

平台使用说明

1 平台使用准备

*操作系统：Windows、Linux、Mac OS X

*网络：校园网，访问地址：114.212.190.95:8082

*推荐浏览器：谷歌浏览器（50 及以上）、火狐浏览器（45 及以上）

2 用户登录

在登录页面输入正确的账号密码，点击登录按钮完成登录操作。

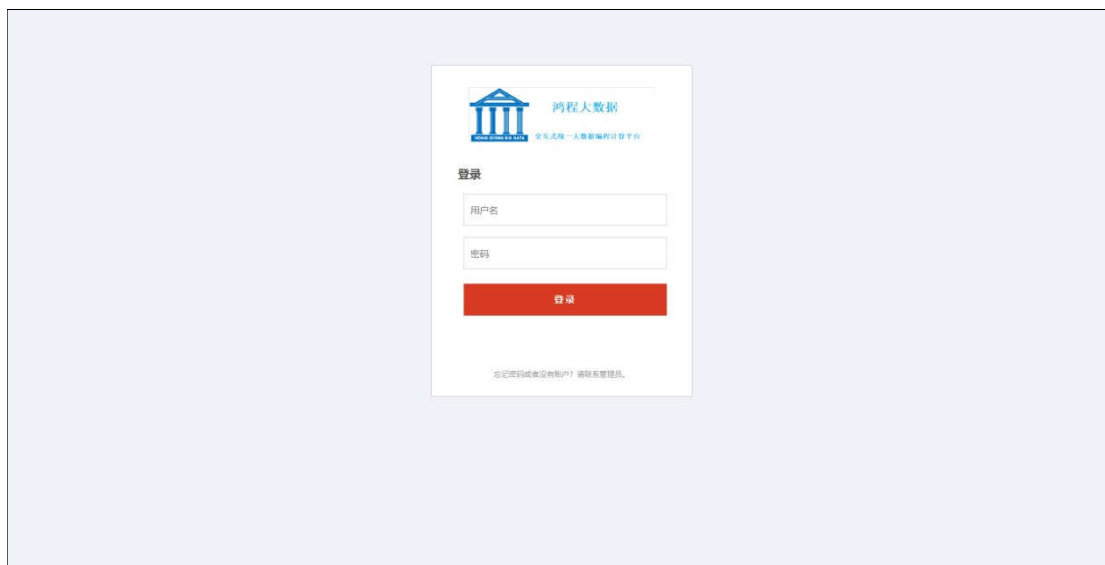


图 1 用户登录页面

3 平台主页面介绍

左侧为功能列表区，右侧为功能使用区。



图 2 系统主页面

首次登录进入主页面，尚未选择功能列表之时，主页面功能区会启动一个 Python3 的交互式编程环境，可用于 Python3 的交互式编程。

1) 如图所示，点击【交互式统一大数据编程计算平台】可以回到主页面

2) 在代码框中可以编写 Python 代码，点击“运行”，运行结果将在主页下方控制台显示

4 普通用户

本节主要介绍了普通用户登录后的使用以及注意事项。

4.1 交互编程与作业提交

本小节将介绍平台下编程工作空间中 Jupyter Notebook 的基本用法以及批量执行 Jar 提交的注意事项，

1、编程工作空间

1) 用户首先需要登录进入编程工作空间，账号密码与平台的账号密码一致。

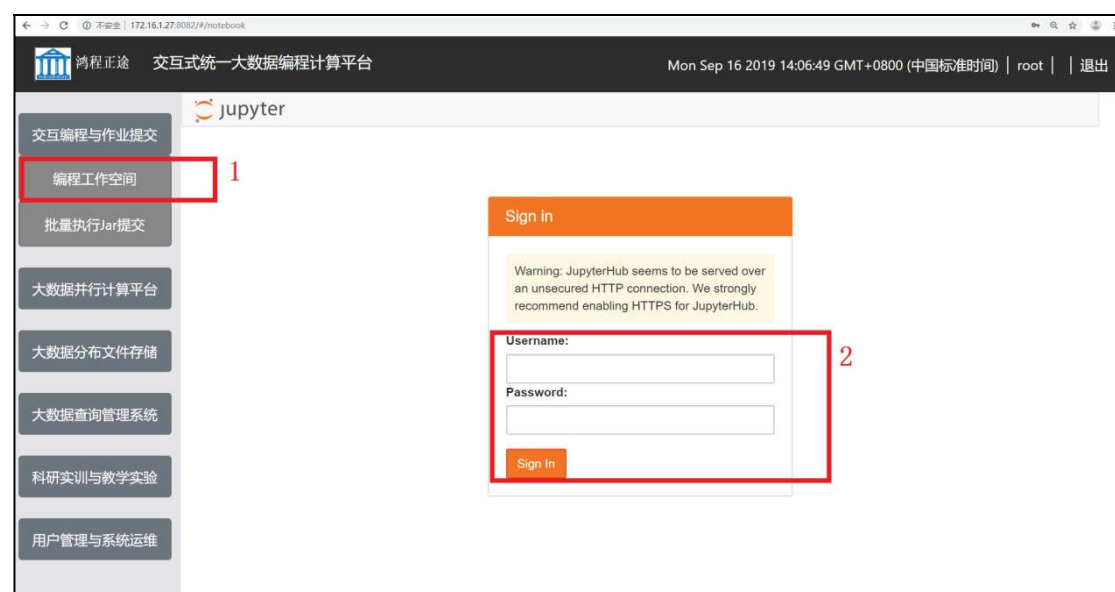


图 3 Notebook 登录页

2) 登录进入后，点击符号 1 处的上传箭头可以从本地上传文件到当前用户的目录下；符号 2 处显示的是当前用户目录下的所有文件；符号 3 处可以选择交互式编程语言，包括 Python 和 Scala，然后进入编程环境进行开发。这些是编程工作空间的主要功能。

