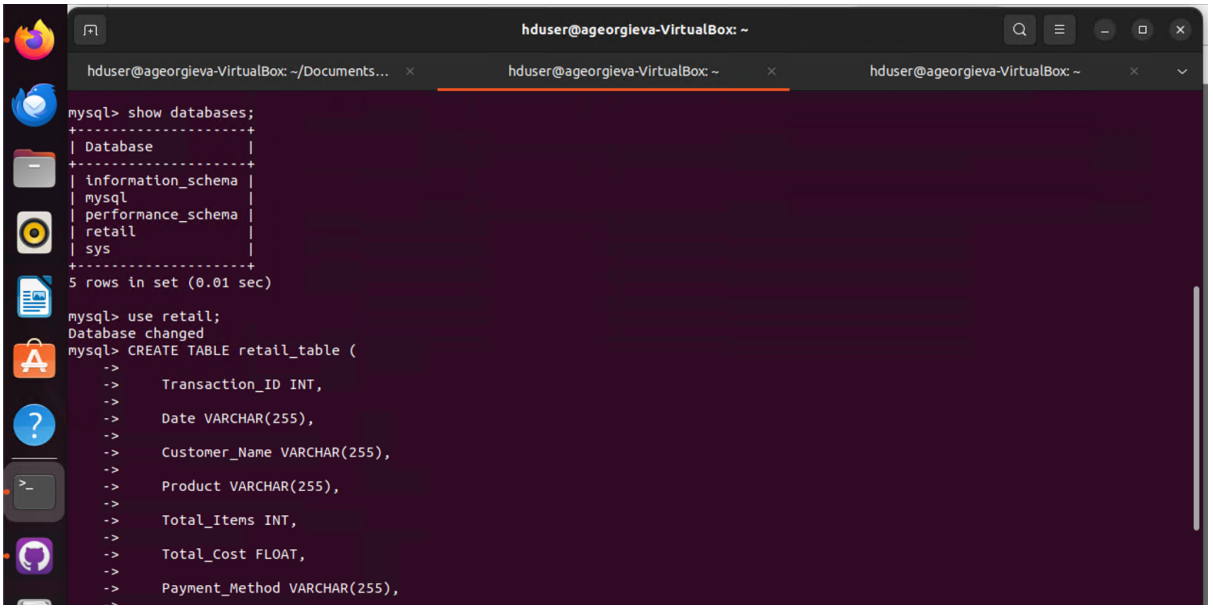
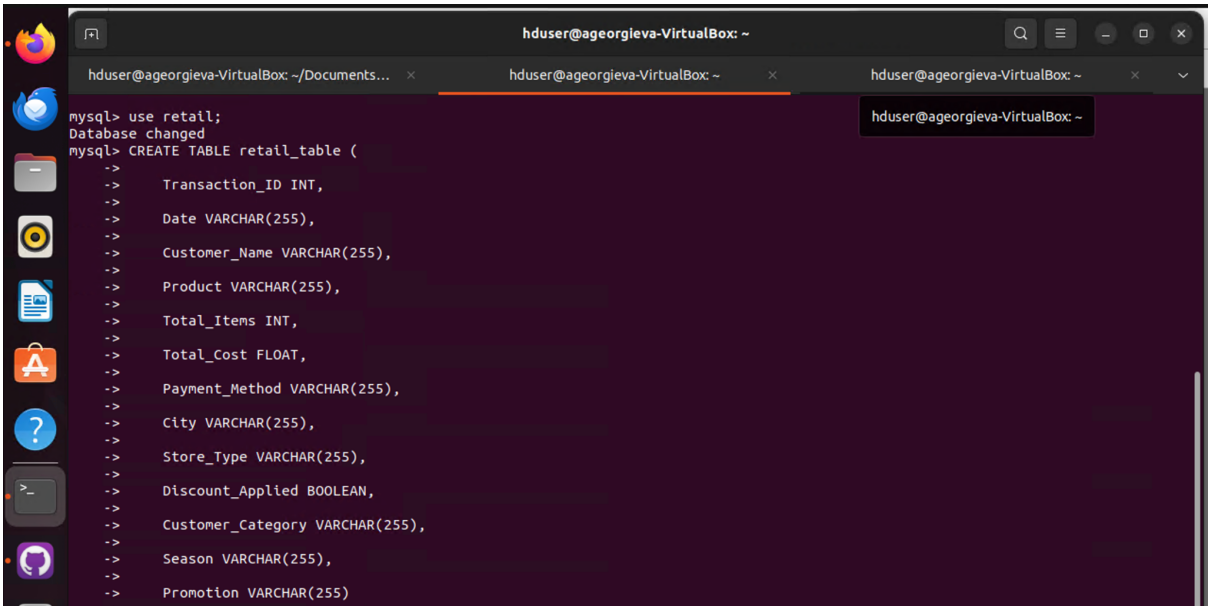


Figure 5: Start Hadoop and move files from Github to HDFS

MySQL queries

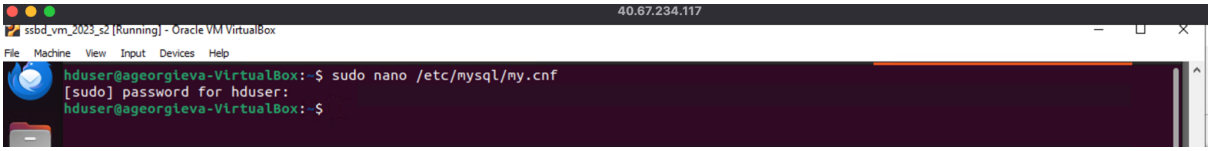


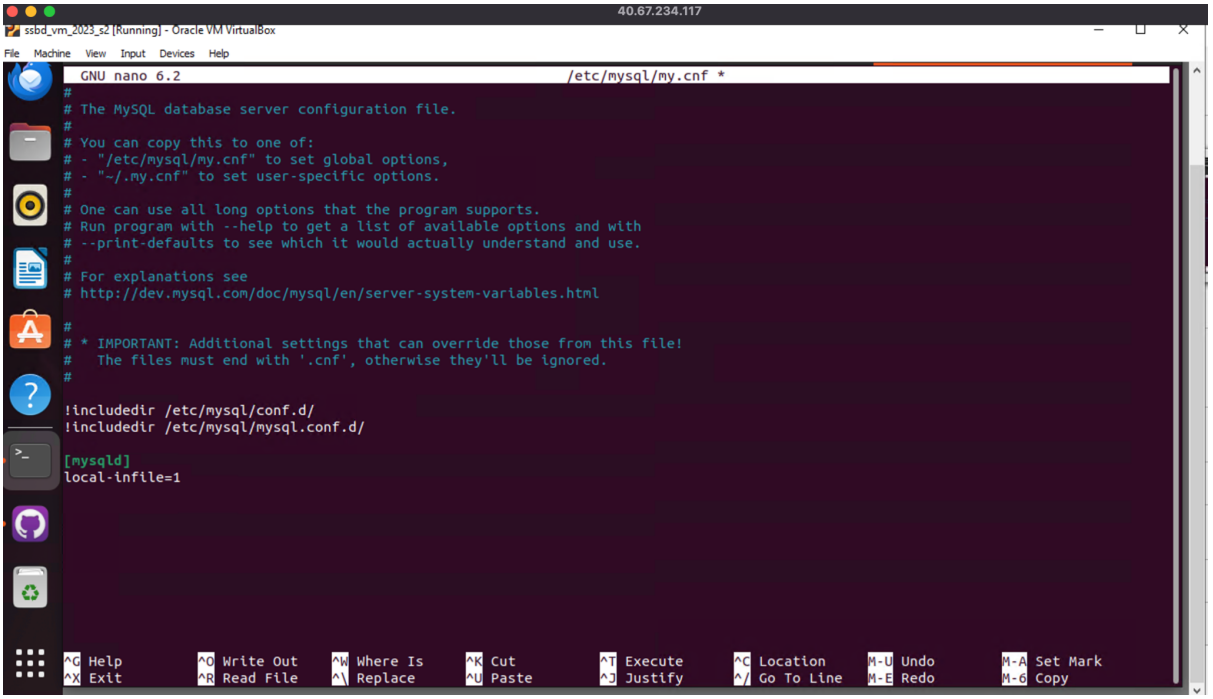

hduser@ageorgieva-VirtualBox: ~
hduser@ageorgieva-VirtualBox: ~/Documents...
hduser@ageorgieva-VirtualBox: ~
hduser@ageorgieva-VirtualBox: ~
mysql> show databases;
+--------------------+
| Database |
+--------------------+
| information_schema |
| mysql |
| performance_schema |
| retail |
| sys |
+--------------------+
5 rows in set (0.01 sec)
mysql> use retail;
Database changed
mysql> CREATE TABLE retail_table (
-> Transaction_ID INT,
-> Date VARCHAR(255),
-> Customer_Name VARCHAR(255),
-> Product VARCHAR(255),
-> Total_Items INT,
-> Total_Cost FLOAT,
-> Payment_Method VARCHAR(255),

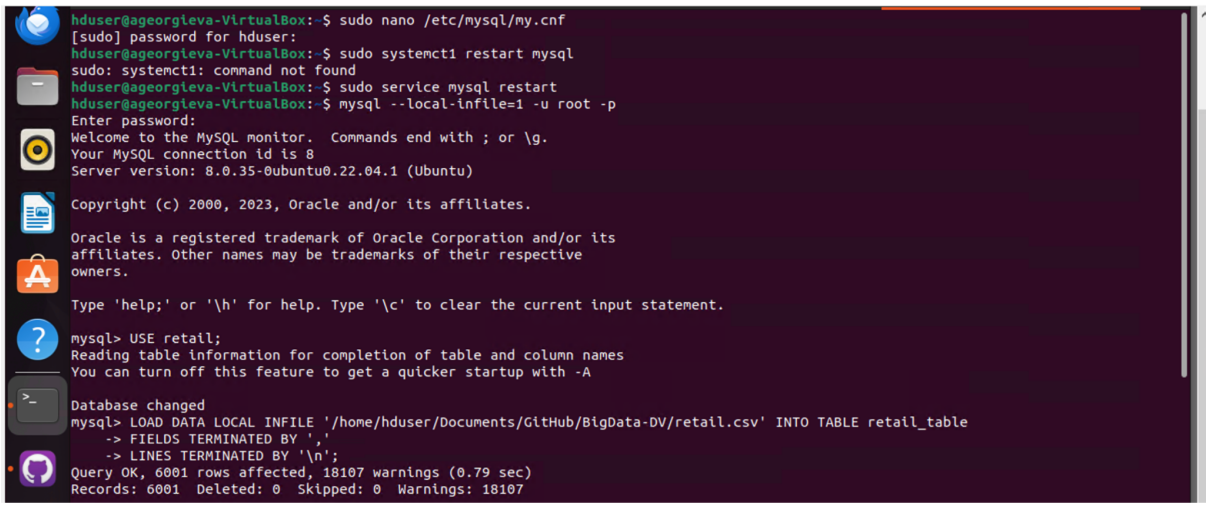

hduser@ageorgieva-VirtualBox: ~
hduser@ageorgieva-VirtualBox: ~/Documents...
hduser@ageorgieva-VirtualBox: ~
hduser@ageorgieva-VirtualBox: ~
mysql> use retail;
Database changed
mysql> CREATE TABLE retail_table (
-> Transaction_ID INT,
-> Date VARCHAR(255),
-> Customer_Name VARCHAR(255),
-> Product VARCHAR(255),
-> Total_Items INT,
-> Total_Cost FLOAT,
-> Payment_Method VARCHAR(255),
-> City VARCHAR(255),
-> Store_Type VARCHAR(255),
-> Discount_Applied BOOLEAN,
-> Customer_Category VARCHAR(255),
-> Season VARCHAR(255),
-> Promotion VARCHAR(255)

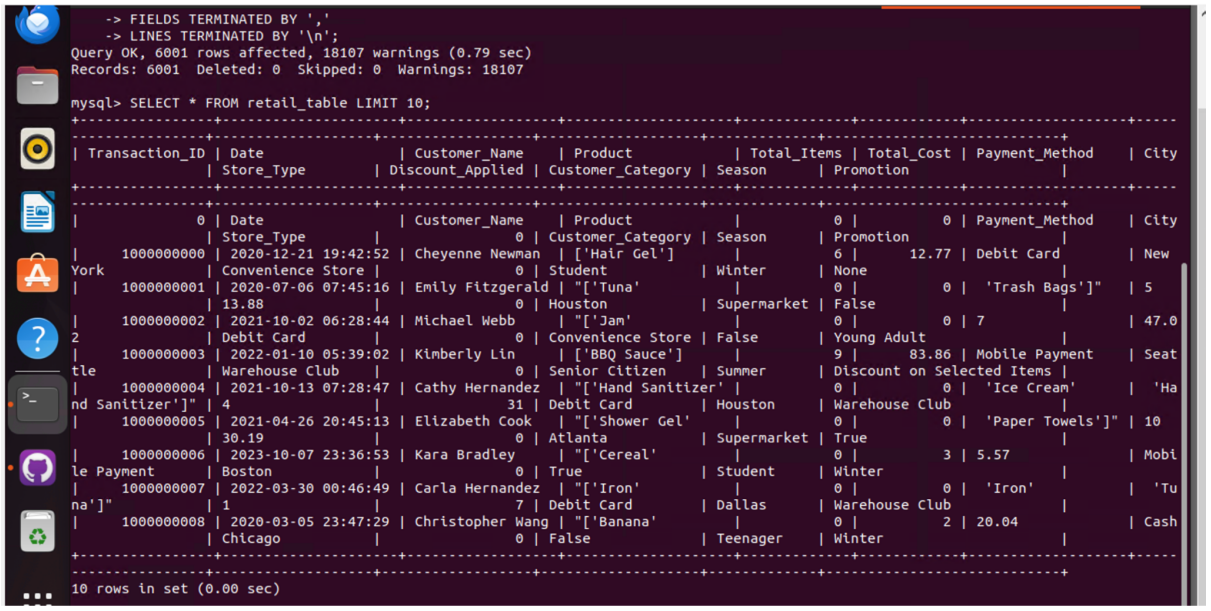

-> Season VARCHAR(255),
-> Promotion VARCHAR(255)
->);
Query OK, 0 rows affected (0.08 sec)


mysql> LOAD DATA INFILE 'hdfs://localhost:9000/user1/retail.csv'
-> INTO TABLE retail_table
-> FIELDS TERMINATED BY ',';
ERROR 1290 (HY000): The MySQL server is running with the --secure-file-priv option so it cannot execute this statement
mysql> SHOW VARIABLES LIKE 'secure_file_priv'';
'>
'> exit
'> \c
'> SHOW VARIABLES LIKE 'secure_file_priv'' \c
mysql> LOAD DATA LOCAL INPATH '/home/hduser/Documents/GitHub/BigData-DV/retail.csv' INTO TABLE retail_table;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the rig
ht syntax to use near 'INPATH '/home/hduser/Documents/GitHub/BigData-DV/retail.csv' INTO TABLE retail_t' at line 1
mysql> LOAD DATA LOCAL INFILE '/home/hduser/Documents/GitHub/BigData-DV/retail.csv' INTO TABLE retail_table
-> FIELDS TERMINATED BY ',';
ERROR 3948 (42000): Loading local data is disabled; this must be enabled on both the client and server sides
mysql>


40.67.234.117
ssbd_vm_2023_s2 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
hduser@ageorgieva-VirtualBox:~$ sudo nano /etc/mysql/my.cnf
[sudo] password for hduser:
hduser@ageorgieva-VirtualBox:~$

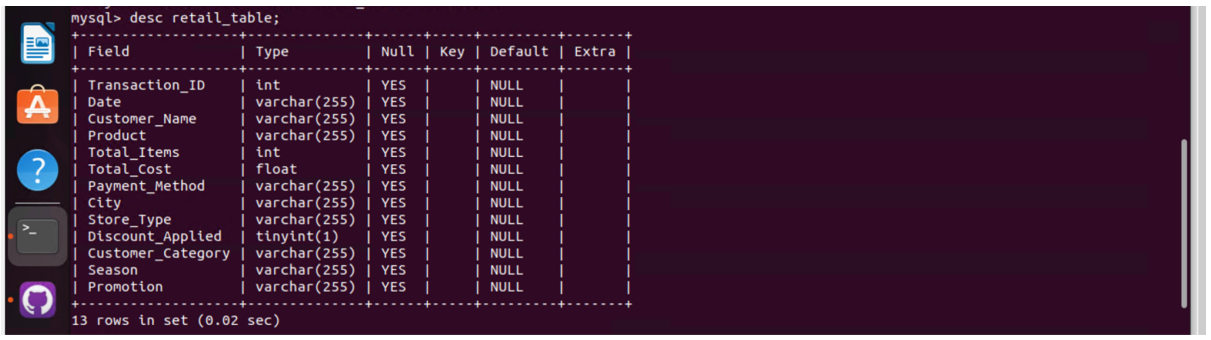

40.67.234.117
ssbd_vm_2023_s2 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
GNU nano 6.2 /etc/mysql/my.cnf *
The MySQL database server configuration file.
You can copy this to one of:
- "/etc/mysql/my.cnf" to set global options,
- "~/.my.cnf" to set user-specific options.
One can use all long options that the program supports.
Run program with --help to get a list of available options and with
--print-defaults to see which it would actually understand and use.
For explanations see
http://dev.mysql.com/doc/mysql/en/server-system-variables.html
* IMPORTANT: Additional settings that can override those from this file!
The files must end with '.cnf', otherwise they'll be ignored.
!includedir /etc/mysql/conf.d/
!includedir /etc/mysql/mysql.conf.d/
[mysqld]
local-infile=1
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location M-U Undo M-A Set Mark
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line M-E Redo M-6 Copy


hduser@ageorgieva-VirtualBox:~$ sudo nano /etc/mysql/my.cnf
[sudo] password for hduser:
hduser@ageorgieva-VirtualBox:~$ sudo systemct1 restart mysql
sudo: systemct1: command not found
hduser@ageorgieva-VirtualBox:~$ sudo service mysql restart
hduser@ageorgieva-VirtualBox:~$ mysql --local-infile=1 -u root -p
Enter password:
Welcome to the MySQL monitor. Commands end with ; or \g.
Your MySQL connection id is 8
Server version: 8.0.35-0ubuntu0.22.04.1 (Ubuntu)
Copyright (c) 2000, 2023, Oracle and/or its affiliates.
Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
mysql> USE retail;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A
Database changed
mysql> LOAD DATA LOCAL INFILE '/home/hduser/Documents/GitHub/BigData-DV/retail.csv' INTO TABLE retail_table
-> FIELDS TERMINATED BY ','
-> LINES TERMINATED BY '\n';
Query OK, 6001 rows affected, 18107 warnings (0.79 sec)
Records: 6001 Deleted: 0 Skipped: 0 Warnings: 18107

