**课题名称：日志分析系统**

一：前提

1.课题概述 ------------------------------------ 3

2.环境搭建 ------------------------------------ 3

3.数据的准备 ------------------------------------ 3

二：数据导入

1.从hdfs导入hbase -------------------------------- 4

2.hive表映射hdfs文件 ------------------------------- 6

三: 基于HIVE的分析

1.页面PV量 ------------------------------------ 8

2.注册数 ------------------------------------ 8

3.IP数 ------------------------------------ 9

4.跳出率 ------------------------------------ 9

5.汇总 ------------------------------------ 10

四: 分析结果 ----------------------------------------------------------- 11

五: 总结 ----------------------------------------------------------- 12

一．前提

1.课题概述

结合HDFS、MapReduce、HBase、Hive、Sqoop框架等相关技术，实现对论坛的日志的分析，计算该论坛的一些关键指标：浏览量PV，注册用户数，IP数，跳出率。同时将分析结果存入mysql以便查看。

实验应满足：提供视图工具供用户使用，指标查询mysql、明细则查询Hbase；这里使用HBase存储明细日志，能够利用ip、时间查询。使用Sqoop把Hive产生的统计结果导出到mysql中。

2.环境搭建

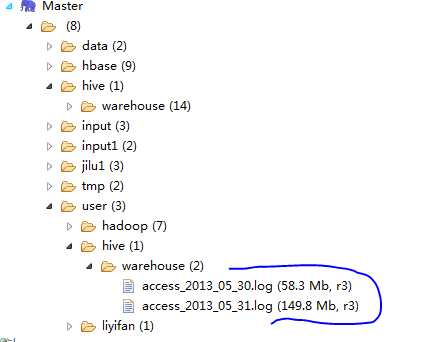
基于三个节点的hadoop集群环境，各个节点安装zookeeper，搭建hbase集群环境，在master节点上安装hive和mysql。

(环境搭建步骤不再赘述)

3.数据的准备

首先可以通过winscp将Windows下两日的日志文件上传到Linux文件系统中，

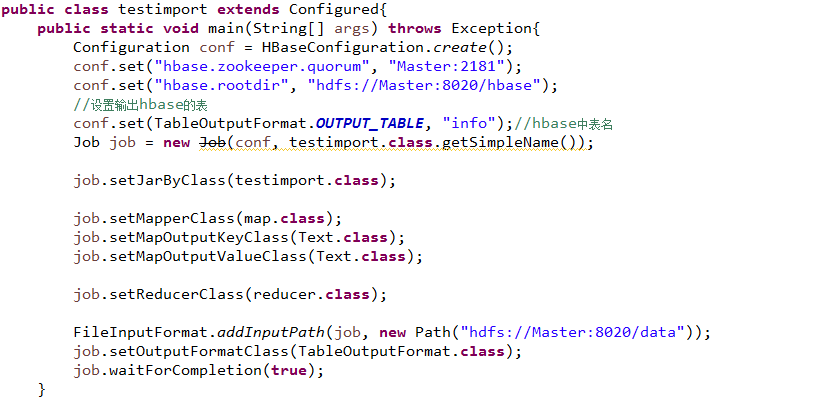
之后可以put命令或者直接使用eclipse DFS中将本地文件上传到hdfs中。上传结果如下图：



