hadoop相关及操作演示

2016年3月19

钱洋

内容简介

- linux系统简介
- xshell及xftp
- linux的常用命令
- hadoop集群的搭建
- hadoop一些常用命令
- hdfs监控UI及MapReduce监控UI
- ecliplse搭载hadoop（运行调试环境搭建）
- MapReduce案例wordcount及特征词匹配
- hive操作
- Sqoop的使用

linux系统简介

Linux是一款免费的操作系统。

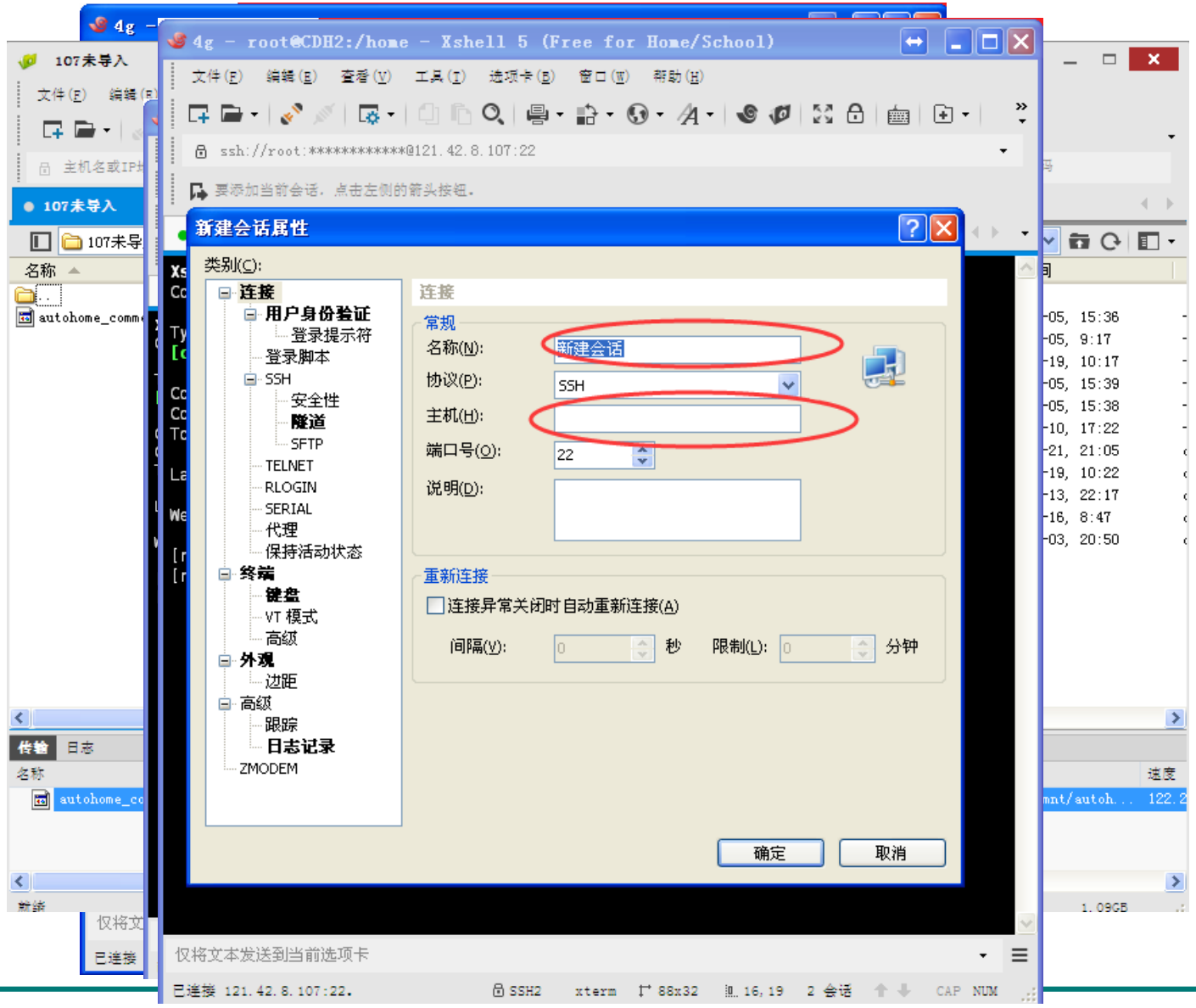
http://www.centoscn.com/CentosSoft/iso/

linux系统简介

linux的好处：
1)开源、免费。linux内核源代码是开放的，任何人可以自由利用，下载、复制、散发都是合法的，不会有人说你用的是盗版。
2）稳定、安全。linux系统以其超强稳定性著称，一台linux服务器，持续运行3年以上不用重新启动，它的运行效率一点不会减慢。而且，病毒也不多。

