



# Table of Contents

[Table of Contents 3](#_heading=h.gjdgxs)

[1. Introduction 5](#_heading=h.30j0zll)

[2. SSBD Assessment Task 5](#_heading=h.1fob9te)

[Question 1: 5](#_heading=h.aoees4tqhce7)

[Question 2: 5](#_heading=h.aejjy3kfdtho)

[Question 3: 5](#_heading=h.9jrv21liu0os)

[Question 4: 5](#_heading=h.ypmaj1omb772)

[Question 5: 6](#_heading=h.jkzhc4wg9gob)

[7. Conclusion 6](#_heading=h.2afmg28)

[8. References 7](#_heading=h.pkwqa1)

[Appendix 1 7](#_heading=h.ru1zxpow32dx)

[Filenames and directories 7](#_heading=h.swzu1ky1ay4k)

# Introduction

This is the integrated assessment for Muhammad and James.

* See the Appendix for list of files
* See Git @ <https://github.com/ciaranquinlan/ssbd-ca2.git>, you are both invited to this repo, everything i did is in there, images, code, markdown.
* All my assessment was run in a PC with the following ubuntu version

| hduser@ciaran-ubun:~/ssbdca/ssbd-ca2/hadoop$ lsb\_release -a  Distributor ID: Ubuntu  Description: Ubuntu 22.04.3 LTS  Release: 22.04  Codename: jammy |
| --- |

# SS

# SSBD Assessment Task

## Question 1: Can you define Big Data?

I would define Big Data as very large (terrabytes+) and complex data sets that cannot be handled or stored for fast retrieval and analysis by traditional data processing software. Big data is characterised in 3V’s

Volume: Massive amounts of data stored on solid state disks with petebytes of data collected from webpages, IOT devices,cameras, sensors,vehicles etc

Velocity: Real time, instantaneous and changing data.

Variety: Structured (transactions, databases,data feeds), semi structured (XML, SGML, JSON), and unstructured (text, images, voice, video) data in multiple data formats.

and some add a further 2 V’s Variability (data in numerous formats) and veracity (reliable and unreliable data).

Banks use big data to combat internal and external fraud, spot customer opportunities, better server regulators and manage risk in the money markets

I have come across these organisations who use big data:

**Dunsink** **Observatory** use big data to analyse the dark sky above Ireland to find new stars, monitor meteors and observe astrological threats.

**Tesla** use big data for analytics in their self-driving cars. This helps improve performance and safety and can identify and predict potential hazards on irish roads, such as other vehicles, pedestrians, and weather challenges.

**Irish** **Revenue** use big data to improve tax management, one example is capturing sharing economy income data from Airbnb. Also they use big data to manage collection of vat & customs fees from Amazon sales. The revenue also use big data to model taxpayer’s behavior and risk profiles.

## Question 2: 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.

Hadoop Distributed File System (HDFS) is part of Apache Hadoop. HDFS is designed for storing and processing large data sets that can be distributed across many servers.

HDFS is different because it is:

* Distributed, meaning it can use many servers in multiple locations that operate together and Scalable, meaning that you can continue to add more servers as your data grows
* Fault Tolerant because it replicates data across multiple nodes so if nodes fail it can still operate..
* Optimised for massive datasets in the terabytes or petabytes in size.
* Write-Once, Read-Many simplifying data consistency with high performance.
* Data Locality: HDFS strives to place data on the same or nearby node where

