**Project: - Physician Compare Dataset**

**Name: - Bhavana Vankhede UB ID: - 1034353**

**MAP REDUCE PROGRAMMING: -**

1. Storing data in Map Reduce using Key, Value pair.

As per requirement, I have used professional id which is PAC\_ID from dataset and Enrollment ID which is Professional Enrollment ID from source dataset, to get detaials about professionals enrolled with multiple offices.

Key PAC\_ID (Physician's ID) and Value (enrollment ID)

In order to run map reduce program, shell wrapper script is created I.e. map\_Reduce.sh script.

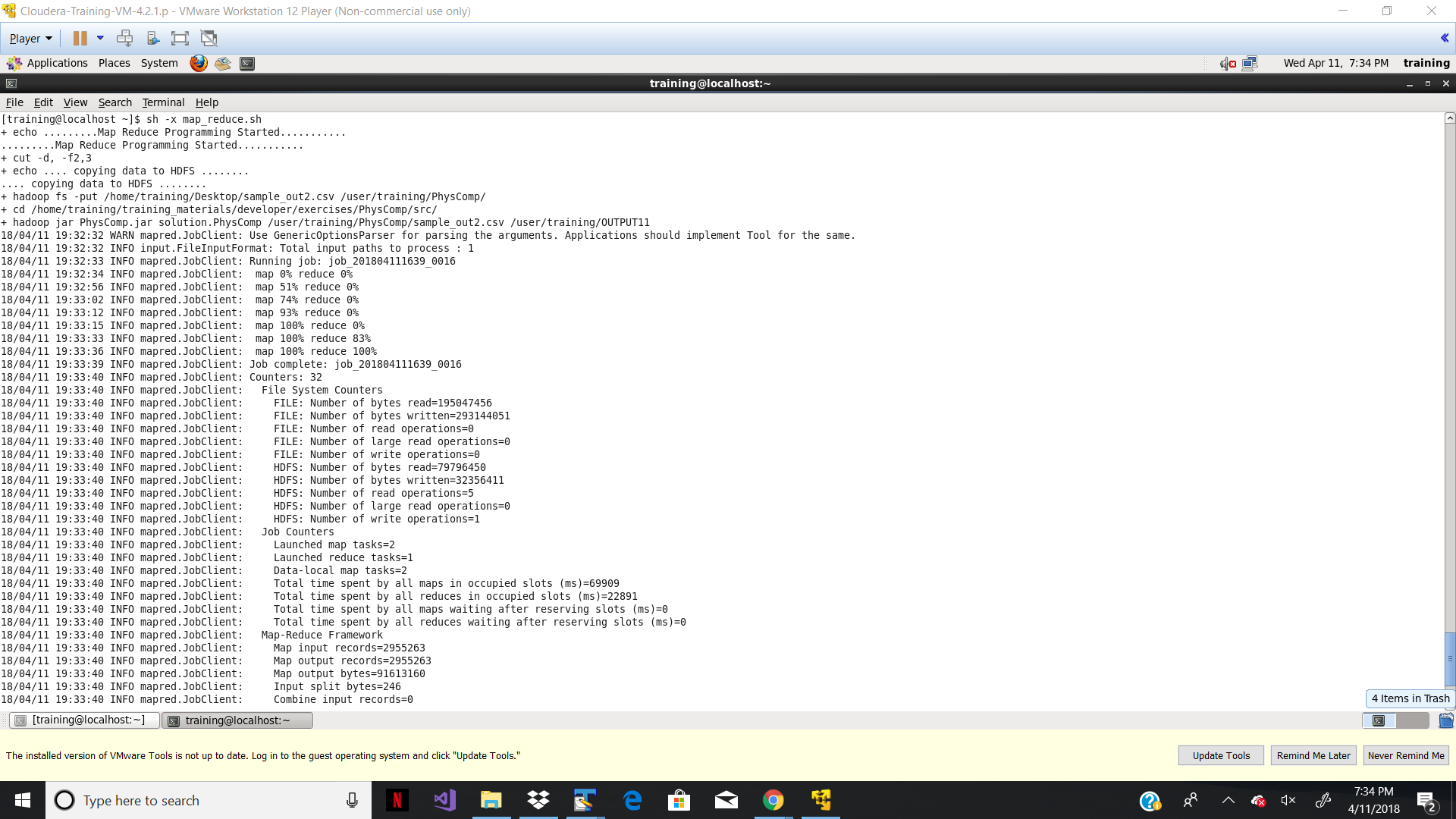
