**Hive Code:**

beeline

!connect jdbc:hive2://babar.es.its.nyu.edu:10000/

use NetID;

**Create temporary functions and set number of mappers and reducers:**

set hive.input.format=org.apache.hadoop.hive.ql.io.HiveInputFormat;

set mapred.map.tasks = 20;

set mapred.reduce.tasks = 20;

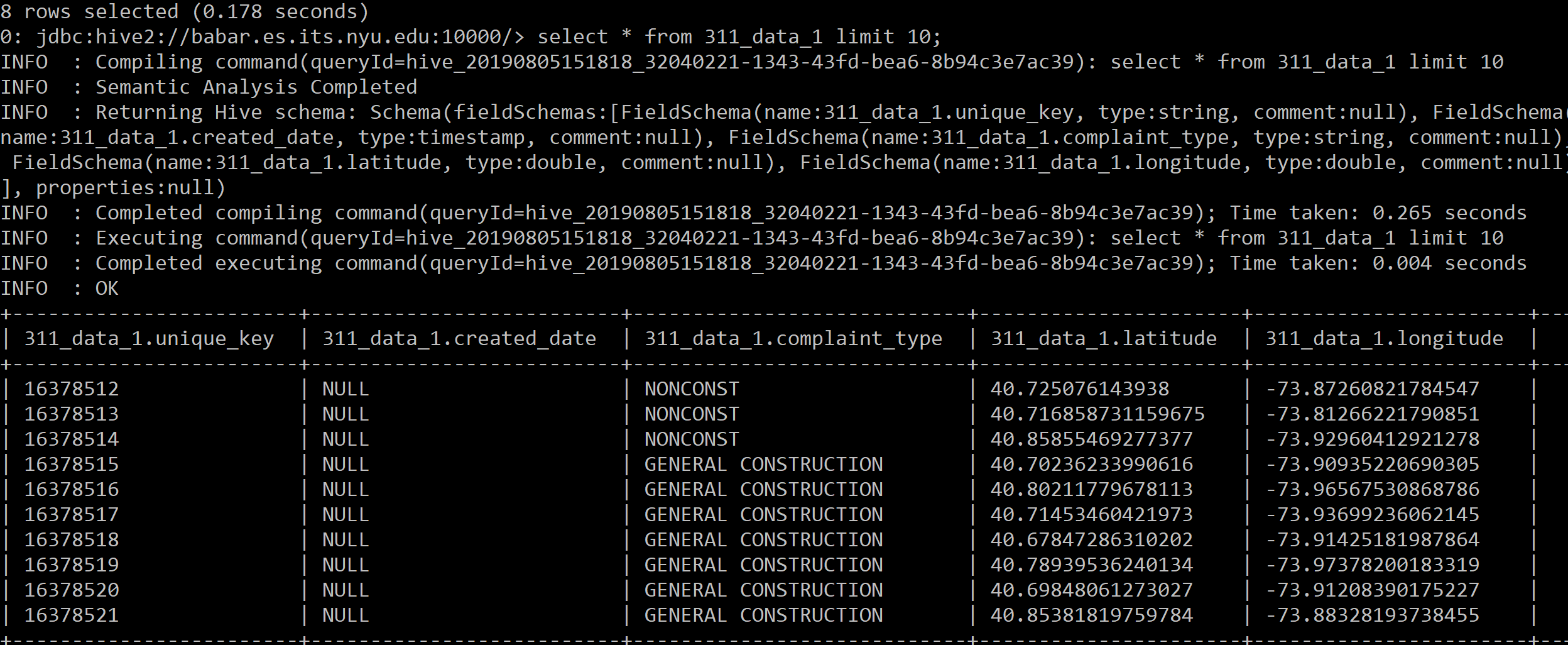
create temporary function st\_polygon as 'com.esri.hadoop.hive.ST\_Polygon';

create temporary function st\_point as 'com.esri.hadoop.hive.ST\_Point';

create temporary function st\_contains as 'com.esri.hadoop.hive.ST\_Contains';

**Creating 311 Data Hive table**

create external table 311\_data\_1 (unique\_key string, created\_date timestamp, complaint\_type string, latitude double, longitude double) row format delimited fields terminated by ',' location '/user/jl9200/rbda/shareHDFS\_project/data2/';



**Taking out only the complaint\_types is Graffiti from the 311 Data Hive Table**

CREATE TABLE 311\_data\_graffiti AS

SELECT \* FROM 311\_data\_1

WHERE complaint\_type=’Graffiti’;

**Create Zillow Neighborhoods table in hive**

create external table zillow1 (name string, geometry string) row format delimited fields terminated by '\t' location '/user/jl9200/rbda/shareHDFS\_project/zillowtxt';

**Performing spatial join on these two datasets ( zillow1 and 311\_data\_graffiti)**

CREATE TABLE 311\_zillow AS

SELECT zillow1.name, count(\*) cnt FROM zillow1

JOIN 311\_data\_graffiti

WHERE st\_contains(st\_polygon(zillow1.geometry), st\_point(311\_data\_graffiti.longitude, 311\_data\_graffiti.latitude))

GROUP BY zillow1.name

ORDER BY cnt desc;



**Export the HIVE TABLE to HDFS**

INSERT OVERWRITE DIRECTORY '/user/jl9200/rbda/shareHDFS\_project/output1' SELECT \* FROM 311\_zillow;

\*\* '/user/jl9200/rbda/shareHDFS\_project/output1' should be a new directory that had never been created before!

**Export table from hdfs to dumbo (This Command is run on dumbo not beeline)**

hdfs dfs -cat 311\_zillow/000000\_0 > 311\_zillow.csv

**Clean Output csv on dumbo (This Command is run on dumbo not beeline)**

python clean\_output.py 311\_zillow.csv 311\_zillow\_output.csv

* **Also Repeat the process above for the created\_date data for 311**

select \* from 311\_zillow\_wtime ORDER BY cnt DESC LIMIT 10;

