Hadoop 测试工具

用户手册

上海心河信息技术有限公司 2018 年 11 月 21 日

目 录

1	系	《统说明	1
2	系	《 统初始化	2
	2.1	主界面	2
	2.2	注册界面	2
	2.3	登录界面	3
	2.4	后台接口配置(仅限管理员)	7
	2.5	注销	8
3	Ą	[}] 入性能测试	9
	3.1	上传文件到服务器文件系统	9
	3.2	上传文件到 HDFS	11
4	数	女据预处理评估	13
	4.1	创建预处理任务	13
	4.2	预处理结果评估	15
	4.3	直接评估	17
	4.4	评估结果	17
5	第	淳法性能测试	19
	5.1	创建任务	19
	5.2	查看结果	21
6	狈	JI试报告管理	27
7	典	中型应用	28
8	用	月户管理(仅限管理员)	30
9	F	引志(仅限管理员)	33

1 系统说明

本平台用于对使用 Hadoop、Spark 等技术搭建的大数据集群环境进行测试,除了常规的性能测试外,还提供了对集群算法的评估方案,通过跟踪集群算法的运行过程评估算法效率,并能够对一些常规类型的算法结果进行评价。

系统主要由导入性能测试、数据预处理评估、算法性能测试、测试报告管理、 典型应用、用户管理和日志七个部分组成。

本手册旨在介绍 Hadoop 测试工具的各项功能、业务流程,用于指导操作人员正确使用。

2 系统初始化

2.1 主界面

打开主界面显示用户欢迎页面(如图 2-1 所示),右上角有"登录"和"注册"按钮。页面底部有关于我们、友情链接及版权信息等内容。

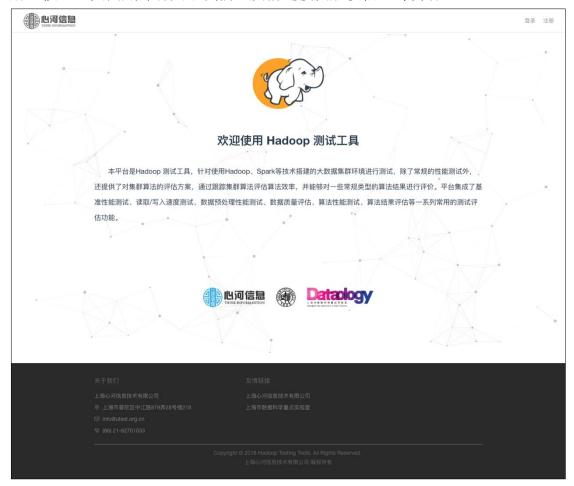


图 2-1 主界面

2.2 注册界面

要使用该系统首先要进行注册,点击页面右上角的"注册"按钮即可跳转到注册页面,如图 2-2 所示。

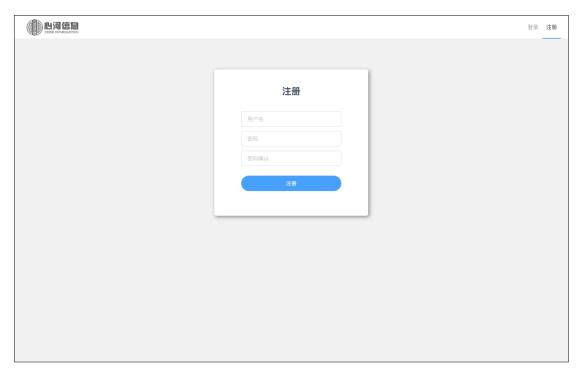


图 2-2 注册页面

用户填写信息不完整或两次密码不一致时,点击"注册"按钮会有相应的错误提示,如图 2-3 所示。

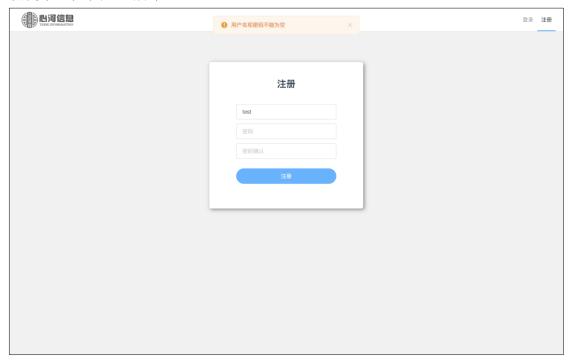


图 2-3 注册错误提示

2.3 登录界面

注册成功后,系统会自动跳转到登录界面。用户输入正确的"用户名"和"密

码"即可登录。登录时输入不完整或用户名、密码不正确会有相应的错误提示,如图 2-4 所示。"记住我"复选框勾选时,可以保持登录状态 7 天,取消勾选将在 8 小时后注销登录状态。

