SparkR Sample - USA Daily Temperatures

In [1]:

```
Sys.getenv("SPARK_HOME")
```

'/Users/skalathur/MyApps/spark'

In [2]:

```
# Set the correct value for SPARK_HOME if not set in your environment
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]:

```
# load the SparkR library (wait until it loads)
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]:

```
# Start the Spark Session, wait until it starts
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//Rtmpi1F4En/backend_port143353f517b08
```

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

In [6]:

```
inputFile <- "/temp/datasets/usa_daily_avg_temps.csv"
```

In [7]:

```
# Read the csv file as a SparkDataFrame
usaDailyTemps <- read.df(inputFile, source = "csv",
                         header='true',
                         inferSchema='true')

usaDailyTemps
```

```
SparkDataFrame[state:string, city:string, month:int, day:int, year:int, avgtemp:
double]
```

In [8]:

```
printSchema(usaDailyTemps)
```

```
root
 |-- state: string (nullable = true)
 |-- city: string (nullable = true)
 |-- month: integer (nullable = true)
 |-- day: integer (nullable = true)
 |-- year: integer (nullable = true)
 |-- avgtemp: double (nullable = true)
```

In [9]:

```
count(usaDailyTemps)
```

1174605

In [10]:

```
head(usaDailyTemps)
```

state	city	month	day	year	avgtemp
Alabama	Birmingham	1	1	1995	50.7
Alabama	Birmingham	1	1	1996	56.8
Alabama	Birmingham	1	1	1997	60.9
Alabama	Birmingham	1	1	1998	35.6
Alabama	Birmingham	1	1	1999	41.0
Alabama	Birmingham	1	1	2000	59.0

Aggregate to find the maximum of avgtemp

In [11]:

```
maxAvgTemp <- summarize(usaDailyTemps, max(usaDailyTemps$avgtemp))
maxAvgTemp
```

```
SparkDataFrame[max(avgtemp):double]
```

In [12]:

```
count(maxAvgTemp)
```

1

In [13]:

```
# collect to local data frame
collect(maxAvgTemp)
```

max(avgtemp)
107.5

In [14]:

```
# Provide the appropriate column name (MaxValue)
maxAvgTemp <- summarize(usaDailyTemps, MaxValue = max(usaDailyTemps$avgtemp))
maxAvgTemp
```

```
SparkDataFrame[MaxValue:double]
```

In [15]:
```
localDf <- collect(maxAvgTemp)
localDf
```

MaxValue
107.5

In [16]:
```
# Filter the SparkDataFrame to find the rows with the max value
maxData <- filter(usaDailyTemps, usaDailyTemps$avgtemp == localDf[1, 'MaxValue'])
maxData
```

```
SparkDataFrame[state:string, city:string, month:int, day:int, year:int, avgtemp:
double]
```

In [17]:
```
# collect to local data frame
collect(maxData)
```

state	city	month	day	year	avgtemp
Arizona	Yuma	7	22	2006	107.5

Aggregate to find the maximum of avgtemp grouping by Year

In [18]:
```
maxTempByYear <- summarize(groupBy(usaDailyTemps, usaDailyTemps$Year),
                           MaxValue = max(usaDailyTemps$avgtemp))
maxTempByYear
```

```
SparkDataFrame[Year:int, MaxValue:double]
```

In [19]:
```
count(maxTempByYear)
```

21

In [20]:
```
collect(maxTempByYear)
```

Year	MaxValue
2003	105.8
2007	104.4
2015	105.1
2006	107.5
2013	104.9
1997	100.6
2014	103.8
2004	101.0
1996	104.3
1998	103.0
2012	103.4
2009	103.3
1995	104.3
2001	104.4
2005	105.5
2000	101.6
2010	103.4
2011	103.1
2008	102.9
1999	100.1
2002	102.6

In [21]:
```
arrange(maxTempByYear, maxTempByYear$Year)
```

```
SparkDataFrame[Year:int, MaxValue:double]
```