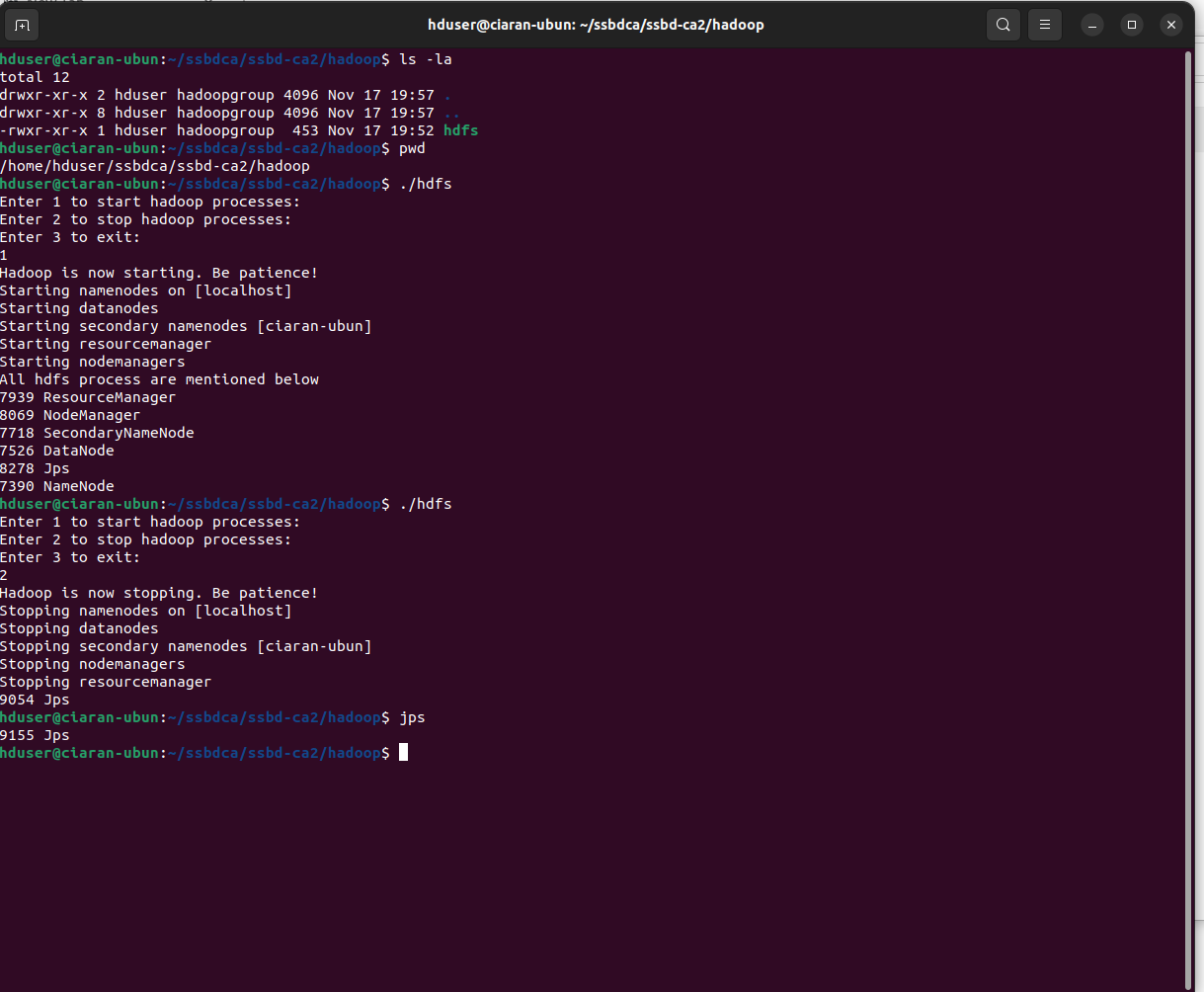
The Hadoop framework consists of several layers, each serving a specific purpose:

* **HDFS** storing and managing large data sets
* **YARN** manages and allocates resources (
* **MapReduce** does data processing tasks.
* **Hive** is for data warehousing with SQL-like query language
* **Pig** a high-level data processing language
* **Spark** not part of the Hadoop core but supports batch processing, real-time streaming, machine learning, and graph processing.

Characteristics of the Hadoop Framework:

* Distributed enabling scalability and fault tolerance.
* Fault Tolerant ensuring data availability if node fail.
* Parallel Processing by splitting data into small chunks for speed processing.
* Scalable accommodating growing data and processing needs.
* Hadoop has a rich ecosystem of tools and libraries.

My deployment of HDFS is visible @ <https://github.com/ciaranquinlan/ssbd-ca2/tree/main/hadoop>

images @ <https://github.com/ciaranquinlan/ssbd-ca2/blob/main/hadoop/Hadoop%20Screenshot%20from%202023-11-17%2020-02-10.png>. It is running on a acer pc with ubuntu 18

## 

## Question 3:

My Demonstration of MySQL and Apache Hive can be see in the folder <https://github.com/ciaranquinlan/ssbd-ca2/tree/main/mysql_hive>

I used a Dataset : Irish names from 2015-2021, 35886 rows x 5 columns

Query: calculate total\_rows, min\_year & max\_year, number of unique\_names

#### Mysql: I created a database and imported the dataset of 35886 rows. I ran a query and the output in 0.9 sec was , see screenshot below:

