**2020年安徽省大数据与人工智能应用初赛**

**学校名称： 皖西学院 队伍名称： zzw 队伍编号：**

# 第六部分：综合题（15分）

Spark作为大数据第三代计算引擎，在企业中被广泛应用，下面请你完成Spark相关题目：

1. 请简答Spark、Mapreduce、Hive三者区别，并说明分别在什么场景下使用三者，请分别举一个案例（5分）

Spark：使用DAG调度器，查询优化器的物理引擎，能够执行引擎，在批处理和流数据获得很高的性能；使用场景：数据量不是特别大，但是要求实时统计分析需求

Mapreduce：主要区别：处理数据适合处理数据的复杂度运算，不适合处理算法复杂度的运算，不适合实时运算、流式运算、DAG有向图运算等；使用场景：在实际应用中处理大量文件，文件中存储了大量的单词，而且一个单词占一行；例如在搜索引擎中统计最流行的K个搜索词以及统计搜索词的频率，帮助优化搜索词的提示等；

Hive：主要区别：操作接口直接采用SQL语法，提高快速开发的能力，操作较为方便，避免了去写MapReduce，减少了开发人员的学习成本，Hive可以自由的扩展集群的规模，一般情况下不需要进行重启服务；使用场景：现在有一批数据，现在要求：每个用户截止到每月为止的最大单月访问次数和累计到该月的总访问次数；

1. 请完成下面相关统计

请你将以下学生成绩数据，存放在Hdfs上，使用Spark读取完成下面分析



学生表字段描述：学号,姓名,年龄,性别,班级

分数表字段描述：学号,科目名,分数

科目表字段描述：科目名,总分

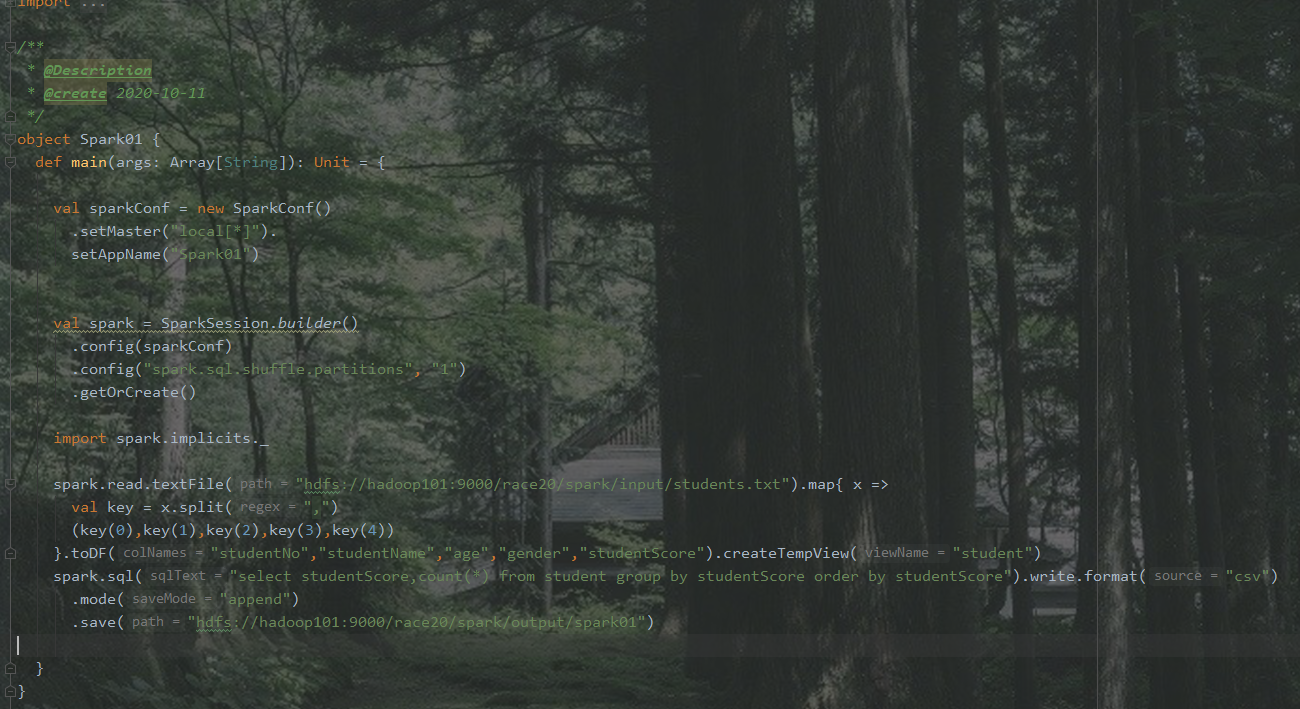
1、使用Spark统计每个班级学生的人数,将统计好的结果保存到文件中（请提供编程代码和截图）（3分）

