Hive 数据抽样的几种方法

在大规模数据量的数据分析及建模任务中，往往针对全量数据进行挖掘分析时会十分耗时和占用集群资源，因此一般情况下只需要抽取一小部分数据进行分析及建模操作。本文就介绍 Hive 中三种数据抽样的方法

块抽样（Block Sampling）

Hive 本身提供了抽样函数，使用 TABLESAMPLE 抽取指定的 行数/比例/大小，举例：

```
CREATE TABLE iteblog AS SELECT * FROM iteblog1 TABLESAMPLE(1000 ROWS);
CREATE TABLE iteblog AS SELECT * FROM iteblog1 TABLESAMPLE (20 PERCENT);
CREATE TABLE iteblog AS SELECT * FROM iteblog1 TABLESAMPLE(1M);
```

缺点
：不随机。该方法实际上是按照文件中的顺序返回数据，对分区表，从头开始抽取，可能造成只有前面几个分区的数据。
优点：速度快。

分桶表抽样（Smapling Bucketized Table）

利用分桶表，随机分到多个桶里，然后抽取指定的一个桶。举例：随机分到10个桶，抽取第一个桶。

```
CREATE TABLE iteblog AS SELECT * FROM iteblog1 TABLESAMPLE (BUCKET 1 OUT OF 10 ON rand());
```

优点：随机，测试发现，速度比方法3的rand()快。

随机抽样

原理：利用 rand() 函数进行抽取，rand() 返回一个0到1之间的 double 值。

使用方法一

```
CREATE TABLE iteblog AS
```

```
SELECT * FROM iteblog1
ORDER BY rand()
limit 10000
```

此时，可以提供真正的随机抽样，但是，需要在单个 reducer 中进行总排序，速度慢。

使用方法二

```
CREATE TABLE iteblog AS
SELECT * FROM iteblog1
SORT BY rand()
limit 10000
```

Hive 提供了 sort by，sort by 提供了单个 reducer
内的排序功能，但不保证整体有序，上面的语句是不保证随机性的。

使用方法三

```
CREATE TABLE iteblog AS
SELECT * FROM iteblog1
where rand()<0.002
distribute by rand()
sort by rand()
limit 10000;
```

where 条件首先进行一次 map 端的优化，减少 reducer
需要处理的数据量，提高速度。distribute by 将数据随机分布，然后在每个 reducer
内进行随机排序，最终取10000条数据（如果数据量不足，可以提高 where 条件的 rand
过滤值）。
缺点：速度慢

使用方法四

```
CREATE TABLE iteblog AS
SELECT * FROM iteblog1
where rand()<0.002
```

```
 cluster by rand()
limit 10000;
```

cluster by 的功能是 distribute by 和 sort by 的功能相结合，distribute by rand() sort by rand() 进行了两次随机，cluster by rand() 仅一次随机，所以速度上会比上一种方法快。

随机结果里面添加分区

上面的几种方法会丢失掉分区信息，我们可以结合动态分区将分区信息加到结果集中，具体如下：

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

INSERT INTO TABLE iteblog partition(thedate)
SELECT * FROM iteblog1 TABLESAMPLE (BUCKET 1 OUT OF 10 ON rand());
```