第三版：Elasticsearch 50 道

简要介绍一下 Elasticsearch?

Elasticsearch 是一个分布式、RESTful 风格的搜索和数据分析引擎，能够解决不断涌现出的各种用例。作为 Elastic Stack 的核心，它集中存储您的数据，帮助您发现意料之中以及意料之外的情况。

ElasticSearch 是基于 Lucene 的搜索服务器。它提供了一个分布式多用户能力的全文搜索引擎，基于 RESTful web 接口。Elasticsearch 是用 Java 开发的，并作为 Apache 许可条款下的开放源码，是当前流行的企业级搜索引擎。

核心特点如下：

1、 分布式的实时文件存储，每个字段都被索引且可用于搜索。

2、 分布式的实时分析搜索引擎，海量数据下近实时秒级响应。

3、 简单的 restful api，天生的兼容多语言开发。

4、 易扩展，处理 PB 级结构化或非结构化数据。

您能否说明当前可下载的稳定 Elasticsearch 版本?

Elasticsearch 当前最新版本是 7.10（2020 年 11 月 21 日）。

为什么问这个问题？ES 更新太快了，应聘者如果了解并使用最新版本，基本能说明他关注 ES 更新。甚至从更广维度讲，他关注技术的迭代和更新。

但，不信你可以问问，很多求职者只知道用了 ES，什么版本一概不知。

安装 Elasticsearch 需要依赖什么组件吗?

ES 早期版本需要 JDK，在 7.X 版本后已经集成了 JDK，已无需第三方依赖。

您能否分步介绍如何启动 Elasticsearch 服务器?

启动方式有很多种，一般 bin 路径下

```
./elasticsearch -d
```

就可以后台启动。

打开浏览器输入 http://ES IP:9200 就能知道集群是否启动成功。

如果启动报错，日志里会有详细信息，逐条核对解决就可以。

能列出 10 个使用 Elasticsearch 作为其搜索引擎或数据库的公司吗?

这个问题，铭毅本来想删掉。但仔细一想，至少能看出求职者的视野够不够开阔。

参与过 Elastic 中文社区活动或者经常关注社区动态的就知道，公司太多了，列举如下（排名不分先后）：

1、阿里

2、腾讯

3、百度

4、京东

5、美团

6、小米

7、滴滴

8、携程

9、 今日头条

10、 贝壳找房

11、 360

12、 IBM

13、 顺丰快递

几乎我们能想到的互联网公司都在使用 Elasticsearch。

关注 TOP 互联网公司的相关技术的动态和技术博客，也是一种非常好的学习方式。

解释一下 Elasticsearch Cluster?

Elasticsearch 集群是一组连接在一起的一个或多个 Elasticsearch 节点实例。

Elasticsearch 集群的功能在于在集群中的所有节点之间分配任务，进行搜索和建立索引。

解释一下 Elasticsearch Node?

节点是 Elasticsearch 的实例。实际业务中，我们会说：ES 集群包含 3 个节点、7 个节点。

这里节点实际就是：一个独立的 Elasticsearch 进程，一般将一个节点部署到一台独立的服务器或者虚拟机、容器中。

不同节点根据角色不同，可以划分为：

主节点

帮助配置和管理在整个集群中添加和删除节点。

数据节点

存储数据并执行诸如 CRUD（创建/读取/更新/删除）操作，对数据进行搜索和聚合的操作。

1、 客户端节点（或者说：协调节点） 将集群请求转发到主节点，将与数据相关的请求转发到数据节点

2、 摄取节点

用于在索引之前对文档进行预处理。

解释一下 Elasticsearch 集群中的 索引的概念 ？

Elasticsearch 集群可以包含多个索引，与关系数据库相比，它们相当于数据库表

解释一下 Elasticsearch 集群中的 Type 的概念 ？

5、X 以及之前的 2.X、1.X 版本 ES 支持一个索引多个 type 的，举例 ES 6.X 中的 Join 类型在早期版本实际是多 Type 实现的。

在 6、0.0 或 更高版本中创建的索引只能包含一个 Mapping 类型。

Type 将在 Elasticsearch 7.0.0 中的 API 中弃用，并在 8.0.0 中完全删除。

你能否在 Elasticsearch 中定义映射?

映射是定义文档及其包含的字段的存储和索引方式的过程。

例如，使用映射定义：

1、 哪些字符串字段应该定义为 text 类型。

2、 哪些字段应该定义为：数字，日期或地理位置 类型。

3、 自定义规则来控制动态添加字段的类型。

Elasticsearch 的 文档是什么?

