A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Derek O’Hara |
| *Student Number* | sbs24018 |
| *Module Title* | Storage Solutions for Big Data |
| *Assessment Title* | CA2 |
| *Assessment Due Date* | 24th November |
| *Date of Submission* |  |

**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.

# Contents

[Contents ii](#_Toc183211678)

[Assessment Task 1](#_Toc183211679)

[Question 1 1](#_Toc183211680)

[Answer 1](#_Toc183211681)

[Question 2 3](#_Toc183211682)

[Answer 3](#_Toc183211683)

[Question 3 7](#_Toc183211684)

[Answer 7](#_Toc183211685)

[Question 4 12](#_Toc183211686)

[Answer 12](#_Toc183211687)

[Question 5 16](#_Toc183211688)

[Answer 16](#_Toc183211689)

[References 20](#_Toc183211690)

[Github 20](#_Toc183211691)

# Assessment Task

## Question 1

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.

(10 Marks)

## Answer

Big data refers to extremely large and complex data sets that cannot be easily managed or analysed with traditional data processing tools, particularly spreadsheets (Oracle, 2024). Big Data includes diverse formats, such as structured, semi-structured, and unstructured data from a wide range of sources like social media, sensors, transactional activities, and more.

**Main Characteristics of Big Data:**

Traditionally, big data was recognised by three characteristics known as the “three Vs”: **variety**, **volume**, and **velocity**. Over the past few years, two additional Vs have emerged: **value** and **veracity**.

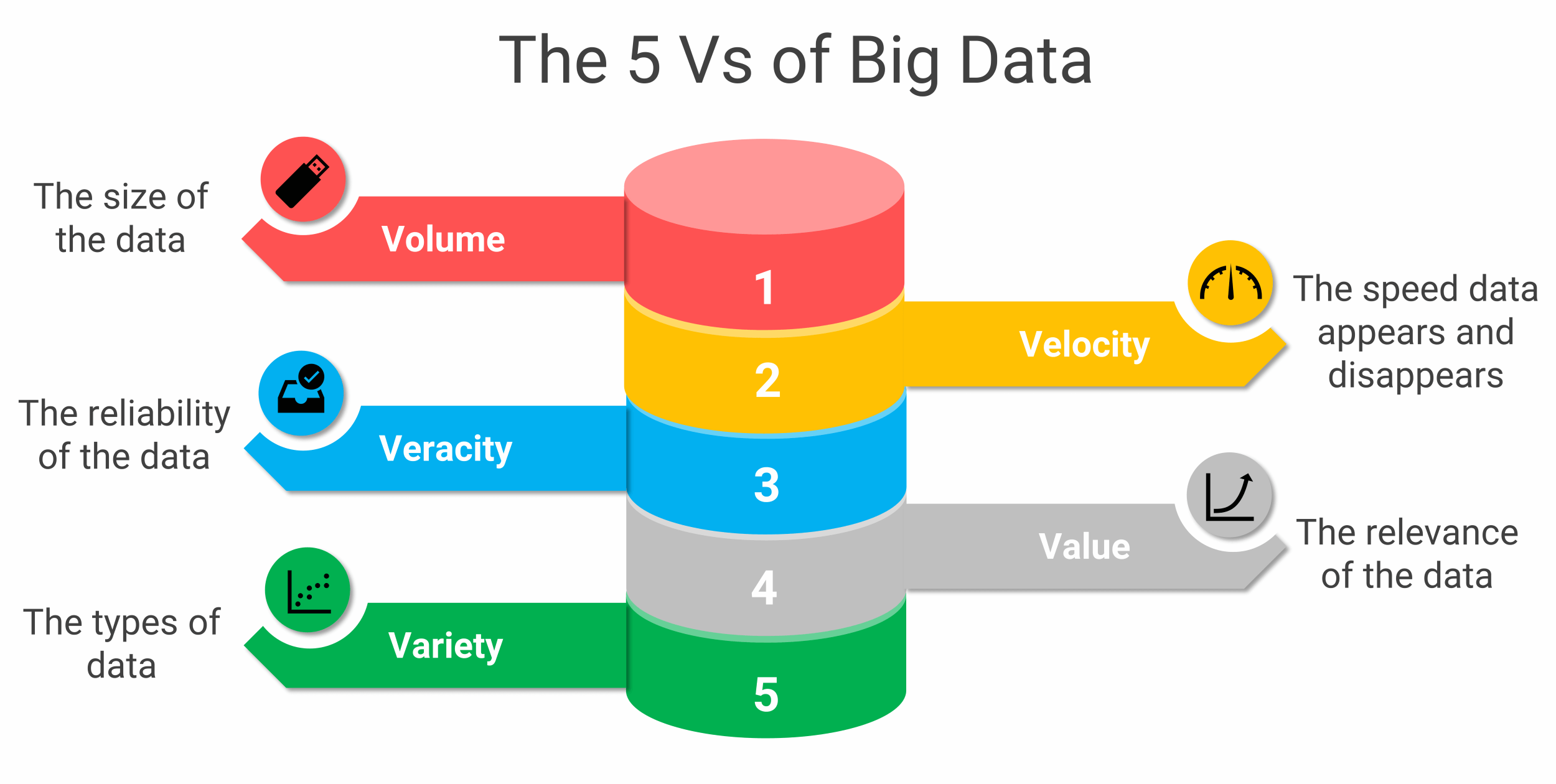


Fig 1. 5Vs of Big Data (Passionned Group. 2024).

