

**ACADEMIC YEAR 2020-2021**

Logo

Description automatically generated

**BIGDATA LABORATORY**

*Report on,*

**Learning Activity II-Programming Assignment**

*Submitted by,*

**Harshith BN (1NT18IS069)**

*Submitted to,*

**Ms. Disha DN,**

Department of Information Science and Engineering Nitte Meenakshi Institute of Technology

Bangalore-064

**DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING**

**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

**(An autonomous institution with A+ Grade by NAAC /UGC, Affiliated to Visvesvaraya Technological University, Belgaum, Approved by UGC/AICTE/Govt. of Karnataka)**

**Yelahanka, Bengaluru-560064**

Source Code:

github.com/1NT18IS069/1NT18IS069\_harshithbn\_A\_bdLab

**Brief note on Hadoop and Map Reduce**

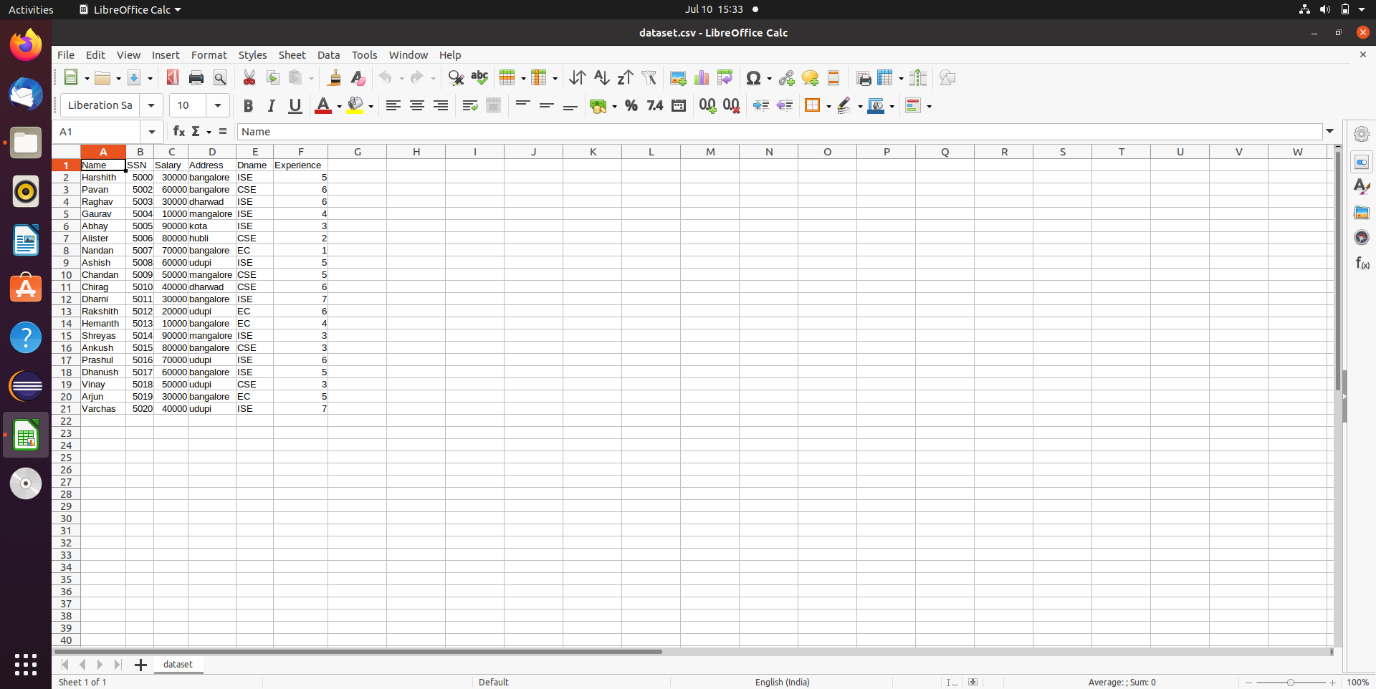
Hadoop is an Apache opensource framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.

The Hadoop Distributed File System (HDFS) is based on the Google File System (GFS) and provides a distributed file system that is designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. It is highly fault-tolerant and is designed to be deployed on low- cost hardware. It provides high throughput access to application data and is suitable for applications having large datasets.

MapReduce is a parallel programming model for writing distributed applications devised at Google for efficient processing of large amounts of data (multi-terabyte datasets), on large clusters (thousands of nodes) of commodity hardware in a reliable, fault-tolerant manner. The MapReduce program runs on Hadoop which is an Apache open-source framework. It is quite expensive to build bigger servers with heavy configurations that handle large scale processing, but as an alternative, you can tie together many commodity computers with single-CPU, as a single functional distributed system and practically, the clustered machines can read the dataset in parallel and provide a much higher throughput.