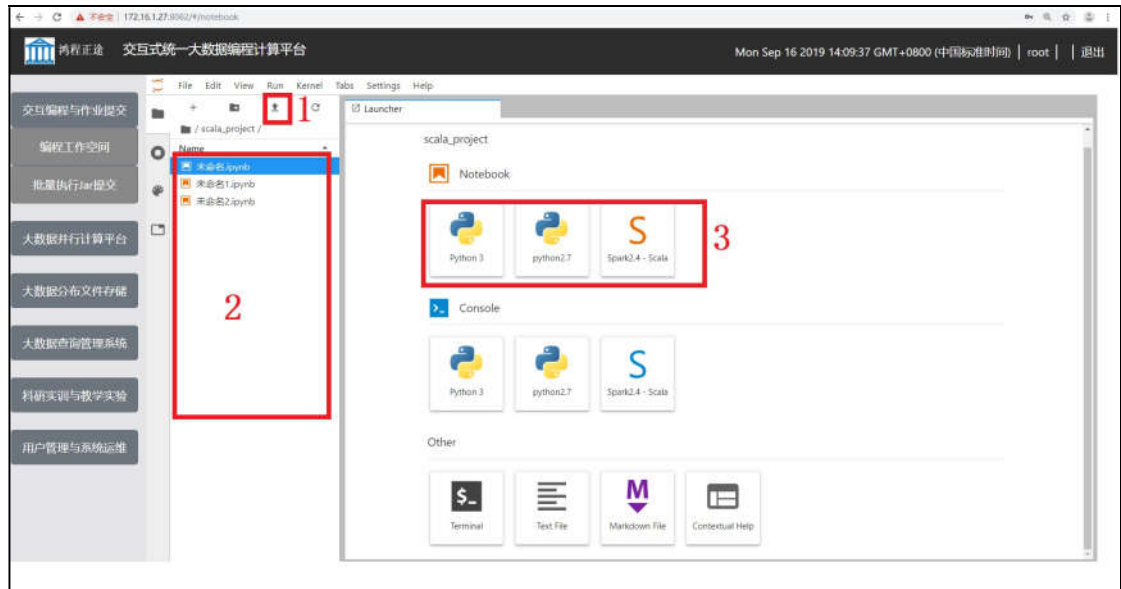


图4 Notebook 首页

案例 1: Scala 编程

如下图所示,是一个简单的使用 Scala 语言的案例。当用户选择 spark2.4-scala 环境时,可以直接使用 SparkContext 的 sc 引用。

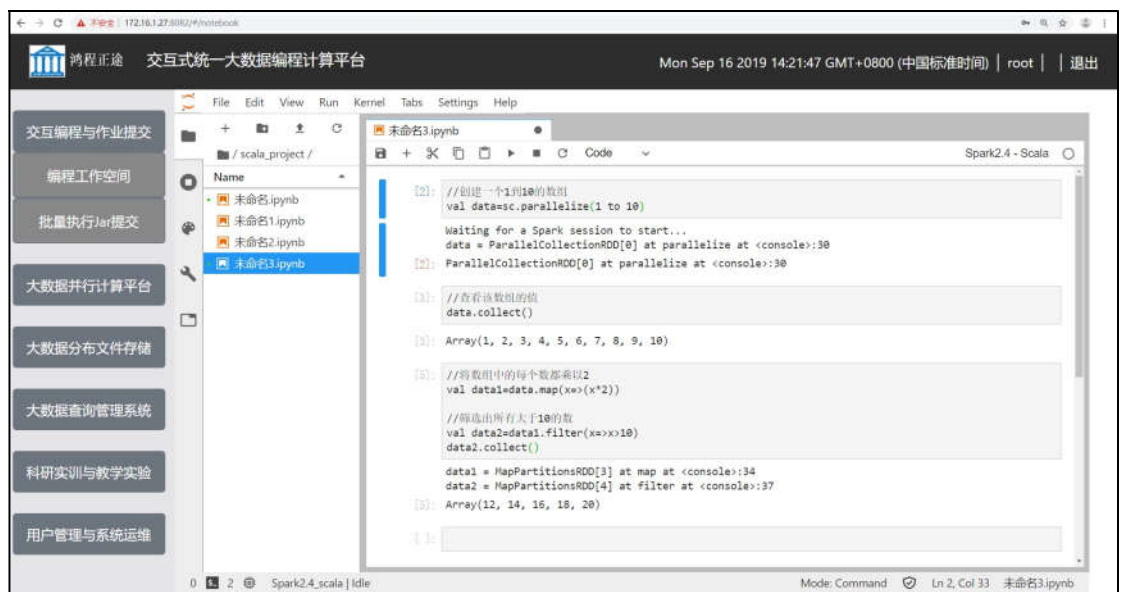


图5 Scala 编程页

案例 2: Python 编程（实现聚类的案例）

注意：当使用 PySpark 开发时，如下图所示，如需要指定 Master 在 Yarn 上运行，设置`master("yarn")`

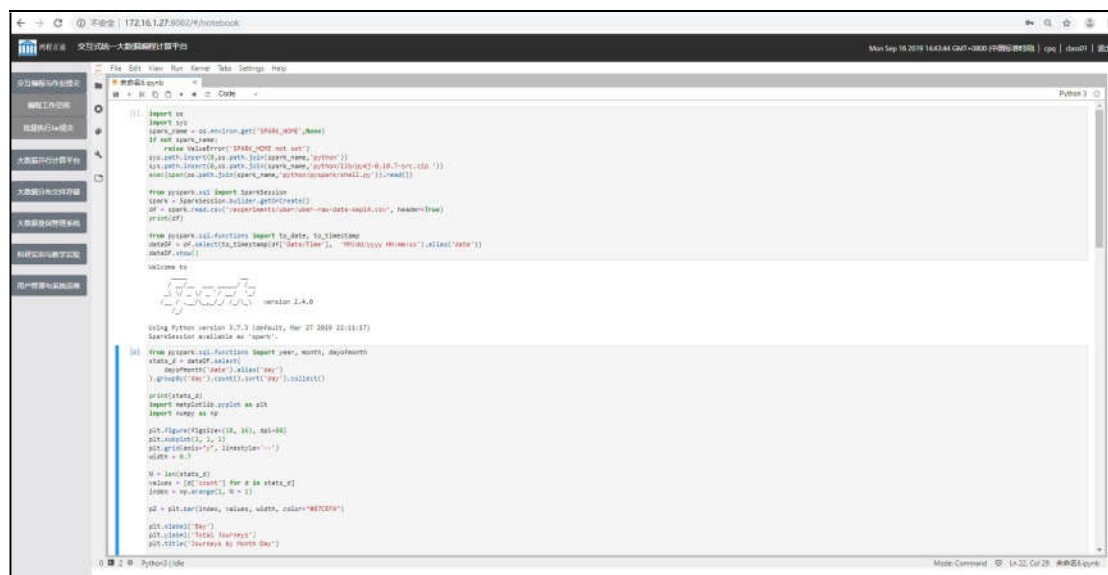


图 6 Python 编程页

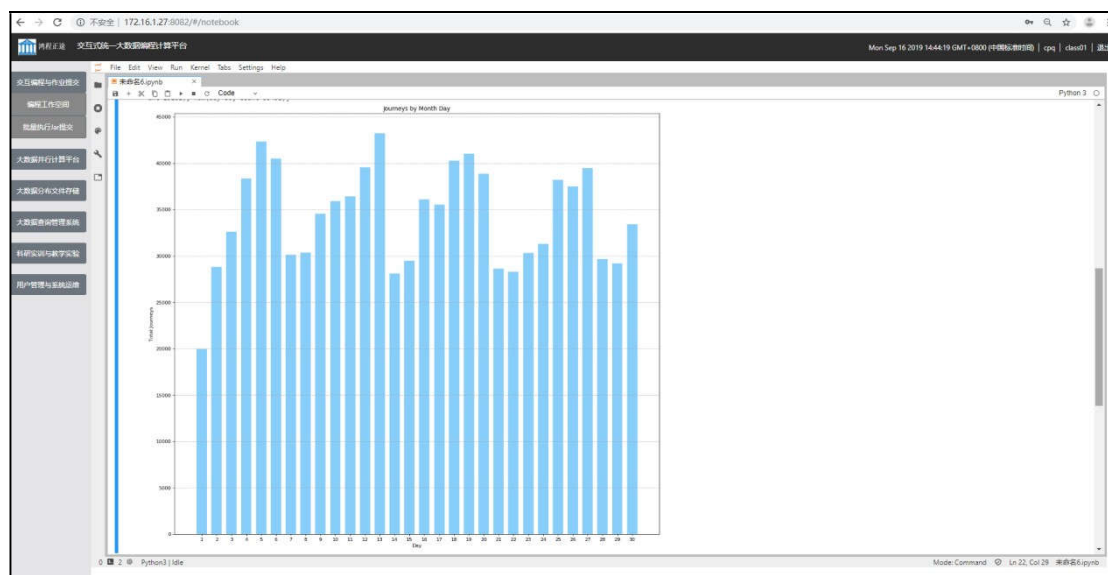


图 7 Python 编程页

2、批量执行 Jar 提交

点击左侧的”批量执行 Jar”提交进入 Jar 包提交页面。

在本页面用户可以提交 MapReduce 或者 Spark 的 Jar 包，前提是用户已经将数据集上传到 HDFS，然后用户只需要数据提交的命令就可以执行。

如图所示，点击符号 1 处的”批量执行 Jar 提交”功能进入页面，在符号 2 或者符号 3 处输入对应的提交命令，然后点击提交，运行结果在符号 4 处的控制台打印出来。



