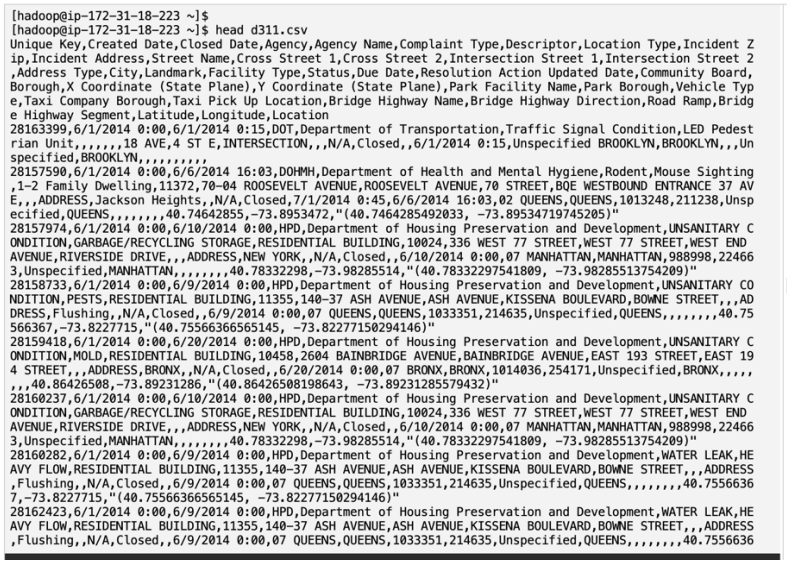
Soudabeh Rafieisakhaei

Download the d311.csv file from the Chapter 11 data set.

