Alluxio创始人及实践先驱联合力荐

Broadview
www.broadview.com.cn

Alluxio

大数据统一存储原理与实践

范斌　顾荣 / 著

深度解密Alluxio核心概念与技术应用

项目PMC&Maintainer凝聚钻研实力与超前视野

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Alluxio

大数据统一存储原理与实践

范斌　顾荣 / 著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

Alluxio 这一以内存为中心的分布式虚拟文件系统，最初诞生于加州大学伯克利分校的 AMPLab，其开源社区在目前大数据生态系统中发展很快。本书以广泛使用的 Alluxio 1.8.1 版本为基础进行编写，是一本全面介绍 Alluxio 相关技术原理与实践案例的书籍。本书主要内容包括 Alluxio 系统快速入门、Alluxio 系统架构及读写工作机制、Alluxio 与底层存储系统的集成、Alluxio 与上层计算框架的集成、Alluxio 基本功能和高级功能的介绍与使用。此外，本书还详细介绍了 Alluxio 的应用案例与生产实践，并详细解读了 Alluxio 的核心框架和技术应用，旨在为大数据从业人员和大数据存储技术爱好者提供一个深入学习的平台，也可用作开源社区开发者指南。

图书在版编目（CIP）数据

Alluxio：大数据统一存储原理与实践 / 范斌，顾荣著. —北京：电子工业出版社，2019.8

ISBN 978-7-121-36782-3

Ⅰ. ①A… Ⅱ. ①范… ②顾… Ⅲ. ①分布式数据处理 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2019）第 108352 号

责任编辑：张春雨　　特约编辑：田学清
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　　印张：13.75　　字数：242 千字
版　　次：2019 年 8 月第 1 版
印　　次：2019 年 8 月第 1 次印刷
定　　价：79.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

推荐序一

如今的世界步入了一个数据革命的时代。随着互联网、人工智能、移动计算、自动驾驶、物联网等新技术的不断进步，人们生成、采集、管理和分析的数据规模正在呈指数级增长，存储和处理这些大规模数据促使人们不断地实现技术的进步，并为人们带来了难以想象的技术革命的重大机遇。在过去的十年中，我们看到了数据处理的技术栈领域产生了很多重要的技术革新。例如，在数据应用层，从最初的MapReduce 框架，衍生出了很多不同的通用化和专用化的系统，如通用数据处理平台 Apache Spark，流式计算系统 Apache Flink、Apache Samza，深度学习系统 TensorFlow、Apache Mahout，图计算系统 GraphLab、GraphX，查询系统 Presto、Apache Hive、Apache Drill，等等。类似地，整个生态系统的存储层也从 Hadoop 分布式文件系统 HDFS 发展并增加了更多的可选项。例如，文件系统、对象存储（Object Store）系统、二进制大对象存储（BLOB Store）系统、键-值对存储（Key-Value Store）系统、NoSQL 数据库等。这些不同类型的系统实现了对性能、速度、成本、易用性、架构等设计上不同的权衡。

随着技术栈复杂程度的不断增加，数据产业的发展也面临更多的机遇和更大的挑战。数据被存储在不同的存储系统中，这使用户和上层数据应用很难高效地发现、访问和使用这些数据。例如，对于系统开发人员而言，需要开展更多的工作以将一个新的计算或存储部件集成到现有的生态系统中；对于应用开发人员而言，高效地

访问不同数据存储系统的方式变得更加复杂；对于终端用户而言，从远程的数据存储系统中访问数据，容易导致性能的损失和语义的不一致；对于系统管理员而言，当底层物理存储和上层所有应用都深度耦合时，添加、删除、升级一个现有计算系统或数据系统，抑或将数据从一个存储系统迁移到另一个存储系统，是非常具有挑战性的。

Alluxio 作为全球首创的分布式虚拟文件系统（Virtual Distributed File System），就在上述背景下应运而生。它统一了数据访问的方式，为上层计算框架和底层存储系统构建了桥梁，使应用可以通过 Alluxio 提供的统一数据访问方式访问底层任意存储系统中的数据。在大数据生态系统中，Alluxio 位于上层大数据计算框架和底层分布式存储系统之间，运行在上层的大数据计算框架可以忽略底层分布式存储系统的细节，直接和 Alluxio 进行交互，Alluxio 透明地将上层大数据框架的数据访问请求转发到底层分布式存储系统中，并将底层多个分布式存储系统中的数据自动缓存到 Alluxio 中，从而提升某些上层大数据计算框架的数据访问速度的数量级。Alluxio（前身 Tachyon）系统曾是我在加州大学伯克利分校 AMPLab 的博士研究课题，并在 2012 年年末完成了该系统的第一个版本，于 2013 年 4 月正式开源，2015 年项目更名为 Alluxio。

