H1B-VISA PETITIONS
(2011-2016)

H1-B CASE STUDY

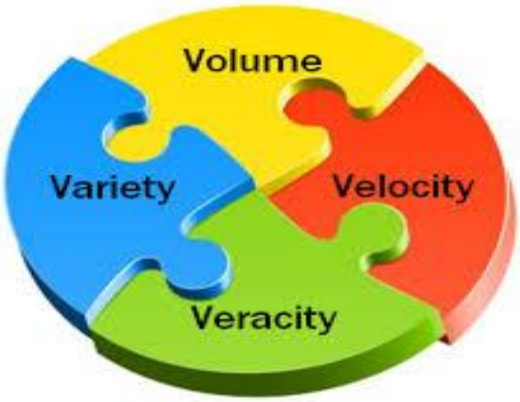
Anchal Sadhu | Big Data and Hadoop

A Brief Introduction about Various Technologies

BIG DATA

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information private

4V'S OF Big Data



Volume

It is estimated that, on an average, 2.3 trillion gigabytes of data are generated every day. Forget analyzing, simply capturing such quantities of data is impractical. Most companies in the US have at least 100,000 gigabytes of data stored; and almost all of them will tell you that they aren't collecting enough data.
The right approach is to fight the urge of making your company's server a data dump. Efforts must be made to employ the right software to filter the relevant data.

Variety

Along with quantity, the diversity of data is equally important. The variety in data can be in terms of the devices or the sources of data generation.

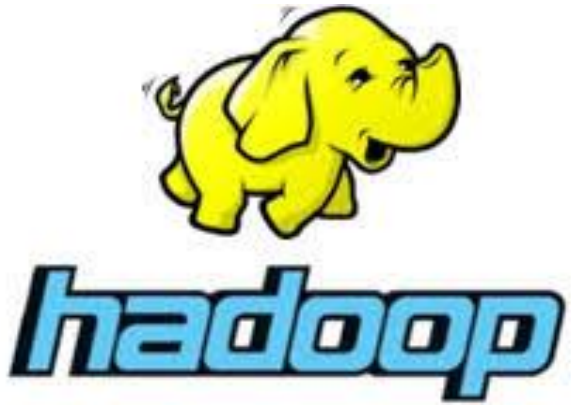
Velocity

Not only is the volume of data ever increasing, but the rate of data generation (from the Internet of Things, social media, etc.) is increasing as well.

Veracity

Not all data is good. In fact, unfiltered data is more likely to be bad than good. Although data quality and usability depend largely on the source, you can never rule out junk.
This unreliability of data makes many business heads reluctant to rely on information analysis. That's the wrong approach.

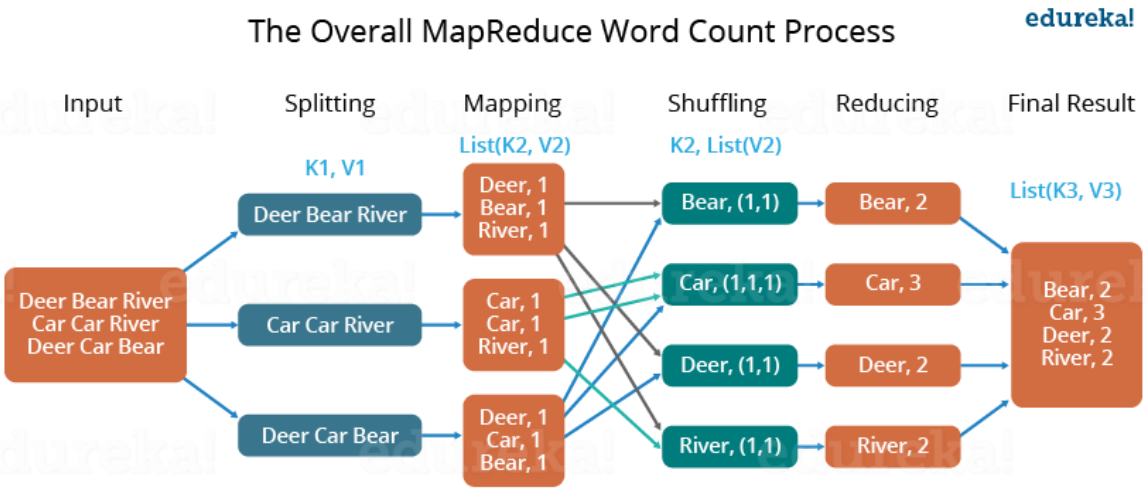
HADOOP



Apache Hadoop is an open-source software framework used for distributed storage and processing of very large data sets.

It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are a common occurrence and should be automatically handled by the frame.

MapReduce

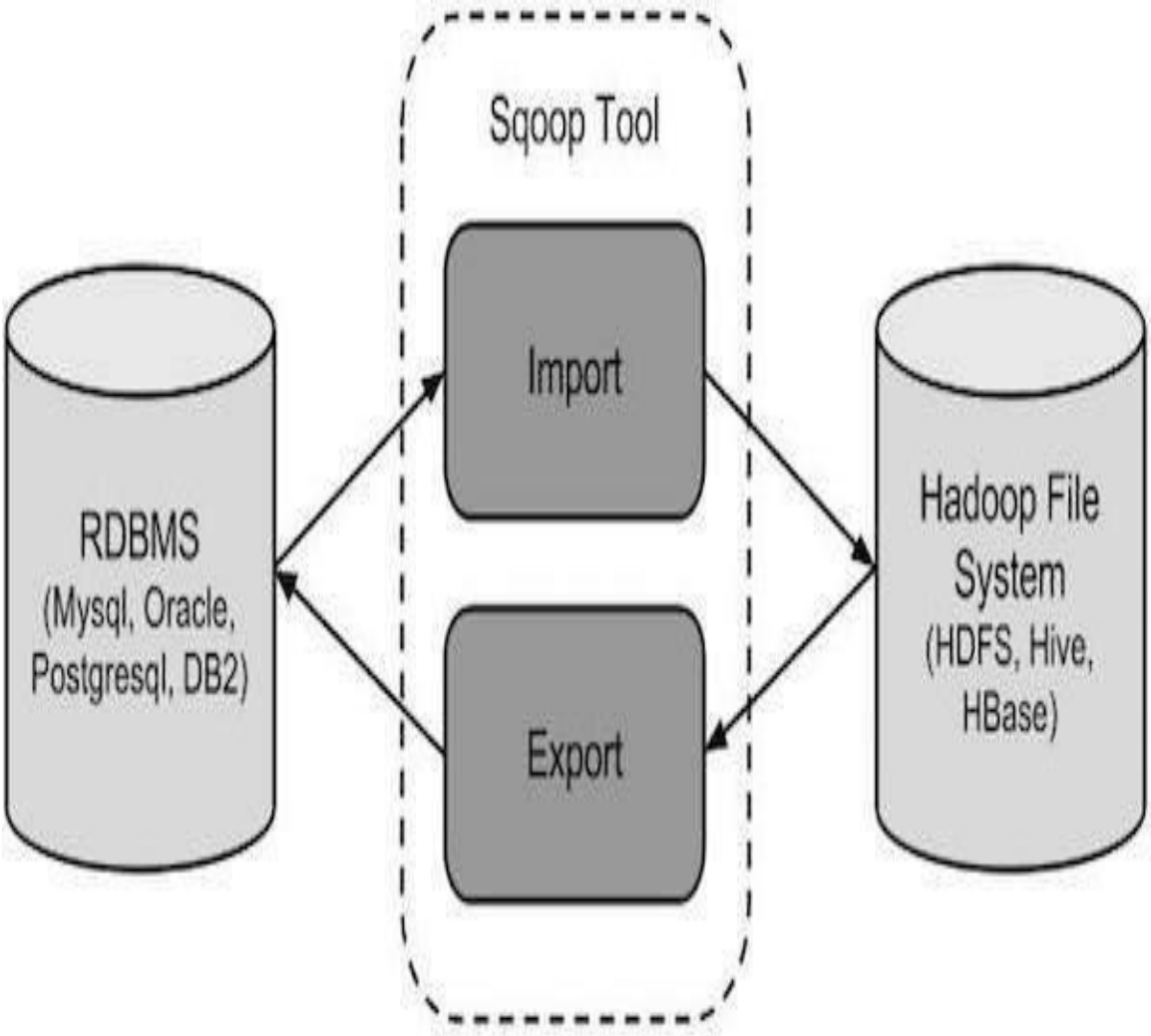


- Hadoop MapReduce (Hadoop Map/Reduce) is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. It is a sub-project of the Apache Hadoop project. The framework takes care of scheduling tasks, monitoring them and re-executing any failed tasks.

- A MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file-system.

SQOOP

SQOOP is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.


Sqoop Tool
Import
Export
RDBMS
(Mysql, Oracle,
Postgresql, DB2)
Hadoop File
System
(HDFS, Hive,
HBase)

HIVE

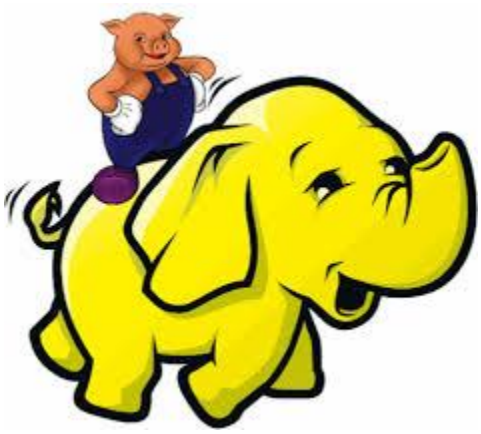
Apache Hive is a data warehouse infrastructure built on top of Hadoop for providing data summarization, query, and analysis.

Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

It is built on top of Hadoop and developed by Facebook. Hive provides a way to query the data using a SQL-like query language called HiveQL (Hive query Language).

Internally, a compiler translates HiveQL statements into MapReduce jobs, which are then submitted to Hadoop framework for execution.

Apache PIG



Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin.

Pig can execute its Hadoop jobs in Map Reduce, Apache Tez, or Apache Spark. Pig Latin abstracts the programming from the Java Map Reduce idiom into a notation which makes Map Reduce programming high level, similar to that of SQL for RDBMS.

Pig Latin can be extended using User Defined Functions (UDFs) which the user can write in Java, Python, JavaScript, Ruby or Groovy and then call directly from the language.

H1-B CASE STUDY

PROJECT OBJECTIVE **---** The H1- B is an employment –based, non-immigrant visa category for temporary foreign workers in the United States. For a foreign national to apply H1-B Visa, an US employer must offer a job and petition for H1-B visa with the US immigration department. This is the most common visa status applied for and held by international students once they complete college/higher education (Maters, Ph.D.) and work in a full-time position.

Hardware Requirements-

8 GB RAM

64Bit OS

Technology Requirements-

- Apache Hadoop
- MapReduce
- Hive
- Pig
- SQOOP

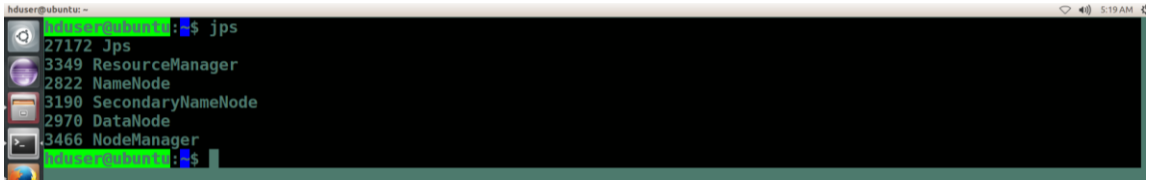
Software Used–

- VMware
- Ubuntu
- Eclipse
- MySQL

Assumptions-

- VMware Workstation – Configurations are set correctly.

- Ubuntu is lying on the Virtual Box and it is powered on
- Hadoop Folder must be extracted and all the services of the Hadoop is running. Configuration to be made in the XML are set.
- Confirmation Box Below that Everything is Set Right.



Datasets Required-

- H1-B CASE Applications Data
- The Dataset has nearly 3 million records.

The Dataset given below—

The columns in the dataset include:

1. **CASE_STATUS:** Status associated with the last significant event or decision. Valid values include "Certified," "Certified-Withdrawn," Denied," and "Withdrawn".

Certified: Employer filed the LCA, which was approved by DOL

Certified Withdrawn: LCA was approved but later withdrawn by employer

Withdrawn: LCA was withdrawn by employer before approval

Denied: LCA was denied by DOL

2. **EMPLOYER_NAME:** Name of employer submitting labour condition application.

3. **SOC_NAME:** The Occupational name associated with the SOC_CODE.

SOC_CODE is the occupational code associated with the job being requested for temporary labour condition, as classified by the Standard Occupational Classification (SOC) System.

4. **JOB_TITLE:** Title of the job

5. **FULL_TIME_POSITION:** Y = Full Time Position; N = Part Time Position

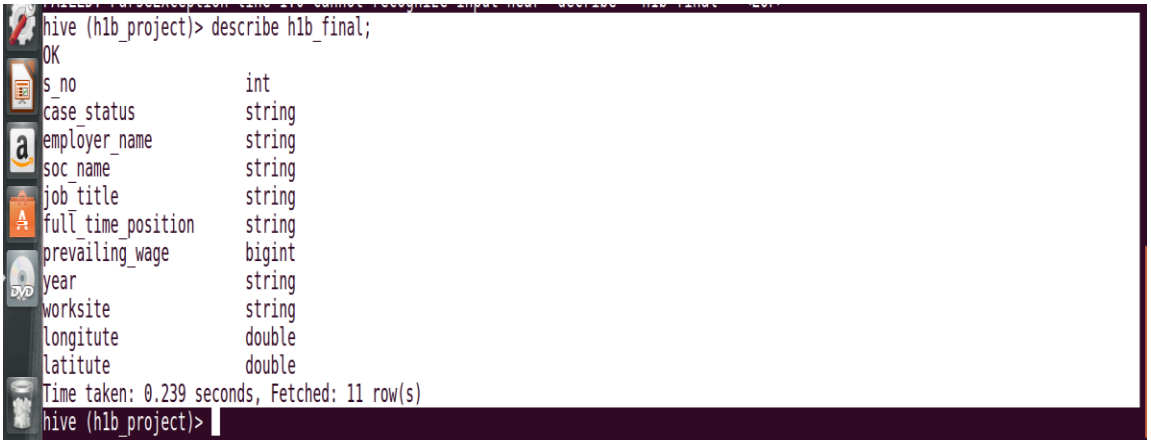
6. **PREVAILING_WAGE:** Prevailing Wage for the job being requested for temporary labour condition. The wage is listed at annual scale in USD. The prevailing wage for a job position is defined as the average wage paid to similarly employed workers in the requested occupation in the area of intended employment. The prevailing wage is based on the employer's minimum requirements for the position.

7. **YEAR:** Year in which the H1B visa petition was filed

8. **WORKSITE:** City and State information of the foreign worker's intended area of employment

9. **LONGITUDE:** longitude of the Worksite

10. **LATITUDE:** latitude of the Worksite



Outcome of this Project:

To generate reports and hence,

We will be performing analysis on the H1B visa applicants between the years 2011-2016. After analysing the data, we can derive the following facts.

1 a) Is the number of petitions with Data Engineer job title increasing over time?
b) Find top 5 job titles who are having highest avg growth in applications. [ALL]

2 a) Which part of the US has the most Data Engineer jobs for each year?
b) find top 5 locations in the US who have got certified visa for each year. [certified]

3)Which industry(SOC_NAME) has the most number of Data Scientist positions? [certified]

4)Which top 5 employers file the most petitions each year? - Case Status - ALL

5) Find the most popular top 10 job positions for H1B visa applications for each year?
a) for all the applications
b) for only certified applications.

6) Find the percentage and the count of each case status on total applications for each year. Create a line graph depicting the pattern of All the cases over the period of time.

7) Create a bar graph to depict the number of applications for each year [All]

8) Find the average Prevailing Wage for each Job for each Year (take part time and full time separate). Arrange the output in descending order - [Certified and Certified Withdrawn.]

9) Which are the employers along with the number of petitions who have the success rate more than 70% in petitions. (total petitions filed 1000 OR more than 1000)?

10) Which are the job positions along with the number of petitions which have the success rate more than 70% in petitions (total petitions filed 1000 OR more than 1000)?

11) Export result for question no 10 to MySQL database.

Let's create a table to load the h1b applicant's data as shown below:

- CREATE TABLE h1b_applications (s_no int, case_status string, employer_name string, soc_name string, job_title string, full_time_position string, prevailing_wage bigint, year string, worksite string, longitute double, latitute double) ROW FORMAT SERDE 'org. apache. hadoop. hive. serde2.OpenCSVSerde'
  WITH SERDEPROPERTIES (
  "separatorChar" = ",",
  "quoteChar" = "\""
) STORED AS TEXTFILE;

- load data local inpath '/home/hduser/Downloads/H1Project/h1b.csv' overwrite into table h1b_applications;

- CREATE TABLE h1b_app2(s_no int, case_status string, employer_name string, soc_name string, job_title string, full_time_position string, prevailing_wage bigint, year string, worksite string, longitute double, latitute double)
  row format delimited
  fields terminated by '\t'
  STORED AS TEXTFILE;

- INSERT OVERWRITE TABLE h1b_app2 SELECT regexp_replace (s_no, "\t", ""), regexp_replace (case_status, "\t", ""), regexp_replace (employer_name,

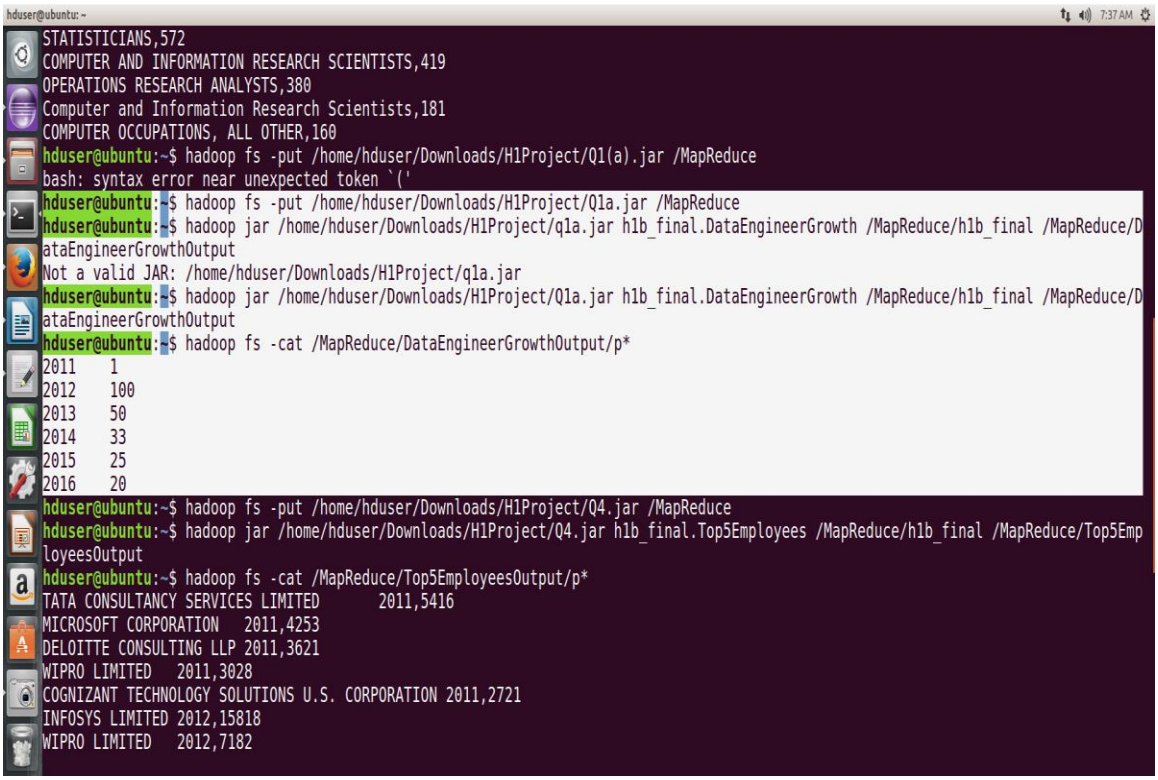
"\t", ""), regexp_replace (soc_name, "\t", ""), regexp_replace (job_title, "\t", ""), regexp_replace (full_time_position, "\t", ""), prevailing_wage, regexp_replace (year, "\t", ""), regexp_replace (worksite, "\t", ""), regexp_replace (longitute, "\t", ""), regexp_replace (latitute, "\t", "") FROM h1b_applications where case_status! = "NA";

- CREATE TABLE h1b_final (s_no int, case_status string, employer_name string, soc_name string, job_title string, full_time_position string, prevailing_wage bigint, year string, worksite string, longitute double, latitute double)
  row format delimited
  fields terminated by '\t'
  STORED AS TEXTFILE;

- INSERT OVERWRITE TABLE h1b_final SELECT s_no,
  case when trim(case_status) = "PENDING QUALITY AND COMPLIANCE REVIEW -
  UNASSIGNED" then "DENIED"
  when trim(case_status) = "REJECTED" then "DENIED"
  when trim(case_status) = "INVALIDATED" then "DENIED"
  else case_status end,
  employer_name, soc_name, job_title, full_time_position,
  case when prevailing_wage is null then 100000
  else prevailing_wage end, year, worksite, longitute, latitute FROM h1b_app2;

1 a) Is the number of petitions with Data Engineer job title increasing over time? (MapReduce)

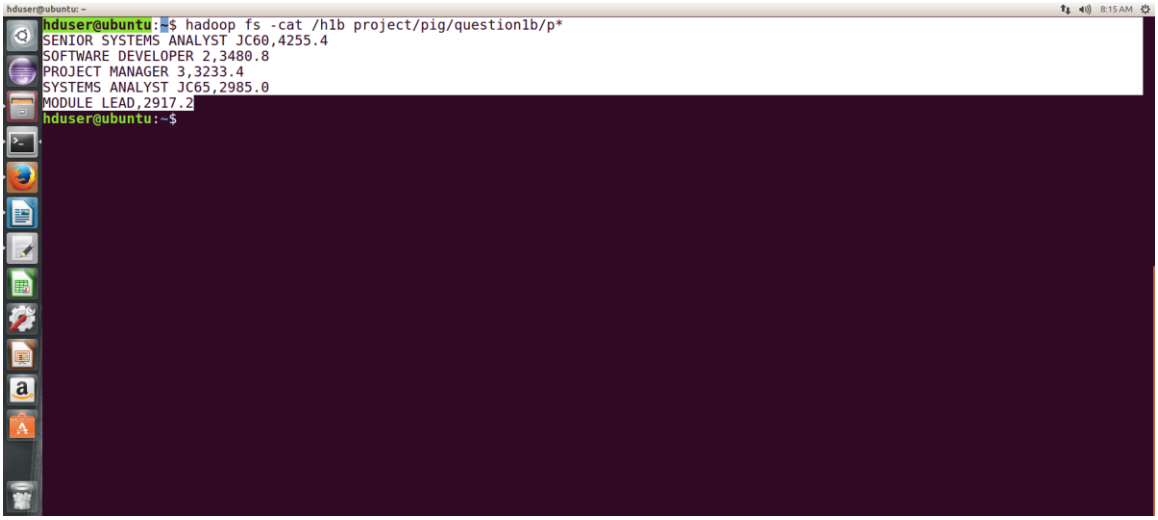
Solution:



---

1 b) Find top 5 job titles who are having highest average growth in applications. [ALL] (PIG)

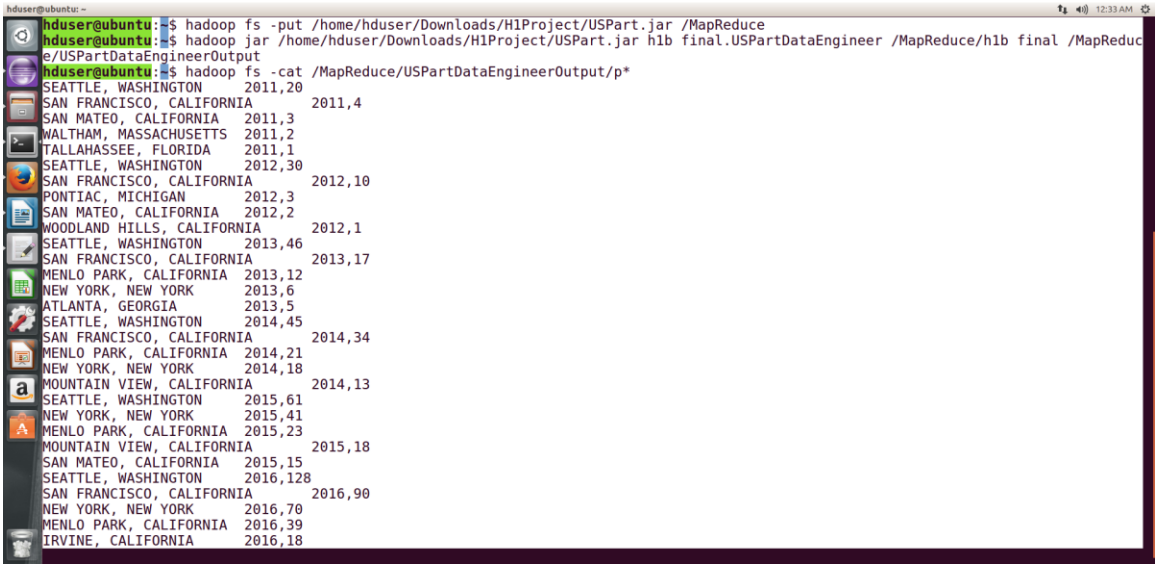
Solution:



---

2 a) Which part of the US has the most Data Engineer jobs for each year? (MapReduce)

