# **1. ES 基础一网打尽**

## **1.1 ES定义**

ES=elaticsearch简写， Elasticsearch是一个开源的高扩展的分布式全文检索引擎，它可以近乎实时的存储、检索数据；本身扩展性很好，可以扩展到上百台服务器，处理PB级别的数据。   
Elasticsearch也使用Java开发并使用Lucene作为其核心来实现所有索引和搜索的功能，但是它的目的是通过简单的RESTful API来隐藏Lucene的复杂性，从而让全文搜索变得简单。

## **1.2 Lucene与ES关系？**

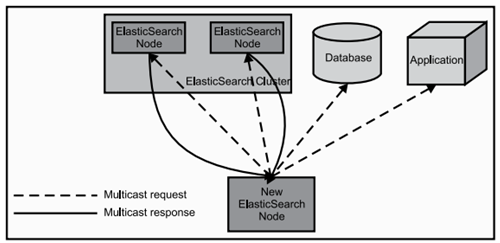
1）Lucene只是一个库。想要使用它，你必须使用Java来作为开发语言并将其直接集成到你的应用中，更糟糕的是，Lucene非常复杂，你需要深入了解检索的相关知识来理解它是如何工作的。

2）Elasticsearch也使用Java开发并使用Lucene作为其核心来实现所有索引和搜索的功能，但是它的目的是通过简单的RESTful API来隐藏Lucene的复杂性，从而让全文搜索变得简单。

## **1.3 ES主要解决问题：**

1）检索相关数据；   
2）返回统计结果；   
3）速度要快。

## **1.4 ES工作原理**

当ElasticSearch的节点启动后，它会利用多播(multicast)(或者单播，如果用户更改了配置)寻找集群中的其它节点，并与之建立连接。这个过程如下图所示：   


## **1.5 ES核心概念**

### **1）Cluster：集群。**

ES可以作为一个独立的单个搜索服务器。不过，为了处理大型数据集，实现容错和高可用性，ES可以运行在许多互相合作的服务器上。这些服务器的集合称为集群。

### **2）Node：节点。**

形成集群的每个服务器称为节点。

### **3）Shard：分片。**

当有大量的文档时，由于内存的限制、磁盘处理能力不足、无法足够快的响应客户端的请求等，一个节点可能不够。这种情况下，数据可以分为较小的分片。每个分片放到不同的服务器上。   
当你查询的索引分布在多个分片上时，ES会把查询发送给每个相关的分片，并将结果组合在一起，而应用程序并不知道分片的存在。即：这个过程对用户来说是透明的。

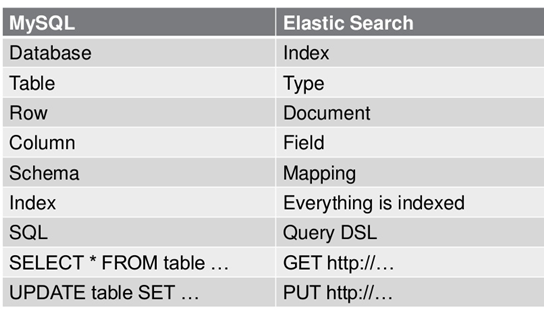
### **4）Replia：副本。**

为提高查询吞吐量或实现高可用性，可以使用分片副本。   
副本是一个分片的精确复制，每个分片可以有零个或多个副本。ES中可以有许多相同的分片，其中之一被选择更改索引操作，这种特殊的分片称为主分片。   
当主分片丢失时，如：该分片所在的数据不可用时，集群将副本提升为新的主分片。

### **5）全文检索。**

全文检索就是对一篇文章进行索引，可以根据关键字搜索，类似于mysql里的like语句。   
全文索引就是把内容根据词的意义进行分词，然后分别创建索引，例如”你们的激情是因为什么事情来的” 可能会被分词成：“你们“，”激情“，“什么事情“，”来“ 等token，这样当你搜索“你们” 或者 “激情” 都会把这句搜出来。

## **1.6 ES数据架构的主要概念（与关系数据库Mysql对比）**

   
（1）关系型数据库中的数据库（DataBase），等价于ES中的索引（Index）   
（2）一个数据库下面有N张表（Table），等价于1个索引Index下面有N多类型（Type），   
（3）一个数据库表（Table）下的数据由多行（ROW）多列（column，属性）组成，等价于1个Type由多个文档（Document）和多Field组成。   
（4）在一个关系型数据库里面，schema定义了表、每个表的字段，还有表和字段之间的关系。 与之对应的，在ES中：Mapping定义索引下的Type的字段处理规则，即索引如何建立、索引类型、是否保存原始索引JSON文档、是否压缩原始JSON文档、是否需要分词处理、如何进行分词处理等。   
（5）在数据库中的增insert、删delete、改update、查search操作等价于ES中的增PUT/POST、删Delete、改\_update、查GET.

## **1.7 ELK是什么？**

ELK=elasticsearch+Logstash+kibana   
elasticsearch：后台分布式存储以及全文检索   
logstash: 日志加工、“搬运工”   
kibana：数据可视化展示。   
ELK架构为数据分布式存储、可视化查询和日志解析创建了一个功能强大的管理链。 三者相互配合，取长补短，共同完成分布式大数据处理工作。

# **2. ES特点和优势**

1）分布式实时文件存储，可将每一个字段存入索引，使其可以被检索到。   
2）实时分析的分布式搜索引擎。   
分布式：索引分拆成多个分片，每个分片可有零个或多个副本。集群中的每个数据节点都可承载一个或多个分片，并且协调和处理各种操作；   
负载再平衡和路由在大多数情况下自动完成。   
3）可以扩展到上百台服务器，处理PB级别的结构化或非结构化数据。也可以运行在单台PC上（已测试）   
4）支持插件机制，分词插件、同步插件、Hadoop插件、可视化插件等。

# **3、ES性能**

## **3.1 性能结果展示**

（1）硬件配置：   
CPU 16核 AuthenticAMD   
内存 总量：32GB   
硬盘 总量：500GB 非SSD

（2）在上述硬件指标的基础上测试性能如下：   
1）平均索引吞吐量： 12307docs/s（每个文档大小：40B/docs）   
2）平均CPU使用率： 887.7%（16核，平均每核：55.48%）   
3）构建索引大小： 3.30111 GB   
4）总写入量： 20.2123 GB   
5）测试总耗时： 28m 54s.