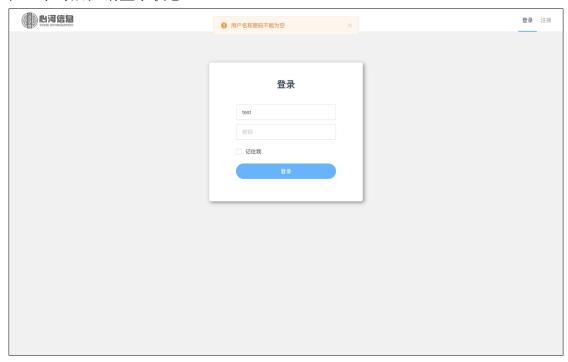


图 2-4 登录错误提示

用户登录成功后的主界面如图 2-5 所示,与未登录时的主界面不同,此时可以在导航栏看到"首页","导入性能测试","数据预处理评估","算法性能测试","测试报告管理","典型应用"六个功能项,右上角显示当前用户的用户名,点击用户名有"注销"选项。相应地,页面底部也会多出"应用案例"和"使用说明"两项功能。

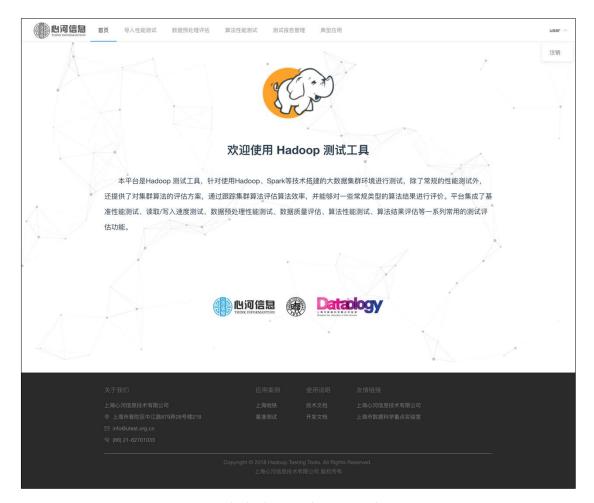


图 2-5 角色为"用户"的用户首页

而且不同角色的用户主界面也有差别,前面介绍的是角色为"用户"的主界面,角色为"管理员"的主界面如图 2-6 所示,导航栏会多出"用户管理"、"日志"两个功能项,点击用户名也多出了"配置"功能项,相应地页面底部也多出了"我的账户"一列,三个选项分别对应上述三个管理员专属功能。

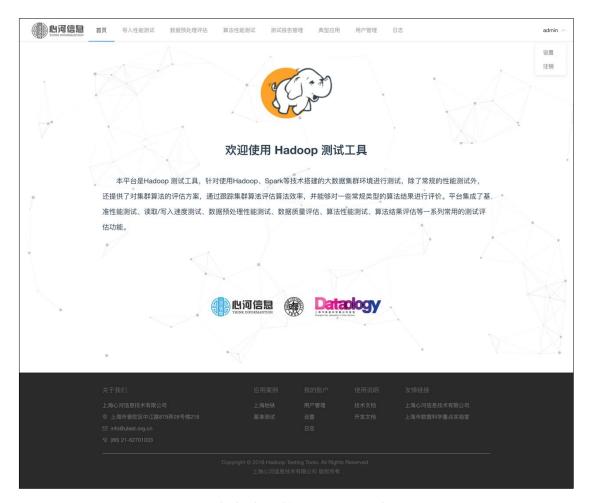


图 2-6 角色为"管理员"的用户首页

角色为"演示"的主界面如图 2-7 所示,和角色为"用户"的界面相同,但系统内各项功能只有查看权限,不能够执行创建任务和删除等操作。新用户注册时为该角色,角色变更操作见第 8 小节中的说明。

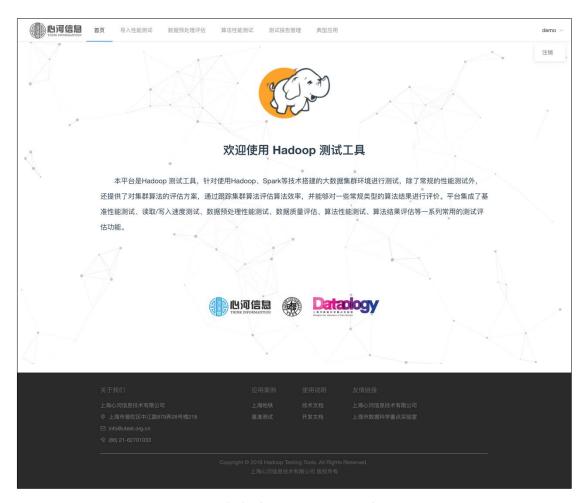


图 2-7 角色为"演示"的用户首页

2.4 后台接口配置(仅限管理员)

角色为"管理员"的用户点击如图 2-6 右上角的用户名,再次点击下拉菜单中的"设置"按钮,即可跳转到后台接口配置界面,如图 2-8 所示的。用户可以配置"Web 后台 IP","集群 Master IP","Spark 端口","HDFS 端口","HDFS 目录","集群 Master 用户名","集群 Master 密码","spark-submit 路径","Hadoop路径","Benchmark jar 包路径","Python 路径"共 11 项接口参数,输入正确的内容,点击"保存"按钮即可成功配置。



图 2-8 后台接口配置

2.5 注销

任何用户点击页面右上角的用户名,再次点击下拉菜单中的"注销"按钮即可退出登录。

3 导入性能测试

本功能下,用户可以将自己的文件(包括数据文件和算法包等)上传到 Web 后台服务器文件系统,进而把文件上传到 HDFS,后面这一过程会记录到耗时和上传速度,用以评估 HDFS 导入数据的性能。

