Hadoop 测试工具

技术文档

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

本平台是 Hadoop 测试工具，针对使用 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。平台集成了基准性能测试、读取/写入速度测试、数据预处理性能测试、数据质量评估、算法性能测试、算法结果评估等一系列常用的测试评估功能。

2 整体架构

本系统完全基于 B/S 结构设计，兼容性好，方便部署，易于扩展和更新迭代，有着单点维护全面升级的优势。开发时基于 Web 前后台分离的思想，前端负责与用户交互、展示各项信息、处理用户输入等业务逻辑。后台接收前端请求，完成相应的操作，返回需要的数据。前后台分离不仅提升了开发效率，也使得项目更易于维护和扩展。

2.1 前端

前端主要使用 Vue 进行开发，Vue 是一套用于构建用户界面的渐进式框架，有着方便的双向数据绑定、指令、组件化等诸多特性，特别适合用于构建数据驱动的 Web 界面。本平台中几乎所有功能都需要从后台获取数据或者向后台提交数据，使 Vue 进行开发使得这一切操作都变得更加方便。组件化的开发模式大大降低了各个功能的耦合程度，使得整个前台条理更加清晰，提升了项目的可维护性。

2.2 后台

本系统后台使用 Spring Boot 框架开发。

Spring Boot 是由 Pivotal 团队提供的全新框架，其设计目的是用来简化新 Spring 应用的初始搭建以及开发过程。该框架使用了特定的方式来进行配置，从而使开发人员不再需要定义样板化的配置。

Spring 是一个轻量级的 Java 开发框架，为了解决企业应用开发的复杂性而创建。框架的主要优势之一就是其分层架构，分层架构允许使用者选择使用哪一个组件，同时为 J2EE 应用程序开发提供集成的框架。Spring 使用基本的 JavaBean 来完成以前只可能由 EJB 完成的事情。简单来说，Spring 是一个分层

的 Java SE/EE full-stack(一站式) 轻量级开源框架。

Spring 有着如下优点：

1) 方便解耦，简化开发

Spring 就是一个大容器，可以将所有对象创建和依赖关系维护，交给 Spring 管理。

2) AOP 编程的支持

Spring 提供面向切面编程，可以方便的实现对程序进行权限拦截、运行监控等功能。

3) 声明式事务的支持

只需要通过配置就可以完成对事务的管理。

4) 方便程序的测试

Spring 支持 Junit4，可以通过注解方便地测试 Spring 程序。

5) 方便集成各种优秀框架

Spring 不排斥各种优秀的开源框架，其内部提供了对各种优秀框架（如：Struts、Hibernate、MyBatis、Quartz 等）的直接支持。

6) 降低 JavaEE API 的使用难度

Spring 对 JavaEE 开发中非常难用的一些 API（如 JDBC、JavaMail、远程调用等）都提供了封装，使这些 API 应用难度大大降低。

2.3 被测算法

本平台旨在测试大数据集群环境的性能和算法效率，算法方面支持 Spark 集群算法。要使本平台能够跟踪到被测算法的运行过程，算法应满足一定的要求，具体开发要求请参照《算法开发文档》。

3 系统功能

本小节系统地介绍平台的各项功能，具体操作流程可以参照《用户手册》。

3.1 导入性能测试

本部分用于测试集群环境的数据导入性能，使用时需要先将本地的文件上传至 Web 后台服务器，再从服务器的文件系统上传至 HDFS。

对于已经上传到文件系统和 HDFS 的文件都提供了根据文件名/上传时间查找、下载、删除等管理操作。

HDFS 文件管理界面底部展示了集群环境的存储空间情况，包括总空间、不可用空间、已用空间和剩余空间，如图 3-1 所示。

图 3-1 HDFS 文件管理页面

3.2 数据预处理及评估

本部分用于测试数据预处理过程的性能，以及评估数据质量。

数据预处理评估需要创建数据预处理任务，如图 3-2 所示，这里用到的“算法”和“数据”均为用户上传到 3.1 中文件系统的文件（扩展名为.jar/.py 的文件会识别为算法，其他识别为数据）。