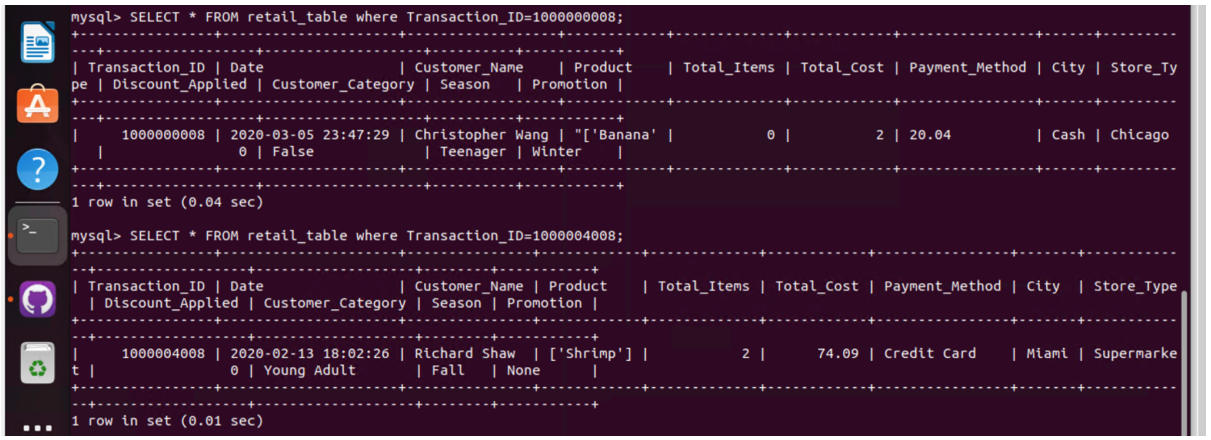


```
    -> FIELDS TERMINATED BY ','
    -> LINES TERMINATED BY '\n';
Query OK, 6001 rows affected, 18107 warnings (0.79 sec)
Records: 6001  Deleted: 0  Skipped: 0  Warnings: 18107

mysql> SELECT * FROM retail_table LIMIT 10;
+----------------+---------------------+-----------------+------------------+-------------+-----------+--------------------+-----
---------------+---------------------+-----------------+------------------+-------------+-----------+--------------------+-----
| Transaction_ID | Date                | Customer_Name   | Product          | Total_Items | Total_Cost | Payment_Method     | City
               | Store_Type          | Discount_Applied | Customer_Category | Season      | Promotion                |
+----------------+---------------------+-----------------+------------------+-------------+-----------+--------------------+-----
---------------+---------------------+-----------------+------------------+-------------+-----------+--------------------+-----
|              0 | Date                | Customer_Name   | Product          |           0 |         0 | Payment_Method     | City
               | Store_Type          |               0 | Customer_Category | Season      | Promotion                |
|     1000000000 | 2020-12-21 19:42:52 | Cheyenne Newman | ['Hair Gel']     |           6 |     12.77 | Debit Card         | New
York           | Convenience Store   |               0 | Student          | Winter      | None                     |
|     1000000001 | 2020-07-06 07:45:16 | Emily Fitzgerald | "['Tuna'        |           0 |         0 |  'Trash Bags']"    | 5
               | 13.88               |               0 | Houston          | Supermarket | False                    |
|     1000000002 | 2021-10-02 06:28:44 | Michael Webb    | "['Jam'          |           0 |         0 | 7                  | 47.0
2              | Debit Card          |               0 | Convenience Store | False       | Young Adult              |
|     1000000003 | 2022-01-10 05:39:02 | Kimberly Lin    | ['BBQ Sauce']    |           9 |     83.86 | Mobile Payment     | Seat
tle            | Warehouse Club      |               0 | Senior Citizen   | Summer      | Discount on Selected Items |
|     1000000004 | 2021-10-13 07:28:47 | Cathy Hernandez | "['Hand Sanitizer' |           0 |         0 |  'Ice Cream'       |  'Ha
nd Sanitizer']" | 4                   |              31 | Debit Card       | Houston     | Warehouse Club           |
|     1000000005 | 2021-04-26 20:45:13 | Elizabeth Cook  | "['Shower Gel'   |           0 |         0 |  'Paper Towels']"  | 10
               | 30.19               |               0 | Atlanta          | Supermarket | True                     |
|     1000000006 | 2023-10-07 23:36:53 | Kara Bradley    | "['Cereal'       |           0 |         3 | 5.57               | Mobi
le Payment     | Boston              |               0 | True             | Student     | Winter                   |
|     1000000007 | 2022-03-30 00:46:49 | Carla Hernandez | "['Iron'         |           0 |         0 |  'Iron'            |  'Tu
na']"          | 1                   |               7 | Debit Card       | Dallas      | Warehouse Club           |
|     1000000008 | 2020-03-05 23:47:29 | Christopher Wang | "['Banana'      |           0 |         2 | 20.04              | Cash
               | Chicago             |               0 | False            | Winter      |                          |
+----------------+---------------------+-----------------+------------------+-------------+-----------+--------------------+-----
---------------+---------------------+-----------------+------------------+-------------+-----------+--------------------+-----
10 rows in set (0.00 sec)
```



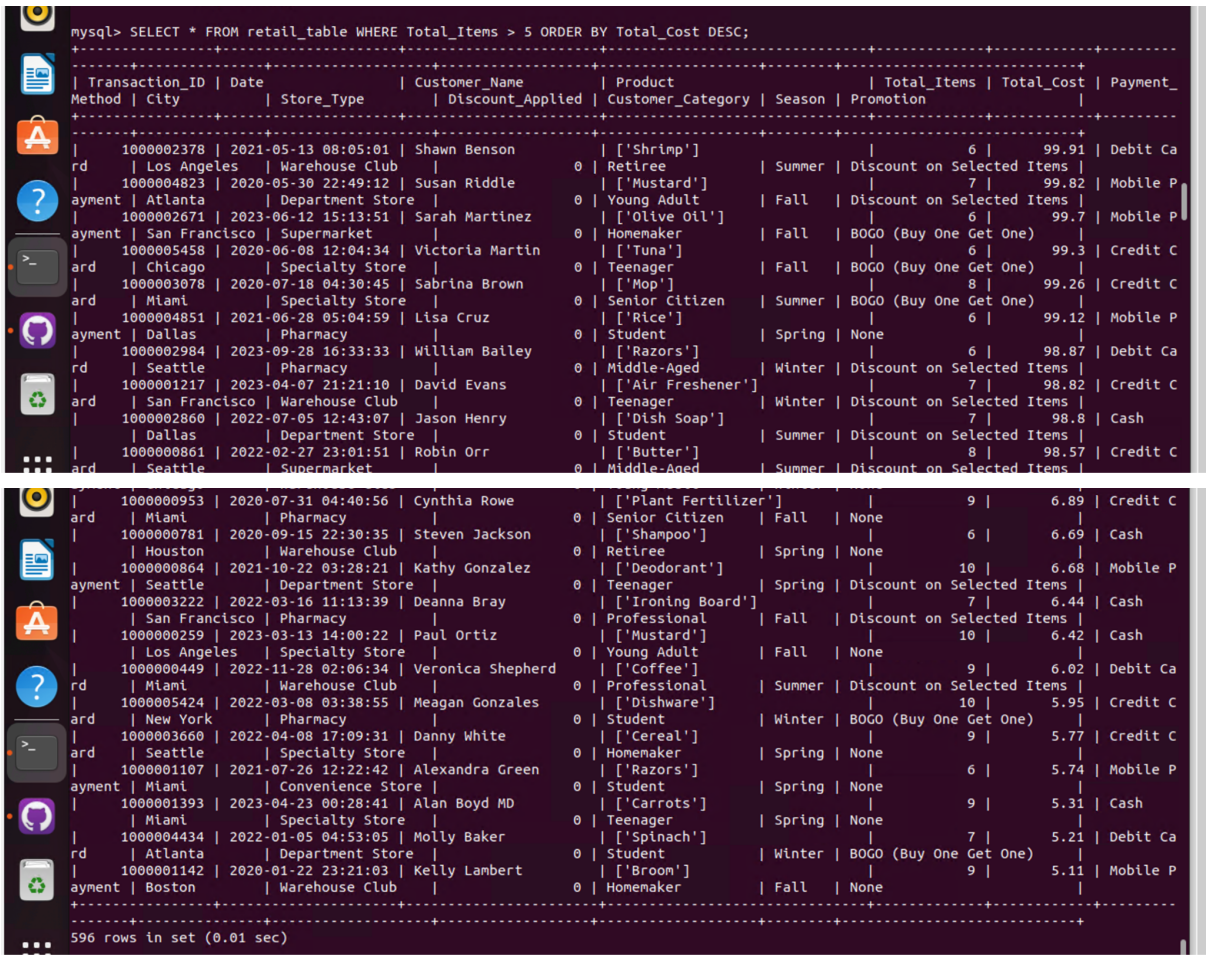
```
mysql> desc retail_table;
+------------------+--------------+------+-----+---------+-------+
| Field            | Type         | Null | Key | Default | Extra |
+------------------+--------------+------+-----+---------+-------+
| Transaction_ID   | int          | YES  |     | NULL    |       |
| Date             | varchar(255) | YES  |     | NULL    |       |
| Customer_Name    | varchar(255) | YES  |     | NULL    |       |
| Product          | varchar(255) | YES  |     | NULL    |       |
| Total_Items      | int          | YES  |     | NULL    |       |
| Total_Cost       | float        | YES  |     | NULL    |       |
| Payment_Method   | varchar(255) | YES  |     | NULL    |       |
| City             | varchar(255) | YES  |     | NULL    |       |
| Store_Type       | varchar(255) | YES  |     | NULL    |       |
| Discount_Applied | tinyint(1)   | YES  |     | NULL    |       |
| Customer_Category | varchar(255) | YES  |     | NULL    |       |
| Season           | varchar(255) | YES  |     | NULL    |       |
| Promotion        | varchar(255) | YES  |     | NULL    |       |
+------------------+--------------+------+-----+---------+-------+
13 rows in set (0.02 sec)
```



```
mysql> SELECT * FROM retail_table where Transaction_ID=1000000008;
+----------------+---------------------+-----------------+----------+-------------+------------+----------------+------+---------
---+-----------------+-----------------+----------+-----------+
| Transaction_ID | Date                | Customer_Name   | Product  | Total_Items | Total_Cost | Payment_Method | City | Store_Ty
pe | Discount_Applied | Customer_Category | Season   | Promotion |
+----------------+---------------------+-----------------+----------+-------------+------------+----------------+------+---------
---+-----------------+-----------------+----------+-----------+
|     1000000008 | 2020-03-05 23:47:29 | Christopher Wang | "['Banana' |           0 |          2 | 20.04          | Cash | Chicago
  |               0 | False             | Teenager | Winter    |
+----------------+---------------------+-----------------+----------+-------------+------------+----------------+------+---------
---+-----------------+-----------------+----------+-----------+
1 row in set (0.04 sec)

mysql> SELECT * FROM retail_table where Transaction_ID=1000004008;
+----------------+---------------------+---------------+---------+-------------+------------+----------------+-------+----------
--+-----------------+-----------------+--------+-----------+
| Transaction_ID | Date                | Customer_Name | Product | Total_Items | Total_Cost | Payment_Method | City  | Store_Type
  | Discount_Applied | Customer_Category | Season | Promotion |
+----------------+---------------------+---------------+---------+-------------+------------+----------------+-------+----------
--+-----------------+-----------------+--------+-----------+
|     1000004008 | 2020-02-13 18:02:26 | Richard Shaw  | ['Shrimp'] |           2 |      74.09 | Credit Card    | Miami | Supermarke
t |               0 | Young Adult       | Fall   | None      |
+----------------+---------------------+---------------+---------+-------------+------------+----------------+-------+----------
--+-----------------+-----------------+--------+-----------+
1 row in set (0.01 sec)
```

```
mysql> SELECT City, SUM(Total_Cost) FROM retail_table GROUP BY City;
+------------------------------+--------------------+
| City                         | SUM(Total_Cost)    |
+------------------------------+--------------------+
| City                         |                  0 |
| New York                     |  7141.990013122559 |
| 5                            |                  0 |
| 47.02                        |                  0 |
| Seattle                      |  6782.249997138977 |
|  'Hand Sanitizer']"          |                  0 |
| 10                           |                  0 |
| Mobile Payment               |               1590 |
|  'Tuna']"                    |                  0 |
| Cash                         |               1761 |
| 88.79                        |                  0 |
| 9                            |                  0 |
| 21.29                        |                  0 |
| 2                            |                  0 |
| Credit Card                  |               1782 |
| Los Angeles                  | 6598.1100125312805 |
```

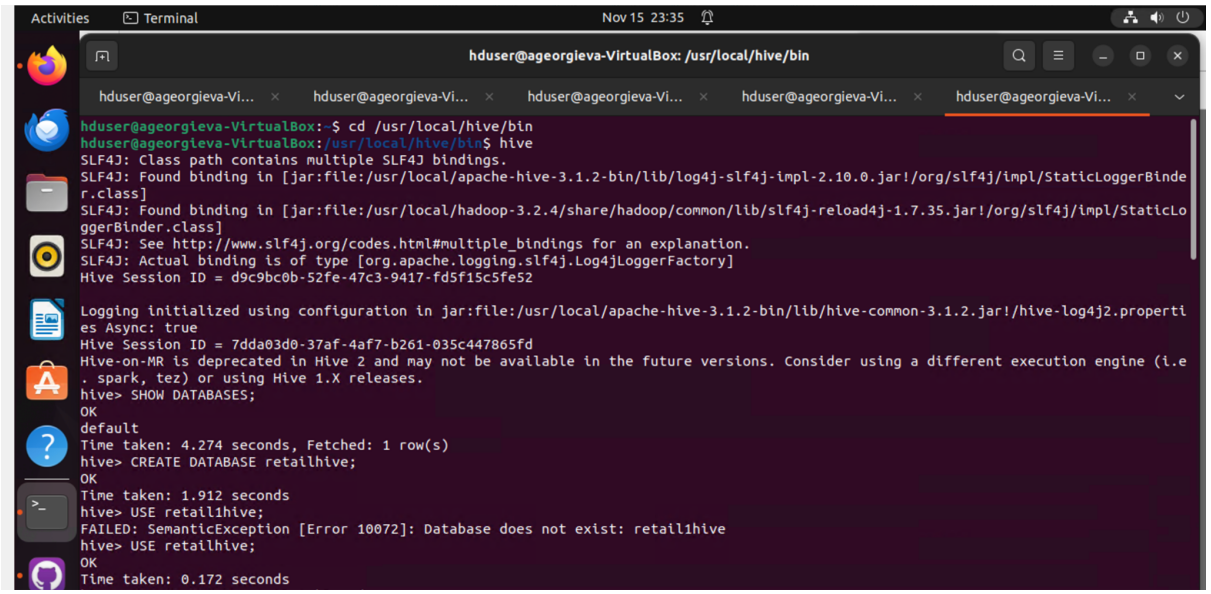


```
mysql> SELECT * FROM retail_table WHERE Total_Items > 5 ORDER BY Total_Cost DESC;
+----------------+---------------------+---------------------+---------------------------+------------+------------+---------
-------+---------------+-----------------+---------------------+-----------------+--------+-------------------------+
| Transaction_ID | Date                | Customer_Name       | Product                   | Total_Items | Total_Cost | Payment_
Method | City          | Store_Type      | Discount_Applied | Customer_Category | Season | Promotion               |
+----------------+---------------------+---------------------+---------------------------+------------+------------+---------
-------+---------------+-----------------+---------------------+-----------------+--------+-------------------------+
|     1000002378 | 2021-05-13 08:05:01 | Shawn Benson        | ['Shrimp']                |          6 |      99.91 | Debit Ca
rd     | Los Angeles   | Warehouse Club  |                   0 | Retiree           | Summer | Discount on Selected Items |
|     1000004823 | 2020-05-30 22:49:12 | Susan Riddle        | ['Mustard']               |          7 |      99.82 | Mobile P
ayment | Atlanta       | Department Store |                   0 | Young Adult       | Fall   | Discount on Selected Items |
|     1000002671 | 2023-06-12 15:13:51 | Sarah Martinez      | ['Olive Oil']             |          6 |       99.7 | Mobile P
ayment | San Francisco | Supermarket     |                   0 | Homemaker         | Fall   | BOGO (Buy One Get One)  |
|     1000005458 | 2020-06-08 12:04:34 | Victoria Martin     | ['Tuna']                  |          6 |       99.3 | Credit C
ard    | Chicago       | Specialty Store |                   0 | Teenager          | Fall   | BOGO (Buy One Get One)  |
|     1000003078 | 2020-07-18 04:30:45 | Sabrina Brown       | ['Mop']                   |          8 |      99.26 | Credit C
ard    | Miami         | Specialty Store |                   0 | Senior Citizen    | Summer | BOGO (Buy One Get One)  |
|     1000004851 | 2021-06-28 05:04:59 | Lisa Cruz           | ['Rice']                  |          6 |      99.12 | Mobile P
ayment | Dallas        | Pharmacy        |                   0 | Student           | Spring | None                    |
|     1000002984 | 2023-09-28 16:33:33 | William Bailey      | ['Razors']                |          6 |      98.87 | Debit Ca
rd     | Seattle       | Pharmacy        |                   0 | Middle-Aged       | Winter | Discount on Selected Items |
|     1000001217 | 2023-04-07 21:21:10 | David Evans         | ['Air Freshener']         |          7 |      98.82 | Credit C
ard    | San Francisco | Warehouse Club  |                   0 | Teenager          | Winter | Discount on Selected Items |
|     1000002860 | 2022-07-05 12:43:07 | Jason Henry         | ['Dish Soap']             |          7 |       98.8 | Cash
       | Dallas        | Department Store |                   0 | Student           | Summer | Discount on Selected Items |
|     1000000861 | 2022-02-27 23:01:51 | Robin Orr           | ['Butter']                |          8 |      98.57 | Credit C
ard    | Seattle       | Supermarket     |                   0 | Middle-Aged       | Summer | Discount on Selected Items |
```



```
|     1000000953 | 2020-07-31 04:40:56 | Cynthia Rowe        | ['Plant Fertilizer']      |          9 |       6.89 | Credit C
ard    | Miami         | Pharmacy        |                   0 | Senior Citizen    | Fall   | None                    |
|     1000000781 | 2020-09-15 22:30:35 | Steven Jackson      | ['Shampoo']               |          6 |       6.69 | Cash
       | Houston       | Warehouse Club  |                   0 | Retiree           | Spring | None                    |
|     1000000864 | 2021-10-22 03:28:21 | Kathy Gonzalez      | ['Deodorant']             |         10 |       6.68 | Mobile P
ayment | Seattle       | Department Store |                   0 | Teenager          | Spring | Discount on Selected Items |
|     1000003222 | 2022-03-16 11:13:39 | Deanna Bray         | ['Ironing Board']         |          7 |       6.44 | Cash
       | San Francisco | Pharmacy        |                   0 | Professional      | Fall   | Discount on Selected Items |
|     1000000259 | 2023-03-13 14:00:22 | Paul Ortiz          | ['Mustard']               |         10 |       6.42 | Cash
       | Los Angeles   | Specialty Store |                   0 | Young Adult       | Fall   | None                    |
|     1000000449 | 2022-11-28 02:06:34 | Veronica Shepherd   | ['Coffee']                |          9 |       6.02 | Debit Ca
rd     | Miami         | Warehouse Club  |                   0 | Professional      | Summer | Discount on Selected Items |
|     1000005424 | 2022-03-08 03:38:55 | Meagan Gonzales     | ['Dishware']              |         10 |       5.95 | Credit C
ard    | New York      | Pharmacy        |                   0 | Student           | Winter | BOGO (Buy One Get One)  |
|     1000003660 | 2022-04-08 17:09:31 | Danny White         | ['Cereal']                |          9 |       5.77 | Credit C
ard    | Seattle       | Specialty Store |                   0 | Homemaker         | Spring | None                    |
|     1000001107 | 2021-07-26 12:22:42 | Alexandra Green     | ['Razors']                |          6 |       5.74 | Mobile P
ayment | Miami         | Convenience Store |                   0 | Student           | Spring | None                    |
|     1000001393 | 2023-04-23 00:28:41 | Alan Boyd MD        | ['Carrots']               |          9 |       5.31 | Cash
       | Miami         | Specialty Store |                   0 | Teenager          | Spring | None                    |
|     1000004434 | 2022-01-05 04:53:05 | Molly Baker         | ['Spinach']               |          7 |       5.21 | Debit Ca
rd     | Atlanta       | Department Store |                   0 | Student           | Winter | BOGO (Buy One Get One)  |
|     1000001142 | 2020-01-22 23:21:03 | Kelly Lambert       | ['Broom']                 |          9 |       5.11 | Mobile P
ayment | Boston        | Warehouse Club  |                   0 | Homemaker         | Fall   | None                    |
+----------------+---------------------+---------------------+---------------------------+------------+------------+---------
-------+---------------+-----------------+---------------------+-----------------+--------+-------------------------+
596 rows in set (0.01 sec)
```

