EXPERIMENT NO: 3

Student Name: Ansari Furqan Ashraf Mohd Shahid

PRN: 211205

Course Name: Big Data Analytics Course Code: CSC702

Dept. name of organization: Computer

Name of organization: M. H. Saboo Siddik College of Engineering

AIM: Implement the Word Count program using Hadoop MapReduce

***Keywords—* Hadoop, Map Reduce, Word Count, JAVA, Cloudera**

***Abstract***

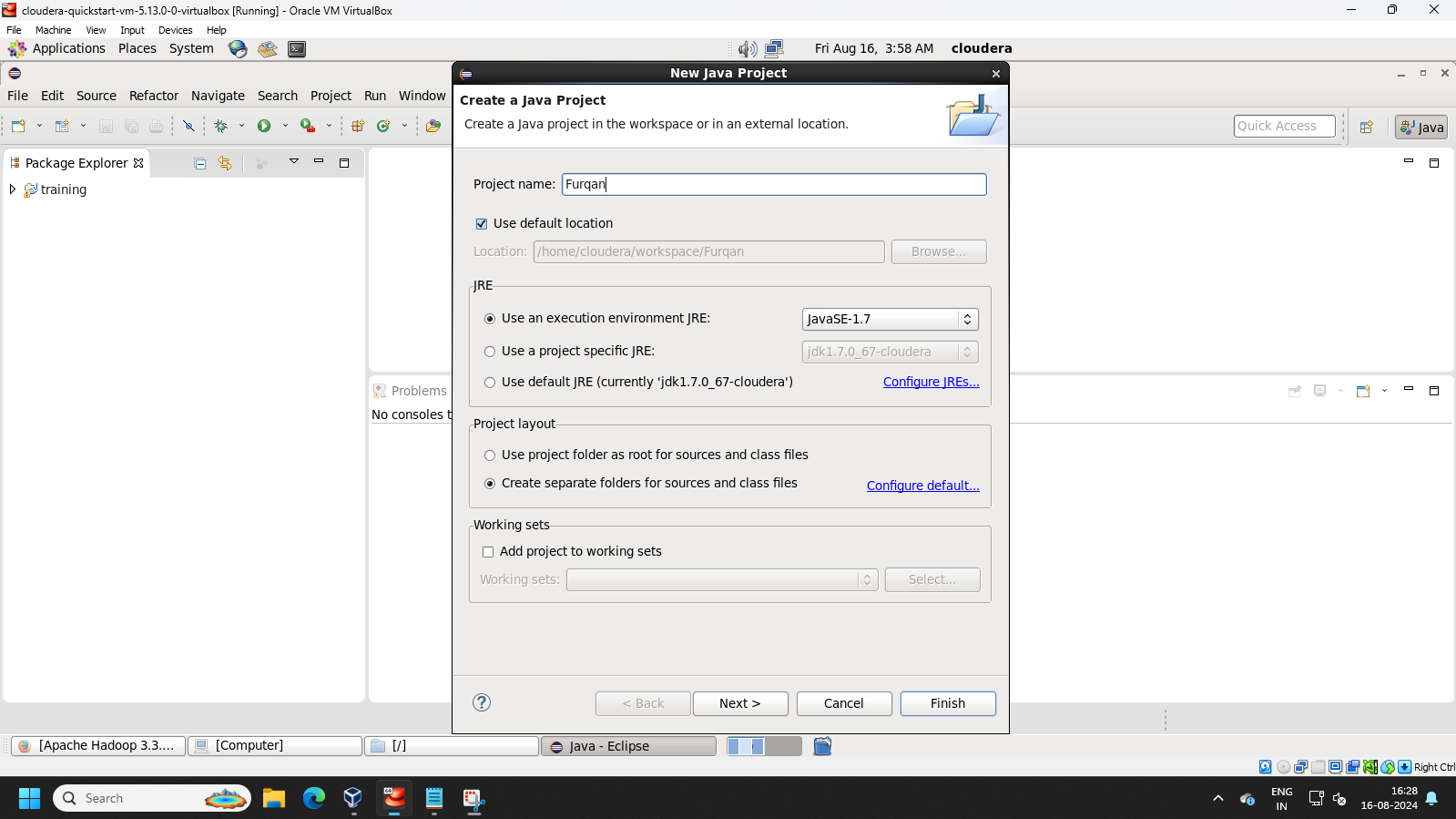
The Word Count Problem is a fundamental task in the realm of big data processing and serves as an introductory exercise for understanding Hadoop's capabilities. In Cloudera’s Hadoop environment, the Word Count Problem involves counting the frequency of each word in a large dataset. This problem is typically approached using Hadoop’s MapReduce framework, which divides the data processing task into two phases: Map and Reduce. In the Map phase, data is split into manageable chunks and processed by mapper functions to produce key-value pairs, where the key is the word and the value is its count. The Reduce phase then aggregates these key-value pairs, summing up the counts for each word to produce the final output. Leveraging Cloudera’s distribution of Hadoop, which includes tools like Apache Hive and Apache Pig, enhances the efficiency and scalability of this process, allowing for the handling of vast amounts of data across distributed clusters. The Word Count Problem not only demonstrates the basic principles of Hadoop’s distributed computing but also illustrates the power of Cloudera’s ecosystem in managing and analyzing big data effectively.