**Command**: -

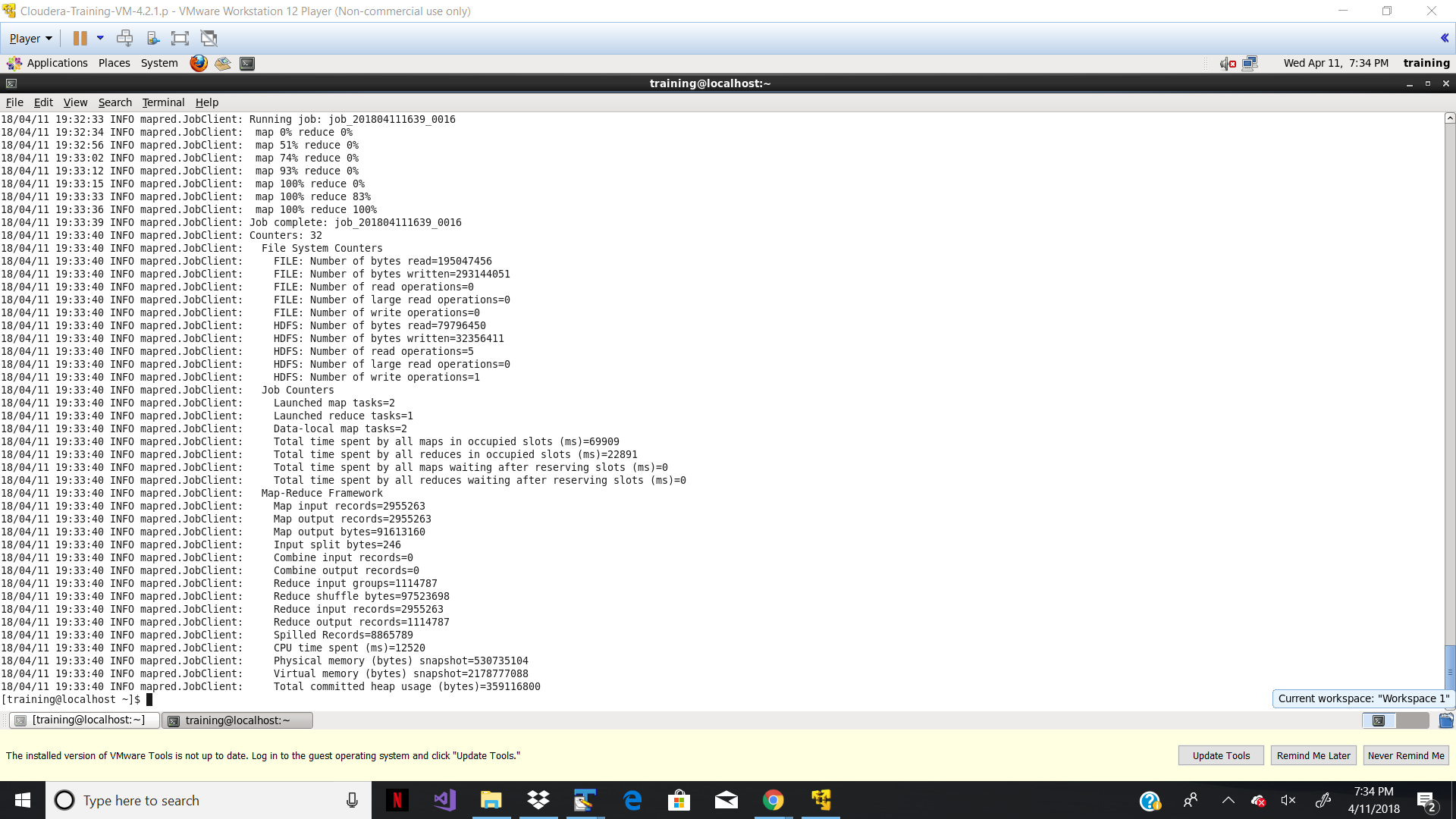
Sh -x map\_reduce.sh

Above script uses below commands to run map reduce program: -

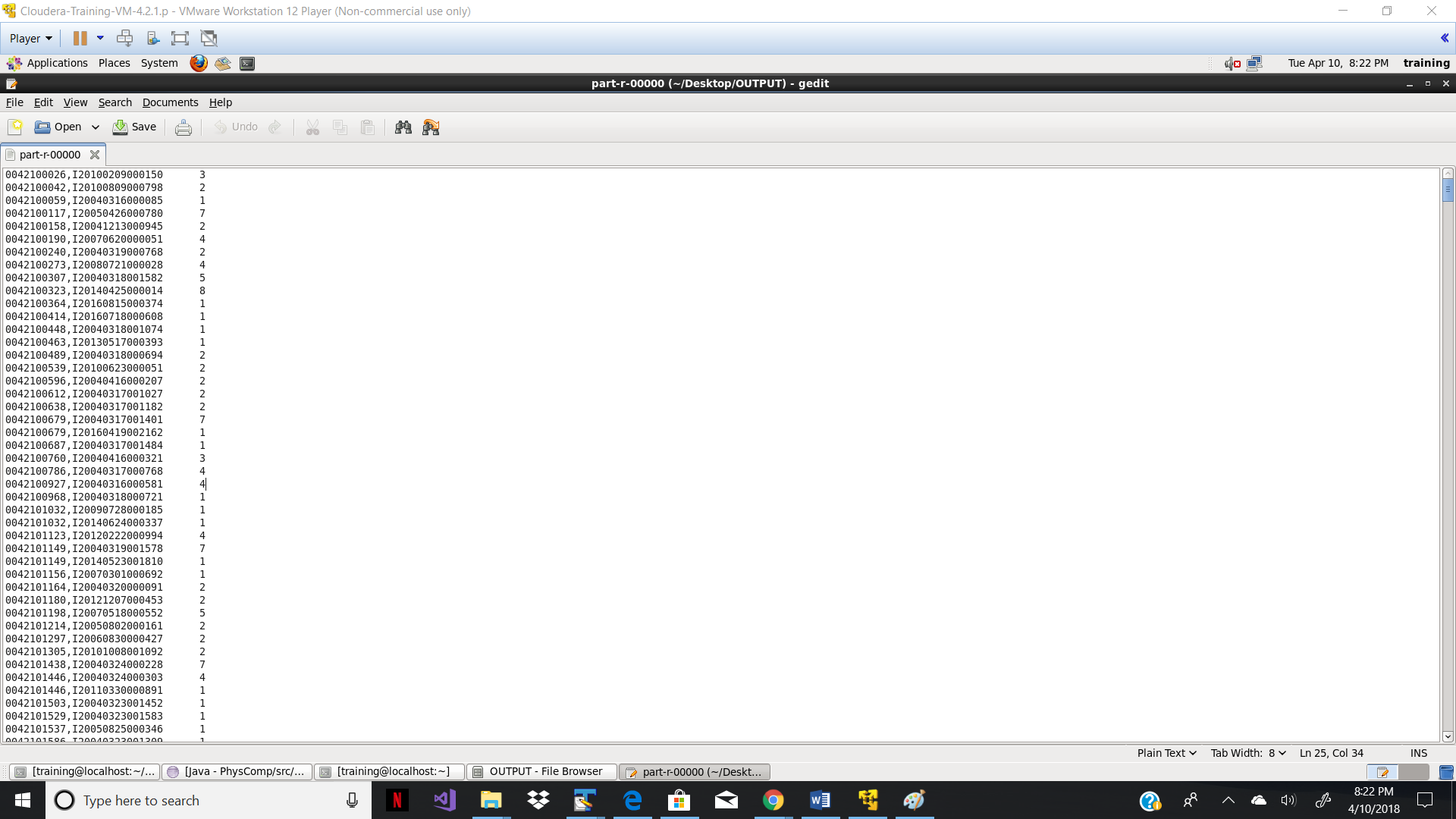
hadoop jar PhysComp.jar solution.PhysComp /user/training/PhysComp/sample\_out.csv /user/training/OUTPUT11

**Execution Details: -**





**OUTPUT: -**



**Hive Programming: -**

1. Folder is created and Physcomp.csv kept on below path: -

Make Directory: -

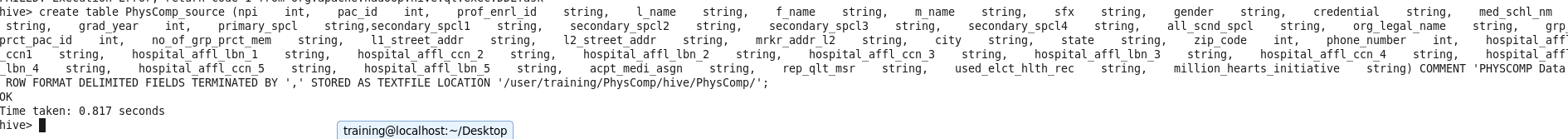
hadoop fs –mkdir /user/training/PhysComp/hive/PhysComp

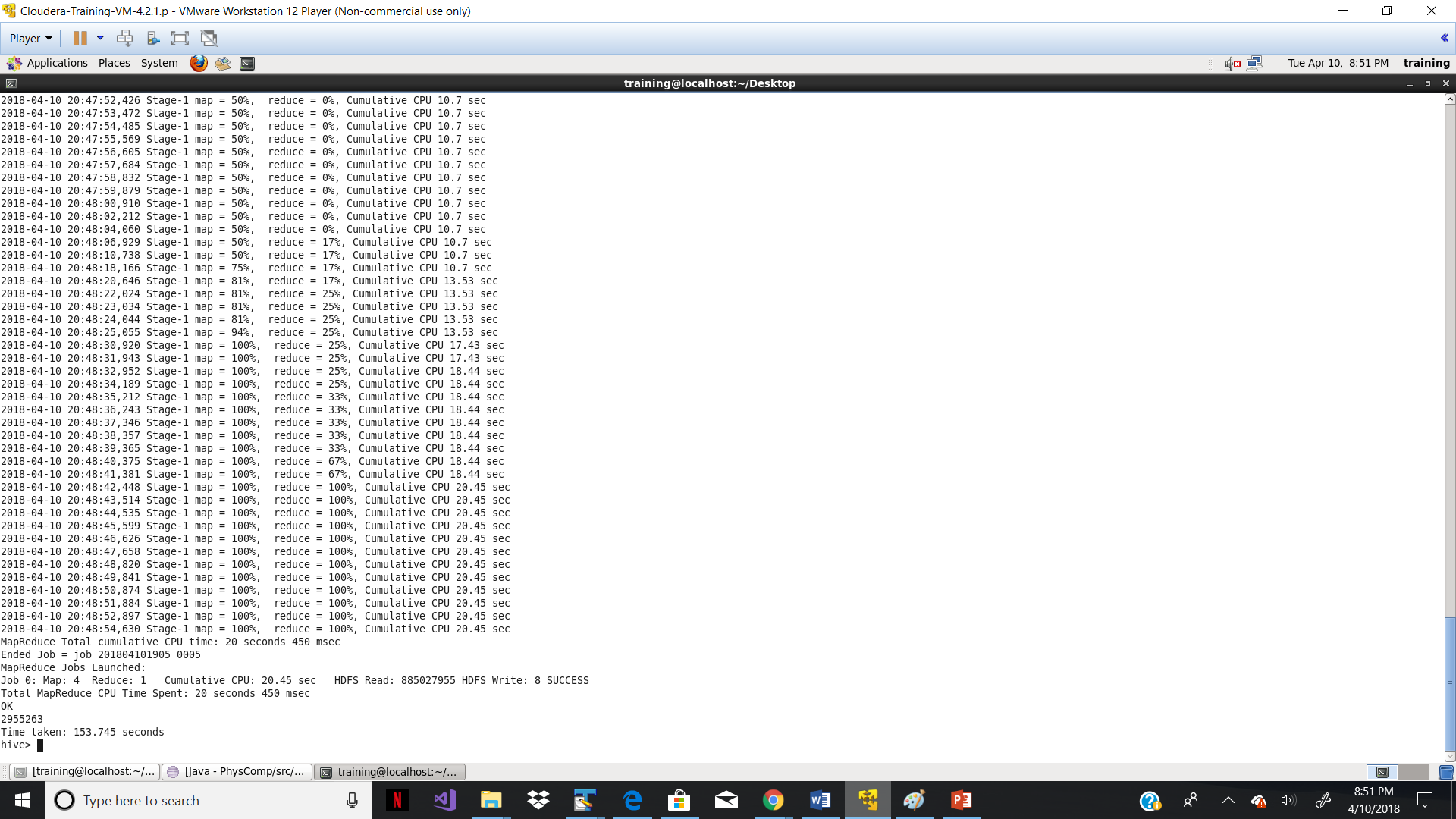
Copy File from Local system to Hadoop File system: -