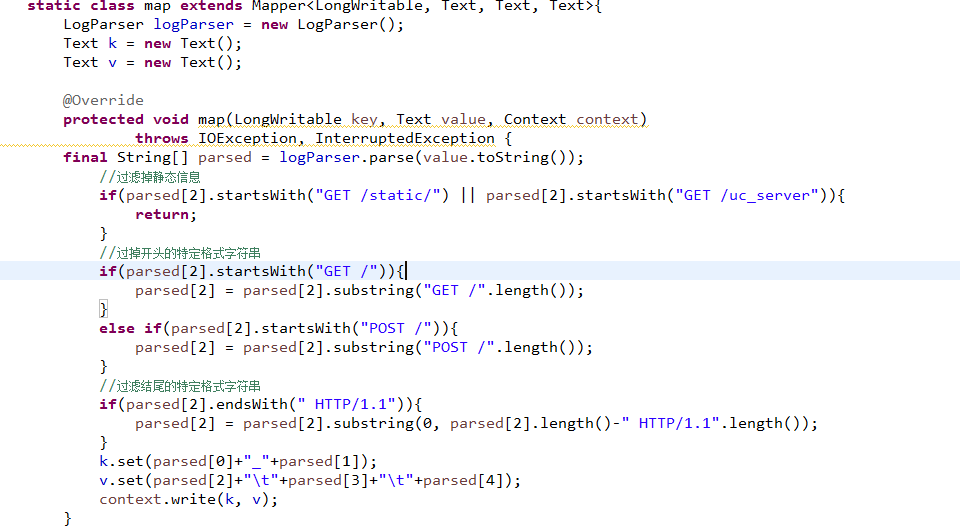
二．数据的导入

1. 使用javaAPI将数据文件从hdfs导入到hbase。

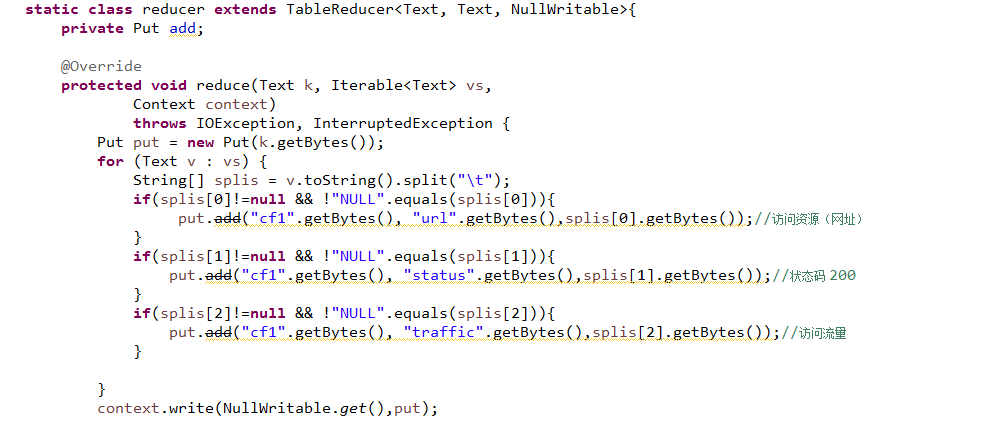
首先在hbase下建立表 info ，准备导入数据。在eclipse中编写mapreducer程序将数据导入到hbase中。程序详细代码如下图：



