

hadoop跨集群之间迁移hive数据

1. 迁移hdfs数据至新集群

```
#!/bin/bash
for DB in `cat db_name.txt`
do
sudo -u hive hadoop distcp -i -overwrite
hdfs://10.15.93.13:8020/user/hive/warehouse/$DB.db
hdfs://10.30.18.11:8020/user/hive/warehouse/$DB.db
done
```

注意：用IP(HA取active的) 不要用nameservice1 ，一个库一库的来，直接用warehouse出错了 不好弄

2. 源集群metastore数据备份导出(mysql导出)

```
mysqldump -u root -p'密码' --skip-lock-tables -h xxx.xxx.xxx.xxx hive
> mysql_hive.sql
```

```
mysqldump -uroot -p --database hive > mysql_hive_data.sql (我的环境操作)
```

3. 新的集群导入metastore数据(MySQL导入)

```
mysql -u root -proot --default-character-set=utf8 hvie <
mysql_hive.sql
```

```
mysql -uroot -p < mysql_data.sql (我的环境操作)
```

4. 升级hive内容库(如果hive版本需要升级操作, 同版本不需要操作)

```
mysql -uroot -proot risk -hxxx.xxx.xxx.xxx < mysqlupgrade-0.13.0-to-
0.14.0.mysql.sql
```

```
mysql -uroot -proot risk -hxxx.xxx.xxx.xxx < mysqlupgrade-0.14.0-to-1.1.0.mysql.sql
```

版本要依据版本序列升序升级,不可跨越版本,如当前是hive0.12打算升级到0.14,需要先升级到0.13再升级到0.14

5. 修改 metastore 内容库的集群信息 (重要)

因为夸集群, hdfs访问的名字可能变化了,所以需要修改下hive库中的表DBS和SDS内容,除非你的集群名字或者HA的名字跟之前的一致这个就不用修改了

6. hue库迁移。

```
mysqldump -uroot -p --database hue > mysql_hue_data.sql
```

```
mysql -u root -proot --default-character-set=utf8 hue < mysql_hue_data.sql
```

distcp 命令是用于集群内部或者集群之间拷贝数据的常用命令。

顾名思义: dist即分布式, 分布式拷贝是也。

```
hadoop distcp [选项] src_url dest_url
```

常用选项

-m 表示启用多少map

-delete: 删除已经存在的目标文件, 不会删除源文件。这个删除是通过FS Shell实现的。所以如果垃圾回收机制启动的话, 删除的目标文件会进入trash。

-i: 忽略失败。这个选项会比默认情况提供关于拷贝的更精确的统计, 同时它还将保留失败拷贝操作的日志, 这些日志信息可以用于调试。最后, 如果一个map失败了, 但并没完成所有分块任务的尝试, 这不会导致整个作业的失败。

-overwrite: 覆盖目标。如果一个map失败并且没有使用-i选项, 不仅仅那些拷贝失败的文件, 这个分块任务中的所有文件都会被重新拷贝。 *所以这就是为什么要使用-i参数。*

