

O'REILLY®



图灵程序设计丛书



# Hadoop安全

## 大数据平台隐私保护

Hadoop Security: Protecting Your Big Data Platform

“Hadoop之父”推荐阅读

[美] Ben Spivey, Joey Echeverria 著

赵双 白波 译



中国工信出版集团



人民邮电出版社  
POSTS & TELECOM PRESS

# 版权信息

书名：Hadoop安全：大数据平台隐私保护

作者：[美] Ben Spivey Joey Echeverria

译者：赵双 白波

ISBN：978-7-115-46771-3

本书由北京图灵文化发展有限公司发行数字版。版权所有，侵权必究。

---

您购买的图灵电子书仅供您个人使用，未经授权，不得以任何方式复制和传播本书内容。

我们愿意相信读者具有这样的良知和觉悟，与我们共同保护知识产权。

如果购买者有侵权行为，我们可能对该用户实施包括但不限于关闭该帐号等维权措施，并可能追究法律责任。

---

图灵社区会员 xiaoqi8866 (this-only-one@163.com) 专享 尊重版权

版权声明

O'Reilly Media, Inc. 介绍

业界评论

序

前言

目标读者

排版约定

使用代码示例

Safari® Books Online

联系我们

致谢

来自Joey的致谢

来自Ben的致谢

来自Eddie的致谢

免责声明

电子书

第1章 引言

1.1 安全概览

1.1.1 机密性

1.1.2 完整性

1.1.3 可用性

1.1.4 验证、授权和审计

1.2 Hadoop安全：简史

1.3 Hadoop组件和生态系统

1.3.1 Apache HDFS

1.3.2 Apache YARN

1.3.3 Apache MapReduce

1.3.4 Apache Hive

- 1.3.5 Cloudera Impala
- 1.3.6 Apache Sentry
- 1.3.7 Apache HBase
- 1.3.8 Apache Accumulo
- 1.3.9 Apache Solr
- 1.3.10 Apache Oozie
- 1.3.11 Apache ZooKeeper
- 1.3.12 Apache Flume
- 1.3.13 Apache Sqoop
- 1.3.14 Cloudera Hue

#### 1.4 小结

### 第一部分 安全架构

## 第 2 章 保护分布式系统

### 2.1 威胁种类

- 2.1.1 非授权访问 / 伪装
- 2.1.2 内在威胁
- 2.1.3 拒绝服务
- 2.1.4 数据威胁

### 2.2 威胁和风险评估

- 2.2.1 用户评估
- 2.2.2 环境评估

### 2.3 漏洞

### 2.4 深度防御

### 2.5 小结

## 第 3 章 系统架构

### 3.1 运行环境

### 3.2 网络安全

#### 3.2.1 网络划分

- 3.2.2 网络防火墙
    - 3.2.3 入侵检测和防御
  - 3.3 Hadoop角色和隔离策略
    - 3.3.1 主节点
    - 3.3.2 工作节点
    - 3.3.3 管理节点
    - 3.3.4 边界节点
  - 3.4 操作系统安全
    - 3.4.1 远程访问控制
    - 3.4.2 主机防火墙
    - 3.4.3 SELinux
  - 3.5 小结
- 第 4 章 Kerberos
  - 4.1 什么是Kerberos
  - 4.2 Kerberos概览
  - 4.3 Kerberos工作流：一个简单示例
  - 4.4 Kerberos信任
  - 4.5 MIT Kerberos
    - 4.5.1 服务端配置
    - 4.5.2 客户端配置
  - 4.6 小结
- 第二部分 验证、授权和审计
- 第 5 章 身份和验证
  - 5.1 身份
    - 5.1.1 将Kerberos主体映射为用户名
    - 5.1.2 Hadoop用户到组的映射
    - 5.1.3 Hadoop用户配置
  - 5.2 身份验证

|       |       |                     |
|-------|-------|---------------------|
|       | 5.2.1 | Kerberos            |
|       | 5.2.2 | 用户名和密码验证            |
|       | 5.2.3 | 令牌                  |
|       | 5.2.4 | 用户模拟                |
|       | 5.2.5 | 配置                  |
| 5.3   |       | 小结                  |
| 第 6 章 |       | 授权                  |
| 6.1   |       | HDFS授权              |
|       |       | HDFS扩展ACL           |
| 6.2   |       | 服务级授权               |
| 6.3   |       | MapReduce和YARN的授权   |
|       | 6.3.1 | MapReduce (MR1)     |
|       | 6.3.2 | YARN (MR2)          |
| 6.4   |       | ZooKeeper ACLs      |
| 6.5   |       | Oozie授权             |
| 6.6   |       | HBase和Accumulo的授权   |
|       | 6.6.1 | 系统、命名空间和表级授权        |
|       | 6.6.2 | 列级别和单元级别授权          |
| 6.7   |       | 小结                  |
| 第 7 章 |       | Apache Sentry (孵化中) |
| 7.1   |       | Sentry概念            |
| 7.2   |       | Sentry服务            |
|       |       | Sentry服务配置          |
| 7.3   |       | Hive授权              |
|       |       | Hive Sentry的配置      |
| 7.4   |       | Impala授权            |
|       |       | Impala的Sentry配置     |
| 7.5   |       | Solr授权              |

