Commands for the project

Step 1: ssh [asolank5@129.150.205.113](mailto:asolank5@129.150.205.113)

Step 2: Entered the password

Step 3: hdfs dfs -mkdir group4

**Step 4: Download the dataset from USPTO**

1. wget <https://bulkdata.uspto.gov/data/patent/assignment/economics/2017/documentid_admin.csv.zip>
2. wget <https://bulkdata.uspto.gov/data/patent/assignment/economics/2017/documentid.csv.zip>
3. wget <https://bulkdata.uspto.gov/data/patent/assignment/economics/2017/assignee.csv.zip>
4. wget <https://bulkdata.uspto.gov/data/patent/assignment/economics/2017/assignment.csv.zip>
5. wget <https://bulkdata.uspto.gov/data/patent/assignment/economics/2017/assignor.csv.zip>
6. wget <https://bulkdata.uspto.gov/data/patent/assignment/economics/2017/assignment_conveyance.csv.zip>

**Step 4: Unzip the files**

1. unzip documentid\_admin.csv.zip
2. unzip documentid.csv.zip
3. unzip assignee.csv.zip
4. unzip assignment.csv.zip
5. unzip assignor.csv.zip
6. unzip assignment\_conveyance.csv.zip

**Step 5:** ls -l to see all the files

**Step 6: Move all the csv files to group4 directory**

1. hdfs dfs -put documentid\_admin.csv group4
2. hdfs dfs -put documentid.csv group4
3. hdfs dfs -put assignee.csv group4
4. hdfs dfs -put assignment.csv.csv group4
5. hdfs dfs -put assignor.csv group4
6. hdfs dfs -put assignment\_conveyance.csv group4

**Step 7: Open another bash and connect using ssh command for beeline**

**Step 8:** !connect jdbc:hive2://cis5200s3-bdcsce-4.compute-608214094.oraclecloud.internal:2181,cis5200s3-bdcsce-2.compute-608214094.oraclecloud.internal:2181,cis5200s3-bdcsce-3.compute-608214094.oraclecloud.internal:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive bdcsce\_admin

**Step 9: beeline**

**Step 10: use asolank5;**

**Step 11: show tables;**

**Step 12: Now create tables for each csv file and load data in it**

1. CREATE EXTERNAL TABLE IF NOT EXISTS assignee (rf\_id INT, ee\_name STRING, ee\_address\_1 STRING, ee\_address\_2 STRING, ee\_city STRING, ee\_state STRING, ee\_postcode INT, ee\_country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',’;

LOAD DATA INPATH '/user/asolank5/group4/assignee.csv' INTO TABLE assignee;

1. CREATE EXTERNAL TABLE IF NOT EXISTS assignment (rf\_id INT, file\_id INT, cname STRING, caddress\_1 STRING, caddress\_2 STRING, caddress\_3 STRING, caddress\_4 STRING, reel\_no INT, frame\_no INT, convey\_text STRING, record\_dt DATE, last\_update\_dt DATE, page\_count INT, purge\_in STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/asolank5/group4/assignment';

LOAD DATA INPATH '/user/asolank5/group4/assignment' INTO TABLE assignment;

1. CREATE EXTERNAL TABLE IF NOT EXISTS assignment\_conveyance (rf\_id INT, convey\_ty STRING, employer\_assign INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/asolank5/group4/assignment\_conveyance';

LOAD DATA INPATH '/user/asolank5/group4/assignment\_conveyance' INTO TABLE assignment\_conveyance;

1. CREATE EXTERNAL TABLE IF NOT EXISTS assignor (rf\_id INT, or\_name STRING, exec\_dt DATE, ack\_dt DATE) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LOCATION '/user/asolank5/Group4/assignor';

LOAD DATA INPATH '/user/asolank5/group4/assignor' INTO TABLE assignor;

1. CREATE EXTERNAL TABLE IF NOT EXISTS documentid (rf\_id INT, title STRING, lang STRING, appno\_doc\_num INT, appno\_date DATE, appno\_country STRING, pgpub\_doc\_num BIGINT, pgpub\_date DATE, pgpub\_country STRING, grant\_doc\_num INT, grant\_date DATE, grant\_country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA INPATH '/user/asolank5/group4/ document\_id.csv' INTO TABLE document\_id;