图 8 提交 Jar 包页

4.2 大数据并行计算平台

本节包含 MapReduce 并行计算和 Spark 并行计算两个功能，下面将依次介绍。

1、MapReduce 并行计算

如图 9 所示：

点击符号 1 可以查看集群中的所有任务，包括历史任务和运行中的任务；

点击符号 2 可以查看集群中的资源池信息，包括资源池的名称，池中运行的任务以及资源占用情况；

符号 3 处显示的是任务的具体信息，包括任务 ID、提交者、任务名称、任务类型、开始时间、结束时间、运行状态、运行进度。

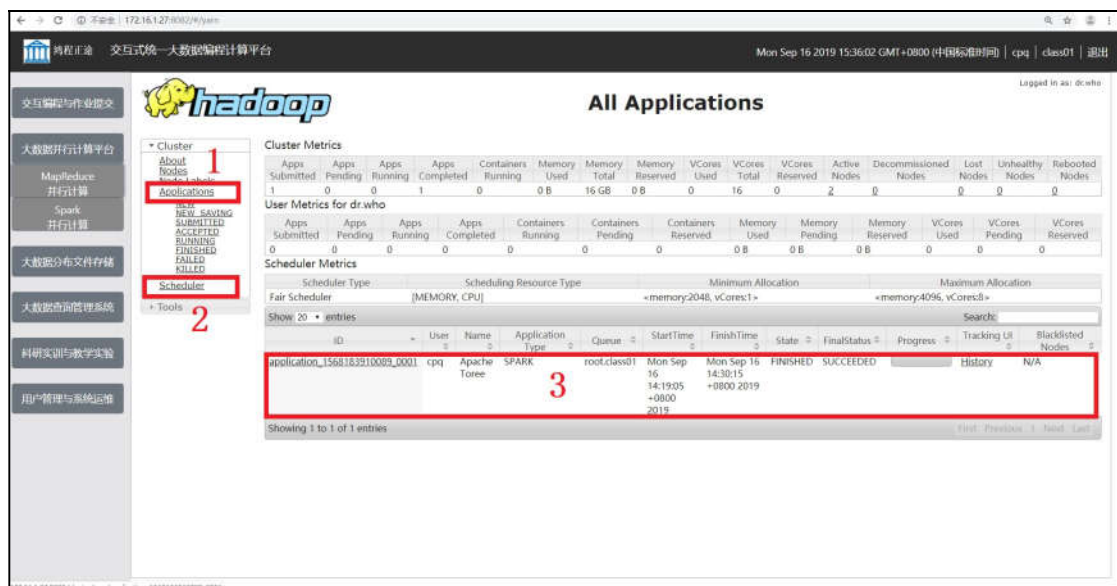


图 9 MapReduce 任务监控页

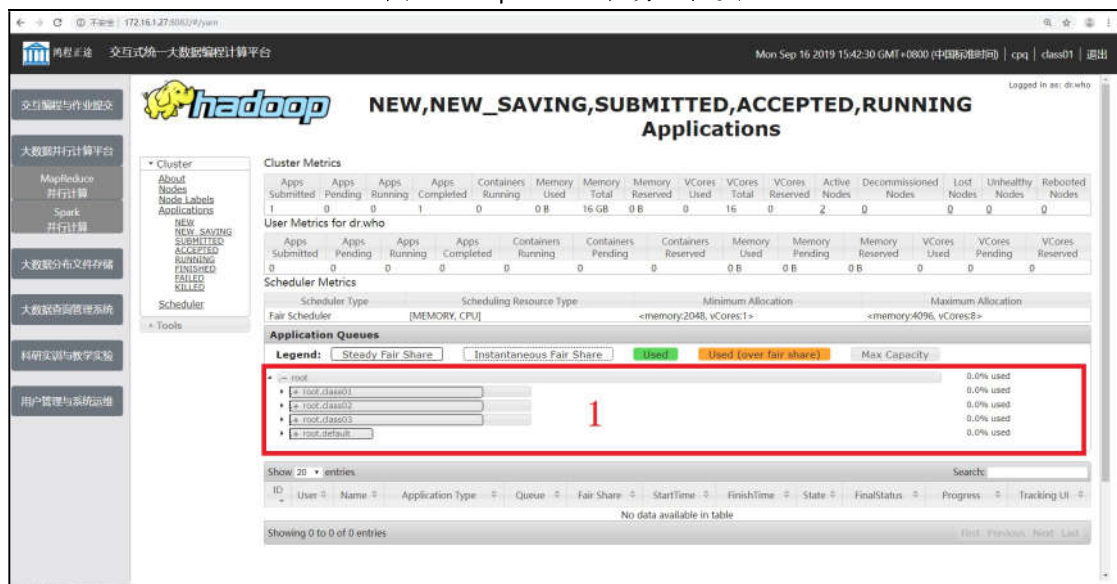


图 10 MapReduce 资源池页

2、Spark 并行计算

如图 11 所示：

符号 1 处所示为所有 Spark 的工作节点信息，包括节点状态信息、可用内存、可用核心数等；

符号 2 处所示为当前正在运行的任务信息，包括任务 ID、任务名称、占用资源数等，用户也可以点击“kill”来终止当前任务；

符号 3 处所示为所有历史的任务信息。

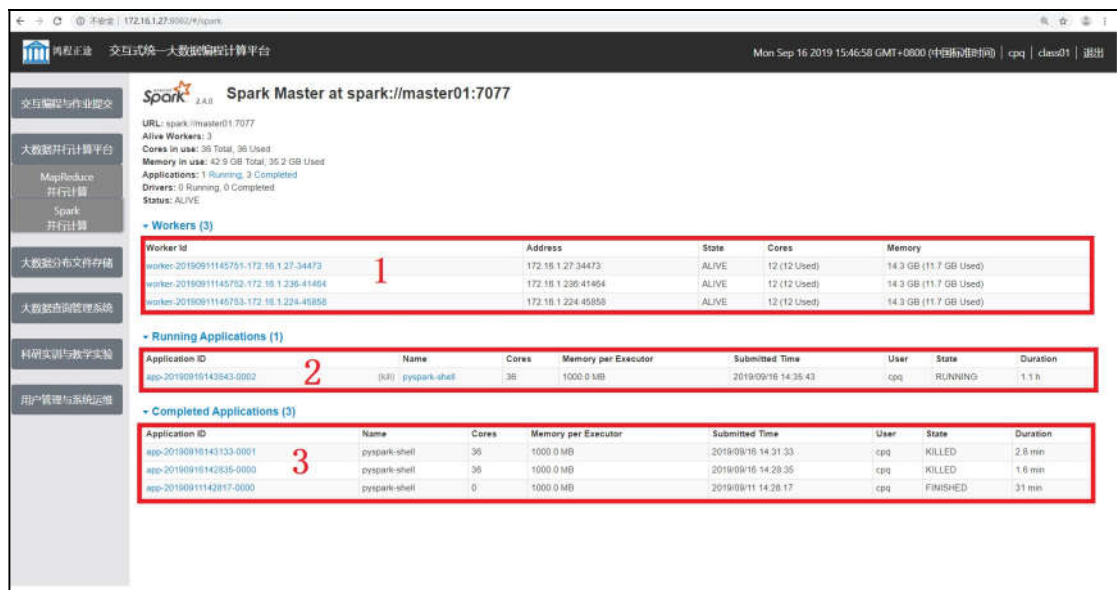


图 11 Spark 任务监控页

4.3 大数据分布文件存储

本节主要介绍在本平台下分布式文件存储系统 HDFS 以及分布式虚拟存储系统 Alluxio 的相关用法。

1、分布式文件存储系统 HDFS

如下图 12 所示。符号 1-6 是本平台下 HDFS 提供的主要功能，下文将依次介绍。

注意：外部读取本平台下 HDFS 文件的路径为：**hdfs://ip:9000/path**

ip 为主节点的 ip，**path** 为目标文件在 HDFS 的路径。

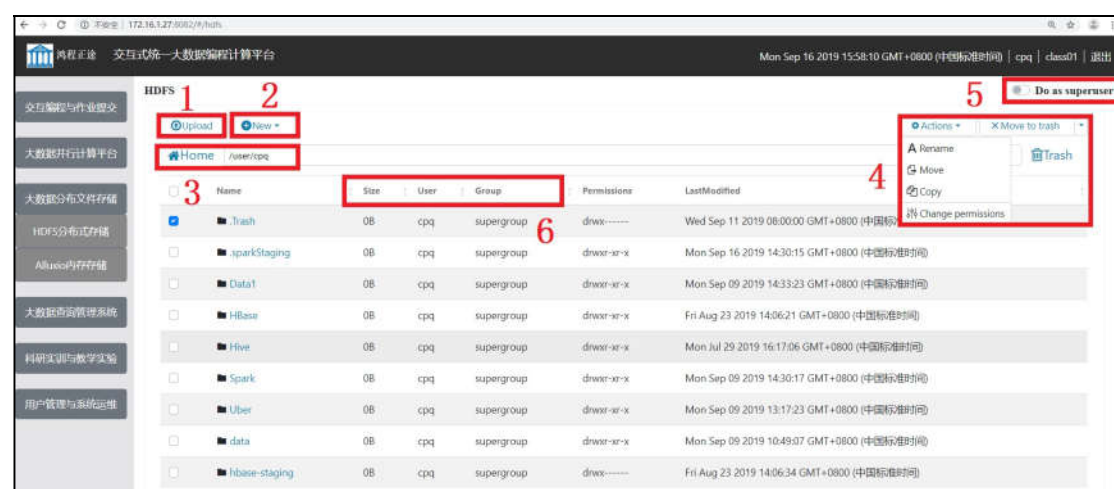


图 12 HDFS 首页

功能 1：上传文件

- 1) 点击图 13 标注 1 处【Upload】
- 2) 在弹出框中点击【select files】
- 3) 文件上传完成时在标注 3 处会出现“success”提示
- 4) 点击标注 4 处【Finish】完成上传

