南华大学

计算机学院

实验报告

（ 2022 ~2023 学年度 第一学期 ）

|  |  |
| --- | --- |
| 课程名称 | 大数据平台及编程实践 |
| **实验名称** | Spark SQL |

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 刘纯龙 | 学号 | 20184351149 |

|  |  |  |  |
| --- | --- | --- | --- |
| 专业 | 软件工程 | 班级 | 软卓01班 |

|  |  |  |  |
| --- | --- | --- | --- |
| 地点 |  | 教师 | 聂沛 |

1. **实验目的与要求**

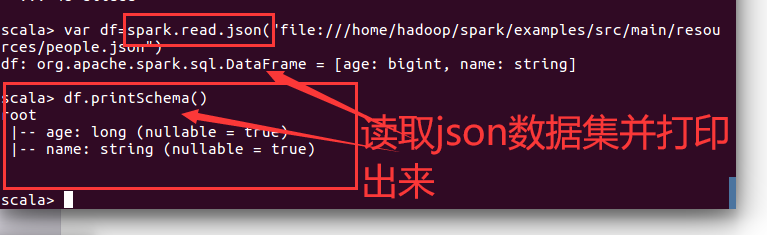
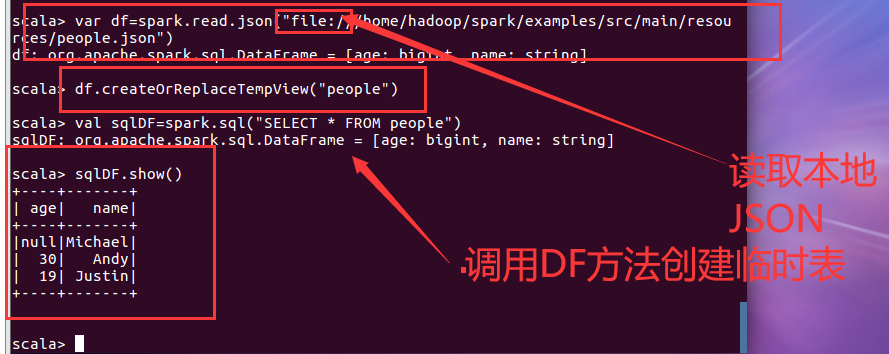
实验目的：掌握RDD与DataFrame的常见操作使用以及DataFrame 的 创建与操作和Spark sql 与hive,mysql的操作与联系

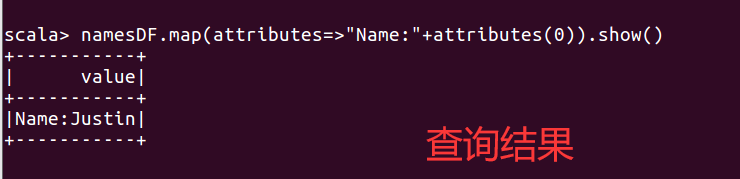
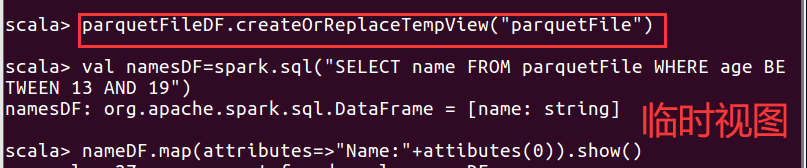
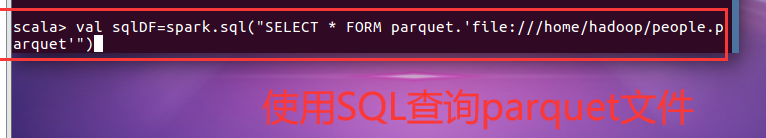
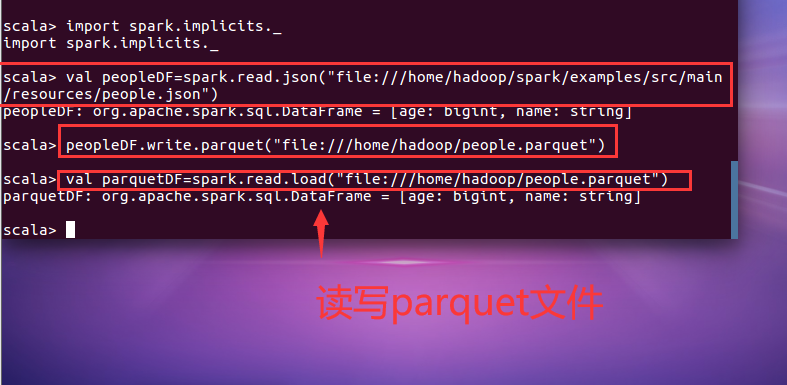
实验要求：配置好环境，对Spark sql指令进行操作

