from pyspark.sql import SparkSession  
from pyspark.sql.functions import split, col, size  
  
  
spark = SparkSession.builder.master(**"local"**).appName(**"city\_report"**).getOrCreate()  
  
df = spark.read.format(**"parquet"**).option(**"header"**,**"true"**).option(**"inferSchema"**,**"true"**) \  
 .load(**"C:**\\**Users**\\**zakeerm**\\**Downloads**\\**PysparkLearning**\\**resources**\\**us\_cities\_dimension.parquet"**)  
df.printSchema()  
df.show()  
  
print(**"cleansing the data"**)  
cityReportDF = df.select(**"city"**,**"state\_id"**,**"state\_name"**,**"county\_name"**,**"population"**,**"zips"**).show()  
  
cityFinalDF = df.select(col(**"city"**).alias(**"CITY"**),  
 col(**"state\_id"**).alias(**"STATE\_ID"**),col(**"state\_name"**).alias(**"STATE\_NAME"**),  
 col(**"county\_name"**).alias(**"COUNTY\_NAME"**)).show()  
  
print(**"Calculate the Number of zips in each city."**)  
cityDF = df.withColumn(**"zips"**, split(**"zips"**, **","**))  
  
result\_df = cityDF.withColumn(**"num\_zips"**, size(**"zips"**))  
  
result\_df.select(**"zips"**,**"num\_zips"**,**"city"**,**"state\_name"**,**"county\_name"**,**"population"**)  
  
result\_df.show()  
result\_df.write.format(**"json"**).mode(**"overwrite"**).option(**"compression"**, **"bzip2"**).save(**"hdfs\_path"**)

Output =>

C:\Users\zakeerm\IdeaProjects\Spark-Scala\venv\Scripts\python.exe C:/Users/zakeerm/Downloads/PysparkLearning/pysparkprograms/city.py

23/11/16 13:11:57 WARN Shell: Did not find winutils.exe: java.io.FileNotFoundException: java.io.FileNotFoundException: HADOOP\_HOME and hadoop.home.dir are unset. -see https://wiki.apache.org/hadoop/WindowsProblems

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

23/11/16 13:11:58 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

root

|-- city: string (nullable = true)

|-- city\_ascii: string (nullable = true)

|-- state\_id: string (nullable = true)

|-- state\_name: string (nullable = true)

|-- county\_fips: integer (nullable = true)

|-- county\_name: string (nullable = true)

|-- lat: double (nullable = true)

|-- lng: double (nullable = true)

|-- population: integer (nullable = true)

|-- density: integer (nullable = true)

|-- timezone: string (nullable = true)

|-- zips: string (nullable = true)

+-------------+-------------+--------+--------------------+-----------+--------------------+-------+---------+----------+-------+-------------------+--------------------+

| city| city\_ascii|state\_id| state\_name|county\_fips| county\_name| lat| lng|population|density| timezone| zips|

+-------------+-------------+--------+--------------------+-----------+--------------------+-------+---------+----------+-------+-------------------+--------------------+

| New York| New York| NY| New York| 36061| New York|40.6943| -73.9249| 18713220| 10715| America/New\_York|11229 11226 11225...|

| Los Angeles| Los Angeles| CA| California| 6037| Los Angeles|34.1139|-118.4068| 12750807| 3276|America/Los\_Angeles|90291 90293 90292...|

| Chicago| Chicago| IL| Illinois| 17031| Cook|41.8373| -87.6862| 8604203| 4574| America/Chicago|60018 60649 60641...|

| Miami| Miami| FL| Florida| 12086| Miami-Dade|25.7839| -80.2102| 6445545| 5019| America/New\_York|33129 33125 33126...|

| Dallas| Dallas| TX| Texas| 48113| Dallas|32.7936| -96.7662| 5743938| 1526| America/Chicago|75287 75098 75233...|

| Philadelphia| Philadelphia| PA| Pennsylvania| 42101| Philadelphia|40.0077| -75.1339| 5649300| 4554| America/New\_York|19154 19151 19150...|

| Houston| Houston| TX| Texas| 48201| Harris|29.7863| -95.3889| 5464251| 1399| America/Chicago|77069 77068 77061...|

| Atlanta| Atlanta| GA| Georgia| 13121| Fulton|33.7627| -84.4224| 5449398| 1441| America/New\_York|30334 30331 30332...|

| Washington| Washington| DC|District of Columbia| 11001|District of Columbia|38.9047| -77.0163| 5379184| 4457| America/New\_York|20010 20011 20012...|

| Boston| Boston| MA| Massachusetts| 25025| Suffolk|42.3188| -71.0846| 4688346| 5532| America/New\_York|02120 02121 02122...|