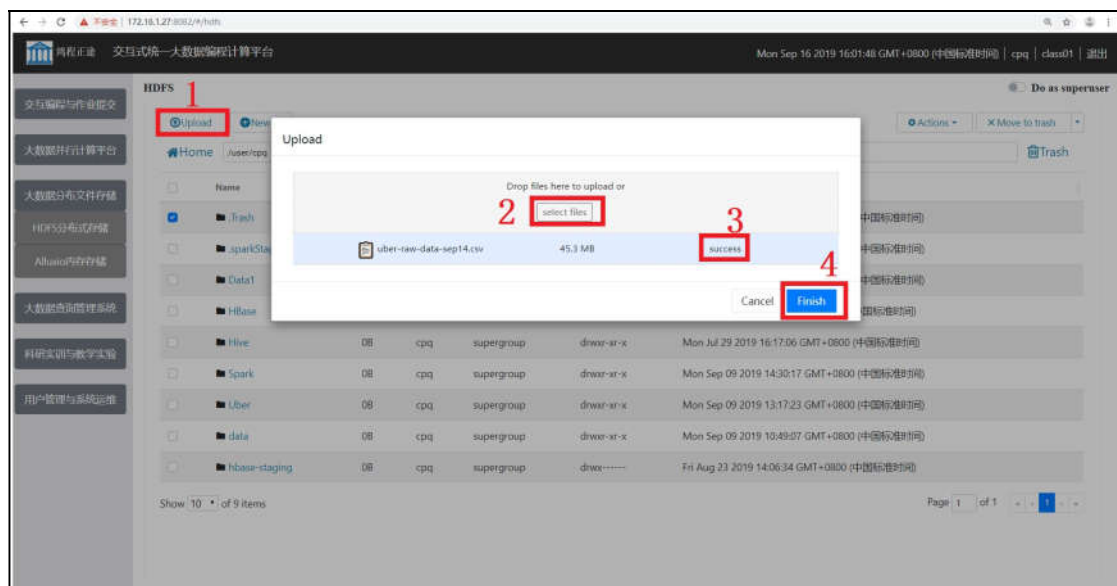


图 13 HDFS 上传文件示例图

功能 2：新建文件/文件夹

- 1) 点击图 9 中标注 1 处【New】
- 2) 在标注 2 处输入文件夹名称
- 3) 点击标注 3【Create】确认创建

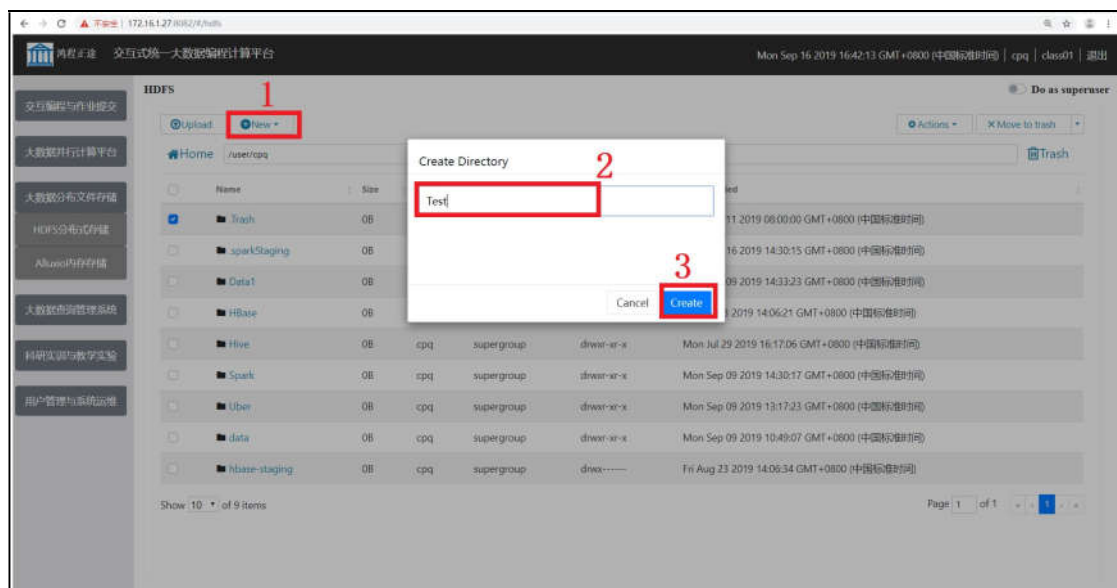


图 14 新建文件夹示例图

功能 3：【Home】键的使用

点击图中的【Home】键可以在任意路径处回到当前用户的家目录，或者在地址栏输入目标路径直接跳转。（此处不做展示）

功能 4：【Action】操作

选中一个文件夹或者文件后，点击下图标注 1 处的【Action】按钮对其进行

相应操作，包括 Rename(更改名称)、Move(移动)、Copy(复制)、Change Permissions(更改权限)，此处只展示更改权限操作。

- 1) Read、Write、Execute 分别为读、写、执行权限
- 2) Sticky: 防删除位，一般只用在目录上，用在文件上起不到什么作用

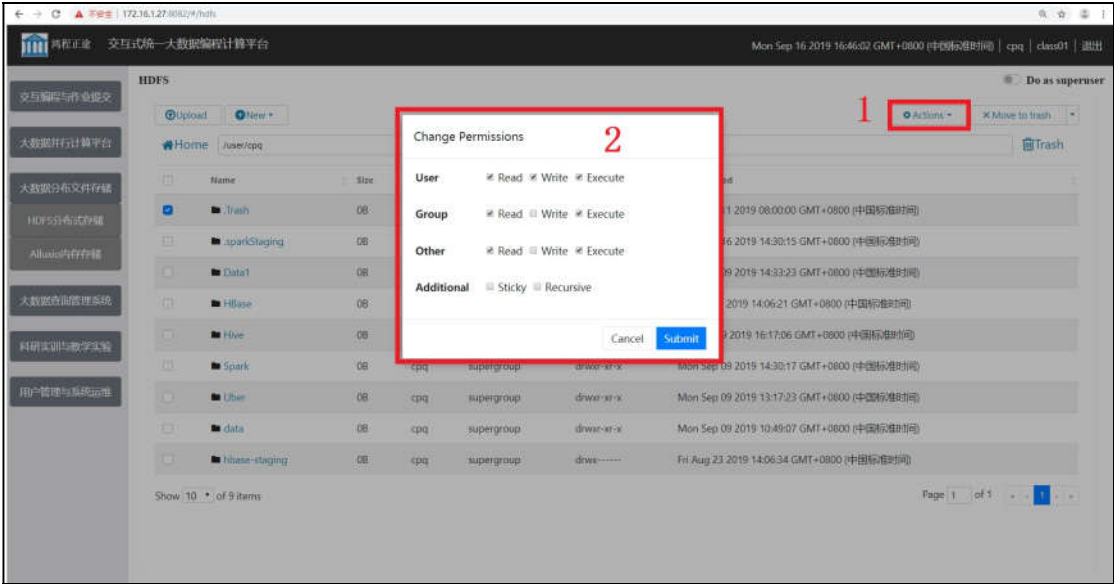


图 15 更改权限示例图

功能 5：垃圾箱功能

1) 点击【Move to trash】可以将文件放入垃圾箱中，或者点击【Delete forever】永久删除；点击垃圾桶按钮【Trash】可以查看垃圾箱。

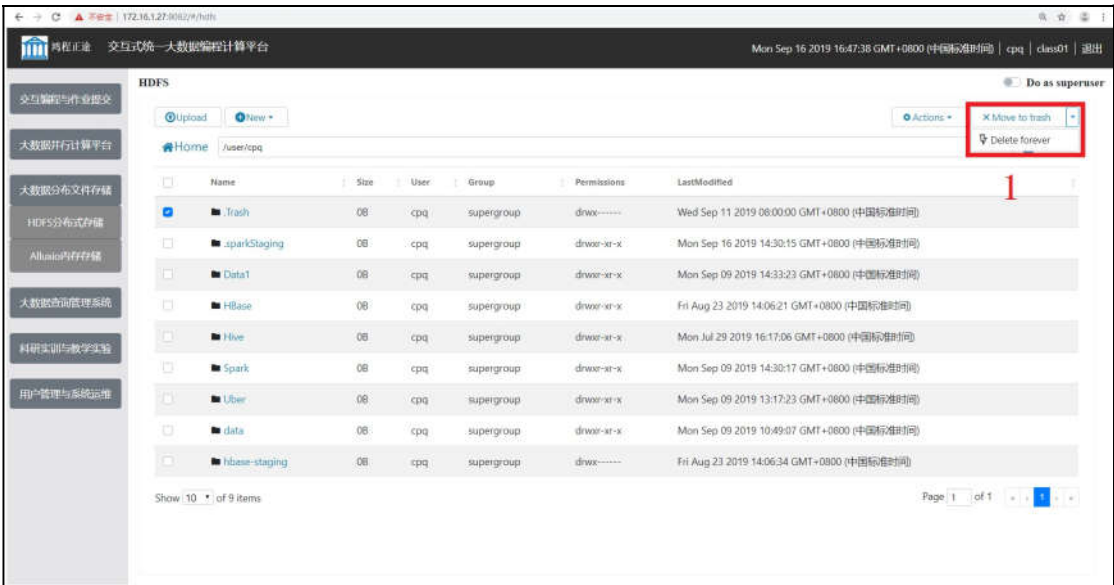


图 16 删除文件示例图

- 2) 垃圾箱中包含【Restore】：还原操作；【Delete forever】：永久删除；【Empty】

trash】：清空操作

