# Elasticsearch-jdbc说明

## 一、elasticsearch-jdbc简介

elasticsearch-jdbc是一个将关系型数据库（RDBMS）数据导入到ElasticSearch库中的一个工具包，支持mysql、oracle、postgrey、csv等存储列式数据的容器。目前最新的版本是2.3.4.1，支持的ElasticSearch的版本为2.3.4。

## 二、elasticsearch-jdbc的简单使用

1. 下载zip包，

wget [http://xbib.org/repository/org/xbib/elasticsearch/importer/elasticsearch-jdbc/<version>/elasticsearch-jdbc-<version>-dist.zip](http://xbib.org/repository/org/xbib/elasticsearch/importer/elasticsearch-jdbc/%3cversion%3e/elasticsearch-jdbc-%3cversion%3e-dist.zip)，将其中的<version换成最新的版本号即可>。

1. 解压

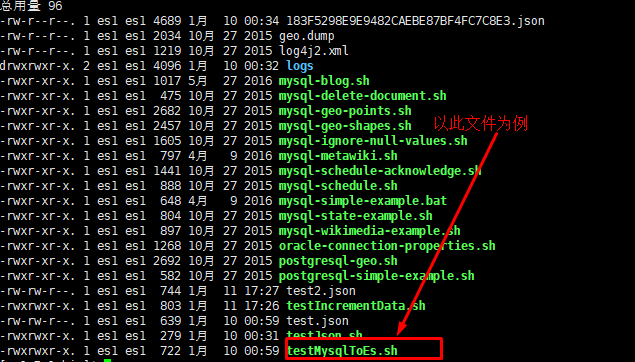
unzip elasticsearch-jdbc-<version>-dist.zip

1. 进入到elasticsearch-jdbc的目录

cd elasticsearch-jdbc-<version>

注：elasticsearch-jdbc默认带了mysql的driver，如果要添加其他数据库的driver，只需将对应的driver加到elasticsearch-jdbc-<version>/lib目录下即可

1. 切换到elasticsearch-jdbc-<version>/bin目录下，在此目录下编写你的运行脚本（不一定非要在bin目录下）



testMysqlToEs.sh脚本的内容如下：

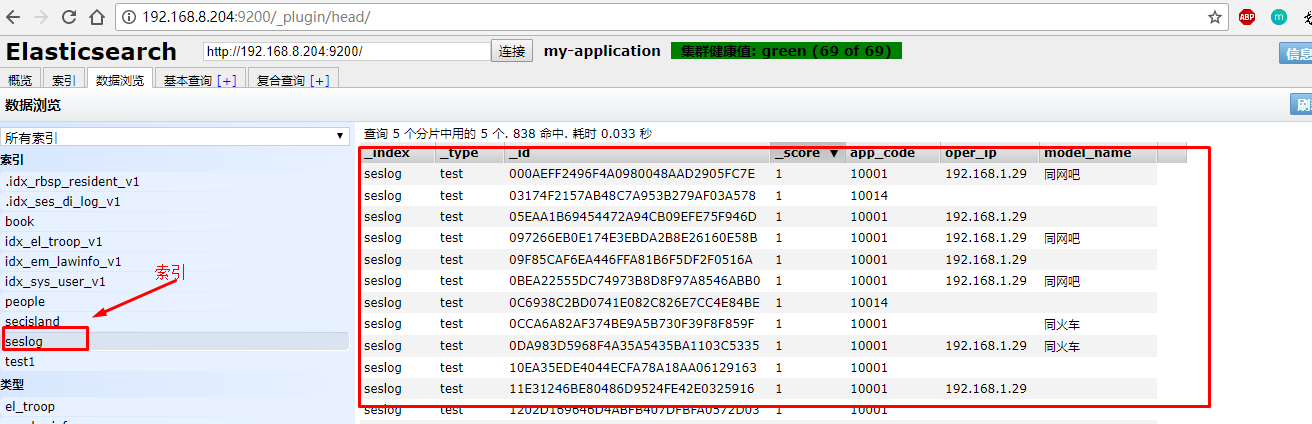
|  |
| --- |
| #!/bin/sh  bin=/home/es1/softs/elasticsearch-jdbc-2.3.4.0/bin  lib=/home/es1/softs/elasticsearch-jdbc-2.3.4.0/lib  echo '{  "type" : "jdbc",  "jdbc" : {  "url" : "jdbc:mysql://192.168.8.201:3306/ses\_v1.1",  "user" : "root",  "password" : "root",  "sql" : "select app\_code,oper\_ip,model\_name, id as \_id from ses\_api\_log",  "index": "seslog",  "type":"test",  "statefile":"test.json",  "ignor\_null\_values":true,  "elasticsearch": {  "cluster": "my-application",  "host":"192.168.8.204",  "port": 9300,  "autodiscover": true  }  }  }' | java \  -cp "${lib}/\*" \  -Dlog4j.configurationFile=${bin}/log4j2.xml \  org.xbib.tools.Runner \  org.xbib.tools.JDBCImporter |

