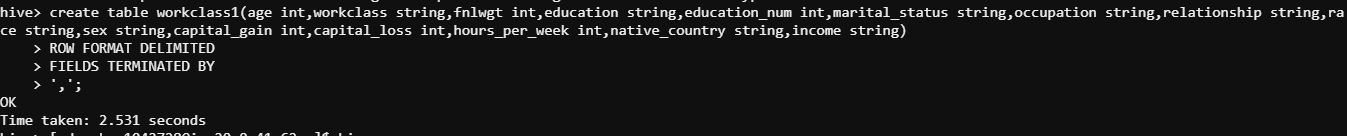
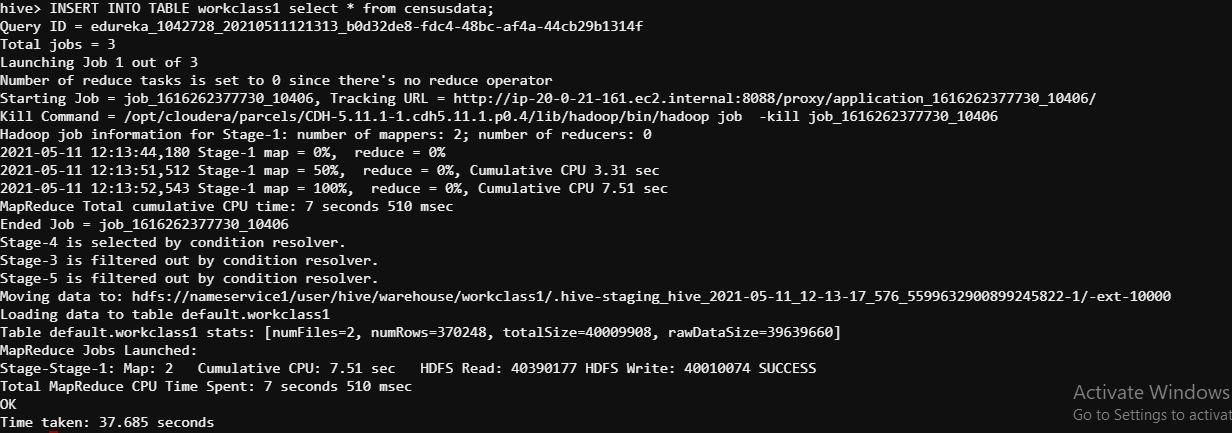
# Project part1