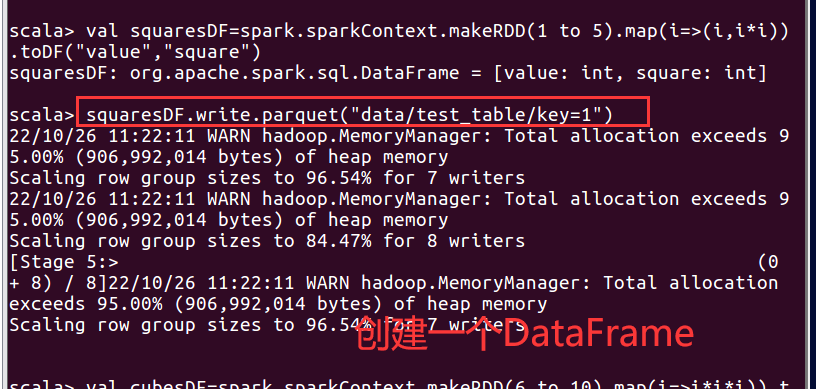
1. **实验内容与步骤**

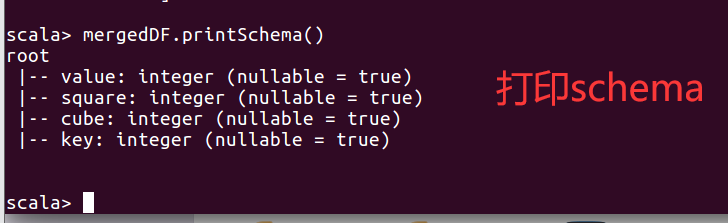
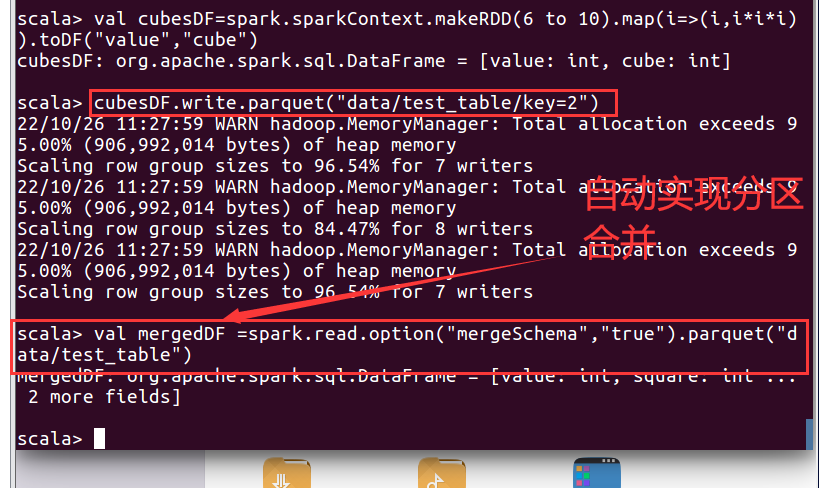
(一)DataFrame的创建

读取本地或hdf上的json，parquet

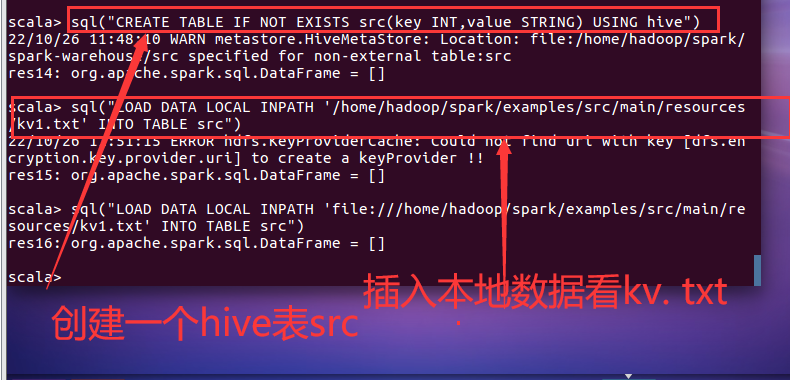
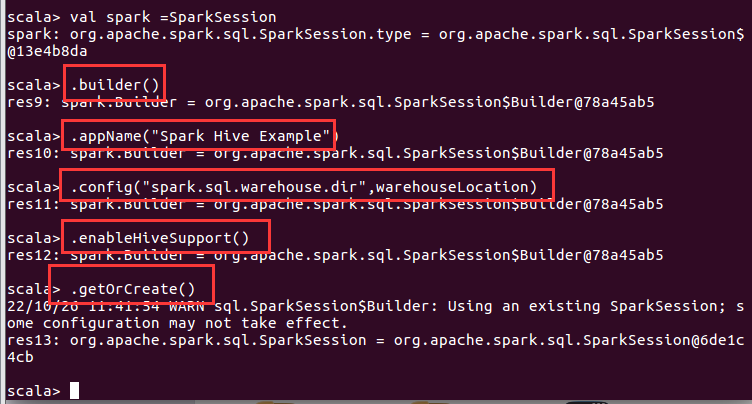




