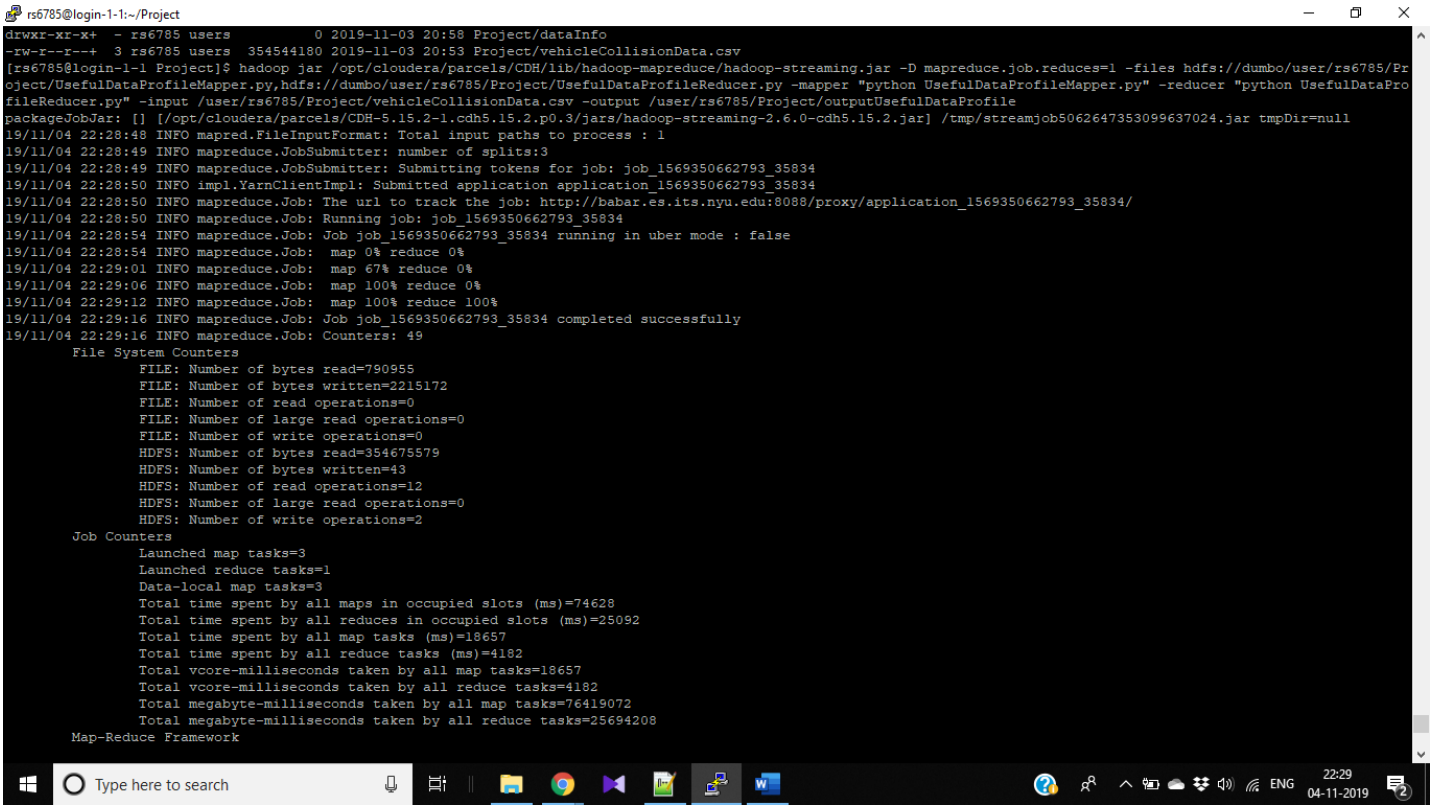
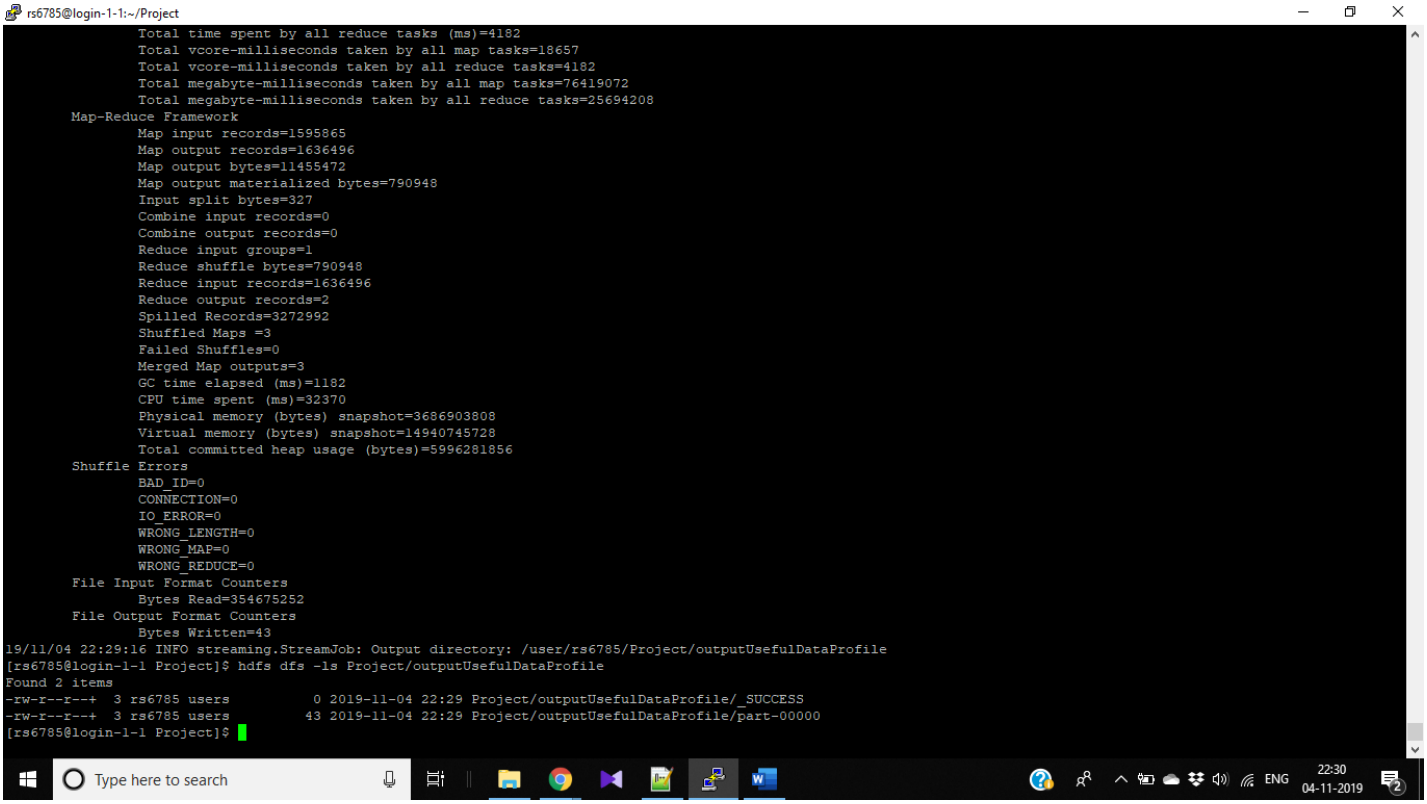
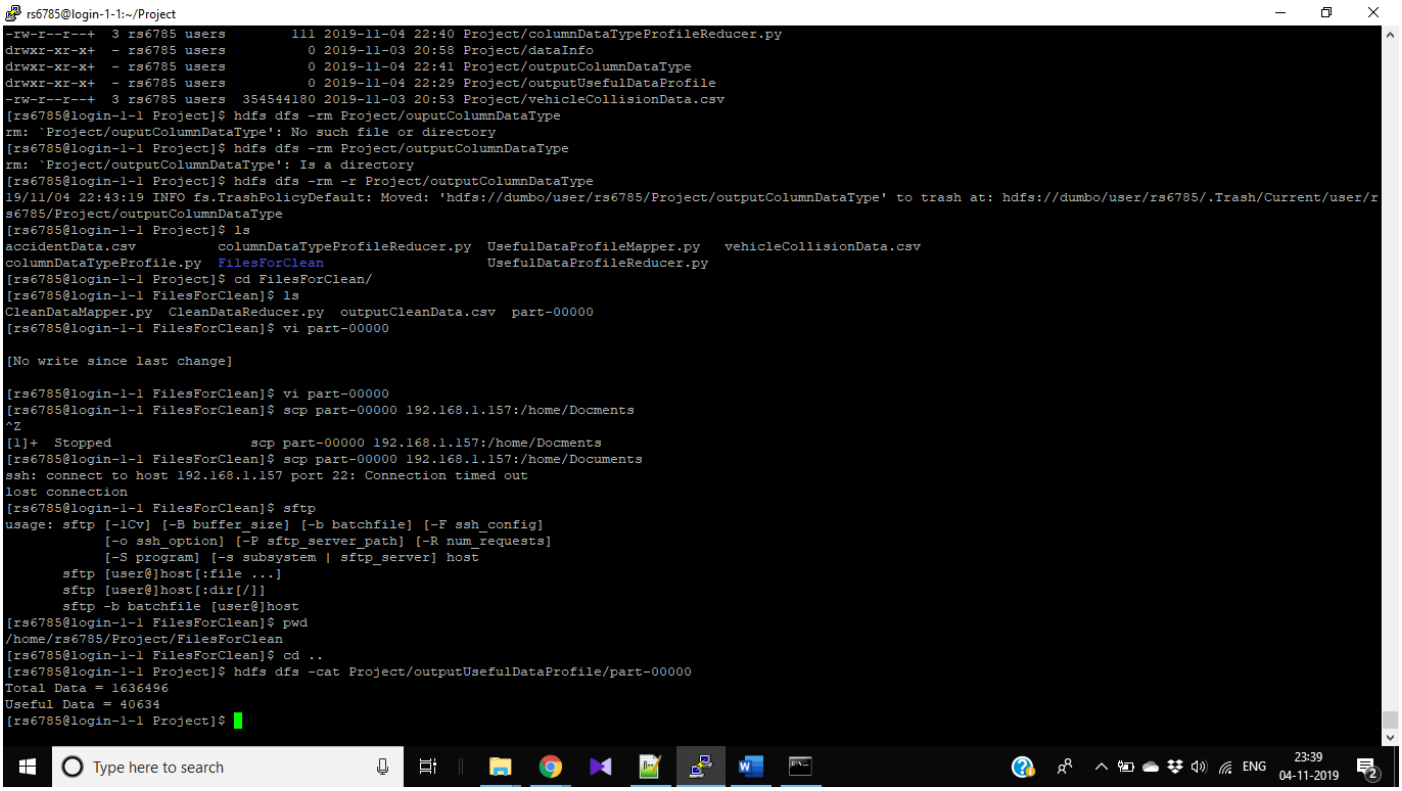
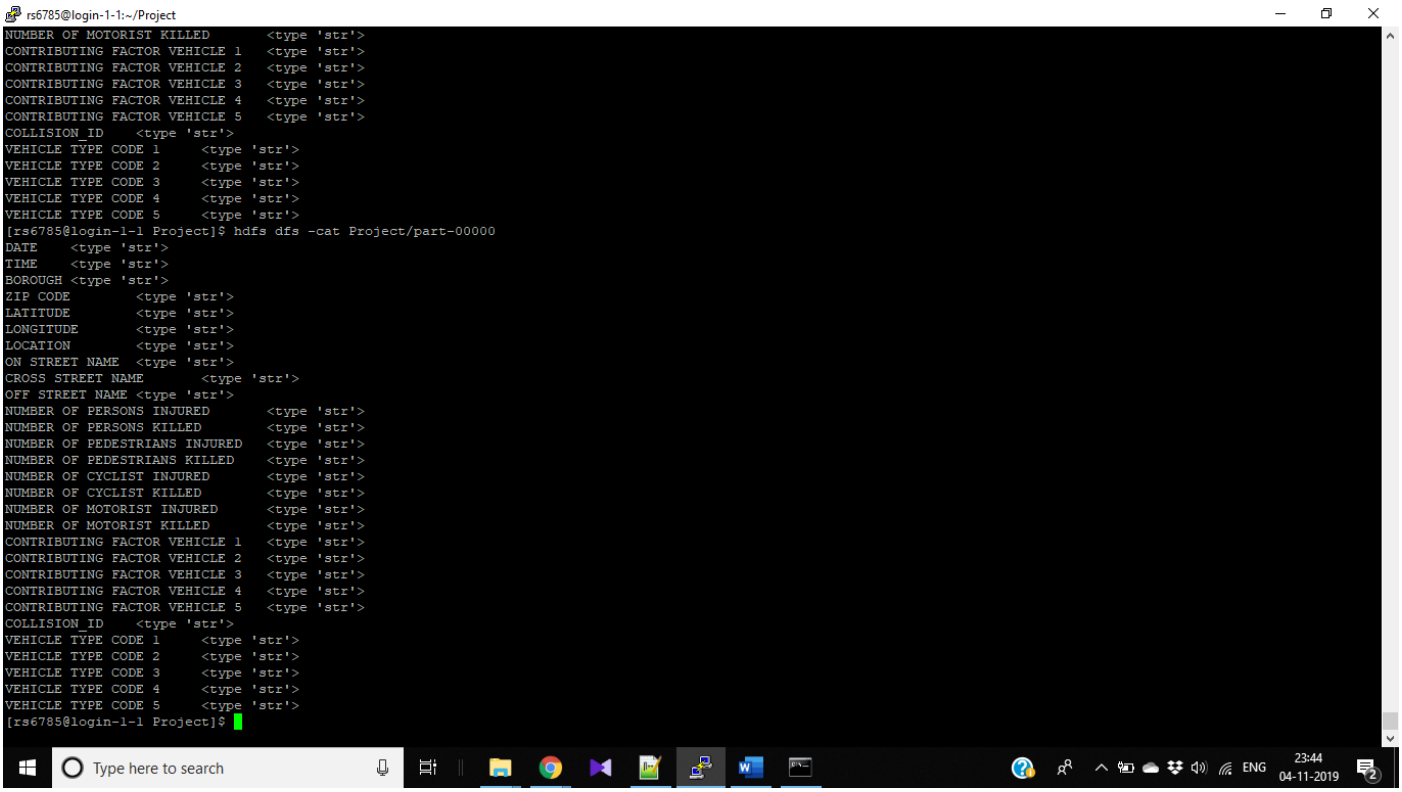
Profile Data

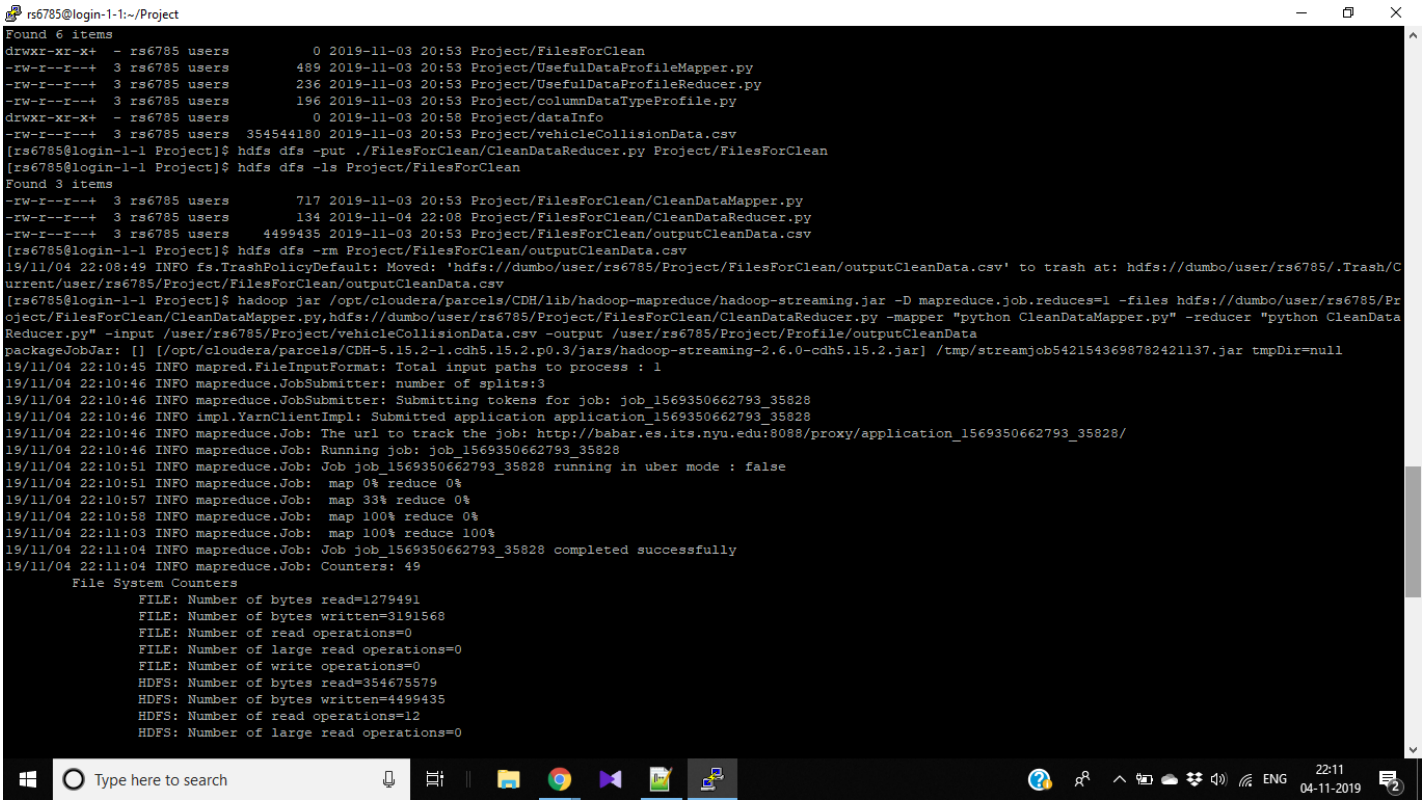


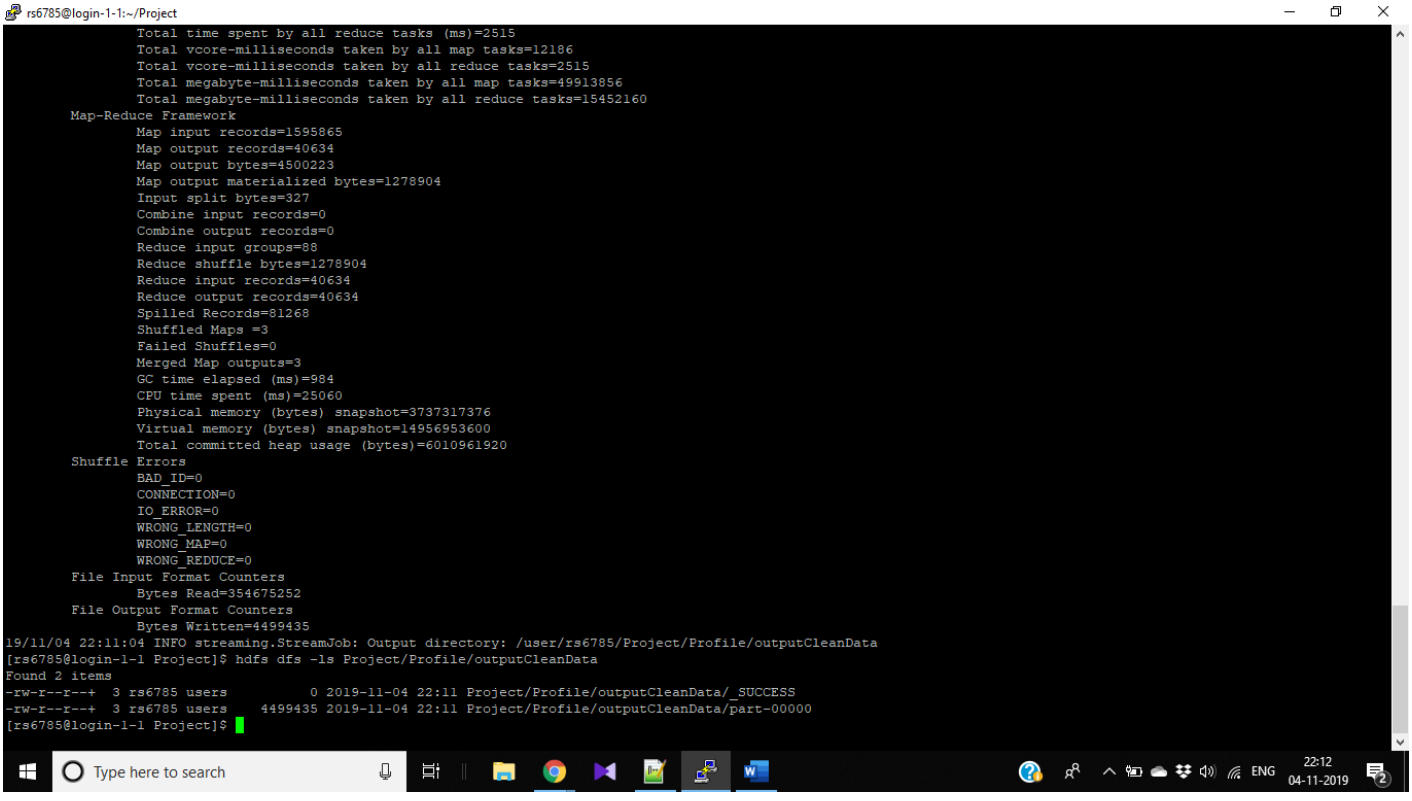


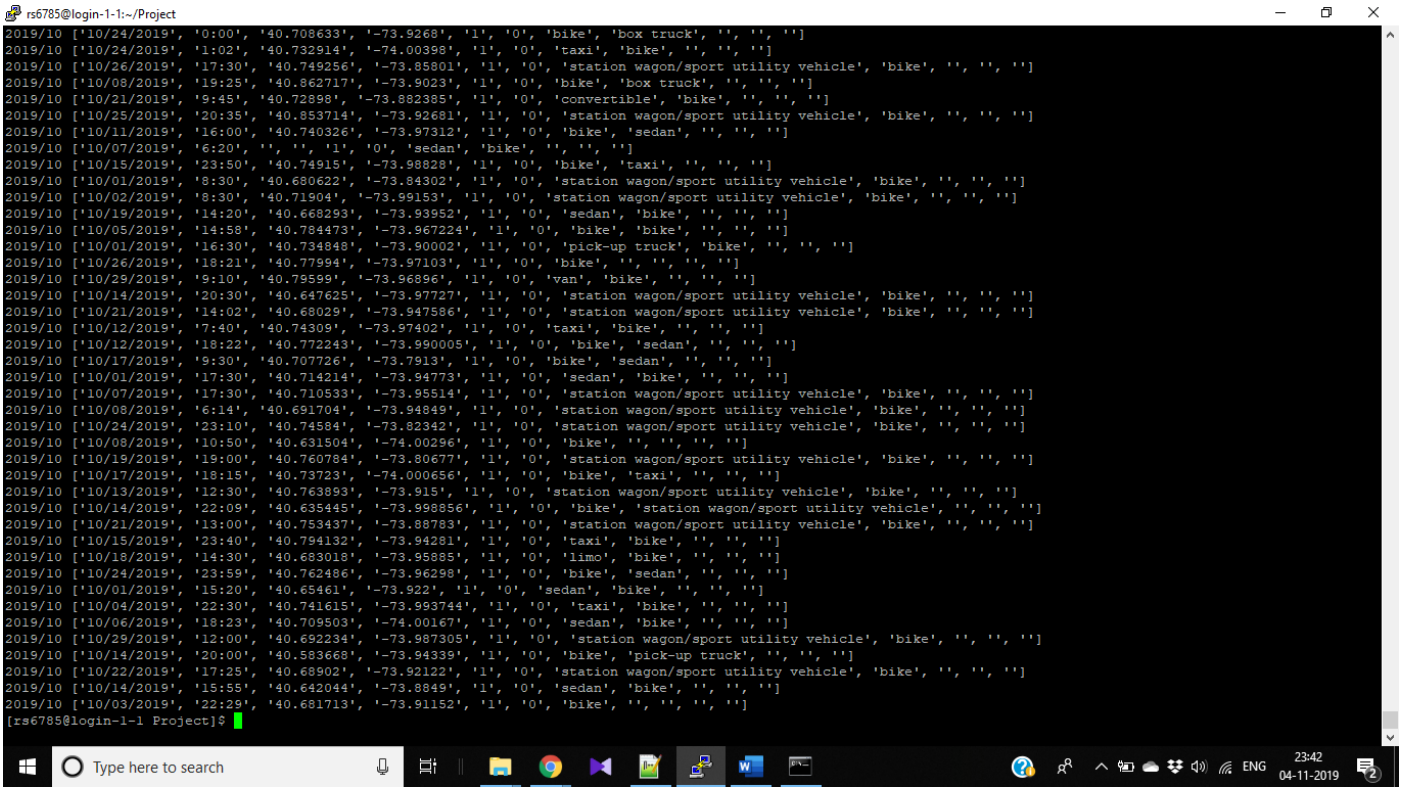

rs6785@login-1-1:~/Project
-rw-r--r--+ 3 rs6785 users 111 2019-11-04 22:40 Project/columnDataTypeProfileReducer.py
drwxr-xr-x+ - rs6785 users 0 2019-11-03 20:58 Project/dataInfo
drwxr-xr-x+ - rs6785 users 0 2019-11-04 22:41 Project/outputColumnDataType
drwxr-xr-x+ - rs6785 users 0 2019-11-04 22:29 Project/outputUsefulDataProfile
-rw-r--r--+ 3 rs6785 users 354544180 2019-11-03 20:53 Project/vehicleCollisionData.csv
[rs6785@login-1-1 Project]$ hdfs dfs -rm Project/ouputColumnDataType
rm: `Project/ouputColumnDataType': No such file or directory
[rs6785@login-1-1 Project]$ hdfs dfs -rm Project/outputColumnDataType
rm: `Project/outputColumnDataType': Is a directory
[rs6785@login-1-1 Project]$ hdfs dfs -rm -r Project/outputColumnDataType
19/11/04 22:43:19 INFO fs.TrashPolicyDefault: Moved: 'hdfs://dumbo/user/rs6785/Project/outputColumnDataType' to trash at: hdfs://dumbo/user/rs6785/.Trash/Current/user/rs6785/Project/outputColumnDataType
[rs6785@login-1-1 Project]$ ls
accidentData.csv columnDataTypeProfileReducer.py UsefulDataProfileMapper.py vehicleCollisionData.csv
columnDataTypeProfile.py FilesForClean UsefulDataProfileReducer.py
[rs6785@login-1-1 Project]$ cd FilesForClean/
[rs6785@login-1-1 FilesForClean]$ ls
CleanDataMapper.py CleanDataReducer.py outputCleanData.csv part-00000
[rs6785@login-1-1 FilesForClean]$ vi part-00000
