常用Hsql指标

# 这里主要讲的是group by的用法

1.使用场景

1.一般出现要求中出现"每"这样的字样的时候需要考虑使用group by 分组

2.使用语法

1.group在hive中select后面的字段和group by对应的要一样

2.他们对应分组的时候如果有多个字段,多个字段之间的顺序也是有讲究的

# 关于topN的求取

1.使用场景

1.分组后再对应的求前几名,前几项,等的需求的时候就是了

2.实现套路

1.select ref\_host,ref\_host\_cnts,concat(month,day,hou),

-- row\_number() over()主要它是用来给排序好的列,再加一列,row\_number()表示的从1开始从小到大一直往下排,对应需要排序的字段等值的时候取前面一个,rank() over()函数如果等值,会给他们同一个数字

row\_number() over (partition by concat(month,day,hour

-- order by 对应的是需要根据这个字段排序的

) order by ref\_host\_cnts desc) as od

from dw\_pvs\_refererhost\_everyhour;

# 流量分析

--------------------------------------------------------------------------------------------

## --计算每小时pvs，注意gruop by语句的语法

select count(\*) as pvs,month,day,hour from ods\_weblog\_detail group by month,day,hour;

select count(\*) as pvs,hour from ods\_weblog\_detail group by hour;

--------------------------------------------------------------------------------------------

# 多维度统计PV总量

--第一种方式：直接在ods\_weblog\_detail单表上进行查询

## 计算该处理批次（一天）中的各小时pvs

drop table if exists dw\_pvs\_everyhour\_oneday;

create table if not exists dw\_pvs\_everyhour\_oneday(month string,day string,hour string,pvs bigint) partitioned by(datestr string);

insert into table dw\_pvs\_everyhour\_oneday partition(datestr='20130918')

select a.month as month,a.day as day,a.hour as hour,count(\*) as pvs from t\_ods\_tmp\_detail a

where a.datestr='20130918' group by a.month,a.day,a.hour;

## --计算每天的pvs

drop table if exists dw\_pvs\_everyday;

create table if not exists dw\_pvs\_everyday(pvs bigint,month string,day string);

insert into table dw\_pvs\_everyday

select count(\*) as pvs,a.month as month,a.day as day from ods\_weblog\_detail a

group by a.month,a.day;

1.1.2 第二种方式：与时间维表关联查询

--维度：日

drop table dw\_pvs\_everyday;

create table dw\_pvs\_everyday(pvs bigint,month string,day string);

insert into table dw\_pvs\_everyday

select count(\*) as pvs,a.month as month,a.day as day from (select distinct month, day from t\_dim\_time) a

join ods\_weblog\_detail b

on a.month=b.month and a.day=b.day

group by a.month,a.day;

--维度：月

drop table dw\_pvs\_everymonth;

create table dw\_pvs\_everymonth (pvs bigint,month string);

insert into table dw\_pvs\_everymonth

select count(\*) as pvs,a.month from (select distinct month from t\_dim\_time) a

join ods\_weblog\_detail b on a.month=b.month group by a.month;

--另外，也可以直接利用之前的计算结果。比如从之前算好的小时结果中统计每一天的

Insert into table dw\_pvs\_everyday

Select sum(pvs) as pvs,month,day from dw\_pvs\_everyhour\_oneday group by month,day having day='18';

--------------------------------------------------------------------------------------------

## 按照来访维度统计pv

-- 统计每小时各来访url产生的pv量，查询结果存入：( "dw\_pvs\_referer\_everyhour" )

drop table if exists dw\_pvs\_referer\_everyhour;

create table if not exists dw\_pvs\_referer\_everyhour

(referer\_url string,referer\_host string,month string,day string,

hour string,pv\_referer\_cnt bigint) partitioned by(datestr string);

insert into table dw\_pvs\_referer\_everyhour partition(datestr='20130918')

select http\_referer,ref\_host,month,day,hour,count(1) as pv\_referer\_cnt

from ods\_weblog\_detail

group by http\_referer,ref\_host,month,day,hour

having ref\_host is not null

order by hour asc,day asc,month asc,pv\_referer\_cnt desc;

--统计每小时各来访host的产生的pv数并排序

-- 以小时和host为分组对象在求count()

drop table dw\_pvs\_refererhost\_everyhour;

create table dw\_pvs\_refererhost\_everyhour(ref\_host string,month string,day string,hour string,ref\_host\_cnts bigint) partitioned by(datestr string);

insert into table dw\_pvs\_refererhost\_everyhour partition(datestr='20130918')

select ref\_host,month,day,hour,count(1) as ref\_host\_cnts

from ods\_weblog\_detail

group by ref\_host,month,day,hour

having ref\_host is not null

order by hour asc,day asc,month asc,ref\_host\_cnts desc;

---------------------------------------------------------------------------

## 统计pv总量最大的来源TOPN

--需求：按照时间维度，统计一天内各小时产生最多pvs的来源topN

--row\_number函数

select ref\_host,ref\_host\_cnts,concat(month,day,hour),

row\_number() over (partition by concat(month,day,hour) order by ref\_host\_cnts desc) as od

from dw\_pvs\_refererhost\_everyhour;

--综上可以得出

drop table dw\_pvs\_refhost\_topn\_everyhour;

create table dw\_pvs\_refhost\_topn\_everyhour(

hour string,

toporder string,

ref\_host string,

ref\_host\_cnts string

)partitioned by(datestr string);

insert into table dw\_pvs\_refhost\_topn\_everyhour partition(datestr='20130918')

select t.hour,t.od,t.ref\_host,t.ref\_host\_cnts from

(select ref\_host,ref\_host\_cnts,concat(month,day,hour) as hour,

row\_number() over (partition by concat(month,day,hour) order by ref\_host\_cnts desc) as od

from dw\_pvs\_refererhost\_everyhour) t where od<=3;

---------------------------------------------------------------------------------------------

## 人均浏览页数

--需求描述：统计今日所有来访者平均请求的页面数。

--总页面请求数/去重总人数

drop table dw\_avgpv\_user\_everyday;

create table dw\_avgpv\_user\_everyday(

day string,

avgpv string);

insert into table dw\_avgpv\_user\_everyday

select '20130918',sum(b.pvs)/count(b.remote\_addr)

from

(select remote\_addr,count(1) as pvs from ods\_weblog\_detail where datestr='20130918'

group by remote\_addr) b;