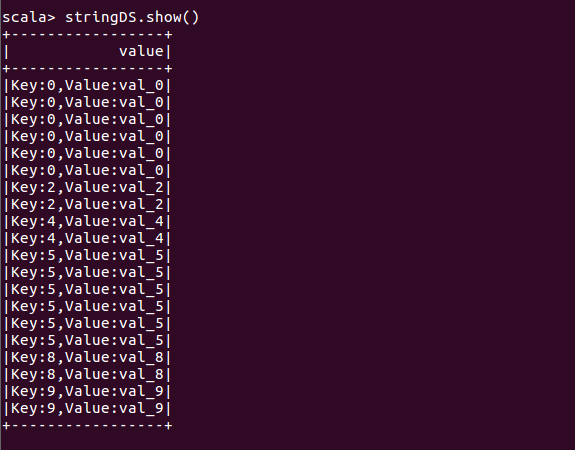
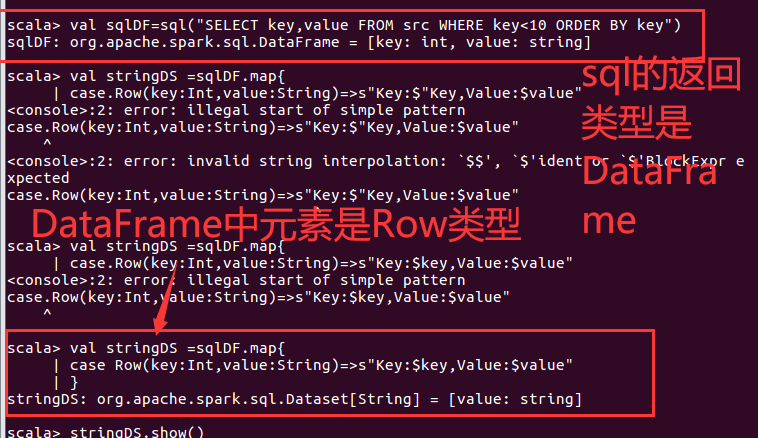