linux的安装比较简单

xshell及xftp



linux的常用命令

hadoop集群的安装

主要部分
1.jdk的安装，所有机子
2.配制ssh的无密码登陆
3.hadoop软件安装（安装一个其他的scp命令，拷贝到其他机器上）

具体细节请看文档

hadoop一些常用命令


```
[root@master home]# cd /root/
[root@master ~]# hadoop fs -ls /input/
Found 7 items
-rw-r--r--   3 root supergroup       4452 2016-01-05 18:12 /input/attw
-rw-r--r--   3 root supergroup      38818 2016-03-11 09:32 /input/carword.txt
drwxr-xr-x   - root supergroup          0 2015-12-26 22:44 /input/comments
-rw-r--r--   3 root supergroup         63 2016-01-05 22:27 /input/kmeansdata
drwxr-xr-x   - root supergroup          0 2015-12-26 13:31 /input/posts
-rw-r--r--   3 root supergroup      13258 2016-03-02 13:47 /input/test
drwxr-xr-x   - root supergroup          0 2016-03-17 12:42 /input/zhangxu
[root@master ~]#
```

命令：hadoop fs -ls /input

查看目录文件

hadoop一些常用命令


```
Last login: Wed Mar 16 13:50:27 2016 from 114.213.234.145
[root@master ~]# cd /home
[root@master home]# ls
autohome_competion_wangjinkun  hadoop  lost+found  zhangxu
[root@master home]# hive

Logging initialized using configuration in jar:file:/usr/local/hive/apache-hive-1.2.1-bin/lib/hive-common
hive> exit;
[root@master home]# hadoop fs -mkdir /input/input/
[root@master home]# hadoop fs -rmr /input/input/
Deleted hdfs://master:9000/input/input
[root@master home]#
```

命令：hadoop fs -mkdir /input/input/
hadoop fs -rmr /input/input/

创建和删除目录及文件

hadoop一些常用命令


```
[root@master ~]# cd /root
[root@master ~]# ls
anaconda-ks.cfg                      autohome_qirui.java          Desktop    dufei
autohome_competion_wangjinkun.java   carword.txt                  Documents  Music
autohome_posturl2015.java            comments_user_demo_stat.java Downloads  Pictures
[root@master ~]# hadoop fs -put /root/carword.txt /input/
[root@master ~]#
```

```
[root@master ~]# hadoop fs -copyFromLocal /root/carword.txt /output/
[root@master ~]# hadoop fs -ls /output/
Found 4 items
drwxr-xr-x   - root supergroup          0 2016-01-19 15:39 /output/autoname
-rw-r--r--   3 root supergroup      38695 2016-03-17 19:15 /output/carword.txt
drwxr-xr-x   - root supergroup          0 2016-03-15 09:32 /output/comm_car2
drwxr-xr-x   - root supergroup          0 2016-03-17 13:20 /output/zhangxu
[root@master ~]#
```

命令：hadoop fs -put /root/carword.txt /output/
hadoop fs -copyFromLocal /root/carword.txt /output/

本地文件系统中复制单个或多个源路径到目标文件系统

hadoop一些常用命令


```
[root@master ~]#
[root@master ~]#
[root@master ~]# rm carword.txt
rm: remove regular file `carword.txt'? y
[root@master ~]# ls
anaconda-ks.cfg            autohome_qirui.java          Desktop    Downloads  Music     post-install      pos
autohome_posturl2015.java  comments_user_demo_stat.java  Documents  dufei      Pictures  post-install.log  pos
[root@master ~]# hadoop fs -get /input/carword.txt /root/
[root@master ~]# ls
anaconda-ks.cfg            carword.txt                  Documents  Music         post-install.log
autohome_posturl2015.java  comments_user_demo_stat.java  Downloads  Pictures      posts_user_behaviour_stat.ja
autohome_qirui.java        Desktop                      dufei      post-install  posts_user_demo_stat.java
[root@master ~]#
```

```
[root@master ~]# cd /root
[root@master ~]# ls
anaconda-ks.cfg            carword.txt                  Documents  Music         post
autohome_posturl2015.java  comments_user_demo_stat.java  Downloads  Pictures      post
autohome_qirui.java        Desktop                      dufei      post-install  post
[root@master ~]# rm carword.txt
rm: remove regular file `carword.txt'? y
[root@master ~]# ls
anaconda-ks.cfg            autohome_qirui.java          Desktop    Downloads  Music
autohome_posturl2015.java  comments_user_demo_stat.java  Documents  dufei      Picture
[root@master ~]# hadoop fs -copyToLocal /input/carword.txt /root/
[root@master ~]# ls
anaconda-ks.cfg            carword.txt                  Documents  Music         post
autohome_posturl2015.java  comments_user_demo_stat.java  Downloads  Pictures      post
autohome_qirui.java        Desktop                      dufei      post-install  post
[root@master ~]#
```