如图 3-3 所示，数据预处理任务创建完成可以在处理结果列表中看到，完成后可以查看预处理过程的耗时，并提供下载预处理结果、删除预处理任务等管理功能。

对于预处理后的结果文件或者用户上传到 3.1 中文件系统中的文件都可以进行数据质量评估，前者通过任务列表中对应的“评估”按钮创建任务，后者通过任务列表下方的“直接评估”按钮创建任务。平台支持对数据中每一列进行约束来评估，列的序号从 0 开始，约束规则支持表 1 中的六种。

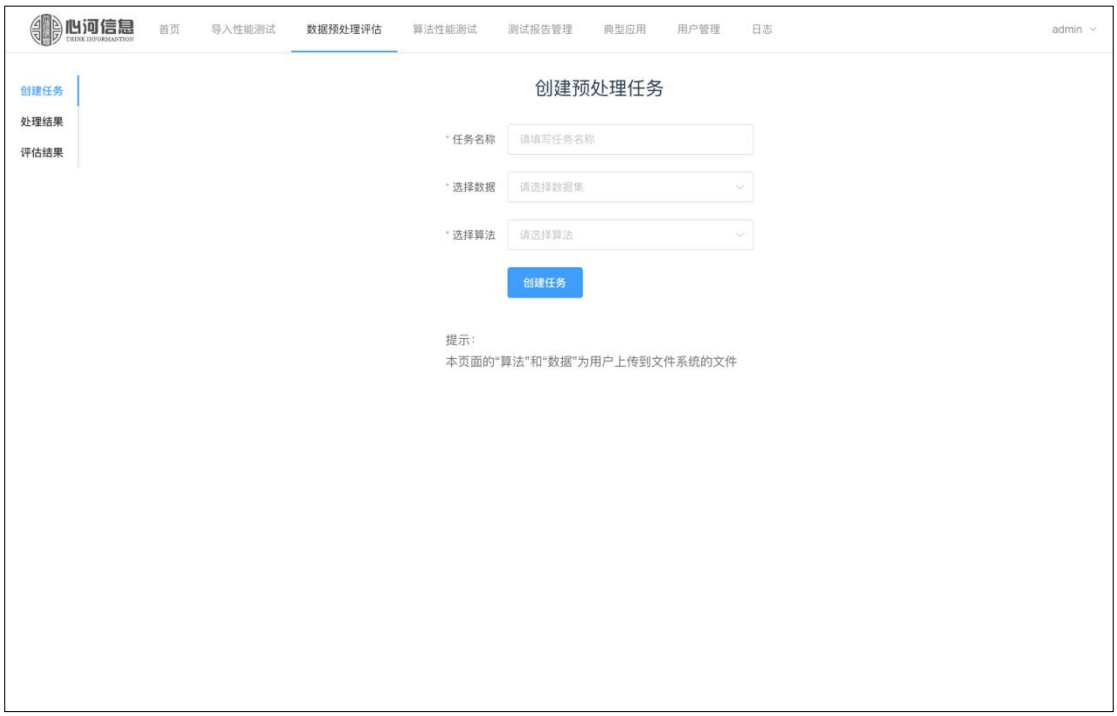

图 3-2 数据预处理评估界面

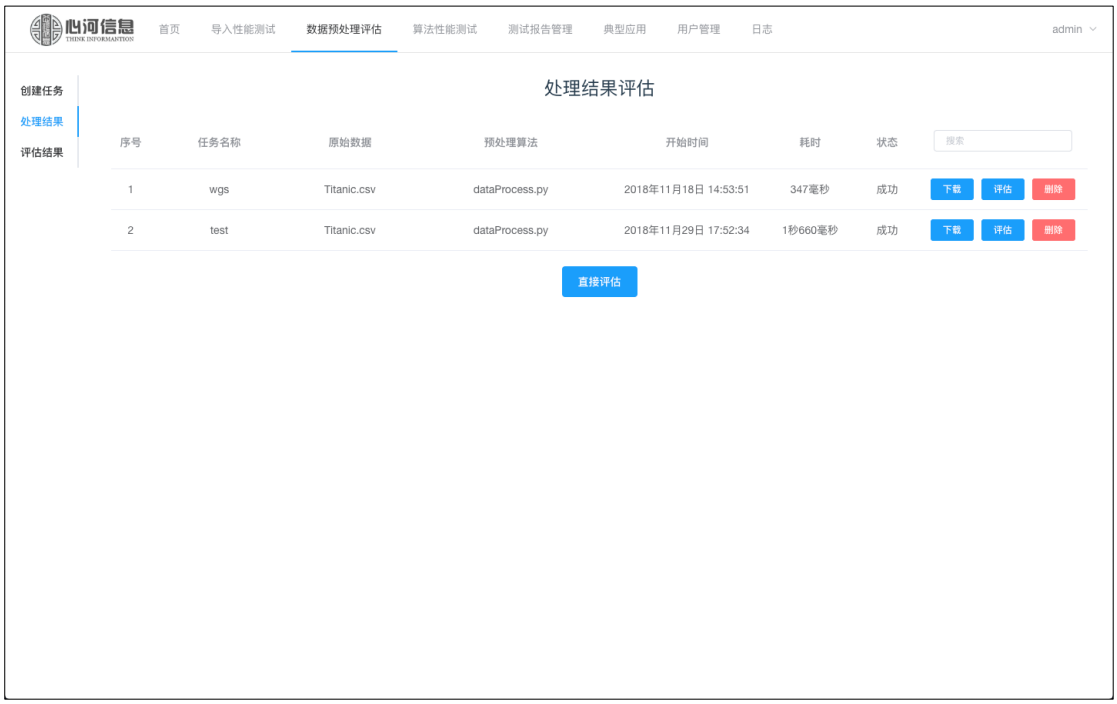

图 3-3 处理结果界面

表 1 数据质量评估约束规则表

规则	说明	示例
=	等于某个具体的值	0
<	小于某个值	200
<=	小于等于某个值	200
>	大于某个值	100
>=	大于等于某个值	100
in	在某个集合内	[0,1,2,3,5]

数据质量评估任务创建完成可以在评估结果列表中看到，完成后可以查看评估过程的耗时和评估结果，点击“查看评估结果”按钮会弹出结果详情模态框，详细地列出创建任务时指定的每一条约束规则以及所有规则汇总的合格行数、合格比例，如图 3-4 所示。

图 3-4 评估结果

3.3 算法性能测试

本部分用于测试集群算法的性能，并提供几类常用算法的结果评估功能。

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，这样有监督的算法任务也可以得出结果。

算法性能测试需要创建算法测试任务，这里用到的“算法”和“数据”均为用户上传到 3.1 中 HDFS 的文件（扩展名为.jar/.py 的文件会识别为算法，其他识别为数据）。