文档是存储在 Elasticsearch 中的 JSON 文档。它等效于关系数据库表中的一行记录。

解释一下 Elasticsearch 的 分片?

当文档数量增加，硬盘容量和处理能力不足时，对客户端请求的响应将延迟。

在这种情况下，将索引数据分成小块的过程称为分片，可改善数据搜索结果的获取。

定义副本、创建副本的好处是什么?

副本是 分片的对应副本，用在极端负载条件下提高查询吞吐量或实现高可用性。

所谓高可用主要指：如果某主分片 1 出了问题，对应的副本分片 1 会提升为主分片，保证集群的高可用。

请解释在 Elasticsearch 集群中添加或创建索引的过程?

要添加新索引，应使用创建索引 API 选项。创建索引所需的参数是索引的配置 Settings，索引中的字段 Mapping 以及索引别名 Alias。

也可以通过模板 Template 创建索引。

在 Elasticsearch 中删除索引的语法是什么?

可以使用以下语法删除现有索引：

```
DELETE <index_name>
```

支持通配符删除：

```
DELETE my_*
```

在 Elasticsearch 中列出集群的所有索引的语法是什么?

```
GET _cat/indices
```

在索引中更新 Mapping 的语法?

```
PUT test_001/_mapping
{
  "properties": {
    "title":{
      "type":"keyword"
    }
  }
}
```

在 Elasticsearch 中 按 ID 检索文档的语法是什么?

```
GET test_001/_doc/1
```

19、解释 Elasticsearch 中的相关性和得分?

当你在互联网上搜索有关 Apple 的信息时。它可以显示有关水果或苹果公司名称的搜索结果。

1. 你可能要在线购买水果，检查水果中的食谱或食用水果，苹果对健康的好处。
2. 你也可能要检查 Apple.com，以查找该公司提供的最新产品范围，检查评估公司的股价以及最近 6 个月，1 或 5 年内该公司在纳斯达克的表现。

同样，当我们从 Elasticsearch 中搜索文档（记录）时，你会对获取所需的相关信息感兴趣。基于相关性，通过 Lucene 评分算法计算获得相关信息的概率。

ES 会将相关的内容都返回给你，只是：计算得出的评分高的排在前面，评分低的排在后面。

计算评分相关的两个核心因素是：词频和逆向文档频率（文档的稀缺性）。

大体可以解释为：单篇文档词频越高、得分越高；多篇文档某词越稀缺，得分越高。

我们可以在 Elasticsearch 中执行搜索的各种可能方式有哪些?

核心方式如下：

方式一：基于 DSL 检索（最常用） Elasticsearch 提供基于 JSON 的完整查询 DSL 来定义查询。

```
GET /shirts/_search
{
  "query": {
```

```
    "bool": {
      "filter": [
        { "term": { "color": "red"    }},
        { "term": { "brand": "gucci" }}
      ]
    }
  }
}
```

方式二：基于 URL 检索

```
GET /my_index/_search?q=user:seina
```

方式三：类 SQL 检索

```
POST /_sql?format=txt
{
  "query": "SELECT * FROM uint-2020-08-17 ORDER BY itemid DESC
LIMIT 5"
}
```

功能还不完备，不推荐使用。

Elasticsearch 支持哪些类型的查询?

查询主要分为两种类型：精确匹配、全文检索匹配。

1. 精确匹配，例如 term、exists、term set、 range、prefix、 ids、 wildcard、regexp、 fuzzy 等。

2. 全文检索，例如 match、match_phrase、multi_match、match_phrase_prefix、query_string 等

精准匹配检索和全文检索匹配检索的不同?

两者的本质区别：

精确匹配用于：是否完全一致？

举例：邮编、身份证号的匹配往往是精准匹配。

全文检索用于：是否相关？

举例：类似 B 站搜索特定关键词如“马保国 视频”往往是模糊匹配，相关的都返回就可以。

请解释一下 Elasticsearch 中聚合?

聚合有助于从搜索中使用的查询中收集数据，聚合为各种统计指标，便于统计信息或做其他分析。聚合可帮助回答以下问题：

1、 我的网站平均加载时间是多少？

2、 根据交易量，谁是我最有价值的客户？

3、 什么会被视为我网络上的大文件？

4、 每个产品类别中有多少个产品？

聚合的分三类：

主要查看 7.10 的官方文档，早期是 4 个分类，别大意啊！

分桶 Bucket 聚合

根据字段值，范围或其他条件将文档分组为桶（也称为箱）。

指标 Metric 聚合

从字段值计算指标（例如总和或平均值）的指标聚合。

管道 Pipeline 聚合

子聚合，从其他聚合（而不是文档或字段）获取输入。

你能告诉我 Elasticsearch 中的数据存储功能吗?