In [22]:

```
collect(arrange(maxTempByYear, maxTempByYear$Year))
```

Year	MaxValue
1995	104.3
1996	104.3
1997	100.6
1998	103.0
1999	100.1
2000	101.6
2001	104.4
2002	102.6
2003	105.8
2004	101.0
2005	105.5
2006	107.5
2007	104.4
2008	102.9
2009	103.3
2010	103.4
2011	103.1
2012	103.4
2013	104.9
2014	103.8
2015	105.1

Aggregate to find the maximum of avgtemp grouping by State

In [23]:

```
maxTempByState <- summarize(groupBy(usaDailyTemps, usaDailyTemps$State),
                            MaxValue = max(usaDailyTemps$avgtemp))
maxTempByState
```

```
SparkDataFrame[State:string, MaxValue:double]
```

In [24]:

```
count(maxTempByState)
```

```
50
```

```
In [25]: collect(maxTempByState)
```

State	MaxValue
Utah	92.2
Hawaii	87.2
Minnesota	92.0
Ohio	91.2
Arkansas	100.7
Oregon	97.3
Texas	98.5
North Dakota	91.7
Pennsylvania	92.9
Connecticut	89.8
Nebraska	93.2
Vermont	87.4
Nevada	105.5
Washington	97.7
Illinois	92.3
Oklahoma	100.4
Delaware	89.7
Alaska	79.5
New Mexico	89.4
West Virginia	92.5
Missouri	96.3
Rhode Island	89.2
Georgia	97.7
Montana	100.1
Michigan	89.4
Virginia	93.5
North Carolina	91.0
Wyoming	87.1
Kansas	96.1
New Jersey	95.6
Maryland	92.8
Alabama	91.5
Arizona	107.5
Iowa	93.0
Massachusetts	90.7
Kentucky	93.2

In [26]:

```
arrange(maxTempByState, maxTempByState$State)
```

```
SparkDataFrame[State:string, MaxValue:double]
```

In [27]:

```
collect(arrange(maxTempByState, maxTempByState$State))
```

State	MaxValue
Alabama	91.5
Alaska	79.5
Arizona	107.5
Arkansas	100.7
California	102.6
Colorado	94.7
Connecticut	89.8
Delaware	89.7
Florida	92.8
Georgia	97.7
Hawaii	87.2
Idaho	94.2
Illinois	92.3
Indiana	94.0
Iowa	93.0
Kansas	96.1
Kentucky	93.2
Louisiana	95.4
Maine	89.1
Maryland	92.8
Massachusetts	90.7
Michigan	89.4
Minnesota	92.0
Mississippi	92.8
Missouri	96.3
Montana	100.1
Nebraska	93.2
Nevada	105.5
New Hampshire	88.0
New Jersey	95.6
New Mexico	89.4
New York	93.7
North Carolina	91.0
North Dakota	91.7
Ohio	91.2
Oklahoma	100.4

Aggregate to find the number of entries grouping by State

In [28]:

```
stateCounts <- summarize(groupBy(usaDailyTemps, usaDailyTemps$state),
                         count = n(usaDailyTemps$state))
stateCounts
```

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

```
In [29]: collect(arrange(stateCounts, desc(stateCounts$count)))
```

state	count
Texas	106736
Ohio	53368
Florida	51495
Pennsylvania	43871
Michigan	38120
California	38120
New York	38120
Oregon	30496
Illinois	30496
Georgia	30496
North Carolina	30496
Alabama	30496
Tennessee	30496
Indiana	30496
Colorado	30496
Louisiana	28670
Arizona	23202
Nebraska	22872
Washington	22872
Alaska	22872
Missouri	22872
Montana	22872
Virginia	22872
Kansas	22872
Kentucky	22872
Wisconsin	22872
Minnesota	15248
Arkansas	15248
North Dakota	15248
Connecticut	15248
Nevada	15248
Oklahoma	15248
West Virginia	15248
Wyoming	15248
New Jersey	15248
Maryland	15248

In [ ]:

```
### Aggregate to find the number of entries grouping by State and City
```

