Hive表与外部表

写在前面的话
，学Hive这么久了，发现目前国内还没有一本完整的介绍Hive的书籍，而且互联网上面的资料很乱，于是我决定写一些关于《Hive的那些事》
序列文章，分享给大家。我会在接下来的时间整理有关Hive的资料，如果对Hive的东西感兴趣，请关注本博客。/archives/tag/hive的那些事

这几天比较忙，公司里面各种事，所以这几天博客没有时间更新，今天特意抽了一点时间来更新本博客，继续我们的《Hive的那些事》。

好了，进入正题。今天我们要探讨的话题是Hive的里面的表与外部表两个概念，以及如何在Hive里面创建表和外部表，
它们之间有什么区别等话题。在本博客的《Hive的数据存储模式》
文章里面我们谈到了Hive的数据存储模式，里面简单的说到Hive中表以及外部表的简单概念，相信很多读者对这些概念还不是很了解，今天就给大家科普一下，希望对大家有所帮助。

相信很多用户都用过关系型数据库，我们可以在关系型数据库里面创建表（create table），这里要讨论的表和关系型数据库中的表在概念上很类似。我们可以用下面的语句在Hive里面创建一个表：

```
hive> create table wyp(id int,
    > name string,
    > age int,
    > tele string)
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '₩t'
    > STORED AS TEXTFILE;
OK
Time taken: 0.759 seconds
```

这样我们就在Hive里面创建了一张普通的表，现在我们给这个表导入数据：

```
hive> load data local inpath '/home/wyp/data/wyp.txt' into table wyp;
Copying data from file:/home/wyp/data/wyp.txt
Copying file: file:/home/hdfs/wyp.txt
Loading data to table default.wyp
Table default.wyp stats: [num_partitions: 0, num_files: 1,
        num_rows: 0, total_size: 67, raw_data_size: 0]
OK
Time taken: 3.289 seconds
hive> select * from wyp;
```

```
OK
1       wyp     25      13188888888888
2       test    30      13888888888888
3       zs      34      899314121
Time taken: 0.41 seconds, Fetched: 3 row(s)
```

注意：/home/wyp/data/路径是Linux本地文件系统路径；而/home/hdfs/是HDFS文件系统上面的路径！从上面的输出我们可以看到数据是先从本地的/home/wyp/data/文件夹下复制到HDFS上的/home/hdfs/wyp.txt(这个是Hive中的配置导致的)文件中！最后Hive将从HDFS上把数据移动到wyp表中！移到表中的数据到底存放在HDFS的什么地方？其实在Hive的${HIVE_HOME}/conf/hive-site.xml配置文件的hive.metastore.warehouse.dir属性指向的就是Hive表数据存放的路径（在我的店电脑里面配置是/user/hive/warehouse），而Hive每创建一个表都会在hive.metastore.warehouse.dir指向的目录下以表名创建一个文件夹，所有属于这个表的数据都存放在这个文件夹里面。所以，刚刚导入到wyp表的数据都存放在/user/hive/warehouse/wyp/文件夹中，我们可以去看看：

```
hive> dfs -ls /user/hive/warehouse/wyp ;
Found 1 items
-rw-r--r-- 3 wyp supergroup   67 2014-01-14 22:23 /user/hive/warehouse/wyp/wyp.txt
```

看到没，上面的命令就是显示HDFS上的/user/hive/warehouse/wyp中的所有内容。如果需要删除wyp表，可以用下面的命令：

```
hive> drop table wyp;
Moved: 'hdfs://mycluster/user/hive/warehouse/wyp' to
    trash at: hdfs://mycluster/user/hdfs/.Trash/Current
OK
Time taken: 2.503 seconds
```

从上面的输出Moved: 'hdfs://mycluster/user/hive/warehouse/wyp' to trash at: hdfs://mycluster/user/hdfs/.Trash/Current我们可以得知，原来属于wyp表的数据被移到hdfs://mycluster/user/hdfs/.Trash/Current文件夹中（如果你的Hadoop没有取用垃圾箱机制，那么drop table wyp命令将会把属于wyp表的所有数据全部删除！），其实就是删掉了属于wyp表的数据。记住这些，因为这些和外部表有很大的不同。同时，属于表wyp的元数据也全部删除了！
我们再来创建一个外部表：

```
hive> create external table exter_table(
```

```
    > id int,
    > name string,
    > age int,
    > tel string)
    > location '/home/wyp/external';
OK
Time taken: 0.098 seconds
```

仔细观察一下创建表和外部表的区别，仔细的同学们一个会发现创建外部表多了external关键字说明以及location '/home/wyp/external'。是的，你说对了！如果你需要创建外部表，需要在创建表的时候加上external关键字，同时指定外部表存放数据的路径（当然，你也可以不指定外部表的存放路径，这样Hive将在HDFS上的/user/hive/warehouse/文件夹下以外部表的表名创建一个文件夹，并将属于这个表的数据存放在这里）：

```
hive> load data local inpath '/home/wyp/data/wyp.txt' into table exter_table;
Copying data from file:/home/wyp/data/wyp.txt
Copying file: file:/home/hdfs/wyp.txt
Loading data to table default.exter_table
Table default.exter_table stats: [num_partitions: 0, num_files:
        1, num_rows: 0, total_size: 67, raw_data_size: 0]
OK
Time taken: 0.456 seconds
```

和上面的导入数据到表一样，将本地的数据导入到外部表，数据也是从本地文件系统复制到HDFS中/home/hdfs/wyp.txt文件中，但是，最后数据不是移动到外部表的/user/hive/warehouse/exter_table文件夹中（除非你创建表的时候没有指定数据的存放路径）！大家可以去HDFS上看看！对于外部表，数据是被移动到创建表时指定的目录（本例是存放在/home/wyp/external文件夹中）！如果你要删除外部表，看看下面的操作：

```
hive> drop table exter_table;
OK
Time taken: 0.093 seconds
```

和上面删除Hive的表对比可以发现，没有输出将数据从一个地方移到任一个地方！那是不是删除外部表的的时候数据直接被删除掉呢？答案不是这样的：

```
hive> dfs  -ls /home/wyp/external;
```

```
Found 1 items
-rw-r--r--  3 wyp supergroup 67 2014-01-14 23:21 /home/wyp/external/wyp.txt
```

你会发现删除外部表的时候，数据并没有被删除，这是和删除表的数据完全不一样的！

最后归纳一下Hive中表与外部表的区别：

1、在导入数据到外部表，数据并没有移动到自己的数据仓库目录下，也就是说外部表中的数据并不是由它自己来管理的！而表则不一样；

2、在删除表的时候，Hive将会把属于表的元数据和数据全部删掉；而删除外部表的时候，Hive仅仅删除外部表的元数据，数据是不会删除的！

那么，应该如何选择使用哪种表呢？在大多数情况没有太多的区别，因此选择只是个人喜好的问题。但是作为一个经验，如果所有处理都需要由Hive完成，那么你应该创建表，否则使用外部表！

（今天就到这吧，已经深夜23：32了，程序员的生活很苦逼的！）