大数据搜索与日志挖掘及可视化方案

高凯 等 编著

—— **ELK Stack:** Elasticsearch、Logstash、Kibana

(第2版)

清华大学出版社

大数据搜索与日志挖掘及可视化方案

高凯 等 编著

—— **ELK Stack:** Elasticsearch、Logstash、Kibana

(第2版)

清华大学出版社
北京

内 容 简 介

对大数据的搜索与挖掘，在当今的“互联网+”时代是很有必要的。本书提出的分布式大数据搜索与日志挖掘及可视化方案是基于 ELK Stack 而提出的，它能有效应对海量大数据所带来的分布式存储与处理、全文检索、日志挖掘、可视化等问题。构建在全文检索开源软件 Lucene 之上的 Elasticsearch，不仅能对海量规模的数据完成分布式索引与检索，还能提供数据聚合分析。据国际权威的数据库产品评测机构 DB-Engines 的统计，在 2016 年 1 月，Elasticsearch 已超过 Solr 等，成为排名第一的搜索引擎类应用；Logstash 能有效处理来源于各种数据源的日志信息；Kibana 能得出可视化分析结果。了解基于 ELK Stack 的大数据搜索与日志挖掘及可视化方案，掌握 Elasticsearch、Logstash、Kibana 的基本使用方法和技巧，很有必要。

全书以模块化的方式进行组织。和第一版相比，本书力求反映 ELK Stack 的最新成果，内容新颖、强调实践。本书也可为高等学校相关专业（如计算机科学与技术、软件工程、物联网、信息管理与信息系统）学生的学习和科研工作提供帮助，同时对于从事大数据搜索与挖掘、日志分析、信息可视化技术的工程技术人员和希望了解网络信息检索技术的人员也具有较高的参考价值和工程应用价值。

图书在版编目(CIP)数据

大数据搜索与日志挖掘及可视化方案：ELK Stack：Elasticsearch、Logstash、Kibana/高凯等编著. —2 版. —北京：清华大学出版社，2016

ISBN 978-7-302-43328-6

Ⅰ. ①大… Ⅱ. ①高… Ⅲ. ①情报检索－研究 ②数据采集－研究 Ⅳ. ①G354 ②TP274

中国版本图书馆 CIP 数据核字(2016)第 051612 号

责任编辑：焦　虹
封面设计：傅瑞学
责任校对：李建庄
责任印制：何　芊

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954

印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：185mm×240mm　　**印　张**：18.25　　**字　数**：411 千字
版　　次：2015 年 6 月第 1 版　2016 年 6 月第 2 版　　**印　次**：2016 年 6 月第1次印刷
印　　数：1～1500
定　　价：49.90 元

产品编号：068922-01

序言

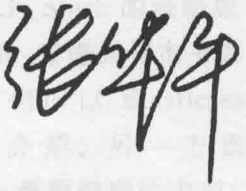
Preface

云计算、智慧城市、移动互联网、大数据与物联网已经成为大数据时代的前瞻技术，实现了人、机器与实物的多维互联互通，监测数据、内容数据、社交数据、关系数据裂变式增长，大数据时代已全方位到来。大数据具有多（体量大）、快（生成速度快）、好（价值大）、省（高效）的特征，传统的信息搜索、数据挖掘与知识呈现技术难以满足当下多样化的需求。大数据的理念与理论已经成为人所共知的科学常识，但是大数据搜索、挖掘与可视化等落地的工程实践尚有较大距离，也是当下的工程急需。

本书从分布式大数据搜索、日志挖掘与可视化三个角度出发，以非结构化文本信息、半结构化的日志数据为处理对象，进行宏观解决方案与微观方法技巧的全面阐释。具体地说，如何利用在全文检索开源软件 Lucene 之上的 Elasticsearch 对大数据进行分布式计算与全文检索；如何利用 Logstash 对日志文件进行智能分析与处理；如何利用 Web 接口 Kibana 对日志进行高效的搜索、可视化、分析等，是本书的论述重点。

