《大数据初识》实验指导手册

《大数据初识》实验是学习本课程必须要通过的一个重要的实践环节。

本指导手册共计有编有6个实验，在实际实验过程中可以按照实际教学进度情况加以增删。每个实验都应做到：

1.做好上机实验的准备工作：应考者对每个实验需要进行预习，了解相关内容和具体要求，熟悉上机操作步骤，学会相关系统软件的使用。

2.搭建工作流：按照实验题目的要求，应考者要事先熟悉涉及到的工作流组件，并初步设计工作流，以减少错误的发生。

3.完成实验报告：做完上机实验，学员要按照实验报告的格式要求写出实验报告。实验报告的内容为：实验名称、实验的目的、实验的内容和要求、实验步骤（工作流及组件界面配置截图）、实验中遇到的问题与个人总结。

目录

实验一　大数据分析平台——BDAP 使用说明

（一）实验目的

1 熟悉 BDAP 平台的基本使用

（二）实验器材与实验准备

1．实验器材

硬件：微机一台

软件：Chrome 浏览器

2．实验准备：

(1) 下载 hosts 文件

(2) 阅读附录（BDAP 使用说明书）中相关使用说明。

（三）实验内容和要求

1. 修改 hosts
2. 登录、注销
3. 本地文件系统与分布式文件系统交互
4. 新建工作流：拖拽组件图标到工作流面板，双击配置，左键单击，拖拽指定连接顺序。
5. 保存、打开、插入、切换、清空工作流。
6. 配置元数据、HDFS 文件系统查看。
7. 运行工作流，通过“输出”组件查看结果。

（四）实验步骤

1. 修改 hosts 文件

将给定的 hosts 文件用记事本打开，将其复制后追加到本地 hosts 文件尾部。

（Windows 系统下，路径为：C:\Windows\System32\drivers\etc\hosts）

（linux 系统下，路径为：/usr/etc/hosts）

2. 登录、注销

（1） 用浏览器访问“mirageX：1337” ip 地址，进入 BDAP 登陆界面;

X 取值为 2,5,8,11,14；各个小班根据助教指示输入正确;