1. Download the dataset named censusdata.csv that is provided in your LMS
2. mid_q1.2.JPG

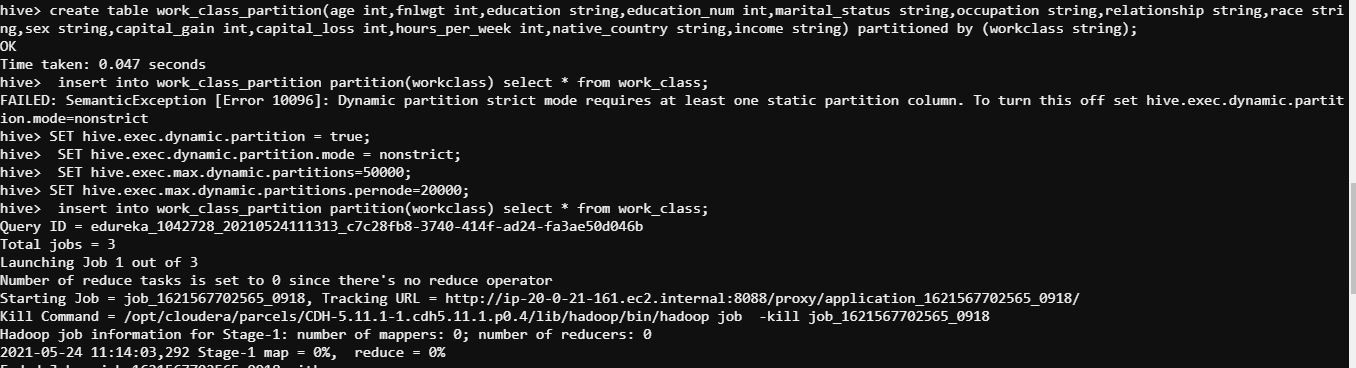
mid_q1.2.1.JPG

## Create an internal table in Hive to store the data

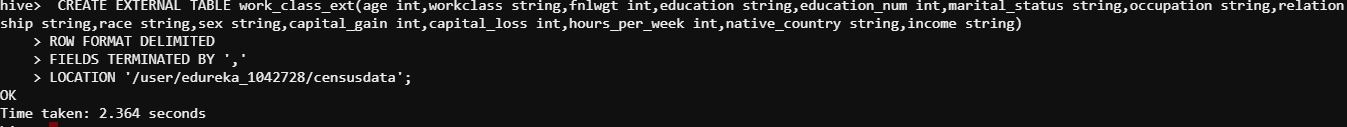


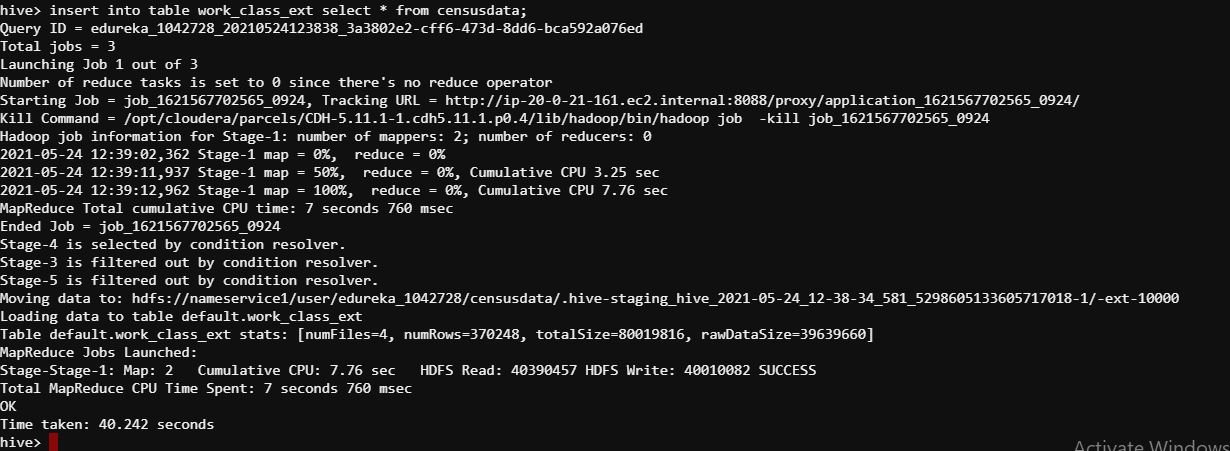


## Create an internal table in Hive with partitions

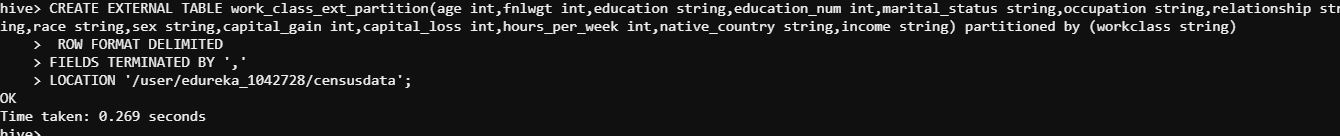


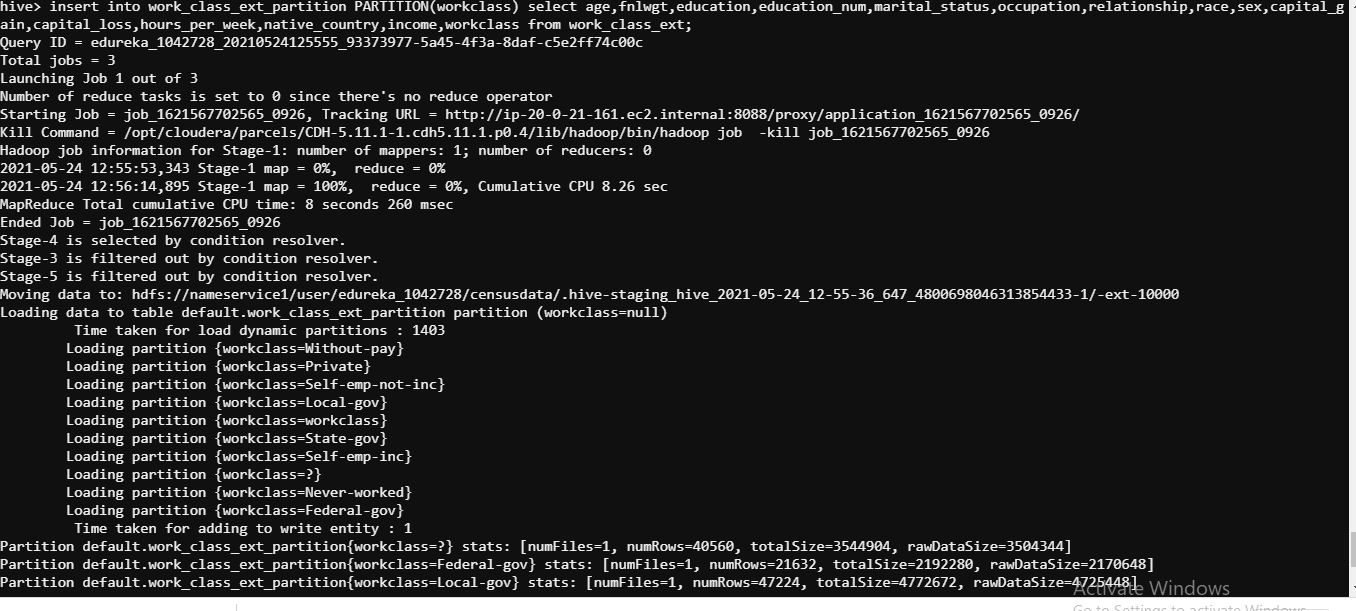
## 5. Create an external table in Hive to hold the same data stored in HDFS