3.1 上传文件到服务器文件系统

用户点击导航栏的"导入性能测试",出现如图 3-1 所示界面。用户可以点击特定文件后面的"上传到 HDFS"按钮,将该文件上传到 HDFS,点击"下载"按钮将文件下载到本地,或者点击"删除"按钮将文件从 Web 后台服务器删除。

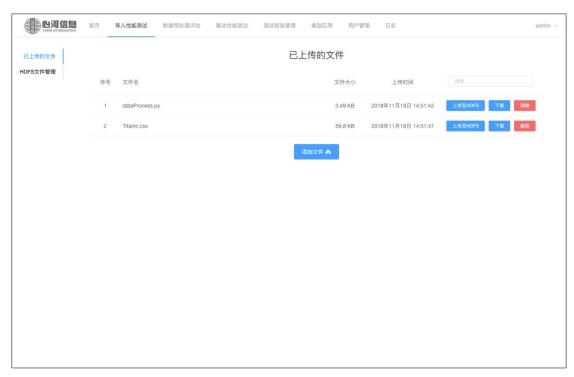


图 3-1 已上传的文件界面

点击图 3-1 中文件列表下方的"添加文件"按钮会出现图 3-2 所示的文件上传框,点击虚线框或者将文件拖进虚线框即可选择要上传的文件,之后点击"上传到服务器"按钮立即开始上传,如图 3-3 所示。支持多文件同时上传,单文件最大 800MB。文件上传成功后,列表会自动刷新,如图 3-4 所示。

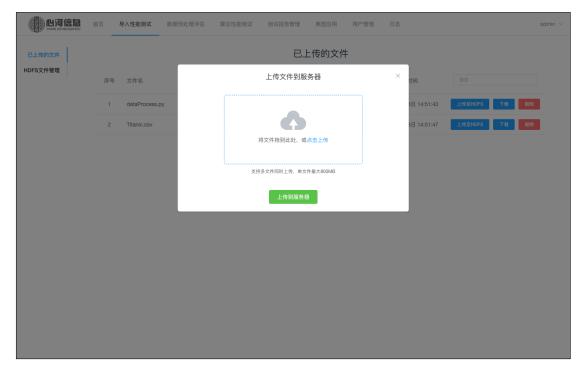


图 3-2 文件上传框

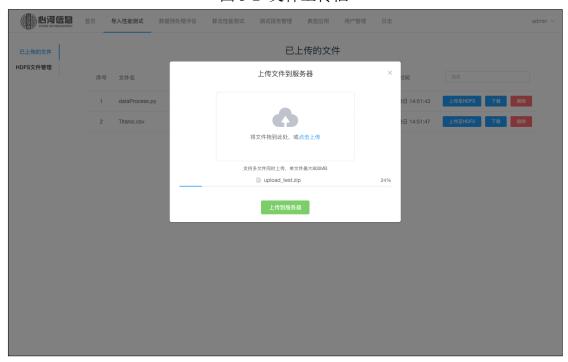


图 3-3 上传文件到 Web 服务器

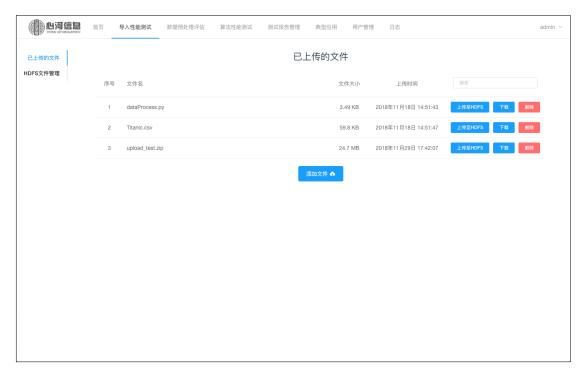


图 3-4 上传完成文件列表自动刷新

3.2 上传文件到 HDFS

用户可以在"已上传的文件"标签页下选择文件上传到 HDFS,上传成功后会提示"上传成功!",如图 3-5 所示:

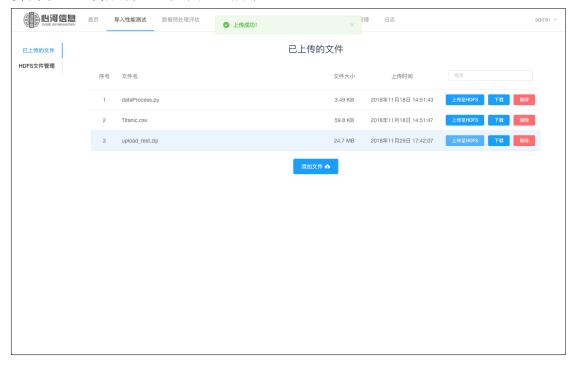


图 3-5 成功上传文件到 HDFS

用户点击左侧"HDFS 文件管理"即可查看已上传到 HDFS 的文件,如图 3-6

所示,在已上传的文件列表可以查看上传该文件的耗时以及上传速度,还可以点击"删除"按钮将特定文件从 HDFS 删除,点击"下载"按钮获取文件到本地。

页面底部以图表的形式显示 HDFS 的总空间、不可用空间、已用空间和剩余空间等信息。

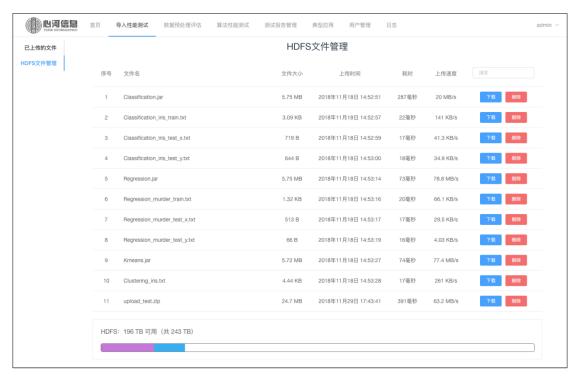


图 3-6 HDFS 文件管理页面

4 数据预处理评估

本功能部分,用户可以创建预处理任务对数据进行预处理,评估数据预处理 效率,评估数据质量。

点击"数据预处理评估"时,出现如图 4-1 所示的界面。

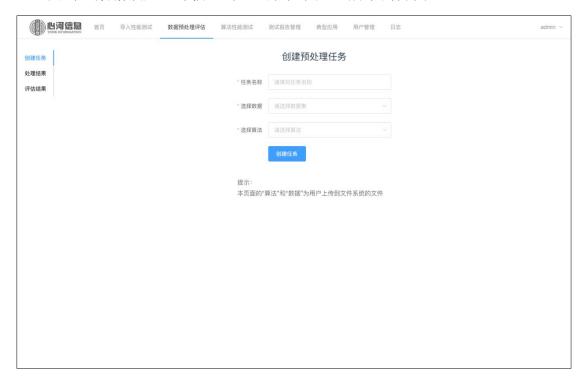


图 4-1 数据预处理评估界面

4.1 创建预处理任务

用户输入任务名称,选择数据和算法之后,点击"创建任务"按钮,可以创建数据预处理任务。用户输入不完整时,会提示错误信息,如图 4-2 所示。所有页面的表单都有完整性验证,后面不再赘述。



图 4-2 错误提示

数据预处理时的数据和算法为用户上传到 Web 后台服务器的文件,均在下 拉框中进行选择,如图 4-3 和图 4-4 所示。



图 4-3 选择预处理数据



图 4-4 选择预处理算法

4.2 预处理结果评估

点击左侧"处理结果"标签即可查看数据预处理任务,点击"下载"按钮可以获取预处理结果,"删除"按钮可以将特定任务删除。另外列表中还展示了预处理任务的耗时以及状态。

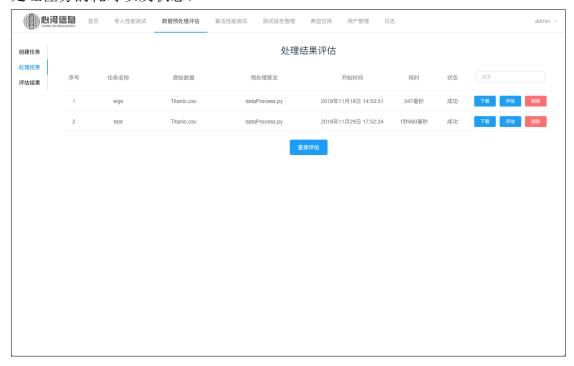


图 4-5 处理结果界面

当用户点击图 4-5 中的"评估"按钮时,弹出如图 4-6 所示对话框。用户需输入任务名称,并点击"添加规则"按钮添加规则,一个评估任务至少有一条规则,且支持多条规则,然后点击"创建任务"按钮即可创建评估任务。点击"重置"按钮将清空任务名称及所有规则。

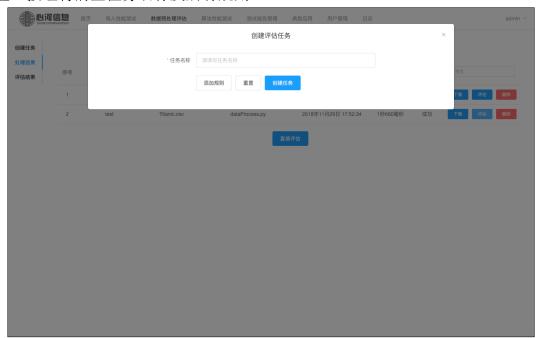


图 4-6 创建评估任务界面

一条评估规则是对数据中某一列的限制,列的序号从 0 开始,支持的规则有 =, <, <=, >, >=和 in,右侧的"删除"按钮可以删除对应的规则。



图 4-7 创建评估任务时添加规则

4.3 直接评估

除了可以对预处理结果文件进行数据质量评估之外,还支持直接对某一数据进行质量评估,通过点击图 4-5 结果列表底部的"直接评估"按钮呼出图 4-8 所示的直接评估弹出框,与前面不同的是,这里要选择一个要评估的数据,该数据也是来自于用户上传到 Web 后台服务器的文件。