比如助教安排 2 班使用 mirage2,则 2 班同学输入 mirage2:1337

（2） 根据学号尾数后三位%20，得到对应用户账号，密码统一为 student

正确输入后点击“登录”按钮，即可登录成功。

例：学号尾数为 123 对应的账号为 student3,密码为 student

（3）点击右上角进行注销操作

3. 本地文件系统与分布式文件系统交互

（1） 本地文件上传到 HDFS（截图）

BDAP 登录成功后

->鼠标左键单击“数据交换”模块，跳转数据交换界面

->单击展开“数据导入”菜单栏

->左键单击“本地导入 FS”按钮,弹出上传至 FS 配置弹框

->配置完（图 1-1）后，点击“确定”按钮，导入成功后会有弹框提示（图 2-1）。

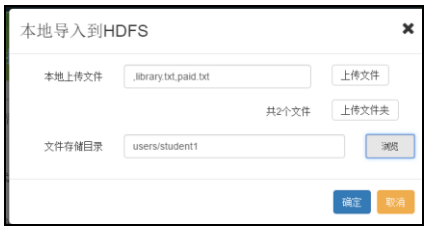

图 1-1

图 2-1

（2） 从 HDFS 下载文件到本地（截图）

BDAP 登录成功

->鼠标左键单击“数据交换”模块，跳转数据交换界面

->单击展开“数据导出”菜单栏

->左键单击“FS 导出本地”按钮,弹出 FS 下载本地配置弹框

->单击“浏览”按钮，弹出 HDFS 目录树（图 1-3），单击选中指定文件，点击“确定”按钮，会有弹框提示（图 1-4）。

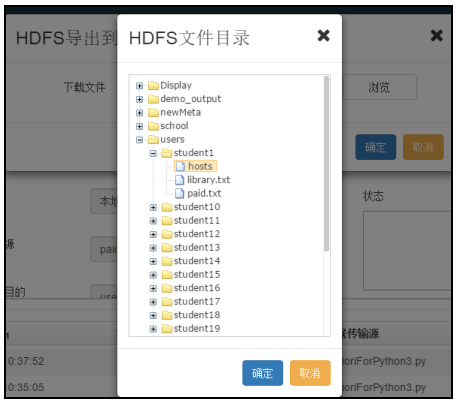

图 1-3

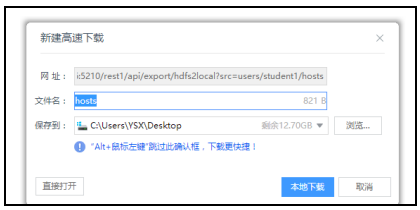

图 1-4

注意设置浏览器允许弹出框。

4. 新建工作流：拖拽组件图标到工作流面板，双击配置，左键单击，拖拽指定连接顺序。

(1) 点击“数据挖掘”按钮，进入数据挖掘界面（图 1-5）。

图 1-5

(2) 点击“清空界面”按钮，清空界面（画布）

图 1-6

(3) 单击左侧组件菜单栏，展开组件列表（图 1-7）；

图 1-7

从左端组件区通过鼠标左键拖动组件到右边画布；（图 1-8~图 1-9）

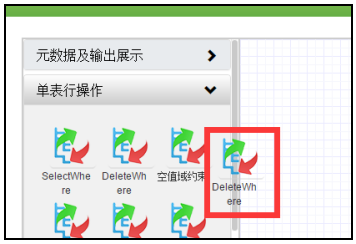

图 1-8

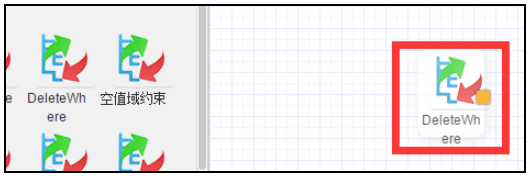

图 1-9

左键双击组件打开其配置界面；（图 1-10）

图 1-10

在配置界面，按界面提示配置完成后，点击“保存”按钮，则参数保存(图 1-11);点击“取消”则不更新变动，切换回工作流画布。

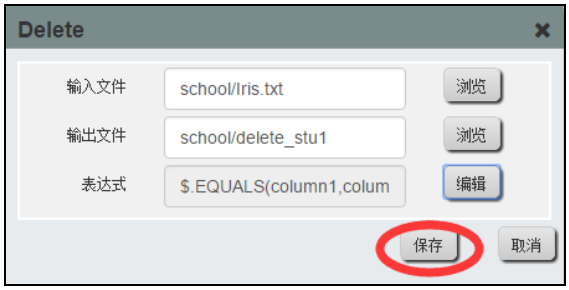

图 1-11

鼠标右键单击组件然后点击“删除”即可删除该组件；(图 1-12)

图 1-12

(4) 组件图标右侧黄色区域是用于建立连接线的锚点；(图 1-13)

图 1-13

当组件 A“isSource”为 ture 时，允许其作为源端通过鼠标点击锚点后长按引出有向连接线，当鼠标移动到组件 B 时，视为发起一次 A->B 连接请求，当 B 符合条件时，可以在画布中看到 A、B 间有一条从 A 到 B 的连接线；（图 1-14）