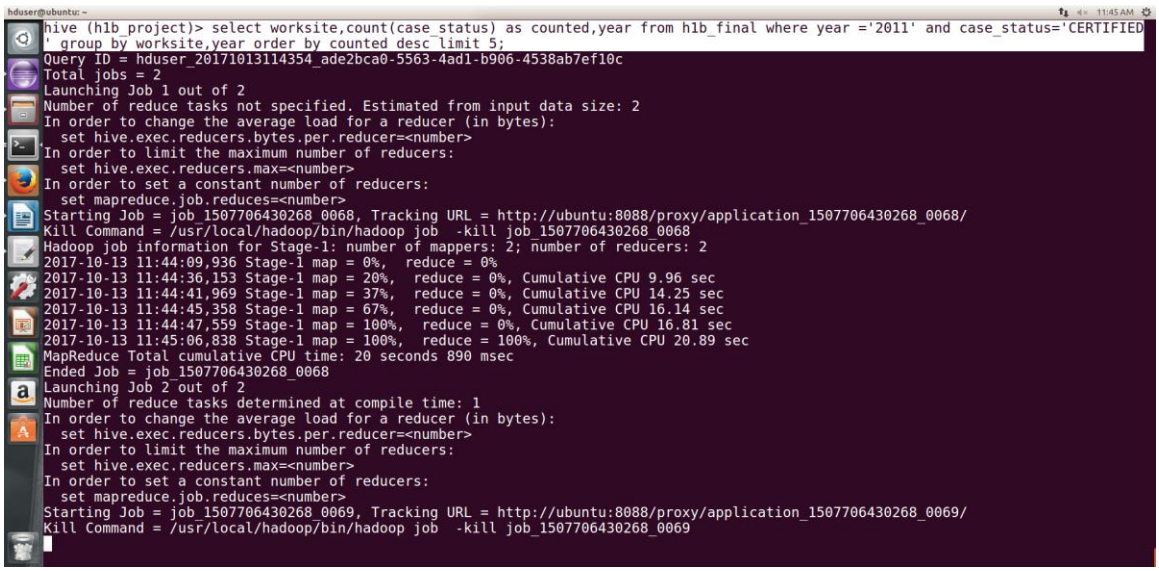
Solution:

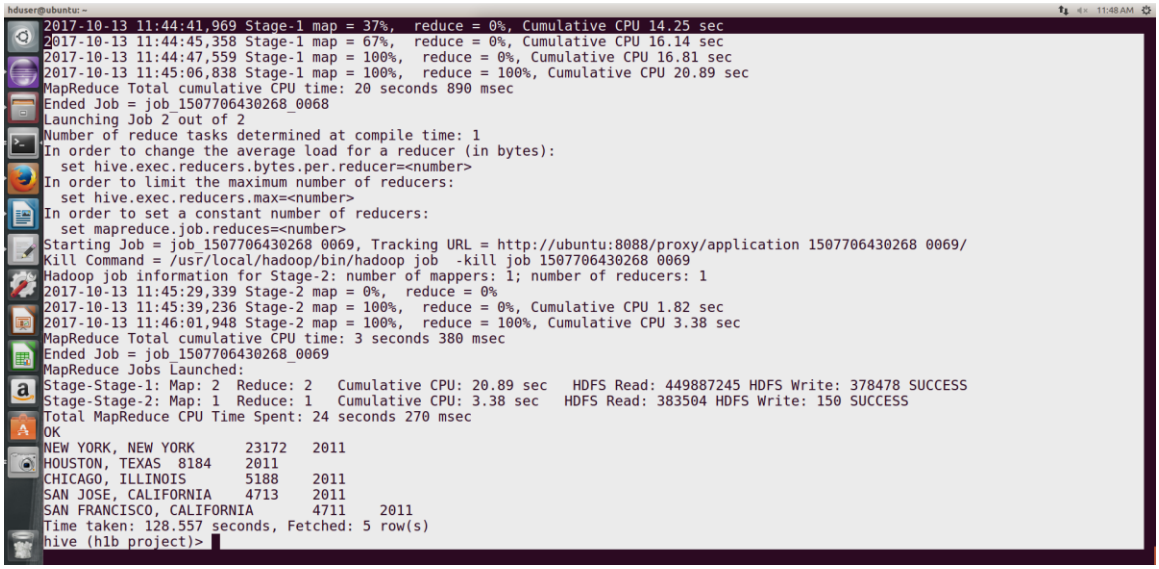


2 b) find top 5 locations in the US who have got certified visa for each year. [certified] (HIVE)

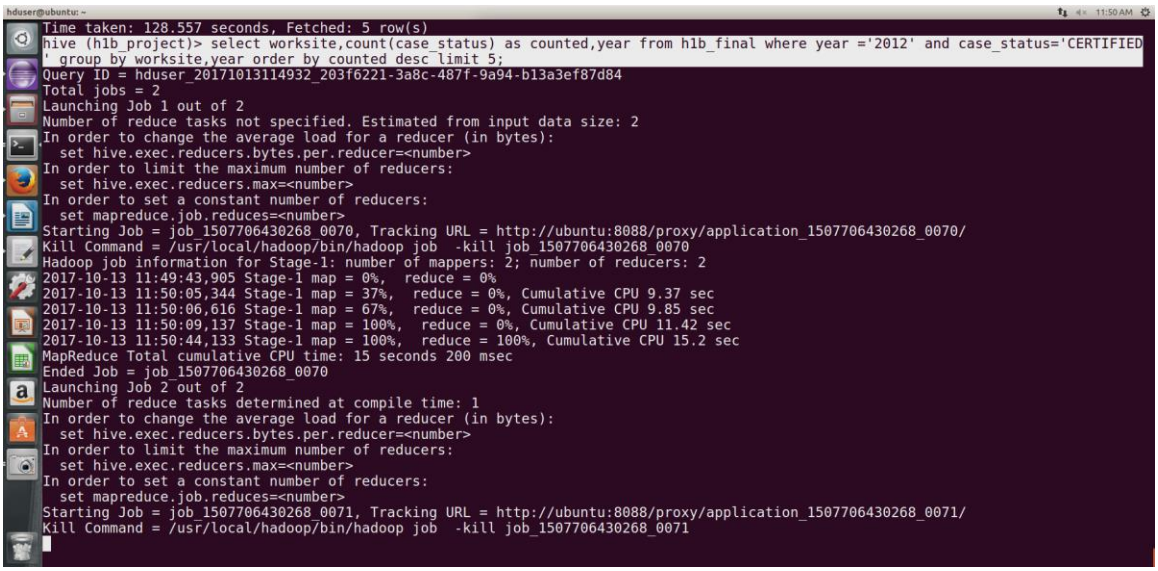
Solution:

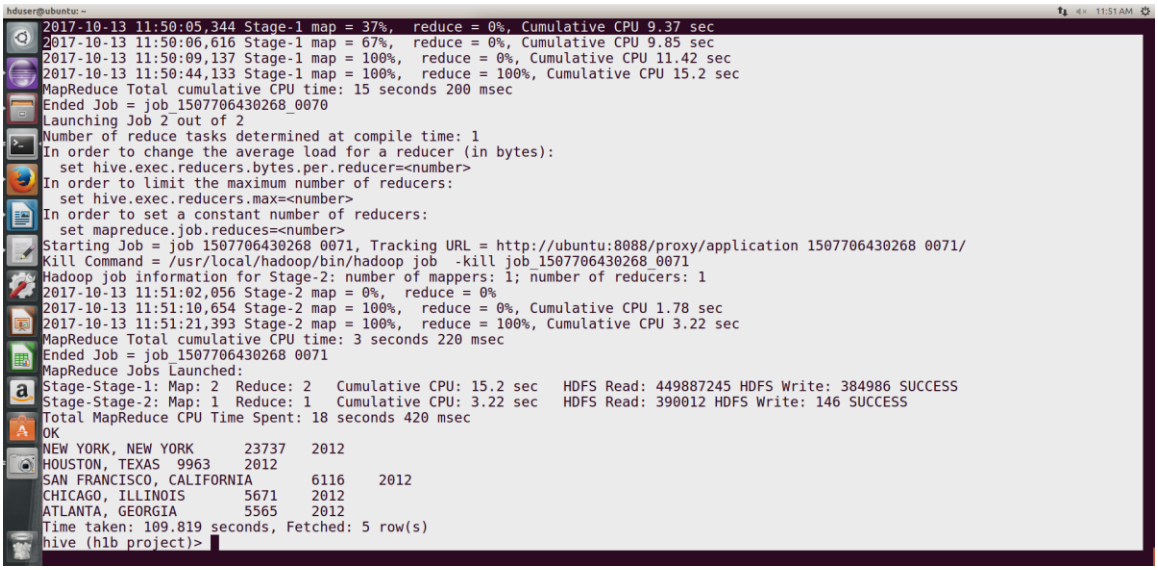
Select worksite, count (case_status) as counted, year from h1b_final where year ='2011' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;



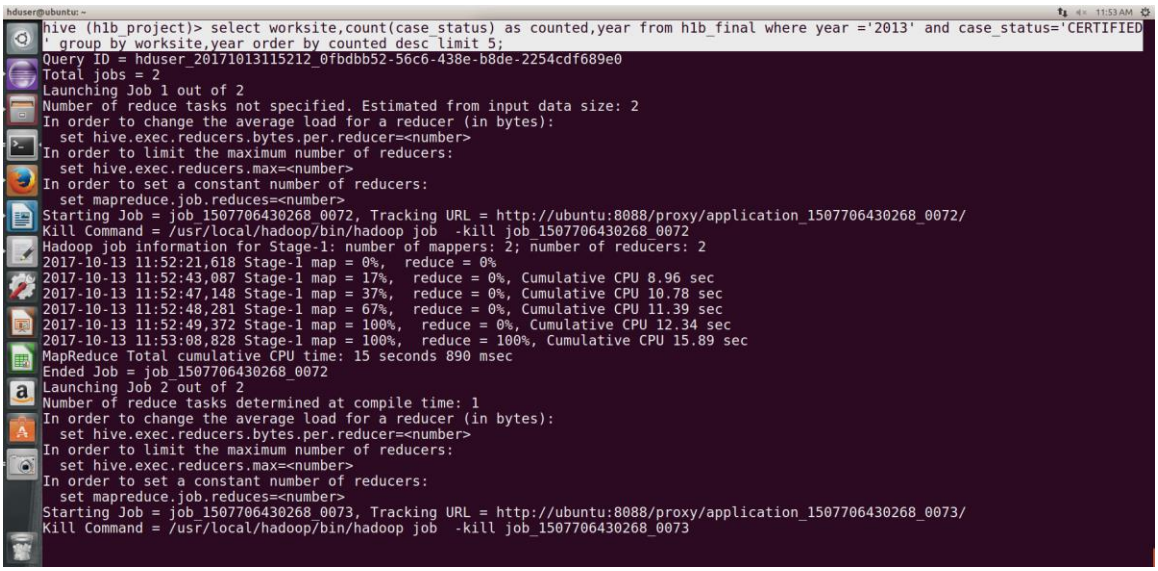


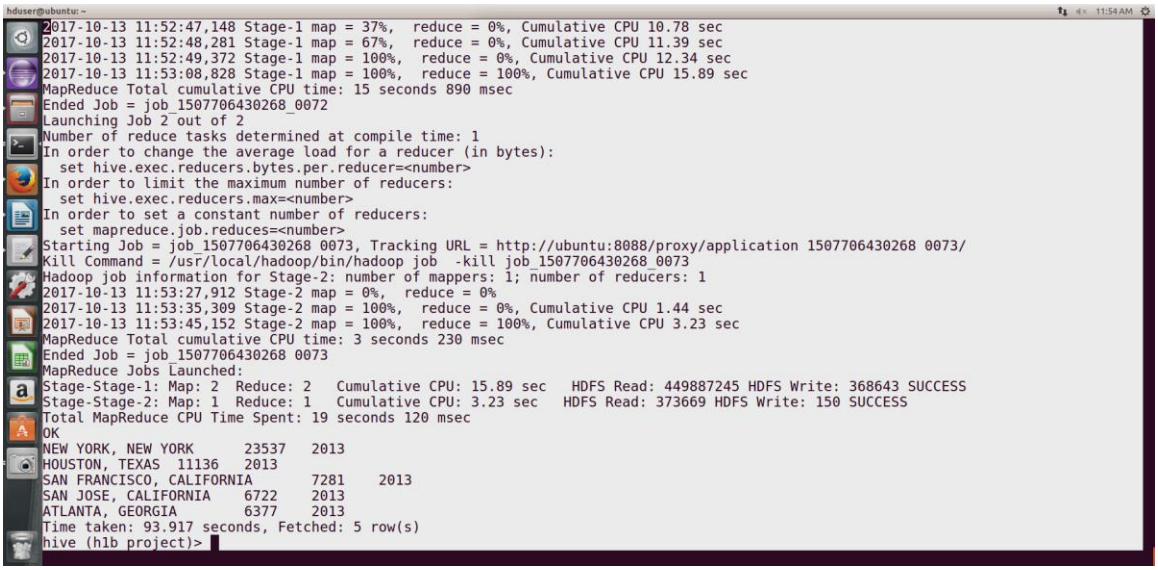
select worksite, count(case_status) as counted, year from h1b_final where year ='2012' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;



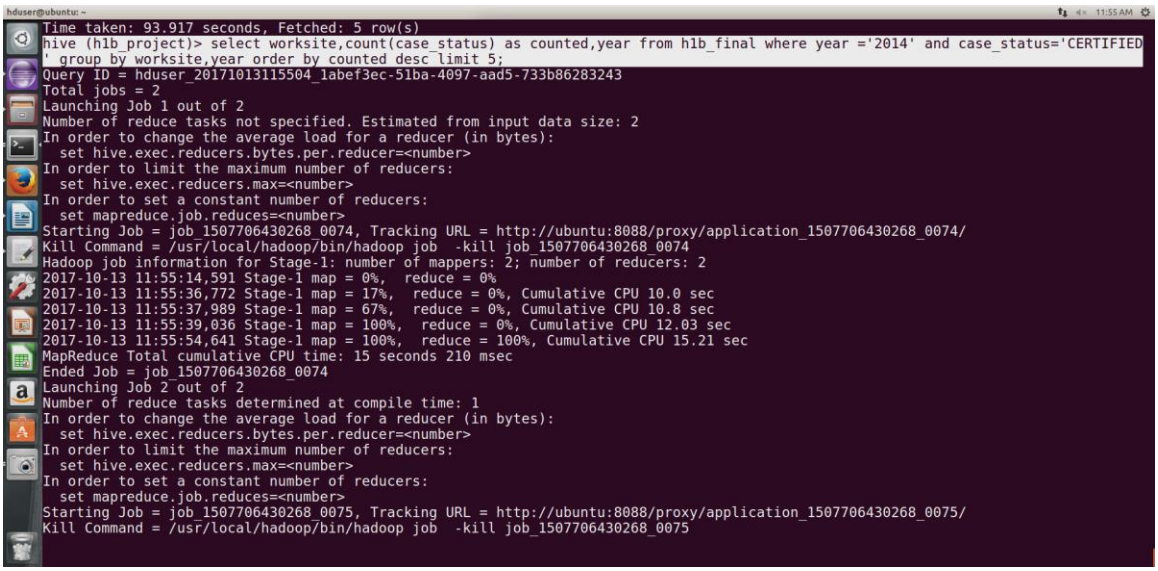


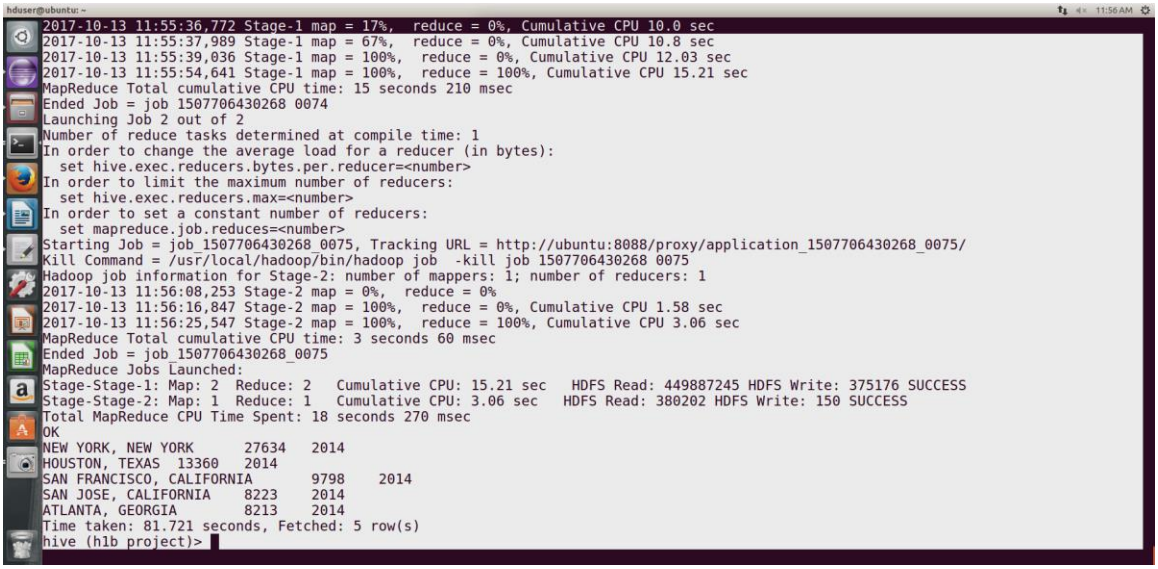
select worksite, count(case_status) as counted, year from h1b_final where year ='2013' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;



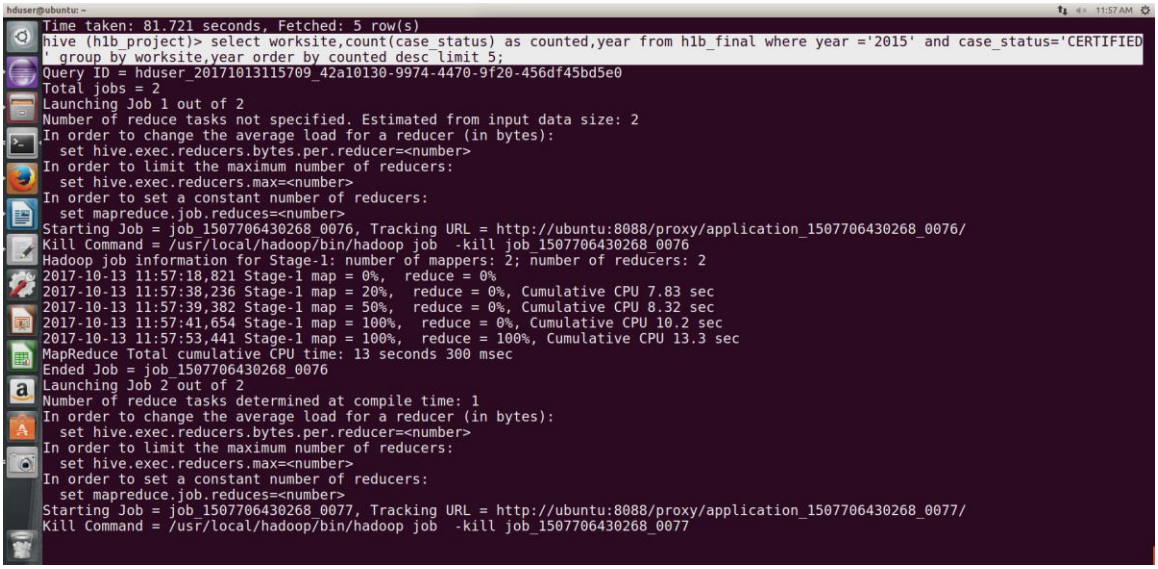


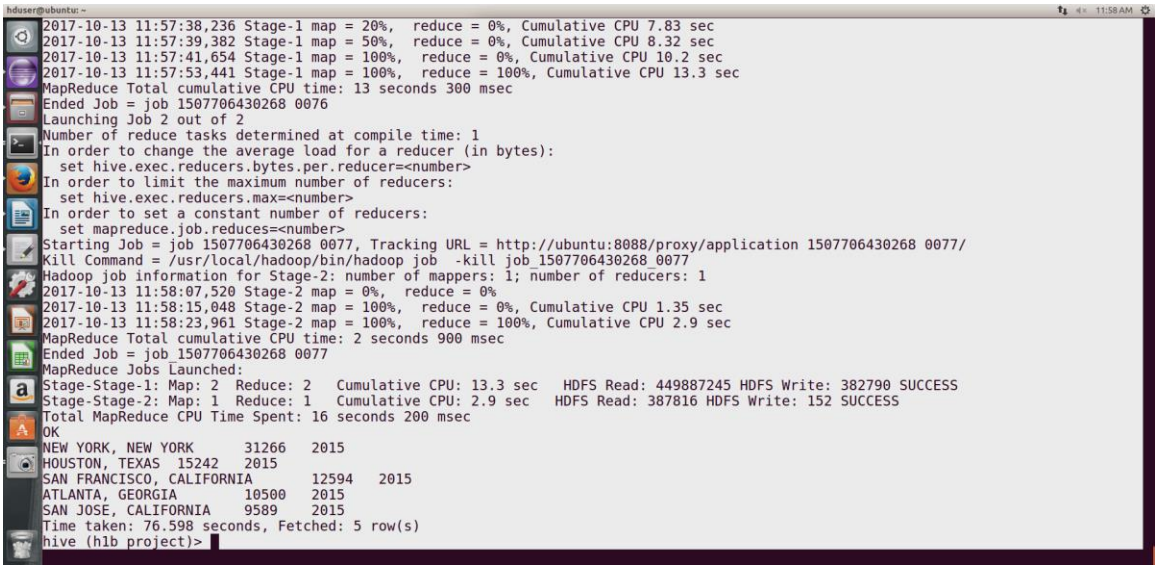
select worksite, count(case_status) as counted, year from h1b_final where year ='2014' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;





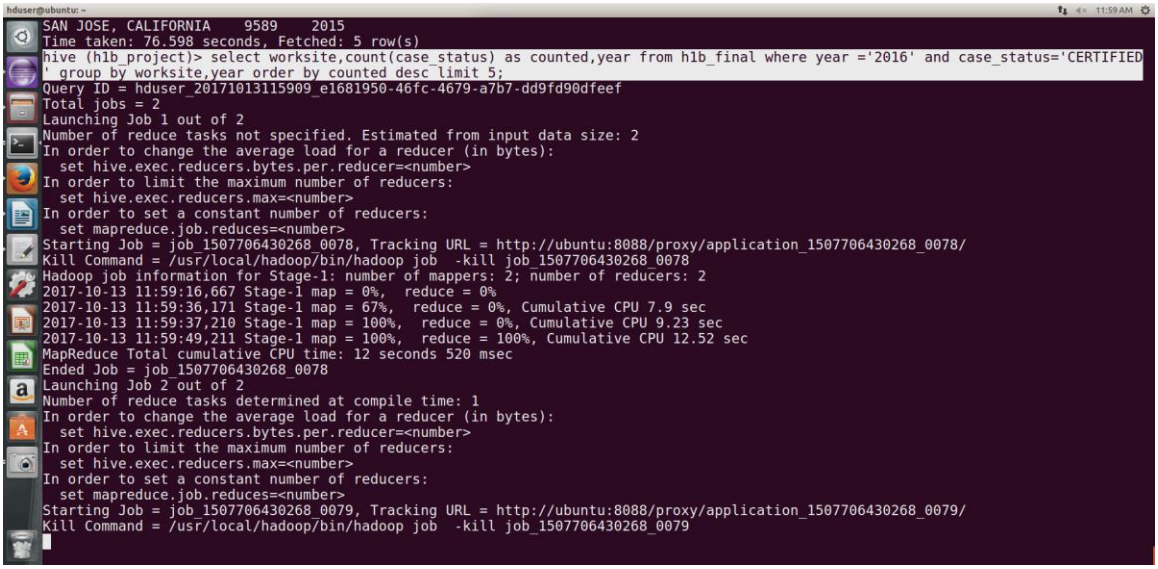
select worksite, count(case_status) as counted, year from h1b_final where year ='2015' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;

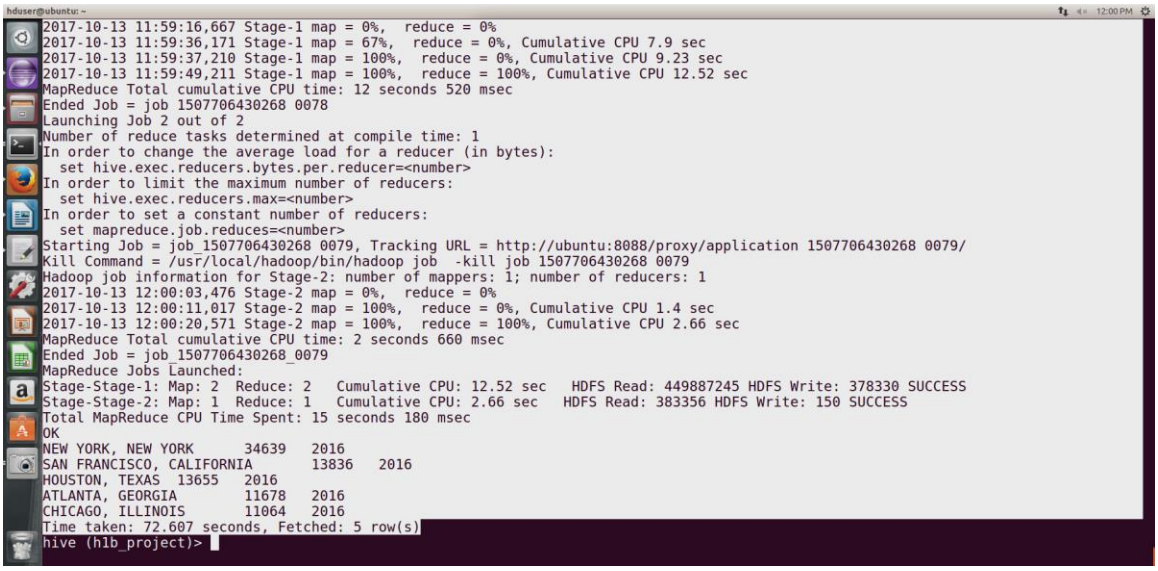

Time taken: 81.721 seconds, Fetched: 5 row(s)
hive (h1b_project)> select worksite,count(case_status) as counted,year from h1b_final where year ='2015' and case_status='CERTIFIED
' group by worksite,year order by counted desc limit 5;
Query ID = hduser_20171013115709_42a10130-9974-4470-9f20-456df45bd5e0
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0076, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0076/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0076
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 11:57:18,821 Stage-1 map = 0%, reduce = 0%
2017-10-13 11:57:38,236 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 7.83 sec
2017-10-13 11:57:39,382 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 8.32 sec
2017-10-13 11:57:41,654 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 10.2 sec
2017-10-13 11:57:53,441 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.3 sec
MapReduce Total cumulative CPU time: 13 seconds 300 msec
Ended Job = job_1507706430268_0076
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0077, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0077/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0077


