# [C:\Users\jwoo5\AppData\Local\Temp\templateTermTutorial.html](http://www.calstatela.edu/centers/hipic) CIS5200 Term Project Tutorial

California State University, Los Angeles

#### Authors: Ashish Solanki, Vinayak Gaur, Rakshit Makkar, Devashish Nayak

#### Instructor: [Jongwook Woo](https://www.linkedin.com/in/jongwook-woo-7081a85)

#### Date: 12/15/2018

**Lab Tutorial**

Ashish Solanki ([asolank5@calstatela.edu](mailto:asolank5@calstatela.edu)),

Vinayak Gaur ([vgaur@calstatela.edu](mailto:vgaur@calstatela.edu)),

Rakshit Makkar ([rmakkar@calstatela.edu](mailto:rmakkar@calstatela.edu)),

Devashish Nayak ([dnayak@calstatela.edu](mailto:dnayak@calstatela.edu))

**Patent Assignment Dataset**

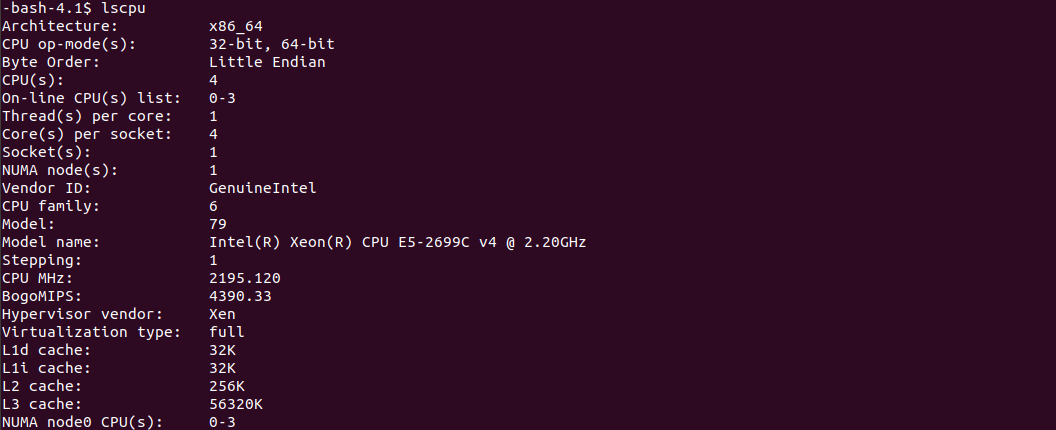
**Objectives**

**List what your objectives are.** In this project, we get to learn

* Analysis of machine learning and AI related patents granted
* HiveQL commands to perform the analysis.
* Visualization

**Platform Spec**

* Hadoop cluster- Oracle Classic Compute Edition
* Hive
* Power BI
* CPU Speed: 2.2 GHz
* # of CPU cores: 8
* # of nodes: 5 nodes
* Total Memory Size: 150 GB
* Data size : 12.2 GB



Step 1: Get data from USPTO website on Hadoop using wget

* Open **ssh** [**[username]@129.150.20**](mailto:dnayak@129.150.206.199)**6.199**
* Entered the **password**:
* Then, we need to upload the dataset from USPTO to our Hadoop local cluster. For that follow the below commands
* Create a directory, we named it Group4

**-bash-4.1$ hdfs dfs -mkdir /user/dnayak/pro\_group4**

* To download the dataset on Hadoop cluster, below are the following commands

1. -bash-4.1$ wget <https://bulkdata.uspto.gov/data/patent/grant/redbook/2018/I20180109.tar>
2. -bash-4.1$ wget https://bulkdata.uspto.gov/data/patent/grant/redbook/2018/I20180109.tar
3. -bash-4.1$ wget https://bulkdata.uspto.gov/data/patent/grant/redbook/2018/I20180116.tar

* Downloaded few more using wget command.

Step 2: Extract the dataset from zip files and create Tables to store the data in Hive Tables

* **Unzip the files using below commands, since these are zip files**

1. -bash-4.1$ unzip I201080109.zip
2. -bash-4.1$ unzip I20180109.tar
3. -bash-4.1$ unzipI20180116.tar

* ls -l to see all the files
* Download these XLM files into your desktop using winscp.
* Process these file using Python to get the CSV file. using this GitHub link.