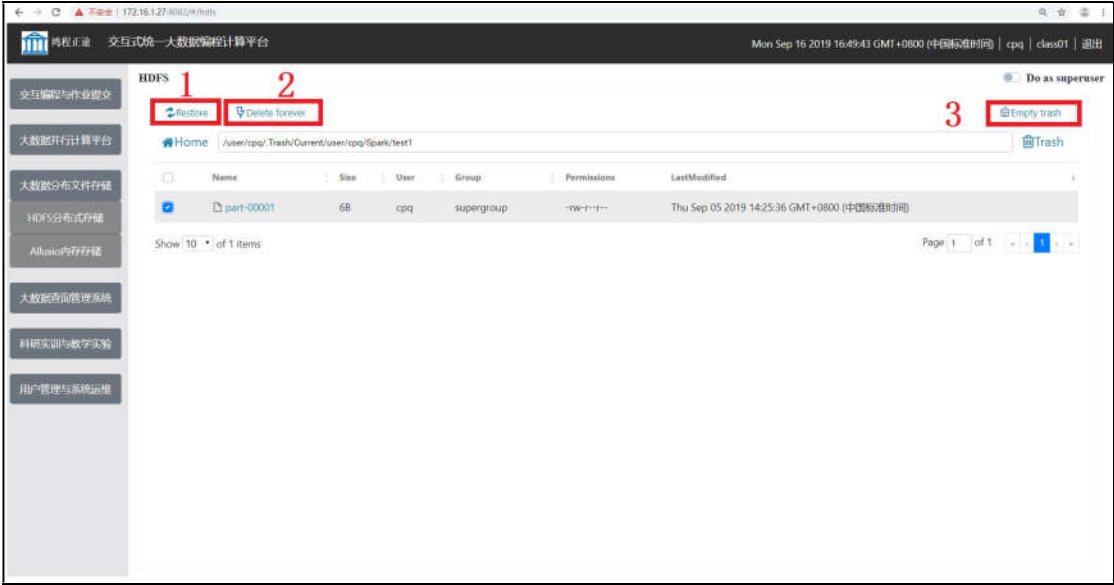


图 17 垃圾箱首页

功能 6:

- 1) 点击图中右上角处的【Do as superuser】可以让当前用户成为底层集群的 root 用户，拥有 root 用户的所有权限，该按钮只有管理员才可以点击。
- 2) 点击图 12 中的标注 6 处所示的上下标注可以将文件按照指定规则排序。

2、分布式虚拟存储系统 Alluxio

本节将介绍在本平台下 Alluxio 的相关用法，点击左侧大数据分布文件存储，选择 Alluxio 内存存储，如下图所示。

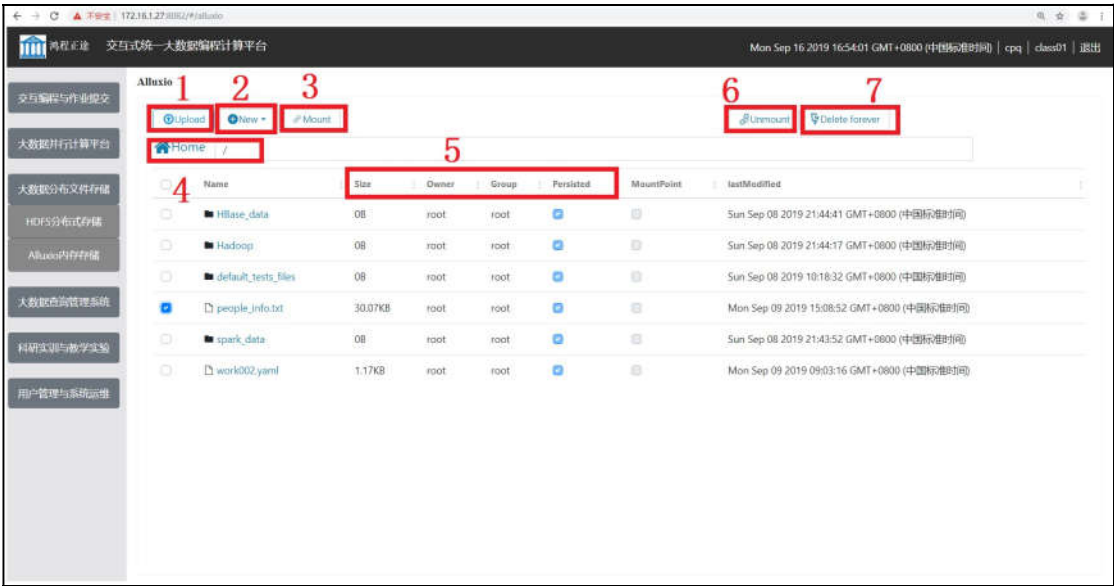


图 18 Alluxio 首页

如图所示，本平台上的 Alluxio 主要包含以下功能：

- 1) 【Upload】上传操作，上传时可以选择 persist 功能
- 2) 【New】新建文件或文件夹
- 3) 【Mount】和【Unmount】挂载与取消挂载
- 4) 【Home】回到家目录
- 5) 【Delete forever】永久删除文件或文件夹
- 6) 【标注 5】处可以按照规则排序功能

功能 1：上传文件。上传文件时可以选择是否勾选 persist，决定是否将仅存于 Alluxio 中的文件或文件夹持久化到底层文件系统中

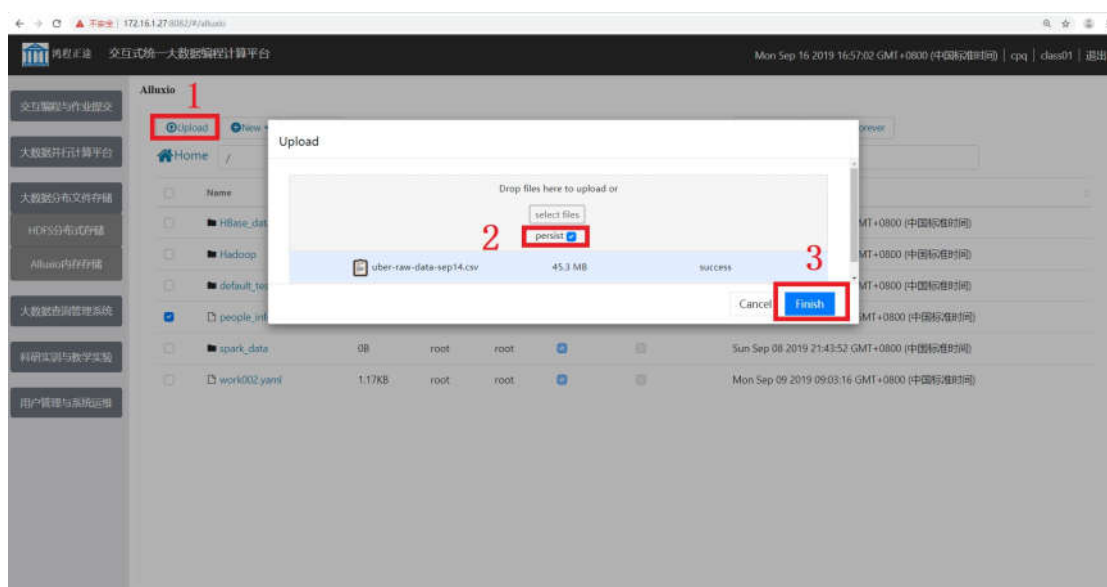


图 19 Alluxio 上传文件页

功能 2：Mount

将底层存储系统的 Src 路径挂载到 Alluxio 中的 Path 路径，将底层存储系统与 Alluxio 自身的存储空间统一管理起来，呈现给应用层一个统一的命名空间，避免了复杂的输入输出逻辑。