2017-10-13 11:57:38,236 Stage-1 map = 20%, reduce = 0%, Cumulative CPU 7.83 sec
2017-10-13 11:57:39,382 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 8.32 sec
2017-10-13 11:57:41,654 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 10.2 sec
2017-10-13 11:57:53,441 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 13.3 sec
MapReduce Total cumulative CPU time: 13 seconds 300 msec
Ended Job = job_1507706430268_0076
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0077, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0077/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0077
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 11:58:07,520 Stage-2 map = 0%, reduce = 0%
2017-10-13 11:58:15,048 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.35 sec
2017-10-13 11:58:23,961 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 2.9 sec
MapReduce Total cumulative CPU time: 2 seconds 900 msec
Ended Job = job_1507706430268_0077
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 13.3 sec HDFS Read: 449887245 HDFS Write: 382790 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 2.9 sec HDFS Read: 387816 HDFS Write: 152 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 200 msec
OK
NEW YORK, NEW YORK 31266 2015
HOUSTON, TEXAS 15242 2015
SAN FRANCISCO, CALIFORNIA 12594 2015
ATLANTA, GEORGIA 10500 2015
SAN JOSE, CALIFORNIA 9589 2015
Time taken: 76.598 seconds, Fetched: 5 row(s)
hive (h1b_project)>

S

select worksite, count(case_status) as counted, year from h1b_final where year ='2016' and case_status='CERTIFIED' group by worksite, year order by counted desc limit 5;

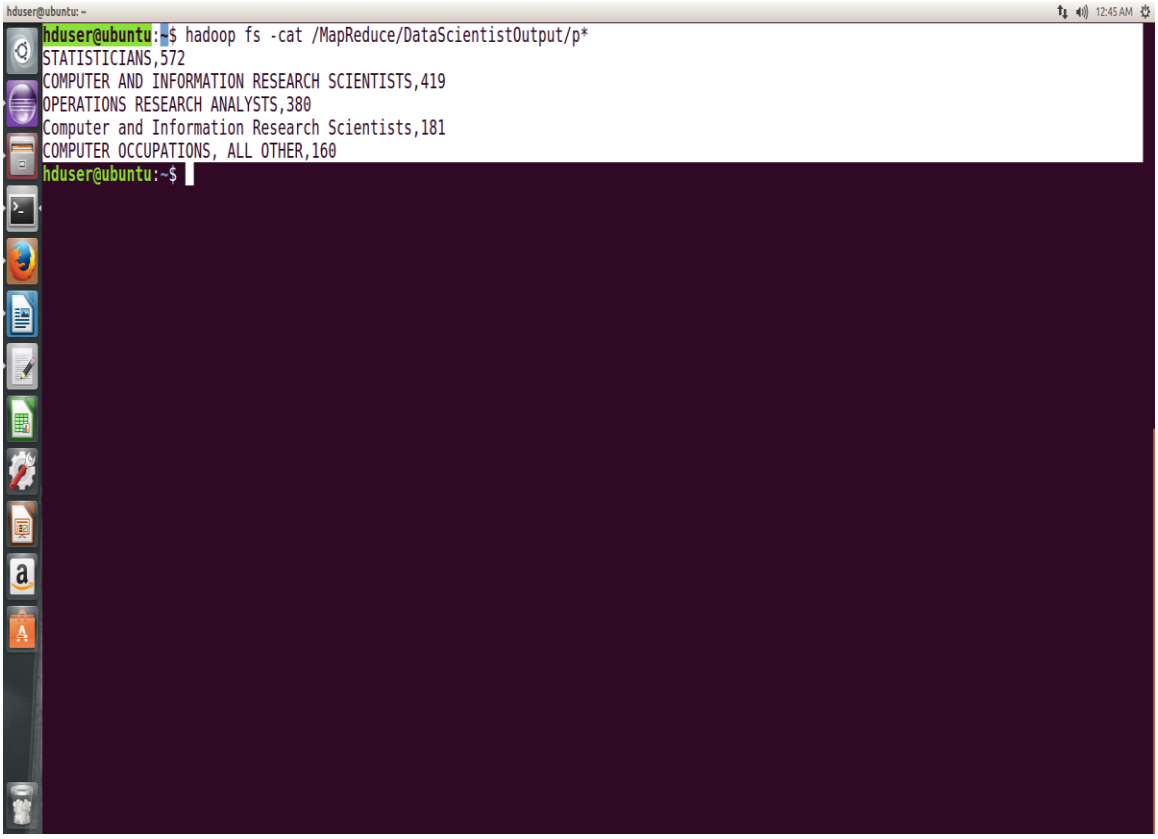




3)Which industry(SOC_NAME) has the most number of Data Scientist positions?

[certified] (MapReduce)

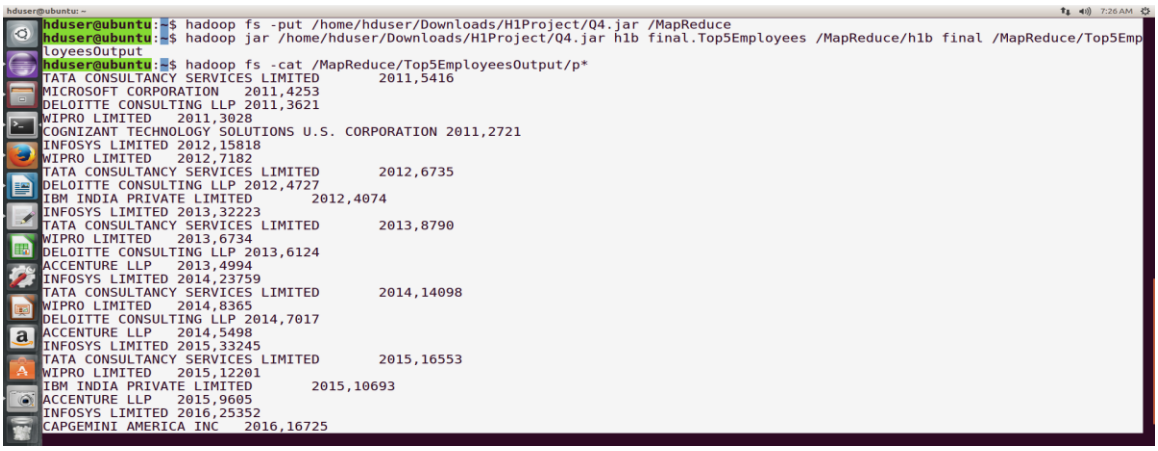
Solution:



---

4)Which top 5 employers file the most petitions each year? - Case Status – ALL (MapReduce)

Solution:

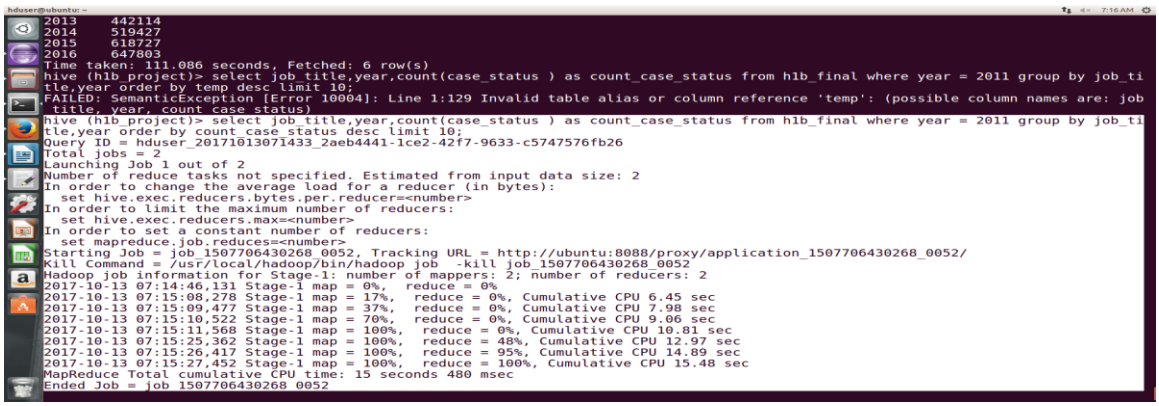


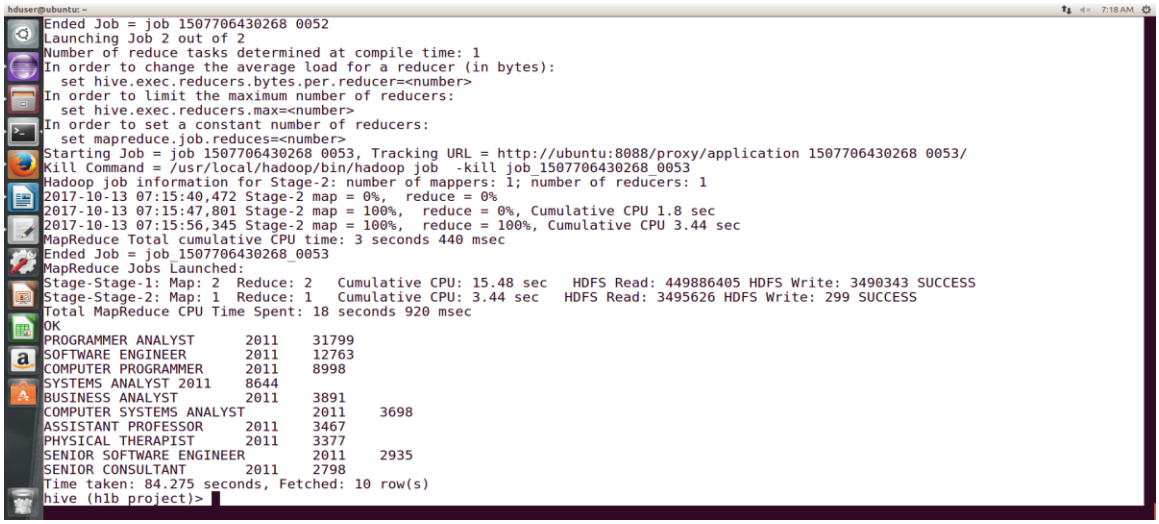
---

5) Find the most popular top 10 job positions for H1B visa applications for each year?
(a) for all the applications (HIVE)

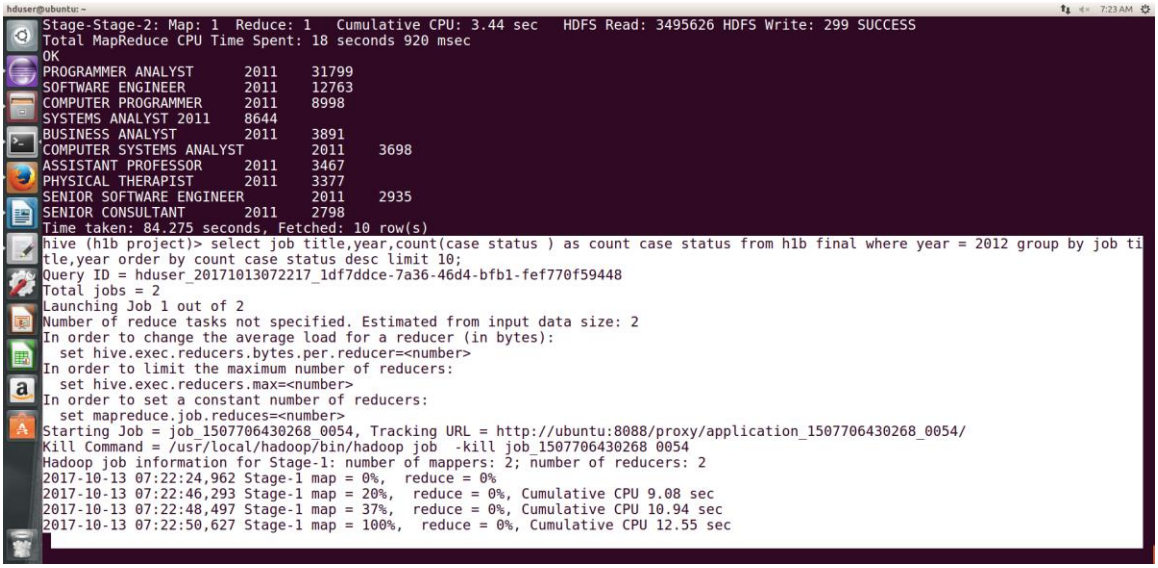
Solution:

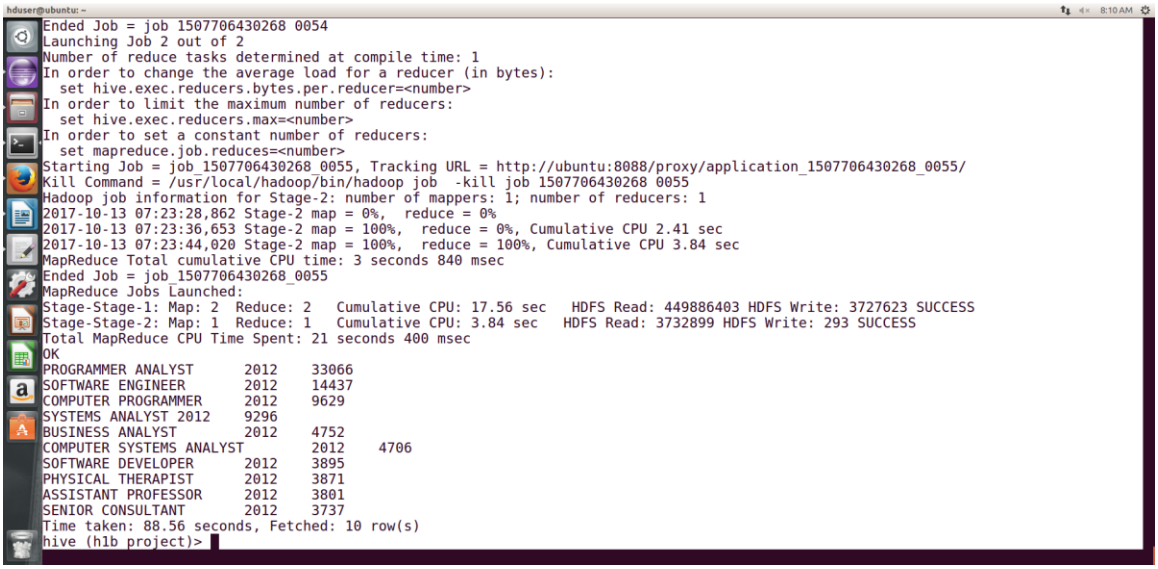
select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2011 group by job_title,year order by count_case_status desc limit 10;



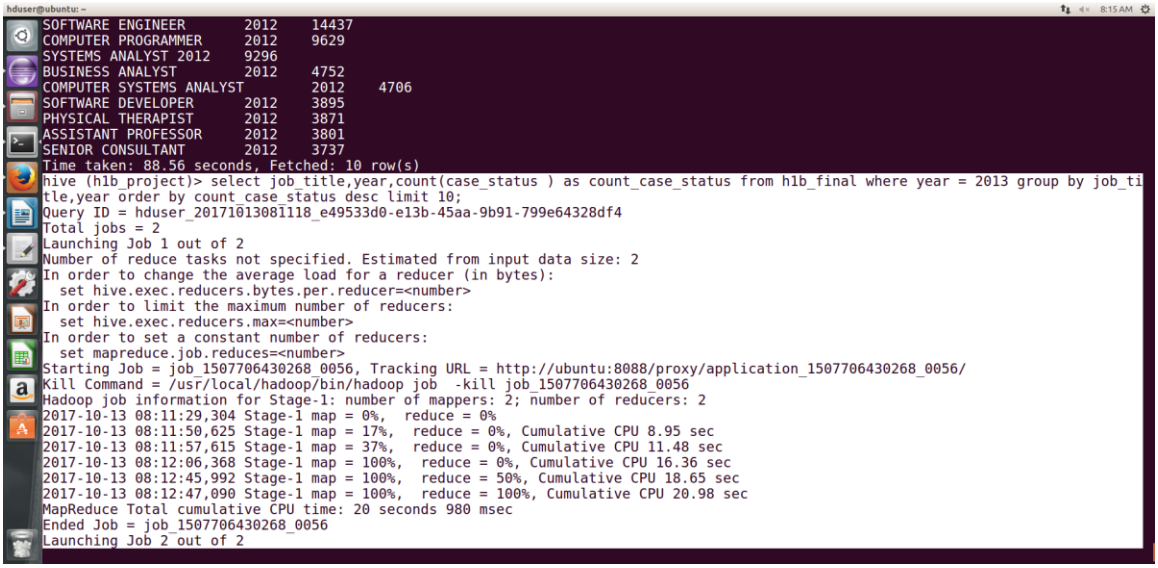


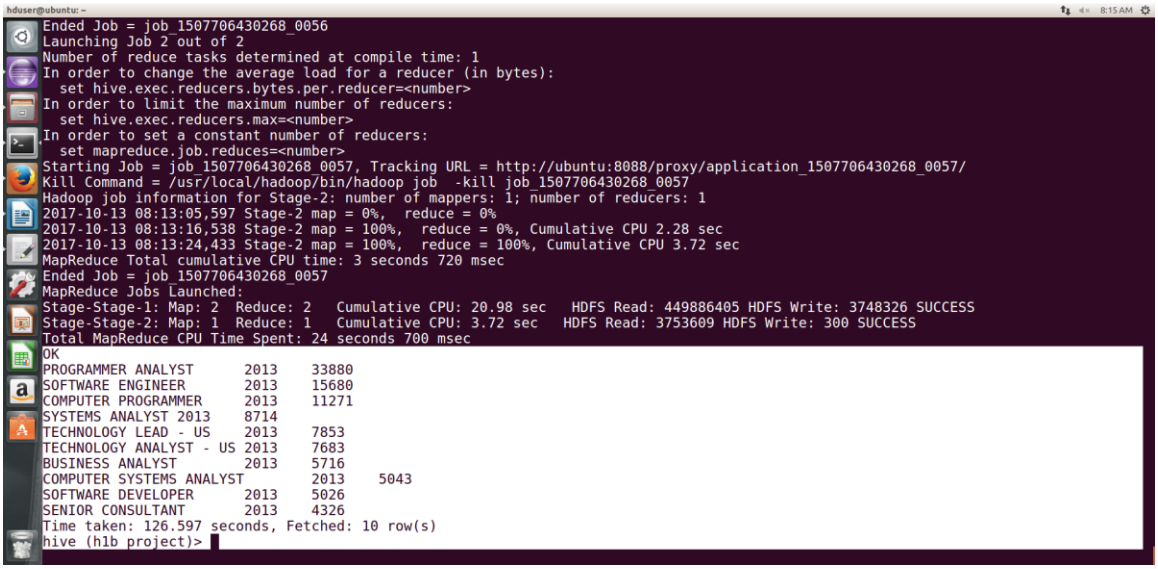
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2012 group by job_title, year order by count_case_status desc limit 10;



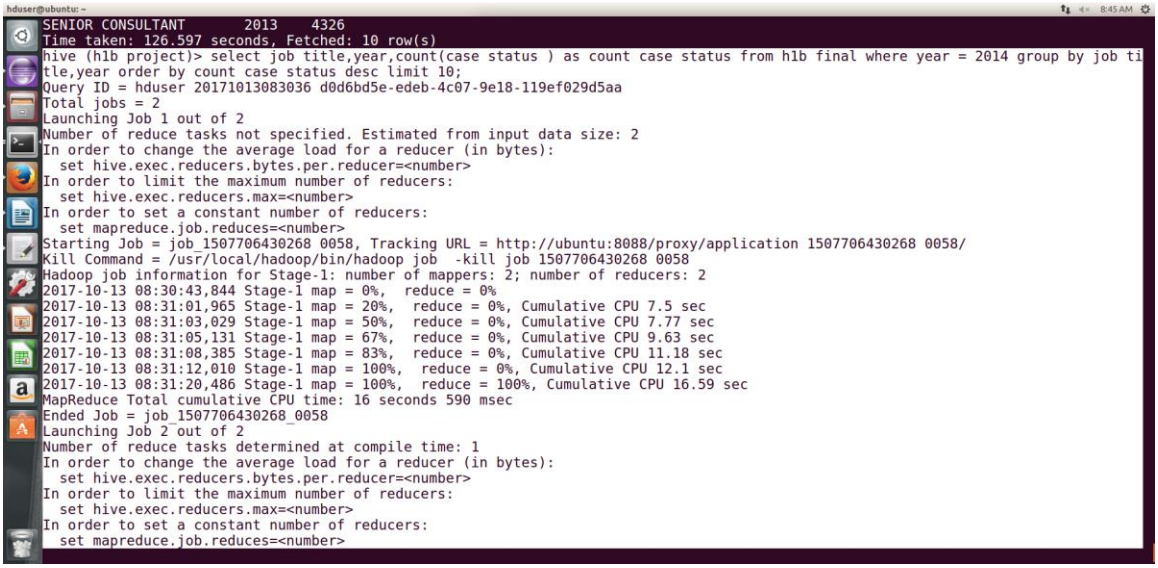


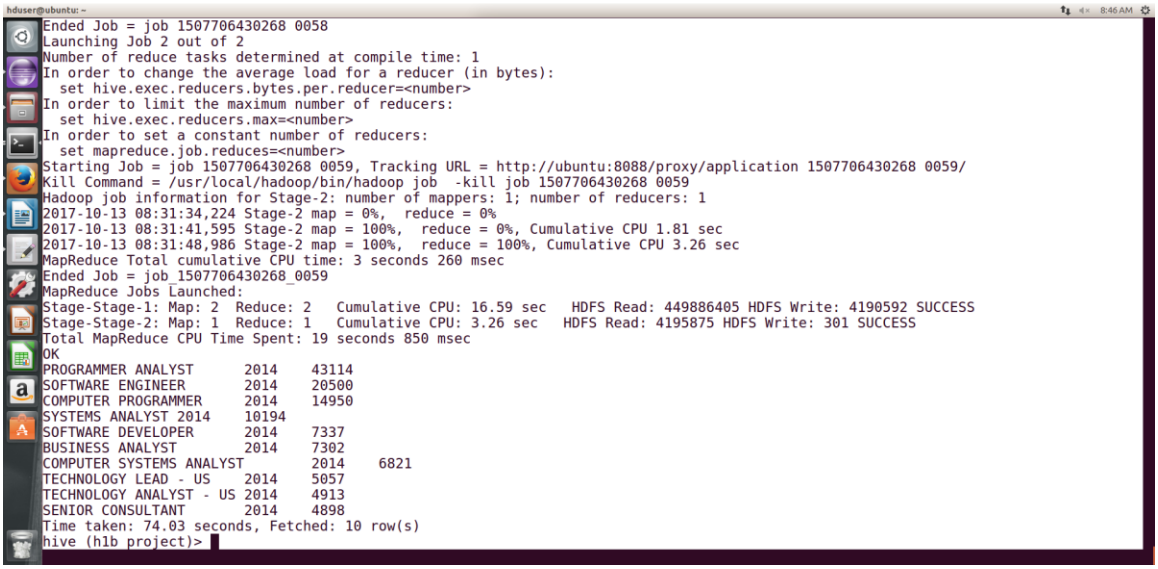
select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2013 group by job_title,year order by count_case_status desc limit 10;



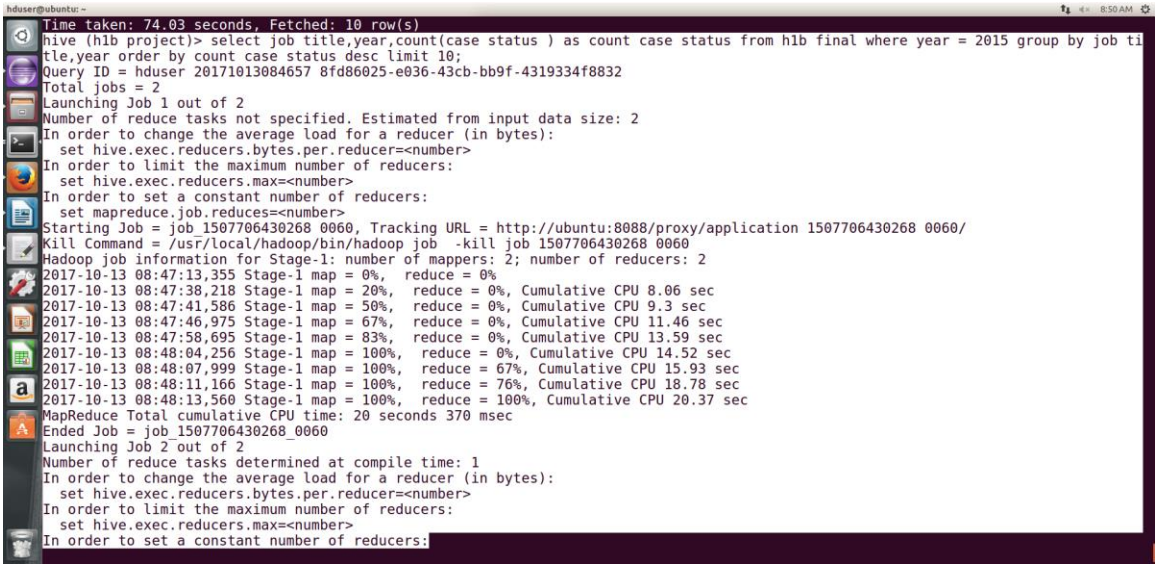


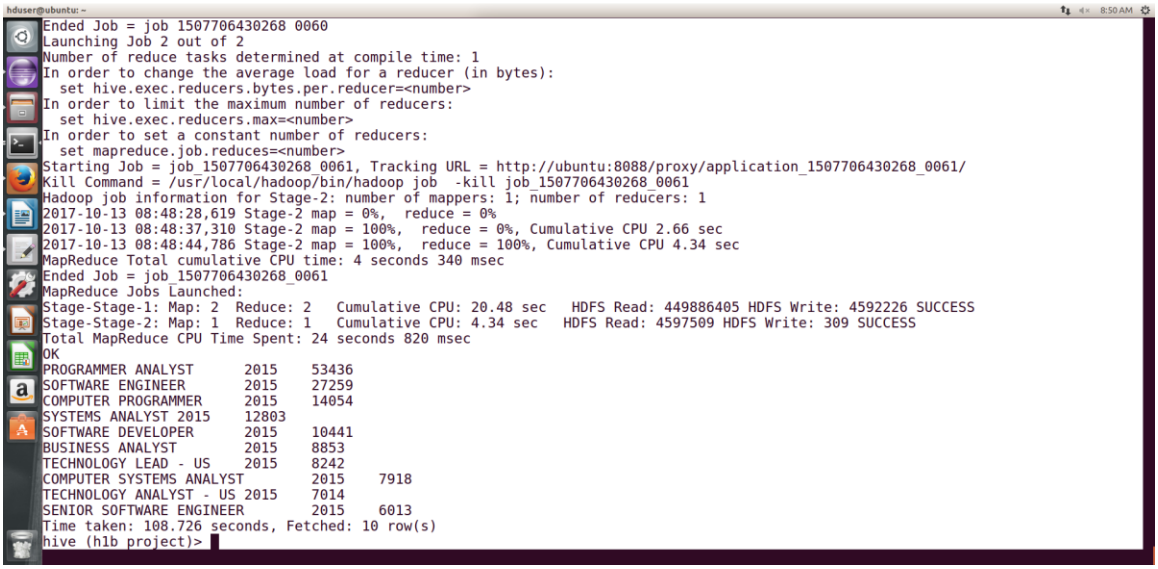
select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2014 group by job_title,year order by count_case_status desc limit 10;