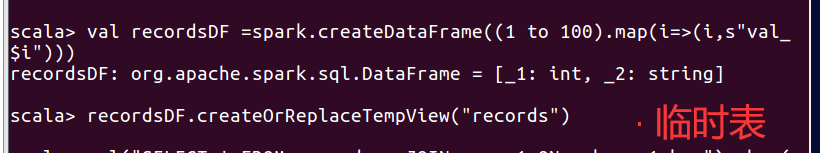
读取和操作Hive的数据



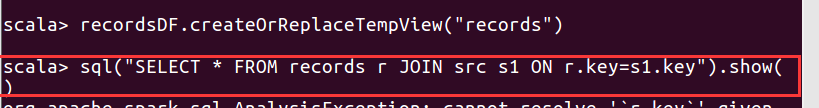


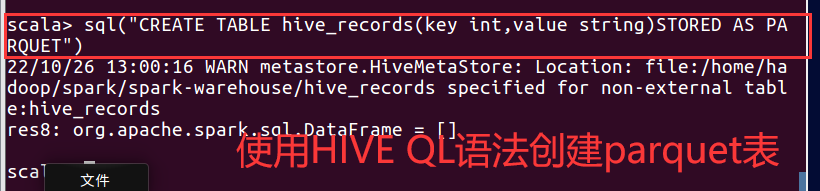


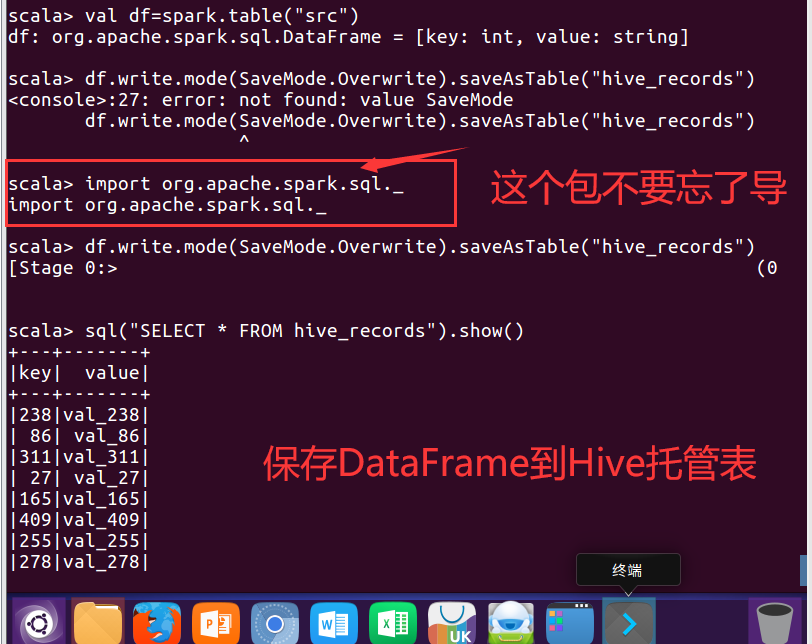
**创建临时表**



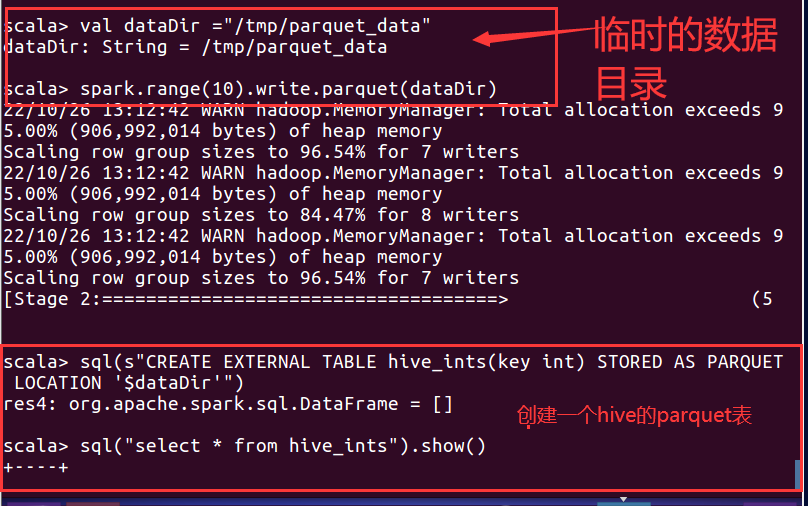
**临时表和HiVE表进行JOIN操作**



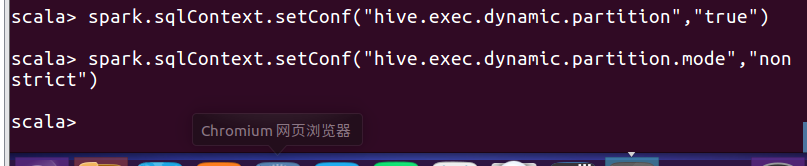


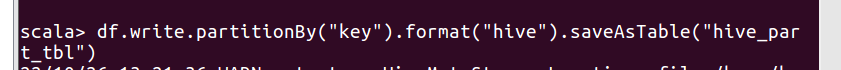


**创建一个Hive额外的parquet表**



**打开HIVE动态分区的标志**

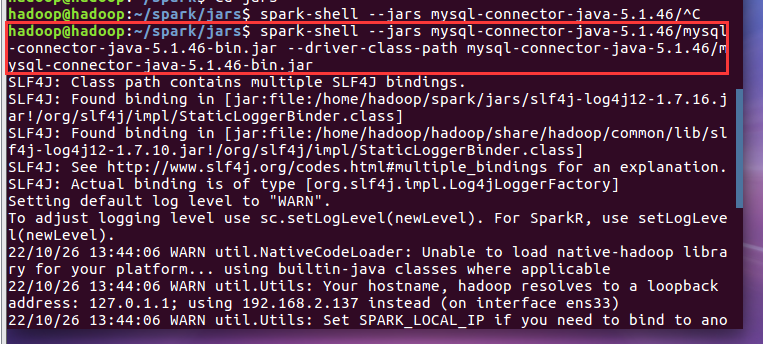




读取和操作MYSQL的数据

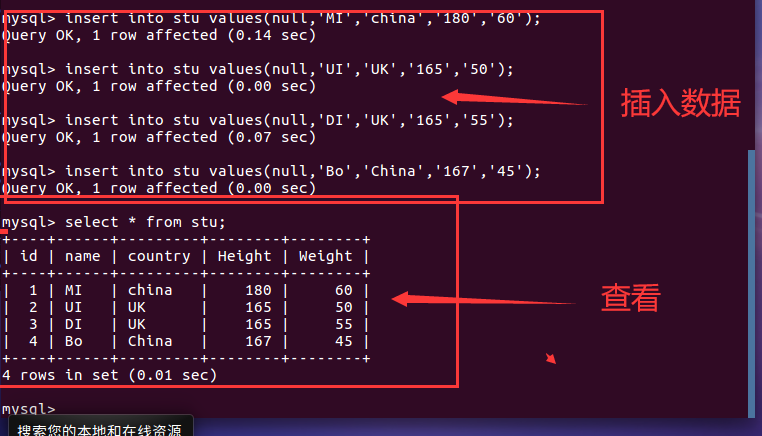
使用JDBC连接MYSQL

**从Spark-shell连接到MYSQL数据库**

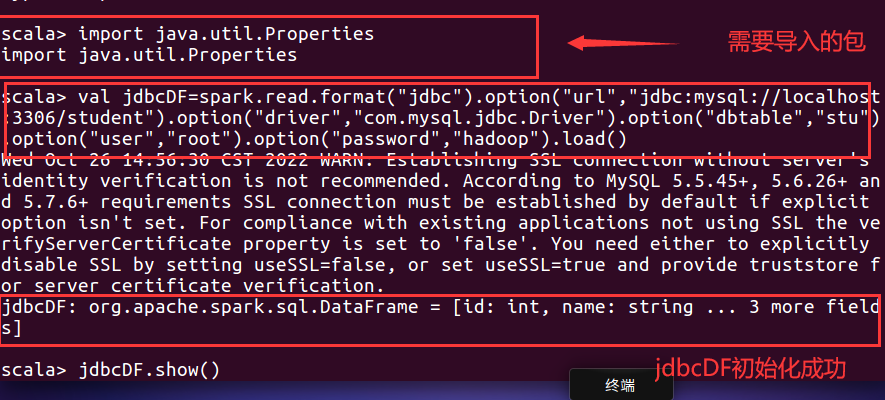