Hadoop fs –put /home/training/Desktop/Sample.csv /user/training/PhysComp/hive/PhysComp/

1. Data Loaded on HIVE tables: -

create table physComp\_source (npi int, pac\_id int, prof\_enrl\_id string, l\_name string, f\_name string, m\_name string, sfx string, gender string, credential string, med\_schl\_nm string, grad\_year int, primary\_spcl string,secondary\_spcl1 string, secondary\_spcl2 string, secondary\_spcl3 string, secondary\_spcl4 string, all\_scnd\_spcl string, org\_legal\_name string, grp\_prct\_pac\_id int, no\_of\_grp\_prct\_mem string, l1\_street\_addr string, l2\_street\_addr string, mrkr\_addr\_l2 string, city string, state string, zip\_code int, phone\_number int, hospital\_affl\_ccn1 string, hospital\_affl\_lbn\_1 string, hospital\_affl\_ccn\_2 string, hospital\_affl\_lbn\_2 string, hospital\_affl\_ccn\_3 string, hospital\_affl\_lbn\_3 string, hospital\_affl\_ccn\_4 string, hospital\_affl\_lbn\_4 string, hospital\_affl\_ccn\_5 string, hospital\_affl\_lbn\_5 string, acpt\_medi\_asgn string, rep\_qlt\_msr string, used\_elct\_hlth\_rec string, million\_hearts\_initiative string) COMMENT 'PHYSCOMP Data' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/training/PhysComp/hive/PhysComp/Sample.csv';





1. Generating Report on basis of Specialization: -

Create a static table only Specialties for data cleansing as so many random values were populated from source dataset: -

create table Specialties as select distinct primary\_spcl from Physcomp\_source where primary\_spcl IN

('ADDICTION MEDICINE','ADVANCED HEART FAILURE AND TRANSPLANT CARDIOLOGY', 'ALLERGY/IMMUNOLOGY', 'ANESTHESIOLOGY', 'ANESTHESIOLOGY ASSISTANT', 'AUDIOLOGIST', 'CARDIAC ELECTROPHYSIOLOGY', 'CARDIAC SURGERY', 'CARDIOVASCULAR DISEASE (CARDIOLOGY)', 'CERTIFIED NURSE MIDWIFE', 'CERTIFIED REGISTERED NURSE ANESTHETIST', 'CHIROPRACTIC', 'CLINICAL NURSE SPECIALIST','CLINICAL PSYCHOLOGIST', 'CLINICAL SOCIAL WORKER', 'COLORECTAL SURGERY (PROCTOLOGY)', 'CRITICAL CARE (INTENSIVISTS)', 'DENTIST', 'DERMATOLOGY', 'DIAGNOSTIC RADIOLOGY', 'EMERGENCY MEDICINE', 'ENDOCRINOLOGY', 'FAMILY PRACTICE', 'GASTROENTEROLOGY',

'GENERAL PRACTICE', 'GENERAL SURGERY', 'GERIATRIC MEDICINE', 'GERIATRIC PSYCHIATRY',

'GYNECOLOGICAL ONCOLOGY', 'HAND SURGERY','HEMATOLOGY','HEMATOLOGY/ONCOLOGY',

'HEMATOPOIETIC CELL TRANSPLANTATION AND CELLULAR TH', 'HOSPICE/PALLIATIVE CARE',

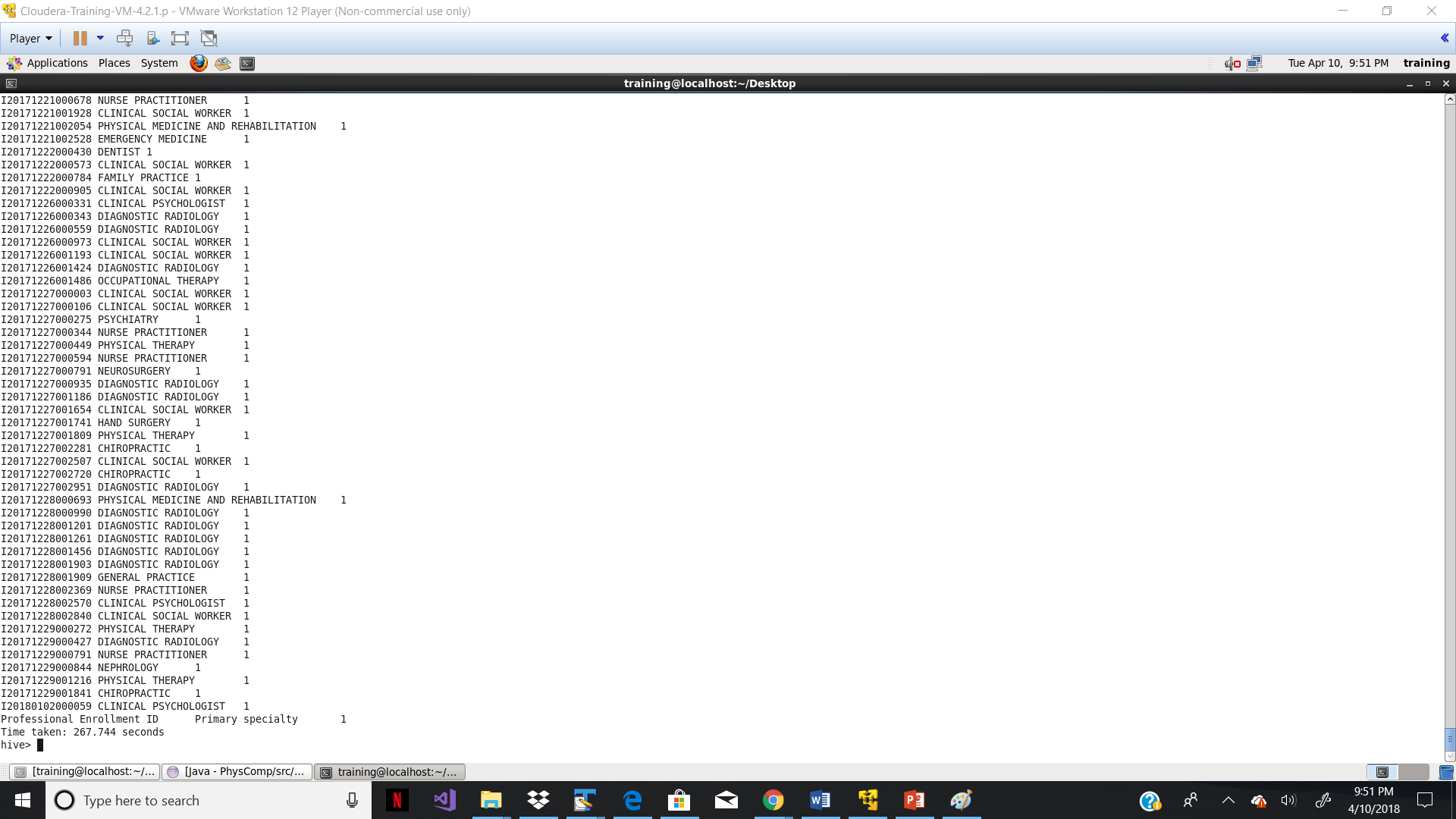
'HOSPITALIST', 'INFECTIOUS DISEASE', 'INTERNAL MEDICINE','INTERVENTIONAL CARDIOLOGY',