<https://github.com/funginstitute/patentprocessor>

* After you convert these file upload it to your server to use it on hadoop server using winscp.
* **Move all the csv files to group4 directory**

1. hdfs dfs -put patent.csv /user/dnayak/project4/patent.csv
2. hdfs dfs -put patdesc.csv /user/dnayak/project4/patdesc.csv
3. hdfs dfs -put lawyer.csv /user/dnayak/project4/lawyer.csv
4. hdfs dfs -put inventor.csv /user/dnayak/project4/inventor.csv
5. hdfs dfs -put citation.csv /user/dnayak/project4/citation.csv
6. hdfs dfs -put assignee.csv /user/dnayak/project4/assignee.csv
7. hdfs dfs -put invpat.csv /user/dnayak/project4/invpat.csv

Step 3: Create Tables to store the data in Hive Tables

* Run the following command to make beeline works

hdfs dfs -chmod -R o+w .

* **Open another bash and connect using ssh command for beeline**
* Open **ssh** [**[username]@129.150.206.199**](mailto:dnayak@129.150.205.113)
* Entered the **password**:
* !connect jdbc:hive2://cis5200s3-bdcsce-4.compute-608214094.oraclecloud.internal:2181,cis5200s3-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5mailto:asolank5@129.150.205.113200s3-bdcsce-3.compute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce\_admin
* To open Beeline, please type **beeline** at bash
* Since this is a common cluster, therefore select your database using following command **use [username];**
* **show tables;**
* **Now create tables for each csv file and load data in it**

1. DROP TABLE IF EXISTS citation;

--create the citation table on comma-separated sensor data

CREATE EXTERNAL TABLE IF NOT EXISTS citation (Patent BIGINT, Cit\_Date DATE,

Cit\_Name STRING,

Cit\_Kind STRING,

Cit\_Country STRING,

Citation STRING,

Category STRING,

CitSeq BIGINT)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/dnayak/project4/citation'

TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA INPATH '/user/dnayak/project4/citation.csv' OVERWRITE INTO TABLE citation;

2. DROP TABLE IF EXISTS assignee;

CREATE EXTERNAL TABLE IF NOT EXISTS assignee (Patent BIGINT, AsgTypee INT,

Assignee STRING,

City STRING,

Country STRING,

Nationality STRING,

Residence STRING,

AsgSeq BIGINT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

LOCATION '/user/dnayak/project4/assignee'

TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA INPATH '/user/dnayak/project4/assignee.csv' OVERWRITE INTO TABLE assignee;

3. DROP TABLE IF EXISTS invpat;

CREATE EXTERNAL TABLE IF NOT EXISTS invpat (Firstname STRING,

Middlename STRING,

Lastname STRING,

Street STRING,

City STRING,

State STRING,

Country STRING,

Zipcode STRING,

Latitude FLOAT,

Longitube FLOAT,

InvSeq BIGINT,

AppYear STRING,

ApplyYear STRING,

GYear INT,

AppDate DATE,

Assignee STRING,

AsgNum STRING,

Class STRING,

Coauthor STRING,

Invnum STRING,

Invnum\_N STRING,

Unique\_Record\_ID STRING

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/dnayak/project4/invpat'

TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA INPATH '/user/dnayak/project4/invpat.csv' OVERWRITE INTO TABLE invpat;

4. DROP TABLE IF EXISTS inventor ;

CREATE EXTERNAL TABLE IF NOT EXISTS inventor (Patent BIGINT,

Firstname STRING,

Middlename STRING,

Lastname STRING,

Street STRING,

City STRING,

State STRING,

Country STRING,

Zipcode STRING,

Nationality STRING,

InvSeq BIGINT

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/dnayak/project4/inventor'

TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA INPATH '/user/dnayak/project4/inventor.csv' OVERWRITE INTO TABLE inventor;

5. DROP TABLE IF EXISTS lawyer ;

CREATE EXTERNAL TABLE IF NOT EXISTS lawyer(Patent VARCHAR(40),

Firstname STRING,

Lastname STRING,

LawCountry STRING,

OrgName STRING,

LawSeq STRING

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/dnayak/project4/lawyer'

TBLPROPERTIES ('skip.header.line.count'='1')

;

LOAD DATA INPATH '/user/dnayak/project4/lawyer.csv' OVERWRITE INTO TABLE lawyer;