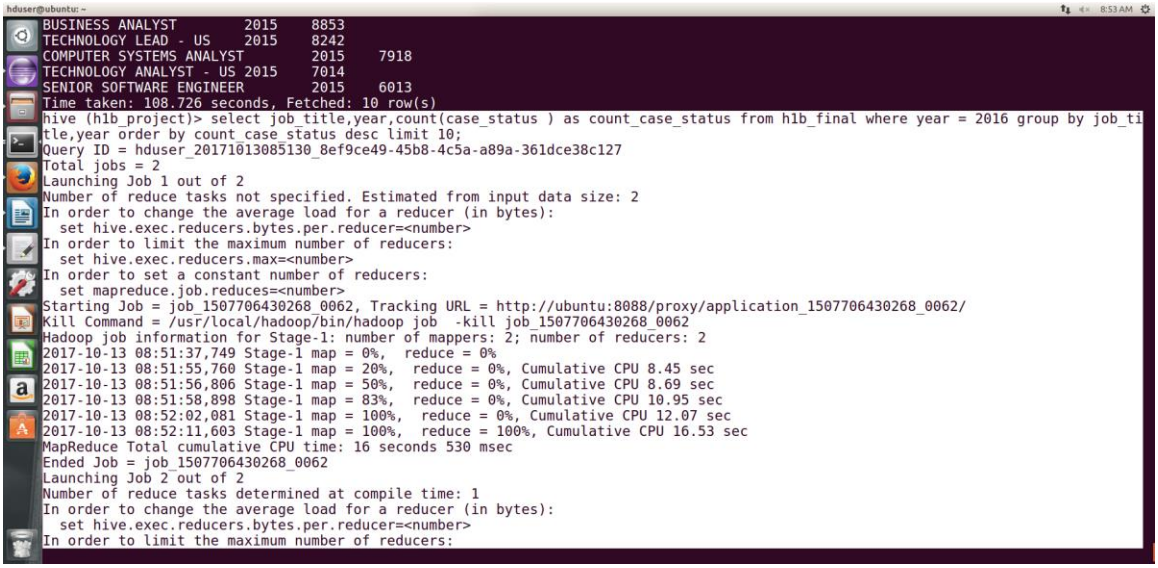


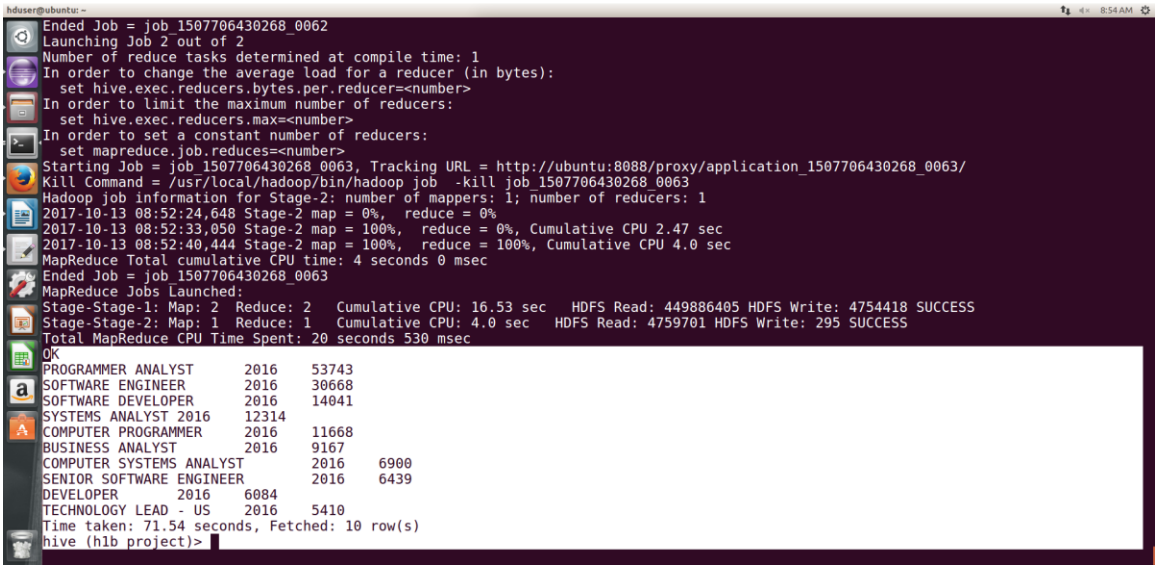
select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2015 group by job_title,year order by count_case_status desc limit 10;





select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2016 group by job_title,year order by count_case_status desc limit 10;



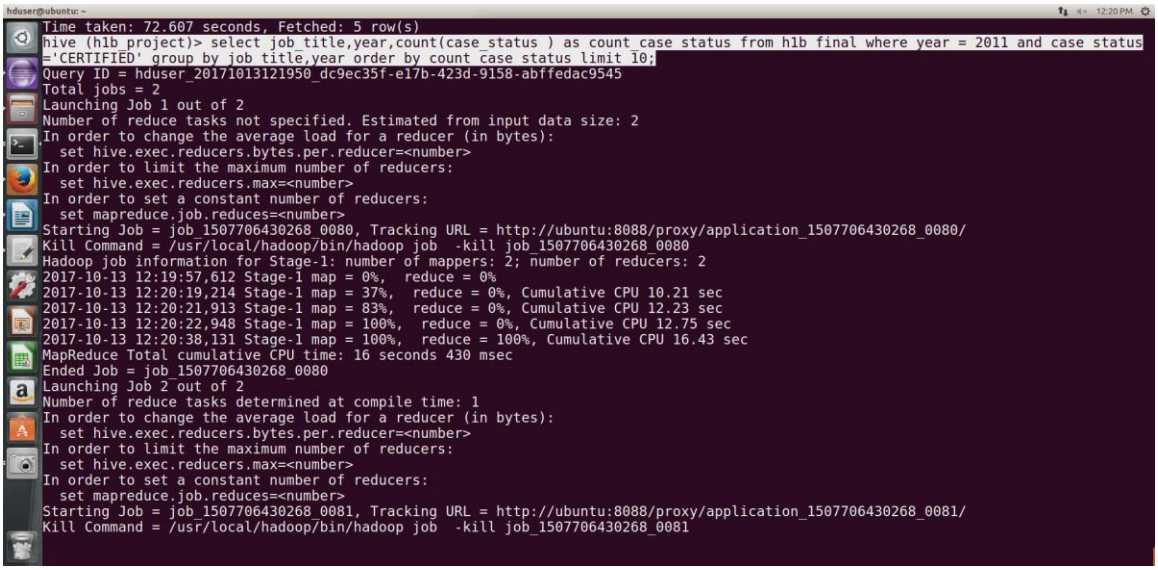


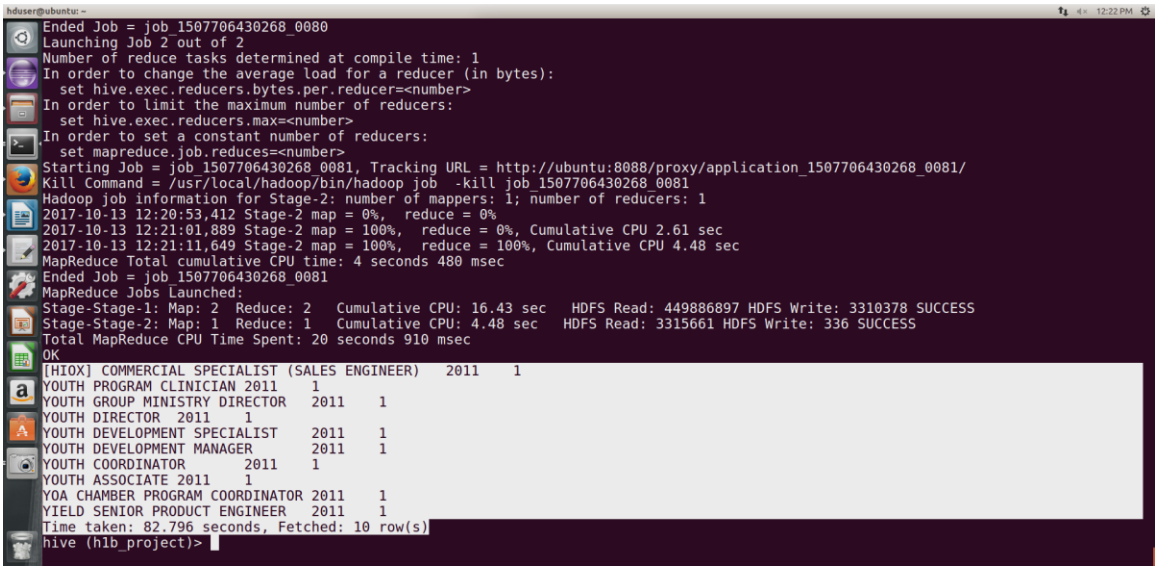
---

5 (b) for only certified applications. (HIVE)

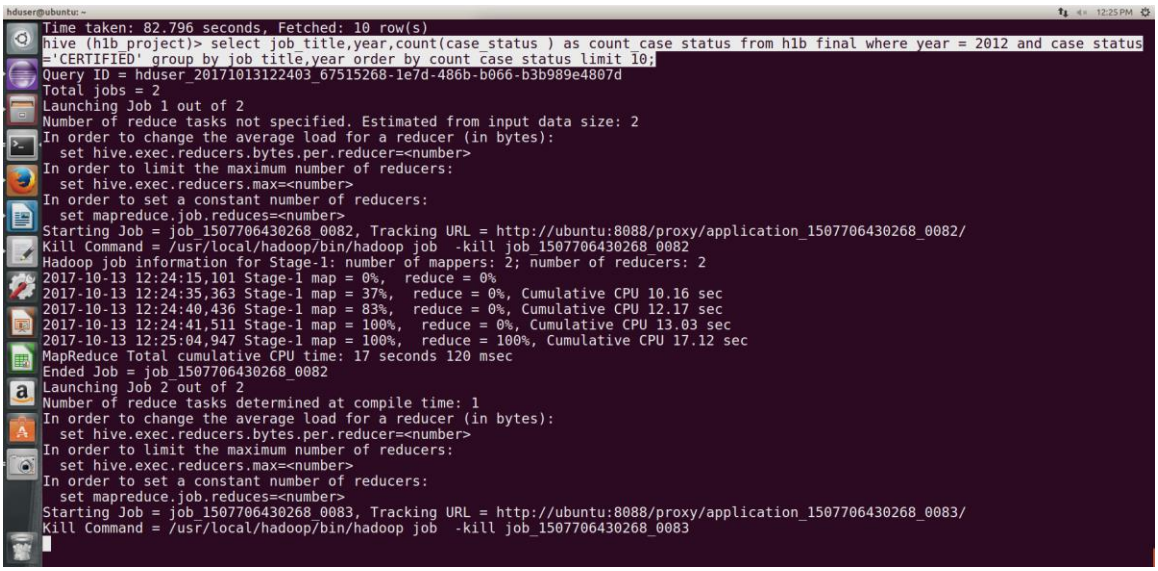
Solution:

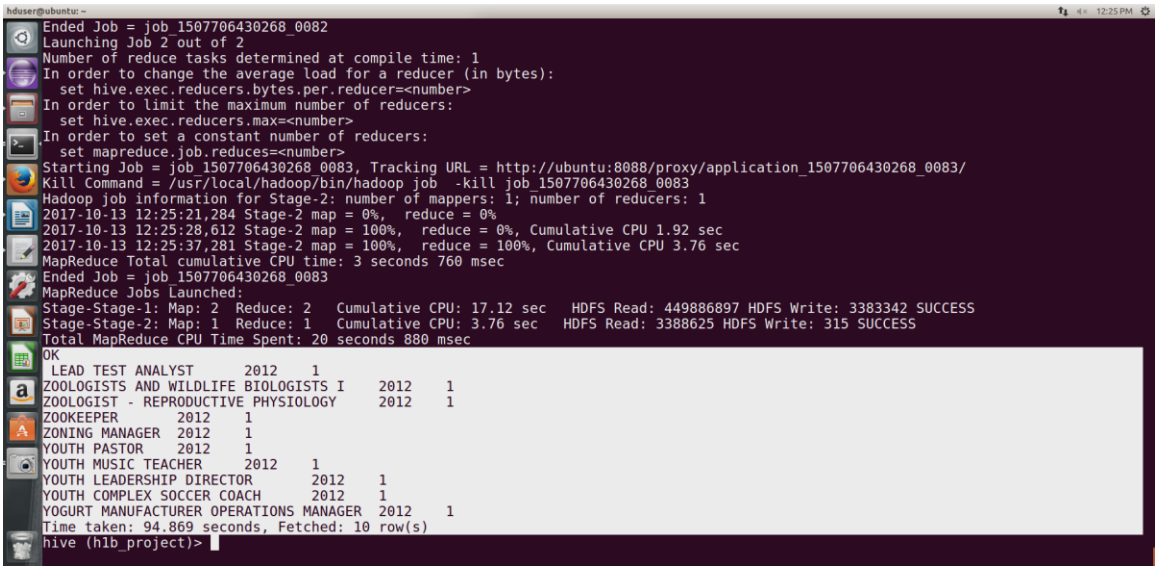
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2011 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;



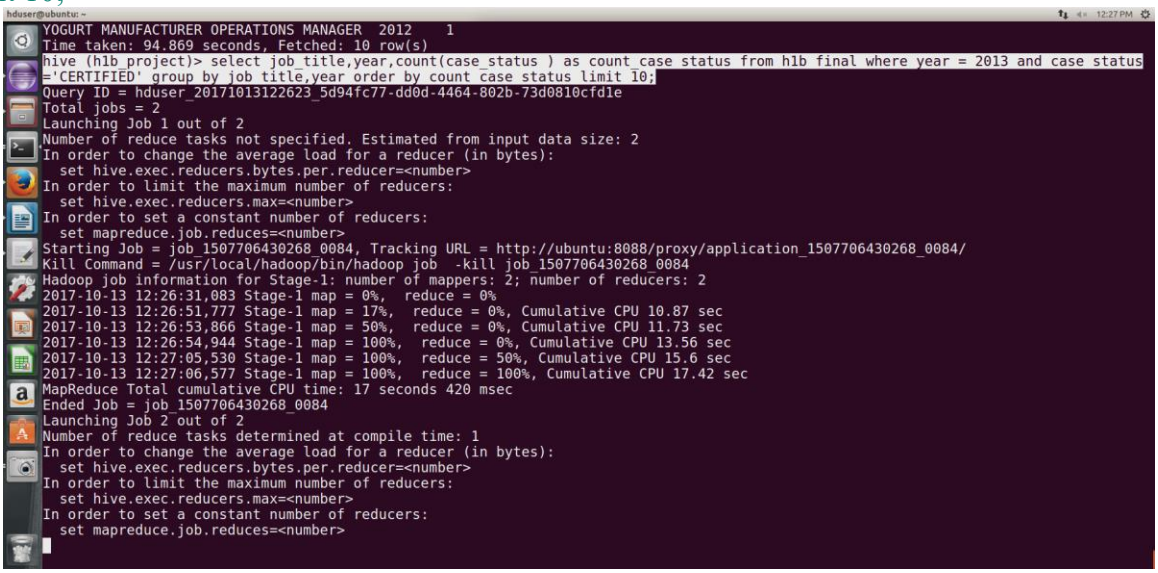


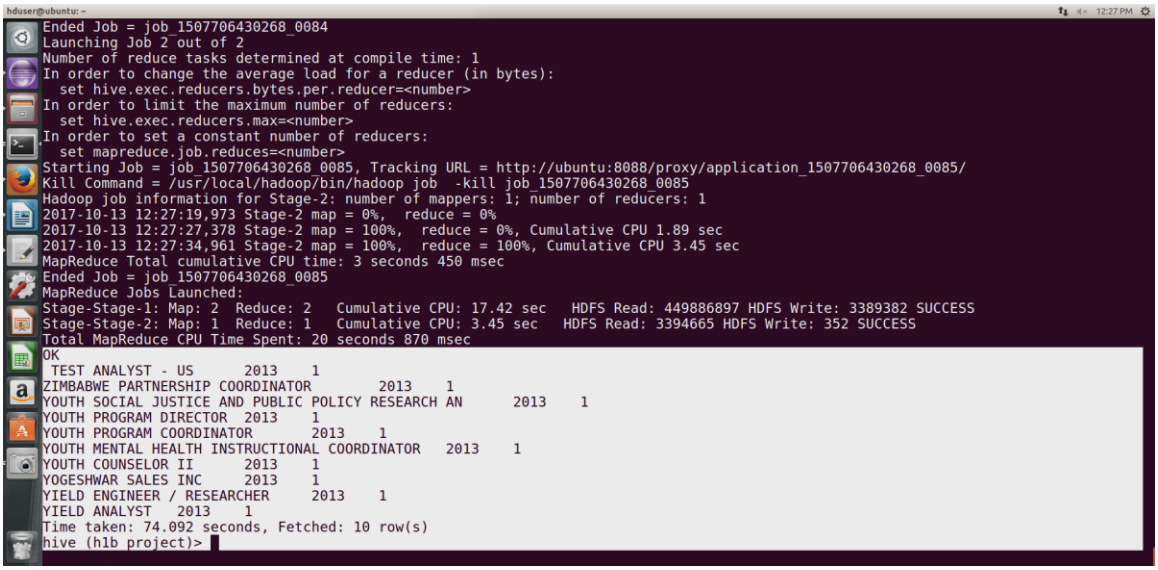
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2012 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;


hduser@ubuntu: ~
12:25 PM
Time taken: 82.796 seconds, Fetched: 10 row(s)
hive (h1b_project)> select job_title,year,count(case_status) as count_case_status from h1b_final where year = 2012 and case_status ='CERTIFIED' group by job_title,year order by count_case_status limit 10;
Query ID = hduser_20171013122403_67515268-1e7d-486b-b066-b3b989e4807d
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 2
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0082, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0082/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0082
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 2
2017-10-13 12:24:15,101 Stage-1 map = 0%, reduce = 0%
2017-10-13 12:24:35,363 Stage-1 map = 37%, reduce = 0%, Cumulative CPU 10.16 sec
2017-10-13 12:24:40,436 Stage-1 map = 83%, reduce = 0%, Cumulative CPU 12.17 sec
2017-10-13 12:24:41,511 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 13.03 sec
2017-10-13 12:25:04,947 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 17.12 sec
MapReduce Total cumulative CPU time: 17 seconds 120 msec
Ended Job = job_1507706430268_0082
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0083, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0083/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0083

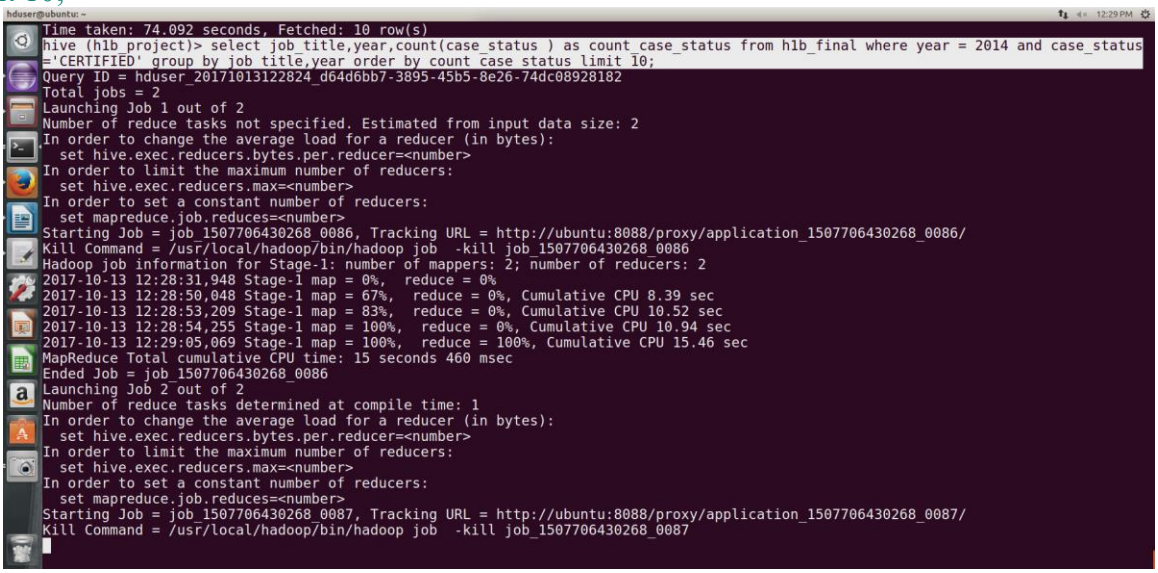

hduser@ubuntu: ~
12:25 PM
Ended Job = job_1507706430268_0082
Launching Job 2 out of 2
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
set mapreduce.job.reduces=<number>
Starting Job = job_1507706430268_0083, Tracking URL = http://ubuntu:8088/proxy/application_1507706430268_0083/
Kill Command = /usr/local/hadoop/bin/hadoop job -kill job_1507706430268_0083
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2017-10-13 12:25:21,284 Stage-2 map = 0%, reduce = 0%
2017-10-13 12:25:28,612 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 1.92 sec
2017-10-13 12:25:37,281 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 3.76 sec
MapReduce Total cumulative CPU time: 3 seconds 760 msec
Ended Job = job_1507706430268_0083
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 2 Cumulative CPU: 17.12 sec HDFS Read: 449886897 HDFS Write: 3383342 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 3.76 sec HDFS Read: 3388625 HDFS Write: 315 SUCCESS
Total MapReduce CPU Time Spent: 20 seconds 880 msec
OK
LEAD TEST ANALYST 2012 1
ZOOLOGISTS AND WILDLIFE BIOLOGISTS I 2012 1
ZOOLOGIST - REPRODUCTIVE PHYSIOLOGY 2012 1
ZOOKEEPER 2012 1
ZONING MANAGER 2012 1
YOUTH PASTOR 2012 1
YOUTH MUSIC TEACHER 2012 1
YOUTH LEADERSHIP DIRECTOR 2012 1
YOUTH COMPLEX SOCCER COACH 2012 1
YOGURT MANUFACTURER OPERATIONS MANAGER 2012 1
Time taken: 94.869 seconds, Fetched: 10 row(s)
hive (h1b_project)>

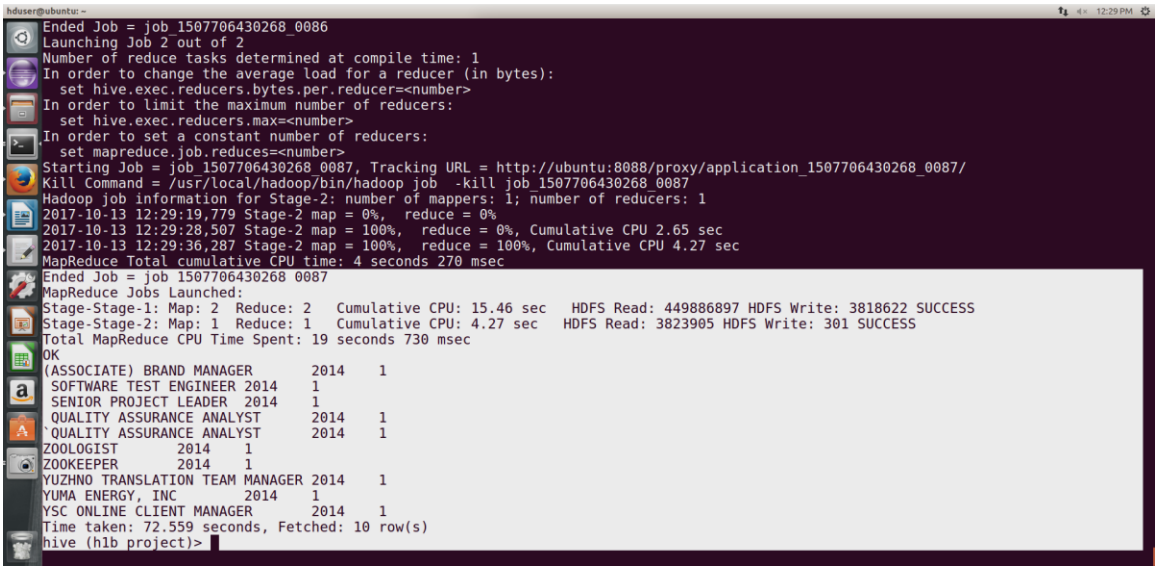
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2013 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;



