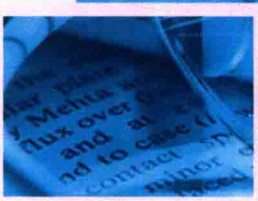
数据馆员的 Hadoop 简明手册



顾立平 袁 慧 **编著**

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

数据馆员的 Hadoop 简明手册

顾立平 袁 慧 编著

科学技术文献出版社
SCIENTIFIC AND TECHNICAL DOCUMENTATION PRESS

·北 京·

图书在版编目(CIP)数据

数据馆员的Hadoop简明手册 / 顾立平, 袁慧编著. —北京：科学技术文献出版社，2017. 10

ISBN 978-7-5189-3013-5

Ⅰ. ①数… Ⅱ. ①顾… ②袁… Ⅲ. ①数据处理软件—技术手册 Ⅳ. ① TP274-62

中国版本图书馆 CIP 数据核字（2017）第 161150 号

数据馆员的Hadoop简明手册

策划编辑：崔灵菲 责任编辑：崔灵菲 责任校对：张吲哚 责任出版：张志平

出 版 者 科学技术文献出版社
地 址 北京市复兴路15号 邮编 100038
编 务 部 （010）58882938，58882087（传真）
发 行 部 （010）58882868，58882874（传真）
邮 购 部 （010）58882873
官方网址 www.stdp.com.cn
发 行 者 科学技术文献出版社发行 全国各地新华书店经销
印 刷 者 虎彩印艺股份有限公司
版 次 2017 年 10 月第 1 版 2017 年 10 月第 1 次印刷
开 本 850 × 1168 1/32
字 数 45千
印 张 2.875
书 号 ISBN 978-7-5189-3013-5
定 价 28.00元

Preface 前言

本手册旨在协助初级数据馆员们能够迅速了解 Hadoop 的知识、用途及整体概貌，作为进一步实践操作之前的入门基础读物。

数据馆员是能够充分实现开放科学政策、措施、服务的一群新型信息管理人员，他们熟悉数据处理、数据分析、数据权益、数据政策，且具有知识产权与开放获取的知识和经验。

Hadoop 是一个开源的框架，它能够使用户在不了解分布式底层细节的情况下，开发分布式程序，以便进行大规模数据集的分布式处理、用于计算机集群进行高速运算，以及面向海量数据的存储。

本手册力求简单、通俗、易懂，既不泛泛之谈，也不过早深入细节，而是力求把握重点。事实上，唯有实践才能真正理解 Hadoop 的有趣之处和局限之处，但在实践之前，或

者考虑选择架构之前，如果有这么一本手册，会容易理解、沟通及评估。

本手册包括 5 个部分。第 1 章概述分布式大数据的基本概念，以及开源软件 Hadoop 的历史、生态体系及主要版本的变化。第 2 章概述核心架构中的计算资源分配、列式计算的工具及索引。第 3 章概述分布式计算的 MapReduce 方案，这也是最为通用的一种方案，能满足海量数据的处理。第 4 章概述如何优化 Hadoop 的案例。最后，附录介绍 Hadoop 家族产品。

本手册旨在将知识模块化，有了整体概述，可以方便读者与其他解决方案进行比较，在实践中遇到问题才能发现需要深入钻研的部分。在掌握全部知识点的基础上，通过搭建、测试、运行、试验之后，读者可以逐步参照其他已有的案例经验和Hadoop 深入源码的著作进行进一步的探索应用。

编著者

2017 年初春于中关村

Contents 目录

>>>>>> 第1章

Hadoop概念

1.1 Hadoop简介

1.1.1 Hadoop是什么

简单而言，Hadoop就是一个开源的框架。这个框架，能够使用户在不了解分布式底层细节的情况下，开发分布式程序。Hadoop能够进行大规模数据集的分布式处理，能够用计算机集群进行高速运算和海量数据的存储。

1.1.2 Hadoop形成的历史

① Hadoop的雏形开始于2002年Apache的Nutch，一个开源Java实现的搜索引擎。它提供了运行自己的搜索引擎所需的全部工具，包括全文搜索和Web爬虫。