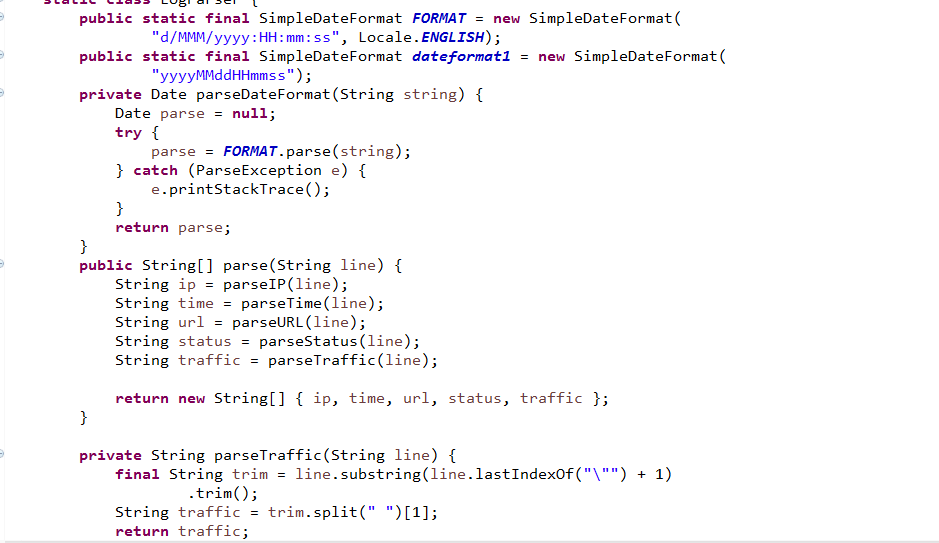
**Map:清洗日志文件中无用字符，将ip和时间作为rowkey.**



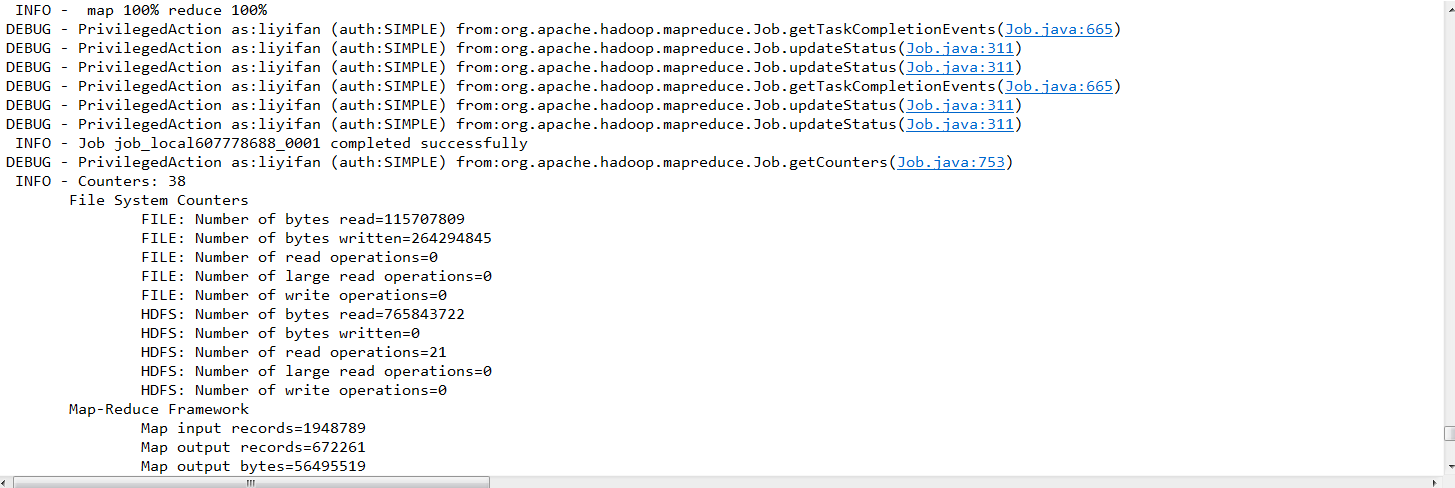
**Reducer:将url,状态码，访问流量作为列**



**清洗：对每个列信息无用字符进行清洗**



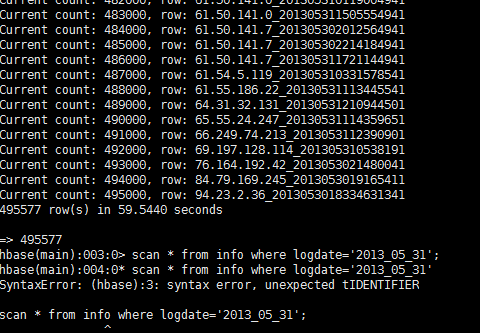
**运行程序导入文件**



**Hbase中查询结果**

**\*\*此处存在错误，数据未完全导入，可以使用export工具导入。\*\***

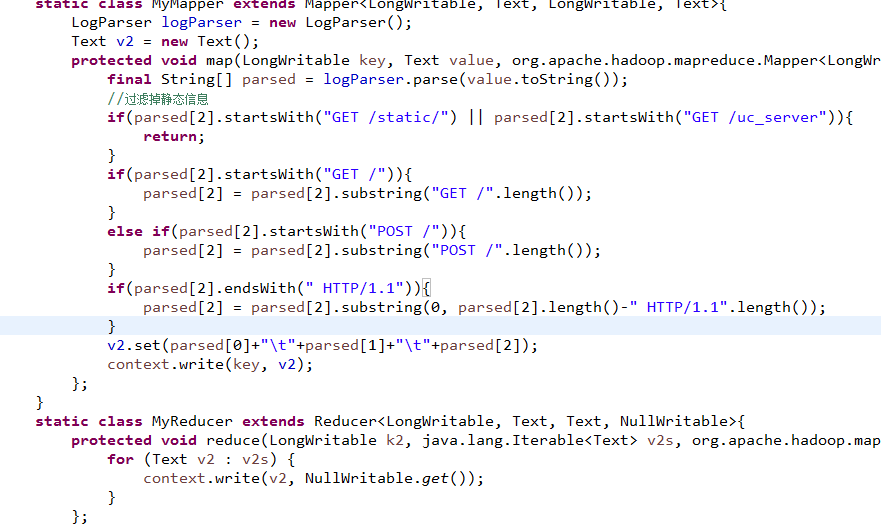
**（因为rowkey设计）**



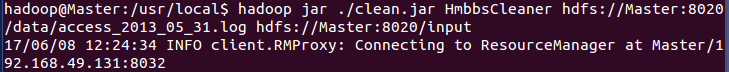
2. 在hive中建表映射hdfs中清洗过的文件

通过编写mapreduce程序打成jar包在终端中运行，清洗数据，本次清洗数据是为了利用hive进行分析，所以把日志文件中的状态码和访问流量信息筛选掉不再使用。

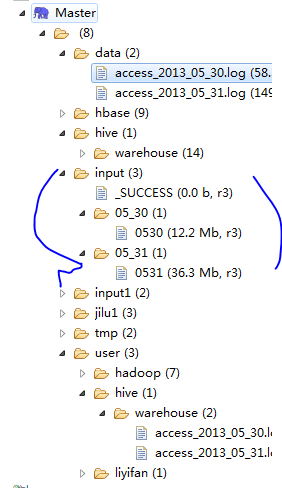
部分代码如下图：



在终端中运行：



清洗两次分别输出到input和input1目录，然后再整合到一起如下图：