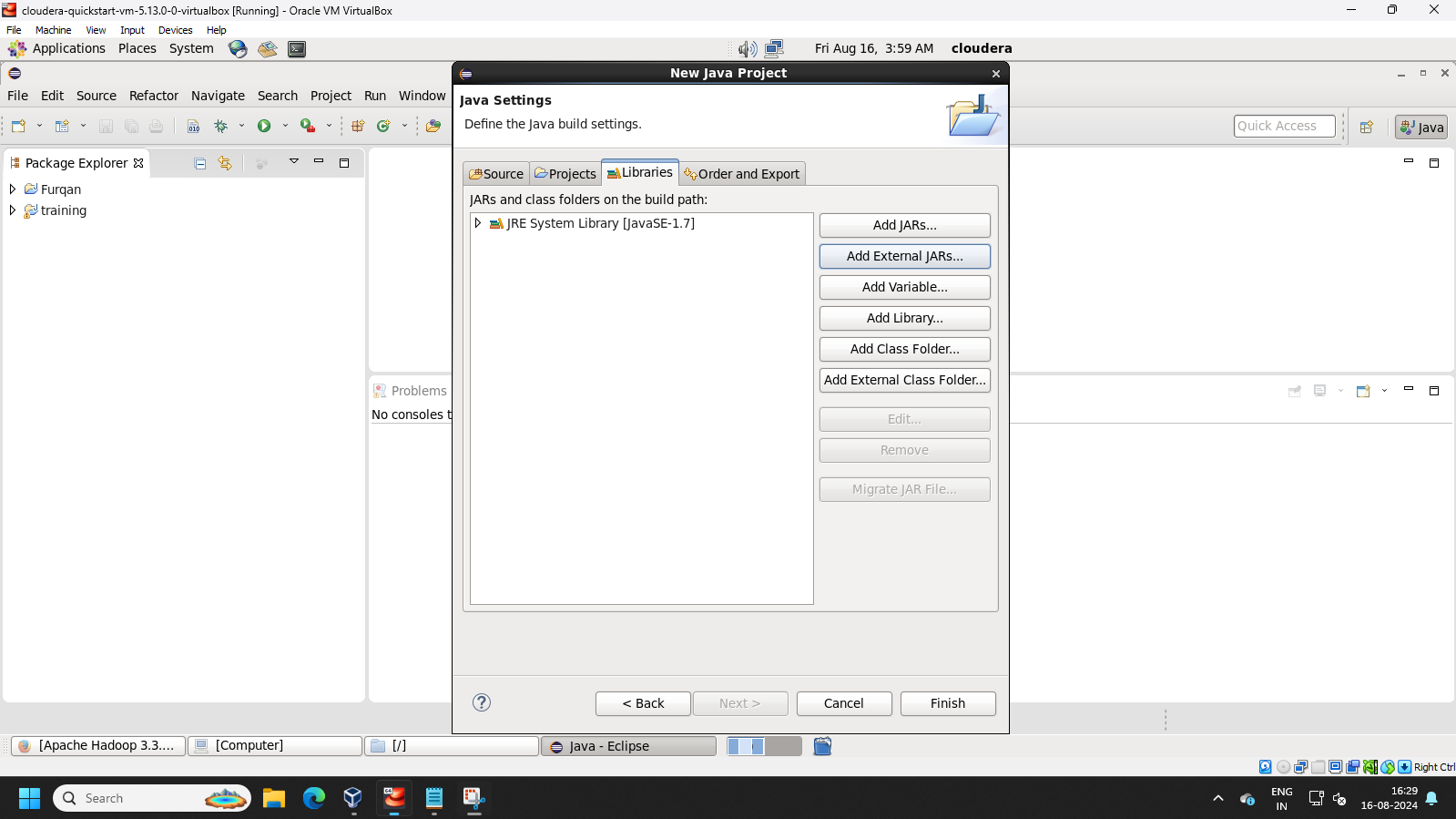
# Introduction

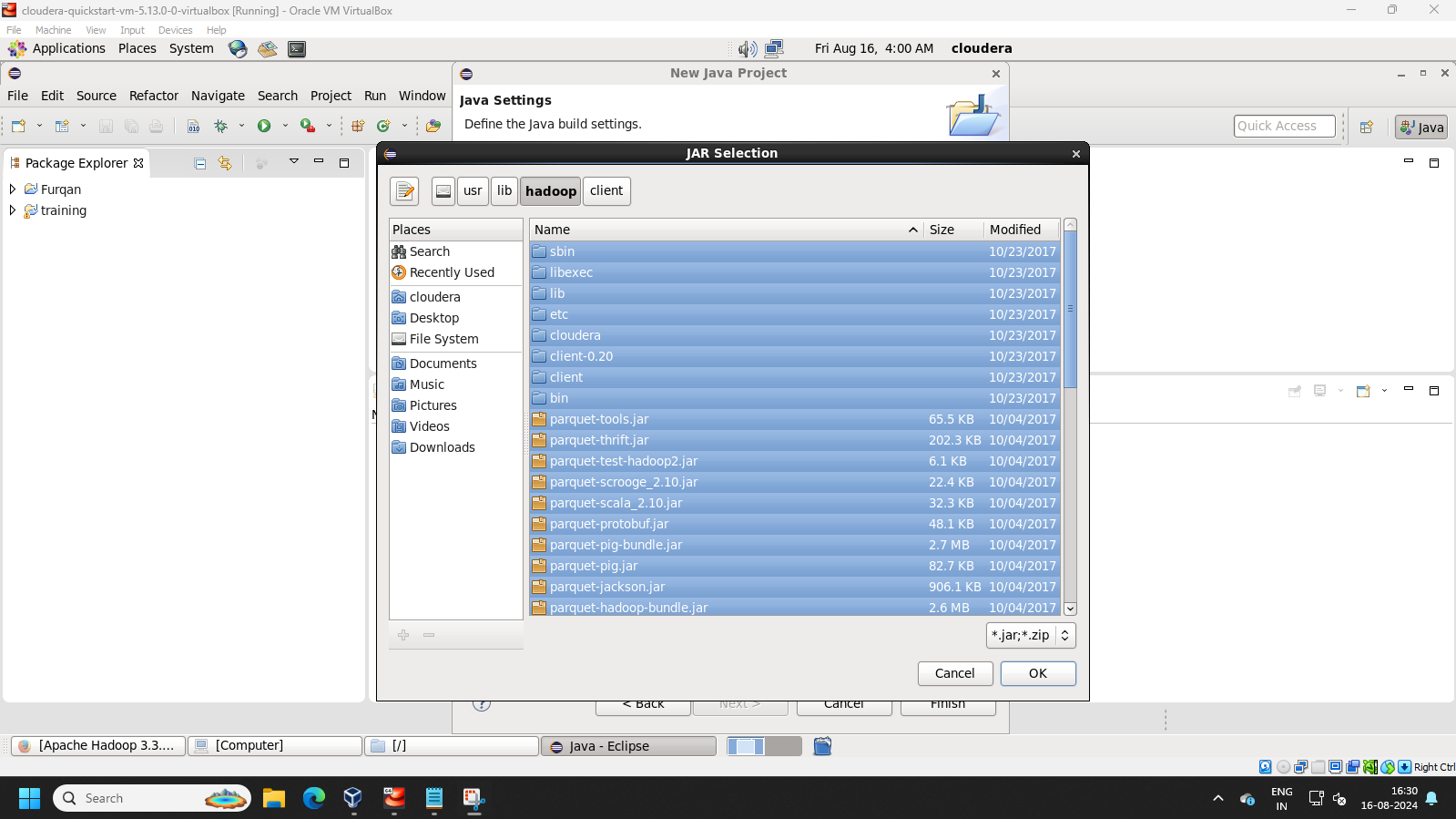
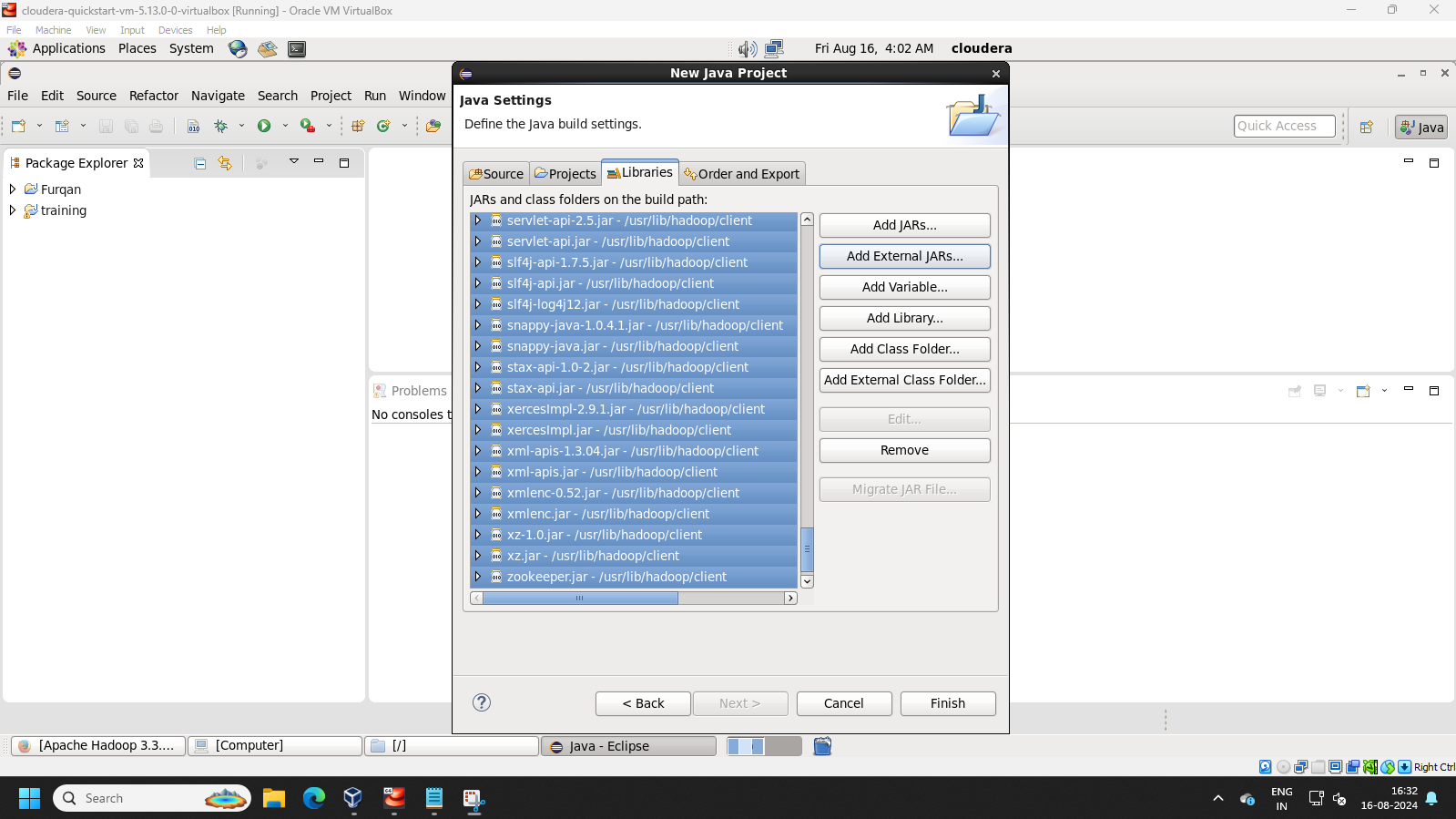
The Word Count Problem is a classic example used to illustrate the principles of distributed computing and data processing with Hadoop. As data volumes continue to grow exponentially, traditional methods of data analysis become insufficient, necessitating the use of scalable and efficient tools. Hadoop, an open-source framework designed for distributed storage and processing, provides a robust solution to this challenge. Within Hadoop, the MapReduce programming model is instrumental in performing large-scale data processing tasks by dividing the work into manageable units. Cloudera, a leading provider of Hadoop-based solutions, offers a comprehensive platform that enhances the Hadoop ecosystem with additional tools and optimizations.

