

直播
公开课

HarmonyOS组件的
分布式适配和应用畅想

时间：2021年1月14日 20:00-22:00

51CTO

HarmonyOS技术社区

董昱

HarmonyOS系统
课程开发者

2021薛大龙软考报名入口
Linux工程师10问10答
考华为认证，做高级网工
网络安全攻防实战系统班

终于有人把Elasticsearch原理讲透了！

小史是一个非科班的程序员，虽然学的是电子专业，但是通过自己的努力成功通过了面试，现在要开始迎接新生活了。

作者：channingbreeze 来源：互联网侦察 | 2019-04-08 12:14

收藏 分享

小史是一个非科班的程序员，虽然学的是电子专业，但是通过自己的努力成功通过了面试，现在要开始迎接新生活了。

随着央视诗词大会的热播，小史开始对诗词感兴趣，最喜欢的就是飞花令的环节。



但是由于小史很久没有背过诗词了，飞一个字很难说出一句，很多之前很熟悉的诗句也想不起来。



- 编辑推荐
- 头条

中台不行了？阿里彻底拆中台了！
- 热点

2020年编程语言排行榜单年终大盘点
- 关注

如何部署Jupyter Notebook用于交互式Python开发？
- 头条

为何还在浪费时间？Python与R语言的数据科学之争
- 头条

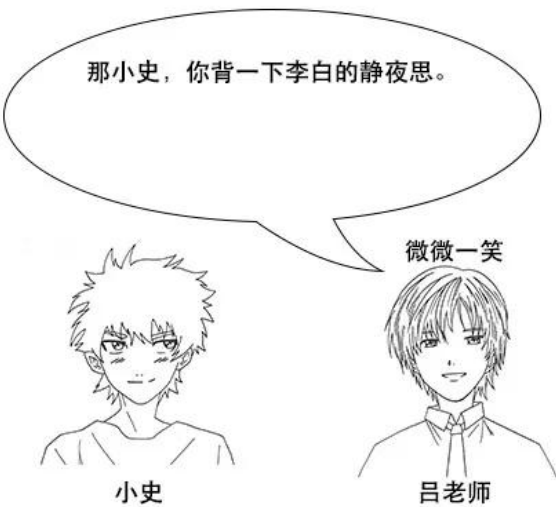
走近来自Canonical的轻量级Kubernetes发行版：Microk8s

- 24H热文 一周话题 本月获赞
- 太厉害了，终于有人能把TCP/IP协议讲的明...
坐在马桶上看算法：快速排序
花了一个星期，我终于把RPC框架整明白了！
图文详解两种算法：深度优先遍历（DFS） ...
为什么我不要求996，公司还能做到市值1万...
终于有人把Elasticsearch原理讲透了！
五大自动化测试的Python框架
用Node.js进行后端开发的特点与优势

- 订阅专栏 +更多
- 这就是5G

这就是5G
5G那些事儿
共15章 | armmay
115人订阅学习
- 16招轻松掌握PPT技巧

16招轻松掌握PPT技巧



20年
建设案例
网络搭建技巧
共20章 | 捷哥CCIE
758人订阅学习

20个局域网建设改造案例
网络搭建技巧
共20章 | 捷哥CCIE
758人订阅学习

51CTO学院

新用户免费领VIP月度会员>>

NEW FEATURES

HTML+CSS
视频教程

JAVA全科工程师

HTML+CSS视频教程

HTML

前端课程html+css

架构师成长之路
HTML+CSS+JS教程

前端html/css/html5/css3/js教程

CTO品牌

+ 更多

CTO研习体验营，19.9元入营

技术经理研习营2020年第一期招募

CTO训练营

申请入营 互联网班 体验营

技术经理

申请加入 能力地图 进化图谱

专题推荐

+ 更多

如何成为一个值得追随的技术管理者

Yii框架深度剖析
Yii框架

直播平台
直播平台在实践中遇到的挑战

Micro service
微服务实战中的那些“坑”
微服务/实战



大佬~别丢下你的粉丝，给他们建个答
Teams中如何控制创建团队权限
Proxmox虚拟机自动备份填坑记
Elasticsearch使用优化之拙见
firewalld防火墙配置IP伪装和端口转

读书 +更多



网管员必读——网络安全（第2版）

本书是在《网管员必读—网络安全》第1版的基础上修改而成的。新版在保留第1版实用内容的基础上增加了大量新的实用内容，同时删除了一些过时...