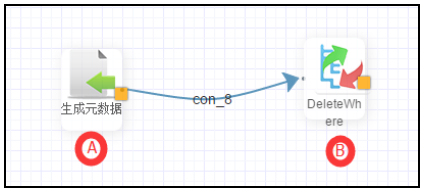

图 1-14

当组件“isDestintion”为 true 时，允许其被其他组件连接；

A-->B 有向连接表示，组件 A 先运行，A 运行完成后再去运行 B；

双击连接线，即可删除该连接线；（图 1-15）

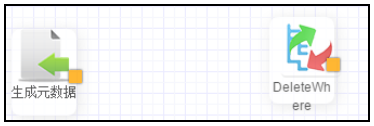

图 1-15

5. 保存、打开、插入、切换、清空工作流。

(1) 点击“保存工作流”按钮（图 1-16），输入工作流名称（图 1-17）；

注意：工作流内容不能为空、同名工作流会直接覆盖、工作流保存成功会有弹框提示（图 1-18）；

图 1-16

图 1-17

图 1-18

(2) 点击“打开工作流”按钮（图 1-19），选中指定工作流，点击“打开”按钮（图 1-20），即可清空工作流面板，打开已保存的工作流（图 1-21）；

图 1-19

图 1-20

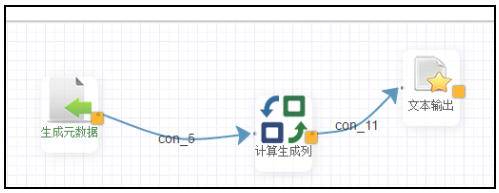

图 1-21

(3) 点击“查看工作流”按钮，选中指定工作流，点击“插入”按钮，即可向现有工作流面板中插入该工作流；

图 1-22

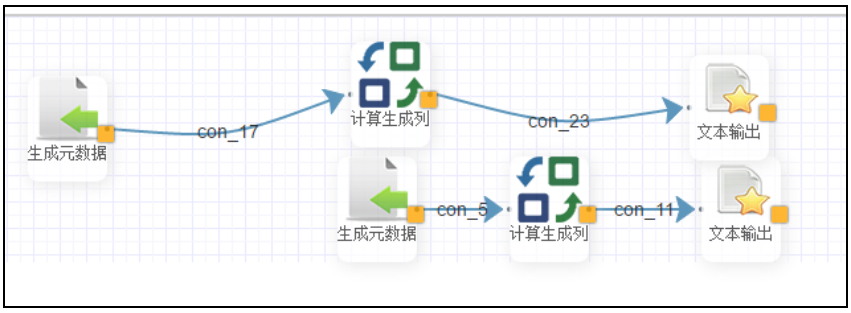

图 1-23

6. 配置元数据

所谓“元数据”，指的是用于描述 HDFS 文件的数据结构信息。

大多情况下，HDFS 文件中一行视为一个结构体，该行经过“分隔符”分隔，得到各个列的字符串格式的数据，再根据列名、列类型的描述信息，转化后用于组件的运算。

所以在使用大多数组件之前，我们要确保该组件输入文件对应的元数据存在，指明分隔符、列名、列类型（通过“统计”组件，我们可以向元数据中添加统计信息）。“生成元数据”组件可以从无到有创建一个新的元数据，也可以修改该组件配置页面的元数据信息，单击“保存”，重新覆盖该元数据信息。

平台组件在其配置面板配置完后，单击该面板的“保存”按钮，会根据输入文件的元数据和当前面板的配置信息去生成输出文件的元数据信息。

(1) 清空画布，拖动“生成元数据”组件到工作流面板

(2) 左键双组件打开配置面板；

(3) 配置面板中通过“浏览”按钮指定输入文件；（图 1-24）

图 1-24

(4) 配置面板中手输分隔符（注意：空格、制表符特殊字符可通过下拉框指定设置）（图 1-25）

图 1-25

(5) 通过检查，即可获得元数据检测到的列名，与列类型；（图 1-26）


源路径
ysxHF/train.csv
浏览
分隔符
,
逗号
检查
1
不带表头
带表头
2
输入元数据信息
3
name
type
column1
integer
column2
nominal
column3
nominal
column4
string
column5
nominal

图 1-26

(6) 可以人为去修改列名、列类型；（图 1-27，图 1-28）

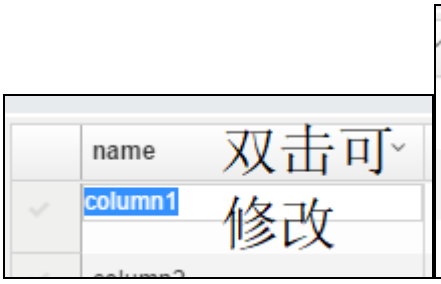

图 1-27

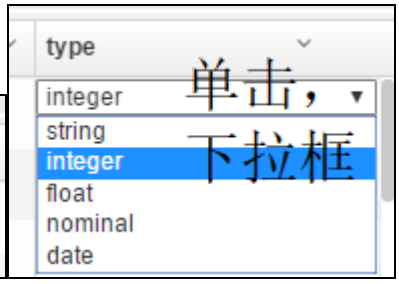

图 1-28

(7) 默认检查类型是“不带表头”，则列名以“column1，column2...”形式命名;(图 1-26)

(8) “带表头”，意味着对输入文件 A.txt.,必须要有 A.txt.title 文件位于 A.txt 所在 HDFS 路径下，且 A.txt.title 为 1 行以分隔符连接的列名字符串；（图 1-29）
举例：三列，列名为 cx,cy,cz,分隔符为”,”，则 A.txt.title 内容为：cx,cy,cz