② 2003年，Google发表了一篇关于谷歌文件系统（GFS）的技术学术论文。GFS也就是Google File System，是Google公司为了存储海量搜索数据而设计的专用文件系统。

③ 2004 年，Nutch 的创始人 Doug Cutting 基于 Google 的 GFS 开发了分布式文件存储系统（NDFS）。

④ 2004 年，Google 又发表了一篇关于 MapReduce 的技术学术论文。MapReduce 是一种编程模型，用于大规模数据集（大于 1 TB）的并行分析运算。

⑤ 2005 年，Doug Cutting 又基于 MapReduce，在 Nutch 搜索引擎实现了该功能。

⑥ 2006 年，Yahoo 雇用了 Doug Cutting，Doug Cutting 将 NDFS 和 MapReduce 升级命名为 Hadoop。Yahoo 还为 Doug Cutting 建立了一个独立的团队，用来研究和发展 Hadoop。

1.1.3 Hadoop 在云计算和大数据中的地位

目前的大数据不仅用来描述大量的数据，还涵盖了数据处理速度。改进的大数据定义：大数据（Big Data），或称劣绅海量资料，指的是所涉及的资料，规模巨大到无法通过目前主流软件工具在合理时间内获取、管理、处理，并整理成为帮助企业经营决策的信息。

大数据研究领域目前分为四大块：大数据技术、大数据工程、大数据科学和大数据应用。云计算是属于大数据技术的范畴。云计算（Cloud Computing）是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态、易扩展且经常是虚拟化的资源。

云计算可以认为包括以下 3 个层次的服务：基础设施即服务（IaaS）、平台即服务（PaaS）和软件即服务（SaaS）。

现在用的 Hadoop 位于云计算的 PaaS 层。

二者关系为：①云计算属于大数据中的大数据技术范畴。②云计算包含大数据。③云计算和大数据是两个领域。④ Hadoop 位于云计算的 PaaS 层。

1.1.4　Hadoop 与 Google FS 的关系

从 Hadoop 的形成历史看，我们不难发现 HDFS 就是在 GFS 的基础上实现的，所以 HDFS 作为 GFS 的一个最重要的实现，HDFS 的设计目标和 GFS 是高度一致的：在架构、块大小、元数据等的实现上，HDFS 与 GFS 大致一致。而 Hadoop 的核心组件之一就是 HDFS，故 GFS 是 Hadoop 发展的基石之一。

GFS 包括一个 Master 节点（元数据服务器）、多个 Chunkserver（数据服务器）和多个 Client（运行各种应用的客户端）。GFS 的工作就是协调成百上千的服务器为各种应用提供服务。

HDFS 是根据 GFS 论文中的概念模型设计实现的，简化了 GFS 中关于并发写的思路。

分布式文件系统的发展历程如图 1–1 所示。

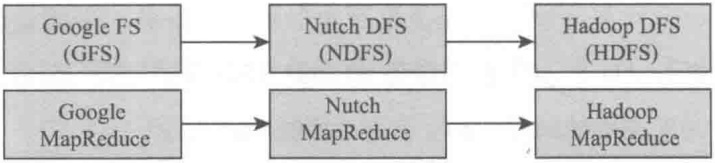


图 1-1 分布式文件系统的发展历程

1.1.5 小结

总体来说，Hadoop 适合应用于大数据存储和大数据分析的应用，适合于几千台到几万台服务器的集群运行，支持 PB 级的存储容量。Hadoop 的典型应用有搜索、日志处理、推荐系统、数据分析、视频图像分析、数据保存等。

1.2 Hadoop 生态系统

1.2.1 Hadoop 组成

2006 年项目成立一开始，“Hadoop”这个单词只代表了两个组件——HDFS 和 MapReduce。现在，这个单词代表的是“核心”项目（即 Core Hadoop 项目）及与之相关的一个不断成长的生态系统。这一点和 Linux 非常类似，都是由一个核心和一个生态系统组成。

