2017 年度职业院校技能大赛

大数据技术与应用赛项 赛题

第一节 赛题

“四合影业”公司计划参与投拍一部电影，名为《青春的竞赛》。为提高票房收入，降低投资风险，需要了解电影市场的情况，包括何种类型的电影票房收入高，不同类型观众对电影的偏好等等。为此，四合影业公司计划聘请“TMS”大数据分析公司，收集并分析电影市场的相关信息，并最终给出分析报告。合作之前，“四合影业”需要“TMS”公司提出可行的技术解决方案。

为完成四合影业的项目，“TMS”公司选用了在业界广泛应用的“Python”语言，作为开发分析程序的基础语言，并综合利用 numpy、pandas、matplotlib、scikit 模块和 MapReduce 技术提高开发效率，由于预计数据量会超过“T”级，“TMS”公司在技术方案中提出在一个高性能工作站集群上利用 Hadoop 平台提高数据处理能力，并利用 Hive 以及 streaming 技术提高效能和简化 MapReduce 过程。但此技术方案需要较高成本，为向“四合影业”展示该技术方案的合理性并达成与“四合影业”的合作，“TMS”公司先用廉价 PC 集群，配置了小规模的技术演示环境，并利用网络爬虫抓取了历年来影音娱乐行业的信息，数据量约为 4G，随后开发了程序对数据进行清洗、整理、计算、表达、分析，力求展示 “TMS”技术方案的合理性和自身出色的技术能力。

作为“TMS”公司的技术人员，你们是这次技术方案展示的核心成员，请按照下面步骤完成本次技术展示任务，并提交技术报告。圆满完成展示并得到预期结果，“TMS”就能获得这个数百万元的项目合同，祝你们成功。

任务一、部署 Hadoop 平台，并根据计算对象调优 Hadoop 平台的性能（15 分）

1、 按下面如下个步骤完成 Hadoop 环境的部署：

1) Hadoop 系统存储于“/usr/local/hadoop” ，要求配置 hadoop.tmp.dir 目录存放位置为“/usr/local/hadoop/tmp”

2) 配置 hadoop 的 dfs.namenode.name.dir 为/usr/local/hadoop/tmp/dfs/name
3) 配置 hadoop 的 dfs.datanode.data.dir 为/usr/local/hadoop/tmp/dfs/data
4) 格式化 NameNode
5) 开启 NameNode 和 DataNode 守护进程

本题要求配置完成后在 Hadoop 平台上运行 jps 命令，要求 jps 运行结果的截屏保存于文件 ans0101.jpg 中

2、 按下面步骤建立用户目录，并导入存于竞赛平台 arg 目录中的数据文件 dat0102.dat，并完成 Hadoop 平台的性能测试：

1) 在 hdfs 中创建用户目录（如果系统用户为 hadoop，请建立/user/hadoop）
2) 在 hdfs 中创建 input 目录，把数据文件上传至 input 目录
3) 运行 hadoop-mapreduce-examples-2.7.3.jar 查询特定字符串出现次数
4) 用 hdfs 命令查看输出结果。

本题要求将第 4 步运行结果的截屏保存于文件 ans0102.jpg 中。

3、 对 Hadoop 平台进行性能调优，设置: yarn.scheduler.maximum-allocation-mb 的值为系统内存减 1024MB，mapreduce.map.memory.mb 的值为 1024MB，设置 mapreduce.map.java.opts 的值为 -Xmx768m，设置 mapreduce.reduce.memory.mb 的值为 2048MB，设置 mapreduce.reduce.java.opts 的值为 –Xmx1536m。重新启动 hadoop。 本题要求提交修改后的配置文件，文件名为在原有文件名加前缀”ans0103_”。（5%）

任务二、数据抓取（30 分）

1、 现在，网络爬虫抓取到约 4G 的数据，保存于 arg 目录的 spider.log 中，但其中既有电影市场放映信息数据也有其他数据，通过分析数据样本，发现从网站“http://www.movie.com/ bor/”抓取的数据包含有效的电影市场数据，数据中有效数据项包括：电影名称、上映日期、上映场次数、院线城市、导演、演员、影片类型、票房收入，请从 spider.log 中筛选出一部分有效数据项，并以规定格式保存于 ans0201.csv 文件中。本题的赛前抽取参数是：数据文件 spider.log、需要保存于 ans0201.csv 文件的有效数据项以及有效数据项的保存格式。

