Soudabeh Rafieisakhaei

Download the accounts.csv and us_zip_codes.csv. files from the Chapter 10 data set.

1. Download the data to your computer and upload them to the HDFS subdirectory SPRK/FinalExam.

```
accounts.csv  building.txt  client.txt  manager.txt  us_zip_codes.csv
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -mkdir SPRK
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -mkdir SPRK/FinalExam
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -copyFromLocal us_zip_codes.csv SPRK/FinalExam
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -copyFromLocal accounts.csv SPRK/FinalExam
```

2. Start spark-shell and create two DataFrames: accountsDF and zipCodesDF. Display the schemas for both DataFrames and provide a screenshot of the results.

```
scala>  val zipcodesDF = spark.read.format("csv").option("header", "true").load("SPRK/FinalExam/us_zip_codes.csv")
zipcodesDF: org.apache.spark.sql.DataFrame = [zip: string, lat: string ... 16 more fields]

scala>  val accountsDF = spark.read.format("csv").option("header", "true").load("SPRK/FinalExam/accounts.csv")
accountsDF: org.apache.spark.sql.DataFrame = [first_name: string, last_name: string ... 7 more fields]

scala> zipcodesDF.printSchema()
root
 |-- zip: string (nullable = true)
 |-- lat: string (nullable = true)
 |-- lng: string (nullable = true)
 |-- city: string (nullable = true)
 |-- state_id: string (nullable = true)
 |-- state_name: string (nullable = true)
 |-- zcta: string (nullable = true)
 |-- parent_zcta: string (nullable = true)
 |-- population: string (nullable = true)
 |-- density: string (nullable = true)
 |-- county_fips: string (nullable = true)
 |-- county_name: string (nullable = true)
 |-- county_weights: string (nullable = true)
 |-- county_names_all: string (nullable = true)
 |-- county_fips_all: string (nullable = true)
 |-- imprecise: string (nullable = true)
 |-- military: string (nullable = true)
 |-- timezone: string (nullable = true)


scala> accountsDF.printSchema()
root
 |-- first_name: string (nullable = true)
 |-- last_name: string (nullable = true)
 |-- company_name: string (nullable = true)
 |-- address: string (nullable = true)
 |-- zip: string (nullable = true)
 |-- phone1: string (nullable = true)
 |-- phone2: string (nullable = true)
 |-- email: string (nullable = true)
 |-- web: string (nullable = true)
```

3. Perform a simple query by selecting two columns from the accountsDF Use an alias for the second. Provide a screenshot of the result.

```
scala> val selectedDF = accountsDF.select(accountsDF("first_name"), accountsDF("last_name").alias("surname"))
selectedDF: org.apache.spark.sql.DataFrame = [first_name: string, surname: string]

scala> selectedDF.show()
+----------+---------+
|first_name|  surname|
+----------+---------+
|     James|     Butt|
| Josephine|  Darakjy|
|       Art|   Venere|
|     Lenna| Paprocki|
|   Donette|   Foller|
|    Simona|  Morasca|
|    Mitsue|  Tollner|
|     Leota| Dilliard|
|      Sage|   Wieser|
|      Kris|  Marrier|
|     Minna|   Amigon|
|      Abel|  Maclead|
|     Kiley|Caldarera|
|  Graciela|     Ruta|
|     Cammy|  Albares|
|    Mattie| Poquette|
|   Meaghan|   Garufi|
|    Gladys|      Rim|
|      Yuki|  Whobrey|
|  Fletcher|    Flosi|
+----------+---------+
only showing top 20 rows
```

4. Perform a query that results in a DataFrame that has just first_name and last_name columns and only includes users whose last name begins with a given letter. (For example, if you choose the letter to be "A," then all users whose last name starts with "A" should be displayed.) Provide a screenshot of the result.

```
scala> val filteredDF = accountsDF.filter(accountsDF("last_name").startsWith("N"))
filteredDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [first_name: string, last_name: string ... 7 more fields]

scala> val filteredresultDF = filteredDF.select("first_name", "last_name")
filteredresultDF: org.apache.spark.sql.DataFrame = [first_name: string, last_name: string]

scala> filteredresultDF.show()
+----------+-----------+
|first_name|  last_name|
+----------+-----------+
|     Bette|      Nicka|
|    Lorrie|     Nestle|
|   Matthew|    Neither|
|  Herminia|Nicolozakes|
|   Adelina|    Nabours|
|     Tarra|     Nachor|
|     Erick|     Nievas|
|     Lenna|   Newville|
|     Rikki|      Nayar|
|      Gary|     Nunlee|
+----------+-----------+
```

