L2-jiawu449-zijfe244

Jiawei Wu & Zijie Feng

2019/5/26

1)

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row
from pyspark.sql import functions as F

# sc.stop()
sc = SparkContext()
sqlContext = SQLContext(sc)

rdd = sc.textFile("/user/x_zijfe/data1/temperature-readings.csv")
parts = rdd.map(lambda l: l.split(";"))
df = parts.map(lambda p: Row(year=p[1].split("-")[0], value=float(p[3]),station=int(p[0]) ))

df = sqlContext.createDataFrame(df)
df.registerTempTable("tempReadings")

df1 = df.filter(df.year.between(1950,2014))
df2 = df.groupBy("year",).agg({"value":"max"}).withColumnRenamed("max(value)","value")
df2 = df2.join(df1, ["year","value"],"inner").select("year", "station", "value")
df2 = df2.dropDuplicates(["year"])
df2 = df2.sort(df2.value.desc())

df3 = df.groupBy("year",).agg({"value":"min"}).withColumnRenamed("min(value)","value")
df3 = df3.join(df1, ["year","value"],"inner").select("year", "station", "value")
df3 = df3.dropDuplicates(["year"])
df3 = df3.sort(df3.value.desc())

max_temperatureSorted = df2.rdd
min_temperatureSorted = df3.rdd

max_temperatureSorted.saveAsTextFile("./results2/2q1/max_temperatureSorted")
min_temperatureSorted.saveAsTextFile("./results2/2q1/min_temperatureSorted")
```

```
# Head part max_temperatureSort
Row(year=u'1975', station=86200, value=36.1)
Row(year=u'1992', station=63600, value=35.4)
Row(year=u'1994', station=117160, value=34.7)
Row(year=u'2010', station=75250, value=34.4)
Row(year=u'2014', station=96560, value=34.4)
Row(year=u'1989', station=63050, value=33.9)
Row(year=u'1982', station=94050, value=33.8)
Row(year=u'1968', station=137100, value=33.7)
Row(year=u'1966', station=151640, value=33.5)
Row(year=u'1983', station=98210, value=33.3)
Row(year=u'2002', station=78290, value=33.3)
Row(year=u'1986', station=76470, value=33.2)
```

```
Row(year=u'1970', station=103080, value=33.2)
Row(year=u'2000', station=62400, value=33.0)
Row(year=u'1956', station=145340, value=33.0)
Row(year=u'1959', station=65160, value=32.8)
Row(year=u'1991', station=137040, value=32.7)
Row(year=u'2006', station=75240, value=32.7)
Row(year=u'1988', station=102540, value=32.6)
Row(year=u'2011', station=172770, value=32.5)
Row(year=u'1999', station=98210, value=32.4)
Row(year=u'2003', station=136420, value=32.2)
Row(year=u'2007', station=86420, value=32.2)
Row(year=u'2008', station=102390, value=32.2)
Row(year=u'1953', station=65160, value=32.2)
Row(year=u'1955', station=97260, value=32.2)
Row(year=u'1973', station=71470, value=32.2)
Row(year=u'2005', station=107140, value=32.1)
Row(year=u'1969', station=71470, value=32.0)
Row(year=u'1979', station=63600, value=32.0)
...

# End part min_temperatureSort
...
Row(year=u'1960', station=155910, value=-40.0)
Row(year=u'1997', station=179960, value=-40.2)
Row(year=u'1994', station=179960, value=-40.5)
Row(year=u'2006', station=169860, value=-40.6)
Row(year=u'2007', station=169860, value=-40.7)
Row(year=u'2013', station=179960, value=-40.7)
Row(year=u'1963', station=181900, value=-41.0)
Row(year=u'1955', station=160790, value=-41.2)
Row(year=u'2003', station=179960, value=-41.5)
Row(year=u'1969', station=181900, value=-41.5)
Row(year=u'1996', station=155790, value=-41.7)
Row(year=u'2010', station=191910, value=-41.7)
Row(year=u'2011', station=179960, value=-42.0)
Row(year=u'1950', station=155910, value=-42.0)
Row(year=u'1951', station=155910, value=-42.0)
Row(year=u'1962', station=181900, value=-42.0)
Row(year=u'1968', station=179950, value=-42.0)
Row(year=u'1982', station=113410, value=-42.2)
Row(year=u'2002', station=169860, value=-42.2)
Row(year=u'1976', station=192830, value=-42.2)
Row(year=u'2014', station=192840, value=-42.5)
Row(year=u'1977', station=179950, value=-42.5)
Row(year=u'1998', station=169860, value=-42.7)
Row(year=u'2012', station=191910, value=-42.7)
Row(year=u'1958', station=159970, value=-43.0)
Row(year=u'1985', station=166870, value=-43.4)
Row(year=u'1959', station=159970, value=-43.6)
Row(year=u'1981', station=166870, value=-44.0)
Row(year=u'2001', station=112530, value=-44.0)
Row(year=u'1965', station=189780, value=-44.0)
Row(year=u'1979', station=112170, value=-44.0)
```