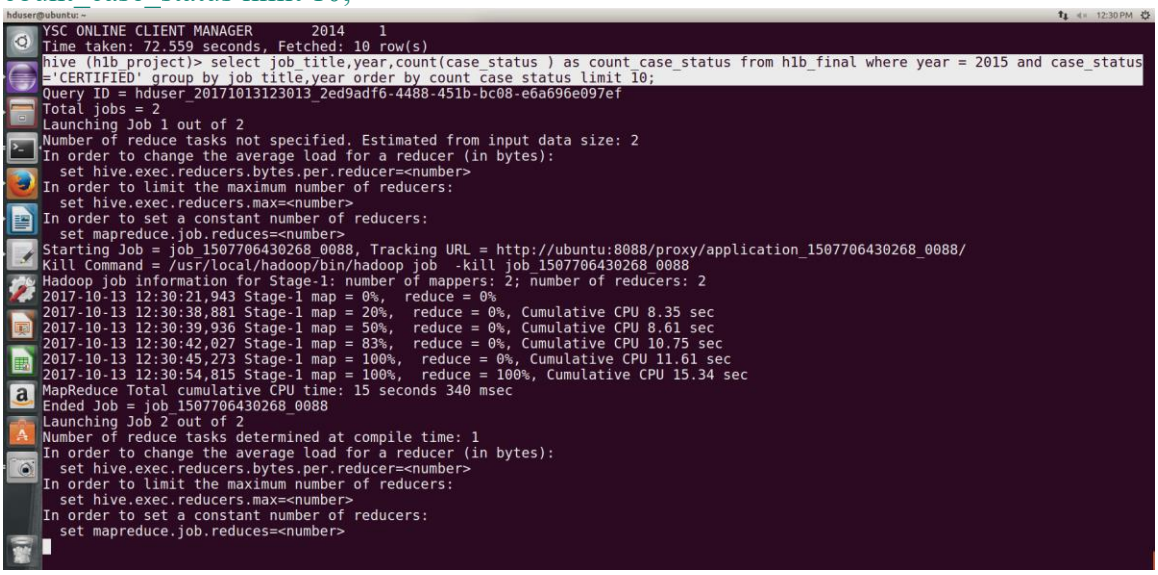


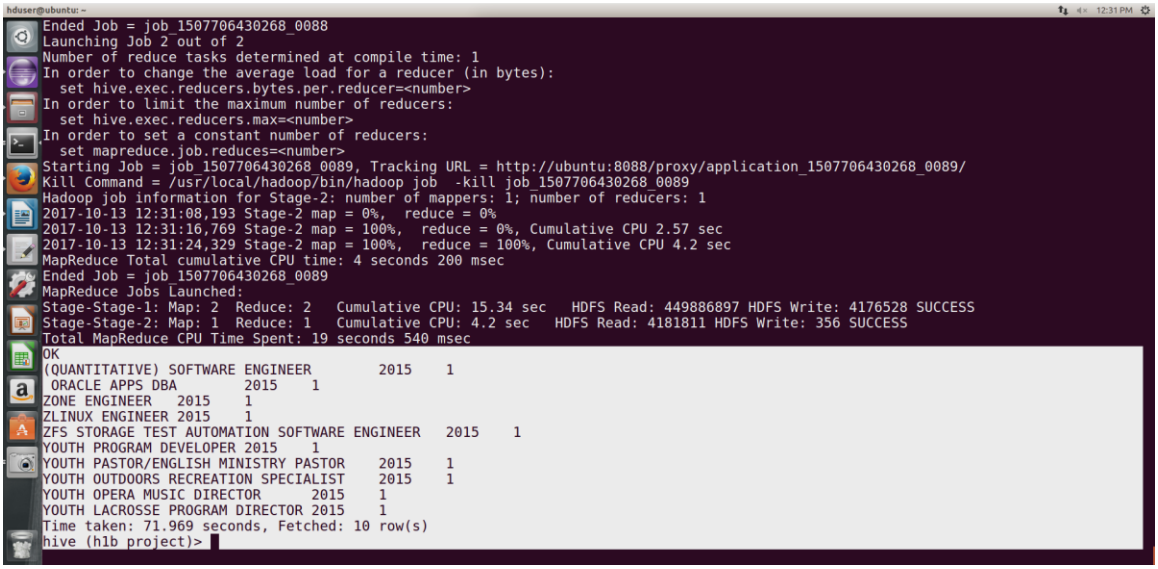
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2014 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;



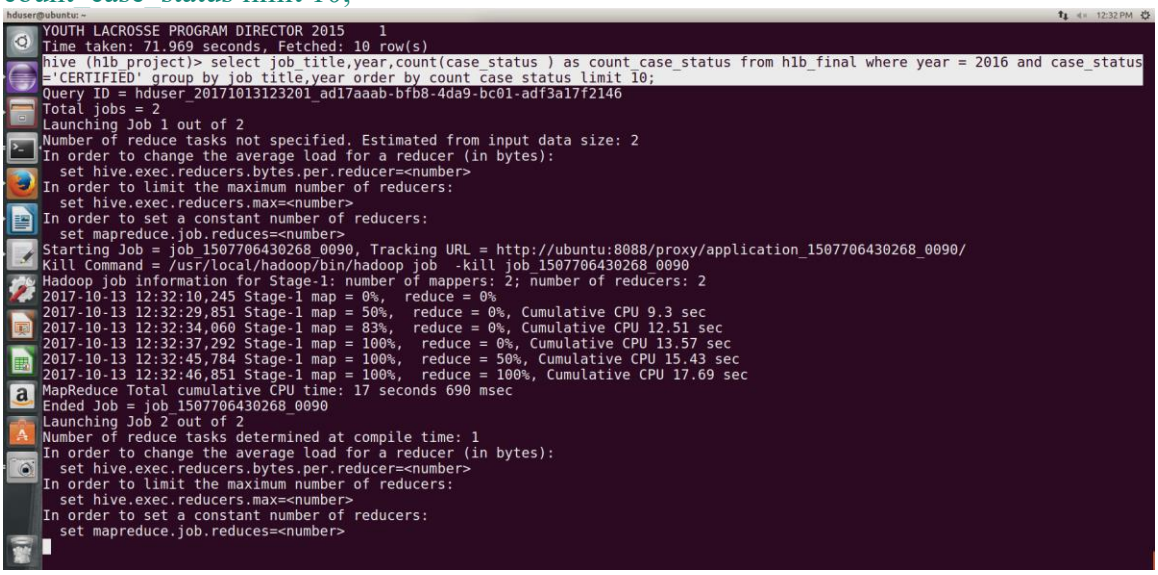


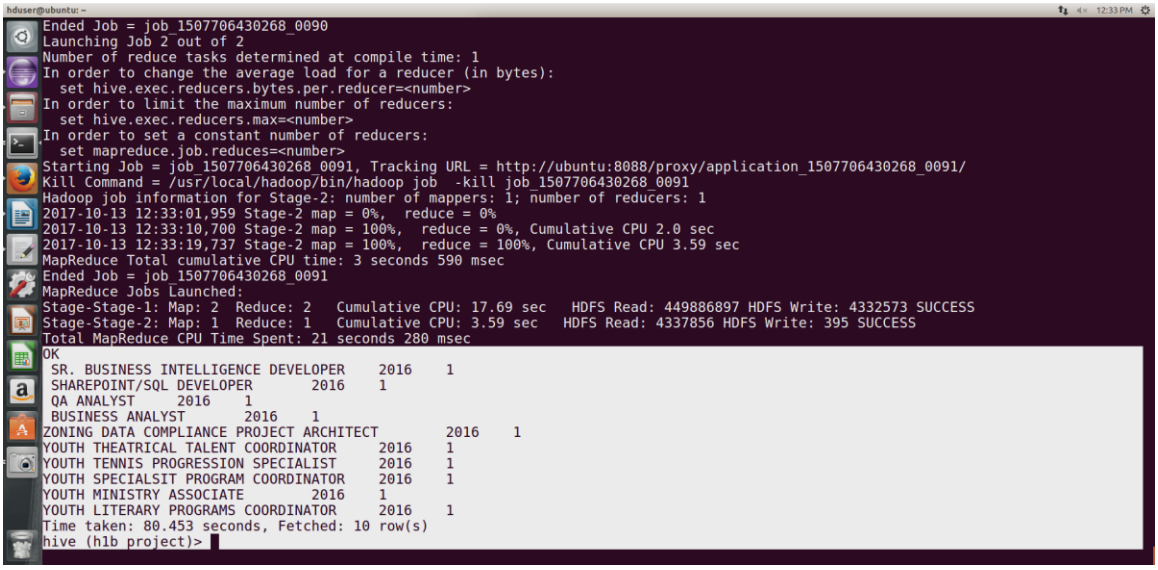
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2015 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;





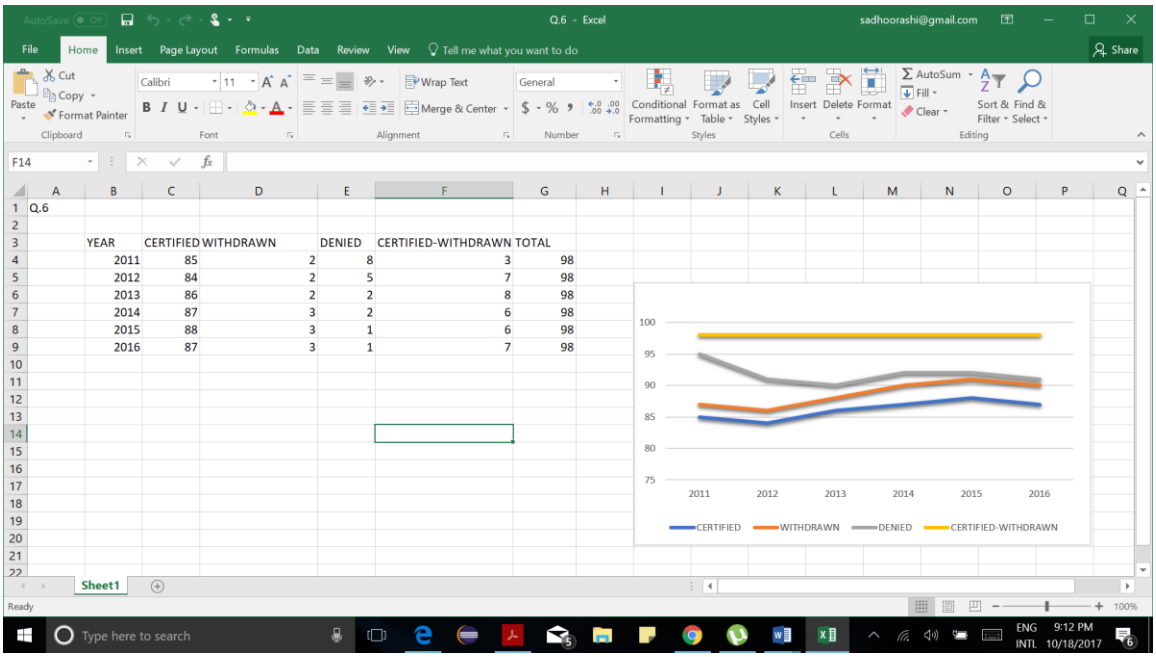
select job_title, year, count (case_status) as count_case_status from h1b_final where year = 2016 and case_status='CERTIFIED' group by job_title, year order by count_case_status limit 10;





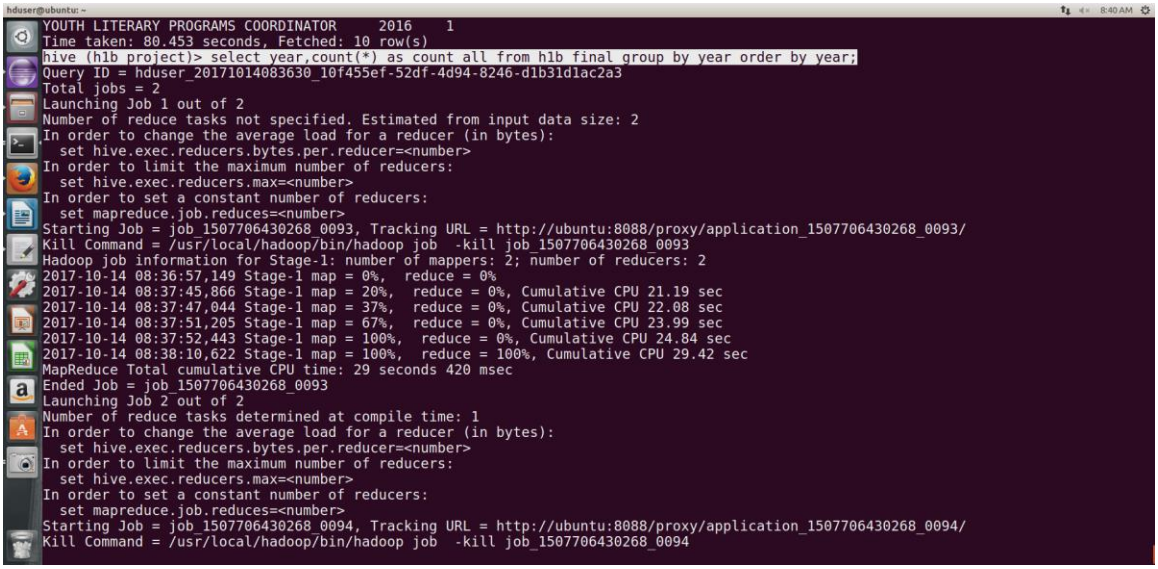
6) Find the percentage and the count of each case status on total applications for each year. Create a line graph depicting the pattern of All the cases over the period of time. (PIG)

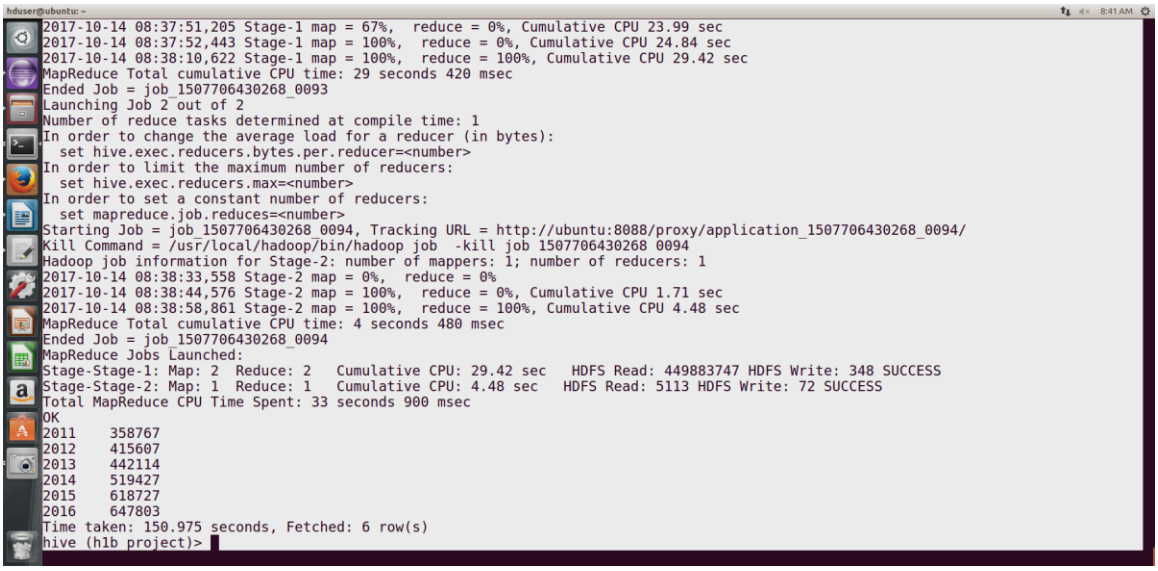
Solution:

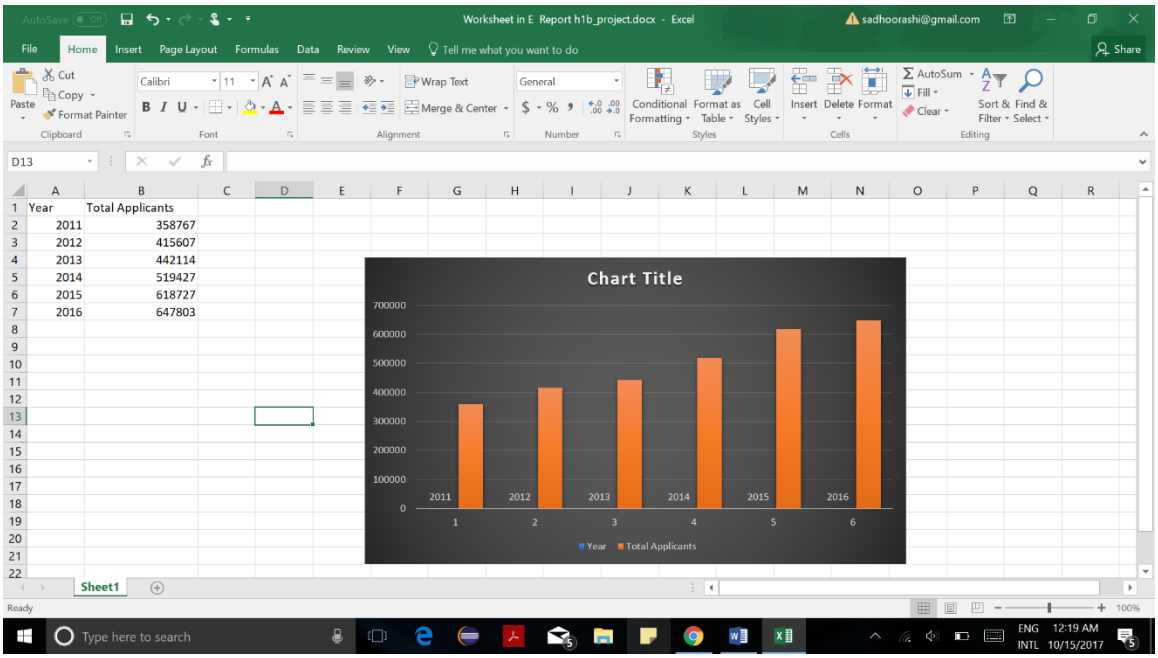


7) Create a bar graph to depict the number of applications for each year [All] (HIVE)

Solution:



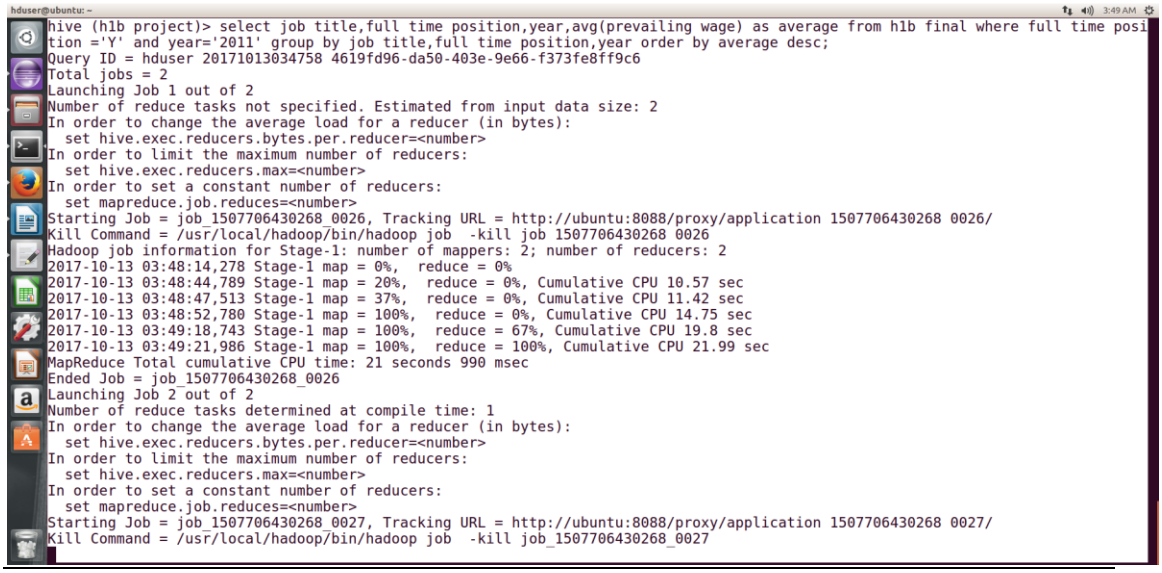


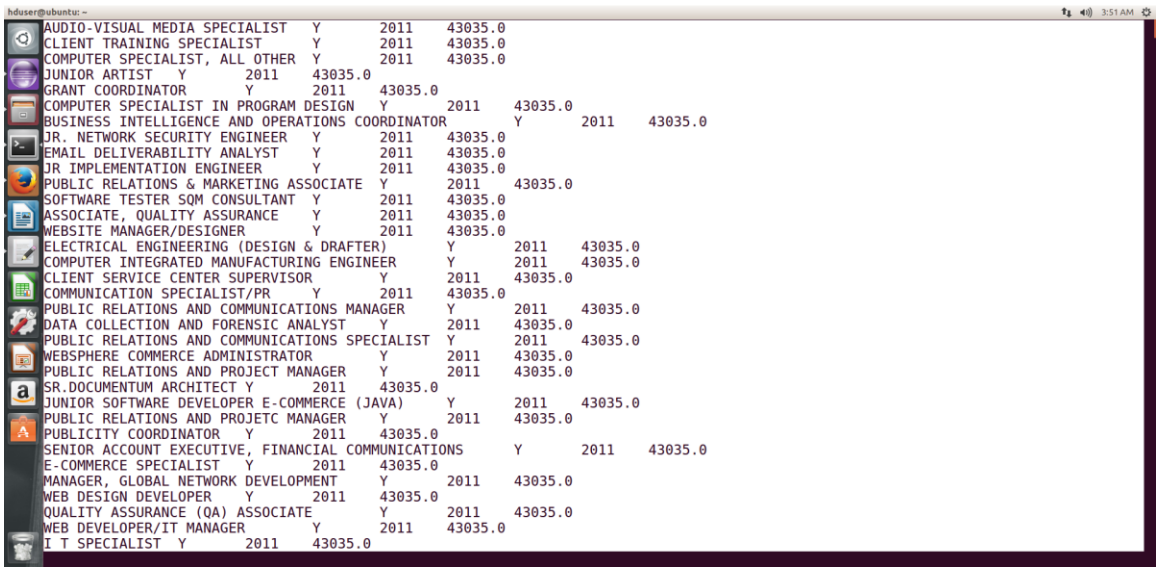

Worksheet in E Report h1b_project.docx - Excel
sadhoorashi@gmail.com
File Home Insert Page Layout Formulas Data Review View Tell me what you want to do
Share
Paste Cut Copy Format Painter
Calibri 11
Wrap Text
General
Merge & Center
Conditional Formatting
Format as Table
Cell Styles
Insert Delete Format
AutoSum Fill Clear
Sort & Filter Find & Select
Clipboard Font Alignment Number Styles Cells Editing
D13
Year Total Applicants
2011 358767
2012 415607
2013 442114
2014 519427
2015 618727
2016 647803
Chart Title
700000
600000
500000
400000
300000
200000
100000
0
2011 2012 2013 2014 2015 2016
1 2 3 4 5 6
Year Total Applicants
Sheet1
Ready
100%
Type here to search
ENG INTL 12:19 AM 10/15/2017

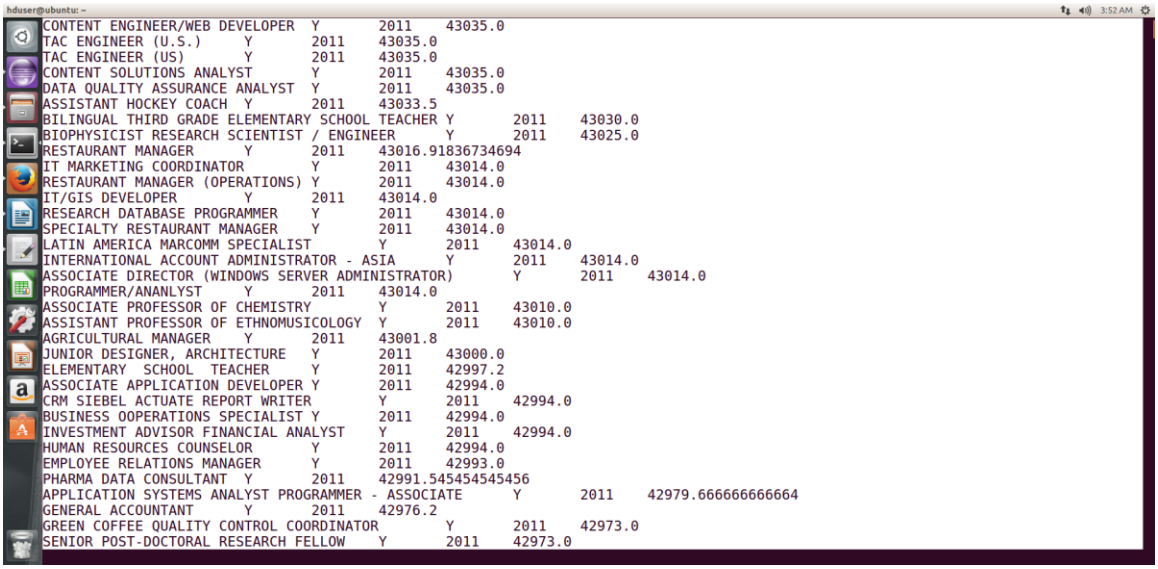
8) Find the average Prevailing Wage for each Job for each Year (take part time and full time separate). Arrange the output in descending order - [Certified and Certified Withdrawn.] (HIVE)

Solution:

