**Creating NYPD Arrests Hive table**

create external table nypd\_arrests1 (arrest\_key string, arrest\_date string, pd\_cd string, pd\_desc string, ky\_cd string, ofns\_desc string, law\_code string, law\_cat\_cd string, arrest\_bd string, arrest\_precinct string, jusrisdiction\_code string, age\_group string, perp\_sex string, perp\_race string, x\_coord\_cd string,

y\_coord\_cd string, latitude string, longitude string) row format delimited fields terminated by '|' location '/user/sds695/cleaned\_nypd\_arrests/';

**Taking out only relevant columns from the NYPD Arrests table**

create table time\_nypd as select SUBSTRING(nypd\_arrests1.arrest\_date,7,4) as year,nypd\_arrests1.latitude as latitude,nypd\_arrests1.longitude as longitude from nypd\_arrests1;

**Creating Zillow Neighborhoods table**

create external table zillow1 (fid int, state string, county string, city string, name string, redionid string, polygons geometry) row format delimited fields terminated by ',' location '/user/sds695/Zillow\_data/';

**Performing spatial join on these two datasets**

CREATE TABLE nypd\_zillow1 AS SELECT zillow1.name, time\_nypd.year, count(\*) cnt FROM zillow1 JOIN time\_nypd WHERE st\_contains(st\_polygon(zillow1.geometry), st\_point(time\_nypd.longitude, time\_nypd.latitude)) GROUP BY zillow1.name,time\_nypd.year ORDER BY cnt desc;

**Exporting the table to hdfs**

INSERT OVERWRITE DIRECTORY '/user/sds695/nypd\_zillow\_year' select \* from nypd\_zillow1;

**Exporting table from hdfs to dumbo(This command is run on dumbo not beeline)**

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