PaaS 平台使用手册

中国电信云公司大数据 PaaS 创新中心

目录

第一章 前言

1.1 目的

中国电信大数据 PaaS(Platform as-a Service)平台是基于开源的架构，面向政企客户和电信内部客户提供大数据收集、存储、计算、分析、展示及管理的能力，通过大数据 PaaS 平台的能力使客户可以很容易的使用开发大数据产品，并且快速开发基于大数据的应用。

本手册内容为大数据 PaaS 平台的用户操作手册，主要包括快速操作指南、Hadoop 集群操作及 Map Reduce 应用创建和使用说明等内容。

1.2 目标读者

本手册内容为支撑算法应用大赛的 PaaS 平台操作手册，请参赛者仔细阅读。

1.3 支持版本

支持版本	中国电信大数据 PaaS V1.1

1.4 阅读前提

阅读本用户手册必须具备如下知识：

1. 了解计算机相关概念和技术
2. 熟练使用 Linux 系统
3. 熟悉 Hadoop 技术栈相关技术

1.5 用户须知

1. **PaaS 平台使用开源 Apache Hadoop 2.6.0 版本；**
2. **PaaS 平台使用 JDK 统一版本为 1.7.0_79；**
3. **用户打包.jar 包时需要指定 mainclass；**
4. 算法大赛 wei 每个选手都是三节点集群，集群有 3 个工作节点；

5. 每个工作节点配置在 yarn 中可用资源为：CPU：4vcore，内存：12G；
6. 每个 container 最小内存为：3G，最大为：12G；
7. 默认的 containers 使用内存大小为：map：3G，reduce：12G；
8. mapreduce.job.reduces (reduce 个数)参数为：3。

第二章 快速操作指南

2.1 登录算法应用大赛官网

1. 在浏览器输入算法应用大赛地址“http://bdg.ctyun.cn/”，页面“登录”按钮如下图所示：

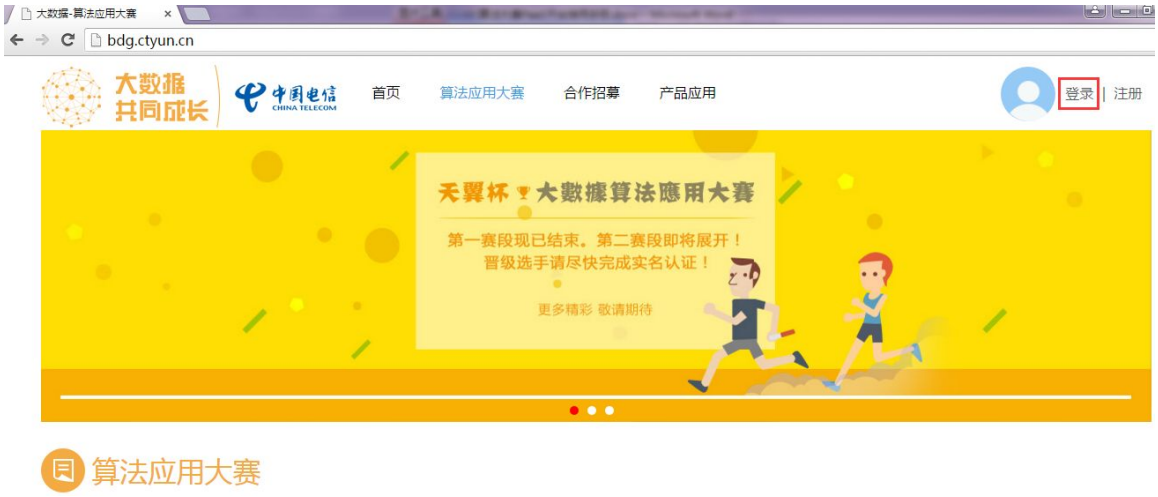

图 1 算法应用大赛登录界面

2. 点击“登录”按钮，在登录页面中输入用户名、密码、验证码（**仅支持队长 ID 登录，且只能同时在 1 台机器上登录**），点击“登录”按钮，如忘记密码，可通过“忘记密码”进行密码重置，登录页面如下图所示：

图 2 算法大赛网站登录页面

2.2 跳转至大数据 PaaS 平台

参数选手登录算法应用大赛官网后，有 2 个途径跳转至 PaaS 平台。分别如下：

1. 通过个人中心跳转：

1.登录算法应用大赛官方网后，默认展示“首页>个人中心”列表中的“个人信息”页；在默认展示列表中，点击“第二赛段入口”按钮，即可进入大数据 PaaS 平台介绍页面，按钮如下图所示：