HIVE queries



```
Activities  Terminal  Nov 15 23:35
hduser@ageorgieva-VirtualBox: /usr/local/hive/bin
hduser@ageorgieva-Vi...  hduser@ageorgieva-Vi...  hduser@ageorgieva-Vi...  hduser@ageorgieva-Vi...  hduser@ageorgieva-Vi...
hduser@ageorgieva-VirtualBox:~$ cd /usr/local/hive/bin
hduser@ageorgieva-VirtualBox:/usr/local/hive/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinde
r.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-3.2.4/share/hadoop/common/lib/slf4j-reload4j-1.7.35.jar!/org/slf4j/impl/StaticLo
ggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = d9c9bc0b-52fe-47c3-9417-fd5f15c5fe52

Logging initialized using configuration in jar:file:/usr/local/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properti
es Async: true
Hive Session ID = 7dda03d0-37af-4af7-b261-035c447865fd
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e
. spark, tez) or using Hive 1.X releases.
hive> SHOW DATABASES;
OK
default
Time taken: 4.274 seconds, Fetched: 1 row(s)
hive> CREATE DATABASE retailhive;
OK
Time taken: 1.912 seconds
hive> USE retail1hive;
FAILED: SemanticException [Error 10072]: Database does not exist: retail1hive
hive> USE retailhive;
OK
Time taken: 0.172 seconds
```

```
hive> CREATE TABLE retailhive (
    > Transaction_ID INT,
    > `Date` STRING,
    > Customer_name STRING,
    > Product STRING,
    > Total_Items INT,
    > Total_Cost FLOAT,
    > Payment_Methods STRING,
    > City STRING,
    > Store_Type STRING,
    > Discount_Applied BOOLEAN,
    > Customer_Category STRING,
    > Season STRING,
    > Promotion STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 6.476 seconds
hive>
```

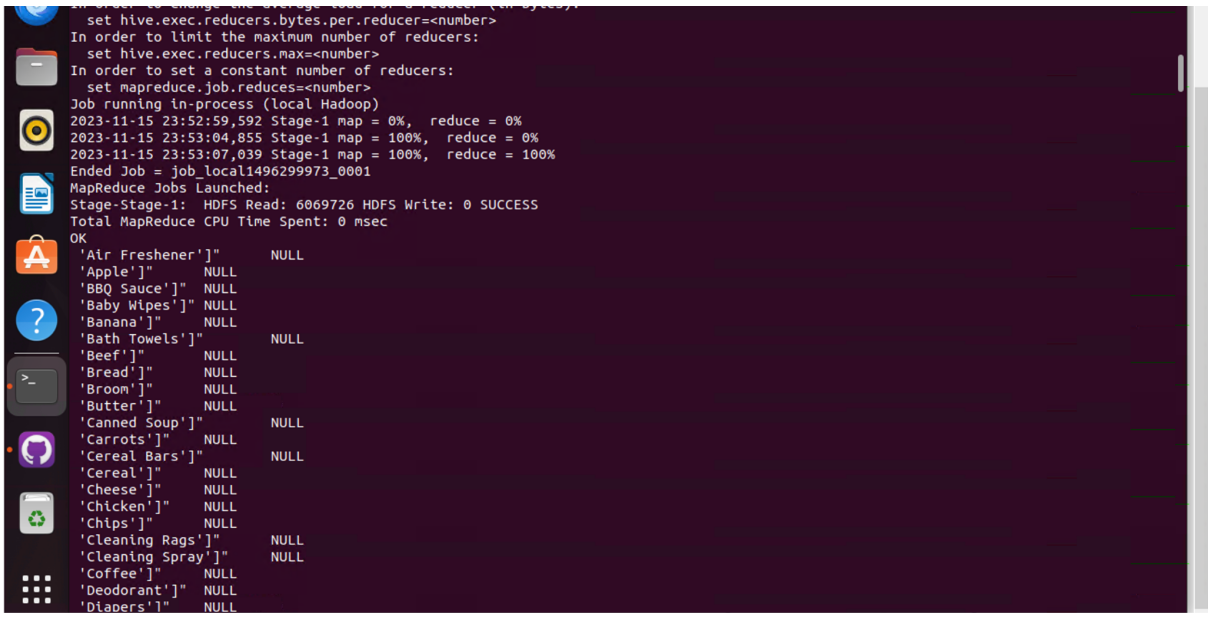
40.67.234.117

n_2023_s2 [Running] - Oracle VM VirtualBox

e View Input Devices Help

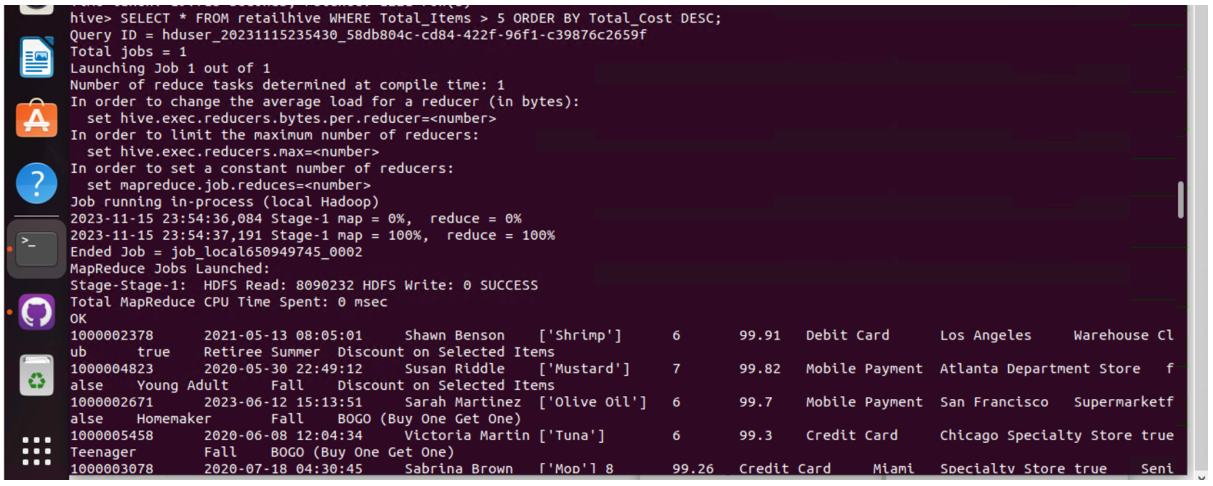
```
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ',';
OK
Time taken: 6.476 seconds
hive> LOAD DATA INPATH 'hdfs://localhost:9000/user1/retail.csv' INTO TABLE retailhive;
Loading data to table retailhive.retailhive
OK
Time taken: 8.219 seconds
hive> SHOW TABLES;
OK
retailhive
Time taken: 0.372 seconds, Fetched: 1 row(s)
hive> SELECT * FROM retailhive LIMIT 10;
OK
NULL    Date    Customer_Name   Product NULL    NULL    Payment_Method  City    Store_Type      NULL    Customer_Category       Seas
on      Promotion
1000000000      2020-12-21 19:42:52     Cheyenne Newman ['Hair Gel']    6       12.77   Debit Card      New York        Convenience
Store   true    Student Winter  None
1000000001      2020-07-06 07:45:16     Emily Fitzgerald        "['Tuna'        NULL    NULL     'Trash Bags']" 5       13.88   NULL
Houston Supermarket     False
1000000002      2021-10-02 06:28:44     Michael Webb    "['Jam' NULL    NULL    7       47.02   Debit Card      NULL    Convenience
Store   False   Young Adult
1000000003      2022-01-10 05:39:02     Kimberly Lin    ['BBQ Sauce']   9       83.86   Mobile Payment  Seattle Warehouse Club  true
Senior Citizen  Summer  Discount on Selected Items
1000000004      2021-10-13 07:28:47     Cathy Hernandez "['Hand Sanitizer'      NULL    NULL     'Ice Cream'     'Hand Sanitizer']"4
NULL    Debit Card      Houston Warehouse Club
1000000005      2021-04-26 20:45:13     Elizabeth Cook  "['Shower Gel'  NULL    NULL     'Paper Towels']"       10      30.19   NULL
Atlanta Supermarket     True
1000000006      2023-10-07 23:36:53     Kara Bradley    "['Cereal'      NULL    3.0     5.57    Mobile Payment  Boston  NULL    True
Student Winter
1000000007      2022-03-30 00:46:49     Carla Hernandez "['Iron'        NULL    NULL     'Iron'  'Tuna']"       1       NULL    Debi
t Card  Dallas  Warehouse Club
1000000008      2020-03-05 23:47:29     Christopher Wang        "['Banana'      NULL    2.0     20.04   Cash    Chicago NULL    Fals
e       Teenager        Winter
Time taken: 18.752 seconds, Fetched: 10 row(s)
hive> DESCRIBE retailhive;
```

```
hive> DESCRIBE retailhive;
OK
transaction_id          int
date                    string
customer_name           string
product                 string
total_items             int
total_cost              float
payment_methods         string
city                    string
store_type              string
discount_applied        boolean
customer_category       string
season                  string
promotion               string
Time taken: 0.767 seconds, Fetched: 13 row(s)
hive> SELECT * FROM retail_table where Transaction_ID=1000000008;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'retail_table'
hive> SELECT * FROM retailhive where Transaction_ID=1000000008;
OK
1000000008      2020-03-05 23:47:29     Christopher Wang        "['Banana'      NULL    2.0     20.04   Cash    Chicago NULL    Fals
e       Teenager        Winter
Time taken: 6.25 seconds, Fetched: 1 row(s)
hive> SELECT * FROM retail_table where Transaction_ID=1000004008;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'retail_table'
hive> SELECT * FROM retailhive where Transaction_ID=1000004008;
OK
1000004008      2020-02-13 18:02:26     Richard Shaw    ['Shrimp']      2       74.09   Credit Card     Miami   Supermarket     true
Young Adult     Fall    None
Time taken: 1.596 seconds, Fetched: 1 row(s)
hive> SELECT City, SUM(Total_Cost) FROM retailhive GROUP BY City;
Query ID = hduser_20231115235247_1291140c-b00b-4420-8f4a-682e67a2942f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
```

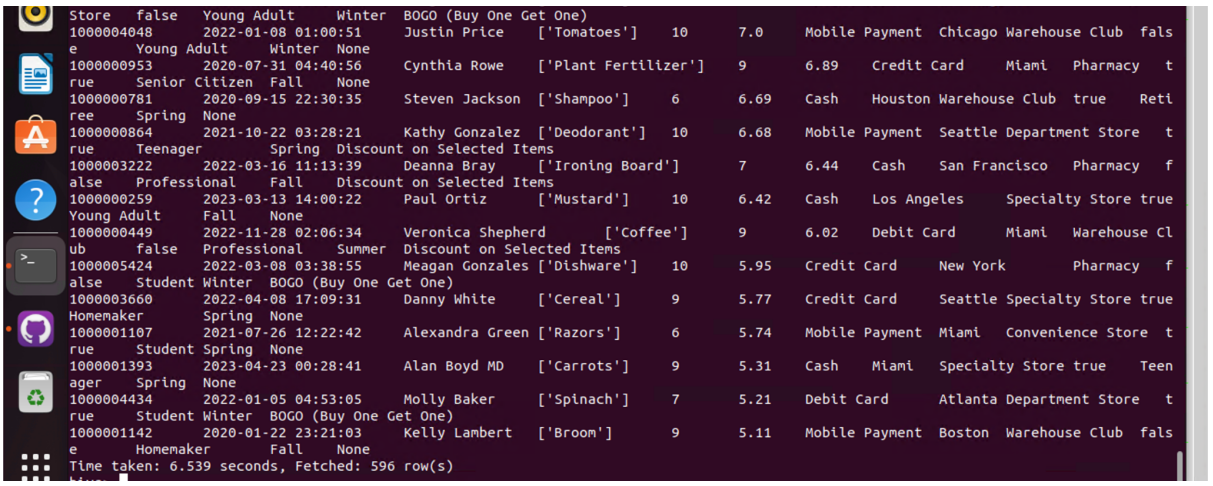


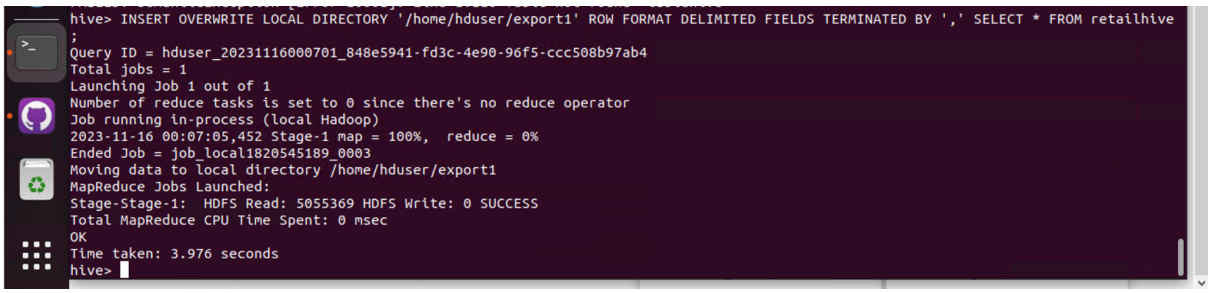
```
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-15 23:52:59,592 Stage-1 map = 0%,  reduce = 0%
2023-11-15 23:53:04,855 Stage-1 map = 100%,  reduce = 0%
2023-11-15 23:53:07,039 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1496299973_0001
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 6069726 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
 'Air Freshener']"      NULL
 'Apple']"      NULL
 'BBQ Sauce']"  NULL
 'Baby Wipes']" NULL
 'Banana']"     NULL
 'Bath Towels']"        NULL
 'Beef']"       NULL
 'Bread']"      NULL
 'Broom']"      NULL
 'Butter']"     NULL
 'Canned Soup']"        NULL
 'Carrots']"    NULL
 'Cereal Bars']"        NULL
 'Cereal']"     NULL
 'Cheese']"     NULL
 'Chicken']"    NULL
 'Chips']"      NULL
 'Cleaning Rags']"      NULL
 'Cleaning Spray']"     NULL
 'Coffee']"     NULL
 'Deodorant']"  NULL
 'Diapers']"    NULL
```

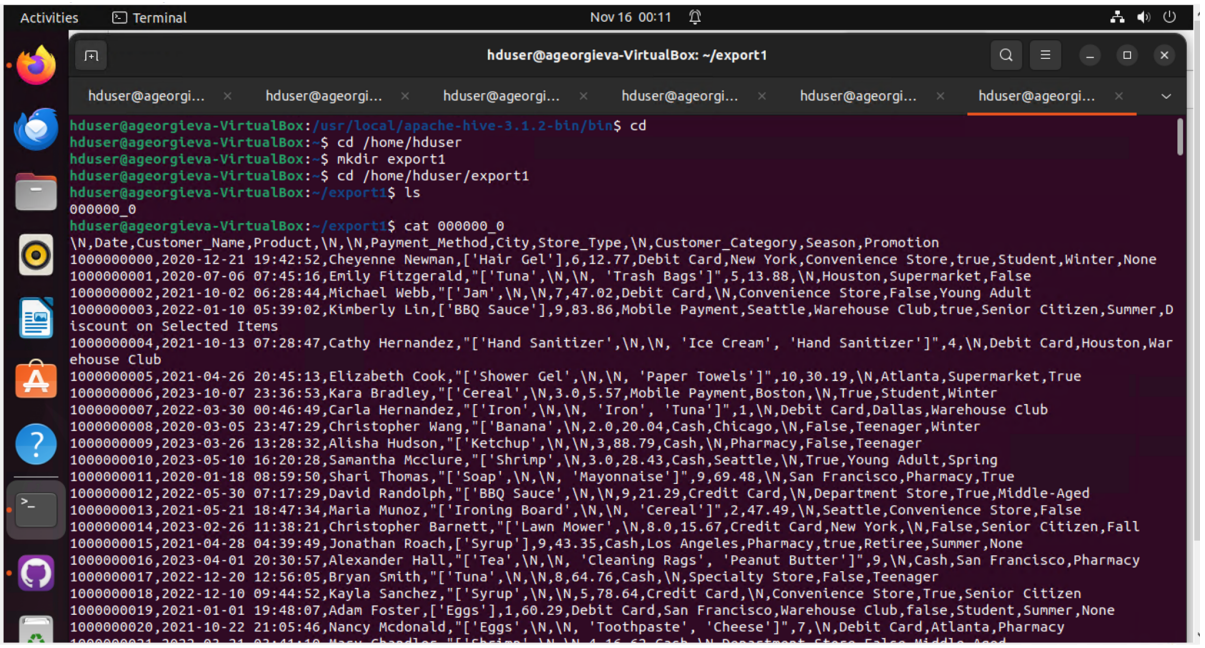
```
99.16   NULL
99.26   NULL
99.31   NULL
99.39   NULL
99.47   NULL
99.53   NULL
99.59   NULL
99.61   NULL
99.68   NULL
99.71   NULL
99.81   NULL
99.82   NULL
Atlanta 5461.569971084595
Boston  6435.270000934601
Cash    1761.0
Chicago 4653.139994621277
City    NULL
Credit Card     1782.0
Dallas  7198.820017337799
Debit Card      1691.0
Houston 5674.839991569519
Los Angeles     6598.1100125312805
Miami   5493.159993648529
Mobile Payment  1590.0
New York        7141.990013122559
San Francisco   5665.0199937820435
Seattle 6782.249997138977
Time taken: 19.715 seconds, Fetched: 1221 row(s)
```

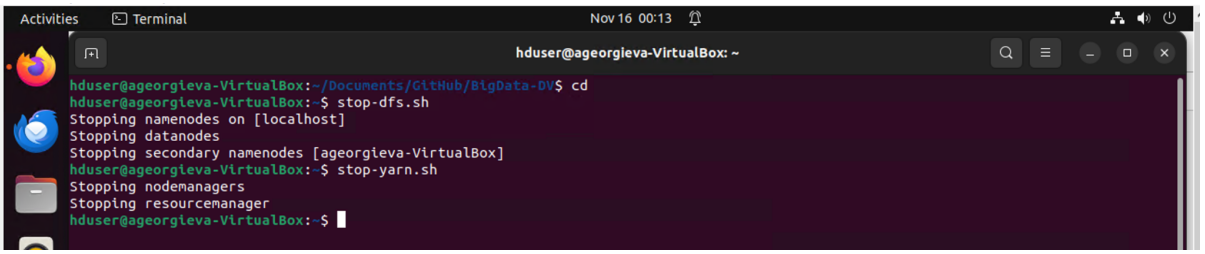


```
hive> SELECT * FROM retailhive WHERE Total_Items > 5 ORDER BY Total_Cost DESC;
Query ID = hduser_20231115235430_58db804c-cd84-422f-96f1-c39876c2659f
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2023-11-15 23:54:36,084 Stage-1 map = 0%,  reduce = 0%
2023-11-15 23:54:37,191 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local650949745_0002
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 8090232 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
1000002378      2021-05-13 08:05:01     Shawn Benson    ['Shrimp']      6       99.91   Debit Card      Los Angeles     Warehouse Cl
ub      true    Retiree Summer  Discount on Selected Items
1000004823      2020-05-30 22:49:12     Susan Riddle    ['Mustard']     7       99.82   Mobile Payment  Atlanta Department Store    f
alse    Young Adult     Fall    Discount on Selected Items
1000002671      2023-06-12 15:13:51     Sarah Martinez  ['Olive Oil']   6       99.7    Mobile Payment  San Francisco   Supermarketf
alse    Homemaker       Fall    BOGO (Buy One Get One)
1000005458      2020-06-08 12:04:34     Victoria Martin ['Tuna']        6       99.3    Credit Card     Chicago Specialty Store true
Teenager        Fall    BOGO (Buy One Get One)
1000003078      2020-07-18 04:30:45     Sabrina Brown   ['Mop'] 8       99.26   Credit Card     Miami   Specialty Store true    Seni
```


