**《时空数据处理与组织课程实习》**

**实验 3实习报告**

**学 院: 遥感信息工程学院**

**班 级: 20F10**

**学 号: 2020302131249**

**姓 名: 马文卓**

**实习地点: 教学实验大楼101**

**指导教师: 李晓雷**

**2022 年 5 月 25 日**

1. **实验目的**
2. 通过实验掌握Spark SQL的基本编程方法；
3. 熟悉RDD到DataFrame的转化方法；
4. 熟悉利用Spark SQL管理来自不同数据源的数据。
5. **实验环境**

基于Windows的Spark环境，采用Python编程。

1. **实验内容**

主要包括三个内容：SparkSQL的基本操作、RDD到DataFrame的转换、使用DataFrame读写mySQL数据库。（具体内容见实验文档）

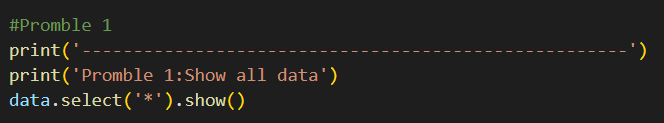
1. **步骤**
2. **SparkSQL的基本操作**

首先导入相关第三方库，并且从json文件中导入数据，创建DataFrame对象data。



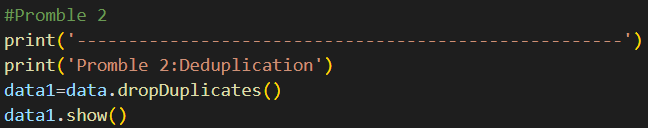
1. **查询所有数据；**

使用select方法进行查询，\*代表所有数据，最后通过show方法展示。



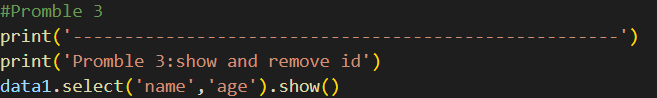
1. **查询所有数据，并去除重复的数据；**

使用dropDuplicates方法进行数据的去重，最后展示



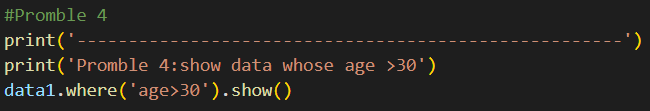
1. **查询所有数据，打印时去除id字段；**

使用select方法查询name、age字段，不查询id字段



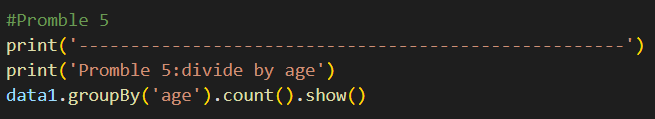
1. **筛选出age>30的记录；**

使用where方法限定查询条件为age>30即可（这里其实也可以用filter方法达到同样的效果）



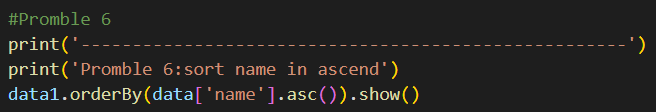
1. **将数据按age分组；**

使用groupBy方法按照age字段进行分组，使用count方法统计各个组别中的数据个数，展示出来



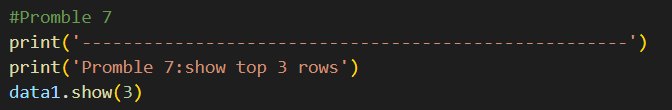
1. **将数据按name升序排列；**

使用oderBy方法对name字段进行排序，后面加上.asc()表示升序（这里最开始出现了排序不正确的现象，后来检查发现是json数据中有一个空格导致的）



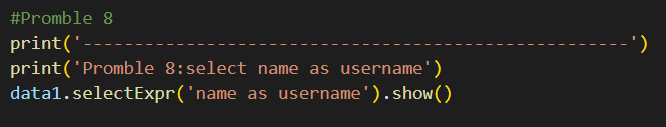
