浅析Elasticsearch

|  |  |  |
| --- | --- | --- |
| 版本 | 作者 | 修订记录 |
| 1.0 | 毛仙斌 | 初版 |

# 简介

Elasticsearch(ES)是一个基于Lucene构建的开源的、分布式、RESTful接口全文搜索引擎，也是一个分布式文档数据库，文档中的每个字段都是被索引的数据并且可以被搜索。ES能够扩展至数以百计的服务器存储以及处理PB级的数据，它可以在很短的时间内存储、搜索和分析大量的数据。

# 优缺点

## 优点

* 横向扩展性好：增加一台服务器，通过设置相同的集群名称即可自动加入集群进行服务。
* 数据分片：支持数据分布式存储。
* 数据副本：支持分片数据复制，保证数据可靠性。
* 查询快：大数据量下有很好的查询速度。
* 基于Java编写，开源。

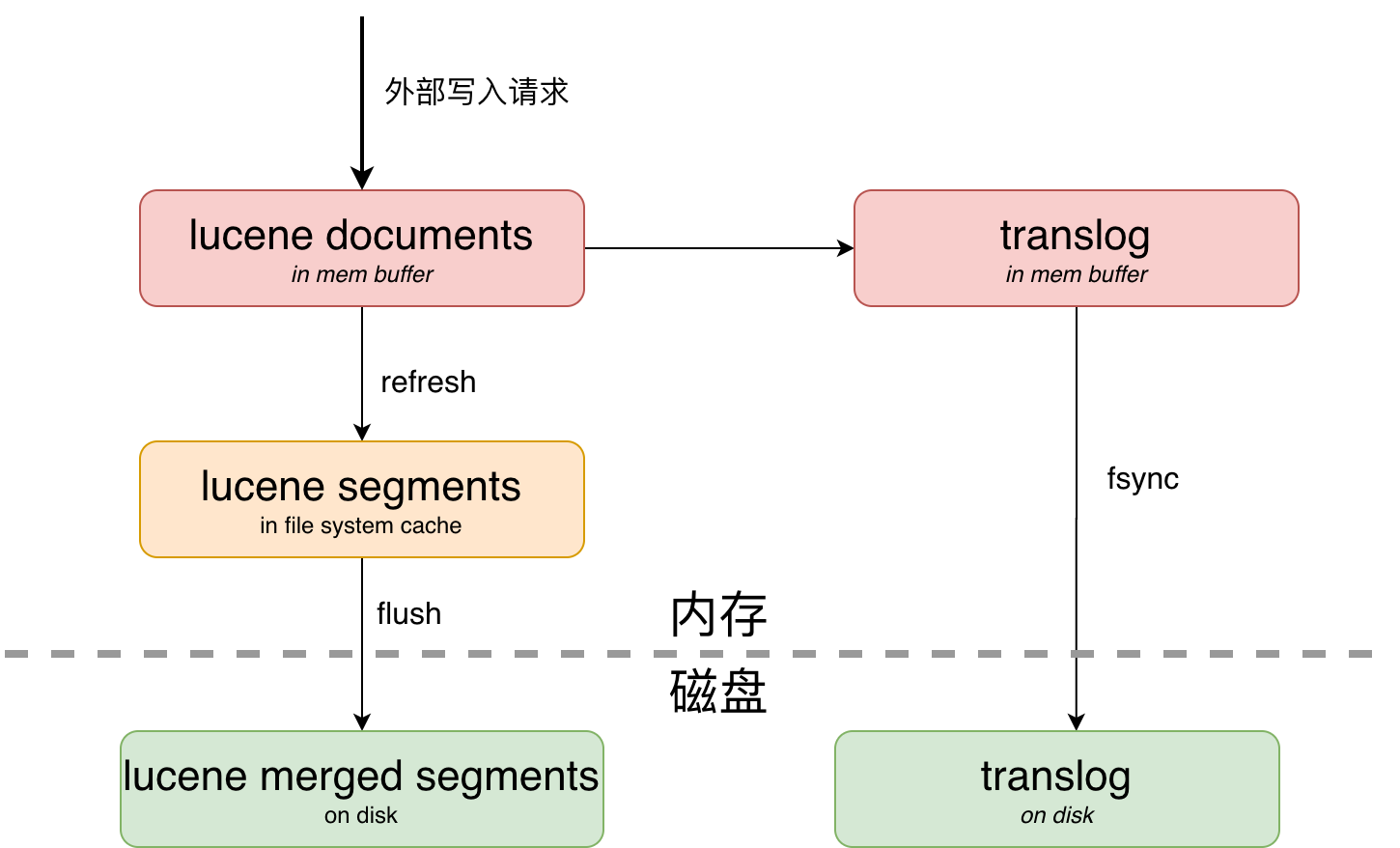
## 缺点

* 读写存在延迟：由于文档数据写入内存后需要等refresh（默认1秒）之后才能被查询到。
* 不满足ACID特性，不支持事务。
* ES自定义了CRUD语法（ES DSL），存在学习成本。
* 对关联性数据支持较差，比较考验文档模型的设计能力。
* 通常要求分配相对较高的内存（4G～32G）。

# 数据可靠性

## 数据存储

1. 文档数据写入请求到达后，首先进行Lucene documents的写入，写入成功后再写入translog。如下图：



https://niyanchun.com/es-data-reliability.html/comment-page-1

