实验三: SPARK RDD 基础编程方法二

胡嘉鑫　　102102145

2023 年 11 月 11 日

目录

1 实验目的

- 理解 Spark 工作流程;
- 掌握 Spark RDD 基础编程方法.

2 实验平台

- OS: Linux
- Hadoop v3.3.5

- JDK v1.8
- SPARK v3.4.0

3 实验步骤

3.1 分析用户行为数据

现在有一份用户行为数据，包括用户 ID、时间戳、事件类型和事件内容。请使用大数据计算工具（如 Spark）进行分析，以解决以下问题。

3.1.1 二次排序和筛选

- 从给定的数据集中，使用二次排序的方法，首先按照用户 ID 升序排序，然后按照时间戳降序排序。输出按照排序条件筛选后的前 10 条记录。
- 从给定的数据集中，使用二次排序的方法，首先按照用户 ID 降序排序，然后按照事件类型升序排序。输出按照排序条件筛选后的前 10 条记录。

图 1: 运行结果

3.1.2 去重和统计

- 从给定的数据集中，找出重复的事件记录（完全相同的记录），并将它们去重。输出去重后的数据集。输出的数据格式：用户 ID 事件类型事件内容

- 对去重后的数据集，统计每种事件类型的数量，并输出事件类型和对应的数量。

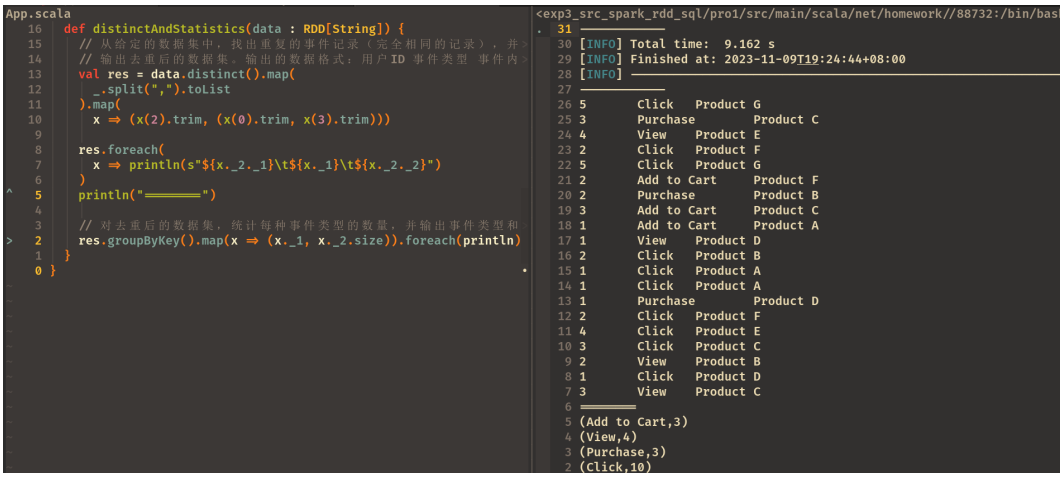

图 2: 运行结果

3.1.3 数据连接和格式化

从用户信息数据集中读取用户 ID 和用户名称。将两个数据集根据用户 ID 进行连接，得到一个新的数据集，其中包含用户信息和相关用户行为记录。

- 对连接后的数据集进行进一步处理，将用户信息和活动详情合并成一个格式化的字符串，例如：”用户名称 - 事件类型 - 事件内容”。
- 输出格式化后的报告，包括每个用户的用户名称、事件类型和事件数量。输出格式为：用户名称–事件类型–事件数量

