Hadoop学习之Hive编程指南学习(四)

1. HiveQL的数据操作
   1. 向管理表中装载数据

Hive没有行级别的数据插入、更新、删除操作，那么向表中装载数据就只能是使用大量的数据装载操作，或通过其他方式仅仅将文件写入到正确的目录下。

Load data local inpath '${env:HOME}/California-employees'

Overwrite into table employees

Partition {country = 'US',state = 'CA'};

如果分区不存在的话，会先创建分区，然后将数据复制到该目录下。如果目标表是非分区表，语句中省掉partition子句。

通常情况下，路径指定的应该是个目录而不是单独文件，hive会将所有文件都复制到这个目录中，这是的用户将更方便的组织数据到多文件中，同时在不修改hive脚本的前提下修改文件命名规则。不管怎样，文件都会被复制到目标表路径下而且文件名保持不变。

命令中使用local则路径应该是本地FS的路径;如果省略掉local，则路径是HDFS上的路径。

PS: 用户不可以使用load data 语句将数据从一个集群的HDFS中转移到另一个集群的HDFS中。

如果用户指定了overwrite，目标文件夹执勤啊的数据被先被删除。

如果目标表示分区表那么需要使用partition子句，而用户还必须为每个分区的键指定值。

Inpath子句中的文件路径不可以包含任何文件夹。

* 1. 通过查询语句向表中插入数据

Insert语句语序用户通过查询语句向表中插入数据。依旧使用emp表做例子，假设还有表staged\_employees表。

Insert overwrite table emp

Partition(country = 'US',state = 'OR')

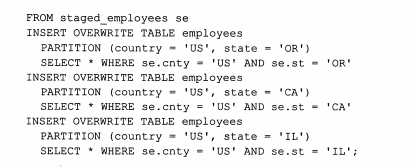
Select \* from staged\_employees se

Where se.cnty ='US' and se.st = 'OR';

这个例子展示了这个功能非常有用的场景：数据存在于某个目录，但对hive是个外部表，现在想将其导入最终的分区表中。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表的话，使用这种方式也是很好的。

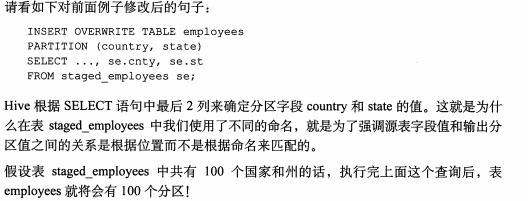
如果stated\_employees表非常大，用户需要对美国65个州执行这些语句，那就意味着需要扫描stated\_employees表65次，但hive有一种方法可以只扫描1次，然后按多种方式划分。

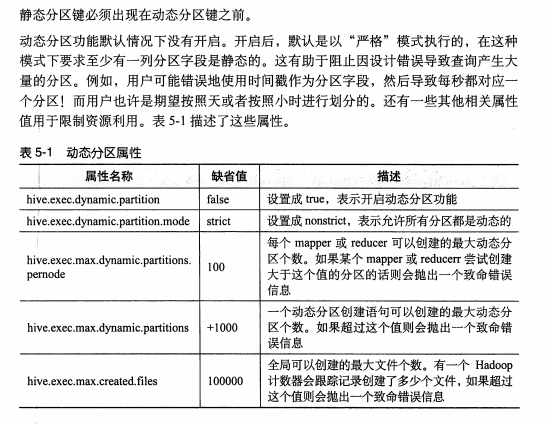
创建表employes分区



但需要创建非常多的分区，用户就要写非常多的sql，不过hive提供了动态分区功能，可以基于查询参数推断出需要创建的分区名称。

（动态分区插入）





* 1. 单个查询语句中创建表并加载数据

Create table ca\_emp

As select name,salary,address

From emp

Where se.state = 'CA';

常用于在大的宽表中选取部分需要的数据集。不能用于外部表。

* 1. 导出数据

如果数据文件时用户需要的格式，只需要复制文件或目录

Hadoop fs –cp source\_path target\_path

否则，使用insert … directory …

Insert overwrite local directory '/tmp/ca\_emp'

Select name,salary,address

From emp

Where se.state ='CA';

路径也可以写全的URL(hdfs://…/…).

不管源表数据如何存储,hive会将所有字段序列化成字符串写入到文件中。