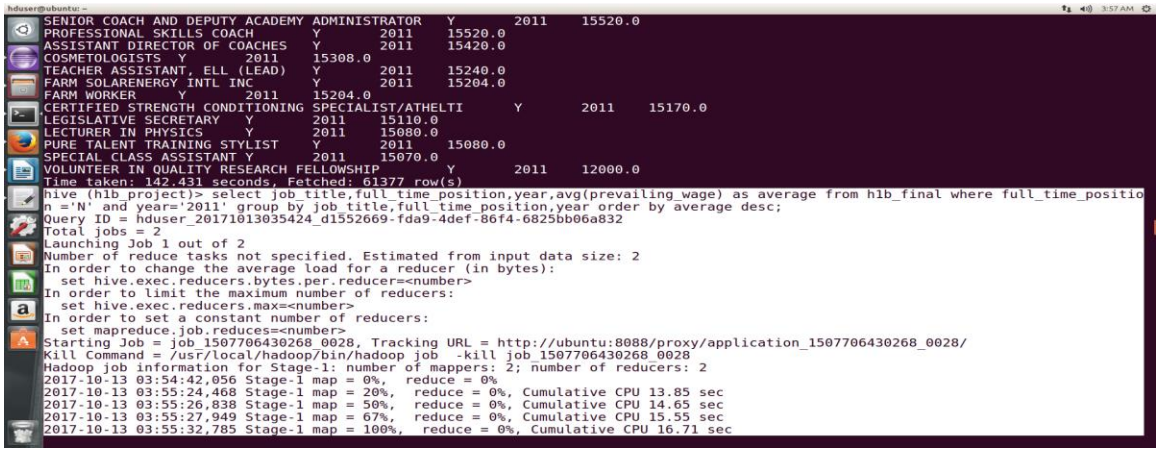
```
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2011' group by job_title, full_time_position, year order by average desc;
```

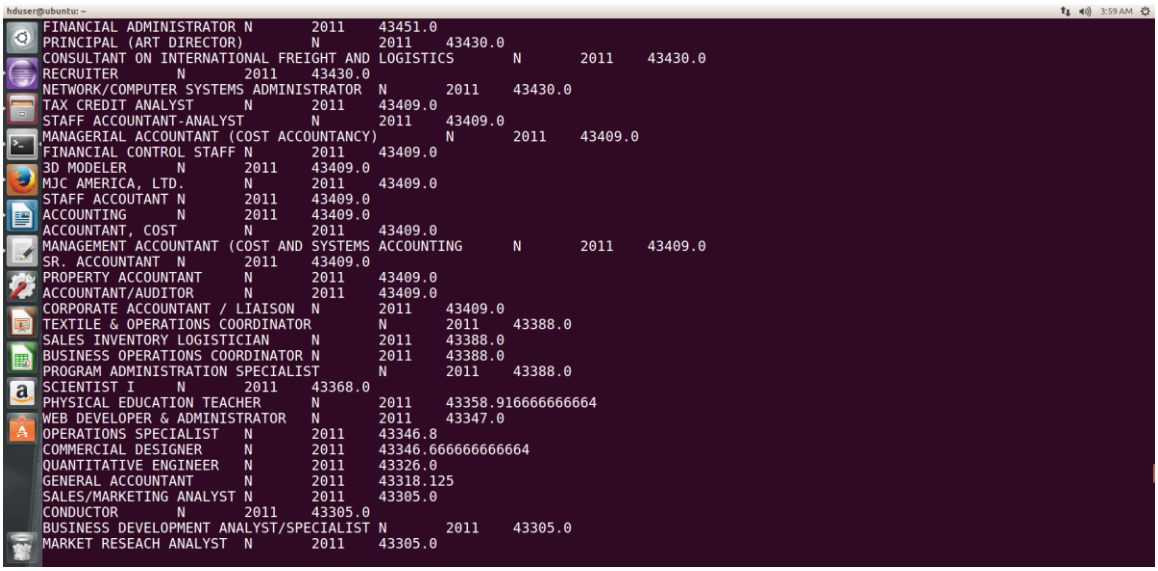


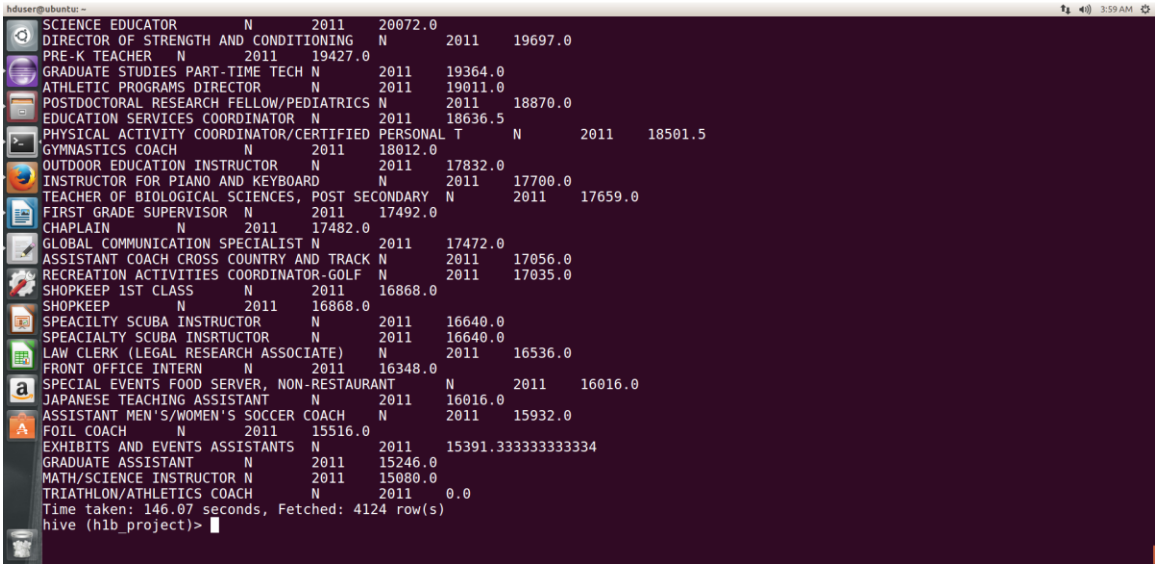

hduser@ubuntu: ~
3:51 AM
AUDIO-VISUAL MEDIA SPECIALIST Y 2011 43035.0
CLIENT TRAINING SPECIALIST Y 2011 43035.0
COMPUTER SPECIALIST, ALL OTHER Y 2011 43035.0
JUNIOR ARTIST Y 2011 43035.0
GRANT COORDINATOR Y 2011 43035.0
COMPUTER SPECIALIST IN PROGRAM DESIGN Y 2011 43035.0
BUSINESS INTELLIGENCE AND OPERATIONS COORDINATOR Y 2011 43035.0
JR. NETWORK SECURITY ENGINEER Y 2011 43035.0
EMAIL DELIVERABILITY ANALYST Y 2011 43035.0
JR IMPLEMENTATION ENGINEER Y 2011 43035.0
PUBLIC RELATIONS & MARKETING ASSOCIATE Y 2011 43035.0
SOFTWARE TESTER SQM CONSULTANT Y 2011 43035.0
ASSOCIATE, QUALITY ASSURANCE Y 2011 43035.0
WEBSITE MANAGER/DESIGNER Y 2011 43035.0
ELECTRICAL ENGINEERING (DESIGN & DRAFTER) Y 2011 43035.0
COMPUTER INTEGRATED MANUFACTURING ENGINEER Y 2011 43035.0
CLIENT SERVICE CENTER SUPERVISOR Y 2011 43035.0
COMMUNICATION SPECIALIST/PR Y 2011 43035.0
PUBLIC RELATIONS AND COMMUNICATIONS MANAGER Y 2011 43035.0
DATA COLLECTION AND FORENSIC ANALYST Y 2011 43035.0
PUBLIC RELATIONS AND COMMUNICATIONS SPECIALIST Y 2011 43035.0
WEBSPHERE COMMERCE ADMINISTRATOR Y 2011 43035.0
PUBLIC RELATIONS AND PROJECT MANAGER Y 2011 43035.0
SR.DOCUMENTUM ARCHITECT Y 2011 43035.0
JUNIOR SOFTWARE DEVELOPER E-COMMERCE (JAVA) Y 2011 43035.0
PUBLIC RELATIONS AND PROJETC MANAGER Y 2011 43035.0
PUBLICITY COORDINATOR Y 2011 43035.0
SENIOR ACCOUNT EXECUTIVE, FINANCIAL COMMUNICATIONS Y 2011 43035.0
E-COMMERCE SPECIALIST Y 2011 43035.0
MANAGER, GLOBAL NETWORK DEVELOPMENT Y 2011 43035.0
WEB DESIGN DEVELOPER Y 2011 43035.0
QUALITY ASSURANCE (QA) ASSOCIATE Y 2011 43035.0
WEB DEVELOPER/IT MANAGER Y 2011 43035.0
I T SPECIALIST Y 2011 43035.0


hduser@ubuntu: ~
3:52 AM
CONTENT ENGINEER/WEB DEVELOPER Y 2011 43035.0
TAC ENGINEER (U.S.) Y 2011 43035.0
TAC ENGINEER (US) Y 2011 43035.0
CONTENT SOLUTIONS ANALYST Y 2011 43035.0
DATA QUALITY ASSURANCE ANALYST Y 2011 43035.0
ASSISTANT HOCKEY COACH Y 2011 43033.5
BILINGUAL THIRD GRADE ELEMENTARY SCHOOL TEACHER Y 2011 43030.0
BIOPHYSICIST RESEARCH SCIENTIST / ENGINEER Y 2011 43025.0
RESTAURANT MANAGER Y 2011 43016.91836734694
IT MARKETING COORDINATOR Y 2011 43014.0
RESTAURANT MANAGER (OPERATIONS) Y 2011 43014.0
IT/GIS DEVELOPER Y 2011 43014.0
RESEARCH DATABASE PROGRAMMER Y 2011 43014.0
SPECIALTY RESTAURANT MANAGER Y 2011 43014.0
LATIN AMERICA MARCOMM SPECIALIST Y 2011 43014.0
INTERNATIONAL ACCOUNT ADMINISTRATOR - ASIA Y 2011 43014.0
ASSOCIATE DIRECTOR (WINDOWS SERVER ADMINISTRATOR) Y 2011 43014.0
PROGRAMMER/ANANLYST Y 2011 43014.0
ASSOCIATE PROFESSOR OF CHEMISTRY Y 2011 43010.0
ASSISTANT PROFESSOR OF ETHNOMUSICOLOGY Y 2011 43010.0
AGRICULTURAL MANAGER Y 2011 43001.8
JUNIOR DESIGNER, ARCHITECTURE Y 2011 43000.0
ELEMENTARY SCHOOL TEACHER Y 2011 42997.2
ASSOCIATE APPLICATION DEVELOPER Y 2011 42994.0
CRM SIEBEL ACTUATE REPORT WRITER Y 2011 42994.0
BUSINESS OOPERATIONS SPECIALIST Y 2011 42994.0
INVESTMENT ADVISOR FINANCIAL ANALYST Y 2011 42994.0
HUMAN RESOURCES COUNSELOR Y 2011 42994.0
EMPLOYEE RELATIONS MANAGER Y 2011 42993.0
PHARMA DATA CONSULTANT Y 2011 42991.545454545456
APPLICATION SYSTEMS ANALYST PROGRAMMER - ASSOCIATE Y 2011 42979.666666666664
GENERAL ACCOUNTANT Y 2011 42976.2
GREEN COFFEE QUALITY CONTROL COORDINATOR Y 2011 42973.0
SENIOR POST-DOCTORAL RESEARCH FELLOW Y 2011 42973.0

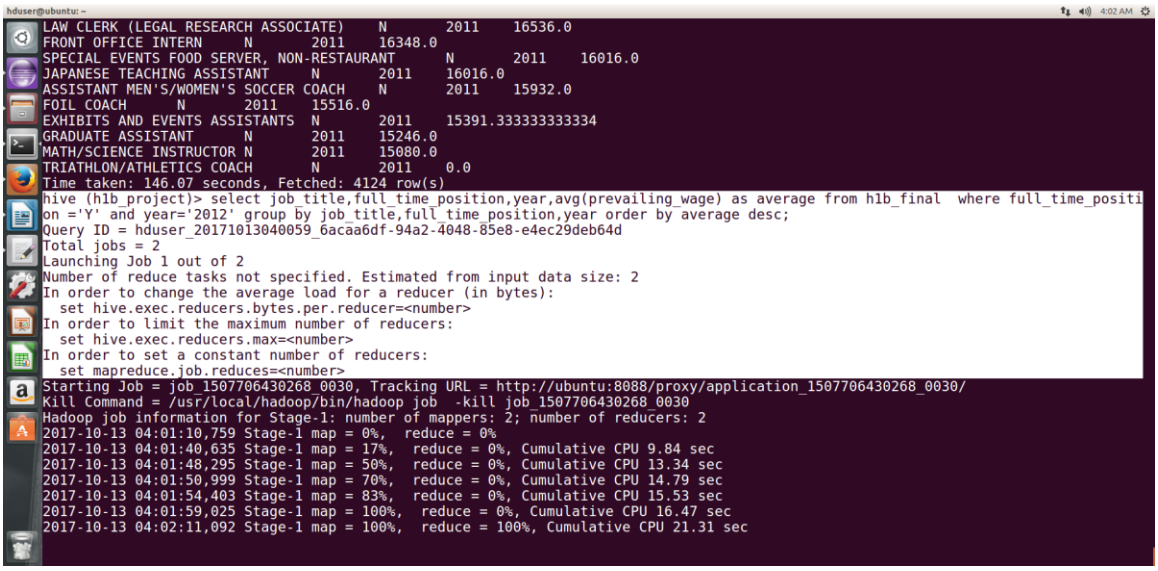
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2011' group by job_title, full_time_position, year order by average desc;

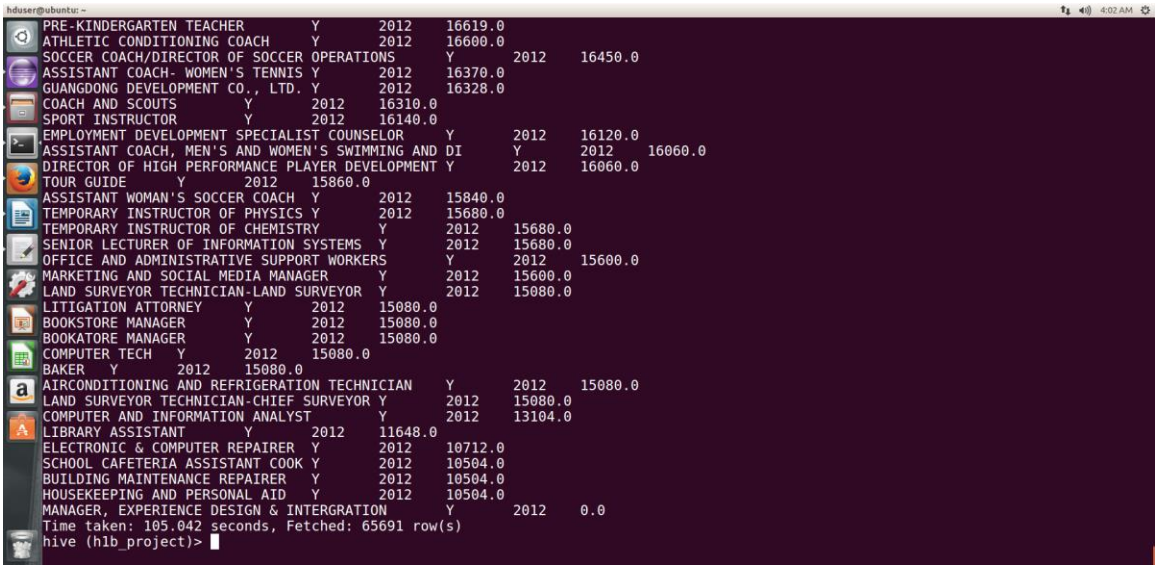




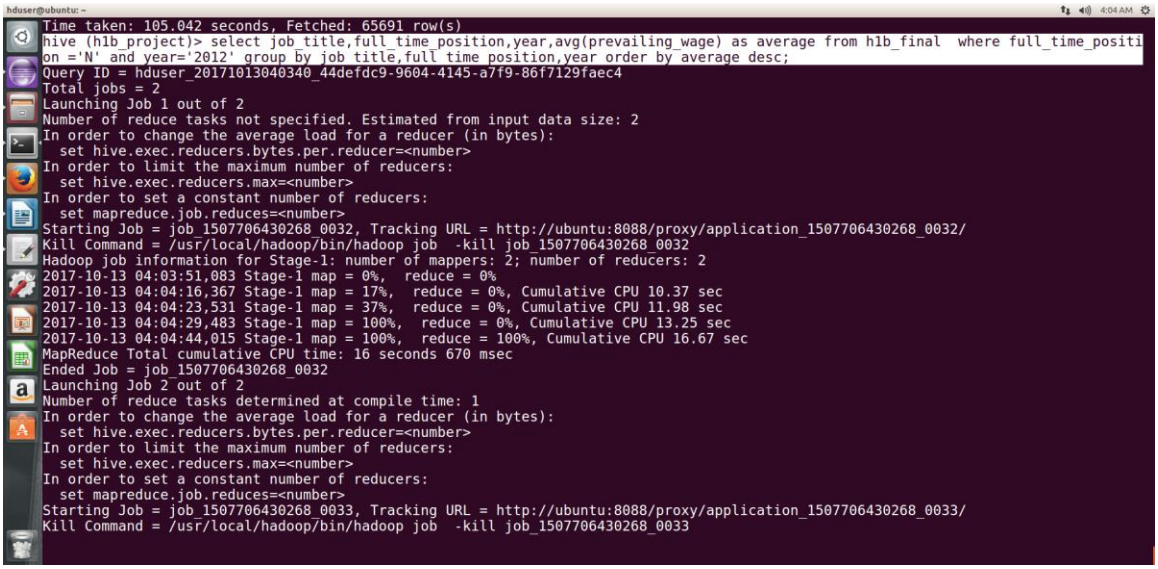

hduser@ubuntu: ~
3:59 AM
SCIENCE EDUCATOR N 2011 20072.0
DIRECTOR OF STRENGTH AND CONDITIONING N 2011 19697.0
PRE-K TEACHER N 2011 19427.0
GRADUATE STUDIES PART-TIME TECH N 2011 19364.0
ATHLETIC PROGRAMS DIRECTOR N 2011 19011.0
POSTDOCTORAL RESEARCH FELLOW/PEDIATRICS N 2011 18870.0
EDUCATION SERVICES COORDINATOR N 2011 18636.5
PHYSICAL ACTIVITY COORDINATOR/CERTIFIED PERSONAL T N 2011 18501.5
GYMNASTICS COACH N 2011 18012.0
OUTDOOR EDUCATION INSTRUCTOR N 2011 17832.0
INSTRUCTOR FOR PIANO AND KEYBOARD N 2011 17700.0
TEACHER OF BIOLOGICAL SCIENCES, POST SECONDARY N 2011 17659.0
FIRST GRADE SUPERVISOR N 2011 17492.0
CHAPLAIN N 2011 17482.0
GLOBAL COMMUNICATION SPECIALIST N 2011 17472.0
ASSISTANT COACH CROSS COUNTRY AND TRACK N 2011 17056.0
RECREATION ACTIVITIES COORDINATOR-GOLF N 2011 17035.0
SHOPKEEP 1ST CLASS N 2011 16868.0
SHOPKEEP N 2011 16868.0
SPEACILTY SCUBA INSTRUCTOR N 2011 16640.0
SPEACIALTY SCUBA INSRTUCTOR N 2011 16640.0
LAW CLERK (LEGAL RESEARCH ASSOCIATE) N 2011 16536.0
FRONT OFFICE INTERN N 2011 16348.0
SPECIAL EVENTS FOOD SERVER, NON-RESTAURANT N 2011 16016.0
JAPANESE TEACHING ASSISTANT N 2011 16016.0
ASSISTANT MEN'S/WOMEN'S SOCCER COACH N 2011 15932.0
FOIL COACH N 2011 15516.0
EXHIBITS AND EVENTS ASSISTANTS N 2011 15391.333333333334
GRADUATE ASSISTANT N 2011 15246.0
MATH/SCIENCE INSTRUCTOR N 2011 15080.0
TRIATHLON/ATHLETICS COACH N 2011 0.0
Time taken: 146.07 seconds, Fetched: 4124 row(s)
hive (h1b_project)>

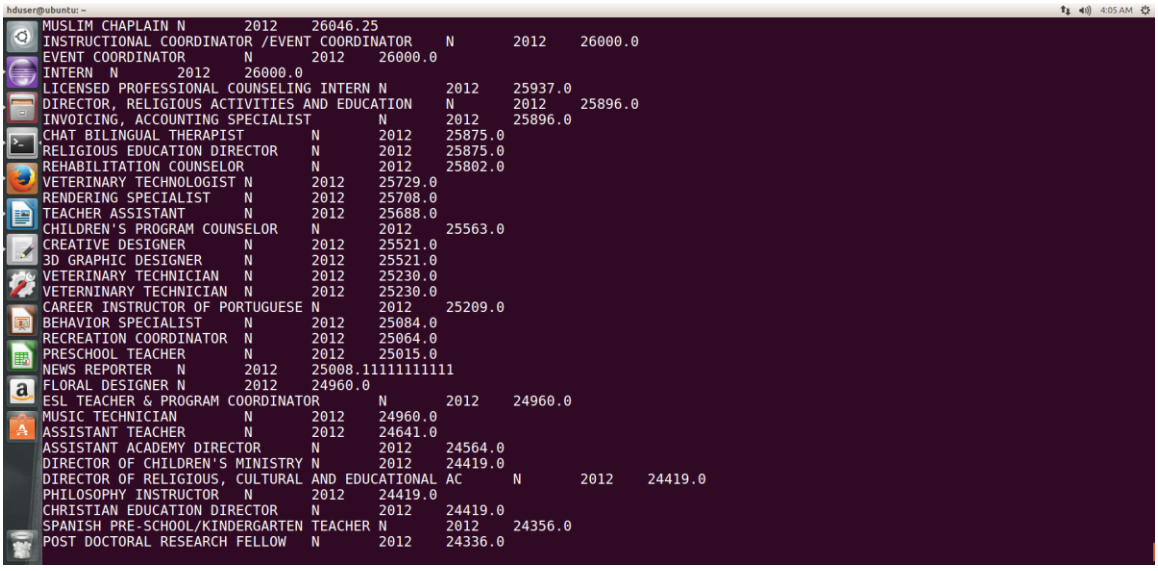
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2012' group by job_title, full_time_position, year order by average desc;



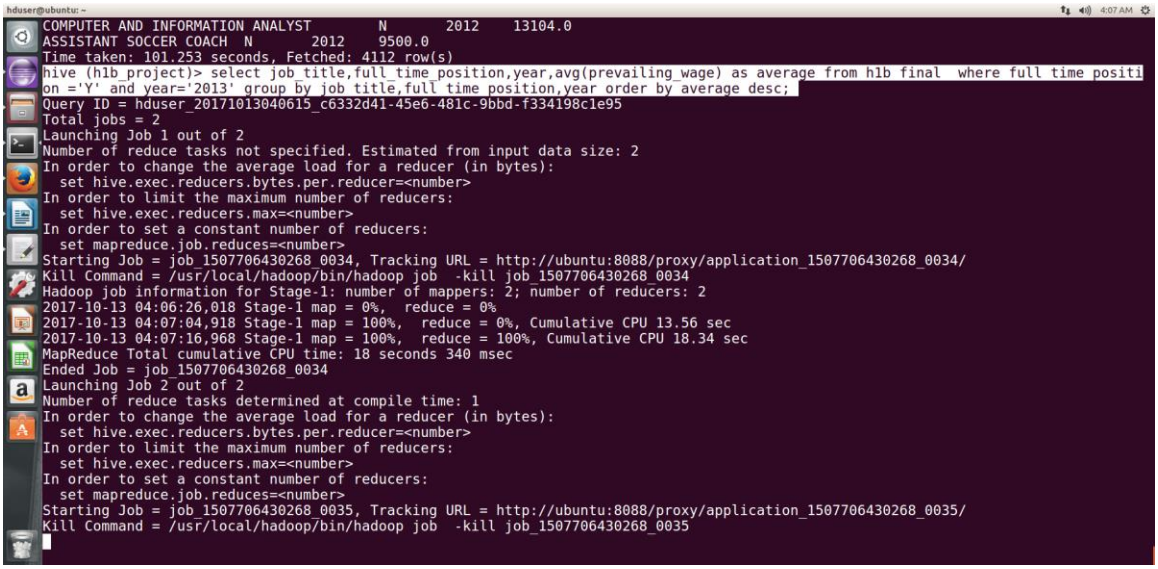


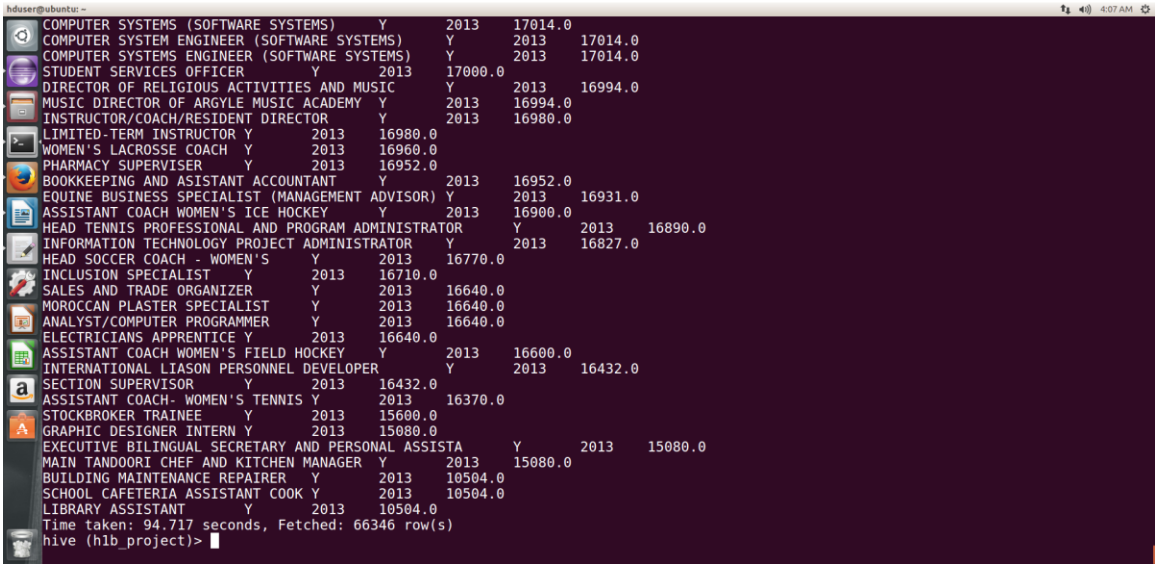
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2012' group by job_title, full_time_position, year order by average desc;



