# Big Data Hadoop Training

Session 16 Assignment 1 Solution:

**Dataset: Emp\_Sal**

**Dataset Details**

**Column1:** Employee\_Id

**Column2:** Employee\_Name

**Column3:** Employee\_Salary

**Column4:** Employee\_Unit (Department\_Name)

1. I have placed the dataset in HDFS @ /tmp/hive/BDDS16A1 directory.

Now, create a table employee inside a database say, BDDS16A1.

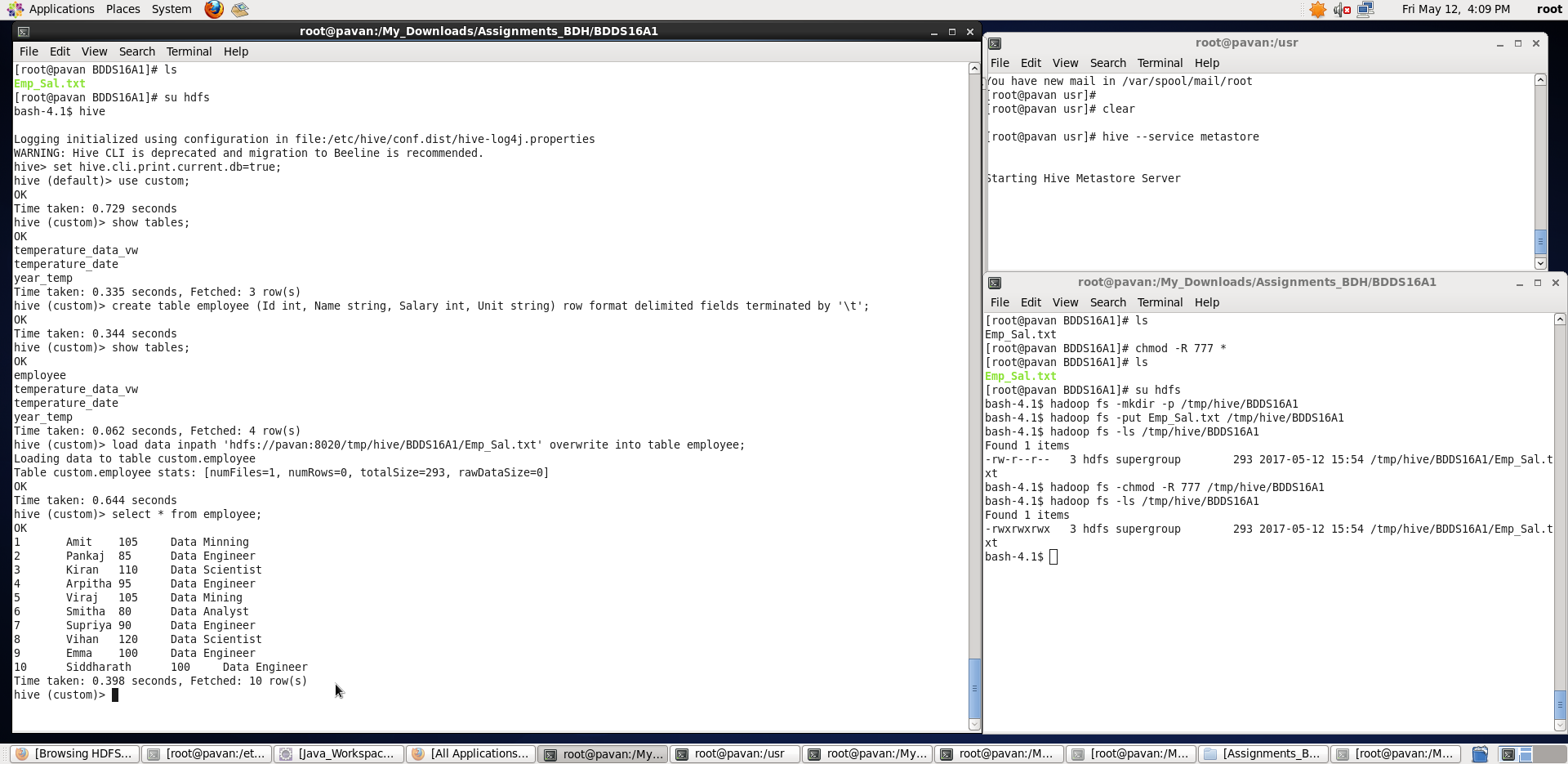
**hive> create table employee (Id int, Name string, Salary int, Unit string) row format delimited fields terminated by '\t';**

We can see a table being created. Use **show** and **describe** commands to see the table.

