Hive动态分区和分桶

- What?Why?How?

动态分区

- hive 动态分区
  - 开启支持动态分区
    - set hive.exec.dynamic.partition=true;
      - 默认：true
    - set hive.exec.dynamic.partition.mode=nostrict;
      - 默认：strict（至少有一个分区列是静态分区）
    - 相关参数
      - set hive.exec.max.dynamic.partitions.pernode;
        - 每一个执行mr节点上，允许创建的动态分区的最大数量(100)
      - set hive.exec.max.dynamic.partitions;
        - 所有执行mr节点上，允许创建的所有动态分区的最大数量(1000)
      - set hive.exec.max.created.files;
        - 所有的mr job允许创建的文件的最大数量(100000)

动态分区

- 加载数据

```
from psn21
insert overwrite table psn22 partition(age, sex)
select id, name, age, sex, likes, address distribute by age, sex;
```

Hive分桶

- hive 分桶
  - 分桶表是对列值取哈希值的方式，将不同数据放到不同文件中存储。
  - 对于hive中每一个表、分区都可以进一步进行分桶。
  - 由列的哈希值除以桶的个数来决定每条数据划分在哪个桶中。
- 适用场景：
  - 数据抽样（ sampling ）

Hive 分桶

- 开启支持分桶
  - set hive.enforce.bucketing=true;
    - 默认：false；设置为true之后，mr运行时会根据bucket的个数自动分配reduce task个数。（用户也可以通过mapred.reduce.tasks自己设置reduce任务个数，但分桶时不推荐使用）
    - 注意：一次作业产生的桶（文件数量）和reduce task个数一致。
- 往分桶表中加载数据
  - insert into table bucket_table select columns from tbl;
  - insert overwrite table bucket_table select columns from tbl;

Hive 分桶

- 桶表 抽样查询
  - select * from bucket_table tablesample(bucket 1 out of 4 on columns);
- TABLESAMPLE语法：
  - TABLESAMPLE(BUCKET x OUT OF y)
  - x：表示从哪个bucket开始抽取数据
  - y：必须为该表总bucket数的倍数或因子

Hive 分桶

- 例：
  - 当表总bucket数为32时
  - TABLESAMPLE(BUCKET 3 OUT OF 8)，抽取哪些数据？
    - 共抽取2（32/16）个bucket的数据，抽取第2、第18（16+2）个bucket的数据
  - TABLESAMPLE(BUCKET 3 OUT OF 256)，抽取哪些数据？
    - ？

Hive 分桶

- 例:

  CREATE TABLE psn31(id INT, name STRING, age INT)

  ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';

  测试数据:

  1,tom,11

  2,cat,22

  3,dog,33

  4,hive,44

  5,hbase,55

  6,mr,66

  7,alice,77

  8,scala,88

Hive 分桶

- 创建分桶表

```
CREATE TABLE psnbucket( id INT, name STRING, age INT)
CLUSTERED BY (age) INTO 4 BUCKETS
ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

- 加载数据：

```
insert into table psnbucket select id, name, age from psn31;
```

- 抽样

```
select id, name, age from psnbucket tablesample(bucket 2 out of 4 on
age);
```