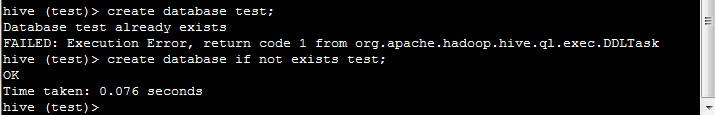
Hadoop学习之Hive编程指南学习(三)

1. HiveQL的数据定义

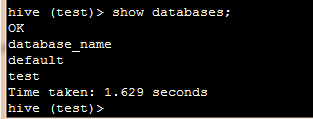
HiveQL是hive查询语言，不完全遵守ANSI SQL标准，hiveQL与mysql的方言接近。但是hive不支持行级插入，更新和删除；不支持事务。

Hive中的DB本质上是表的目录或者命名空间。但好处是：对于有很多组和用户的大集群这是很有用的，可以避免表命名冲突。还会使用数据库将生产表组织成逻辑组。

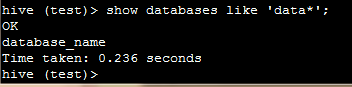
1. 创建数据库
   1. hive (test)> create database test; --直接创建数据库，如果存在会打印警告信息
   2. hive (test)> create database if not exists test; --只有在数据库不存在时才创建。



* 1. hive (test)> show databases; --显示已经存在的数据库



* 1. hive (test)> show databases like ‘data\*’;-- 数据库太多也可以使用正则表达式筛选

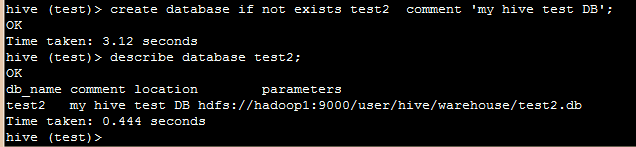


* 1. Hive会为每个DB创建一个目录。DB中的table将会以DB目录中子目录的形式存储，除了default(没有目录)。

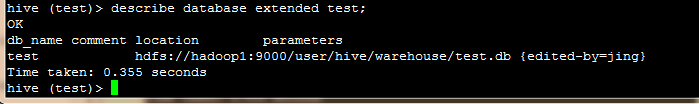
数据库所在目录位于属性hive.metastore.warehouse.dir下面，是该目录的子目录。

hive (test)> create database test location '/my/preferred/dirctory'; --修改数据库默认位置。

添加描述信息,使用describe database;还能显示出数据库的位置



* 1. hive (test)> describe database extended test; --显示带属性的数据库信息

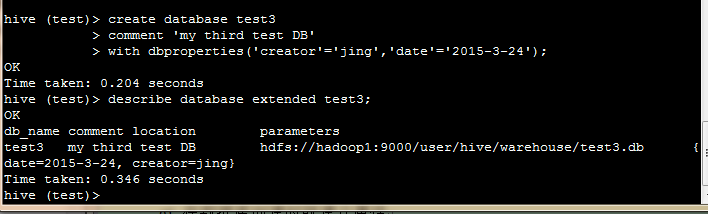


* 1. 在数据库创建时就建立属性和备注

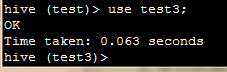
hive (test)> create database test3

> comment 'my third test DB'

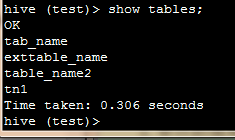
> with dbproperties('creator'='jing','date'='2015-3-24');



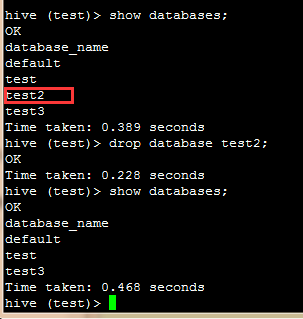
* 1. 切换数据库 hive (test)> use test3;



* 1. 显示当前数据库中的表 hive (test)> show tables;

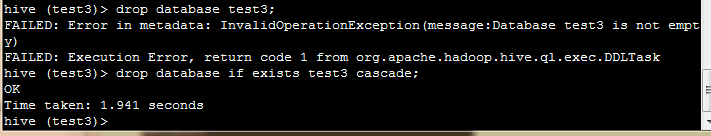


* 1. 删除数据库



删除存在表的数据库

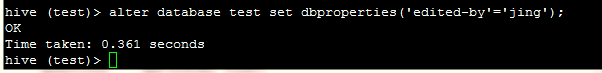
hive (test3)> drop database if exists test3 cascade;



