Readme

by Elaine Shan, Jason Yu

Directories and files:

data_ingest

- This directory is empty since no commands like curl were used to get data. All we did was downloading the datasets from the websites (links listed at the bottom) then uploading those to HDFS.

etl_code

- This directory contains the MapReduce code to clean our datasets.

- **cleanBicycle** contains the code used to clean the dataset: ***Bicycle_Counts.csv***
  A successful run of this code yields the file: ***bi.csv***

- **sumBicycle** contains the code used to further cleaning/profiling the file: ***bi.csv,*** by summing bicycles of every 15 mins from a single day to get total number of bicycles that day.
  A successful run of this code yields the file: ***bi_sum.csv***

- **cleanCovidCase** contains the code used to clean the dataset:
  COVID-19_Daily_Counts_of_Cases__Hospitalizations__and_Deaths.csv
  A successful run of this code yields the file: ***covid.csv***

profiling_code

- This directory contains the MapReduce code to profile our datasets.
- **CountRecs** contains the code used to count the total number of records of an input file.
  A successful run of this code yields the total number of records in the terminal.

ana_code:

- This directory contains the MapReduce code for analytics.
- **bi_change** contains the code used to calculate % Change of bicycle numbers from pre-pandemic equivalent day.
  A successful run of this code yields the file: ***bi_change.csv***.
- A file named ***hive_analytics.txt*** containing all the hive queries we made for analytics.

Screenshots:

- This directory contains Screenshots that show our analytics running.

Building and running the codes:

cleanCovidCase

```
cd ~/Project_Final_Code/etl_code/cleanCovidCase
rm *.class *.jar
javac -classpath `yarn classpath` -d . cleanCovidCaseMapper.java
javac -classpath `yarn classpath` -d . cleanCovidCaseReducer.java
javac -classpath `yarn classpath`:. -d . cleanCovidCase.java
jar -cvf cleanCovidCase.jar *.class
hadoop jar cleanCovidCase.jar cleanCovidCase project_data/COVID-19_Daily_Counts_of_Cases__Hospitalizations__and_Deaths.csv
/user/jy2575/hw/output
hdfs dfs -cp hw/output/part-r-00000 project_data/covid.csv
hdfs dfs -cat project_data/covid.csv
hdfs dfs -rm -r -f hw/output
```

cleanBicycle

```
cd ~/Project_Final_Code/etl_code/cleanBicycle
rm *.class *.jar
javac -classpath `yarn classpath` -d . cleanBicycleMapper.java
javac -classpath `yarn classpath` -d . cleanBicycleReducer.java
javac -classpath `yarn classpath`:. -d . cleanBicycle.java
jar -cvf cleanBicycle.jar *.class
hadoop jar cleanBicycle.jar cleanBicycle project_data/Bicycle_Counts.csv /user/jy2575/hw/output
hdfs dfs -cp hw/output/part-r-00000 project_data/bi.csv
hdfs dfs -cat project_data/bi.csv
hdfs dfs -rm -r -f hw/output
```

CountRecs

```
cd ~/Project_Final_Code/profiling_code/CountRecs
rm *.class *.jar
javac -classpath `yarn classpath` -d . CountRecsMapper.java
javac -classpath `yarn classpath` -d . CountRecsReducer.java
javac -classpath `yarn classpath`:. -d . CountRecs.java
jar -cvf CountRecs.jar *.class
hadoop jar CountRecs.jar CountRecs project_data/bi.csv /user/jy2575/hw/output
hdfs dfs -cat hw/output/part-r-00000
hdfs dfs -rm -r -f hw/output
```

sumBicycle

```
cd ~/Project_Final_Code/etl_code/sumBicycle
rm *.class *.jar
rm -r -f output
javac -classpath `yarn classpath` -d . sumBicycleMapper.java
javac -classpath `yarn classpath` -d . sumBicycleReducer.java
javac -classpath `yarn classpath`:. -d . sumBicycle.java
jar -cvf sumBicycle.jar *.class
hadoop jar sumBicycle.jar sumBicycle project_data/bi.csv /user/jy2575/hw/output
hdfs dfs -cp hw/output/part-r-00000 project_data/bi_sum.csv
hdfs dfs -cat project_data/bi_sum.csv
hdfs dfs -rm -r -f hw/output
```

bi_change

```
cd ~/Project_Final_Code/ana_code/bi_change
rm *.class *.jar
rm -r -f output
javac -classpath `yarn classpath` -d . bi_changeMapper.java
javac -classpath `yarn classpath` -d . bi_changeReducer.java
javac -classpath `yarn classpath`:. -d . bi_change.java
jar -cvf bi_change.jar *.class
hadoop jar bi_change.jar bi_change project_data/bi_sum.csv /user/jy2575/hw/output
hdfs dfs -cp hw/output/part-r-00000 project_data/bi_change.csv
```