图 3 登录算法应用大赛页面-1

2. 在大数据 PaaS 平台介绍页面，选手可下载浏览 PaaS 相关介绍文档及选手操作手册。点击“PaaS 平台入口”按钮，即可进入平台页面，如下图所示：

2.通过“算法大赛”页面跳转：

参赛选手登录算法应用大赛官网后，如从默认页面切换过其他页面，需要再次进入“算法应用大赛”页面，进行大数据 PaaS 平台转跳，具体操作步骤如下：

1. 在算法应用大赛官网，点击“算法应用大赛”页面链接，如下图：

图 4 首页中算法应用大赛

2. 进入算法应用大赛分页后，点击“比赛日程”按钮，如下图：

图 5 算法应用大赛-比赛日程按钮

3. 在赛制介绍页面，点击“第二赛段入口”按钮，页面跳转大数据 PaaS 平台介绍页面至如下图点击“PaaS 平台入口”按钮即可进入平台：

图 6 赛制介绍页面中第二赛段入口跳转按钮

通过以上 2 种途径，点击“PaaS 平台”按钮，在新页面中展示大数据 PaaS 平台用户概览界面：

PaaS 平台概览页面展示信息，简介如下：

A. 首行-展示 PaaS 平台标识及文字“中国电信大数据云-飞龙”；

B. 右上角-显示参赛选手队名；

C. 左侧-列表形式展示 PaaS 平台整体功能菜单栏；

D. 居中-图形化显示 PaaS 平台整体资源情况；

E. 底栏-版权信息：@2015 中国电信云计算分公司版权所有 京 ICP 备。

PaaS 平台具体页面展示，如下图所示：

图 7 PaaS 平台概览界面

2.3 Hadoop 集群基本操作

跳转至 PaaS 平台后，参赛选手需查看集群服务，打开 HDFS 文件存储，确认比赛所需分析的数据源。具体操作方式介绍如下：

1. 点击左侧菜单栏“我的服务实例-> 大数据实例”按钮进行查看，菜单栏位置如下图所示：

图 8 Hadoop 集群查看菜单

2. 在大数据服务实例列表中，选择具体集群的大数据工具集列表中的“HDFS Browser”按钮，可进入 HDFS Browser 页面，按钮如下图：

图 9 Hadoop 集群中 HDFS Browser

3. 进入 Hadoop 集群 HDFS 的指定目录下（game-data），下载比赛源数据，具体路径如下：

图 10 比赛所需分析源数据目录

4. 在 HDFS 文件列表中选定文件夹，点击文件内的具体某个文件，页面弹出 HDFS 文件下载窗口，点击页面中的“Download”按钮进行本地下载，下载文件大小没有限制，具体操作如下图所示：

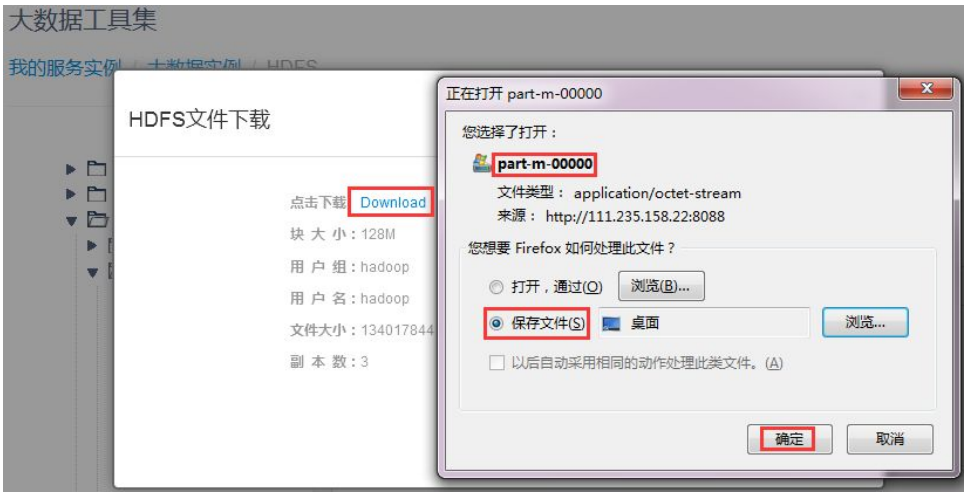

图 11 HDFS 文件下载页面

2.4 MapReduce 应用创建

创建前的准备工作，分为两步：

a) 确定待处理的数据文件在 HDFS 的具体对应目录；

b) 自编程的 Java 程序已按要求完成打包。

创建立即执行的简单 MR 过程，分步骤介绍如下：

1. 点击左侧菜单栏的“应用列表-> hadoop 应用”，如下图所示：

图 12 菜单栏中 hadoop 应用

2. 进入 MapReduce 应用创建界面，如下图所示：

图 13 hadoop 应用可申请页面

3. 点击列表中 MapReduce 应用信息行对应的“创建”按钮，进入创建应用步骤 1“上传应用”界面；

A. 选择应用部署的 Hadoop 服务实例，下拉列表中选择；

B. 输入 MapReduce 应用的名称；

C. 选择应用包对应文件所在的存储路径，进行应用包上传（注：应用文件需*.jar 格式，**用户打包.jar 包时需要指定 mainclass**；

D. PaaS 平台 Hadoop 版本是 Hadoop2.6.0，请基于 Hadoop2.6.0 编写 MapReduce 程序）。

图 14 创建 MapReduce 应用步骤-1

4. 点击下一步，进入步骤 2“参数配置”界面，分别设置 input，output 路径，如下图所示：

A. 设置 input 路径：

点击“加载 HDFS”选择 HDFS 目录中具体路径下的数据文件；

图 15 MapReduce 应用 input 路径

B. 设置 output 路径：

点击“加载 HDFS”选择 HDFS 目录中的具体路径，再填写新文件夹名称，点击“确定”按钮，完成新路径创建；

注意：输出路径不能重复。

图 16 MapReduce 应用 output 路径

5. 点击“下一步”，进入步骤 3“应用参数”设置界面，如下图所示：

图 17 应用参数设置界面

注意 1：普通 MR 不勾选调度设置；

注意 2：mapreduce.map.class 和 mapredce.reduce.class 处填写 jar 包中实际的 mapclass 和 reduceclass 名称。

4 个参数填写完成后，如下图所示：

图 18 普通 MR 填写 class 后显示

6. 点击“下一步，”进入步骤 4“创建应用”界面，如下图所示：

图 19 确认 MR 应用配置界面

创建 MR 应用之前，再次确认应用配置，具体包括：应用名称；应用文件；应用参数等。

7. 确认信息后，点击“完成”按钮，等待应用创建完成，提示信息如下：

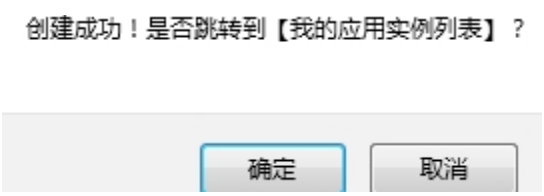

图 20 应用创建成功提示

以下为 wordCount 示例，仅供参考：

```
public class WordCountMain
{
   public static void main(String[] args)
      throws IOException, InterruptedException, ClassNotFoundException
   {
      Configuration conf = new Configuration();
      Job job = Job.getInstance(conf);
      job.setJarByClass(WordCountMain.class);

      job.setMapperClass(WordCountMapper.class);
      job.setMapOutputKeyClass(Text.class);
      job.setMapOutputValueClass(LongWritable.class);

      FileInputFormat.setInputPaths(job, new Path[] { new Path("hdfs://mycluster" +
args[0]) });

      job.setReducerClass(WordCountReduce.class);
      job.setMapOutputKeyClass(Text.class);
      job.setMapOutputValueClass(LongWritable.class);

      FileOutputFormat.setOutputPath(job, new Path("hdfs://mycluster" + args[1]));
```

```
        job.waitForCompletion(true);
    }
}
```