Now, load the dataset into the table:

**hive> load data inpath 'hdfs://pavan:8020/tmp/hive/BDDS16A1/Emp\_Sal.txt' overwrite into table employee;**

**hive> select \* from employee; -- Used to view the entire contents of the table**

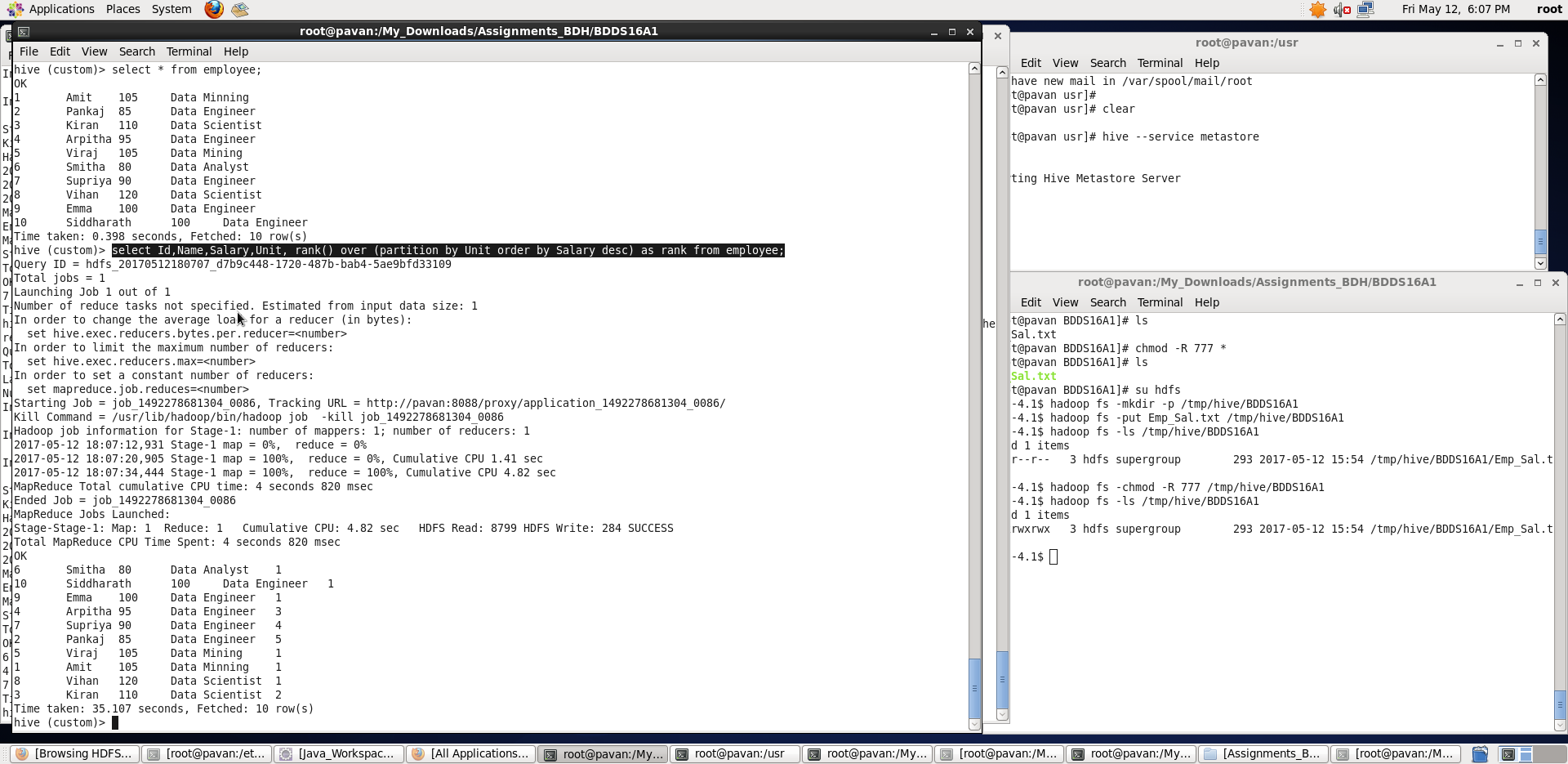


Once, we have successfully loaded the data into table, we will perform the below specified tasks

1. **Get a list of employees who receive a salary less than 100, compared to their immediate employee with higher salary in the same unit**

**A.** First, we will partition the dataset based on Unit and inside the partition we will rank the records in decreasing order of their salaries.