6. DROP TABLE IF EXISTS patent ;

CREATE EXTERNAL TABLE IF NOT EXISTS patent(Patent BIGINT,

Kind STRING,

Claims BIGINT,

AppType STRING,

AppNum STRING,

Gdate DATE,

GYear DATE,

AppDate DATE,

AppYear DATE,

pattype STRING

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/dnayak/project4/patent'

TBLPROPERTIES ('skip.header.line.count'='1');

LOAD DATA INPATH '/user/dnayak/project4/patent.csv' OVERWRITE INTO TABLE patent;

7. DROP TABLE IF EXISTS patdesc ;

CREATE EXTERNAL TABLE IF NOT EXISTS patdesc(Patent STRING,

Abstract STRING,

Title STRING

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/user/dnayak/project4/patdesc'

TBLPROPERTIES ('skip.header.line.count'='2');

LOAD DATA INPATH '/user/dnayak/project4/patdesc.csv' OVERWRITE INTO TABLE patdesc;

* **show tables;**

****

Step 4: Running Hive queries to sort, clean and get the output

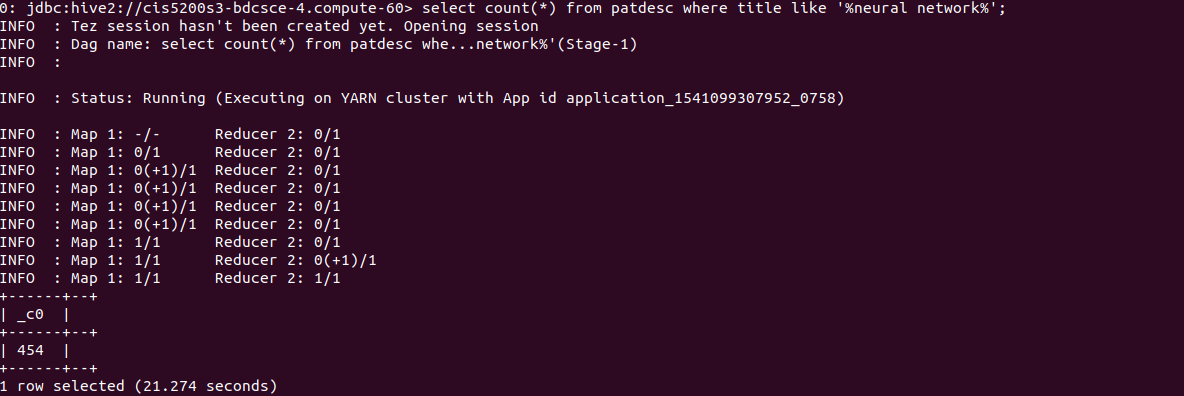
* **To replace empty string with null** 
  1. alter table assignee set tblproperties ('serialization.null.format'='');
  2. alter table patdesc set tblproperties ('serialization.null.format'='');
  3. alter table lawyer set tblproperties ('serialization.null.format'='');
  4. alter table inventor set tblproperties ('serialization.null.format'='');
  5. alter table patent set tblproperties ('serialization.null.format'='');
* Up until here, data cleaning part is completed. Now we need to generate desired output and to store them in table format in order to create visualization using Power BI. For doing that, below are the following queries to first get the output and then creating new table with different name and then storing output result in that created table