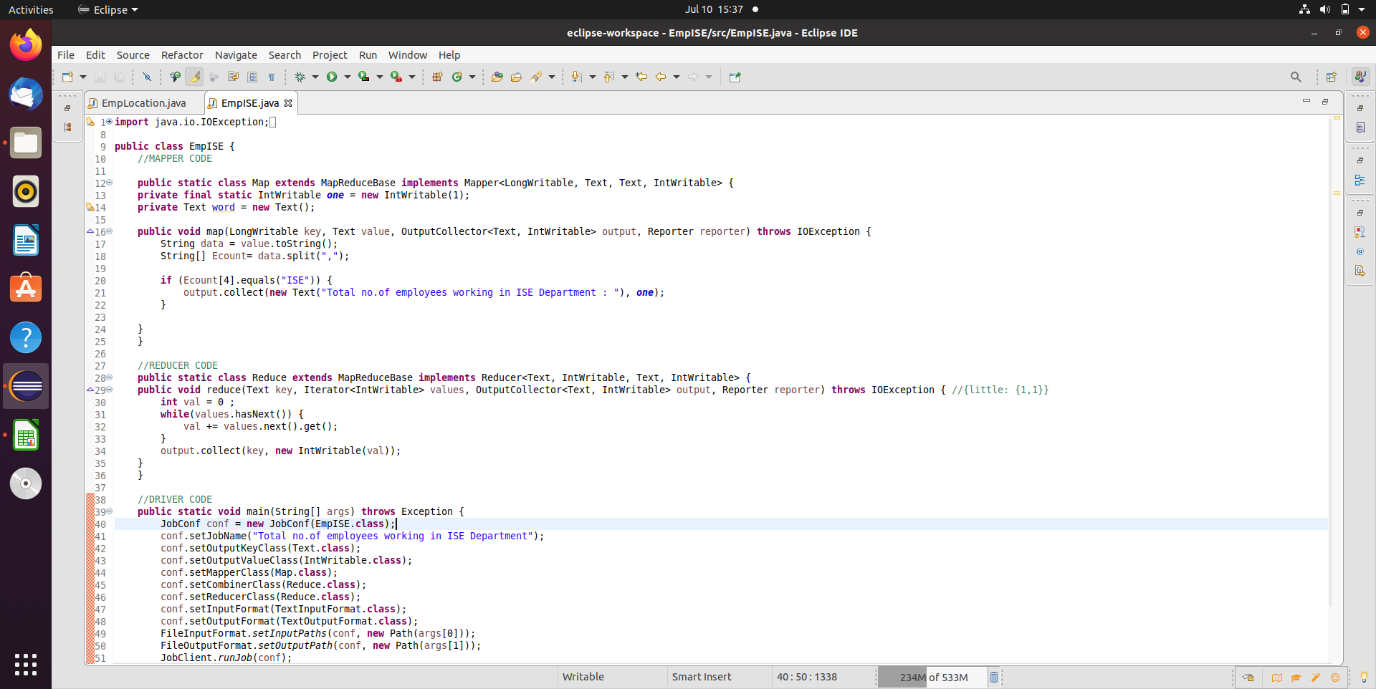
**EXERCISE 1**

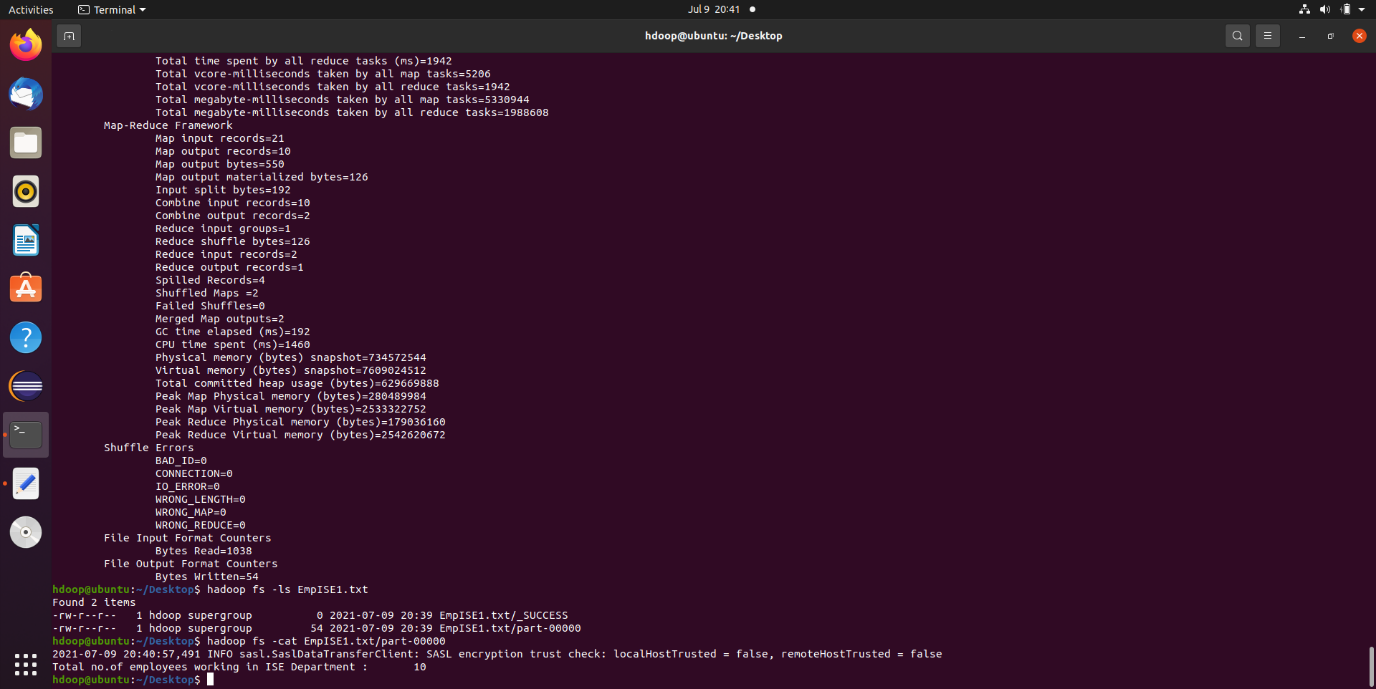
DATASET:



