**CCT College Dublin**

**Assessment Cover Page**

| **Module Title:** | Storage Solutions for big data |
| --- | --- |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | Dr. Muhammad Iqbal |
| **Student Full Name:** | Kavi Patak |
| **Student Number:** | sba22391 |
| **Assessment Due Date:** | 26 May 2024 |
| **Date of Submission:** | 26 May 2024 |

****

**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. |
| --- |

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Storage Solutions for Big Data:

Continuous Assessment 2

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

by Kavi Patak

sba22391

Storage Solutions for Big data

Lecturer: Dr. Muhammad Iqbal

CCT College, Dublin

May 26, 2024

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Table of Contents

[**Storage Solutions for Big Data: 2**](#_3iddv1ojtzjk)

[**Table of Contents 3**](#_ytyrg5mbj69k)

[Question 1: 4](#_usiluud1xtj5)

[Define Big Data and outline its key characteristics. Discuss the potential for banks to increase profits through big data processing and analysis. Identify three businesses that have successfully leveraged big data storage solutions in recent times. 4](#_l90b5kk1iqyh)

[Question 2: 4](#_6v3q9s8f4v4a)

[Provide the screenshots based on your VM (hduser@studentusername:~$) or Google cloud platform (we check the Google cloud has been used or not, if not used, zero marks) for the following processes 4](#_9r0rujli3652)

[a) Start all five processes of hadoop distributed file system 4](#_wpiu62odcn4e)

[b) Find the text dataset of size (500 MB at least) and create a folder on hadoop named CA2. Copy the dataset from VM to the hdfs directory (CA2) the data. 4](#_o57biulv9hxn)

[c) Execute Mapper and Reducer programmes by using loaded input file. 4](#_hkzik1kyrpsj)

[d) Display the frequency of each word obtained from the dataset 4](#_gb9yzkj45w4b)

[e) Download the output file from hadoop and upload it on Moodle. If you could not provide the screenshots for all commands or related files, no marks will be awarded. 4](#_bcug5tipbkkn)

[**References 6**](#_kya26suve4eo)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

### Question 1:

#### Define Big Data and outline its key characteristics.

Big Data may be defined in many ways, depending on who you ask. However, most definitions of big data reflect the growing technological ability to capture, aggregate, and process an ever increasing volume, velocity, and variety of data (Iqbal, n.d.). Volume, Velocity and Variety, have become known as the 3 V’s of Big Data. Big Data comes in a variety of forms including pictures, video, audio, and social media, and hence a variety of formats, whether structured, semi structured, or unstructured. Volume and variety refer to data that is too large or complex for storage and analysis by traditional methods. Since the advent of the Internet and proceeding developments in technologies like the Internet of Things(IoT), Sensors, GPS devices, Machine Learning (ML) and Artificial Intelligence (AI), the volume and variety of data being produced is on an upward and seemingly unbound trajectory (Agnellutti, 2014). This data is being produced at an exponential rate, with velocity referring to this, and the requirement for its instant processing for immediate use and inferred insight. Other interpretations of Big Data include the addition of veracity and value, with veracity referring to the messiness and trustworthiness of data, while value refers to the potential insight or gain, hidden within the data (Yaqoob et al., 2016).

#### Discuss the potential for banks to increase profits through big data processing and analysis.

The banking industry, like most others, stands to benefit greatly from the storage, processing and analysis of big data.

It is an extremely competitive industry, with most banks offering similar financial products and services. The opportunity exists for banks to harness big data through their daily transactions, e-banking, real-time feeds, social media posts and customer service records, and from this develop analytic models to optimise outcomes and derive insights for a competitive advantage (Bedeley, 2014). Specific areas in which banks may benefit include through customer segmentation and personalised services, including target marketing campaigns. By leveraging AI for the analysis of transactional data, banks can improve their fraud detection capabilities as well as develop models for advanced risk assessments and credit scoring, for loan and service approvals. Additionally, predictive models may aid in analysing the financial landscape for economic indicators thereby aiding informed investment decisions.Through the processing and analysis of big data, banks may improve every day operational efficiency, including optimising processes, reducing expenditure and resource management. Finally, the analysis of big data such as social media would lead to insights into customer behaviour, brand loyalty, competitor actions, customer services interactions and customer experiences through sentiment analysis.

#### Identify three businesses that have successfully leveraged big data storage solutions in recent times.

The first business that not only leverages big data storage solutions, but offers its own is Google. Google uses Big data to better understand its users and to improve its services (Kumari, 2021). Google also -provides storage solutions through google cloud platform and workspace .

Amazon is another powerhouse in the e-commerce sector that both utilises and provides Big Data solutions. Amazon collects and stores boundless data related to products and customers (Kumari, 2021). They too provide a cloud computing platform with a diverse array of services through Amazon Web Services (AWS)(Emergen Research, https://www.emergenresearch.com, 2024).

A third company leveraging big data storage solutions would be IBM. IBM is a global technology and consulting giant that has played an influential role in the ever evolving technological landscape (Emergen Research, https://www.emergenresearch.com, 2024). They are one of the biggest vendors of big data related products and services, with their storage solutions known for their scalability, integrated data management and analytics capabilities (www.linkedin.com, n.d.).

### Question 2: Word Count

#### Provide the screenshots based on your VM (hduser@studentusername:~$) or Google cloud platform (we check the Google cloud has been used or not, if not used, zero marks) for the following processes

#### a) Start all five processes of hadoop distributed file system

# Open Terminal

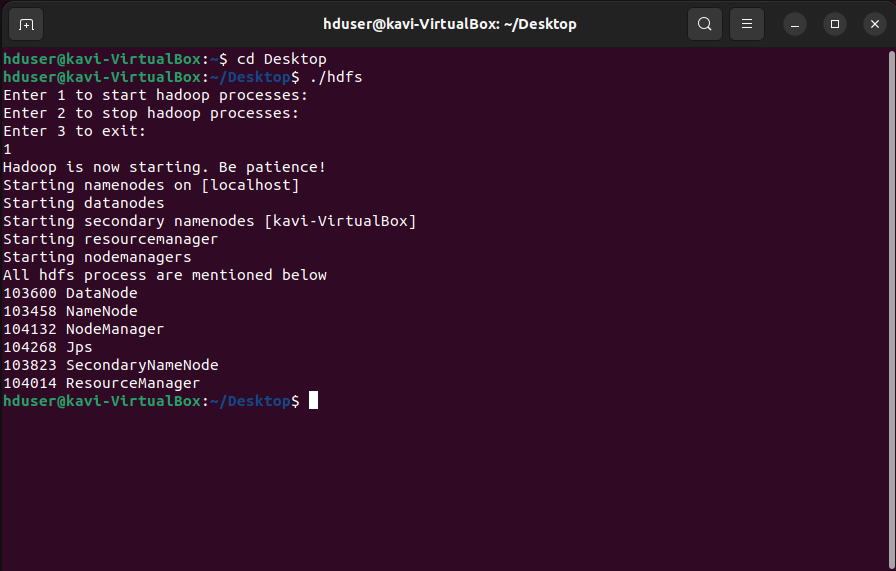
Ctrl + Alt + t

# Change to Desktop

Cd Desktop

# Start HDFS

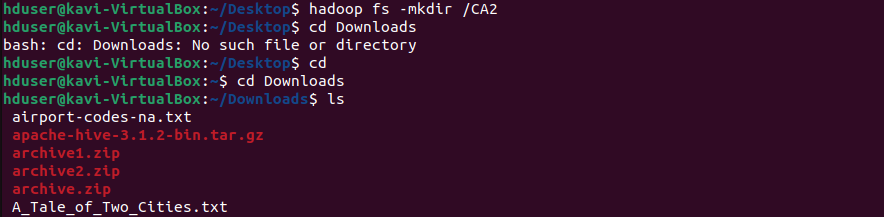
./hdfs



#### b) Find the text dataset of size (500 MB at least) and create a folder on hadoop named CA2. Copy the dataset from VM to the hdfs directory (CA2) the data.