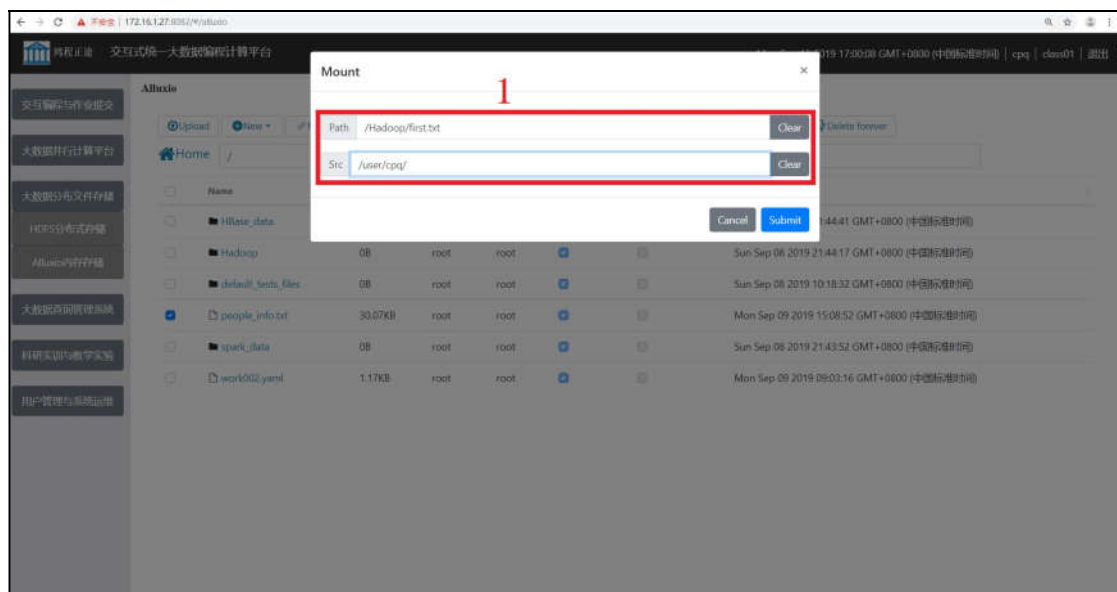


图 20 Mount 挂载操作图

功能 3：查看文件信息

点击文件可以查看文件的具体格式信息和文本内容，也可以将文件下载到本地。（只支持查看纯文本格式的文件，其他格式会出现乱码）

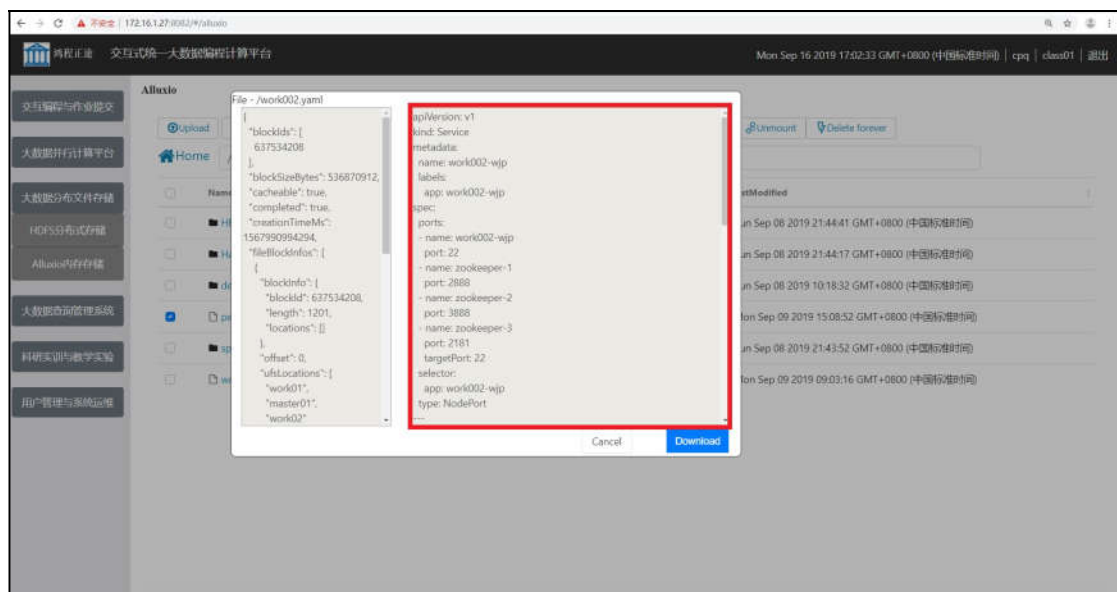


图 21 Alluxio 查看文件详情

4.4 大数据查询管理系统

本节将介绍在本平台下相关数据库工具的用法，包括 HBase、Hive、Presto，点击左侧大数据查询管理系统，选择 SQL On Hadoop。

首先看一下主页面，主要分为两个区域，左侧的展示区和右侧的功能区。

注意：请管理员用户为每个用户手动创建库，库名=用户名。Create database username

1、以操作 Hive 为例：

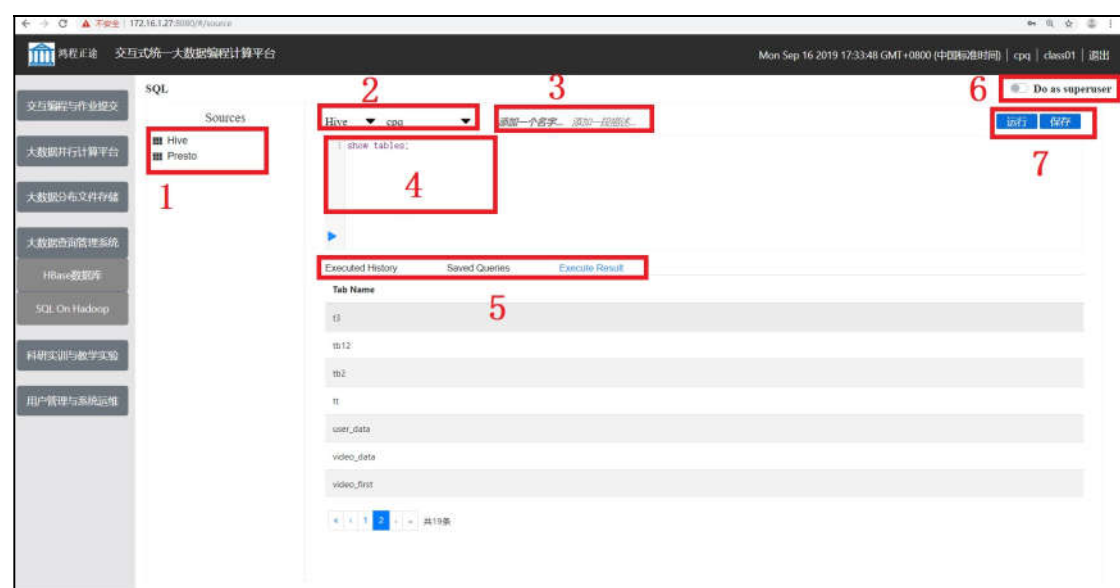


图 22 Hive 首页

1) 标注 1 处为展示区，可以展示 Hive 或者 Presto 下所拥有的库和库里面所有的表。

2) 标注 2 处选择要操作的工具为 Hive 或者 Presto，然后再选择对应的库，我们在这里选择 Hive 和下面的库 cpq。

3) 标注 3 处可以将你执行的语句添加一段说明后，点击 7 处【保存】起来，以便后续查看或者执行。

4) 标注 4 里面输入所要执行的语句，点击 7 处的【运行】按钮运行，结果显示在 5 处的 Execute Result 中。例如我们在这里输入的语句为：show tables，结果显示在图中。

5) 红框 5 中包含 3 个功能，分别为：Executed History、Saved Queries、Execute Result。分别表示执行的历史，保存的语句和执行的结果，另外在 Executed

History 右侧的下载箭头可以将查询结果下载到本地。

6) 在 Executed History 、 Saved Queries 中双击语句可以快速将语句复现在 4 处的执行框中，以便快速的执行语句。

2、分布式 NoSQL 数据库 HBase

HBase 是一种面向列的分布式数据库，在本平台中，我们对其进行了可视化呈现的方式，用户可以方便直观的看到表中的内容以及它们之间的关系。

点击左侧大数据查询管理系统，选择 HBase 数据库，如图所示，是 HBase 的首页展示。

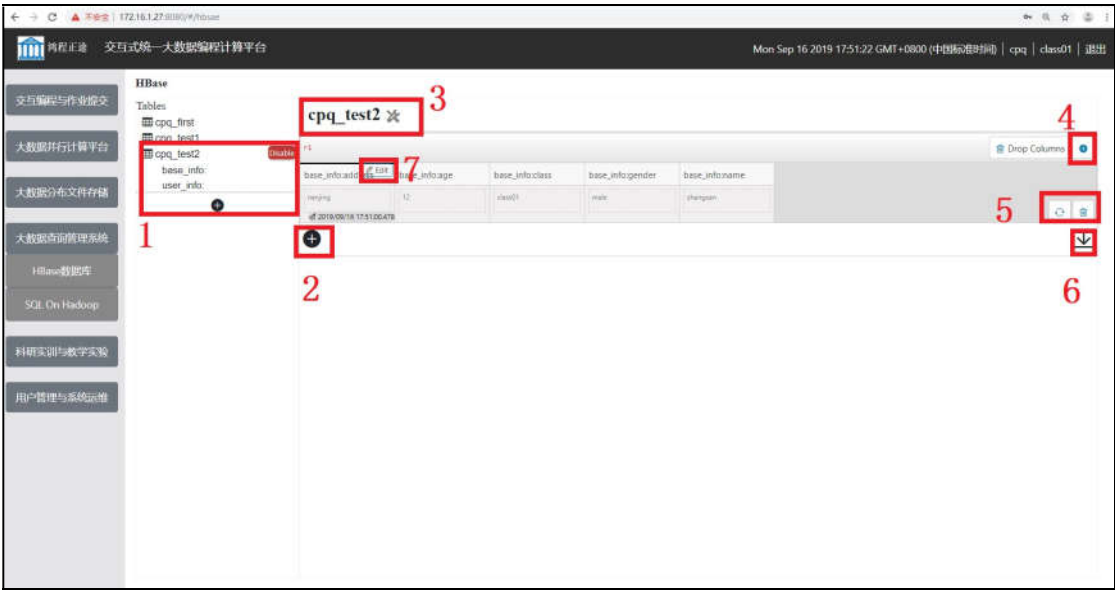


图 23 HBase 首页

主要功能点如下：

- 1) 标注 1 处显示当前用户的所有可用表，可以点击【+】号按钮新增表，或者点击 Disable 将表停用然后删除。
- 2) 表的内容显示在右侧框中，可以点击标注 2 处的+号新增一行。
- 3) 点击标注 3 处的工具按钮可以按条件查询表中的内容。
- 4) 点击标注 4 处的【+】号按钮可以新增一列。
- 5) 点击标注 5 处的按钮可以刷新表内容，或者将这一行删除。
- 6) 点击标注 6 处的标注按钮可以将整张表下载到本地。
- 7) 点击标注 7 处的按钮可以实时更改列值，可以选择保存或者取消。