自 2013 年 4 月 Alluxio 开源以来，已有超过 200 个机构、1000 多位贡献者参与到 Alluxio 系统的开发中，其中包括阿里巴巴、百度、卡内基梅隆大学、谷歌、IBM、英特尔、加州大学伯克利分校、腾讯、京东、雅虎等大学、科研院所和企业。到今天为止，上百家公司的生产线中已经部署了 Alluxio，其中有的集群已经超过了 1000 个节点。随着 Alluxio 开源项目的快速发展和应用需求的日益旺盛，我们于 2015 年创立了 Alluxio 公司，并且获得 Andreessen Horowitz、Mark Leslie（Veritas Founding CEO）、Jack Xu（网易、新浪前 CTO）、Sujal Patel（Isilon 创始人）等人的投资。未来，我们将立志于让 Alluxio 成为大数据及其他水平扩展应用的事实上的统一数据层。

我很高兴看到，这本系统、深入介绍 Alluxio 项目技术原理和应用实践的书籍即将付梓。本书的两位作者范斌博士和顾荣博士是分布式系统领域的专家，也是 Alluxio 项目管理委员会的成员和源码的维护者。其中，范斌博士于 2015 年从谷歌

离职之后全身心致力于 Alluxio 开源项目的技术架构、开发与推广，目前在 Alluxio 社区代码贡献排名中排第二位。顾荣博士从 2013 年就开始向 Alluxio 社区贡献源代码，此后他在南京大学 PASA 大数据实验室担任助理教授，继续从事大数据系统方面的研究，在 Alluxio 上开展了很多有意义的研究工作，并且一直努力推动 Alluxio 社区在国内的发展。范斌和顾荣在 Alluxio 社区方面都是非常著名的技术专家，为 Alluxio 开源社区的发展做出了重要贡献。相信他们完成的这本著作能够很好地帮助需要学习 Alluxio 技术的广大读者。最后，我也要特别感谢一直对 Alluxio 开源项目给予关心与支持的朋友们，我们将一如既往地努力投入，在不断完善 Alluxio 软件的同时，让我们开源社区的运转更加高效，期待后续创作出更多高质量的文章和书籍，以飨读者。

李浩源

Alluxio 开源项目主席、Alluxio 公司创始人、董事长兼 CTO

2019 年 4 月，于美国硅谷

推荐序二

The big data revolution is changing how every industry operates. Organizations and companies are leveraging tremendous amounts of data to create value. For example, Internet companies use data to provide better targeted advertisements and user experiences. Financial institutions process data to detect potential fraud in real time. Manufacturing powerhouses study data to track, understand, and design locomotive and airplane engines better. Autonomous cars depend on data to function and to ensure the safety of passengers. People use data to make decisions or facilitate the decision-making process in some way.

The big data revolution has brought a lot of challenges and opportunities in distributed computer systems. There are significant innovations in distributed computation frameworks, such as Hadoop and Spark, and distributed storage systems, such as HDFS and Alluxio. The large-scale data processing stack has been reshaped by the big data ecosystem. In the big data ecosystem, organizations usually rely on multiple storage systems and computation frameworks in their data processing pipelines. This brings the significant challenges in data sharing and management, performance and flexibility.

To address these challenges, the Alluxio project proposes an architecture with Virtual

Distributed File System (VDFS) as a data unification layer between the computing layer and the storage layer. A data unification layer brings significant value into the ecosystem. It can improve data accessibility, performance, and data management, but also the convenience to plug future systems into the ecosystem, therefore making it easier and quicker for the industry to adopt innovations.

Alluxio is an open-source project started at UC Berkeley AMPLab in December 2012. In the over six years of development, this project has grown to be an important part in the big data ecosystem. Alluxio has been deployed at hundreds of leading companies in production, serving critical workloads. Its open-source community has attracted more than 900 contributors worldwide from over 200 companies. I am very glad to see this book to be published. The authors of this book——Bin Fan and Rong Gu are both Alluxio experts. They were also Alluxio topic speakers in the past Strata + Hadoop World conferences. I believe their Alluxio book will be very helpful to the Alluxio users and developers!

