Directory:
data_ingest: contains commands for data ingestion for moving three datasets: yellow taxi, citibike, weather, from local to dumbo, dumbo to hdfs
file name: Command_data_ingest

etl_code: contains pig script which clean the data, drops the unnecessary columns, retain useful columns, eliminate all the records with one or more missing values from the three datasets: yellow taxi, citibike, weather.
file name: CitiBike.pig, Weather.pig, yellow_taxi.pig

profilling_code: contains the MapReduce code files for from the three datasets: yellow taxi, citibike, weather to do further operations on data. In Mapper we format the key, match the filtered desire value. In the Reducer, we aggregate the data to make it doable for analytics.
file name: under process_bike_data directory: CitiBike.java, CitiBikeMapper.java, CitiBikeReducer.java
under process_taxi_data directory: YellowTaxi.java, YellowTaxiMapper.java, YellowTaxiReducer.java
under weather directory: Weather.java, WeatherMapper.java, WeatherReducer.java
WeatherTest.txt

code_iterations: contains Analytics Script 1209.docx which has impala code we used for analytics the MR result from profiling code
file name: Analytics Script 1209.docx

screentshots: contains Screen shots that show our analytic impala code running

How to Build and Run Code **etl_code:**
pig yellow_taxi.pig
pig CitiBike.pig
pig Weather.pig

How to Build and Run Code **profilling_code:**
How to build and run citibike mapreduce program
1. login to the dumbo account of jl8456
2. cd /home/jl8456/PROJECT
3. Run the following commands

javac -classpath `yarn classpath` -d . CitibikeMapper.java
javac -classpath `yarn classpath` -d . CitibikeReducer.java
javac -classpath `yarn classpath`:. -d . CitiBike.java
jar -cvf CitiBike.jar *.class
hadoop jar CitiBike.jar CitiBike
/user/jl8456/PROJECT/city_bike_data/cleaned_bike/giant_city_bike_cleaned.txt
/user/jl8456/MRCB2

How to build and run yellow taxi mapreduce program

1. login to the dumbo account of yy1316
2. cd /home/yy1316/Project
3. Run the following commands

javac -classpath `yarn classpath` -d . YellowTaxiMapper.java
javac -classpath `yarn classpath` -d . YellowTaxiReducer.java
javac -classpath `yarn classpath`:. -d . YellowTaxi.java
jar -cvf YellowTaxi.jar *.class
hadoop jar YellowTaxi.jar YellowTaxi /user/yy1316/Project/cleaned_taxi
/user/yy1316/Project/stat_taxi

How to build and run weather mapreduce program

How to get access to the Map Reduce Functions for Weather data Profiling
1. login to the dumbo account of xl1638
2. cd /home/xl1638/weather
3. Run the command written below:

//MapReduce running commands:
//Building the Path:
javac -classpath `yarn classpath` -d . WeatherMapper.java
javac -classpath `yarn classpath` -d . WeatherReducer.java
javac -classpath `yarn classpath`:. -d . Weather.java
jar -cvf Weather.jar *.class

//Run the program/jar files:
hadoop jar Weather.jar Weather /user/xl1638/PROJECT/weather/WeatherTest.txt
/user/yy1316/Project/weather_result/

How to Build and Run Code **code_iterations:**
How to run Impala Script:
impala-shell
connect compute-1-1;
use yy1316;
run the commands from the Analytics Script 1209.docx

where to find results of a run:

Pigged result path for FHV data:
hdfs://dumbo/user/yy1316/Project/cleaned_taxi/taxi_cleaned_1
hdfs://dumbo/user/yy1316/Project/cleaned_taxi/taxi_cleaned_2
hdfs://dumbo/user/yy1316/Project/cleaned_taxi/taxi_cleaned_3
hdfs://dumbo/user/yy1316/Project/cleaned_taxi/taxi_cleaned_4
hdfs://dumbo/user/yy1316/Project/cleaned_taxi/taxi_cleaned_5
hdfs://dumbo/user/yy1316/Project/cleaned_taxi/taxi_cleaned_6

MRed result for FHV data:
hdfs://dumbo/user/jl8456/PROJECT/yellow_taxi/latest_taxi_cleaned_giant.txt

Pigged result path for citibike:
hdfs://dumbo/user/jl8456/PROJECT/city_bike_data/cleaned_bike/giant_city_bike_cleaned.txt

Pigged and MRed result path for cleaned citybike file:
hdfs://dumbo/user/jl8456/MRCB2/part-r-00000

Pigged result path for weather data:
/user/xl1638/PROJECT/weather/WeatherTest.txt

MRed result path for weather data:
/user/yy1316/Project/weather_result/weather_output.txt

Impala result for the four EXTERNAL_TABLE:
hdfs://dumbo/user/hive/warehouse/yy1316.db/bike_taxi_stat_per_day
hdfs://dumbo/user/hive/warehouse/yy1316.db/bike_taxi_stat_in_all
hdfs://dumbo/user/hive/warehouse/yy1316.db/bike_taxi_weather_combine_day
hdfs://dumbo/user/hive/warehouse/yy1316.db/bike_taxi_weather_combine_all
query results are on Analytics Script 1209.docx

Input Data:
Raw input data - FHV:
hdfs://dumbo/user/jl8456/PROJECT/yellow_taxi/yellow_tripdata_2018-01.csv
hdfs://dumbo/user/jl8456/PROJECT/yellow_taxi/yellow_tripdata_2018-02.csv
hdfs://dumbo/user/jl8456/PROJECT/yellow_taxi/yellow_tripdata_2018-03.csv
hdfs://dumbo/user/jl8456/PROJECT/yellow_taxi/yellow_tripdata_2018-04.csv
hdfs://dumbo/user/jl8456/PROJECT/yellow_taxi/yellow_tripdata_2018-05.csv
hdfs://dumbo/user/jl8456/PROJECT/yellow_taxi/yellow_tripdata_2018-06.csv

Raw input data - citibike:
hdfs://dumbo/user/jl8456/PROJECT/city_bike_data/201801-citibike-tripdata.csv
hdfs://dumbo/user/jl8456/PROJECT/city_bike_data/201802-citibike-tripdata.csv
hdfs://dumbo/user/jl8456/PROJECT/city_bike_data/201803-citibike-tripdata.csv
hdfs://dumbo/user/jl8456/PROJECT/city_bike_data/201804-citibike-tripdata.csv
hdfs://dumbo/user/jl8456/PROJECT/city_bike_data/201805-citibike-tripdata.csv
hdfs://dumbo/user/jl8456/PROJECT/city_bike_data/201806-citibike-tripdata.csv

Raw input data – weather:
hdfs://dumbo/user/jl8456/PROJECT/weather/central_park_weather.csv

Processed input for FHV MR:
hdfs://dumbo/user/yy1316/Project/cleaned_taxi

Processed input for Citibike MR:

hdfs://dumbo//user/jl8456/PROJECT/city_bike_data/cleaned_bike/giant_city_bike_cleaned.txt

Processed input for Weather MR:
hdfs://dumbo/user/jl8456/PROJECT/weather/cleaned_weather

Impala Input:
/user/yy1316/Project/weather_data
/user/yy1316/Project/taxi_data
/user/yy1316/Project/bike_data