```
Row(year=u'1986', station=167860, value=-44.2)
Row(year=u'1971', station=166870, value=-44.3)
Row(year=u'1980', station=191900, value=-45.0)
Row(year=u'1956', station=160790, value=-45.0)
Row(year=u'1967', station=166870, value=-45.4)
Row(year=u'1987', station=123480, value=-47.3)
Row(year=u'1978', station=155940, value=-47.7)
Row(year=u'1999', station=192830, value=-49.0)
Row(year=u'1966', station=179950, value=-49.4)
```

2)

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row
from pyspark.sql import functions as F
# sc.stop()
sc = SparkContext()
sqlContext = SQLContext(sc)

rdd = sc.textFile("/user/x_zijfe/data1/temperature-readings.csv")
parts = rdd.map(lambda l: l.split(";"))
df = parts.map(lambda p: Row(year=p[1].split("-")[0], month=p[1].split("-")[1],
                             value=float(p[3]),station=int(p[0]) ))
df = sqlContext.createDataFrame(df)
df.registerTempTable("t")

df1 = sqlContext.sql("""SELECT DISTINCT year, month, station
                        FROM t WHERE year BETWEEN 1950 AND 2014 AND value>=10.0""")
df1.registerTempTable("larger")

c = sqlContext.sql("""SELECT year, month, COUNT(station) as count
                        FROM larger GROUP BY year, month ORDER BY count DESC""")
# c.show()

cou = c.rdd
cou.saveAsTextFile("./results2/2q2")
```

```
# Head part
Row(year=u'1972', month=u'10', count=378)
Row(year=u'1973', month=u'06', count=377)
Row(year=u'1973', month=u'05', count=377)
Row(year=u'1972', month=u'08', count=376)
Row(year=u'1973', month=u'09', count=376)
Row(year=u'1972', month=u'05', count=376)
Row(year=u'1972', month=u'09', count=375)
Row(year=u'1972', month=u'06', count=375)
Row(year=u'1971', month=u'08', count=375)
Row(year=u'1972', month=u'07', count=374)
Row(year=u'1971', month=u'09', count=374)
Row(year=u'1971', month=u'06', count=374)
Row(year=u'1973', month=u'08', count=373)
Row(year=u'1971', month=u'05', count=373)
Row(year=u'1974', month=u'08', count=372)
Row(year=u'1974', month=u'06', count=372)
Row(year=u'1974', month=u'05', count=370)
Row(year=u'1973', month=u'07', count=370)
Row(year=u'1974', month=u'09', count=370)
Row(year=u'1970', month=u'08', count=370)
Row(year=u'1971', month=u'07', count=370)
Row(year=u'1970', month=u'09', count=369)
Row(year=u'1976', month=u'05', count=369)
Row(year=u'1975', month=u'09', count=369)
Row(year=u'1970', month=u'06', count=369)
Row(year=u'1975', month=u'06', count=368)
```

```
Row(year=u'1976', month=u'06', count=368)
Row(year=u'1975', month=u'05', count=367)
Row(year=u'1975', month=u'08', count=367)
Row(year=u'1970', month=u'05', count=366)
Row(year=u'1976', month=u'09', count=365)
Row(year=u'1977', month=u'06', count=364)
Row(year=u'1976', month=u'08', count=363)
Row(year=u'1967', month=u'05', count=363)
...
```

3)

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row
from pyspark.sql import functions as F

# sc.stop()
sc = SparkContext()
sqlContext = SQLContext(sc)

rdd = sc.textFile("/user/x_zijfe/data1/temperature-readings.csv")
parts = rdd.map(lambda l: l.split(";"))
df = parts.map(lambda p: Row(year=p[1].split("-")[0], month=p[1].split("-")[1],\
                             value=float(p[3]),station=int(p[0]) ))
df = sqlContext.createDataFrame(df)
df.registerTempTable("tempReadings")

df1 = df.filter(df.year.between(1960,2014))
df1 = df1.groupBy("year","month","station").agg({"value":"avg"})
df1 = df1.select("year","month","station", F.bround("avg(value)", 1).alias('avg_value'))
df1 = df1.sort(df1.avg_value.desc())