[No write since last change]
[rs6785@login-1-1 FilesForClean]$ vi part-00000
[rs6785@login-1-1 FilesForClean]$ scp part-00000 192.168.1.157:/home/Docments
^Z
[1]+ Stopped scp part-00000 192.168.1.157:/home/Docments
[rs6785@login-1-1 FilesForClean]$ scp part-00000 192.168.1.157:/home/Documents
ssh: connect to host 192.168.1.157 port 22: Connection timed out
lost connection
[rs6785@login-1-1 FilesForClean]$ sftp
usage: sftp [-1Cv] [-B buffer_size] [-b batchfile] [-F ssh_config]
[-o ssh_option] [-P sftp_server_path] [-R num_requests]
[-S program] [-s subsystem | sftp_server] host
sftp [user@]host[:file ...]
sftp [user@]host[:dir[/]]
sftp -b batchfile [user@]host
[rs6785@login-1-1 FilesForClean]$ pwd
/home/rs6785/Project/FilesForClean
[rs6785@login-1-1 FilesForClean]$ cd ..
[rs6785@login-1-1 Project]$ hdfs dfs -cat Project/outputUsefulDataProfile/part-00000
Total Data = 1636496
Useful Data = 40634
[rs6785@login-1-1 Project]$
Type here to search
ENG
23:39
04-11-2019


rs6785@login-1-1:~/Project
NUMBER OF MOTORIST KILLED <type 'str'>
CONTRIBUTING FACTOR VEHICLE 1 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 2 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 3 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 4 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 5 <type 'str'>
COLLISION_ID <type 'str'>
VEHICLE TYPE CODE 1 <type 'str'>
VEHICLE TYPE CODE 2 <type 'str'>
VEHICLE TYPE CODE 3 <type 'str'>
VEHICLE TYPE CODE 4 <type 'str'>
VEHICLE TYPE CODE 5 <type 'str'>
[rs6785@login-1-1 Project]$ hdfs dfs -cat Project/part-00000
DATE <type 'str'>
TIME <type 'str'>
BOROUGH <type 'str'>
ZIP CODE <type 'str'>