**Potential for banks to increase profits through Big Data:**

Banks can increase their profits significantly by ramping up their Big Data capabilities in the following ways:

* **Customer profiling:** By analysing transactional and behavioural data, banks can offer personalised financial products and services, boosting customer satisfaction and loyalty.
* **Fraud detection:** Real-time Big Data processing enables banks to identify potential fraudsters swiftly, minimising losses.
* **Decisions on lending:** Big Data can streamline loan approvals and decision-making processes, which can be optimised or even automated.
* **Compliance with regulations:** Big data analytics and BI tools go a long way in making the task of compliance with regulations easier.
* **Cybersecurity:** While cyber-attacks have been growing, banks are looking towards big data analytics and AI tools to improve their cybersecurity because threats from inside the organisations have also started emerging. (V. Zubenko, 2023)

**Three businesses that have successfully leveraged big data storage solutions:**

* **Uber:** Big data in route optimisation, demand estimation, and dynamic pricing are applied. Gathering enormous pieces of information on GPS, traffic patterns, and user behaviour, Uber analyses the data in real time to fit ride-matching. It actively predicts rider demand to study scheduling. This helps in improving the experience of users and reduces costs. (ProjectPro, 2024)
* **Spotify:** Spotify uses big data in recommending music and curated playlists to users, each according to taste. Using machine learning by analysing user listening habits, playlists, and genre preferences, Spotify predicts and recommends songs that their users are likely to enjoy. This personalisation keeps them on the platform for longer periods of time and raises subscription revenue. (Y. Müstecaplıoğlu, 2024)
* **John Deere:** Agricultural machinery company John Deere installs big data and IoT devices in its agricultural machinery. These devices will give actionable data with respect to crop health, soil quality, and weather conditions. This sets of information, therefore, aids farmers right from planting to harvest with optimised yield and minimal wastage. They help farmers speed up profitability and efficiency by providing actionable insights into data and, in the process, enhance their relationship with customers and their service offerings. (John Deere, 2024)

## Question 2

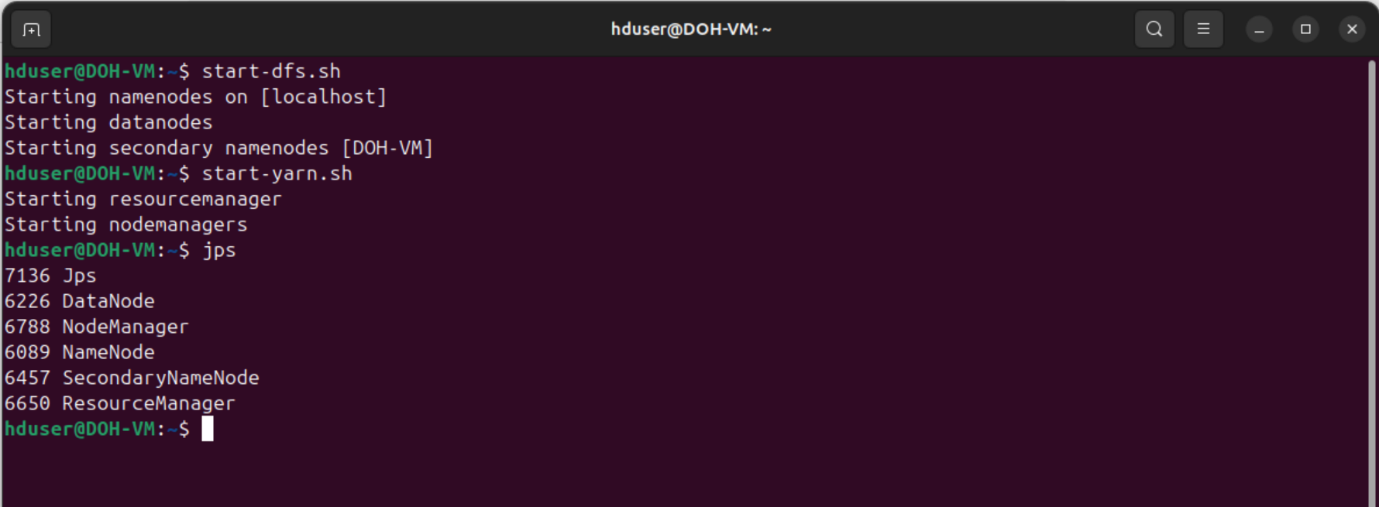
Provide the screenshots based on your VM (hduser@studentusername:~$) or Google cloud platform (we check the Google cloud has been used or not, if not used, zero marks) for the following processes

1. Start all five processes of hadoop distributed file system
2. Find the text dataset of size (600 MB at least) and create a folder on hadoop named CA2. Copy the dataset from VM to the hdfs directory (CA2) the data.
3. Execute Mapper and Reducer programmes by using loaded input file.
4. Display the frequency of each word obtained from the dataset
5. 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.

(30 Marks)