1. Open the file to understand its structure and identify column names.



```
[hadoop@ip-172-31-18-223 ~]$
[hadoop@ip-172-31-18-223 ~]$ head d311.csv
Unique Key,Created Date,Closed Date,Agency,Agency Name,Complaint Type,Descriptor,Location Type,Incident Z
ip,Incident Address,Street Name,Cross Street 1,Cross Street 2,Intersection Street 1,Intersection Street 2
,Address Type,City,Landmark,Facility Type,Status,Due Date,Resolution Action Updated Date,Community Board,
Borough,X Coordinate (State Plane),Y Coordinate (State Plane),Park Facility Name,Park Borough,Vehicle Typ
e,Taxi Company Borough,Taxi Pick Up Location,Bridge Highway Name,Bridge Highway Direction,Road Ramp,Bridg
e Highway Segment,Latitude,Longitude,Location
28163399,6/1/2014 0:00,6/1/2014 0:15,DOT,Department of Transportation,Traffic Signal Condition,LED Pedest
rian Unit,,,,,,18 AVE,4 ST E,INTERSECTION,,,N/A,Closed,,6/1/2014 0:15,Unspecified BROOKLYN,BROOKLYN,,,Un
specified,BROOKLYN,,,,,,,,,,
28157590,6/1/2014 0:00,6/6/2014 16:03,DOHMH,Department of Health and Mental Hygiene,Rodent,Mouse Sighting
,1-2 Family Dwelling,11372,70-04 ROOSEVELT AVENUE,ROOSEVELT AVENUE,70 STREET,BQE WESTBOUND ENTRANCE 37 AV
E,,,ADDRESS,Jackson Heights,,N/A,Closed,7/1/2014 0:45,6/6/2014 16:03,02 QUEENS,QUEENS,1013248,211238,Unsp
ecified,QUEENS,,,,,,,,,40.74642855,-73.8953472,"(40.7464285492033, -73.89534719745205)"
28157974,6/1/2014 0:00,6/10/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY C
ONDITION,GARBAGE/RECYCLING STORAGE,RESIDENTIAL BUILDING,10024,336 WEST 77 STREET,WEST 77 STREET,WEST END
AVENUE,RIVERSIDE DRIVE,,,ADDRESS,NEW YORK,,N/A,Closed,,6/10/2014 0:00,07 MANHATTAN,MANHATTAN,988998,22466
3,Unspecified,MANHATTAN,,,,,,,,,40.78332298,-73.98285514,"(40.78332297541809, -73.98285513754209)"
28158733,6/1/2014 0:00,6/9/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY CO
NDITION,PESTS,RESIDENTIAL BUILDING,11355,140-37 ASH AVENUE,ASH AVENUE,KISSENA BOULEVARD,BOWNE STREET,,,AD
DRESS,Flushing,,N/A,Closed,,6/9/2014 0:00,07 QUEENS,QUEENS,1033351,214635,Unspecified,QUEENS,,,,,,,,40.75
566367,-73.8227715,"(40.75566366565145, -73.82277150294146)"
28159418,6/1/2014 0:00,6/20/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY C
ONDITION,MOLD,RESIDENTIAL BUILDING,10458,2604 BAINBRIDGE AVENUE,BAINBRIDGE AVENUE,EAST 193 STREET,EAST 19
4 STREET,,,ADDRESS,BRONX,,N/A,Closed,,6/20/2014 0:00,07 BRONX,BRONX,1014036,254171,Unspecified,BRONX,,,,,
,,,40.86426508,-73.89231286,"(40.86426508198643, -73.89231285579432)"
28160237,6/1/2014 0:00,6/10/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY C
ONDITION,GARBAGE/RECYCLING STORAGE,RESIDENTIAL BUILDING,10024,336 WEST 77 STREET,WEST 77 STREET,WEST END
AVENUE,RIVERSIDE DRIVE,,,ADDRESS,NEW YORK,,N/A,Closed,,6/10/2014 0:00,07 MANHATTAN,MANHATTAN,988998,22466
3,Unspecified,MANHATTAN,,,,,,,,,40.78332298,-73.98285514,"(40.78332297541809, -73.98285513754209)"
28160282,6/1/2014 0:00,6/9/2014 0:00,HPD,Department of Housing Preservation and Development,WATER LEAK,HE
AVY FLOW,RESIDENTIAL BUILDING,11355,140-37 ASH AVENUE,ASH AVENUE,KISSENA BOULEVARD,BOWNE STREET,,,ADDRESS
,Flushing,,N/A,Closed,,6/9/2014 0:00,07 QUEENS,QUEENS,1033351,214635,Unspecified,QUEENS,,,,,,,,40.7556636
7,-73.8227715,"(40.75566366565145, -73.82277150294146)"
28162423,6/1/2014 0:00,6/9/2014 0:00,HPD,Department of Housing Preservation and Development,WATER LEAK,HE
AVY FLOW,RESIDENTIAL BUILDING,11355,140-37 ASH AVENUE,ASH AVENUE,KISSENA BOULEVARD,BOWNE STREET,,,ADDRESS
,Flushing,,N/A,Closed,,6/9/2014 0:00,07 QUEENS,QUEENS,1033351,214635,Unspecified,QUEENS,,,,,,,,40.7556636
```