# Creating a new folder

hadoop fs -mkdir /CA2



# Change Directory

cd

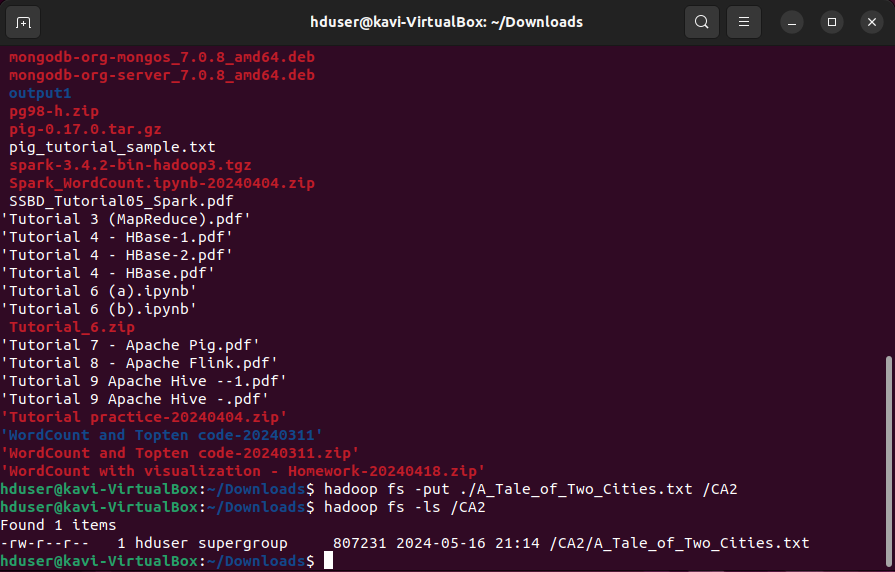
cd Downloads

# Move the text file from Downloads to CA2

hadoop fs -put ./A\_Tale\_of\_Two\_Cities.txt /CA2

# List the contents of CA2 to confirm move

hadoop fs -ls /CA2



#### c) Execute Mapper and Reducer programmes by using loaded input file.

# Change Directory

cd

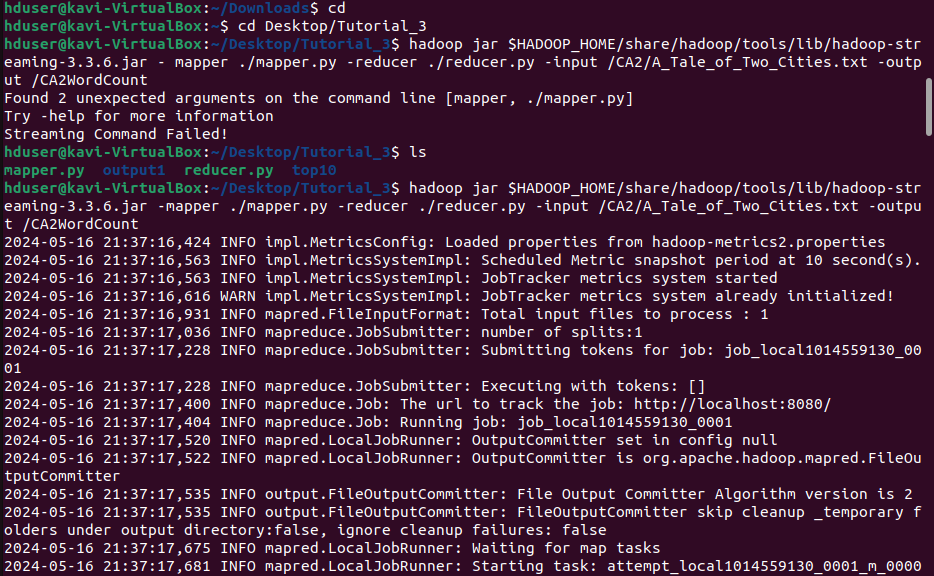
cd Desktop/Tutorial\_3

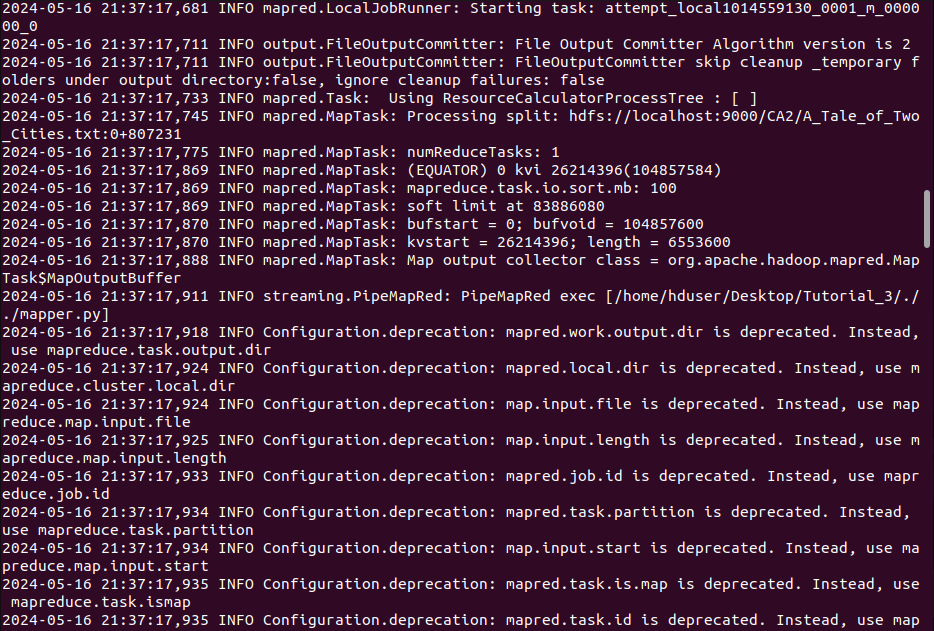
# Confirming mapper and reducer files are listed