MR Mapper 程序：

```
public class WordCountMapper extends Mapper<LongWritable, Text, Text, LongWritable>
{
    Text wordText = new Text();
    LongWritable valueLong = new LongWritable();

    protected void map(LongWritable key, Text value, Mapper<LongWritable, Text, Text,
LongWritable>.Context context)
        throws IOException, InterruptedException
    {
        String line = value.toString();
        String[] words = line.split(",");
        for (int i = 0; i < words.length; ++i) {
            String word = words[i];
            this.wordText.set(word);
            this.valueLong.set(1L);
            context.write(this.wordText, this.valueLong);
        }
    }
}
```

MR Reduce 程序：

```
public class WordCountReduce extends Reducer<Text, LongWritable, Text, LongWritable>
{
```

```
    LongWritable valueLong = new LongWritable();
    protected void reduce(Text key, Iterable<LongWritable> values, Reducer<Text,
LongWritable, Text, LongWritable>.Context context)
        throws IOException, InterruptedException
    {
        long counter = 0L;
        for (LongWritable l : values) {
            counter += l.get();
        }
        context.write(key, new LongWritable(counter));
    }
}
```

2.5 MapReduce 任务状态查看以及结果文件下载

参赛选手使用 PaaS 提交 MR 任务后，可以点击菜单栏“我的应用实例”-> “hadoop 应用实例”查看任务的运行状态。（注：MR 任务的执行结果可在页面上直接下载。）

