**2020年安徽省大数据与人工智能应用初赛**

**学校名称： 队伍名称： 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）

|  |
| --- |
| **Mapreduce**：hadoop集群上的每台计算机都有自己的cpu，充分利用这些cpu进行并行计算。可以理解为一个计算任务被拆分为多个部分，分配到集群下的计算机上，多台机器并行计算然后再将结果汇总。  hive:基于hadoop的数据仓库工作，可以将结构性的数据映射成一张数据库表，提供HiveQL语句（类sql），并将其转化为mapreduce任务运行在hadoop上。  spark：Spark是基于内存计算的大数据并行计算框架。Spark基于内存计算，提高了在大数据环境下数据处理的实时性，同时保证了高容错性和高可伸缩性，允许用户将Spark部署在大量廉价硬件之上，形成集群、  SparkSQL与Hive的区别: SparkSQL替换的是Hive的查询引擎，Hive是一种基于HDFS的数据仓库，并且提供了基于SQL模型的，针对存了大数据的数据仓库，进行分布式交互查询的查询引擎，所以SparkSQL暂时并不能完全替代Hive，实际上，在生产环境中，SparkSQL也是针对Hive数据仓库中的数据进行查询，Spark本身自己不提供存储，自然不可能替代Hive作为数据仓库的功能。这也印证了，SparkSQL只是替代hive的查询引擎，但是hive构建数据仓库的地位，SparkSQL并不能替代。 SparkSQL一个优点，较于Hive的查询引擎，就是速度快，原因很简单，hivesql执行底层还是用MapReduce来实现，必须经过shuffle过程（性能损耗最大，走磁盘），所以速度是很缓慢的，一些复杂的HiveSQL需要一个小时以上的时间才能执行完。但是SparkSQL由于其底层基于Spark自身的基于内存的特点，所以速度是Hive查询引擎的几倍以上。 Spark与Mapreduce的区别 1）从功能上来说，MapReduce能够完成的各种离线批处理功能，以及常见算法（比如二次排序、TopN等），基于Spark RDD的核心编程都可以实现，而且是更好地、更容易地实现。 2）从速度上来说，基于Spark RDD编写的离线批处理程序，运行速度是MapReduce的数倍，速度上有非常明显的优势 hive理解：本质就是MapReduce，简化了MapReduce任务的开发。让使用sql语言的人可以很快的进行大数据的开发  应用场景及举例： Hive是一个建立在分布式存储系统（这里指HDFS）上的SQL引擎。其应用场景就是数据仓库。基于Hadoop做一些数据清洗（ETL）、报表、数据分析等等。  举例：日志分析：大部分互联网公司使用hive进行日志分析，包括百度、淘宝等。 Spark基于内存计算的准Mapreduce，在离线数据处理中，比较适合改动比较小的业务场景。基于实时数据流的大数据处理（Streaming Data Processing），通常在数百毫秒到数秒之间 ；   举例：Yahoo 将 Spark 用在 Audience Expansion 中的应用，进行点击预测和即席查询等;     MapReduce的典型应用场景中，目前日志分析用的比较多，还有做搜素的索引，机器学习算法包mahout也是之一，当然它能做的东西还有很多，比如数据挖掘、信息提取。   举例：Google建立了基于MapReduce的搜索索引系统。从本质上而言，这个索引是由序列的批处理操作组成的。它通过把对数据集的大规模操作分发给网络上的每个节点进行运算，而每个节点会周期性的把完成的工作和状态的更新报告回主计算. |

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

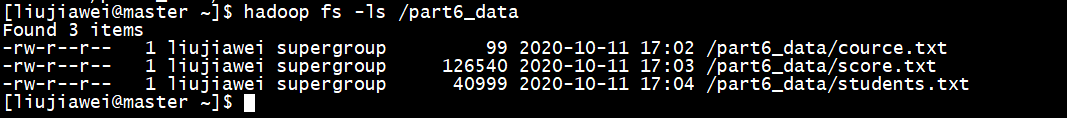
1、使用Spark统计每个班级学生的人数,将统计好的结果保存到文件中（请提供编程代码和截图）（3分）

输出样式:

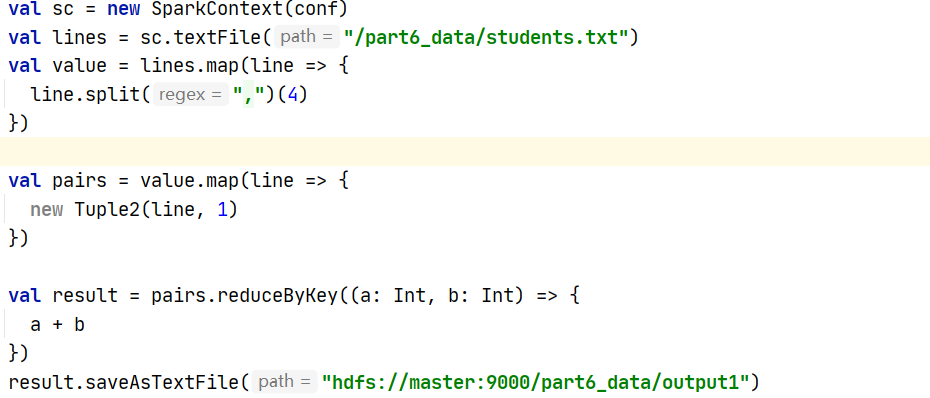
班级,人数

文科一班,39

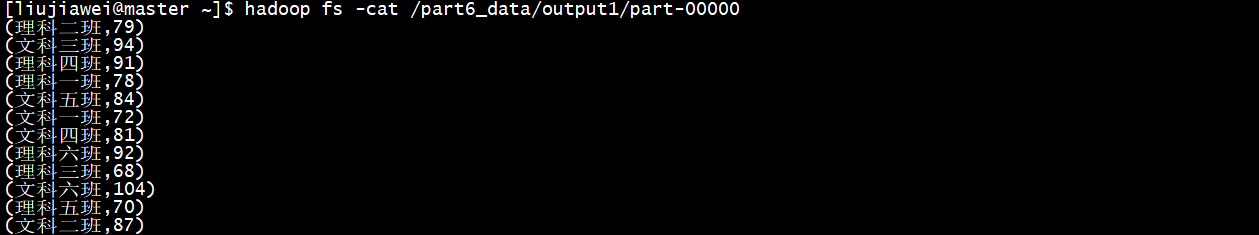
Hdfs数据



**本题我使用Spark-shell写的**



**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

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

|  |
| --- |
| list=[] #存放学生成绩 mfdict = dict() with open(r&apos;cource.txt&apos;,&apos;r&apos;,encoding=&apos;utf-8&apos;) as f1:     list1 = f1.readlines()     for item in list1:         list2=item.split(&apos;,&apos;)         mfdict[list2[0]] = int(list2[1])  del mfdict[&apos;\ufeff语文&apos;] mfdict[&apos;语文&apos;]=150 with open(r&apos;score.txt&apos;,&apos;r&apos;,encoding=&apos;utf-8&apos;) as f:     list1 = f.readlines()     for item in list1:         list2=item.split(&apos;,&apos;)         list.append(list2) list3 = [] with open(r&apos;score.txt&apos;,&apos;r&apos;,encoding=&apos;utf-8&apos;) as f:     list1 = f.readlines()     for item in list1:         list2=item.split(&apos;,&apos;)         list3.append(list2) set1=set() for item in list:     set1.add(item[0]) print(set1) for item in list:     if int(item[2])<mfdict[item[1]]\*0.6:         if item[0] in set1:             set1.remove(item[0]) print(set1) |

全部及格的学生学号