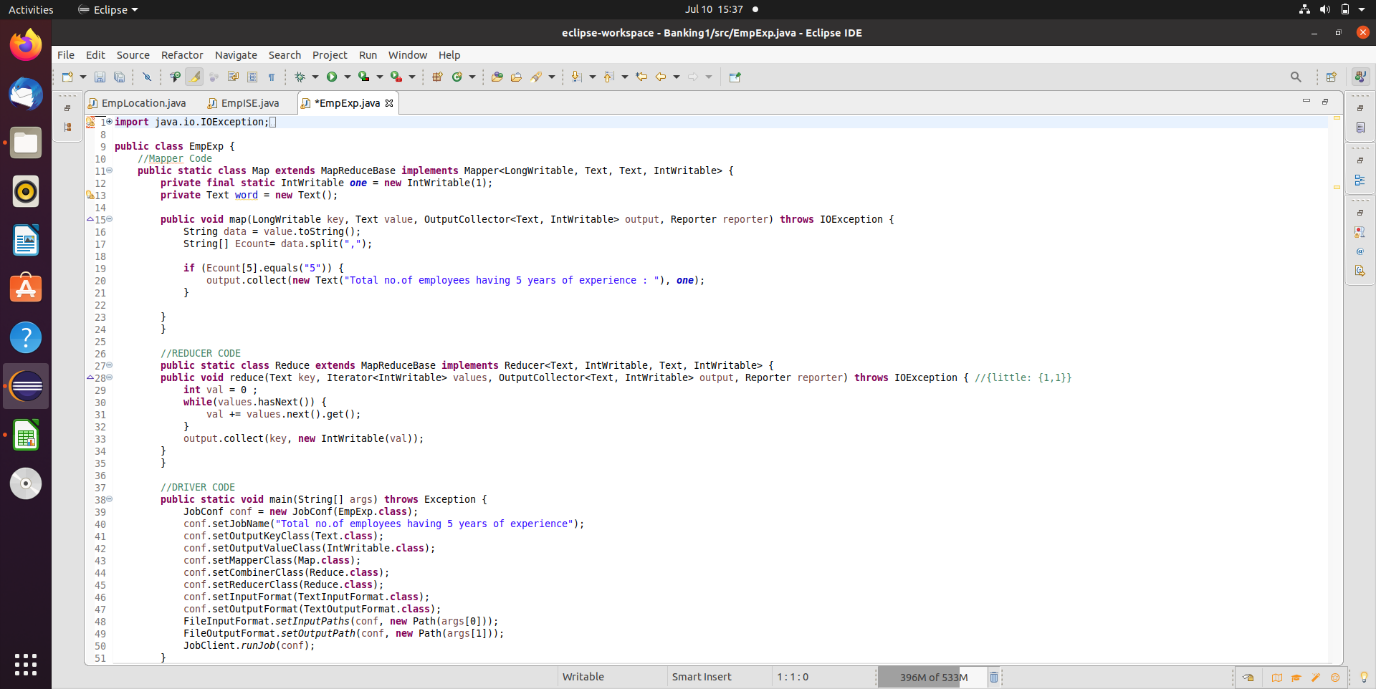
OUTPUT’s

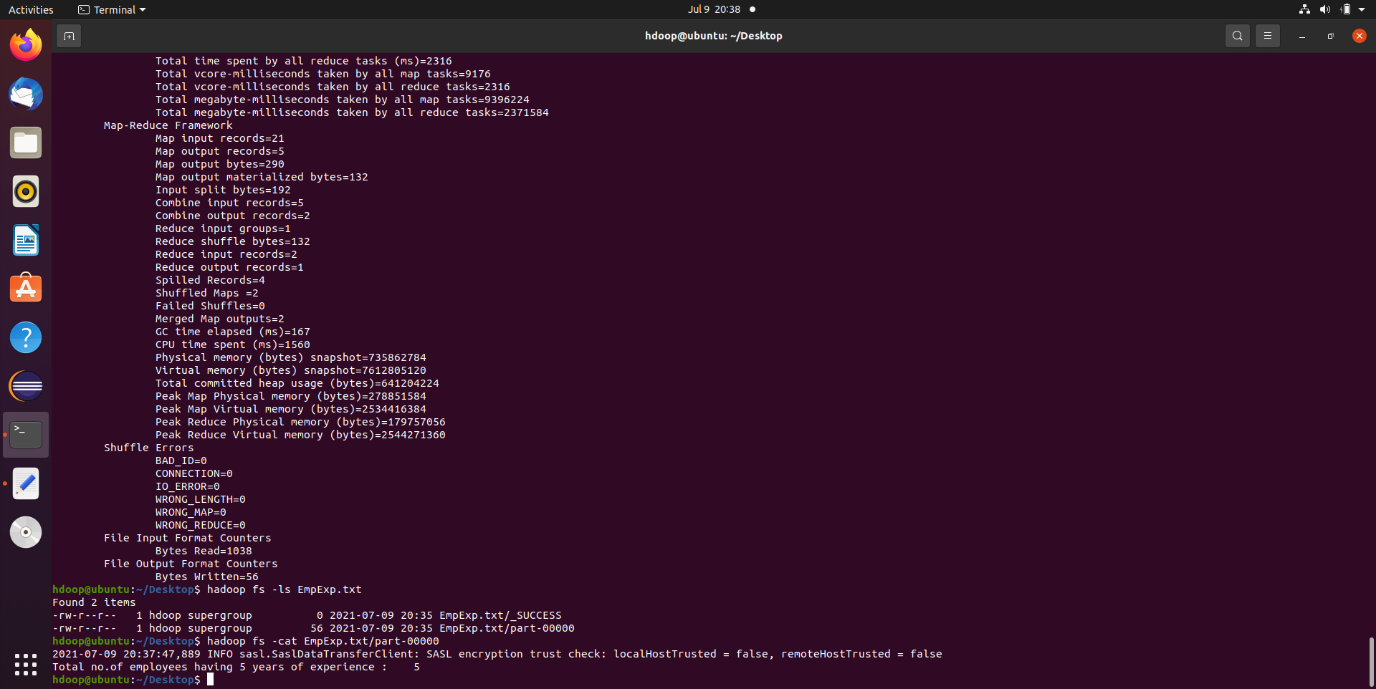
1. Total number of employees who work in ISE department



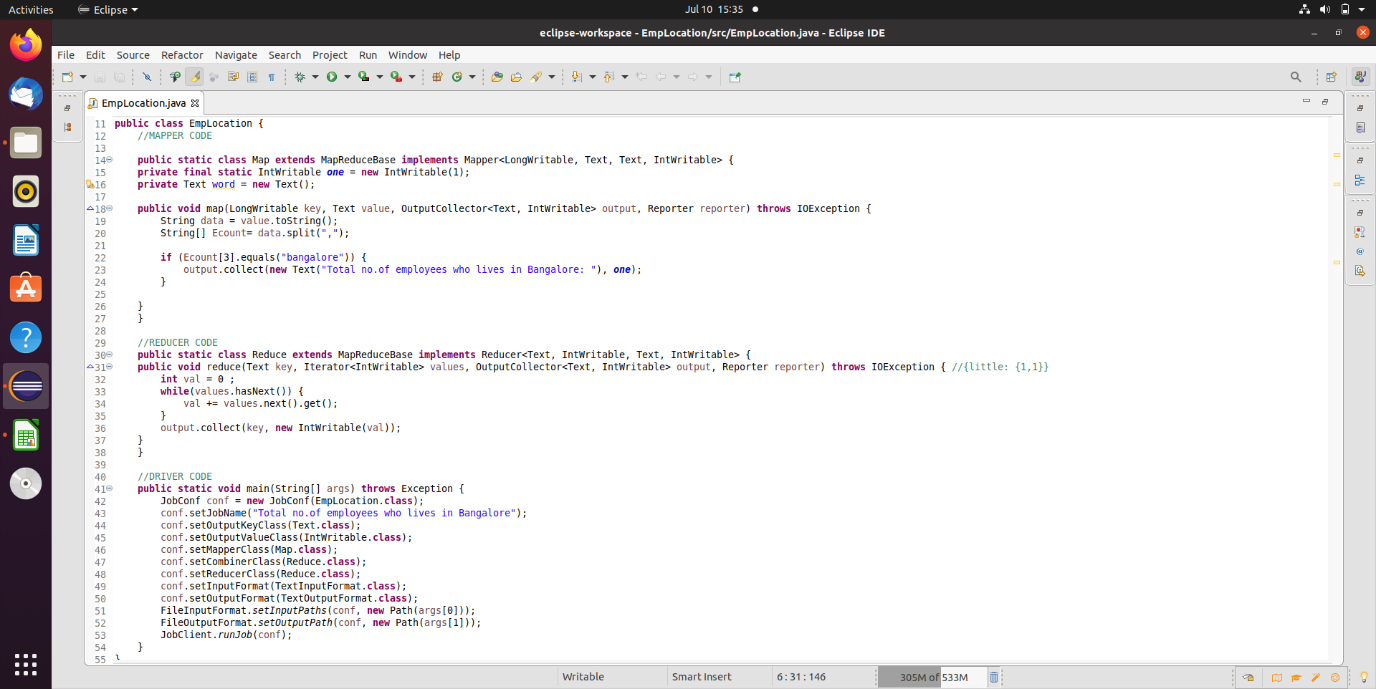


1. Total number of employees with experience=5 years





1. Count the number of employees who lives in Bangalore





**HIVE**

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy.

Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

Hive is not

* + A relational database
  + A design for OnLine Transaction Processing (OLTP)
  + A language for real-time queries and row-level updates

