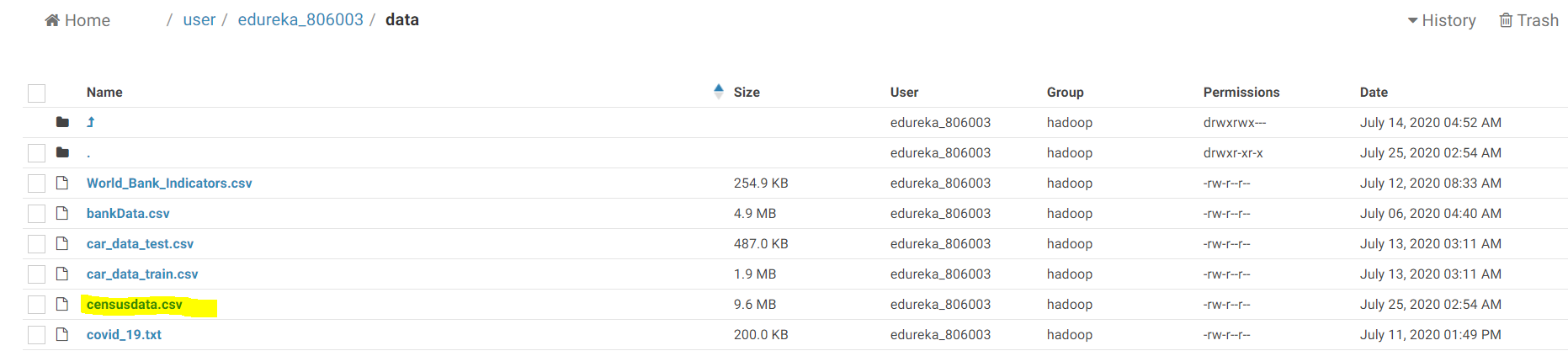
1. **Load Data into HDFS**

hdfs dfs -copyFromLocal /C:/Users/AYAN/Desktop/DATA SCIENCE PG/PROJECTS/INCOME CLASSIFIER/censusdata/censusdata.csv /user/edureka\_806003/data