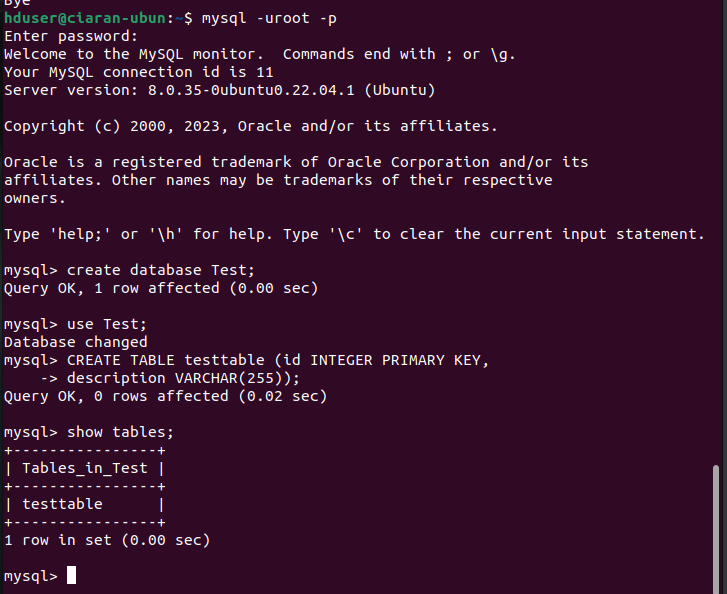
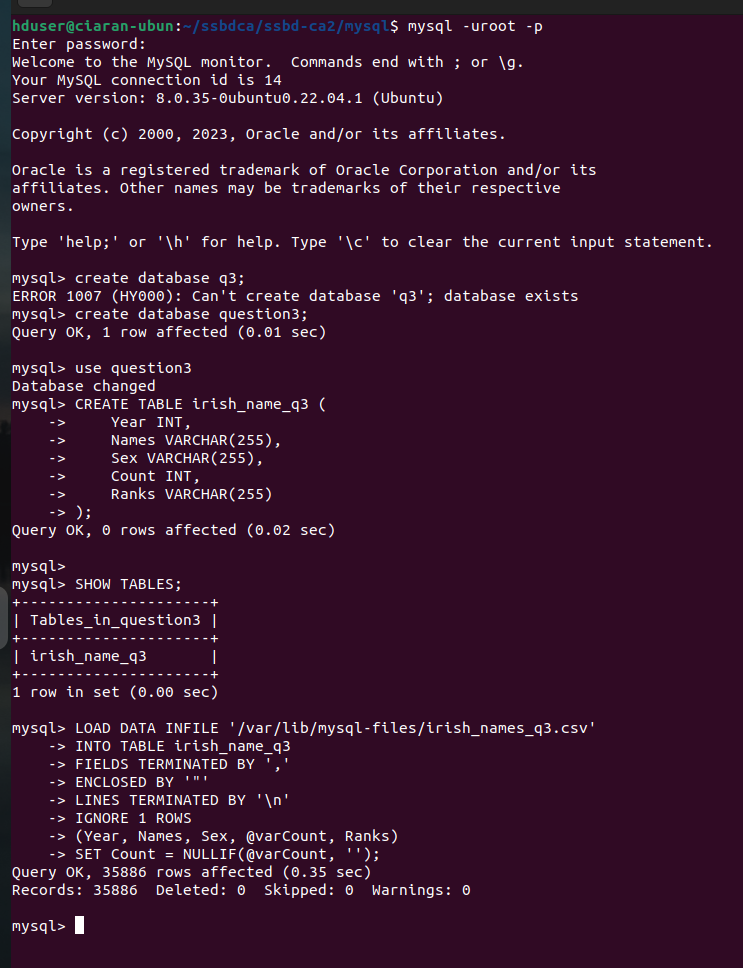
#### +------------+----------+----------+--------------+

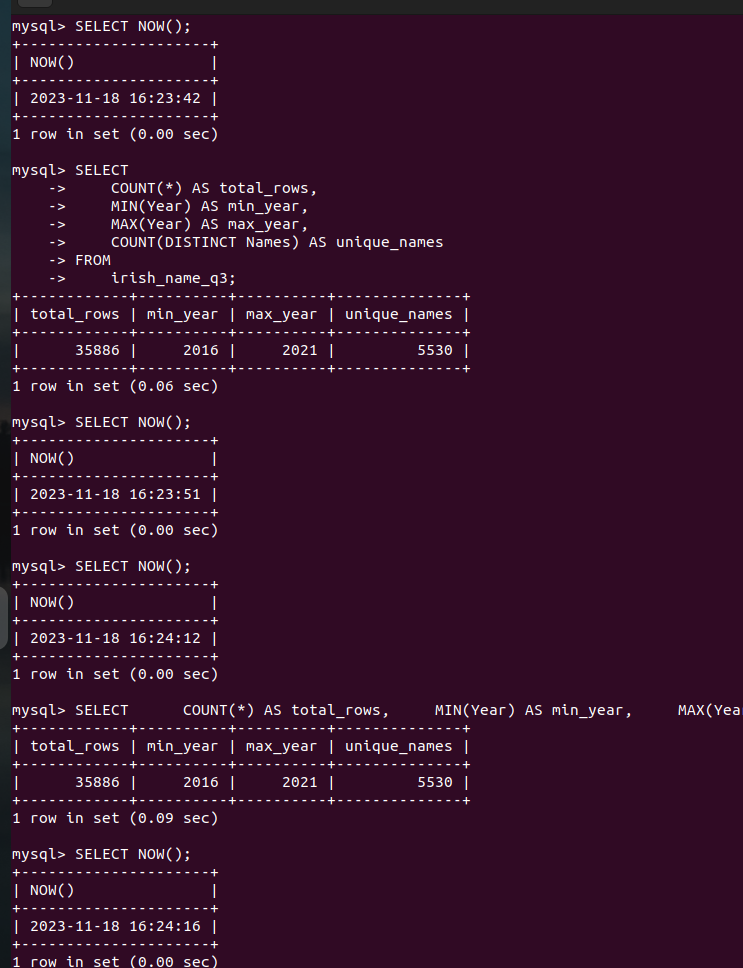
| total\_rows | min\_year | max\_year | unique\_names |

