



资深技术专家手把手教你配置高效的Hadoop集群，充分利用Hadoop平台的优势
全面讲解Hadoop大数据平台安全的实现方案和最佳实践，为Hadoop生态系统实现强健
的安全保障



技术丛书



Hadoop Cluster Deployment, Securing Hadoop

Hadoop集群与安全

(美) Danil Zburivsky Sudheesh Narayanan 著

刘杰 沈鑫 ◎ 译



机械工业出版社
China Machine Press

技术丛书

Hadoop Cluster Deployment, Securing Hadoop

Hadoop集群与安全

(美) Danil Zburivsky Sudheesh Narayanan 著

刘杰 沈鑫 ◎ 译



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

Hadoop 集群与安全 / (美) 扎伯瑞福斯基 (Zburivsky, D.), (美) 娜拉燕南 (Narayanan, S.) 著; 刘杰, 沈鑫译. —北京: 机械工业出版社, 2014.10

(大数据技术丛书)

书名原文: Hadoop Cluster Deployment, Securing Hadoop

ISBN 978-7-111-48062-4

I. H… II. ① 扎… ② 娜… ③ 刘… ④ 沈… III. 数据处理软件 IV. TP274

中国版本图书馆 CIP 数据核字 (2014) 第 220820 号

本书版权登记号: 图字: 01-2014-2029

Danil Zburivsky: Hadoop Cluster Deployment (ISBN: 978-1-78328-171-8)

Sudheesh Narayanan: Securing Hadoop (ISBN: 978-1-78328-525-9)

Copyright © 2013 Packt Publishing. First published in the English language under the title “Hadoop Cluster Deployment” and “Securing Hadoop”.

All rights reserved.

Chinese simplified language edition published by China Machine Press.

Copyright © 2014 by China Machine Press.

本书中文简体字版由 Packt Publishing 授权机械工业出版社独家出版。未经出版者书面许可, 不得以任何方式复制或抄袭本书内容。

Hadoop 集群与安全



出版发行: 机械工业出版社 (北京市西城区百万庄大街 22 号 邮政编码: 100037)

责任编辑: 吴 怡

责任校对: 殷 虹

印 刷: 北京市荣盛彩色印刷有限公司

版 次: 2014 年 11 月第 1 版第 1 次印刷

开 本: 186mm×240mm 1/16

印 张: 11.5

书 号: ISBN 978-7-111-48062-4

定 价: 49.00 元

凡购本书, 如有缺页、倒页、脱页, 由本社发行部调换

客服热线: (010) 88378991 88361066

投稿热线: (010) 88379604

购书热线: (010) 68326294 88379649 68995259

读者信箱: hzsj@hzbook.com

版权所有·侵权必究

封底无防伪标均为盗版

本书法律顾问: 北京大成律师事务所 韩光 / 邹晓东

信息安全是国家安全的重要组成部分，而数据安全则是信息安全的關鍵。在大数据时代，数据的集中为数据深入分析提供了可能，也带来了安全隐患。一方面，大数据系统集成了几十个复杂的分布式系统，系统的复杂性也带来新的安全漏洞，大规模集群环境对网络安全的防护带来了挑战；另一方面，数据的深入分析将挖掘出核心商业知识或者个人隐私信息，如何保护这些数据成为大数据项目成败的关键。本书详细介绍了大数据平台 Hadoop 生态系统的集群搭建，并围绕这个平台提出一套完整的大数据安全体系架构，给出了可以直接实施的解决方案。

本书由两个部分组成，第 1 ~ 6 章从基本的 Hadoop 知识开始，逐步带领读者构建一个完整的 Hadoop 网络，同时在每一个章节中都会列举完成实际项目所需要的基本概念和知识，最后通过实际的动手操作让读者对所学到的知识有更深层次的了解，有助于读者日后能够自行根据需求进行开发和设计。此外作者还提供了一份详细的补充资源列表，其中包括许多实际项目示例的网站以及交流社区，通过这些额外的学习资料，读者可以进一步拓展自己在 Hadoop 网络的知识面。第 7 章 ~ 13 章涵盖认证授权、数据加密、访问控制、审计日志等 Hadoop 安全的各个方面，提供保障基于 Hadoop 的大数据平台安全的实现方案和最佳实践。

