SparkSQL R Sample - USA Zip Codes (JSON)

In [1]:
```
Sys.getenv("SPARK_HOME")
```

'/Users/skalathur/MyApps/spark'

In [2]:
```
if (nchar(Sys.getenv("SPARK_HOME")) < 1) {
  Sys.setenv(SPARK_HOME = "/Users/skalathur/MyApps/spark")
}
```

In [3]:
```
Sys.setenv(SPARK_LOCAL_IP="localhost")
```

In [4]:
```
library(SparkR, lib.loc = c(file.path(Sys.getenv("SPARK_HOME"), "R", "lib")))
```

```
Attaching package: 'SparkR'

The following objects are masked from 'package:stats':

    cov, filter, lag, na.omit, predict, sd, var, window

The following objects are masked from 'package:base':

    as.data.frame, colnames, colnames<-, drop, endsWith, intersect,
    rank, rbind, sample, startsWith, subset, summary, transform, union
```

In [5]:
```
sparkR.session(master = "local[*]", sparkConfig = list(spark.driver.memory = "2g"
))
```

```
Spark package found in SPARK_HOME: /Users/skalathur/MyApps/spark

Launching java with spark-submit command /Users/skalathur/MyApps/spark/bin/spark
-submit   --driver-memory "2g" sparkr-shell /var/folders/s3/hy6_p79n3w1fw802t6ps
40qr0000gp/T//RtmpuvLqbh/backend_port159e413044506

Java ref type org.apache.spark.sql.SparkSession id 1
```

In [6]:
```
inputFile <- "/temp/datasets/usa_zipcodes.json"
```

In [7]:
```
usaZipCodes <- read.df(inputFile, source = "json",
                           inferSchema='true')

usaZipCodes
```

```
SparkDataFrame[_id:string, city:string, loc:array<double>, pop:bigint, state:str
ing]
```

In [8]:
```
printSchema(usaZipCodes)
```

```
root
 |-- _id: string (nullable = true)
 |-- city: string (nullable = true)
 |-- loc: array (nullable = true)
 |    |-- element: double (containsNull = true)
 |-- pop: long (nullable = true)
 |-- state: string (nullable = true)
```

In [9]:
```
count(usaZipCodes)
```

29467

In [10]:
```
head(usaZipCodes)
```

_id	city	loc	pop	state
01001	AGAWAM	-72.62274, 42.07021	15338	MA
01002	CUSHMAN	-72.51565, 42.37702	36963	MA
01005	BARRE	-72.10835, 42.40970	4546	MA
01007	BELCHERTOWN	-72.41095, 42.27510	10579	MA
01008	BLANDFORD	-72.93611, 42.18295	1240	MA
01010	BRIMFIELD	-72.18846, 42.11654	3706	MA

In [11]:
```
persist(usaZipCodes, "MEMORY_AND_DISK")
```

```
SparkDataFrame[_id:string, city:string, loc:array<double>, pop:bigint, state:str
ing]
```

In [12]:
```
createOrReplaceTempView(usaZipCodes, "usaZipCodesTable")
```

In [13]:
```
# Keep only the zip codes with population > 100

query <- "SELECT * FROM usaZipCodesTable WHERE pop > 100"
query
```

'SELECT * FROM usaZipCodesTable WHERE pop > 100'

In [14]:
```
usaZipCodes <- sql(query)
usaZipCodes
```

```
SparkDataFrame[_id:string, city:string, loc:array<double>, pop:bigint, state:str
ing]
```

In [15]:
```
createOrReplaceTempView(usaZipCodes, "usaZipCodesTable")
```

In [16]:
```
query <- "SELECT max(pop) as MaxPop, min(pop) as MinPop from usaZipCodesTable"
query
```

'SELECT max(pop) as MaxPop, min(pop) as MinPop from usaZipCodesTable'

In [17]:
```
maxAndMin <- sql(query)
maxAndMin
```

```
SparkDataFrame[MaxPop:bigint, MinPop:bigint]
```

