从基础到实践 基础知识全面覆盖 实践操作循序渐进

从理论到应用 理论讲解详尽具体 动手应用实操实练

从入门到进阶 内容编排由浅入深 进阶案例综合指导

重点推荐

Hadoop离线分析实战

聂强 付雯◎主编

北京理工大学出版社
BEIJING INSTITUTE OF TECHNOLOGY PRESS

Hadoop 离线分析实战

主　编　付　雯　聂　强

副主编　李俊翰　武春岭

　　李清莲　童世华

参　编　卢　山　段　科

　　李　茂

北京理工大学出版社

BEIJING INSTITUTE OF TECHNOLOGY PRESS

内容简介

本书包含7个项目，项目一介绍数据管理技术的发展，项目二介绍Hadoop平台的搭建与安装，项目三介绍数据清洗技术，项目四介绍数据仓库Hive的使用，项目五介绍Flume的应用，项目六介绍海量数据传输工具Sqoop，项目七介绍Azkaban调度器。

全书以典型案例贯穿，采用任务驱动方式逐步进行教学设计，结合大赛、职业技能证书展开编写工作，知识点由浅入深、覆盖面广，适合大数据技术相关专业教学使用，同时，对专业爱好者来说也是一本不错的入门级参考资料。

图书在版编目（CIP）数据

Hadoop离线分析实战/聂强，付雯主编. --北京：
北京理工大学出版社，2021.8

ISBN 978-7-5682-9489-8

Ⅰ.①H… Ⅱ.①聂…②付… Ⅲ.①数据处理软件
Ⅳ.①TP274

中国版本图书馆CIP数据核字（2021）第017709号

出版发行／北京理工大学出版社有限责任公司
社　　址／北京市海淀区中关村南大街5号
邮　　编／100081
电　　话／（010）68914775（总编室）
　　（010）82562903（教材售后服务热线）
　　（010）68944723（其他图书服务热线）
网　　址／http://www.bitpress.com.cn
经　　销／全国各地新华书店
印　　刷／唐山富达印务有限公司
开　　本／787毫米×1092毫米　1/16
印　　张／18.5
字　　数／425千字
版　　次／2021年8月第1版　2021年8月第1次印刷
定　　价／79.00元

责任编辑／王玲玲
文案编辑／王玲玲
责任校对／刘亚男
责任印制／施胜娟

Foreword 前言

近年来，大数据、人工智能、云计算等技术的应用，使得智能设备和智能应用普及，改善了人们的生活，加速了科技的发展。随着新技术应用到人们的日常生活中，产生了大量的数据信息，科学、合理地对数据进行管理和利用，能够为人们今后的生活和工作带来更大便利。利用 Hadoop 离线分析技术对数据进行管理，使之能够发挥更大价值，让决策者能够迅速地把握用户的关键需求，并及时响应用户的需求变化。未来几年，智能设备的应用和普及将会更加深入人们的生活，未来的数据分析将对实时性要求越来越高。

针对大数据的迅猛发展，本书结合实际应用案例，通过梳理数据管理技术发展过程，引入了大数据的相关概念，并对 Hadoop 的整体技术生态进行了介绍。

本书以 Hadoop 生态圈技术作为大数据离线分析的工具，采用项目驱动的编写方式，精心设计了 7 个项目引导学习，覆盖了 Hadoop 平台的搭建、安装技术、数据清洗的方法、数据仓库的应用及 Flume、Sqoop、Azkaban 的部署应用。通过简单项目开展项目部署，内容深入浅出。具体为：