LATITUDE <type 'str'>
LONGITUDE <type 'str'>
LOCATION <type 'str'>
ON STREET NAME <type 'str'>
CROSS STREET NAME <type 'str'>
OFF STREET NAME <type 'str'>
NUMBER OF PERSONS INJURED <type 'str'>
NUMBER OF PERSONS KILLED <type 'str'>
NUMBER OF PEDESTRIANS INJURED <type 'str'>
NUMBER OF PEDESTRIANS KILLED <type 'str'>
NUMBER OF CYCLIST INJURED <type 'str'>
NUMBER OF CYCLIST KILLED <type 'str'>
NUMBER OF MOTORIST INJURED <type 'str'>
NUMBER OF MOTORIST KILLED <type 'str'>
CONTRIBUTING FACTOR VEHICLE 1 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 2 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 3 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 4 <type 'str'>
CONTRIBUTING FACTOR VEHICLE 5 <type 'str'>
COLLISION_ID <type 'str'>
VEHICLE TYPE CODE 1 <type 'str'>
VEHICLE TYPE CODE 2 <type 'str'>
VEHICLE TYPE CODE 3 <type 'str'>
VEHICLE TYPE CODE 4 <type 'str'>
VEHICLE TYPE CODE 5 <type 'str'>
[rs6785@login-1-1 Project]$
Type here to search
ENG
23:44
04-11-2019

Clean Data


rs6785@login-1-1:~/Project
Found 6 items
drwxr-xr-x+ - rs6785 users   0 2019-11-03 20:53 Project/FilesForClean
-rw-r--r--+ 3 rs6785 users   489 2019-11-03 20:53 Project/UsefulDataProfileMapper.py
-rw-r--r--+ 3 rs6785 users   236 2019-11-03 20:53 Project/UsefulDataProfileReducer.py
-rw-r--r--+ 3 rs6785 users   196 2019-11-03 20:53 Project/columnDataTypeProfile.py
drwxr-xr-x+ - rs6785 users   0 2019-11-03 20:58 Project/dataInfo
-rw-r--r--+ 3 rs6785 users 354544180 2019-11-03 20:53 Project/vehicleCollisionData.csv
[rs6785@login-1-1 Project]$ hdfs dfs -put ./FilesForClean/CleanDataReducer.py Project/FilesForClean
