Spatial Data Science & Engineering Assignment 3

Task:

To implement **ST_GeomFromGeoHash** method in Apache Sedona using UDF. The method should be accessible to be executed from Spark Sql.

Data set Used:
airbnb_Chicago 2015

Steps to generate data:
In Data generator folder there is a python file which converts airbnb_Chicago 2015 dataset into csv file containing, location and hash value of geometry.

The Data generator also generates bench_mark_file_to_test which will contain the expected output of the scala program.

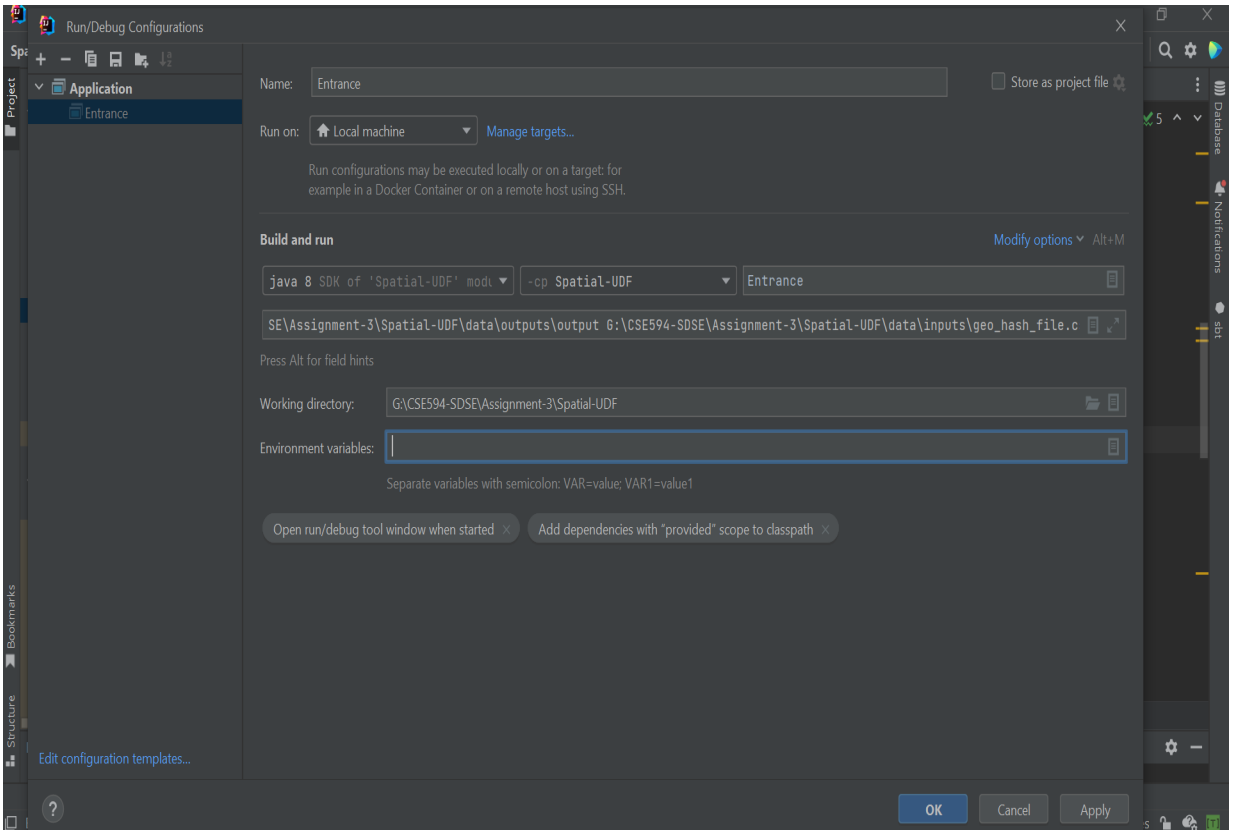
Steps:-

1) Open dataGenerator folder.
2) Load airbnb_Chicago 2015 shape file into inputs folder of data generator (dataGenerator\inputs).
3) Run tester.py file.
4) geo_hash_file.csv and bench_mark_file_to_test files will be generated in the outputs folder of (data generator.dataGenerator\outputs)

Steps to test the scala code:

1) Copy the geo_hash_file.csv file to \Spatial-UDF\data\inputs
2) Copy bench_mark_file_to_test to \Spatial-UDF\data\true-outputs folder.
3) Run Entrance.scala program using the following parameters.
   Output Directory (G:\CSE594-SDSE\Assignment-3\Spatial-UDF\data\outputs\output)
   Input Directory (G:\CSE594-SDSE\Assignment-3\Spatial-UDF\data\inputs\geo_hash_file.csv)

Incase you are using intellij your run configuration will look like this.



4) Now you will be able to see the output in \data\outputs\output folder
5) Use an online comparator like https://text-compare.com/ to compare if both the output file and bench_mark_file_to_test are the same (Only difference is spacing that is because of the way scala and pandas dataframe are handled).