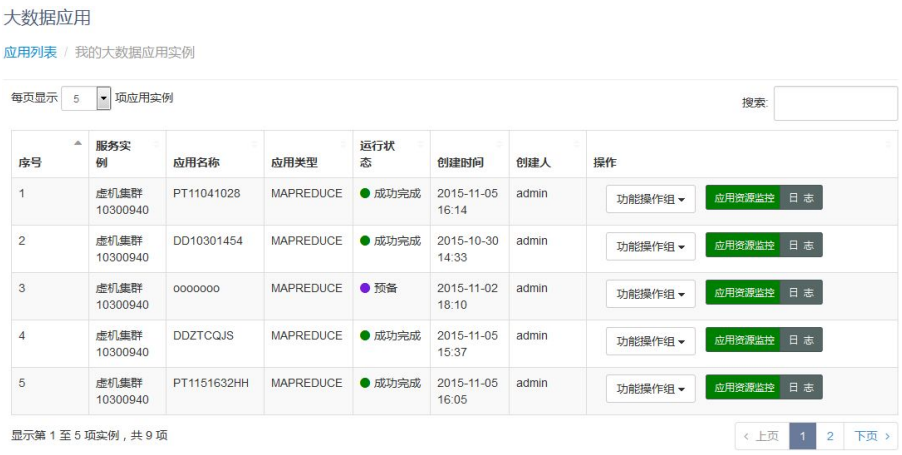

图 21 大数据应用实例列表中查看 MR 运行状态

2.6 MapReduce 执行结果如何上传到大数据共同成长平台

1.选手将结果从 PaaS 上下载下来，进入大数据共同成长平台，点击算法应用大赛>比赛日程。

2.点击上传预测结果，选取上传文件，点击上传后，完成上传，上传完成后可在个人中心>比赛信息中查看。

第三章 Hadoop 集群服务

PaaS 平台提供一整套大数据集群 Hadoop 服务，当前版本增加了对虚拟机集群的支撑。具体集群操作分章节介绍。

当前算法应用大赛组织方，已为每个参赛队部署完成 3 节点的 Hadoop 集群服务。

3.1 Hadoop 集群服务

对应各自的服务实例，可通过点击左侧菜单栏“我的服务实例-> 大数据实例”按钮，进行查看，菜单栏位置如下图所示：

图 22 Hadoop 集群查看菜单

点击“大数据实例”按钮后，进入“我的服务实例”列表，依次呈现：服务实例名称，服务类型，创建时间，服务状态，操作等信息。如下图所示：

图 23Hadoop 集群服务列表

3.2 Hadoop 集群资源监控

Hadoop 集群提供基于 Ambari 的资源监控信息查看，资源监控中主要检测集群整体最近 1 小时内的数据测量值变化。

点击服务实例列表中“操作”栏内的“资源监控组”下拉框按钮，选择“资源监控”按钮，操作按钮如下图所示：

图 24 集群资源监控按钮

点击 Hadoop 集群对应的“资源监控”按钮，页面即可查看资源监控界面，如下图所示：

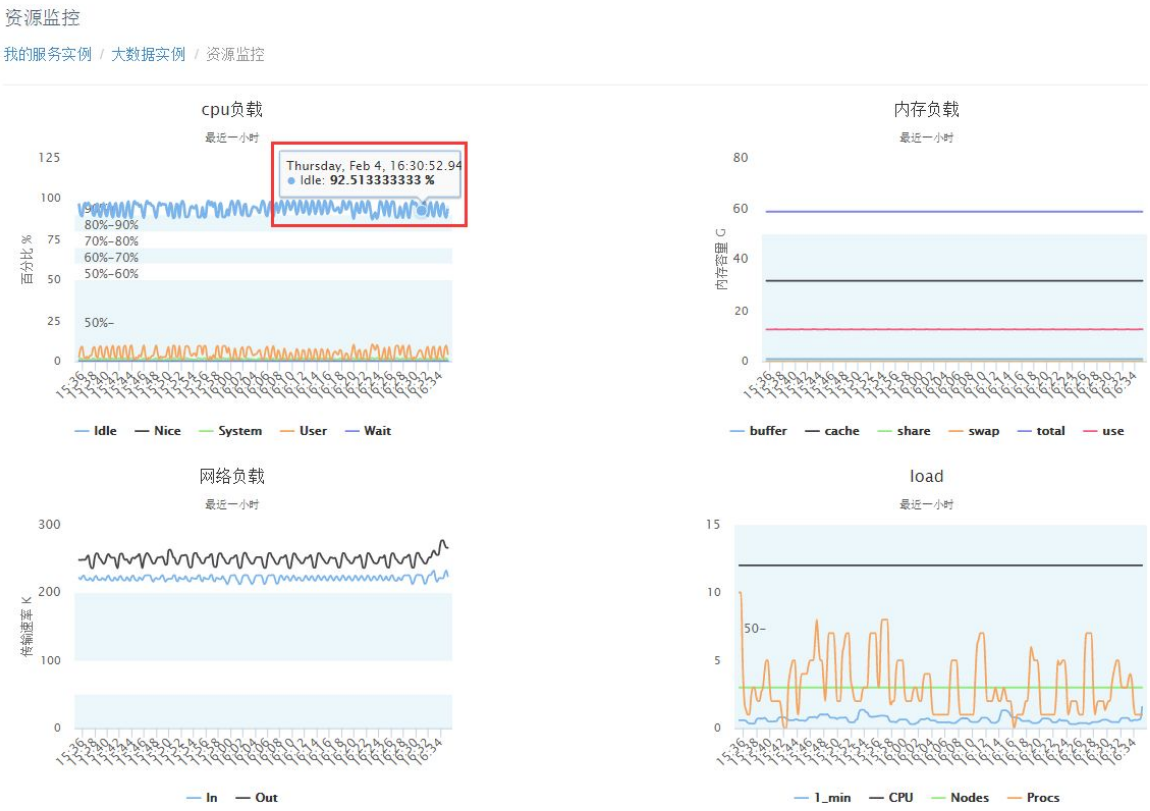

图 25 Hadoop 集群资源监控界面

测量项包括：CPU 负载，内存负载，网络 I/O 负载，集群负载（CPU 核数，Name Nodes

数量，进程数等），HDFS 磁盘存储信息等。

图表中数据可实时查看测量值，检测集群整体最近 1 小时内的数据测量值变化，光标放置图表中测量点时可查看该处具体的测量数值。

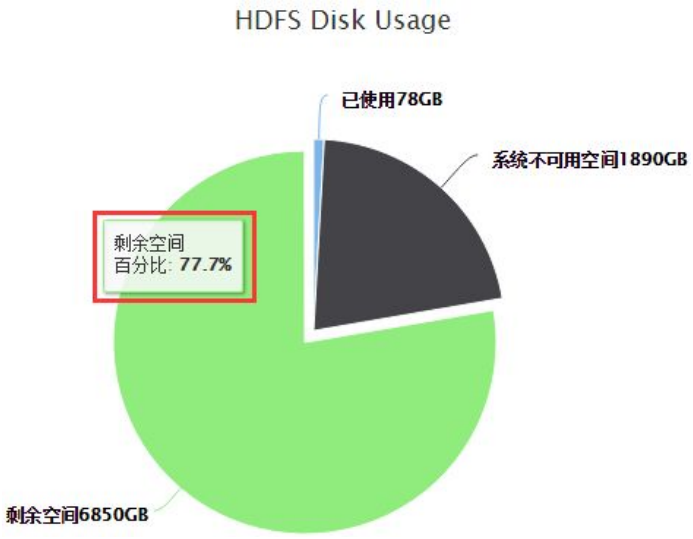

图 26 Hadoop 集群资源监控界面-HDFS Disk 使用情况

3.3 Hadoop 服务大数据工具集

PaaS 平台满足用户对集群中存储数据的多种分析处理功能，提供大数据工具集服务，大数据工具集主要包括：HDFS Browser，Hive explorer 工具，详细功能分章节介绍。

点击服务实例操作列表中点击“大数据工具集”，即可展开工具集列表，如下图所示：

图 27 Hadoop 集群大数据工具集列表

3.3.1 HDFS Browser(HDFS 浏览器)

在大数据服务实例列表中，选择具体集群的大数据工具集列表中的“HDFS Browser”按钮，可进入 HDFS Browser 页面。

图 28 大数据工具集中 HDFS Browser 项

3.3.1.1 浏览 HDFS 文件系统

1. HDFS Browser 文件浏览，可在页面左侧以树状图形式展示 HDFS 中各文件夹结构及内容，可点击文件夹前面的箭头进行展开和收起操作，当前目录可显示在右侧的“当前目录”栏内，如下图所示：

图 29 HDFS 文件浏览界面

2. 文件系统除了浏览文件列表，还可自定义新建文件夹。通过文件浏览确定文件夹的父目录，在“子目录”输入框内填写自定义文件名，再点击“创建文件夹”按钮，新文件夹可实时创建，HDFS 文件列表中也可自动刷新显示。操作过程如下图：