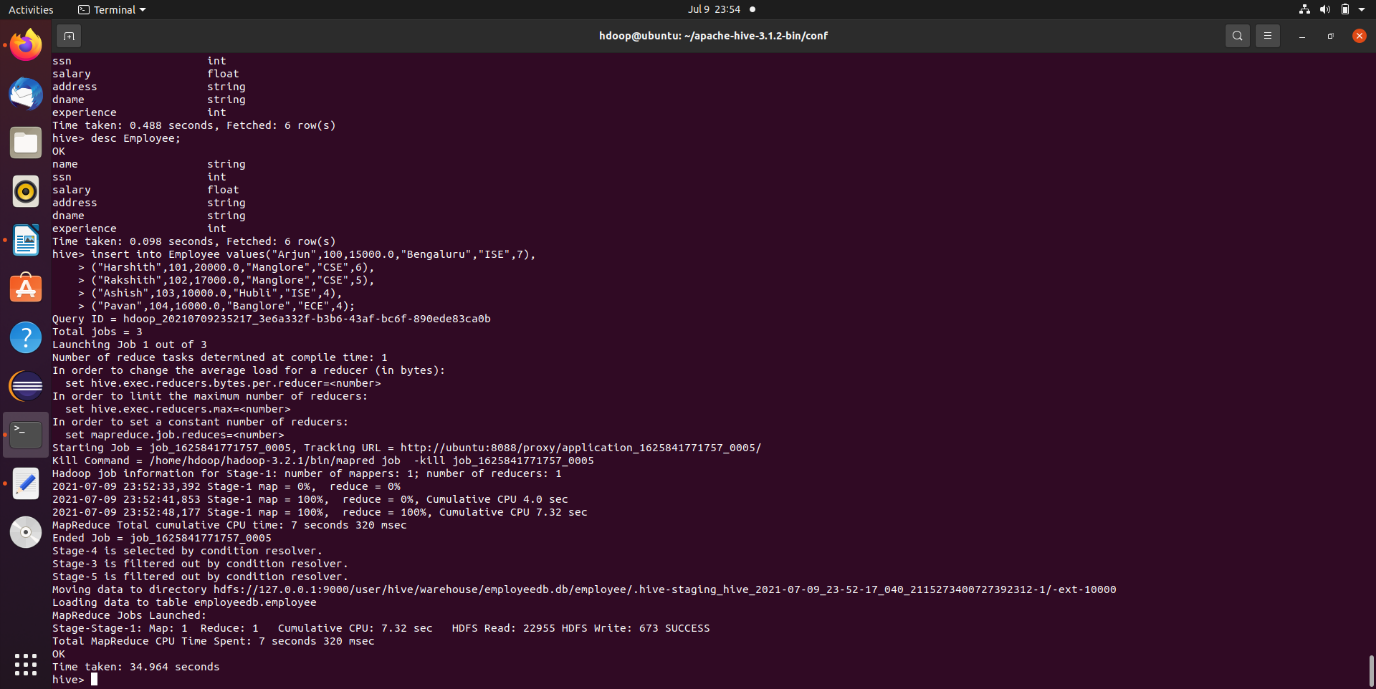
Features of Hive

* + It stores schema in a database and processed data into HDFS.
  + It is designed for OLAP.
  + It provides SQL type language for querying called HiveQL or HQL.
  + It is familiar, fast, scalable, and extensible.

**EXERCISE 2**

OUTPUT’S

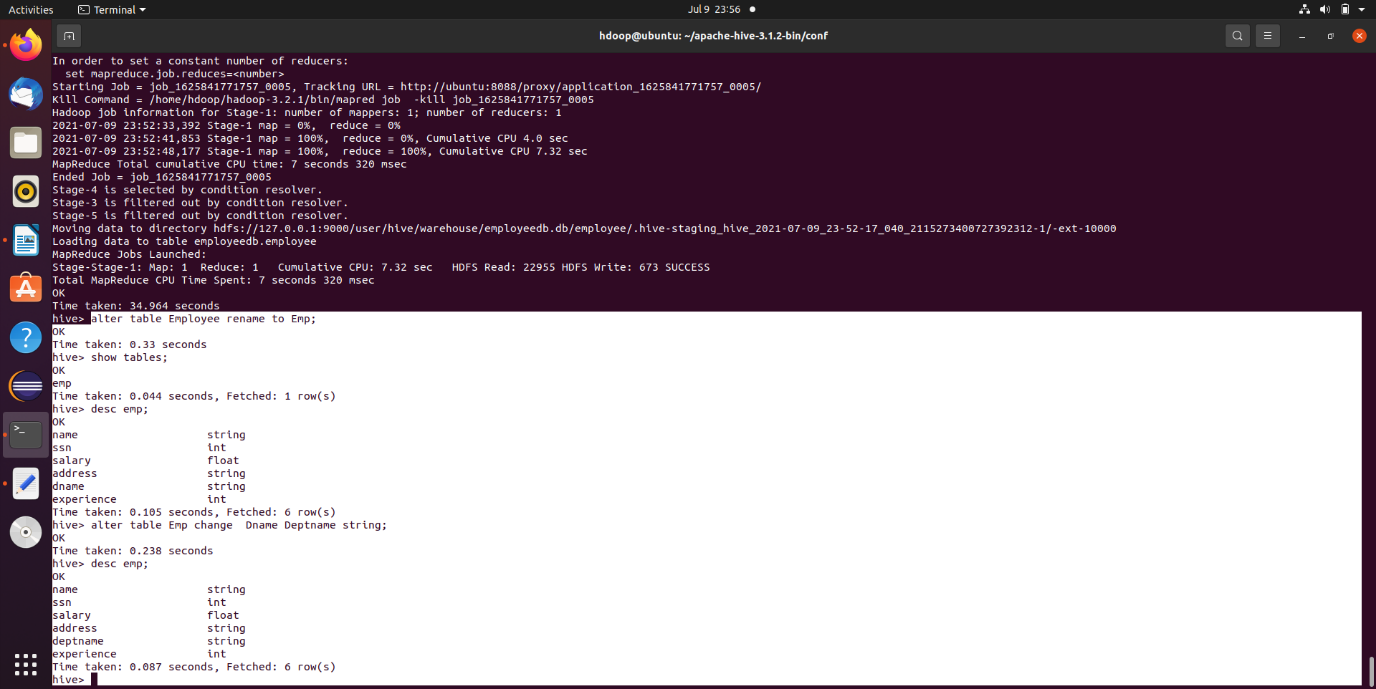
1. Insert 5 records using INSERT command.



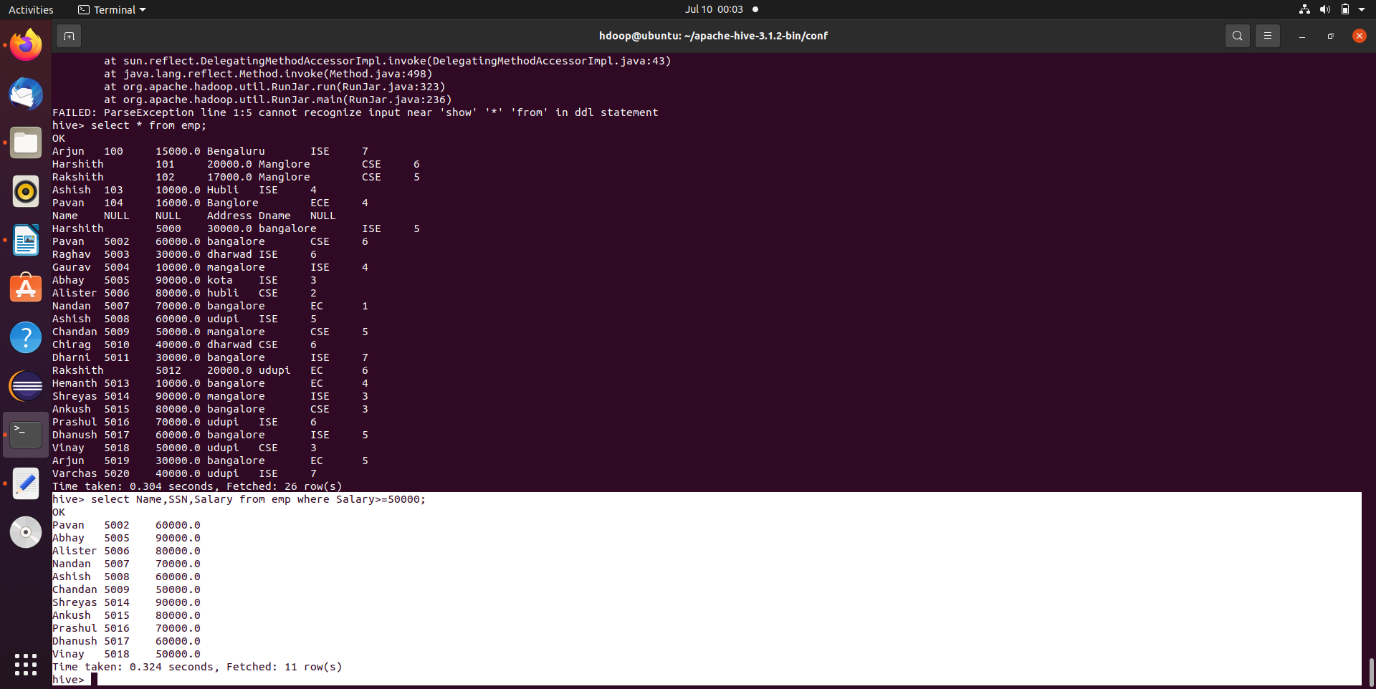
2. Demonstrate the Alter command for the following cases,

