工作原理等内容见<http://www.itnose.net/detail/6504530.html>

通过http restful api操作elasticsearch搜索数据：

<http://blog.csdn.net/stark_summer/article/details/48830493>

lucene原理等内容：

<http://www.cnblogs.com/forfuture1978/category/300665.html>

主要讲下es与hive的集成问题：

环境：JDK：1.8.0\_91(1.7或者1.8都可以，建议环境安装1.8)、hadoop：2.7.3、hive2.1.1（hadoop与hive环境没注意，当前也是支持的，版本不是太低基本集成没有问题）。获取elasticsearch-hadoop的jar包（注意版本），将此jar包拷贝到hive的默认lib目录。然后在hive的配置文件site.sh中配置以下内容：

<property>

<name>hive.aux.jars.path</name> <value>file:///srv/hive-2.1.1/lib/elasticsearch-hadoop-5.1.1.jar</value> <description>A comma separated list (with no spaces) of the jar files</description> </property>

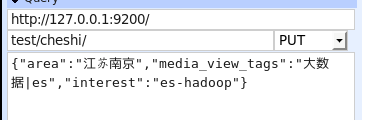
因为5.1.1兼容过去版本，支持es2.x，所以选用5.1.1的jar包

后启动相应环境，按照相应的方法进行建表进行hive与es的集成。

参考：<http://www.tuicool.com/articles/iQBBFr2>

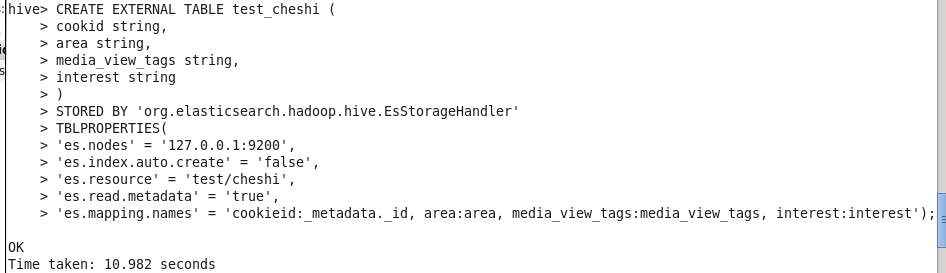
hive读取es中数据实例：

**在es中导入相应数据：**

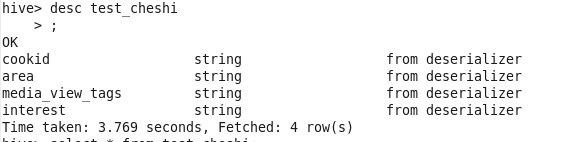


C:\Users\Test11\AppData\Local\YNote\data\qq6E8AD8E4A94E2028F37A9AB748A114CA\a4ebf3570c814f7f8f518fbc47496eb2\clipboard.png

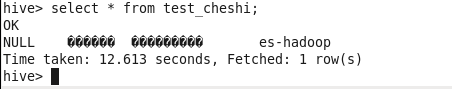
启动相应的es、hadoop、hive等内容后，**在hive中建表：**



**查询表结构：**



**执行查询语句：**



乱码的原因是：

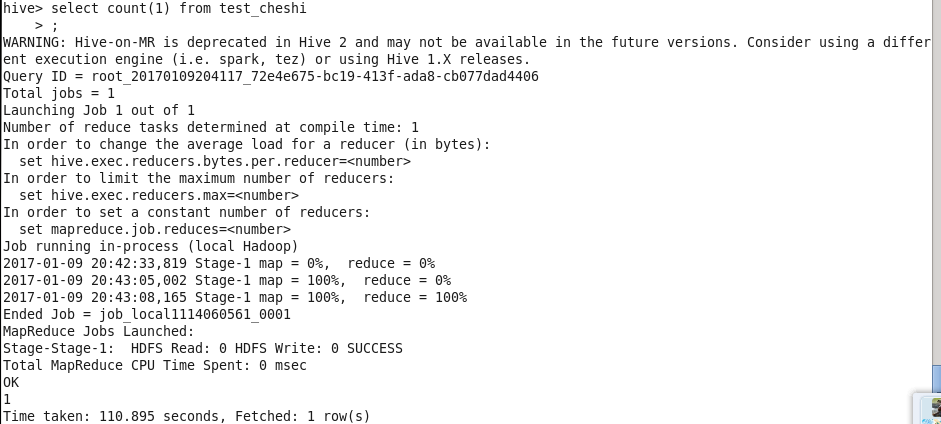
hive在将数据写入hdfs时候，会把数据格式转换为utf-8格式的。如果你导入hive表的源数据不是utf-8格式的，hive在进行写hdfs转换格式的时候会出现乱码，所有你查询出来的中文也是乱码。