Store false Young Adult Winter BOGO (Buy One Get One)
1000004048 2022-01-08 01:00:51 Justin Price ['Tomatoes'] 10 7.0 Mobile Payment Chicago Warehouse Club fals
e Young Adult Winter None
1000000953 2020-07-31 04:40:56 Cynthia Rowe ['Plant Fertilizer'] 9 6.89 Credit Card Miami Pharmacy t
rue Senior Citizen Fall None
1000000781 2020-09-15 22:30:35 Steven Jackson ['Shampoo'] 6 6.69 Cash Houston Warehouse Club true Reti
ree Spring None
1000000864 2021-10-22 03:28:21 Kathy Gonzalez ['Deodorant'] 10 6.68 Mobile Payment Seattle Department Store t
rue Teenager Spring Discount on Selected Items
1000003222 2022-03-16 11:13:39 Deanna Bray ['Ironing Board'] 7 6.44 Cash San Francisco Pharmacy f
alse Professional Fall Discount on Selected Items
1000000259 2023-03-13 14:00:22 Paul Ortiz ['Mustard'] 10 6.42 Cash Los Angeles Specialty Store true
Young Adult Fall None
1000000449 2022-11-28 02:06:34 Veronica Shepherd ['Coffee'] 9 6.02 Debit Card Miami Warehouse Cl
ub false Professional Summer Discount on Selected Items
1000005424 2022-03-08 03:38:55 Meagan Gonzales ['Dishware'] 10 5.95 Credit Card New York Pharmacy f
alse Student Winter BOGO (Buy One Get One)
1000003660 2022-04-08 17:09:31 Danny White ['Cereal'] 9 5.77 Credit Card Seattle Specialty Store true
Homemaker Spring None
1000001107 2021-07-26 12:22:42 Alexandra Green ['Razors'] 6 5.74 Mobile Payment Miami Convenience Store t
rue Student Spring None
1000001393 2023-04-23 00:28:41 Alan Boyd MD ['Carrots'] 9 5.31 Cash Miami Specialty Store true Teen
ager Spring None
1000004434 2022-01-05 04:53:05 Molly Baker ['Spinach'] 7 5.21 Debit Card Atlanta Department Store t
rue Student Winter BOGO (Buy One Get One)
1000001142 2020-01-22 23:21:03 Kelly Lambert ['Broom'] 9 5.11 Mobile Payment Boston Warehouse Club fals
e Homemaker Fall None
Time taken: 6.539 seconds, Fetched: 596 row(s)


hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/hduser/export1' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' SELECT * FROM retailhive
;
Query ID = hduser_20231116000701_848e5941-fd3c-4e90-96f5-ccc508b97ab4
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2023-11-16 00:07:05,452 Stage-1 map = 100%, reduce = 0%
Ended Job = job_local1820545189_0003
Moving data to local directory /home/hduser/export1
MapReduce Jobs Launched:
Stage-Stage-1: HDFS Read: 5055369 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 3.976 seconds
hive>


Activities
Terminal
Nov 16 00:11
hduser@ageorgieva-VirtualBox: ~/export1
hduser@ageorgieva-VirtualBox:/usr/local/apache-hive-3.1.2-bin/bin$ cd
hduser@ageorgieva-VirtualBox:~$ cd /home/hduser
hduser@ageorgieva-VirtualBox:~$ mkdir export1
hduser@ageorgieva-VirtualBox:~$ cd /home/hduser/export1
hduser@ageorgieva-VirtualBox:~/export1$ ls
000000_0
hduser@ageorgieva-VirtualBox:~/export1$ cat 000000_0
\N,Date,Customer_Name,Product,\N,\N,Payment_Method,City,Store_Type,\N,Customer_Category,Season,Promotion
1000000000,2020-12-21 19:42:52,Cheyenne Newman,['Hair Gel'],6,12.77,Debit Card,New York,Convenience Store,true,Student,Winter,None
1000000001,2020-07-06 07:45:16,Emily Fitzgerald,"['Tuna',\N,\N, 'Trash Bags']",5,13.88,\N,Houston,Supermarket,False
1000000002,2021-10-02 06:28:44,Michael Webb,"['Jam',\N,\N,7,47.02,Debit Card,\N,Convenience Store,False,Young Adult
1000000003,2022-01-10 05:39:02,Kimberly Lin,['BBQ Sauce'],9,83.86,Mobile Payment,Seattle,Warehouse Club,true,Senior Citizen,Summer,D
iscount on Selected Items
1000000004,2021-10-13 07:28:47,Cathy Hernandez,"['Hand Sanitizer',\N,\N, 'Ice Cream', 'Hand Sanitizer']",4,\N,Debit Card,Houston,War
ehouse Club
1000000005,2021-04-26 20:45:13,Elizabeth Cook,"['Shower Gel',\N,\N, 'Paper Towels']",10,30.19,\N,Atlanta,Supermarket,True
1000000006,2023-10-07 23:36:53,Kara Bradley,"['Cereal',\N,3.0,5.57,Mobile Payment,Boston,\N,True,Student,Winter
1000000007,2022-03-30 00:46:49,Carla Hernandez,"['Iron',\N,\N, 'Iron', 'Tuna']",1,\N,Debit Card,Dallas,Warehouse Club
1000000008,2020-03-05 23:47:29,Christopher Wang,"['Banana',\N,2.0,20.04,Cash,Chicago,\N,False,Teenager,Winter
1000000009,2023-03-26 13:28:32,Alisha Hudson,"['Ketchup',\N,\N,3,88.79,Cash,\N,Pharmacy,False,Teenager
1000000010,2023-05-10 16:20:28,Samantha Mcclure,"['Shrimp',\N,3.0,28.43,Cash,Seattle,\N,True,Young Adult,Spring
1000000011,2020-01-18 08:59:50,Shari Thomas,"['Soap',\N,\N, 'Mayonnaise']",9,69.48,\N,San Francisco,Pharmacy,True
1000000012,2022-05-30 07:17:29,David Randolph,"['BBQ Sauce',\N,\N,9,21.29,Credit Card,\N,Department Store,True,Middle-Aged
1000000013,2021-05-21 18:47:34,Maria Munoz,"['Ironing Board',\N,\N, 'Cereal']",2,47.49,\N,Seattle,Convenience Store,False
1000000014,2023-02-26 11:38:21,Christopher Barnett,"['Lawn Mower',\N,8.0,15.67,Credit Card,New York,\N,False,Senior Citizen,Fall
1000000015,2021-04-28 04:39:49,Jonathan Roach,['Syrup'],9,43.35,Cash,Los Angeles,Pharmacy,true,Retiree,Summer,None
1000000016,2023-04-01 20:30:57,Alexander Hall,"['Tea',\N,\N, 'Cleaning Rags', 'Peanut Butter']",9,\N,Cash,San Francisco,Pharmacy
1000000017,2022-12-20 12:56:05,Bryan Smith,"['Tuna',\N,\N,8,64.76,Cash,\N,Specialty Store,False,Teenager
1000000018,2022-12-10 09:44:52,Kayla Sanchez,"['Syrup',\N,\N,5,78.64,Credit Card,\N,Convenience Store,True,Senior Citizen
1000000019,2021-01-01 19:48:07,Adam Foster,['Eggs'],1,60.29,Debit Card,San Francisco,Warehouse Club,false,Student,Summer,None
1000000020,2021-10-22 21:05:46,Nancy Mcdonald,"['Eggs',\N,\N, 'Toothpaste', 'Cheese']",7,\N,Debit Card,Atlanta,Pharmacy


Activities
Terminal
Nov 16 00:13
hduser@ageorgieva-VirtualBox: ~
hduser@ageorgieva-VirtualBox:~/Documents/GitHub/BigData-DV$ cd
hduser@ageorgieva-VirtualBox:~$ stop-dfs.sh
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [ageorgieva-VirtualBox]
hduser@ageorgieva-VirtualBox:~$ stop-yarn.sh
Stopping nodemanagers
Stopping resourcemanager
hduser@ageorgieva-VirtualBox:~$

Question 4:

Word Count: 103

Explain Apache Flink architecture and illustrate with your own conceptual diagram (Use of online/ book images is prohibited, Use draw.io to create the image). What is Apache Storm, and how does it differ from other distributed computing systems? Consider a text file comprising at least 20,000 words and write a wordcount program (Java/ Python) to count the frequency of words and related aggregation functions.
(20 Marks)

Apache Flink consists of Master Node and Slave nodes. The master node achieves distributed computing by sharing the workload among different slave nodes, which increases the processing of data performance significantly (Flink Architecture, 2023).

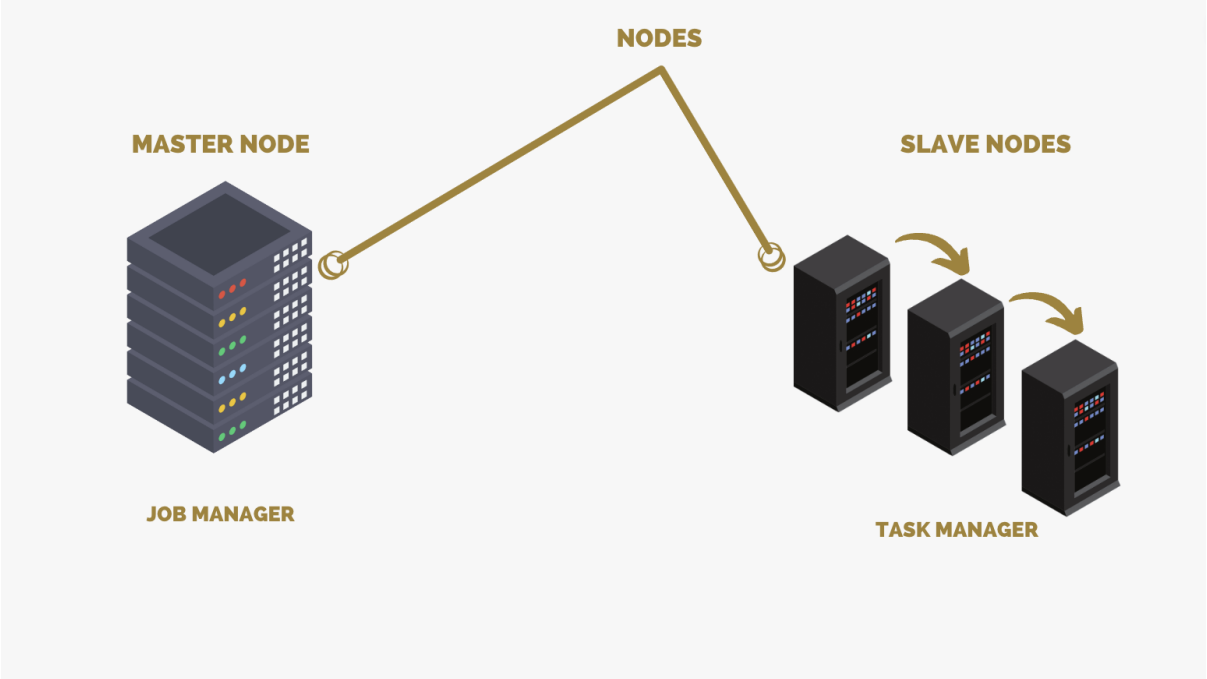
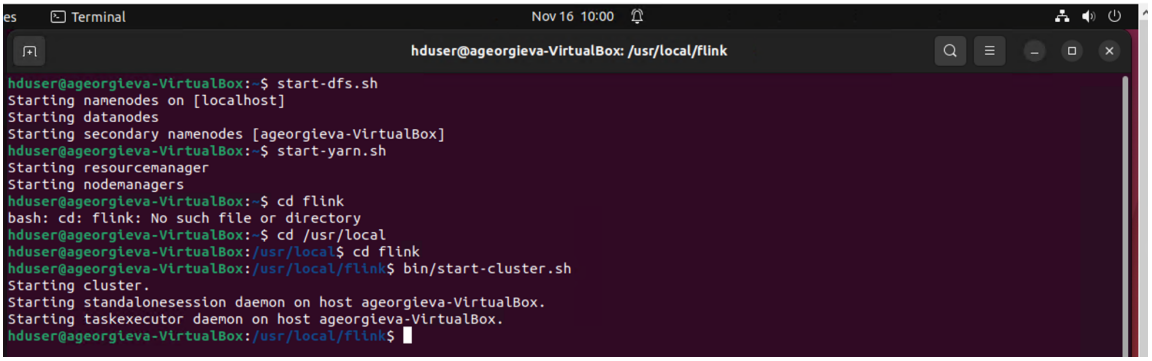
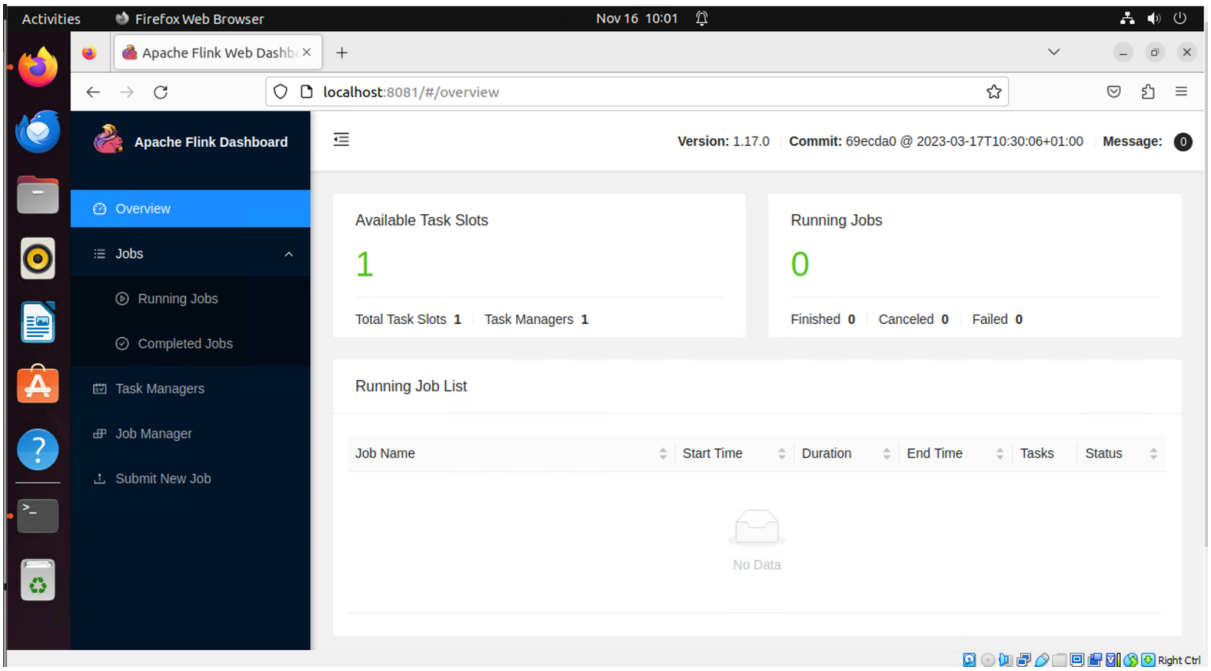
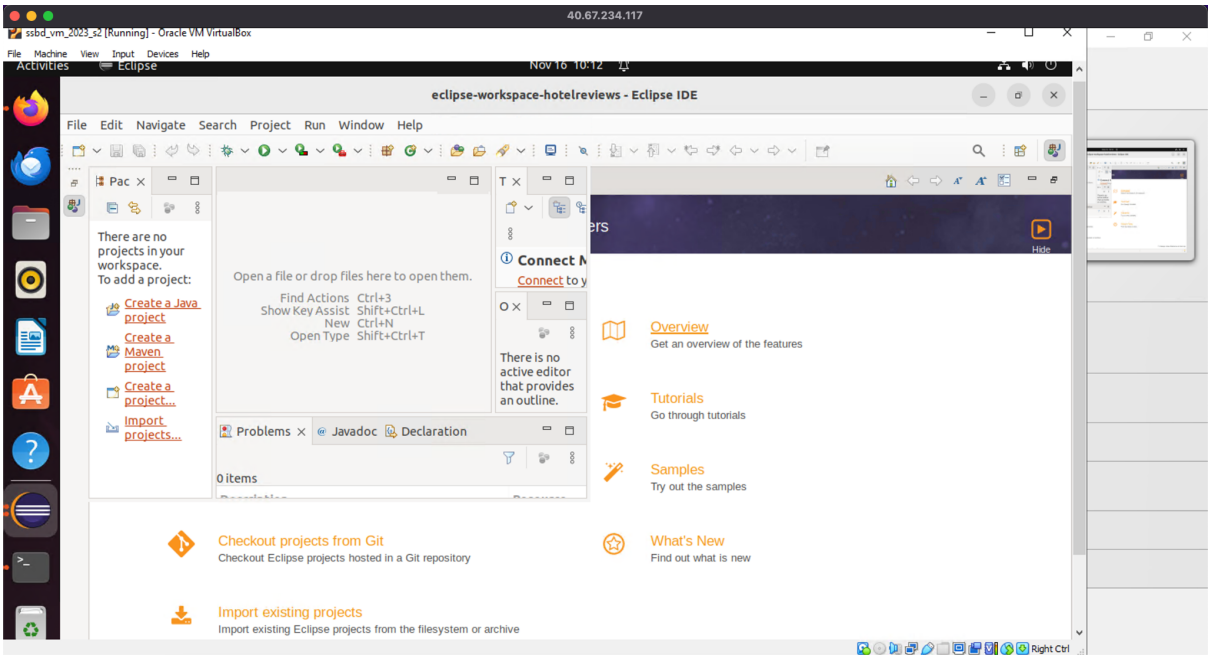