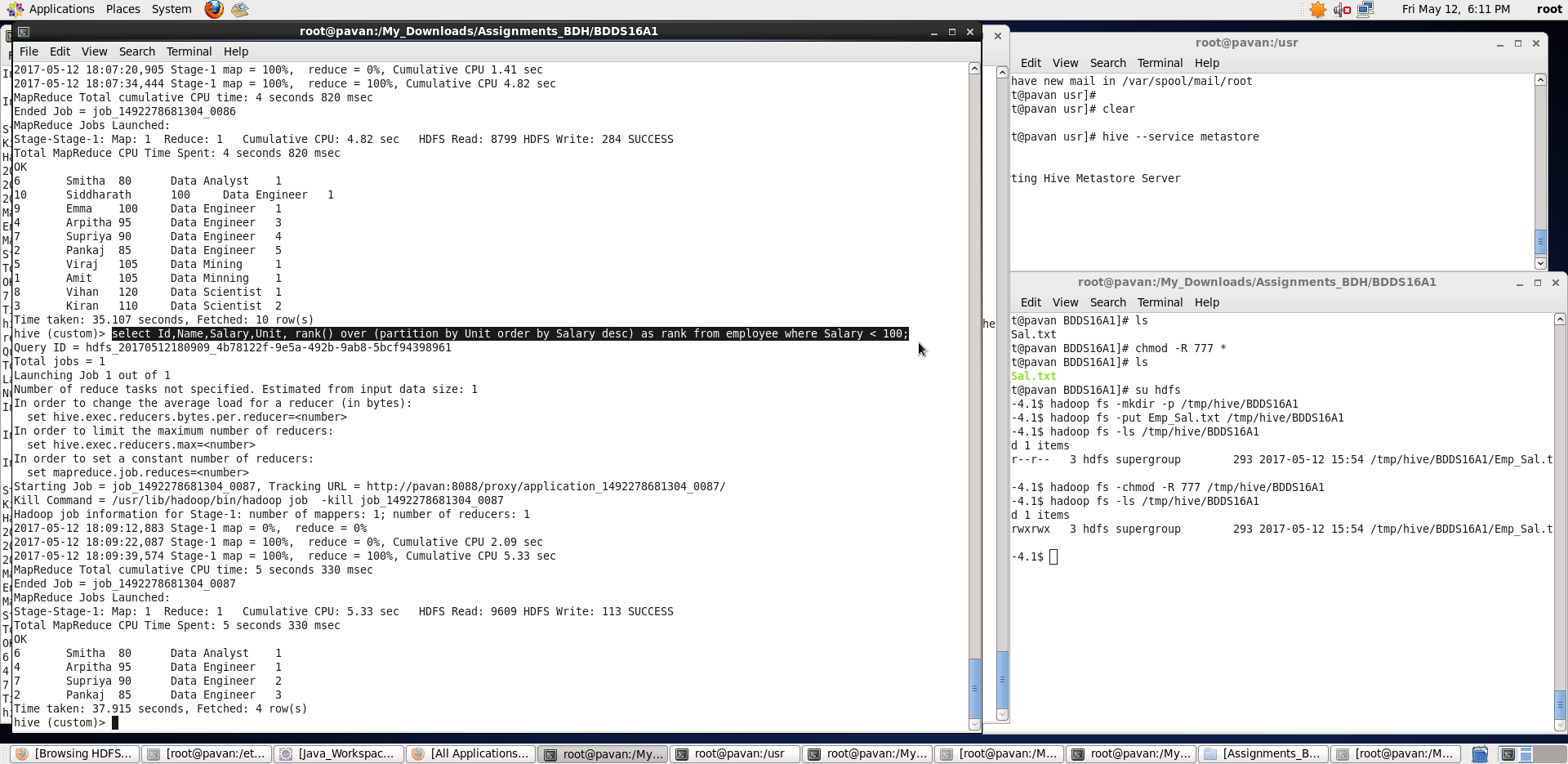
**hive> select Id,Name,Salary,Unit, rank() over (partition by Unit order by Salary desc) as rank from employee;**



We can see “rank” being displayed as the last column corresponding to each Unit.