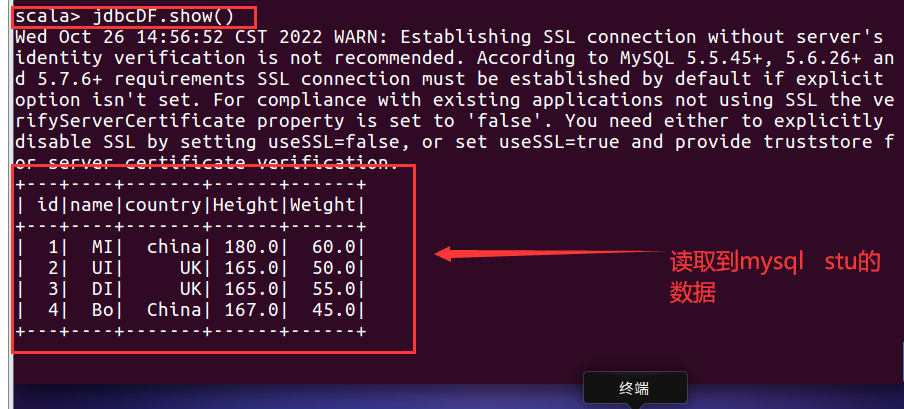






**实现与mysql的连接，读取数据(配置url，用户，密码，jdbc驱动包等等**

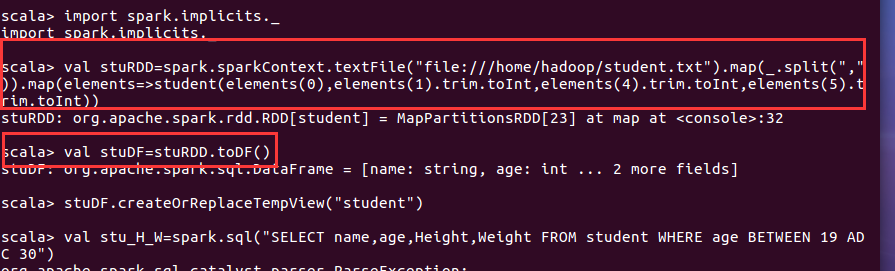




**需要写入数据的临时表**

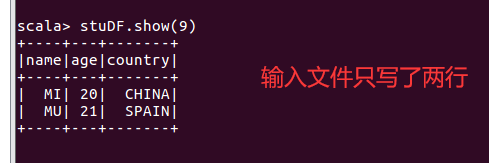


1. **RDD转化为DataFrame**
   * + 1. **反射机制,将rdd隐式转化成DataFrame**





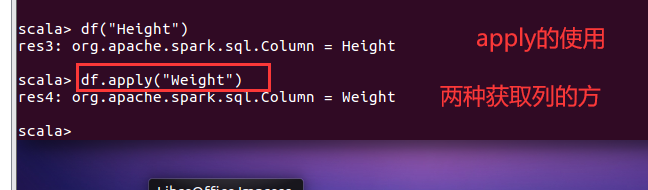
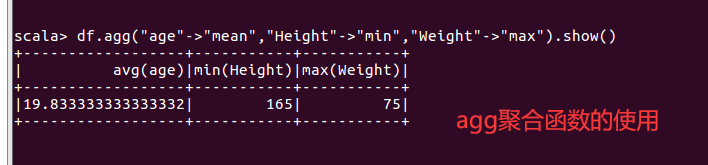
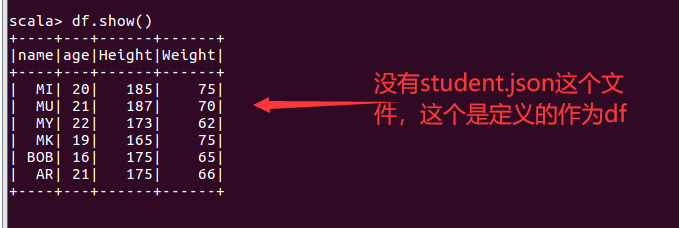
* + - 1. **Rdd显式转化为DataFrame**

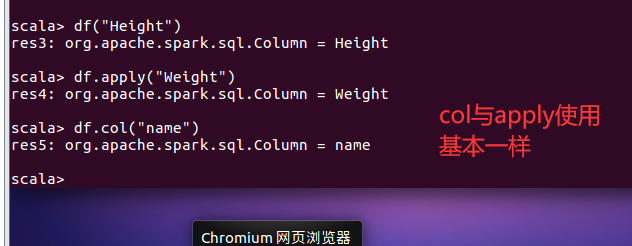




(二)DataFrame操作

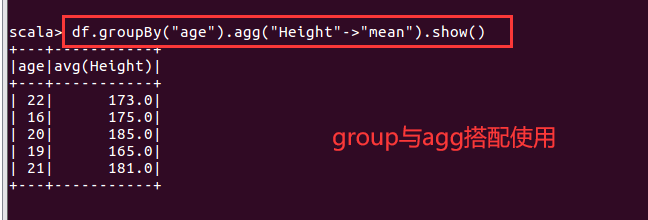
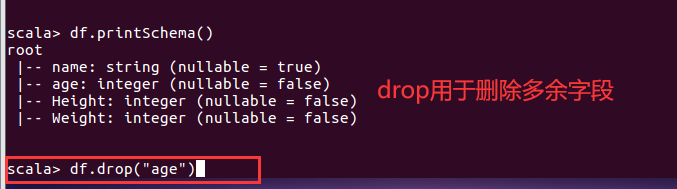
Transformation操作