1. **Create an internal table in Hive to store the data**

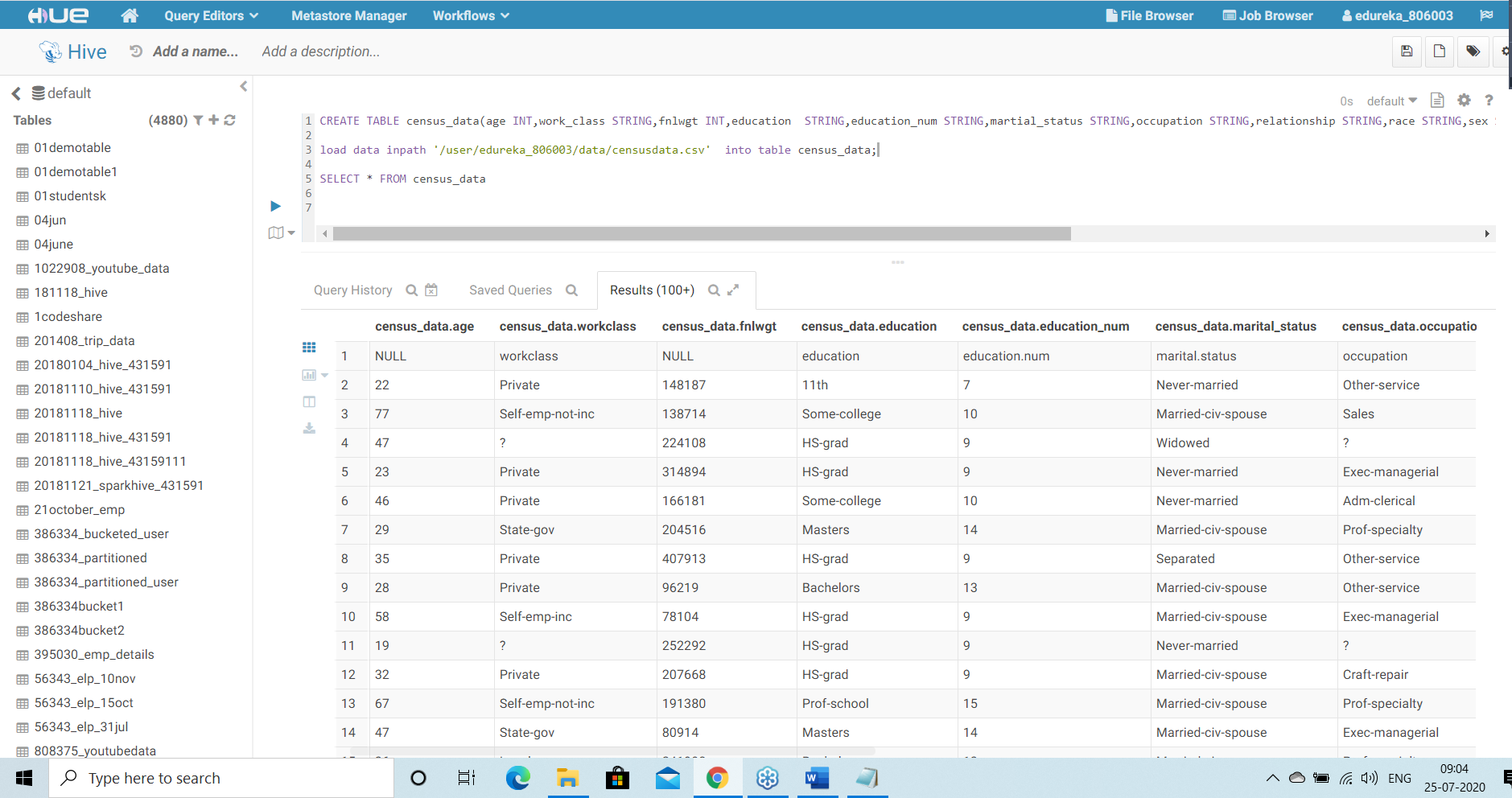
**a. Create the table structure**

**b. Load the data from HDFS into the Hive table**

CREATE TABLE census\_data(age INT,work\_class STRING,fnlwgt INT,education STRING,education\_num STRING,martial\_status STRING,occupation STRING,relationship STRING,race STRING,sex STRING,capital\_gain INT,capital\_loss INT, hours\_per\_week INT,native\_country STRING,income STRING);

load data inpath '/user/edureka\_806003/data/censusdata.csv' into table census\_data;

SELECT \* FROM census\_data



1. **Create an internal table in Hive with partitions**

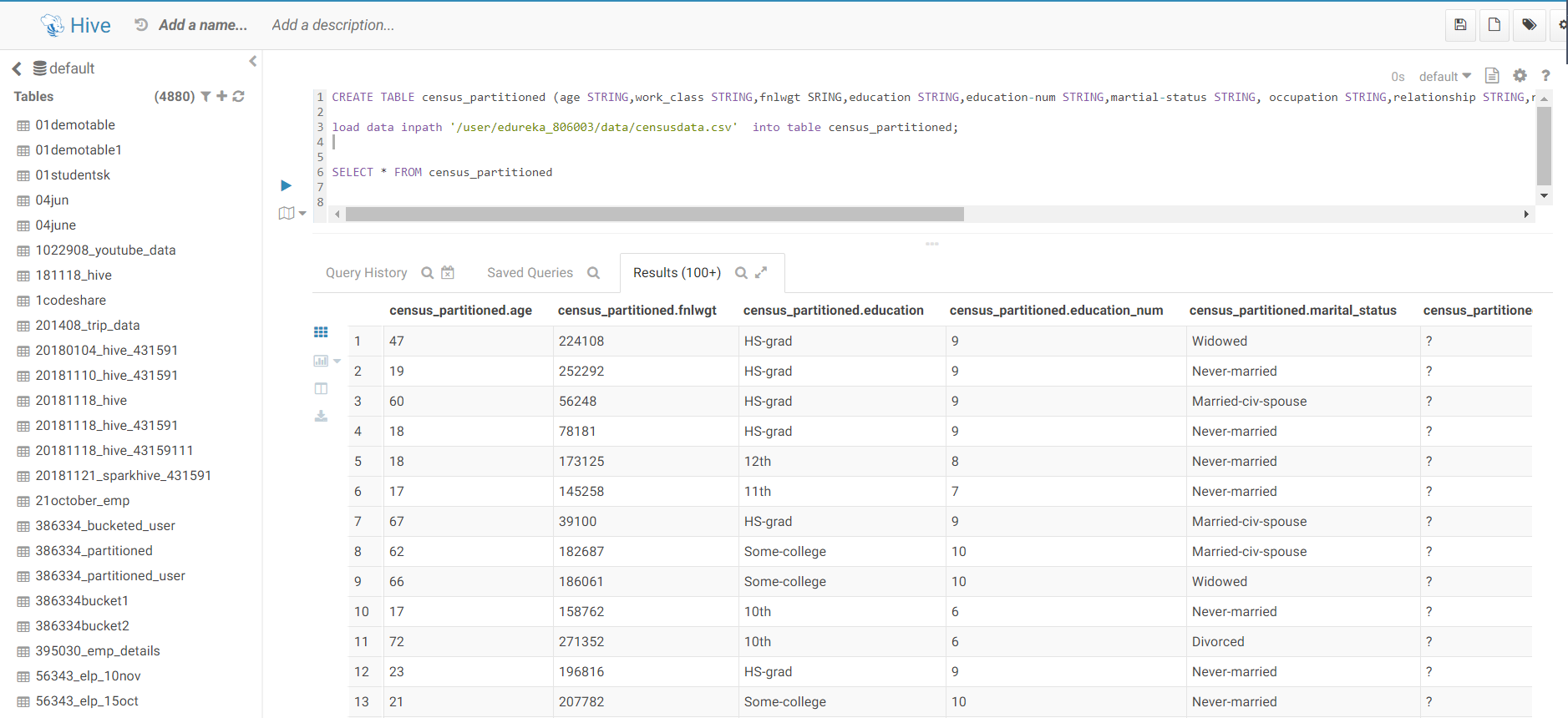
**a. Create a Partition Table in Hive using “workclass” as the Partition Key**

**b. Load data from the staging table (Table created in Step 3) into this table**

CREATE TABLE census\_partitioned (age INT,work\_class STRING,fnlwgt INT,education STRING,education\_num STRING,martial\_status STRING,occupation STRING,relationship STRING,race STRING,sex STRING,capital\_gain INT,capital\_loss INT, hours\_per\_week INT,native\_country STRING,income STRING) PARTITIONED BY(work\_class STRING);

load data inpath '/user/edureka\_806003/data/censusdata.csv' into table census\_partitioned;