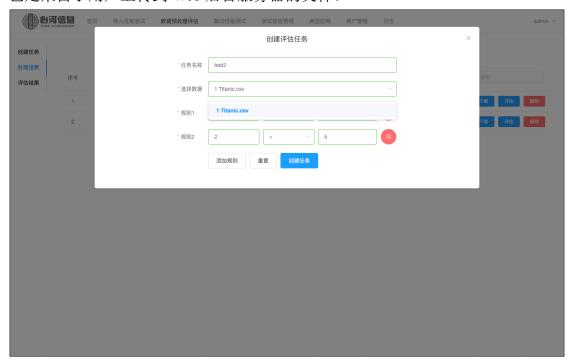


图 4-8 直接评估

4.4 评估结果

用户在 4.2 或 4.3 中创建评估任务之后,点击左侧的"评估结果"标签可以看到评估结果。用户可以看到评估任务的状态以及耗时。点击"查看评估结果"按钮出现图 4-10 所示的结果详情弹出框,点击"删除"按钮可以删除对应的评估任务。



图 4-9 评估结果查看界面

图 4-10 所示为评估结果详情弹出框,其中的一行信息对应创建评估任务时的一条规则结果(包括行数和比例),最后还有所有规则累计的最终结果。



图 4-10 评估结果

5 算法性能测试

本功能主要用于测试集群算法性能,以及评估算法结果。无监督和有监督的算法创建测试任务有一些差别,无监督的算法可以直接从原始数据得出结果,而有监督的学习要经过模型训练、结果预测两步才能得出结果。为了统一流程,对于有监督的算法过程,创建任务时直接指定训练集以及无标签的测试集,这样一来有监督的算法任务也可以得出结果。

5.1 创建任务

用户点击导航栏的"算法性能测试"时,出现如图 5-1 所示的界面。无监督的算法用户输入任务名称,选择算法和数据,如果所选的算法还有需要用户指定的参数,需要点击"添加参数"按钮添加一个或多个参数(如图 5-4),最后点击"创建任务"按钮便可提交任务。

注意: 我们提供的 demo 程序 KMeans 算法,需要添加一个参数 k 来指定聚类的簇数量,如图 5-4 中所示。

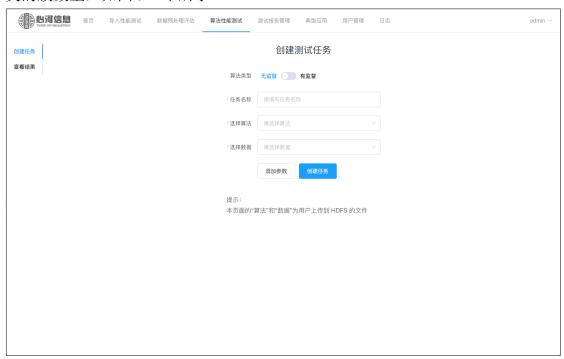


图 5-1 创建测试任务

创建测试任务时,所选算法和数据均为用户上传到 HDFS 的文件,如图 5-2 和图 5-3。



图 5-2 选择算法



图 5-3 选择数据



图 5-4 添加参数

对于有监督的算法要将"算法类型"选为"有监督",这时会多出一个"测试数据"选择框,如图 5-5 所示,用于选择无标签的测试数据。



图 5-5 有监督算法选择测试数据

5.2 查看结果

用户在 5.1 中创建任务之后,可以点击左侧"查看结果"标签来查看测试结

果。用户可以查看任务的状态以及耗时。点击"下载"按钮可以将算法结果下载到本地,"删除"按钮可以删除对应的任务,点击"评估"按钮会根据任务所用算法类型的不同,出现如图 5-7 (无监督)或图 5-8 (有监督)所示的算法结果评估对话框,选择和任务算法对应的评估方案(及数据),会将算法结果以可视化的形式展现出来,并且以表格的形式呈现相应的评价指标及结果。