1. 修改数据库

可以使用alter database 命令为数据库设置dbproperties设置键值对属性值，用来描述数据库的属性信息。数据库的其他元数据心心都是不可更改的，包括数据库名和数据库所在目录位置。而且也不可以删除或者重置数据库属性。

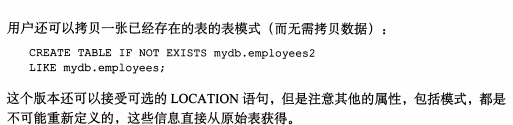
alter database test set dbproperties('edited-by'='jing');



1. 创建表
   1. 简述

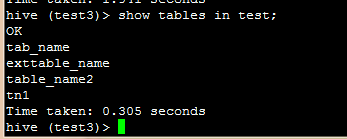


Hive会自动增加两个表属性：last\_modified\_by （保存最后修改表的用户的用户名），last\_modified\_time（保存最后修改的新纪元时间秒）

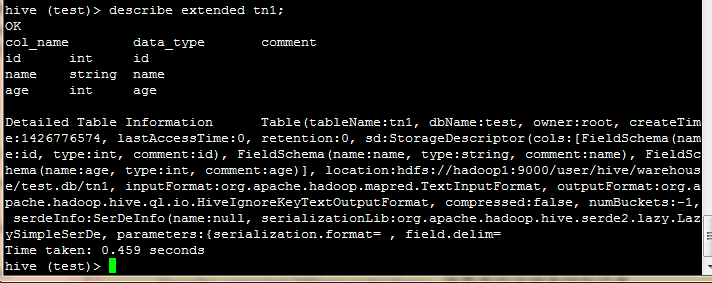


显示指定数据库的表

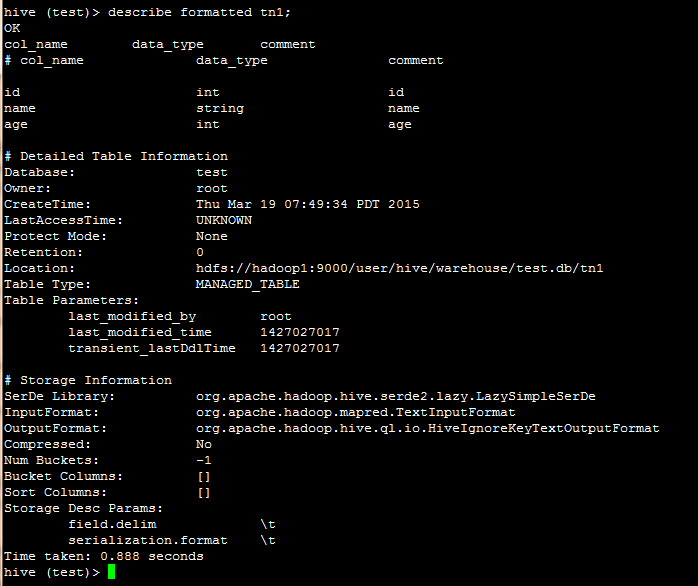
数据库当前位置在test3，但显示的test中的表



Describe extended DBName.tableName 查看表的详细表结构信息



Formatted关键字代替extended可以提供更加刻度的和荣昌的输出信息。



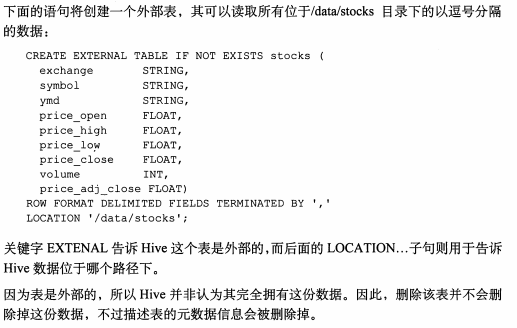
* 1. 管理表