订阅51CTO邮刊

[点击这里查看样刊](#)

立即订阅







倒排索引







吕老师：但是我让你说出带“前”字的诗句，由于没有索引，你只能遍历脑海中所有诗词，当你的脑海中诗词量大的时候，就很难在短时间内得到结果了。



key

前

→

value

床前明月光
疑是地上霜

倒排索引，以“前”字
作为key，诗句内容作为
value。

微微一笑


你的意思是“前”作为索引？

瞬间理解

小史


吕老师

没错，这样的话，让你背带“前”字的诗，你是不是马上就能背出来了？


小史

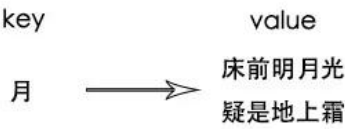
微微一笑

吕老师



索引量爆炸





key		value
床	→	床前明月光 疑是地上霜
前	→	床前明月光 疑是地上霜
明	→	床前明月光 疑是地上霜
月	→	床前明月光 疑是地上霜
光	→	床前明月光 疑是地上霜
疑	→	床前明月光 疑是地上霜
是	→	床前明月光 疑是地上霜
地	→	床前明月光 疑是地上霜
上	→	床前明月光 疑是地上霜
霜	→	床前明月光 疑是地上霜



✕





✕

key		value
床	➡	静夜思
前	➡	静夜思
明	➡	静夜思
月	➡	静夜思
光	➡	静夜思
疑	➡	静夜思
是	➡	静夜思
地	➡	静夜思
上	➡	静夜思
霜	➡	静夜思

value不存诗句改存诗
题，数据量就会减少很
多。这里，诗题可以理
解为数据正向索引。

微微一笑



哦，我明白了，这样的话，数据量确实
减少一些了。

瞬间理解



小史



吕老师



key		value
静夜思	→	床前明月光 疑是地上霜 举头望明月 低头思故乡
望庐山瀑布	→	日照香炉生紫烟 遥看瀑布挂前川 飞流直下三千尺 疑是银河落九天
月下独酌	→	花间一壶酒 独酌无相亲 举杯邀明月 对影成三人





静夜思	带“前”字
望庐山瀑布	带“前”字
静夜思	带“月”字
月下独酌	带“月”字



key	value
前	静夜思 望庐山瀑布
月	静夜思 月下独酌

“前”字可以索引到两首诗，“月”字也可以索引到两首诗。

微微一笑



哦，明白了，这样根据一个“前”字，就可以方便地找出所有带“前”字的诗了。

瞬间理解



小史



吕老师

搜索引擎原理

小史，有没有感觉你这个过程和咱们百度或者谷歌搜索类似呀？

微微一笑



小史



吕老师

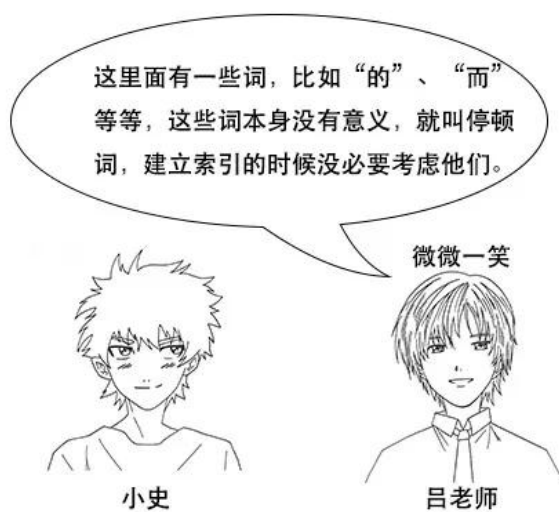






面试官的每一个问题，都是有考察点的。但是对于这样的软问题，你可以好好回答来主导方向，要体现你的深度思考，体现你在工作中如何创造价值，而不是浮于问题表面。

的、而这种没有意义的词可以认为是停顿词





1、爬取

面试官的每一个问题，都是有考察点的。但是对于这样的软问题，你可以好好回答来主导方向，要体现你的深度思考，体现你在工作中如何创造价值，而不是浮于问题表面。

2、分词

3、建立反向索引

- | | | | |
|-------|------|-----|-------|
| 面试官 | 面试官 | ——> | 面试软技巧 |
| 软问题 | 软问题 | ——> | 面试软技巧 |
| 考察 | 考察 | ——> | 面试软技巧 |
| 回答 | 回答 | ——> | 面试软技巧 |
| 深度思考 | 深度思考 | ——> | 面试软技巧 |
| 创造价值 | 创造价值 | ——> | 面试软技巧 |
| | | | |



搜索引擎三大过程，爬取内容、进行分词、建立反向索引。

嗯，是这样。



小史

微微一笑



吕老师

Elasticsearch 简介





吕老师：但是 Lucene 还是一个库，必须要懂一点搜索引擎原理的人才能用的好，所以后来又有人基于 Lucene 进行封装，写出了 Elasticsearch。



elasticsearch将对搜索引擎的操作都封装成了restful的api, 通过http请求就能对其进行操作。

微微一笑



小史



吕老师

同时, 它还考虑了海量数据, 实现了分布式, 是一个可以存储海量数据的分布式搜索引擎。

微微一笑



小史



吕老师

那它也是基于hdfs的吗?