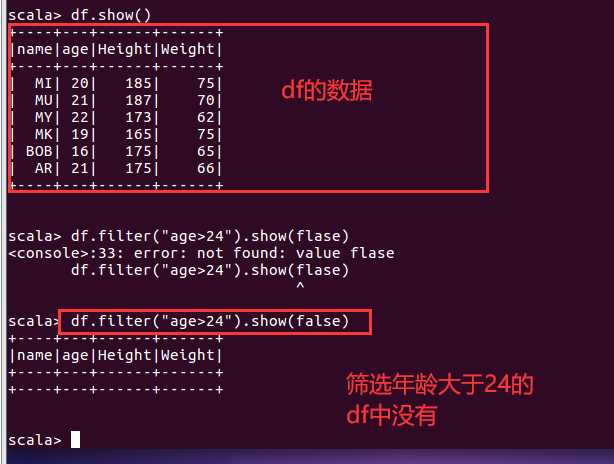


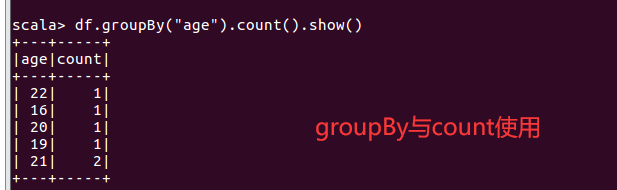


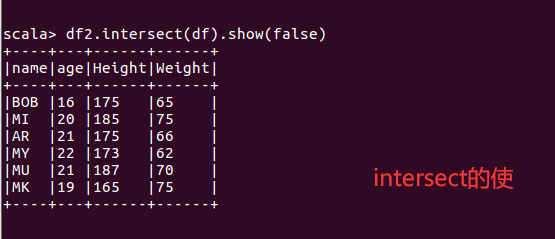


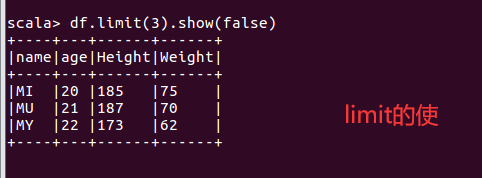


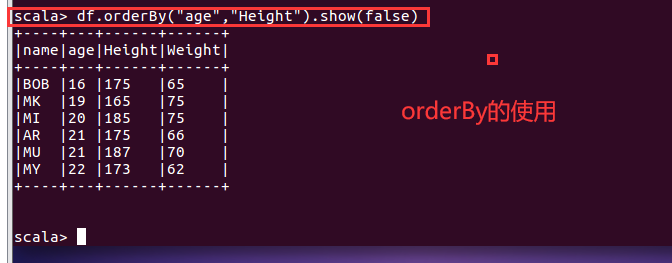




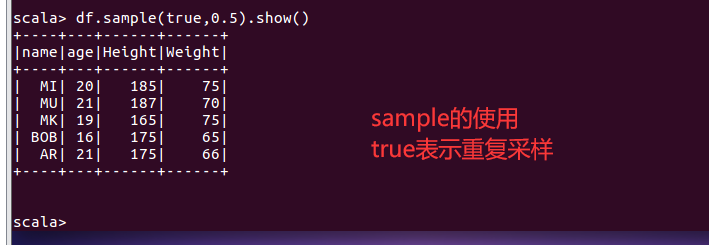


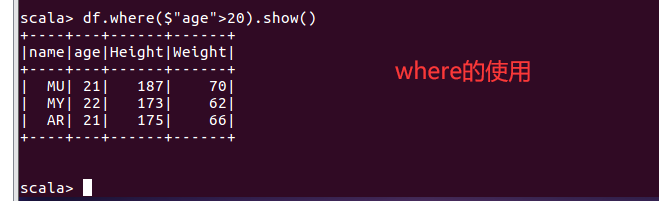


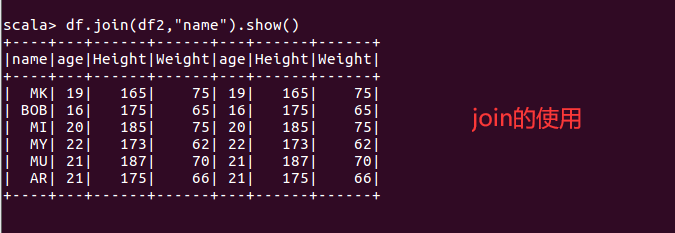




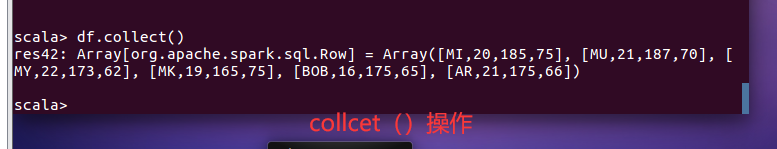


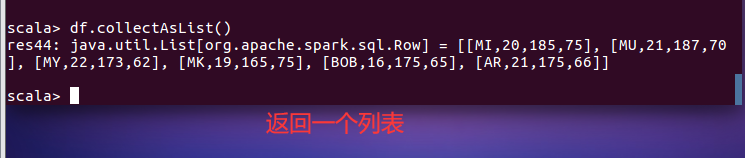




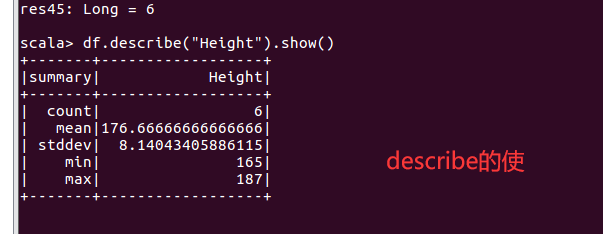