SELECT \* FROM census\_partitioned



1. **Create an external table in Hive to hold the same data stored in HDFS**

CREATE EXTERNAL TABLE census\_data\_external(age INT,work\_class STRING,fnlwgt INT,education STRING,education\_num STRING,martial\_status STRING,occupation STRING,relationship STRING,race STRING,sex STRING,capital\_gain INT,capital\_loss INT, hours\_per\_week INT,native\_country STRING,income STRING);

load data inpath '/user/edureka\_806003/data/censusdata.csv' into table census\_data\_external;



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

CREATE EXTERNAL TABLE census\_data\_partition\_external(age INT,work\_class STRING,fnlwgt INT,education STRING,education\_num STRING,martial\_status STRING,occupation STRING,relationship STRING,race STRING,sex STRING,capital\_gain INT,capital\_loss INT, hours\_per\_week INT,native\_country STRING,income STRING) PARTITIONED BY(work\_class\_partition STRING);

load data inpath '/user/edureka\_806003/data/censusdata.csv' into table census\_data\_partition\_external;



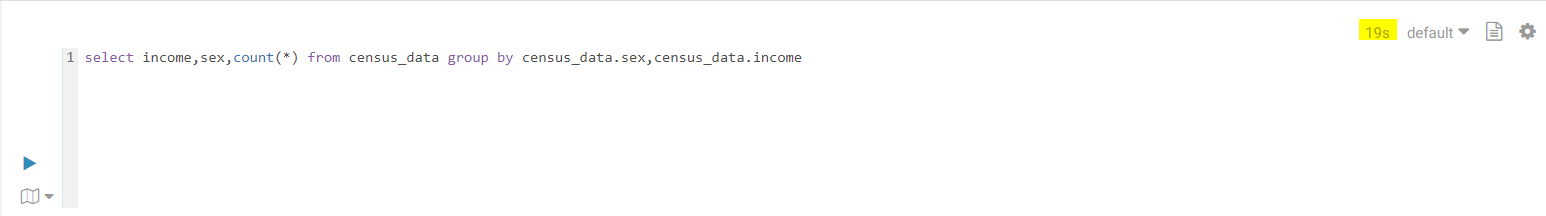
1. **For each of the four tables created above, perform the following operations**

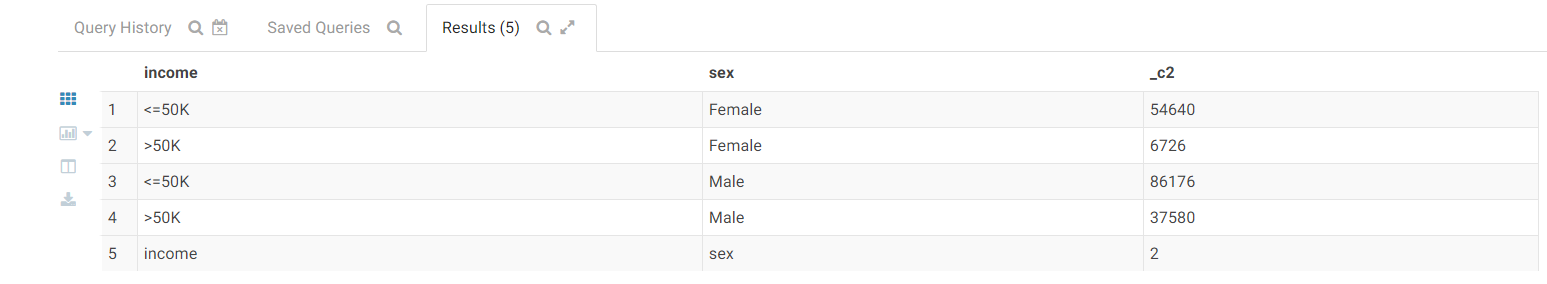
**• Find out the number of adults based on income and gender. Note the time taken for getting the result**

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

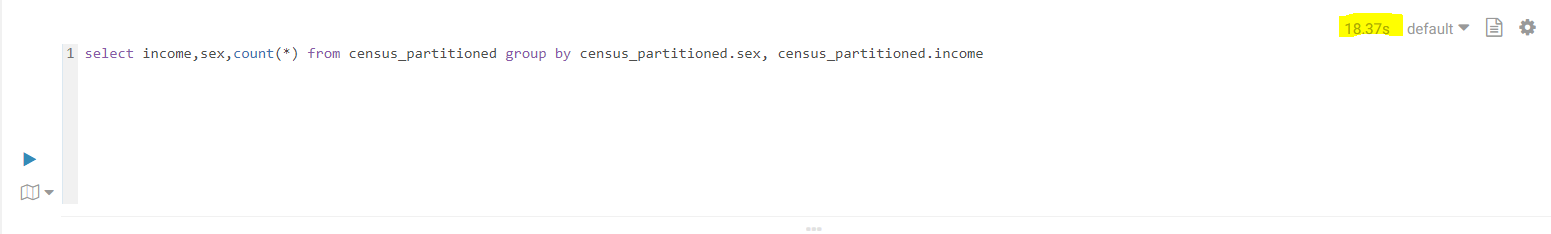
**• Write your observations by comparing the time taken for executing the commands between: a. Internal & External Tables b. Partitioned & Non-partitioned Tables**