Elasticsearch 是一个搜索引擎，输入写入 ES 的过程就是索引化的过程，数据按照既定的 Mapping 序列化为 Json 文档实现存储。

什么是 Elasticsearch Analyzer?

分析器用于文本分析，它可以是内置分析器也可以是自定义分析器。

你可以列出 Elasticsearch 各种类型的分析器吗?

Elasticsearch Analyzer 的类型为内置分析器和自定义分析器。

Standard Analyzer

标准分析器是默认分词器，如果未指定，则使用该分词器。

它基于 Unicode 文本分割算法，适用于大多数语言。

Whitespace Analyzer

基于空格字符切词。

Stop Analyzer

在 simple Analyzer 的基础上，移除停用词。

Keyword Analyzer

不切词，将输入的整个串一起返回。

自定义分词器的模板

自定义分词器的在 Mapping 的 Setting 部分设置：

```
PUT my_custom_index
{
 "settings":{
  "analysis":{
  "char_filter":{},
  "tokenizer":{},
  "filter":{},
```

```
  "analyzer":{}
  }
 }
}
```

脑海中还是上面的三部分组成的图示。其中：

“char_filter”:{},——对应字符过滤部分；

“tokenizer”:{},——对应文本切分为分词部分；

“filter”:{},——对应分词后再过滤部分；

“analyzer”:{}——对应分词器组成部分，其中会包含：1. 2. 3。

如何使用 Elasticsearch Tokenizer?

Tokenizer 接收字符流（如果包含了字符过滤，则接收过滤后的字符流；否则，接收原始字符流），将其分词。同时记录分词后的顺序或位置(position)，以及开始值（start_offset）和偏移值(end_offset-start_offset)。

token filter 过滤器 在 Elasticsearch 中如何工作?

针对 tokenizers 处理后的字符流进行再加工，比如：转小写、删除（删除停用词）、新增（添加同义词）等。

Elasticsearch 中的 Ingest 节点如何工作?

ingest 节点可以看作是数据前置处理转换的节点，支持 pipeline 管道 设置，可以使用 ingest 对数据进行过滤、转换等操作，类似于 logstash 中 filter 的作用，功能相当强大。

REST API 在 Elasticsearch 方面有哪些优势?

REST API 是使用超文本传输协议的系统之间的通信，该协议以 XML 和 JSON 格式传输数据请求。

REST 协议是无状态的，并且与带有服务器和存储数据的用户界面分开，从而增强了用户界面与任何类型平台的可移植性。它还提高了可伸缩性，允许独立实现组件，因此应用程序变得更加灵活。

REST API 与平台和语言无关，只是用于数据交换的语言是 XML 或 JSON。

借助：REST API 查看集群信息或者排查问题都非常方便。

在安装 Elasticsearch 时，请说明不同的软件包及其重要性?

这个貌似没什么好说的，去官方文档下载对应操作系统安装包即可。

部分功能是收费的，如机器学习、高级别 kerberos 认证安全等选型要知悉。

Elasticsearch 支持哪些配置管理工具?

1、 Ansible

2、 Chef

3、 Puppet

4、 Salt Stack

是 DevOps 团队使用的 Elasticsearch 支持的配置工具。

您能解释一下 X-Pack for Elasticsearch 的功能和重要性吗?

X-Pack 是与 Elasticsearch 一起安装的扩展程序。

X-Pack 的各种功能包括安全性(基于角色的访问,特权/权限,角色和用户安全性),监视，报告，警报等。

可以列出 X-Pack API 吗?

付费功能只是试用过（面试时如实回答就可以）。

7.1 安全功能免费后，用 X-pack 创建 Space、角色、用户，设置 SSL 加密，并且为不同用户设置不同的密码和分配不同的权限。

其他如：机器学习、 Watcher、 Migration 等 API 用的较少。

能列举过你使用的 X-Pack 命令吗?

7.1 安全功能免费后，使用了：setup-passwords 为账号设置密码，确保集群安全。

在 Elasticsearch 中 cat API 的功能是什么?

cat API 命令提供了 Elasticsearch 集群的分析、概述和运行状况，其中包括与别名，分配，索引，节点属性等有关的信息。

这些 cat 命令使用查询字符串作为其参数，并以 JSON 文档格式返回结果信息。

Elasticsearch 中常用的 cat 命令有哪些?