从工程实践的角度掌握 ElasticSearch、Logstash、Kibana 的基本使用方法和技巧，很有必要。目前，国内专门针对 Elasticsearch、Logstash、Kibana 进行介绍的书很少。本书是目前国内较早综合介绍 ELK 架构的图书，涉及范围广泛，内容新颖，条理清晰，组织合理。

高凯老师是我多年的朋友，我们都在大数据搜索与挖掘方向上从事教学、科研与开发工作。他严谨的治学态度、理论联系实际的做法以及敬业的态度也一直为我所学习。非常荣幸能够有这个机会来为高老师的新著作序。认真拜读后，我以为本书实战性很强，是大数据搜索与挖掘所需的上乘之作，是大数据“知著、见微、晓意”的必备工具，值得推荐！



（张华平 博士，副教授，北京理工大学大数据搜索挖掘实验室主任，ICTCLAS 及 NLPIR 分词软件发明者。）

Foreword

第2版 前言

本书第1版《实战 Elasticsearch、Logstash、Kibana——分布式大数据搜索与日志挖掘及可视化解决方案》从出版发行到现在，虽过去短短的半年时间，但在这期间，伴随着《中共中央关于制定国民经济和社会发展第十三个五年规划的建议》中国家大数据战略的实施，伴随着海量数据管理技术在国民经济以及互联网+、物联网、移动计算等各个领域的广泛应用，分布式大数据搜索与日志挖掘及可视化解决方案正日益受到各行各业人员的普遍关注。开源的、基于 Lucene 的全文搜索引擎 Elasticsearch 以其独到的分布式数据处理能力，正发挥着越来越重要的作用。根据国际权威的数据库产品评测机构 DB-Engines 统计，在 2016 年 1 月，Elasticsearch 已超过 Solr 等，成为排名第一的搜索引擎类应用。

ELK Stack 是以 Elasticsearch、Logstash、Kibana 三个开源软件为主的大数据处理工具集，也是目前开源的最流行的大数据分析解决方案，它为编程人员提供了一个分布式可扩展的信息存储和全文检索机制、基于 Logstash 的日志处理机制、基于 Kibana 的挖掘结果可视化的机制。不仅如此，ELK Stack 还有 Shield（安全和管理插件，如权限控制、加密通信、审计等）、Watcher（性能监控平台等）、Beats（官方提供了用来收集日志的 Filebeat、用来收集系统基础设置数据的 Topbeat、统计收集网络信息的 Packetbeat）等中间件。在实时大数据处理的应用中，上述软件通常配合使用。因此，从实战的角度掌握 Elasticsearch、Logstash、Kibana 等软件的基本使用方法和技巧，很有必要。

考虑到部分读者对本书第 1 版的修改意见，我们对其中的部分内容进行了必要的补充和修改、完善。一方面，对 ELK Stack 的最新版本进行了简述，力求反映 ELK Stack 的最新成果；同时，考虑到与本书第 1 版的内容衔接，对部分使用上无差异的操作，仍旧以 Elasticsearch、Logstash、Kibana 的经典版本为基础进行介绍。另一方面，对 Elasticsearch 中涉及索引、检索、统计、Java 实现、集群管理的内容（主要涉及第 1 版中的第 2～6 章的内容），给出了实例。同第 1 版一样，本书第 2 版仍强调实践和面向初学者，并通过实战讲解的方式，让读者更好

地了解 ELK Stack 的应用。全书涵盖 ELK Stack 简介、文档索引与处理、信息检索与过滤、信息统计与分析、基于 Java 客户端的 Elasticsearch 功能实现、Elasticsearch 配置与管理、基于 Logstash 的网络日志处理、基于 Kibana 的分析结果可视化、应用实例等内容。本书介绍的基于 ELK Stack 架构的分布式大数据搜索与日志挖掘及可视化是入门方案，对有一定基础的中、高级使用者亦有一定的参考和工程应用价值。