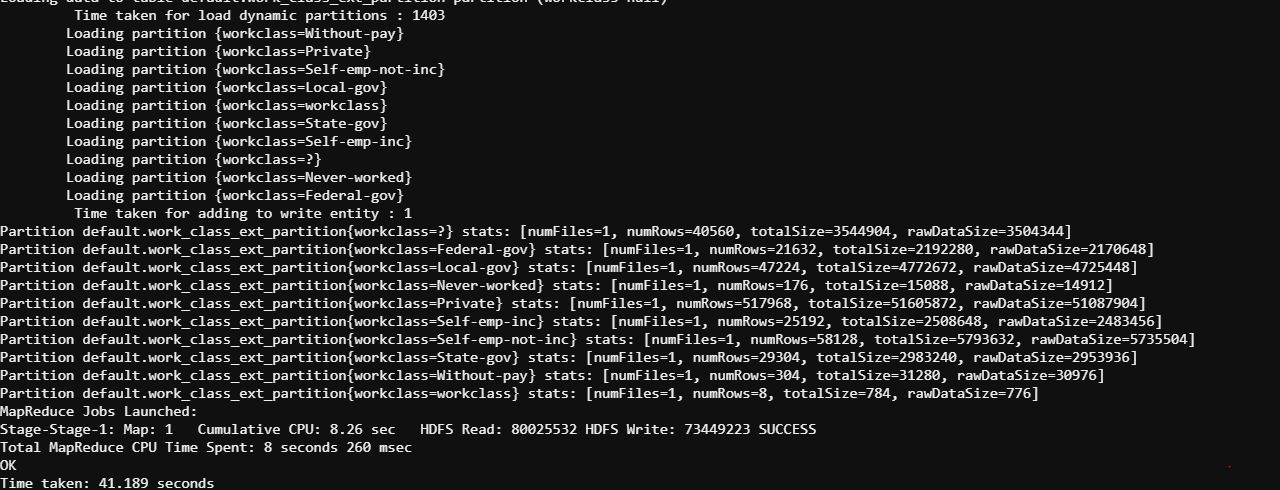




## 6. Create an external table in Hive with partitions using “workclass” as Partition Key







## 7 a. Find out the number of adults based on income and gender. Note the time taken for getting the result

1) select sex, income, count(age) from workclass1 where age>=18 group by sex, income;

2) select sex, income, count(age) from work\_class\_partition where age>=18 group by sex, income;

3 )select sex, income, count(age) from work\_class\_ext where age>=18 group by sex, income;

4) select sex, income, count(age) from work\_class\_ext\_partition where age>=18 group by sex, income;

7 b. Find out the number of adults based on income and workclass. Note the time taken for getting the result

1. select workclass, income ,count(age) from workclass1 where age>=18 group by workclass, income;

2) select workclass, income ,count(age) from work\_class\_partition where age>=18 group by workclass,

income;

3) select workclass, income ,count(age) from work\_class\_ext where age>=18 group by workclass, income;

4) select workclass, income ,count(age) from work\_class\_ext\_partition where age>=18 group by workclass, income;

# Write your observations by comparing the time taken for executing the commands between

a.Internal & External Tables

b.Partitioned & Non-partitioned Tables

|  |  |  |
| --- | --- | --- |
| Internal | Income and gender | Income and workclass |
| 29.64 | 31.89 |
| External | 24.66 | 25.83 |

|  |  |  |
| --- | --- | --- |
| Partitioned | Income and gender | Income and workclass |
| 22.71 | 25.36 |
| Non-partitioned | 30.81 | 26.54 |

**8)** In internal table If we drop the managed table or partition, the table data and the metadata associated with that table will be deleted from the HDFS and if we drop the external table, then only the metadata associated with the table will get deleted, the table data remains untouched by Hive.