In [18]:
```
localDf <- collect(maxAndMin)
localDf
```

MaxPop	MinPop
112047	101

Number of zip codes in each state

In [19]:
```
query <- "SELECT state, count(*) as Count FROM usaZipCodesTable GROUP BY state"
query
```

'SELECT state, count(*) as Count FROM usaZipCodesTable GROUP BY state'

In [20]:
```
zipCodesByState <- sql(query)
zipCodesByState
```

```
SparkDataFrame[state:string, Count:bigint]
```

In [21]:
```
count(zipCodesByState)
```

51

In [22]:
```
collect(zipCodesByState)
```

state	Count
SC	347
AZ	260
LA	457
MN	877
NJ	535
DC	22
OR	363
VA	802
RI	69
KY	791
WY	123
NH	214
MI	869
NV	96
WI	706
ID	225
CA	1475
CT	260
NE	572
MT	290
NC	698
VT	238
MD	415
DE	53
MO	989
IL	1232
ME	395
ND	376
WA	474
MS	359
AL	564
IN	675
OH	1007
TN	575
IA	914
NM	231

In [26]:
```
query <- "SELECT state, count(*) as Count FROM usaZipCodesTable
            GROUP BY state ORDER BY state"
query
```

'SELECT state, count(*) as Count FROM usaZipCodesTable
GROUP BY state ORDER BY state'

In [27]:
```
collect(sql(query))
```

state	Count
AK	169
AL	564
AR	569
AZ	260
CA	1475
CO	397
CT	260
DC	22
DE	53
FL	820
GA	631
HI	78
IA	914
ID	225
IL	1232
IN	675
KS	707
KY	791
LA	457
MA	470
MD	415
ME	395
MI	869
MN	877
MO	989
MS	359
MT	290
NC	698
ND	376
NE	572
NH	214
NJ	535
NM	231
NV	96
NY	1546
OH	1007

10 Most populous zip codes

In [25]:
```
collect(sql("SELECT * FROM usaZipCodesTable ORDER BY pop DESC LIMIT 10"))
```

_id	city	loc	pop	state
60623	CHICAGO	-87.71570, 41.84902	112047	IL
11226	BROOKLYN	-73.95699, 40.64669	111396	NY
10021	NEW YORK	-73.95880, 40.76848	106564	NY
10025	NEW YORK	-73.96831, 40.79747	100027	NY
90201	BELL GARDENS	-118.17205, 33.96918	99568	CA
60617	CHICAGO	-87.55601, 41.72574	98612	IL
90011	LOS ANGELES	-118.25819, 34.00786	96074	CA
60647	CHICAGO	-87.70432, 41.92090	95971	IL
60628	CHICAGO	-87.62428, 41.69344	94317	IL
90650	NORWALK	-118.08177, 33.90564	94188	CA

Most populous states

In [28]:
```
query <- "SELECT state, sum(pop) as TotalPop FROM usaZipCodesTable
                GROUP BY state ORDER BY TotalPop DESC"
query
```

'SELECT state, sum(pop) as TotalPop FROM usaZipCodesTable
GROUP BY state ORDER BY TotalPop DESC'

In [29]:
```
popByState <-  sql(query)
popByState
```

```
SparkDataFrame[state:string, TotalPop:bigint]
```

In [30]:
```
count(popByState)
```

51

In [31]:
```
collect(popByState)
```

state	TotalPop
CA	29758155
NY	17988283
TX	16984340
FL	12937753
PA	11880512
IL	11430349
OH	10847077
MI	9295060
NJ	7729991
NC	6628251
GA	6478100
VA	6186121
MA	6016214
IN	5544061
MO	5113794
WI	4891317
TN	4876062
WA	4866199
MD	4781093
MN	4374503
LA	4219523
AL	4040533
KY	3683669
AZ	3664722
SC	3486578
CO	3293351
CT	3286943
OK	3144943
OR	2841361
IA	2776234
MS	2572971
KS	2476552
AR	2350382
WV	1791298
UT	1722387
NE	1578207

In [32]:
```
# Stop the SparkSession now
sparkR.session.stop()
```