Community Experience Distilled

Elasticsearch集成Hadoop最佳实践

将Elasticsearch集成进Hadoop生态系统，有效地可视化及分析数据

[美] Vishal Shukla 著
贾传青 译

清华大学出版社

Elasticsearch 集成 Hadoop 最佳实践

[美] Vishal Shukla 著
贾传青 译

清華大學出版社
北京

内 容 简 介

ElasticSearch是一个开源的分布式搜索引擎，具有高可靠性，支持非常多的企业级搜索用例。Elasticsearch Hadoop作为一个完美的工具，用来连接 Elasticsearch 和 Hadoop 的生态系统。通过Kibana技术，Elasticsearch Hadoop很容易从Hadoop 生态系统中获得大数据分析的结果。

本书全面介绍Elasticsearch Hadoop技术用于大数据分析以及数据可视化的方法。内容共分7章，包括Hadoop、Elasticsearch、 Marvel和 Kibana 安装；通过编写 MapReduce 作业，把Hadoop数据导入Elasticsearch；全面分析 Elasticsearch本质，如全文本搜索分析、 查询、 筛选器和聚合；使用 Kibana创建各种可视化和交互式仪表板，并使用Storm和 Elasticsearch分类现实世界的流数据以及相关的其他主题。

本书适合从事大数据分析人员、大数据应用开发的人员参考，也适合高等院校及培训机构相关专业的师生教学参考。

图书在版编目（CIP）数据

Elasticsearch集成Hadoop最佳实践/(美)尔玛·舒克拉(Vishal Shukla)著；贾传青译. —北京：清华大学出版社，2017

书名原文: Elasticsearch for Hadoop

ISBN 978-7-302-46967-4

Ⅰ.①E… Ⅱ.①尔… ②贾… Ⅲ.①互联网络—信息检索 Ⅳ.①G254.928

中国版本图书馆CIP数据核字（2017）第078166号

责任编辑：夏毓彦
封面设计：王 翔
责任校对：闫秀华
责任印制：李红英

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
印 装 者：清华大学印刷厂
经 销：全国新华书店
开 本：185mm×230mm **印 张**：13 **字 数**：333 千字
版 次：2017 年 6 月第 1 版 **印 次**：2017 年 6 月第1 次印刷
印 数：1～3000
定 价：55.00 元

产品编号：070289-01

有关人员

作者
Vishal Shukla

审稿
Vincent Behar
Elias Abou Haydar
Yi Wang

组稿编辑
Vivek Anantharaman
Larissa Pinto

内容开发编辑
Pooja Mhapsekar

技术编辑
Siddhesh Ghadi

文字编辑
Relin Hedly

项目协调
Suzanne Coutinho

校对
Safis Editing

图片
Disha Haria

生产协调
Nilesh R. Mohite

封面
Nilesh R. Mohite

关于作者

Vishal Shukla是Brevitaz系统（http://brevitaz.com）的CEO，一名真诚的技术传道者。他是一位充满激情的软件科学家，同时也是一名大数据专家。Vishal在设计模块化企业系统方面拥有丰富的经验，自从大学时代到编写本书时已经过去了11年，Vishal都一直很喜欢基于JVM的代码开发。他也信奉设计思想和可持续软件开发，在各个行业都拥有丰富的企业系统架构经验。他还热衷于大数据工程、分析及机器学习等技术。

Vishal创办了Brevitaz系统，该公司为全球客户提供大规模可扩展可持续的大数据方案及面向分析的企业应用开发。依靠专业的大数据架构及大数据技术能力，Brevitaz团队对用户原有系统重新设计开发，将其改造成可扩展的最先进的系统。为了给用户提供高质量的产品，Brevitaz把Scrum、测试驱动开发、持续集成及持续交付等敏捷实践吸纳进自己的文化中。

他是一位音乐及艺术爱好者。他喜欢唱歌、乐器、画画等，在空闲时间他喜欢玩板球、乒乓球、游泳等运动。

读者可以通过vishal.shukla@brevitaz.com或者链接到https://in.linkedin.com/in/vishalshu联系Vishal，也可以关注Vishal的Twitter @vishal1shukla2。

致谢

由于撰写本书总是要在公司加班到深夜，在此我要特别感谢我亲爱的妻子Sweta Bhatt　Shukla和我即将出生的宝贝。妻子从不抱怨我没有时间陪她，一直鼓励我完成本书。我衷心感谢我的妹妹Krishna Meet Bhavesh Shah 和Arpit Panchal，她们在本书的细节上进行了详细的检查，提出了很多宝贵的意见，给予了我很多帮助。另外，也要衷心感谢我的兄弟、偶像Pranav Shukla以及我的家人和朋友对我的支持和指导。