Step 5: To store the desired output in new tables and save them as csv

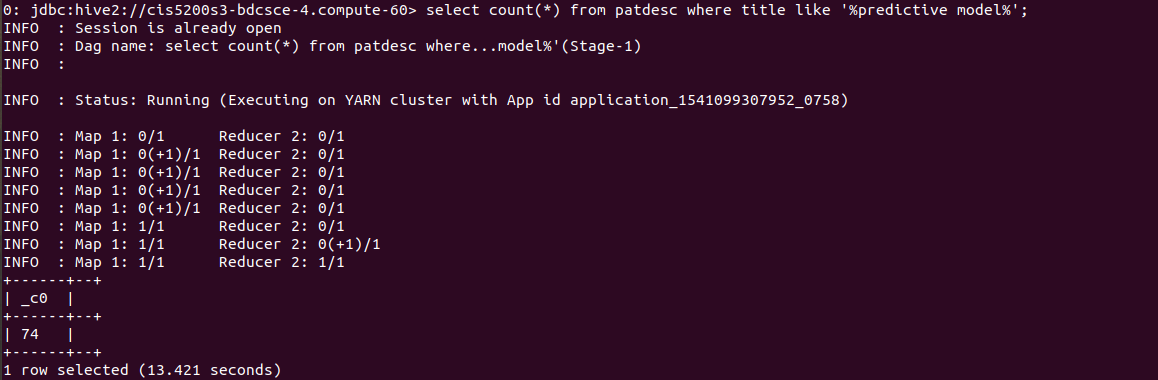
**Analysis:**

* **Here we are checking in patent description table for these keyword.**

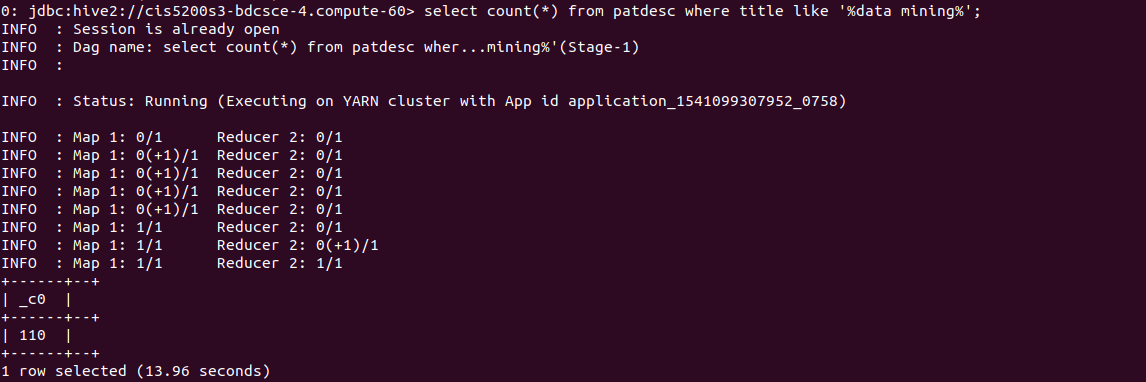
0: jdbc:hive2://cis5200s3-bdcsce-4.compute-60> select count(\*) from patdesc where title like '%neural network%';



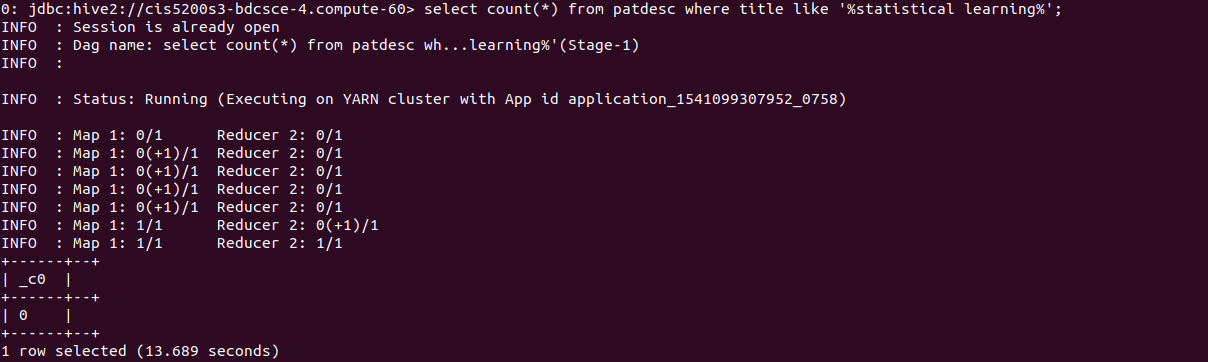
0: jdbc:hive2://cis5200s3-bdcsce-4.compute-60> select count(distinct title) from patdesc where title like '%predictive model%';