命令：hadoop fs -get

hadoop fs -getmerge（合并文件，多到1）

hadoop fs -copyToLocal

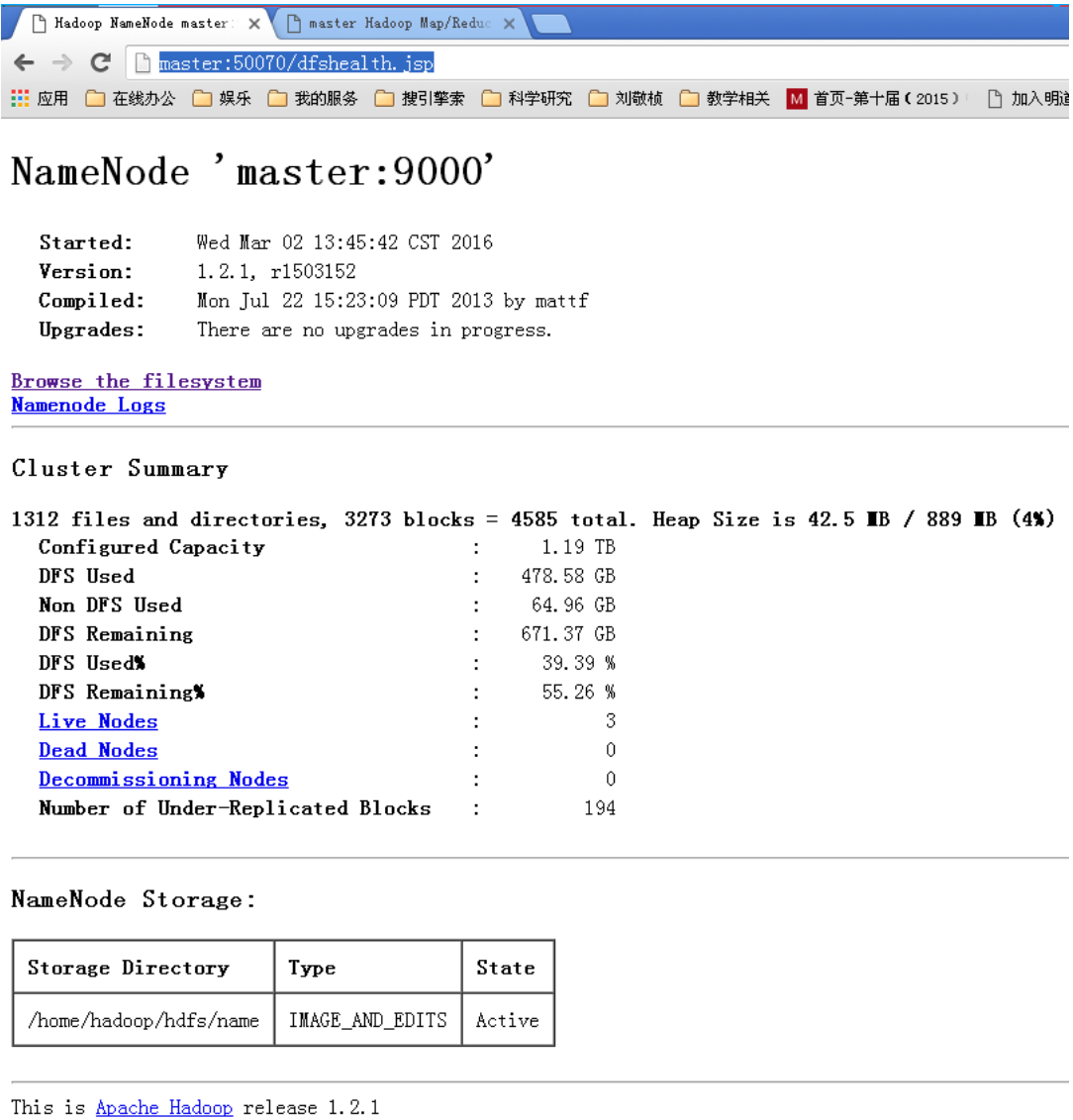
下载文件到本地

hadoop一些常用命令

查看文件内容：
命令：**hadoop fs -cat /input/carword.txt**

杀死正在运行的**Job（http://master:50030/）**
命令：**hadoop job –kill job名**

hdfs监控UI



地址：**http://master:50070/**

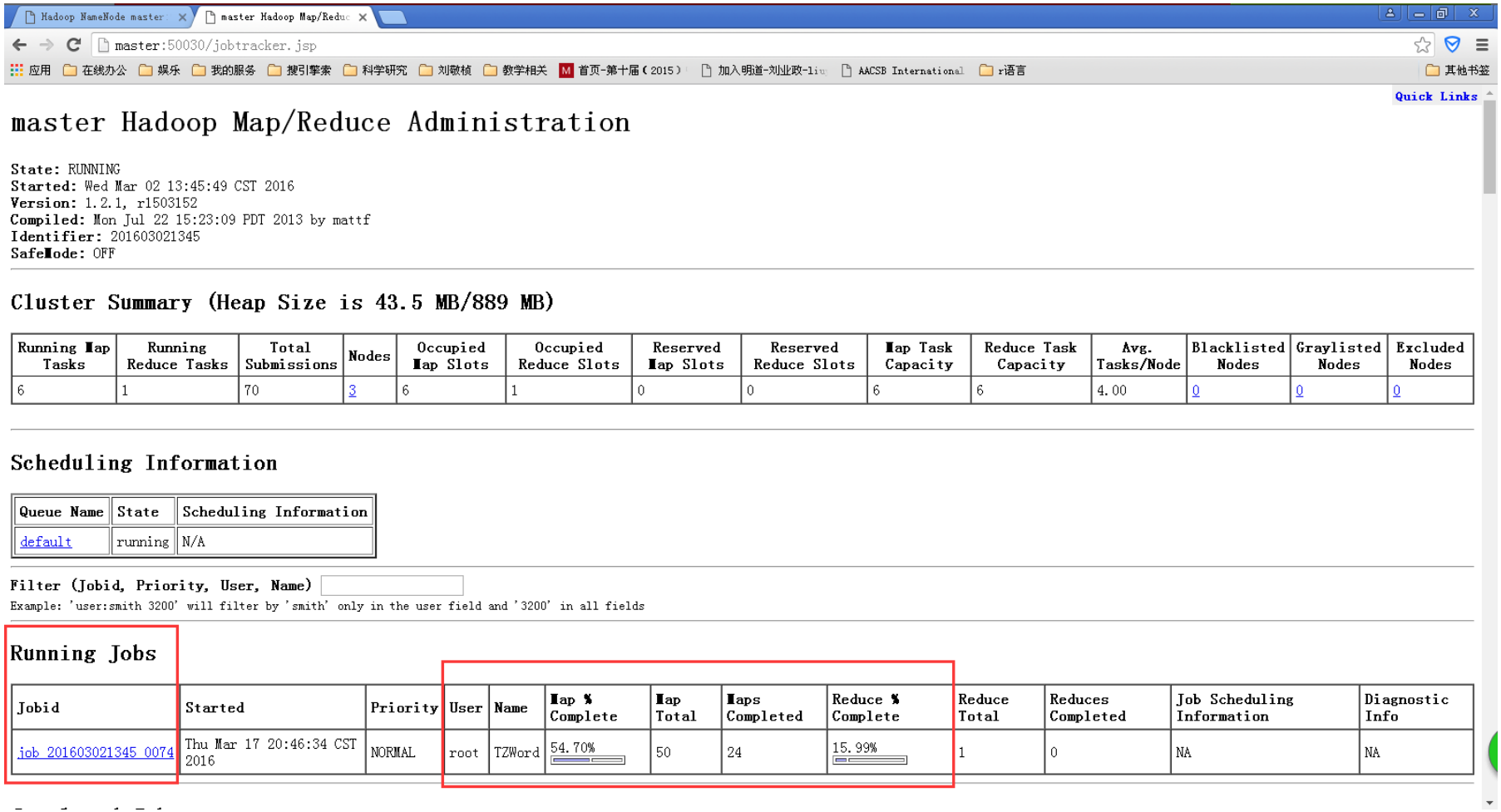
通常遇到的问题：装好hadoop之后，windows上打开hdfs监控UI浏览文件（browse the filesystem）出错

修改windows下的hosts文件（C:\WINDOWS\system32\drivers\etc）

MapReduce监控UI

地址：**http://master:50030/**

MapReduce监控UI



Hadoop NameNode master | master Hadoop Map/Reduc

master:50030/jobtracker.jsp

应用 | 在线办公 | 娱乐 | 我的服务 | 搜引擎索 | 科学研究 | 刘敬楠 | 教学相关 | 首页-第十届(2015) | 加入明道-刘业政-liu | AACSB International | r语言 | 其他书签

Quick Links

master Hadoop Map/Reduce Administration

State: RUNNING
Started: Wed Mar 02 13:45:49 CST 2016
Version: 1.2.1, r1503152
Compiled: Mon Jul 22 15:23:09 PDT 2013 by mattf
Identifier: 201603021345
SafeMode: OFF

Cluster Summary (Heap Size is 43.5 MB/889 MB)