1. **取出前3行数据；**

使用show函数，传入3作为参数，表示显示前三行数据



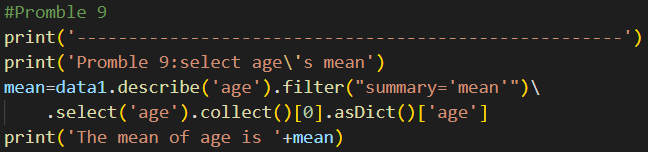
1. **查询所有记录的name列，并为其取别名为username；**

使用selectExpr函数对name字段取别名为username，展示出来



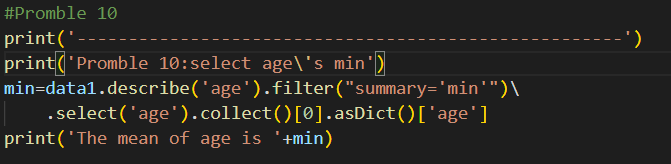
1. **查询年龄age的平均值；**

对去重后的数据data1使用decribe函数取出age字段的数据，然后使用filter函数选择求均值（summary=’mean’），得到一个仅新的DataFrame。再使用select选择其age字段，使用collect函数将其转为数组，取其中的第一个元素（类型为Row），在使用asDict函数将其转化为字典，最后取其age的值，即得到均值，输出即可。



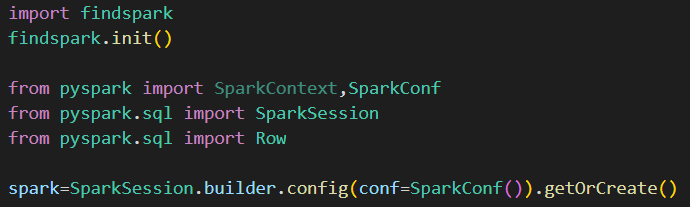
1. **查询年龄age的最小值。**

这里查询age的最小值和第（9）问中的求均值类似，唯一的不同就是将summary=’mean’改为summary=’min’即可。



1. **RDD到DataFrame的转换**

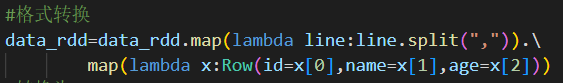
在所有操作之前依旧是导入第三方库，创建sparkSession对象。



然后从txt文件中导入数据，创建RDD对象data\_rdd，这里使用的是spark.sparkContext.textFile函数。



然后进行格式转换，使用map函数进行映射（其中的匿名函数以逗号为分隔符将数据分开），再使用map进行映射（匿名函数表示把每一行数据生成为一个Row对象，其表头为id、name、age）。实际上这里也可以不加表头信息，而是新建一个StructType来存储表头，然后在创建DataFrame的时候再将数据和表头同时传入。这里我们选择直接添加表头。

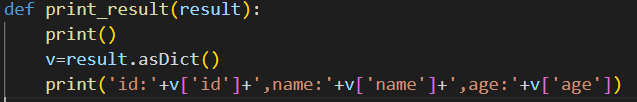


使用spark.createDataFrame函数，传入之前处理的数据，生成DataFrame对象data。



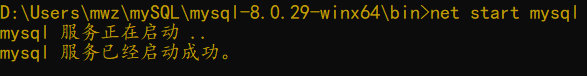
最后按照要求格式输出（自己编写了格式输出函数print\_result），使用foreach函数调用即可。

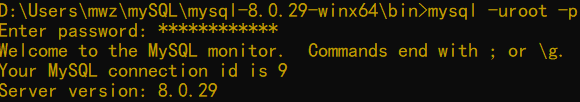




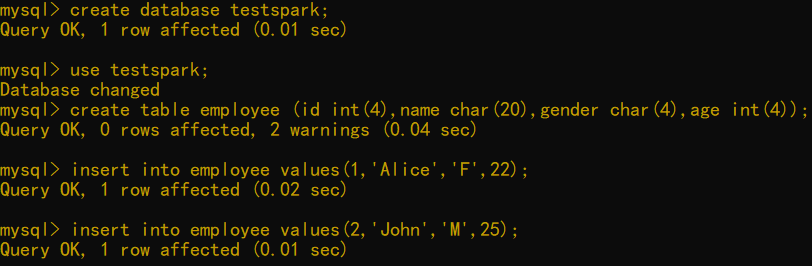
1. **利用DataFrame读写mySQL的数据**
2. **准备工作**

