1. Open your computer’s command prompt or terminal and SSH into oracle server.

ssh [your\_username]@129.150.64.74

1. Change directory to a folder that can handle large files. For this example we will be using /dev/shm directory.

$ cd /dev/shm

1. Download the data set from amazon S3 and extract.

$ wget <https://s3-us-west-1.amazonaws.com/cis.group.project/311_Service_Requests.zip>

$ unzip 311\_Service\_Requests.zip

1. Create folders in hadoop for data set and upload.

$ hdfs dfs -mkdir Services

$ hdfs dfs -put 311\_Service\_Requests\_from\_2010\_to\_Present.csv Services

1. Change file permissions so that it can be used. Be sure to include the . at the end.

$ hdfs dfs -chmod -R o+w .

1. Check to see if files are inside the Services folder.

$ hdfs dfs -ls Services

1. Type beeline to open the Hive interface.

$ beeline

1. Run the following command to access the class database.

!connect jdbc:hive2://bigdai-nov-bdcsce-1:2181,bigdai-nov-bdcsce-2:2181,bigdai-nov-bdcsce-3:2181/;serviceDiscoveryMode=zooKeeper;zooKeeperNamespace=hiveserver2?tez.queue.name=interactive  bdcsce\_admin

1. View and use your database.

beeline> show databases;

beeline> use [your\_username];

1. Copy and paste the following code to create a table name Service. The table created is from the CSV file using the Hive query.

DROP TABLE IF EXISTS Services;

--create the building table on comma-separated building data

CREATE EXTERNAL TABLE IF NOT EXISTS Services(Unique\_Key string,

Created\_Date date,

Closed\_Date date,

Agency string,

Agency\_Name string,

Complaint\_Type string,

Descriptor string,

Location\_Type string,

Incident\_Zip string,

Incident\_Address string,

Street\_Name string,

Cross\_Street\_1 string,

Cross\_Street\_2 string,

Intersection\_Street\_1 string,

Intersection\_Street\_2 string,

Address\_Type string,

City string,

Landmark string,

Facility\_Type string,

Status string,

Due\_Date date,

Resolution\_Description string,

Resolution\_Action\_Updated\_Date date,

Community\_Board string,

BBL string,

Borough string,

X\_Coordinate\_State\_Plane int,

Y\_Coordinate\_State\_Plane int,

Open\_Data\_Channel\_Type string,

Park\_Facility\_Name string,

Park\_Borough string,

Vehicle\_Type string,

Taxi\_Company\_Borough string,

Taxi\_Pick\_Up\_Location string,

Bridge\_Highway\_Name string,

Bridge\_Highway\_Direction string,

Road\_Ramp string,

Bridge\_Highway\_Segment string,

Latitude double,

Longitude double,

Location string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED AS TEXTFILE LOCATION '/user/bascenc8/Services'

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

1. .We will then create a complaints table using specific columns from the data set and overwrite  with a new Hive query.

DROP TABLE IF EXISTS Complaints;

CREATE EXTERNAL TABLE IF NOT EXISTS Complaints(agency\_name string, Complaint\_type string,  NumberofComplaints int)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ', '

STORED AS TEXTFILE LOCATION '/user/bascenc8/Services/complaints.txt';

INSERT OVERWRITE TABLE Complaints

select agency\_name, complaint\_type, count(complaint\_type) as NumberofComplaints from services group by agency\_name, complaint\_type having count(complaint\_type) > 5 order by NumberofComplaints DESC;

1. Check your database tables to see if you have the following tables: Services, Complaints

beeline> show tables;

1. Let’s view some data from the newly created table. Enter the following code.

SELECT \*

FROM complaints

LIMIT 10;

SELECT \*

FROM complaints

WHERE agency\_name

LIKE 'New York City Police Department'

LIMIT 10;

SELECT complaint\_type, numberofcomplaints

FROM complaints

WHERE agency\_name

LIKE 'New York City Police Department'

LIMIT 10;

SELECT complaint\_type, numberofcomplaints

FROM Complaints

WHERE agency\_name

LIKE 'New York City Police Department'

LIMIT 10;

SELECT agency\_name, COUNT(agency\_name) AS complaints\_handled

FROM complaints

GROUP BY agency\_name

ORDER BY complaints\_handled DESC

LIMIT 10;