**2019年安徽省大数据与人工智能应用网络赛题**

**队伍编号：**

**说明：本次比赛为操作题。请在相应的虚拟机上作答，请遵守考试规则。**

# 第一部分：大数据平台部署（10分）

**场景说明**

目前，大数据技术及应用已经深入到各行各业，各高校也在积极建设信息化、智慧化校园，那么大数据中心建设必不可少，它可以服务于学校上游和下游，解决数据孤岛问题。假设现在某大学正在假设大数据中心，首先搭建大数据基础平台，用于数据存储和批量计算以及实时查询，现在请你完成Hadoop和Hive两个大数据框架的搭建及运维：

任务一、安装环境准备（5分）

### 修改各主机的主机名，并配置 hosts实现主机名和ip地址映射 ，要求主机名称为队伍名称英文全拼。截图hosts文件配置内容 （0.5分）

**结果截图：**

### 配置SSH免密登录（截图成功ssh登录到子节点截图）（0.5分）

**使用ssh登陆子节点 结果截图：**

### 安装配置JDK（截图java -version命令）（0.5分）

**执行命令 java -version 结果截图：**

### 解压Hadoop安装包并修改配置文件(0.5分)

**结果截图：**

### 配置环境变量（截图profile文件配置的关键位置）（0.5分）

**执行命令 # vi /etc/profile 结果截图：**

### 拷贝hadoop到其他的机器上（截图拷贝命令）（0.5分）

**结果截图：**

### 初始化hadoop集群（截图执行命令）（1分）

**结果截图：**

### 启动Hadoop集群（启动成功后执行jps命令，截图主节点和子节点的进程）（1分）

**执行命令 jps 主节点截图：**

**从节点1截图：**

**从节点2截图：**

## 任务二、配置Hive集群环境（5分）

Hive是基于Hadoop的一个数据仓库工具，其运行依赖Hadoop和Mysql，其中Hadoop为其提供Hdfs文件系统，Mysql为其提供元数据存储。下面我们开始搭建Hive。

### Mysql安装（截图mysql搭建成功show databases命令结果界面2分）

**进入mysql查询 执行 show databases; 语句 结果截图：**

### Hive搭建（关键步骤截图2分）

1. 添加环境变量 执行命令 # vi /etc/profile

**结果截图：**

1. 修改配置文件

根据环境部署的需求，修改主节点上的hive配置文件：hive-env.sh和hive-site.xml。

**结果截图：**

### 启动测试hive（截图搭建成功show databases命令结果界面1分）

hadoop启动情况下，执行hive命令

# hive

测试输入

hive> show databases;

**结果截图 ：**

# 第二部分：大数据预处理部分（25分）



1 该数据集log.log是某APP用户日志，请你使用MapReduce程序对数据进行预处理清洗提取。（15分）

1.1 原始数据预处理（编写相关代码及部分结果截图8分）

规则一：如果没有数据中没有uid、platform、app\_version、pid四个字段同时出现的数据，请过滤掉

规则二：将数据中字段locationcity的值为0的全部替换为1

**代码文件：**

**结果截图：**

1.2请你将原始数据中用户的"uid","platform","app\_version","pid","cityid" 五个字段和期对应的值提取出来。（编写相关代码及部分结果截图7分）

**代码文件：**

**结果截图：**

2 该题为数据处理常用处理方式，进行两个数据集的关联，原始数据集依然使用上述题1的log.log数据，另一数据集为cityid.txt。请完成数据关联。（10分）



2.1请将1.2题目的输出结果中city的值对应编号关联匹配数据cityid.txt，将城市编码替换为城市名称输出，截图关联后的部分结果

**代码文件：**

**结果截图：**

# 第三部分：大数据分析（20分）



1、 示例数据集dns\_log.txt是某公司网站访问的日志，其中第1个字段为访问的源ip地址，第2个字段为访问的网站地址，分隔符为竖线 ’|’ 。

1.1 利用MapReduce程序进行简单的过滤清洗，将字段个数不满足5个的数据过滤掉，并且将网站地址中为 ”www.taobao.com” 的标记为购物网替换为“ShoppingAction ”，最后将清洗过滤后的数据全部输出（5分）

**代码文件：**

**结果截图：**

1.2 统计输出该原始数据集中网站被访问次数最多的前5名的网站地址和次数。（5分）。

使用到的json解析相关包：



**代码文件：**

**结果截图：**

**2、基于hive的数据仓库离线计算编程（10分）**

Hive作为基于hadoop的海量数据处理工具之一，在企业中有广泛应用。可以将复杂的MapReduce计算逻辑使用SQL逻辑处理。

2.1 创建一个数据库，以你的组名命名，创建成功后使用use命令切换为该库，

并执行set hive.cli.print.current.db=true;截图作为答案（2分）

**Sql语句：**

**结果截图：**

2.2 根据该题中原始数据建一个hive内部表，命名格式为:你的组名\_日期\_表名称，如：anhuidaxue\_001\_20190831\_dnsdata,文件存储格式为 textfile，文件存储位置为数据在你hdfs中你存放数据的路径。截图建表成功（2分）