## Answer

1. Start all five processes of Hadoop distributed file system.

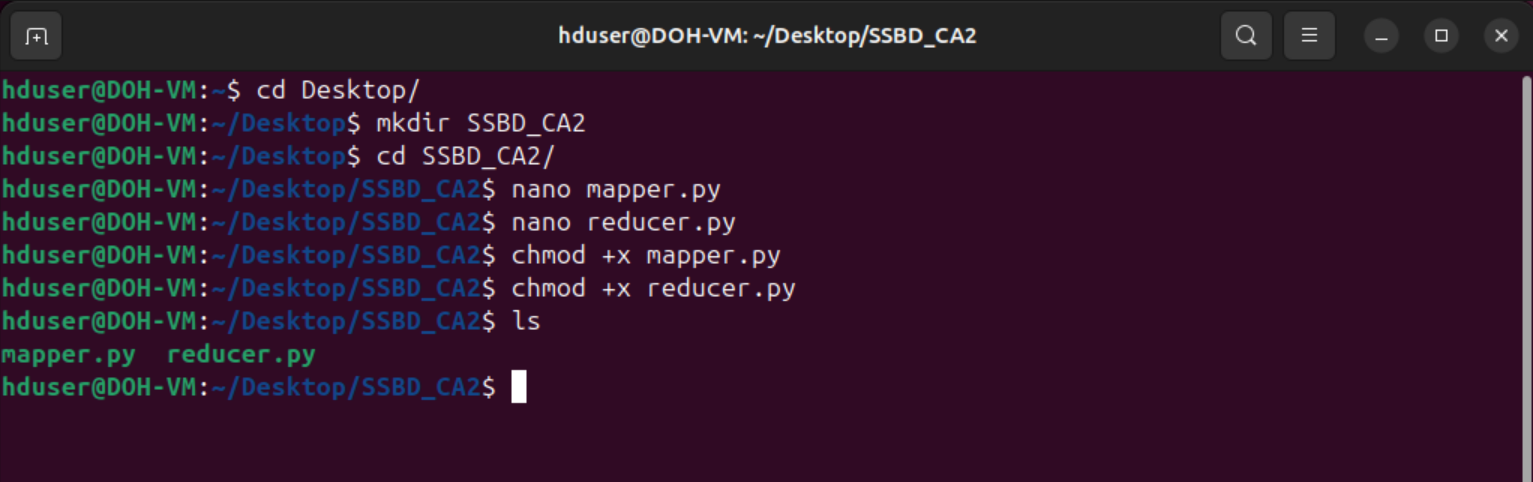


1. I found “[The Enron email dataset](https://www.kaggle.com/datasets/wcukierski/enron-email-dataset)” which consists of around 500,000 emails generated by employees of the Enron Corporation. The dataset comprises of two columns “file” & “message” and is approximately 1.39GB in size.