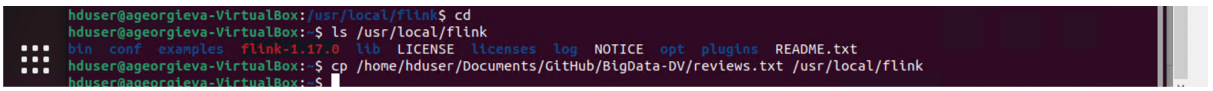
Figure 6: Apache Flink Architecture

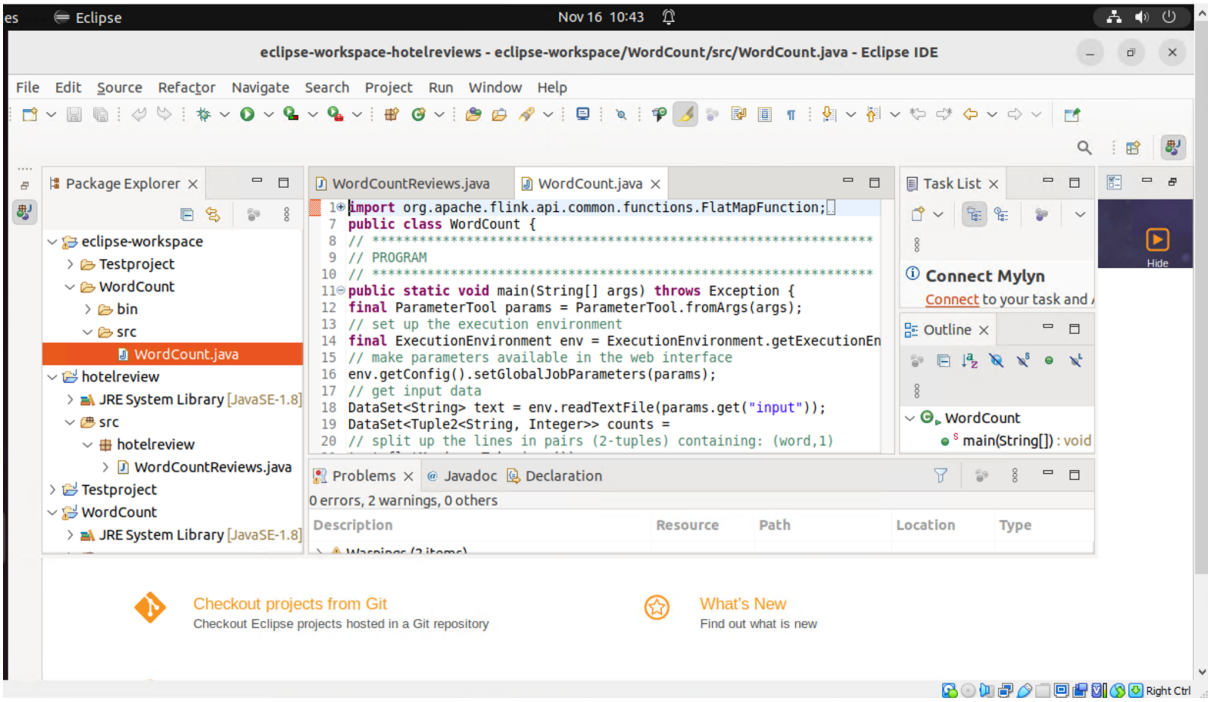
Apache Flink

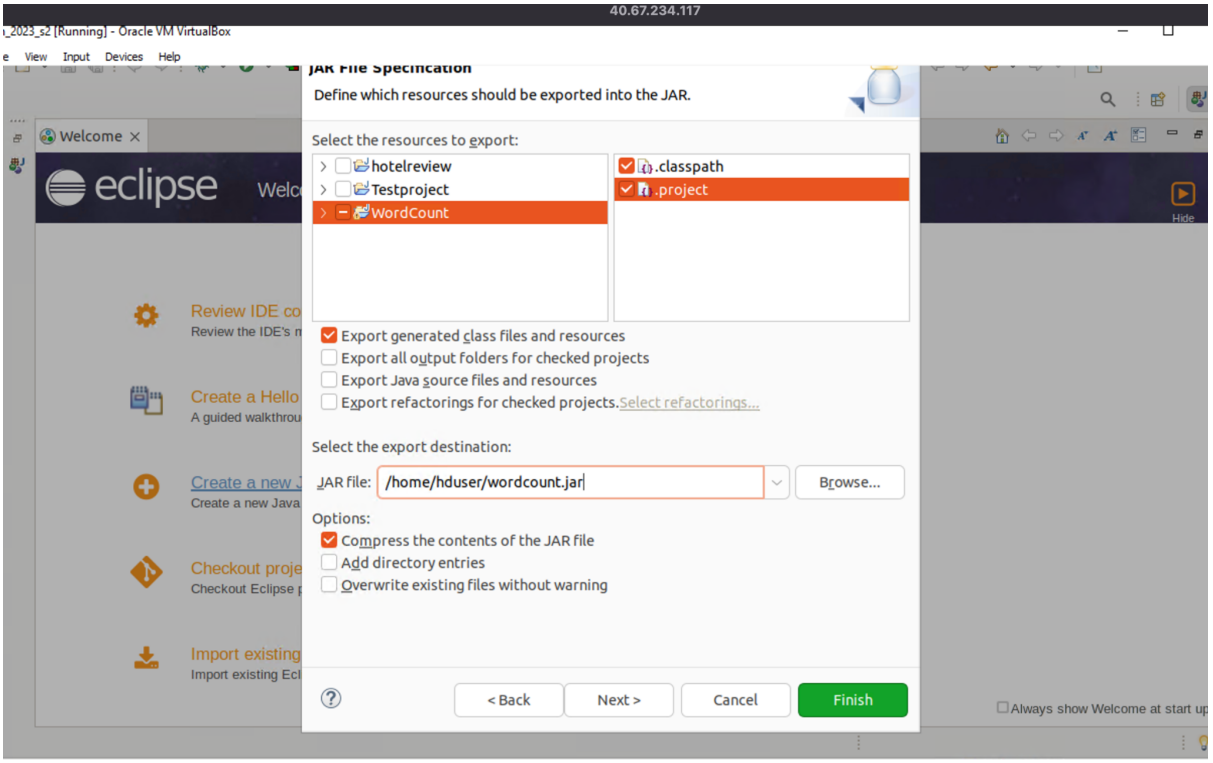



Activities
Firefox Web Browser
Nov 16 10:01
Apache Flink Web Dashb
localhost:8081/#/overview
Apache Flink Dashboard
Version: 1.17.0
Commit: 69ecda0 @ 2023-03-17T10:30:06+01:00
Message:
Overview
Jobs
Running Jobs
Completed Jobs
Task Managers
Job Manager
Submit New Job
Available Task Slots
1
Total Task Slots 1
Task Managers 1
Running Jobs
0
Finished 0
Canceled 0
Failed 0
Running Job List
Job Name
Start Time
Duration
End Time
Tasks
Status
No Data
Right Ctrl


40.67.234.117
ssbd_vm_2023_s2 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Eclipse
Nov 16 10:12
eclipse-workspace-hotelreviews - Eclipse IDE
File Edit Navigate Search Project Run Window Help
There are no projects in your workspace.
To add a project:
Create a Java project
Create a Maven project
Create a project...
Import projects...
Open a file or drop files here to open them.
Find Actions Ctrl+3
Show Key Assist Shift+Ctrl+L
New Ctrl+N
Open Type Shift+Ctrl+T
There is no active editor that provides an outline.
Problems
Javadoc
Declaration
0 items
Overview
Get an overview of the features
Tutorials
Go through tutorials
Samples
Try out the samples
What's New
Find out what is new
Checkout projects from Git
Checkout Eclipse projects hosted in a Git repository
Import existing projects
Import existing Eclipse projects from the filesystem or archive
Right Ctrl

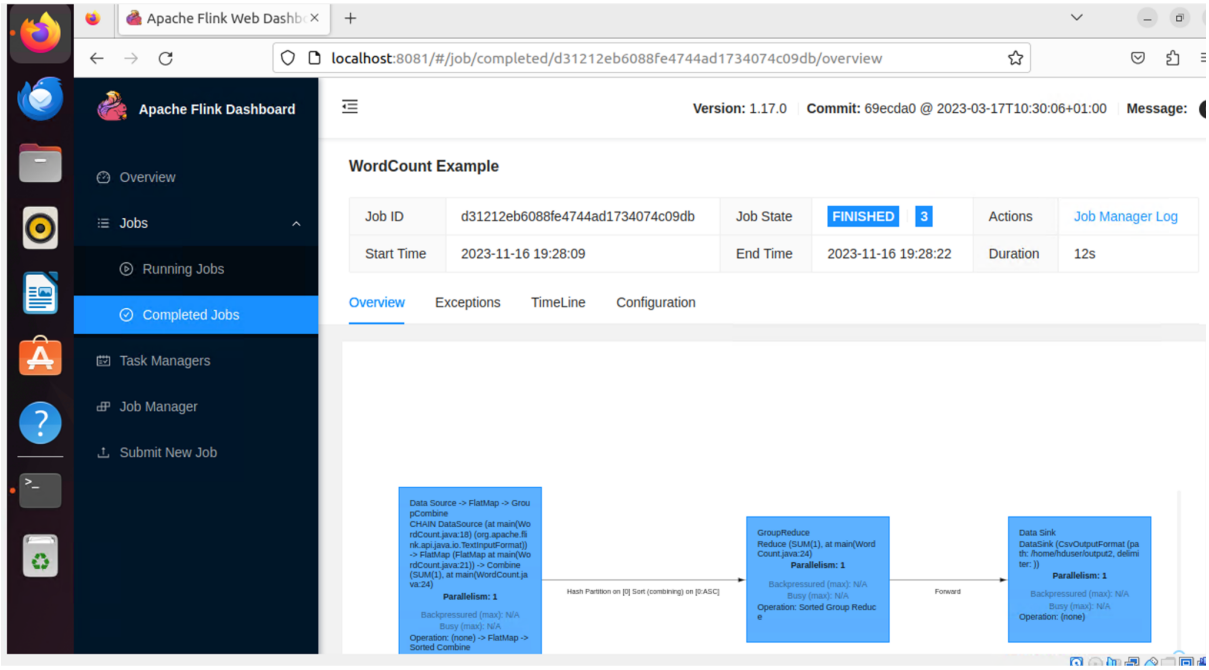

hduser@ageorgieva-VirtualBox:/usr/local/flink$ cd
hduser@ageorgieva-VirtualBox:~$ ls /usr/local/flink
bin conf examples flink-1.17.0 lib LICENSE licenses log NOTICE opt plugins README.txt
hduser@ageorgieva-VirtualBox:~$ cp /home/hduser/Documents/GitHub/BigData-DV/reviews.txt /usr/local/flink
hduser@ageorgieva-VirtualBox:~$

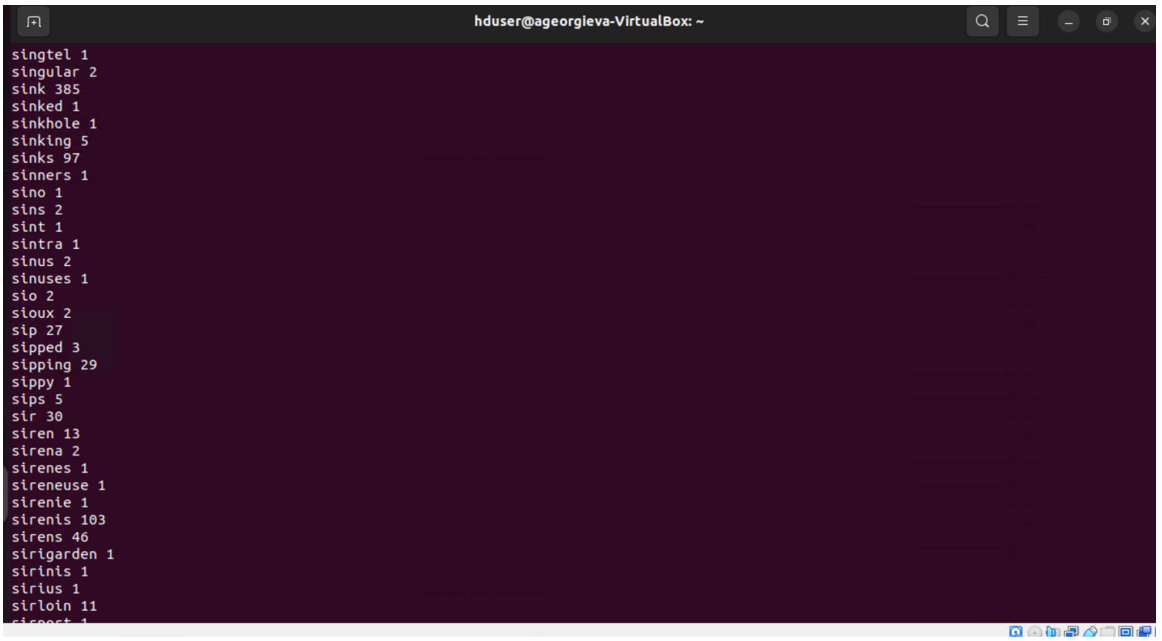


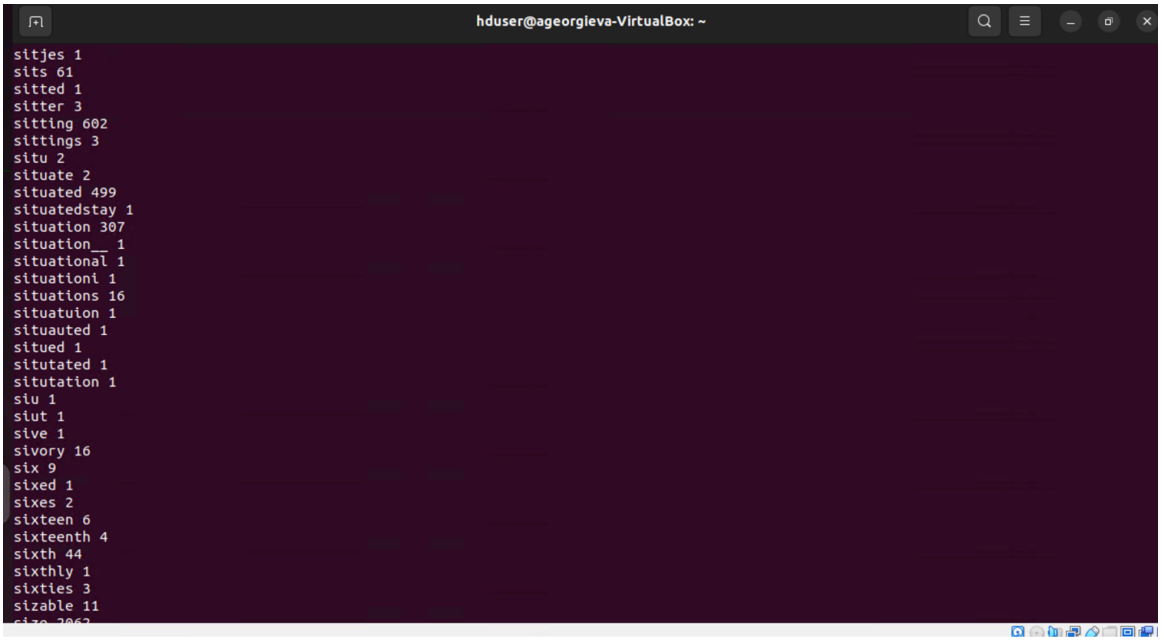



```
hduser@ageorgieva-VirtualBox:/usr/local/flink$ ./bin/flink run /home/hduser/wordcount.jar --input ./reviews.txt --output /home/hduse
r/output2
Job has been submitted with JobID d31212eb6088fe4744ad1734074c09db
Program execution finished
Job with JobID d31212eb6088fe4744ad1734074c09db has finished.
Job Runtime: 12511 ms

hduser@ageorgieva-VirtualBox:/usr/local/flink$
```

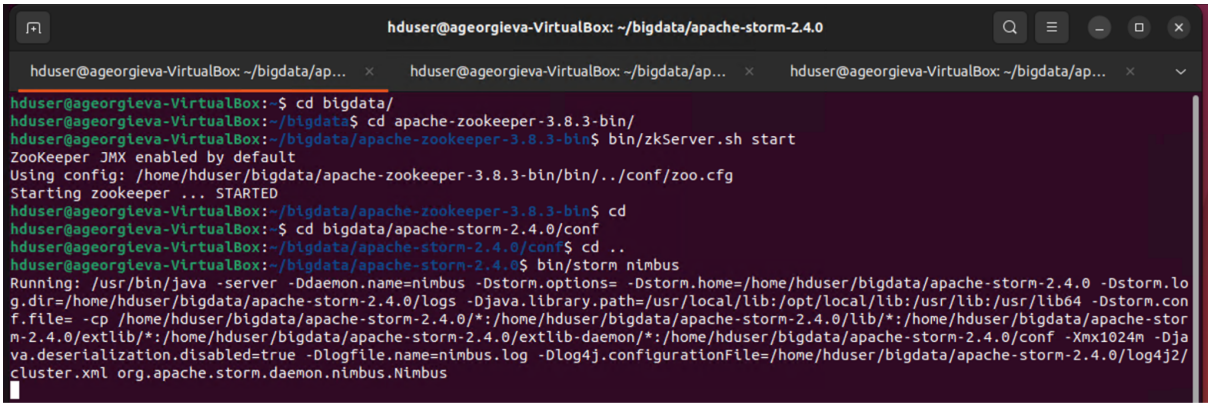

Apache Flink Web Dashb
localhost:8081/#/job/completed/d31212eb6088fe4744ad1734074c09db/overview
Apache Flink Dashboard
Version: 1.17.0
Commit: 69ecda0 @ 2023-03-17T10:30:06+01:00
Message:
Overview
Jobs
Running Jobs
Completed Jobs
Task Managers
Job Manager
Submit New Job
WordCount Example
Job ID
d31212eb6088fe4744ad1734074c09db
Job State
FINISHED
3
Actions
Job Manager Log
Start Time
2023-11-16 19:28:09
End Time
2023-11-16 19:28:22
Duration
12s
Overview
Exceptions
TimeLine
Configuration
Data Source -> FlatMap -> GroupCombine
CHAIN DataSource (at main(WordCount.java:18) (org.apache.flink.api.java.io.TextInputFormat)) -> FlatMap (FlatMap at main(WordCount.java:21)) -> Combine (SUM(1), at main(WordCount.java:24)
Parallelism: 1
Backpressured (max): N/A
Busy (max): N/A
Operation: (none) -> FlatMap -> Sorted Combine
Hash Partition on [0] Sort (combining) on [0:ASC]
GroupReduce
Reduce (SUM(1), at main(WordCount.java:24)
Parallelism: 1
Backpressured (max): N/A
Busy (max): N/A
Operation: Sorted Group Reduce
Forward
Data Sink
DataSink (CsvOutputFormat (path: /home/hduser/output2, delimiter:))
Parallelism: 1
Backpressured (max): N/A
Busy (max): N/A
Operation: (none)

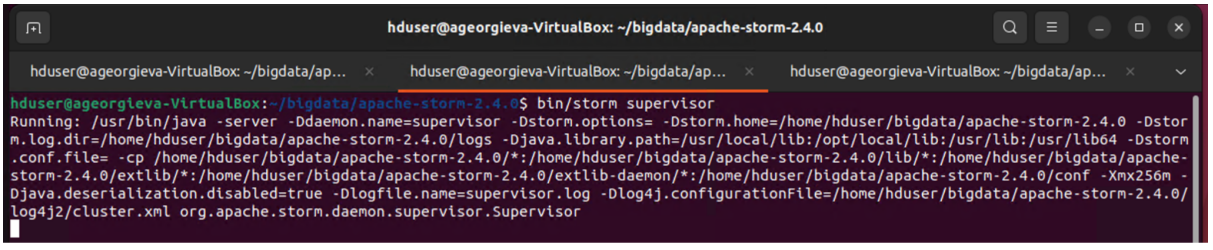

hduser@ageorgieva-VirtualBox: ~
singtel 1
singular 2
sink 385
sinked 1
sinkhole 1
sinking 5
sinks 97
sinners 1
sino 1
sins 2
sint 1
sintra 1
sinus 2
sinuses 1
sio 2
sioux 2
sip 27
sipped 3
sipping 29
sippy 1
sips 5
sir 30
siren 13
sirena 2
sirenes 1
sireneuse 1
sirenie 1
sirenis 103
sirens 46
sirigarden 1
sirinis 1
sirius 1
sirloin 11

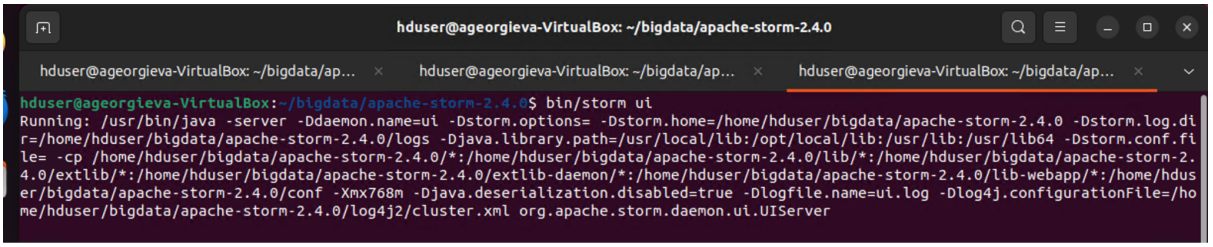


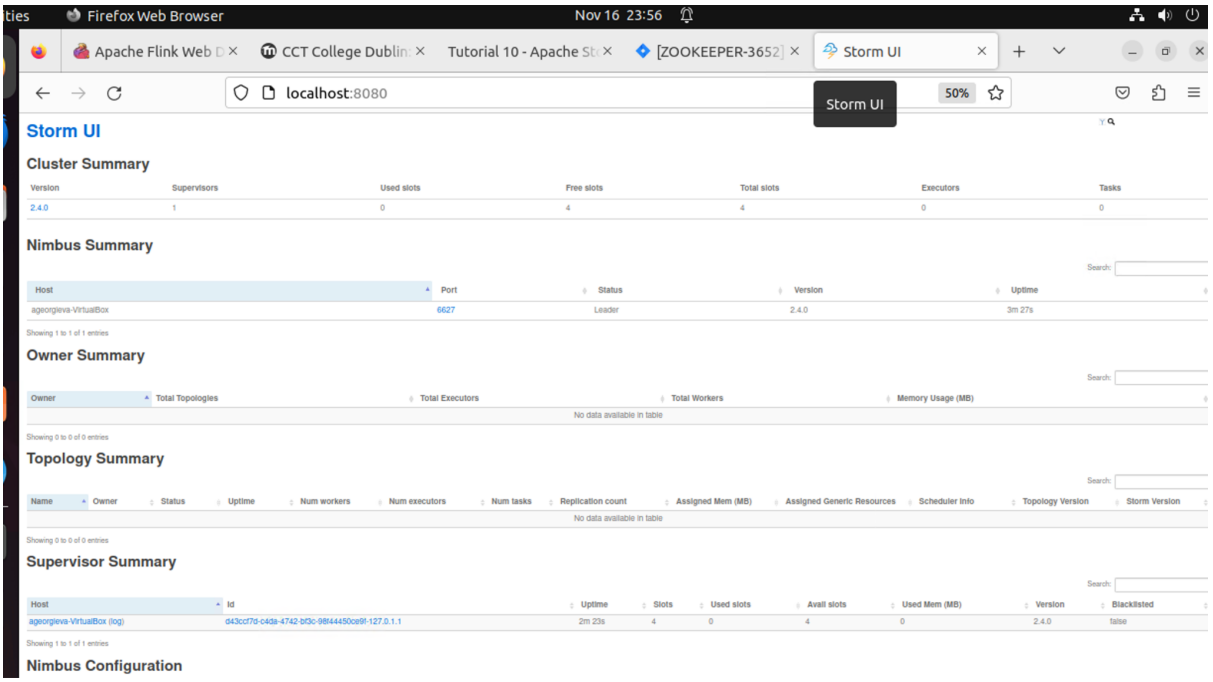
Apache Storm

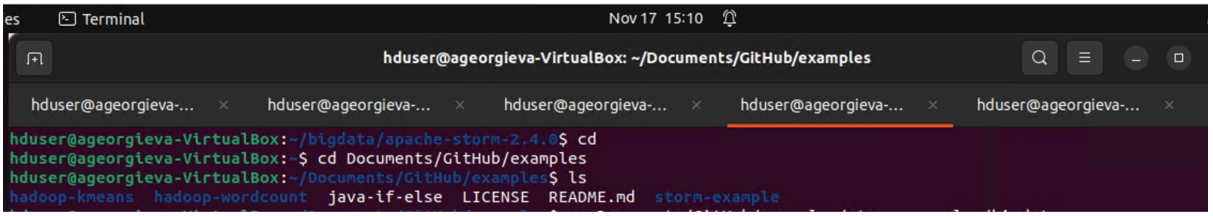
Apache Storm, like Hadoop, is a distributed network for processing data at scale. Apache Storm does stream processing, making it ideal for real-time applications like stock market updates, for example. While Hadoop is suitable for batch processing or tasks that are not time-sensitive, such as monthly salary processing. Storm can speed up the processing of data in a fast manner within seconds, while Hadoop can take minutes or hours (Apache Storm, 2023).