****

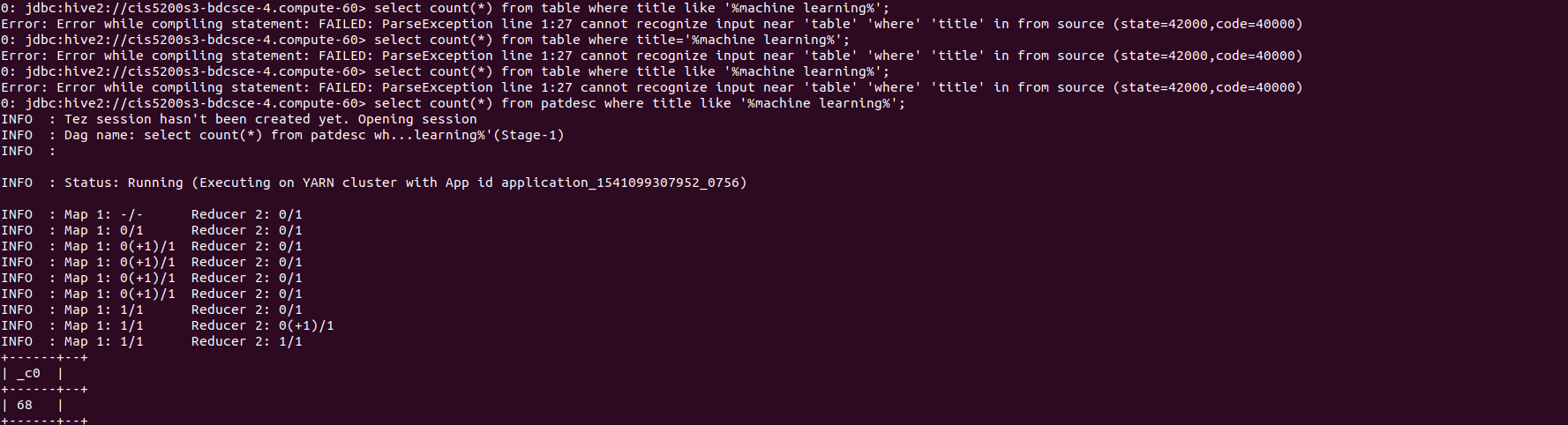
0: jdbc:hive2://cis5200s3-bdcsce-4.compute-60> select count(\*) from patdesc where title like '%deep learning%';



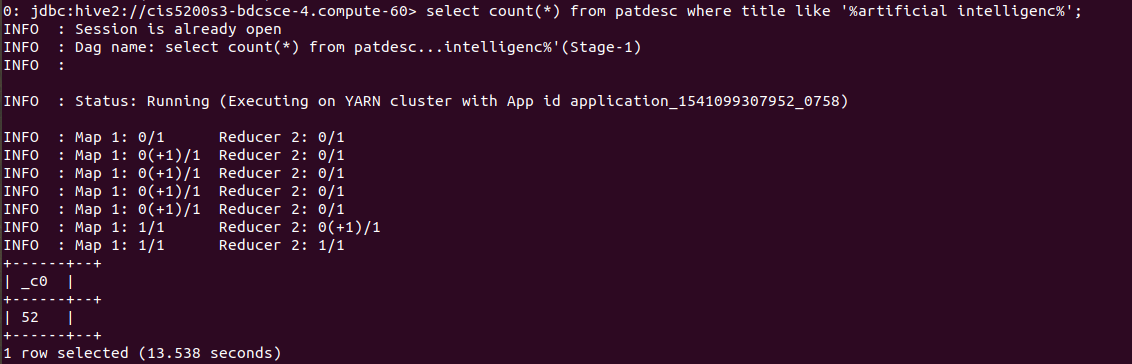
0: jdbc:hive2://cis5200s3-bdcsce-4.compute-60> select count(\*) from patdesc where title like '%statistical learning%';



0: jdbc:hive2://cis5200s3-bdcsce-4.compute-60> select count(distinct title) from patdesc where title like '%machine learning%';



0: jdbc:hive2://cis5200s3-bdcsce-4.compute-60> select count(\*) from patdesc where title like '%artificial intelligence%';



* **Please type below queries one by one to get the output and store them in tables**

1. The first Query we are taking output of the analysis into new table where we have created the table with two columns 1st one is for Keywords and 2nd column is for counts.