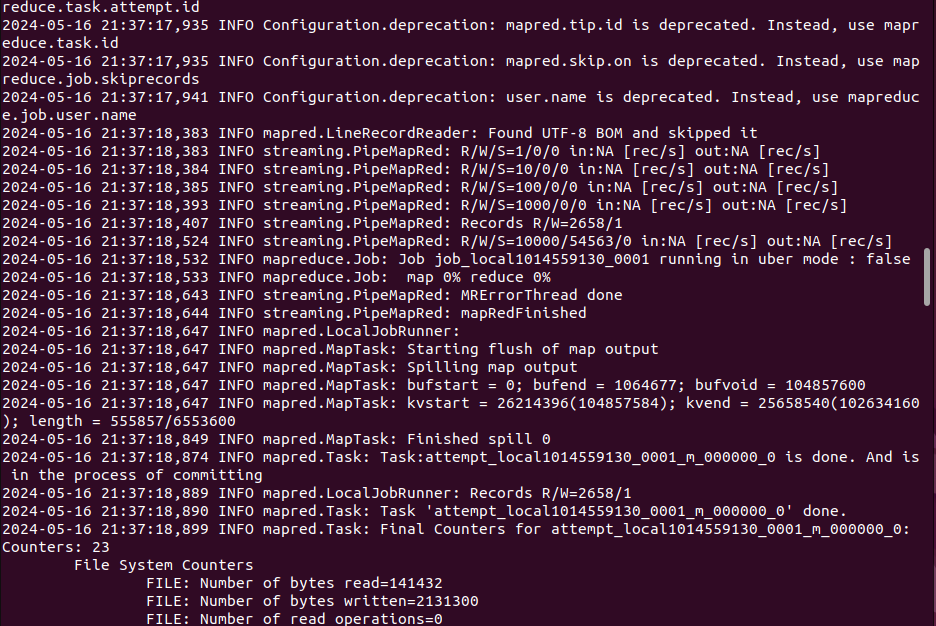
ls

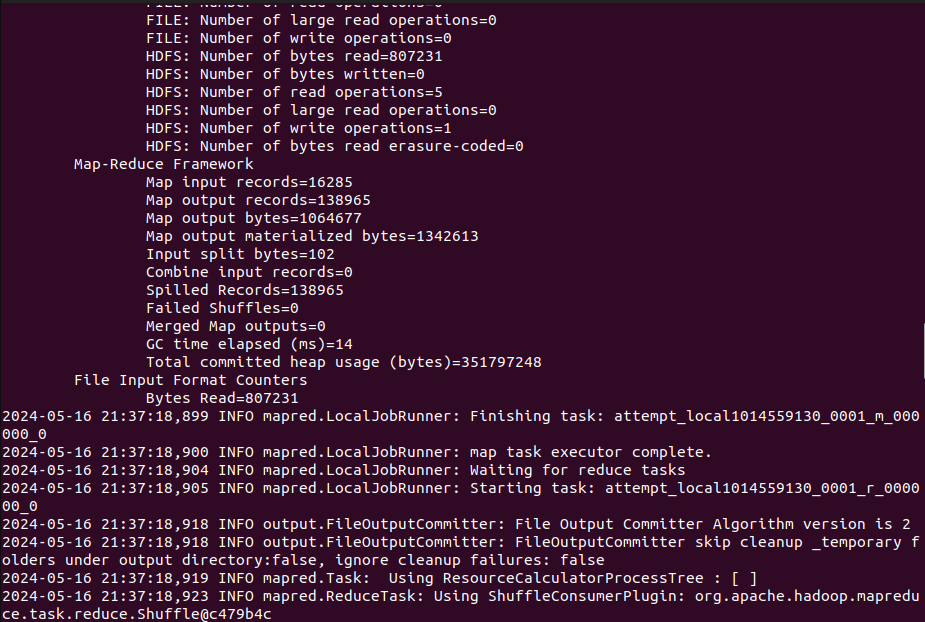
# Initiating Hadoop Streaming

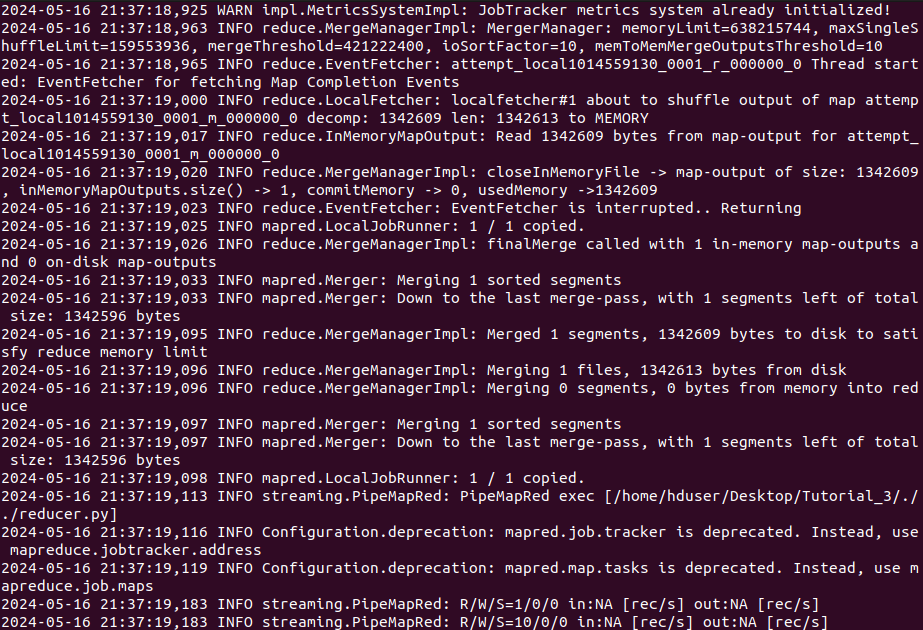
Hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -mapper ./mapper.py -reducer ./reducer.py -input /CA2/A\_Tale\_of\_Two\_Cities.txt -output /CA2WordCount

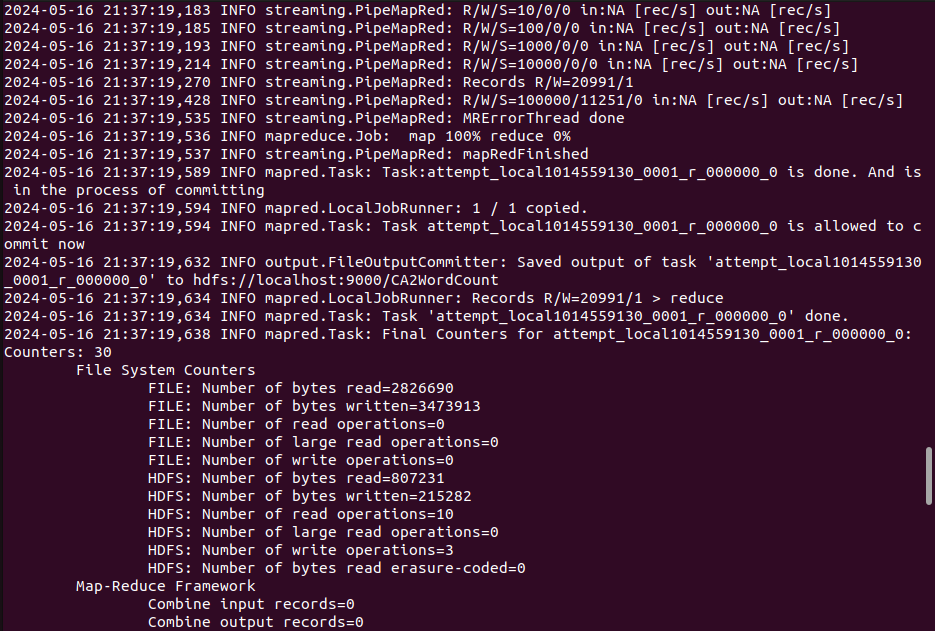


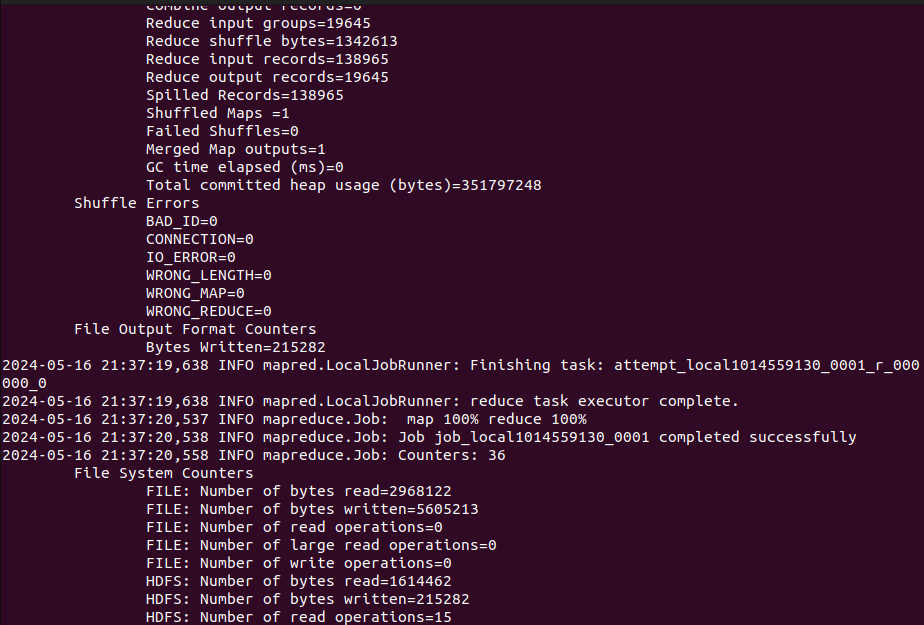


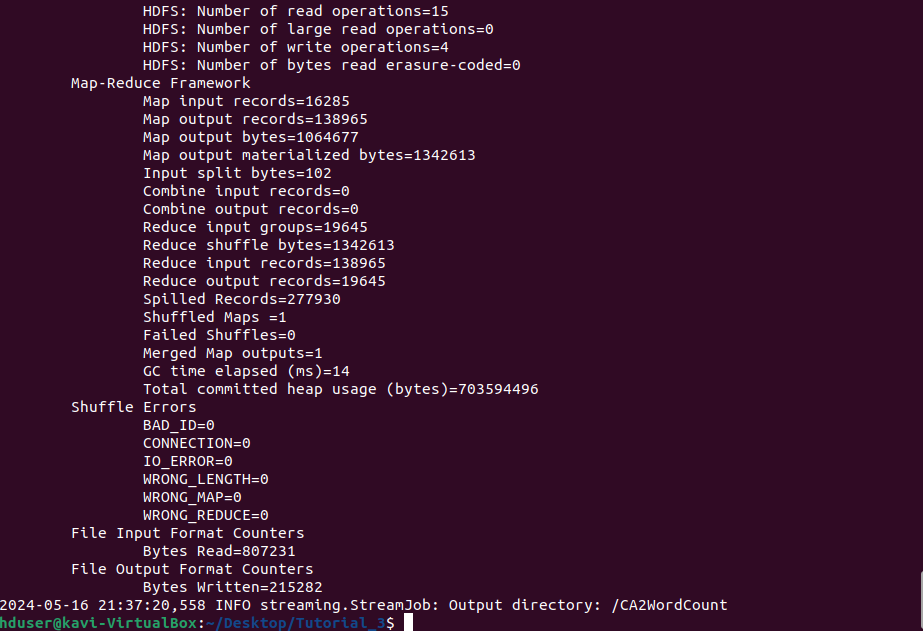












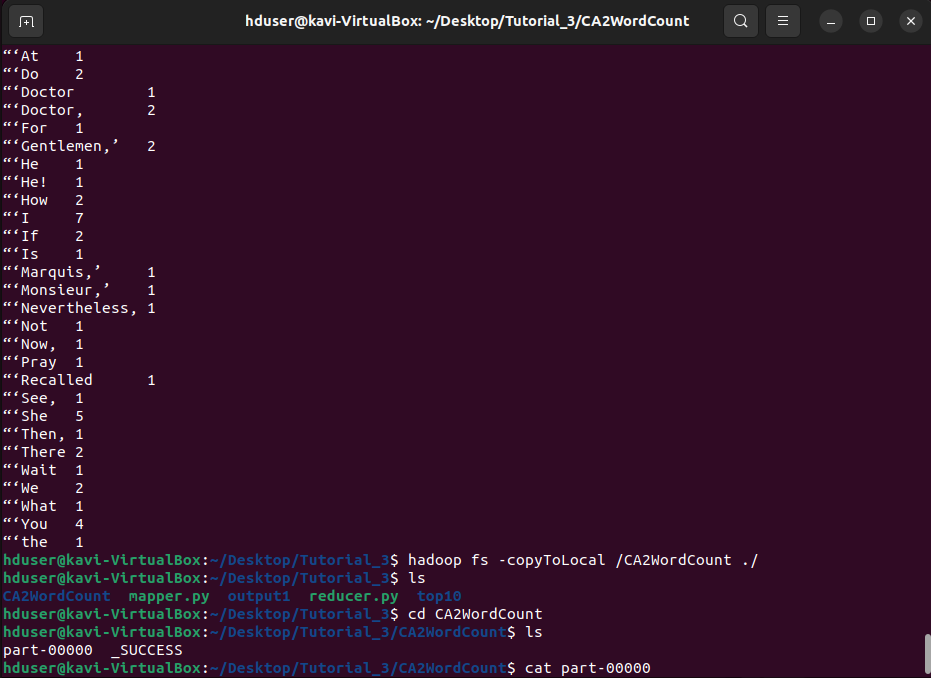
#### d) Display the frequency of each word obtained from the dataset

# Checking contents of CA2WordCount (output file)

hadoop fs -ls /CA2WordCount

# Checking word count

hadoop fs -cat /CA2WordCount/part-00000



#### e) Download the output file from hadoop and upload it on Moodle. If you could not provide the screenshots for all commands or related files, no marks will be awarded.