Hadoop 的核心是 HDFS 和 MapReduce，Hadoop 2.0 还包括 YARN。

还有一些 HBase、Hive、Pig、ZooKeeper、Ambari、

Hcatalog、Mahout、Avro、Sqoop、Flume、Oozie、Thrift、Storm、Spark 等，都是 Hadoop 上的一些软件或应用。

Hadoop 家族产品（生态系统中的其他产品）见附录。

图 1–2 展示了 Hadoop 生态系统的核心组件。

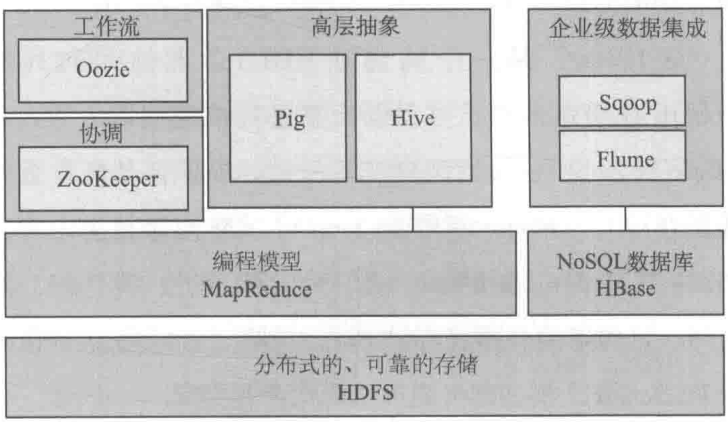


图 1–2　Hadoop 生态系统的核心组件

下面简单介绍各个组件的作用。

① HDFS（Hadoop Distribute File System）：是 Hadoop 生态系统的基础组件 Hadoop 分布式文件系统。它是其他一些工具的基础，HDFS 的机制是将大量数据分布到计算机集群上，数据一次写入，但可以多次读取，用于分析。HDFS 让 Hadoop 可以最大化利用磁盘。

② MapReduce：是 Hadoop 的主要执行框架，用于分

布式并行数据处理编程模型，将作业分为 mapping 阶段和 reduce 阶段。开发人员为 Hadoop 编写 MapReduce 作业，并使用 HDFS 中存储的数据，而 HDFS 可以保证快速的数据访问。鉴于 MapReduce 作业的特性，Hadoop 以并行的方式将处理过程移向数据。MapReduce 使得 Hadoop 可以最大化利用 CPU。

③ HBase：是一个构建在 HDFS 之上的面向列的 NoSQL 数据库，用于对大量数据进行快速读取 / 写入。HBase 将 ZooKeeper 用于自身的管理，以保证其所有组件都正在运行。HBase 使得 Hadoop 可以最大化利用内存。HBase 是 Hadoop Database，是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用 HBase 技术可在廉价 PC Server 上搭建起大规模结构化存储集群。

④ ZooKeeper：是 Hadoop 的分布式协调服务。ZooKeeper 被设计成可以在机器集群上运行，是一个具有高度可用性的服务，用于 Hadoop 操作的管理，而且很多 Hadoop 组件都依赖于它。ZooKeeper 是 Hadoop 的正式子项目，是针对大型分布式系统的可靠协调系统，提供的功能包括：配置维护、名字服务、分布式同步、组服务等。ZooKeeper 的目标就是封装好复杂易出错的关键服务，将简单易用的接口和性能高效、功能稳定的系统提供给用户。

⑤ Oozie：是一个可扩展的 Workflow 系统，已经被集

成到 Hadoop 软件栈中，用于协调多个 MapReduce 作业的执行任务。它能够处理大量的复杂性，基于外部事件来管理执行任务。Oozie 是一个开源的工作流和协作服务引擎，基于 Apache Hadoop 的数据处理任务。Oozie 是可扩展的、可伸缩的、面向数据的服务，运行在 Hadoop 平台上。Oozie 包括一个离线的 Hadoop 处理的工作流解决方案，以及一个查询处理 API。