a. Rename the table name to “Emp”.

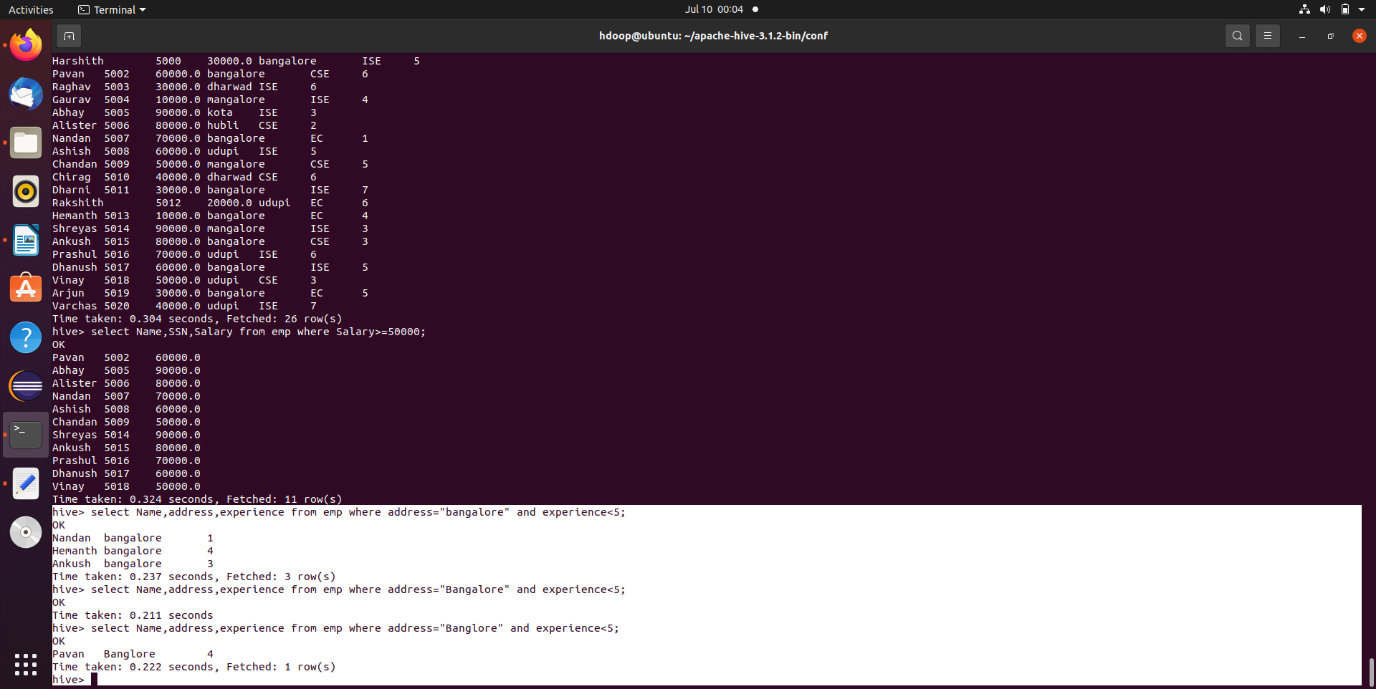
b. Rename the column name “Dname” to “Dept\_name”.



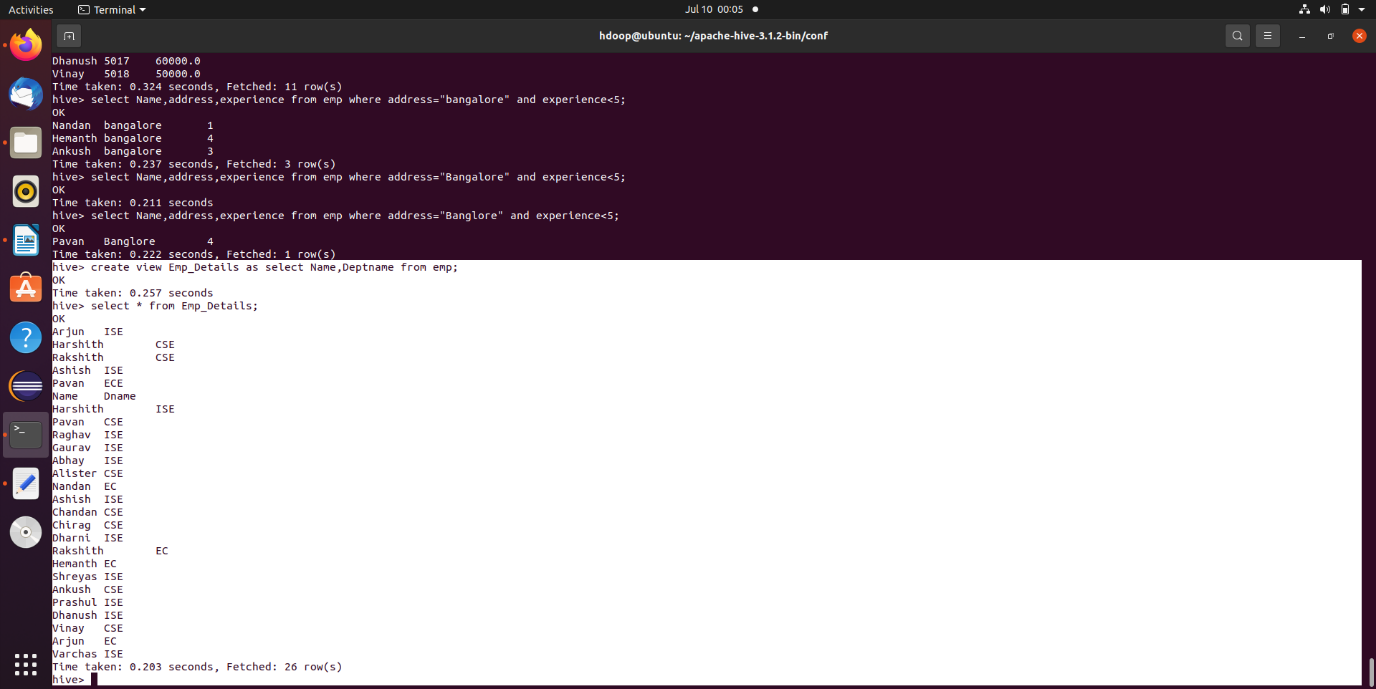
3. Retrieve all the employees who’s salary is not less than 50000.



