Chicago Taxi Trip Dashboard Implementation

# Set up EC2 Instance for this project

EC2 Instance for Lab1-4 can be used but need to attach a EBS volume of 200GB+ for this project.

# Run load\_data.sh to load data and put them into HDFS

Under root folder

$ git clone <https://github.com/CynthiaHu/W205_Summer.git>

It takes about 5-6 hours to download all data (44GB) and copy them to HDFS, mainly depending on the network speed.

$ su – w205

$ cd W205\_Summer/"Course Project"/1\_Loading\_and\_Modeling

$ . Load\_data.sh

# Save all scripts in /data/taxi\_trip Directory

$ cd ~

$ cd W205\_Summer/"Course Project"/1\_Loading\_and\_Modeling

$ cp hive\_trip\_ddl.sql /data/taxi\_trip

$ cd ~

$ cd W205\_Summer/"Course Project"/2\_Transforming

$ cp trip\_transforming.sql /data/taxi\_trip

$ cd ~

$ cd W205\_Summer/"Course Project"/3\_Aggregation\_and\_Analysis

$ cp trip\_aggregation.sql /data/taxi\_trip

$ cd ~

$ cd W205\_Summer/"Course Project"/4\_Forecast

$ cp python\_setup.sh /data/taxi\_trip

$ cp trip\_forecast.py /data/taxi\_trip

# Run hive\_trip\_ddl.sql to create base Hive tables

It’s quite fast.

$ cd /data/taxi\_trip

$ hive -f hive\_trip\_ddl.sql

# Run trip\_transforming.sql to transform datasets

It takes about 1 hour in Hive.

$ hive -f trip\_transforming.sql

# Run trip\_aggregation.sql to aggregate datasets for visualization and forecast

It takes about 5 hours to run.

$ hive -f trip\_aggregation.sql

# Set up python and run trip\_forecast.py

Make sure Python 2.7 or higher version is used. If not, please use python\_setup.sh to install python3 and appropriate version of pip.

Install Python3 as a root user. Go to the directory where below script is saved.

$ . python\_setup.sh

Next, use Python3 for PySpark and execute the python script

$ cd /data

$ su - w205

$ export PYSPARK\_PYTHON=python3

$ /data/spark15/bin/spark-submit /data/taxi\_trip/trip\_forecast.py

# Connect tableau to the Hive Server in EC2 instance

Start Hive server on EC2 instance as w205 user

$ hive --service hiveserver2 --hiveconf hive.server2.thrift.port=10000 &

Open Tableau and use your instance IP for connection.