In [30]:

```
stateCityCounts <- summarize(groupBy(usaDailyTemps, usaDailyTemps$state, usaDaily
Temps$city),
                             count = n(usaDailyTemps$state))
stateCityCounts
```

```
SparkDataFrame[state:string, city:string, count:bigint]
```

```
In [31]: collect(arrange(stateCityCounts, asc(stateCityCounts$state)))
```

state	city	count
Alabama	Huntsville	7624
Alabama	Birmingham	7624
Alabama	Montgomery	7624
Alabama	Mobile	7624
Alaska	Fairbanks	7624
Alaska	Anchorage	7624
Alaska	Juneau	7624
Arizona	Tucson	7624
Arizona	Phoenix	7624
Arizona	Yuma	4380
Arizona	Flagstaff	3574
Arkansas	Fort Smith	7624
Arkansas	Little Rock	7624
California	San Francisco	7624
California	Fresno	7624
California	San Diego	7624
California	Los Angeles	7624
California	Sacramento	7624
Colorado	Pueblo	7624
Colorado	Colorado Springs	7624
Colorado	Denver	7624
Colorado	Grand Junction	7624
Connecticut	Hartford Springfield	7624
Connecticut	Bridgeport	7624
Delaware	Wilmington	5751
Florida	Tallahassee	7624
Florida	Orlando	7624
Florida	Daytona Beach	5751
Florida	Jacksonville	7624
Florida	Miami Beach	7624
⋮	⋮	⋮
Tennessee	Knoxville	7624
Texas	San Antonio	7624
Texas	Wichita Falls	7624
Texas	Abilene	7624
Texas	Dallas Ft Worth	7624

In [32]:
```
collect(arrange(stateCityCounts, asc(stateCityCounts$state),
                asc(stateCityCounts$city)))
```

state	city	count
Alabama	Birmingham	7624
Alabama	Huntsville	7624
Alabama	Mobile	7624
Alabama	Montgomery	7624
Alaska	Anchorage	7624
Alaska	Fairbanks	7624
Alaska	Juneau	7624
Arizona	Flagstaff	3574
Arizona	Phoenix	7624
Arizona	Tucson	7624
Arizona	Yuma	4380
Arkansas	Fort Smith	7624
Arkansas	Little Rock	7624
California	Fresno	7624
California	Los Angeles	7624
California	Sacramento	7624
California	San Diego	7624
California	San Francisco	7624
Colorado	Colorado Springs	7624
Colorado	Denver	7624
Colorado	Grand Junction	7624
Colorado	Pueblo	7624
Connecticut	Bridgeport	7624
Connecticut	Hartford Springfield	7624
Delaware	Wilmington	5751
Florida	Daytona Beach	5751
Florida	Jacksonville	7624
Florida	Miami Beach	7624
Florida	Orlando	7624
Florida	Tallahassee	7624
⋮	⋮	⋮
Tennessee	Nashville	7624
Texas	Abilene	7624
Texas	Amarillo	7624
Texas	Austin	7624
Texas	Brownsville	7624

Number of cities for each state in the dataset

```
In [33]: collect(summarize(groupBy(stateCityCounts, stateCityCounts$state),
                           count = n(stateCityCounts$state)))
```

state	count
Utah	1
Hawaii	1
Minnesota	2
Ohio	7
Oregon	4
Arkansas	2
Texas	14
North Dakota	2
Pennsylvania	6
Connecticut	2
Nebraska	3
Vermont	1
Nevada	2
Washington	3
Illinois	4
Oklahoma	2
Delaware	1
Alaska	3
New Mexico	1
West Virginia	2
Missouri	3
Rhode Island	1
Georgia	4
Montana	3
Virginia	3
Michigan	5
North Carolina	4
Wyoming	2
Kansas	3
New Jersey	2
Maryland	2
Alabama	4
Arizona	4
Iowa	2
Massachusetts	1
Kentucky	3

Create a subset SparkDataFrame for Boston

```
In [34]: bostonDailyTemps <- subset(usaDailyTemps, usaDailyTemps$city == 'Boston')
         bostonDailyTemps
```

```
SparkDataFrame[state:string, city:string, month:int, day:int, year:int, avgtemp:
double]
```

```
In [35]: count(bostonDailyTemps)
```

