

hive做离线数据处理

2017年11月20日

11:01

**实现步骤**

1.启动hive

进入hive的bin目录

执行：./hive

2.创建数据库

hive>create database weblog;

3.创建外部分区表管理数据

>use weblog;

>create external table flux (url string,urlname string,title string,chset string,scr string,col string,lg string,je string,ec string,fv string,cn string,ref string,uagent string,stat\_uv string,stat\_ss string,cip string) PARTITIONED BY (reportTime string) row format delimited fields terminated by '|' location '/weblog';

创建分区

>alter table flux add partition (reportTime='2018-02-23') location '/weblog/reportTime=2018-02-23';

4.建立数据清洗表： dataclear 指定的分割符 : |

去除多余字段，只保留需要的字段，并将会话信息拆开保存

所需要的字段为：

reportTime、url、urlname、uvid、ssid、sscount、sstime、cip

report time 
unname 
uvid 
ssid 
sscount 
s s time 
CIP 
string 
string 
string 
string 
string 
string 
string 
string 

5.建立完dataclear之后，需要从flux（总表中）将相关字段的值插入到dataclear表。

线索：会用到hive的内置函数：**split 函数 ——比如：select split("hello,world",",")[1];**

>create table dataclear(reportTime string,url string,urlname string,uvid string,ssid string,sscount string,sstime string,cip string)row format delimited fields terminated by '|';

>insert overwrite table dataclear

select reporttime,url,urlname,stat\_uv,split(stat\_ss,"\_")[0],split(stat\_ss,"\_")[1],split(stat\_ss,"\_")[2],cip

from flux;

5.业务处理

**1）pv**

hive>select count(\*) as pv from dataclear where reportTime = '2018-02-23';

**2）uv**

uv - 独立访客数 - 一天之内所有的访客的数量 - 一天之内uvid去重后的总数

hive>select count(distinct uvid) as uv from dataclear where reportTime = '2018-02-23';

**3）vv**

vv - 独立会话数 - 一天之内所有的会话的数量 - 一天之内ssid去重后的总数

hive>select count(distinct ssid) as vv from dataclear where reportTime = '2018-02-23';

**4）br**

br - 跳出率 - 一天内跳出的会话总数/会话总数

hive>

select tab1.a/tab2.b from (select count(\*)as a from dataclear where sscount=0 and reportTime="2018-01-08")as tab1,

(select count(distinct ssid)as b from dataclear where reportTime="2018-01-08")as tab2;

**5）newip**

newip - 新增ip总数 - 一天内所有ip去重后在历史数据中从未出现过的数量

hive>select count(distinct dataclear.cip) from dataclear where dataclear.reportTime = '2018-02-23'

and cip **not in**

(select dc2.cip from dataclear as dc2 where dc2.**reportTime < '2018-02-23'**);

**6）newcust**

newcust - 新增客户数 - 一天内所有的uvid去重后在历史数据中从未出现过的总数

hive>select count(distinct dataclear.uvid) from dataclear where dataclear.reportTime='2018-02-23'

and uvid not in

(select dc2.uvid from dataclear as dc2 where dc2.reportTime < '2018-02-23');

**7）avgtime**

avgtime - 平均访问时常 - 一天内所有会话的访问时常的平均值

**注：** 一个会话的时长 = 会话中所有访问的时间的最大值 - 会话中所有访问时间的最小值

hive>select **avg**(atTab.usetime) as avgtime from

(select **max**(sstime) - **min**(sstime) as usetime from dataclear where reportTime='2018-02-23' **group by** ssid) as atTab;

**8）avgdeep**

avgdeep - 平均访问深度 - 一天内所有会话访问深度的平均值

一个会话的访问深度=一个会话访问的所有url去重后的个数

比如会话①：url <http://demo/a.jsp> <http://demo/b.jsp> <http://demo/a.jsp> 则访问深度是2

hive>select round(avg(deep),4) as avgdeep from

(select count(distinct urlname) as deep from dataclear where reportTime='2017-07-09' group by ssid) as adTab;

**创建业务表并插入数据**

hive> create table tongji(reportTime string,pv int,uv int,vv int, br double,newip int, newcust int, avgtime double,avgdeep double) row format delimited fields terminated by '|';

hive>insert overwrite table tongji select '2017-07-09',tab1.pv,tab2.uv,tab3.vv,tab4.br,tab5.newip,tab6.newcust,tab7.avgtime,tab8.avgdeep from (select count(\*) as pv from dataclear where reportTime = '2017-07-09') as tab1,

(select count(distinct uvid) as uv from dataclear where reportTime = '2017-07-09') as tab2,

(select count(distinct ssid) as vv from dataclear where reportTime = '2017-07-09') as tab3,

(select round(br\_taba.a/br\_tabb.b,4)as br from (select count(\*) as a from (select ssid from dataclear where reportTime='2017-07-09' group by ssid having count(ssid) = 1) as br\_tab) as br\_taba,

(select count(distinct ssid) as b from dataclear where reportTime='2017-07-09') as br\_tabb) as tab4,

(select count(distinct dataclear.cip) as newip from dataclear where dataclear.reportTime = '2017-07-09' and cip not in (select dc2.cip from dataclear as dc2 where dc2.reportTime < '2017-07-09')) as tab5,

(select count(distinct dataclear.uvid) as newcust from dataclear where dataclear.reportTime='2017-07-09' and uvid not in (select dc2.uvid from dataclear as dc2 where dc2.reportTime < '2017-07-09')) as tab6,

(select round(avg(atTab.usetime),4) as avgtime from (select max(sstime) - min(sstime) as usetime from dataclear where reportTime='2017-07-09' group by ssid) as atTab) as tab7,

(select round(avg(deep),4) as avgdeep from (select count(distinct urlname) as deep from dataclear where reportTime='2017-07-09' group by ssid) as adTab) as tab8;