Folders:

* analytics\_code:
  + analysis\_with\_merged\_dataset:
    - Part1: These are queries that can be run on Hive/Impala to analyze how timing impacts the traffic violations in NYC
    - Part2: These are queries that can be run on Hive/Impala to analyze how weather conditions impact speeding violations in NYC
    - Part3: These are queries that can be run on Hive/Impala to analyze how the location, weather, and timing impacts the traffic violations in NYC.
* screenshots:

Here we took a screen shot for every analysis query/job.

* Code\_iterations:

Here we included all various job/analysis we have conducted throughout the project but haven't used them in the final project submission.

* profiling\_code:
  + traffic\_speed: The HMR job can be utilized to find the ranges and the limits of each field in the raw dataset.

To run the job:

hadoop jar profileData.jar ProfileData /user/ry856/sharedData/InputDatasets/trafficSpeed /user/ry856/sharedData/profiling/trafficSpeed/outputProfile1

* + weather: The HMR job can be utilized to find the ranges and the limits of each field in the raw dataset.

To run this job: take column VSB for example

Hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -D mapreduce.job.reduces=1 -files hdfs://dumbo/user/yy1420/project/VSB\_mapper.py hdfs://dumbo/user/yy1420/project/VSB\_reducer.py - mapper “python VSB\_mapper.py” -reducer “python VSB\_reducer.py

-input /user/yy1420/project/WeatherData\_CentralPark.txt -output /user/yy1420/project/profile1

All codes and output resides in the directory /user/yy1420/project

* + Speed Limit: The HMR job can be utilized to find the ranges and limits of each field in the raw speed limit dataset.

To run this job: hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar \-D mapreduce.job.reduces=1 \-files hdfs://dumbo/user/bm2515/class1/python\_code/Speed\_Limit\_profiling\_Mapper.py,hdfs://dumbo/user/bm2515/class1/python\_code/speed\_Limit\_profiling\_Reducer.py \-mapper "python Speed\_Limit\_profiling\_Mapper.py" \-reducer "python Speed\_Limit\_profiling\_Reducer.py" \-input /user/bm2515/class1/speedlimit.csv \-output /user/bm2515/class1/output

All code and output resides in the directory /user/bm2515/HadoopProject

* etl\_code:
  + To\_merge: Under this folder we have two folders.
    - One folder is to represent Hive queries to merge between the weather dataset and the traffic speed dataset on the date fields. We joined both table on the rows where the dates fields (all the way to the hours- e.g. yyyyMMddHH) matched.

To execute the code, please run the queries hive.hql file in Hive.

* + - The next folder is to merge the previously merged dataset with the speed limit dataset. Here we had to write a HMR job to do it and load the speed\_limit dataset for every reduce job.

to complete the job, run this command:

hadoop jar mergeData.jar MergeData /user/ry856/sharedData/mergedData/trafficSpeed\_and\_weather /user/ry856/sharedData/mergedData/trafficSpeed\_and\_weather

* + Traffic\_speed: The ETL was used to drop some columns and change the format of the remaining as the first data cleaning HMR job.

To run the job:

hadoop jar cleanData.jar CleanData /user/ry856/sharedData/InputDatasets/trafficSpeed /user/ry856/sharedData/dataToMerge/TrafficSpeed

* + weather: The ETL was used to extract the columns to our interest and also drop the rows with missing values as the first data cleaning HMR job.

To run this job:

Hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -D mapreduce.job.reduces=0 -files hdfs://dumbo/user/yy1420/project/DataCleaning\_Mapper.py - mapper “python DataCleaning\_Mapper.py” -input /user/yy1420/project/WeatherData\_CentralPark.txt -output /user/yy1420/project/test1

* + Speed Limit: The ETL was used to drop some columns and change the format of the remaining columns into meaningful and consistent entries.

To run this job: hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar \-D mapreduce.job.reduces=0 \-files hdfs://dumbo/user/bm2515/class1/python\_code/SpeedLimit\_Clean\_Mapper

.py \-mapper "python SpeedLimit\_Clean\_Mapper.py" \-input /user/bm2515/class1/speedlimit.csv \-output /user/bm2515/class1/output

* Data\_ingest:

under this folder, you should find three folders pertaining to each dataset (traffic\_speed, weather and speed\_limit).Under each folder, you should find all the steps taken to ingest the data onto HDFS on the Dumbo cluster.