In the context of the Word Count Problem, Cloudera’s distribution facilitates efficient processing by leveraging its advanced Hadoop ecosystem, which includes components like Apache HDFS for distributed storage, and Apache YARN for resource management. By distributing the word counting task across a cluster of machines, Hadoop and Cloudera’s technologies enable the processing of vast datasets with high performance and fault tolerance. This introduction to the Word Count Problem within the Cloudera environment sets the stage for exploring how

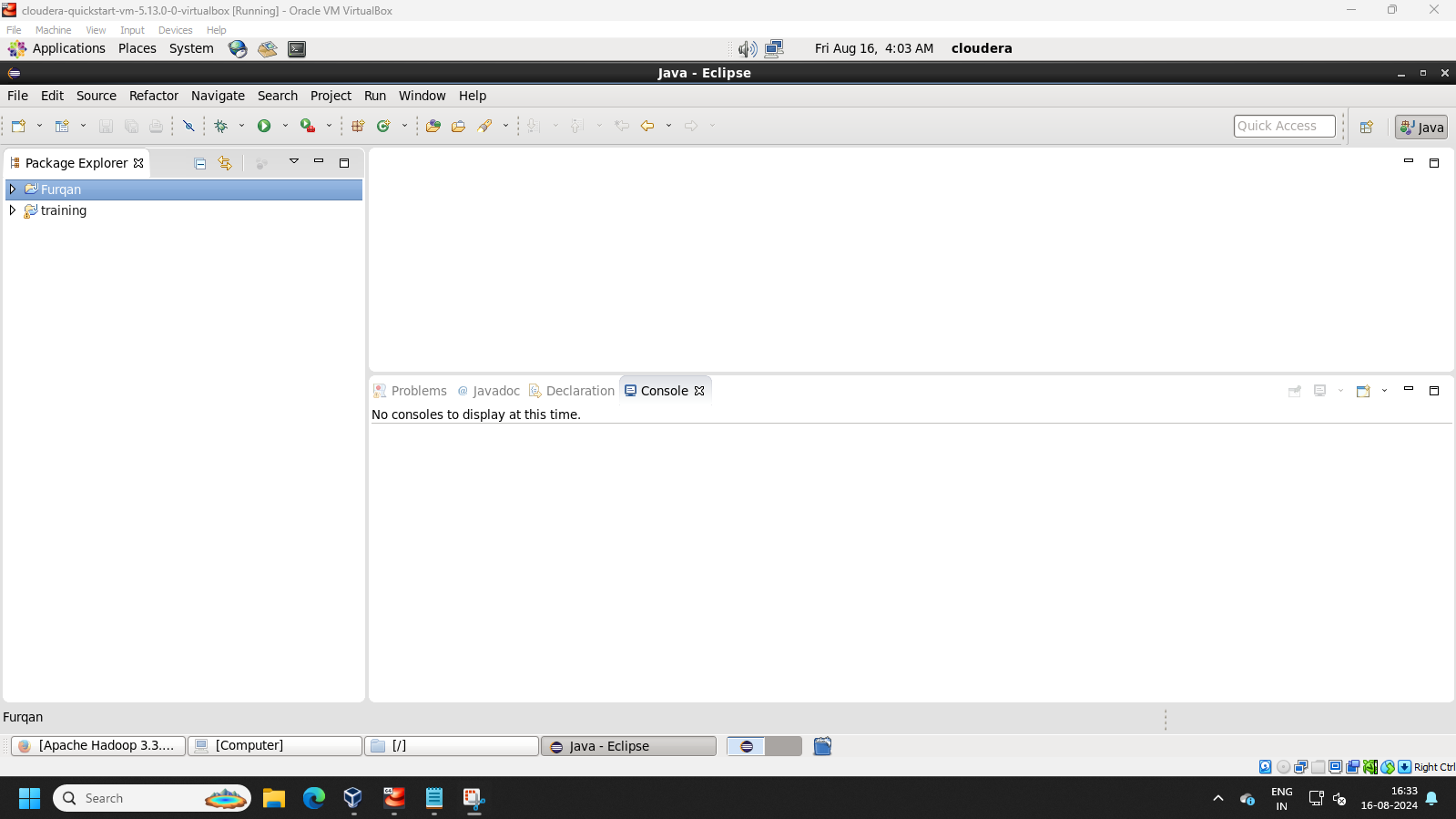
Hadoop’s MapReduce model operates in practice and highlights the advantages of using Cloudera’s platform for handling large-scale data processing tasks.