图 5-6 测试结果界面

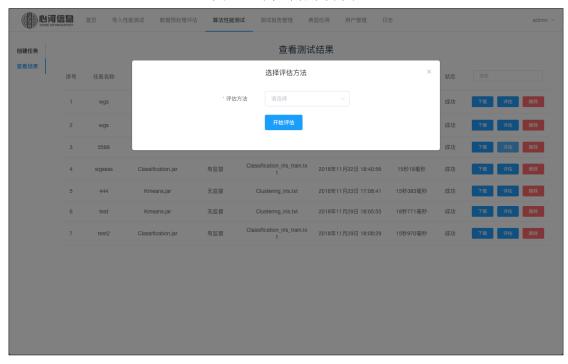


图 5-7 无监督算法结果评估方法选择

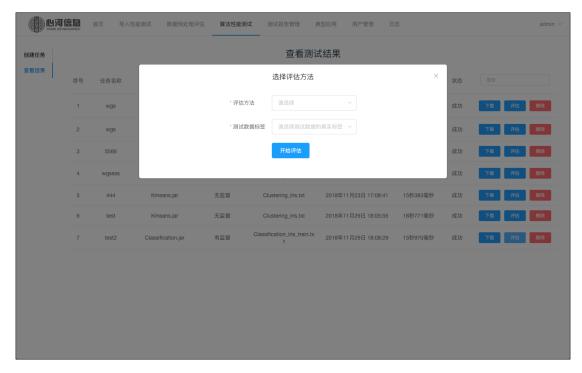


图 5-8 有监督算法结果评估方法和测试数据真实标签选择

当前平台支持三种算法的评价方法,其中无监督算法支持聚类(图 5-9),有监督算法支持回归和分类(图 5-11)。

评估无监督的算法结果只需选择对应的评估方法(图 5-9)。

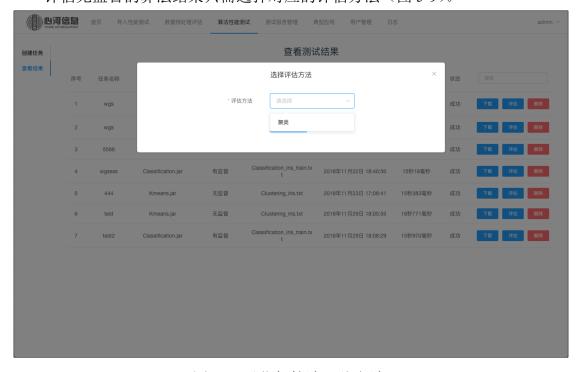


图 5-9 无监督算法评估方法

聚类算法的结果评价页面上方以可视化的形式展示每个簇及各自的簇中心点,下方的评价指标有 Compactness、Separation、Davies-Bouldin Index 和 Dunn

Validity Index o

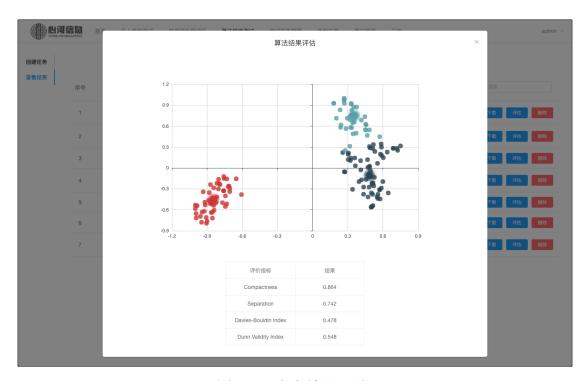


图 5-10 聚类结果评估

评估有监督的算法结果除了要选择对应的评估方法(图 5-11)外还需选择测试数据真实值或真实标签(图 5-12),用于与预测结果进行对比得出评估结果和指标。

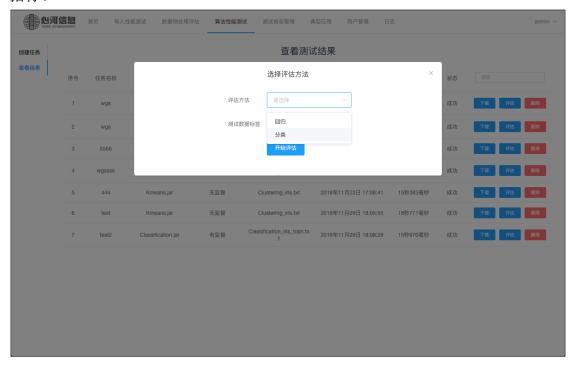


图 5-11 有监督算法评估方法选择



图 5-12 有监督算法测试数据真实标签选择

分类算法的结果评价页面上方以柱状图的形式展示各个类别预测正确/错误的数量和百分比,下方的评价指标有 Hamming loss。

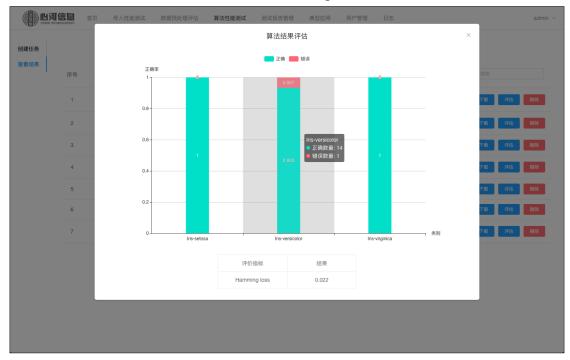


图 5-13 分类结果评估

回归算法的结果评价页面上方以折线图的形式展示测试数据的真实值和预测值,可以直观地观察出算法结果的好坏,下方的评价指标有 MAE、SSE、MSE、RMSE 和 R-Square。