图 1-29

(9) 通过点击“保存”按钮，即可将本页面的元数据信息覆盖原有元数据，也就是对默认检测得到的元数据进行手动修改后提交到 HDFS。

7. HDFS 文件系统

(1) 单击“数据管理”按钮，显示“文件系统”界面，左键双击即可展开或收起目录树；（图 1-30，图 1-31）

图 1-30

(2)　　左键单击文件，选中该文件；（图 1-31）

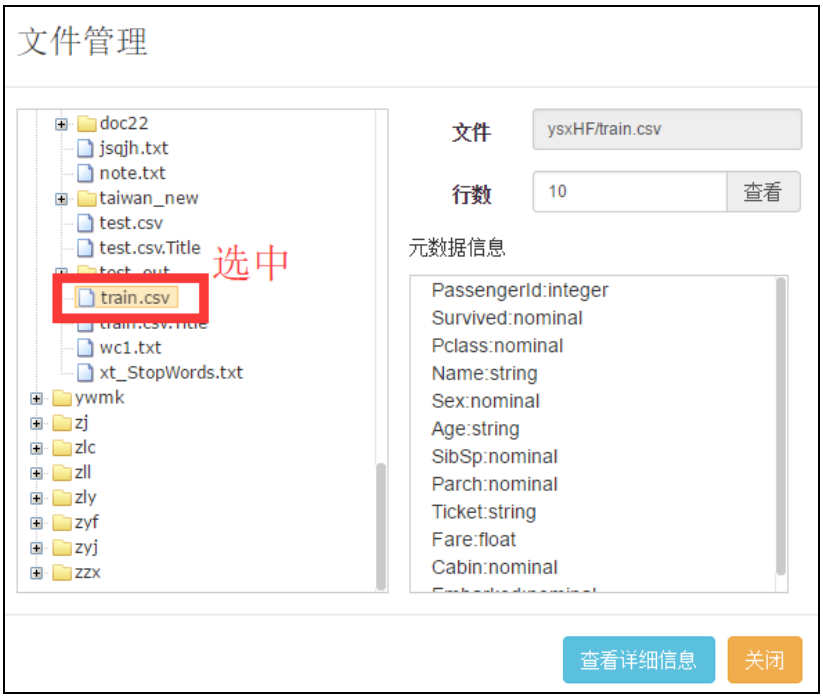

图 1-31

(3)　　单击“查看”按钮（图 1-32），即可查看“行数”对应的文件内容（图 1-33）；

图 1-32

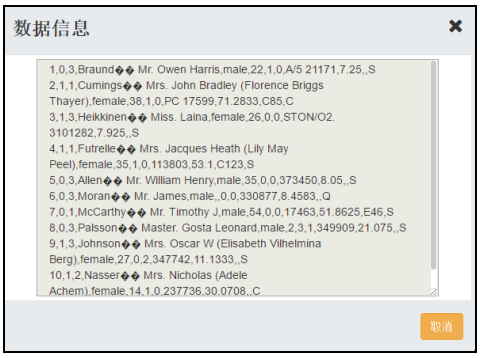

图 1-33

(4)　　若已为该文件配置过“元数据”，则可通过“查看详细信息”按钮（图 1-34）查看对应的元数据信息（图 1-35）；

图 1-34

图 1-35

8. 打开示例工作流 demo_etl_generateColumns（图 1-36，图 1-37）

“计算生成列”组件配置面板如图 1-38 所示，新列（name:“newCol”,type:“String”,epxr:“$.CONCAT(column1,column2)”），即 column1,column2 列的元素进行字符串拼接操作追加为新列；