Running Map Tasks	Running Reduce Tasks	Total Submissions	Nodes	Occupied Map Slots	Occupied Reduce Slots	Reserved Map Slots	Reserved Reduce Slots	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes	Graylisted Nodes	Excluded Nodes
6	1	70	3	6	1	0	0	6	6	4.00	0	0	0

Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)

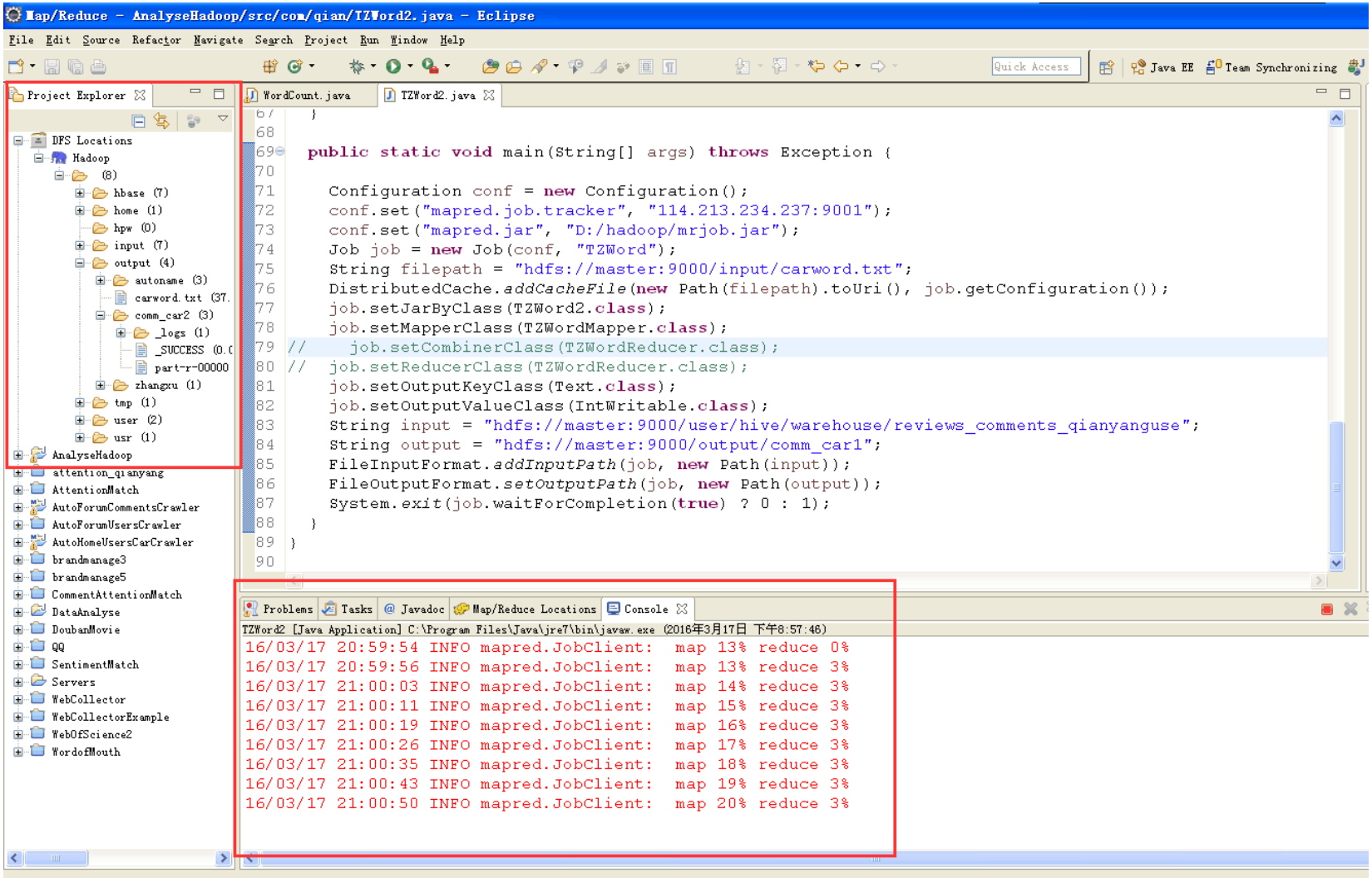
Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

Jobid	Started	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201603021345_0074	Thu Mar 17 20:46:34 CST 2016	NORMAL	root	TZWord	54.70%	50	24	15.99%	1	0	NA	NA