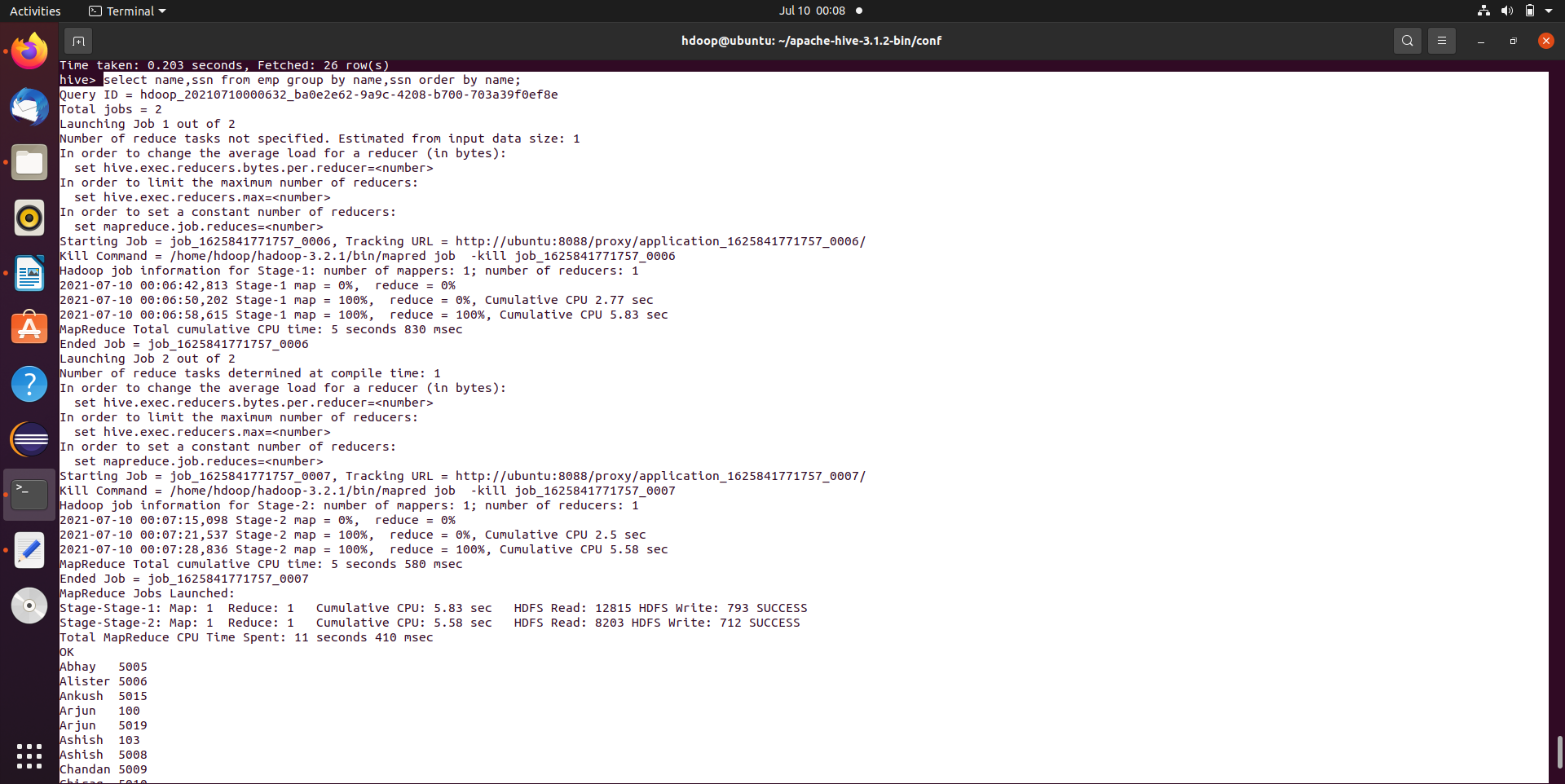
4. Extract all employees who live in Bangalore but having less than 5 years of experience