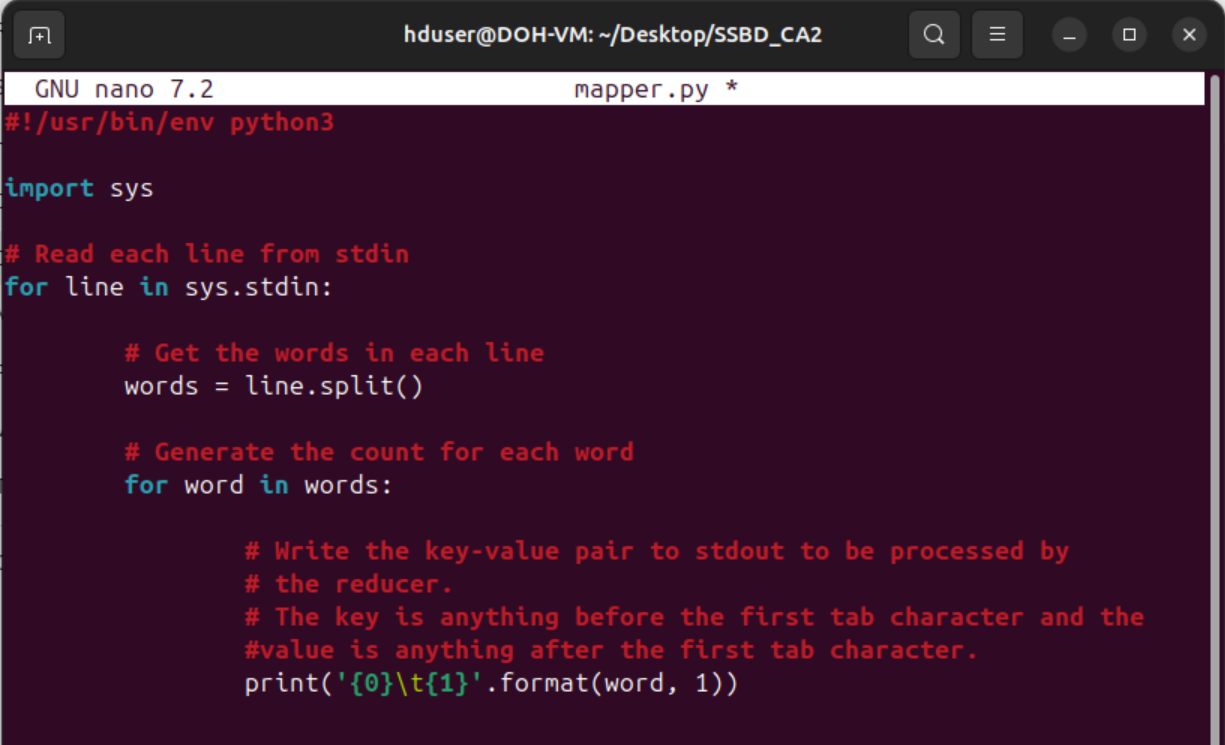
A screen shot of a computer

Description automatically generated

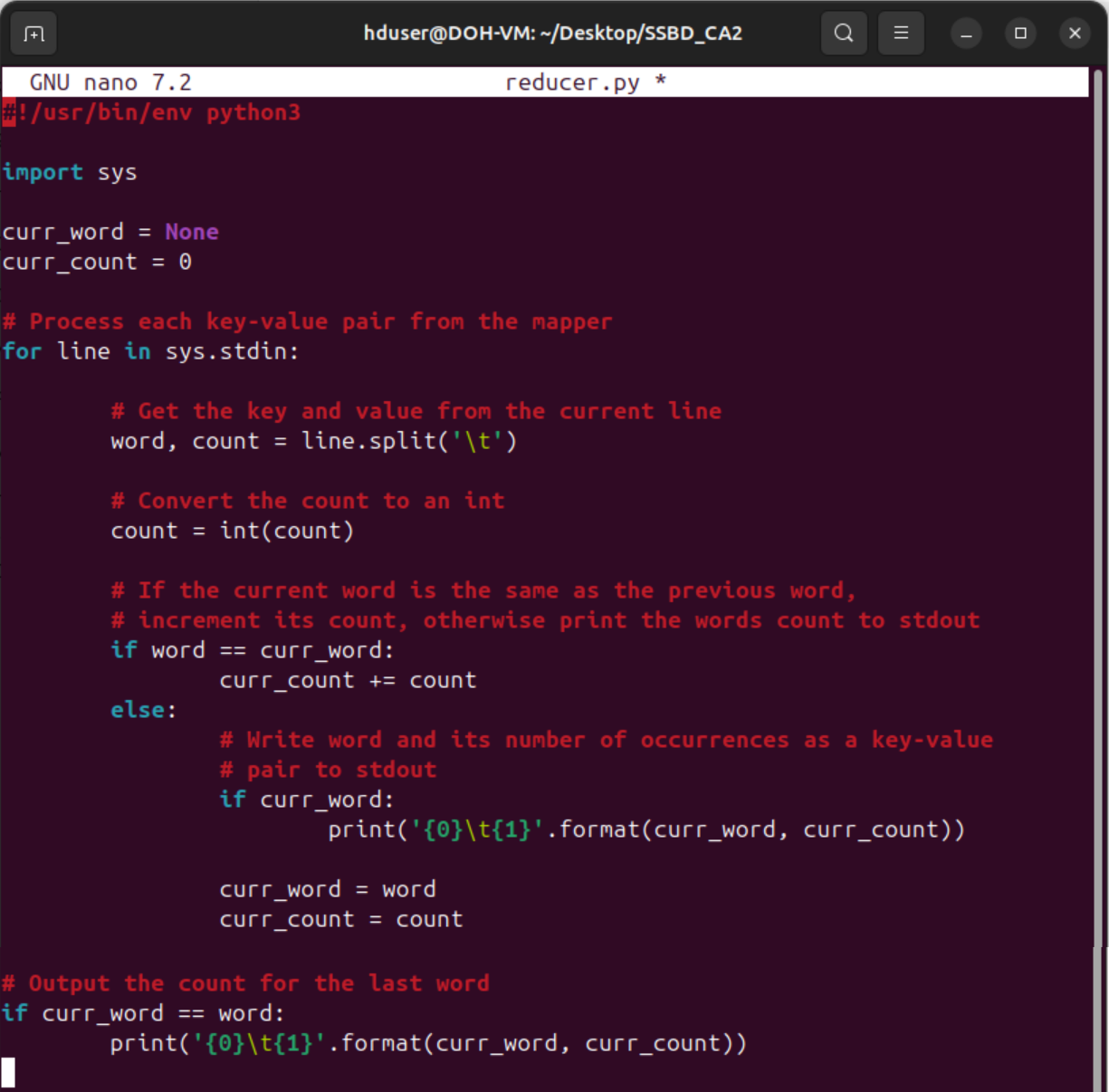
1. Execute Mapper and Reducer programmes by using loaded input file.



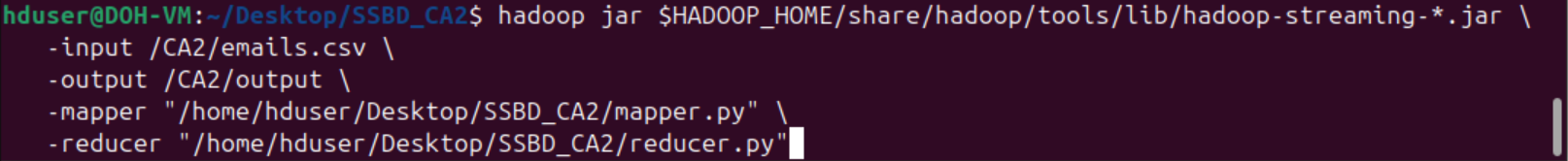
Creating the mapper.py file.



Creating the reducer.py file.