[rs6785@login-1-1 Project]$ hdfs dfs -ls Project/FilesForClean
Found 3 items
-rw-r--r--+ 3 rs6785 users   717 2019-11-03 20:53 Project/FilesForClean/CleanDataMapper.py
-rw-r--r--+ 3 rs6785 users   134 2019-11-04 22:08 Project/FilesForClean/CleanDataReducer.py
-rw-r--r--+ 3 rs6785 users   4499435 2019-11-03 20:53 Project/FilesForClean/outputCleanData.csv
[rs6785@login-1-1 Project]$ hdfs dfs -rm Project/FilesForClean/outputCleanData.csv
19/11/04 22:08:49 INFO fs.TrashPolicyDefault: Moved: 'hdfs://dumbo/user/rs6785/Project/FilesForClean/outputCleanData.csv' to trash at: hdfs://dumbo/user/rs6785/.Trash/Current/user/rs6785/Project/FilesForClean/outputCleanData.csv
[rs6785@login-1-1 Project]$ hadoop jar /opt/cloudera/parcels/CDH/lib/hadoop-mapreduce/hadoop-streaming.jar -D mapreduce.job.reduces=1 -files hdfs://dumbo/user/rs6785/Project/FilesForClean/CleanDataMapper.py,hdfs://dumbo/user/rs6785/Project/FilesForClean/CleanDataReducer.py -mapper "python CleanDataMapper.py" -reducer "python CleanDataReducer.py" -input /user/rs6785/Project/vehicleCollisionData.csv -output /user/rs6785/Project/Profile/outputCleanData
packageJobJar: [] [/opt/cloudera/parcels/CDH-5.15.2-1.cdh5.15.2.p0.3/jars/hadoop-streaming-2.6.0-cdh5.15.2.jar] /tmp/streamjob5421543698782421137.jar tmpDir=null
19/11/04 22:10:45 INFO mapred.FileInputFormat: Total input paths to process : 1
19/11/04 22:10:46 INFO mapreduce.JobSubmitter: number of splits:3
19/11/04 22:10:46 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1569350662793_35828
19/11/04 22:10:46 INFO impl.YarnClientImpl: Submitted application application_1569350662793_35828