| Phoenix| Phoenix| AZ| Arizona| 4013| Maricopa|33.5722|-112.0891| 4219697| 1253| America/Phoenix|85008 85009 85006...|

| Seattle| Seattle| WA| Washington| 53033| King|47.6211|-122.3244| 3789215| 3469|America/Los\_Angeles|98109 98108 98104...|

|San Francisco|San Francisco| CA| California| 6075| San Francisco|37.7562| -122.443| 3592294| 7256|America/Los\_Angeles|94130 94131 94132...|

| Detroit| Detroit| MI| Michigan| 26163| Wayne|42.3834| -83.1024| 3506126| 1864| America/Detroit|48209 48208 48201...|

| San Diego| San Diego| CA| California| 6073| San Diego|32.8312|-117.1225| 3220118| 1686|America/Los\_Angeles|92109 92108 92103...|

| Minneapolis| Minneapolis| MN| Minnesota| 27053| Hennepin|44.9635| -93.2678| 2977172| 3071| America/Chicago|55403 55409 55408...|

| Tampa| Tampa| FL| Florida| 12057| Hillsborough|27.9942| -82.4451| 2908063| 1353| America/New\_York|33637 33629 33621...|

| Denver| Denver| CO| Colorado| 8031| Denver|39.7621|-104.8759| 2876625| 1831| America/Denver|80264 80230 80231...|

| Brooklyn| Brooklyn| NY| New York| 36047| Kings|40.6501| -73.9496| 2559903| 14221| America/New\_York|11212 11213 11216...|

| Queens| Queens| NY| New York| 36081| Queens|40.7498| -73.7976| 2230722| 7885| America/New\_York|11361 11362 11363...|

+-------------+-------------+--------+--------------------+-----------+--------------------+-------+---------+----------+-------+-------------------+--------------------+

only showing top 20 rows

cleansing the data

+-------------+--------+--------------------+--------------------+----------+--------------------+

| city|state\_id| state\_name| county\_name|population| zips|

+-------------+--------+--------------------+--------------------+----------+--------------------+

| New York| NY| New York| New York| 18713220|11229 11226 11225...|

| Los Angeles| CA| California| Los Angeles| 12750807|90291 90293 90292...|

| Chicago| IL| Illinois| Cook| 8604203|60018 60649 60641...|

| Miami| FL| Florida| Miami-Dade| 6445545|33129 33125 33126...|

| Dallas| TX| Texas| Dallas| 5743938|75287 75098 75233...|

| Philadelphia| PA| Pennsylvania| Philadelphia| 5649300|19154 19151 19150...|

| Houston| TX| Texas| Harris| 5464251|77069 77068 77061...|

| Atlanta| GA| Georgia| Fulton| 5449398|30334 30331 30332...|

| Washington| DC|District of Columbia|District of Columbia| 5379184|20010 20011 20012...|

| Boston| MA| Massachusetts| Suffolk| 4688346|02120 02121 02122...|

| Phoenix| AZ| Arizona| Maricopa| 4219697|85008 85009 85006...|

| Seattle| WA| Washington| King| 3789215|98109 98108 98104...|

|San Francisco| CA| California| San Francisco| 3592294|94130 94131 94132...|

| Detroit| MI| Michigan| Wayne| 3506126|48209 48208 48201...|

| San Diego| CA| California| San Diego| 3220118|92109 92108 92103...|

| Minneapolis| MN| Minnesota| Hennepin| 2977172|55403 55409 55408...|

| Tampa| FL| Florida| Hillsborough| 2908063|33637 33629 33621...|

| Denver| CO| Colorado| Denver| 2876625|80264 80230 80231...|

| Brooklyn| NY| New York| Kings| 2559903|11212 11213 11216...|

| Queens| NY| New York| Queens| 2230722|11361 11362 11363...|

+-------------+--------+--------------------+--------------------+----------+--------------------+

only showing top 20 rows

+-------------+--------+--------------------+--------------------+

| CITY|STATE\_ID| STATE\_NAME| COUNTY\_NAME|

+-------------+--------+--------------------+--------------------+

| New York| NY| New York| New York|

| Los Angeles| CA| California| Los Angeles|

| Chicago| IL| Illinois| Cook|

| Miami| FL| Florida| Miami-Dade|

| Dallas| TX| Texas| Dallas|

| Philadelphia| PA| Pennsylvania| Philadelphia|

| Houston| TX| Texas| Harris|

| Atlanta| GA| Georgia| Fulton|

| Washington| DC|District of Columbia|District of Columbia|

| Boston| MA| Massachusetts| Suffolk|

| Phoenix| AZ| Arizona| Maricopa|

| Seattle| WA| Washington| King|

|San Francisco| CA| California| San Francisco|

| Detroit| MI| Michigan| Wayne|

| San Diego| CA| California| San Diego|

| Minneapolis| MN| Minnesota| Hennepin|

| Tampa| FL| Florida| Hillsborough|

| Denver| CO| Colorado| Denver|

| Brooklyn| NY| New York| Kings|

| Queens| NY| New York| Queens|

+-------------+--------+--------------------+--------------------+

only showing top 20 rows

Calculate the Number of zips in each city.