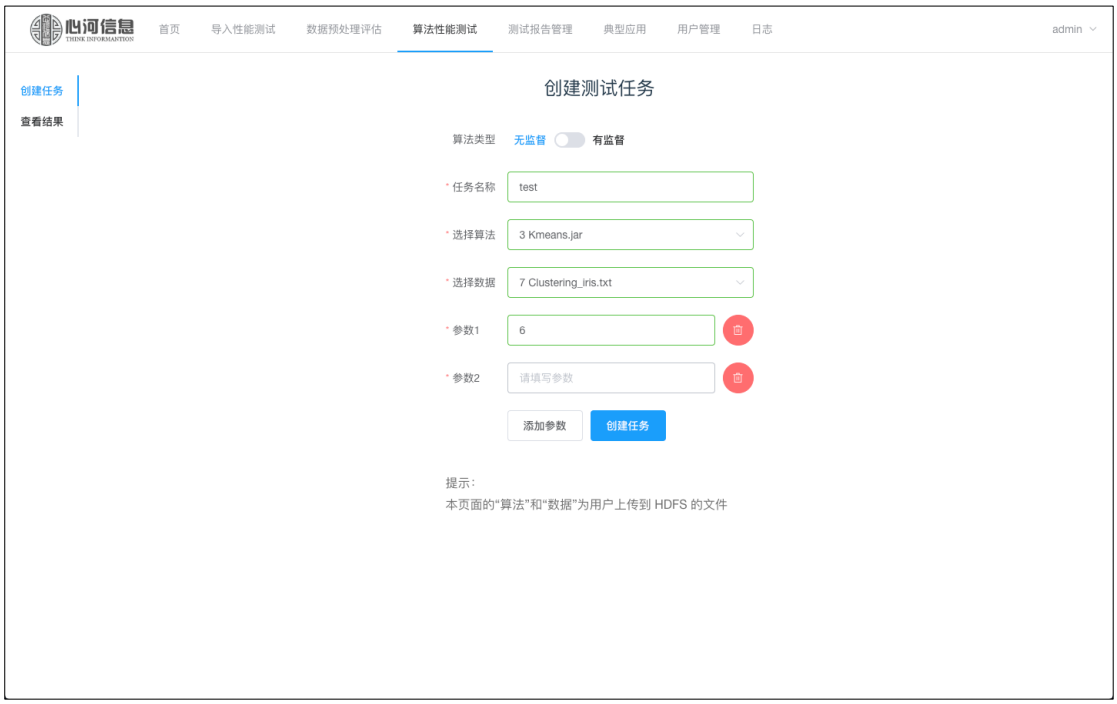

图 3-5 创建无监督算法任务

图 3-6 有监督算法选择测试数据

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，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 3-5 所示。

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，用于选择无标签的测试数据，如图 3-6 所示。

算法测试任务创建完成可以在任务结果列表中看到，完成后可以查看算法运行过程的耗时，并提供下载算法结果、删除算法测试任务等管理功能。

另外平台提供了聚类、分类、回归三种常用算法的结果评估方法，对于前面创建的算法任务，选择相应的评估方法平台会给出相应的评估结果，包括可视化的图形和各项数值指标。表 2 展示了对于这三类算法的具体评估内容。

表 2 三种算法评估方法表

算法	图	评价指标
聚类	将原始数据的多维度特征降到二维，在平面图中以不同的颜色展示每个簇中所有的点及各自的簇中心点	Compactness：紧凑度
		Separation：分离度
		Davies-Bouldin Index：戴维森堡丁指数
		Dunn Validity Index：邓恩指数
分类	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	Hamming loss：汉明损失
回归	以折线图的形式展示测试数据的真实值和预测值	MAE：平均绝对误差
		SSE：误差平方和
		MSE：均方误差
		RMSE：均方根误差
		R-Square：确定系数