# Copying the CA2WordCount (output file) to local drive

Hadoop fs -copyToLocal /CA2WordCount ./

# Listing contents of Tutorial\_3 folder to confirm copy

ls

# Changing directory to CA2WordCount

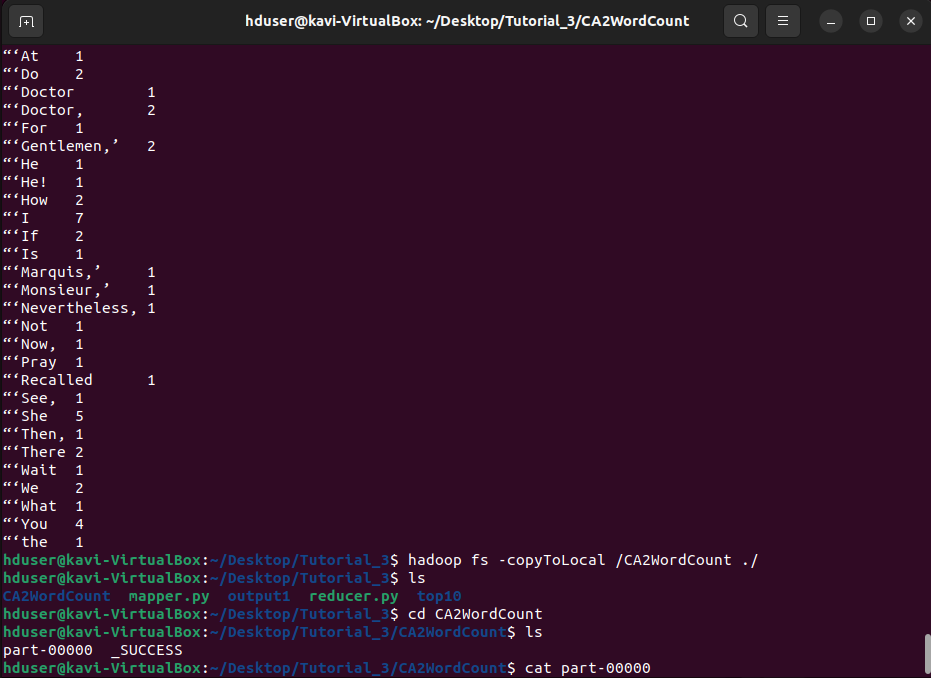
Cd CA2WordCount

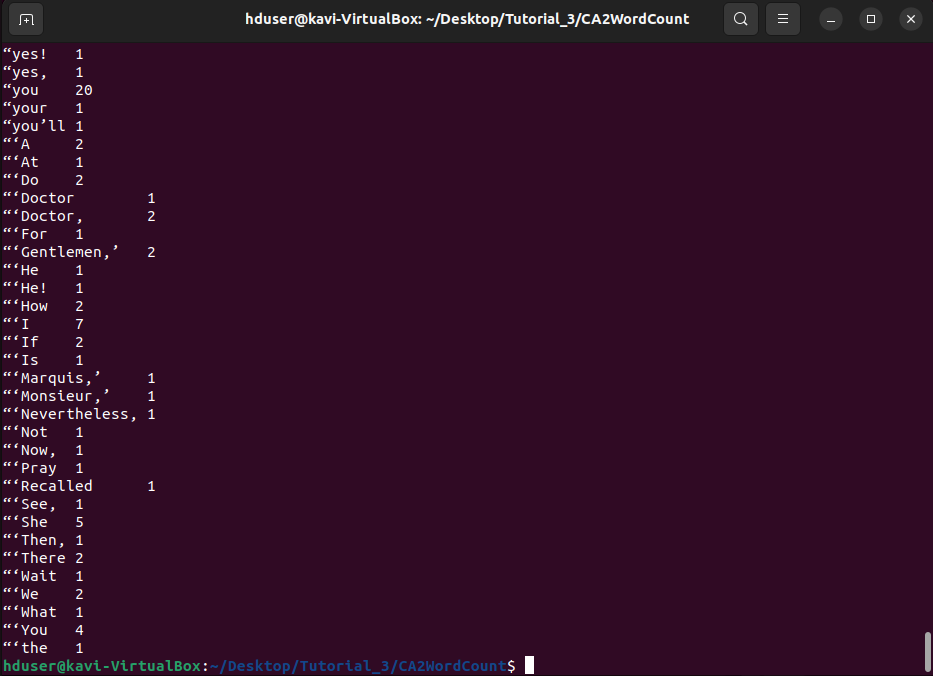
# Listing contents of CA2WordCount

ls

# Checking word count

cat part-00000





#### 

#### 

### Question 2: Top Ten

#### Provide the screenshots based on your VM (hduser@studentusername:~$) or Google cloud platform (we check the Google cloud has been used or not, if not used, zero marks) for the following processes

#### a) Start all five processes of hadoop distributed file system

# Open Terminal

Ctrl + Alt + t

# Change to Desktop

Cd Desktop

# Start HDFS

./hdfs

#### b) Find the text dataset of size (500 MB at least) and create a folder on hadoop named CA2. Copy the dataset from VM to the hdfs directory (CA2) the data.

# Creating a new folder

hadoop fs -mkdir /CA2

# Change Directory

cd

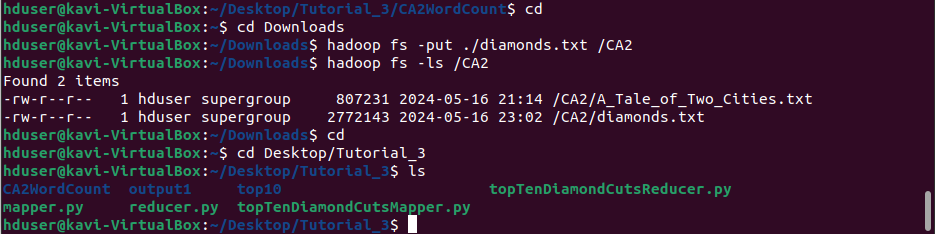
cd Downloads

# Move the text file from Downloads to CA2

hadoop fs -put ./diamonds.txt /CA2

# List the contents of CA2 to confirm move

hadoop fs -ls /CA2



#### c) Execute Mapper and Reducer programmes by using loaded input file.

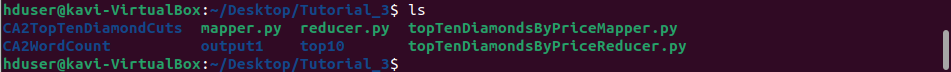
# Change Directory

cd

cd Desktop/Tutorial\_3

# Confirming mapper and reducer files are listed

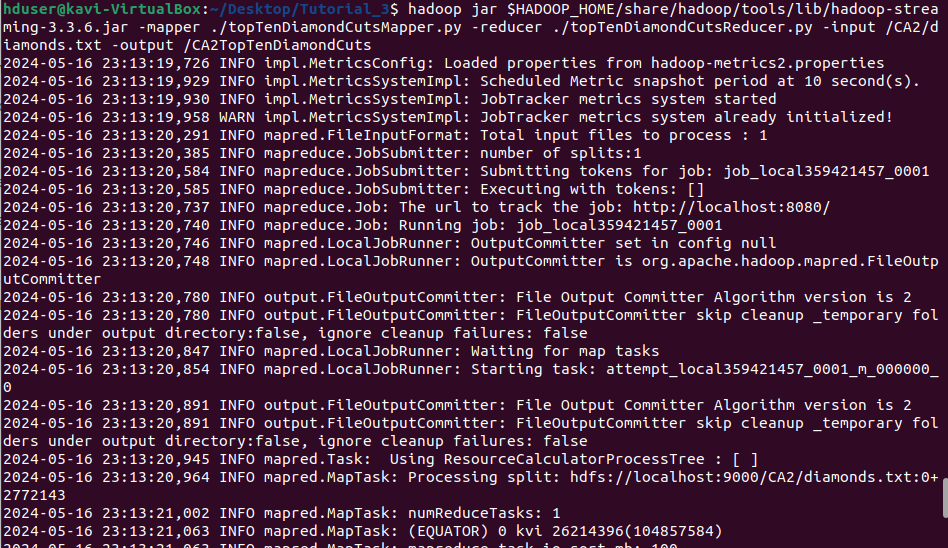
ls



# Initiating Hadoop Streaming

Hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-3.3.6.jar -mapper ./topTenDiamondCutsMapper.py -reducer ./topTenDiamondCutsReducer.py -input /CA2/A\_Tale\_of\_Two\_Cities.txt -output /CA2TopTenDiamondCuts