hdfs dfs -cat project_data/bi_change.csv
hdfs dfs -rm -r -f hw/output

Hive queries (the same as ***hive_analytics.txt***)

```
beeline --silent
!connect jdbc:hive2://hm-1.hpc.nyu.edu:10000/
Use jy2575;

create external table bi_sum (`date` date, `count` int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
load data inpath 'hdfs://horton.hpc.nyu.edu:8020/user/jy2575/project_data/bi_sum.csv' overwrite into table bi_sum;

select * from bi_sum;

Select count(*) from bi_sum where `date` <= '2021-08-01' and `date` >= '2020-03-01';

select sum(count) from bi_sum where `date` <= '2021-08-01' and `date` >= '2020-03-01';
select sum(count) from bi_sum where `date` <= '2020-08-01' and `date` >= '2019-03-01';

select avg(count) from bi_sum where `date` <= '2021-08-01' and `date` >= '2020-03-01';
select avg(count) from bi_sum where `date` <= '2020-08-01' and `date` >= '2019-03-01';

select avg(count) from bi_sum where `date` <= '2021-07-01' and `date` >= '2021-04-01';

select max(count) from bi_sum where `date` <= '2021-08-01' and `date` >= '2020-03-01';
Select * from bi_sum where `count`=66489;

select min(count) from bi_sum where `date` <= '2021-08-01' and `date` >= '2020-03-01';
Select * from bi_sum where `count`=913;


create external table covid (`date` date, `count` int) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' STORED AS TEXTFILE;
load data inpath 'hdfs://horton.hpc.nyu.edu:8020/user/jy2575/project_data/covid.csv' overwrite into table covid;

select * from covid;

select sum(count) from covid where `date` <= '2021-08-01' and `date` >= '2020-03-01';

select avg(count) from covid where `date` <= '2021-08-01' and `date` >= '2020-03-01';

select avg(count) from covid where `date` <= '2021-07-01' and `date` >= '2021-04-01';

select max(count) from covid where `date` <= '2021-08-01' and `date` >= '2020-03-01';

select min(count) from covid where `date` <= '2021-08-01' and `date` >= '2020-03-01';


Select avg(bi_sum.count) from bi_sum join covid on (bi_sum.`date`=covid.`date`) where bi_sum.`date` <= '2021-08-01' and bi_sum.`date` >= '2020-03-01'
and covid.count>=1500;
Select avg(bi_sum.count) from bi_sum join covid on (bi_sum.`date`=covid.`date`) where bi_sum.`date` <= '2021-08-01' and bi_sum.`date` >= '2020-03-01'
and covid.count>=3000;
Select avg(bi_sum.count) from bi_sum join covid on (bi_sum.`date`=covid.`date`) where bi_sum.`date` <= '2021-08-01' and bi_sum.`date` >= '2020-03-01'
and covid.count>=4500;


Select corr(bi_sum.count,covid.count) from bi_sum join covid on (bi_sum.`date`=covid.`date`) where bi_sum.`date` <= '2021-08-01' and bi_sum.`date` >=
'2020-03-01';
```

Input and output files:

All the input files could be found in the HDFS directory of
/user/jy2575/project_data
It should be something like this:

```
[jy2575@hlog-1 bi_change]$ hdfs dfs -ls project_data
Found 3 items
-rw-rwx---+  3 jy2575 jy2575  174382053 2021-10-30 13:17 project_data/Bicycle_Counts.csv
-rw-rwx---+  3 jy2575 jy2575     115204 2021-10-30 13:17 project_data/COVID-19_Daily_Counts_of_Cases__Hospitalizations__and_Deaths.csv
-rw-rwx---+  3 jy2575 jy2575      53248 2021-10-30 13:17 project_data/MTA_recent_ridership_data_20211026.csv
```

And after successfully building and running all the codes with the commands in the previous page, the output files should also be present in the HDFS directory of
/user/jy2575/project_data
It should be something like this:

```
[jy2575@hlog-1 bi_change]$ hdfs dfs -ls project_data
Found 6 items
-rw-rwx---+  3 jy2575 jy2575  174382053 2021-10-30 13:17 project_data/Bicycle_Counts.csv
-rw-rwx---+  3 jy2575 jy2575     115204 2021-10-30 13:17 project_data/COVID-19_Daily_Counts_of_Cases__Hospitalizations__and_Deaths.csv
-rw-rwx---+  3 jy2575 jy2575      53248 2021-10-30 13:17 project_data/MTA_recent_ridership_data_20211026.csv
-rw-rw----+  3 jy2575 jy2575   92476942 2021-11-21 15:35 project_data/bi.csv
-rw-rw----+  3 jy2575 jy2575      10072 2021-11-21 17:31 project_data/bi_change.csv
-rw-rw----+  3 jy2575 jy2575      54292 2021-11-21 17:31 project_data/bi_sum.csv
```

Dataset sources:

https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3
https://data.cityofnewyork.us/Transportation/Bicycle-Counts/uczf-rk3c
https://new.mta.info/coronavirus/ridership