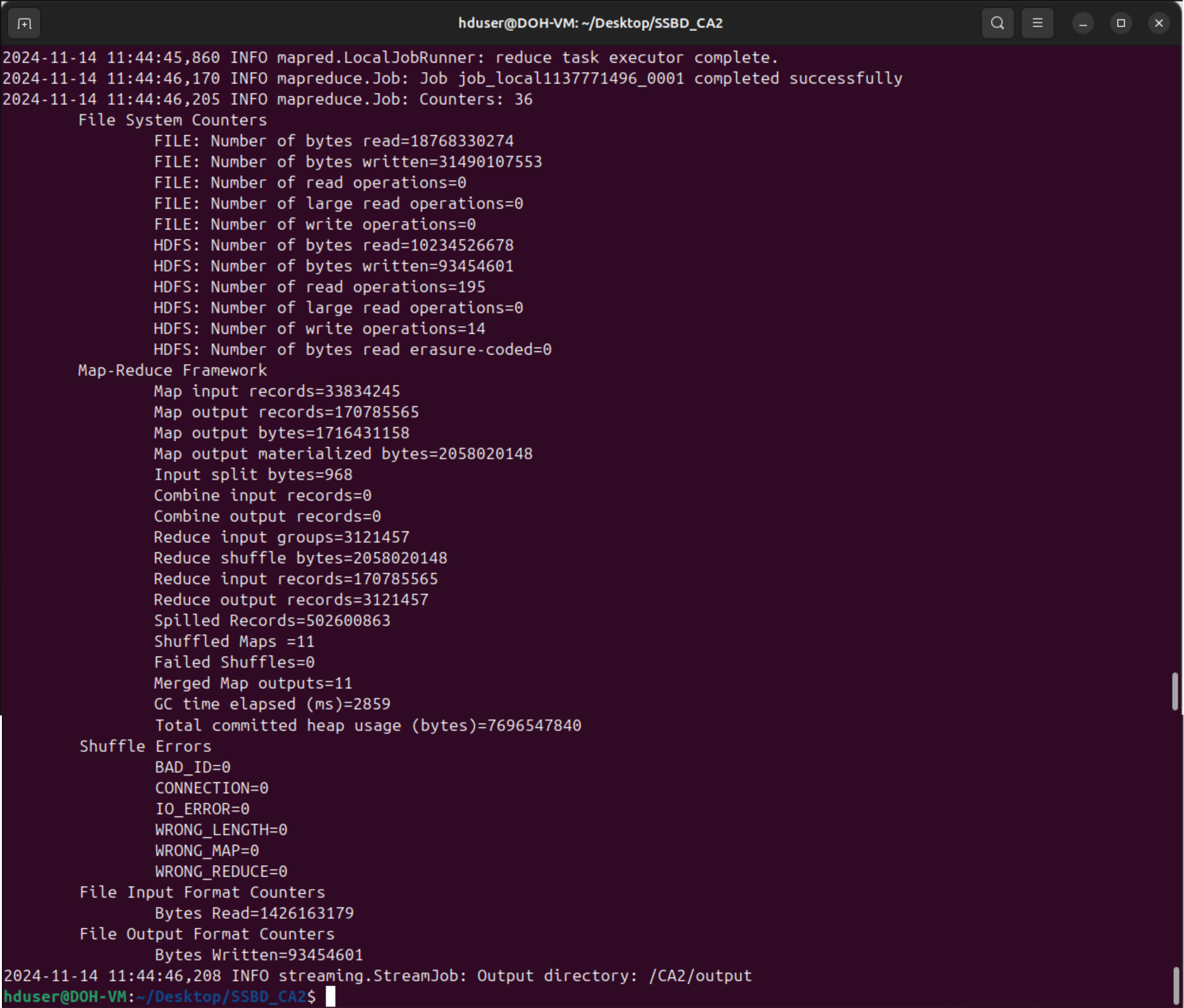


Execute the command to run the Hadoop Streaming job:



The job:

* Reads the input data from emails.csv in the /CA2/ directory in HDFS.
* The mapper.py script breaks down the data into key-value pairs.
* The key-value pairs are then grouped and processed by the reducer.py script to produce the final output.
* Results are saved in the /CA2/output directory in HDFS.



1. Display the frequency of each word obtained from the dataset.

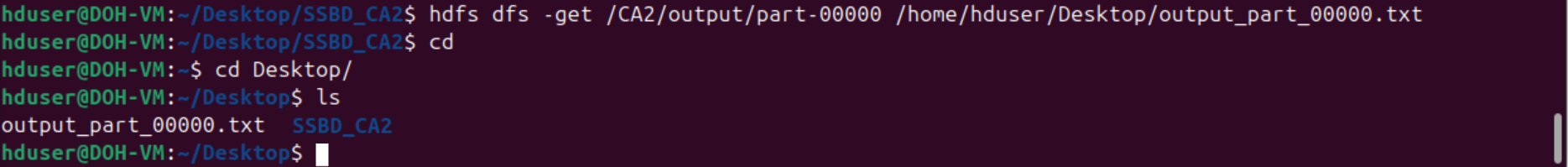
Due to a terminal buffer limit constraint, I couldn’t display the full contents of the output. A more expanded view of the output is below from the output text file.

  
A screen shot of a computer

Description automatically generated

1. Download the output file from hadoop and upload it on Moodle.

I downloaded the part-00000 file from HDFS to my VM desktop as output\_part\_00000.txt and will upload to Moodle.



## Question 3

Discuss and 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)

## Answer

The comparison of the architecture and performance of MySQL and Apache Hive can be broken down into several aspects:

|  |  |  |
| --- | --- | --- |
| **Aspect** | **MySQL** | **Hive** |
| **Architecture** | Relational database (RDBMS) | Data warehouse system built on top of Hadoop. |
| **Processing Model** | Row-based processing. | Batch processing via MapReduce or Tez (columnar storage). |
| **Execution Engine** | Single-node processing for queries. | Distributed processing using Hadoop. |
| **Storage** | Stores data directly in database files. | Reads data from HDFS or similar distributed storage. |
| **Schema** | Strict schema enforcement. | Schema-on-read (data can be queried without enforcement). |
| **Performance** | Low-latency queries, optimised for OLTP. | High-latency but scales well for OLAP and large datasets. |
| **Data Size** | Best suited for small to medium datasets. | Handles petabyte-scale datasets. |
| **Use Cases** | Transactional systems, small-medium datasets. | Data warehousing, analytics on large datasets. |

(M Rathbone, 2015)

To perform a query comparison between MySQL and Apache Hive, I picked the “[Steam Store Games](https://www.kaggle.com/datasets/nikdavis/steam-store-games?resource=download)” dataset (Kaggle, 2019). The dataset is 5.82 MB, consists of 18 columns and is the combined data of 27,000 games scraped from Steam and SteamSpy APIs.

For the comparison exercise, I will:

* Create separate Hive and MYSQL Databases.
* Create the same table on each Database.
* Load the dataset into each Database/table.
* Execute and compare the same 3 queries on each Database:

|  |  |  |
| --- | --- | --- |
| **Query Number** | **Purpose** | **Query** |
| 1 | Fetch the top 10 rows of the table. | SELECT \* FROM TABLE LIMIT 10 |
| 2 | Fetch the number of entries in the table. | SELECT COUNT (\*) FROM TABLE |
| 3 | Fetch the top 10 games with the highest number of positive ratings. | SELECT NAME, POSITIVE\_RATINGS FROM STEAM\_GAMES ORDER BY POSITIVE\_RATINGS DESC LIMIT 10; |

**Hive example**

Steps involved:

* Start the Hadoop Distributed File System (HDFS).
* Create a directory /user/hive/steam\_data in HDFS and upload the steam.csv dataset from the local filesystem into the created HDFS directory.
* Launch Hive.
* Create the Hive Database and the table.
* Load the data from the dataset into the newly created table.
* Execute the 3 test queries in the Hive Database.