**DROP TABLE IF EXISTS patentcount ;**

**CREATE EXTERNAL TABLE IF NOT EXISTS patentcount(Title STRING, Count BIGINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','LOCATION '/user/dnayak/project4/patentcount';**

**INSERT INTO TABLE patentcount (title, count) SELECT 'Machine Learning', count(distinct title) as count FROM dnayak.patdesc WHERE title like '%machine learning%';**

**INSERT INTO TABLE patentcount (title, count) SELECT 'Predictive Model', count(distinct title) as count FROM dnayak.patdesc WHERE title like '%predictive model%';**

**INSERT INTO TABLE patentcount (title, count) SELECT 'Data Mining', count(distinct title) as count FROM dnayak.patdesc WHERE title like '%data mining%';**

**INSERT INTO TABLE patentcount (title, count) SELECT 'Artificial Intelligence', count(distinct title) as count FROM dnayak.patdesc WHERE title like '%artificial intelligence%';**

**INSERT INTO TABLE patentcount (title, count) SELECT 'Deep Learning', count(distinct title) as count FROM dnayak.patdesc WHERE title like '%deep learning%';**

**INSERT INTO TABLE patentcount (title, count) SELECT 'Statistical Learning', count(distinct title) as count FROM dnayak.patdesc WHERE title like '%statistical learning%';**

**INSERT INTO TABLE patentcount (title, count) SELECT 'Neural Network', count(distinct title) as count FROM dnayak.patdesc WHERE title like '%neural network%';**

**Downloading Data into your PC**

After the Hive tables are created, you can download it to your lab (or personal PC/Laptop) as follows.

1. Open another terminal with git bash, minty, or putty, which is to connect the Oracle cloud to

download the output file 000000\_0 at the HDFS path “/user/jwoo5/tmp/tweets\_sentiment ”:

$ ssh jusername@ipaddress

password:

-bash-4.1$ cd patentcount/

-bash-4.1$ ls -al

total 40

drwxrwxr-x 2 dnayak dnayak 4096 Dec 5 11:38 .

drwx------ 10 dnayak dnayak 4096 Dec 5 11:37 ..

-rw-r--r-- 1 dnayak dnayak 20 Dec 5 11:37 000000\_0

-rw-r--r-- 1 dnayak dnayak 20 Dec 5 11:37 000000\_0\_copy\_1

-rw-r--r-- 1 dnayak dnayak 15 Dec 5 11:37 000000\_0\_copy\_2

-rw-r--r-- 1 dnayak dnayak 27 Dec 5 11:37 000000\_0\_copy\_3

-rw-r--r-- 1 dnayak dnayak 16 Dec 5 11:37 000000\_0\_copy\_4

-rw-r--r-- 1 dnayak dnayak 24 Dec 5 11:37 000000\_0\_copy\_5

-rw-r--r-- 1 dnayak dnayak 19 Dec 5 11:37 000000\_0\_copy\_6

2. Download the 2 output files to the local file systems:

-bash-4.1$ hdfs dfs -get /user/dnayak/project4/patentcount patentcount

-bash-4.1$ cd patentcount/

-bash-4.1$ ls -al

total 40

drwxrwxr-x 2 dnayak dnayak 4096 Dec 5 11:38 .

drwx------ 10 dnayak dnayak 4096 Dec 5 11:37 ..

-rw-r--r-- 1 dnayak dnayak 20 Dec 5 11:37 000000\_0

-rw-r--r-- 1 dnayak dnayak 20 Dec 5 11:37 000000\_0\_copy\_1

-rw-r--r-- 1 dnayak dnayak 15 Dec 5 11:37 000000\_0\_copy\_2

-rw-r--r-- 1 dnayak dnayak 27 Dec 5 11:37 000000\_0\_copy\_3

-rw-r--r-- 1 dnayak dnayak 16 Dec 5 11:37 000000\_0\_copy\_4

-rw-r--r-- 1 dnayak dnayak 24 Dec 5 11:37 000000\_0\_copy\_5

-rw-r--r-- 1 dnayak dnayak 19 Dec 5 11:37 000000\_0\_copy\_6

3. Now you have to merge the two output files to a file named “senti\_out.csv” using the following:

-bash-4.1$ cat 000000\_0 000000\_0\_copy\_1 000000\_0\_copy\_2 000000\_0\_copy\_3 000000\_0\_copy\_4 000000\_0\_copy\_5 000000\_0\_copy\_6 > patentcount.csv

-bash-4.1$ ls -al

total 40

drwxrwxr-x 2 dnayak dnayak 4096 Dec 5 11:38 .

drwx------ 10 dnayak dnayak 4096 Dec 5 11:37 ..