Ben Lorica

Chief Data Scientist at O'Reilly Media

Chair of Strata Data, and the Artificial Intelligence Conference

前　言

随着计算机和信息技术的迅猛发展和普及应用，行业数据呈爆炸式增长，全球已经进入了“大数据”时代。大数据给全球带来了重大的发展机遇，大规模数据资源蕴含着巨大的社会价值和商业价值，有效地管理这些数据，挖掘数据的深度价值，对国家治理、社会管理、企业决策和个人生活将带来巨大的影响。然而，大规模数据资源给人们带来新的发展机遇的同时，也带来很多新的技术挑战。格式多样、形态复杂、规模庞大的行业大数据给传统的计算技术带来了很多技术困难。传统的数据库等信息处理技术已经难以有效应对大规模数据的处理需求。大数据广泛且强烈的应用需求极大地推动了大数据技术的快速发展，促进了大数据处理相关的基础理论方法、关键技术及系统平台的长足发展。

大数据处理的第一个基本问题是，如何有效地存储和管理海量的大数据。大数据存储管理是进行后续大数据计算分析和提供大数据应用服务的重要基础。分布式存储是目前公认并有效的大数据存储管理方法，在大数据系统中处于基础地位，在行业大数据应用中发挥着重要的作用。本书将介绍近些年来在大数据存储领域发展得如火如荼的分布式存储系统 Alluxio。Alluxio 是全球首创的以内存为中心（Memory-Centric）的分布式虚拟文件系统，已在全球数百个公司部署应用，并在超过 1000 个节点的集群上运行。

本书以广泛使用的 Alluxio 1.8.1 版本为基础进行编写，全面介绍了 Alluxio 的相

关技术原理与实践案例，以及 Alluxio 的核心原理和架构技术。本书从概念和原理上对 Alluxio 的核心框架和相关技术应用进行了详细解读，并介绍了 Alluxio 技术在互联网公司的使用案例，以及就如何向开源社区贡献源代码进行了简要介绍，具有较好的前沿性和一定的国际视野。

本书目的

Alluxio 项目自 2013 年开源以来得到了长足的发展，贡献者和用户数量不断增多。但是放眼国内，很少有完整、系统地介绍 Alluxio 相关技术使用原理和实践应用案例的书籍。本书的两位作者均为 Alluxio 项目管理委员会成员和源码维护者，在社区的日常工作中经常需要回答很多关于 Alluxio 的技术问题，他们发现用户很多时候苦于没有完整的 Alluxio 中文学习资料。因此，他们决定一起写一本关于 Alluxio 的书籍，为大数据从业人员和大数据存储技术爱好者提供一个深入学习的平台，帮助 Alluxio 的用户能够更加全面和透彻地了解 Alluxio 的基本原理，从而更加容易地使用 Alluxio。

内容快览

全书一共分为 8 章，各章的内容简介如下。

第 1 章 Alluxio 系统快速入门：本章介绍了 Alluxio 项目的背景，包括系统功能简介、项目发展历史；还介绍了 Alluxio 软件的获取或编译方式，以及搭建部署流程。

第 2 章 Alluxio 系统架构及读写工作机制：本章阐述了 Alluxio 的系统架构与功能组件，并介绍了 Alluxio 内部的读数据和写数据的工作运行原理，使读者对 Alluxio 的总体架构和运行流程有一定的认识。

第 3 章 Alluxio 与底层存储系统的集成：本章介绍了 Alluxio 与当前主流的分布式存储系统进行集成的方法，这些底层存储系统具体包括 HDFS、Secure HDFS、AWS S3、Google GCS、Azure BLOB Store。

第 4 章 Alluxio 与上层计算框架的集成：本章首先介绍了 Alluxio 提供给管理员和用户的命令行及其含义，然后阐述了 Alluxio 与主流的上层大数据计算框架进行对接集成的方法。上层计算框架包括 Hadoop MapReduce、Spark、Hive、Presto、TensorFlow。

第 5 章 Alluxio 基本功能的介绍与使用：本章介绍了 Alluxio 提供给用户的基本配置与管理功能，包括 Alluxio 系统环境与属性的配置、Alluxio 底层文件系统的配置与管理、Alluxio 缓存资源的配置与管理，还介绍了 Alluxio 系统 Web 用户界面的查看与使用方法。

第 6 章 Alluxio 高级功能的介绍与使用：本章介绍了 Alluxio 提供给用户的高级功能，具体包括 Alluxio 的安全认证与权限控制、Alluxio 的内置 Metrics 系统、Alluxio 文件系统日志的使用与维护、Alluxio 系统的异常排查。

