# 数据库内容脱敏

## 安装Python及Python版本的MySQL驱动

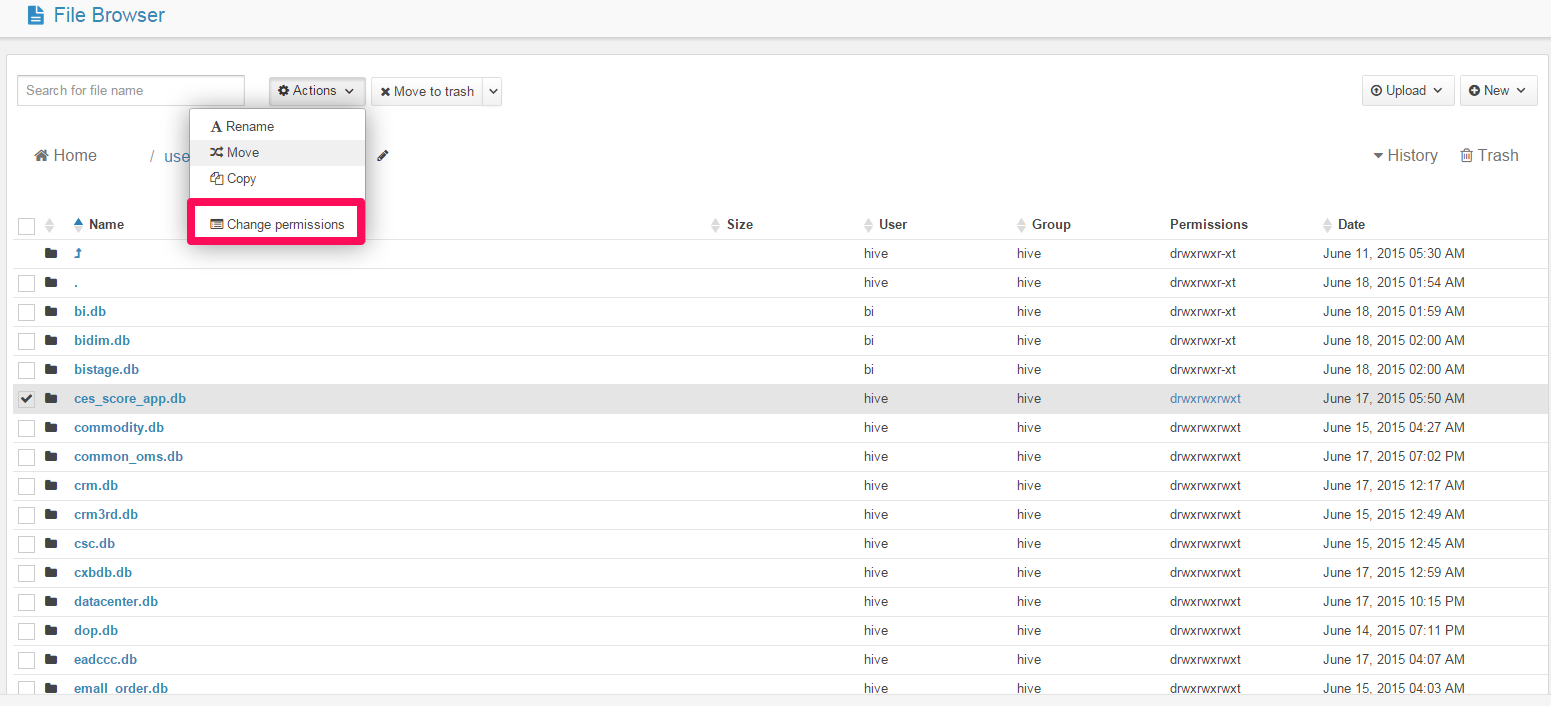
1. 安装Python 2.7. Python的安装目录设置为D:/Python27。
2. 设置Path环境变量为D:/Python27及D:\Python27\Scripts。
3. 使用pip install mysql-connector-python来安装Python版本的MySQL驱动。