2、 网页 “http://movie.xtime.com/FilmId/ ”中包含观众对电影的评分信息，请编写程序抓取网页（网页样本保存于 task0202 目录中）上电影的评分信息并计算其统计信息（统计方法指对某部电影的评分求极值或求平均值），本题的赛前抽取参数是统计方法以及网页样本，请参赛学生将本题的答案保存于 ans0202.txt 文件中，注意 ans0202.txt 文件中只能包含一个浮点型数字，保留 4 位小数，文件样例如下：

```
1.2345
```

3、 向 Hadoop 平台提交日志文件 dat0203.log，并使用 streaming 和 MapReduce 机制编制程序，统计日志文件 dat0203.log 的数据中一共包含多少部电影？本题的赛前抽取参数是 dat0203.log 文件，请参赛学生用 hdfs 命令查看输出的结果，截屏保存于图片 ans0203.jpg，并用 hdfs 命令把输出文件传输到本地，修改文件名为 ans0203.txt

4、 根据本题给定的数据文件 dat0204.log 编写 Hive 命令建立数据表，并将 dat0204.log 导入所建立的数据表，然后编写 Hive 查询语句获取 2014 全年上映电影的数据记录，并将查询结果导入 Hadoop 平台的 result 目录。本题赛前抽取参数是 dat0204.log 文件，请参赛学生将完成本题要求的所有命令按步骤顺序以分行的形式保存于 ans0204.txt 中，ans0204.txt 的文件样例如下：

```
statement 1
statement 2
statement 3
statement 4
```

任务三、本阶段的任务是：film_log3.csv 中包含了不同地区、不同影院的电影票房信息，你的小组通过编程完成对文件 film_log3.csv 中电影信息数据的清洗和整理，并完成数据计算、分析和表达任务。（20 分）

本竞赛任务的赛前抽取参数是：电影名称 A、B、C 和地名 M 市、N 市以及数据文件 film_log3.csv，选手可在竞赛环境的 arg0300.txt 文件中获得 A、B、C、M、N 的值。本任务阶段，需要参赛学生提交每个小题涉及到的所有 ansXXXX.jpg、ansXXXX.py、ansXXXX.dat 文件（XXXX 相关指数字）。

1、 编程统计并输出影片 A 的上映天数和日平均票房（文件中的所有涉及地区总平均），程序源代码保存成 ans0301.py，并将结果保存于 ans0301.dat，要求 ans0301.dat 只包含 1 个 long 型数据和一个 1 个浮点型数据，浮点数据以万元为单位，保留 6 位小数，2 个数以英文逗号分隔，不换行，文件样例如下：

123, 23.123456

2、 编程绘制一个直方图，在图中输出影片 A、B、C 的周平均票房（文件中的所有涉及地区周票房总平均），Y 轴表示票房收入，单位万元；X 轴表示电影名称，电影名称的排列从左至右以 A、B、C 为准，要求将输出的直方图保存成图像文件 ans0302.jpg，程序源代码保存成 ans0302.py，另外，将三部电影各自的票房总收入按自高到低的顺序存入 ans0302.dat 文件中，要求 ans0302.dat 中只包含 3 个浮点型票房数据，以万元为单位，保留 6 位小数，数据以英文逗号分隔，不换行，文件样例如下：

23.123456,20.654321,18.123456

对本题周票房的说明如下：若某部电影从某月 2 日开始上映，则从当月 2 日到 8 日为其第 1 周票房，9 日至 15 日为其第 2 周票房，不满 1 周按 1 周计算以此类推。

3、 编程，在一个折线图中，画出影片 A、B、C 各自的周票房（文件中的所有涉及地区总周票房）收入变化，要求将输出的折线图保存成图像文件 ans0303.jpg，程序源代码保存成 ans0303.py，Y 轴表示票房收入，单位为“万元”；X 轴表示时间，以“0、1、2、3...n”的非负整数作为刻度值，单位为“周”，要求：

1) 折线图中含图例；
2) 三部电影用不同的颜色和线型表达；
3) 将电影 A 第一周的票房收入，电影 B 第二周的票房收入，电影 C 第三周的票房收入顺序存入 ans0304.dat 文件中，注意 ans0303.dat 只包含 3 个浮点型票房数据，以万元为单位，保留 6 位小数，数据以英文逗号分隔，不换行，文件样例如下：

   23.123456,20.654321,18.123456

4) 对本题周票房的说明如下：若某部电影从某月 2 日开始上映，则从当月 2 日到 8 日为其第一周票房，9 日至 15 日为其第 2 周票房，以此类推。

4、 编程，在一个子图系统中，用两个水平排列的折线型子图画出 M 市和 N 市 2016 年 1 至 3 月的上映电影的票房总收入趋势，要求将输出的完整子图保存成图像文件 ans0304.jpg；程序源代码保存成 ans0304.py，要求：