图 3: 运行结果

3.2 分析校运动会数据

有一份运动会竞赛结果的数据集，其中包括比赛项目、班级、运动员、成绩和名次。使用 Spark SQL 进行分析。

3.2.1 计算所有比赛项目的平均成绩

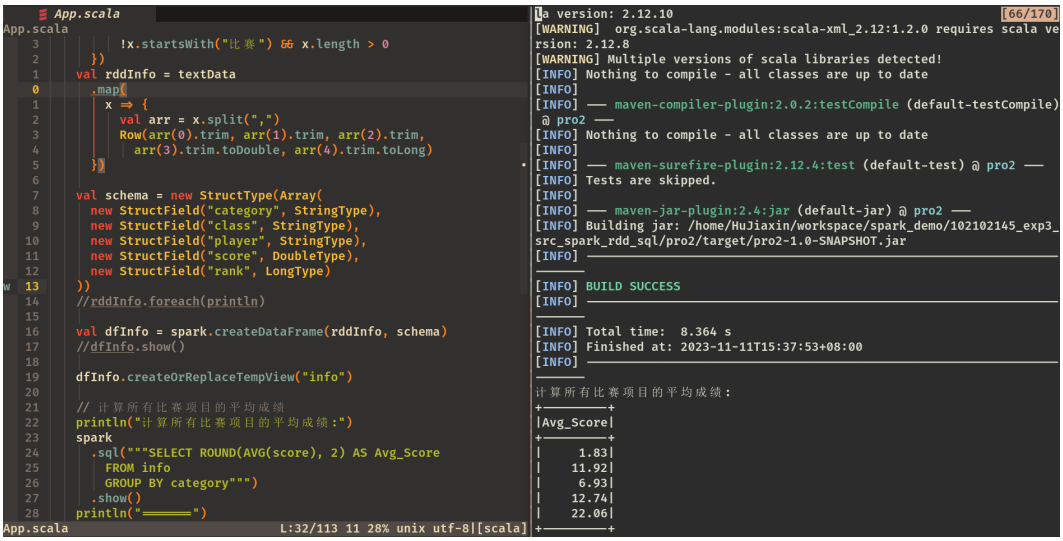

图 4: 运行结果

3.2.2 统计每个班级的名次总数

- 统计每个班级获得的第一名、第二名和第三名的次数。
- 列出获得第一名次数超过 2 次的班级。

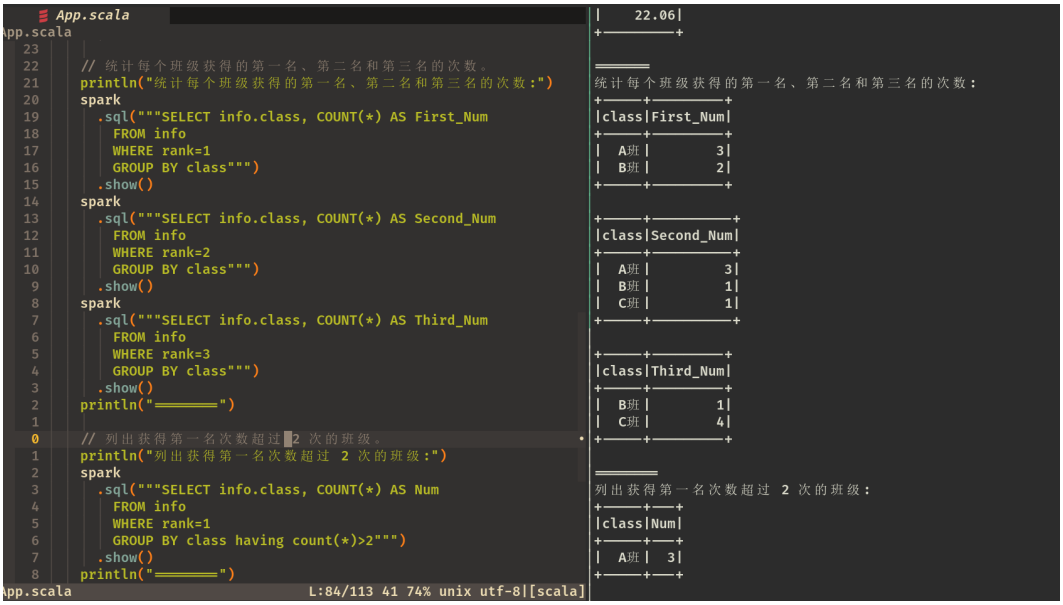

图 5: 运行结果

3.2.3 筛选并统计特定项目的成绩

- 筛选出所有田径项目（如 100 米短跑、200 米短跑）的比赛结果。
- 统计在这些田径项目中获得前三名的个人数量。

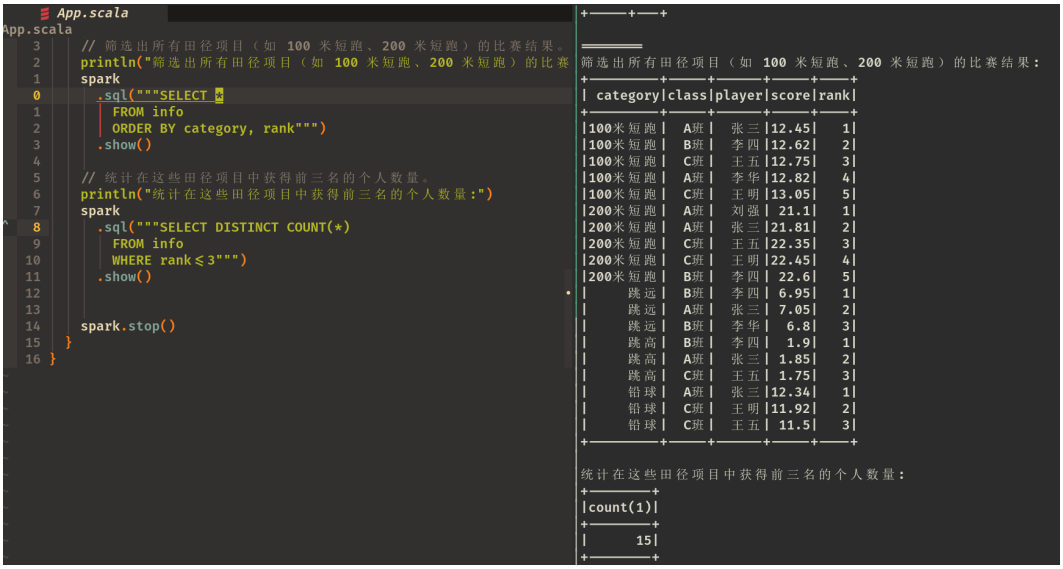

图 6: 运行结果

4 出现的问题及其解决方案

没有问题.