19/11/04 22:10:46 INFO mapreduce.Job: The url to track the job: http://babar.es.its.nyu.edu:8088/proxy/application_1569350662793_35828/
19/11/04 22:10:46 INFO mapreduce.Job: Running job: job_1569350662793_35828
19/11/04 22:10:51 INFO mapreduce.Job: Job job_1569350662793_35828 running in uber mode : false
19/11/04 22:10:51 INFO mapreduce.Job: map 0% reduce 0%
19/11/04 22:10:57 INFO mapreduce.Job: map 33% reduce 0%
19/11/04 22:10:58 INFO mapreduce.Job: map 100% reduce 0%
19/11/04 22:11:03 INFO mapreduce.Job: map 100% reduce 100%
19/11/04 22:11:04 INFO mapreduce.Job: Job job_1569350662793_35828 completed successfully
19/11/04 22:11:04 INFO mapreduce.Job: Counters: 49
File System Counters
FILE: Number of bytes read=1279491
FILE: Number of bytes written=3191568
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=354675579
HDFS: Number of bytes written=4499435
HDFS: Number of read operations=12
HDFS: Number of large read operations=0
Type here to search
ENG
22:11
04-11-2019


rs6785@login-1-1:~/Project
Total time spent by all reduce tasks (ms)=2515
Total vcore-milliseconds taken by all map tasks=12186
Total vcore-milliseconds taken by all reduce tasks=2515
Total megabyte-milliseconds taken by all map tasks=49913856
Total megabyte-milliseconds taken by all reduce tasks=15452160
Map-Reduce Framework
Map input records=1595865
Map output records=40634
Map output bytes=4500223
Map output materialized bytes=1278904
Input split bytes=327
Combine input records=0
Combine output records=0
Reduce input groups=88
Reduce shuffle bytes=1278904
Reduce input records=40634
Reduce output records=40634
Spilled Records=81268
Shuffled Maps =3
Failed Shuffles=0
Merged Map outputs=3
GC time elapsed (ms)=984
