Hive进行数据处理

2019年3月10日　　17:30

1. 网站流量分析项目中的数据清洗

a. 清洗目标

只保留需要的字段

将会话信息拆分 为 会话编号 会话页面数 会话时间

url urlname ref uagent uvid ssid sscoutn sstime cip

b. 创建外部分区表管理已经在HDFS的流量数据

```sql
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 '/flux';
```

c. 增加flux的分区信息

```sql
alter table flux add partition(reportTime='2018-09-17') location '/flux/reportTime=2018-09-17';
```

d. 创建数据清洗表dataclear

```sql
create table dataclear (url string,urlname string,ref string,uagent string,uvid string,ssid string,sscoutn string,sstime string,cip string) partitioned by (reportTime string) row format delimited fields terminated by '|';
```

e. 从zebra表中导入数据到dataclear表，在这个过程中完成数据清洗

```sql
insert into dataclear partition(reportTime='2018-09-17') select url,urlname,ref,uagent,stat_uv,split(stat_ss,'_')[0],split(stat_ss,'_')[1],split(stat_ss,'_')[2],cip from flux where reportTime = '2018-09-17';
```

2. 利用Hive实现业务指标的计算

a. PV

访问量，一天之内访问的总量，有多少条日志就是多少个访问量。

```sql
select count(*) as pv from dataclear where reportTime='2018-09-17';
```

b. UV

独立访客数，一天之内用户的总数，将一天内所有日志的uvid去重后计数。

```sql
select count(distinct uvid) as uv from dataclear where reportTime='2018-09-17';
```

c. VV

会话总数，一天之内会话的总的数量，将一天内所有的日志的ssid去重后计数。

```sql
select count(distinct ssid) as vv from dataclear where reportTime='2018-09-17';
```

d. BR

跳出率，一天之内跳出的会话占总的会话的比率。一天内跳出会话的总数/会话的总数。

跳出的会话总数

```sql
select count(br_tab.ssid) from (select ssid from dataclear where reportTime='2018-09-17' group by ssid having count(*) = 1) as br_tab;
```

会话的总数就是vv

```sql
select count(distinct ssid) from dataclear where reportTime='2018-09-17';
```

计算跳出率

```sql
select round(br_left_tab.br_count / br_right_tab.vv_count,4) as br from (select count(br_tab.ssid) as br_count from (select ssid from dataclear where reportTime='2018-09-17' group by ssid having count(*) = 1) as br_tab) as br_left_tab, (select count(distinct ssid) as vv_count from dataclear where reportTime='2018-09-17') as br_right_tab;
```

e. NewIP

新增IP总数，一天之内新IP的数量。

将一天所有日志的IP去重 后 检查在历史数据从未出现过的数量。

```
select count(distinct dataclear.cip) as newip from dataclear where dataclear.reportTime='2018-09-17' and dataclear.cip not in (select distinct inner_dataclear_tab.cip from dataclear as inner_dataclear_tab where datediff('2018-09-17',inner_dataclear_tab.reportTime)>0);
```

f. NewCust

新增客户总数，一天之内新用户的数量。

将一天内所有日志的uvid去重 后 检查从未在历史数据中出现过的数量。

```
select count(distinct dataclear.uvid) as newcust from dataclear where dataclear.reportTime='2018-09-17' and dataclear.uvid not in (select inner_dataclear_tab.uvid from dataclear as inner_dataclear_tab where datediff('2018-09-17',inner_dataclear_tab.reportTime)>0);
```

g. AvgTime

平均访问时长，一天之内所有会话访问时长的平均值

将一天内所有日志按照会话分组后，求会话内部最后一次访问的时间减去第一次访问的时间就是会话时长，求其平均值。

```
select avg(avgtime_tab.use_time) as avgtime from (select max(sstime) - min(sstime) as use_time from dataclear where reportTime='2018-09-17' group by ssid) as avgtime_tab;
```

h. AvgDeep

平均访问深度，一天内所有会话访问深度的平均值。

将一天内所有日志按照会话分组后，统计每个会话访问的页面去重后的总数为会话的访问深度，再求这些会话访问深度的平均值。

```
select round(avg(avgdeep_tab.deep),4) as avgdeep from (select
count(distinct urlname) as deep from dataclear where
reportTime='2018-09-17' group by ssid) as avgdeep_tab;
```

3. 将计算结果存入统计表 - 方案1

a. 创建tongji1表

```
create table tongji1 (reportTime string,pv int,uv int,vv int,br
double,newip int,newcust int,avgtime double,avgdeep double) row
format delimited fields terminated by '|';
```