5. Query the accountsDF DataFrame using groupBy with count to find out the total number of people sharing each last name. Display only five records. Provide a screenshot of the result.

```
scala> val countByLastName = accountsDF.groupBy("last_name").count()
countByLastName: org.apache.spark.sql.DataFrame = [last_name: string, count: bigint]

scala> countByLastName.show(5)
+----------+-----+
| last_name|count|
+----------+-----+
|    Coyier|    1|
|Eschberger|    1|
|   Hirpara|    1|
|     Galam|    1|
|   Mallett|    1|
+----------+-----+
only showing top 5 rows
```

6. Create a new DataFrame that joins the two original DataFrames—accountsDF and zipCodesDF—by the zip code. Display the first ten records. Provide a screenshot of the result.

```
scala> val joinedDF = accountsDF.join(zipcodesDF,"zip")
joinedDF: org.apache.spark.sql.DataFrame = [zip: string, first_name: string ... 24 more fields]

scala> joinedDF.show(10)
24/05/14 07:42:16 WARN SparkStringUtils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.sql.debug.maxToStringFields'.
+-----+----------+---------+--------------------+------------------+------------+------------+--------------------+--------------------+--------+---------+--------+--------+----------+----+-----------+-------
---+-------+-----------+-----------+--------------+-----------------+---------------+--------+--------+----------------+
|  zip|first_name|last_name|        company_name|           address|      phone1|      phone2|               email|                 web|     lat|      lng|    city|state_id|state_name|zcta|parent_zcta|populat
ion|density|county_fips|county_name|county_weights|county_names_all|county_fips_all|imprecise|military|        timezone|
+-----+----------+---------+--------------------+------------------+------------+------------+--------------------+--------------------+--------+---------+--------+--------+----------+----+-----------+-------
---+-------+-----------+-----------+--------------+-----------------+---------------+--------+--------+----------------+
|10002|     Ozell|   Shealy|     Silver Bros Inc|      8 Industry Ln|212-332-8435|212-880-8865| oshealy@hotmail.com|http://www.silver...|40.71586|-73.98613|New York|      NY|  New York|TRUE|       NULL|     74
993|35781.9|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10003|     Brock| Bolognia|         Orinda News|     4486 W O St #1|212-402-9216|212-617-5063| bbolognia@yahoo.com|http://www.orinda...| 40.7318|-73.98911|New York|      NY|  New York|TRUE|       NULL|     54
682|37524.3|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10004|     Mirta|  Mallett|Stephen Kennerly ...|  7 S San Marcos Rd|212-870-1286|212-745-6948|mirta_mallett@gma...|http://www.stephe...|40.69465|-74.02106|New York|      NY|  New York|TRUE|       NULL|      3
028| 2214.8|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10009|     Tawna|   Buvens|H H H Enterprises...|3305 Nabell Ave #679|212-674-9610|212-462-9157|     tawna@gmail.com|http://www.hhhent...|40.72664|-73.97858|New York|      NY|  New York|TRUE|       NULL|     57
925|38399.1|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10011|     Layla|  Springe|  Chadds Ford Winery|   229 N Forty Driv|212-260-3151|212-253-7448|layla.springe@cox...|http://www.chadds...|40.74187|-74.00052|New York|      NY|  New York|TRUE|       NULL|     50
472|29744.0|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10011|      Jose| Stockham|Tri State Refuele...|    128 Bransten Rd|212-675-8570|212-569-4233|      jose@yahoo.com|http://www.trista...|40.74187|-74.00052|New York|      NY|  New York|TRUE|       NULL|     50
472|29744.0|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10011|    Willow|    Kusko|           U Pull It|  90991 Thorburn Ave|212-582-4976|212-934-5167|    wkusko@yahoo.com|http://www.upulli...|40.74187|-74.00052|New York|      NY|  New York|TRUE|       NULL|     50
472|29744.0|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10013|     Cyril| Daufeldt|Galaxy Internatio...|       3 Lawton St|212-745-8484|212-422-5427|cyril_daufeldt@da...|http://www.galaxy...|40.72014|-74.00476|New York|      NY|  New York|TRUE|       NULL|     28
709|19437.4|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10013|    Derick|   Dhamer|Studer, Eugene A Esq|   87163 N Main Ave|212-304-4515|212-225-9676|     ddhamer@cox.net|http://www.studer...|40.72014|-74.00476|New York|      NY|  New York|TRUE|       NULL|     28
709|19437.4|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
|10016|      Jess| Chaffins|New York Public L...|        18 3rd Ave|212-510-4633|212-428-9538|jess.chaffins@cha...|http://www.newyor...|40.74517|-73.97834|New York|      NY|  New York|TRUE|       NULL|     51
057|39315.3|      36061|   New York| {'36061':100}|        New York|          36061|    FALSE|   FALSE|America/New_York|
+-----+----------+---------+--------------------+------------------+------------+------------+--------------------+--------------------+--------+---------+--------+--------+----------+----+-----------+-------
---+-------+-----------+-----------+--------------+-----------------+---------------+--------+--------+----------------+
only showing top 10 rows
```