1. 运行测试脚本文件

sh testMysqlToEs.sh

注:在bin目录下会生成一个logs文件，这个目录下的jdbc.log文件记录的是本次运行的信息。

1. 到ES集群中查看导入的数据情况



注：详细的内容可以查看https://github.com/lissic/elasticsearch-jdbc

## 三、elasticsearch-jdbc中常用相关参数说明（以上边shell脚本中的参数为例）

1. 这里列出的是jdbc中的参数

|  |  |
| --- | --- |
| 参数名 | 说明 |
| strategy | jdbc导入的策略，当前实现了“standard”“column”两种方式 |
| url | 连接数据库的url |
| user | 连接数据库的用户名 |
| password | 连接数据库的密码 |
| sql | 要执行的sql语句，这里sql参数还有很多的子参数，包括statement、write、callable、parameter等等，还包括一些内置的变量，如$now,$state,$metrics.counter, $metrics.lastexecutionstart, $metrics.lastexecutionendd等等 |

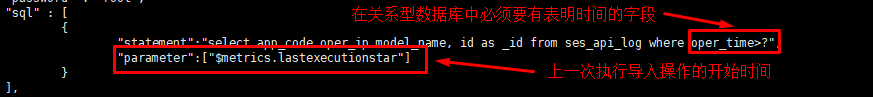
注：详细的参数可以参考：https://github.com/lissic/elasticsearch-jdbc

1. type：定义连接的类型，默认就是这个类型，也许在后边的版本中还支持其他的类型
2. index：定义导入到ES库中的索引名
3. type：定义索引中的type名称
4. statefile：定义结果输出文件名
5. ignore\_null\_true：是否忽略关系型数据库中的空值
6. elasticsearh：定义连接的es集群信息，其中cluster说明集群的名称，host说明集群中master节点的主机，port说明通信端口，一般为9300，autodiscover说明时候自动发现集群中的其他节点。

其他详细的参数可以参考: https://github.com/lissic/elasticsearch-jdbc

## 四、elasticsearch-jdbc如何同步增量数据

这里需要在脚本中的sql脚本中配置参数，如下：



Elasticsearch-jdbc是基于时间进行增量导入的，所以在关系型数据库中的表结构要有标明时间的字段，elasticsearch-jdbc会根据自身执行脚本的时间，比如执行开始时间，执行结束时间等时间点来划分那些是新数据，那些是老数据。

## 五、elasticsearch-jdbc如何进行定时任务

Elasticsearch-jdbc是基于Quartz调度器进行任务调度的，遵循cron时间调度表达式，详细可参考<https://github.com/lissic/elasticsearch-jdbc> Time scheduled execution部分。

通过在脚本中设置“schedule”参数进行定时任务的定制。

## 六、关于神盾易搜服务的思考

首先，神盾的易搜服务是基于elasticsearch-jdbc这个工具包的，通过反编译di服务源码发现，他是通过CommandLineInterpreter的java类将要执行的脚本以参数方式传给CommandLineInterpreter这个类，然后调用run方法执行相应的脚本。基于这个思路，可以自己封装一个执行elasticsearch-jdbc执行器，并调用它的api实现关系型数据库到es库的数据导入，增量数据更新，调用存储过程，发布定时任务等操作。