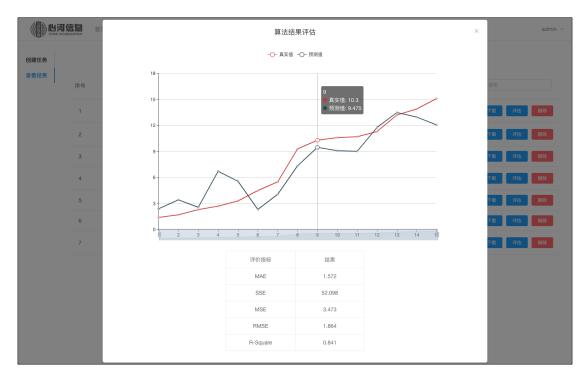


图 5-14 回归结果评估

6 测试报告管理

如图 6-1 所示,本模块是提供测试报告的保存和管理功能。具体操作方法请参照 3.1 中的介绍。

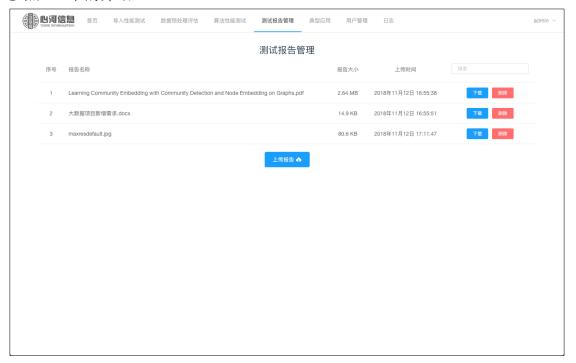


图 6-1 测试报告管理

7 典型应用

本部分用于展示一些典型应用,目前有上海地铁站点数据聚类和基准测试两项内容,如图 7-1 所示。

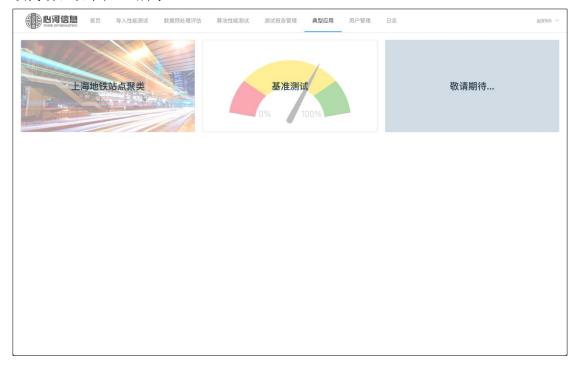


图 7-1 典型应用

上海地铁站点功能聚类,首先将客流数据用 LDA 的思想抽取站点的向量特征,然后通过聚类算法将站点分为多个不同功能的集合。如图 7-2 所示,本页面将这一过程的耗时情况,以及最终结果以图表的形式展示出来。点击表格下方的"重新计算"会重新执行这一过程,并将结果重新显示在页面上。由于数据量较大,特征提取算法复杂,执行一次要耗时 40 多秒。

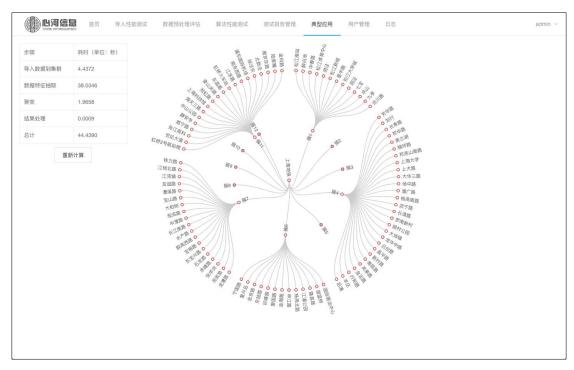


图 7-2 上海地铁站点聚类

基准测试是 Hadoop 官方 demo 里的一项测试内容,目的是测试集群环境的 I/O 性能,图 7-3 是这一测试的结果,点击界面下方的"重新测试"按钮将重新执行测试过程。

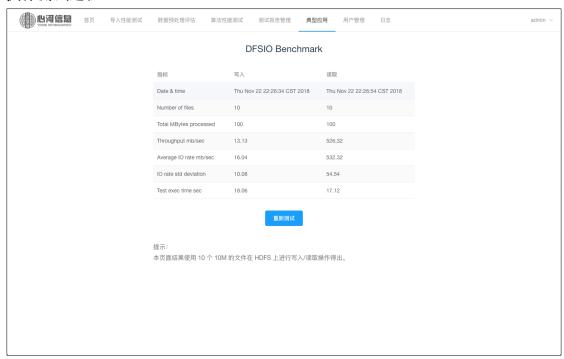


图 7-3 基准测试

8 用户管理(仅限管理员)

本平台的用户一共有三种角色——管理员、用户、演示。"用户"角色可以使用系统的普通功能;"管理员"角色除了用户的权限外还有"后台接口配置、用户管理、日志"三个高级页面;"演示"角色和"用户"角色能进入的页面一样,但所有页面都没有新增、删除等更改操作,只能查看和下载,如图 8-1 和图 8-2 所示。新注册的账号角色为"演示",权限最低。

已上传的文件 DFS文件管理	HDFS文件管理							
DISCHERE	序号	文件名	文件大小	上传时间	耗时	上传速度	投资	
	1	Classification.jar	5.75 MB	2018年11月18日 14:52:51	287毫秒	20 MB/s	下载	
	2	Classification_iris_train.txt	3.09 KB	2018年11月18日 14:52:57	22毫秒	141 KB/s	下载	
	3	Classification_iris_test_x.txt	719 B	2018年11月18日 14:52:59	17毫秒	41.3 KB/s	下载	
	4	Classification_iris_test_y.txt	644 B	2018年11月18日 14:53:00	18毫秒	34.9 KB/s	下虹	
	5	Regression.jar	5,75 MB	2018年11月18日 14:53:14	73毫秒	78.8 MB/s	下觀	
	6	Regression_murder_train.txt	1.32 KB	2018年11月18日 14:53:16	20毫秒	66.1 KB/s	下载	
	7	Regression_murder_test_x.txt	513 B	2018年11月18日 14:53:17	17毫秒	29.5 KB/s	下館	
	8	Regression_murder_test_y.txt	66 B	2018年11月18日 14:53:19	16毫秒	4.03 KB/s	下虹	
	9	Kmeans.jar	5.72 MB	2018年11月18日 14:53:27	74毫秒	77.4 MB/s	下概	
	10	Clustering_iris,txt	4.44 KB	2018年11月18日 14:53:28	17毫秒	261 KB/s	下载	
	LIDEO	:: 196 TB 可用 (共 243 TB)						