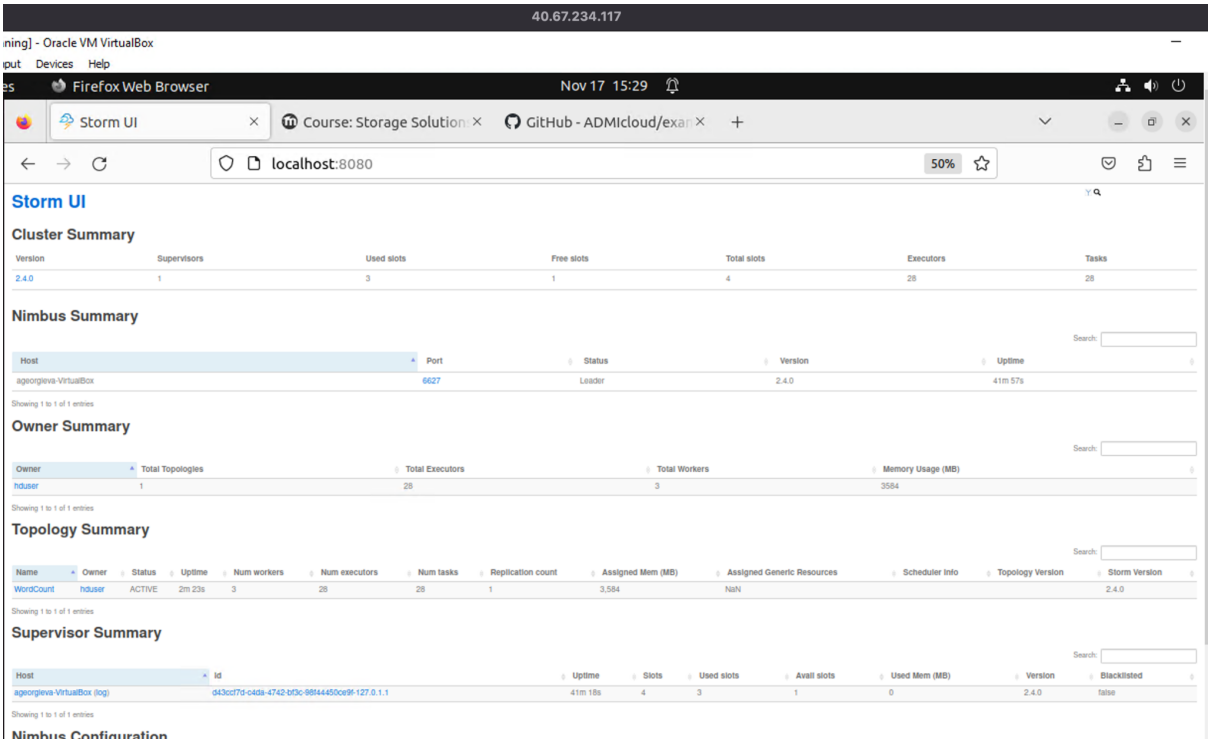

```
hduser@ageorgieva-VirtualBox:~/Documents/GitHub/examples$ cp -r ~/Documents/GitHub/examples/storm-example ~/bigdata/
```

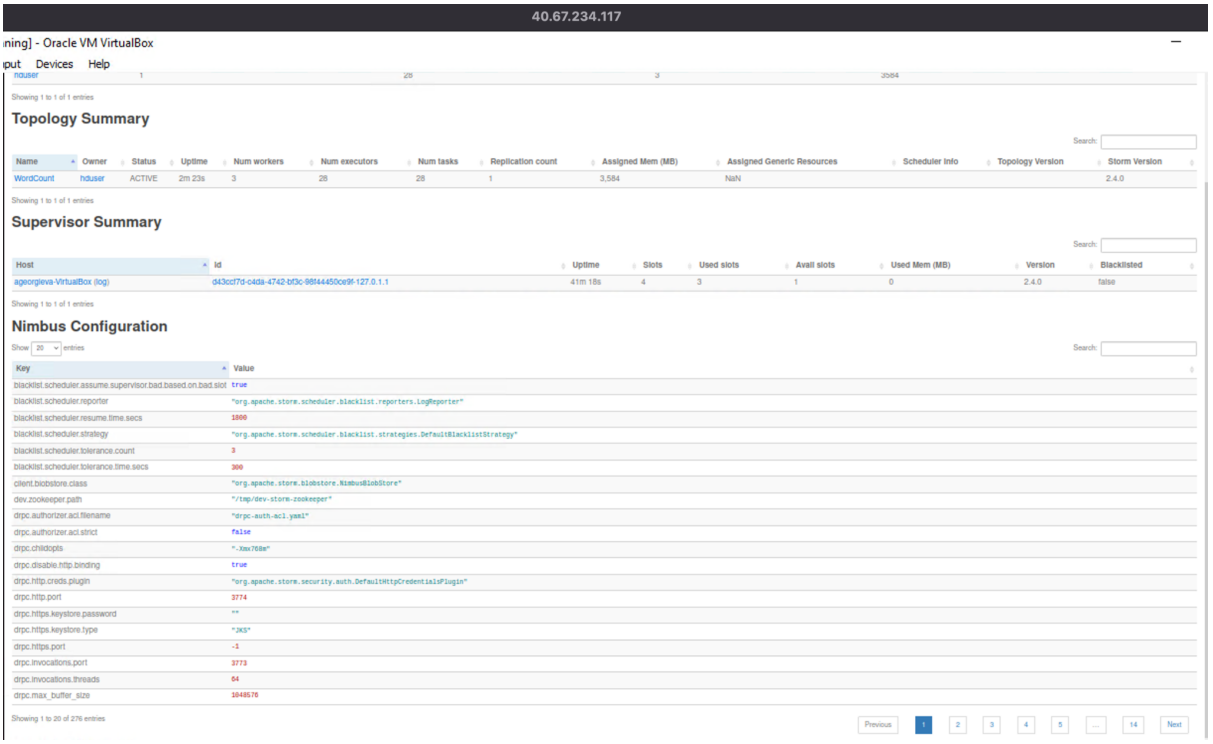
```
hduser@ageorgieva-VirtualBox:~/Documents/GitHub/examples$ cd
hduser@ageorgieva-VirtualBox:~$ cd bigdata/
hduser@ageorgieva-VirtualBox:~/bigdata$ cd storm-example
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ mvn clean install
[INFO] Scanning for projects...
[INFO]
[INFO] ----------------------< admicloud:storm-example >-----------------------
[INFO] Building storm-example 1.0
[INFO] --------------------------------[ jar ]---------------------------------
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/plugins/maven-resources-plugin/2.6/maven-resources-plugin-2.6.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/apache/maven/plugins/maven-resources-plugin/2.6/maven-resources-plugin-2.6.pom (8.1 kB at 3.4 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/plugins/maven-resources-plugin/2.6/maven-resc
```

```
[INFO] --- maven-install-plugin:2.4:install (default-install) @ storm-example ---
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.pom (2.
5 kB at 45 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus/3.1/plexus-3.1.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus/3.1/plexus-3.1.pom (19 kB at 327 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.pom (1.1
kB at 20 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-components/1.1.7/plexus-components-1.
1.7.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-components/1.1.7/plexus-components-1.1
.7.pom (5.0 kB at 89 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.jar
Downloading from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.jar
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-digest/1.0/plexus-digest-1.0.jar (12 k
B at 100 kB/s)
Downloaded from central: https://repo.maven.apache.org/maven2/org/codehaus/plexus/plexus-utils/3.0.5/plexus-utils-3.0.5.jar (23
0 kB at 676 kB/s)
[INFO] Installing /home/hduser/bigdata/storm-example/target/storm-example-1.0.jar to /home/hduser/.m2/repository/admicloud/stor
m-example/1.0/storm-example-1.0.jar
[INFO] Installing /home/hduser/bigdata/storm-example/pom.xml to /home/hduser/.m2/repository/admicloud/storm-example/1.0/storm-e
xample-1.0.pom
[INFO] Installing /home/hduser/bigdata/storm-example/target/storm-example-1.0-jar-with-dependencies.jar to /home/hduser/.m2/rep
ository/admicloud/storm-example/1.0/storm-example-1.0-jar-with-dependencies.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  01:05 min
[INFO] Finished at: 2023-11-17T15:19:40Z
[INFO] ------------------------------------------------------------------------
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$
```

```
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  01:05 min
[INFO] Finished at: 2023-11-17T15:19:40Z
[INFO] ------------------------------------------------------------------------
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ cd
hduser@ageorgieva-VirtualBox:~$ /home/hduser/bigdata/apache-storm-2.4.0/bin/storm jar /home/hduser/bigdata/storm-example/target
/storm-example-1.0-jar-with-dependencies.jar admicloud.storm.wordcount.WordCountTopology WordCount
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hduser/bigdata/apache-storm-2.4.0/lib/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/impl/Stati
cLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hduser/bigdata/storm-example/target/storm-example-1.0-jar-with-dependencies.jar!/org/sl
f4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Running: /usr/bin/java -client -Ddaemon.name= -Dstorm.options= -Dstorm.home=/home/hduser/bigdata/apache-storm-2.4.0 -Dstorm.log
.dir=/home/hduser/bigdata/apache-storm-2.4.0/logs -Djava.library.path=/usr/local/lib:/opt/local/lib:/usr/lib:/usr/lib64 -Dstorm
.conf.file= -cp /home/hduser/bigdata/apache-storm-2.4.0/*:/home/hduser/bigdata/apache-storm-2.4.0/lib-worker/*:/home/hduser/big
data/apache-storm-2.4.0/extlib/*:/home/hduser/bigdata/storm-example/target/storm-example-1.0-jar-with-dependencies.jar:/home/hd
user/bigdata/apache-storm-2.4.0/conf:/home/hduser/bigdata/apache-storm-2.4.0/bin: -Dstorm.jar=/home/hduser/bigdata/storm-exampl
e/target/storm-example-1.0-jar-with-dependencies.jar -Dstorm.dependency.jars= -Dstorm.dependency.artifacts={} admicloud.storm.w
ordcount.WordCountTopology WordCount
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hduser/bigdata/apache-storm-2.4.0/lib-worker/log4j-slf4j-impl-2.17.1.jar!/org/slf4j/imp
l/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hduser/bigdata/storm-example/target/storm-example-1.0-jar-with-dependencies.jar!/org/sl
f4j/impl/StaticLoggerBinder.class]
```

```
6da7c828-3b4a-450d-8f4f-21a426c40ad7.jar
Start uploading file '/home/hduser/bigdata/storm-example/target/storm-example-1.0-jar-with-dependencies.jar' to '/home/hduser/b
igdata/apache-storm-2.4.0/storm-local/nimbus/inbox/stormjar-6da7c828-3b4a-450d-8f4f-21a426c40ad7.jar' (165695 bytes)
[==================================================] 165695 / 165695
File '/home/hduser/bigdata/storm-example/target/storm-example-1.0-jar-with-dependencies.jar' uploaded to '/home/hduser/bigdata/
apache-storm-2.4.0/storm-local/nimbus/inbox/stormjar-6da7c828-3b4a-450d-8f4f-21a426c40ad7.jar' (165695 bytes)
15:26:32.809 [main] INFO  o.a.s.StormSubmitter - Successfully uploaded topology jar to assigned location: /home/hduser/bigdata/
apache-storm-2.4.0/storm-local/nimbus/inbox/stormjar-6da7c828-3b4a-450d-8f4f-21a426c40ad7.jar
15:26:32.811 [main] INFO  o.a.s.StormSubmitter - Submitting topology WordCount in distributed mode with conf {"storm.zookeeper.
topology.auth.scheme":"digest","storm.zookeeper.topology.auth.payload":"-5678361208442557000:-8914009818931553625","topology.wo
rkers":3,"topology.debug":true}
15:26:34.113 [main] INFO  o.a.s.StormSubmitter - Finished submitting topology: WordCount
hduser@ageorgieva-VirtualBox:~$
```


40.67.234.117
Firefox Web Browser
Nov 17 15:29
Storm UI
Course: Storage Solution
GitHub - ADMIcloud/exa
localhost:8080
50%
Storm UI
Cluster Summary
Version
Supervisors
Used slots
Free slots
Total slots
Executors
Tasks
2.4.0
1
3
1
4
28
28
Nimbus Summary
Host
Port
Status
Version
Uptime
ageorgieva-VirtualBox
6627
Leader
2.4.0
41m 57s
Owner Summary
Owner
Total Topologies
Total Executors
Total Workers
Memory Usage (MB)
hduser
1
28
3
3584
Topology Summary
Name
Owner
Status
Uptime
Num workers
Num executors
Num tasks
Replication count
Assigned Mem (MB)
Assigned Generic Resources
Scheduler Info
Topology Version
Storm Version
WordCount
hduser
ACTIVE
2m 23s
3
28
28
1
3,584
NaN
2.4.0
Supervisor Summary
Host
Id
Uptime
Slots
Used slots
Avail slots
Used Mem (MB)
Version
Blacklisted
ageorgieva-VirtualBox (log)
d43ccf7d-c4da-4742-bf3c-98f44450ce9f-127.0.1.1
41m 18s
4
3
1
0
2.4.0
false
Nimbus Configuration


40.67.234.117
Topology Summary
Supervisor Summary
Nimbus Configuration
Key
Value
blacklist.scheduler.assume.supervisor.bad.based.on.bad.slot
blacklist.scheduler.reporter
blacklist.scheduler.resume.time.secs
blacklist.scheduler.strategy
blacklist.scheduler.tolerance.count
blacklist.scheduler.tolerance.time.secs
client.blobstore.class
dev.zookeeper.path
drpc.authorizer.acl.filename
drpc.authorizer.acl.strict
drpc.childopts
drpc.disable.http.binding
drpc.http.creds.plugin
drpc.http.port
drpc.https.keystore.password
drpc.https.keystore.type
drpc.https.port
drpc.invocations.port
drpc.invocations.threads
drpc.max_buffer_size

40.67.234.117

ning] - Oracle VM VirtualBox

put Devices Help

localhost:8080/topology.html?id=WordCount-1-1700234793 50%

Storm UI

Search WordCount-1-1700234793: Search Search Archived Logs:

Topology summary

Name	Id	Owner	Status	Uptime	Num workers	Num executors	Num tasks	Replication count	Assigned Mem (MB)	Scheduler Info	Topology Version	Storm Version
WordCount	WordCount-1-1700234793	hduser	ACTIVE	4m 38s	3	28	28	1	0			2.4.0

Topology actions

Activate Deactivate Rebalance Kill Debug Stop Debug Change Log Level

Topology stats

Window	Emitted	Transferred	Complete latency (ms)	Acked	Failed
10m 0s	41,720	27,120	0	0	0
3h 0m 0s	41,720	27,120	0	0	0
1d 0h 0m 0s	41,720	27,120	0	0	0
All time	41,720	27,120	0	0	0

Spouts (All time)

Search:

Id	Executors	Tasks	Emitted	Transferred	Complete latency (ms)	Acked	Failed	Error Host	Error Port	Last error	Error Time
spout	5	5	6,120	6,120	0.000	0	0				

Showing 1 to 1 of 1 entries

Bolts (All time)

Search:

Id	Executors	Tasks	Emitted	Transferred	Capacity (last 10m)	Execute latency (ms)	Executed	Process latency (ms)	Acked	Failed	Error Host	Error Port	Last error	Error Time
count	12	12	14,600	0	0.624	62.382	14,540	41.355	14,580	0				
split	8	8	21,000	21,000	0.836	346.709	3,300	341.429	3,260	0				

Showing 1 to 2 of 2 entries

Worker Resources

Search: Toggle Components

Host	Supervisor Id	Port	Uptime	Num executors	Assigned Mem (MB)	Components
ageorgieva-VirtualBox	d43ccf7d-c4da-4742-bf3c-98f44450ce9f-127.0.1.1	6701	3m 6s	9	1152	3 components
ageorgieva-VirtualBox	d43ccf7d-c4da-4742-bf3c-98f44450ce9f-127.0.1.1	6700	3m 7s	10	1280	3 components