本书的特色是考虑了企业信息架构的需求，讲解了在各个环节如何与已有的企业系统集成，形成安全的大数据平台。本书还十分完整地介绍了 Hadoop 生态系统各个子系统的安全防护方案，包含 MapReduce、HDFS、Hive、Pig、HBase、Oozie 等。本书写作浅显易懂，很适合 Hadoop 系统运维人员作为操作手册。通过阅读本书可以让你从另一个角度来了解 Hadoop 生态系统，来理解大数据。

本书的作者是一位精通语法的语言天才，原书的语言诙谐幽默，但又不失技术类书籍的专业性。本书由沈鑫与刘杰合作翻译，沈鑫翻译了第 1 ~ 6 章，刘杰翻译了第 7 ~ 12 章。译者在翻译的过程中力求在最大程度上还原原书的精华之处，即使如此可能还会有不尽人意的地方，希望各位读者批评指正。

刘杰致谢：

很高兴可以翻译市面上第一本介绍 Hadoop 安全的书籍。在本书成稿之时，正好我家宝宝诞生，这也算是双喜临门吧。十分感谢支持我的家人、同事和学生。

沈鑫致谢：

我要感谢我的父母和教导过我的老师，感谢他们无私地把知识传授给我。

About the Author 作者简介

Danil Zburivsky 是一名致力于开源技术的数据库工程师。他的职业起点是 MySQL 数据库管理员，目前的职业是全球数据基础构架管理公司 Pythian 的一名咨询师。在 Pythian 公司，他负责为金融、娱乐以及通信领域的客户创建各种 Hadoop 集群。

他的研究爱好包括使用 Python 编写有趣的程序、机器人学以及机器学习。他是各种行业会议的客座讲师。

感谢我的妻子体谅我花费了盛夏的大部分夜晚来编写本书。同样我还要感谢 Pythian 公司的同事，尤其是 Alan Gardner、Cyril Ganchev 和 Yanick Champoux，他们是本书的审校者，为本书做出了很多贡献。

Sudheesh Narayannan 是一位在大数据解决方案咨询与实施领域具有丰富经验的实践者和技术战略家。他在信息管理、商务智能、大数据分析以及云应用和 J2EE 应用开发领域有超过 15 年的 IT 经验。他擅长架构、设计与开发大数据产品、云管理平台和高度可扩展平台服务等，他钻研的技术包括 Hadoop 平台及相关生态系统、NoSQL 数据库（MongoDB、Cassandra 和 HBase）、文本分析（GATE 和 OpenNLP）、机器学习（Mahout、Weka 和 R），以及复杂事件处理等。

他目前是 Genpact 公司的助理副总裁和首席大数据架构师。专注于推动创新，建立知识产权资产、框架与解决方案。入职 Genpact 前，他是 Infosys BigDataEdge 产品的联合发明人和首席架构师。

我十分感谢我的妻子 Smita 和儿子 Aryan，感谢他们在我写作期间的奉献与支持，感谢我的爸爸、妈妈和妹妹一直鼓励我为社区做出回馈。这本书离不开他们的鼓励和持续的支持。

特别感谢 Rupak 与 Debika 牺牲个人周末时间和我一起实验 Hadoop 安全的一些思路。

我要感谢 Shwetha、Sivaram、Ajay、Manpreet 和 Venky 为本书不断提供反馈意见，帮助我在整个过程持续改进写作。

此外，我要诚挚地感谢我的老师 N.C. Jain 教授，我的领导 Paddy、Vishnu Bhat、Sandeep Bhagat、Jaikrishnan、Anil D'Souza 和 KNM Rao，感谢他们对我的指导和培养，我才有能力来写这本书。

About the Author 审校者简介

Skanda Bhargav 毕业于 Visvesvaraya 理工大学（位于印度卡纳塔克邦的贝尔高姆），他的主修专业是计算机科学以及工程学。他目前就职于总部位于班加罗尔的一家跨国公司。他是 Apache Hadoop 的一名注册开发工程师，他的研究兴趣是大数据和 Hadoop。