在命令行中输入【net start mysql】启动mysql服务，然后使用【mysql -uroot -p】连接数据库，输入密码即可进入。

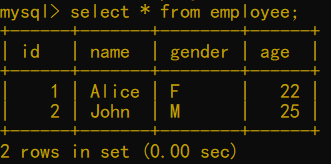




然后创建数据库testspark，创建表employee，使用insert命令向employ表中插入两条数据。

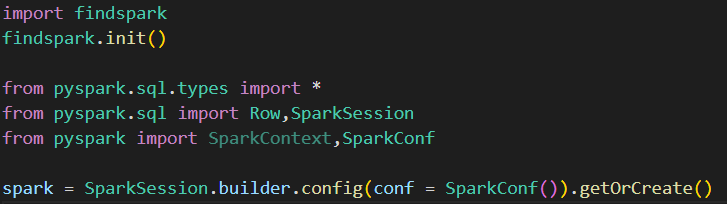


使用select \* from employee语句查询表中所有数据，查看是否添加成功。



1. **插入数据**

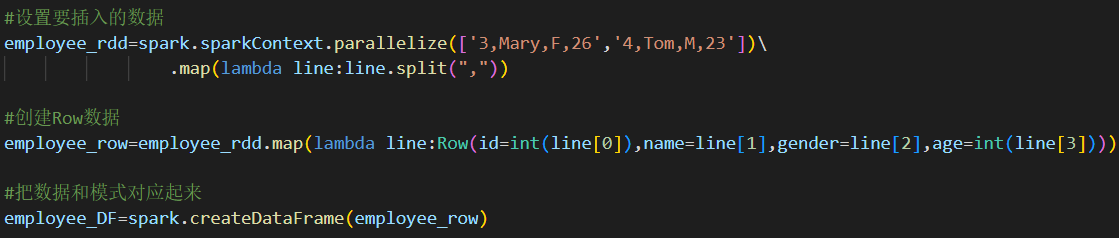
首先导入第三方库，创建SparkSession对象spark。



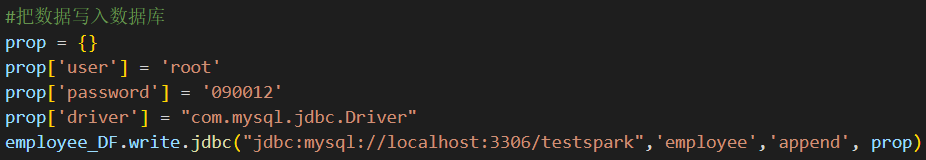
然后使用spark.sparkContext.parallesize函数将要插入的数据转为RDD类型，再使用map映射以逗号为分隔符将数据分开，得到employee\_rdd。

