|  |
| --- |
| 试卷编号：A(开卷) |

桂林电子科技大学试卷

2022-2023 学年第 1 学期 课号 2212261

课程名称 大数据处理技术 适用班级（或年级、专业） 2020级智能科学与技术

考试时间 150 分钟 班级 学号 姓名

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 题号 | 一 | 二 | 三 | 四 | 五 | 六 | 七 | 八 | 九 | 十 | 成绩 |
| 满分 | 4 | 8 | 8 | 20 | 20 | 20 | 20 |  |  |  | 100 |
| 得分 |  |  |  |  |  |  |  |  |  |  |  |
| 评卷人 |  |  |  |  |  |  |  |  |  |  |  |

**交卷时，请将试卷、代码、数据打包发送到： 54pe@163.com**

# 问答题（20分）（请将答案写在问题下方）

1. 当你打开百度搜索引擎时时，百度首页会百度热搜信息。请你用所学大数据相关知识，阐述百度的热搜功能，可能应用到的哪些大数据技术？（4分）



答：

1. 请列举出本课程所学习到的**大数据存储框架**（不低于三个），分别从架构、适用场景和优势与不足等方面进行对这些框架进行对比分析（8分）

答：

1. 请列举出本课程所学习到的**大数据处理框架**（不低于三个），请分别从每个框架的架构、适应场景，数据处理方面的数据来源、处理过程、处理速度等方面进行分析对比。（8分）

答：

# 程序设计题（80分）

1. **HDFS基本操作**（20分）

以下程序，请使用Python语言编程实现。（将关键代码和运行结果截图写在问题下方）

1. 在服务器上创建一个04a.txt文件（也可以在本地电脑上创建），在文件的第一行写上你的学号和姓名。（2分）

**运行截图**：将服务器文件，用cat命令，打印在终端并截图。

1. 编写Python程序，在HDFS中创建目录，目录路径为：/exam/。（4分）

**运行截图：**

**关键代码：**

1. 编写Python程序，将修改好的文件上传至HDFS中，目录和文件名为 /exam/04a.txt。（6分）

**运行截图：**

**关键代码：**

1. 编写Python程序，查看并打印HDFS中/exam/04a.txt文件的属性信息，包括文件大小、创建时间。（4分）

**运行截图：**

**关键代码：**

1. 编写程序，从HDFS中下载/exam/04a.txt文件到本地电脑~/exam目录。（4分）

**运行截图：**

**关键代码：**

1. **MapReduce编程**（20分）

以下程序，请使用Python语言编写实现。（将关键代码和运行结果截图写在问题下方）

1. 在服务器上使用wget命令，下载待统计的文本文件到服务器上，并命名为5.txt。文件地址：（2分）

<https://labeye.oss-cn-beijing.aliyuncs.com/hadoop/exam/5.txt>

**运行截图**：

1. 编写程序，将下载的文件上传到HDFS的/exam/05.txt下。（2分）
2. 编写程序，统计HDFS的/exam/05.txt文件中单词出现的次数，并按单词的顺序进行排序。期望得到如下结果：（12分）

Chinese 10

English 12

… …

1. 将分析结果保存到HDFS的/output/05/目录下。（4分）

**运行截图**：

**关键代码：**

1. **Spark Shell编程**（20分）

以下程序，使用Spark Shell进行编程，每小题完成后，将运行结果界面截图保存到此试卷。

* 1. 使用Putty等SSH远程客户端软件，连接至Spark集群的主节点。（2分）

**运行截图**：

* 1. 通过命令行进入Spark Shell。（2分）

**关键命令**：

**运行截图**：

* 1. 加载HDFS中目录/exam/05.txt的文件。（4分）

**关键命令**：

**运行截图**：

* 1. 编写Shell代码，对05a.txt的单词进行统计。（8分）

**关键命令**：

**运行截图**：

* 1. 将统计分析结果保存到HDFS的/output/06/目录下。（4分）

**关键命令**：

**运行截图**：

1. **综合设计**（20分）

以下程序，请使用Python语言编写实现。（将关键代码和运行结果截图写在问题下方）

**说明：**已知以下文件存储员工信息，信息包括有：工员、姓名、性别、年龄、月工资。

文件地址：

https://labeye.oss-cn-beijing.aliyuncs.com/hadoop/exam/07.txt

**0001 李梅 女 23 2040.00**

**0002 陈霞 女 20 1879.70**

**要求：**请编写MapReduce程序，下载以上文件，分别统计男生、女生各多少人，每个年龄各有多少人，所有人的平均工资。将结果保存到/output/07/目录下。

**关键代码**：

**运行截图**：