

## 1. Elasticsearch 与 solr 的对比

### 1.1. 官网资料

- es: <https://www.elastic.co/guide/en/elasticsearch/reference/current/index.html>  
(从 tutorial 到详细的 api 使用)
- solr: <http://mirror.bit.edu.cn/apache/lucene/solr/ref-guide/apache-solr-ref-guide-6.6.pdf> (900 多页的文档, 十分详尽)

### 1.2. 分布式架构 & 分布式部署

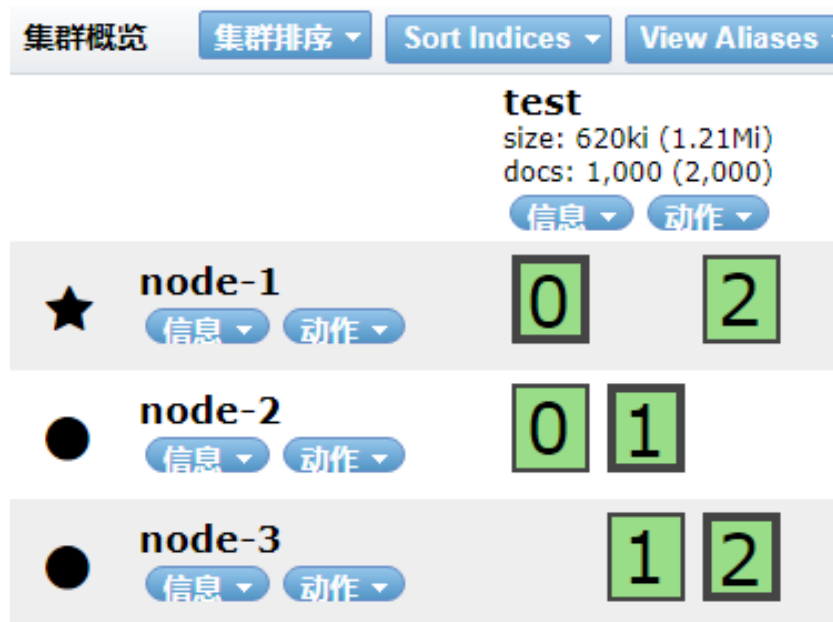
#### 1.2.1. 分布式架构

- es: 基于自身的 node 特性, 不需要接入其他服务构建集群, 自身有一个“自动发现”的特性, 可以通过一个端口检测其他节点并于其他节点通信。相关配置如下:  
discovery.zen.ping.unicast.hosts (设置集群中 master 节点的初始列表, 可以通过这些节点来自动发现新加入集群的节点)  
discovery.zen.minimum\_master\_nodes (设置这个参数来保证集群中的节点可以知道其它 N 个有 master 资格的节点, 若少于这个数, 集群其他节点也不会工作)  
支持主节点自动选举。管理 web 界面 (Kibana 的 x-pack 插件或 head 插件)
- solr: 基于 zookeeper 建立集群, 便于管理。通过 zk 自动发现集群你得节点。支持主节点自动选举 (zk 特性)。管理 web 界面 (solr 自带)

#### 1.2.2. 数据分片与备份

- es: 可以指定特定数据项的数据分片数以及数据备份数。在数据导入过程中配置。相关配置项如下:

```
“number_of_shards” : 3,  
“number_of_replicas” : 1
```



如上图所示，表示数据进行了三分片两备份。

- solr: 与 es 相同，同样可指定。相关配置项如下：  
-shards 2 -replicationFactor 2

### 1.2.3. 集群监控

- es: 通过 Http Restful Api 检测集群健康状况
- solr: 通过 JMX 可与 Prometheus 结合监控状态

## 1.3. 数据导入

### 1.3.1. 数据格式

- es: 仅支持 json 的导入，但是由第三方很多的序列化方法可供使用。
- solr: 支持多种格式，json、csv、xml

### 1.3.2. 数据从 mysql 导入，同步

- es:
  - 1、通过 elasticsearch-jdbc 插件可导入。（已测试可导入）。
  - 2、同样的，可通过这个插件进行[数据同步](#)（近实时）。原理：通过

logstash 监听数据库的某个表。可自定义检查时间间隔以及 sql 来定制需要同步的字段，实现数据库的操作 (insert、update 等) 自动同步至 es。(未测试)

### 3、支持 kettle 导入

- solr: 目前找到方法为定时与手动方法，可增量、全量同步数据。配置 db-data-config.xml，可定制导入字段，通过 web 管理界面或代码 (solrj) 实现导入。

## 1.4. 搜索语法、包装成本、Java 支持 & 中文分词支持

注：原生的 es 和 solr 对中文的分词可以说是跟没有一样，只是简单的把字分开而已，es 可以设置单字或双字模式。

- es:
  - 1、Http Restful Api
  - 2、Client: 对 Java 支持很好，通过连接串拿到 es\_client。(官方多语言支持,Java、Python、javascript...)
  - 3、如果不使用 Client 方法，专门使用 Http Restful Api es 提供了如下两种请求格式：

REST request URI, 例如:

```
GET /bank/_search?q=*&sort=account_number:asc&pretty
```

REST request body, 例如:

```
GET /bank/_search
{
  "query": { "match_all": {} },
  "sort": [
    { "account_number": "asc" }
  ]
}
```

- 4、es 有一套成熟的 DSL，可以做十分复杂的过滤、聚合、查找、分页的 query。
- 5、中文分词：elasticsearch-analysis-ik、elasticsearch-hanlp 等第三方

插件，并且插件很多，可对比效果使用

6、在 title 里面添加 zh 和 en 两个 fields，这样对中文，英文都分别使用分词器，title.zh 使用 ik 分词，title.en 使用英文分词，搜索可同时覆盖。

- solr:

- 1、Http Restful Api

- 2、Client: solrj(maven:solr-solrj)（官方只支持 Java）

- 3、conf 下的 managed-schema.xml 配置文件中配置中文分词：IK Analyzer/mmseg4j（可配词库）

- 4、有全文本搜索过滤主要还是查找功能。

Enjoy! By Hanbing.

Using madoko.