[(251条消息) 大数据综合项目--网站流量日志数据分析系统(详细步骤和代码)\_为java献身的博客-CSDN博客\_网站流量日志数据分析系统](https://blog.csdn.net/weixin_54028135/article/details/125072042)

1．创建数据仓库

/etc/init.d/mysqld start

在hadoop01启动集群以及hive服务，然后在任意一台使用beeline远程连接hive服务

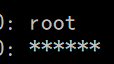
Hadoop001: cd /export/servers/apache-hive-1.2.1-bin/

bin/hiveserver2

Hadoop002: cd /export/servers/apache-hive-1.2.1-bin/

bin/beeline

!connect jdbc:hive2://Hadoop001:10000



create database weblog;

use weblog;

create table ods\_weblog\_origin(

valid string,

remote\_addr string,

remote\_user string,

time\_local string,

request string,

status string,

body\_bytes\_sent string,

http\_referer string,

http\_user\_agent string ) partitioned by(datestr string) row format delimited fields terminated by '\001';

show tables;查看表

2.将上次运行处理出来的结果上传

mkdir /root/weblog/

put "F:\Hadoop\download\weblog\part-m-00000"

cd /root/weblog/

hadoop fs -mkdir -p /weblog/preprocessed

hadoop fs -put part-m-00000 /weblog/preprocessed

回到hadoop02:hive

load data inpath '/weblog/preprocessed/' overwrite into table ods\_weblog\_origin partition(datestr='20130918');

select \* from ods\_weblog\_origin limit 3;

3.生成明细宽表

create table ods\_weblog\_detail(

valid string, --有效标识

remote\_addr string, --来源IP

remote\_user string, --用户标识

time\_local string, --访问完整时间

daystr string, --访问日期

timestr string, --访问时间

month string, --访问月

day string, --访问日

hour string, --访问时

request string, --请求的url

status string, --响应码

body\_bytes\_sent string, --传输字节数

http\_referer string, --来源url

ref\_host string, --来源的host

ref\_path string, --来源的路径

ref\_query string, --来源参数query

ref\_query\_id string, --来源参数query值

http\_user\_agent string --客户终端标识

) partitioned by(datestr string);

查看：show tables;

创建中间临时表t\_ods\_tmp\_referurl

create table t\_ods\_tmp\_referurl as SELECT a.\*,b.\*

FROM ods\_weblog\_origin a LATERAL VIEW parse\_url\_tuple(regexp\_replace(http\_referer, "\"", ""), 'HOST', 'PATH','QUERY', 'QUERY:id') b as host, path, query, query\_id;

创建临时中间表t\_ods\_tmp\_detail

create table t\_ods\_tmp\_detail as select b.\*,substring(time\_local,0,10) as daystr,

substring(time\_local,12) as tmstr,

substring(time\_local,6,2) as month,

substring(time\_local,9,2) as day,

substring(time\_local,11,3) as hour

from t\_ods\_tmp\_referurl b;

加载数据到明细宽表前启用动态分区

set hive.exec.dynamic.partition=true;

set hive.exec.dynamic.partition.mode=nonstrict;

生成明细宽表 向ods\_weblog\_detail表，加载数据

insert overwrite table ods\_weblog\_detail partition(datestr)

select distinct otd.valid,otd.remote\_addr,otd.remote\_user,

otd.time\_local,otd.daystr,otd.tmstr,otd.month,otd.day,otd.hour,

otr.request,otr.status,otr.body\_bytes\_sent,

otr.http\_referer,otr.host,otr.path,

otr.query,otr.query\_id,otr.http\_user\_agent,otd.daystr

from t\_ods\_tmp\_detail as otd,t\_ods\_tmp\_referurl as otr

where otd.remote\_addr=otr.remote\_addr

and otd.time\_local=otr.time\_local

and otd.body\_bytes\_sent=otr.body\_bytes\_sent

and otd.request=otr.request;

