SparkR 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//RtmpU1hhYU/backend_port14428531fffa5

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]:
```
collect(subset(usaZipCodes, usaZipCodes$pop <= 100))
```

_id	city	loc	pop	state
01338	BUCKLAND	-72.76412, 42.61517	16	MA
01350	MONROE	-72.96016, 42.72389	97	MA
02163	CAMBRIDGE	-71.14188, 42.36400	0	MA
02713	CUTTYHUNK	-70.87854, 41.44360	98	MA
02815	CLAYVILLE	-71.67059, 41.77776	45	RI
03232	EAST HEBRON	-71.76791, 43.69697	47	NH
03291	WEST NOTTINGHAM	-71.11101, 43.13397	27	NH
03817	CHOCORUA	-71.24072, 43.89085	70	NH
03838	GLEN	-71.19246, 44.10178	84	NH
04013	BUSTINS ISLAND	-70.04225, 43.79602	0	ME
04019	CLIFF ISLAND	-70.10710, 43.69555	87	ME
04109	CUSHING ISLAND	-70.20220, 43.67497	28	ME
04235	FRYE	-70.56532, 44.59948	28	ME
04278	RUMFORD CENTER	-70.70006, 44.59233	92	ME
04279	RUMFORD POINT	-70.70028, 44.55710	36	ME
04442	GREENVILLE JUNCT	-69.63753, 45.48839	99	ME
04563	CUSHING	-69.27206, 43.98674	12	ME
04567	SMALL POINT	-69.84116, 43.73172	66	ME
04570	SQUIRREL ISLAND	-69.63097, 43.80903	3	ME
04645	ISLE AU HAUT	-68.62060, 44.05606	46	ME
04673	SARGENTVILLE	-68.70522, 44.33450	43	ME
04678	SOUTH GOULDSBORO	-68.03041, 44.47163	58	ME
04737	CLAYTON LAKE	-69.62645, 46.62980	54	ME
04764	OXBOW	-68.52179, 46.40196	76	ME
04852	MONHEGAN	-69.31643, 43.76422	88	ME
04985	WEST FORKS	-69.98406, 45.38390	86	ME
05073	TAFTSVILLE	-72.46733, 43.62982	35	VT
05405	UNIV OF VERMONT	-73.20020, 44.47773	0	VT
05447	EAST BERKSHIRE	-72.70656, 44.92980	94	VT
05748	HANCOCK	-72.91329, 43.91253	98	VT
⋮	⋮	⋮	⋮	⋮
99147	LINCOLN	-118.48101, 47.78204	31	WA
99345	PATERSON	-119.75587, 45.99114	94	WA
99402	ASOTIN	-117.00155, 46.13432	89	WA
99563	CHEVAK	-164.77646, 61.58398	0	AK
99569	CLARKS POINT	-158.45124, 58.84921	68	AK

In [12]:
```
# Keep only the zip codes with population > 100

usaZipCodes <- subset(usaZipCodes, usaZipCodes$pop > 100)
usaZipCodes
```

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

In [13]:
```
maxAndMin <- summarize(usaZipCodes, MaxPop = max(usaZipCodes$pop),
                       MinPop = min(usaZipCodes$pop))
maxAndMin
```

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

In [14]:
```
localDf <- collect(maxAndMin)
localDf
```

MaxPop	MinPop
112047	101

Number of zip codes in each state

In [15]:
```
zipCodesByState <-  summarize(groupBy(usaZipCodes, usaZipCodes$state),
                              Count = n(usaZipCodes$state))

zipCodesByState
```

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

In [16]:
```
count(zipCodesByState)
```

```
51
```

In [17]:
```
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 [18]:
```
collect(arrange(zipCodesByState, zipCodesByState$state))
```

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

In [19]:
```
collect(arrange(zipCodesByState, desc(zipCodesByState$Count)))
```

state	Count
TX	1628
NY	1546
CA	1475
PA	1434
IL	1232
OH	1007
MO	989
IA	914
MN	877
MI	869
FL	820
VA	802
KY	791
KS	707
WI	706
NC	698
IN	675
GA	631
WV	618
OK	576
TN	575
NE	572
AR	569
AL	564
NJ	535
WA	474
MA	470
LA	457
MD	415
CO	397
ME	395
ND	376
OR	363
SD	362
MS	359
SC	347

10 Most populous zip codes

In [20]:
```
arrange(usaZipCodes, desc(usaZipCodes$pop))
```

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

In [21]:
```
head(arrange(usaZipCodes, desc(usaZipCodes$pop)), n = 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 [22]:
```
popByState <-  summarize(groupBy(usaZipCodes, usaZipCodes$state),
                                TotalPop = sum(usaZipCodes$pop))

popByState
```

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

In [23]:
```
count(popByState)
```

```
51
```

In [24]:
```
collect(popByState)
```

state	TotalPop
SC	3486578
AZ	3664722
LA	4219523
MN	4374503
NJ	7729991
DC	606868
OR	2841361
VA	6186121
RI	1003419
KY	3683669
WY	452722
NH	1109024
MI	9295060
NV	1201366
WI	4891317
ID	1005753
CA	29758155
CT	3286943
NE	1578207
MT	797589
NC	6628251
VT	562524
MD	4781093
DE	666168
MO	5113794
IL	11430349
ME	1227026
ND	637322
WA	4866199
MS	2572971
AL	4040533
IN	5544061
OH	10847077
TN	4876062
IA	2776234
NM	1513103

In [25]:
```
head(arrange(popByState, desc(popByState$TotalPop)))
```

state	TotalPop
CA	29758155
NY	17988283
TX	16984340
FL	12937753
PA	11880512
IL	11430349

In [26]:
```
# Stop the SparkSession now
sparkR.session.stop()
```