rdd1 = df1.rdd
rdd1.saveAsTextFile("./results2/2q3")
```

```
# End part
Row(year=u'1966', month=u'02', station=192710, avg_value=-23.06666666666668)
Row(year=u'1985', month=u'02', station=167860, avg_value=-23.067142857142848)
Row(year=u'1985', month=u'12', station=156940, avg_value=-23.088709677419374)
Row(year=u'1968', month=u'01', station=169930, avg_value=-23.091397849462357)
Row(year=u'1985', month=u'02', station=162980, avg_value=-23.135714285714283)
Row(year=u'1966', month=u'01', station=181900, avg_value=-23.173118279569906)
Row(year=u'1967', month=u'01', station=181900, avg_value=-23.188172043010763)
Row(year=u'1985', month=u'02', station=147570, avg_value=-23.25267857142857)
Row(year=u'1966', month=u'02', station=166810, avg_value=-23.28214285714285)
Row(year=u'1985', month=u'02', station=191900, avg_value=-23.353124999999984)
Row(year=u'1987', month=u'01', station=181900, avg_value=-23.39032258064518)
Row(year=u'1987', month=u'01', station=169880, avg_value=-23.42379032258065)
Row(year=u'1987', month=u'01', station=170790, avg_value=-23.432608695652167)
Row(year=u'1985', month=u'02', station=172790, avg_value=-23.470714285714287)
Row(year=u'1966', month=u'02', station=179950, avg_value=-23.53303571428571)
Row(year=u'1985', month=u'02', station=160890, avg_value=-23.598809523809525)
Row(year=u'1985', month=u'02', station=167980, avg_value=-23.60535714285712)
Row(year=u'1985', month=u'02', station=182930, avg_value=-23.626785714285717)
Row(year=u'1966', month=u'02', station=191900, avg_value=-23.63214285714286)
Row(year=u'1985', month=u'02', station=166810, avg_value=-23.635714285714283)
Row(year=u'1985', month=u'02', station=179950, avg_value=-23.641836734693854)
Row(year=u'1985', month=u'02', station=183980, avg_value=-23.701886792452836)
Row(year=u'1993', month=u'12', station=166900, avg_value=-23.8)
Row(year=u'1985', month=u'02', station=159970, avg_value=-23.961607142857144)
Row(year=u'1966', month=u'02', station=189780, avg_value=-23.977678571428566)
Row(year=u'1966', month=u'02', station=192830, avg_value=-24.025000000000002)
Row(year=u'1966', month=u'02', station=181900, avg_value=-24.092857142857138)
Row(year=u'1966', month=u'02', station=166870, avg_value=-24.228571428571428)
```

```
Row(year=u'1985', month=u'02', station=183760, avg_value=-24.480803571428574)
Row(year=u'1966', month=u'02', station=159970, avg_value=-24.69241071428571)
Row(year=u'1966', month=u'02', station=167860, avg_value=-24.727678571428577)
Row(year=u'1985', month=u'02', station=166870, avg_value=-24.767346938775503)
Row(year=u'1985', month=u'02', station=169880, avg_value=-25.408035714285717)
Row(year=u'1985', month=u'02', station=192830, avg_value=-26.19081632653062)
Row(year=u'1985', month=u'02', station=181900, avg_value=-26.673809523809528)
```

4)

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row
from pyspark.sql import functions as F

sc = SparkContext()
sqlContext = SQLContext(sc)

rdd = sc.textFile("/user/x_zijfe/data1/temperature-readings.csv")
parts = rdd.map(lambda l: l.split(";"))
df = parts.map(lambda p: Row(date=p[1].split("-"), temp=float(p[3]),station=int(p[0]) ))
df1 = sqlContext.createDataFrame(df)
df1.registerTempTable("tempReadings")

rdd = sc.textFile("/user/x_zijfe/data1/precipitation-readings.csv")
parts = rdd.map(lambda l: l.split(";"))
df = parts.map(lambda p: Row(date=p[1].split("-"), prec=float(p[3]),station=int(p[0]) ))
df2 = sqlContext.createDataFrame(df)
df2.registerTempTable("precReadings")
# max temperature for each station
df1 = df1.groupBy("station").max("temp").withColumnRenamed("max(temp)","temp")
# max daily precipitation for each station
df2 = df2.groupBy("station","date").sum("prec").withColumnRenamed("sum(prec)","prec")
df2 = df2.groupBy("station").max("prec").withColumnRenamed("max(prec)","prec")

cond = [df1.station==df2.station]
df_final = df1.join(df2, cond, "inner").select(df1.station, df1.temp, df2.prec)
df_final = df_final.filter(df_final.temp.between(25,30) & df_final.prec.between(100,200))
rdd_station = df_final.rdd

rdd_station.saveAsTextFile("./results2/2q4")
```

The output is nothing.