7624

```
In [36]: bostonAvgTempsByYear <- summarize(groupBy(bostonDailyTemps, bostonDailyTemps$Year
         ),
                                        Average = avg(bostonDailyTemps$avgtemp))
         bostonAvgTempsByYear
```

```
SparkDataFrame[Year:int, Average:double]
```

In [37]:

```
collect(
  arrange(bostonAvgTempsByYear, bostonAvgTempsByYear$Year)
  )
```

Year	Average
1995	51.32027
1996	47.71749
1997	50.83863
1998	51.51562
1999	52.33945
2000	50.36148
2001	52.42822
2002	50.41205
2003	49.73014
2004	50.52514
2005	50.97726
2006	53.02055
2007	51.12219
2008	50.95355
2009	50.32247
2010	53.47205
2011	53.22384
2012	53.86749
2013	51.69753
2014	50.95452
2015	52.46959

In [38]:

```
bostonAvgTempsByMonth <- summarize(groupBy(bostonDailyTemps, bostonDailyTemps$Mon
th),
                                 Average = avg(bostonDailyTemps$avgtemp))
bostonAvgTempsByMonth
```

```
SparkDataFrame[Month:int, Average:double]
```

In [39]:

```
collect(
  arrange(bostonAvgTempsByMonth, bostonAvgTempsByMonth$Month)
)
```

Month	Average
1	29.76667
2	31.47032
3	37.57604
4	47.08413
5	57.57803
6	66.10714
7	73.55038
8	71.68909
9	65.05762
10	54.73456
11	44.89366
12	34.99742

In [40]:

```
bostonAvgTempsByYearAndMonth <- summarize(groupBy(bostonDailyTemps, bostonDailyTe
mps$Year, bostonDailyTemps$Month),
                                  Average = avg(bostonDailyTemps$avgtemp))
bostonAvgTempsByYearAndMonth
```

```
SparkDataFrame[Year:int, Month:int, Average:double]
```

```
In [41]: collect(
           arrange(bostonAvgTempsByYearAndMonth, bostonAvgTempsByYearAndMonth$Year, boston
         AvgTempsByYearAndMonth$Month)
         )
```

Year	Month	Average
1995	1	34.51935
1995	2	28.57500
1995	3	38.03871
1995	4	45.42000
1995	5	56.69677
1995	6	68.47667
1995	7	75.57419
1995	8	72.52581
1995	9	62.93667
1995	10	58.07742
1995	11	42.20333
1995	12	31.04194
1996	1	30.04516
1996	2	30.71034
1996	3	26.99355
1996	4	27.20000
1996	5	51.57419
1996	6	66.72667
1996	7	71.46452
1996	8	70.37419
1996	9	63.72667
1996	10	53.51613
1996	11	40.15000
1996	12	39.25484
1997	1	29.02258
1997	2	36.32500
1997	3	36.59032
1997	4	45.82333
1997	5	55.48710
1997	6	67.87333
⋮	⋮	⋮
2013	6	68.80000
2013	7	76.05161
2013	8	71.67419
2013	9	64.59667
2013	10	56.28065

In [42]:
```
bostonYears <- select(bostonDailyTemps, 'year')
bostonYears
```

```
SparkDataFrame[year:int]
```

In [43]:
```
distinctBostonYears <- distinct(bostonYears)
distinctBostonYears
```

```
SparkDataFrame[year:int]
```

In [44]:
```
yearsDF <- collect(distinct(bostonYears))
yearsDF
```

year
2003
2007
2015
2006
2013
1997
2014
2004
1996
1998
2012
2009
1995
2001
2005
2000
2010
2011
2008
1999
2002

In [45]:
```
yearsDF[order(yearsDF$year), ]
```

```
1995   1996   1997   1998   1999   2000   2001   2002   2003   2004   2005   2006   2007
2008   2009   2010   2011   2012   2013   2014   2015
```

In [46]:
```
# Stop the SparkSession now
sparkR.session.stop()
```