再将RDD对象的每一行转为Row对象，使用map映射对每一行进行操作，将字段名称和字段的值对应起来（注意这里有两个字符串转整型的操作）。

最后使用spark.createDataFrame函数将上面得到的数据传入，生成对应的DataFrame格式的数据employee\_DF。

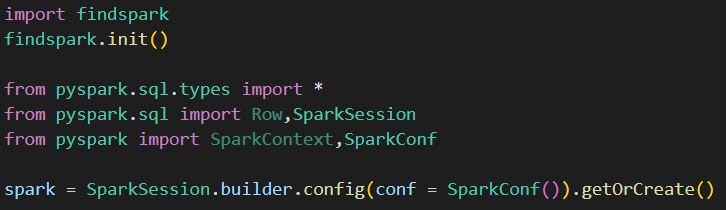


最后把数据写入数据库。首先把用户名、密码、驱动、数据库、表名指定完成。然后使用writer.jdbc函数将数据写入其中。

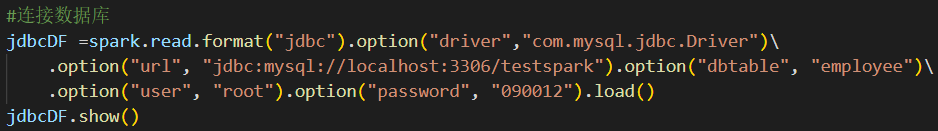


1. **读取数据并计算**

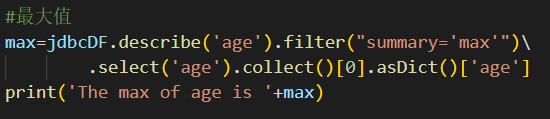
依然首先导入相关第三方库，创建SparkSession。



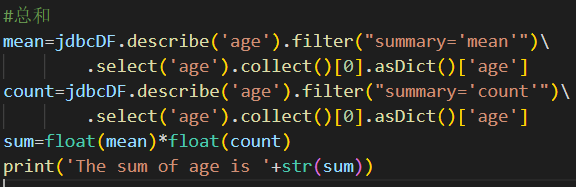
然后连接数据库。这里注意用户名、数据库名、表名指定正确。（连接数据库时出过一点问题，后来查阅相关资料发现，使用conda虚拟环境的pyspark连接mysql时，其驱动不仅要放在spark安装目录的jars文件夹下，还要放在相应环境<比如我这里环境是BG>的第三方库<packages>文件夹下的pyspark文件夹下的jars文件夹中，二者缺一不可）



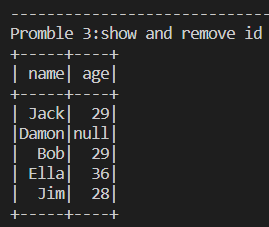
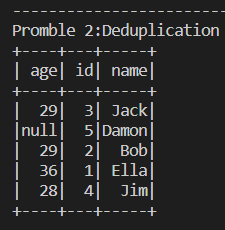
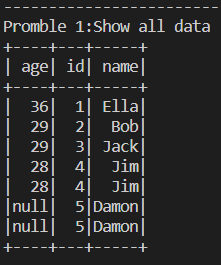
求年龄的最大值，过程和第一题的第（9）问类似，只是将summary=‘mean’换为summary=‘max’即可

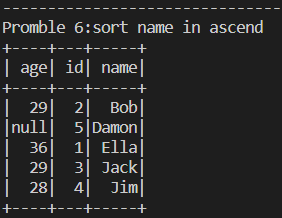
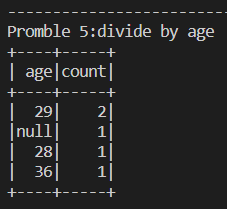
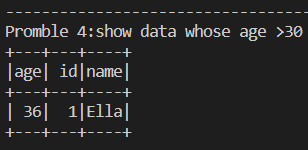


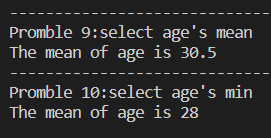
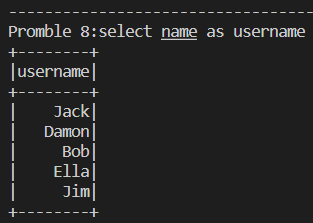
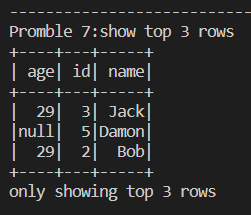
求年龄的和，由于spark.DataFrame的summary中没有提供相应的计算，我通过均值\*数量的方式求和。步骤与上面类似。



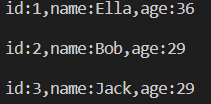
1. **结果**
2. **SparkSQL的基本操作**



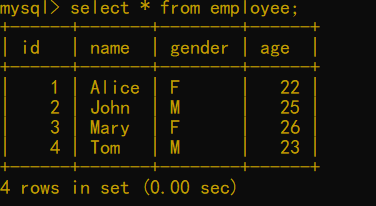




1. **RDD到DataFrame的转换**



**（三）利用DataFrame读写mySQL的数据**





1. **感悟**

本次实验主要是熟悉了sparkSQL的操作以及RDD和DataFrame的转换等，对于大数据的相关操作有了更加深刻的理解。

个人感觉这种循序渐进的教学模式是必需的也是效果良好的，从最开始的基础操作出发，再到后面去解决一些问题，能够更好的帮助我掌握这些技能。

当然在这其中也遇到一些问题。

①安装mySQL的时候配置文件出现了错误，在搜索资料之后，将其中注释删除就可以了

②在安装mysql的驱动时，发现仅仅在spark的安装目录下存放是不够的，还需要在相应的conda环境的pyspark的jars文件夹下存放，才可运行。