我要感谢家人在求学阶段给予的支持和信任。我的朋友们给予了我追求完美的信心。我很高兴上帝保佑我身边有那么多善良的人，如果没有这些我可能无法取得今天这样的成就。

Yanick Champoux 目前是大数据领域的一名解决方案架构师。在业余时间，他研究 Perl，种植胡姬花，创作漫画书。

Cyril Ganchev 是一名系统管理员、数据库管理员和软件开发工程师，居住在保加利亚索非亚。在 2005 年获得索非亚理工大学计算机系统和技术的硕士学位。

2002 年他在索非亚理工大学求学期间，就在一家网吧担任系统管理员。2004 年他开始在保加利亚最大的 IT 公司 Information Services Plc 中担任软件开发工程师。他参与了保加利亚政府、保加利亚国家银行以及保加利亚国家税务局等多个组织的不同项目。他开发过统计保加利亚政府选举结果的程序。

直到 2012 年他都在远程为加拿大公司 Pythian 工作，他的职位是 Oracle 数据库管理员。在 2013 年他转投了另一个研究大数据和 NoSQL 的新团队。

他拥有 Oracle 高级 PL/SQL 认证开发专家以及 Oracle 11g 数据库管理员认证专员。

我要衷心感谢父母给予的一贯支持。

Alan Gardner 是一名解决方案架构师和侧重于大数据系统设计的开发工程师。这些系统融合了来自 Hadoop、Apache Kafka 和 Storm 以及数据科学的技术。Alan 乐于将他的项目和经验在用户组与会议上进行分享。在业余时间他还研究函数式编程、移动开发和网络编程。

Alan 还是渥太华开发者社区的资深会员，他向各种不同的公司提供咨询，帮助非技术股东组织开发者会议。在渥太华 Drones 组织的编程马拉松中，开发者可以通过网络交流想法并且在编写飞行机器人时锻炼他们的技巧。

我要感谢 Paul White、Alex Gorbachev 和 Mick Saunders 帮助我在职业生涯的不同阶段选择正确的道路，同样也感谢 Jasmin 一贯的支持。

Mark Kerzner 同时拥有法学、数学和计算机学位。他自 2008 年开始为 Hadoop 生态系统设计软件。他是 SHMsoft 公司总裁，为各个领域提供 Hadoop 应用，联合编写了多本 Hadoop 相关教程。他也是多本书籍的联合作者，多项专利发明人。

我十分感谢我的同事 Sujee Maniyam 对我的帮助，还有我多才多艺的家人。

Nitin Pawar 职业生涯开始于软件和工具开发，然后转型为操作和解决方案开发、过程管理、大数据分析等。目前，他作为大数据系统架构师致力于解决客户管理相关的问题。他的主要工作围绕第一代 Hadoop 生态系统展开。

Preface 前言

在过去的数年中，Hadoop 已经成为数据集成平台的标准解决方案。各企业都在数据基础设施架构中融入新的技术，这要求系统工程师和数据管理员要迅速地掌握新组件的各个方面。由于 Hadoop 是一款包含多种软件的产品，其中集成了多种不同的开源项目，因此掌握这些技术并不简单，需要进行正确安装和配置才能确保 Hadoop 平台的健壮性和稳定性。

许多已有的 Hadoop 发行版本通过某种类型的图形界面来简化 Hadoop 的安装。这种方法极大地缩短了构建完整 Hadoop 集群所需的时间，同时也降低了管理集群配置的复杂性。自动化安装和配置的问题在于隐藏了许多有关 Hadoop 组件协同工作的重要细节，比如某些组件需要其他组件的原因或者哪些是最为关键的配置参数，等等。

本书前 6 章讲解如何手动安装和配置所有主要 Hadoop 组件。通过设置一个完全可运行的集群，你将会对 Hadoop 内部的运行机制产生一定的了解并且有助于你调试任何可能发生的问题。你也可以将第 1 ~ 6 章用作主要 Hadoop 组件和综合配置选项的快速参考。在编写本书的过程中，我自己在实际配置 Hadoop 集群时也参阅了本书来查找某个特定的变量或者在配置操作系统时寻找最佳步骤。这更加让我坚信，本书会其他充满热情和经验丰富的 Hadoop 管理员与开发者提供帮助。

目前，许多组织在生产环境中实现 Hadoop。当一个组织开始大数据之旅，大数据的安全就成为一个重要的问题，而保障敏感数据的安全具有最高的优先级。企业安全部门会担心将 Hadoop 集成到企业系统后带来的安全隐患。本书第 7 ~ 13 章则提供保障基于 Hadoop 的大数据平台安全的实现方案和最佳实践，覆盖 Kerberos 安全协议和 Hadoop 安全机制的设计思想，并且包括在企业内部来确保 Hadoop 及其生态系统安全的详细方法。这部分的目标是通过深入研究大数据安全参考架构，从企业视角提供端到端的大数据安全，并且详细介绍如何集成各种技术构建一个安全的大数据平台。