项目一从对数据的认识、数据的使用、数据的管理入手，通过案例让读者认识 Hadoop 技术。

项目二详细介绍了 Hadoop 平台搭建的基本知识，并介绍了验证 Hadoop 平台的基本方法。

项目三介绍了 HDFS 分布式文件系统体系架构和文件操作、MapReduce 分布式计算系统基本运行框架、YARN 的平台资源调度方式。

项目四详细介绍了 Hive 的搭建、组件的新增工作，以及如何用 Hive 来存储数据并执行查询分析。

项目五介绍了 Flume 组件的安装及其运行机制，同时介绍了日志采集的流程和方法。

项目六介绍了海量数据传输的原理和方法，以及 Sqoop 组件的安装和数据的导入/导出。

项目七介绍了 Azkaban 调度器的安装部署、数据库的导入及工作流的构建等内容。

本书由重庆电子工程职业学院的聂强、付雯教授主编，重庆电子工程职业学院的武春岭、李俊翰、童世华、李清莲副主编，中山职业技术学院段科、重庆翰海睿智大数据科技有

限公司总经理卢山、四川柒零众邦信息技术有限公司董事长李茂参编，重庆电子工程职业学院江恒也参与了本书的编写工作。

本书中引入大量的具备极高操作性的可运行案例代码，进一步降低了读者的实操难度。

尽管我们尽了最大努力，但书中难免有不妥之处，欢迎各界专家和读者朋友们给予宝贵意见。

编　者

Contents 目录

项目一　认识数据管理 …… 1
项目描述 …… 1
项目分析 …… 1
任务 1　认识数据管理技术 …… 1
任务 2　初识 Hadoop …… 3
项目二　Hadoop 平台的搭建与安装 …… 9
项目描述 …… 9
项目分析 …… 9
任务 1　Hadoop 平台搭建基础 …… 10
任务 2　Hadoop 集群规划 …… 12
任务 3　运行平台搭建 …… 14
任务 4　安装配置支持软件 …… 22
任务 5　安装配置 Hadoop …… 36
任务 6　验证 Hadoop …… 57
项目三　数据清洗技术 …… 68
项目描述 …… 68
项目分析 …… 69
任务 1　HDFS 分布式文件系统的体系架构和文件操作 …… 69
任务 2　MapReduce 分布式计算系统的基本运行框架 …… 82
任务 3　YARN 分布式资源管理平台的资源调度 …… 92
项目四　数据仓库——Hive 的搭建与应用 …… 100
项目描述 …… 100
项目分析 …… 100
任务 1　Hive 搭建的准备工作 …… 101
任务 2　Hive 组件的新增工作 …… 112
任务 3　用 Hive 来存储数据并执行查询分析 …… 122

项目五　Flume 的应用 …… 133
项目描述 …… 133
项目分析 …… 133
任务 1　Flume 组件的安装 …… 133
任务 2　Flume 的运行机制 …… 136
任务 3　Flume 应用案例 …… 160
项目六　Sqoop——海量数据传输工具使用 …… 179
项目描述 …… 179
项目分析 …… 179
任务 1　Sqoop 组件的安装 …… 180
任务 2　Sqoop 的数据导入与导出 …… 187
项目七　Azkaban 调度器 …… 239
项目描述 …… 239
项目分析 …… 239
任务 1　Azkaban 的安装部署 …… 239
任务 2　导入数据库 …… 268
任务 3　验证 Azkaban …… 281
任务 4　构建工作流 …… 282

项目一
认识数据管理

项目描述

数字化时代的到来，使得数据量越来越多，因此需要对这些数据进行有效、科学的管理。数据管理是通过软件技术来实施的特殊操作。数据的多样化使许多行业在数据管理方面的难度加大，这就促使企业在开展数据管理工作时，创新管理方法，从而合理、科学地利用数据进行项目开发和管理。

项目分析

通过本项目的学习，了解数据管理技术及 Hadoop 分布式平台架构技术。

任务1　认识数据管理技术

1. 早期的数据管理技术

美国统计学家赫尔曼·霍尔瑞斯为了统计 1890 年的人口普查数据，发明了一台电动器来读取卡片上的洞数，这台设备使美国用了 1 年时间就完成了原本需要耗时 8 年的人口普查工作。这就是最早所说的数据管理技术，即通过大量的分类、比较和表格绘制的机器运行数百万穿孔卡片来进行数据的处理，其运行结果在纸上打印出来或制成新的穿孔卡片。此时的数据管理就是对穿孔卡片进行物理上的存储和处理。

2. 早期的数据库技术

最早出现的是网状数据库系统，1961 年美国通用电气公司开发出世界上第一个网状数据库系——继承数据存储（Integrated Data Store，IDS），奠定了网状数据库的基础，并在当时得到了广泛的应用。

层次数据库系统是紧随网状数据库系统出现的，最著名的层次数据库是 IBM 公司在 1968 年开发的 IMS（Information Management System）。由于网状数据库模型对于层次和非层