## 数据脱敏

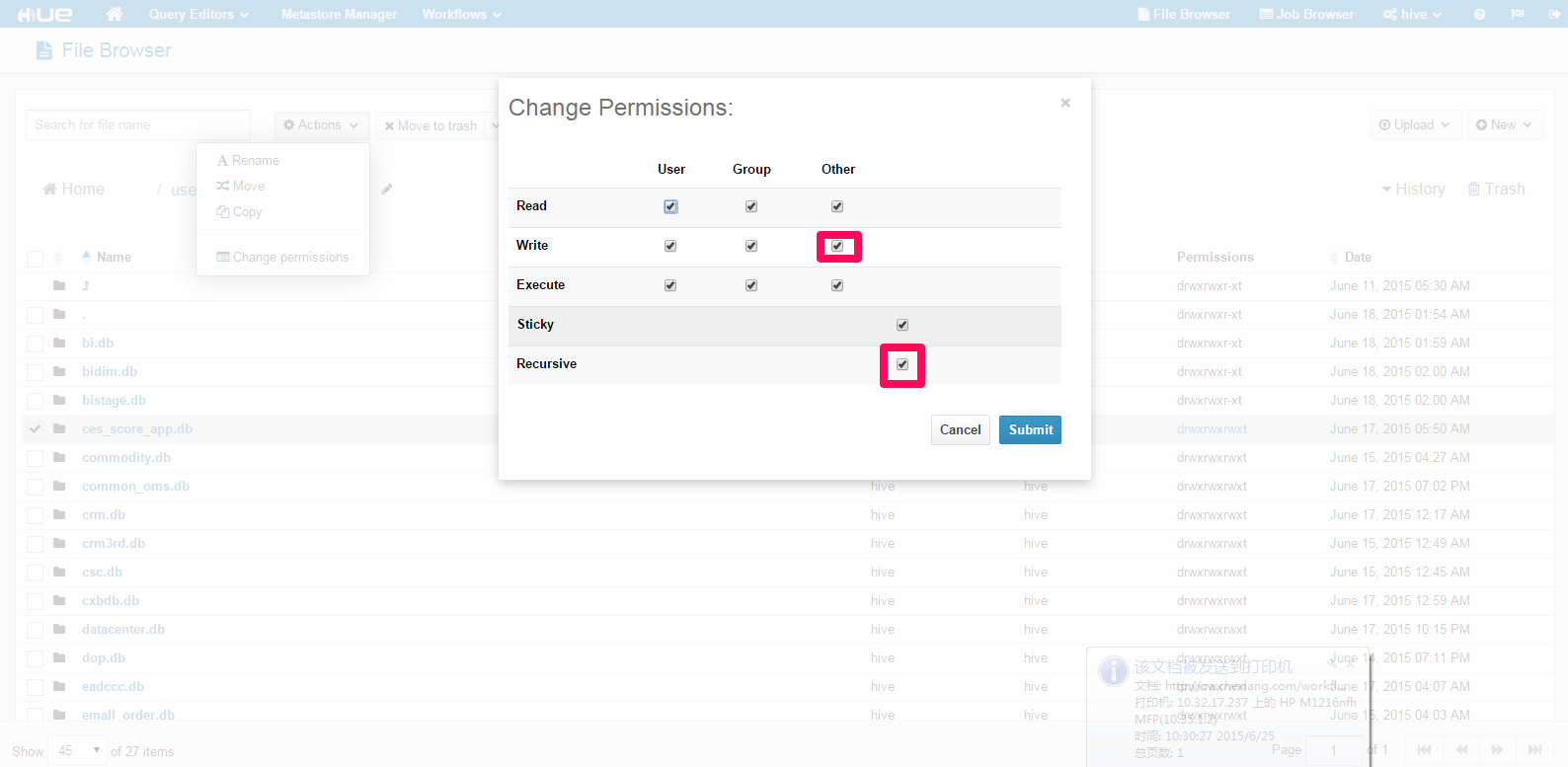
调用：python sqlencrypt.py生成对应数据库的脱敏脚本。在Hue的“Hive Editor”的Setting Tab也中添加UDFjar包：/user/bi/lib/udf\_encrypt.jar。添加UDF的Name为：md5，UDF的类名为：com.saic.udf.EncryptCommon。执行生成的脱敏脚本，将对应库的数据脱敏到bimigration库中。

## 在Hadoop测试环境中重建对应库

1. 使用Hive/Hive登录到Hadoop测试集群上。使用create dababase database\_name来创建对应的数据仓库。
2. 使用File Browser切换到对应的/user/hive/warehouse目录下。



点击“Change Permission”，进入如下页面：



勾选途中红框的部分，点击Submit按钮。

1. 使用application/application账号登录到Hadoop集群。使用命令：hadoop distcp hftp://namenode01.hadoop:50070/user/hive/warehouse/bimigration.db/\* hdfs://test-hadoop03.dds.com/user/hive/warehouse/carmall.db将hadoop prd环境脱敏数据抽取到Hadoop测试环境。
2. 使用Python命令：python table\_generate.py生成表创建语句，在Hive Editor中执行表生成语句。

到此，数据脱敏就完成了。

# 日志内容脱敏

1. 使用Hadoop命令：

hadoop distcp hftp://namenode01.hadoop:50070/user/hive/warehouse/datacenter.db/t\_accesslog/year=2015/month=06/day=2015-06-18/\* hdfs://test-hadoop02.dds.com/user/hive/warehouse/datacenter.db/t\_accesslog/year=2015/month=06/day=2015-06-18

将对应的Hadoop PRD环境上的分区文件拷贝到Hadoop测试环境对应目录。

1. 按照Hadoop PRD上的表分区情况，在Hadoop测试环境的Hive上重建对应的表结构。表语句如下：

create table datacenter.t\_accesslog (source\_type string,server\_addr string,remote\_addr string,time\_local string,msec string,url string,request string,status string,connection string,request\_time string,body\_bytes\_sent string,http\_referer string,http\_user\_agent string,proxy\_forwarded\_for string,cookie string,city\_id string,ngx\_uid string)

partitioned by (year string, month string, day string)

stored as

inputformat 'com.hadoop.mapred.DeprecatedLzoTextInputFormat'

outputformat 'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat';

1. 给表增加新的分区

ALTER TABLE DATACENTER.t\_accesslog ADD PARTITION (YEAR = '2015', MONTH = '06' , DAY = '2015-06-18')