图 30 创建文件夹操作页面

点击“创建文件夹”按钮后，页面弹出提示信息，如下：

图 31 创建文件夹提示信息

点击“确定”按钮，系统弹出文件创建成功提示信息，如下：

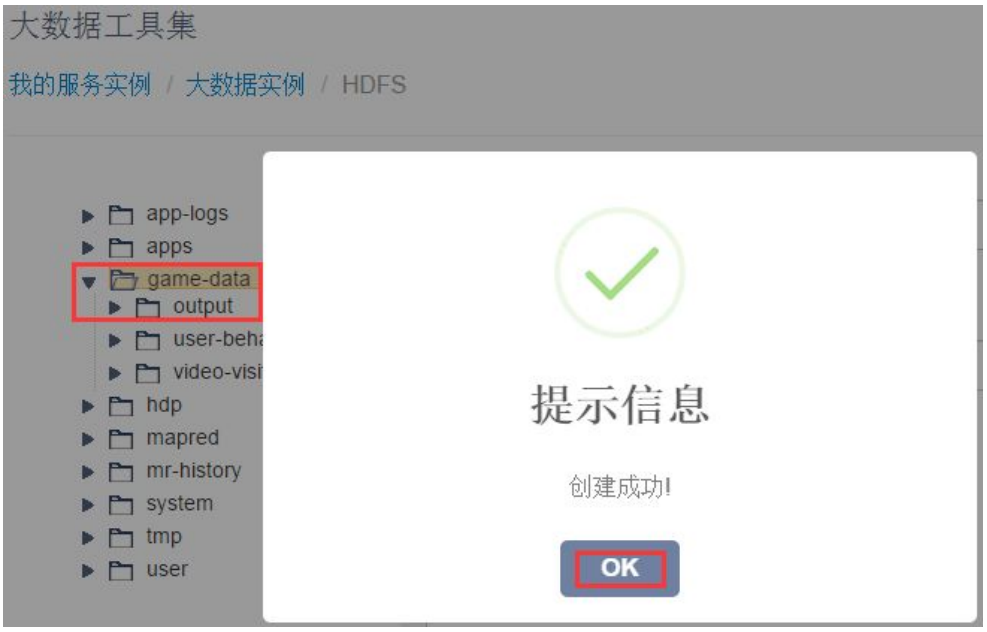

图 32 文件夹创建成功提示

点击“OK”按钮后，查看 HDFS 文件列表，新建文件夹已可显示在目标路径下。

图 33 新文件夹创建成功后显示

3.3.1.2 文件上传与下载

1. 文件上传操作：在 HDFS 文件列表中选定文件夹，点击“上传文件”，选择目标文件，进行文件上传，如下图所示：

图 34 HDFS 上传文件操作页面

选择本地具体文件进行上传操作，上传进度和完成会有提示信息。

注：当前平台上传文件要求大小不超过 20M。

2. 文件下载操作：在 HDFS 文件列表中选定文件夹，点击文件内的具体某个文件，页面弹出 HDFS 文件下载窗口，点击页面中的“Download”按钮进行本地下载，下载文件大小没有显示，具体操作如下图所示：

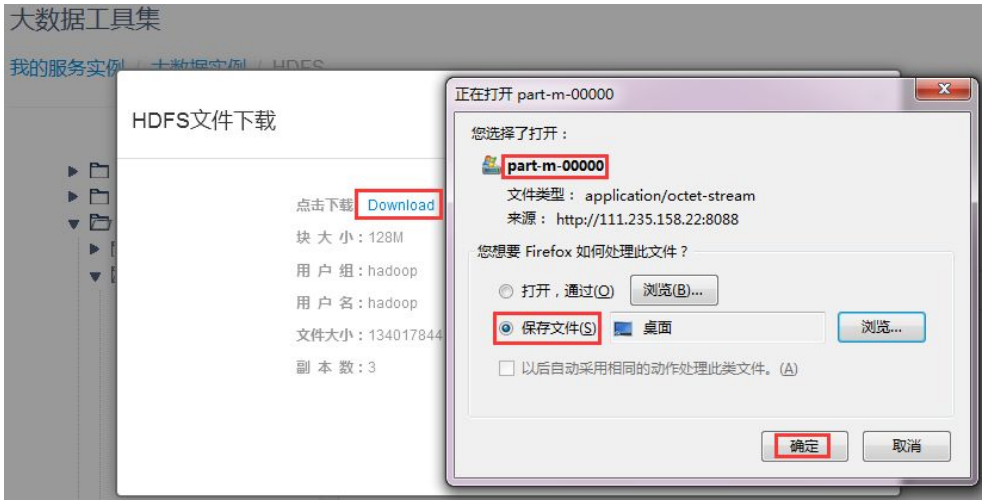

图 35 HDFS 文件下载页面

3.3.2 Hive Explorer(Hive 探索工具)

在大数据服务实例列表中，选择具体集群的大数据工具集列表中的“HDFS Browser”按钮，进入 HDFS Browser 页面。

图 36 大数据工具集中 Hive Explorer 项

3.3.2.1 浏览 Hive 数据仓库

Hive Explorer 页面左侧“选择数据库”的下拉列表可呈现用户创建的不同数据库，默认为“default”数据库，且数据库内无数据表，可点击下拉框选择进入不同数据库，查看数据库内包含表信息。

图 37 Hive 中数据库选择默认为 default 且为空

3.3.2.2 基于 Hive SQL 的查询

Hive SQL 支持操作主要包括：创建数据库，创建普通的内部表，创建外部表，建表同时导入 HDFS 中数据，创建 csv 格式表并导出数据，个性化 SQL 查询等。

选定操作数据库，可在查询输入窗口可输入自定义 Hive 语句并执行，输入完成后点击“执行”按钮即可。

Hive Explorer 还支持查询输入框内容重置，内容保存，限制查询条数，操作记录中 SQL 语句复用。

注 1：查询输入窗口中编辑 SQL 语句时，具体表名可点击左侧列表中数据表获得；

注 2：查询输入框中 SQL 语句结尾不能有分号“;”