2. Create a subdirectory RDD/FinalExam in HDFS and upload the csv file to that subdirectory. Start the Spark Shell.

```
[hadoop@ip-172-31-28-13 ~]$ hadoop fs -mkdir RDD
[hadoop@ip-172-31-28-13 ~]$ hadoop fs -mkdir RDD/FinalExam
[hadoop@ip-172-31-28-13 ~]$ hadoop fs -copyFromLocal d311.csv RDD/FinalExam
```

```
[hadoop@ip-172-31-28-13 ~]$ hadoop fs -cat RDD/FinalExam/d311.csv
```

```
EEEEEEEEEEEEEEEEEEE MMMMMMM           MMMMMMM RRRRRRRRRRRRRR
E::::::::::::::::::E M:::::::M         M:::::::M R::::::::::::::R
EE:::::EEEEEEEEE:::E M::::::::M       M::::::::M R:::::RRRRRR:::::R
  E::::E       EEEEE M:::::::::M     M:::::::::M RR::::R      R::::R
  E::::E             M::::::M:::M   M:::M::::::M   R:::R      R::::R
  E:::::EEEEEEEEEE   M:::::M M:::M M:::M M:::::M   R:::RRRRRR:::::R
  E::::::::::::::E   M:::::M  M:::M:::M  M:::::M   R:::::::::::RR
  E:::::EEEEEEEEEE   M:::::M   M:::::M   M:::::M   R:::RRRRRR:::::R
  E::::E             M:::::M    M:::M    M:::::M   R:::R      R::::R
  E::::E       EEEEE M:::::M     MMM     M:::::M   R:::R      R::::R
EE:::::EEEEEEEE:::E M:::::M             M:::::M   R:::R      R::::R
E::::::::::::::::::E M:::::M             M:::::M RR::::R      R::::R
EEEEEEEEEEEEEEEEEEE MMMMMMM             MMMMMMM RRRRRRR      RRRRRR

[hadoop@ip-172-31-18-203 ~]$ spark-shell
May 15, 2024 5:29:05 PM org.apache.spark.launcher.Log4jHotPatchOption staticJavaAgentOption
WARNING: spark.log4jHotPatch.enabled is set to true, but /usr/share/log4j-cve-2021-44228-hotpatch/jdk17/L
og4jHotPatchFat.jar does not exist at the configured location

Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
24/05/15 17:29:18 WARN Client: Neither spark.yarn.jars nor spark.yarn.archive is set, falling back to upl
oading libraries under SPARK_HOME.
Spark context Web UI available at http://ip-172-31-18-203.ec2.internal:4040
Spark context available as 'sc' (master = yarn, app id = application_1715793901468_0001).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.0-amzn-0
      /_/

Using Scala version 2.12.17 (OpenJDK 64-Bit Server VM, Java 17.0.11)
Type in expressions to have them evaluated.
Type :help for more information.

scala>
```

3. Check if there is any header in the file. If there is a header in the first row, then remove it.

```
scala> val rdd=sc.textFile("RDD/FinalExam/d311.csv")
rdd: org.apache.spark.rdd.RDD[String] = RDD/FinalExam/d311.csv MapPartitionsRDD[1] at textFile at <console>:23

scala> rdd.first
res0: String = Unique Key,Created Date,Closed Date,Agency,Agency Name,Complaint Type,Descriptor,Location Type,Incident Zip,Incident
 Address,Street Name,Cross Street 1,Cross Street 2,Intersection Street 1,Intersection Street 2,Address Type,City,Landmark,Facility
Type,Status,Due Date,Resolution Action Updated Date,Community Board,Borough,X Coordinate (State Plane),Y Coordinate (State Plane),P
ark Facility Name,Park Borough,Vehicle Type,Taxi Company Borough,Taxi Pick Up Location,Bridge Highway Name,Bridge Highway Direction
,Road Ramp,Bridge Highway Segment,Latitude,Longitude,Location

scala> val header=rdd.first
header: String = Unique Key,Created Date,Closed Date,Agency,Agency Name,Complaint Type,Descriptor,Location Type,Incident Zip,Incide
nt Address,Street Name,Cross Street 1,Cross Street 2,Intersection Street 1,Intersection Street 2,Address Type,City,Landmark,Facilit
y Type,Status,Due Date,Resolution Action Updated Date,Community Board,Borough,X Coordinate (State Plane),Y Coordinate (State Plane)
,Park Facility Name,Park Borough,Vehicle Type,Taxi Company Borough,Taxi Pick Up Location,Bridge Highway Name,Bridge Highway Directi
on,Road Ramp,Bridge Highway Segment,Latitude,Longitude,Location

scala> val dataRDD = rdd.filter(line => line !=header)
dataRDD: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at filter at <console>:24
```