1. CREATE EXTERNAL TABLE IF NOT EXISTS documented\_admin (rf\_id INT,appno\_doc\_num INT, grant\_doc\_num INT, admin\_appl\_id\_for\_grant INT, amin\_pat\_no\_for\_appno INT, error STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

LOAD DATA INPATH '/user/asolank5/group4/documentid\_admin.csv' INTO TABLE documentid\_admin;

**Step 13: Running Hive queries to sort, clean and get the output**

1. //to replace empty string with null

alter table assignment\_conveyance set tblproperties ('serialization.null.format'='');

alter table assignee set tblproperties ('serialization.null.format'='');

alter table documentid\_admin set tblproperties ('serialization.null.format'='');

alter table documentid set tblproperties ('serialization.null.format'='');

alter table assignor set tblproperties ('serialization.null.format'='');

alter table assignment set tblproperties ('serialization.null.format'='');

1. select count(ee\_country) Total\_Count\_Of\_Country, ee\_country from assignee GROUP BY ee\_country;
2. SELECT COUNT(rf\_id) Total\_RFID, convey\_ty from assignment\_conveyance GROUP BY convey\_ty;
3. SELECT ac.convey\_ty, ae.ee\_country FROM assignment\_conveyance ac JOIN assignee ae ON (ac.rf\_id = ae.rf\_id) WHERE ae.ee\_country NOT LIKE "%NOT PROVIDED%" LIMIT 20;
4. SELECT ac.convey\_ty, ae.ee\_country FROM assignment\_conveyance ac JOIN assignee ae ON (ac.rf\_id = ae.rf\_id) WHERE ae.ee\_country NOT LIKE "%NOT PROVIDED%" LIMIT 20;

**Step 14: To store the desired output in new tables and save them as csv**

1. CREATE TABLE IF NOT EXISTS correct\_patents\_assigned ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/asolank5/group4/patent\_assignment/' AS select count(ee\_country), ee\_country from assignee where rf\_id in (select rf\_id from assignment\_conveyance where convey\_ty like "%correct%") AND ee\_country not like "%NOT PROVIDED%" group by ee\_country;
2. CREATE TABLE IF NOT EXISTS rf\_id\_ConveyanceCorrect ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION

'/user/asolank5/group4/rf\_id\_ConveyanceCorrect/' AS

select count(ee\_country), ee\_country from assignee where rf\_id in (select rf\_id from assignment\_conveyance where convey\_ty like "%correct%") AND ee\_country not like "%NOT PROVIDED%" group by ee\_country;

1. CREATE TABLE IF NOT EXISTS rf\_id\_ConveyanceMerger ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/asolank5/group4/rf\_id\_ConveyanceMerger/' AS

select count(rf\_id), ee\_country from assignee where rf\_id in (select rf\_id from assignment\_conveyance where convey\_ty like "%merger%") AND ee\_country not like "%NOT PROVIDED%" group by ee\_country;

1. CREATE TABLE IF NOT EXISTS rf\_id\_ConveyanceGovern ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/asolank5/group4/rf\_id\_ConveyanceGovern/' AS

select count(rf\_id), ee\_country from assignee where rf\_id in (select rf\_id from assignment\_conveyance where convey\_ty like "%govern%") AND ee\_country not like "%NOT PROVIDED%" group by ee\_country;

1. CREATE TABLE IF NOT EXISTS rf\_id\_ConveyanceSecurity ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/asolank5/group4/rf\_id\_ConveyanceSecurity/' AS

select count(rf\_id), ee\_country from assignee where rf\_id in (select rf\_id from assignment\_conveyance where convey\_ty like "%security%") AND ee\_country not like "%NOT PROVIDED%" group by ee\_country;

1. CREATE TABLE IF NOT EXISTS rfId\_BasedOnCountry ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED AS TEXTFILE LOCATION '/user/asolank5/group4/rfId\_BasedOnCountry/' AS select count(ee\_country) Total\_Count\_Of\_Country, ee\_country from assignee GROUP BY ee\_country;