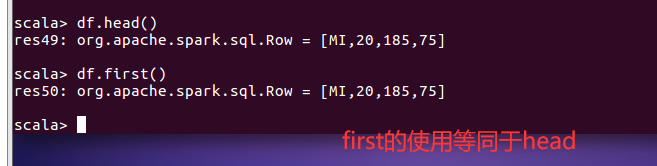
2.Action操作



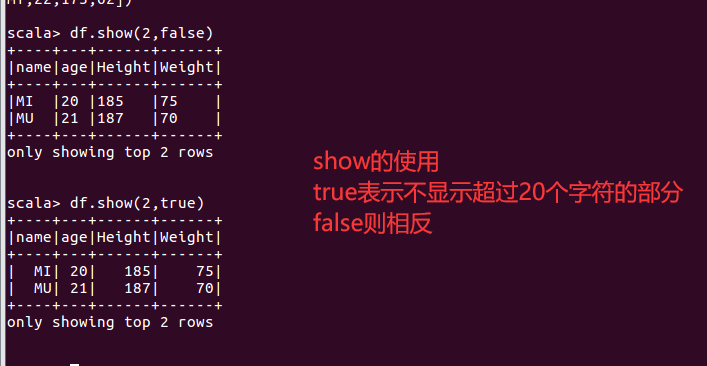


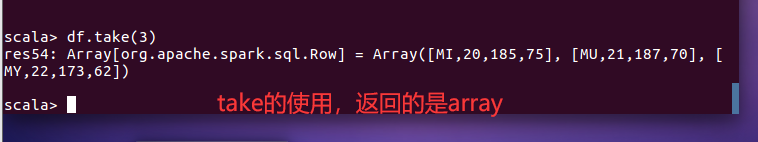




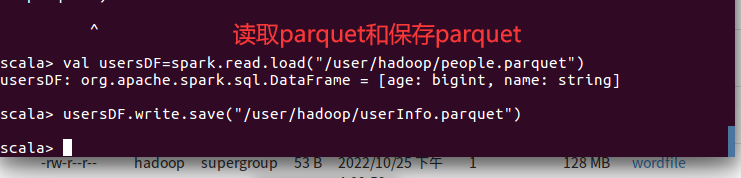


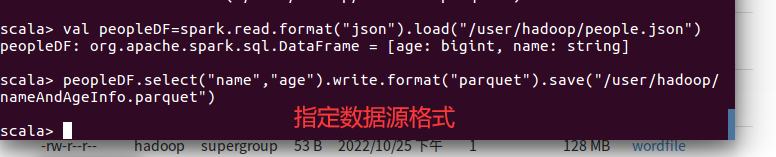






**4.保存操作**



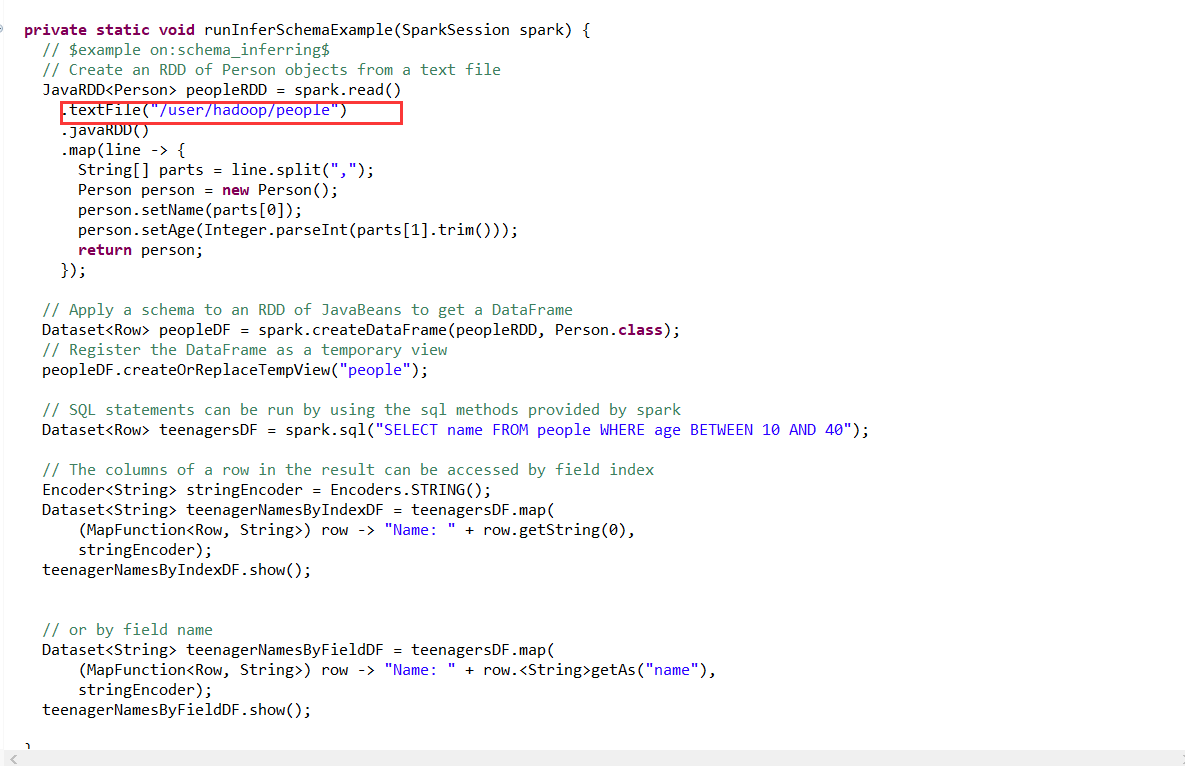


(三) JAVA SPARK SQL 简单实例

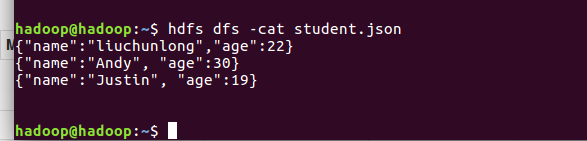
1.代码部分

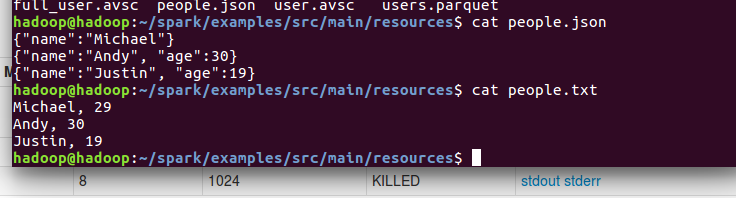




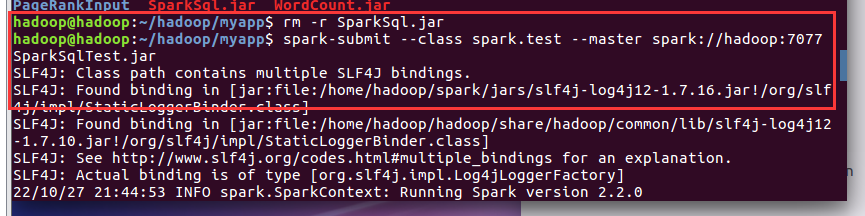