5)

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row

sc = SparkContext()
sqlContext = SQLContext(sc)

rdd = sc.textFile("/user/x_zijfe/data1/precipitation-readings.csv")
parts = rdd.map(lambda l: l.split(";"))
df = parts.map(lambda p: Row(year=p[1].split("-")[0], \
        month=p[1].split("-")[1], prec=float(p[3]),station=int(p[0]) ))
df = sqlContext.createDataFrame(df)
df.registerTempTable("precReadings")

rdd = sc.textFile("/user/x_zijfe/data1/stations-Ostergotland.csv")
parts = rdd.map(lambda l: l.split(";"))
df1 = parts.map(lambda p: Row(station=int(p[0])))
df1 = sqlContext.createDataFrame(df1)
df1.registerTempTable("stations")

df = df.filter(df.year.between(1933, 2016))
dfavg = df.groupBy("year","month","station").sum("prec")\
          .withColumnRenamed("sum(prec)","prec").sort(df.year.desc(), df.month.desc())
df1 = df1.join(dfavg, ["station"], "inner")\
          .select(dfavg.prec, dfavg.year, dfavg.month, df1.station)
df1 = df1.groupBy("year","month").avg("prec").withColumnRenamed("avg(temp)","temp")\
          .sort(df1.year.desc(), df1.month.desc())

monthAve = df1.rdd
#print(monthAve.take(10))
monthAve.saveAsTextFile("./results2/2q5")
```

```
# Head part
Row(year=u'2016', month=u'07', avg(prec)=0.0)
Row(year=u'2016', month=u'06', avg(prec)=47.6625)
Row(year=u'2016', month=u'05', avg(prec)=29.250000000000007)
Row(year=u'2016', month=u'04', avg(prec)=26.900000000000006)
Row(year=u'2016', month=u'03', avg(prec)=19.962500000000002)
Row(year=u'2016', month=u'02', avg(prec)=21.5625)
Row(year=u'2016', month=u'01', avg(prec)=22.325000000000003)
Row(year=u'2015', month=u'12', avg(prec)=28.925)
Row(year=u'2015', month=u'11', avg(prec)=63.88750000000002)
Row(year=u'2015', month=u'10', avg(prec)=2.2625)
Row(year=u'2015', month=u'09', avg(prec)=101.29999999999998)
Row(year=u'2015', month=u'08', avg(prec)=26.987500000000004)
Row(year=u'2015', month=u'07', avg(prec)=119.09999999999998)
Row(year=u'2015', month=u'06', avg(prec)=78.66250000000001)
Row(year=u'2015', month=u'05', avg(prec)=93.22499999999998)
Row(year=u'2015', month=u'04', avg(prec)=15.3375)
Row(year=u'2015', month=u'03', avg(prec)=42.61250000000001)
Row(year=u'2015', month=u'02', avg(prec)=24.824999999999996)
Row(year=u'2015', month=u'01', avg(prec)=59.11250000000003)
Row(year=u'2014', month=u'12', avg(prec)=35.46250000000001)
```

```
Row(year=u'2014', month=u'11', avg(prec)=52.425000000000054)
Row(year=u'2014', month=u'10', avg(prec)=72.13749999999999)
Row(year=u'2014', month=u'09', avg(prec)=48.45)
Row(year=u'2014', month=u'08', avg(prec)=90.81249999999997)
Row(year=u'2014', month=u'07', avg(prec)=22.9875)
Row(year=u'2014', month=u'06', avg(prec)=75.1375)
Row(year=u'2014', month=u'05', avg(prec)=58.0)
Row(year=u'2014', month=u'04', avg(prec)=31.762500000000003)
Row(year=u'2014', month=u'03', avg(prec)=36.56250000000001)
Row(year=u'2014', month=u'02', avg(prec)=43.71250000000002)
Row(year=u'2014', month=u'01', avg(prec)=62.575000000000074)
...
```

6)