=>鼠标右键单击组件然后点击“从此处开始运行”（图 1-39）；

图 1-40，图 1-41 为 spark 任务运行状态截图（Running –> Finished）

=>通过“文本输出”组件查看结果（图 1-42）；

图 1-36

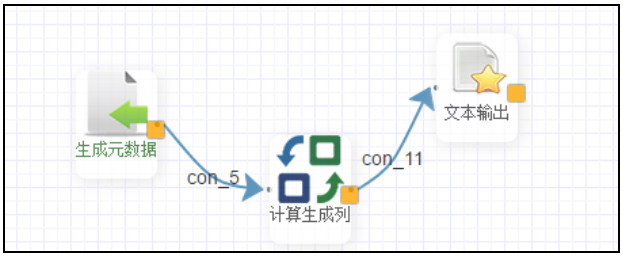

图 1-37

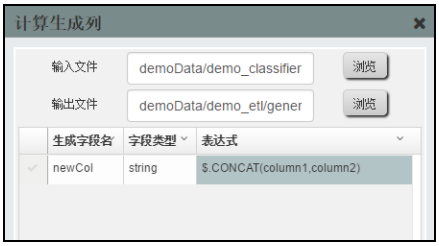

图 1-38

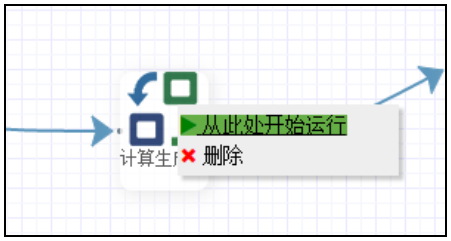

图 1-39

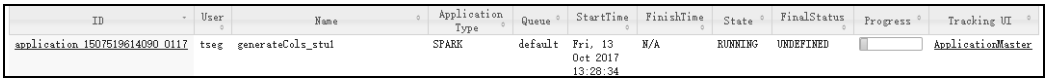

图 1-40

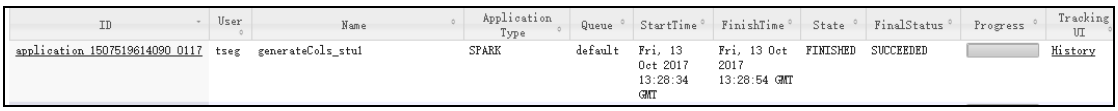

图 1-41

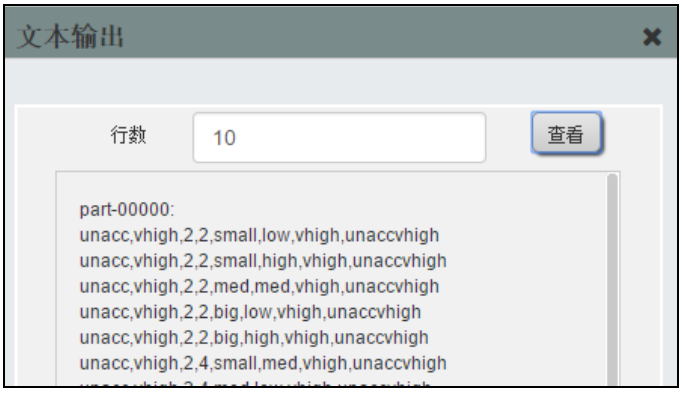

图 1-42

实验二　BDAP 常见的数据清洗操作

（一）实验目的

1 熟悉 BDAP 平台的常见的数据清洗操作。

（二）实验器材与实验准备

1．实验器材

硬件：微机一台

软件：Chrome 浏览器

2． 实验准备:

（1） 阅读附录（BDAP 使用说明书）中相关使用说明。

（三）实验内容和要求

1. 熟悉 BDAP 平台常见的清洗类组件，比如“列投影”,“casewhen”,“threeClean”,“字符数值化”,“归一化”,“统计”,“计算生成列”,“空值域约束”、“groupBy”等。

（四）实验步骤

1. 查看示例工作流（不用运行，之前已运行过，有输出结果）。

2. 任选其二，截图实验结果（输入文件片段、组件配置面板信息、输出组件显示信息），并简要用语言描述数据清洗的过程，工作流挖掘算法（分类、聚类、关联规则）不用描述。

3. 参考示例工作流，搭建自己的数据清洗工作流。注意：组件的输出路径均要重新制定为自己目录下的根路径。比如用 mirage5:1337 登录 student1 的同学（学号后三位为 123）的组件 caseWhen 输出应当放在前缀为“student1_121/”的路径下，可以命名为“student1_121/123/caseWhen_123”。截图，并用语言描述清洗过程。

（五） 提示说明

输入文件为 1.csv。6 列：序号,学校名称,主管部门,所在地,办学层次,备注。
文件内容示例如图 2-1
若以逗号作为分隔符，合法的行切分后应当有 7 列。
我们的清洗目标是：剔除不合法的行；
剔除为空的第 7 列；
对于合法的行，我们将其“备注”列的空值填充为“公立”；