案例 1：新增一张表

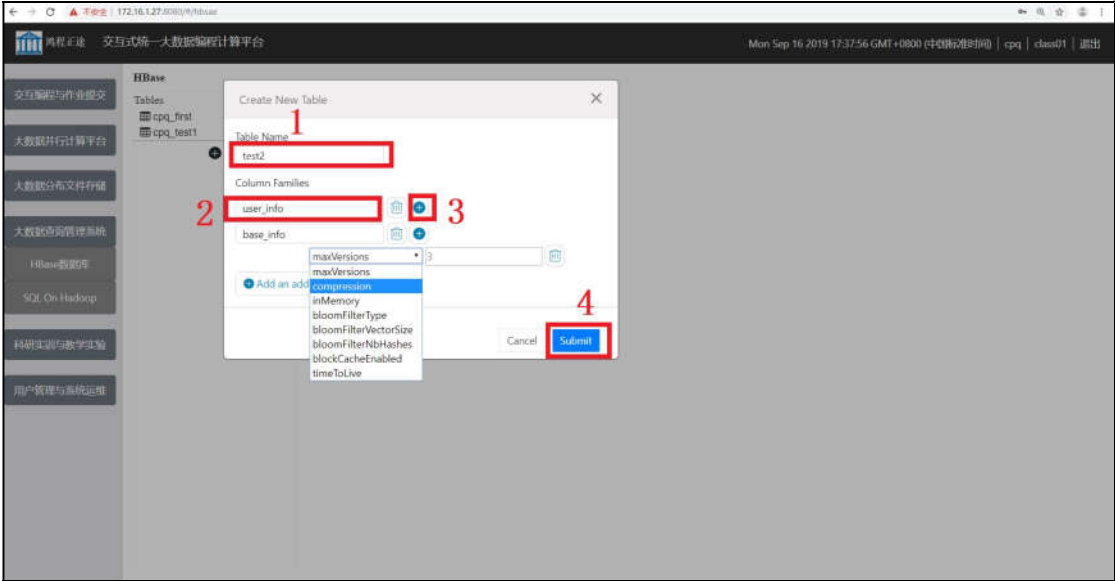


图 24 HBase 建表页

- 1) 标注 1 中为表名，标注 2 中为列族名，可以添加或者删除。
- 2) 点击标注 3 处+号可以为该列族指定条件，然后点击标注 4 处的【Submit】提交。

案例 2：新增一行数据

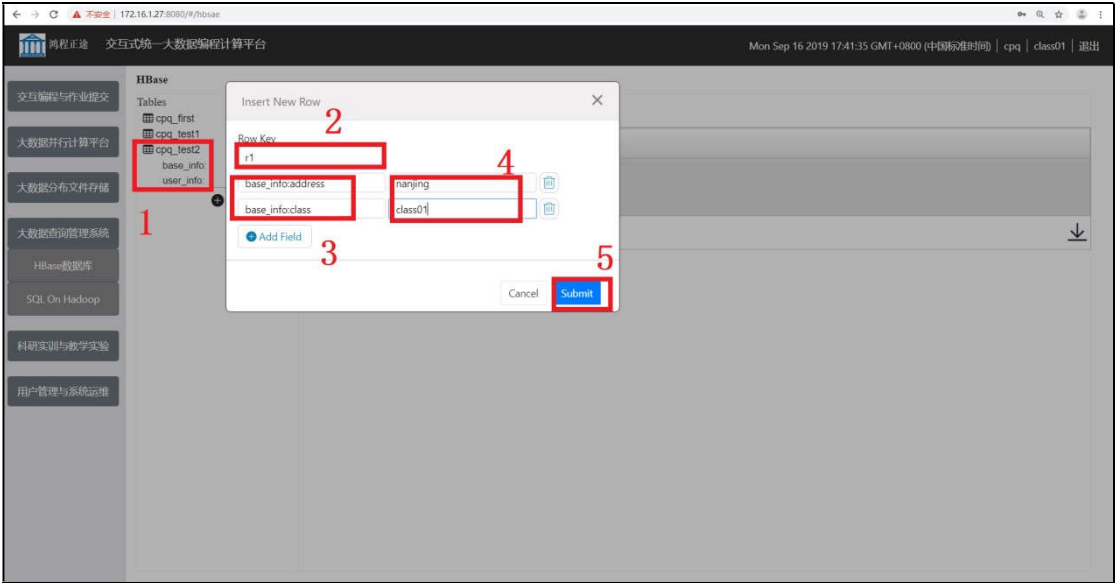


图 25 新增一行数据

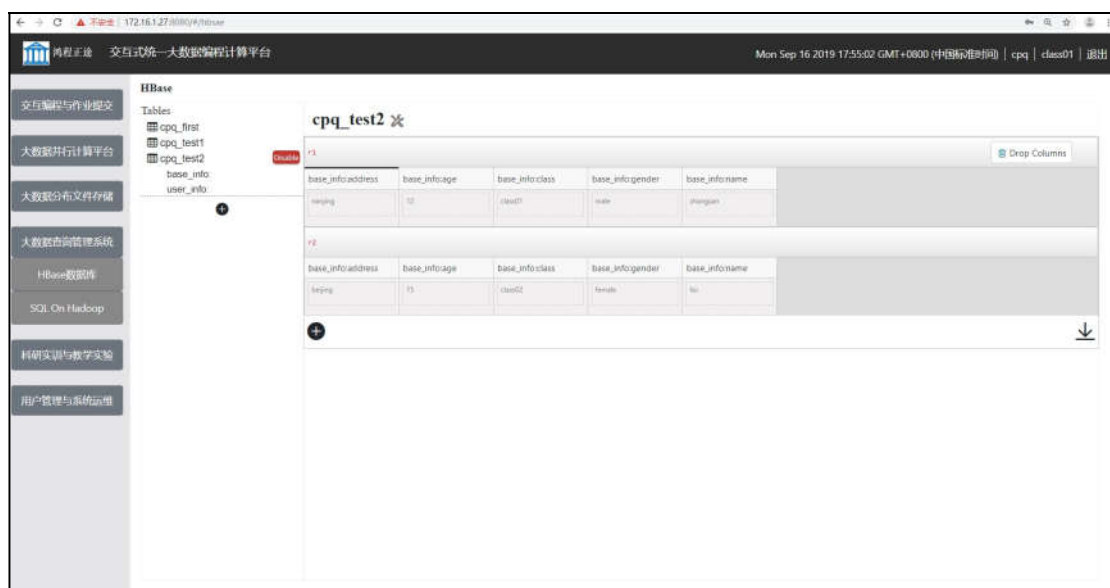
- 1) 可以看到，左侧标注 1 中新增了一张 cpq_test2 表，是案例一新建的表，由于控制不同用户只能使用自己建的表，我们对其在前面指定了用户名的操作。

- 2) 标注 2 中为指定的行键名。
- 3) 标注 3 中为列族名和列名，格式为：**列族名：列名**。
- 4) 标注 4 中输入列的值，然后点击 5 处的【Submit】提交。

案例 3：条件查询

对新增的 cpq_test2 表的内容进行条件查询。

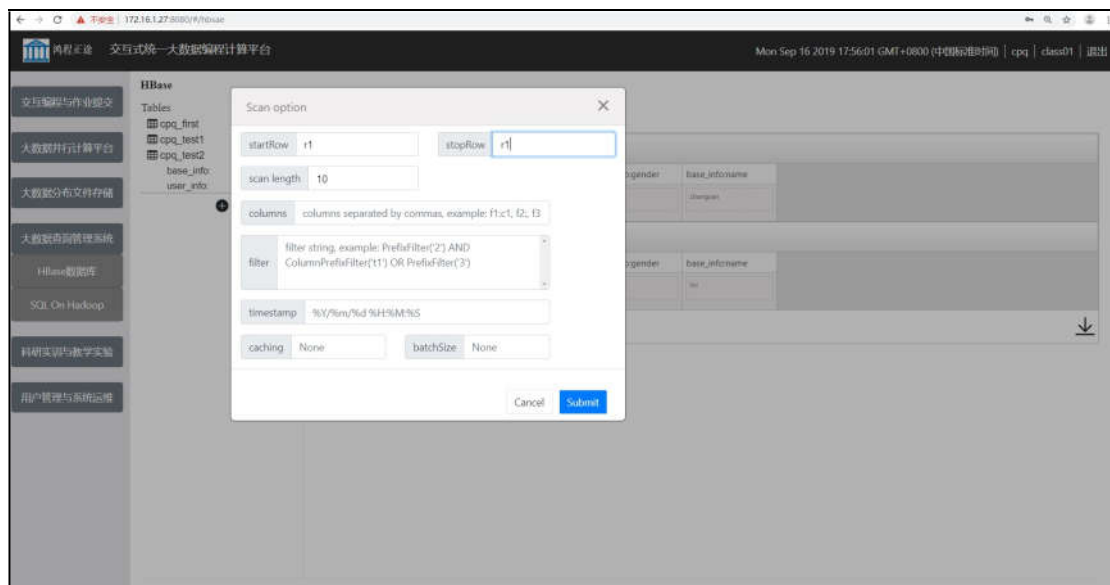
查询前：



base_info.address	base_info.age	base_info.class	base_info.gender	base_info.name
nanjing	12	class01	male	zhongqian
nanjing	13	class02	female	liu