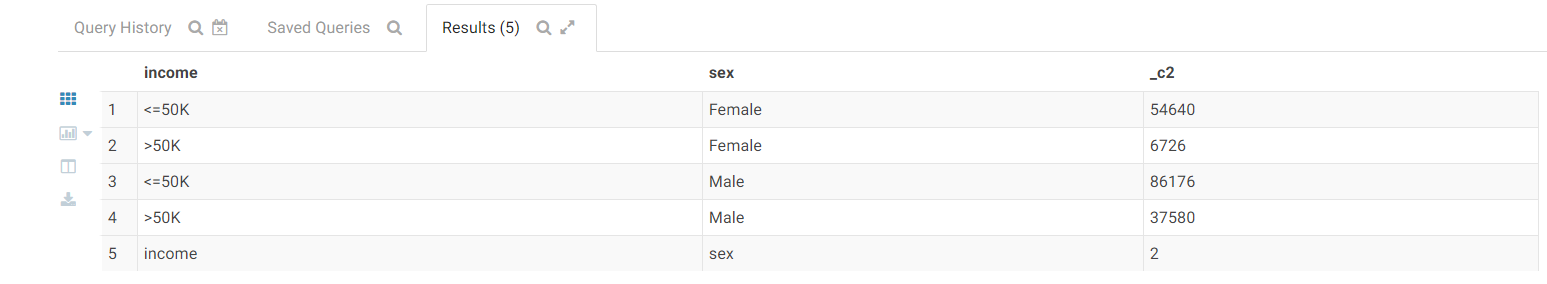
select income,sex,count(\*) from census\_data group by census\_data.sex,census\_data.income



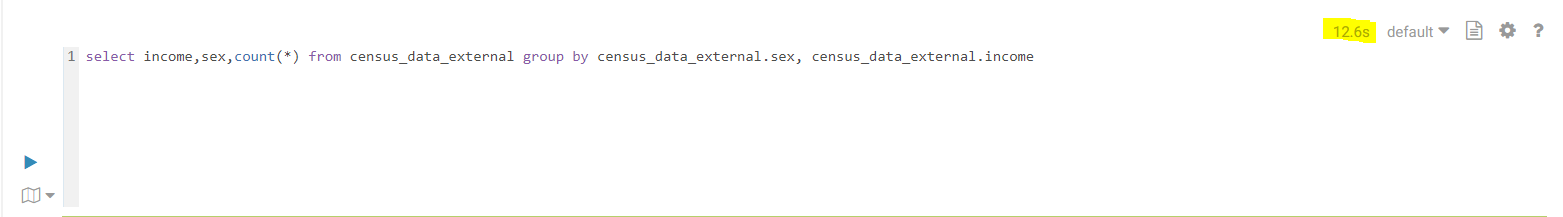


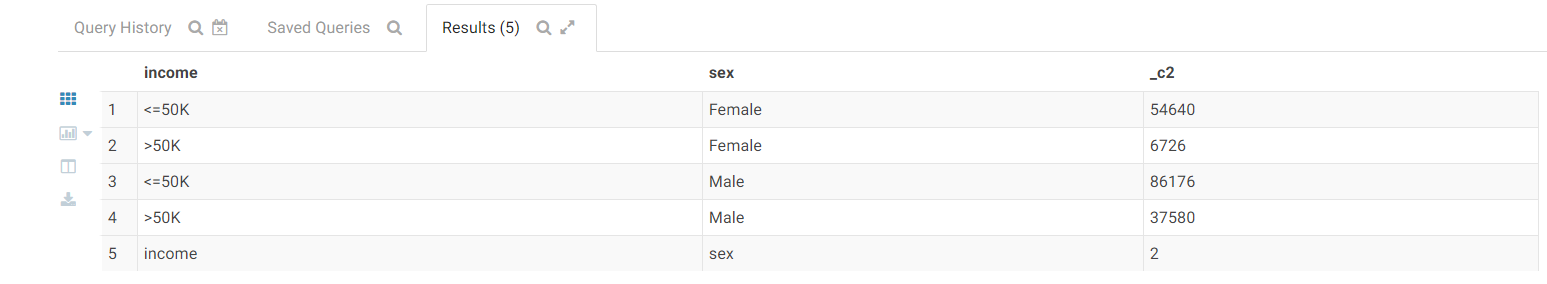
select income,sex,count(\*) from census\_partitioned group by census\_partitioned.sex, census\_partitioned.income