第 7 章 Alluxio 的应用案例与生产实践：本章阐述了 Alluxio 在陌陌、京东、携程、去哪儿网、百度等大型互联网公司的应用与生产实践案例。

第 8 章 Alluxio 的开源社区开发者指南：本章介绍了源代码的规范、单元测试流程及向 Alluxio 开源社区贡献源代码的具体流程。

写作分工

本书第 1 章、第 5 章、第 6 章、第 8 章由范斌完成，第 2 章、第 3 章、第 4 章由顾荣完成，第 7 章由富羽鹏、陈浩骏、毛宝龙、郭建华、徐磊、刘少山完成。

致谢

能够完成本书需要感谢很多人。首先，我们要衷心地感谢 Alluxio 开源社区的广大贡献者和用户，没有你们的支持就没有 Alluxio 项目的今天，也就没有本书的出版问世。感谢本书第 7 章的来自众多互联网公司的工程师作者，感谢你们在繁忙的工

作之余撰写 Alluxio 在贵公司团队的实践应用案例。感谢为本书撰写序言的李浩源博士和 Ben Lorica，他们在百忙之中阅读了书籍的样稿并提出了很多中肯的建议。感谢南京大学 PASA 大数据实验室黄宜华教授、袁春风教授，以及实验室众多同学对于本书的主编顾荣在 Alluxio 开源项目工作上的认可与大力支持。感谢本书编辑及其他工作人员，你们认真严谨的工作为本书的出版奠定了坚实的基础。最后，感谢我们的家人，整本书籍编写周期较长，感谢你们在背后的默默支持，并且对于我们很多节假日未能陪同给予极大的理解与宽容。

由于作者水平有限，书中的疏漏和不妥之处在所难免，敬请读者批评指正，并将反馈意见发送到邮箱 gurong@nju.edu.cn 或 binfan@alluxio.com，以便我们再版时及时修正错误。

目　录

第 1 章　Alluxio 系统快速入门 1

1.1　Alluxio 背景概述 1

1.1.1　Alluxio 系统功能简介 4

1.1.2　Alluxio 项目发展历史 5

1.2　获取/编译 Alluxio 软件 6

1.2.1　下载预编译的 Alluxio 可执行包 6

1.2.2　编译 Alluxio 源代码 6

1.3　Alluxio 的搭建部署及程序运行 10

1.3.1　单机模式 10

1.3.2　集群模式 13

1.3.3　高可用集群模式 16

第 2 章　Alluxio 系统架构及读写工作机制 22

2.1　Alluxio 的构架简介与基本特征 22

2.1.1　提升远程存储读写性能 23

2.1.2　统一持久化数据访问接口 24

2.1.3　数据的快速复用和共享 26

2.2　Alluxio 的系统功能组件 27

2.2.1 Alluxio Master 组件 27
2.2.2 Alluxio Worker 组件 29
2.2.3 Alluxio Client 组件 30
2.3 Alluxio 读写场景的行为分析 31
2.3.1 Alluxio 的读场景数据流 31
2.3.2 Alluxio 的写场景数据流 37
第 3 章 Alluxio 与底层存储系统的集成 40
3.1 配置 HDFS 作为 Alluxio 底层存储 40
3.1.1 准备步骤与基本配置流程 41
3.1.2 高级参数配置 43
3.1.3 使用 HDFS 在本地运行 Alluxio 44
3.2 配置 Secure HDFS 作为 Alluxio 底层存储 44
3.2.1 准备步骤与基本配置流程 45
3.2.2 使用安全认证模式 HDFS 在本地运行 Alluxio 46
3.3 配置 AWS S3 作为 Alluxio 底层存储 47
3.3.1 准备步骤与基本配置流程 47
3.3.2 高级参数配置 49
3.3.3 使用 S3 在本地运行 Alluxio 51
3.4 配置 Google GCS 作为 Alluxio 底层存储 52
3.4.1 准备步骤与基本配置流程 52
3.4.2 高级参数配置 53
3.4.3 使用 GCS 本地运行 Alluxio 54
3.5 配置 Azure BLOB Store 作为 Alluxio 底层存储系统 55
3.5.1 准备步骤与基本配置流程 55
3.5.2 使用 Azure BLOB Store 本地运行 Alluxio 57
第 4 章 Alluxio 与上层计算框架的集成 58
4.1 Alluxio 的管理员操作命令 58