地址：**http://master:50030/**

eclipse搭载hadoop



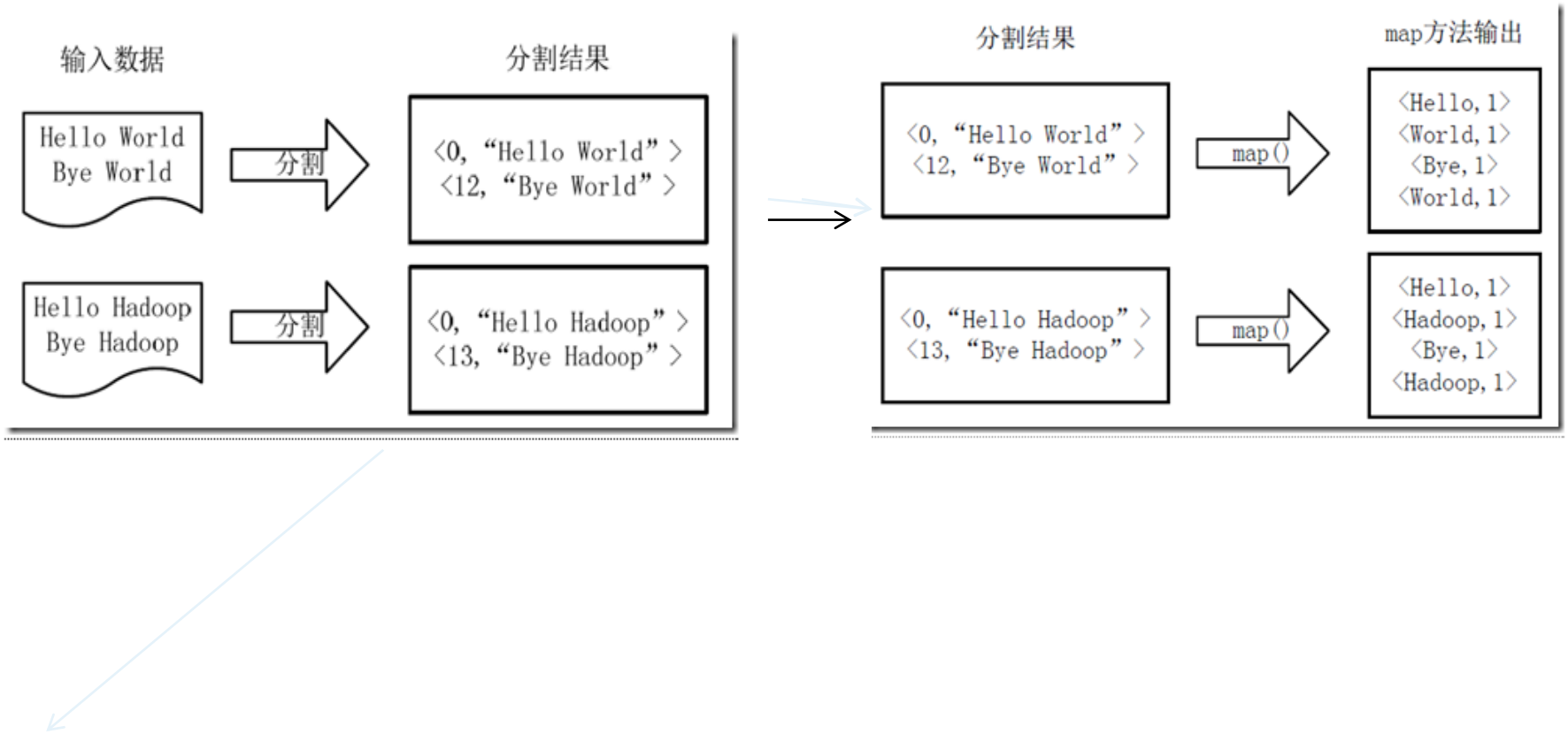
ecliplse搭载hadoop

MapReduce程序直接运行在hadoop集群上，省去打包上传到集群的步骤，Eclipse的Console会打印相关信息——数据集较小

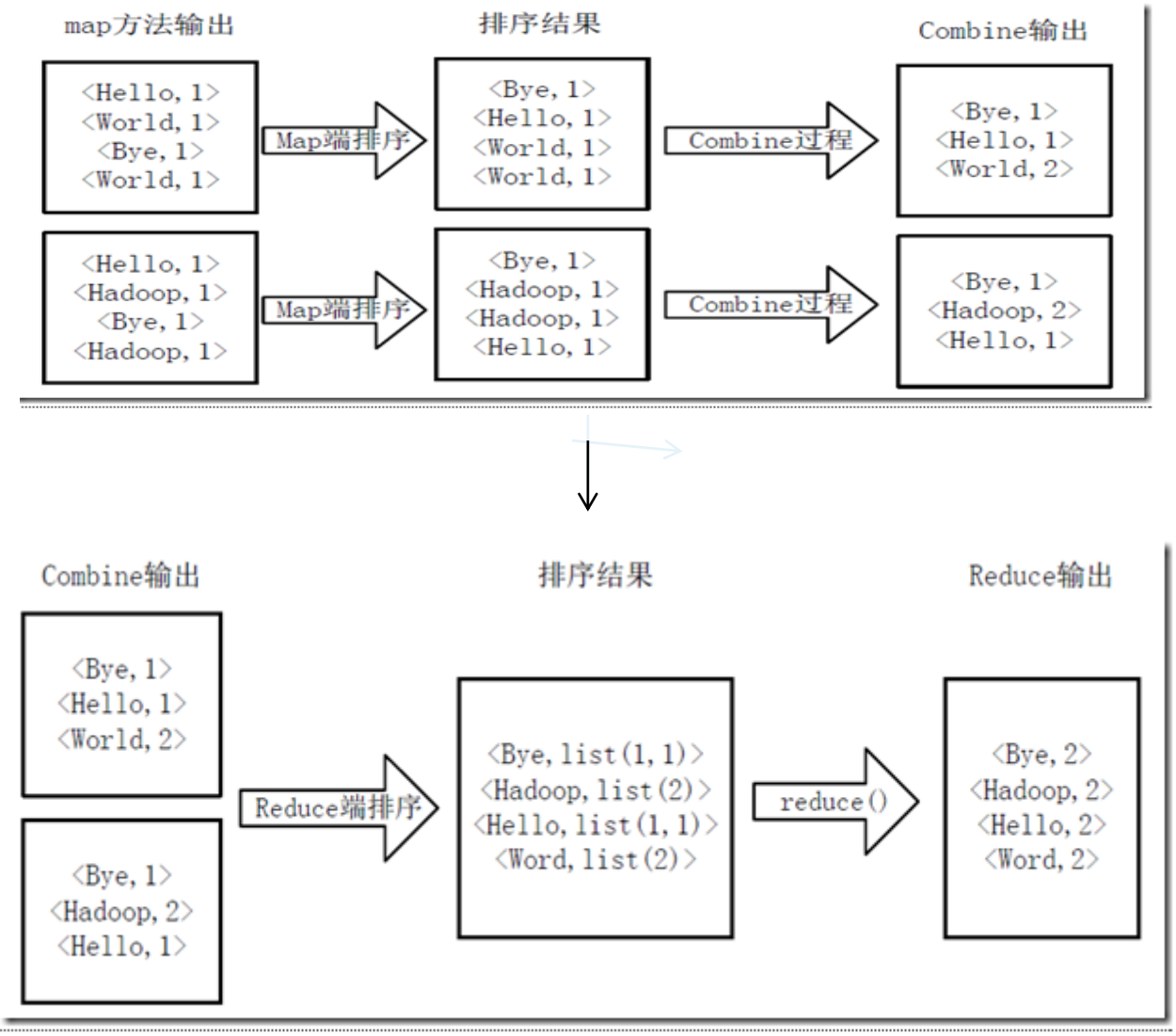
1.将hadoop-eclispe插件（网上下载）放入eclipse安装目录的plugins文件夹下
2.解压hadoop-1.2.1-bin.tar.gz，将hadoop-core-1.2.1.jar替换为hadoop-core-1.2.1-forwinlocal.jar
3.启动Eclipse，打开Map/Reduce视图。window-Preferences-Hadoop Map/Reduce下选择hadoop-1.2.1-bin.tar.gz的解压目录
4.在视图正下方的Map/Reduce Locations页面New Hadoop Location，填入hadoop集群配置文件core-site.xml和mapred-site.xml中指定的Map/Reduce Master和DFS Master的IP地址和端口号。其余选项默认即可
5.在左侧的Project Explorer中即可看到DFS Location中有了我们刚才新增的hdfs。可以新建文件夹删除文件看是否已经可用。

注意：为解决部分权限问题，需要将windows的用户名Administrator或者自定义的用户名修改为hadoop集群的用户名称，如root，或hadoop等。