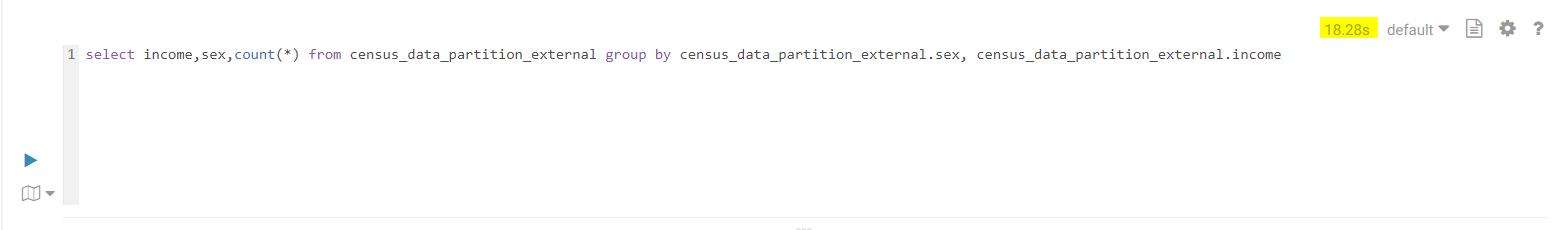


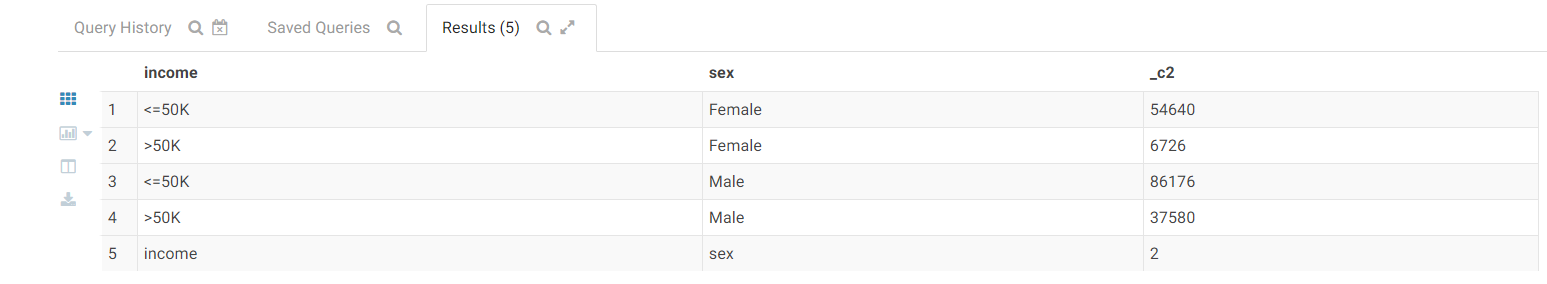
select income,sex,count(\*) from census\_data\_external group by census\_data\_external.sex, census\_data\_external.income





select income,sex,count(\*) from census\_data\_partition\_external group by census\_data\_partition\_external.sex, census\_data\_partition\_external.income