注 3：当查询到数据较多时，数据列较多时可通过拖动上下、左右滚动条查看查询结果。

1. Hive 中建表，并导入 HDFS 中数据，举例执行页面如下：

注意：LOCATION ‘’ 内需填写 HDFS 中的具体路径。

图 38 Hive 中创建表并导入 HDFS 数据

2. 点击“执行”按钮后，执行成功提示如下：

图 39 操作成功后提示信息

3. 表创建完成后，刷新页面，选择数据库下拉列表中某个数据库，并可查看其中已创建表的列信息，列数量，列名称等信息，如下图：

图 40 查看创建完成的数据表

4. 执行简单查询后，结果展示如下：

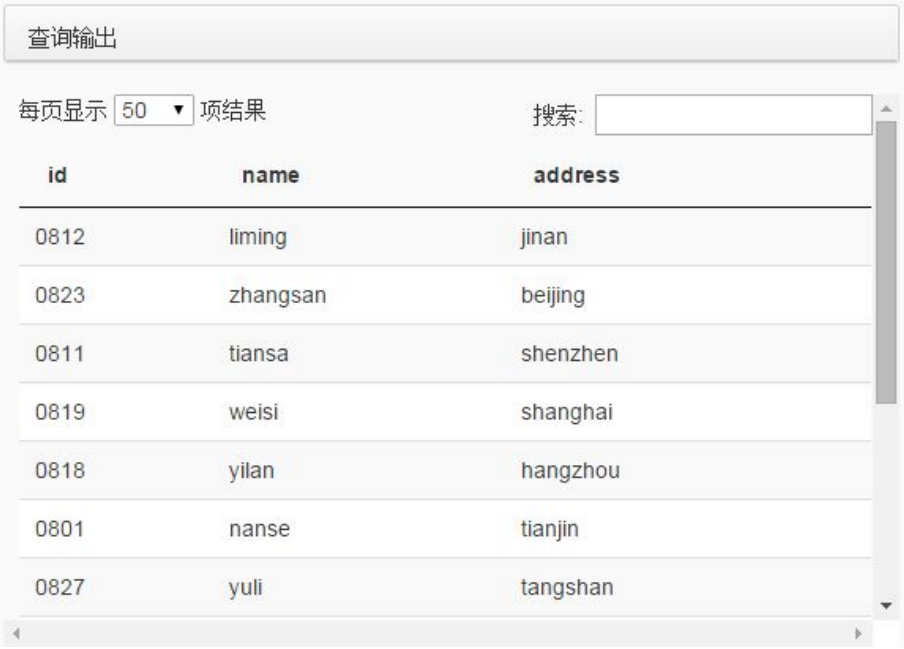

图 41 查询结果输出界面

5. 在操作窗口中输入的 Hive 语句支持保存功能，点击“保存”按钮，弹出 hive sql 保存为 txt 文件的提示框，如下图所示：

图 42 Hive SQL 语句保存提示框

6. 操作记录表中可记录用户对数据库的操作信息：

注 1：通过“重置”按钮，可清空查询输入窗口中的 SQL 语句；

注 2：点击“操作记录”中的任意 sql 语句，可将其自动复制到查询输入窗口中；

图 43 Hive 操作记录表

3.3.2.3 支持 Hive UDF

用户编写的 UDF 程序，可在 Hive Explorer 页面添加并在查询输入窗口中使用。具体操作如下：

1. 点击 Hive Explorer 界面右侧 UDF “+增加”按钮；

图 44 UDF 添加按钮

2. 展开 UDF 添加模块，点击“上传文件”按钮，在弹出选择路径中选择具体的 UDF 程序；

图 45 上传 UDF 文件按钮

3. 选择自定义的 UDF jar 包进行上传后，上传进度有提示显示；

图 46 UDF 文件上传进度

4. UDF 上传成功后，在“UDF 类全路径”框内输入该 UDF 文件中的类名全路径，如下图中 UDF 的类全路径为：hadoop.HiveUDF

图 47 查看 UDF 中类的全路径

在“UDF 类全路径”框内输入效果如下图：

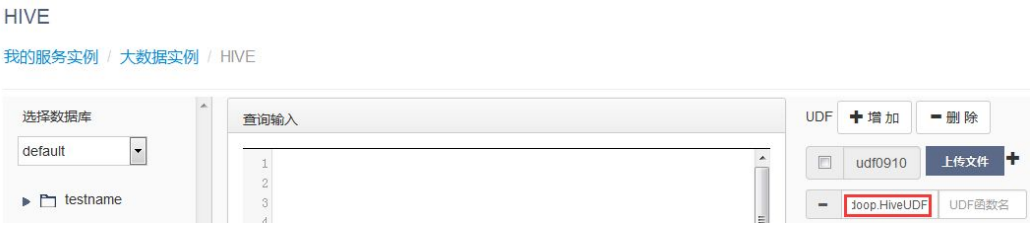

图 48 将 UDF 类全路径写入框内

5. 并在“UDF 函数名”框内填写用户自定义的函数名称；

图 49 输入自定义函数名

6. 勾选 UDF 文件前的可勾选框，并在查询窗口输入包含 UDF 自定义函数的 SQL 语句；

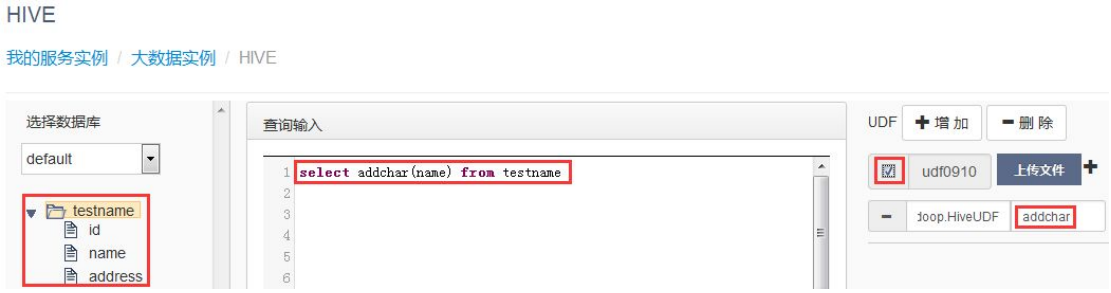

图 50 输入 SQL 语句包含 UDF 函数

7. 点击“执行”按钮，SQL 语句被执行，支持完成后，查看查询结果；

图 51 使用简单 UDF 函数查询结果

8. 如 1 个 UDF 中含有多个 class，可点击“上传文件”后的“+”按钮添加多个 class，并分别命名自定义函数名，SQL 语句中可使用多个函数；

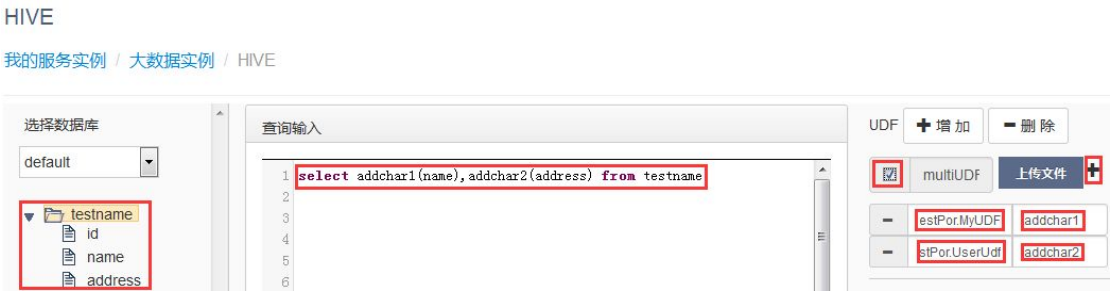

图 52 UDF 中含有多个函数

图 53 SQL 中含有多个函数的查询结果

9. UDF 添加后，可点击“—删除”按钮选中 UDF 函数进行删除，删除提示如下图：

图 54 UDF 删除时提示信息

第四章 Hadoop MapReduce 应用

创建前的准备工作：

A. 确定待处理的数据文件在 HDFS 的具体对应目录；

B. 自编程的 Java 程序已按要求完成打包。

4.1 Hadoop MapReduce 应用创建

4.1.1 创建普通 MR

1. 点击左侧菜单栏的“应用列表-> hadoop 应用”