+-------------+-------------+--------+--------------------+-----------+--------------------+-------+---------+----------+-------+-------------------+--------------------+--------+

| city| city\_ascii|state\_id| state\_name|county\_fips| county\_name| lat| lng|population|density| timezone| zips|num\_zips|

+-------------+-------------+--------+--------------------+-----------+--------------------+-------+---------+----------+-------+-------------------+--------------------+--------+

| New York| New York| NY| New York| 36061| New York|40.6943| -73.9249| 18713220| 10715| America/New\_York|[11229 11226 1122...| 1|

| Los Angeles| Los Angeles| CA| California| 6037| Los Angeles|34.1139|-118.4068| 12750807| 3276|America/Los\_Angeles|[90291 90293 9029...| 1|

| Chicago| Chicago| IL| Illinois| 17031| Cook|41.8373| -87.6862| 8604203| 4574| America/Chicago|[60018 60649 6064...| 1|

| Miami| Miami| FL| Florida| 12086| Miami-Dade|25.7839| -80.2102| 6445545| 5019| America/New\_York|[33129 33125 3312...| 1|

| Dallas| Dallas| TX| Texas| 48113| Dallas|32.7936| -96.7662| 5743938| 1526| America/Chicago|[75287 75098 7523...| 1|

| Philadelphia| Philadelphia| PA| Pennsylvania| 42101| Philadelphia|40.0077| -75.1339| 5649300| 4554| America/New\_York|[19154 19151 1915...| 1|

| Houston| Houston| TX| Texas| 48201| Harris|29.7863| -95.3889| 5464251| 1399| America/Chicago|[77069 77068 7706...| 1|

| Atlanta| Atlanta| GA| Georgia| 13121| Fulton|33.7627| -84.4224| 5449398| 1441| America/New\_York|[30334 30331 3033...| 1|

| Washington| Washington| DC|District of Columbia| 11001|District of Columbia|38.9047| -77.0163| 5379184| 4457| America/New\_York|[20010 20011 2001...| 1|

| Boston| Boston| MA| Massachusetts| 25025| Suffolk|42.3188| -71.0846| 4688346| 5532| America/New\_York|[02120 02121 0212...| 1|

| Phoenix| Phoenix| AZ| Arizona| 4013| Maricopa|33.5722|-112.0891| 4219697| 1253| America/Phoenix|[85008 85009 8500...| 1|

| Seattle| Seattle| WA| Washington| 53033| King|47.6211|-122.3244| 3789215| 3469|America/Los\_Angeles|[98109 98108 9810...| 1|

|San Francisco|San Francisco| CA| California| 6075| San Francisco|37.7562| -122.443| 3592294| 7256|America/Los\_Angeles|[94130 94131 9413...| 1|

| Detroit| Detroit| MI| Michigan| 26163| Wayne|42.3834| -83.1024| 3506126| 1864| America/Detroit|[48209 48208 4820...| 1|

| San Diego| San Diego| CA| California| 6073| San Diego|32.8312|-117.1225| 3220118| 1686|America/Los\_Angeles|[92109 92108 9210...| 1|

| Minneapolis| Minneapolis| MN| Minnesota| 27053| Hennepin|44.9635| -93.2678| 2977172| 3071| America/Chicago|[55403 55409 5540...| 1|

| Tampa| Tampa| FL| Florida| 12057| Hillsborough|27.9942| -82.4451| 2908063| 1353| America/New\_York|[33637 33629 3362...| 1|

| Denver| Denver| CO| Colorado| 8031| Denver|39.7621|-104.8759| 2876625| 1831| America/Denver|[80264 80230 8023...| 1|

| Brooklyn| Brooklyn| NY| New York| 36047| Kings|40.6501| -73.9496| 2559903| 14221| America/New\_York|[11212 11213 1121...| 1|

| Queens| Queens| NY| New York| 36081| Queens|40.7498| -73.7976| 2230722| 7885| America/New\_York|[11361 11362 1136...| 1|

+-------------+-------------+--------+--------------------+-----------+--------------------+-------+---------+----------+-------+-------------------+--------------------+--------+

only showing top 20 rows

Process finished with exit code 0