**Time taken:**

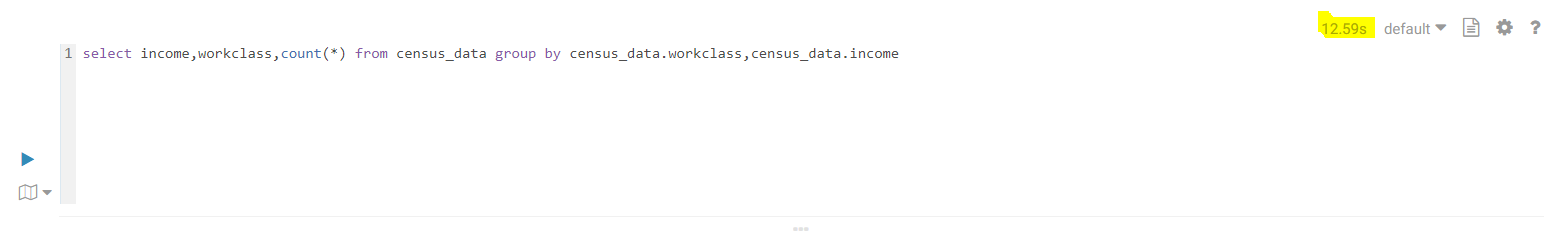
**Internal Non-Partitioned:- 19 sec**

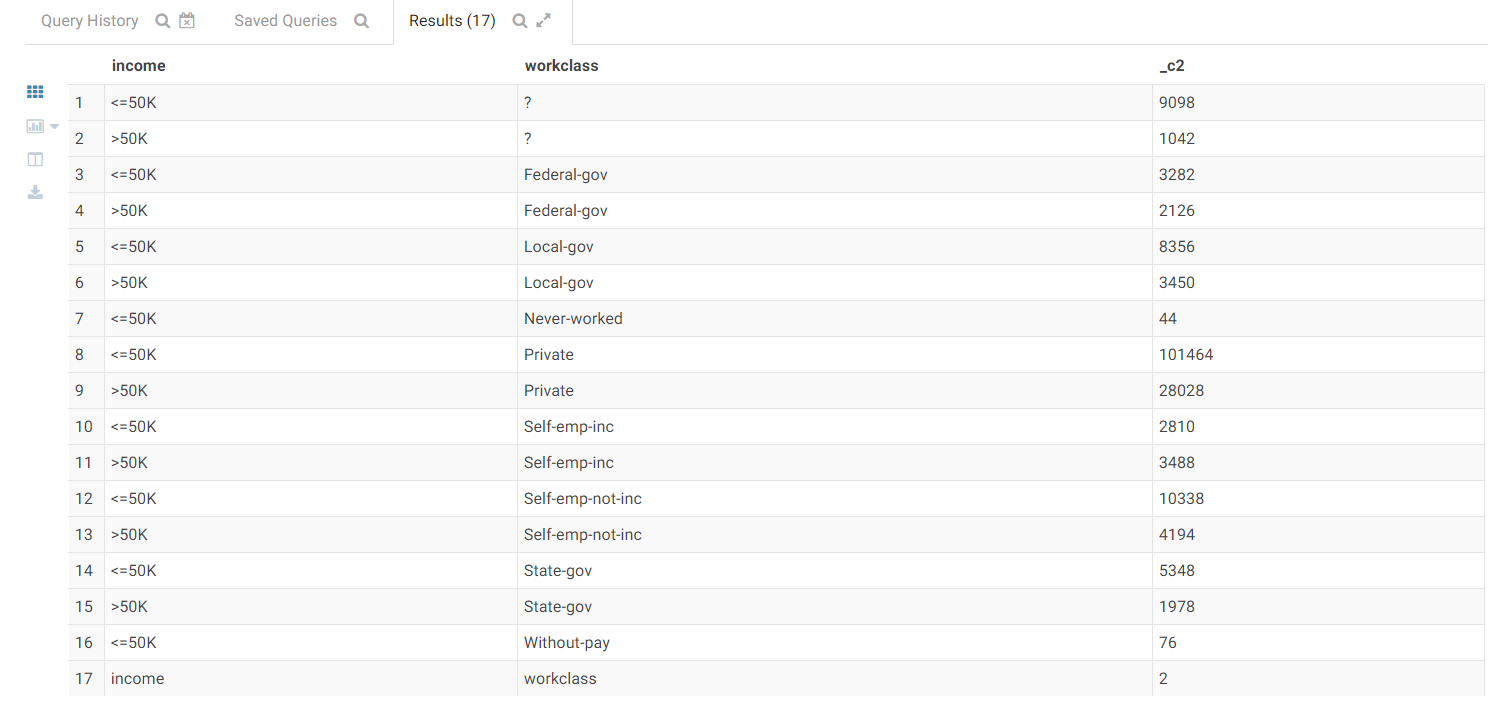
**Internal Partitioned:- 18.37 sec**

**External Non-Partitioned:- 12.6 sec**

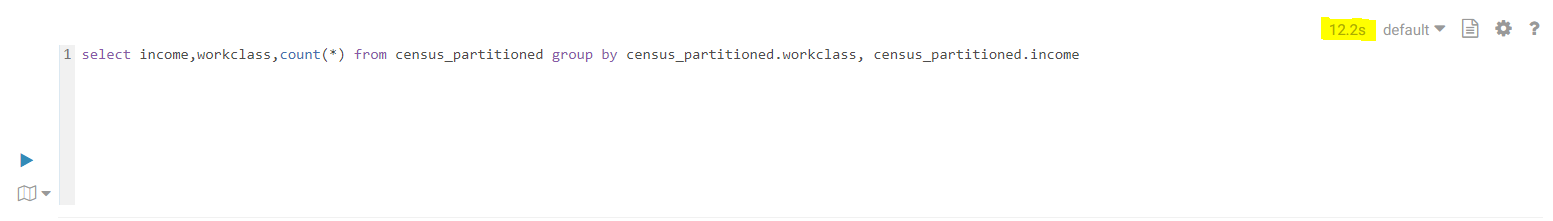
**External Partitioned:- 18.28 sec**

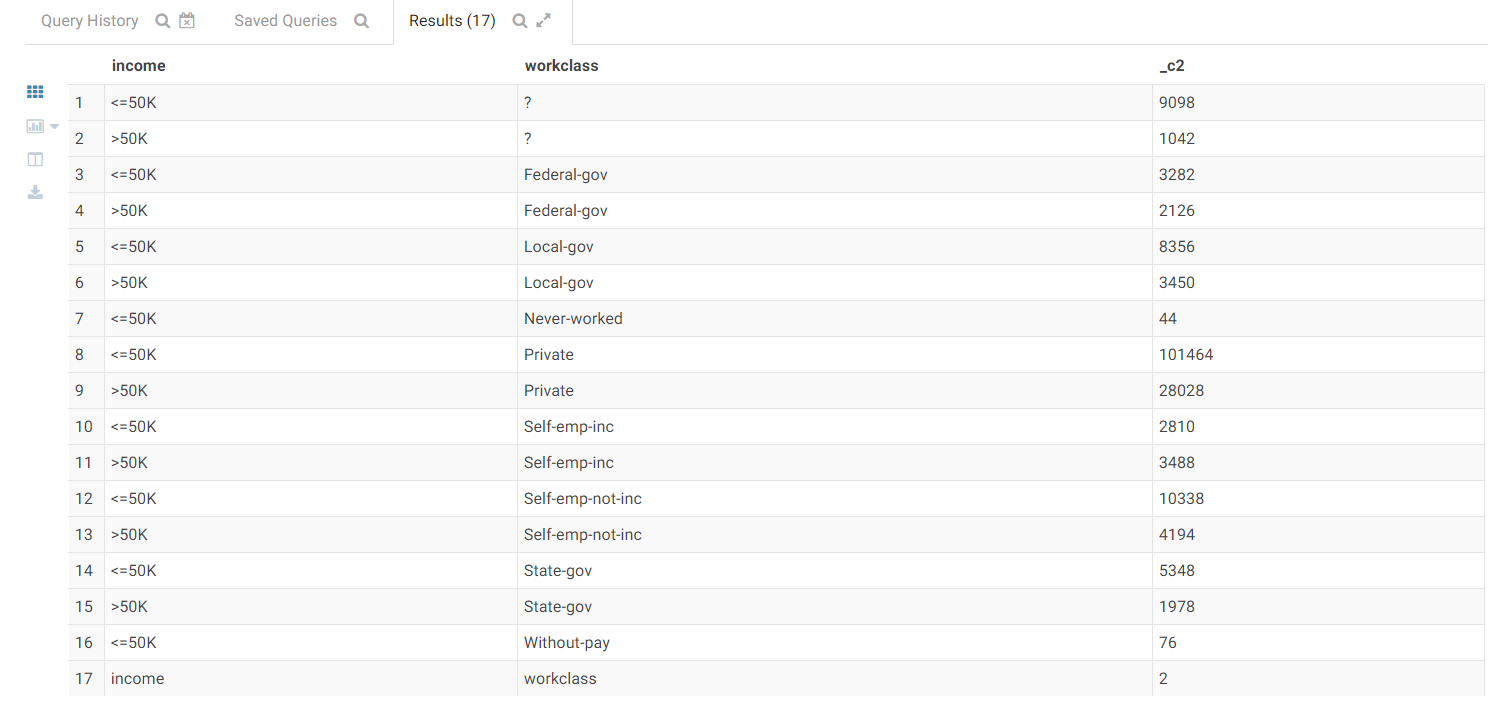
select income,workclass,count(\*) from census\_data group by census\_data.workclass,census\_data.income