CPU time spent (ms)=25060
Physical memory (bytes) snapshot=3737317376
Virtual memory (bytes) snapshot=14956953600
Total committed heap usage (bytes)=6010961920
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=354675252
File Output Format Counters
Bytes Written=4499435
19/11/04 22:11:04 INFO streaming.StreamJob: Output directory: /user/rs6785/Project/Profile/outputCleanData
[rs6785@login-1-1 Project]$ hdfs dfs -ls Project/Profile/outputCleanData
Found 2 items
-rw-r--r--+ 3 rs6785 users   0 2019-11-04 22:11 Project/Profile/outputCleanData/_SUCCESS
-rw-r--r--+ 3 rs6785 users   4499435 2019-11-04 22:11 Project/Profile/outputCleanData/part-00000
[rs6785@login-1-1 Project]$
Type here to search
ENG
22:12
04-11-2019


rs6785@login-1-1:~/Project
2019/10 ['10/24/2019', '0:00', '40.708633', '-73.9268', '1', '0', 'bike', 'box truck', '', '', '']
2019/10 ['10/24/2019', '1:02', '40.732914', '-74.00398', '1', '0', 'taxi', 'bike', '', '', '']
2019/10 ['10/26/2019', '17:30', '40.749256', '-73.85801', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/08/2019', '19:25', '40.862717', '-73.9023', '1', '0', 'bike', 'box truck', '', '', '']
2019/10 ['10/21/2019', '9:45', '40.72898', '-73.882385', '1', '0', 'convertible', 'bike', '', '', '']
2019/10 ['10/25/2019', '20:35', '40.853714', '-73.92681', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/11/2019', '16:00', '40.740326', '-73.97312', '1', '0', 'bike', 'sedan', '', '', '']
2019/10 ['10/07/2019', '6:20', '', '', '1', '0', 'sedan', 'bike', '', '', '']
2019/10 ['10/15/2019', '23:50', '40.74915', '-73.98828', '1', '0', 'bike', 'taxi', '', '', '']