'INTERVENTIONAL PAIN MANAGEMENT', 'INTERVENTIONAL RADIOLOGY', 'MASS IMMUNIZATION ROSTER BILLER', 'MAXILLOFACIAL SURGERY', 'MEDICAL ONCOLOGY', 'MEDICAL TOXICOLOGY', 'NEPHROLOGY',

'NEUROLOGY','NEUROPSYCHIATRY','NEUROSURGERY','NUCLEAR MEDICINE','NURSE PRACTITIONER','OBSTETRICS/GYNECOLOGY','OCCUPATIONAL THERAPY','OPHTHALMOLOGY','OPTOMETRY','ORAL SURGERY','ORTHOPEDIC SURGERY','OSTEOPATHIC MANIPULATIVE MEDICINE','OTOLARYNGOLOGY','PAIN MANAGEMENT','PATHOLOGY','PEDIATRIC MEDICINE','PERIPHERAL VASCULAR DISEASE','PHYSICAL MEDICINE AND REHABILITATION','PHYSICAL THERAPY','PHYSICIAN ASSISTANT','PLASTIC AND RECONSTRUCTIVE SURGERY','PODIATRY','PREVENTATIVE MEDICINE','PSYCHIATRY','PULMONARY DISEASE','RADIATION ONCOLOGY','REGISTERED DIETITIAN OR NUTRITION PROFESSIONAL','RHEUMATOLOGY','SLEEP MEDICINE','SPEECH LANGUAGE PATHOLOGIST','SPORTS MEDICINE','SURGICAL ONCOLOGY','THORACIC SURGERY','UROLOGY','VASCULAR SURGERY')

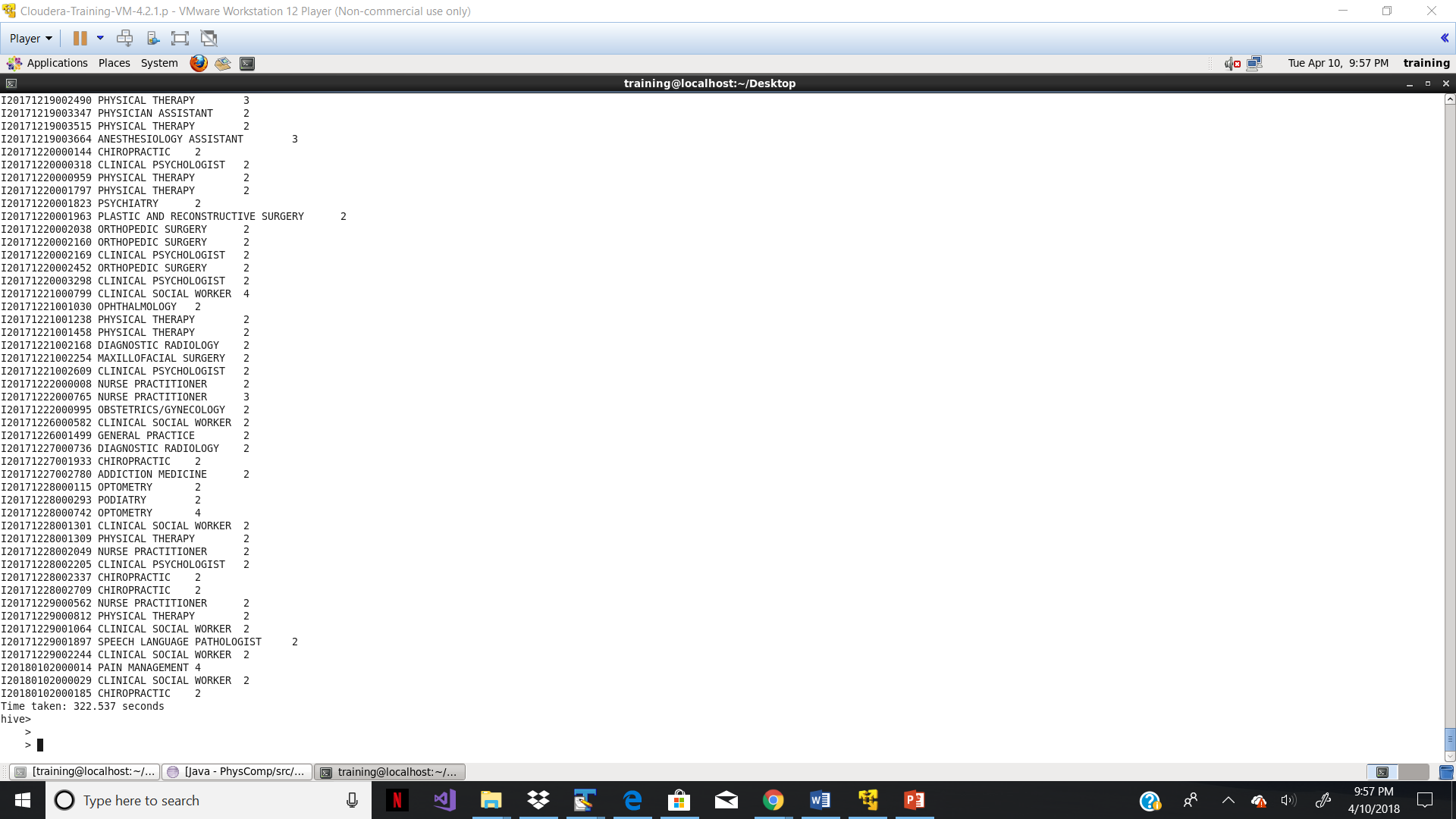
**To get Professionals with same specialties with single specialization: -**

select a.prof\_enrl\_id, a.primary\_spcl,count(\*) from physcomp\_source a join Specialties b on (a.primary\_spcl=b.primary\_spcl) group by a.prof\_enrl\_id,a.primary\_spcl having count(\*) =1;



**To get Professionals with same specialties with multiple specialization: -**

select a.prof\_enrl\_id, a.primary\_spcl,count(\*) from physcomp\_source a join Specialties b on (a.primary\_spcl=b.primary\_spcl) group by a.prof\_enrl\_id,a.primary\_spcl having count(\*) >1;



**--> Static Partition: -**

I have use static portioning to store the data on basis of graduation year of physicians.