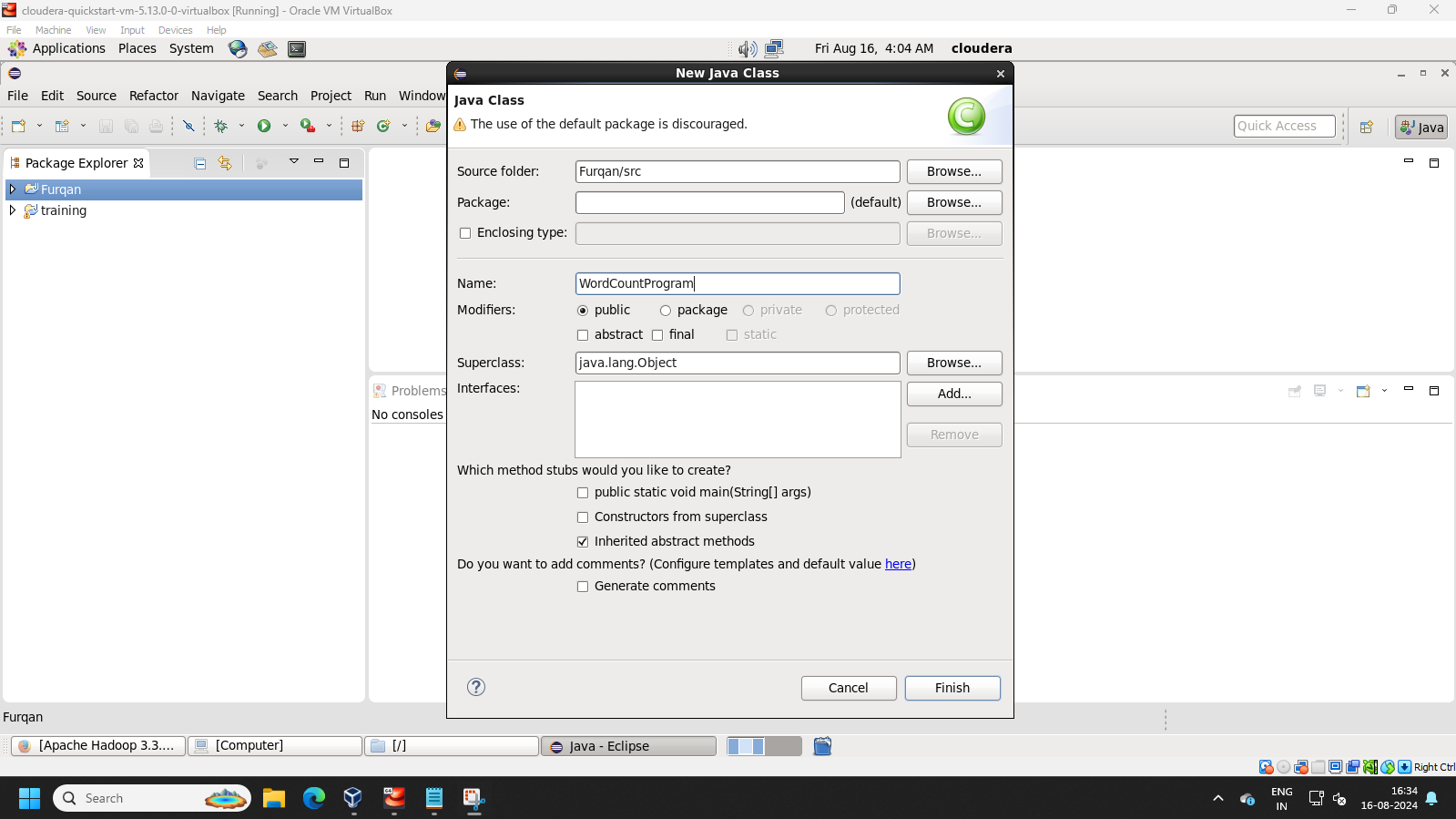
1. Project Name
2. Adding External Jars



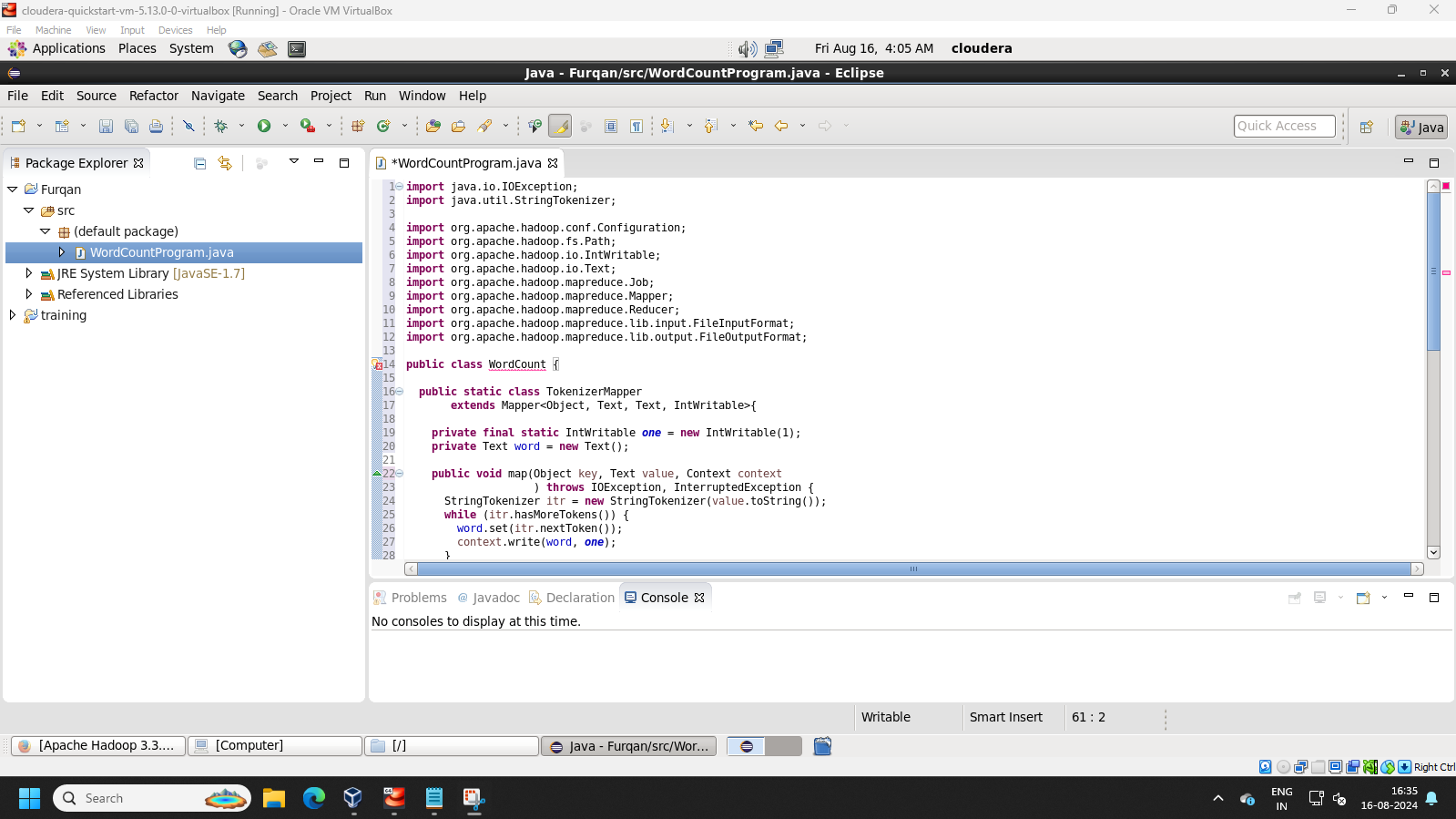
  


1. Creating a class WordCountProgram

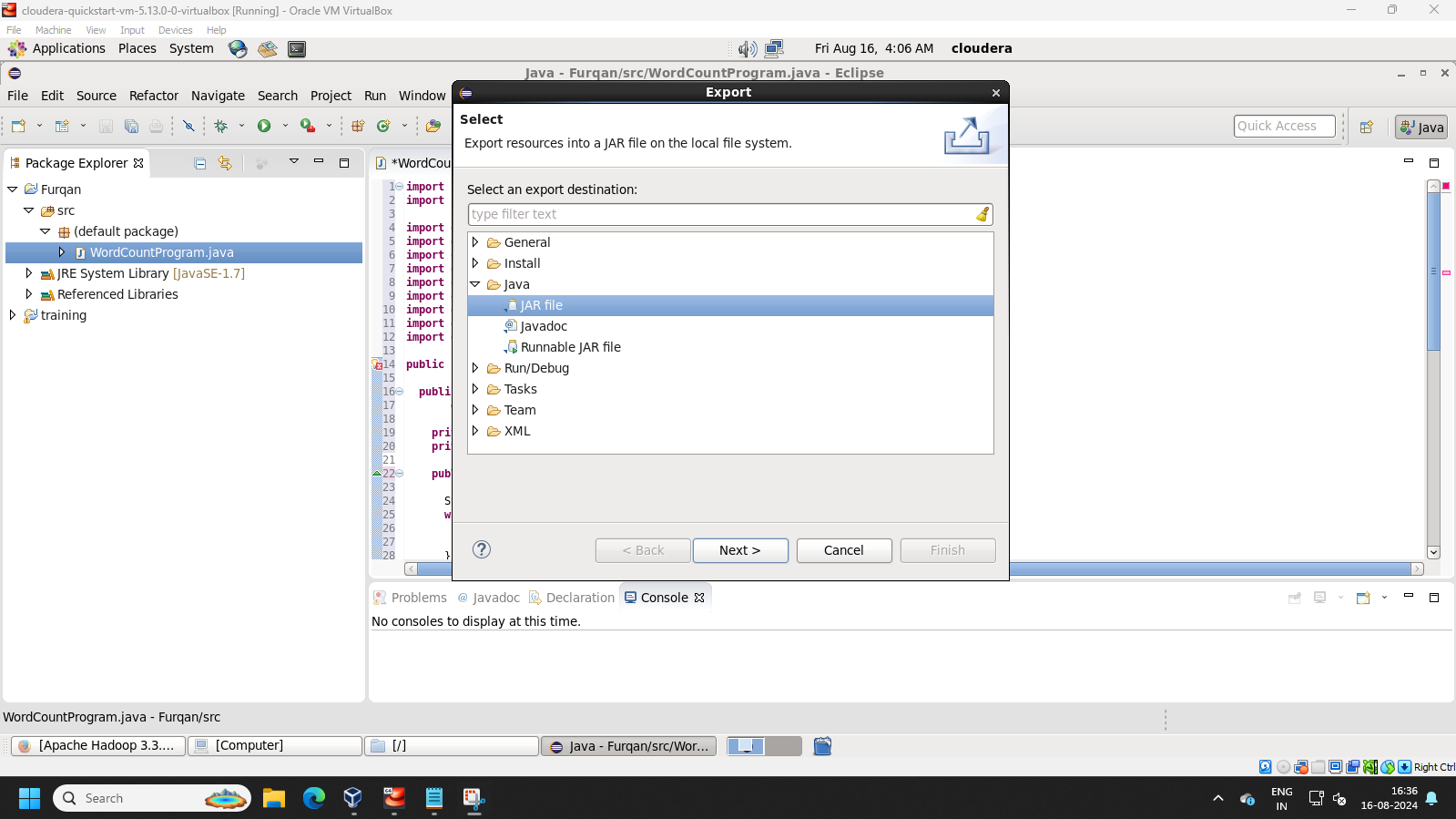


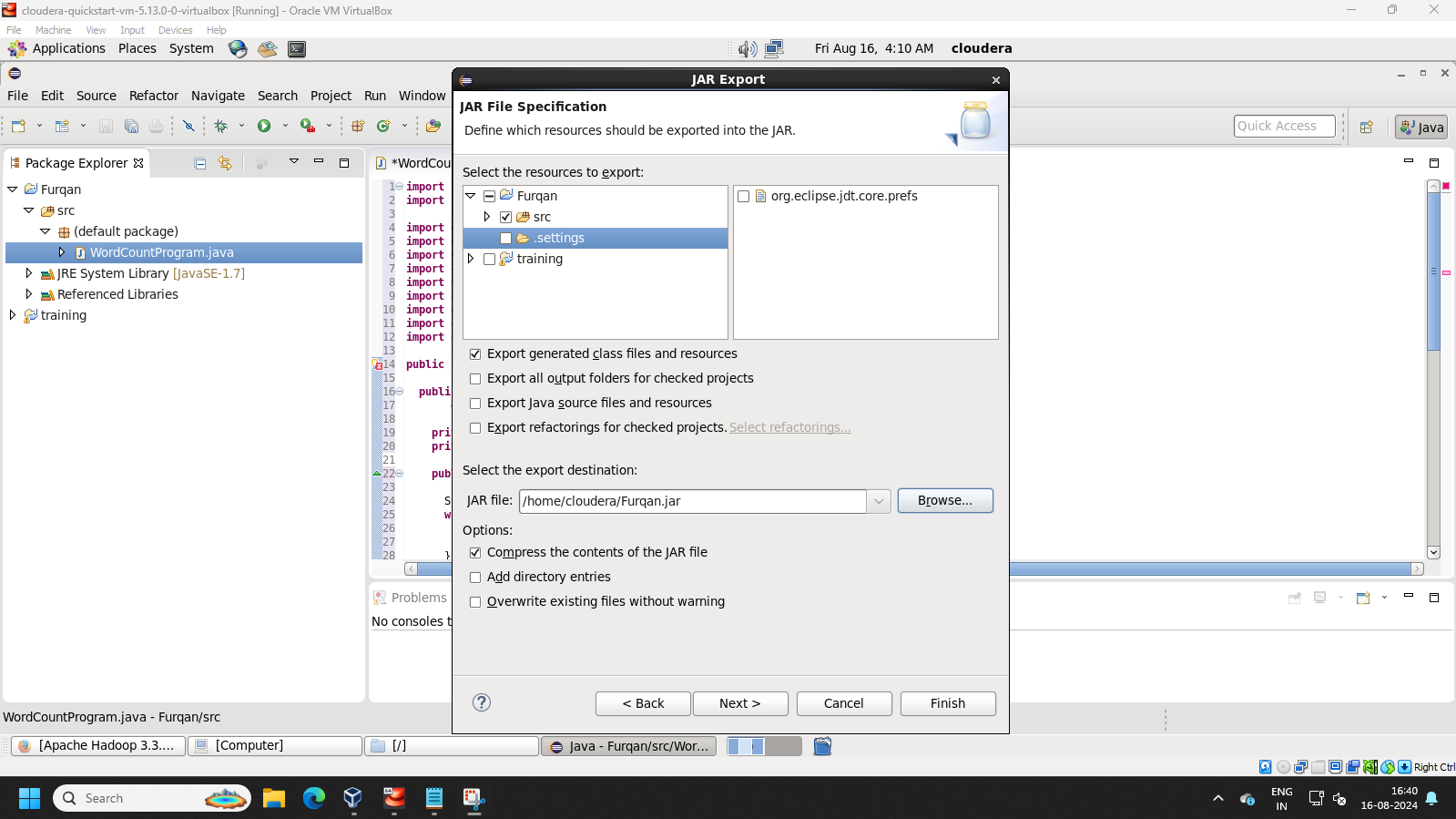


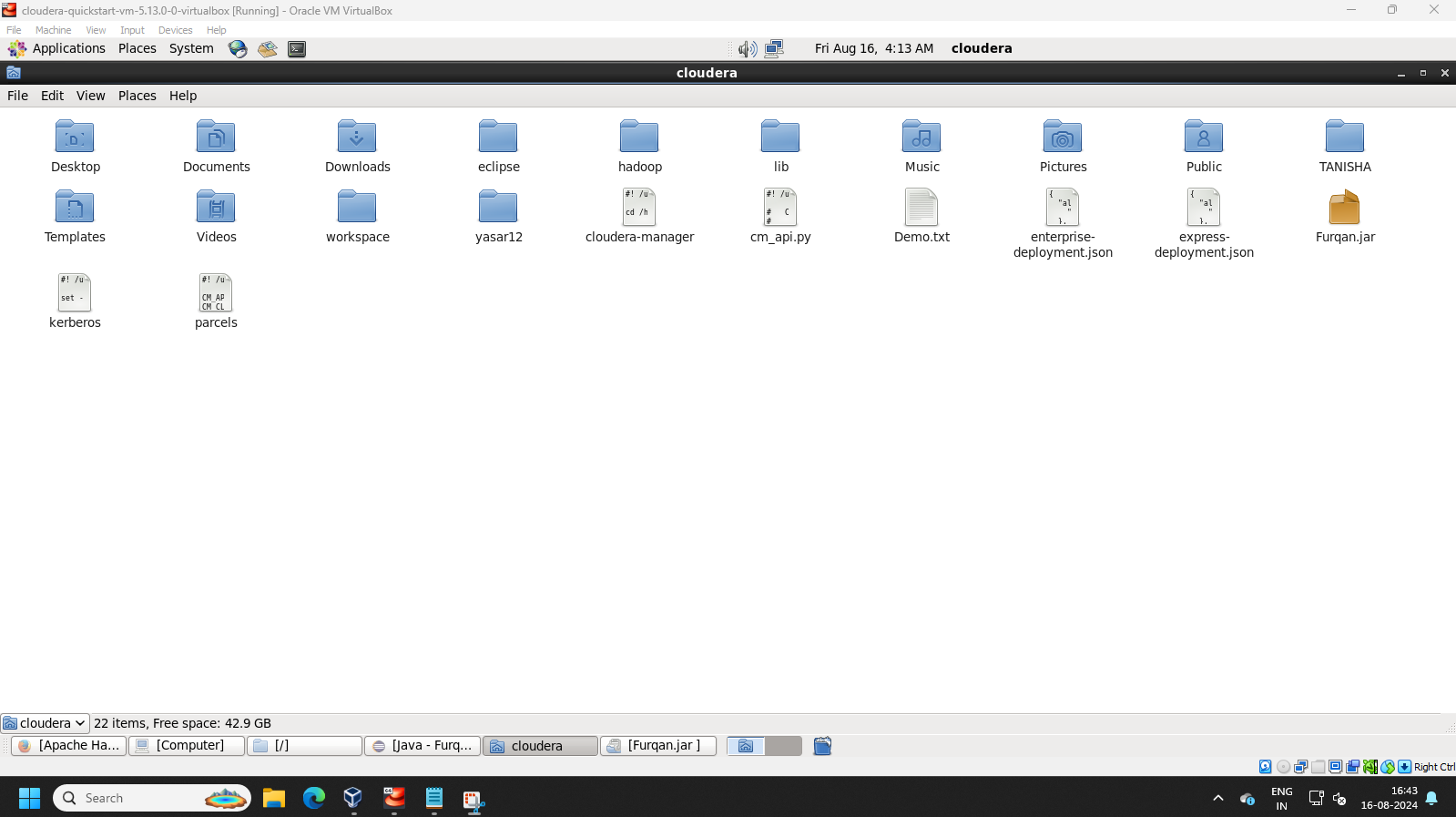
1. Copy Code



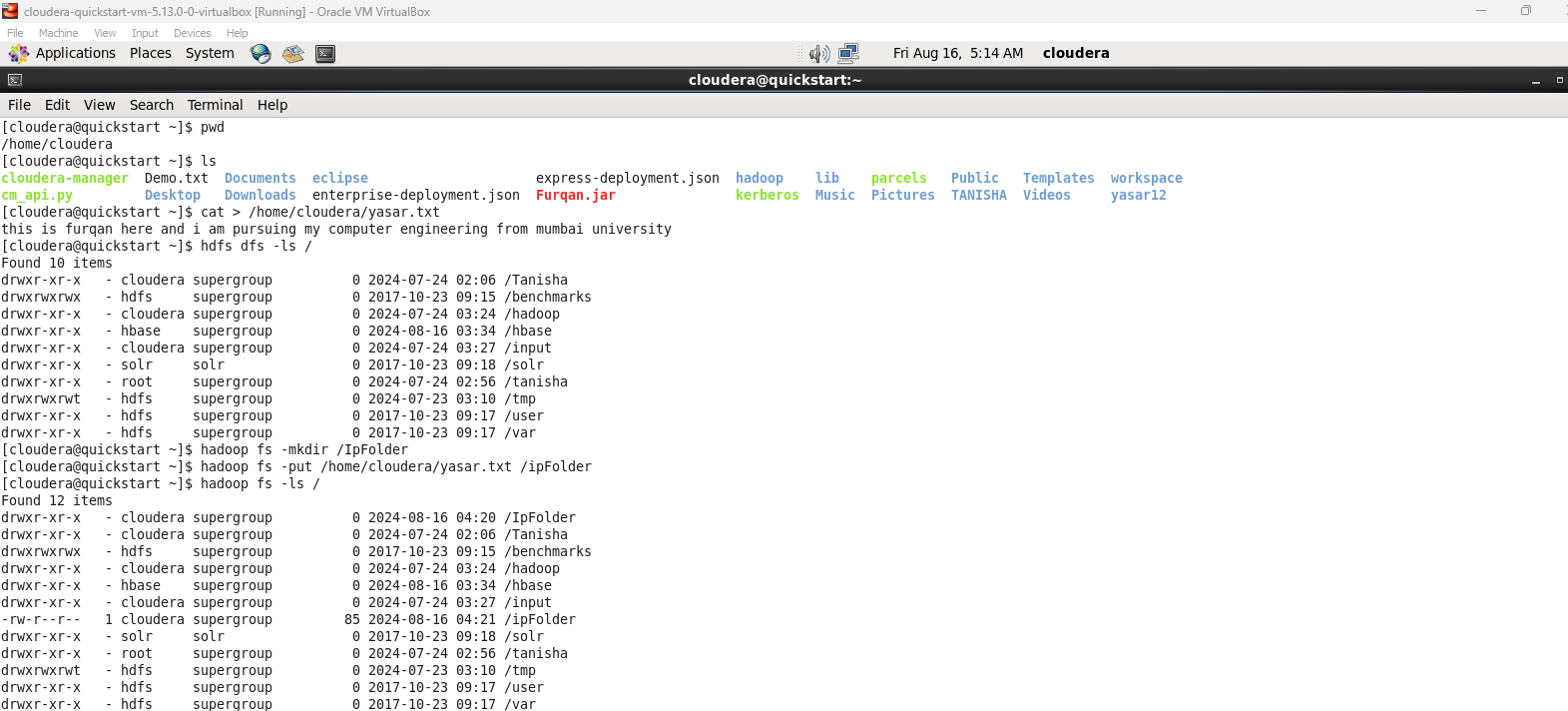