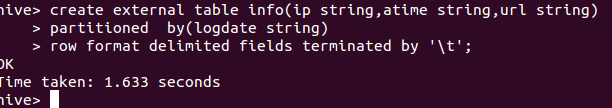
* 1. 终端输入hive进入hive后建立外部表info，设置标签 logdate,location映射位置指定为/input，即清洗后文件所在位置（下图少写了location，实际中有更正。）

CREATE EXTERNAL TABLE info(ip string, atime string, url string)

PARTITIONED BY (logdate string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t'

LOCATION '/data\_clean';

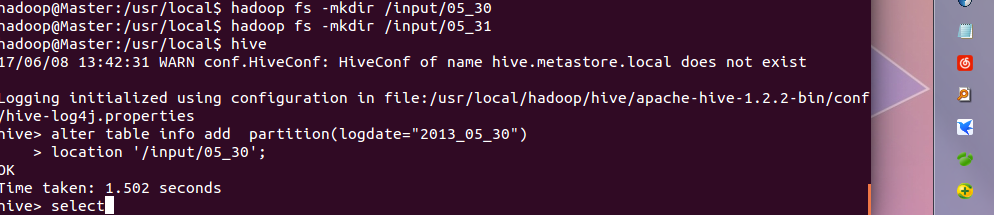


随即就是建立表分区，根据标签logdate建立两个分区。如下图：

ALTER TABLE websiteInfo ADD PARTITION(logdate='2013\_05\_31')

LOCATION '/data\_clean/2013\_0531';

ALTER TABLE websiteInfo ADD PARTITION(logdate='2013\_05\_30')

LOCATION '/data\_clean/2013\_0530';  