## Solr的Sentry配置

### 7.6 Sentry特权模型

#### 7.6.1 SQL特权模型

#### 7.6.2 Solr特权模型

### 7.7 Sentry策略管理

#### 7.7.1 SQL命令

#### 7.7.2 SQL策略文件

#### 7.7.3 Solr策略文件

#### 7.7.4 策略文件的验证和校验

#### 7.7.5 从策略文件迁移

### 7.8 小结

## 第8章 审计

### 8.1 HDFS审计日志

### 8.2 MapReduce审计日志

### 8.3 YARN审计日志

### 8.4 Hive审计日志

### 8.5 Cloudera Impala审计日志

### 8.6 HBase审计日志

### 8.7 Accumulo审计日志

### 8.8 Sentry审计日志

### 8.9 日志聚合

### 8.10 小结

## 第三部分 数据安全

## 第9章 数据保护

### 9.1 加密算法

### 9.2 静态数据加密

#### 9.2.1 加密和密钥管理

#### 9.2.2 HDFS静态数据加密

- 9.2.3 MapReduce2中间数据加密
  - 9.2.4 Impala磁盘溢出加密
  - 9.2.5 全盘加密
  - 9.2.6 文件系统加密
  - 9.2.7 Hadoop中重要数据的安全考虑
- 9.3 动态数据加密
  - 9.3.1 传输层安全
  - 9.3.2 Hadoop动态数据加密
- 9.4 数据销毁和删除
- 9.5 小结
- 第 10 章 数据导入安全
  - 10.1 导入数据的完整性
  - 10.2 数据导入的机密性
    - 10.2.1 Flume加密
    - 10.2.2 Sqoop加密
  - 10.3 导入工作流
  - 10.4 企业架构
  - 10.5 小结
- 第 11 章 数据提取和客户端访问安全
  - 11.1 Hadoop命令行接口
  - 11.2 保护应用安全
  - 11.3 HBase
    - 11.3.1 HBase shell
    - 11.3.2 HBase REST网关
    - 11.3.3 HBase Thrift网关
  - 11.4 Accumulo
    - 11.4.1 Accumulo shell
    - 11.4.2 Accumulo代理服务

|        |                     |
|--------|---------------------|
| 11.5   | Oozie               |
| 11.6   | Sqoop               |
| 11.7   | SQL访问               |
| 11.7.1 | Impala              |
| 11.7.2 | Hive                |
| 11.8   | WebHDFS/HttpFS      |
| 11.9   | 小结                  |
| 第 12 章 | Cloudera Hue        |
| 12.1   | Hue HTTPS           |
| 12.2   | Hue身份验证             |
| 12.2.1 | SPNEGO后端            |
| 12.2.2 | SAML后端              |
| 12.2.3 | LDAP后端              |
| 12.3   | Hue授权               |
| 12.4   | Hue SSL客户端配置        |
| 12.5   | 小结                  |
| 第四部分   | 综合应用                |
| 第 13 章 | 案例分析                |
| 13.1   | 案例分析：Hadoop 数据仓库    |
| 13.1.1 | 环境搭建                |
| 13.1.2 | 用户体验                |
| 13.1.3 | 小结                  |
| 13.2   | 案例分析：交互式HBase Web应用 |
| 13.2.1 | 设计与架构               |
| 13.2.2 | 安全需求                |
| 13.2.3 | 集群配置                |
| 13.2.4 | 实现中的注意事项            |
| 13.2.5 | 小结                  |

后记

统一授权

数据管控

原生数据保护

结语

关于作者

关于封面

# 版权声明

Copyright © 2015 Joseph Echeverria and Benjamin Spivey.