```
from pyspark import SparkContext
from pyspark.sql import SQLContext, Row

# sc.stop()
sc = SparkContext()
sqlContext = SQLContext(sc)
rdd = sc.textFile("/user/x_zijfe/data1/temperature-readings.csv")
parts = rdd.map(lambda l: l.split(";"))
df = parts.map(lambda p: Row(year=p[1].split("-")[0], \
            month=p[1].split("-")[1], temp=float(p[3]),station=int(p[0]) ))
df = sqlContext.createDataFrame(df)
df.registerTempTable("precReadings")
rdd = sc.textFile("/user/x_zijfe/data1/stations-Ostergotland.csv")
parts = rdd.map(lambda l: l.split(";"))
dfs = parts.map(lambda p: Row(station=int(p[0]) ))
dfs = sqlContext.createDataFrame(dfs)
dfs.registerTempTable("stations")

dfavg = df.groupBy("year","month","station").avg("temp")\
          .withColumnRenamed("avg(temp)","temp")
dfavg = dfavg.join(dfs, "station", "inner")
dfavg = dfavg.groupBy("year","month").avg("temp").withColumnRenamed("avg(temp)","temp")

df1 = dfavg.filter(df.year.between(1950, 2014))
df2 = dfavg.filter(df.year.between(1950, 1980))
df2 = df2.groupBy("month").avg("temp").withColumnRenamed("avg(temp)","temp")

df1 = df1.sort(df1.year.desc(), df1.month.desc())
df2 = df2.sort(df2.month.desc())

ave1 = df1.rdd
ave2 = df2.rdd
final = ave1.collect()
final = [list(x) for x in final]  # change Row to list
ave2 = dict(ave2.collect())

for i in range(len(final)):
    month = final[i][1]
    final[i][2] -= ave2[month]

lll = sc.parallelize(final)   # list to rdd
sqlContext = SQLContext(sc)   # rdd  to DF
df = sqlContext.createDataFrame(lll)
df = df.sort(df._1.desc(), df._2.desc()).withColumnRenamed("_1","year") \
                                        .withColumnRenamed("_2","month")\
                                        .withColumnRenamed("_3","difference")
fff = df.rdd   # DF to rdd
fff.saveAsTextFile("./results2/2q6")
```

```
# Head part (differences from 2012 to 2014)
Row(year=u'2014', month=u'12', difference=0.8090050128011183)
Row(year=u'2014', month=u'11', difference=2.095774582278379)
```

```
Row(year=u'2014', month=u'10', difference=1.5369843525571039)
Row(year=u'2014', month=u'09', difference=0.04528868836290911)
Row(year=u'2014', month=u'08', difference=-0.7553156355884632)
Row(year=u'2014', month=u'07', difference=2.168129032266794)
Row(year=u'2014', month=u'06', difference=-1.8788251330868313)
Row(year=u'2014', month=u'05', difference=0.15765101764698208)
Row(year=u'2014', month=u'04', difference=2.1211036869225035)
Row(year=u'2014', month=u'03', difference=4.222140505337414)
Row(year=u'2014', month=u'02', difference=5.25560805504952)
Row(year=u'2014', month=u'01', difference=0.9193686872017373)
Row(year=u'2013', month=u'12', difference=3.8778419122262773)
Row(year=u'2013', month=u'11', difference=0.9942484465096872)
Row(year=u'2013', month=u'10', difference=0.6741854602767265)
Row(year=u'2013', month=u'09', difference=-1.015200385195305)
Row(year=u'2013', month=u'08', difference=-0.38568783683974317)
Row(year=u'2013', month=u'07', difference=0.14340585614450063)
Row(year=u'2013', month=u'06', difference=-0.6412691757865563)
Row(year=u'2013', month=u'05', difference=1.4171806693274593)
Row(year=u'2013', month=u'04', difference=-0.7561818989841265)
Row(year=u'2013', month=u'03', difference=-3.5385474271177197)
Row(year=u'2013', month=u'02', difference=0.6025816048169408)
Row(year=u'2013', month=u'01', difference=-0.6509672815448817)
Row(year=u'2012', month=u'12', difference=-3.5723414709038575)
Row(year=u'2012', month=u'11', difference=1.3384372851086863)
Row(year=u'2012', month=u'10', difference=-1.4712725765528827)
Row(year=u'2012', month=u'09', difference=-0.47859623178403154)
Row(year=u'2012', month=u'08', difference=-0.8105206956571305)
Row(year=u'2012', month=u'07', difference=-0.7777453374623384)
Row(year=u'2012', month=u'06', difference=-3.1633179928387314)
Row(year=u'2012', month=u'05', difference=0.7109398802281373)
Row(year=u'2012', month=u'04', difference=-0.5505924661981503)
Row(year=u'2012', month=u'03', difference=4.3529041318145945)
Row(year=u'2012', month=u'02', difference=0.06863220126275316)
Row(year=u'2012', month=u'01', difference=1.4824009320507694)
...
```