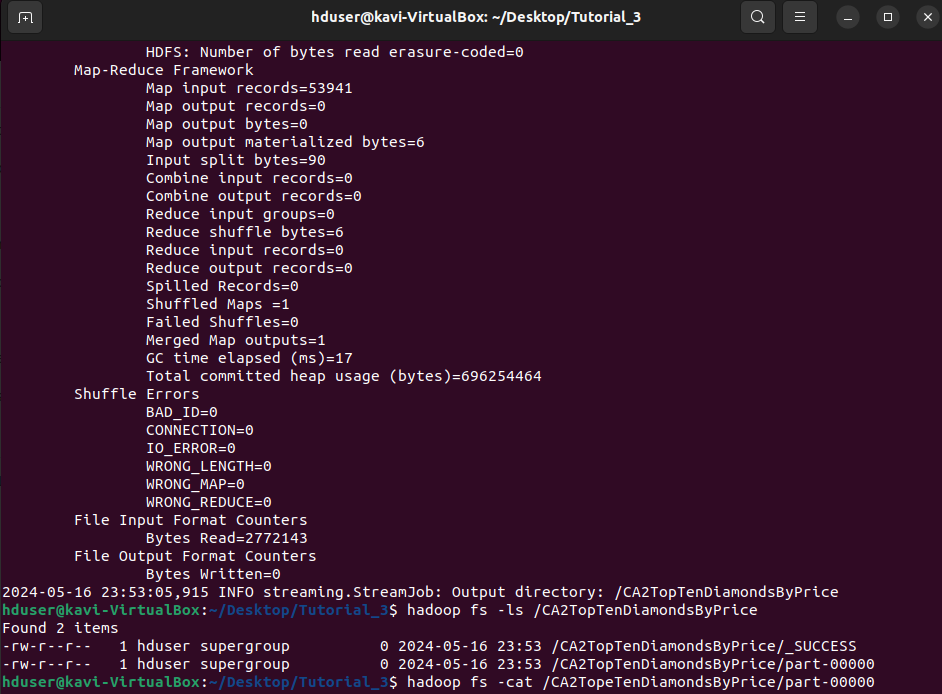
#### d) Display the frequency of each word obtained from the dataset

# Checking contents of CA2TopTenDiamondsByPrice (output file)

hadoop fs -ls /CA2TopTenDiamondsByPrice

# Checking word count

hadoop fs -cat /CA2TopTenDiamondsByPrice/part-00000



#### e) Download the output file from hadoop and upload it on Moodle. If you could not provide the screenshots for all commands or related files, no marks will be awarded.

# Copying the CA2WordCount (output file) to local drive

hadoop fs -copyToLocal /CA2TopTenDiamondsByPrice ./

# Listing contents of Tutorial\_3 folder to confirm copy

ls

# Changing directory to CA2TopTenDiamondsByPrice

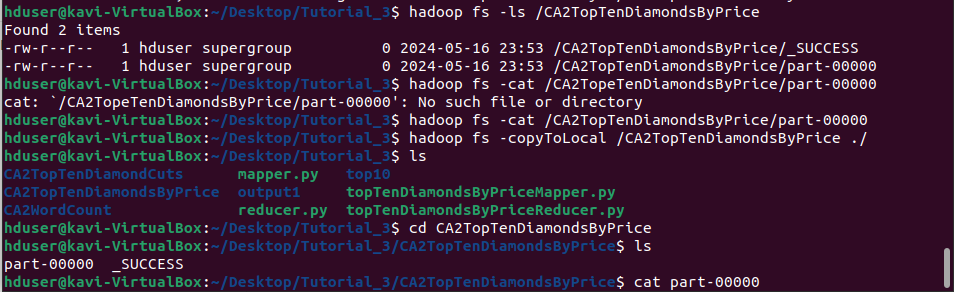
cd CA2TopTenDiamondsByPrice

# Listing contents of CA2TopTenDiamondsByPrice

ls

# Checking top ten

cat part-00000



### Question 3:

#### Discuss and demonstrate a comparison of MySQL and Apache Hive based on the architecture and performance.

MySQL is a commonly used open-source relational database management system (RDBMS), that supports predefined data types like date and float. MySQL provides Structured Query Language (SQL) statements including DDL (Data Definition Language) and DML (Data Manipulation Language). DDL provides the ability to define, create and modify database objects such as tables, indexes, and views. DML allows for manipulating data in a database, like inserting, updating, and deleting records. MySQL supports secondary indexing, which is a way to efficiently access records in a database without the use of the primary key. It also offers user-defined functions and the integration of map-reduce. MySQL follows ACID properties in Atomicity, Consistency, Isolation, and Durability (db-engines.com, n.d.). Other features include sharding or horizontal partitioning that separates large databases into smaller, faster, more easily manageable database shards as well as various replication methods.

Apache Hive on the other hand provides data warehouse functionality, built on top of Hadoop, for querying and managing large distributed datasets. It is a system or framework for managing and querying unstructured data as if it was structured. With regards to data types, Hive is slightly more limited and does not support spatial data (Quora, n.d.). This can however be facilitated with the use of external libraries. Similarly to MySQL, Hive offers SQL like syntax through HiveQL (HQL). Hive too supports secondary indexing and user defined functions to manipulate data and support data-mining, while using map reduce for execution (Iqbal, n.d.). Apache Hives architecture consists of three core parts in Hive Clients, Hive Services and Hive Storage and Computing, and seven main components in the Metastore, Driver, Command Line Interface (VLI), Web Interface, Thrift Server, JDBC (Java database connectivity) and ODBC (Open Database Connectivity) (www.guru99.com, n.d.).

For communicating with different applications, Hive provides different drivers.Thrift client is provided for communication with Thrift based applications, JDBC is provided for Java applications, with ODBC drivers provided for other applications. These Clients and drivers themselves communicate with Hive server within Hive Services.

Client interactions with Hive including query related operations can be performed through Hive Services. DDL operations are performed through CLI with all drivers communicating with Hive server and to the main driver in Hiver services. This main driver communicates with all applications and will process different requests to the Meta store and field systems for further processing.

Metastore, File system, Job Client and other services then communicate with Hive storage. Hive's Meta storage database stores Metadata information of tables that are created in Hive, while query results and any data loaded in tables is stored in HDFS (www.guru99.com, n.d.).

With regards to performance, both tools perform similar actions in retrieving data but in very different ways. Hive is designed as a convenient interface for querying data within HDFS while MySQL is intended for online operations requiring many reads and writes. Hive applies serialisation and deserialisation adaptors which allow users to redefine tables to match data without actually touching the data using a method called ‘schema on read’ (Iqbal, n.d.). This makes it less suitable for online tasks which require heavy read and write traffic (Rathbone, 2015). MySQL on the other hand defines table schemas before adding data using ‘schema on write’, which allows for data reading and writing in an optimal way. Hive provides eventual consistency while MySQL offers immediate consistency. Hive is better utilised when performing complex querying and analytics on large datasets already stored on Hadoop. It offers extensibility, interoperability and performance through query optimisation and hash joining.

MySQL performs better on smaller datasets, for frequently updating and modifying records and for online transactions (Rathbone, 2015).