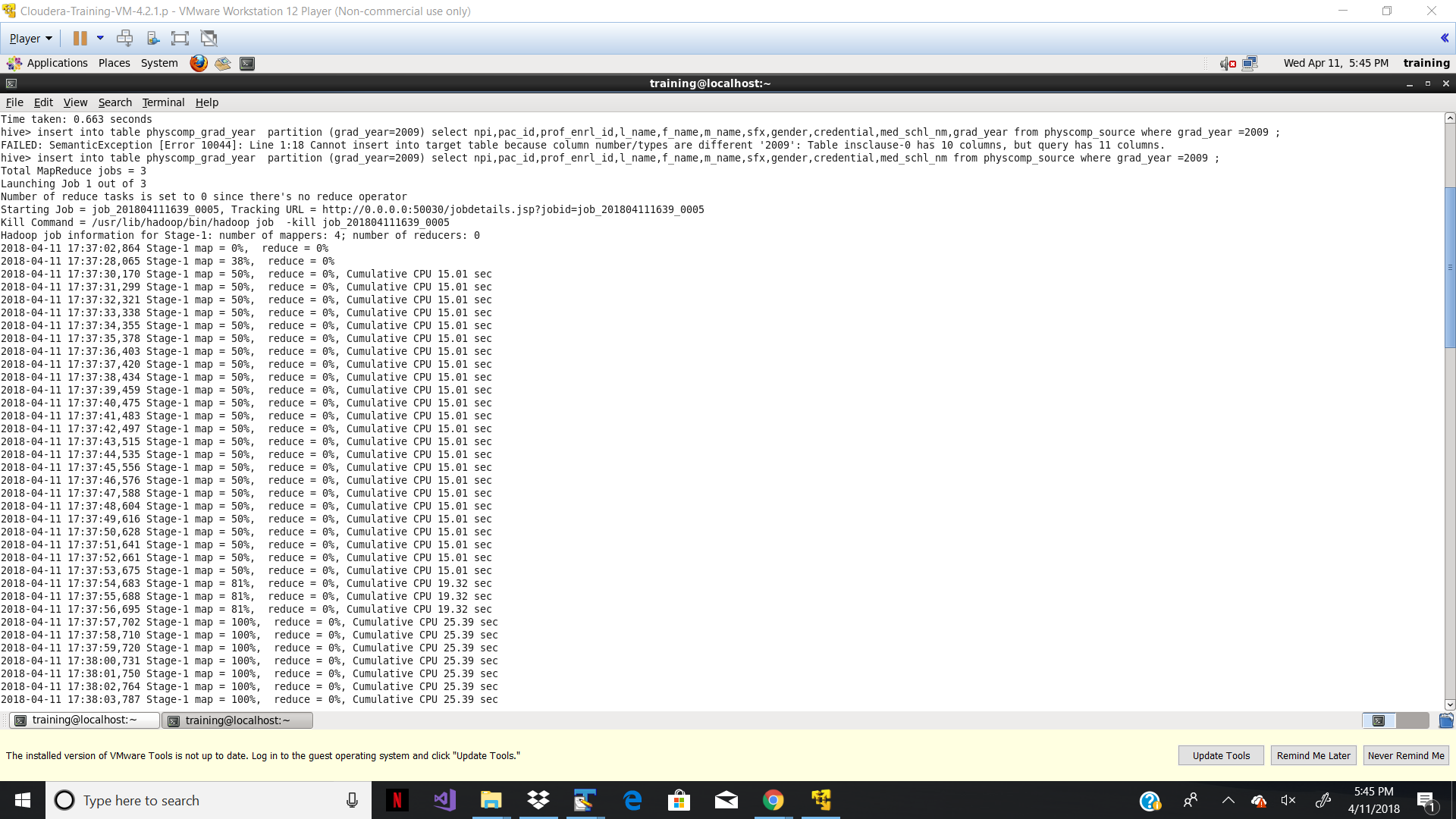
For that created separate table than source table.

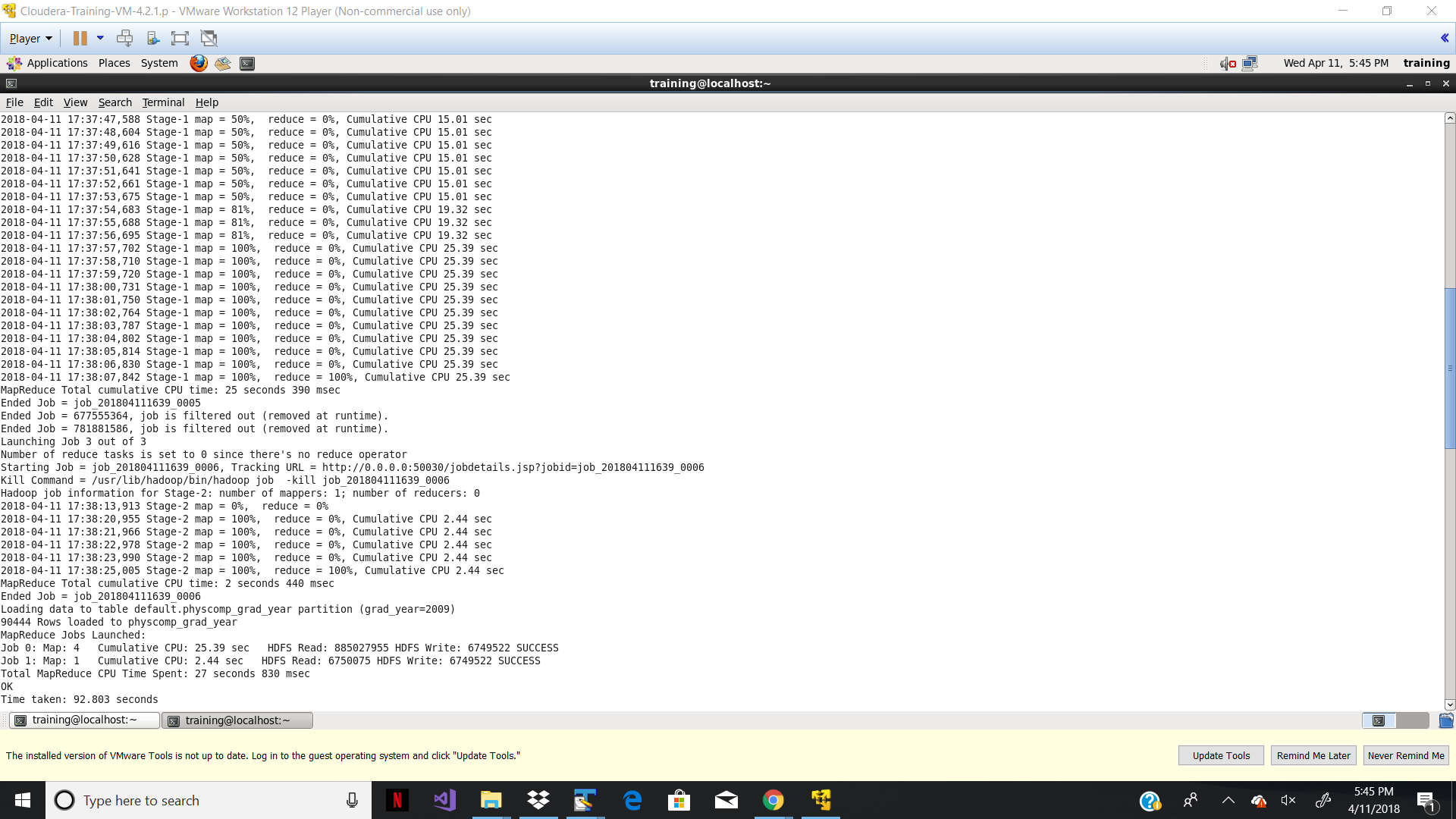
Table Creation Scripts: -

create table physcomp\_grad\_year(npi int,pac\_id int,prof\_enrl\_id string, l\_name string, f\_name string, m\_name string, sfx string , gender string, credential string, med\_schl\_nm string )partitioned by (grad\_year int);