提出疑问



小史



吕老师



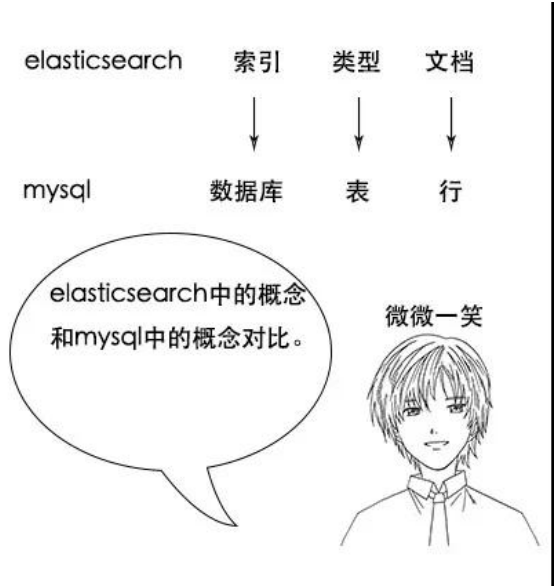
Elasticsearch 基本概念







吕老师：类型是用来定义数据结构的，你可以认为是 MySQL 中的一张表。文档就是最终的数据了，你可以认为一个文档就是一条记录。





吕老师：比如一首诗，有诗题、作者、朝代、字数、诗内容等字段，那么首先，我们可以建立一个名叫 Poems 的索引，然后创建一个名叫 Poem 的类型，类型是通过 Mapping 来定义每个字段的类型。

比如诗题、作者、朝代都是 Keyword 类型，诗内容是 Text 类型，而字数是 Integer 类型，
***就是把数据组织成 Json 格式存放进去了。

索引

poems

类型

```
"poem": {
  "properties": {
    "title": {
      "type": "keyword",
    },
    "author": {
      "type": "keyword",
    },
    "dynasty": {
      "type": "keyword"
    },
    "words": {
      "type": "integer"
    },
    "content": {
      "type": "text"
    }
  }
}
```

文档

```
{
  "title": "静夜思",
  "author": "李白",
  "dynasty": "唐",
  "words": "20",
  "content": "床前明月光，疑是地上霜。举
头望明月，低头思故乡。"
}
```

类型相当于表结构的描述，描述每个字段的类型，文档已json形式描述一行数据。

微微一笑





吕老师：这个问题问得好，这涉及到分词的问题，Keyword 类型是不会分词的，直接根据字符串内容建立反向索引，Text 类型在存入 Elasticsearch 的时候，会先分词，然后根据分词后的内容建立反向索引。





吕老师：之前我们说过，Elasticsearch 把操作都封装成了 HTTP 的 API，我们只要给 Elasticsearch 发送 HTTP 请求就行。

比如使用 `curl -XPUT 'http://ip:port/poems'`，就能建立一个名为 Poems 的索引，其他操作也是类似的。



Elasticsearch 分布式原理



吕老师：没错，Elasticsearch 也是会对数据进行切分，同时每一个分片会保存多个副本，其原因和 HDFS 是一样的，都是为了保证分布式环境下的高可用。

★ master

信息

动作

014

● slave1

信息

动作

234

● slave2

信息

动作

0123

绿色的表示数据块，其实elasticsearch中数据块也是备份存储至多个节点中的。

微微一笑

所以elasticsearch也是master-slave架构咯？

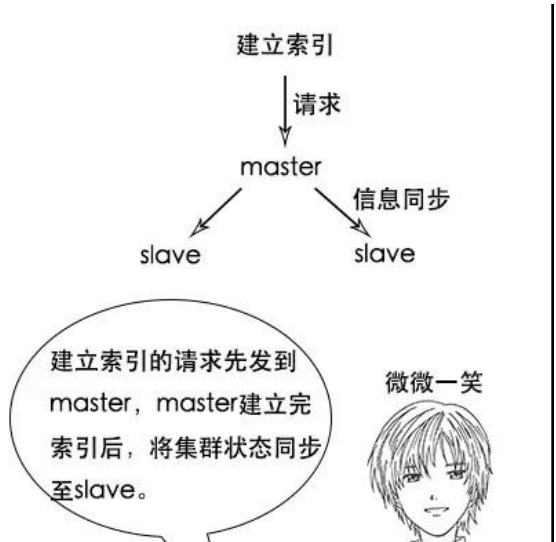
提出疑问

小史

吕老师



吕老师：没错，在 Elasticsearch 中，节点是对等的，节点间会通过自己的一些规则选取集群的 Master，Master 会负责集群状态信息的改变，并同步给其他节点。