#### 

#### 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).

Setting up Hive

# Open Terminal

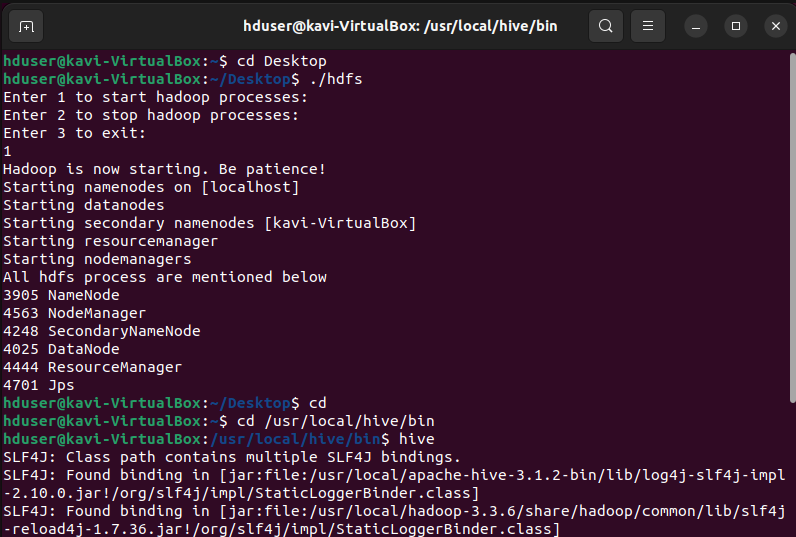
Ctrl + Alt + t

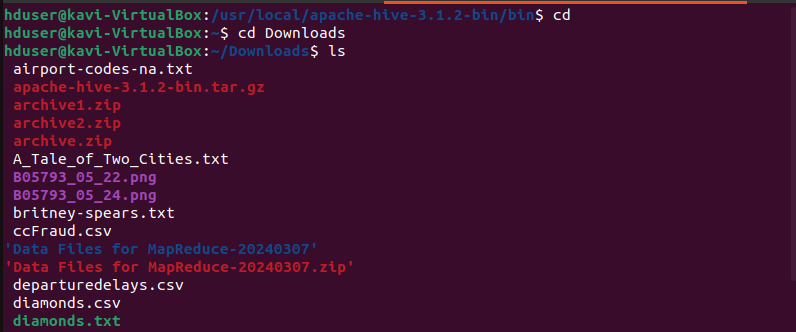
# Change to Desktop

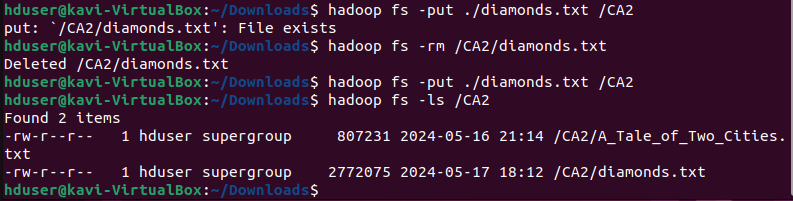
Cd Desktop

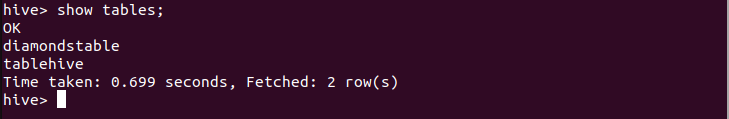
# Start HDFS

./hdfs

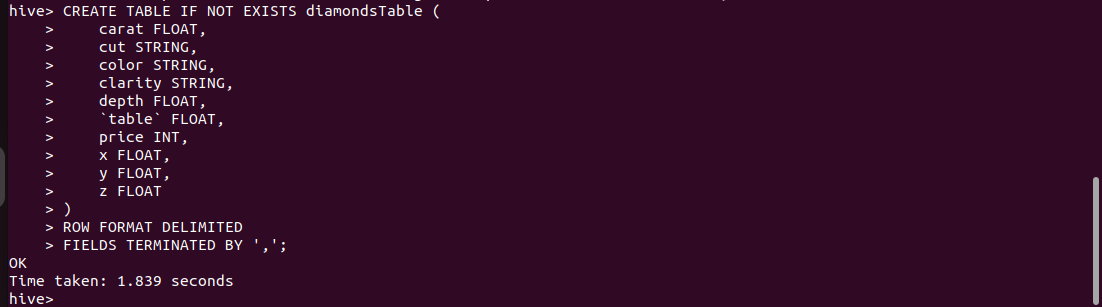




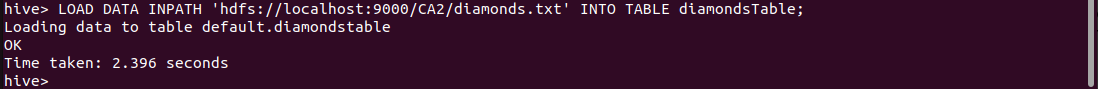




CREATE TABLE IF NOT EXISTs diamondsTable (carat FLOAT, cut STRING, color STRING, clarity STRING, depth FLOAT, ‘table’ FLOAT, price INT, x FLOAT, y FLOAT, z FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';



LOAD DATA INPATH 'hdfs://localhost:9000/CA2/diamonds.txt' INTO TABLE diamondsTable;



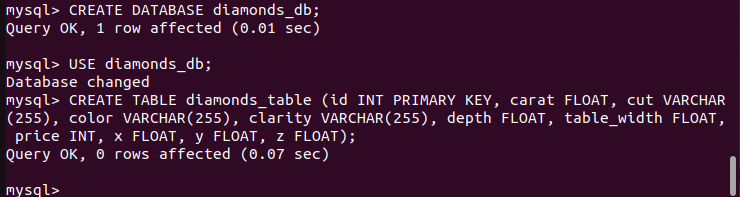
##### 

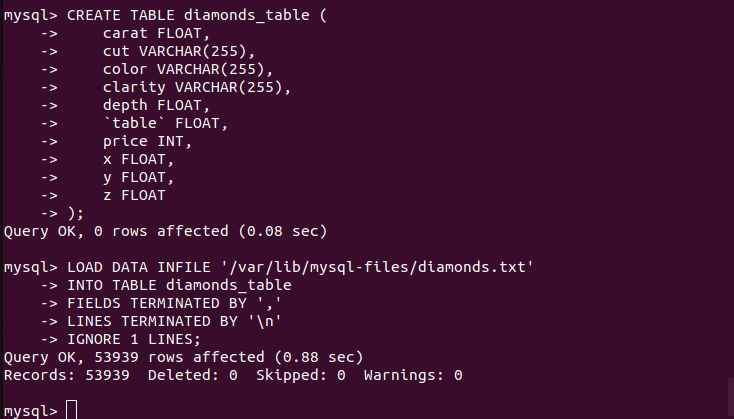
##### 

##### 

##### 

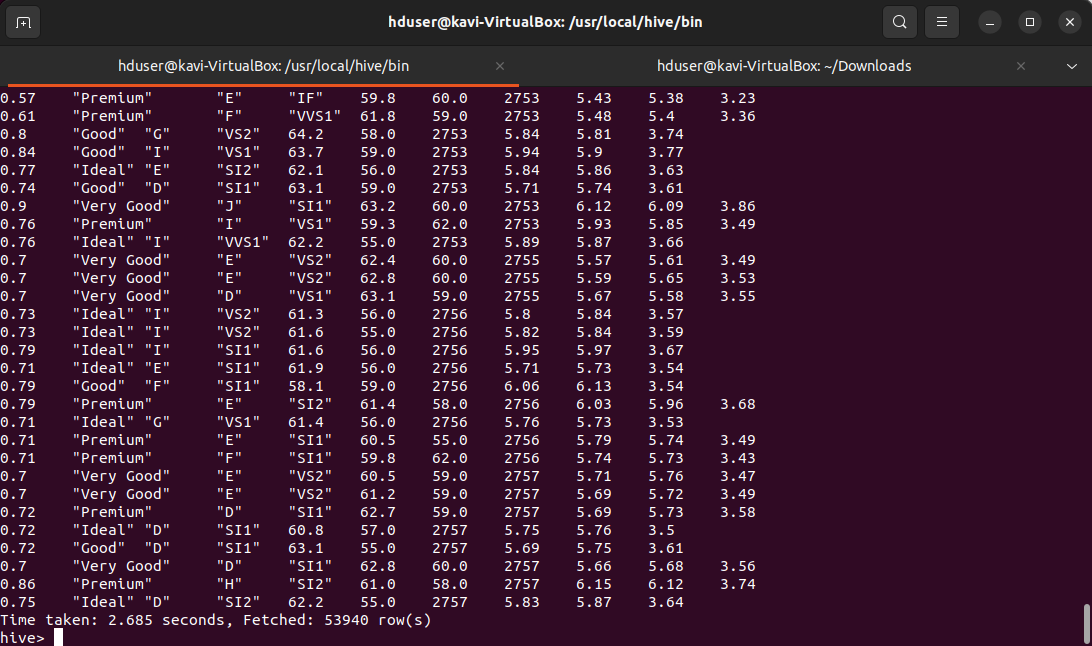
Setting up MySQL Database and Table





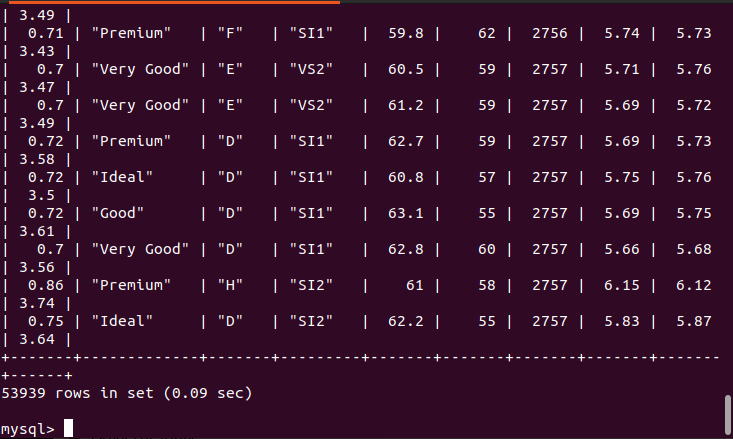
Select All Comparison

SELECT \* FROM diamondsTable;



SELECT \* FROM diamonds\_table;