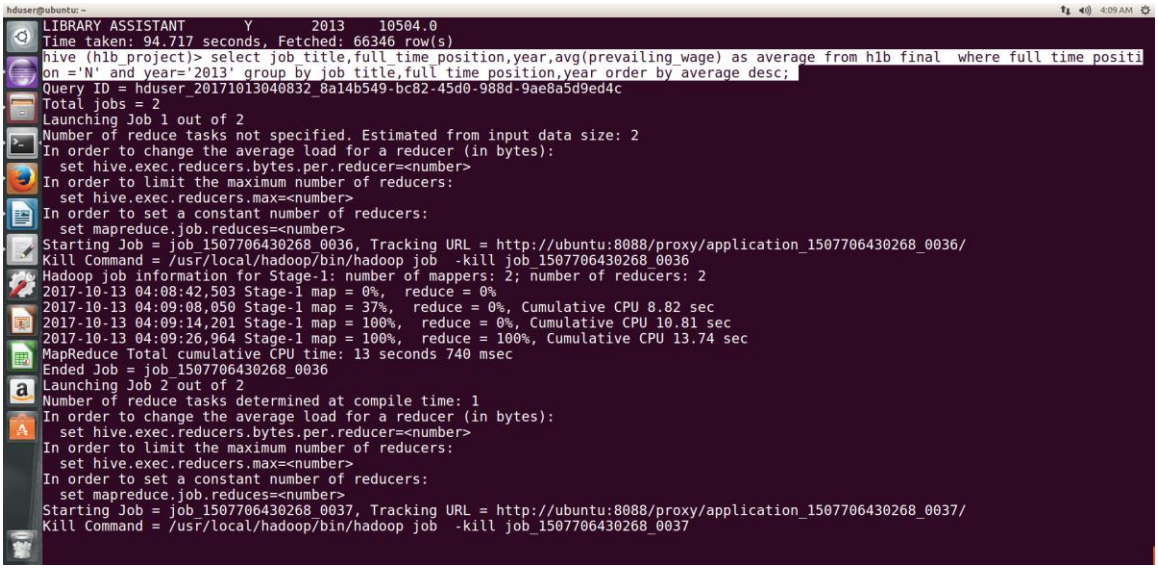


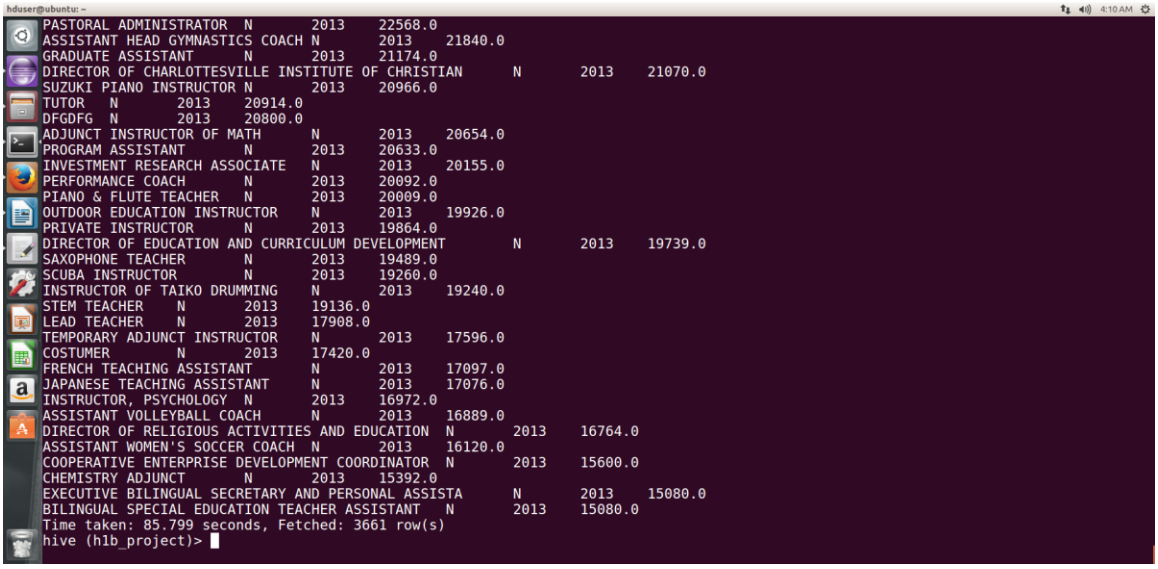
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2013' group by job_title, full_time_position, year order by average desc;



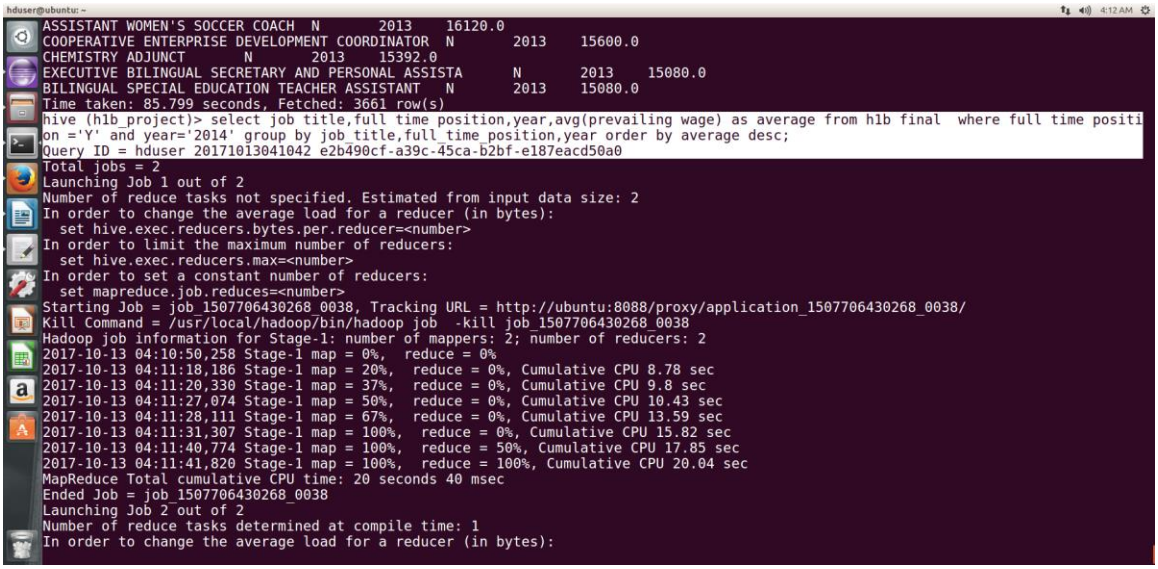


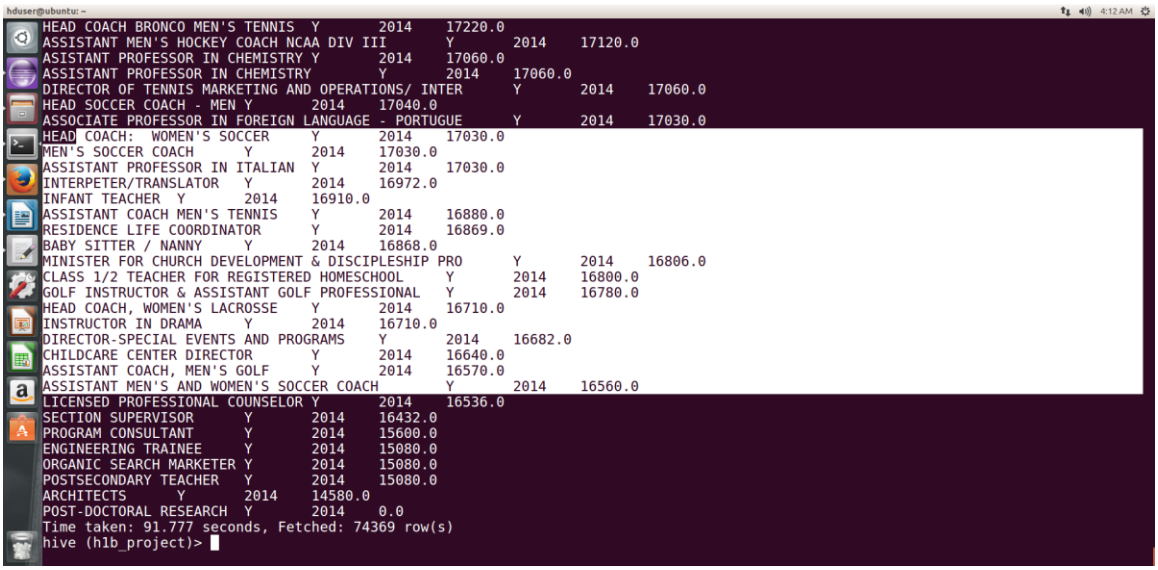
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2013' group by job_title, full_time_position, year order by average desc;



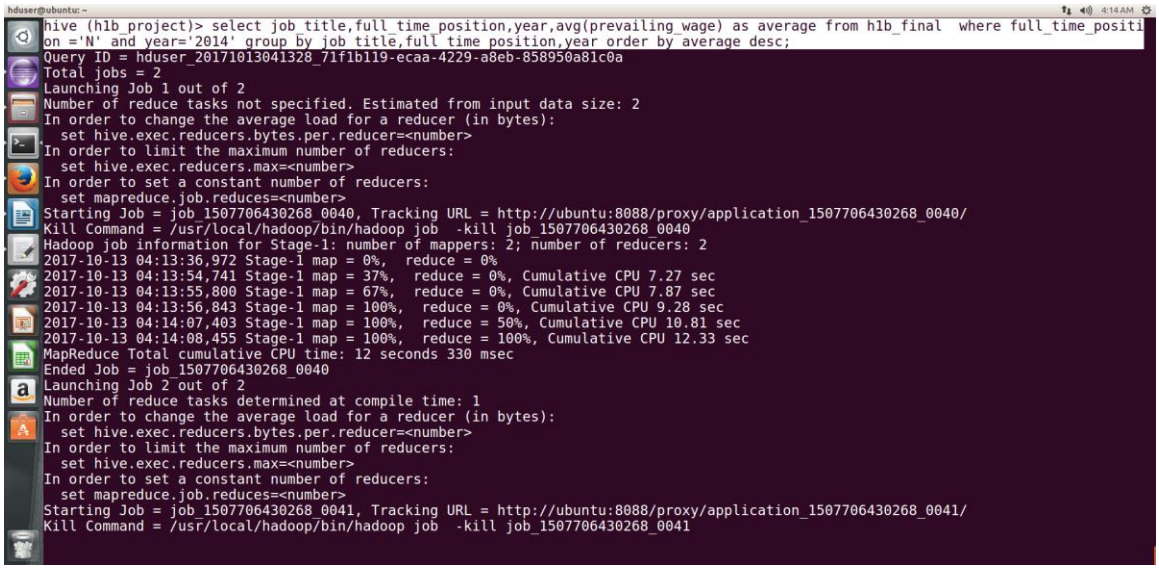


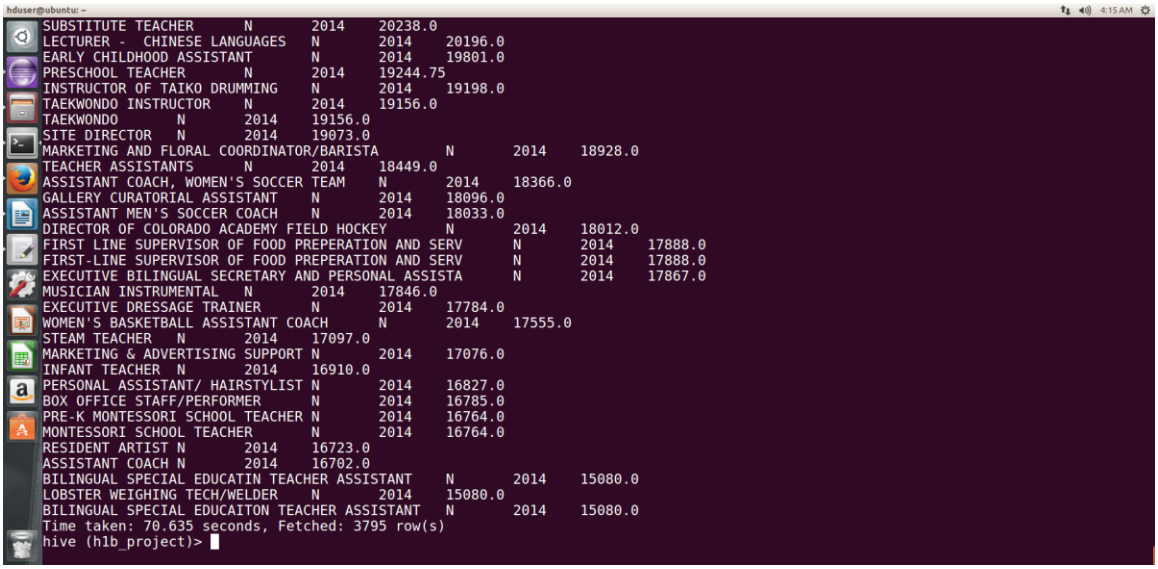
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2014' group by job_title, full_time_position, year order by average desc;



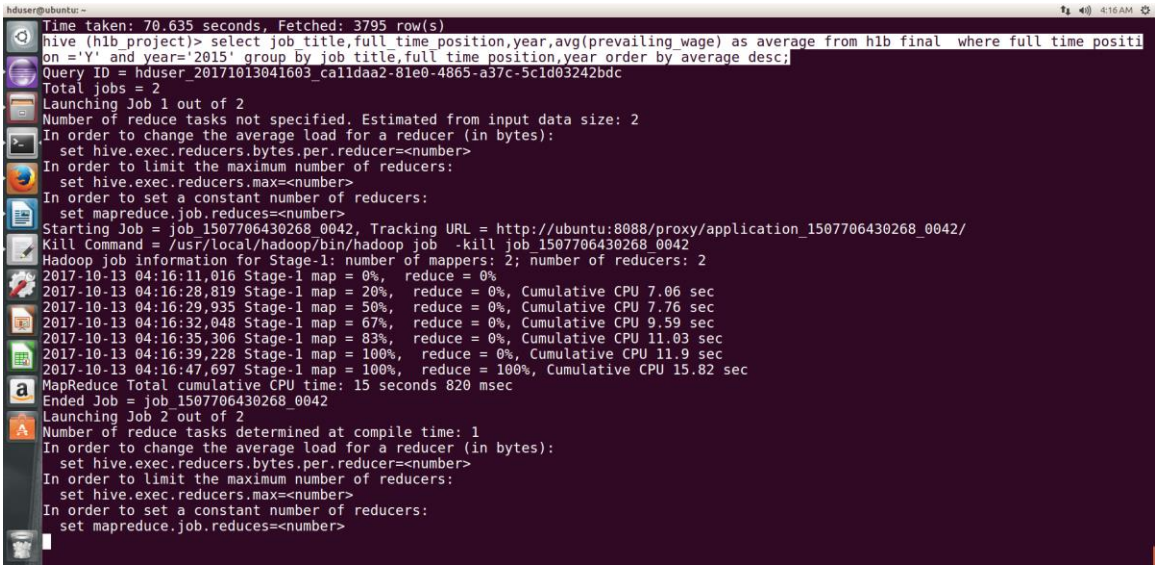


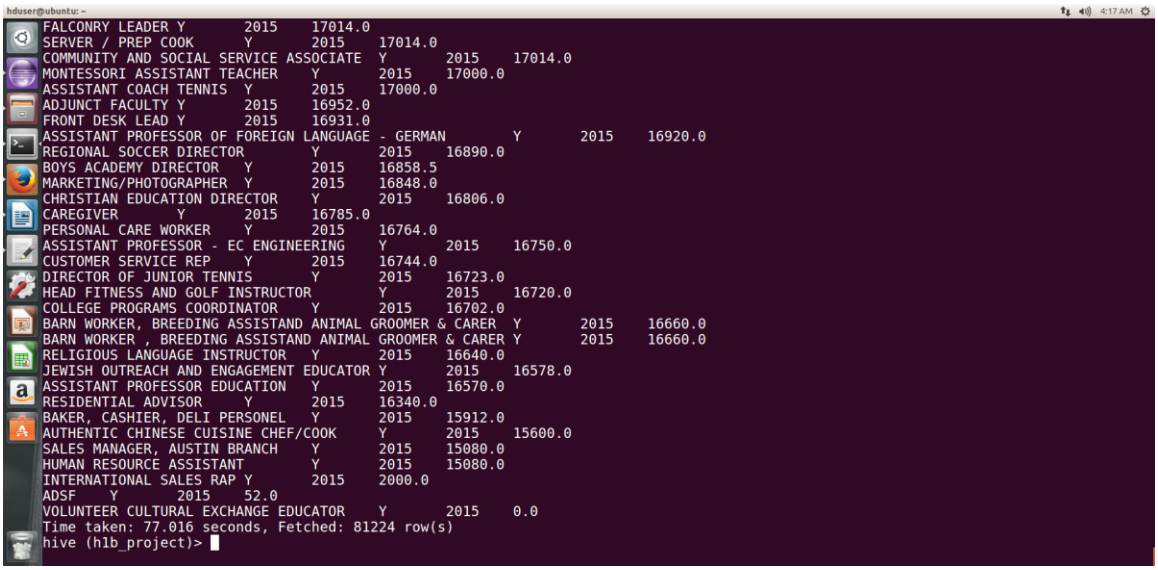
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2014' group by job_title, full_time_position, year order by average desc;



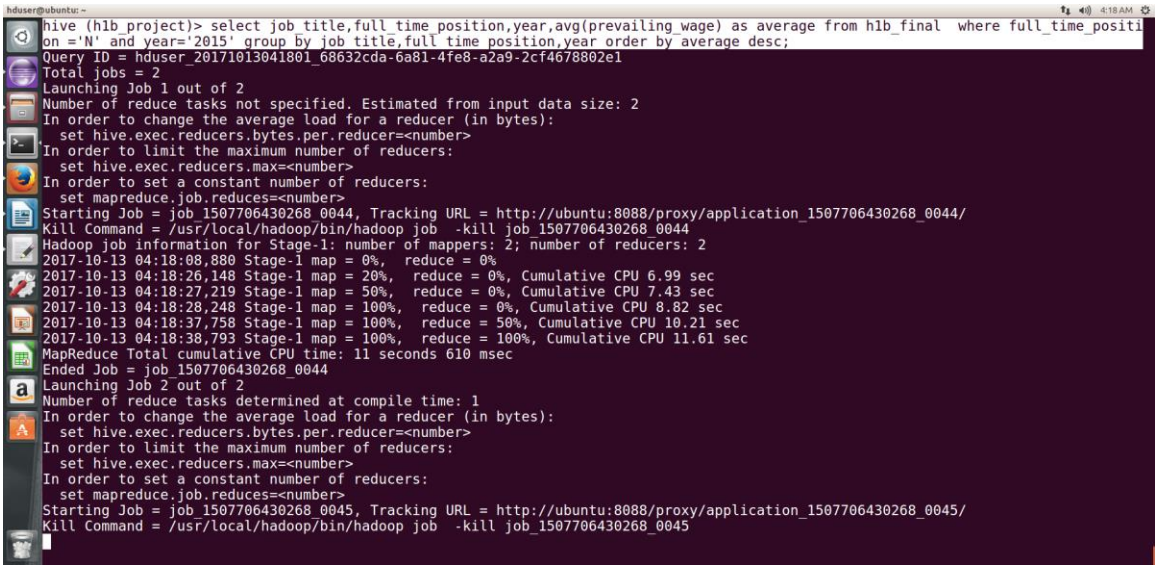


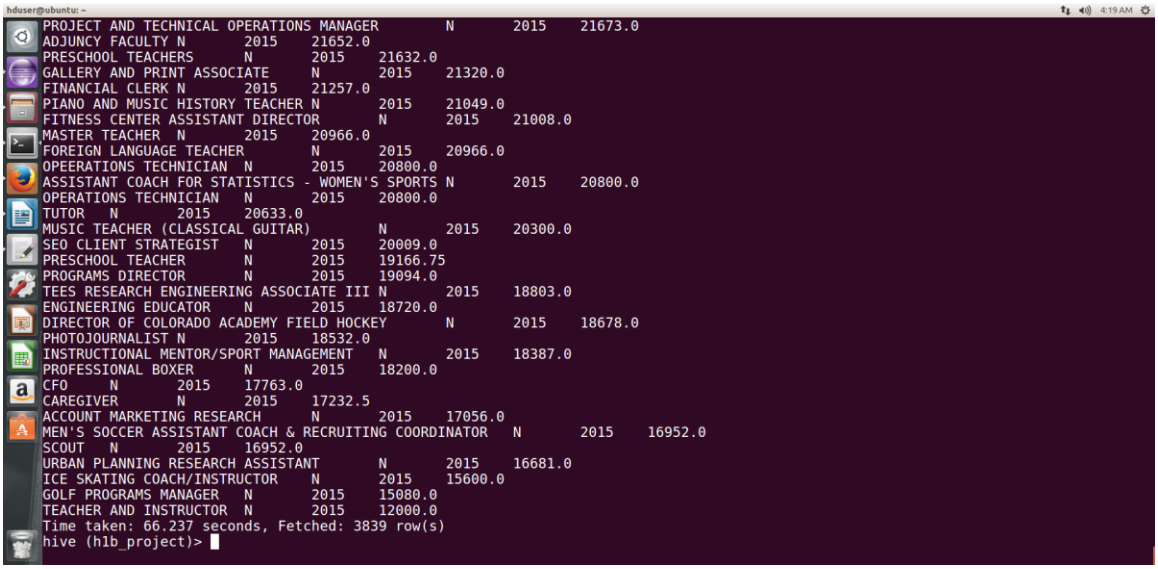
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2015' group by job_title, full_time_position, year order by average desc;



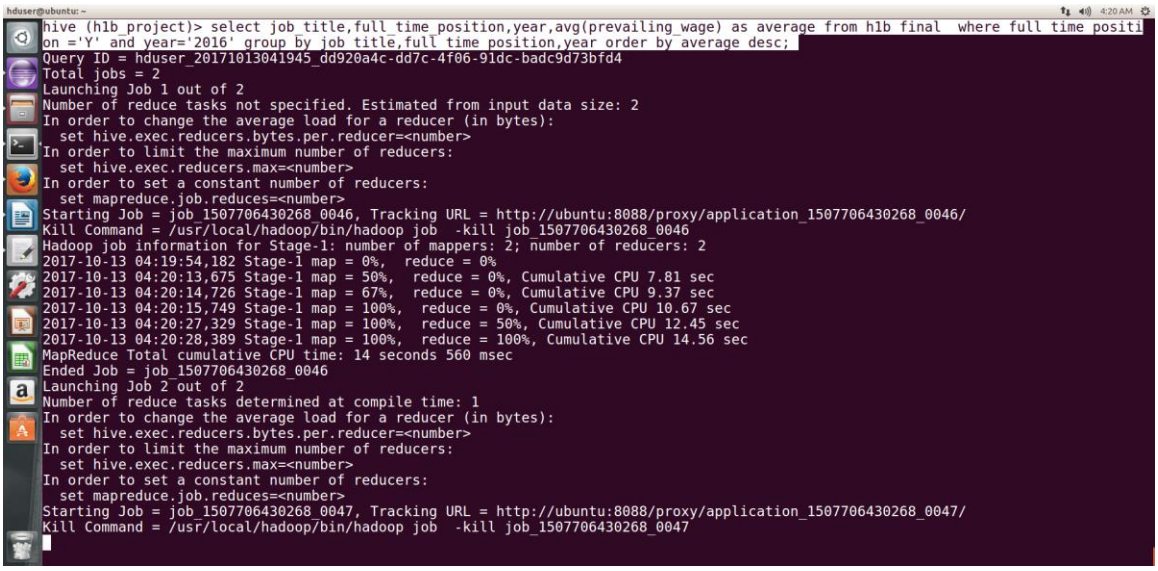


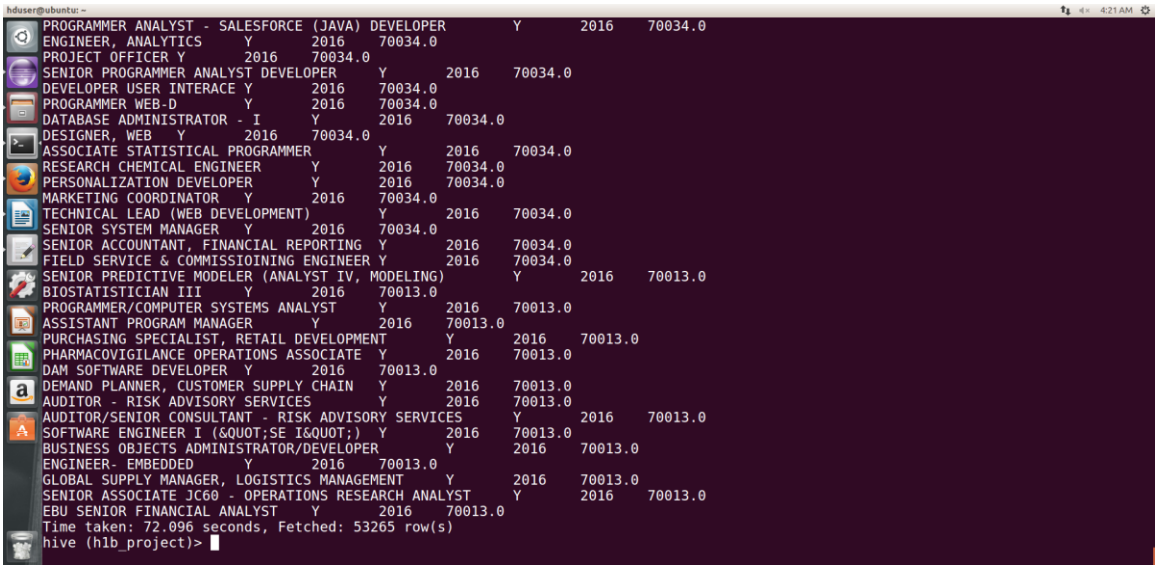
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2015' group by job_title, full_time_position, year order by average desc;