输出样式:

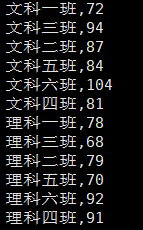
班级,人数

文科一班,39

编程代码：



运行结果：



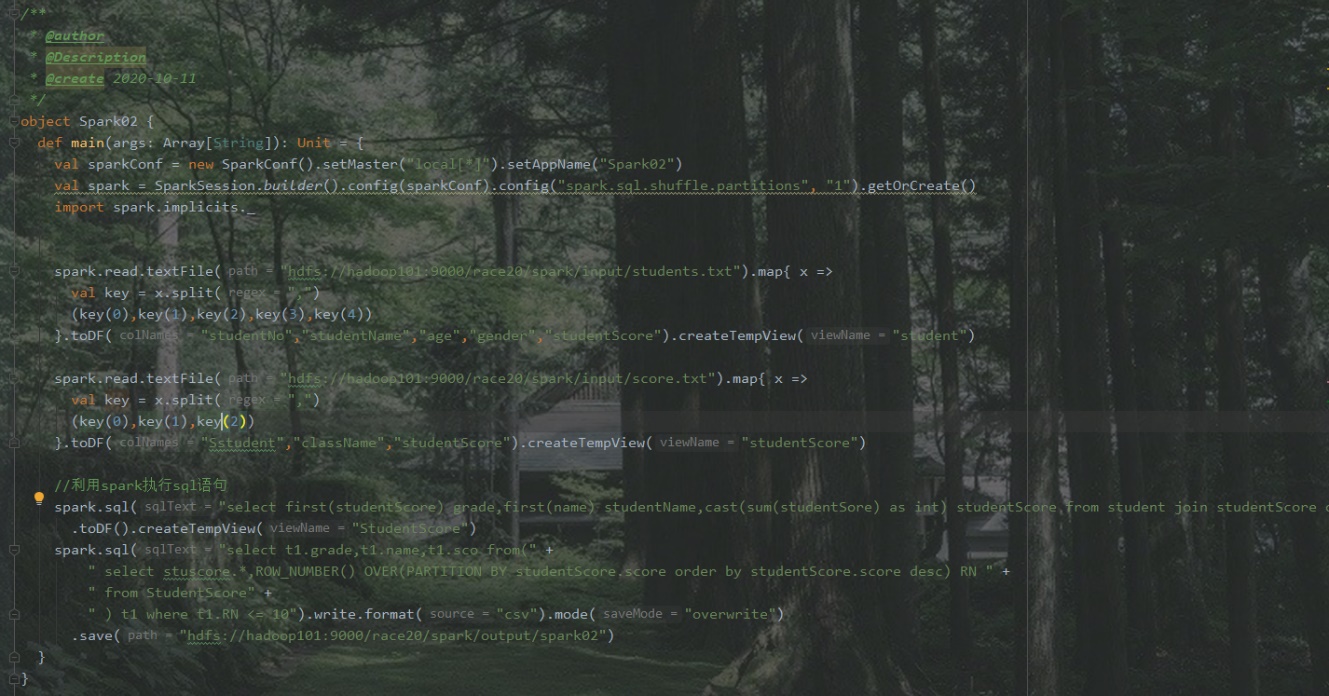
2、使用Spark sql统计每个班级总分排名前十的学生，将统计好的结果保存到文件中 （请提供编程代码和截图）（3分）

输出样式：

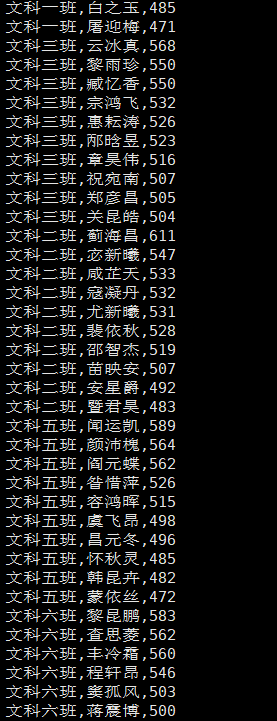
班级,姓名,总分

文科一班,张三,400

编程代码：



运行结果：



3、请使用你擅长的一种编程语言和框架统计每科都及格的学生 （请提供编程代码和结果截图）（4分）

输出样式

学号,姓名,班级,科目名,分数

1500100001,施笑槐,文科六班,语文,80

代码截屏：



运行结果：