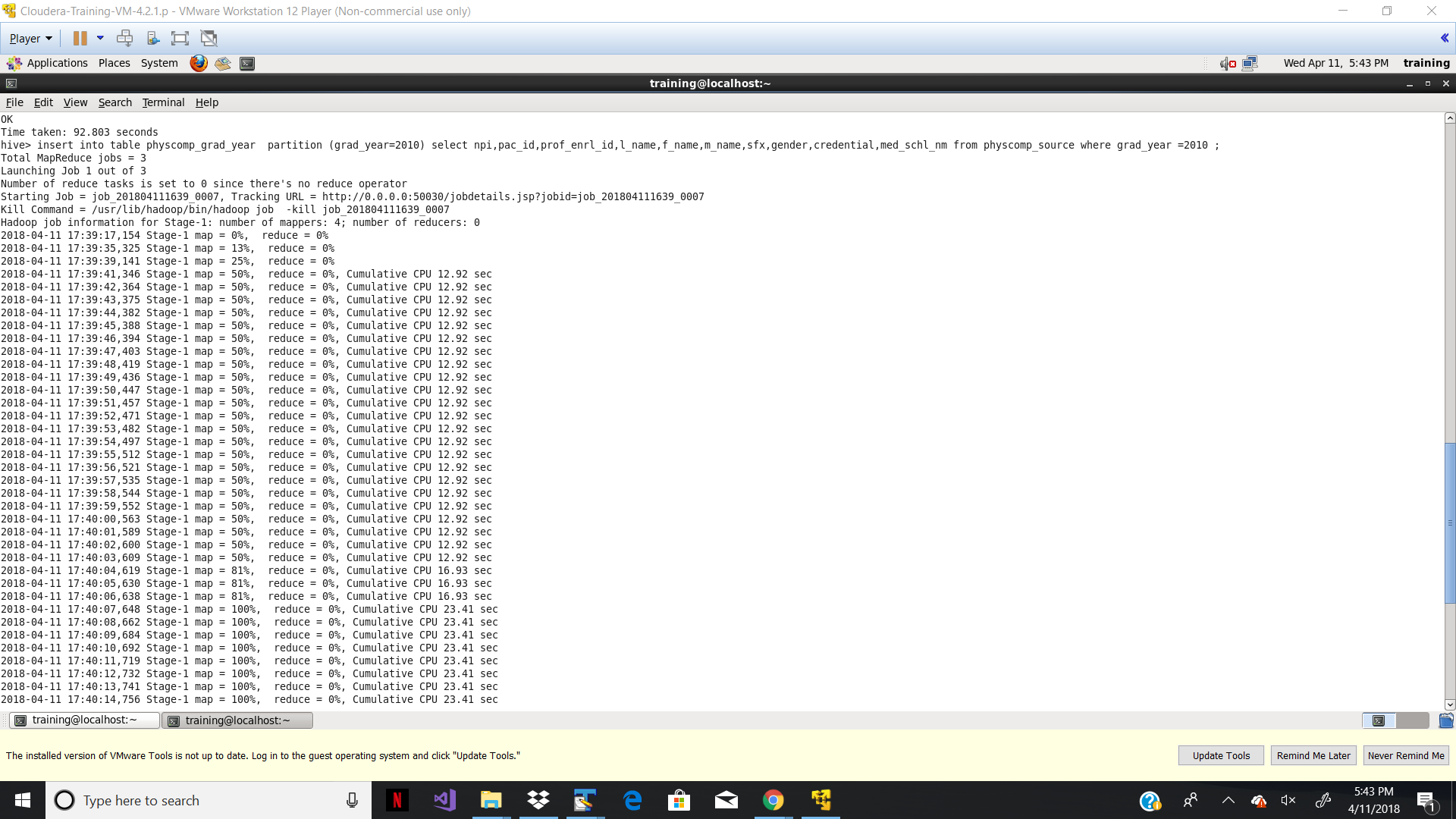
Insert data on basis of partitions: -

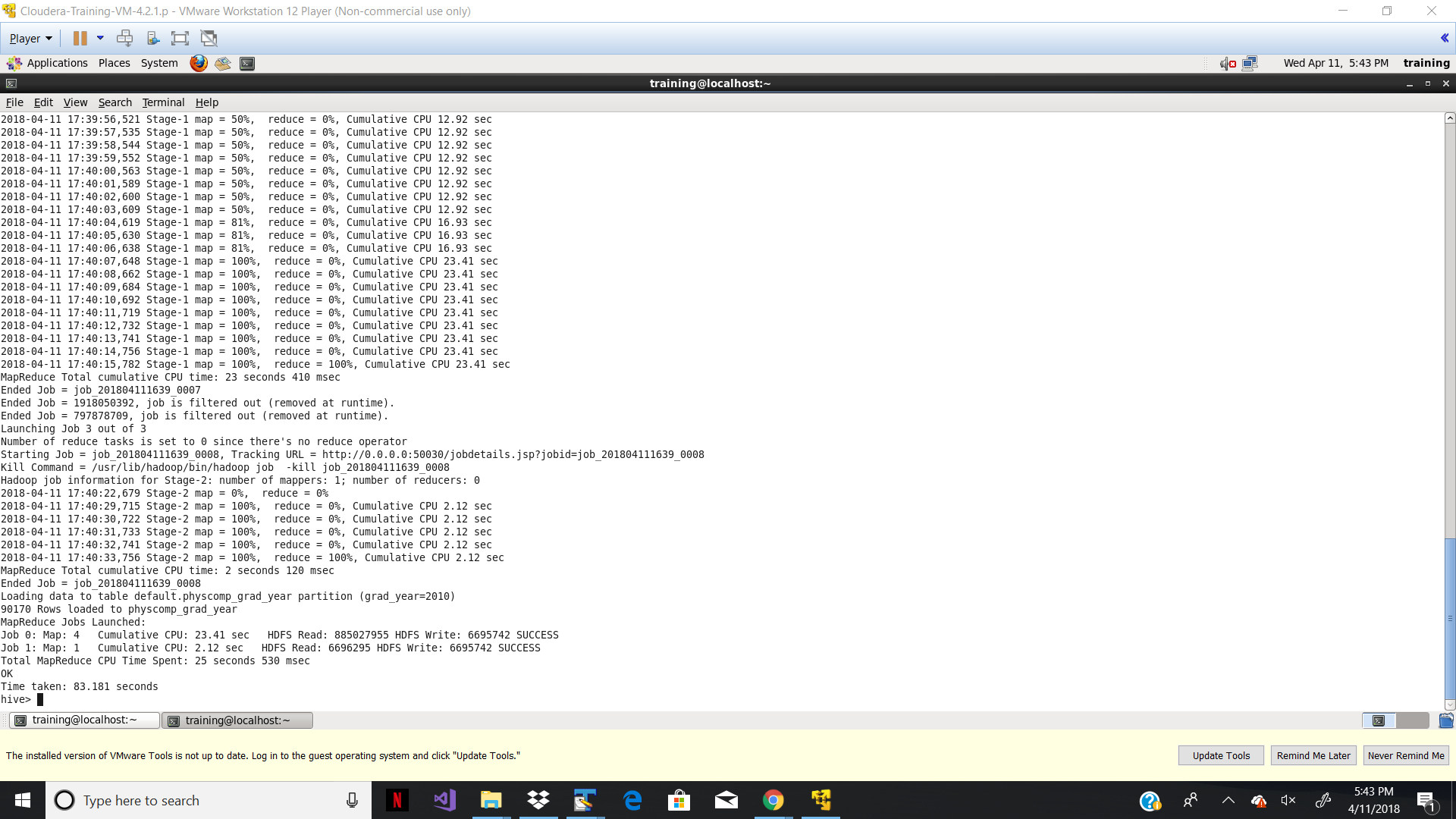
hive> insert into table physcomp\_grad\_year partition (grad\_year=2009) select npi,pac\_id,prof\_enrl\_id,l\_name,f\_name,m\_name,sfx,gender,credential,med\_schl\_nm,grad\_year from physcomp\_source where grad\_year =2009 ;