默认创建的是管理表，即内部表。Hive控制着数据的生命周期，数据存储在hive.Metastore.warehouse.dir目录的子目录下。

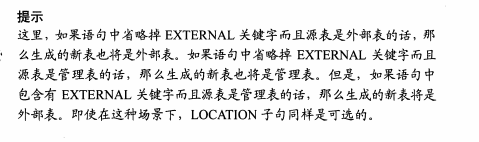
当删除一个管理表时，也会删除表中的数据。

但是管理表不方便和其他工作表共享数据。这时就需要外部表指向这份数据。

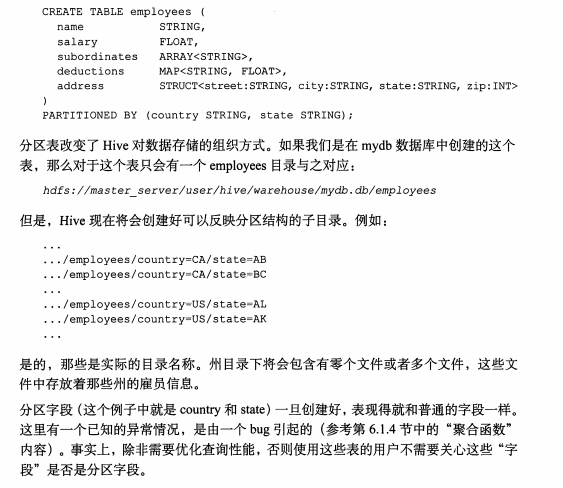
* 1. 外部表

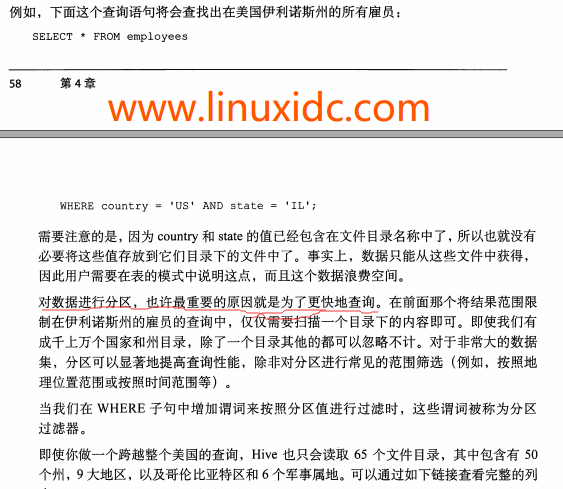


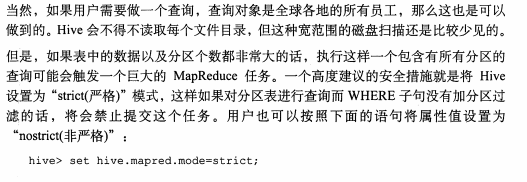
管理表和外部表的区别：有些hiveQL语法结构不适用于外部表。



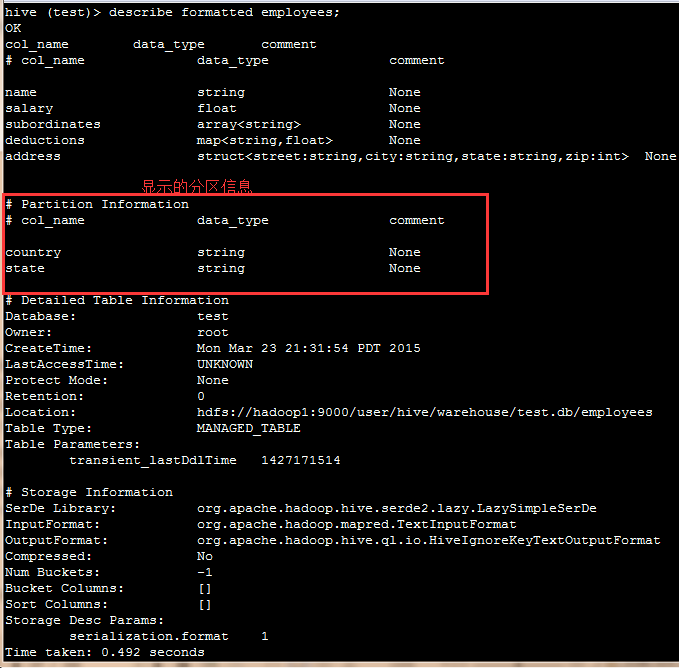
* 1. 分区表和管理表



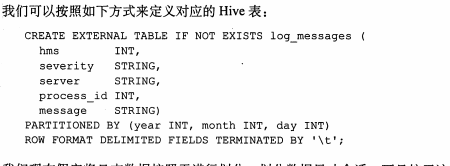




* 1. 用describe formatted table\_name;显示分区键



* 1. 外部分区表

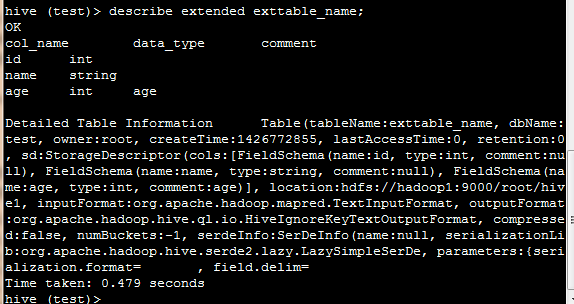


Hive不关心一个分区对应的分区目录是否存在或者分区目录下是否有文件，如果分区目录不存在或者分区目录下没有问及那，则查询将没有返回结果。当用户想在另一个进程开始往分区中写数据之前创建好分区时，这样做是很方便的。数据一旦存在，对这份数据的查询就会有返回结果。