1. Creating jar file and exporting

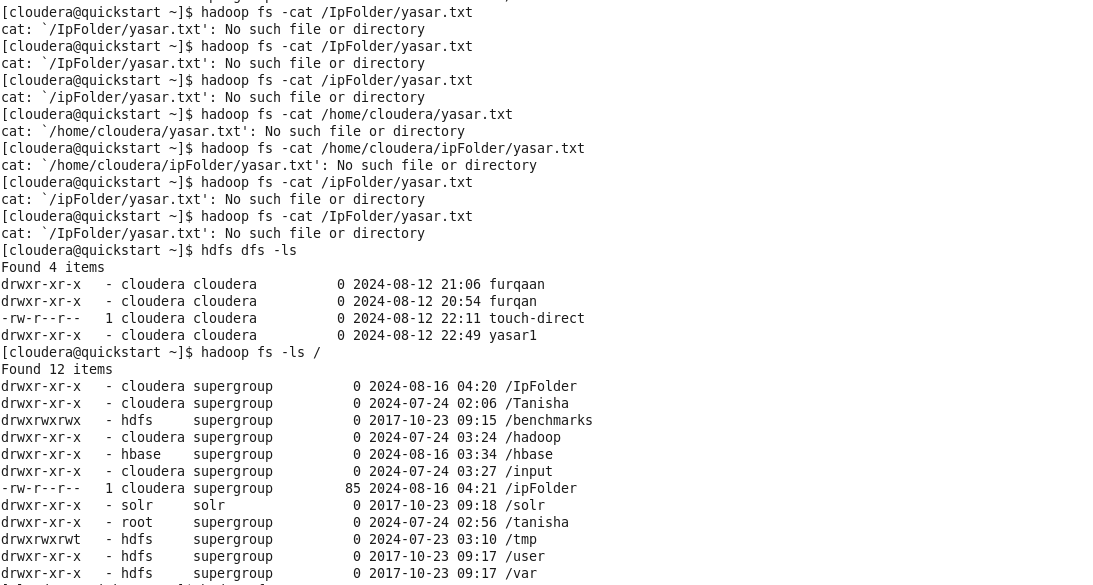


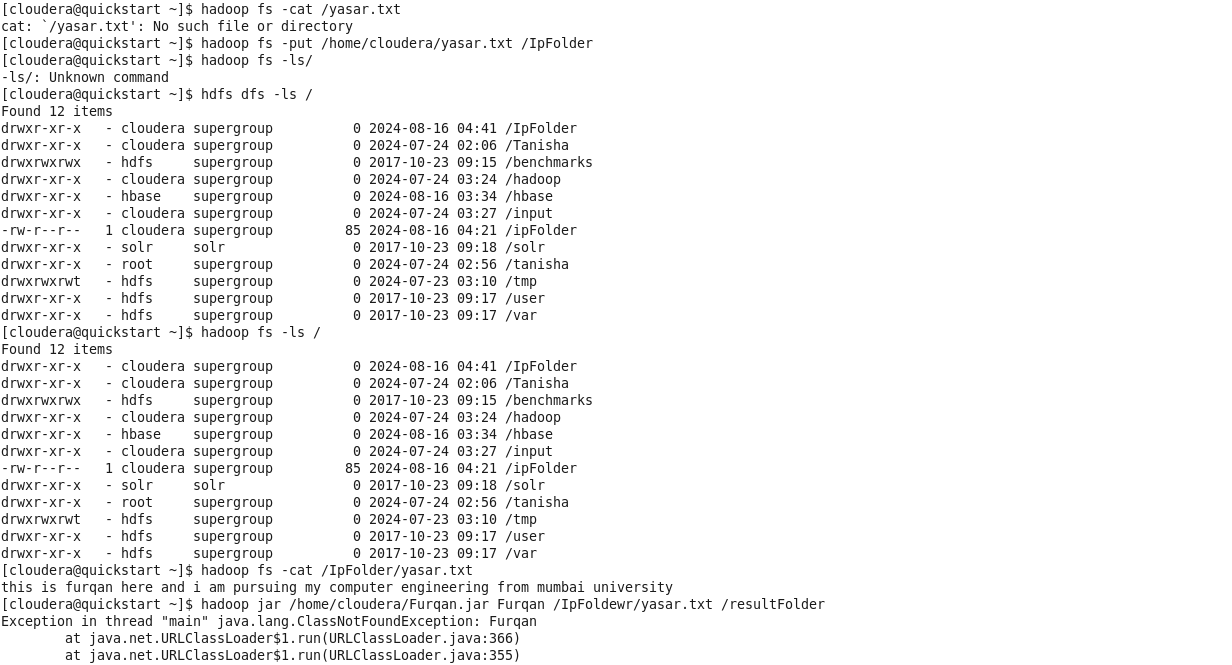


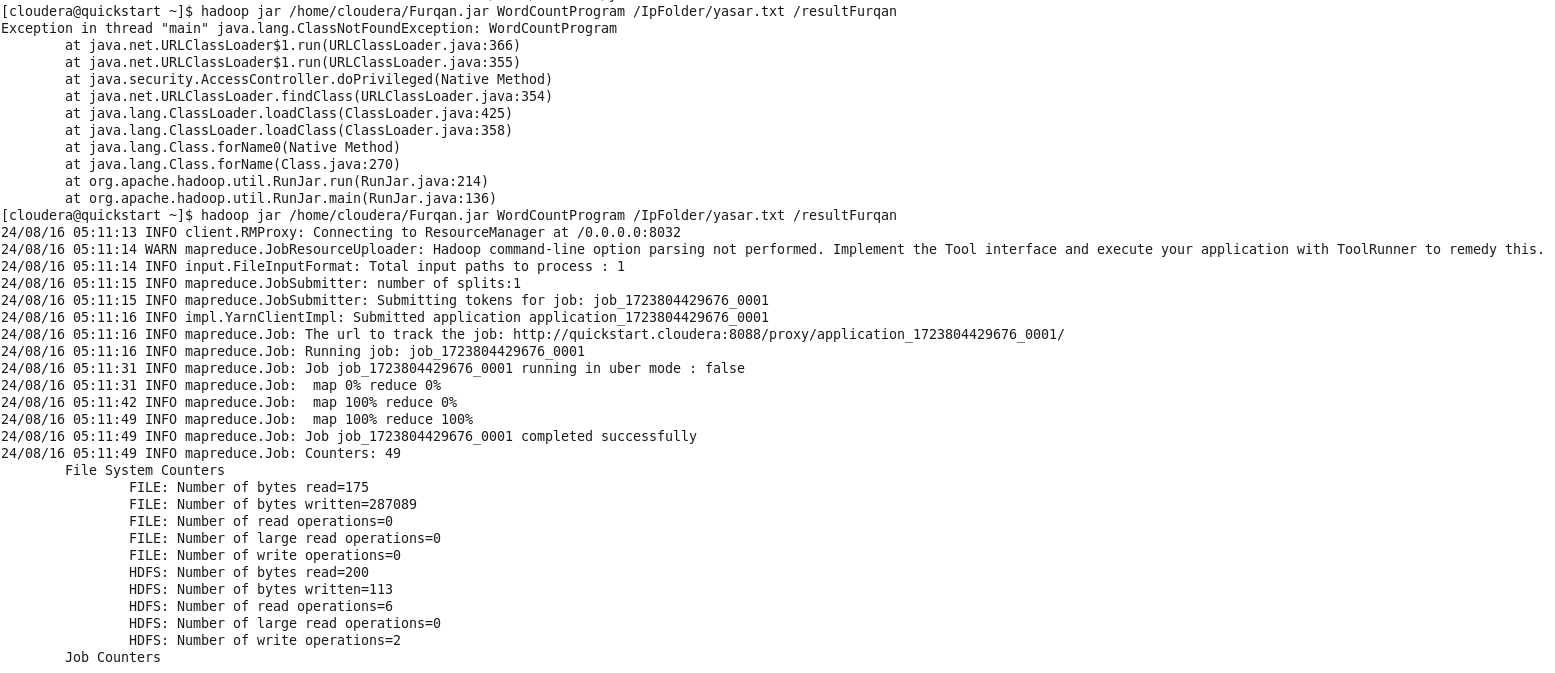


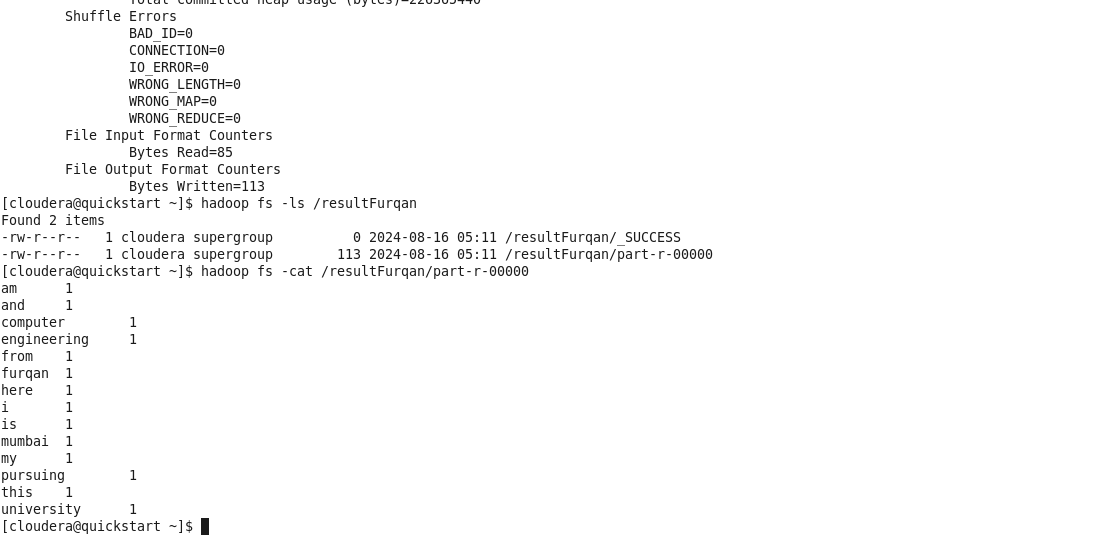
1. Open Terminal to execute the program











# Conclusion

The conclusion of the MapReduce process in Hadoop underscores its pivotal role in enabling large-scale data processing by efficiently distributing and parallelizing tasks across a cluster of nodes. By breaking down a job into smaller map and reduce tasks, MapReduce allows for the processing of vast datasets in a fault-tolerant and scalable manner. The map phase is responsible for filtering and sorting data, while the reduce phase aggregates and consolidates the results, ultimately producing the final output. This framework's ability to handle massive volumes of data with reliability and speed has made it a cornerstone technology in the realm of big data, powering everything from search engines to data analytics platforms.

REFERENCES

1. Apache Hadoop, "MapReduce Tutorial," [Online]. Available: [https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/](https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html) [MapReduceTutorial.html](https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html). [Accessed: 10-Aug-2024].
2. Exp 5 - Map Reducer - WordCount.docx, Google Classroom, 2024. [Accessed: 30-July-2024].