4.1.1 操作命令列表 59
4.1.2 操作命令示例 59
4.2 Alluxio 的用户操作命令 61
4.2.1 操作命令列表 62
4.2.2 操作命令示例 65
4.3 Alluxio 与 Hadoop 操作命令行的集成 78
4.3.1 前期准备与配置 78
4.3.2 具体使用示例 79
4.4 Alluxio 与 Hadoop MapReduce 的集成 79
4.4.1 前期准备与配置 80
4.4.2 具体使用示例 82
4.5 Alluxio 与 Spark 的集成 83
4.5.1 前期准备与配置 83
4.5.2 使用 Alluxio 作为输入/输出源 85
4.5.3 Alluxio 与 Spark 集成常见问题分析与解决 86
4.6 Alluxio 与 Hive 的集成 89
4.6.1 安装并配置 Hive 环境 89
4.6.2 使用 Alluxio 存储部分 Hive 表 90
4.6.3 使用 Alluxio 作为默认文件系统（存储全部数据） 93
4.6.4 检查 Hive 和 Alluxio 的集成情况（支持 Hive 2.x） 95
4.7 Alluxio 与 Presto 的集成 96
4.7.1 前期准备 96
4.7.2 部署分发 Alluxio 客户端 jar 包 98
4.7.3 Presto 操作命令示例 98
4.8 Alluxio 与 TensorFlow 的集成 100
4.8.1 深度学习面临的数据挑战 100
4.8.2 基于 Alluxio 解决深度学习存储问题的分析 101
4.8.3 安装并配置 Alluxio FUSE 102

4.8.4 TensorFlow 使用 Alluxio FUSE 管理访问数据 103

第 5 章 Alluxio 基本功能的介绍与使用 105

5.1 Alluxio 系统环境与属性的配置 105

5.1.1 Alluxio 系统组件参数的配置 106

5.1.2 Alluxio 客户端组件参数的配置 109

5.1.3 Alluxio 参数配置的相关工具 112

5.2 Alluxio 底层文件系统的配置与管理 113

5.2.1 Alluxio 挂载底层存储 113

5.2.2 Alluxio 与底层存储的元数据一致性保证 116

5.3 Alluxio 缓存资源的配置与管理 120

5.3.1 配置 Alluxio 缓存存储资源 121

5.3.2 Alluxio 缓存数据的载入、驻留及释放 126

5.3.3 配置 Alluxio 缓存数据的生存时间 127

5.4 Alluxio 系统 Web 用户界面的查看与使用 128

5.4.1 Alluxio Master Web 界面介绍 128

5.4.2 Alluxio Worker Web 界面介绍 134

第 6 章 Alluxio 高级功能的介绍与使用 137

6.1 Alluxio 的安全认证与权限控制 137

6.1.1 Alluxio 安全认证模式的介绍 138

6.1.2 Alluxio 访问权限控制的介绍 139

6.1.3 Alluxio 用户模拟功能的介绍 141

6.1.4 Alluxio 审计日志功能的介绍 142

6.2 Alluxio 的内置 Metrics 系统 143

6.3 Alluxio 文件系统日志的使用与维护 145

6.4 Alluxio 系统的异常排查 148

第 7 章 Alluxio 的应用案例与生产实践 152
7.1 陌陌基于 Alluxio 加速 Spark SQL 查询 152
7.1.1 Alluxio 缓存应用背景简介 153
7.1.2 陌陌应用场景结合 Alluxio 的分析 153
7.1.3 基于 Alluxio 的陌陌 Ad Hoc 查询系统架构 155
7.1.4 基于 Alluxio 的查询性能评估与分析 156
7.1.5 陌陌在 Alluxio 实战方面的后续实践 158
7.2 京东基于 Alluxio 和 Presto 构建交互式查询引擎 158
7.2.1 京东大数据平台的业务问题背景 159
7.2.2 JDPresto on Alluxio 架构与特性的介绍 160
7.2.3 JDPresto on Alluxio 的性能评估与分析 161
7.2.4 JDPresto on Alluxio 的应用总结 164
7.3 Alluxio 在携程实时计算平台中的应用与实践 165
7.3.1 携程实时计算的应用背景 165
7.3.2 基于 Alluxio 的跨集群数据共享方案与性能评估 168
7.4 去哪儿网利用 Alluxio 提升异地存储访问性能 169
7.4.1 去哪儿网流式处理背景简介 170
7.4.2 原有系统架构及相关问题分析 171
7.4.3 基于 Alluxio 改进后的系统架构介绍与性能评估 172
7.5 百度基于 Alluxio 加速远程数据读取 176
7.5.1 百度跨机房数据查询问题的描述 177
7.5.2 使用 Alluxio 缓存远端数据的方案与效果 177
7.5.3 使用 Alluxio 分层存储的方案与效果 178
7.5.4 基于 Alluxio 提速远程数据访问的总结 180
第 8 章 Alluxio 的开源社区开发者指南 181
8.1 Alluxio 的源代码规范 181
8.1.1 源代码风格要求 182