A screenshot of a computer

Description automatically generated

A black and purple background

Description automatically generated with medium confidence

A computer screen with text

Description automatically generated with medium confidence

A black background with a purple backdrop

Description automatically generated with medium confidence

A computer screen with white text

Description automatically generated

**MYSQL example**

Steps involved:

* I move the dataset steam.csv from my Downloads folder to /var/lib/mysql/ and then to /tmp/ and update the file permissions.
* Start MYSQL and log in as root user.
* Create the MYSQL Database and the table.
* Load the data from the dataset into the newly created table.
* Execute the 3 test queries in the MYSQL Database.

A black background with a purple background

Description automatically generated with medium confidence

A black and purple background

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated

A black and purple background

Description automatically generated with medium confidence

**Results:**

|  |  |  |
| --- | --- | --- |
| **Query Number** | **Hive performance** | **MYSQL performance** |
| 1 | 2.209 sec | 0.00 sec |
| 2 | 3.980 sec | 0.02 sec |
| 3 | 1.722 sec | 0.02 sec |

**Conclusion:**

The above results demonstrate MySQL significantly outperforms Apache Hive in terms of query execution time, making it better suited for small to medium-sized datasets requiring low-latency operations.

## Question 4

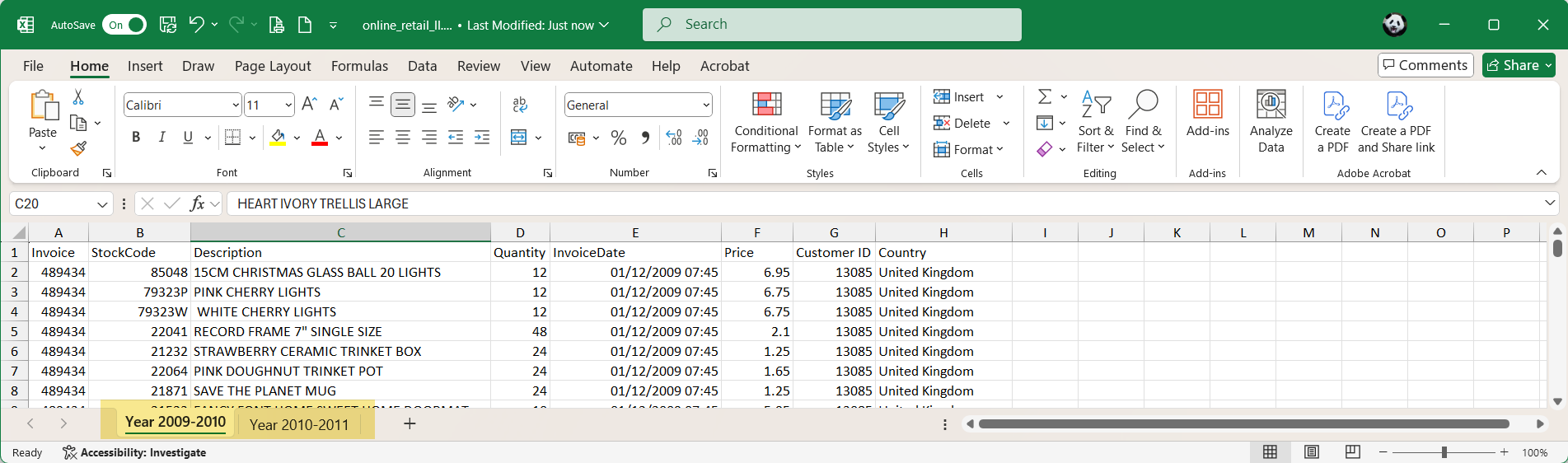
Using Apache Pig and the provided dataset with columns Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country, write a Pig script to calculate the total sales generated by customers in each country, as well as identify the top 10 products sold overall. Your script should 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.

(20 Marks)

## Answer

**Step 1:**

The dataset comes in the form of an excel file with two sheets covering the years 2009 to 2011. To handle the dataset in Apache Pig, I combined the two sheets into one .csv file in Jupyter Notebook.



A screenshot of a computer program

Description automatically generated

**Step 2:**

I initialised the HDFS and YARN, uploaded the dataset to HDFS, and set up a folder CA2\_Q4 on the vm local desktop. I then created the script sales\_analysis.pig in CA2\_Q4.

A screenshot of a computer

Description automatically generated

**Step 3:**

The sales\_analysis.pig script processes the dataset by loading it from HDFS, filtering out invalid entries, calculating total sales per row, aggregating sales by country and product, identifying the top 10 products by sales, and storing the results back into HDFS.