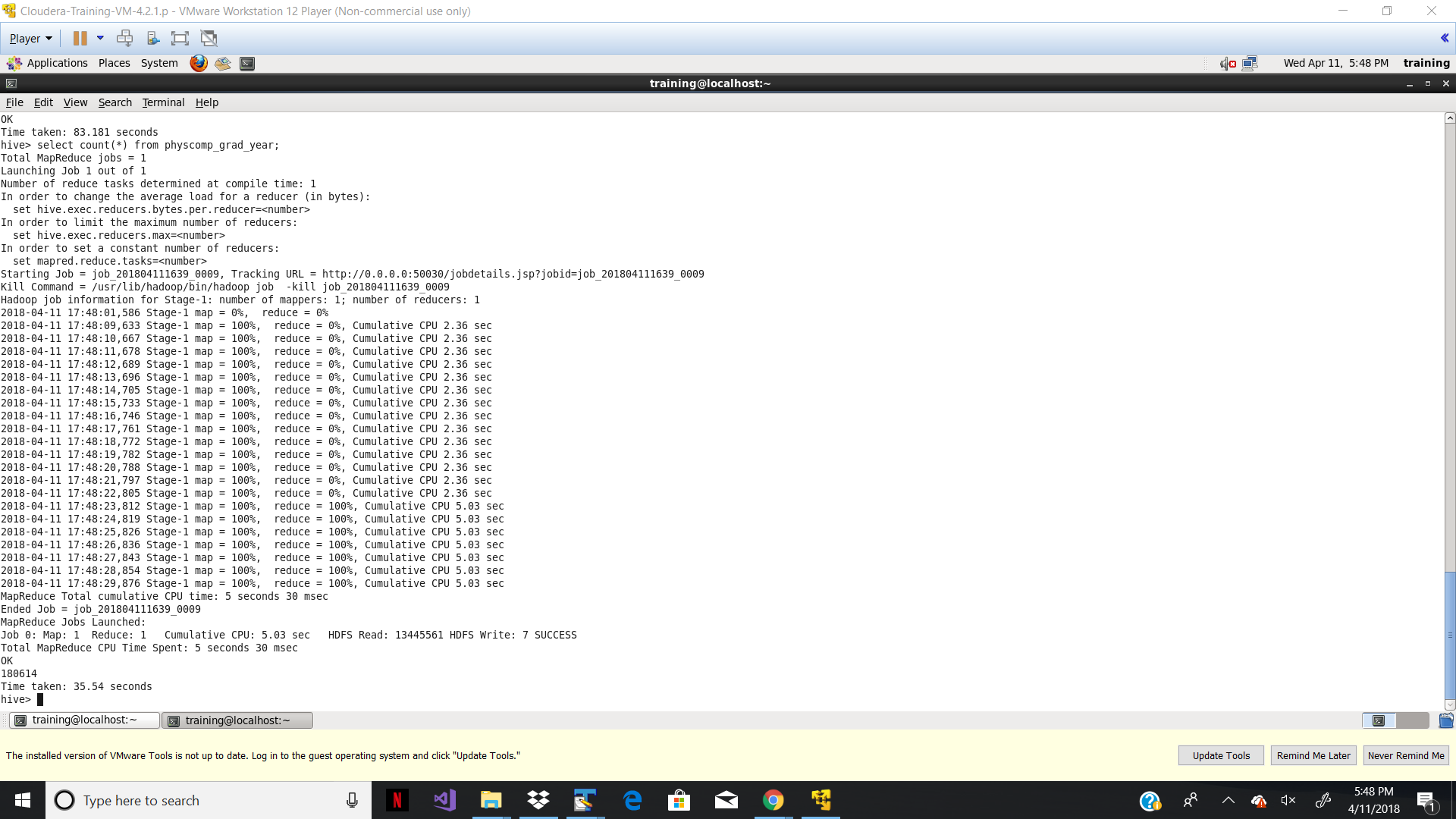


hive> insert into table physcomp\_grad\_year partition (grad\_year=2010) select npi,pac\_id,prof\_enrl\_id,l\_name,f\_name,m\_name,sfx,gender,credential,med\_schl\_nm,grad\_year from physcomp\_source where grad\_year =2010;





**Data inserted in Physcomp\_grad\_year tables on basis of partitioning: -**



* **-Dynamic Partition: -**

There are couple of data issues from source datasets. While loading data from source dataset, it was unable to load. Hence, I have tried to implement dynamic partition on very few datasets.

For this I have used table which is created in static partition.

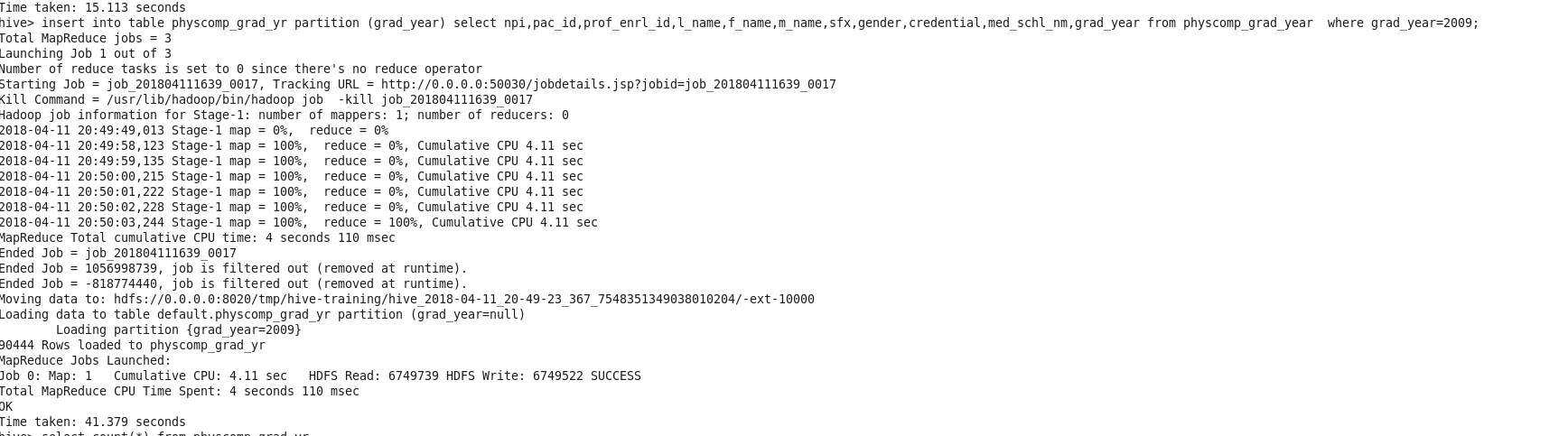
**For Dynamic Partition separate table created: -**