全书由高凯提出写作大纲。第 1 章、第 6 章和第 7 章中的部分内容由高凯撰写，其余各章由高莘撰写，最后由高凯完成全书通稿和审校工作，书中部分实验数据集亦由高凯提供。在本书的写作过程中，也得到了多方面的支持与帮助。第 2～6 章中的实例部分分别由何晓艺、张姗姗、孟天宏、刘多星等参加编写。同时，我们也参考了相关文献和互联网上众多热心网友提供的素材。本书的顺利完成也得益于参阅了大量的相关工作及研究成果，在此谨向这些文献的作者、热心网友，以及为本书提供帮助的老师，特别是那些由于篇幅所限未及在参考文献中提及的相关文献的作者和网站，致以诚挚的谢意和崇高的敬意。

由于我们的学识、水平均有限，书中不妥之处在所难免，恳请广大读者批评指正。

编　者

第1版 前言

Foreword

建立在分布式系统之上的大数据搜索与挖掘应用，是当今 IT 业的研究与工程实践热点之一。在 DB-Engines 公布的 2015 年度最受欢迎的数据库系统中，Elasticsearch 名列前茅。作为开源分布式检索与数据处理平台，Elasticsearch 不仅仅是一个数据库，它还是一个基于 Lucene 构建的开源、分布式、RESTful 信息检索框架。基于 Elasticsearch＋Logstash＋Kibana 的信息处理架构，为编程人员提供了一种分布式可扩展的信息存储和全文检索机制以及基于 Logstash 的日志处理机制、基于 Kibana 的挖掘结果可视化机制。它不仅能对海量规模的数据完成分布式索引与检索，还能提供数据聚合分析和可视化。因此，从实战的角度掌握 Elasticsearch、Logstash、Kibana 的基本使用方法和技巧，很有必要。

大数据这个术语的出现，大概可追溯到基于 Lucene 的 Apache 开源项目 Nutch。从 2009 年开始，大数据开始成为互联网行业的流行词汇，也吸引了越来越多的关注。物联网、云计算、移动互联网、手机与平板电脑、PC 以及遍布各个角落的各种各样的传感器，无一不是大数据的来源方或承载方。可以说，大数据就在我们身边。从阿里巴巴、1 号店、京东商城等电子商务数据，到 QQ 等即时聊天内容，再到 Google、Bing、百度，又到社会网络与微博、微信等，都在生产、承载着大数据。随着信息处理量的增大，对大数据的分布式存储、快速搜索与挖掘显得特别必要。例如，挖掘用户的行为习惯和喜好，从凌乱纷繁的大数据背后找到符合用户兴趣和习惯的产品和服务，并对产品和服务进行有针对性的调整和优化，本身就蕴含着巨大的商机。但是，传统的基于关系型数据库管理系统的方法，在高效处理大数据时显得有些力不从心。虽然开源的全文检索工具 Lucene 能处理非结构化和半结构化的信息，但其某些版本在分布式处理方面的不足限制了它在大数据方面的应用。我们希望找到一个快速的分布式信息检索解决方案，希望它是一个零配置和易于上手的全文检索模式，希望它能够简单地使用 JSON 通过 HTTP 索引数据，更希望它支持分布式处理并支持系统扩展，能够实时搜索，并且稳定、可靠。

Elasticsearch 是一个基于 Lucene 的开源分布式信息检索架构和全文搜索工具。构建在 Elasticsearch 基础上的日志处理工具 Logstash 和信息可视化组件 Kibana，能有效衔接并高效处理由 Elasticsearch 索引的分布式数据，

三者优势互补，各司其职，共同完成网络大数据分布式存储、倒排索引、全文检索、Web 日志处理、挖掘结果可视化这一整套的信息处理流程。目前，国内这方面的资料很少，仅有的几部译著所提及的 Elasticsearch 版本较低，且没有任何有关 Logstash 和 Kibana 的书籍。因此，我们萌发了一个想法，将 Elasticsearch、Logstash、Kibana（统称为 ELK）联袂奉献给广大软件开发者，帮助他们尽快熟悉 ELK 架构，并构建自己的 Web 应用程序，完成对分布式信息的检索与分析工作。