7. Save the results of the last DataFrame in HDFS in the SPRK/FinalExam Exit Spark and perform a cat HDFS command to display the records in the saved file.

```
scala> joinedDF.write.save("joined_saved")

scala> joinedDF.write.json("SPRK/FinalExam/joined_saved")

scala> joinedDF.write.format("csv").save("SPRK/FinalExam/joined_save_csv")

scala> sys.exit
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -ls SPRK/FinalExam
Found 5 items
-rw-r--r--   1 hadoop hdfsadmingroup      71025 2024-05-14 06:25 SPRK/FinalExam/accounts.csv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-05-14 07:33 SPRK/FinalExam/joined_save
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-05-14 07:46 SPRK/FinalExam/joined_save_csv
drwxr-xr-x   - hadoop hdfsadmingroup          0 2024-05-14 07:46 SPRK/FinalExam/joined_saved
-rw-r--r--   1 hadoop hdfsadmingroup    5998595 2024-05-14 06:24 SPRK/FinalExam/us_zip_codes.csv
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -ls SPRK/FinalExam/joined_save_csv
Found 3 items
-rw-r--r--   1 hadoop hdfsadmingroup          0 2024-05-14 07:46 SPRK/FinalExam/joined_save_csv/_SUCCESS
-rw-r--r--   1 hadoop hdfsadmingroup      83341 2024-05-14 07:46 SPRK/FinalExam/joined_save_csv/part-00000-479aff69-a508-4339-9b32-4870a5a042ea-c000.csv
-rw-r--r--   1 hadoop hdfsadmingroup      32608 2024-05-14 07:46 SPRK/FinalExam/joined_save_csv/part-00001-479aff69-a508-4339-9b32-4870a5a042ea-c000.csv
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -cat SPRK/FinalExam/joined_save_csv
cat: `SPRK/FinalExam/joined_save_csv': Is a directory
[hadoop@ip-172-31-31-211 ~]$ hadoop fs -cat SPRK/FinalExam/joined_save_csv/part-*
10002,Ozell,Shealy,Silver Bros Inc,8 Industry Ln,212-332-8435,212-880-8865,oshealy@hotmail.com,http://www.silverbrosinc.com,40.71586,-73.98613,New York,NY,New York,TRUE,,74993,35781.9,36061,New York,{'36061':10
0},New York,36061,FALSE,FALSE,America/New_York
10003,Brock,Bolognia,Orinda News,4486 W O St #1,212-402-9216,212-617-5063,bbolognia@yahoo.com,http://www.orindanews.com,40.7318,-73.98911,New York,NY,New York,TRUE,,54682,37524.3,36061,New York,{'36061':100},Ne
w York,36061,FALSE,FALSE,America/New_York
10004,Mirta,Mallett,Stephen Kennerly Archts Inc Pc,7 S San Marcos Rd,212-870-1286,212-745-6948,mirta_mallett@gmail.com,http://www.stephenkennerlyarchtsincpc.com,40.69465,-74.02106,New York,NY,New York,TRUE,,302
8,2214.8,36061,New York,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10009,Tawna,Buvens,H H H Enterprises Inc,3305 Nabell Ave #679,212-674-9610,212-462-9157,tawna@gmail.com,http://www.hhhenterprisesinc.com,40.72664,-73.97858,New York,NY,New York,TRUE,,57925,38399.1,36061,New Yor
k,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10011,Layla,Springe,Chadds Ford Winery,229 N Forty Driv,212-260-3151,212-253-7448,layla.springe@cox.net,http://www.chaddsfordwinery.com,40.74187,-74.00052,New York,NY,New York,TRUE,,50472,29744.0,36061,New York
,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10011,Jose,Stockham,Tri State Refueler Co,128 Bransten Rd,212-675-8570,212-569-4233,jose@yahoo.com,http://www.tristaterefuelerco.com,40.74187,-74.00052,New York,NY,New York,TRUE,,50472,29744.0,36061,New York,{'