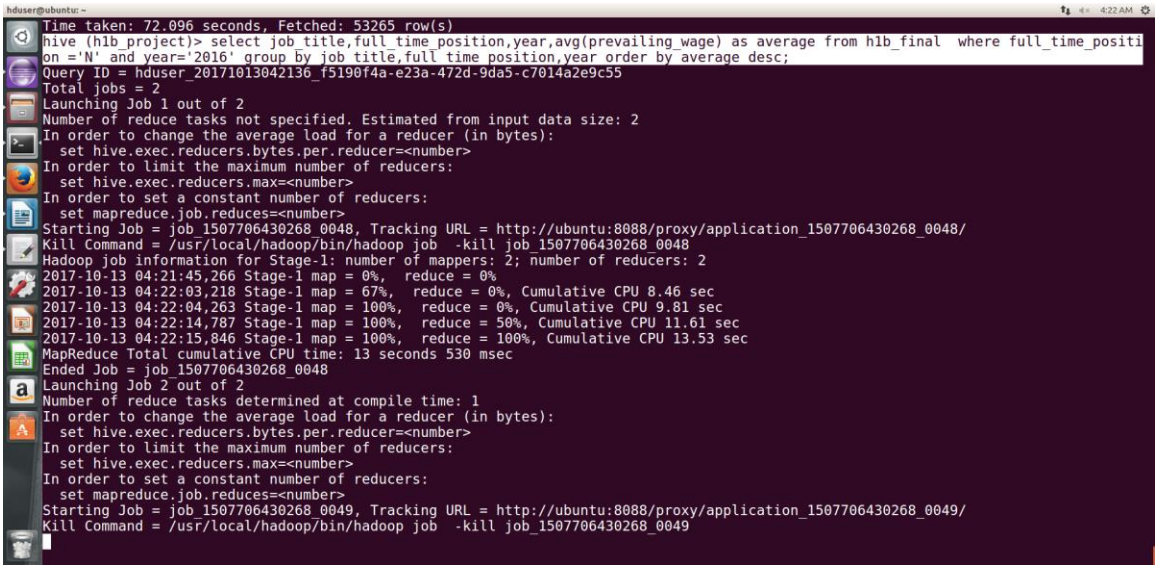


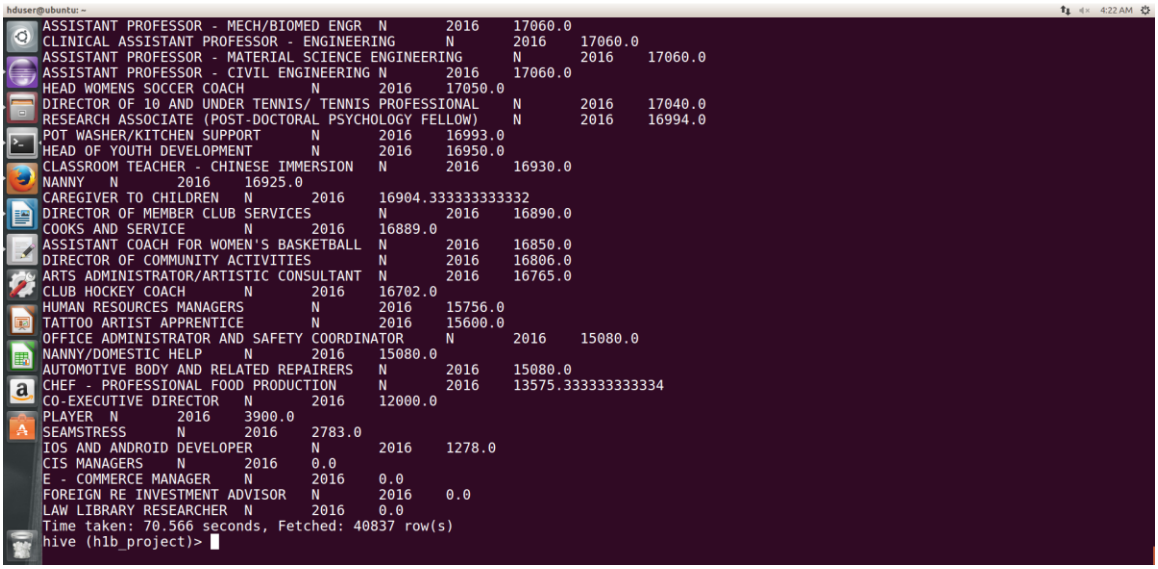
select job_title, full_time_position, year, avg(prevailing_wage) as average from h1b_final where full_time_position ='Y' and year='2016' group by job_title, full_time_position, year order by average desc;





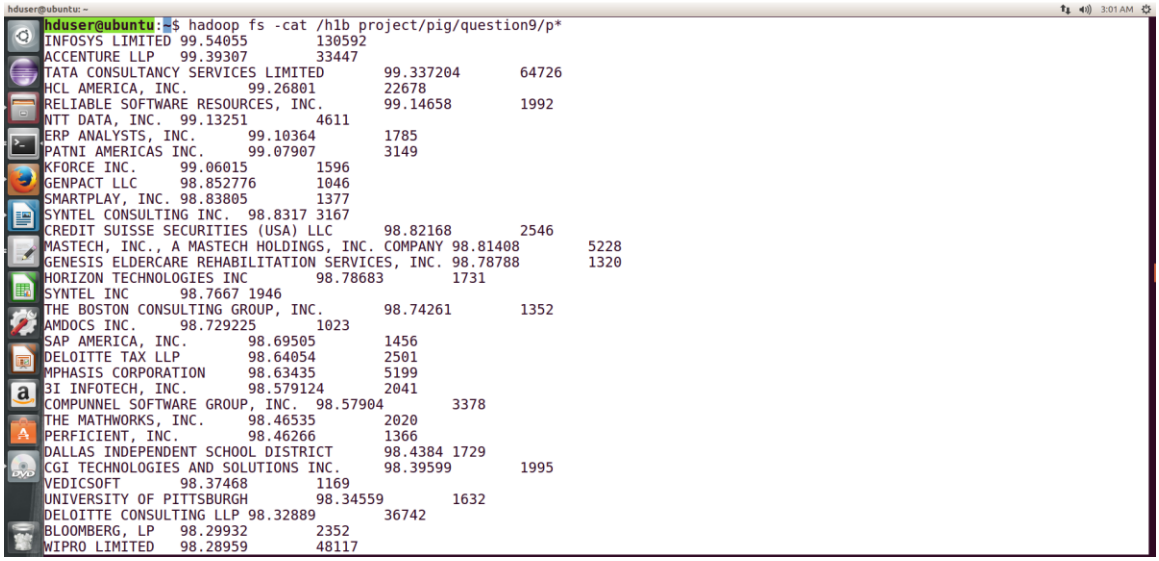
select job_title, full_time_position, year, age(prevailing_wage) as average from h1b_final where full_time_position ='N' and year='2016' group by job_title, full_time_position, year order by average desc;

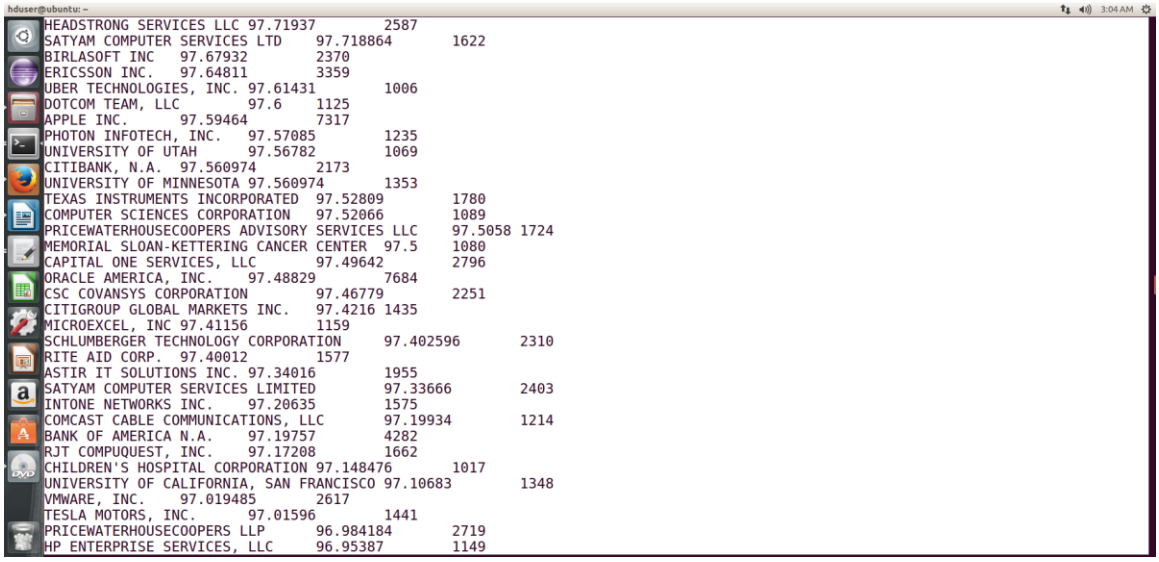




9) Which are the employers along with the number of petitions who have the success rate more than 70% in petitions. (total petitions filed 1000 OR more than 1000)? (PIG)

Solution:

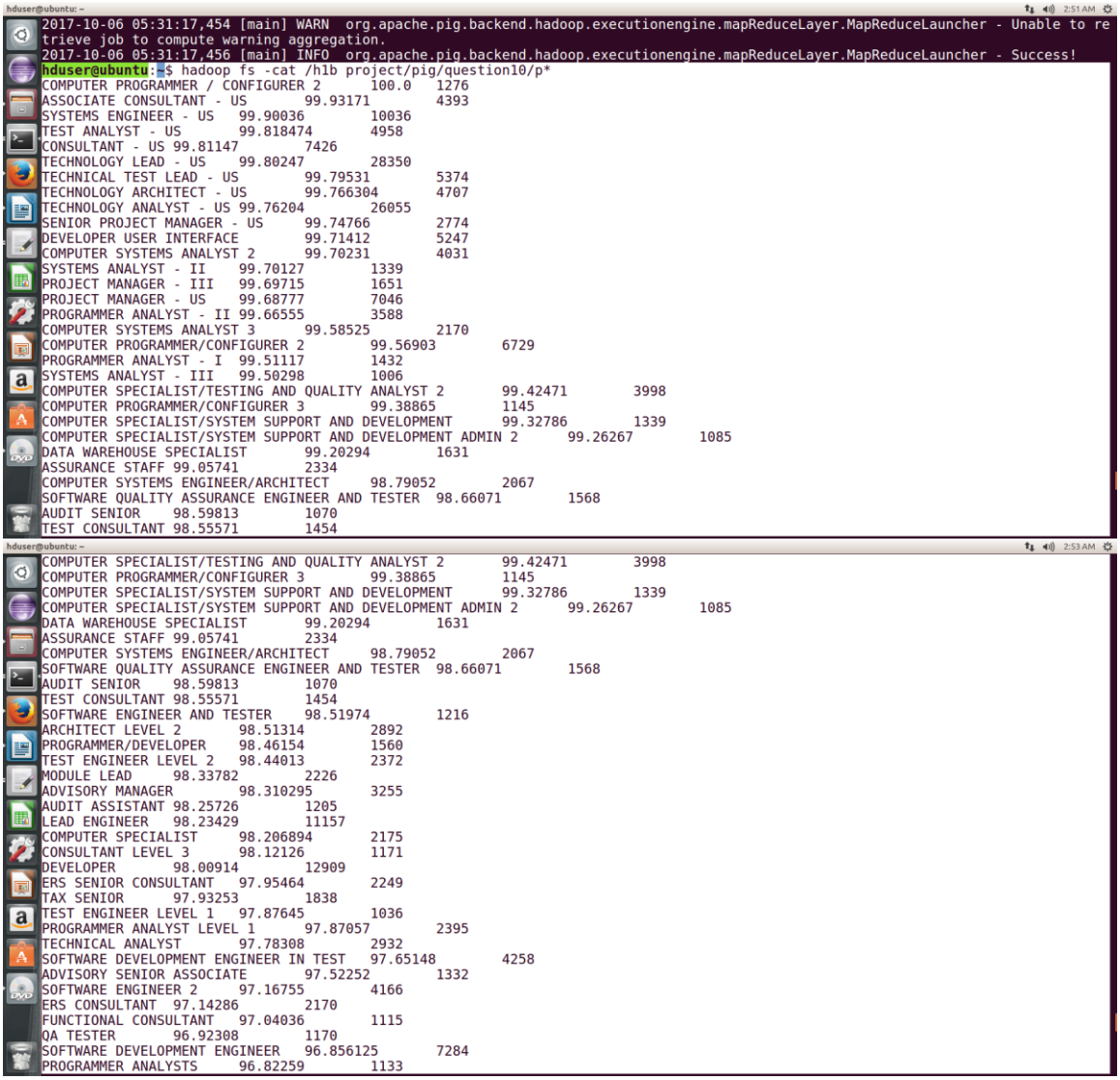




---

10) Which are the job positions along with the number of petitions which have the success rate more than 70% in petitions (total petitions filed 1000 OR more than1000)? (PIG)

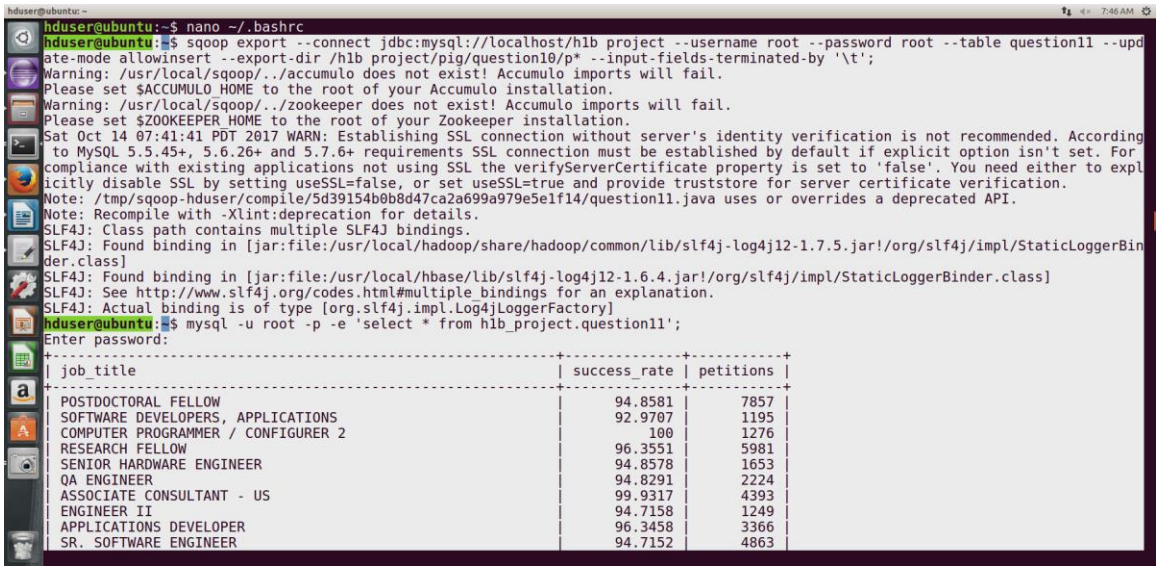
Solution:



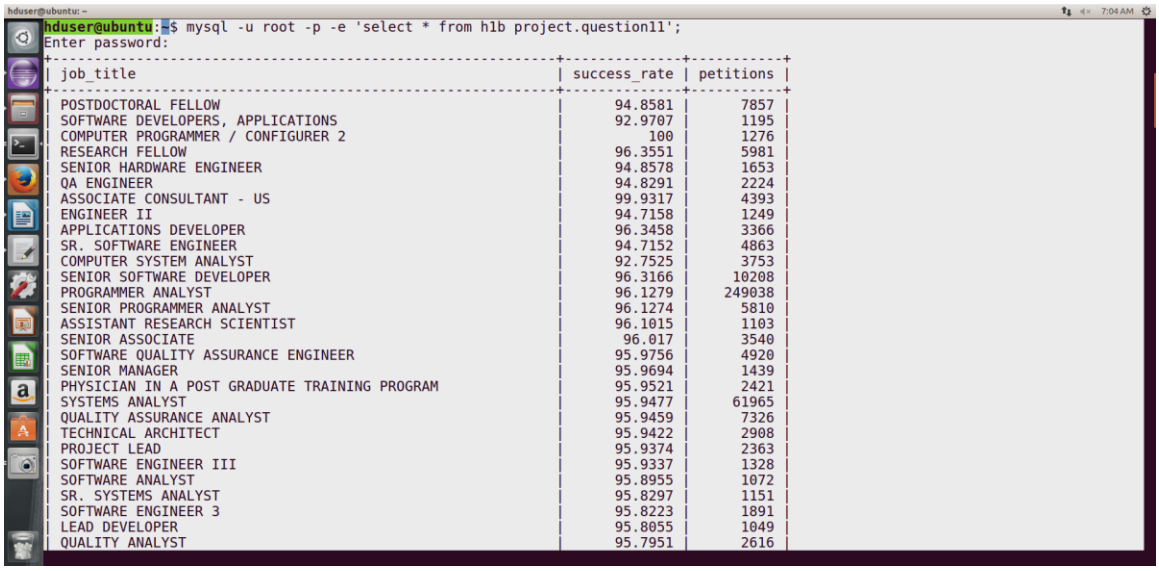
---

11) Export result for question no 10 to MySQL database. (SQOOP)

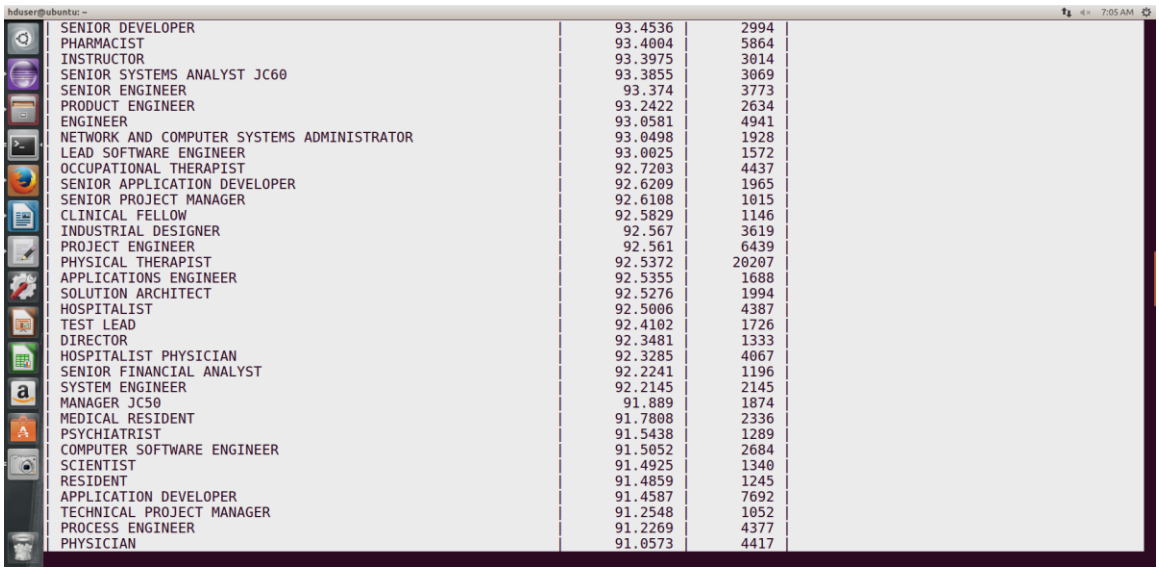
Solution:



```
hduser@ubuntu:~$ nano ~/.bashrc
hduser@ubuntu:~$ sqoop export --connect jdbc:mysql://localhost/h1b project --username root --password root --table question11 --upd
ate-mode allowinsert --export-dir /h1b project/pig/question10/p* --input-fields-terminated-by '\t';
Warning: /usr/local/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Sat Oct 14 07:41:41 PDT 2017 WARN: Establishing SSL connection without server's identity verification is not recommended. According
 to MySQL 5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established by default if explicit option isn't set. For
compliance with existing applications not using SSL the verifyServerCertificate property is set to 'false'. You need either to expl
icitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate verification.
Note: /tmp/sqoop-hduser/compile/5d39154b0b8d47ca2a699a979e5e1f14/question11.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBin
der.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
hduser@ubuntu:~$ mysql -u root -p -e 'select * from h1b_project.question11';
Enter password:
+-----------------------------------------------------------+--------------+-----------+
| job_title                                                 | success_rate | petitions |
+-----------------------------------------------------------+--------------+-----------+
| POSTDOCTORAL FELLOW                                       |      94.8581 |      7857 |
| SOFTWARE DEVELOPERS, APPLICATIONS                         |      92.9707 |      1195 |
| COMPUTER PROGRAMMER / CONFIGURER 2                        |          100 |      1276 |
| RESEARCH FELLOW                                           |      96.3551 |      5981 |
| SENIOR HARDWARE ENGINEER                                  |      94.8578 |      1653 |
| QA ENGINEER                                               |      94.8291 |      2224 |
| ASSOCIATE CONSULTANT - US                                 |      99.9317 |      4393 |
| ENGINEER II                                               |      94.7158 |      1249 |
| APPLICATIONS DEVELOPER                                    |      96.3458 |      3366 |
| SR. SOFTWARE ENGINEER                                     |      94.7152 |      4863 |
```



```
hduser@ubuntu:~$ mysql -u root -p -e 'select * from h1b project.question11';
Enter password:
+-----------------------------------------------------------+--------------+-----------+
| job_title                                                 | success_rate | petitions |
+-----------------------------------------------------------+--------------+-----------+
| POSTDOCTORAL FELLOW                                       |      94.8581 |      7857 |
| SOFTWARE DEVELOPERS, APPLICATIONS                         |      92.9707 |      1195 |
| COMPUTER PROGRAMMER / CONFIGURER 2                        |          100 |      1276 |
| RESEARCH FELLOW                                           |      96.3551 |      5981 |
| SENIOR HARDWARE ENGINEER                                  |      94.8578 |      1653 |
| QA ENGINEER                                               |      94.8291 |      2224 |
| ASSOCIATE CONSULTANT - US                                 |      99.9317 |      4393 |
| ENGINEER II                                               |      94.7158 |      1249 |
| APPLICATIONS DEVELOPER                                    |      96.3458 |      3366 |
| SR. SOFTWARE ENGINEER                                     |      94.7152 |      4863 |
| COMPUTER SYSTEM ANALYST                                   |      92.7525 |      3753 |
| SENIOR SOFTWARE DEVELOPER                                 |      96.3166 |     10208 |
| PROGRAMMER ANALYST                                        |      96.1279 |    249038 |
| SENIOR PROGRAMMER ANALYST                                 |      96.1274 |      5810 |
| ASSISTANT RESEARCH SCIENTIST                              |      96.1015 |      1103 |
| SENIOR ASSOCIATE                                          |       96.017 |      3540 |
| SOFTWARE QUALITY ASSURANCE ENGINEER                       |      95.9756 |      4920 |
| SENIOR MANAGER                                            |      95.9694 |      1439 |
| PHYSICIAN IN A POST GRADUATE TRAINING PROGRAM             |      95.9521 |      2421 |
| SYSTEMS ANALYST                                           |      95.9477 |     61965 |
| QUALITY ASSURANCE ANALYST                                 |      95.9459 |      7326 |
| TECHNICAL ARCHITECT                                       |      95.9422 |      2908 |
| PROJECT LEAD                                              |      95.9374 |      2363 |
| SOFTWARE ENGINEER III                                     |      95.9337 |      1328 |
| SOFTWARE ANALYST                                          |      95.8955 |      1072 |
| SR. SYSTEMS ANALYST                                       |      95.8297 |      1151 |
| SOFTWARE ENGINEER 3                                       |      95.8223 |      1891 |
| LEAD DEVELOPER                                            |      95.8055 |      1049 |
| QUALITY ANALYST                                           |      95.7951 |      2616 |
```



SENIOR DEVELOPER	93.4536	2994
PHARMACIST	93.4004	5864
INSTRUCTOR	93.3975	3014
SENIOR SYSTEMS ANALYST JC60	93.3855	3069
SENIOR ENGINEER	93.374	3773
PRODUCT ENGINEER	93.2422	2634
ENGINEER	93.0581	4941
NETWORK AND COMPUTER SYSTEMS ADMINISTRATOR	93.0498	1928
LEAD SOFTWARE ENGINEER	93.0025	1572
OCCUPATIONAL THERAPIST	92.7203	4437
SENIOR APPLICATION DEVELOPER	92.6209	1965
SENIOR PROJECT MANAGER	92.6108	1015
CLINICAL FELLOW	92.5829	1146
INDUSTRIAL DESIGNER	92.567	3619
PROJECT ENGINEER	92.561	6439
PHYSICAL THERAPIST	92.5372	20207
APPLICATIONS ENGINEER	92.5355	1688
SOLUTION ARCHITECT	92.5276	1994
HOSPITALIST	92.5006	4387
TEST LEAD	92.4102	1726
DIRECTOR	92.3481	1333
HOSPITALIST PHYSICIAN	92.3285	4067
SENIOR FINANCIAL ANALYST	92.2241	1196
SYSTEM ENGINEER	92.2145	2145
MANAGER JC50	91.889	1874
MEDICAL RESIDENT	91.7808	2336
PSYCHIATRIST	91.5438	1289
COMPUTER SOFTWARE ENGINEER	91.5052	2684
SCIENTIST	91.4925	1340
RESIDENT	91.4859	1245
APPLICATION DEVELOPER	91.4587	7692
TECHNICAL PROJECT MANAGER	91.2548	1052
PROCESS ENGINEER	91.2269	4377
PHYSICIAN	91.0573	4417

CONCLUSION

Therefore, from the given dataset of H1-B Applicants within the years 2011-2016, which contained around 3 billion/30 lakh records, we have done a complete analysis on various factors and criteria which has some predictive nature. This predictive analysis on the H1-B visa dataset has revolved around finding top Occupation, States, Employers and Industries that contribute to highest number of H1-B visa applications. This clearly indicates the trends of H1-B Visa filings which has a high correlation with the employer's acceptance rate.