40.67.234.117

nning] - Oracle VM VirtualBox

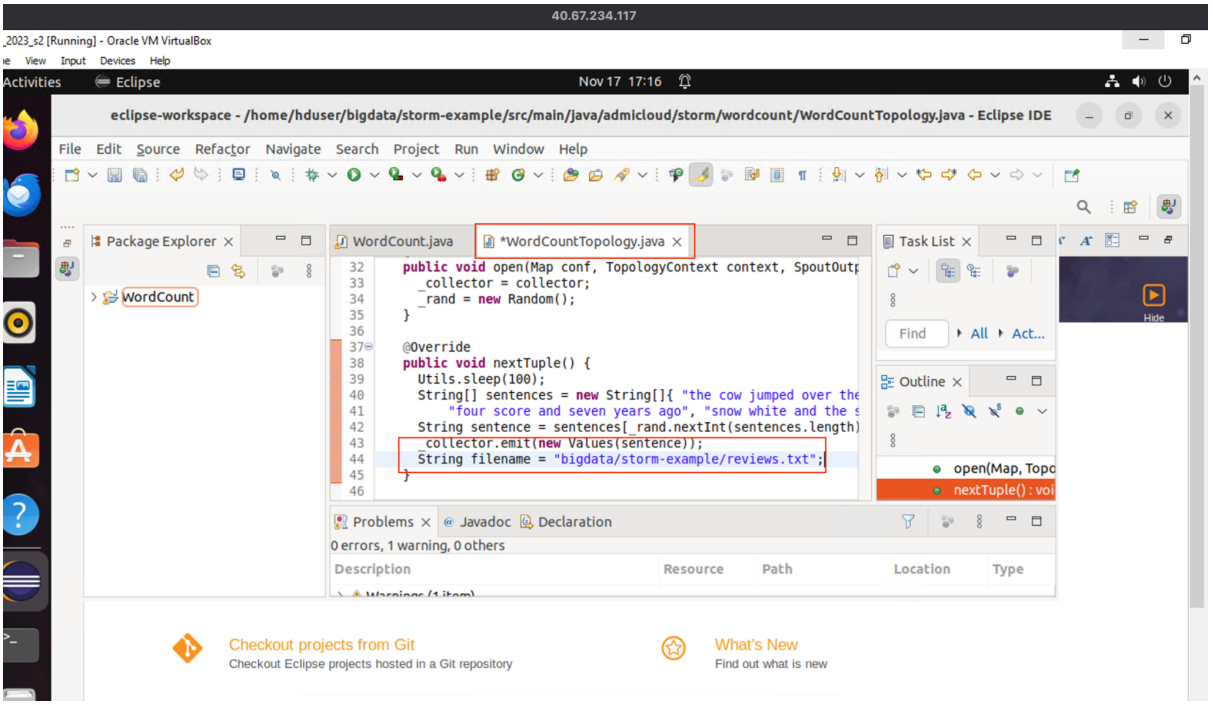
nput Devices Help

hduser@ageorgieva-VirtualBox: ~/bigdata

hduser@ageorgieva-... hduser@ageorgieva-... hduser@ageorgieva-... hduser@ageorgieva-... hduser@ageorgieva-...

```
hduser@ageorgieva-VirtualBox:~/bigdata/apache-storm-2.4.0$ cd
hduser@ageorgieva-VirtualBox:~$ cd Documents/GitHub/BigData-DV
hduser@ageorgieva-VirtualBox:~/Documents/GitHub/BigData-DV$ ls
'Data Visualisation - Dashboard.ipynb'   retail.csv     Retail_Transactions_Dataset.csv   tripadvisor_hotel_reviews.csv
 HotelReviews.ipynb                      Retail.ipynb   reviews.txt
hduser@ageorgieva-VirtualBox:~/Documents/GitHub/BigData-DV$ cp -r ~/Documents/GitHub/BigData-DV/reviews.txt ~/bigdata/
hduser@ageorgieva-VirtualBox:~/Documents/GitHub/BigData-DV$ cd
hduser@ageorgieva-VirtualBox:~$ cd bigdata/
hduser@ageorgieva-VirtualBox:~/bigdata$ ls
apache-storm-2.4.0  apache-zookeeper-3.8.3-bin  examples  reviews.txt  storm-example
hduser@ageorgieva-VirtualBox:~/bigdata$
```

```
hduser@ageorgieva-VirtualBox:~$ cd bigdata/
hduser@ageorgieva-VirtualBox:~/bigdata$ ls
apache-storm-2.4.0  apache-zookeeper-3.8.3-bin  examples  reviews.txt  storm-example
hduser@ageorgieva-VirtualBox:~/bigdata$ cd storm-example/
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ ls
pom.xml  src  target
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ cd ..
hduser@ageorgieva-VirtualBox:~/bigdata$ mv reviews.txt ~/bigdata/storm-example
hduser@ageorgieva-VirtualBox:~/bigdata$ cd storm-example/
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ ls
pom.xml  reviews.txt  src  target
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ mvn clean package
[INFO] Scanning for projects...
[INFO]
[INFO] ----------------------< admicloud:storm-example >-----------------------
[INFO] Building storm-example 1.0
[INFO] --------------------------------[ jar ]---------------------------------
[INFO]
[INFO] --- maven-clean-plugin:2.5:clean (default-clean) @ storm-example ---
[INFO] Deleting /home/hduser/bigdata/storm-example/target
[INFO]
[INFO] --- maven-resources-plugin:2.6:resources (default-resources) @ storm-example ---
[INFO] Using 'UTF-8' encoding to copy filtered resources.
[INFO] skip non existing resourceDirectory /home/hduser/bigdata/storm-example/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.1:compile (default-compile) @ storm-example ---
[INFO] Changes detected - recompiling the module!
[INFO] Compiling 1 source file to /home/hduser/bigdata/storm-example/target/classes
[INFO]
[INFO] --- maven-resources-plugin:2.6:testResources (default-testResources) @ storm-example ---
[INFO] Using 'UTF-8' encoding to copy filtered resources.
```




```
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ hdfs dfs -put reviews.txt /user1/
put: `/user1/reviews.txt': File exists
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$ hdfs dfs -ls /user1/
Found 12 items
-rw-r--r--   1 hduser supergroup  113709538 2023-10-27 16:39 /user1/0.csv
-rw-r--r--   1 hduser supergroup  113794820 2023-10-27 16:39 /user1/1.csv
-rw-r--r--   1 hduser supergroup   68084878 2023-10-27 16:39 /user1/2.csv
-rw-r--r--   1 hduser supergroup   24834870 2023-11-10 21:58 /user1/Loan_default.csv
-rw-r--r--   1 hduser supergroup          0 2023-10-27 16:39 /user1/_SUCCESS
-rw-r--r--   1 hduser supergroup     135287 2023-09-20 23:38 /user1/britney-spears.txt
drwxr-xr-x   - hduser supergroup          0 2023-11-13 00:44 /user1/data
-rw-r--r--   1 hduser supergroup      12048 2023-11-10 22:43 /user1/df.ipynb
-rw-r--r--   1 hduser supergroup         57 2023-09-27 18:41 /user1/employees.csv
-rw-r--r--   1 hduser supergroup   30287534 2023-10-27 16:32 /user1/filtereddf.csv.zip
-rw-r--r--   1 hduser supergroup   14884042 2023-11-15 22:39 /user1/reviews.txt
-rw-r--r--   1 hduser supergroup  305211208 2023-11-02 14:30 /user1/united.csv
hduser@ageorgieva-VirtualBox:~/bigdata/storm-example$
```

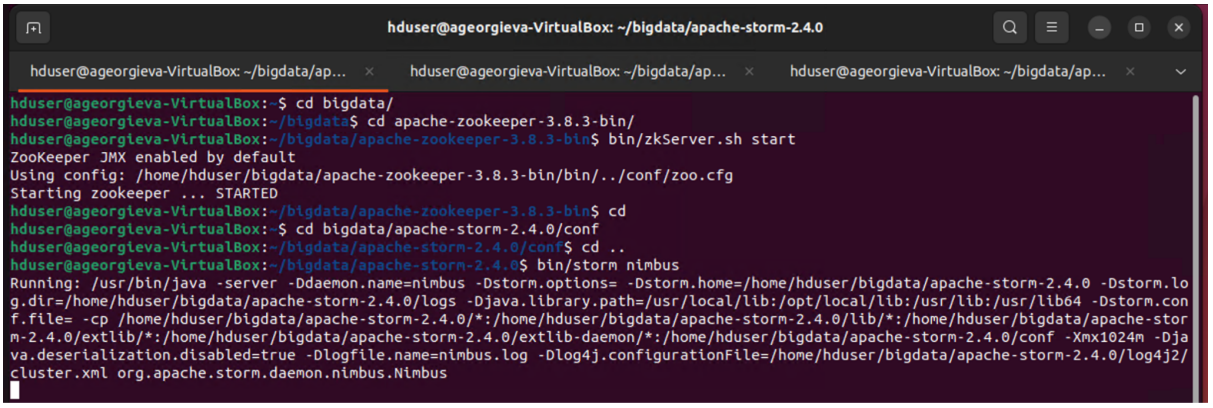
Question 5:

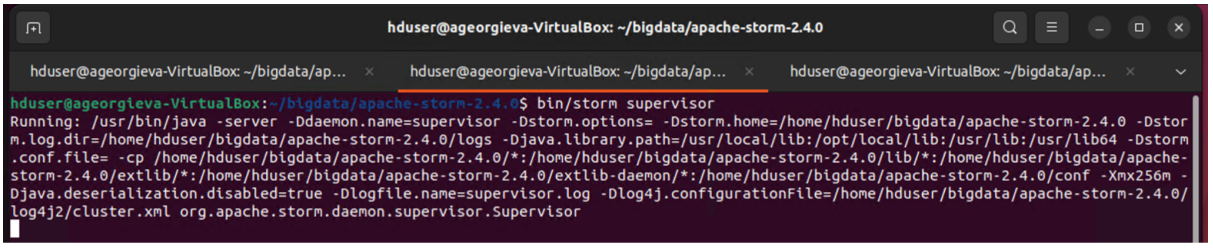
Word Count: 122

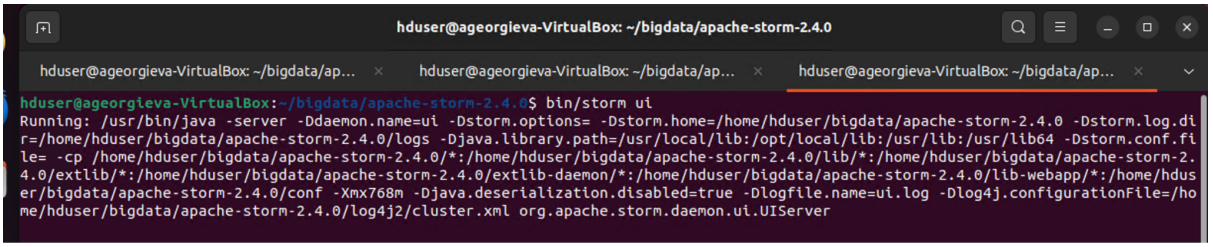
Why is Apache Storm useful for Stream processing specifically? Distinguish the characteristics of Apache storm as compared to Hadoop. What is the role of Apache Zookeeper in Apache Storm deployment. Provide the screenshot of your VM to show working of Storm UI including Cluster, Nimbus and Owner summary.
(20 Marks)

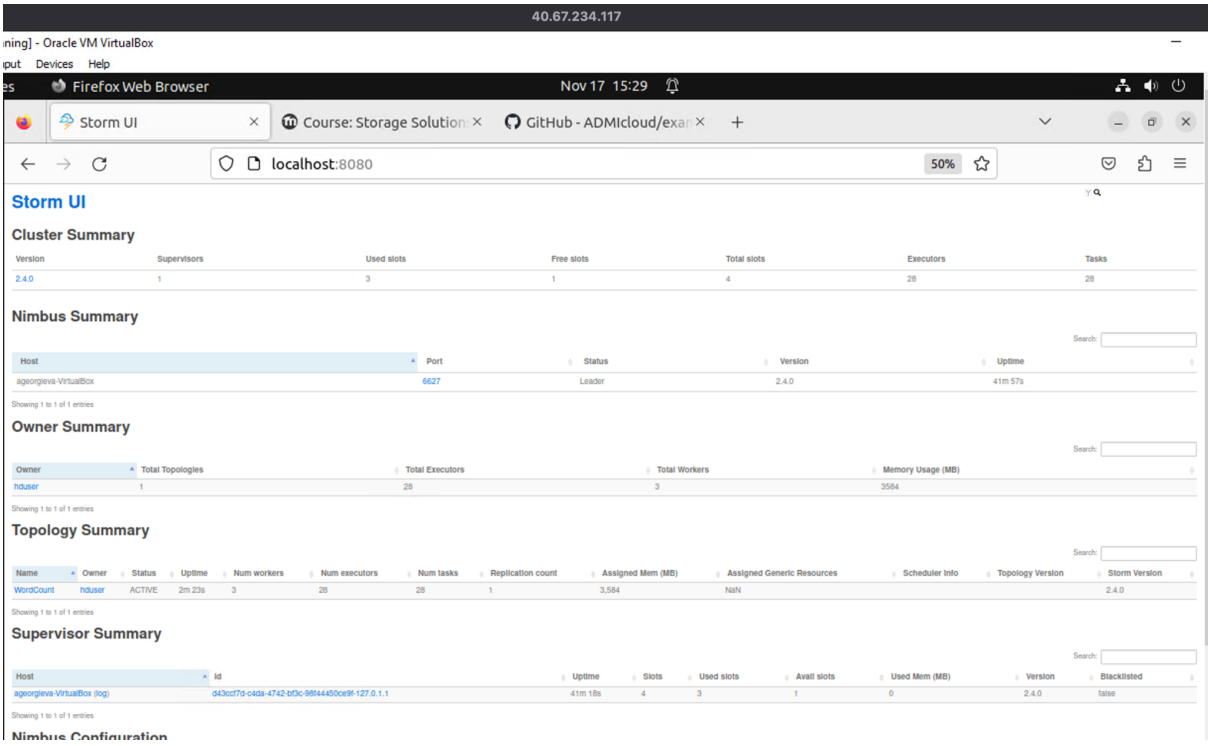
Storm has the capability to process millions of records of data per second per node. It could be useful for cyber security analytics, data monetization, customer service management and threat detection. Apache Storm has three nodes - Nimbus which is the master node, zookeeper, which coordinates Storm cluster and Supervisor, starts and ceases processes of workers depending on the signals from Nimbus. Hadoop excels in batch processing tasks, fitting well with applications where instant data processing isn't a necessity (Cloudera, 2023).

Apache Zookeeper in Apache Storm deployment is acting as a coordinator between clusters and sharing of data. Since Nimbus is stateless, Zookeeper keeps its activity in check and in general is responsible for the state of both Nimbus and Supervisor. (TutorialsPoint, 2023)









Data Visualisation

Word Count: 795

The inspiration for the design and dashboard creation came from a tutorial authored by (Saraswat, 2023)(Programming is Fun, 2023). When designing this dashboard, it was imperative to me that the navigation between filter elements is seamless, providing users the ability to adjust an individual chart as they wish. This approach is intended to offer clear insights into how each filter can influence the target variable. To conceptualise the layout and the user interface of the dashboard, I utilised draw.io for creating a detailed sketch, ensuring a user-friendly and intuitive design.

Figure 7: Dashboard - draw.io

The dataset (Goyal, 2021) in question was previously employed in a machine learning project for Class Assignment 1 last Semester and I decided to use the same for the current project. Its lack of a time series component notwithstanding, I believed its utilisation could underscore the significance of robust storage solutions for data processing within the banking sector, as highlighted in the Big Data Task One. Financial institutions' management teams require comprehensive insights into customer demographics, including age, income, dependents, and their interaction with various banking products. This mirrors the approach taken by companies like American Express, which skillfully profile their customers, who are likely to churn, by utilising advanced data solutions. Such a panoramic customer understanding is vital for any business operation in the banking field.

The dataset's cleanliness is remarkable, devoid of nulls or unknown entries, rendering preprocessing unnecessary. Since there is no time series data, there was no necessity of data types conversion. It wasn't until the deployment of the Random Forest Classification model that I proceeded with the encoding of categorical and numerical columns.

I kept a *'light'* theme with *#5E18EB* set as primary colour. I used the same colour for *'Attrited customers'* and *#A579FF* for *'Existing Customers'*. I designed a dashboard with three columns. The first column contains filters for all the dataset features and it performs comparisons with the target variable. The second column features the first chart, which is responsive and updates based on filter selections. The remaining charts are interactive or generated through a Random Forest classification matrix.

The initial project step involved creating a Jupyter notebook and pushing it to the following GitHub repository: **https://github.com/anag23039/BigData-DV.git**. I also created a suitable project title and included a Streamlit credit card icon to enhance the theme. Users have the option to upload a new dataset that must have the same number of columns and column names.

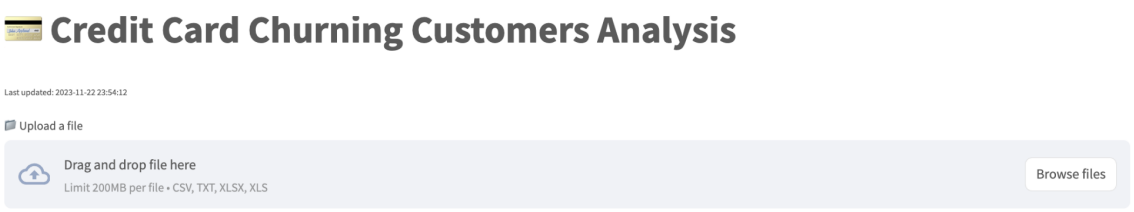


Figure 8: Upload button

I introduced filters to enhance interactivity with the dashboard and maintained a consistent Streamlit theme throughout the entirety of the project.

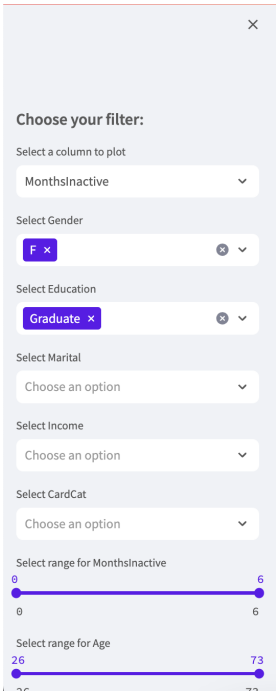


Figure 9: Filters

The first chart in column two is both interactive and responsive to filter selection, capable of displaying bar charts and histograms. Users can view data points for selected features when hovering, making the analysis concept easy to understand.