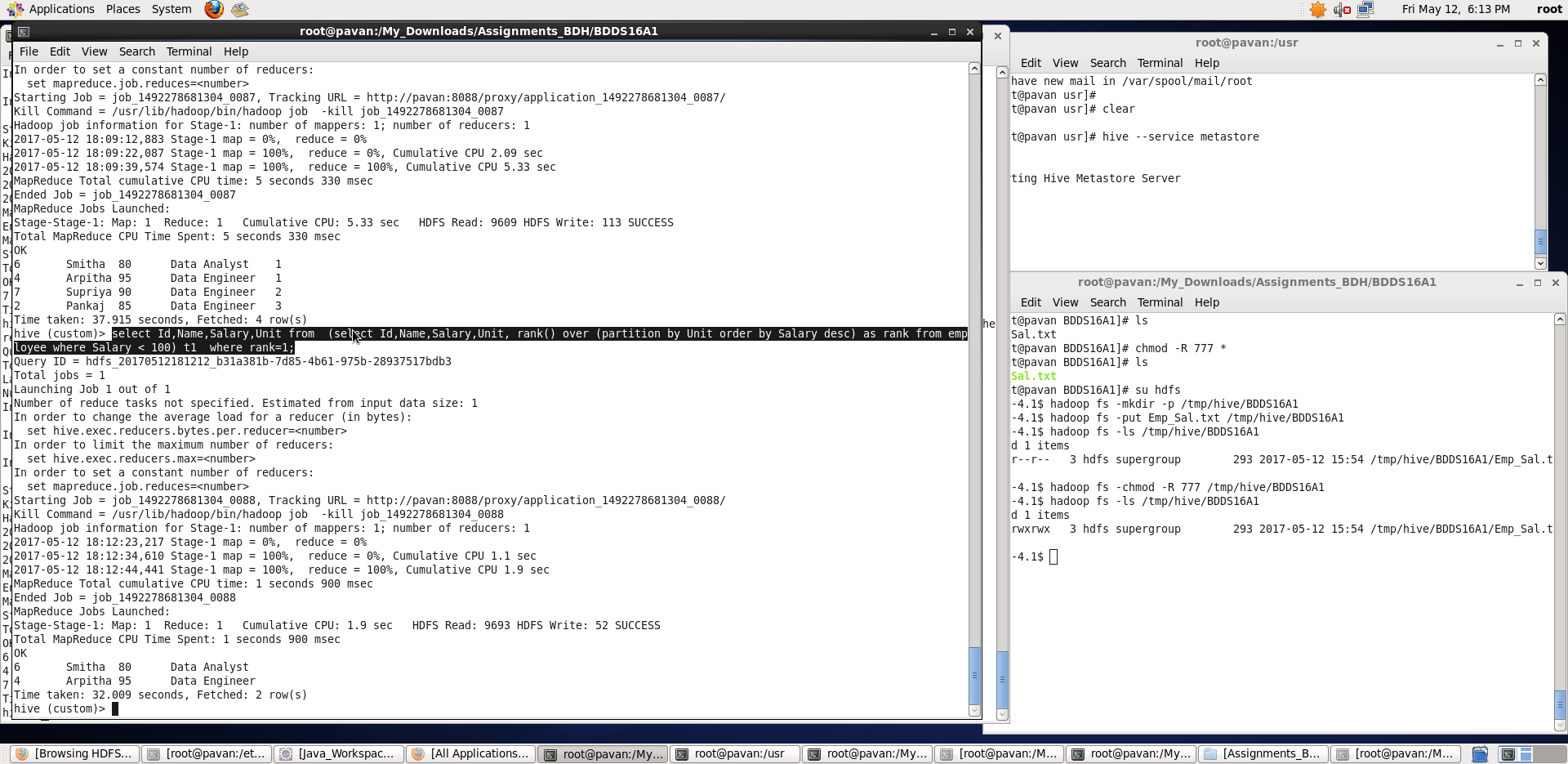
Now, we will filter out those records whose salary is less than 100.

**hive> select Id,Name,Salary,Unit, rank() over (partition by Unit order by Salary desc) as rank from employee where Salary < 100;**



Finally, listing the top record in each Unit (whose rank=1) who will be the employees receiving a salary less than 100, and also compared to their immediate employee with higher salary in the same unit.