4. Create an RDD that reads the csv file and displays the first 10 elements. Provide a screenshot of the results. Use the count action to return the number of items in the RDD.

```
scala> val firstTen = rdd.take(10)
firstTen: Array[String] = Array(Unique Key,Created Date,Closed Date,Agency,Agency Name,Complaint Type,Descriptor,Location Type,Incident Zip,Incident Address,Street Name,Cross Street 1,Cross Street 2,Intersectio
n Street 1,Intersection Street 2,Address Type,City,Landmark,Facility Type,Status,Due Date,Resolution Action Updated Date,Community Board,Borough,X Coordinate (State Plane),Y Coordinate (State Plane),Park Facili
ty Name,Park Borough,Vehicle Type,Taxi Company Borough,Taxi Pick Up Location,Bridge Highway Name,Bridge Highway Direction,Road Ramp,Bridge Highway Segment,Latitude,Longitude,Location, 28163399,6/1/2014 0:00,6/1
/2014 0:15,DOT,Department of Transportation,Traffic Signal Condition,LED Pedestrian Unit,,,,,,,18 AVE,4 ST E,INTERSECTION,,,N/A,Closed,,6/1/2...

scala> firstTen.foreach(println)
Unique Key,Created Date,Closed Date,Agency,Agency Name,Complaint Type,Descriptor,Location Type,Incident Zip,Incident Address,Street Name,Cross Street 1,Cross Street 2,Intersection Street 1,Intersection Street 2
,Address Type,City,Landmark,Facility Type,Status,Due Date,Resolution Action Updated Date,Community Board,Borough,X Coordinate (State Plane),Y Coordinate (State Plane),Park Facility Name,Park Borough,Vehicle Typ
e,Taxi Company Borough,Taxi Pick Up Location,Bridge Highway Name,Bridge Highway Direction,Road Ramp,Bridge Highway Segment,Latitude,Longitude,Location
28163399,6/1/2014 0:00,6/1/2014 0:15,DOT,Department of Transportation,Traffic Signal Condition,LED Pedestrian Unit,,,,,,,18 AVE,4 ST E,INTERSECTION,,,N/A,Closed,,6/1/2014 0:15,Unspecified BROOKLYN,BROOKLYN,,,Un
specified,BROOKLYN,,,,,,,,,,
28157590,6/1/2014 0:00,6/6/2014 16:03,DOHMH,Department of Health and Mental Hygiene,Rodent,Mouse Sighting,1-2 Family Dwelling,11372,70-04 ROOSEVELT AVENUE,ROOSEVELT AVENUE,70 STREET,BQE WESTBOUND ENTRANCE 37 AV
E,,,ADDRESS,Jackson Heights,,N/A,Closed,7/1/2014 0:45,6/6/2014 16:03,02 QUEENS,QUEENS,1013248,211238,Unspecified,QUEENS,,,,,,,,40.74642855,-73.8953472,"(40.7464285492033, -73.89534719745205)"
28157974,6/1/2014 0:00,6/10/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY CONDITION,GARBAGE/RECYCLING STORAGE,RESIDENTIAL BUILDING,10024,336 WEST 77 STREET,WEST 77 STREET,WEST END
AVENUE,RIVERSIDE DRIVE,,,ADDRESS,NEW YORK,,N/A,Closed,,6/10/2014 0:00,07 MANHATTAN,MANHATTAN,988998,224663,Unspecified,MANHATTAN,,,,,,,,40.78332298,-73.98285514,"(40.78332297541809, -73.98285513754209)"
28158733,6/1/2014 0:00,6/9/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY CONDITION,PESTS,RESIDENTIAL BUILDING,11355,140-37 ASH AVENUE,ASH AVENUE,KISSENA BOULEVARD,BOWNE STREET,,,AD
DRESS,Flushing,,N/A,Closed,,6/9/2014 0:00,07 QUEENS,QUEENS,1033351,214635,Unspecified,QUEENS,,,,,,,,40.75566367,-73.8227715,"(40.75566366565145, -73.82277150294146)"
28159418,6/1/2014 0:00,6/20/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY CONDITION,MOLD,RESIDENTIAL BUILDING,10458,2604 BAINBRIDGE AVENUE,BAINBRIDGE AVENUE,EAST 193 STREET,EAST 19
4 STREET,,,ADDRESS,BRONX,,N/A,Closed,,6/20/2014 0:00,07 BRONX,BRONX,1014036,254171,Unspecified,BRONX,,,,,,,,40.86426508,-73.89231286,"(40.86426508198643, -73.89231285579432)"
28160237,6/1/2014 0:00,6/10/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY CONDITION,GARBAGE/RECYCLING STORAGE,RESIDENTIAL BUILDING,10024,336 WEST 77 STREET,WEST 77 STREET,WEST END
AVENUE,RIVERSIDE DRIVE,,,ADDRESS,NEW YORK,,N/A,Closed,,6/10/2014 0:00,07 MANHATTAN,MANHATTAN,988998,224663,Unspecified,MANHATTAN,,,,,,,,40.78332298,-73.98285514,"(40.78332297541809, -73.98285513754209)"
28160282,6/1/2014 0:00,6/9/2014 0:00,HPD,Department of Housing Preservation and Development,WATER LEAK,HEAVY FLOW,RESIDENTIAL BUILDING,11355,140-37 ASH AVENUE,ASH AVENUE,KISSENA BOULEVARD,BOWNE STREET,,,ADDRESS
,Flushing,,N/A,Closed,,6/9/2014 0:00,07 QUEENS,QUEENS,1033351,214635,Unspecified,QUEENS,,,,,,,,40.75566367,-73.8227715,"(40.75566366565145, -73.82277150294146)"
28162423,6/1/2014 0:00,6/9/2014 0:00,HPD,Department of Housing Preservation and Development,WATER LEAK,HEAVY FLOW,RESIDENTIAL BUILDING,11355,140-37 ASH AVENUE,ASH AVENUE,KISSENA BOULEVARD,BOWNE STREET,,,ADDRESS
,Flushing,,N/A,Closed,,6/9/2014 0:00,07 QUEENS,QUEENS,1033351,214635,Unspecified,QUEENS,,,,,,,,40.75566367,-73.8227715,"(40.75566366565145, -73.82277150294146)"
28162488,6/1/2014 0:00,6/20/2014 0:00,HPD,Department of Housing Preservation and Development,UNSANITARY CONDITION,MOLD,RESIDENTIAL BUILDING,10458,2604 BAINBRIDGE AVENUE,BAINBRIDGE AVENUE,EAST 193 STREET,EAST 19
4 STREET,,,ADDRESS,BRONX,,N/A,Closed,,6/20/2014 0:00,07 BRONX,BRONX,1014036,254171,Unspecified,BRONX,,,,,,,,40.86426508,-73.89231286,"(40.86426508198643, -73.89231285579432)"
```

```
scala> val count = rdd.count()
count: Long = 518910

scala> println("Number of items in RDD: " + count)
Number of items in RDD: 518910
```