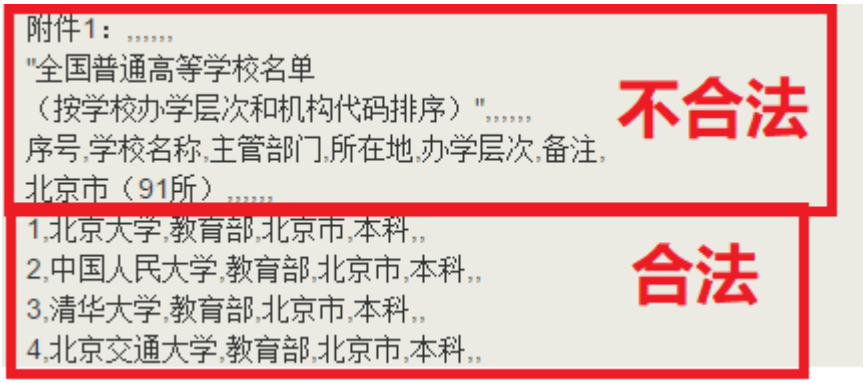

图 2 -1

分析：
(1) 一行是否合法可以根据其第一列的取值类型来判断：为数值型则合法，否则不合法；
(2) 可以用“过滤类型检查及去重”组件的类型检查功能检查第一列是否为数值型；
(3) 在使用(2)之前，必须确保输入文件的元数据设置为 7 列，且第一列类型为 integer 类型；
(4) 可以通过“空值域约束”组件对第 6 列（“备注”列）进行非空约束，设置替换值为“公立”。
(5) 可以通过“列投影”组件，保留前 6 列。
示例工作流为 demo_etl_example3kzy,如图 2-2 所示：
各组件配置面板如图 2-3 ~ 图 2-6 所示。

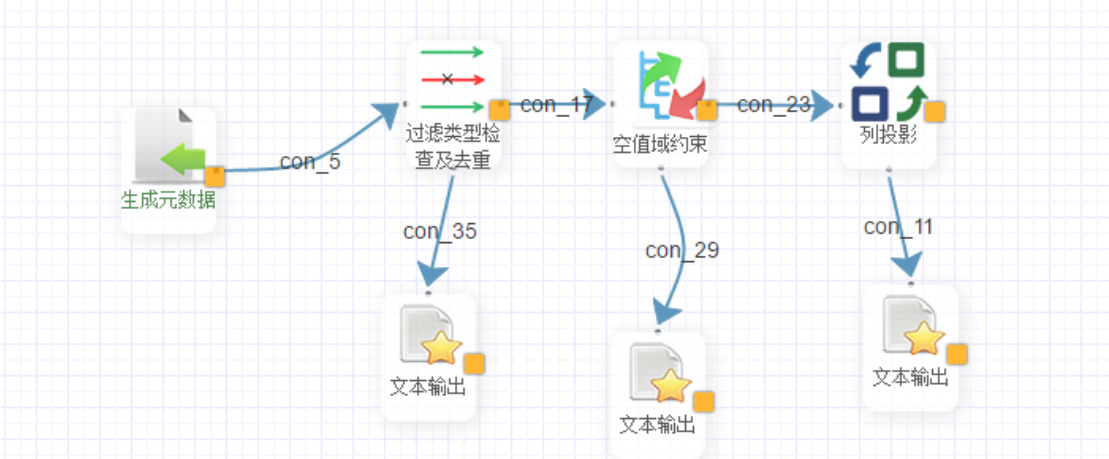

图 2-2

图 2-3

图 2-4

图 2-5

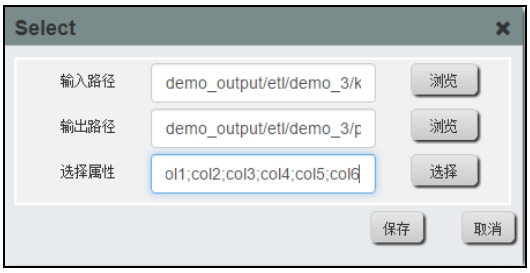

图 2-6

实验三　BDAP 之分类算法实验

（一）实验目的

1．学习 BDAP 平台的分类算法组件的使用方式;
2．通过平台对分类算法的输入、输出有个略直观的初步认知;

（二）实验器材与实验准备

1．实验器材
硬件：微机一台
软件：Chrome 浏览器‘

2．实验准备:
(1) 阅读附录（BDAP 使用说明书）中相关分类算法组件使用说明;
(2) 自行去初步了解一到两个分类算法原理（无需深究其背后的数学原理);