本书的内容

第 1 章“构建 Hadoop 集群”回顾主要的 Hadoop 组件以及选择规划和集群硬件的方法，同时还介绍不同的 Hadoop 发行版本。

第 2 章“安装和配置 Hadoop”提供安装和配置主要 Hadoop 组件的详细步骤：NameNode 节点（包括高可用方案）、JobTracker 节点、DataNode 节点以及 TaskTracker。

第 3 章“配置 Hadoop 生态系统”介绍 Sqoop、Hive 和 Impala 的配置步骤。

第 4 章“Hadoop 安全”讲解确保不同 Hadoop 组件安全的方法，同时还对在 Hadoop 中配置 Kerberos 进行介绍。

第 5 章“监控 Hadoop 集群”指导读者逐步将集群开发至实际应用阶段。

第 6 章“在云端使用 Hadoop”介绍如何在虚拟环境中使用 Hadoop，其中包括 Elastic MapReduce 和 Whirr 的应用。

第 7 章“Hadoop 平台安全概述”阐述基于 Hadoop 的大数据平台安全保障带来的关键挑战和需求，提供大数据安全的企业视角，详细介绍大数据安全参考架构。

第 8 章“Hadoop 安全体系设计”详细介绍 Hadoop 安全保障系统的内部设计细节，讲解 Kerberos 安全机制的关键概念。重点介绍理解本书其他章节所需要的各种术语和概念。

第 9 章“配置一个安全 Hadoop 集群”介绍配置 Kerberos 并建立安全 Hadoop 集群的详细步骤。

第 10 章“Hadoop 生态系统安全保障”详细介绍 Hadoop 生态系统组件之间的交互和通信协议，及带来的安全隐患，提供构建安全的大数据生态系统的详细步骤。

第 11 章“集成 Hadoop 与企业安全系统”介绍如何集成 Hadoop 安全模型与企业已有的安全系统，如何集中化管理 Hadoop 生态系统中的用户访问控制。

第 12 章“Hadoop 中敏感数据的安全保护”提供保护 Hadoop 生态系统中敏感数据的详细实现方法，介绍大数据平台中适用的各种数据加密技术。

第 13 章“安全事件与审计日志”深入介绍大数据平台中安全事件监控系统，提供实现

安全流程和策略的最新实践。

附录“Hadoop 安全机制解决方案”介绍用于构建安全的 Hadoop 大数据生态系统所需要的各种商业的或开源的技术，深入介绍这些技术的细节及它们在大数据安全参考架构中的位置和作用。

阅读本书需要的条件

本书中的示例对于软件并没有特定的要求。如果计划在沙盒环境下安装 Hadoop 集群，那么读者可能需要某种虚拟化软件（比如 Virtual Box）。

为了实践本书提供的例子，则需要一个 Hadoop 集群。你需要一个多节点 Linux 集群（最少两个节点安装有 CentOS 6.2 操作系统或类似系统）。推荐采用 Cloudera CDH4.1 来安装 Hadoop 集群。也可以采用任何最新版本的 Apache Hadoop 发布版本。需要从 MIT 网站下载并安装 Kerberos 5 Release 1.11.3 ([http:// web.mit.edu/kerberos/krb5-1.11/](http://web.mit.edu/kerberos/krb5-1.11/))。

适合本书的读者群

本书十分适合 Hadoop 实践者（大数据架构师、开发人员、系统管理员）阅读，他们具有一定的 Hadoop 工作经验并且期望保障 Hadoop 的安全。本书也适合于为企业设计端到端安全保障的大数据平台的大数据架构师。本书同时适合作为实现和配置安全 Hadoop 系统管理员的参考手册。

约定

本书在排版中会采用一些特殊字体与正文区别，下面是一些示例及其说明。

代码片段如下所示：

```
{
  "access_id": "Your Access Key ID",
  "private_key": "Your AWS Secret Access Key",
  "keypair": "emr-keys",
  "key-pair-file": "/path/to/key-file/emr-keys.pem",
  "log_uri": "s3n://emr-logs-x123/",
  "region": "us-east-1"
}
```