另一个好处是：将数据写入到一个专用的目录中，并与其他目录中的数据存在明显的区别。同时不管用户是将旧数据转移到一个存档位置还是直接删除掉，新数据被篡改的风险都被降低，因为新数据的数据子集位于不同的目录下。

Hive不控制外部表的数据(不论有没有分区)，即使表被删除，数据也不删除。

查看外部表的详细信息与查看管理表命令类似



* 1. 自定义表的存储格式

使用stored as textfile显示指定，默认存储的是文本文件格式。

* 1. 删除表

Drop table if exists table\_name



Hadoop有回收站功能，默认关闭。如果开启回收站，数据会被转移到用户所在DFS中的用户根目录下的.Trash目录下，也就是HDFS中的/user/$USER/.Trash目录。

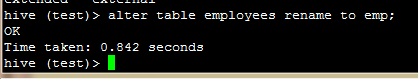
如果想启动该功能，只需要在配置属性fs.trash.interval的值设置为合理的正整数即可。该值是“回收站检查点”间的时间间隔，单位是分钟。

* 1. 修改表

一般通过alter table语句修改，会修改元数据，不会修改数据本身。可用于修改表模式中出现的错误、改变分区路径等。

* + 1. 表重命名

Alter table old\_name rename to new\_name;



* + 1. 增加、修改和删除表分区

Alter table table\_name add if not exists

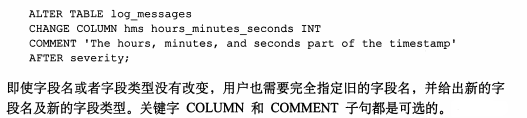
Partition(year =2015,month =1,day =1) location '/log/2015/01/01'

Partition(year =2015,month =1,day =2) location '/log/2015/01/02'

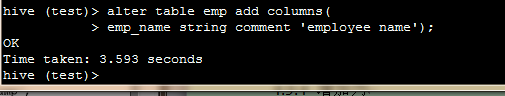
Partition(year =2015,month =1,day =3) location '/log/2015/01/03';

* + 1. 修改列属性

用户可对某个字段重命名，修改位置、类型和注释

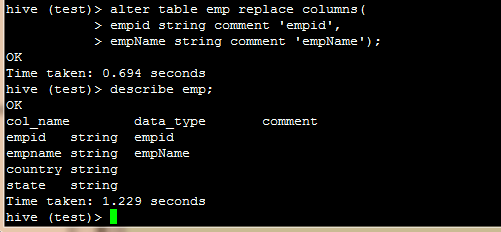


* + 1. 增加列



Comment 是备注，是可选的。

* + 1. 删除或者替换列

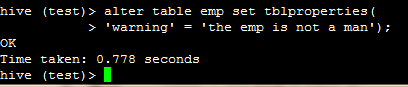


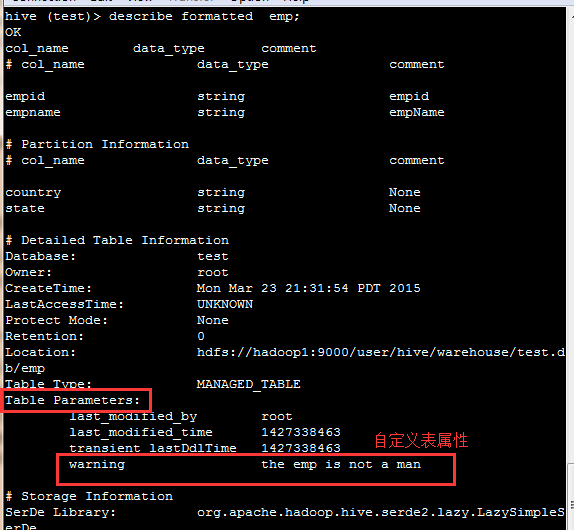
这样会替换掉原来就有的列，但是分区是不会删除的。只有表的元数据信息改变了。

* + 1. 修改表属性

Alter table table\_name set tblproperties(

‘notes’ = ‘the emp is leave ’);





1. 等等
   1. 嗯嗯