+------------+----------+----------+--------------+

| 35886 | 2016 | 2021 | 5530 |

+------------+----------+----------+--------------+





Hive: <https://github.com/ciaranquinlan/ssbd-ca2/tree/main/mysql_hive>

I started hadoop, ran the hive and imported the dataset, i ran the same query in the hive and it took longer to process than the mysql query. here is the hive screenshots, also in the mysql-hive directory.

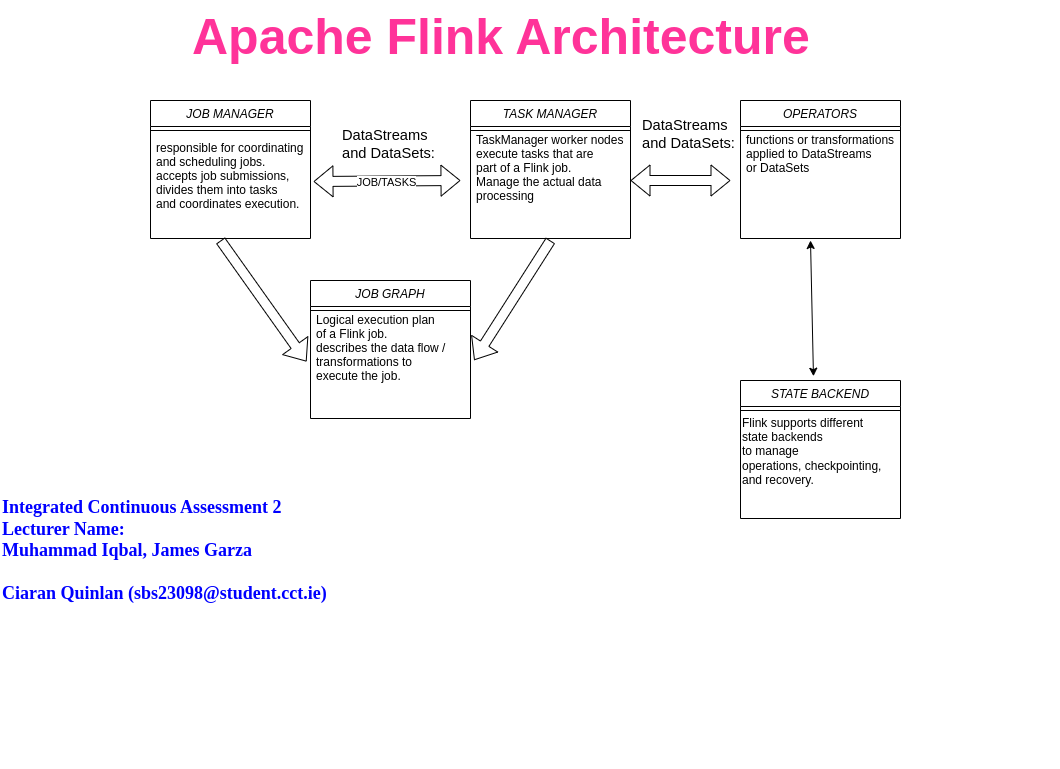
## 

## 

## Question 4:

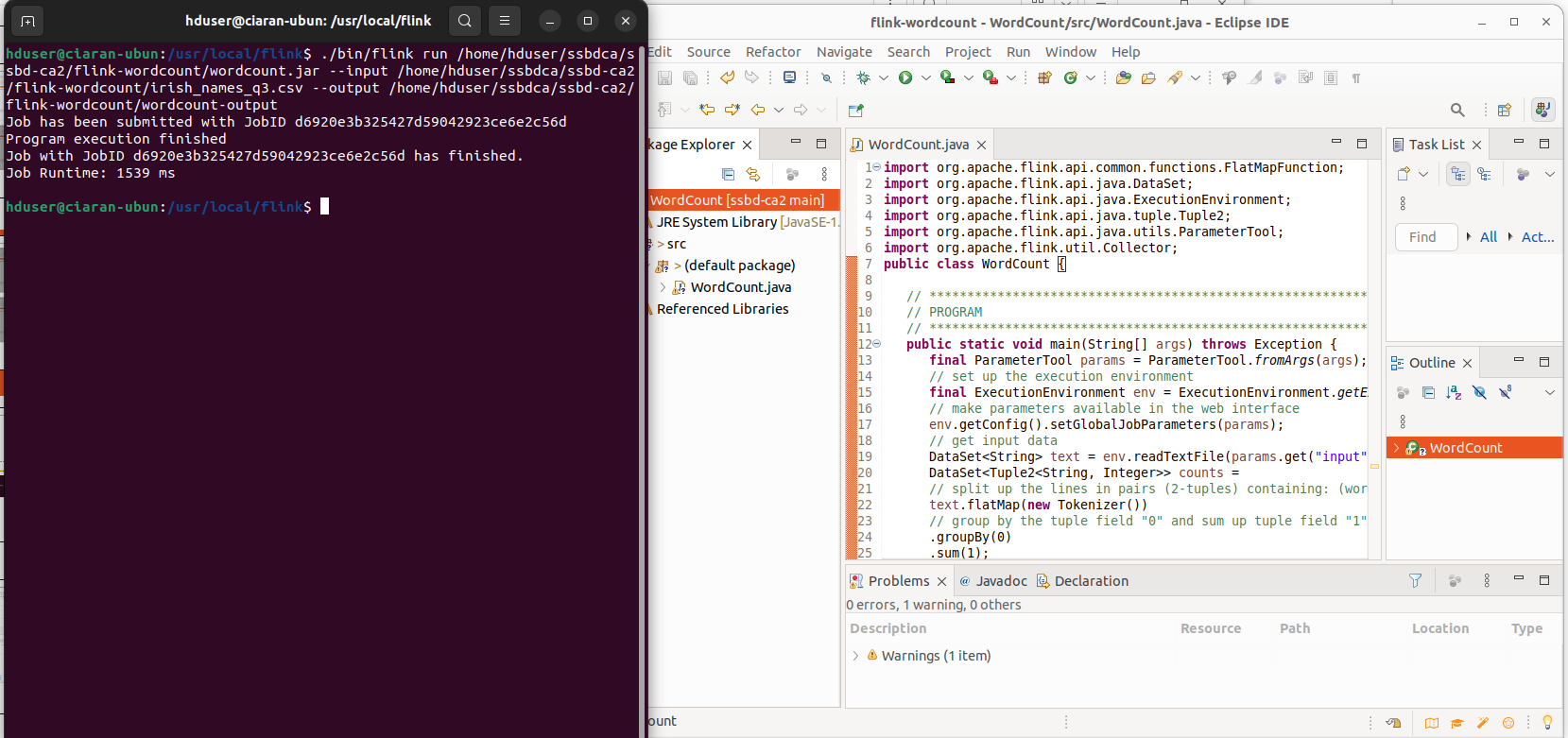
Apache Flink is a distributed stream processing framework and consists of the following components:

JobManager, TaskManagers, JobGraph, Operators, State Backend.   
This is my Apache Flink Architecture:



My flink was installed and up and running see <https://github.com/ciaranquinlan/ssbd-ca2/tree/main/flink-wordcount>

the file are in this directory and I successfully ran the wordcount java and got the output file created.



## Question 5:

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.

Why Apache Storm for Stream Processing:

Apache Storm is for real-time stream processing and it can process data streams that require low-latency and continuous processing.

Apache Storm is useful for stream processing because it has:

* Low Latency, that is Storm can run quickly when data processing and deliver real-time responsesto applications such as fraud detection, live analytics and location recommendations.
* Event Processing: Storm handles events and messages in real time.
* Scalability: being part of the Apache hadoop ecosystem, storm can scale quickly.
* Fault Tolerant, Storm ensures data processing continues even in the event of node failures.
* Ease of Use, Storm makes it easier for developers to build real-time applications without dealing with event and error handling.

Here are some key distinctions between Apache Storm and Hadoop:

Storm is designed for real-time stream processing, Hadoop is primarily designed for batch processing. Storm uses a micro-batch processing model as events or messages arrive.

Hadoop uses a batch processing model

Storm is best suited for real-time analytics, event-driven processing, and immediate responses to data streams.

Hadoop is ideal for batch-oriented tasks, offline data analysis, and processing of historical data.