3.4 测试报告管理

本部分提供测试报告管理，用于保存整理相关测试文档等资料。支持上传、根据报告名称/上传时间查找、下载、删除等操作。

3.5 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。

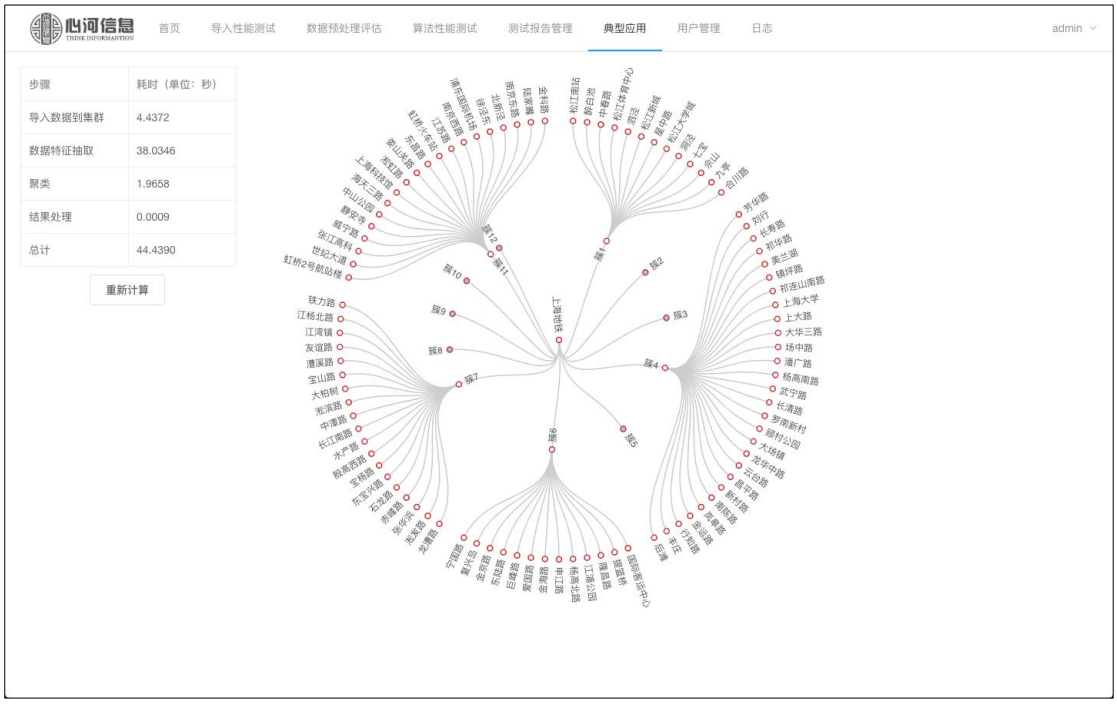

图 3-7 上海地铁站点聚类

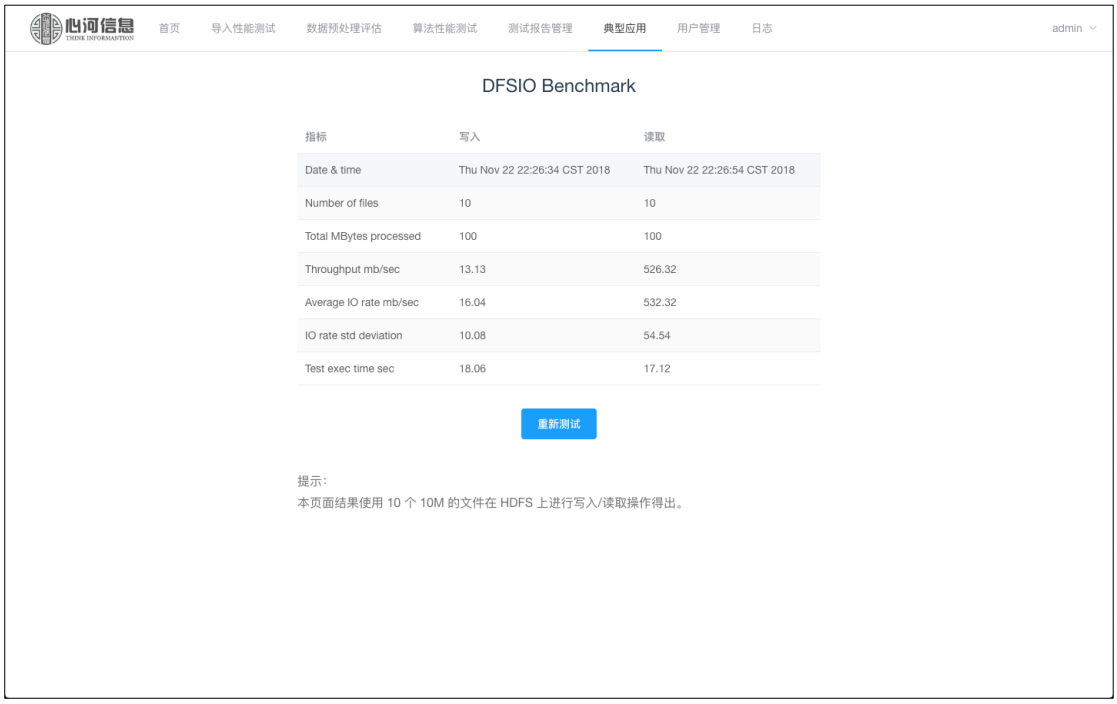

图 3-8 基准测试

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 3-7 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。

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 3-8 页面中展示的是这一测试的结果，点击界面下方的“重新测试”按钮将重新执行测试过程，这一过程实际上执行了表 4 中的两条命令。

表 4 基准测试命令表

测试内容	命令
写入	/home/admin/installation/hadoop-2.7.6/bin/hadoop jar /home/admin/installation/hadoop-2.7.6/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.7.6-tests.jar TestDFSIO -write -nrFiles 10 -fileSize 10MB
读取	/home/admin/installation/hadoop-2.7.6/bin/hadoop jar /home/admin/installation/hadoop-2.7.6/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.7.6-tests.jar TestDFSIO -read -nrFiles 10 -fileSize 10MB

3.6 用户管理（仅限管理员）

本平台的用户分为三种角色——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。新注册的账号角色为“演示”，权限最低。

“管理员”可以修改其他账户的角色以及删除账户。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。

3.7 日志（仅限管理员）

“管理员”角色可以查看平台的操作日志，便于排查问题。日志包括“登录”、“上传文件至文件系统”、“上传文件至 HDFS”、“删除文件系统文件”、“删除 HDFS 文件”、“创建数据预处理任务”、“删除数据预处理任务”、“创建预处理评估任务”、“删除预处理评估任务”、“创建算法测试任务”、“删除算法测试任务”、“上传测试报告”和“删除测试报告”共 13 项操作。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。