**hive> select Id,Name,Salary,Unit from (select Id,Name,Salary,Unit, rank() over (partition by Unit order by Salary desc) as rank from employee where Salary < 100) t1 where rank=1;**

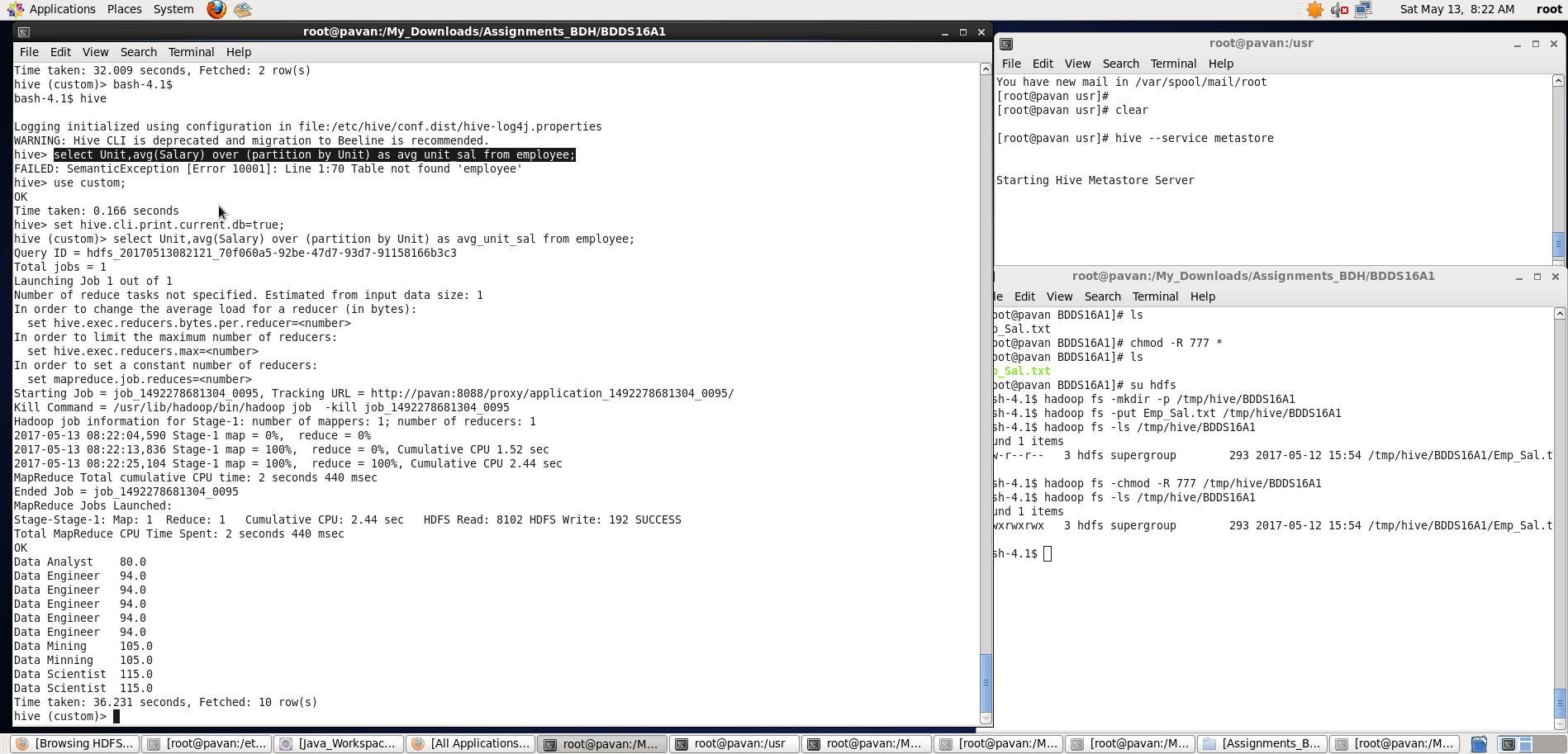


Thus, I have listed out the employees who receive a salary less than 100, compared to their immediate

employee with higher salary in the same unit.