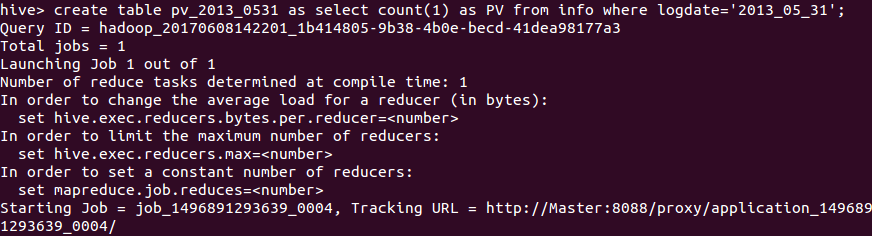
三．基于HIVE的分析

（这里记录出31号的分析结果，30日操作雷同不再赘述）

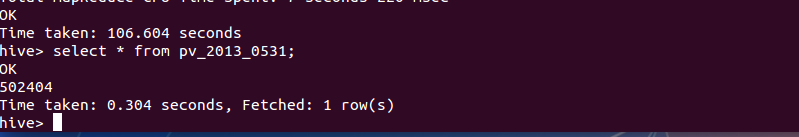
3.1、浏览量PV

页面浏览量即为PV(Page View)，是指所有用户浏览页面的总和，一个独立用户每打开一个页面就被记录1 次。这里，我们只需要统计日志中的记录个数即可。

CREATE TABLE pv\_2013\_0530 AS SELECT COUNT(1) AS PV FROM info WHERE logdate='2013\_0530';



浏览量如下图：502404浏览量

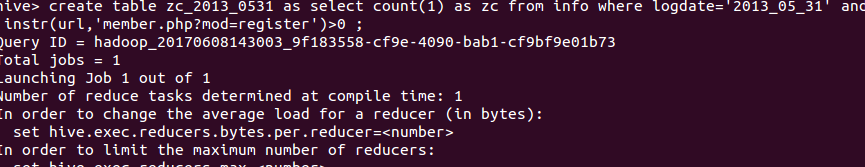


3.2、注册用户数

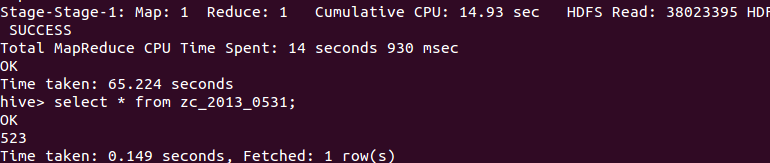
该论坛的用户注册页面为member.php，而当用户点击注册时请求的又是member.php?mod=register的url。因此，这里我们只需要统计出日志中访问的URL是member.php?mod=register的即可。

CREATE TABLE zc\_2013\_0531 AS SELECT COUNT(1) AS REGUSER

FROM info WHERE logdate='2013\_05\_31' AND INSTR(url,'member.php?mod=register')>0;



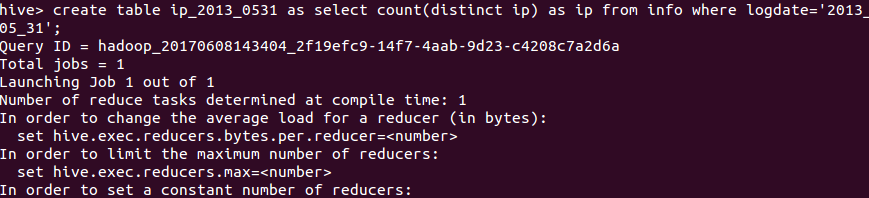
注册结果如下图：523个注册记录



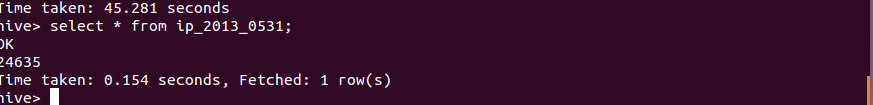
3.3、IP数

一天之内，访问网站的不同独立 IP 个数加和。其中同一IP无论访问了几个页面，独立IP 数均为1。因此，这里我们只需要统计日志中处理的独立IP数即可。

CREATE TABLE ip\_2013\_0530 AS SELECT COUNT(DISTINCT ip) AS IP FROM info WHERE logdate='2013\_05\_30';



结果如下图，24635个ip数

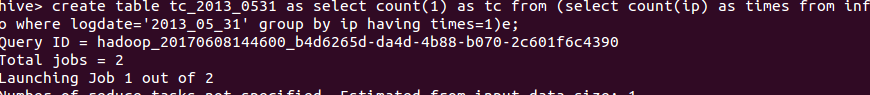


3.4、跳出率

跳出率是指只浏览了一个页面便离开了网站的访问次数占总的访问次数的百分比，即只浏览了一个页面的访问次数 / 全部的访问次数汇总。这里，我们可以将这里得出的跳出用户数/PV数即可得到跳出率。其中，我们可以通过用户的IP进行分组，如果分组后的记录数只有一条，那么即为跳出用户。将这些用户的数量相加，就得出了跳出用户数。