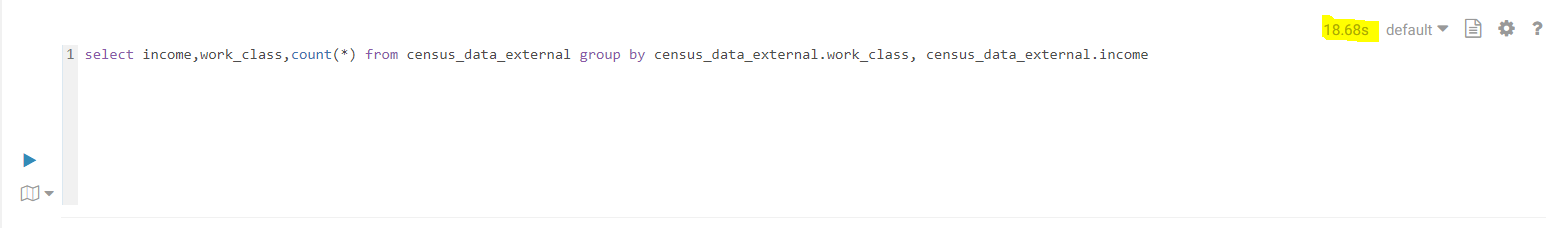


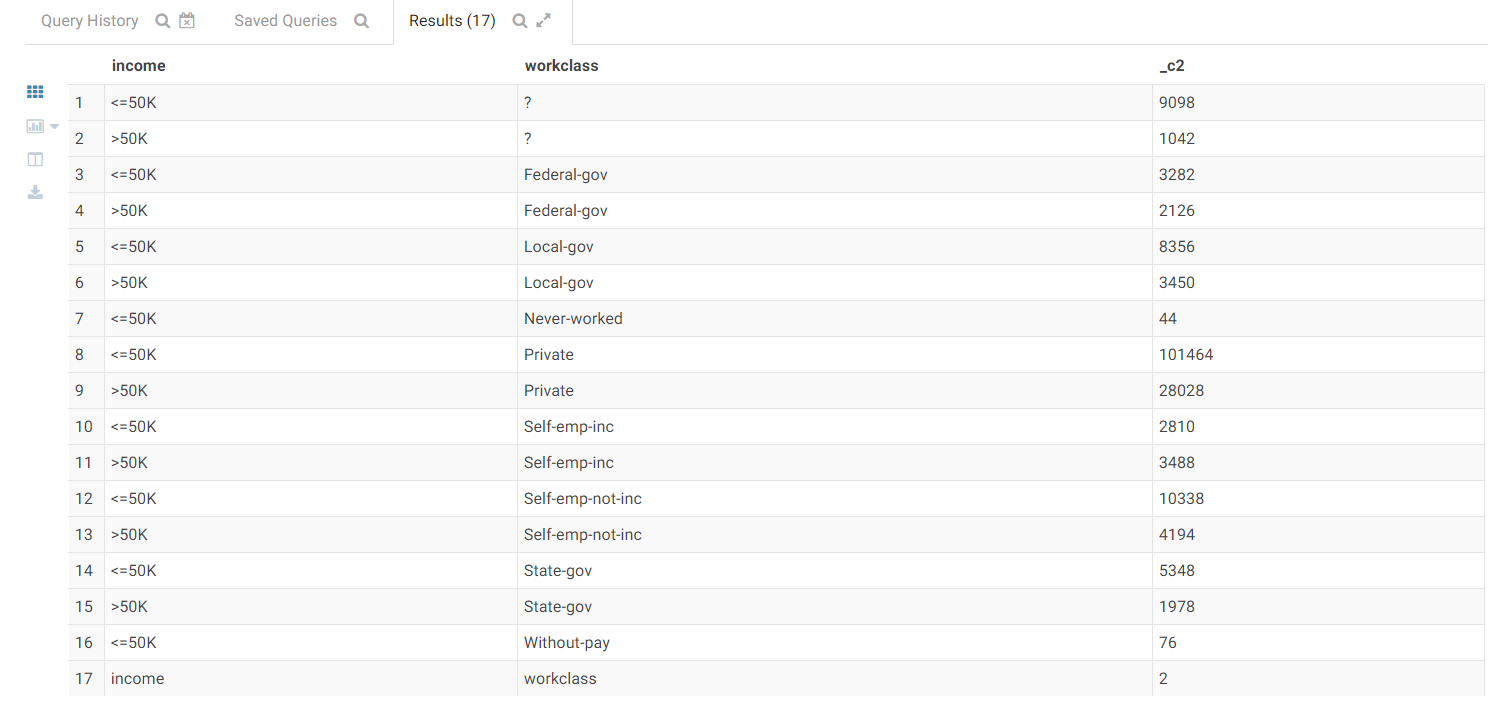
select income,workclass,count(\*) from census\_partitioned group by census\_partitioned.workclass, census\_partitioned.income