-rw-r--r-- 1 dnayak dnayak 20 Dec 5 11:37 000000\_0

-rw-r--r-- 1 dnayak dnayak 20 Dec 5 11:37 000000\_0\_copy\_1

-rw-r--r-- 1 dnayak dnayak 15 Dec 5 11:37 000000\_0\_copy\_2

-rw-r--r-- 1 dnayak dnayak 27 Dec 5 11:37 000000\_0\_copy\_3

-rw-r--r-- 1 dnayak dnayak 16 Dec 5 11:37 000000\_0\_copy\_4

-rw-r--r-- 1 dnayak dnayak 24 Dec 5 11:37 000000\_0\_copy\_5

-rw-r--r-- 1 dnayak dnayak 19 Dec 5 11:37 000000\_0\_copy\_6

-rw-rw-r-- 1 dnayak dnayak 141 Dec 5 11:38 patentcount.csv

4. Open another terminal with git bash, minty, or putty in order to read/import the output file

using your lab computer (or your PC/Laptop) - you have to download the file to your lab

computer (or your PC/Laptop). file located at /home/dnayak/patentcount/patentcount.csv .

scp dnayak@129.150.205.113:/home/dnayak/patentcount/patentcount.csv /home/drsherman/Downloads/

dnayak@129.150.205.113's password:

patentcount.csv 100% 141 2.6KB/s 00:00

**Do same process for other files.**

1. The patents grants for the top companies which are in it industries like google, amazon...

**DROP TABLE IF EXISTS frq\_assignee ;**

**CREATE EXTERNAL TABLE IF NOT EXISTS frq\_assignee (Patent STRING, Count BIGINT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ','LOCATION '/user/dnayak/project4/frq\_assignee';**

**INSERT INTO TABLE frq\_assignee (patent, count) select 'Google Inc.', count(distinct patent) as count from dnayak.assignee where assignee like '%Google Inc.%' and patent is not NULL;**

**INSERT INTO TABLE frq\_assignee (patent, count) select 'Amazon.com', count(distinct patent) as count from dnayak.assignee where assignee like '%Amazon.com%' and patent is not NULL;**

**INSERT INTO TABLE frq\_assignee (patent, count) select 'Facebook', count(distinct patent) as count from dnayak.assignee where assignee like '%Facebook%' and patent is not NULL;**

**INSERT INTO TABLE frq\_assignee (patent, count) select 'AT&T', count(distinct patent) as count from dnayak.assignee where assignee like '%AT&T%' and patent is not NULL;**

**INSERT INTO TABLE frq\_assignee (patent, count) select 'Apple Inc.', count(distinct patent) as count from dnayak.assignee where assignee like '%Apple Inc.%' and patent is not NULL;**

**INSERT INTO TABLE frq\_assignee (patent, count) select 'SONY Corporation', count(distinct patent) as count from dnayak.assignee where assignee like '%SONY Corporation%' and patent is not NULL;**

**INSERT INTO TABLE frq\_assignee (patent, count) select 'Samsung', count(distinct patent) as count from dnayak.assignee where assignee like '%Samsung%' and patent is not NULL;**

**“Do same process as above to get the output file.”**

1. Like the above query, to find out the patents assignee related to machine learning and artificial intelligence, below is the query to get the output thought which next when we are doing the visualization we make it according to countries visuilaization.

**CREATE TABLE IF NOT EXISTS patent\_con ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/dnayak/project4/patent\_con' AS**

**select a.\* from assignee a,patdesc p where a.patent=p.patent and (p.title like '%machine learning%' or p.title like '%artificial intelligenc%' or p.title like '%deep learning%' or p.title like '%data mining%' or p.title like '%data mining%' or p.title like '%stastistical learning%' or p.title like '%neural network%' or p.title like '%predictive model%');**

**“Do same process as above to get the output file.”**

1. Query to obtain the Inventor related to ML and AI .

**CREATE TABLE IF NOT EXISTS patent\_inv ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/dnayak/project4/patent\_inv' AS**

**select a.\* from inventor a,patdesc p where a.patent=p.patent and (p.title like '%machine learning%' or p.title like '%artificial intelligenc%' or p.title like '%deep learning%' or p.title like '%data mining%' or p.title like '%data mining%' or p.title like '%stastistical learning%' or p.title like '%neural network%' or p.title like '%predictive model%');**

**“Do same process as above to get the output file.”**

**Visualization Part :**

References

1. URL of Data Source,
2. https://github.com/funginstitute/patentprocessor
3. URL of your Github, https://github.com/Devs02/CSI-5200-03-Project-Group-4-part2