本书强调实践，内容新颖，条理清晰，组织合理。通过实战讲解的方式，让读者更好地了解 ELK 架构的实现细节。全书内容涵盖 ELK 简介、文档索引与处理、信息检索与过滤、信息统计与分析、基于 Java 客户端的 Elasticsearch 功能实现、Elasticsearch 配置与管理、基于 Logstash 的网络日志处理、基于 Kibana 的分析结果可视化、应用实例等多个部分。

全书由高凯提出写作大纲，第 1 章和第 6 章由高凯撰写并完成全书通稿和审校工作，其余各章均由高莘撰写。其中，第 1 章概述 Elasticsearch、Logstash、Kibana 的主要功能，对涉及的一些概念进行简介，并从实用的角度出发，通过对实例的讲解，介绍索引、检索的实现机制；第 2 章对 Elasticsearch 中的索引、映射等进行说明；第 3 章介绍 Elasticsearch 中的检索功能；第 4 章介绍基于 Facets、Aggregations 的数据聚合与统计功能；第 5 章从工程实践的角度，介绍面向 Java 客户端的 Elasticsearch 部分功能的设计与实现；第 6 章介绍 Elasticsearch 的配置及一些高级功能、监控等的使用；第 7 章介绍日志处理及 Logstash 的应用；第 8 章介绍基于 Kibana 的可视化技术；第 9 章给出一个综合应用实例，该实例从网页采集、处理、存储、索引、日志处理、可视化展示等入手，介绍了基于 ELK 的分布式信息检索与日志挖掘解决方案。

本书的顺利完成也得益于参阅了大量的相关工作及研究成果，部分内容源自 Elasticsearch、Logstash、Kibana 的官方文档。在写作过程中，也参考了相关文献和互联网上众多热心网友提供的素材，在此谨向这些文献的作者、热心网友以及为本书提供帮助的老师，特别是那些由于篇幅所限未及在参考文献中提及的相关文献的作者和网站，致以诚挚的谢意和崇高的敬意。

由于我们的学识、水平均有限，书中不妥之处在所难免，恳请广大读者批评指正。

编　者

Contents

目录

Chapter 1

第1章 概 述

"Elastic provides a growing platform of open source projects and commercial products designed to search, analyze, and visualize your data, allowing you to get actionable insight in real time. Our products are architected to seamlessly work together as a standalone solution or easily integrate into your existing infrastructure.

At the heart of it all are Elasticsearch, Kibana, Beats, and Logstash, four open source projects that, when combined, are known as the Elastic Stack. Our commercial extensions, Shield, Watcher, and Marvel, add even more value to our open source solutions, offering a growing list of capabilities, including security, alerting, and monitoring."——https://www.elastic.co/products

随着大数据、大型电商网站以及 Web 2.0 技术的普及应用，越来越多的软件开发者需处理海量异构信息的实时索引、检索、日志挖掘、可视化等和信息检索与大数据搜索、挖掘相关的业务。虽然 Lucene 是许多互联网公司的标准信息检索工具之一，但它无法在一个合理的时间内索引和检索海量的大数据，不提供实时检索，不具备良好的可扩展性，一般也不适合针对大数据的搜索、挖掘和云计算环境。

ELK Stack 是以 Elasticsearch、Logstash、Kibana、Beats 等几个开源软件为主的大数据处理工具集，也是目前开源的最流行的大数据分析解决方案之一。根据国际权威的数据库产品评测机构 DB-Engines 统计(http://db-engines.com/en)，在 2016 年 1 月，Elasticsearch 已超过 Solr 等，成为排名第一的搜索引擎类应用。

以 Elasticsearch、Logstash、Kibana 三个开源软件为主的数据处理工具链——即 ELK Stack——为编程人员提供了一个分布式的可扩展的信息存储和基于 Lucene 的信息检索机制、基于 Logstash 的日志处理机制、基于 Kibana 的挖掘结果可视化架构。在一个典型的使用场景中，可以由 Logstash 处理日志等信息，并由它充当"数据搬运工"的角色；用