图 20 查询前表内容

指定查询条件：只查询第一行数据



Scan option

startRow: r1 stopRow: r1

scan length: 10

columns: columns separated by commas, example: f1:c1, f2, f3

filter: filter string, example: PrefixFilter('2') AND ColumnPrefixFilter('11') OR PrefixFilter('3')

timestamp: %Y/%m/%d %H:%M:%S

caching: None batchSize: None

Cancel Submit

图 26 指定查询条件图

查询结果如下：

Figure 27 shows the query results for the 'cpq_test2' table in the HBase web interface. The table structure is as follows:

base_info.address	base_info.age	base_info.class	base_info.gender	base_info.name
wangqiang	11	class01	male	zhangsan

图 27 查询结果图

案例 4：下载表内容

提供三种格式的下载，分别为：CSV、Excel、JSON

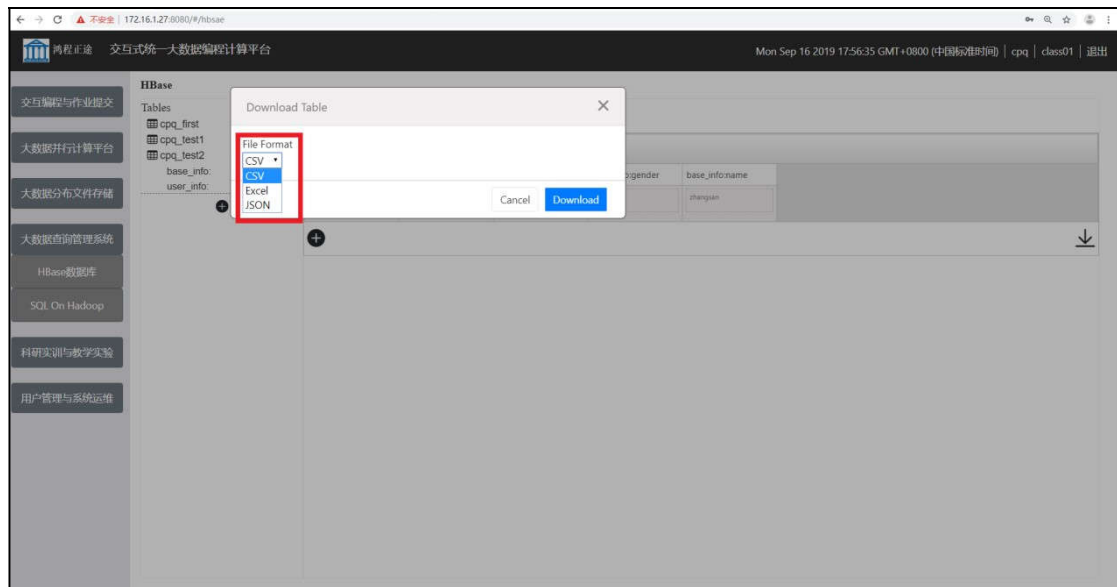


图 28 Hbase 下载表内容

4.5 科研实训与教学实验

本模块包含两个部分，分别是实验手册与编程 API 模块。本平台提供了大量的实验手册，从开发编程语言教程到大数据系统实验、再到大数据行业案例，一共包含了数百个文档。

编程 API 提供了包括 Python、Hadoop、Spark 在内的官方文档，用户在学习这些技术的过程中，如果遇到任何问题，都可以通过查看这些官方文档来解决。

1、实验手册

如下图 29 所示：

符号 2 处的框内可以选择相应的实验列表并查看；

符号 3 处的框内，点击文件名可以下载文件到本地；

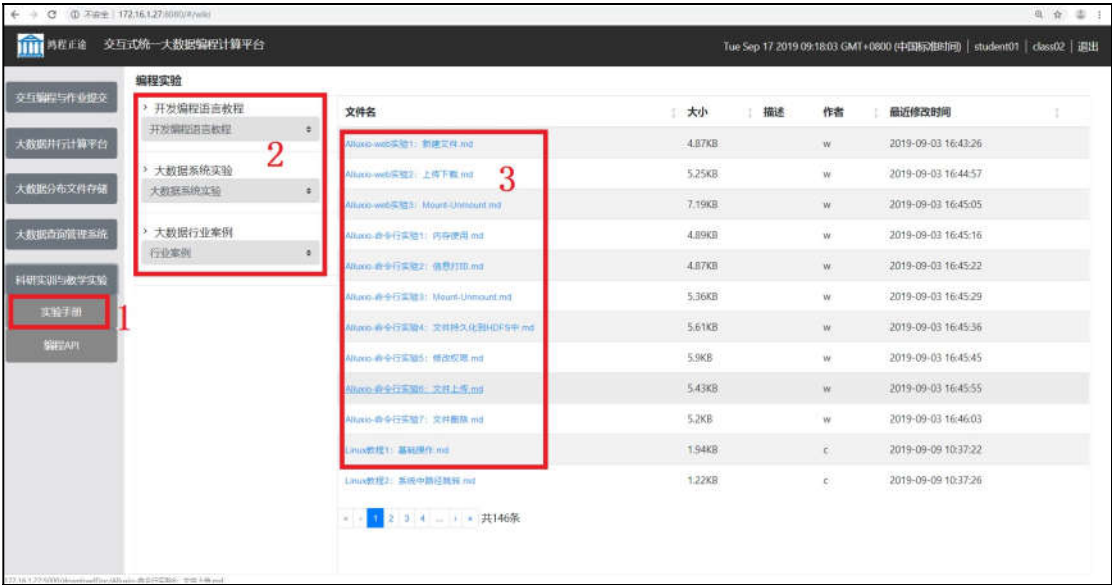


图 29 实验手册页

如下图 30 所示，在符号 1 处选择 SQL 教程，在 2 处会显示所有 SQL 教程的列表，点击教程名字会在 3 处显示具体的文本内容。



图 30 实验内容页

2、编程 API

如下图 31 所示，在符号 1 处选择编程 API，然后在 2 处选择 API 文档，就可以查看相应的官方文档。

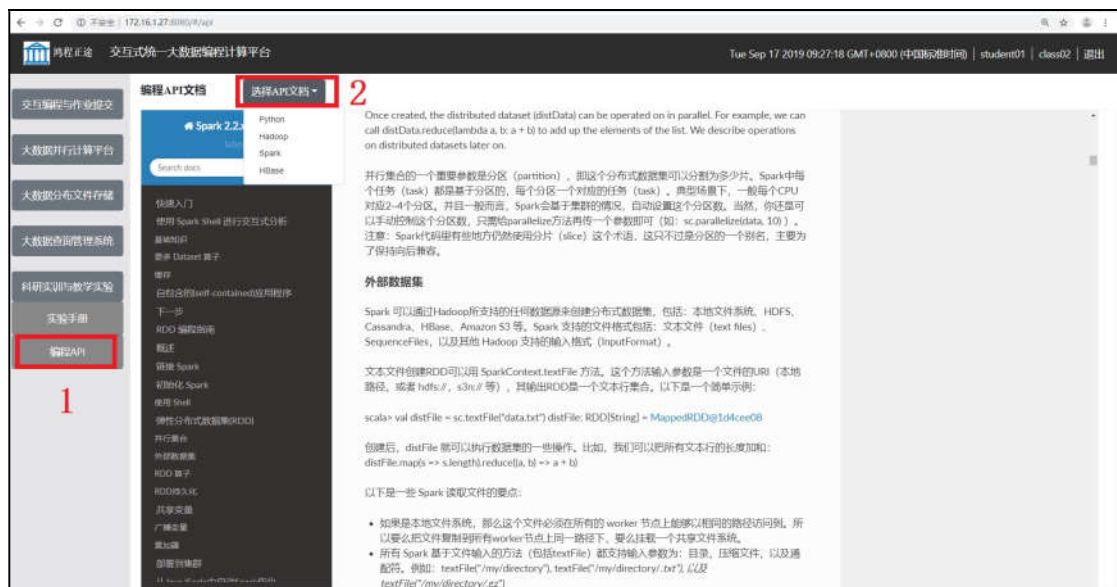


图 31 编程 API 页