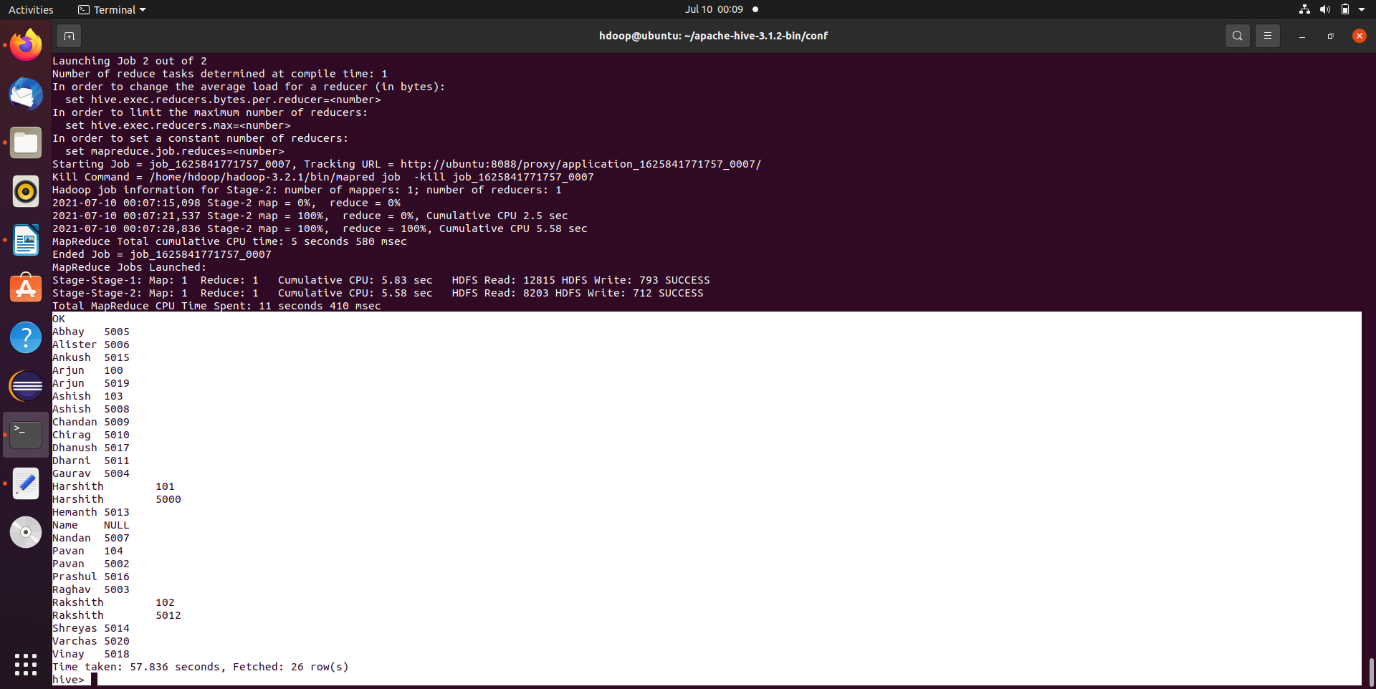


5. Create separate view containing Name, Dept\_name of employees

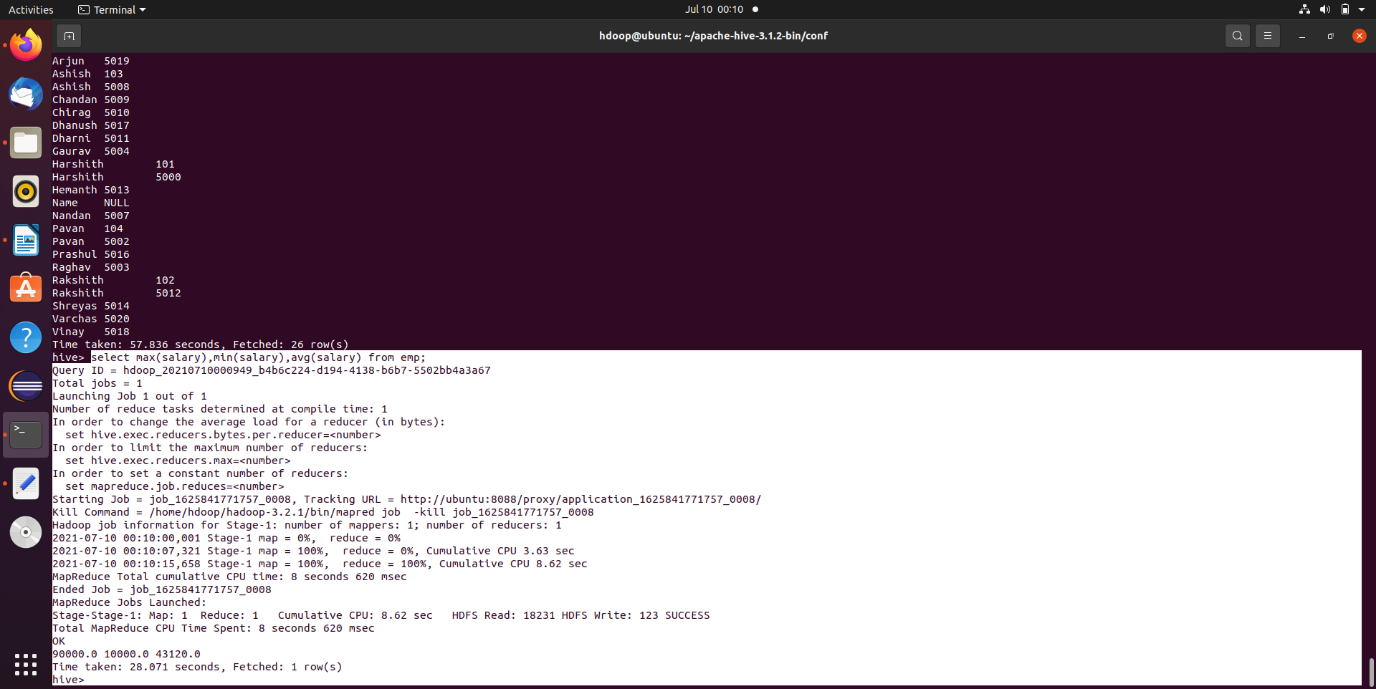


6. Display Name and SSN and use group by SSN and order by Name

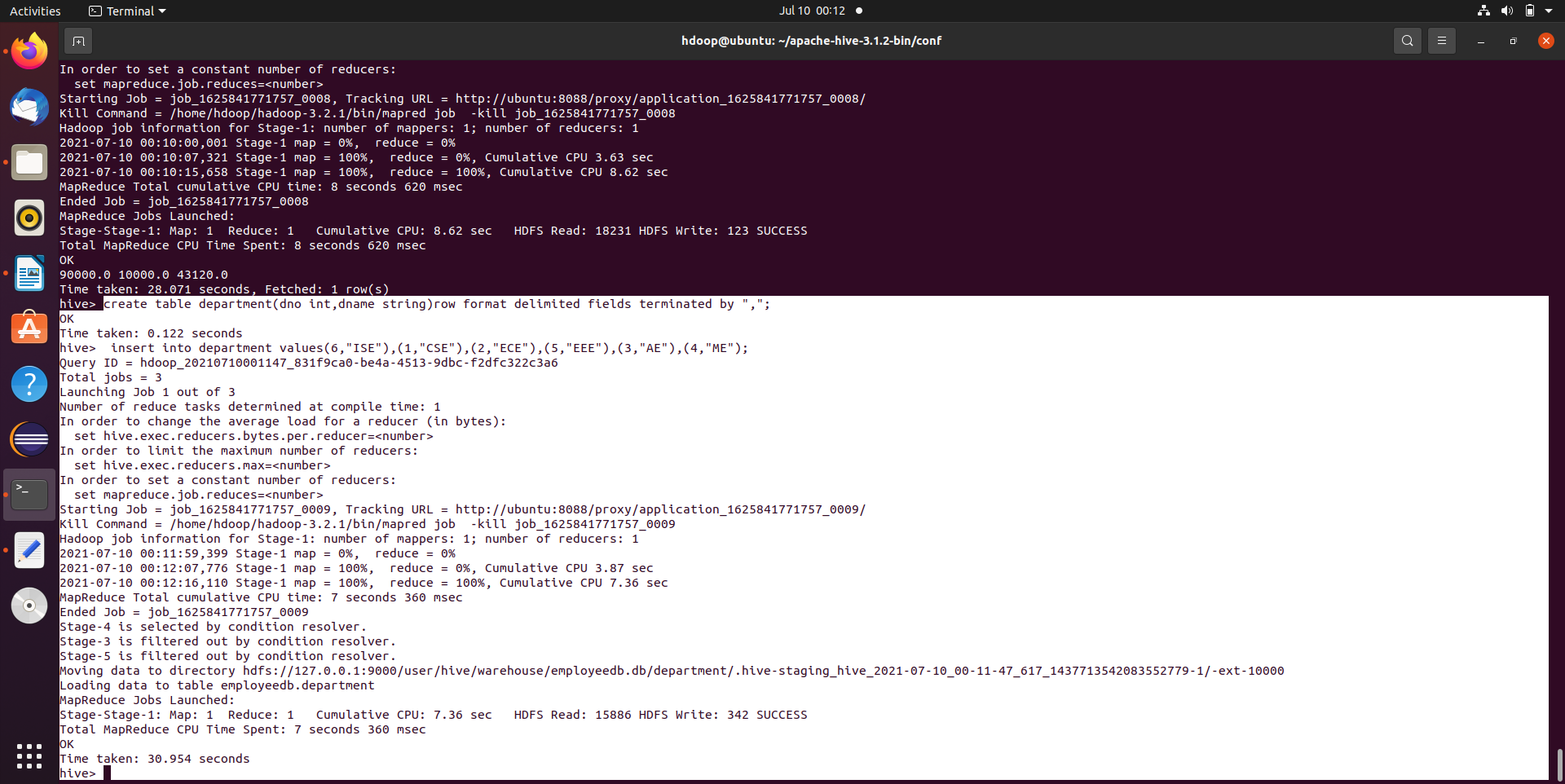




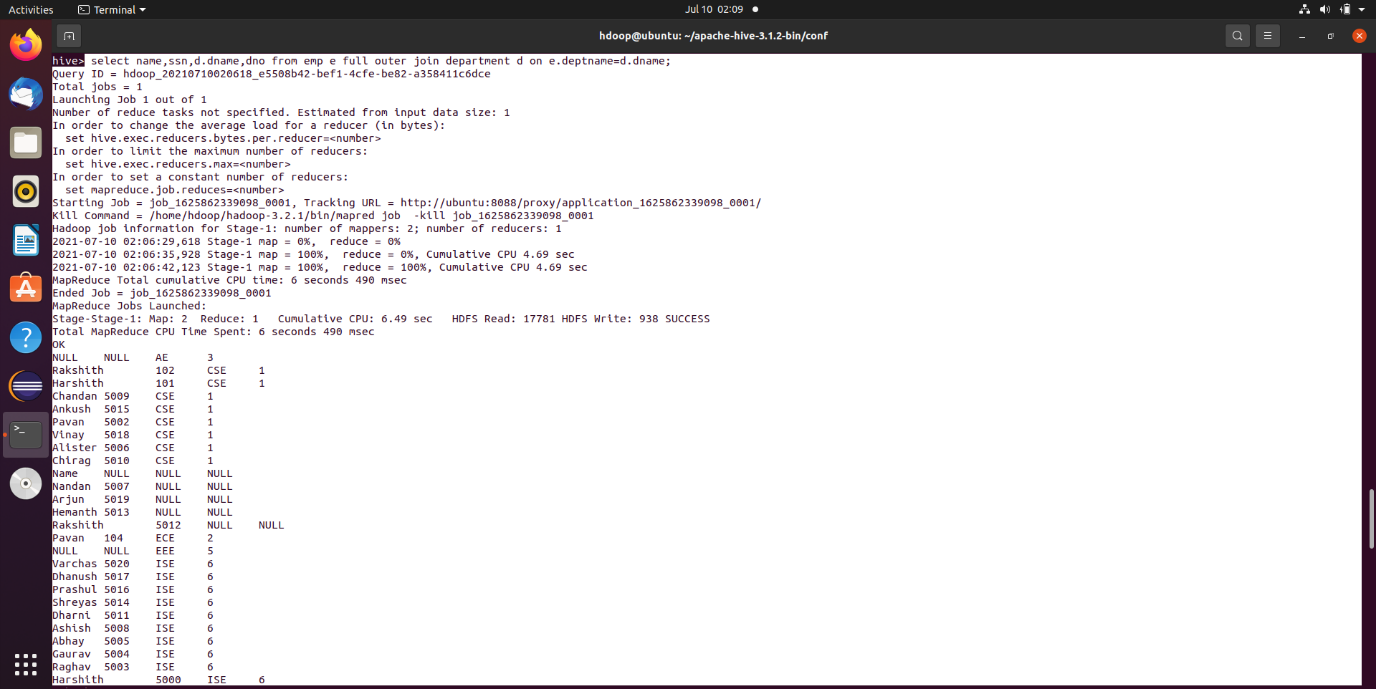