图 55 菜单栏中 hadoop 应用

2. 进入 MapReduce 应用创建界面，如下图所示：

图 56 hadoop 应用可申请页面

3. 点击列表中 MapReduce 应用信息行对应的“创建”按钮，进入创建应用步骤 1“上传应用”界面；

A. 选择应用部署的 Hadoop 服务实例，下拉列表中选择；

B. 输入 MapReduce 应用的名称；

C. 选择应用包对应文件所在的存储路径，进行应用包上传（注：应用文件需*.jar 格式，**用户打包.jar 包时需要指定 mainclass**；

D. PaaS 平台 Hadoop 版本是 Hadoop2.6.0，**请基于 Hadoop2.6.0 编写 MapReduce 程序**。

图 57 创建 MapReduce 应用步骤 1

4. 点击“下一步”，进入步骤 2“参数配置”界面，如下图所示：

图 58 创建 MapReduce 应用步骤 2

设置 input 路径：

点击“加载 HDFS”选择 HDFS 目录中具体路径下的数据文件；

图 59 设置 MapReduce 应用 input 路径

设置 output 路径：

点击“加载 HDFS”选择 HDFS 目录中的具体路径，再填写新文件夹名称，点击“确定”按钮，完成新路径创建；

注意：输出路径不能重复。

图 60 设置 MapReduce 应用 output 路径

5. 点击“下一步”，进入步骤 3“应用参数”设置界面，如下图所示：

图 61 应用参数设置界面

注意 1：普通 MR 不勾选调度设置；

注意 2：mapreduce.map.class 和 mapredce.reduce.class 处填写 jar 包中实际的 mapclass 和 reduceclass 名称。

4 个参数填写完成后，如下图所示：

图 62 填写 4 个 class 后界面

6. 点击“下一步，”进入步骤 4“创建应用”界面，如下图所示：

图 63 确认 MR 应用配置界面

创建 MR 应用之前，再次确认应用配置，具体包括：应用名称；应用文件；应用参数等。

7. 确认信息后，点击“完成”按钮，等待应用创建完成，提示信息如下：

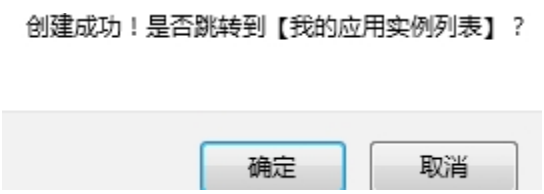

图 64 应用创建成功提示

4.1.2 创建调度 MR

1. 按照普通 MR 创建过程，操作至步骤 5；
2. 在“应用参数”设置页面，分别配置调度和填写 class 信息；
   A. 设置是否启动任务调度配置，通过勾选与非勾选方式进行配置：；
   B. 设置任务调度后，配置任务调度频率：以分钟为单位；
   C. 任务开始、结束时间：通过时间控件选择日期、具体时分秒设置时间；