图 8-1 "演示"账户下的界面 1



图 8-2 "演示"账户下的界面 2

如图 8-3 为"管理员"角色的用户管理界面,"管理员"可以修改其他账户的角色以及删除账户,如图 8-4 所示。为了防止误操作,删除用户需要再次确认,如图 8-5。另外为了防止意外,系统禁止管理员修改自己的账户角色或删除自己的账户,要修改或删除某个管理员账户需使用另一个管理员账户进行操作。。

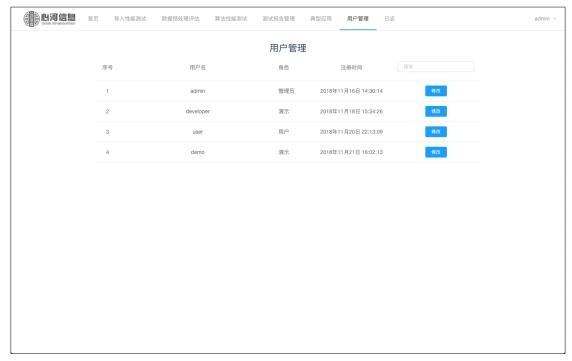


图 8-3 用户管理界面

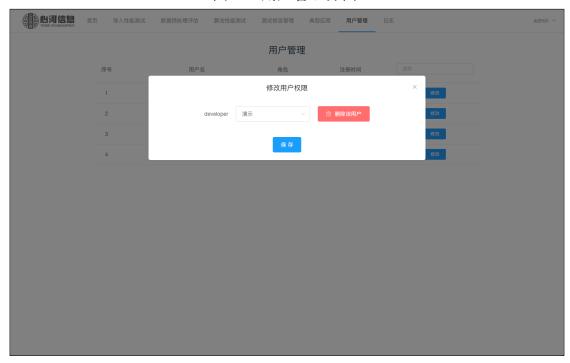


图 8-4 修改用户

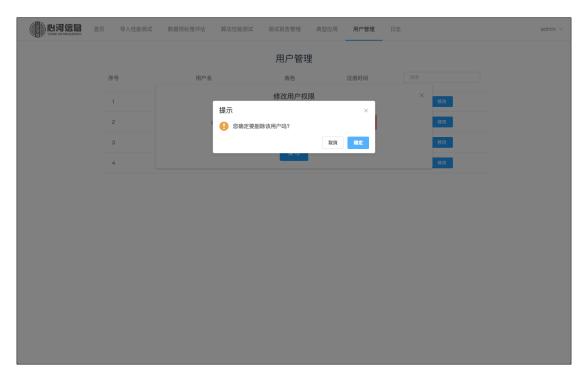


图 8-5 删除确认

9 日志(仅限管理员)

如图 9-1 所示,"管理员"角色可以查看平台的操作日志,便于排查问题。 点击列表下方的"清空日志"按钮可以清空系统日志。为了防止误操作,清空日 志需要再次确认,如图 9-2。

			系统日志		
序号	时间	操作人	类型	内容	夏索
61	2018年11月29日 17:29:09	admin	登录	-	
62	2018年11月29日 17:36:43	admin	登录	-	
63	2018年11月29日 17:42:06	admin	上传文件至文件系统	upload_test.zip	
64	2018年11月29日 17:43:41	admin	上传文件至 HDFS	upload_test.zip	
65	2018年11月29日 17:49:35	admin	删除 HDFS 文件	upload_test.zip	
66	2018年11月29日 17:49:42	admin	上传文件至 HDFS	upload_test.zip	
67	2018年11月29日 17:49:54	admin	删除 HDFS 文件	upload_test.zip	
68	2018年11月29日 17:50:24	admin	删除文件系统文件	upload_test.zip	
69	2018年11月29日 17:52:31	admin	创建数据预处理任务	test	
70	2018年11月29日 17:55:16	admin	创建预处理评估任务	test	
71	2018年11月29日 17:57:54	admin	创建预处理评估任务	test2	
72	2018年11月29日 17:59:15	admin	删除预处理评估任务	test2	
73	2018年11月29日 18:00:21	admin	创建预处理评估任务	test2	
74	2018年11月29日 18:05:53	admin	创建算法测试任务	test	
75	2018年11月29日 18:08:27	admin	创建算法测试任务	test2	

图 9-1 日志界面

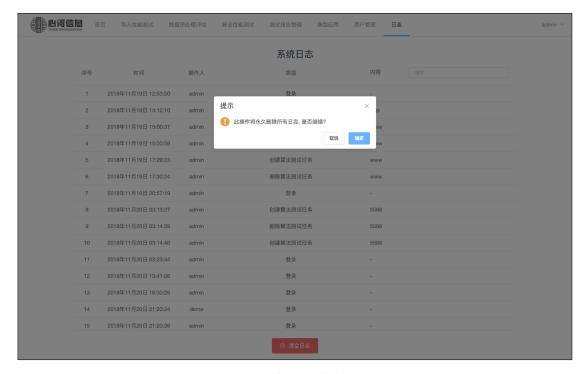


图 9-2 清空日志确认