面试时说几个核心的就可以，包含但不限于：

含义	命令
别名	GET _cat/aliases?v
分配相关	GET _cat/allocation
计数	GET _cat/count?v
字段数据	GET _cat/fielddata?v
运行状况	GET _cat/health?
索引相关	GET _cat/indices?v
主节点相关	GET _cat/master?v
节点属性	GET _cat/nodeattrs?v
节点	GET _cat/nodes?v
待处理任务	GET _cat/pending_tasks?v
插件	GET _cat/plugins?v
恢复	GET _cat / recovery?v
存储库	GET _cat /repositories?v

段	GET _cat /segments?v
分片	GET _cat/shards?v
快照	GET _cat/snapshots?v
任务	GET _cat/tasks?v
模板	GET _cat/templates?v
线程池	GET _cat/thread_pool?v

您能解释一下 Elasticsearch 中的 Explore API 吗?

没有用过，这是 Graph （收费功能）相关的 API。

点到为止即可，类似问题实际开发现用现查，类似问题没有什么意义。

https://www.elastic.co/guide/en/elasticsearch/reference/current/graph-explore-api.html

迁移 Migration API 如何用作 Elasticsearch?

迁移 API 简化了 X-Pack 索引从一个版本到另一个版本的升级。

点到为止即可，类似问题实际开发现用现查，类似问题没有什么意义。

https://www.elastic.co/guide/en/elasticsearch/reference/current/migration-api.html

如何在 Elasticsearch 中 搜索数据?

Search API 有助于从索引、路由参数引导的特定分片中查找检索数据。

你能否列出与 Elasticsearch 有关的主要可用字段数据类型?

1、 字符串数据类型，包括支持全文检索的 text 类型 和 精准匹配的 keyword 类型。

2、 数值数据类型，例如字节，短整数，长整数，浮点数，双精度数，half_float，scaled_float。

3、 日期类型，日期纳秒 Date nanoseconds，布尔值，二进制（Base64 编码的字符串）等。

4、 范围（整数范围 integer_range，长范围 long_range，双精度范围 double_range，浮动范围 float_range，日期范围 date_range）。

5、 包含对象的复杂数据类型，nested 、Object。

6、 GEO 地理位置相关类型。

7、 特定类型如：数组（数组中的值应具有相同的数据类型）

详细说明 ELK Stack 及其内容?

ELK Stack是一系列搜索和分析工具(Elasticsearch),收集和转换工具(Logstash)以及数据管理及可视化工具（Kibana）、解析和收集日志工具（Beats 未来是Agent）以及监视和报告工具（例如 X Pack）的集合。

相当于用户基本不再需要第三方技术栈，就能全流程、全环节搞定数据接入、存储、检索、可视化分析等全部功能。

Kibana 在 Elasticsearch 的哪些地方以及如何使用?

Kibana 是 ELK Stack 一日志分析解决方案的一部分。

它是一种开放源代码的可视化工具，可以以拖拽、自定义图表的方式直观分析数据，极大降低的数据分析的门槛。

未来会向类似：商业智能和分析软件 - Tableau 发展。

logstash 如何与 Elasticsearch 结合使用?

logstash 是 ELK Stack 附带的开源 ETL 服务器端引擎，该引擎可以收集和处理来自各种来源的数据。

最典型应用包含：同步日志、邮件数据，同步关系型数据库（MySQL、Oracle）数据，同步非关系型数据库（MongoDB）数据，同步实时数据流 Kafka 数据、同步高性能缓存 Redis 数据等。

Beats 如何与 Elasticsearch 结合使用?

Beats是一种开源工具,可以将数据直接传输到 Elasticsearch 或通过 logstash，在使用 Kibana 进行查看之前，可以对数据进行处理或过滤。

传输的数据类型包含：审核数据，日志文件，云数据，网络流量和窗口事件日志等。

如何使用 Elastic Reporting ？

收费功能，只是了解，点到为止。

Reporting API 有助于将检索结果生成 PDF 格式，图像 PNG 格式以及电子表格 CSV 格式的数据，并可根据需要进行共享或保存。

您能否列出 与 ELK 日志分析相关的应用场景?

1、 电子商务搜索解决方案

2、 欺诈识别

3、 市场情报

4、 风险管理

5、 安全分析 等。