首先统计出跳出用户数：

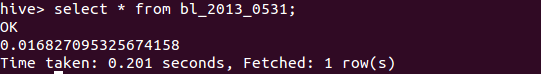
CREATE TABLE tc\_2013\_0531 AS SELECT COUNT(1) AS tc FROM (SELECT COUNT(ip) AS times FROM info WHERE logdate='2013\_05\_31' GROUP BY ip HAVING times=1)e;



再将跳出用户数与pv数相除：

create table bl\_2013\_0531 as select bl\_2013\_0531.bl/pv\_2013\_0531.PV as bl from pv\_2013\_0531, tc\_2013\_0531;



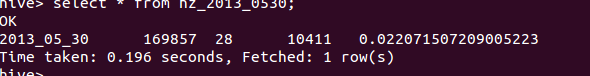
统计跳出率如下图：  


3.5分别建立汇总表

CREATE TABLE hz\_2013\_0531 AS SELECT '2013\_05\_31', a.pv, b.zc, c.ip, d.bl FROM pv\_2013\_0531 a JOIN zc\_2013\_0531 b ON 1=1 JOIN ip\_2013\_0531 c ON 1=1 JOIN bl\_2013\_0531 d ON 1=1;



查看该表信息



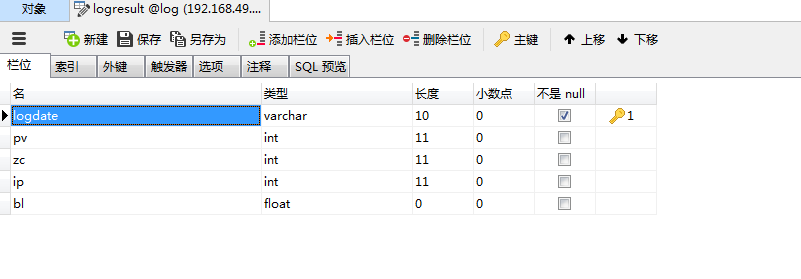
即分别存储两日四个所需指标的信息。

至此，hive中分析结束，接下来利用sqoop将hive中汇总表导入mysql。

四．分析结果

4.1以上已经对日志文件进行分析并汇总指标，下面在mysql中建表。

可以使用navicat for mysql工具建表，五个字段存储四个指标和日期信息。如下图：

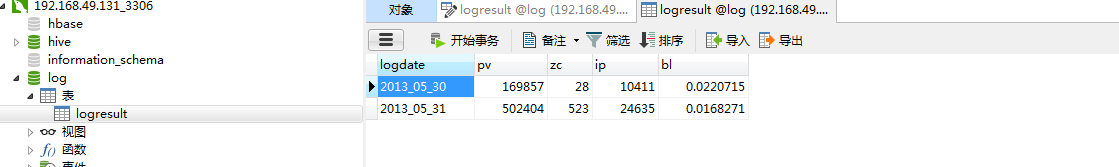


4.2在终端中运行sqoop命令

sqoop export --connect jdbc:mysql://Master:3306/log --username root --password 123456 --table logresult --fields-terminated-by '\001' --export-dir '/hive/warehouse/hz\_2013\_0530'



查看mysql中导出结果



导入成功，至此，实验完成。

五．总结

1. 在数据导入hbase时数据少导，原因是因为rowkey设计为ip和访问时间，但存在同一ip在同一秒级时间访问到多个url，所以导入hbase时数据就会减少。可以通过更改rowkey设计来完善。
2. 将数据通过hive分析时，我是选择的映射hdfs中清洗过的文件，也可以通过修改配置，hbase和hive集成，使hive映射hbase。
3. Hive建表时重点就是location 的映射位置，要填写正确。
4. 使用hql对数据分析时其实要求已经给出分析思路，难度不是很大。
5. 最后就是使用sqoop将汇总后的信息导入mysql中方便查询。