我还要感谢我的导师和同事们，是他们的帮助和影响才成就了今天的我。虽然每个人对我的帮助都不可或缺，可能我也无法一一列举，但是我还是要特别感谢Thomas Hirsch、Sven Boeckelmann、Abhay Chrungoo、Nikunj Parmar、Kuntal Shah、Vinit Yadav、Kruti Shukla、Brett Connor和Lovato Claiton。

关于审稿

Vincent Behar是一位充满激情的软件开发人员。他所在公司的搜索引擎索引了160亿的网页。在这样的大数据环境中，他所使用的工具包括Hadoop、MapReduce和Cascading。在公司的ELK环境中，他们使用大规模的Elasticsearch多租户集群来完成索引和搜索需求。因此，下一步他们打算将这两类技术进行集成，未来也会考虑将Elasticsearch与Spark进行集成。

Elias Abou Haydar是巴黎iGraal的数据科学家。他拥有巴黎狄德罗大学计算机科学的硕士学位，在硕士期间主要的方向是分布式系统和算法。作为LIAFA、CNRS的研究员，他主要是为图像分割应用研究分布式图算法。在实习期间他发现了Elasticsearch并且深深地爱上了这一技术。

Yi Wang目前是数据分析公司Trendalytics的首席软件工程师。他主要负责选定、设计并实现数据收集、数据可视化及数据分析的整个流程。他拥有北京大学的物理学硕士学位和哥伦比亚大学的计算机科学硕士学位，具有数学、化学和生物学混合学术背景。

www.PacktPub.com

支持文件、电子书、折扣优惠等

你可以访问www.PacktPub.com来获得和图书配套的文件及资源。

你知道吗？Packt为每本出版的图书提供了PDF、ePub格式的电子版，你可以到www.PacktPub.com下载最新版的电子书。同时，作为一名纸质书用户，你可以在电子书上享受折扣。如果你想了解更多信息，可以通过service@packtpub.com联系我们。

在www.PacktPub.com，你可以免费阅读技术文章，注册即可获得新闻推送，还可以收到Packt图书和电子书的独家折扣及优惠信息。

https://www2.packtpub.com/books/subscription/packtlib

你希望关于IT的问题得到即时的答案吗？PackLib是Packt在线的数字图书馆。在这里，你可以搜索、阅读到Packt所有的图书。

为什么订阅

- 可搜索所有Packt出版的图书
- 可对内容进行复制、粘贴和标记
- 可通过网页浏览器直接访问

Packt免费账户

如果你拥有www.PacktPub.com的Packt账户，就可以使用它访问PacktLib，同时还可以免费阅读9本图书。你只需要使用登录凭据直接登录即可。

前言

在2004年到2006年期间，关于Hadoop的核心组件的讨论都是围绕MapReduce的。Hadoop天生具有分布式运算能力和水平扩展能力，这些特性使其在各个行业被广泛应用。那些超大型的组织认识到Hadoop带来的价值，包括处理TB和PB级数据、采集处理社交数据、利用廉价的商业硬件存储海量数据等。然而，大数据解决方案除了这些以外，还需要解决数据处理的实时性问题，尤其是对非结构化数据的实时性处理。

Elasticsearch是一款高效的分布式搜索及分析引擎，可以让你实时了解你的海量数据。它丰富的查询能力可以帮助你进行复杂的全文检索、基于地理位置的分析及异常检测等。Elasticsearch-Hadoop也被简称为ES-Hadoop，是Elasticsearch和Hadoop的连接器，通过它可以非常方便地在Hadoop生态系统和Elasticsearch之间进行数据交互。你也可以将流式数据从Apache Storm或者Apache Spark写入Elasticsearch进行实时分析。

本书的目标是让你获得真正的利用Hadoop和Elasticsearch的能力。我将带你一步一步地对海量数据进行数据发现和数据探索。你将学习如何将Elasticsearch与Pig、Hive、Cascading、Apache Storm和Apache Spark等Hadoop生态系统工具进行无缝集成。通过本书的学习，你可以使用Elasticsearch创建自己的分析报表。通过强大的数据分析和可视化平台Kibana，你可以对要展示的图形、大小、颜色等进行控制。

在本书中我使用了不少很有意思的数据集，通过这些数据集你将获得真实的数据探索体验。因此，你可以使用我们介绍的工具和技术非常快速地构建基于特定行业的解决方案。我衷心希望阅读本书能够给你带来有趣的学习体验。

