课堂实验

一、基本信息

文档编号		文档版本	1.0
实验名称	YARN 作业管理		
所属课程	第 4 讲 分布式资源管理框架 YARN	认证等级	数据工程师（初级）
授课形式	上机实验	实验批次	第 4 次 / 共 11 次

二、实验目的

- 掌握在 YARN 上提交和执行作业。
- 掌握 YARN 的作业管理命令。
- 理解 YARN 的工作原理和作业执行过程。

三、实验准备

- 下载并安装 TDH Client。
- 在本地新建 wordcount.txt，文件中输入若干单词，单词间用空格分隔。
- 实验目录与命名规划

  （1）本地目录

  工作目录：/mnt/disk1/{student_name}

  （2）HDFS 目录

  工作目录：/tmp/{student_name}

  YARN 作业输入目录：/tmp/{student_name}/yarn_data/wordcount_input

  YARN 作业输出目录：/tmp/{student_name}/yarn_data/wordcount_output

  （3）{student_name}为变量，代表学员姓名全拼
- 文件服务器

  （1）IP：172.16.140.111

  （2）目录：/mnt/disk1/de_training

四、实验内容

1、启动 TDH Client

- 任务：执行 TDH Client 的 init.sh 脚本，启动 TDH Client。
- 步骤

```
Linux：
  // 执行 TDH Client 的 init.sh 脚本
  1. source {TDH_Client_install_dir}/init.sh
```

2、上传作业输入文件

- 任务：将本地文件 wordcount.txt 上传至 HDFS 作业输入目录中。
- 步骤

```
Linux：
  // 将 Hadoop 当前用户切换为 yarn，进行访问授权
  1. export HADOOP_USER_NAME=yarn
  // 在 HDFS 中创建作业输入目录
  2. hadoop fs -mkdir -p /tmp/{student_name}/yarn_data/wordcount_input
  // 将 wordcount.txt 上传到作业输入目录
  3. hadoop fs -put {wordcount.txt_filepath} /tmp/{student_name}/yarn_data/wordcount_input
```

3、执行作业

- 任务：将 Hadoop Mapreduce 样例程序 hadoop-mapreduce-examples-2.7.2-transwarp-5.1.2.jar 的 WordCount 作业提交给 YARN，并执行得出结果。
- 步骤

```
Linux：
  // 切换目录
  1. cd {TDH_Client_install_dir}/hadoop/hadoop-mapreduce
  /* 向 YARN 提交并执行作业。wordcount 是 main 函数所在类的路径，
     /tmp/{student_name}/yarn_data/wordcount_input 为作业输入目录，
     /tmp/{student_name}/yarn_data/wordcount_output 为作业输出目录 */
  2. hadoop jar hadoop-mapreduce-examples-2.7.2-transwarp-5.1.2.jar wordcount
     /tmp/{student_name}/yarn_data/wordcount_input
     /tmp/{student_name}/yarn_data/wordcount_output
```

4、查看作业输出结果

- 任务：查看 WordCount 作业的输出结果，先看输出目录是否创建，再看生成的输出文件是

否完成了词频统计。

- 步骤

```
Linux：
  // 查看输出目录是否创建
  1. hadoop fs -ls /tmp/{student_name}/yarn_data/wordcount_output
  // 查看输出文件内容
  2. hadoop fs -cat
     /tmp/{student_name}/yarn_data/wordcount_output/{wordcount_output_filename}
```

5、管理作业

- 任务：先向 YARN 提交一个运行时间较长的 tpcds 数据制造作业，再通过 YARN 命令查看和停止作业。
- 步骤

（1）运行 tpcds 造数作业

```
Linux：
  // 登录文件服务器（172.16.140.111），拷贝 tpcds-5.x.tar.gz 到集群节点
  1. scp /mnt/disk1/de_training/tpcds-5.x.tar.gz 172.16.140.85:/mnt/disk1/{student_name}
  // 登录集群节点，解压 tpcds-5.x.tar.gz，生成 tpcds 目录
  2. cd /mnt/disk1/{student_name}
  3. tar -xzvf tpcds-5.x.tar.gz
  // 将 Hadoop 当前用户切换为 hdfs，进行访问授权
  4. export HADOOP_USER_NAME=hdfs
  // 执行造数作业
  5. cd tpcds/bin
  6. ./gen-data.sh
```

（2）管理 YARN 作业

```
Linux：
  // 列出运行的作业，复制 tpcds 作业的 Application Id
  1. yarn application -list
  // 查看 tpcds 作业的状态
  2. yarn application -status {Application_Id}
  // 停止 tpcds 作业
  3. yarn application -kill {Application_Id}
```