36061':100},New York,36061,FALSE,FALSE,America/New_York
10011,Willow,Kusko,U Pull It,90991 Thorburn Ave,212-582-4976,212-934-5167,wkusko@yahoo.com,http://www.upullit.com,40.74187,-74.00052,New York,NY,New York,TRUE,,50472,29744.0,36061,New York,{'36061':100},New Yor
k,36061,FALSE,FALSE,America/New_York
10013,Cyril,Daufeldt,Galaxy International Inc,3 Lawton St,212-745-8484,212-422-5427,cyril_daufeldt@daufeldt.com,http://www.galaxyinternationalinc.com,40.72014,-74.00476,New York,NY,New York,TRUE,,28709,19437.4,
36061,New York,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10013,Derick,Dhamer,"Studer, Eugene A Esq",87163 N Main Ave,212-304-4515,212-225-9676,ddhamer@cox.net,http://www.studereugeneaesq.com,40.72014,-74.00476,New York,NY,New York,TRUE,,28709,19437.4,36061,New York,{
'36061':100},New York,36061,FALSE,FALSE,America/New_York
10016,Jess,Chaffins,New York Public Library,18 3rd Ave,212-510-4633,212-428-9538,jess.chaffins@chaffins.org,http://www.newyorkpubliclibrary.com,40.74517,-73.97834,New York,NY,New York,TRUE,,51057,39315.3,36061,
New York,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10016,Haydee,Denooyer,Cleaning Station Inc,25346 New Rd,212-792-8658,212-782-3493,hdenooyer@denooyer.org,http://www.cleaningstationinc.com,40.74517,-73.97834,New York,NY,New York,TRUE,,51057,39315.3,36061,New Y
ork,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10025,Alishia,Sergi,Milford Enterprises Inc,2742 Distribution Way,212-860-1579,212-753-2740,asergi@gmail.com,http://www.milfordenterprisesinc.com,40.79857,-73.96659,New York,NY,New York,TRUE,,92805,48594.3,3606
1,New York,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10038,Fausto,Agramonte,Marriott Hotels Resorts Suites,5 Harrison Rd,212-313-1783,212-778-3063,fausto_agramonte@yahoo.com,http://www.marriotthotelsresortssuites.com,40.7092,-74.00284,New York,NY,New York,TRUE,,2
2800,26251.9,36061,New York,{'36061':100},New York,36061,FALSE,FALSE,America/New_York
10309,Timothy,Mulqueen,Saronix Nymph Products,44 W 4th St,718-332-6527,718-654-7063,timothy_mulqueen@mulqueen.org,http://www.saronixnymphproducts.com,40.53132,-74.22056,Staten Island,NY,New York,TRUE,,33531,188
6.3,36085,Richmond,{'36085':100},Richmond,36085,FALSE,FALSE,America/New_York
10468,Bok,Isaacs,Nelson Hawaiian Ltd,6 Gilson St,718-809-3762,718-478-8568,bok.isaacs@aol.com,http://www.nelsonhawaiianltd.com,40.86805,-73.90009,Bronx,NY,New York,TRUE,,78647,27442.2,36005,Bronx,{'36005':100},
Bronx,36005,FALSE,FALSE,America/New_York
10536,Leslie,Threets,C W D C Metal Fabricators,2 A Kelley Dr,914-861-9748,914-396-2615,leslie@cox.net,http://www.cwdcmetalfabricators.com,41.26983,-73.68725,Katonah,NY,New York,TRUE,,10853,161.8,36119,Westchest
er,{'36119':100},Westchester,36119,FALSE,FALSE,America/New_York
10553,Nana,Wrinkles,"Ray, Milbern D",6 Van Buren St,914-855-2115,914-796-3775,nana@aol.com,http://www.raymilbernd.com,40.90854,-73.82173,Mount Vernon,NY,New York,TRUE,,10429,5670.5,36119,Westchester,{'36119':10
0},Westchester,36119,FALSE,FALSE,America/New_York
```