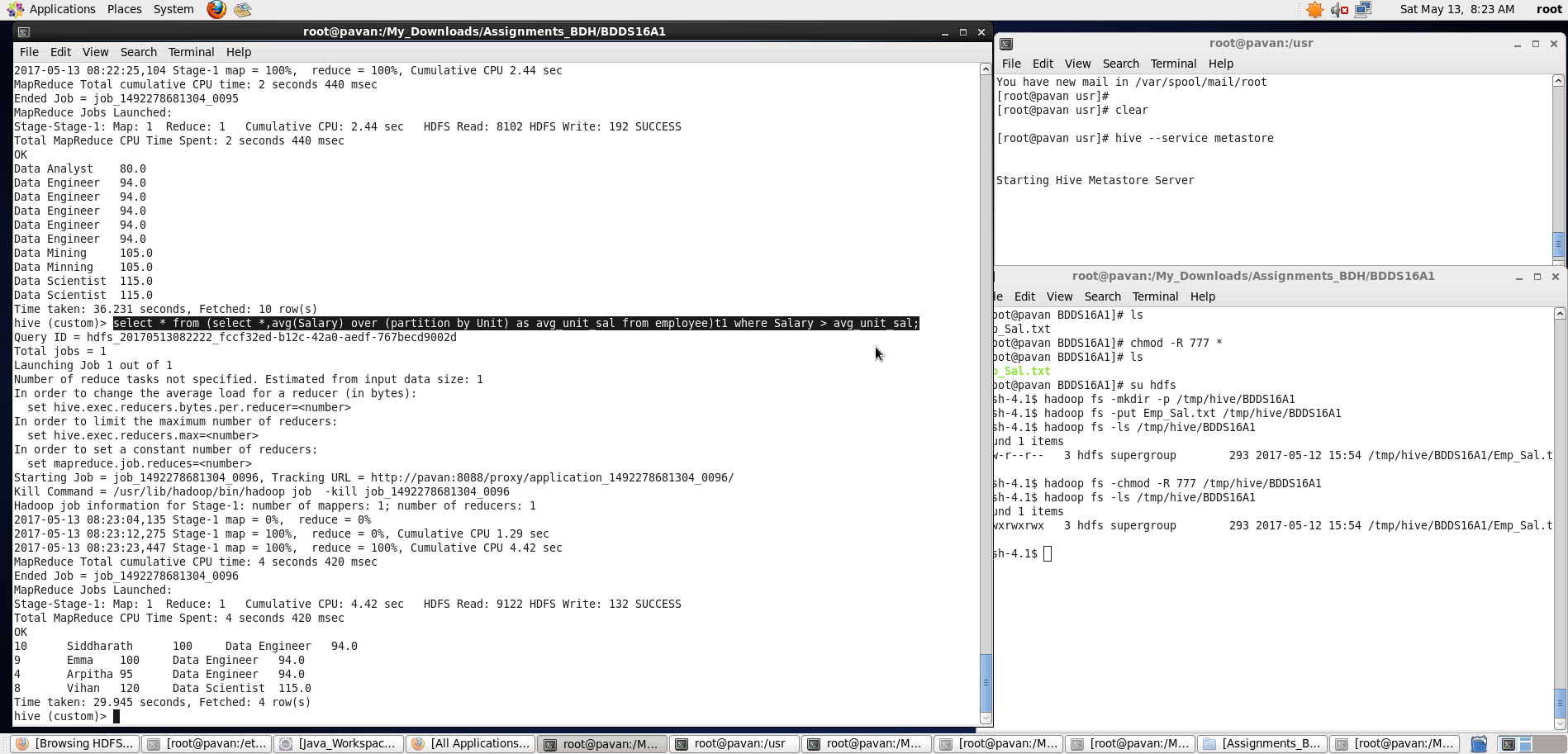
1. **List of all employees who draw higher salary than the average salary of that department**
2. Fist, we will partition the records based on their Unit and find the average salary corresponding to each Unit using **avg()** function.

**hive> select Unit,avg(Salary) over (partition by Unit) as avg\_unit\_sal from employee;**



Now, filtering those records whose salary > avg\_salary of that particular Unit is as follows:

**hive> select \* from (select \*,avg(Salary) over (partition by Unit) as avg\_unit\_sal from employee)t1 where Salary > avg\_unit\_sal;**



We can see the final output of those employees who draw higher salary than the average salary of that department. Observe that the last column shows the avg\_salary of that particular department.

Thus, we have performed required tasks on the dataset by loading into a hive table.