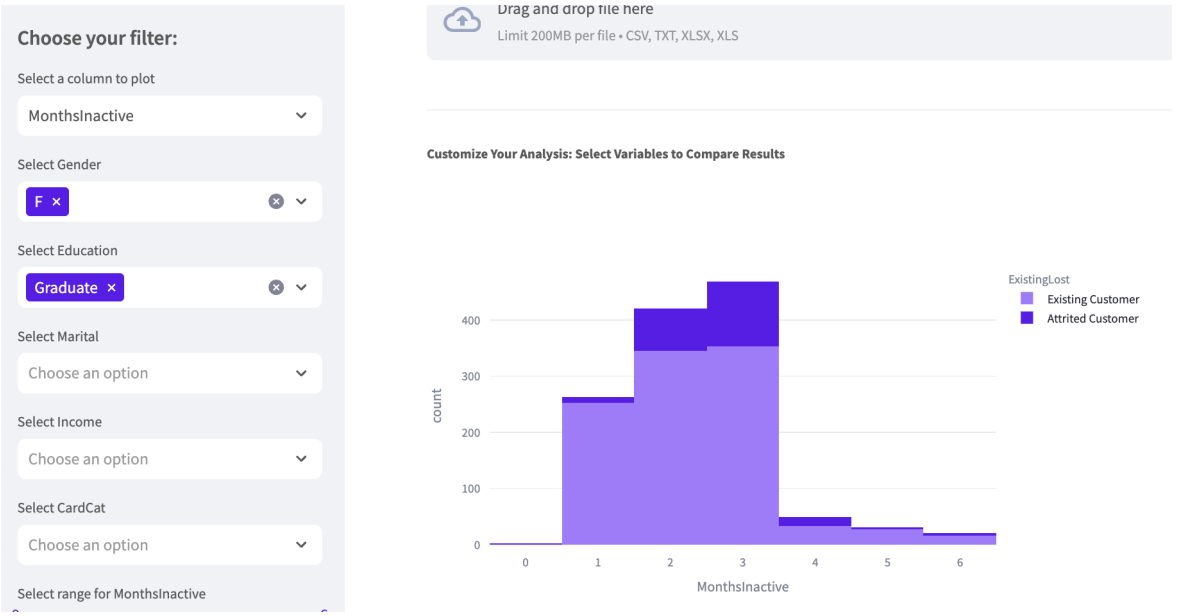


Figure 10: Customise Your Analysis: Select Variables to Compare Results

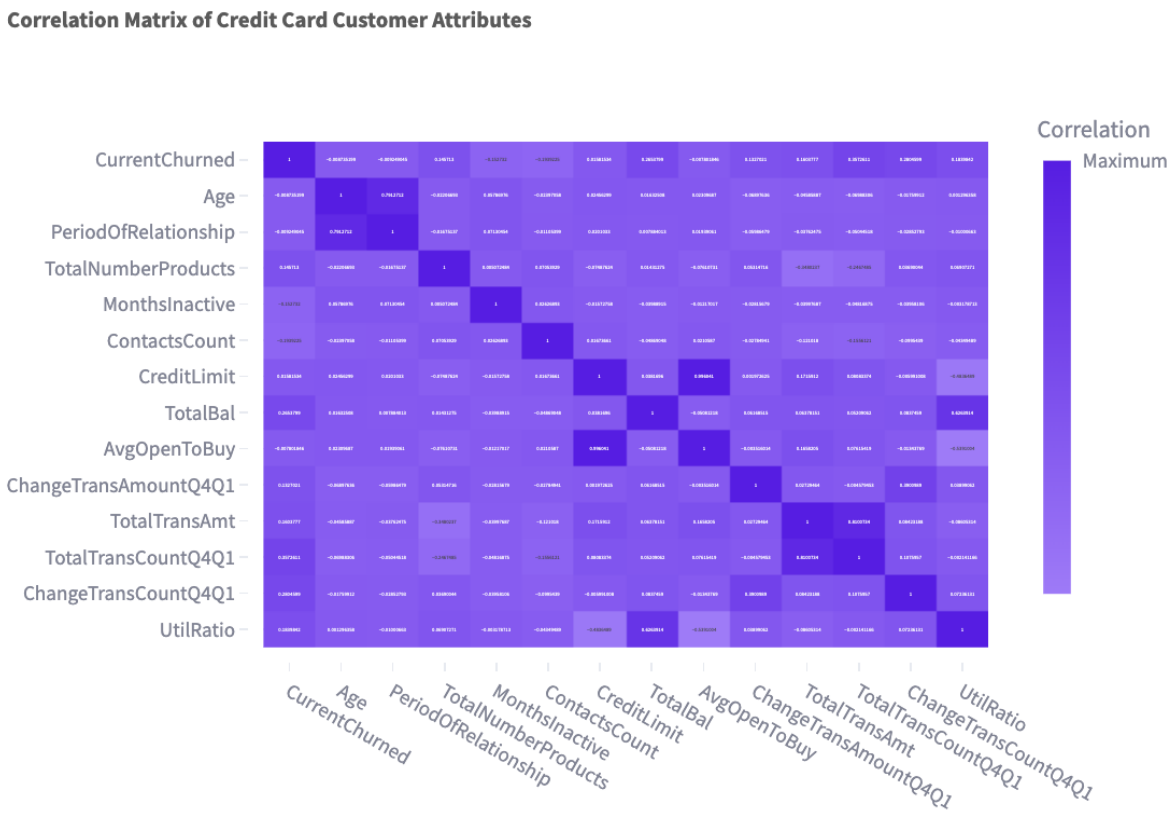


Figure 11: Correlation Matrix of Credit Card Customer Attributes

Additionally, a heatmap was incorporated into the project. This heatmap is interactive, displaying correlations between all numerical columns and the '*ExistingLost*' target variable. Users can select specific sections of the heatmap for detailed zoom-in.

In column two, a scatter plot was introduced, illustrating the connection between *'Months inactive,' 'Period Of Relationship,'* and the target variables *'Existing customers'* and *'Attrited customers'*. The plot effectively highlights that when customers remain inactive for two to three months with a bank relationship lasting between 20 and 47 months, there is a significant churn risk.

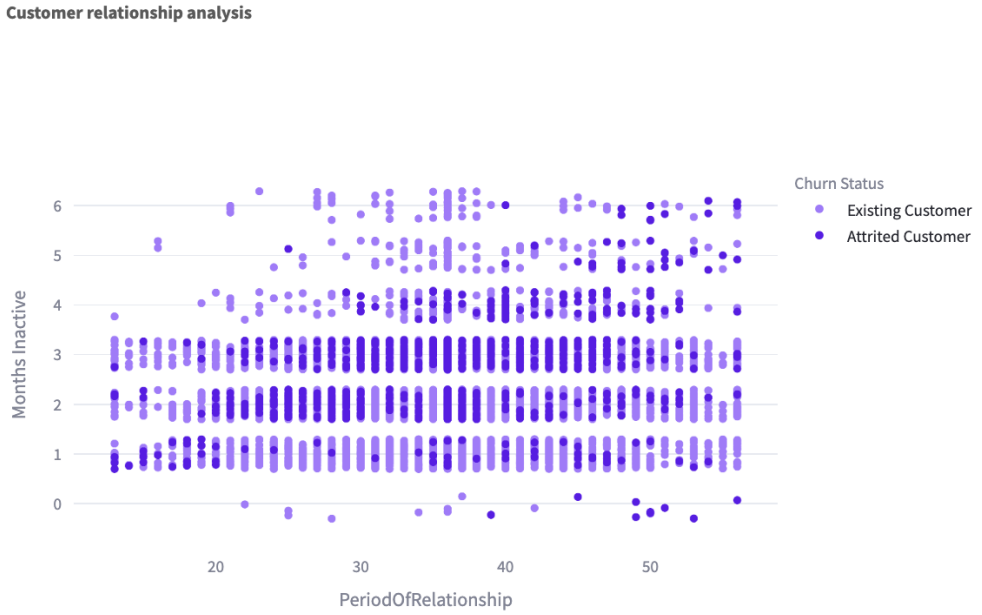


Figure 12: Customer relationship analysis

In column three, I placed a simple interactive chart displaying the distribution of *'ExistingLost'* against *Credit Card Type*. This chart is essential, because a user can infer that Blue cards are the most used cards and it is probably necessary to do some promotional marketing activities for *'Gold', 'Silver' and 'Platinum'* cards.

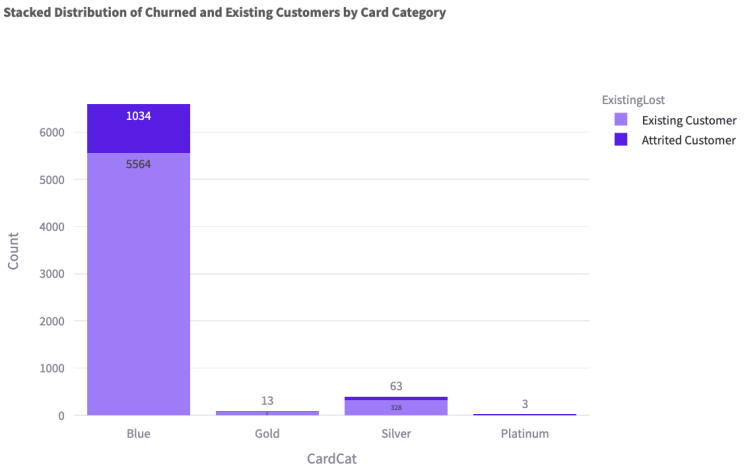


Figure 13: Stacked Distribution of Churned and Existing Customers by Card Category

In the final section, I included a Confusion Matrix based on Random Forest Classification and positioned the accuracy metric alongside it. This is crucial because it ensures that the model is readily available for prediction purposes when training similar datasets, making it a valuable resource.

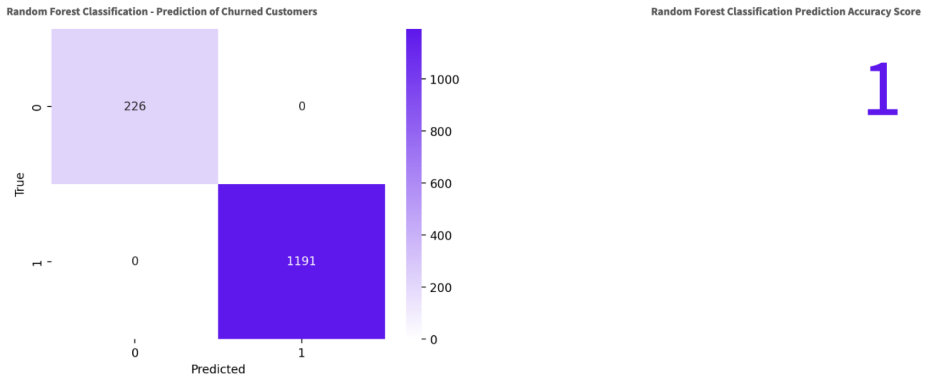


Figure 14: Random Forest Classification - Prediction of Churned Customers

During testing, I utilised the *.py* format to execute the Streamlit session in a web browser.

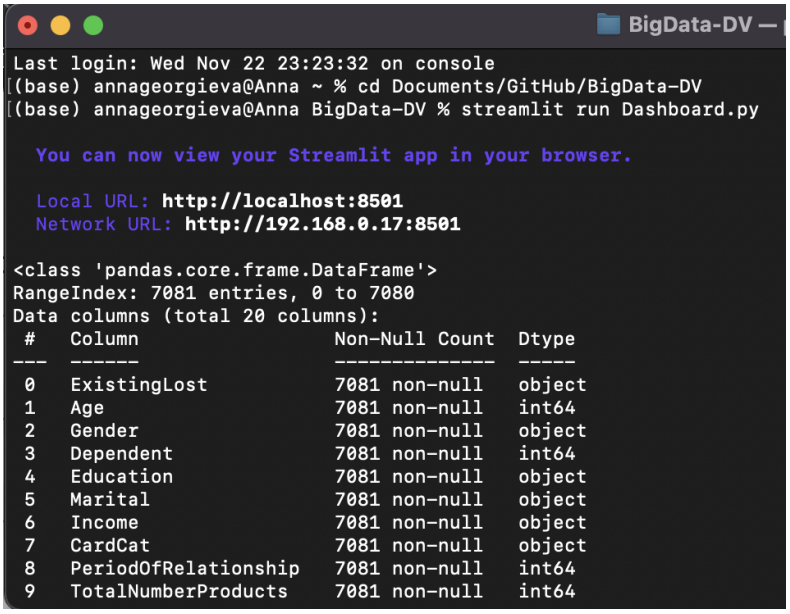


Figure 15: Terminal session - launching Streamlit Dashboard

I generated two files: *a config.toml* containing Streamlit theme settings, and a *requirements.txt* listing all essential libraries for the project. I could then deploy the application in Streamlit, by granting access to GitHub repository - BigData-DV.

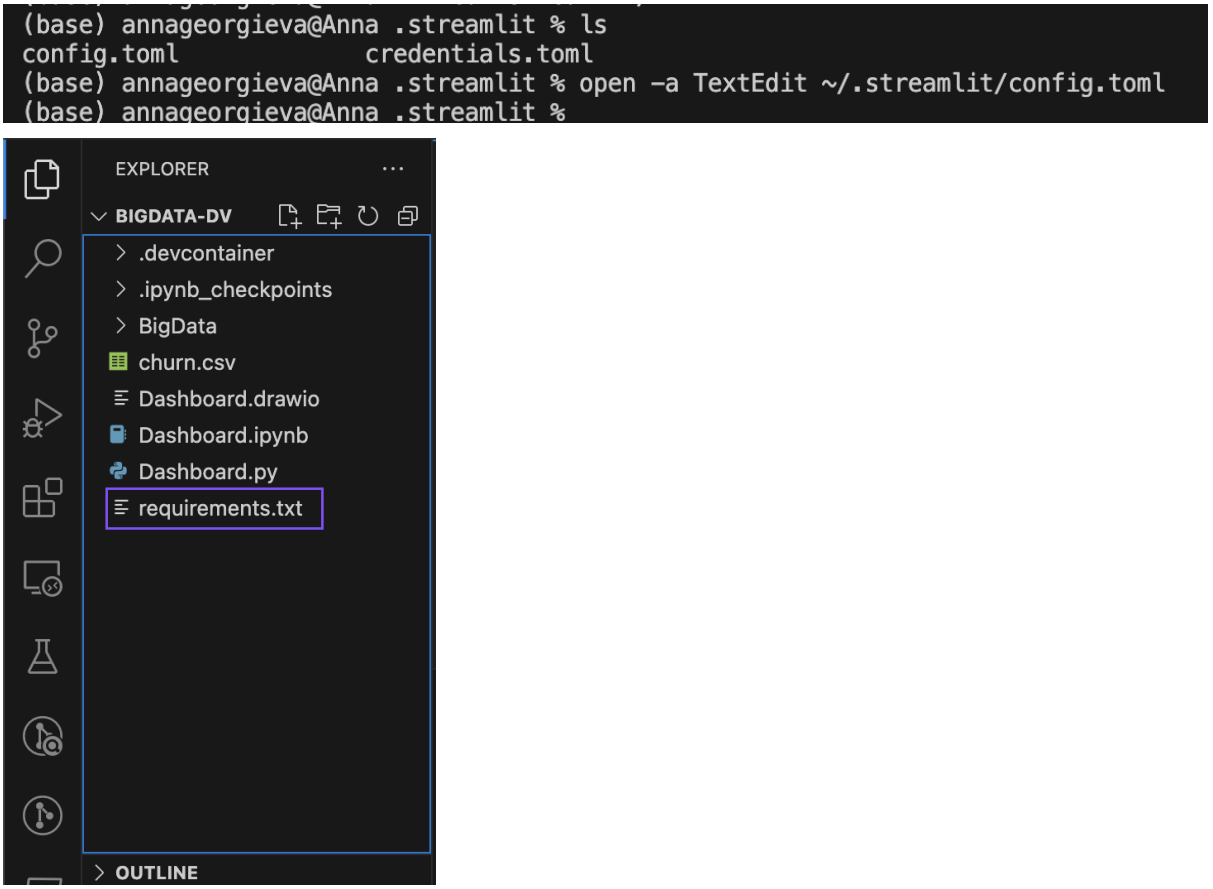


Figure 16: config.toml and requirements.txt created

App link - https://sbs23039.streamlit.app/

The dashboard includes various visualisations and interactive filters that allow users to shift through the data related to credit card transactions and customer behaviour. This dashboard is colour-coded to denote the strength of the correlation between variables. The visualisations are suitable for financial institutions to understand customer behaviour and predict churn, which can inform strategies to improve customer retention. The interactive elements suggest that the user can filter the data by various demographics and account information to explore different hypotheses or to focus on specific customer segments.

Reference list

Works Cited

Apache Storm. “Apache Storm.” *Storm.apache.org*, 2023, storm.apache.org/. Accessed 17 Nov. 2023.

Chen, Tsangyao. “Run SQL Queries from Linux Command Line.” *Www.youtube.com*, 2022, www.youtube.com/watch?v=iQYGaRQO3tQ. Accessed 15 Nov. 2023.

Cloudera. “Apache Storm.” *Cloudera*, 2023, www.cloudera.com/products/open-source/apache-hadoop/apache-storm.html. Accessed 17 Nov. 2023.

Code With Arjun. “Mysql Queries Using Terminal in Ubuntu | Linux | Mysql -u Root -p | Basic Queries in MySQL.” *Www.youtube.com*, 2021, www.youtube.com/watch?v=wQzng2eRdnw. Accessed 15 Nov. 2023.

Community Cloudera. “Sqoop2 Import from HDFS to MySQL Database Error.” *Community.cloudera.com*, 18 Dec. 2014, community.cloudera.com/t5/Support-Questions/Sqoop2-Import-from-HDFS-to-MySQL-database-error/m-p/22836. Accessed 16 Nov. 2023.

Costley, Jennifer, and Peter Lankford. *Big Data Cases in Banking and Securities a Report from the Front Lines Sponsored By*. 2014.

DataFlair Team. “Hadoop Architecture in Detail - HDFS, Yarn & MapReduce - DataFlair.” *DataFlair*, 28 Feb. 2019, data-flair.training/blogs/hadoop-architecture/. Accessed 15 Nov. 2023.

Duggal, Nikita. “Top 7 Benefits of Big Data & Analytics | Simplilearn.” *Simplilearn.com*, 13 Jan. 2022, www.simplilearn.com/benefits-of-big-data-and-analytics-article. Accessed 14 Nov. 2023.

Flink Architecture. “Flink Architecture.” *Nightlies.apache.org*, 2023, nightlies.apache.org/flink/flink-docs-master/docs/concepts/flink-architecture/. Accessed 17 Nov. 2023.

Goyal, S. (2021). Credit Card customers. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers [Accessed 22 Apr. 2023].

Guru99. “What Is Hive? Architecture & Modes.” *Www.guru99.com*, 25 Sept. 2023, www.guru99.com/introduction-hive.html#main. Accessed 16 Nov. 2023.

Larxel. “Trip Advisor Hotel Reviews.” *Www.kaggle.com*, 2020, www.kaggle.com/datasets/andrewmvd/trip-advisor-hotel-reviews. Accessed 14 Nov. 2023.

Lutkevich, Ben. “What Are the 3 V’s of Big Data? | Definition from TechTarget.” *WhatIs.com*, Mar. 2023, www.techtarget.com/whatis/definition/3Vs#:~:text=The%203%20V. Accessed 17 Nov. 2023.

micronimics. “Advantages and Disadvantages of HDFS and Traditional File Systems.” *Data Recovery in Ahmedabad, Best Data Recovery in India*, 24 Dec. 2022, www.micronicsindia.com/advantages-and-disadvantages-of-hdfs-and-traditional-file-systems/#:~:text=HDFS%20is%20a%20distributed%20file. Accessed 15 Nov. 2023.

MySQL. “MySQL :: MySQL 8.0 Reference Manual :: 16.11 Overview of MySQL Storage Engine Architecture.” *Dev.mysql.com*, 2023, dev.mysql.com/doc/refman/8.0/en/pluggable-storage-overview.html.

---. “MySQL :: MySQL HeatWave: OLTP.” *Www.mysql.com*, 2023, www.mysql.com/products/mysqlheatwave/oltp/#:~:text=MySQL%20powers%20the%20most%20demanding. Accessed 16 Nov. 2023.

O’ Neill, Eleanor. “10 Companies Using Big Data.” *Icas.com*, 5 Sept. 2019, www.icas.com/news/10-companies-using-big-data#:~:text=2. Accessed 15 Nov. 2023.

Patil, Prasad. “Retail Transactions Dataset.” *Www.kaggle.com*, 2023, www.kaggle.com/datasets/prasad22/retail-transactions-dataset. Accessed 14 Nov. 2023.

Programming is Fun. “Python Adidas Sales Dashboard Using Streamlit and Plotly-II.” *Www.youtube.com*, 2023, www.youtube.com/watch?v=6Eu2b34alsE&t=1380s. Accessed 22 Nov. 2023.

Quora. “Under What Circumstances Will a Hadoop/Hive Database Perform Slower (on Reads) than a Traditional RDBMS?” *Quora*, 2019, www.quora.com/Under-what-circumstances-will-a-Hadoop-Hive-database-perform-slower-on-reads-than-a-traditional-RDBMS. Accessed 16 Nov. 2023.

Reddy, Sandhya. “What Is Hadoop and Modules of Hadoop? Complete Overview.” *Medium*, 10 Jan. 2020, medium.com/quick-code/what-is-hadoop-and-modules-of-hadoop-complete-overview-2d0b501982e2. Accessed 15 Nov. 2023.

Rosencrance, Linda . “What Is HDFS? Hadoop Distributed File System Overview.” *SearchDataManagement*, 2021, www.techtarget.com/searchdatamanagement/definition/Hadoop-Distributed-File-System-HDFS.

Saraswat, Abhisheak. “PythonStreamlit/App.py at Main · AbhisheakSaraswat/PythonStreamlit.” *GitHub*, 2023, github.com/AbhisheakSaraswat/PythonStreamlit/blob/main/app.py. Accessed 22 Nov. 2023.

Turkington, et al. “Hadoop: Data Processing and Modelling.” *Ebscohost.com*, 2016, eds.s.ebscohost.com/eds/ebookviewer/ebook/ZTAyMG13d19fMTM0NTIwNF9fQU41?sid=7b0d96d8-d032-462b-a3a4-d5d0a15d0257%40redis&vid=1&format=EB&rid=1. Accessed 15 Nov. 2023.

TutorialsPoint. “Apache Storm - Cluster Architecture.” *Www.tutorialspoint.com*, 2023, www.tutorialspoint.com/apache_storm/apache_storm_cluster_architecture.htm#:~:text=Apache%20ZooKeeper%20is%20a%20service. Accessed 17 Nov. 2023.