⑥ Pig：是对 MapReduce 编程复杂性的抽象，Pig 平台包含用于分析 Hadoop 数据集的执行环境和脚本语言 (Pig Latin)。它的编译器将 Pig Latin 翻译为 MapReduce 程序序列。Pig Latin 语言的编译器会把类 SQL 的数据分析请求转换为一系列经过优化处理的 MapReduce 运算。

⑦ Hive：是数据仓库工具，由 Facebook 贡献。Hive 是一个基于 Hadoop 的数据仓库平台。通过 Hive，我们可以方便地进行 ETL 的工作。Hive 定义了一个类似于 SQL 的查询语言：HQL，能够将用户编写的 QL 转化为相应的 MapReduce 程序，类似于 SQL 的高级语言，用于执行对存储在 Hadoop 中数据的查询，Hive 允许不熟悉 MapReduce 的开发人员编写数据查询语句，它将会翻译为 Hadoop 中的 MapReduce 作业。类似于 Pig，Hive 是一个抽象层，适合于较熟悉 SQL 而不是 Java 编程的数据库分析师。

⑧ Sqoop：是一个连通性工具，用于在关系型数据库

和数据仓库 Hadoop 之间移动数据。Sqoop 利用数据库来描述导入 / 导出数据的模式，并使用 MapReduce 实现并行操作和容错。Sqoop 是一个用来将 Hadoop 和关系型数据库中的数据相互转移的工具，可以将一个关系型数据库（如 MySQL、Oracle、Postgres 等）中的数据导入到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导入到关系型数据库中。

⑨ Flume：是一个分布式的、具有可靠性和高可用性的服务，用于从单独的机器上将大量数据高效收集、聚合并移动到 HDFS 中。它基于一个简单灵活的架构，提供流式数据操作。它借助于简单可扩展的数据模型，允许将来自企业中多台机器上的数据移到 Hadoop 中。

⑩ Ambari：是一个集群的安装和管理工具，用来创建、管理、监视 Hadoop 的集群。Ambari 自身也是一个分布式架构的软件，由两部分组成：Ambari Server 和 Ambari Agent。简单来说，用户通过 Ambari Server 通知 Ambari Agent 安装对应的软件；Agent 会定时发送各个机器中每个软件模块的状态给 Ambari Server，最终这些状态信息会呈现在 Ambari 的 GUI 中，方便用户了解到集群的各种状态，并进行相应的维护。

⑪ Mahout：是一个分布式机器学习算法的集合，包括被称为 Taste 的分布式协同过滤的实现、分类、聚类等。Mahout 最大的优点就是基于 Hadoop 实现，把很多以前运

行于单机上的算法，转化为了 MapReduce 模式，这样大大提升了算法可处理的数据量和处理性能。

⑫ YARN：是 Hadoop 2.0 中的资源管理系统，基本设计思想是将 MRv1 中的 JobTracker 拆分成两个独立的服务：一个全局的资源调度器 ResourceManager 和每个应用程序特有的应用程序管理器 ApplicationMaster，该调度器是一个 “纯调度器”，不再参与任何与具体应用程序逻辑相关的工作，而仅根据各个应用程序的资源需求进行分配，资源分配的单位用一个资源抽象概念“Container”来表示，Container 封装了内存和 CPU。此外，调度器是一个可插拔的组件，用户可根据自己的需求设计新的调度器，YARN 自身提供了 Fair Scheduler 和 Capacity Scheduler 这两个调度器。应用程序管理器负责管理整个系统中所有的应用程序，包括应用程序的提交、与调度器协商资源以启动 ApplicationMaster、监控 ApplicationMaster 运行状态并在失败时重新启动等。

⑬ Knox：是一款基于开源 Android 平台的安全解决方案。通过物理手段和软件体系相结合的方式全面增强了安全性，同时完美兼容安卓和谷歌生态系统，为企业及员工个人带来行业领先的企业移动安全解决方案。这个项目就像在 Hadoop 集群中的服务器周围构造一个大的虚拟围栏，对于可用的 Hadoop 服务只有一个安全网关可以进入。