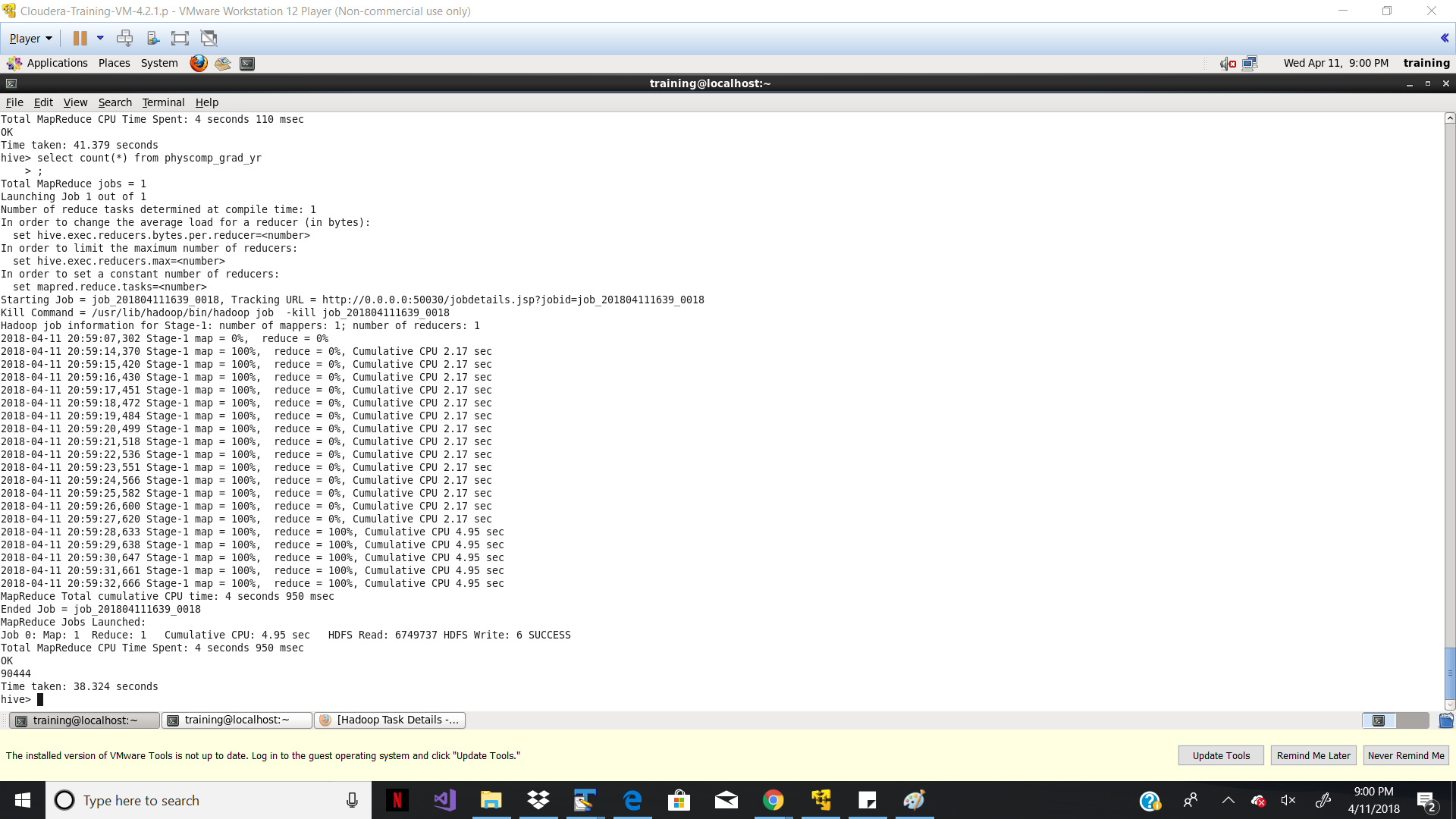
hive> create table physcomp\_grad\_yr(npi int,pac\_id int,prof\_enrl\_id string, l\_name string, f\_name string, m\_name string, sfx string , gender string, credential string, med\_schl\_nm string ) partitioned by (grad\_year int);

**Data Inserted in Physcomp\_grad\_yr from Physcomp\_grad\_year-**

insert into table physcomp\_grad\_yr partition (grad\_year) select npi,pac\_id,prof\_enrl\_id,l\_name,f\_name,m\_name,sfx,gender,credential,med\_schl\_nm,grad\_year from physcomp\_grad\_year where grad\_year=2009;

**Output:-**





**Pig Relation: -**

To get the number of Female and Male physician, following script can be used: -

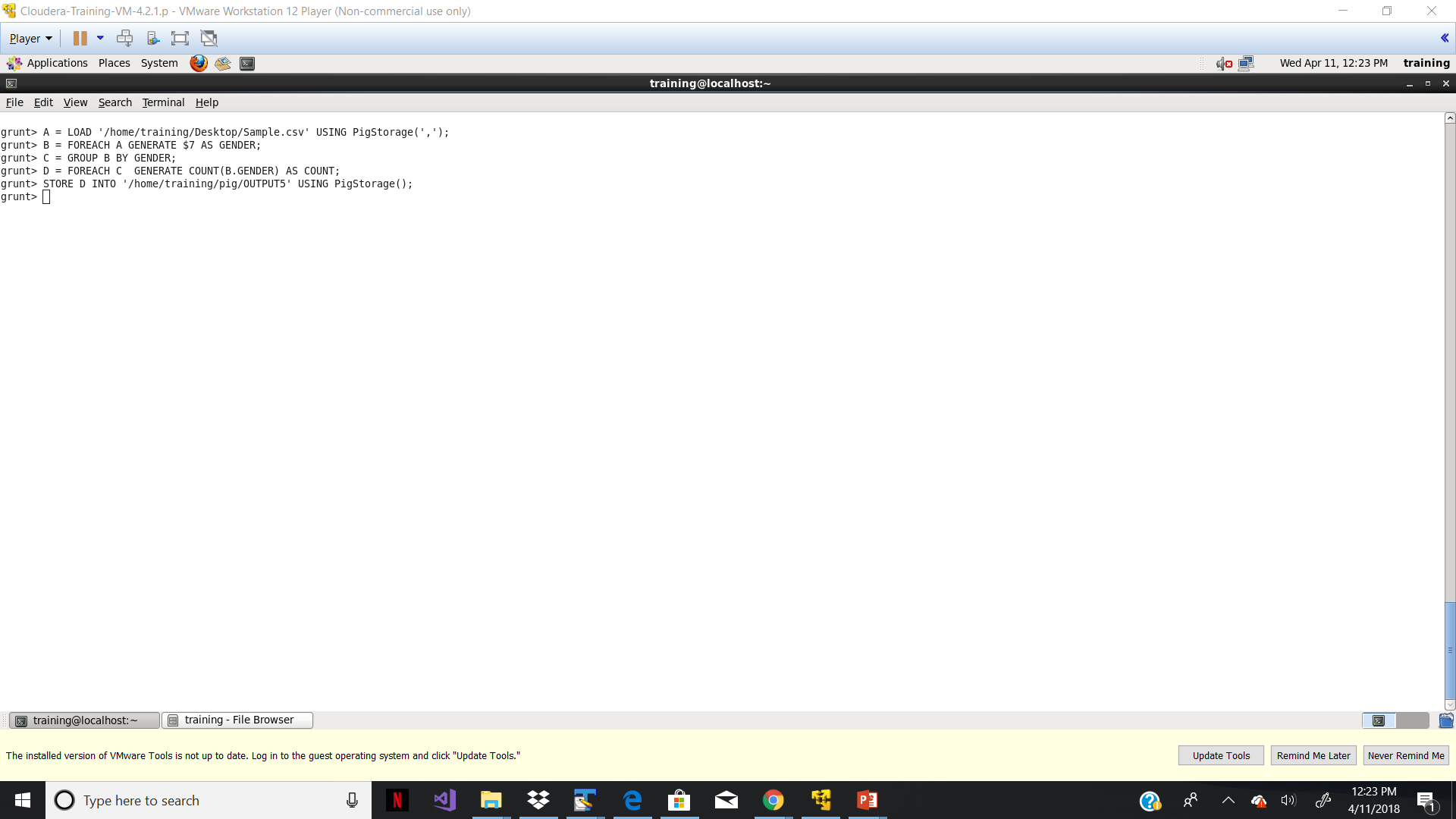
A = LOAD '/home/training/Desktop/Sample.csv' USING PigStorage(',');

B = FOREACH A GENERATE $7 AS GENDER;

C = GROUP B BY GENDER;

D = FOREACH C GENERATE COUNT(B.GENDER) AS COUNT;

STORE D INTO '/home/training/pig/OUTPUT4' USING PigStorage();



**Output: -**