　　解决办法如下：

　　1、把源文件，用editplus等编辑软件打开，将文件转换为urf-8格式，保存。再重新导入到hive表中，问题解决。

　　2、源文件只要是文本格式，如csv，txt，log等文本格式，均可用此种方法转化。前提是你终端也要设置为utf-8格式。

执行SELECT COUNT(1) FROM test\_cheshi;Hive还是通过MapReduce来执行，每个分片使用一个Map任务：



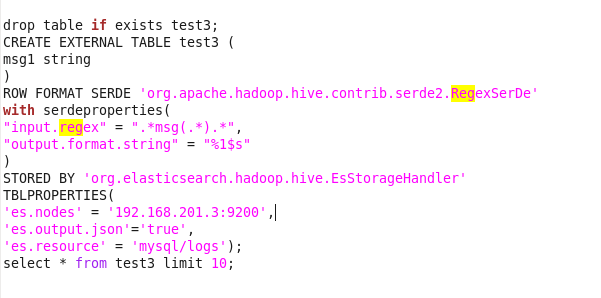
耗时间久

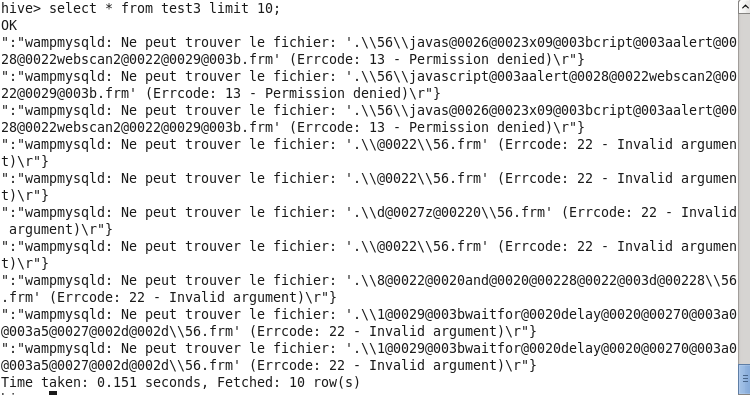
在上述建表过程中如果加入'es.query' = '?q=\_id:1' 这个内容的话在查询的时候只会返回该id对应的文档内容。（但是因数据设置只有一条，所以就不做展示了）。