**translog**在ES中称为事务日志，它记录了每一次的索引、删除、更新或批量请求，每个分片都会存储translog。translog落盘（fsync）有两种策略，一种是对每一次请求都同步刷盘，能够保证数据可靠性，这是默认的方式。另一种是异步落盘，默认每5秒或者当translog达到512MB时刷盘一次，以吞吐量换取数据可靠性。

translog特点是追加写，多数时候能实现顺序读写，对于磁盘来讲，顺序写性能能够媲美内存的随即写，因此在没有很高吞吐量要求的场景，默认的同步写能够很好满足数据可靠性要求。

## 数据副本

在ES集群场景下，每个文档存储时ES会首先存储在主分片（primary shard）中，然后复制到不同的副本分片（replica shard）。副本分片不能和主分片部署在相同节点，当主分片节点服务宕机，副本分片节点保存了相同的数据，可以直接对外提供服务。

# 服务高可用

ES对于集群有很好的支持，集群节点的发现支持组播发现和配置文件发现，集群搭建非常方便。受益与此，ES服务的高可用性和故障转移能力得到了保障。

脑裂问题（split-brain）：ES集群中的两个或多个节点都认为自己是master，导致原有集群被分割为两个独立集群。此时两个集群接收了不同数据，将导致数据混乱或丢失。官方推荐设置最小选主数量为(N/2)+1，集群节点最少3个。此外通过适当增加discovery.zen.ping.timeout的值可以降低节点脱离的误判。

# 部署与监控

ES在CentOS中支持rpm包部署和tar包解压直接运行方式部署服务。

## 部署

### rpm

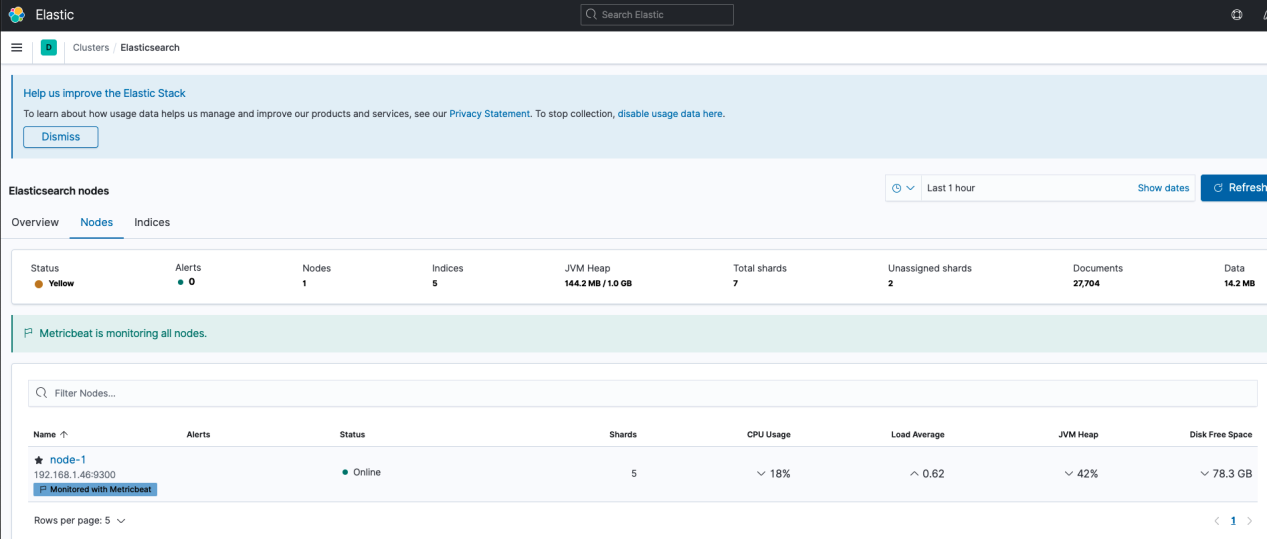
通过下载rpm包进行rpm命令安装，安装后会自动创建一个elasticsearch用户，通过systemctl命令即可起停ES服务，默认启动端口9200。