本书主要内容

第1章 环境部署

本部分将介绍Java、Hadoop、Elasticsearch及相关插件安装部署的详细步骤。运行

示例作业WordCount，将数据导入Elasticsearch，测试安装的环境是否正常运行。

第2章 初识ES-Hadoop

本部分将介绍WordCount示例程序是如何开发出来的。我们将通过解决现实世界中真实的问题来介绍Elasticsearch和Hadoop。

第3章 深入理解Elasticsearch

本部分将介绍Elasticsearch实现全文检索和分析的技术细节。通过实际的案例，你将学习到如何将数据写入Elasticsearch、在Elasticsearch中如何实现搜索，以及Elasticsearch聚集操作的API。

第4章 利用Kibana进行大数据可视化

本部分将通过真实的案例向你展示如何使用不同形状和颜色的图表来展示海量数据。另外，还演示了如何对数据进行探索，并通过动态仪表盘对数据进行可视化。

第5章 实时分析

本部分将介绍如何通过Elasticsearch和Apache Storm对流式的推文数据进行实时数据分析。我们也将了解到如何通过Elasticsearch来对数据进行挖掘以实现异常检测。

第6章 ES-Hadoop配置

本部分将介绍Elasticsearch和ES-Hadoop库如何在分布式环境中运行，以及如何为了特定的需求进行参数调整。我们还提供了一个用于生产环境部署集群的检查列表。

第7章 与Hadoop生态系统集成

本部分将介绍如何使用Elasticsearch与Pig、Hive、Cascading及Apache Spark等Hadoop生态系统技术进行集成。

附录 配置

本部分将提供ES-Hadoop各配置参数的简短说明。

学习本书的准备工作

本书大部分篇幅用于介绍具体实际案例。你可以通过对本书中的案例进行实际操作来学习，还可以通过本书来学习相关软件和工具安装的具体操作。

如果要搭建本书的实验环境运行相关示例，你需要基于Linux的物理机或者虚拟机。本书中所有的命令和案例都在Ubuntu 14.04发行版上测试通过。而且，除了替换掉Ubuntu特有的操作系统命令外，这些案例在其他的Linux平台上也应该可以正常运行。对于Windows用户，最好的办法是使用VirtualBox或者VMware安装一个Ubuntu虚拟机。

本书读者对象

本书适用于具有Hadoop基本概念的Java程序员，即使之前没有Elasticsearch使用经验也不会影响本书阅读。

排版规范

在本书中我们使用了不同的文本样式来区分不同类型的信息。下面是文本样式示例及相关解释。

代码段样式如下：

```
{
  "properties": {
    "skills": {
      "type": "string",
      "analyzer": "simple"
    }
  }
}
```

当我们希望某些代码可以引起你的注意时会对相关的内容进行加粗：

```
    StatusListener listener = new StatusListener() {
    public void onStatus(Status status) {
            queue.offer(status);
        }
    ...
...
```

命令行的输入或者输出使用如下样式：

```
SELECT c.city, s.skill, avg(c.experience)
```

这个图标表示警告或者需要特别注意的内容。

这个图标表示提示或者技巧。

读者反馈

欢迎读者提出反馈。你对本书有什么想法，喜欢它哪些方面，不喜欢它哪些方面，都可以告诉我们。要写出真正有用的图书，你的反馈是很重要的。

如果你有意见要进行反馈，可以发送邮件到feedback@packtpub.com，并在邮件主题中注明书名（英文书名*Elasticsearch for Hadoop*）。

如果你有某个领域的专业知识，而且有兴趣编写一本书或者为书的编写提供帮助，请参考我们的作者指南www.packtpub.com/authors。

客户支持

现在，你已经成为一位令我们自豪的Packt图书拥有者，我们将全力帮你利用好本书。

下载示例代码

你可以使用你的账户从http://www.packtpub.com下载所购买的图书示例代码。如果你从其他地方购买本书，则需要到http://www.packtpub.com/support网站上进行注册，我们会通过电子邮件将文件发送给你。

本书中国读者代码下载地址（注意数字和字母大小写）如下：

https://pan.baidu.com/s/1pLM5iob（密码: 2w7t）

如果下载有问题，请联系电子邮箱booksaga@163.com，邮件主题为“elasticsearch”。

下载本书的彩色图片

我们还为本书提供了PDF格式的彩色插图。跟纸质书上的黑白图片相比，彩色图片能够让你更直观地理解作者要表达的内容。你可以在https://www.packtpub.com/sites/default/files/downloads/8999OS_ColorImages.pdf下载这个文件。

勘误表

虽然我们尽可能地保证本书内容的正确性，但是疏漏在所难免。如果你发现了本书中的文字错误或者代码错误就请告知我们，我们将感激不尽。如果你这么做，就可以帮助到其他遇到同样问题的读者，同时也可以帮助我们改进本书的后续版本。当你

发现了错误时，请访问http://www.packtpub.com/submit-errata，选择要提交的书，单击Errata Submission Form，并输入详细的内容。勘误一经核实，你的提交就会被接受，此勘误将上传到我们的网站或者添加到现有的勘误表中。