注：配置任务开始具体时间时，需确认设置时间晚于当前时间。

图 65 MR 任务调度配置频率、日期

图 66 MR 任务调度配置时间

3. 点击“下一步，”进入步骤 4“创建应用”界面，如下图所示：

图 67 确认 MR 应用配置界面

创建 MR 应用之前，再次确认应用配置，具体包括：应用名称；应用文件；应用参数等。

4. 确认信息后，点击“完成”按钮，等待应用创建完成，提示信息如下：

图 68 应用创建成功提示

4.2 Hadoop MapReduce 应用实例操作

MapReduce 创建完成后，可进入应用实例列表查看状态和执行操作，打开 hadoop 应用实例列表路径如下图所示：

图 69 菜单栏中 hadoop 应用实例

点击菜单栏“我的应用实例”-> “hadoop 应用实例”，进入大数据应用实例列表界面，如下图所示：

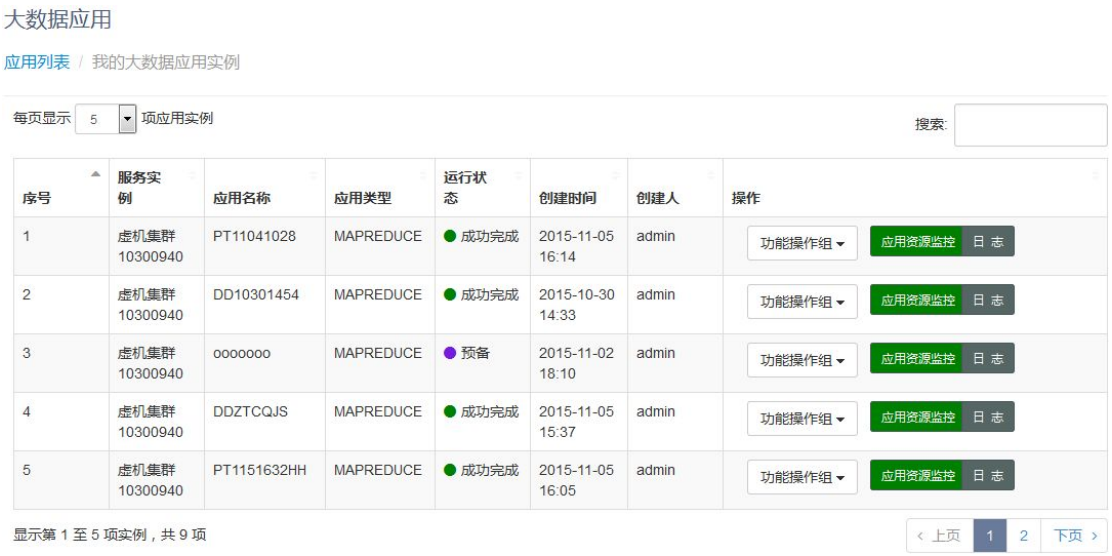

图 70 大数据应用实例列表

1. 应用实例搜索、页面展示数量可选功能：通过列表上方操作区实现，左上方为页面展示数量可调整，右上方为应用实例实时搜索；

2. 功能操作组功能：支持针对每个 hadoop 应用的重配置，重启，定时调度暂停，定时调度唤醒，废弃任务（针对非运行中的应用）操作，如下图所示：

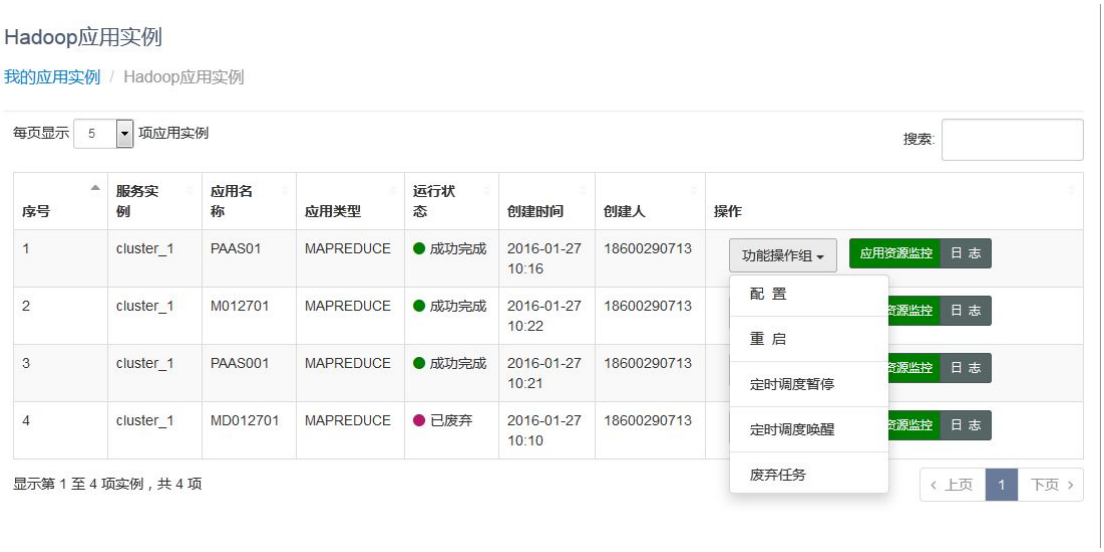

图 71 应用列表中操作列表

Hadoop 应用实例运行状态说明：

- 运行中：正在运行的普通或调度任务。
- 运行完成：成功运行完成的普通或调度任务。
- 调度暂停：已暂停的调度任务。
- 预备：未到达开始运行时间的调度任务。
- 废弃：运行失败或对非运行状态的任务进行了废弃操作。

功能操作组操作说明：

- 配置：对已存在的任务进行重新配置，即修改操作（无法修改应用名称及 jar 包）。
- 重启：对已执行完成的任务进行重启操作，即重新执行任务。
- 定时调度暂停：对执行中的调度任务进行暂停操作。
- 定时调度唤醒：对已暂停的调度任务进行唤醒操作。
- 废弃任务：对非运行中的任务进行废弃操作。

1. 应用资源监控：展示集群中数据节点执行应用的资源使用情况，当前应用以及所有应用的执行情况统计。

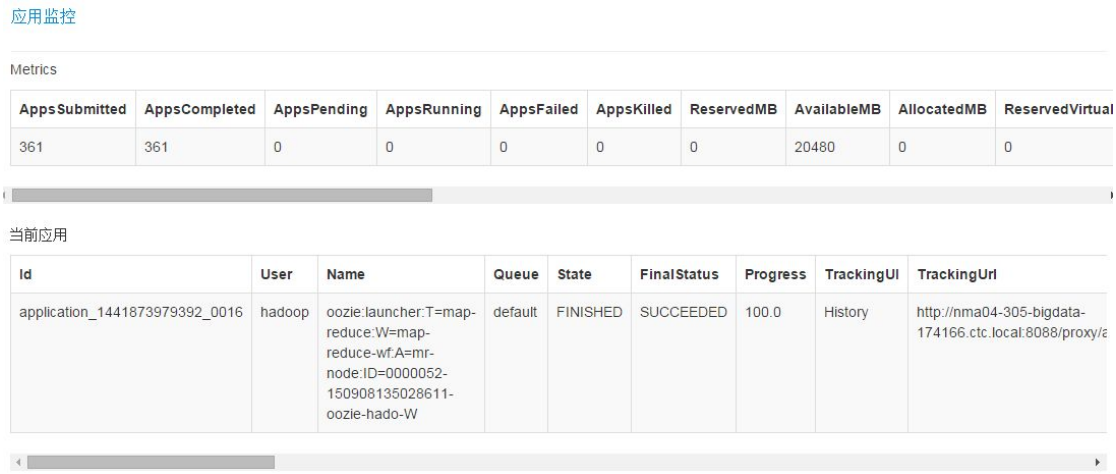

图 72 MR 应用资源监控界面

2. 日志：展示具体应用实例的基本信息，实例详情日志，日志中主要是调度器 oozie 的具体调度日志。

图 73 MR 应用实例详情日志

3. 运行成功后，通过大数据工具集中 HDFS 查看 output 路径下执行结果。包含“_SUCCESS”执行成功标示和“part-r-0000”等结果文件。

第五章　常见问题的说明及解决

此次算法大赛，各参赛选手主要使用 PaaS 平台完成两个功能：①使用 PaaS 平台运行算法程序；②从 PaaS 平台下载算法结果。

为保证比赛的顺利进行，中国电信云公司已对 PaaS 平台进行了充分的测试，并针对一些常见异常问题做了汇总，以供各位选手参考学习。

关于常见问题的详细介绍，请见：http://bdg.ctyun.cn/的“常见问题”。