A screenshot of a computer screen

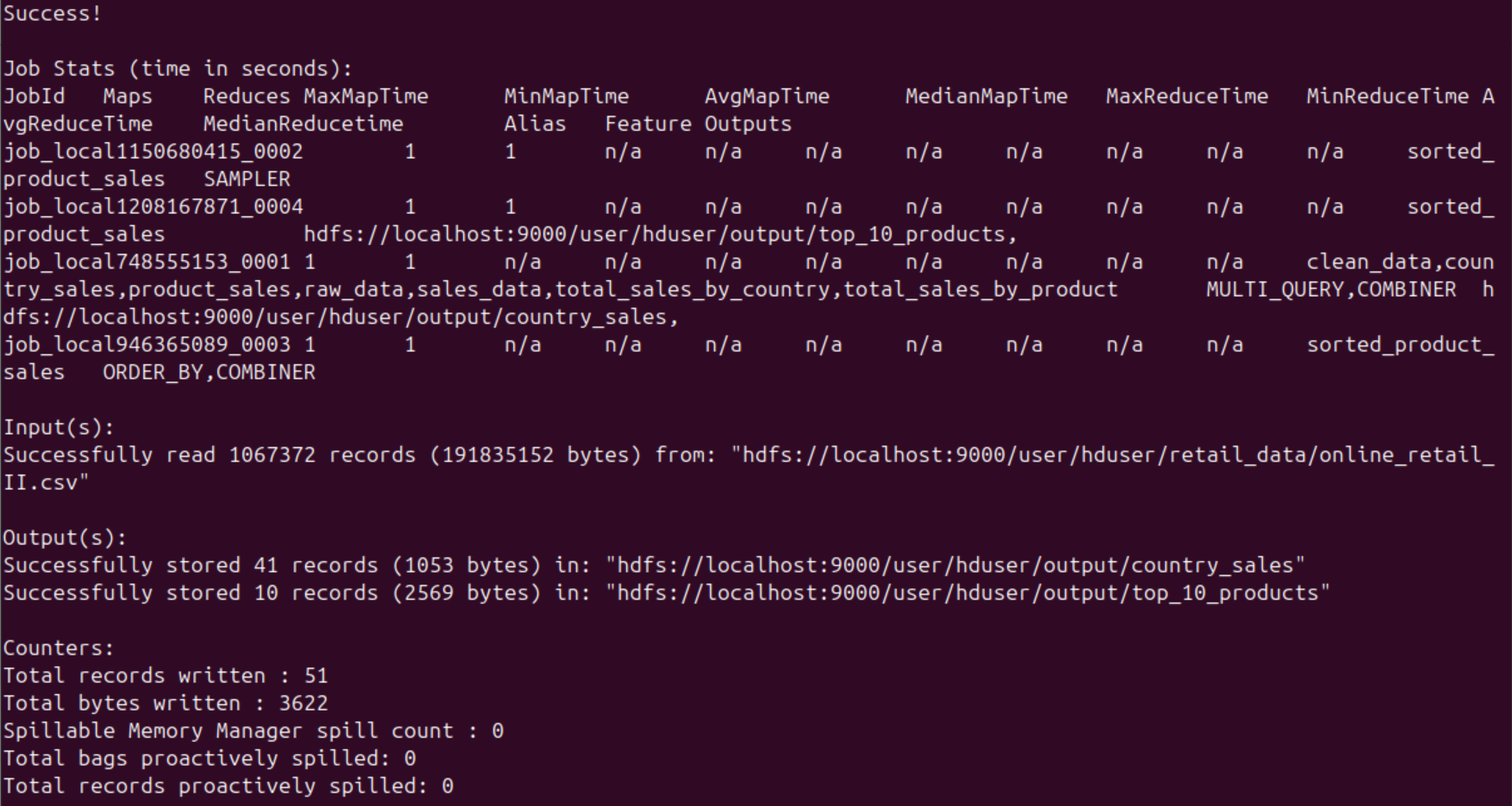
Description automatically generated

**Step 4:**

I ran the pig script in local mode, which completed successfully, processing data and storing the calculated sales by country and the top 10 products in the specified HDFS directories.

A screen shot of a computer

Description automatically generated



**Step 5:**

The outputs confirm the successful execution of the Pig script, generating and storing the top 10 products by sales and total sales by country in their respective HDFS directories.

A computer screen shot of a program

Description automatically generated

A screenshot of a computer program

Description automatically generated

## 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 25,000 words and write a wordcount program (Java/ Python) to count the frequency of words and related aggregation functions.

(20 Marks)

## Answer

**Apache Flink** operates in a Master-Slave architecture:

* **Master Node:** Central manager of the cluster, known as the Job Manager.
* **Worker Nodes:** Worker nodes called Node Managers.

**Workflow:**

* The client submits jobs or applications to the Master Node.
* The Master Node divides the job into tasks and distributes them to the Slave Nodes.
* Tasks are executed in parallel, enabling distributed computing.

**Key Features:**

* This architecture allows Flink to process data at extremely high speeds.
* The Job Manager coordinates job execution across the cluster. (Apache Flink, n.d.)

A screenshot of a computer screen

Description automatically generated

**Apache Storm** is an open-source, real-time distributed computation system designed for processing unbounded streams of data with low latency, offering scalability, fault tolerance, and support for any programming language. Unlike batch-oriented systems like Hadoop, Storm focuses on continuous stream processing, making it ideal for real-time analytics and event-driven applications (Apache Storm, n.d.).

For the **wordcount** taskI downloaded the text of "[War and Peace](https://www.gutenberg.org/ebooks/2600)" by Leo Tolstoy, available at [Project Gutenberg](https://www.gutenberg.org/). I created the SSBDCA2 java project and the WordCount.java class.

A screenshot of a computer

Description automatically generated

The **WordCount.java** class:

* Reads **warandpeace.txt** from the explicit path to the **CA2\_Q5** folder created on the Desktop.
* Splits, normalizes and removes non-word characters using a custom tokenizer function.
* Aggregates the total number of words in the book and prints the result to the console.
* Groups words, aggregates their occurrences, sorts them in descending order and writes the result to **output.csv** in the **CA2\_Q5** folder.
* Extracts the top 10 most frequent words and writes them to another CSV file **top\_10\_words.csv** in the **CA2\_Q5** folder.

I then exported the **wordcount.java** jar file to the **CA2\_Q5** folder

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated A screenshot of a phone

Description automatically generated

I then started the Flink cluster and ran the wordcount.java jar file.

* Both steps of the job completed.
* The total word count of 586,871 was printed to the console.
* The full word count aggregates were written to **output.csv**.
* The top ten words were written to **top\_10\_words.csv**.