2019/10 ['10/01/2019', '8:30', '40.680622', '-73.84302', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/02/2019', '8:30', '40.71904', '-73.99153', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/19/2019', '14:20', '40.668293', '-73.93952', '1', '0', 'sedan', 'bike', '', '', '']
2019/10 ['10/05/2019', '14:58', '40.784473', '-73.967224', '1', '0', 'bike', 'bike', '', '', '']
2019/10 ['10/01/2019', '16:30', '40.734848', '-73.90002', '1', '0', 'pick-up truck', 'bike', '', '', '']
2019/10 ['10/26/2019', '18:21', '40.77994', '-73.97103', '1', '0', 'bike', '', '', '', '']
2019/10 ['10/29/2019', '9:10', '40.79599', '-73.96896', '1', '0', 'van', 'bike', '', '', '']
2019/10 ['10/14/2019', '20:30', '40.647625', '-73.97727', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/21/2019', '14:02', '40.68029', '-73.947586', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/12/2019', '7:40', '40.74309', '-73.97402', '1', '0', 'taxi', 'bike', '', '', '']
2019/10 ['10/12/2019', '18:22', '40.772243', '-73.990005', '1', '0', 'bike', 'sedan', '', '', '']
2019/10 ['10/17/2019', '9:30', '40.707726', '-73.7913', '1', '0', 'bike', 'sedan', '', '', '']
2019/10 ['10/01/2019', '17:30', '40.714214', '-73.94773', '1', '0', 'sedan', 'bike', '', '', '']
2019/10 ['10/07/2019', '17:30', '40.710533', '-73.95514', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/08/2019', '6:14', '40.691704', '-73.94849', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/24/2019', '23:10', '40.74584', '-73.82342', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/08/2019', '10:50', '40.631504', '-74.00296', '1', '0', 'bike', '', '', '', '']