b. 将计算的结果写入tongji1表

```
insert into tongji1 select
'2018-09-17',tab1.pv,tab2.uv,tab3.vv,tab4.br,tab5.newip,tab6.newcust,tab
7.avgtime,tab8.avgdeep from (select count(*) as pv from dataclear
where reportTime='2018-09-17') as tab1, (select count(distinct uvid) as
uv from dataclear where reportTime='2018-09-17') as tab2, (select
count(distinct ssid) as vv from dataclear where
reportTime='2018-09-17') as tab3, (select round(br_left_tab.br_count /
br_right_tab.vv_count,4) as br from   (select count(br_tab.ssid) as
br_count from (select ssid from dataclear where
reportTime='2018-09-17' group by ssid having count(*) = 1) as br_tab)
as br_left_tab, (select count(distinct ssid) as vv_count from dataclear
where reportTime='2018-09-17') as br_right_tab) as tab4, (select
count(distinct dataclear.cip) as newip from dataclear where
dataclear.reportTime='2018-09-17' and dataclear.cip not in (select
distinct inner_dataclear_tab.cip from dataclear as inner_dataclear_tab
where datediff('2018-09-17',inner_dataclear_tab.reportTime)>0)) as tab5,
(select count(distinct dataclear.uvid) as newcust from dataclear where
dataclear.reportTime='2018-09-17' and dataclear.uvid not in (select
inner_dataclear_tab.uvid from dataclear as inner_dataclear_tab where
datediff('2018-09-17',inner_dataclear_tab.reportTime)>0)) as tab6,
(select avg(avgtime_tab.use_time) as avgtime from (select max(sstime) -
min(sstime) as use_time from dataclear where reportTime='2018-09-17'
group by ssid) as avgtime_tab) as tab7, (select
round(avg(avgdeep_tab.deep),4) as avgdeep from (select count(distinct
urlname) as deep from dataclear where reportTime='2018-09-17' group
by ssid) as avgdeep_tab) as tab8;
```

**这种方式通过连接查询实现 将多个查询结果插入一张tongji1表，实现了效果，但是过多的表的连接效率低下，且任意一个mr出错，整个程序要重新计算，可靠性较低。

4. 将计算结果存入统计表 - 方案2

创建过度用表tongji1_temp

```sql
create table tongji1_temp (reportTime string,field string,value double) row format delimited fields terminated by '|';
```

执行各个指标的运算，将结果存入tongji1_temp

```sql
insert into tongji1_temp  select '2018-09-17','pv',t1.pv from (select count(*) as pv from dataclear where reportTime='2018-09-17') as t1;


insert into tongji1_temp  select '2018-09-17','uv',t2.uv from (select count(distinct uvid) as uv from dataclear where reportTime='2018-09-17') as t2;


insert into tongji1_temp  select '2018-09-17','vv',t3.vv from (select count(distinct ssid) as vv from dataclear where reportTime='2018-09-17') as t3;


insert into tongji1_temp  select '2018-09-17','br',t4.br from (select round(br_left_tab.br_count / br_right_tab.vv_count,4) as br from   (select count(br_tab.ssid) as br_count from (select ssid from dataclear where reportTime='2018-09-17' group by ssid having count(*) = 1) as br_tab) as br_left_tab, (select count(distinct ssid) as vv_count from dataclear where reportTime='2018-09-17') as br_right_tab) as t4;



insert into tongji1_temp  select '2018-09-17','newip',t5.newip from (select count(distinct dataclear.cip) as newip from dataclear where dataclear.reportTime='2018-09-17' and dataclear.cip not in (select distinct inner_dataclear_tab.cip from dataclear as inner_dataclear_tab where datediff('2018-09-17',inner_dataclear_tab.reportTime)>0)) as t5;
insert into tongji1_temp  select '2018-09-17','newcust',t6.newcust from
(select count(distinct dataclear.uvid) as newcust from dataclear where
dataclear.reportTime='2018-09-17' and dataclear.uvid not in (select
inner_dataclear_tab.uvid from dataclear as inner_dataclear_tab where
datediff('2018-09-17',inner_dataclear_tab.reportTime)>0)) as t6;



insert into tongji1_temp  select '2018-09-17','avgtime',t7.avgtime from
(select avg(avgtime_tab.use_time) as avgtime from (select max(sstime) -
min(sstime) as use_time from dataclear where reportTime='2018-09-17'
group by ssid) as avgtime_tab) as t7;


insert into tongji1_temp  select '2018-09-17','avgdeep',t8.avgdeep from
(select round(avg(avgdeep_tab.deep),4) as avgdeep from (select
count(distinct urlname) as deep from dataclear where
reportTime='2018-09-17' group by ssid) as avgdeep_tab) as t8;
```

将tongji1_temp表中的数据导入到tongji1表中：

```
insert into tongji1 select '2018-09-17',t1.pv,t2.uv,t3.vv,t4.br,t5.newip,
t6.newcust, t7.avgtime, t8.avgdeep from  (select value as pv from
tongji1_temp where field='pv' and reportTime='2018-09-17') as t1,
(select value as uv from tongji1_temp where field='uv' and
reportTime='2018-09-17') as t2, (select value as vv from tongji1_temp
where field='vv' and reportTime='2018-09-17') as t3, (select value as br
from tongji1_temp where field='br' and reportTime='2018-09-17') as t4,
(select value as newip from tongji1_temp where field='newip' and
reportTime='2018-09-17') as t5, (select value as newcust from tongji1
_temp where field='newcust' and reportTime='2018-09-17') as t6,
(select value as avgtime from tongji1_temp where field='avgtime' and
reportTime='2018-09-17') as t7, (select value as avgdeep from tongji1
_temp where field='avgdeep' and reportTime='2018-09-17') as t8;
```