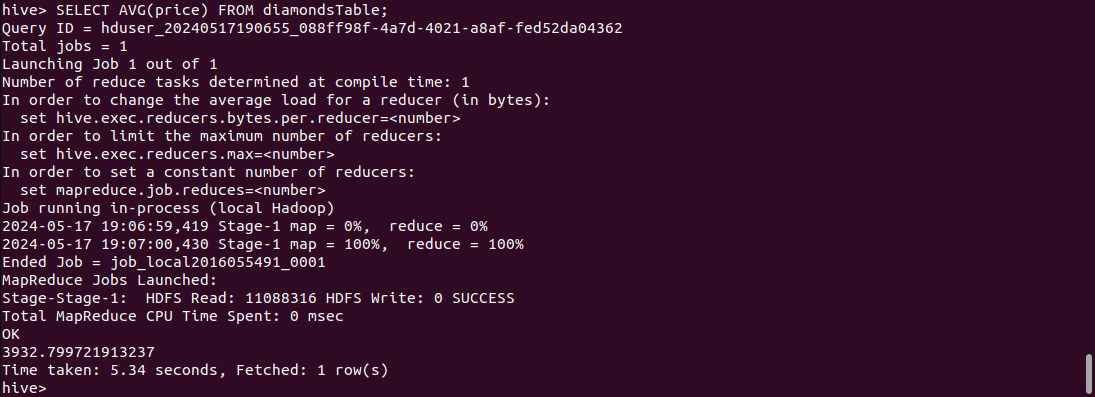


##### 

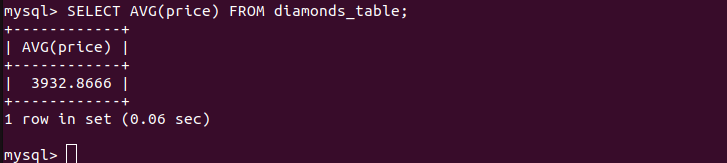
Aggregation Comparison

SELECT AVG(price) FROM diamondsTable;

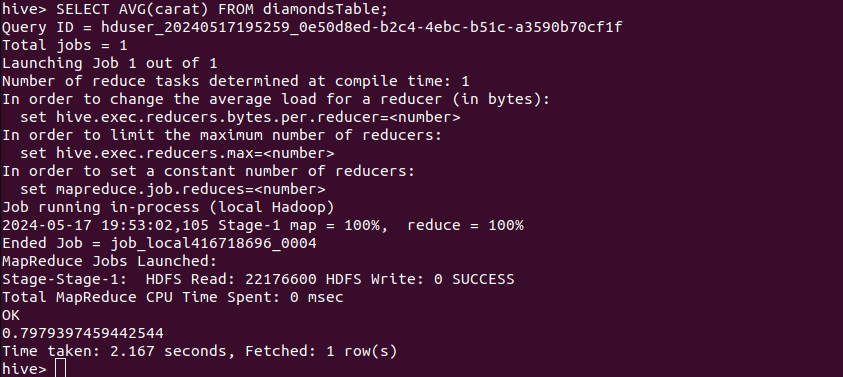




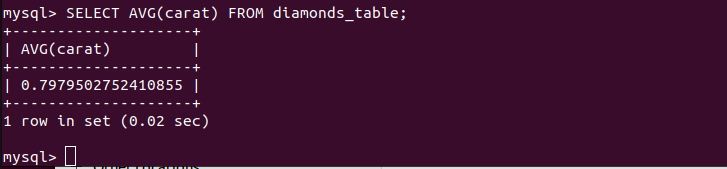
SELECT AVG(price) FROM diamonds\_table;



SELECT AVG(carat) FROM diamondsTable;

****

SELECT AVG(carat) FROM diamonds\_table;

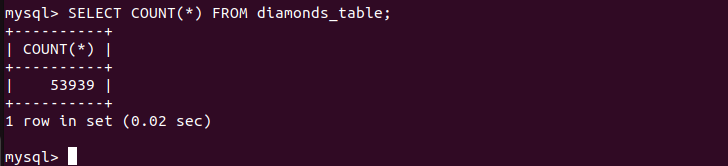


SELECT COUNT(\*) FROM diamondsTable;

##### 

##### 

SELECT COUNT(\*) FROM diamonds\_table;

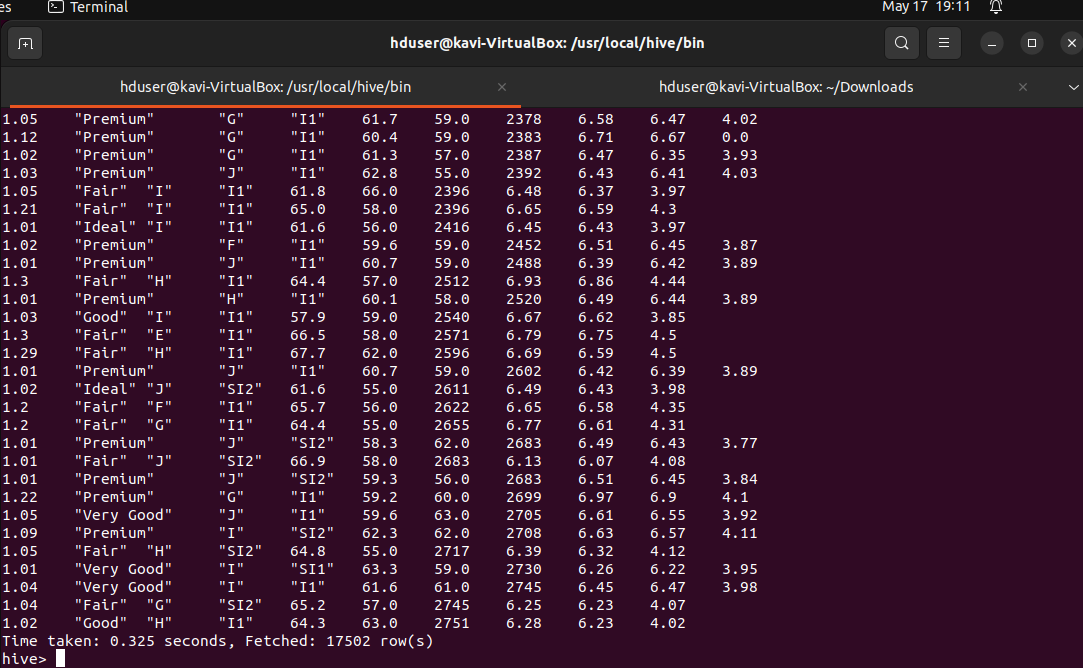


##### 

##### 

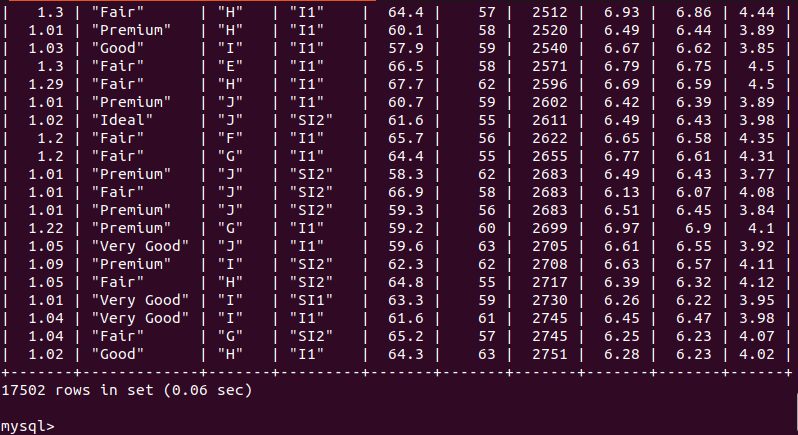
##### Filtering Comparison

SELECT \* FROM diamondsTable WHERE carat > 1.0;



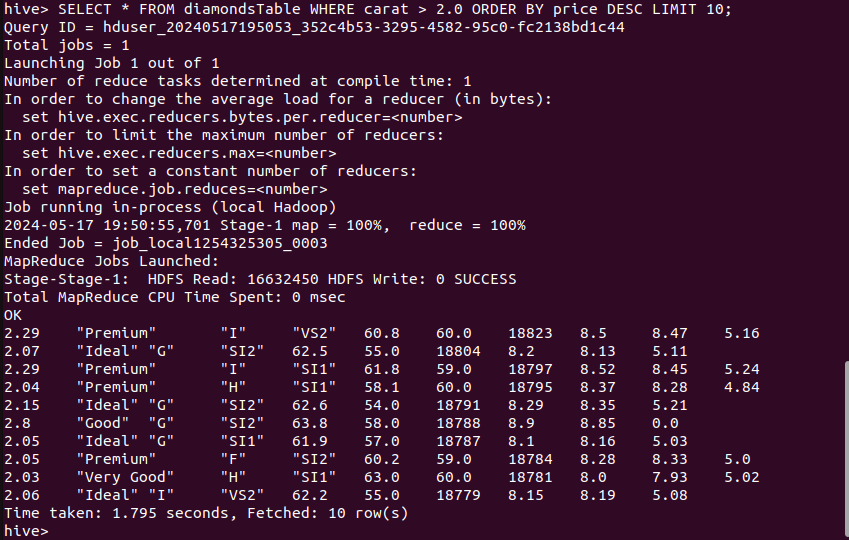
SELECT \* FROM diamonds\_table WHERE carat > 1.0;



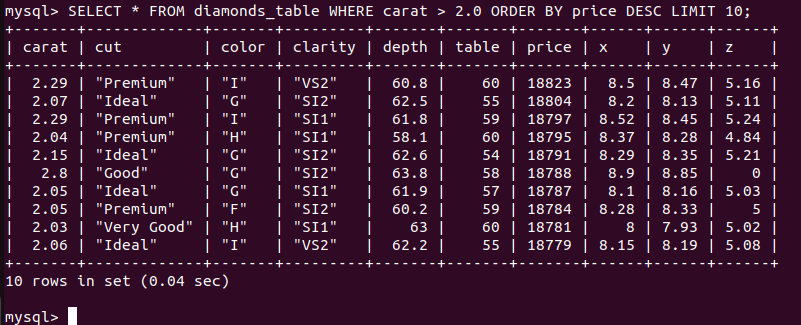


SELECT \* FROM diamondsTable WHERE carat > 2.0 ORDER BY price DESC LIMIT 10;