如果我们希望读者对某些代码块引起重视，则会设置为粗体：

```
{
"access_id": "Your Access Key ID",
"private_key": "Your AWS Secret Access Key",
"keypair": "emr-keys",
"key-pair-file": "/path/to/key-file/emr-keys.pem",
"log_uri": "s3n://emr-logs-x123/",
"region": "us-east-1"
}
```

任何命令行的输入或者输出如下所示：

```
# hdfs dfs -mkdir /warehouse
# hdfs dfs -chmod a+w /warehouse
```

特殊段落的标识如下：



这里给出警告或者重要注意事项。



这里给出提示或者技巧。

译者序

作者简介

审校者简介

前言

第 1 章 构建 Hadoop 集群	1
1.1 选择 Hadoop 集群硬件	2
1.1.1 选择 DataNode 硬件	3
1.1.2 低存储密度集群	4
1.1.3 高存储密度集群	5
1.1.4 NameNode 和 JobTracker 硬件配置	6
1.1.5 网关和其他辅助服务	8
1.1.6 网络配置	8
1.1.7 Hadoop 硬件总结	9
1.2 Hadoop 发行版	10
1.2.1 Hadoop 版本	10
1.2.2 选择 Hadoop 发行版	11
1.2.3 Cloudera Hadoop 发行版	11
1.2.4 Hortonworks Hadoop 发行版	12
1.2.5 MapR	12
1.3 为 Hadoop 集群选择操作系统	13
1.4 小结	14

第 2 章 安装和配置 Hadoop	15
2.1 在 Hadoop 集群中配置操作系统	15
2.1.1 选择和设置文件系统	15
2.1.2 设置 Java 开发包	16
2.1.3 其他操作系统设定	17
2.1.4 设置 CDH 存储库	18
2.2 设置 NameNode	18
2.2.1 JournalNode 节点、ZooKeeper 以及故障转移控制器	22
2.2.2 Hadoop 配置文件	23
2.2.3 NameNode 高可用方案配置	25
2.2.4 JobTracker 配置	31
2.2.5 DataNode 配置	36
2.3 小结	47
第 3 章 配置 Hadoop 生态系统	48
3.1 托管 Hadoop 生态项目	48
3.2 Sqoop	49
3.2.1 安装和配置 Sqoop	49
3.2.2 Sqoop 导入示例	50
3.2.3 Sqoop 导出示例	52
3.3 Hive	52
3.3.1 Hive 架构	53
3.3.2 安装 Hive Metastore	54
3.3.3 安装 Hive 客户端	56
3.3.4 安装 Hive Server	57
3.4 Impala	59
3.4.1 Impala 架构	59
3.4.2 安装 Impala state store	60
3.4.3 安装 Impala server	60
3.5 小结	63
第 4 章 Hadoop 安全	64
4.1 Hadoop 安全概述	64

4.2 Hadoop 分布式文件系统安全	65
4.3 MapReduce 安全	66
4.4 Hadoop 服务级别验证	68
4.5 Hadoop 和 Kerberos	69
4.5.1 Kerberos 概述	70
4.5.2 Hadoop 中的 Kerberos	71
4.6 小结	76
第 5 章 监控 Hadoop 集群	77
5.1 监控策略介绍	77
5.2 Hadoop 参数	78
5.2.1 JMX 参数	79
5.2.2 使用 Nagios 监控 Hadoop	80
5.2.3 监控 Hadoop 分布式文件系统	81
5.2.4 NameNode 校验	81
5.2.5 JournalNode 检查	83
5.2.6 ZooKeeper 检查	83
5.3 监控 MapReduce	84
5.4 使用 Ganglia 监控 Hadoop	85
5.5 小结	86
第 6 章 在云端使用 Hadoop	87
6.1 Amazon Elastic MapReduce	87
6.1.1 安装 EMR 命令行接口	88
6.1.2 选择 Hadoop 版本	89
6.1.3 启动 EMR 集群	89
6.2 使用 Whirr	93
6.3 小结	94
第 7 章 Hadoop 平台安全概述	95
7.1 为什么需要保障 Hadoop 生态系统的安全	96
7.2 确保 Hadoop 生态系统安全面临的挑战	96