A screenshot of a computer

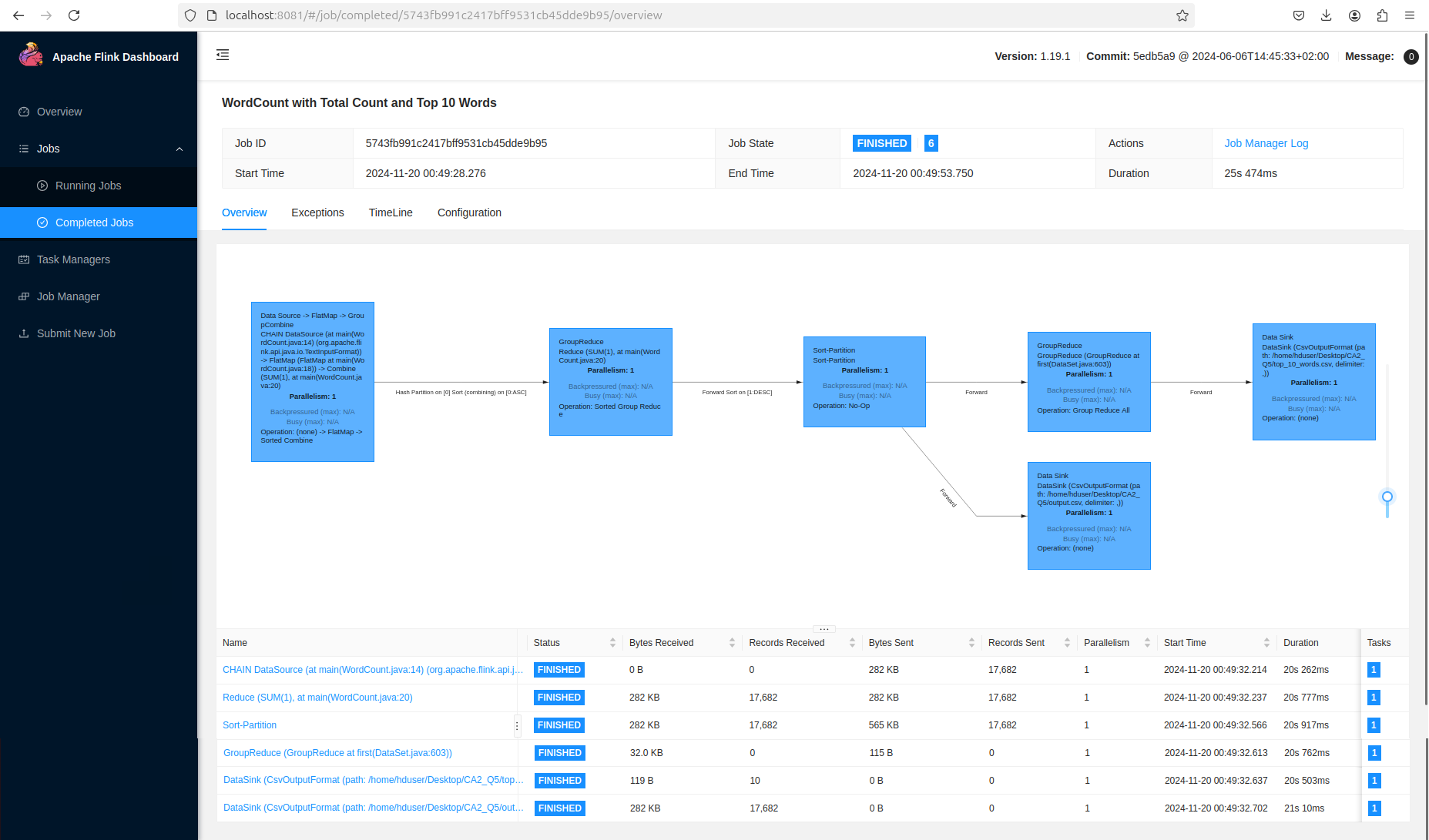
Description automatically generated

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

We can also see the job statistics in the Apache Flink Dashboard.



# References

Oracle. (2024). What Is Big Data? Available at: <https://www.oracle.com/ie/big-data/what-is-big-data/>

Fig 1. 5Vs of Big Data: Passionned Group. (2024). Big Data Analytics: Definition, Pros and Cons, Applications, Opportunities, and Risks. Available at: <https://www.passionned.com/big-data/>

Valentyn. Zubenko (2023) How big data changes the scope of modern banking. Available at: <https://www.avenga.com/magazine/how-big-data-changes-banking/>

Projector. (2024). How Uber Uses Data Science to Reinvent Transportation. Available at: <https://www.projectpro.io/article/how-uber-uses-data-science-to-reinvent-transportation/290>

Yaprak Müstecaplıoğlu (2024) How Spotify Uses Data to Enhance Your Music Experience. Available at: <https://medium.com/@dss_99180/how-spotify-uses-data-to-enhance-your-music-experience-2f26a7437aa8>

John Deere. (2024). Data Management in Precision Ag Technology. Available at: <https://www.deere.com/en/technology-products/precision-ag-technology/data-management/>

Matthew Rathbone (2015) Apache Hive vs MySQL - What are the key differences? Available at: <https://blog.matthewrathbone.com/2015/12/08/hive-vs-mysql.html#:~:text=Hive%20has%20serialization%20and%20deserialization,add%20data%20to%20the%20store>.

Kaggle (2019) Steam Store Games (Clean dataset). Available at: <https://www.kaggle.com/datasets/nikdavis/steam-store-games?resource=download>

Apache Flink (n.d.) Flink architecture. Available at: <https://nightlies.apache.org/flink/flink-docs-master/docs/concepts/flink-architecture/>.

(Leo Tolstoy, 2022) War and Peace by graf Leo Tolstoy. Project Gutenberg. Available at: <https://www.gutenberg.org/ebooks/2600>

# Github

<https://github.com/derekoharacct/Storage-Solutions-Big-Data.git>