7.Retrieve Maximum salary, minimum salary and Average salary of the employees



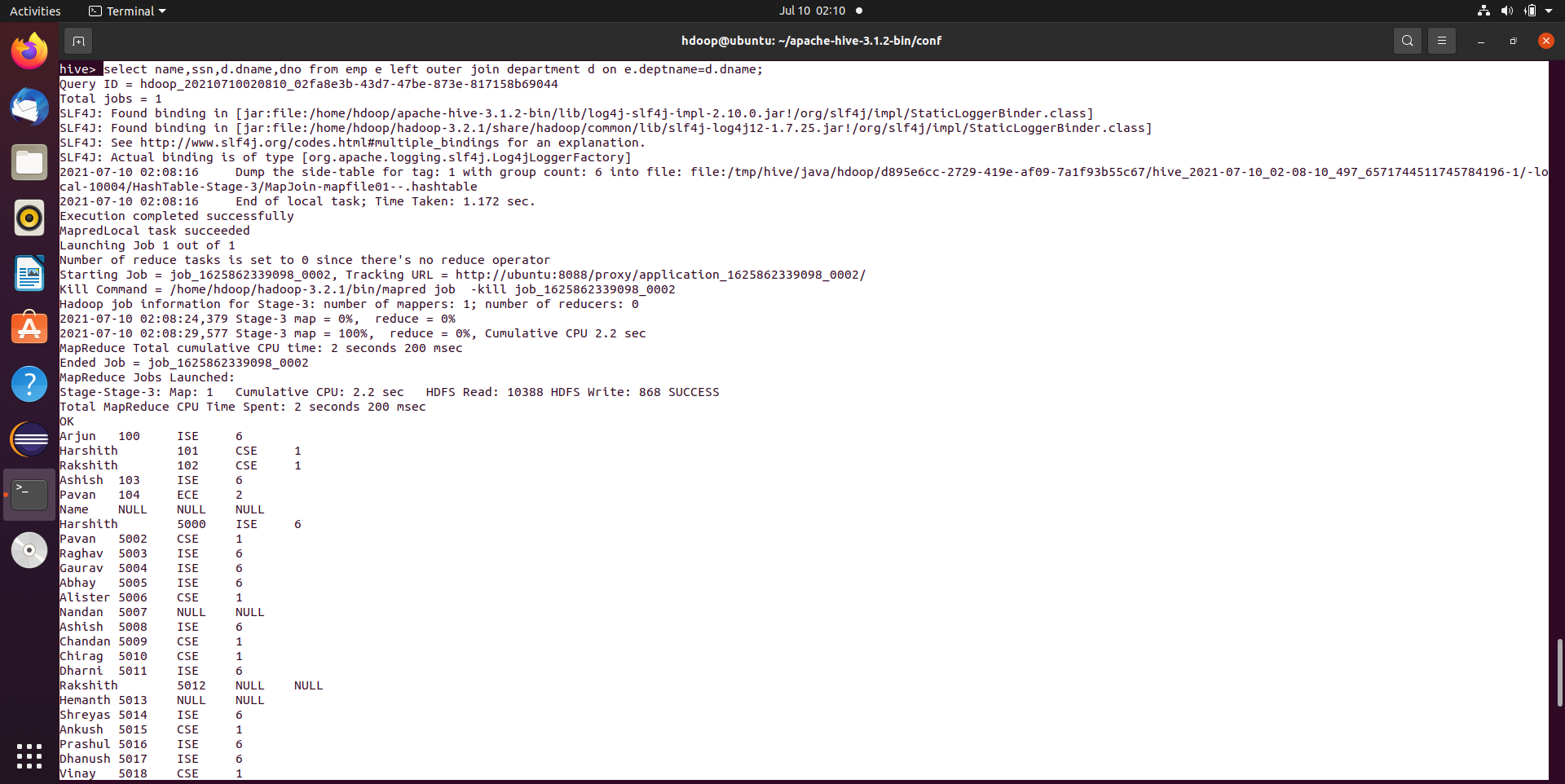
8. Create Another table called Department with the following fields (Dname = Dept\_name and perform the following joins (outer, left outer, right outer) over Dname



9. Full Outer



10. Left Outer



11.Right Outer