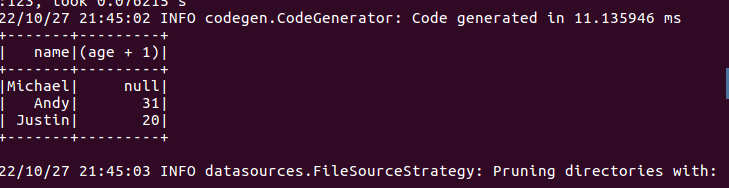
2.输入文件内容

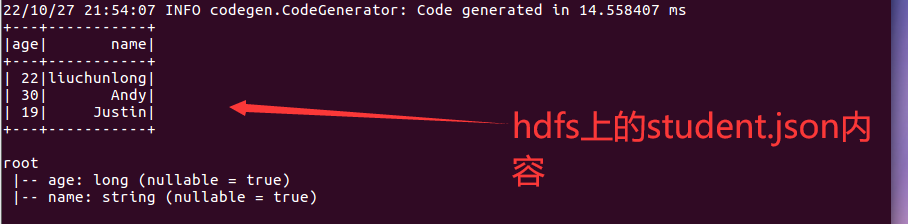


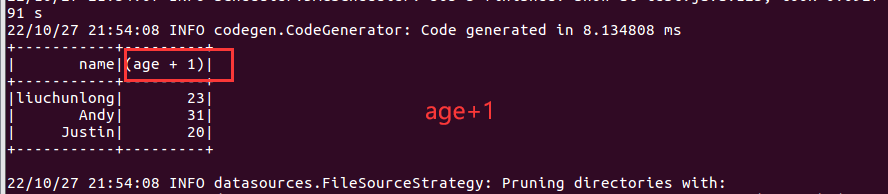


3.运行效果截图











实验总结

本次实验首先了解了DataFrame的创建，读取本地的和HDFS上的json,parquet文件等等。然后学习DataFrame的操作，Transform操作和Acrtion操作等等，值得注意的是要认真去了解各方法的返回值类型，绝大部分是Dataframe类型。

其中还要访问Hive和mysql数据，访问mysql时需要开启mysql服务，还要指定mysql的连接包，开启spark-shell.

最后java spark sql简单实例程序的运行，遇到了不少问题，首先程序中path的填写要注意默认的是hdfs的地址，如果要读本地的要指定file://,其次还有worder分配不到资源的问题，包括核心和内存，通过网上查询慢慢解决了，收益匪浅。