Hive6 数据操作

> Hive中没有行级别的数据插入、数据更新和删除操作，唯一途径就是使用“大量”的数据装载操作。

向管理表中装载数据

```
LOAD DATA LOCAL INPATH '${env:HOME}/california-employees'
OVERWRITE INTO TABLE employees
PARTITION (country = 'US', state = 'CA');
```

- partition：如果分区目录不存在的话，这个命令会先创建分区目录，然后再将数据拷贝到该目录下。如果目标表是非分区表，那么语句中省略PARTITION子句
- local： 如果使用了LOCAL这个关键字，那么这个路径应该为本地文件系统路径，数据将会被拷贝到目标位置。如果省略掉LOCAL这个关键字，那么这个路径应该是分布式文件系统中的路径。
- OVERWRITE： 如果制定了这个关键字，那么目标文件中之前存在的数据将会被先删除掉，如果没有这个关键字，仅仅会把新增的文件加到目标文件中而不会删除之前的数据

通过查询语句向表中插入数据

```
INSERT OVERWRITE TABLE employees
PARTITION (country = 'US', state = 'OR')
SELECT * FROM staged_employees se
WHERE se.cnty = 'US' AND se.st = 'OR';
```

- OVERWRITE: 之前分区中的内容将会被覆盖掉,否则以追加的方式写入文件
- 场景：数据已经存在于某个目录下，对于Hive来说其为一个外部表，而现在想将其导入到最终的分区表中。如果用户想将源表数据导入到一个具有不同记录格式的目标。

如果staged_employees 非常大，而且用户需要对 65个州都执行这些语句，那么也就意味着需要扫描表65次。Hive提供了另一种插入方法。

```
FROM staged_employees se
INSERT OVERWRITE TABLE employees
    PARTITION (country = 'US', state = 'OR')
    SELECT * FROM WHERE se.cnty = 'US' AND se.st = 'OR';
INSERT OVERWRITE TABLE employees
    PARTITION (country = 'US', state = 'CA')
    SELECT * FROM WHERE se.cnty = 'US' AND se.st = 'CA';
INSERT OVERWRITE TABLE employees
    PARTITION (country = 'US', state = 'IL')
    SELECT * FROM WHERE se.cnty = 'US' AND se.st = 'IL';
```

动态分区插入

> 需要创建非常多的分区，那么用户就需要写非常多的SQL，Hive提供了一个动态分区功能，其可以基于查询参数推断出需要创建的分区名称。

```
INSERT OVERWRITE TABLE employees
PARTITION (country, state)
SELECT ..., se.cnty, se.st
FROM staged_employees se;
```

假设表staged_employees 中共有100个国家和州，执行完上面的查询后，表employees就有100个分区。

也可以混合静态和动态,**其中，静态分区必须出现在动态分区键之前**

```
INSERT OVERWRITE TABLE employees
PARTITION (country = 'US', state)
SELECT ..., se.cnty, se.st
FROM staged_employees se
WHERE se.cnty = 'US';
```

动态分区的属性值：

属性名称	缺省值	描述
hive.exec.dynamic.partition	false	设置为true，表示开启动态分区功能
hive.exec.dynamic.partition.mode	strict	设置为nonstrict,表示允许所有分区都是动态的
hive.exec.max.dynamic.partitions.pernode	100	每个mapper或reducer可以创建的最大动态分区个数，如果某个mapper或reducerr尝试打羽这个值会报错
hive.exec.max.dynamic.partitions	+1000	一个动态分区创建语句可以创建的最大动态分区个数
hive.exec.max.created.files	100000	全局可以创建最大文件数

单个查询语句中创建表并加载数据

```
CREATE TABLE ca_employees
AS SELECT name, salary, adress
FROM employees
WHERE se.state = 'CA';
```

这个功能常用于从一个大的宽表中选取部分需要的数据集

导出数据

```
hadoop fs -cp source_path target_path
```

或者

```
INSERT OVERWRITE LOCAL DIRECTORY '/tmp/ca_employees'
SELECT name, salsry, adress
FROM employees
WHERE se.state = 'CA';
```

可以指定多个输出文件夹目录的：

```
FROM staged_employees se
INSERT OVERWRITE LOCAL DIRECTORY '/tmp/ca_employees'
    SELECT *  WHERE se.cty = 'US' and se.st = 'OR';
INSERT OVERWRITE LOCAL DIRECTORY '/tmp/ca_employees'
    SELECT *  WHERE se.cty = 'US' and se.st = 'CA';
INSERT OVERWRITE LOCAL DIRECTORY '/tmp/ca_employees'
    SELECT *  WHERE se.cty = 'US' and se.st = 'IL';
```