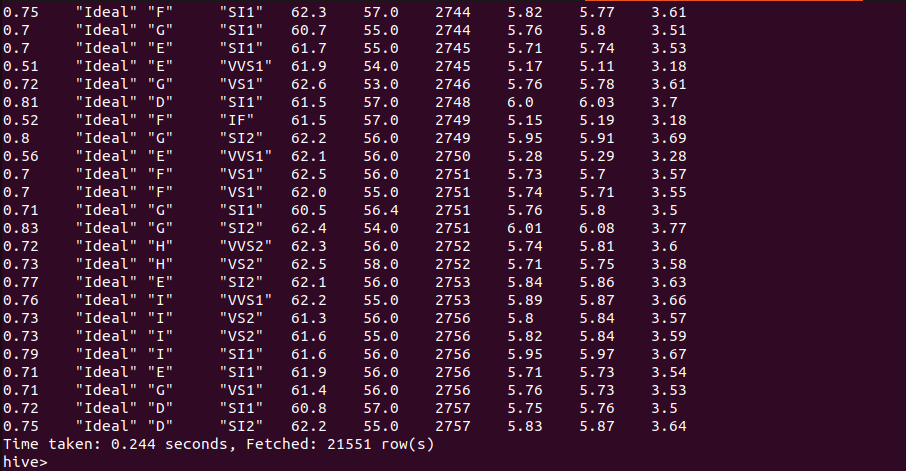


SELECT \* FROM diamonds\_table WHERE carat > 2.0 ORDER BY price DESC LIMIT 10;



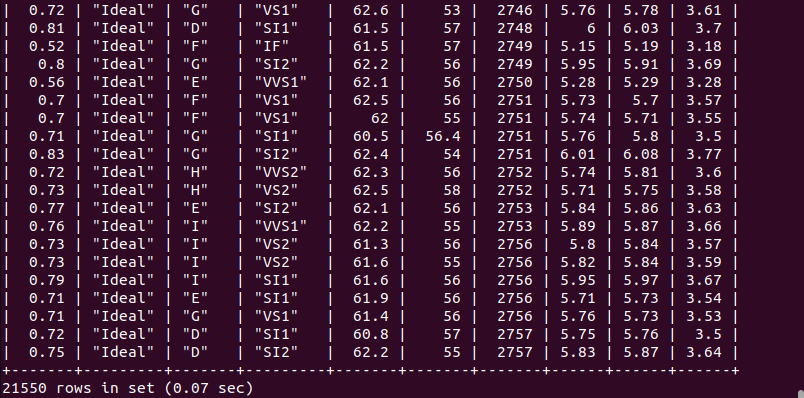
SELECT \* FROM diamondsTable WHERE cut = '"Ideal"';





SELECT \* FROM diamonds\_table WHERE cut = '”Ideal”';



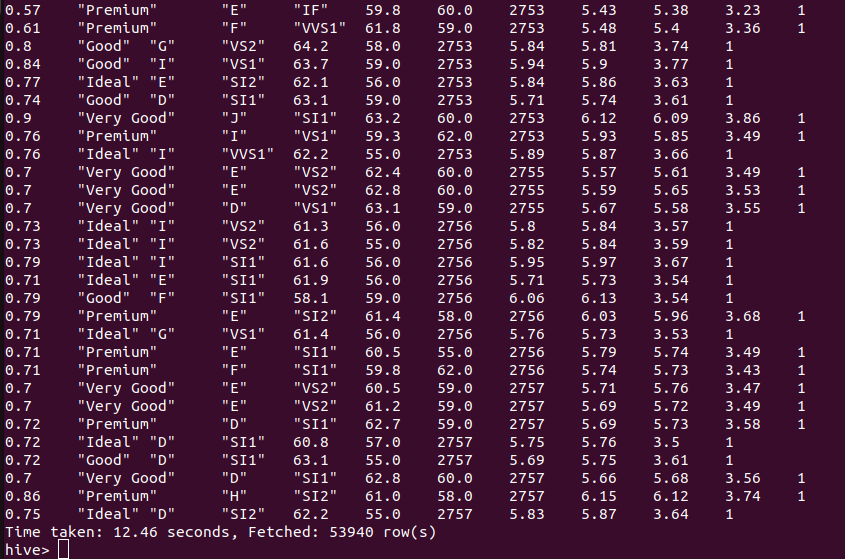


##### Joining Comparison

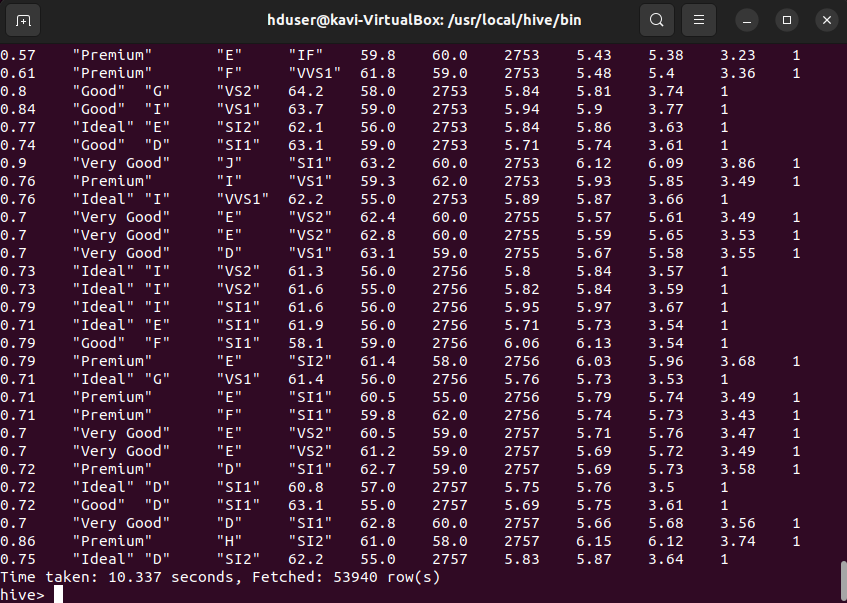
SELECT d.\*, dummy.\* FROM diamondsTable d JOIN (SELECT 1 AS dummy\_col) dummy ON 1=1;



First execution

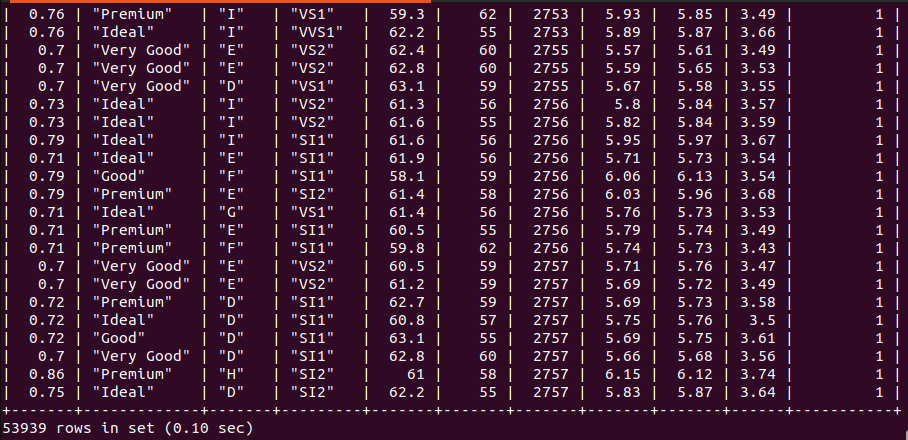


Second execution



SELECT d.\*, dummy.\* FROM diamonds\_table d JOIN (SELECT 1 AS dummy\_col) dummy ON 1=1;





Across all queries, whether for aggregation, filtering, or for joins, MySQL consistently outperforms Apache Hive. The overall results are as expected, with MySQL proving more efficient for querying a relatively small dataset of 53939 rows and 10 columns. Both tools have similar querying structure as we discussed earlier, with Hive providing SQL like syntax for querying HDFS. With this, Hive's architecture introduces additional overheads due to the requirement of translating SQL like queries into MapReduce Jobs. MySQL on the other hand, benefits from its query optimisation, execution engine and relational database design. We find Hive's architecture and MapReduce execution to be considerably slower for join queries as tested with a dummy table above. Apache Hive is however designed for large scale data processing and analysis, executable on Hadoop clusters, for which it should outperform MySQL.

### Question 4:

#### Using Apache Pig and the provided dataset with columns Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country, write a Pig script to calculate the total sales generated by customers in each country, as well as identify the top 10 products sold overall. Your script should compute the total sales for each country and then list the top 10 products based on the total sales amount. Ensure to handle any null or invalid values appropriately in the dataset. Provide the Screenshots for the execution of Pig Script on VM and explain the purpose of all these steps.

### Question 5:

#### 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.

# References

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**