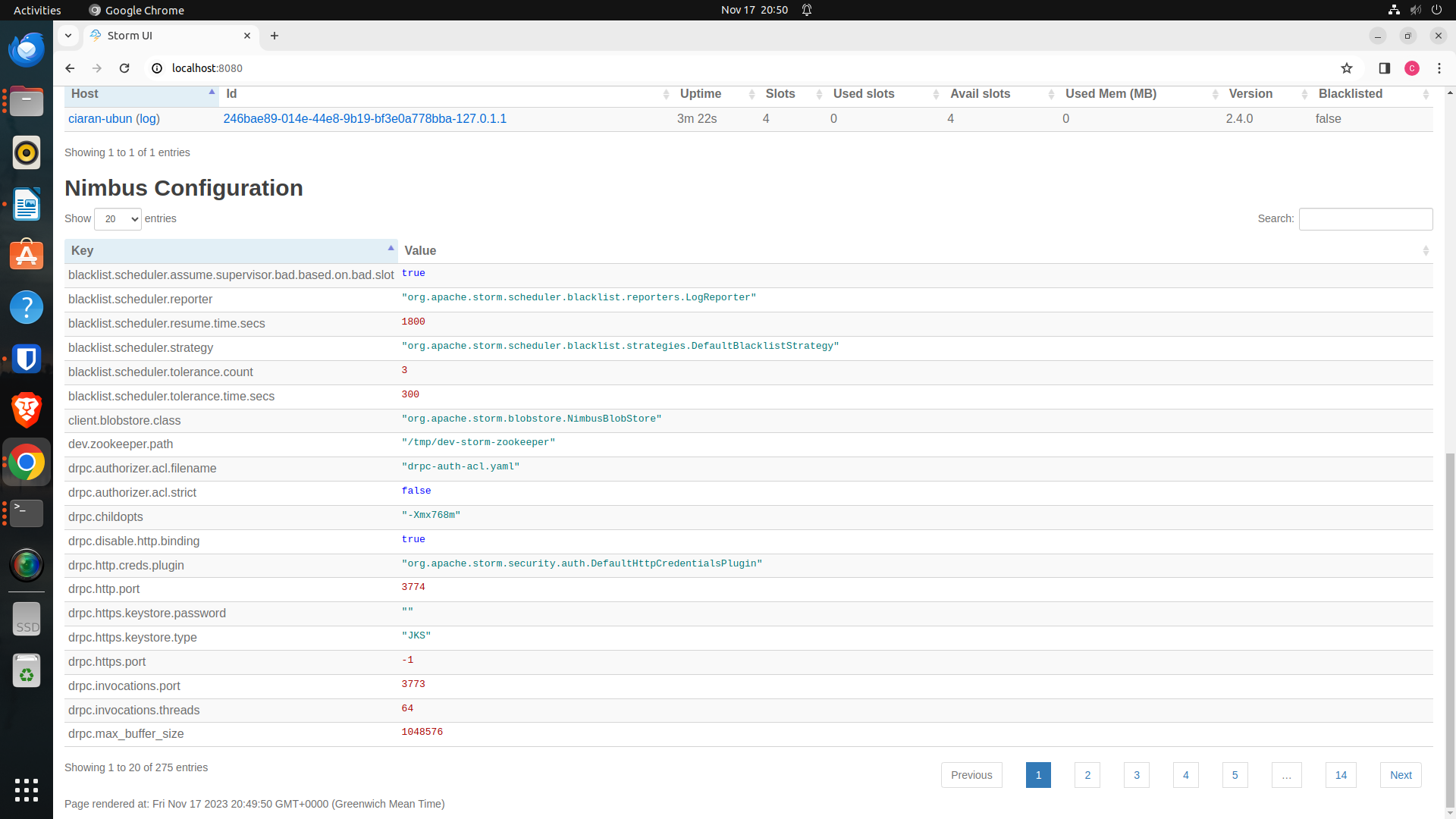
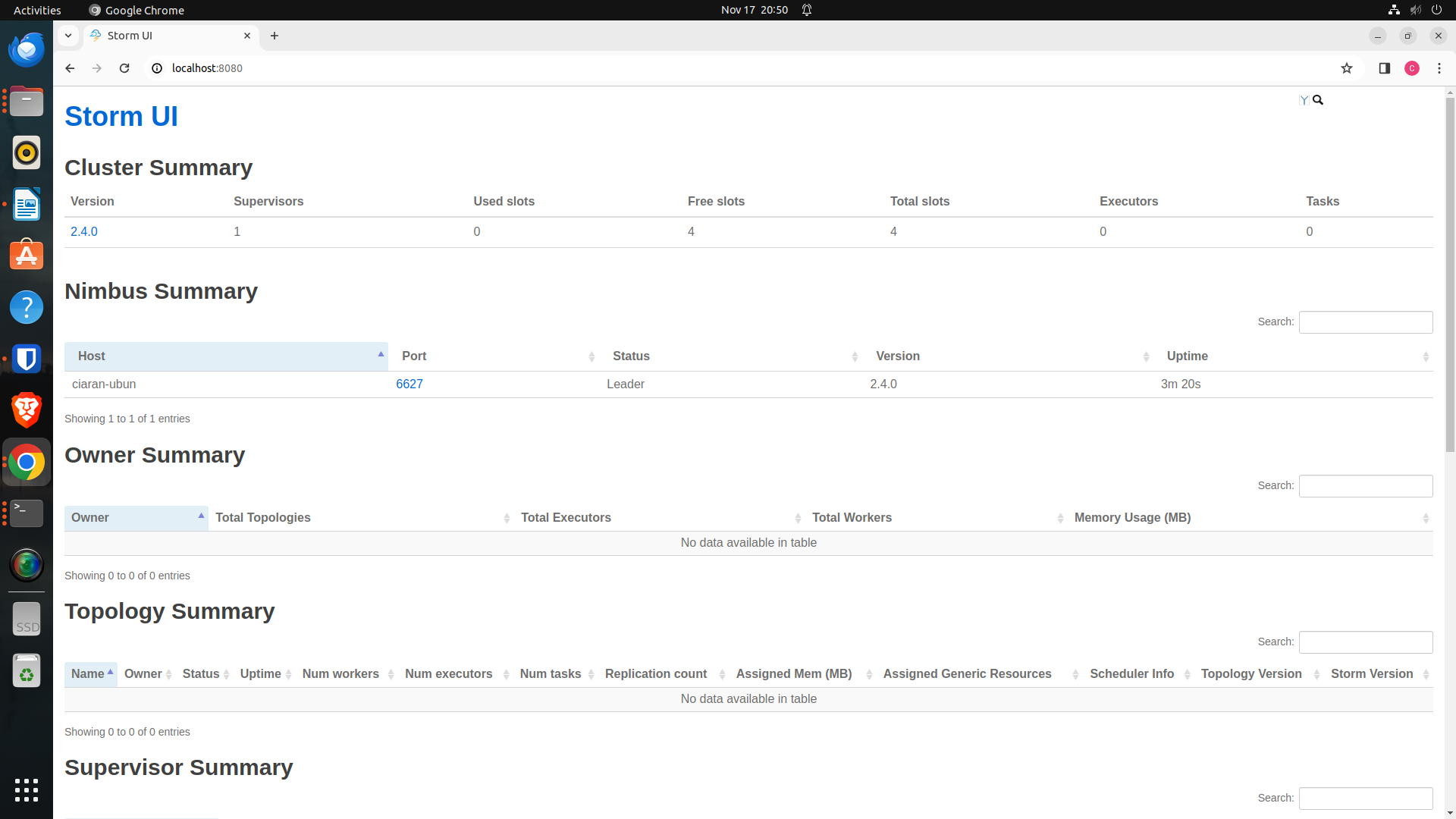
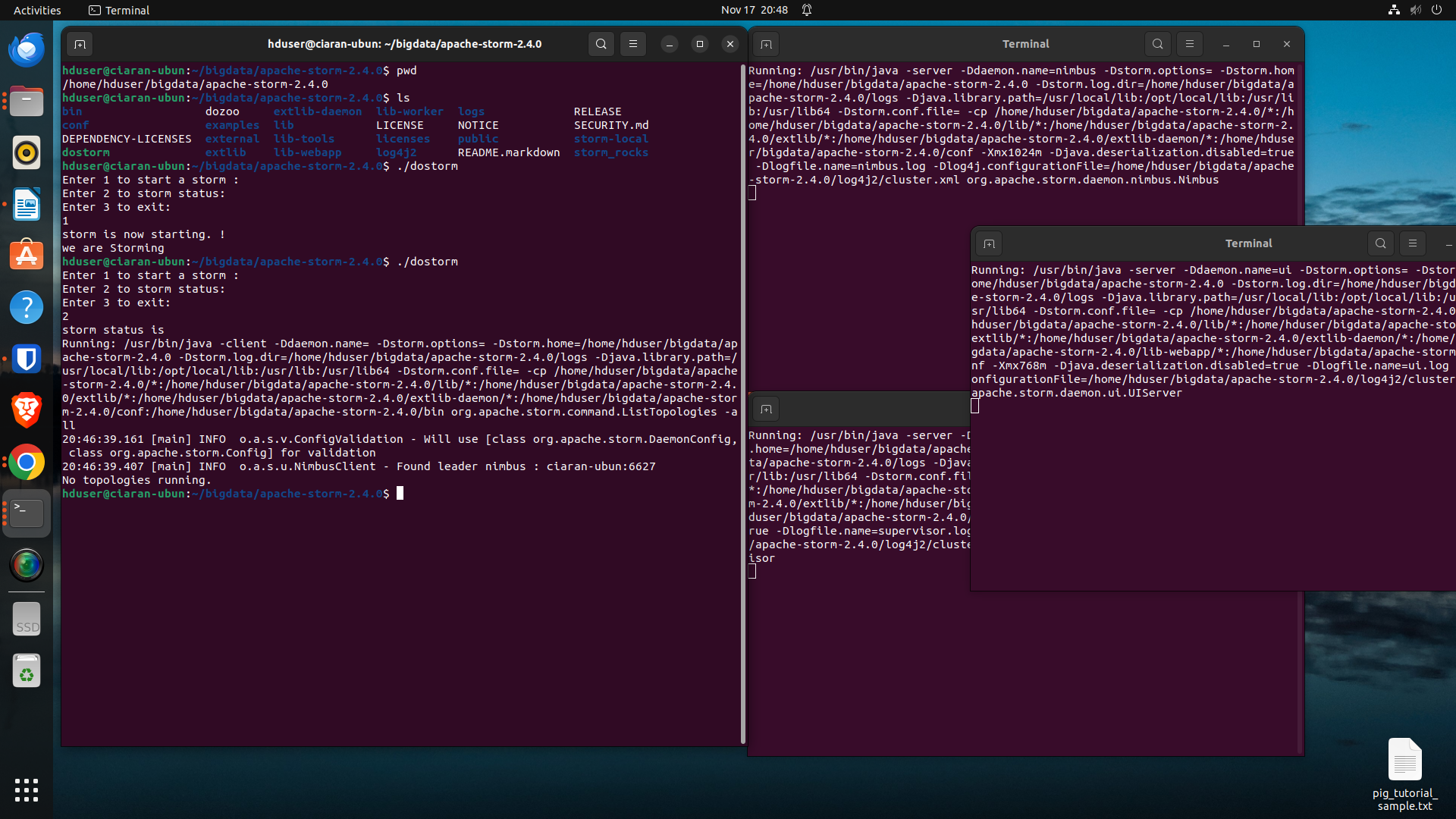
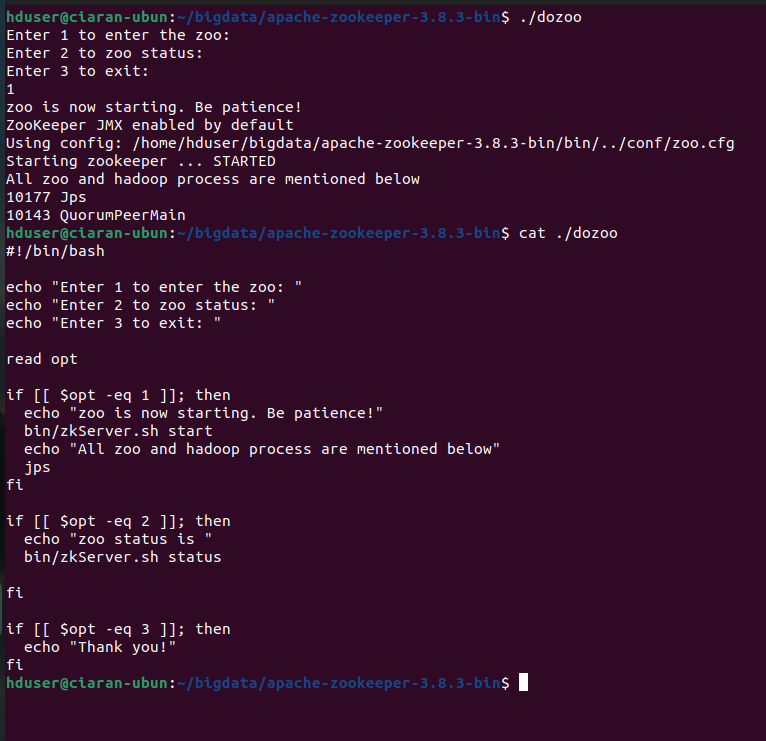
Fault Tolerance:

Apache ZooKeeper plays a critical role in Apache Storm deployments by providing coordination, configuration management, and distributed synchronisation among Storm components. Distributed State Management: Storm uses ZooKeeper for managing distributed states, such as tracking the progress of topologies and worker nodes. It helps maintain the state of tasks and components in the cluster. In summary, Apache ZooKeeper acts as the central nervous system of an Apache Storm cluster, providing coordination and management services to ensure the cluster's proper functioning and fault tolerance.

Screenshot of your VM to show working of Storm UI including Cluster, Nimbus and Owner summary.

All the files are in my storm directory in the repo : <https://github.com/ciaranquinlan/ssbd-ca2/tree/main/storm>

I started Apache zoo with my dozoo script - 1, then started all 3 of the storm processes - 2 and went to localhost:8080 to view the storm process running on my ubuntu 3 and 4. All picture in the repo storm directory..



# Conclusion

In conclusion, the go

# References

## Appendix 1

### Filenames and directories