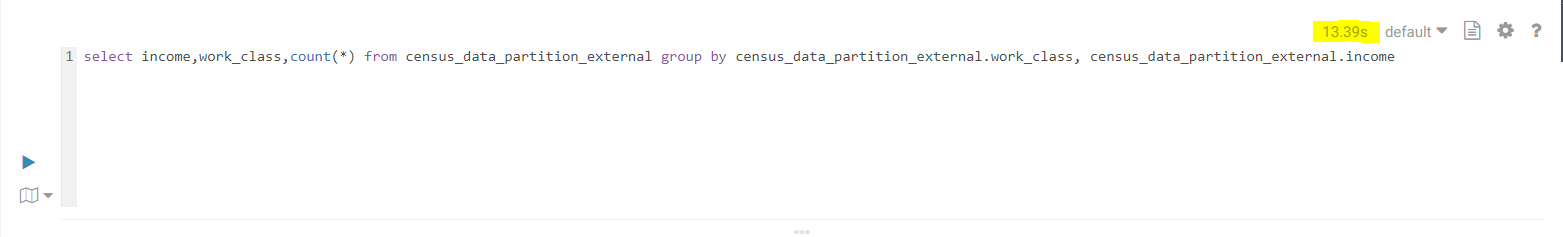


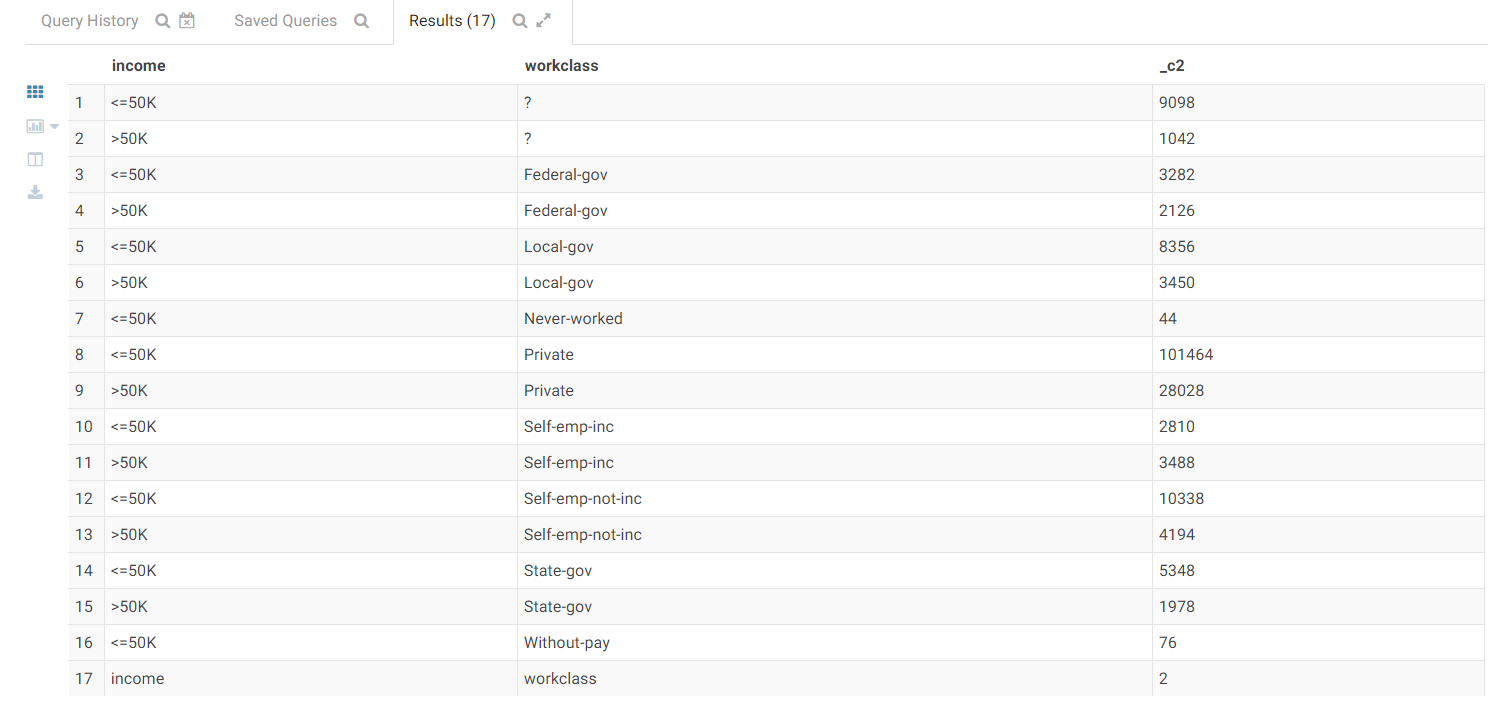
select income,work\_class,count(\*) from census\_data\_external group by census\_data\_external.work\_class, census\_data\_external.income





select income,work\_class,count(\*) from census\_data\_partition\_external group by census\_data\_partition\_external.work\_class, census\_data\_partition\_external.income





**Time taken:**

**Internal Non-Partitioned:- 12.59 sec**

**Internal Partitioned:- 12.2 sec**

**External Non-Partitioned:- 18.68 sec**

**External Partitioned:- 13.39 sec**