Simplified Chinese Edition, jointly published by O'Reilly Media, Inc. and Posts & Telecom Press, 2017. Authorized translation of the English edition, 2017 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 O'Reilly Media, Inc. 出版 2015 。

简体中文版由人民邮电出版社出版，2017。英文原版的翻译得到 O'Reilly Media, Inc. 的授权。此简体中文版的出版和销售得到出版权和销售权的所有者——O'Reilly Media, Inc. 的许可。

版权所有，未得书面许可，本书的任何部分和全部不得以任何形式重制。

# O'Reilly Media, Inc. 介绍

O'Reilly Media 通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自 1978 年开始，O'Reilly 一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly 的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly 为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了 *Make* 杂志，从而成为 DIY 革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly 的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly 现在还将先锋专家的知识传递给普通的计算机用户。无论是通过图书出版、在线服务或者面授课程，每一项 O'Reilly 的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

# 业界评论

“**O'Reilly Radar** 博客有口皆碑。”

——*Wired*

“**O'Reilly** 凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“**O'Reilly Conference** 是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本 **O'Reilly** 的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“**Tim** 是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照 **Yogi Berra** 的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，**Tim** 似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

# 序

近来，“Hadoop 安全”成为了一个充满矛盾的名词。雅虎和 Facebook 等公司创建和使用的这个大数据平台的早期版本，在保护其存储的数据方面并没有花费太多精力。实际上也没有必要这么做，因为 Hadoop 里几乎没有什么敏感数据，只有对黑客毫无吸引力的状态更新和新闻报道。而且，找到它们也不需要花费很多精力。

然而，当这种平台被更多传统企业使用时，便开始涉及更多传统企业数据。金融交易、个人银行账号和税务信息、医疗记录等诸如此类的信息，正是黑客们所追逐的。Hadoop 如今广泛用于零售、银行和医疗保健行业，它也逐渐引来了小偷们的注意。

如果说数据是有吸引力的目标，那么大数据的规模和吸引力都是最大的。与之前的任何预处理系统相比，Hadoop 能从更多地方收集更多数据，然后将它们结合起来，用更多方法进行分析。Hadoop 正是通过这种方式创造了巨大的价值。

显然，“Hadoop 安全”意义重大。

本书的两位作者在将安全概念引入 Hadoop 平台方面做出过突出贡献。他们在书中阐述了 Hadoop 从早期开放的消费互联网时代到现在作为敏感数据可信平台的演变历程，回顾了 Hadoop 安全的历史，既涵盖了其优点和演进过程，也涉及了新的业务问题。身份验证、加密、密钥管理和商业实践在内的诸多主题及其在实际环境下的应用也在书中有详细讨论。

10 年前，Hadoop 被 Facebook 选作图片存储软件，它发展到今天的过程非常有趣。如今，Hadoop 为数据处理和分析提供了更强的能力、更多的方法、更大的规模和更好的表现。因此，无论分开来看还是整体观之，Hadoop 也需要更多保护。

本书最好的地方在于，不仅仅对问题进行描述，还给出了处理建议。它能清晰并且详细地告诉读者，搭建和使用 Hadoop 的资深开发者们如何安全地管理大数据。本书给读者提供了如何使用——并且安全地使用

——这个先进的平台以分析、处理和理解数据的最佳建议。

——**Mike Olson**

**Cloudera** 公司首席战略官、联合创始人

# 前言

虽然 Apache Hadoop 仍然是一个相对较新的技术，但这并没有限制它迅速地被业界采用，并爆发式地出现了很多相关工具，正是这些工具组成了 Hadoop 广阔的生态系统。对于 Hadoop 用户，这当然是个令人振奋的时代。Hadoop 给公司带来了前所未有的增值机会，同时也给负责数据访问安全和系统合规的技术人员带来了许多挑战。目前，已经有丰富的信息能够帮助使用 Hadoop 构建解决方案的开发者，以及部署运维 Hadoop 的管理员。然而，关于如何设计和实现 Hadoop 安全部署的指导信息依然匮乏。

本书全面而深入地介绍了 Hadoop 的众多安全特性，并使用通用的计算机安全概念进行组织。第 1 章是介绍性内容，随后分为 4 大部分：第一部分是安全架构，第二部分是验证、授权和安全审计，第三部分是数据安全，第四部分是归纳总结。这些部分涵盖了从设计安全架构的前期步骤，到实现通用的安全访问控制和数据保护。最后介绍了几个用例，融合了书中的很多概念。

# 目标读者

本书的目标人群是管理大数据平台安全的 Hadoop 管理员，以及需要在大型企业架构中设计集成 Hadoop 安全规划的安全架构师。书中介绍了很多 Hadoop 安全概念，包括验证、授权、审计、加密和系统架构。

第 1 章是对贯穿全书的一些安全概念的综述，以及对 Hadoop 生态系统的简介。如果刚刚接触 Hadoop，建议你阅读《Hadoop 技术详解》和《Hadoop 权威指南》。本书假设读者熟悉 Linux、计算机网络以及一般的系统架构知识。对于那些在保护分布式系统安全方面没有经验的管理员，第 2 章会给出这部分内容的概述。有经验的安全架构师如果不想回顾这方面知识，可以跳过这一章。总体而言，本书不要求读者具备编程背景，而尽量将内容重点集中于 Hadoop 安全在架构和操作方面的实现。