2019/10 ['10/19/2019', '19:00', '40.760784', '-73.80677', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/17/2019', '18:15', '40.73723', '-74.000656', '1', '0', 'bike', 'taxi', '', '', '']
2019/10 ['10/13/2019', '12:30', '40.763893', '-73.915', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/14/2019', '22:09', '40.635445', '-73.998856', '1', '0', 'bike', 'station wagon/sport utility vehicle', '', '', '']
2019/10 ['10/21/2019', '13:00', '40.753437', '-73.88783', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/15/2019', '23:40', '40.794132', '-73.94281', '1', '0', 'taxi', 'bike', '', '', '']
2019/10 ['10/18/2019', '14:30', '40.683018', '-73.95885', '1', '0', 'limo', 'bike', '', '', '']
2019/10 ['10/24/2019', '23:59', '40.762486', '-73.96298', '1', '0', 'bike', 'sedan', '', '', '']
2019/10 ['10/01/2019', '15:20', '40.65461', '-73.922', '1', '0', 'sedan', 'bike', '', '', '']
2019/10 ['10/04/2019', '22:30', '40.741615', '-73.993744', '1', '0', 'taxi', 'bike', '', '', '']
2019/10 ['10/06/2019', '18:23', '40.709503', '-74.00167', '1', '0', 'sedan', 'bike', '', '', '']
2019/10 ['10/29/2019', '12:00', '40.692234', '-73.987305', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/14/2019', '20:00', '40.583668', '-73.94339', '1', '0', 'bike', 'pick-up truck', '', '', '']
2019/10 ['10/22/2019', '17:25', '40.68902', '-73.92122', '1', '0', 'station wagon/sport utility vehicle', 'bike', '', '', '']
2019/10 ['10/14/2019', '15:55', '40.642044', '-73.8849', '1', '0', 'sedan', 'bike', '', '', '']
2019/10 ['10/03/2019', '22:29', '40.681713', '-73.91152', '1', '0', 'bike', '', '', '', '']
[rs6785@login-1-1 Project]$
Type here to search
ENG
23:42
04-11-2019