**Sql语句：**

**结果截图：**

2.3 将上面MapReduce程序题目1使用SQL处理，写出SQL语句，截图含有jobid运行的日志（2分）。

**Sql语句：**

**结果截图：**

2.4 将上面MapReduce程序题目2使用SQL处理，写出SQL语句，截图含有jobid运行的日志（2分）。

**Sql语句：**

**结果截图：**

2.5 统计该原始数据中每个网站访问次数的TOP3的ip地址和次数，也就是每个网站访问次数最多的前3名的ip地址和次数。截图含有jobid运行的日志以及sql（2分）。

**Sql语句：**

**结果截图：**

# 第四部分：可视化（15分）

**近些年空气污染在我国很多地区非常严重，其中PM2.5作为衡量空气质量的一个重要指标，当前数据为2018年12月份1号-10号全国大部分城市的站点检测数据。接下来我们将这些数据进行可视化来看下PM2.5的分布情况。**

****

数据：

1. 利用Bar图输出城市（city）合肥、黄山、芜湖的PM2.5指数，Y轴表示平均数，；X轴表示城市名称，（5分）

要求将输出的直方图保存成图像文件，程序源代码截图，

**代码文件：**

**结果截图：**

1. 利用折线图，画出城市，各城市随时间的pm2.5值变化

要求将输出的折线图保存成图像文件Y轴表示pm2.5的值，X轴表示日期，城市

要求：（5分）

1. 折线图中含图例；

不同的城市用不同的颜色表达；

**代码文件：**

**结果截图：**

3、利用柱线混合图形画出合肥市，高新区，庐阳区的pm2.5的值。要求将输出的折线图保存成图像文件，X轴表示时间，柱状表示 高新区 的pm2.5的值，线形表示 庐阳区的pm2.5的值。（5分）

要求：

1. 图形中含有图例

**代码文件：**

**结果截图：**

# 第五部分 人工智能（10分）

目前人工智能技术广泛应用于各个场景，如分类聚类，图片图像识别，音视频检测等等。下面请你搭建人工智能中深度学习常用的框架之一 **TensorFLow 并完成简单测试任务，本次实验在Linux环境下安装**TensorFlow**。**

1. TensorFlow安装（7分）

1 Linux安装python3环境

1）解压并安装依赖包

**命令行：**

**结果截图：**

2）

指定安装目录，开始安装

**命令行：**

**结果截图：**

3）建立软连接

**命令行：**

**结果截图：**

2 使用pip3安装pip3 tensorflow

**命令行：**

**结果截图：**

1. TensorFlow测试，将测试结果截图（3分）

进入python命令下，测试tensorflow：

　　　　import tensorflow as tf

　　　　sess = tf.Session()

　　　　hello=tf.constant('Hello,Tensorflow!')

　　　　print(sess.run(hello))

回车，若在终端下显示 Hello,Tensorflow! 则表示安装tensorflow成功

**结果截图：**

# 第六部分：综合题（20分）

随着移动互联网、物联网等新技术的迅速发展，人类进入数据时代。大数据带来的信息风暴正深刻改变我们的生活、工作和思维方式，对网络舆情管理也带来深刻影响。在电影行业大数据舆情管理有着广泛的应用，可以通过舆情分析来判断当前电影的热度和影响度等。



id, created\_at, attitudes\_count, comments\_count, reposts\_count, screen\_name, follow\_count,followers\_count, gender, text

字段含义：id,发布时间,点赞数,评论数,转发数,作者昵称,关注,粉丝数,博主性别,博客内容

1 该数据集log\_movie.txt是通过网络爬虫从某主流媒体网站爬取下来的部分关于电影《烈火英雄》、《上海堡垒》的数据，请将该数据上传到hdfs中，目录可以自行创建指定（2分）。

上传到hdfs根目录：

**命令行：**

**结果截图：**

2 使用Spark将数据集中数据关于“烈火英雄”相关的数据提取出来，并回写到hdfs中（请提供编程代码）（3分）。

由于权限问题，先在HDFS中创建一个文件夹test并赋权

# hdfs dfs -mkdir

# hdfs dfs -chmod

**代码文件：**

**生成的结果文件截图：**

**查看结果文件内容截图：**

**# hdfs dfs -cat**

3 使用Spark统计《烈火英雄》相关文章关于评论人性别的占比（请提供完整代码和结果截图5分）。

**代码文件：**

**结果截图：**

4 统计关于《上海堡垒》话题的舆情走势，已知其上映时间为2019-08-09。（本题可以使用你擅长的编程语言和框架）

4.1 讨论热度的计算，按照日期对数据集中每日总评论数计算，按照日期排序（请提供完整代码和结果截图3分）

**代码文件：**

**结果截图：**

4.2 传播影响的计算，计算出转发次数最多的前5条文章，并列出该文章博主的粉丝数（请提供完整代码和结果截图3分）

**代码文件：**

**结果截图：**

4.3 针对于上述指标统计，请简要结合你分析的数据结果，描述当前舆情发展态势（4分）。

**描述：**