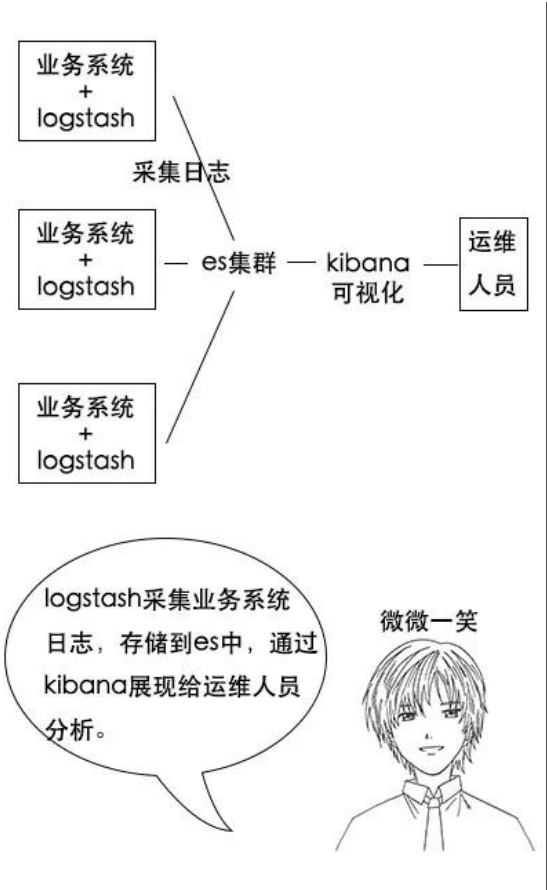




ELK 系统



吕老师：其实很多公司都用 Elasticsearch 搭建 ELK 系统，也就是日志分析系统。其中 E 就是 Elasticsearch，L 是 Logstash，是一个日志收集系统，K 是 Kibana，是一个数据可视化平台。





吕老师：分析日志的用处可大了，你想，假如一个分布式系统有 1000 台机器，系统出现故障时，我要看下日志，还得一台一台登录上去查看，是不是非常麻烦？



吕老师：但是如果日志接入了 ELK 系统就不一样。比如系统运行过程中，突然出现了异常，在日志中就能及时反馈，日志进入 ELK 系统中，我们直接在 Kibana 就能看到日志情况。如果再接入一些实时计算模块，还能做实时报警功能。



总结

小史学完了 Elasticsearch，在笔记本上写下了如下记录：

反向索引又叫倒排索引，是根据文章内容中的关键字建立索引。

搜索引擎原理就是建立反向索引。

Elasticsearch 在 Lucene 的基础上进行封装，实现了分布式搜索引擎。

Elasticsearch 中的索引、类型和文档的概念比较重要，类似于 MySQL 中的数据库、表和行。

Elasticsearch 也是 Master-slave 架构，也实现了数据的分片和备份。

Elasticsearch 一个典型应用就是 ELK 日志分析系统。

写完，又高高兴兴背诗去了。

半亩方塘一鉴开，天光云影共徘徊。

问渠那得清如许？为有源头活水来。



【编辑推荐】

- 1. 法国程序员没有“996” 40多岁最受人尊敬
- 2. 程序员的软件开发指南：零 Bug 策略
- 3. Python 之父再度发声：我们能为中国的“996” 程序员做什么？
- 4. 不要996！程序员创建955.WLB不加班公司名单，GitHub周榜第二
- 5. 程序员发量越多说明薪资越低？今天你脱发了吗？

【责任编辑：武晓燕 TEL：（010）68476606】

点赞 814

Elasticsearch

程序员

Lucene

分享:

大家都在看

猜你喜欢



三大流行数据库
(MongoDB、



实力再次印证，新华三
全面挺进核心路由器市



从塞班到鸿蒙 —— 手机
操作系统这二十年历程



2020年值得推荐的六大
服务器管理软件和工具

好课推荐



HTML+CSS视频教程



前端课程html+css



前端html/css/html5/c
ss3/js教程

51CTO旗下网站： 领先的IT技术网站 51CTO | 中国专业CIO网站 CIOage | 中国数字医疗领军网站 HC3i