（三）实验内容和要求

1. 查看示例 demo,理解工作流组件面板如何配置
2. 自己选择自认为更合适的训练和预测参数，运行工作流，查看输出结果，并用自己的语言去分析。

（四）实验步骤

1. 选择一个示例 demo 工作流
2. 阅读工作流原始输入文件格式说明
3. 查看清洗组件的配置面板，明晓分类算法组件的输入数据格式。
4. 根据个人主观判断，选择合适的数据及训练属性来配置工作流（截图），运行，查看结果（截图），并做出分析

（五）提示说明

以“demo_lineR_test”工作流为例（图 3-1），通过清洗手段得到规范的训练集与测试集；“线性回归”是先猜一个线性模型，y = a1x1+a2x2+...anxn+b;

通过训练数据(X1,y1),(X2,y2),...来求出（a1,a2,...an）的值，及得到一个线性回归模型；

通过带标签的测试数据去评估模型的好坏；

可以通过模型去预测不带标签数据的标签值；

“线性回归”组件 A 用于训练模型，配置面板如图 3-2 所示；

“线性回归”组件 B 用于测试模型，配置面板如图 3-3 所示；

“线性回归”组件 B 的测试结果如图 3-4 所示：

训练方差 = 0.4478650026615373 ，比较大

part-00000:

id，标签值，预测标签值

```
457,  0.0,    0.05856841371254419
474,  1.0,    0.03042402980788308
513,  1.0,    0.05798934747527328
```