删除表：

hive> DROP TABLE IF EXISTS employee; 为删除employee表

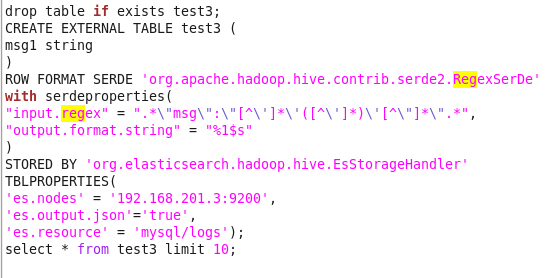
在建表时截取：



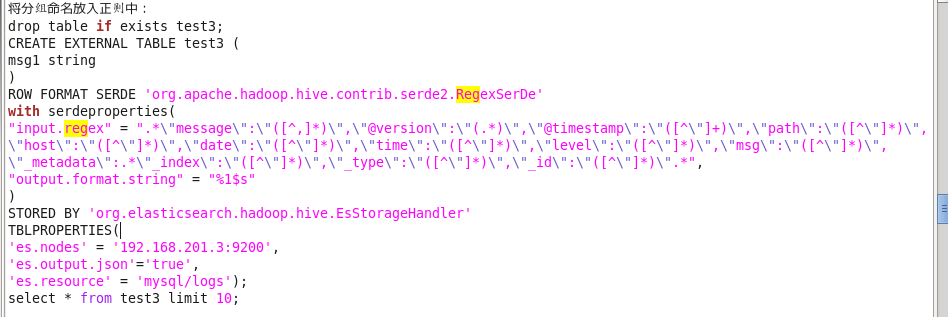


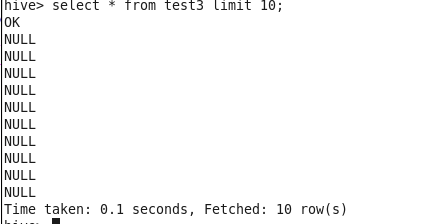
截取msg里的path字段：

input.regex里的正则要匹配输入的全部内容，表的具体内容见输出定义

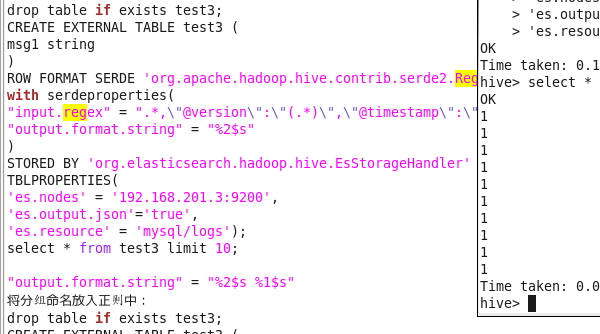


复杂正则无法识别：



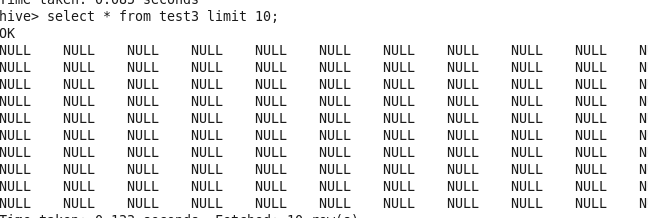


**原因：并不是无法识别，而是设置了多少分组，在上面就要定义好多少字段！如果字段内容少定义的话，查询的时候只会显示前面定义好的几个字段内容。当然，上面这个正则没有识别的原因在于分组过多。**



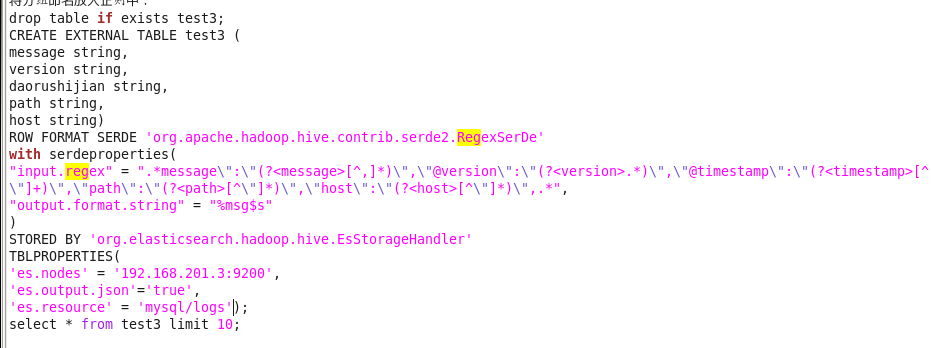
**且数据截取全部分组时候：**





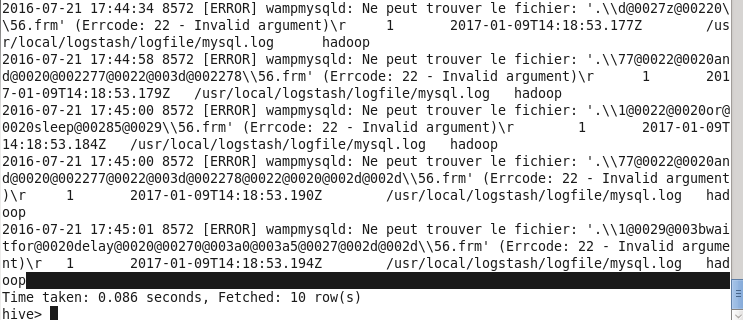
**未返回结果**

**设置分组为前5个时候：**



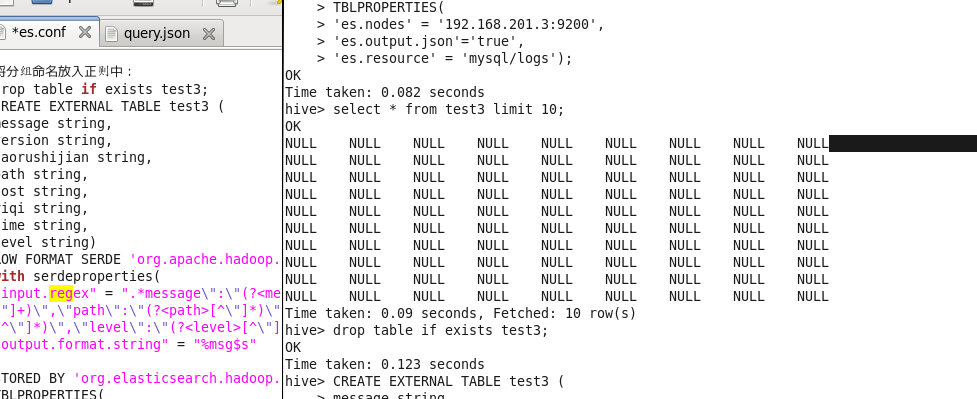
**成功。**

**这个output.format.string要不要对结果都没影响。**



**设置分组7个的时候出现问题，原因是将字段名设置为date与hive中某个内容冲突，将其字段名字修改之后成功**

**字段加到第9个msg时出现问题：**



**换成其他内容也是查询为空，说明分组最多只支持8个，目前来看应该是这种情况。**

**将正则换成最简单的分组，最多支持9个分组，再多读取就是null**

