

基于Spark的气象数据处理与分析

Author: dsy Time: 2019-11-09

1. 导入库

```
In [1]: from pyspark.sql import SparkSession
from pyspark.sql import functions as F
from pyspark.sql.types import DecimalType, TimestampType
import os
import math
```

1. 计算各个城市过去24小时累积雨量

```
In [2]: def passed_rain_analyse(filename): #计算各个城市过去24小时累积雨量

    spark = SparkSession.builder.master("local[*]").appName("passed_rain_analyse").getOrCreate()
    df = spark.read.load(filename,
                          format="csv",
                          header="true")

    df_rain = df.select(df['province'], df['city_name'], df['city_code'], df['rain1h']).coalesce(100) \
        .filter(df['rain1h'] < 1000) \
        .groupBy("province", "city_name", "city_code") \
        .agg(F.sum("rain1h").alias("rain24h")) \
        .sort(F.desc("rain24h")) # 分组、求和、排序

    df_rain.show()
    spark.stop()
```

```
In [3]: passed_rain_analyse("data/weather_noall.csv")
```

| province | city_name | city_code | rain24h |
|----------|-----------|-----------|---------|
| 内蒙古自治区 | 商都 | 53385 | 274.70 |
| 广西壮族自治区 | 天等 | 59227 | 266.20 |
| 广西壮族自治区 | 忻城 | 59038 | 200.90 |
| 福建省 | 建阳 | 58734 | 151.70 |
| 海南省 | 中沙 | 59979 | 118.80 |
| 海南省 | 海口 | 59758 | 118.80 |
| 广西壮族自治区 | 东兴 | 59626 | 110.50 |
| 广东省 | 龙川 | 59107 | 107.90 |
| 广西壮族自治区 | 田东 | 59224 | 107.80 |
| 广西壮族自治区 | 平果 | 59228 | 102.80 |
| 广东省 | 增城 | 59294 | 102.10 |
| 广西壮族自治区 | 陆川 | 59457 | 101.00 |
| 广西壮族自治区 | 防城 | 59631 | 91.90 |
| 广西壮族自治区 | 上林 | 59235 | 90.80 |
| 广东省 | 河源 | 59293 | 84.40 |
| 广西壮族自治区 | 巴马 | 59027 | 80.80 |
| 云南省 | 金平 | 56987 | 78.40 |
| 广东省 | 广州 | 59287 | 76.70 |
| 福建省 | 东山 | 59321 | 75.90 |
| 广西壮族自治区 | 崇左 | 59425 | 74.40 |

only showing top 20 rows

2. 计算各个城市当日平均气温

```
In [4]: def passed_temperature_analyse(filename):
spark = SparkSession.builder.master("local[*]").appName("passed_temperature_analyse")
df = spark.read.load(filename,
                      format="csv",
                      header="true")
df_temperature = df.select( #选择需要的列
    df['province'],
    df['city_name'],
    df['city_code'],
    df['temperature'].cast(DecimalType(scale=2)),
    F.date_format(df['time'], "yyyy-MM-dd").alias("date"), #得到日期数据
    F.hour(df['time']).alias("hour") #得到小时数据
)
# 筛选四点时次
df_4point_temperature = df_temperature.filter(df_temperature['hour'].isin([2, 8, 12, 18]))

df_avg_temperature = df_4point_temperature.groupBy("province", "city_name", "city_code")\
    .agg(F.count("temperature"), F.avg("temperature").alias("avg_temperature"))\
    .filter("count(temperature) = 4")\
    .sort(F.asc("avg_temperature"))\
    .select("province", "city_name", "city_code", "date", F.format_number(' avg_temperature', 2))

# df_avg_temperature.printSchema()
df_avg_temperature.show()
spark.stop()
```

```
In [5]: passed_temperature_analyse("data/weather_noall.csv")
```

| province | city_name | city_code | date | avg_temperature |
|----------|-----------|-----------|------------|-----------------|
| 青海省 | 泽库 | 52968 | 2019-05-28 | 4.82 |
| 四川省 | 峨眉山 | 56385 | 2019-05-28 | 5.48 |
| 西藏自治区 | 普兰 | 55437 | 2019-05-28 | 5.75 |
| 青海省 | 河南 | 56065 | 2019-05-28 | 5.80 |
| 青海省 | 天峻 | 52745 | 2019-05-28 | 6.98 |
| 青海省 | 杂多 | 56018 | 2019-05-28 | 7.05 |
| 青海省 | 玛沁 | 56043 | 2019-05-28 | 7.25 |
| 青海省 | 刚察 | 52754 | 2019-05-28 | 7.42 |
| 青海省 | 玉树 | 56029 | 2019-05-28 | 7.62 |
| 青海省 | 海晏 | 52853 | 2019-05-28 | 7.90 |
| 甘肃省 | 正宁 | 53935 | 2019-05-28 | 8.05 |
| 甘肃省 | 华亭 | 53927 | 2019-05-28 | 8.45 |
| 甘肃省 | 西峰 | 53923 | 2019-05-28 | 8.45 |
| 黑龙江省 | 伊春 | 50774 | 2019-05-28 | 8.58 |
| 宁夏回族自治区 | 西吉 | 53903 | 2019-05-28 | 8.58 |
| 甘肃省 | 临潭 | 56081 | 2019-05-28 | 8.75 |
| 四川省 | 阿坝 | 56171 | 2019-05-28 | 8.92 |
| 宁夏回族自治区 | 固原 | 53817 | 2019-05-28 | 8.98 |
| 甘肃省 | 崆峒 | 53915 | 2019-05-28 | 9.30 |
| 甘肃省 | 东乡 | 52981 | 2019-05-28 | 9.40 |

only showing top 20 rows