结论： 单拿 Fare 列作为训练属性得到模型效果不是很理想。

可以初步认为但从收入来看，收入的高低跟标签值存活率关系不大；

也有可能是数据清洗过程中丢失过多数据造成的；

也有可能是 Fare 列归一化之前，Fare 极值过高、过低导致归一化的效果不好导致的；

输入文件的格式为：

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
```

输入文件示例：

```
1,0,3,Braund_ Mr. Owen Harris,male,22,1,0,A/5 21171,7.25,,S
2,1,1,Cumings_ Mrs. John Bradley (Florence Briggs Thayer),female,38,1,0,PC 17599,71.2833,C85,C
```

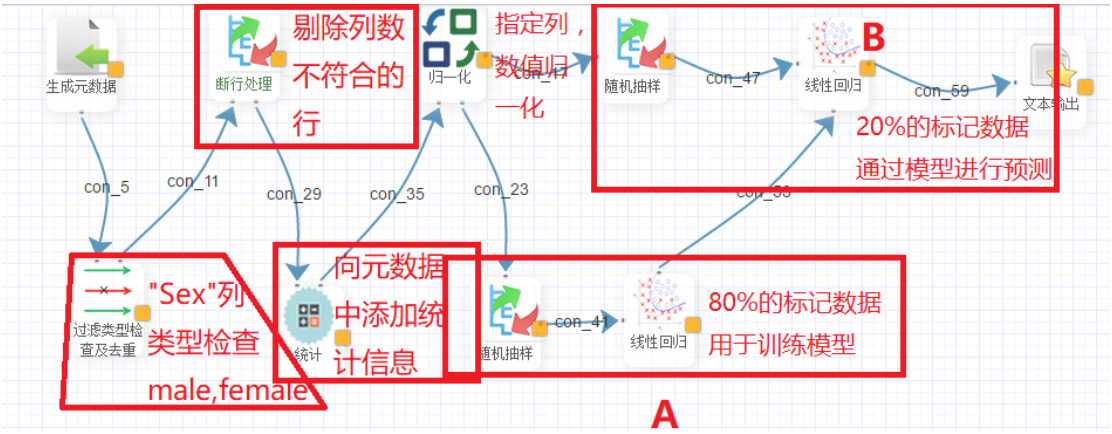

图 3-1

图 3-2

图 3-3

图 3-4

以“demo_lineR_predict”工作流为例（图 3-5），通过实验二的清洗手段得到规范的训练集与测试集；

“线性回归”组件 A 用于训练模型，配置面板如图 3-6 所示；

“线性回归”组件 B 用于模型预测，配置面板如图 3-7 所示；

“线性回归”组件 B 的测试结果如图 3-8 所示：

part-00000:

892,0.010328103888866734

893,0.009234241968792103

894,0.012779531296096215

可以根据训练模型去挑一个合适的阈值 a,大于 a 的标记为“1”，反之标记为“0”

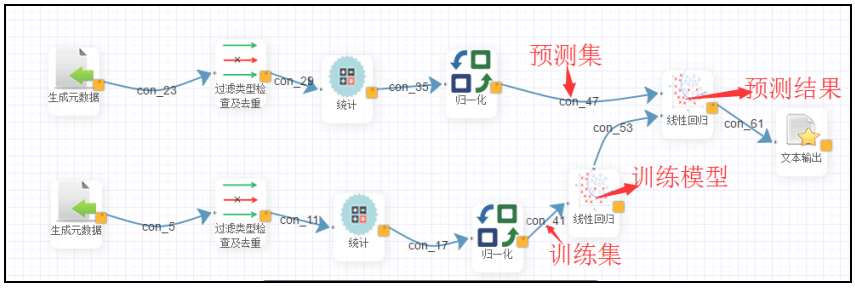

图 3-5

图 3-6

图 3-7

图 3-8

（六）实验结果分析

指明原始数据格式、清洗过程、训练集、测试集、模型评价。

实验四　BDAP 之聚类算法实验

（一）实验目的

1．学习 BDAP 平台的聚类算法组件的使用方式;
2．通过平台对分类算法的输入、输出有个略直观的初步认知;

（二）实验器材与实验准备

1．实验器材
硬件：微机一台
软件：Chrome 浏览器
3．实验准备:
(3) 阅读附录（BDAP 使用说明书）中相关聚类算法组件使用说明;
(4) 自行去初步了解一到两个聚类算法原理（无需深究其背后的数学原理);

（三）实验内容和要求

3. 查看示例 demo,理解工作流组件面板如何配置
4. 自己选择自认为更合适的训练和预测参数，运行工作流，查看输出结果，并用自己的语言去分析。

（四）实验步骤

1. 选择一个示例 demo 工作流
2. 阅读工作流原始输入文件格式说明
3. 查看清洗组件的配置面板，明晓聚类算法组件的输入数据格式。
4. 根据个人主观判断，选择合适的数据及训练属性来配置工作流（截图），运行，查看结果（截图），并做出分析

（五）实验结果分析

指明原始数据格式、清洗过程、训练集、测试集、模型评价。

实验五　BDAP 之关联规则算法实验

（一）实验目的

1．学习 BDAP 平台的关联规则算法组件的使用方式;
2．通过平台对关联规则算法的输入、输出有个略直观的初步认知;

（二）实验器材与实验准备

1．实验器材
硬件：微机一台
软件：Chrome 浏览器
4．实验准备:
(5) 阅读附录（BDAP 使用说明书）中相关分类算法组件使用说明;
(6) 自行去初步了解一到两个分类算法原理（无需深究其背后的数学原理);

（三）实验内容和要求

5. 查看示例 demo,理解工作流组件面板如何配置
6. 自己选择自认为更合适的训练和预测参数，运行工作流，查看输出结果，并用自己的语言去分析。

（四）实验步骤

1. 选择一个示例 demo 工作流
2. 阅读工作流原始输入文件格式说明
3. 查看清洗组件的配置面板，明晓关联规则算法组件的输入数据格式。
4. 根据个人主观判断，选择合适的数据及训练属性来配置工作流（截图），运行，查看结果（截图），并做出分析

（五）实验结果分析

指明原始数据格式、清洗过程、训练集、测试集、模型评价。

实验六　BDAP——开放性数据挖掘实验（选做）

（一）实验目的

1　初步掌握 BDAP 使用方法后，运用平台组件来挖掘自己感兴趣的数据

（二）实验器材与实验准备

1．实验器材

硬件：微机一台

软件：Chrome 浏览器

2．实验准备：

（1）阅读附录（BDAP 使用说明书）中相关使用说明；

（2）熟悉所给“智慧”校园数据格式或中诚信征信 2017 首届风云杯建模大赛数据集（也可以选择自己感兴趣的数据，通过 BDAP 上传到平台 HDFS）；

（3）自己根据所了解的数据挖掘相关知识和 BDAP 现有工具来制定挖掘策略；

（三）实验内容和要求

1. 选择自己感兴趣的原始数据，制定可选的挖掘策略，做出自己的分析判断。

（四）实验步骤

1. 选择自己感兴趣的原始数据(若非提供的数据，需自行上传到 HDFS);
2. 根据自己制定的挖掘策略来搭建工作流；
3. 运行工作流、记录实验结果；
4. 分析总结；

（五）实验结果分析

实验涉及的数据介绍？自己的猜想？工作流设计思路？结果与猜想对比分析？