### 为什么要分桶

跟MR中的HashPartitioner的原理一模一样

MR中：按照key的hash值去模除以reductTask的个数

Hive中：按照分桶字段的hash值去模除以分桶的个数

Hive也是 针对某一列进行桶的组织。Hive采用对列值哈希，然后除以桶的个数求余的方式决定该条记录存放在哪个桶当中。

### 2.数据分桶的作用:

好处：

1、方便抽样

2、提高join查询效率

（1）获得更高的查询处理效率。桶为表加上了额外的结构，Hive 在处理有些查询时能利用这个结构。具体而言，连接两个在（包含连接列的）相同列上划分了桶的表，可以使用 Map 端连接 （Map-side join）高效的实现。比如JOIN操作。对于JOIN操作两个表有一个相同的列，如果对这两个表都进行了桶操作。那么将保存相同列值的桶进行JOIN操作就可以，可以大大较少JOIN的数据量。

（2）使取样（sampling）更高效。在处理大规模数据集时，在开发和修改查询的阶段，如果能在数据集的一小部分数据上试运行查询，会带来很多方便。

3.

hive> create table t6(name string,sal bigint,city string)

clustered by(sal,city) into 3 buckets

row format delimited fields terminated by ","

stored as textfile;

启动自动分桶

hive> set hive.enforce.bucketing = true;

\*\*\*\*\*\*\*

可以设置reduce数量

hive> set mapreduce.job.reduces = 3;

\*\*\*\*\*\*\*

导入数据

hive> insert into table t6 select \* from t55;

查看结果

[hadoop@h201 hhh]$ hadoop fs -lsr /hive1/t6

按分桶抽样数据

hive> select \* from t6 tablesample(bucket 1 out of 2 on sal,city);

(桶id默认从1开始)

数据分桶存在的一些缺陷：

如果通过数据文件LOAD 到分桶表中，会存在额外的MR负担

实际生产中分桶策略使用频率较低，更常见的还是使用数据分区