如果你想查看之前提交的勘误，请访问https://www.packtpub.com/books/content/support，输入书的名称搜索即可。

侵权行为

版权内容在互联网上的盗版是所有媒体需要面对的问题。Packt很重视版权保护和许可证。如果你发现我们的图书在互联网上以任何形式非法复制传播，请立即为我们提供地址或者网站名称，以便我们进行补救。

请把可疑盗版材料链接发送到copyright@packtpub.com。

非常感谢你对作者的保护，以及对我们持续为你提供有价值内容的能力的保护。

问题

如果你对本书内容有疑问，可以随时联系questions@packtpub.com，我们将竭尽全力为你解决。

目录

第1章 环境部署 1

1.1 安装部署Hadoop集群 ····· 1

Java安装和配置 ····· 2

用户添加和配置 ····· 2

SSH认证配置 ····· 3

Hadoop下载 ····· 4

环境变量配置 ····· 4

Hadoop配置 ····· 5

配置core-site.xml ····· 6

配置hdfs-site.xml ····· 6

配置yarn-site.xml ····· 6

配置mapred-site.xml ····· 7

格式化HDFS ····· 7

启动Hadoop进程 ····· 8

1.2 安装Elasticsearch及相关插件 ····· 8

下载Elasticsearch ····· 9

配置Elasticsearch ····· 9

安装Head插件 ····· 11

安装Marvel插件 ····· 11

启动Elasticsearch ····· 12

1.3 运行WordCount示例 ····· 13

下载编译示例程序 ····· 13

将示例文件上传到HDFS ····· 13
运行第一个作业 ····· 14
1.4 使用Head 和 Marvel浏览数据 ····· 16
使用Head浏览数据 ····· 16
初识Marvel ····· 18
使用Sense浏览数据 ····· 19
小结 ····· 21

第2章 初识ES-Hadoop 22

2.1 理解WordCount程序 ····· 23
理解Mapper ····· 23
理解Reducer ····· 24
理解Driver ····· 25
使用旧的API——org.apache.hadoop.mapred ····· 28
2.2 实际案例——网络数据监控 ····· 28
获取并理解数据 ····· 28
明确问题 ····· 29
解决方案 ····· 30
解决方案1——预聚合结果 ····· 30
解决方案2——直接查询聚合结果 ····· 32
2.3 开发MapReduce作业 ····· 33
编写Mapper类 ····· 34
编写Driver ····· 37
编译作业 ····· 38
上传数据到HDFS ····· 41
运行作业 ····· 41
查看TOP N结果 ····· 42
2.4 将数据从Elasticsearch写回HDFS ····· 44
了解Twitter数据集 ····· 44
导入Elasticsearch ····· 45
创建MapReduce作业 ····· 46
编写Tweets2HdfsMapper ····· 46

运行示例 ····· 50
确认输出 ····· 50
小结 ····· 52
第3章 深入理解Elasticsearch 53
3.1 理解搜索 ····· 53
观念转换 ····· 54
索引 ····· 54
类型 ····· 55
文档 ····· 55
字段 ····· 55
3.2 与Elasticsearch交互 ····· 56
Elasticsearch的CRUD ····· 56
创建文档 ····· 56
获取文档 ····· 57
更新文档 ····· 58
删除文档 ····· 58
创建索引 ····· 58
映射 ····· 59
数据类型 ····· 60
创建映射 ····· 61
索引模板 ····· 62
3.3 控制索引过程 ····· 63
什么是反转索引 ····· 63
输入数据分析 ····· 64
停止词 ····· 64
大小写 ····· 65
词根 ····· 65
同义词 ····· 65
分析器 ····· 65
3.4 Elastic查询 ····· 67
编写查询语句 ····· 68
URI查询 ····· 68

match_all查询 ……68
term查询 ……68
boolean查询 ……70
match查询 ……71
range查询 ……72
wildcard查询 ……73
过滤器 ……73
3.5 聚合查询 …… 75
执行聚合查询 …… 76
terms聚合 ……76
histogram聚合 ……78
range聚合 ……78
geo distance聚合 ……79
嵌套聚合 …… 81
自测题 ……82
小结 …… 82
第4章 利用Kibana进行大数据可视化 83
4.1 安装部署 …… 83
Kibana安装 …… 84
准备数据 …… 84
自测题 …… 85
启动Kibana …… 86
4.2 数据发现 …… 87
4.3 数据可视化 …… 90
饼图 …… 91
堆积柱状图 …… 94
使用堆积柱状图完成日期直方图 …… 96
面积图 …… 97
饼图组图 …… 98
环形图 …… 98
瓦片地图 …… 99
自测题 …… 100