1) 左子图为 M 市票房总收入趋势，右子图为 N 市票房总收入趋势，Y 轴表示票房收入，单位为“万元”，X 轴表示时间，以“0、1、2、3”作为刻度值；
2) 两子图均有说明子图内容的标题（如：M 2016 1-3 BOR）；
3) 将以下 6 个数据分 2 行按顺序存入 ans0304.dat 文件中，要求 ans0305.dat 只包含浮点型数据，以万元为单位，保留 6 位小数，需要保存的票房数据是：第 1 行 3 个数据，按顺序分别是 M 市电影市场 2016 年 1、2、3 月票房总收入，第 2 行三个数据，按顺序分别是 N 市电影市场 2016 年 1、2、3 月票房总收入，同行数据以英文逗号分隔，文件样例如下：

2023.123456,2000.654321,1988.123456
2303.123456,2100.654321,17898.123456

任务四、根据现有数据，编写分析报告，分析电影市场情况并预测观众群对“四合影业”计划投拍的电影“青春的竞赛”的评分。（30 分）

请从 arg04 子目录中选取需要的数据文件，依据观影俱乐部的观众评分（评分为 10 分制），利用统计图表分析说明影片类型、导演等因素对观众的影响，以及导演擅长的电影类型，最后预测某观影俱乐部中的 5 位会员对于《青春的竞赛》的评分范围，本赛题需要提交分析报告和相关程序，本题的赛前抽取参数是 5 个会员 ID（保存于 id04.txt 文件中）和数据文件。

分析报告和所提交的程序的要求：

1、 利用 WPS 或 WORD 软件完成分析报告，文件名为 anl0400.doc 或 anl0400.docx，报告中需要明确描述分析方法，分析过程。
2、 分析报告中用明确的表格显示以下数据，评分最高值，评分最低值，评分中位数，评分均值 。
3、 分析报告中至少包含三种图，分别能够表达“各种类型片票房收入比较”，“导演票房收入比较”，“导演执导过的影片类型”的内容。
4、 提交支撑程序名为 ans0400.py，要求程序运行后不可做任何人为操作，自动完成以下任务：

1) 在一个子图系统中输出要求 3 中所提及的三种图，该子图水平排列，顺序以要求 3 所列顺序为准，每个子图的具体形式不限。程序能够有提示地输出 4 个数据：评分最高值，评分最低值，评分中位数，评分均值。

2) 要求按次序将分析得出的评分最高值，评分最低值，评分中位数，评分均值，存入 ans0400.dat 文件中，要求 ans0400.dat 只包含所要求的 4 个浮点型数据，每个数据保留 2 位小数，英文逗号分隔，不分行，文件样例如下：

9.12,2.65,6.12, 5.68

团队综合素质（5 分）

团队分工明确合理、操作规范、文明竞赛。

第二节 学生竞赛须知

重要声明

参赛学生必须严格按照赛题要求的文件名保存所提交的文件，除最终报告所有提交的报告外，参赛学生提交的所有数据文件只包含数字和英文逗号。所有数据文件将由评分系统自动评分，数据文件中不得含有学校名称、参赛学生姓名、座位号等信息，否则 0 分处理。

参数说明

网上的公开样题给出了所有参数文件和参数的样本，以供学生练习，但为方便参赛校下载练习，网上公布的参数文件规模为“M”级（兆），赛题涉及的参数和参数文件将在开赛前的 12 小时内由执委会从参数库中随机抽取，并保存在竞赛环境的相应目录中。所有参数均保存在 arg 目录中，以下的文件名和目录名，均是针对 arg 目录的相对路径。

1、任务一的参数：

1) 无参数，竞赛环境的 hadoop_scr 目录中提供了 Hadoop Vxxx 的安装文件
2) 有 1 个参数文件：dat0102.dat
3) 与第 2 小题相同，1 个参数文件：dat0102.dat

2、任务二的参数：

1) 目录 task0201 中有 1 个参数文件 spider.log，文件格式参数见相同目录 readme.txt 文件
2) 目录 task0202 有 1 个网页样本

3) 目录 task0203 有 1 个参数文件：dat0203.log

4) 目录 task0204 有 1 个参数文件：dat0204.log

3、任务三的参数：

1 个参数文件，film_log3.csv；

选手可在竞赛环境的 arg0300.txt 文件中获得 5 个文本参数 A、B、C、M、N

4、任务四的参数：

多个数据文件保存于 arg04 子目录中，参赛学生自己根据需要选取；

5 个会员 ID 保存于 id04.txt 文件中