### tar包解压

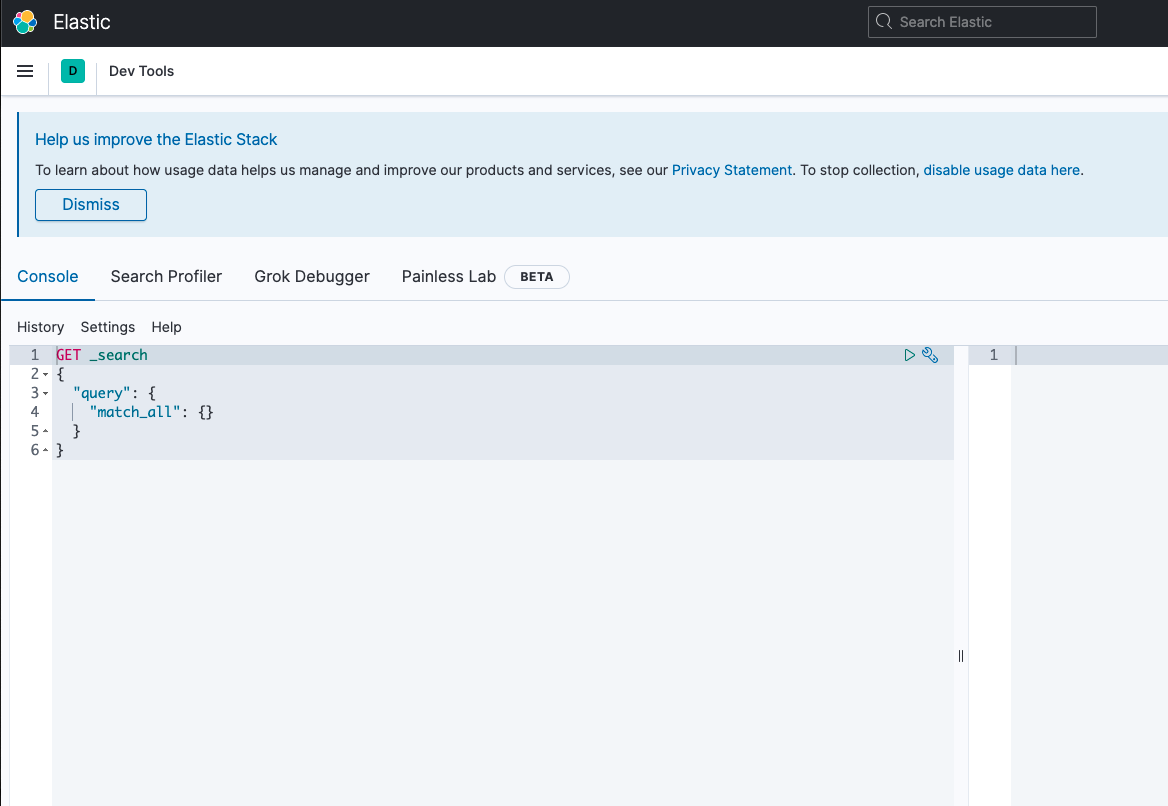
下载并解压tar.gz包后，需要手动创建一个用户（root用户不允许启动ES），并修改/etc/security/limits.conf，然后在解压目录执行bin/elasticsearch即可（可 nohup...&后台执行）。

## 监控

官方提供了Kibana服务，它可以和运行的ES服务建立连接，监控ES集群的运行健康状况，以及各项指标。如下图：



Kibana提供了Dev Tools控制台，可以在控制台中调用ES API对数据进行查询和修改：



# 适用性分析