MapReduce案列wordcount



MapReduce案例wordcount



MapReduce案例特征词

Hive操作

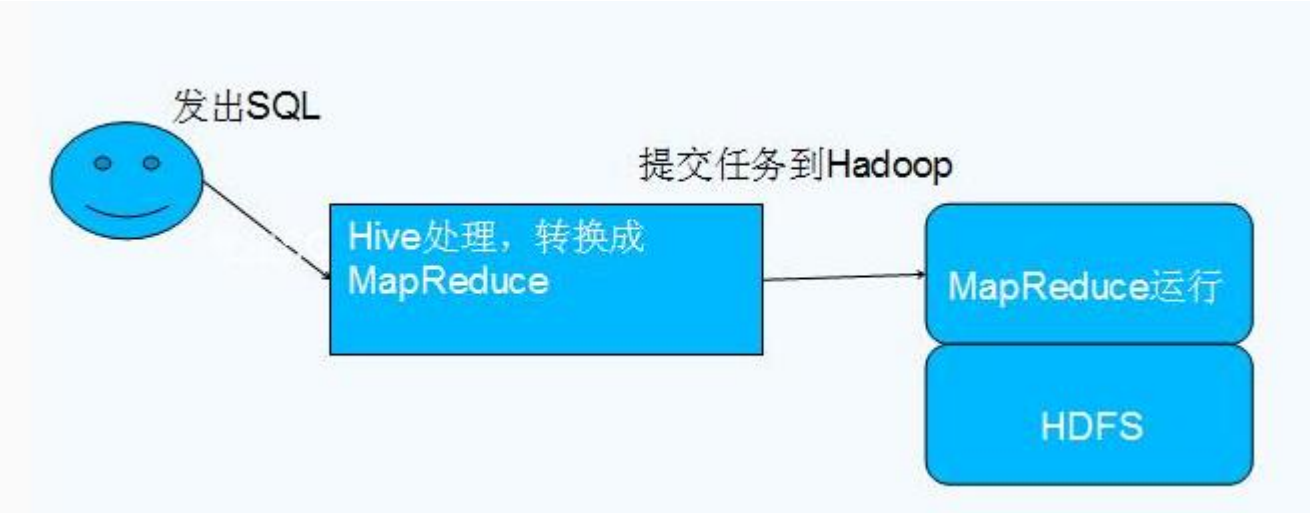
什么是Hive:Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供类SQL查询功能。

为什么使用Hive:Hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供类SQL查询功能。

（1）好写，好操作。类sql语句。

（2）避免了去写MapReduce，减少开发人员的学习成本。

Hive操作



Hive与MapReduce的关系

Hive操作


```
[root@master ~]#
[root@master ~]# hive

Logging initialized using configuration in jar:file:/usr/local/hive/a
hive> show tables;
OK
autohome_competion_wangjinkun
autohome_reviews
comments_qian
comments_tmp_qy2
comments_user_demo_stat
forum_addresses_s
forum_comments_2015sum
forum_comments_s
forum_comments_s_2014
forum_comments_s_2015
forum_comments_s_others
forum_posts_2015sum
forum_posts_s
forum_posts_s_2014
forum_posts_s_2015
forum_posts_s_others
forum_users_s
forum_users_s_age
posts_user_behaviour_stat
```

Sqoop操作


```
[root@master ~]# sqoop list-databases -connect jdbc:mysql://114.213.234.104:3306/ -username root -password 112233
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Warning: $HADOOP_HOME is deprecated.

16/03/18 21:04:51 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
16/03/18 21:04:51 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
16/03/18 21:04:51 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
information_schema
brandetl_db
brandmanage_db
db_carcrawler1
exercise
forum_crawler
forum_crawler_1
forum_crawler_2
forum_crawler_3
forum_crawler_4
forum_crawler_5
lastfm_sun
mam_db
mysql
performance_schema
sakila
```

Sqoop操作


```
[root@master ~]# sqoop list-databases -connect jdbc:mysql://114.213.234.104:3306/ -username root -password 112233
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Warning: $HADOOP_HOME is deprecated.

16/03/18 21:04:51 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
16/03/18 21:04:51 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
16/03/18 21:04:51 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
information_schema
brandetl_db
brandmanage_db
db_carcrawler1
exercise
forum_crawler
forum_crawler_1
forum_crawler_2
forum_crawler_3
forum_crawler_4
forum_crawler_5
lastfm_sun
mam_db
mysql
performance_schema
sakila
```

Sqoop操作


```
zhangxu
[root@master ~]# sqoop list-tables -connect jdbc:mysql://114.213.234.104:3306/brandetl_db -username root -password 112233
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/local/sqoop/sqoop-1.4.6.bin__hadoop-1.0.0/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Warning: $HADOOP_HOME is deprecated.

16/03/18 21:08:29 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
16/03/18 21:08:29 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
16/03/18 21:08:29 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
addresso
attention_qianyang
attentionaveragecount_qianyang
attentionword
authorities
auto_forum_comments_before
auto_forum_commentso
auto_forum_dazhong3
auto_forum_info_stand
auto_forum_luhu
auto_forum_luhu2
auto_forum_luhu3
auto_forum_postso
auto_forum_users_s
auto_forum_userso
autohome_comments2015
autohome_comments2015_1
autohome_commentsbefore2015
autohome_postbefore2015
autohome_qirui
brand_user_area_dong
brand_userarea_dong
car_info_benchmark
car_info_benchmark_1
carinfo
carsalesdata
china_province_city_workbench
```