5. Create a new RDD that captures only the Agency, City, and Descriptor.

```
scala> val selectedRDD = dataRDD.map(line => line.split(',')).map(values => Row(values(3), values(16), values(6)))
selectedRDD: org.apache.spark.rdd.RDD[org.apache.spark.sql.Row] = MapPartitionsRDD[4] at map at <console>:24
```

6. Display the first few elements of the new RDD. Provide a screenshot of the result.

```
scala> selectedRDD.take(5).foreach(println)
[DOT,,LED Pedestrian Unit]
[DOHMH,Jackson Heights,Mouse Sighting]
[HPD,NEW YORK,GARBAGE/RECYCLING STORAGE]
[HPD,Flushing,PESTS]
[HPD,BRONX,MOLD]
```

7. Create a new RDD that captures City and Descriptor, where the descriptor contains the word "Sidewalk". Provide a screenshot of the result.

```
scala> val sidewalkRDD = dataRDD  .map(line => line.split(','))  .filter(values => values(6).contains("Sidewalk"))  .map(values =>
(values(16), values(6)))
sidewalkRDD: org.apache.spark.rdd.RDD[(String, String)] = MapPartitionsRDD[7] at map at <console>:24

scala> sidewalkRDD.take(5).foreach(println)
(NEW YORK,Broken Sidewalk)
(NEW YORK,E3 Dirty Sidewalk)
(FRESH MEADOWS,Blocked Sidewalk)
(COLLEGE POINT,Blocked Sidewalk)
(STATEN ISLAND,Blocked Sidewalk)
```

8. Save the results of the RDD from #7 back into the cluster. Open another terminal and verify that the results are stored in the cluster. Provide a screenshot of the result.

```
scala> sidewalkRDD.saveAsTextFile("RDD/FinalExam/sidewalk_results")
```

```
[hadoop@ip-172-31-21-129 ~]$ hadoop fs -ls RDD/FinalExam
Found 2 items
-rw-r--r--   1 hadoop hdfsadmingroup  188659283 2024-05-15 05:35 RDD/FinalExam/d311.csv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-05-15 07:21 RDD/FinalExam/sidewalk_results
```