3.8 平台接口配置（仅限管理员）

系统运行需要用到一系列的接口，具体说明如下表：

表 3 后台接口配置说明表

接口	说明	默认值
Web 后台 IP	用于集群算法向后台报告运行状态	192.168.1.2
集群 Master IP	-	192.168.1.2
Spark 端口	-	7077
HDFS 端口	-	9000
HDFS 目录	用于向 HDFS 写入/读取文件	test
集群 Master 用户名	-	admin
集群 Master 密码	-	123456
spark-submit 路径	提交集群算法任务	/home/admin/installation/spark-2.2.1-bin-hadoop2.7/bin/spark-submit
Hadoop 路径	-	/home/admin/installation/hadoop-2.7.6/bin/hadoop
Benchmark jar 包路径	基准测试 jar 包	/home/admin/installation/hadoop-2.7.6/share/hadoop/mapreduce/hadoop-mapreduce-client-jobclient-2.7.6-tests.jar
Python 路径	Python 环境	/home/admin/anaconda3/bin/python

4 部署

4.1 集群环境

每次开机后集群需要手动启动，需要按顺序启动如下三个服务：HDFS, Yarn, Spark，命令分别为 HDFS：start-dfs.sh，Yarn：start-yarn.sh，Spark：$SPARK_HOME/sbin/start-all.sh。

可以使用下面的命令全部启动：

start-dfs.sh && start-yarn.sh && $SPARK_HOME/sbin/start-all.sh

相应地，如果想手动关闭集群服务，只需倒序停止上述三个服务，命令为：$SPARK_HOME/sbin/stop-all.sh && stop-yarn.sh && stop-dfs.sh。

4.2 数据库

本系统后台使用 MySQL 做数据存储，初次部署应安装 MySQL，启动服务，并创建名为“xinhe”的数据库。重启服务器 MySQL 服务会自动自动，无需任何操作。

4.3 前端

部署时前端为发布好的静态文件（放置在/home/admin/hadoop_testing/dist），使用 Nginx 进行部署，Nginx 服务配置了开机自启，服务器重启后前端部分不需要任何操作。Nginx 配置文件路径为：/usr/local/nginx/conf/nginx.conf。其中配置了如下内容：

```
server {
    listen          80;
    server_name    localhost;
    location / {
        root     /home/admin/hadoop_testing/dist;
        index    index.html index.htm;
    }
}
```

4.4 后台

后台为打包好的 Java 程序，路径为 /home/admin/hadoop_testing/hadoop_testing-1.0.0.jar，为了方便部署编写了启动和停止脚本，使用方法为：

启动：

cd /home/admin/hadoop_testing/ && bash start.sh

停止：

cd /home/admin/hadoop_testing/ && bash stop.sh

注意：由于后台使用固定的 8090 端口，启动前必须先停止已经启动的进程。

4.5 总结

服务器重启后启动项目的操作如下：

1. 启动集群

start-dfs.sh && start-yarn.sh && $SPARK_HOME/sbin/start-all.sh

2. 启动项目后台

cd /home/admin/hadoop_testing/ && bash start.sh

5 其他说明

5.1 demo 数据介绍

Iris 鸢尾花数据集，是一类多重变量分析的数据集。通过花萼长度，花萼宽度，花瓣长度，花瓣宽度 4 个属性预测鸢尾花卉属于 Setosa， Versicolour，Virginica 三个种类中的哪一类。数据来源：http://archive.ics.uci.edu/ml/datasets/Iris。

Murder 犯罪率数据集，是一个回归问题的数据集。数据为美国 50 个州的人口、收入、文盲率、平均寿命、高中毕业率、全年气温低于 0℃的天数、总面积以及犯罪率。意在通过州的多种特征预测该州的犯罪率。详细介绍：https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/state.html。

Titanic 泰坦尼克号乘客数据集，是一个二分类问题的数据集。其中包含乘客编号、姓名、性别、年龄、船上兄弟姐妹/配偶数量、船上父母/子女数量、船票类型、票号、票价、船舱号、登船港口等特征，以及是否幸存的标签，意在使用上述特征预测乘客是否幸存。数据来源：https://www.kaggle.com/c/titanic/。