次结构的事务都具备良好的兼容性，因此层次数据库系统不如网状数据库系统使用广泛。

网状数据库和层次数据库虽然已解决了数据集中和数据共享问题，但在数据独立性和数据抽象级别上仍存在较大不足。1970 年，IBM 研究员 E. F. Codd 博士发表了“A Relational Model of Data for Large Shared Data Banks”论文，提出关系模型的概念。关系型模型建立后，IBM 公司在 1979 年实现了 SQL 数据库管理系统。1976 年，霍尼韦尔公司开发了第一个商用关系数据库系统 Multics Relational Data Store（MRDS）。

经过长时间的发展，数据库技术越来越成熟、完善，代表产品有 Oracle 公司的 Oracle、MySQL，IBM 公司的 DB2、Informix，微软公司的 MS SQL Server 等。

3. 决策支持系统和数据仓库

决策支持系统是诞生于 20 世纪 60 年代，辅助决策者通过数据、模型和知识，以人机交互方式进行半结构化或非结构化决策的计算机应用系统，是管理信息系统向更高一级发展而产生的先进信息管理系统。其为决策者提供分析问题、建立模型、模拟决策过程和方案的环境，调用各种信息资源和分析工具，帮助决策者提高决策水平和质量。

1988 年，为解决企业集成问题，IBM 公司的研究员 Barry Devlin 和 Paul Murphy 提出了数据仓库（Data Warehouse）的概念。它是决策支持系统和联机分析应用数据源的结构化数据环境，是一个面向主题的（Subject Oriented）、集成的（Integrated）、相对稳定的（Relative Stable）、反映历史变化（Time Variant）的数据集合，用于支持管理决策（Decision Making Support），服务于数据管理。

4. 数据挖掘和商业智能

数据挖掘指通过分析大量的数据来展示数据之间隐藏的关系、模式和趋势，从而为决策者提供新的知识。这里“挖掘”用来比喻在海量数据中寻找有用知识，很困难，也很难得。

数据挖掘是数据量快速增长的直接产物。自从有了数据仓库，数据挖掘如虎添翼，在实业界不断产生化腐朽为神奇的故事。最脍炙人口的当属啤酒和尿布。Wal - Mart（沃尔玛）拥有世界上最大的数据仓库，在一次购物车分析之后，研究人员发现，跟尿布一起搭配购买最多的商品居然是啤酒。经过大量的跟踪调查，研究人员发现，在美国，一些年轻的父亲经常被妻子“派”到超市去购买婴儿尿布，有 30% ~ 40% 的新生爸爸会顺便买点啤酒犒劳自己。沃尔玛随后对啤酒和尿布进行了捆绑销售，不出意料，销售量双双增加。

1989 年，高德纳 IT 咨询公司（Gartner Group）提出了商业智能的概念。商业智能（Business Intelligent，BI）即一系列以数据为支持、辅助商业决策的技术和方法。

商业智能是利用数据仓库、数据挖掘技术对客户数据进行系统的储存和管理，并通过各种数据统计、分析工具对客户数据进行分析，提供各种分析报告，为企业的各种经营活动提供决策信息。

5. 大数据

1980 年，著名未来学家阿尔文·托夫勒的《第三次浪潮》一书中，将大数据称为“第三

次浪潮的华彩乐章”。但是在相当长的一段时间里，大数据技术并没有得到实质性的发展。

自 1995 年起，随着信息技术的发展，大量信息产生，随之出现了的结构化、非结构化数据，彻底打乱了传统数据管理技术、数据挖掘、商业智能的节奏。传统数据管理技术面临着前所未有的压力和挑战，数据挖掘领域和商业智能领域对于高效处理海量数据的技术需求极为迫切。以 Hadoop 为核心的大数据相关技术此刻飞速发展，大数据技术走向成熟。

2011 年，美国咨询公司麦肯锡发表了研究报告“Big data：The next frontier for innovation，competition，and productivity”，报告首次提出“大数据时代的到来”。

目前，被广泛引用的数据均来自国际数据公司（IDC）发布的研究报告“The Digital Universe of Opportunities：Rich Data and the Increasing Value of the Internet of Things”，该报告称：2013 年数字世界项目统计得出，全球数据总量为 4.4 ZB。同时预测，在 2020 年将达到 44 ZB。

研究机构 Gartner 对大数据给出的定义为：大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

麦肯锡全球研究对大数据所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模（Volume）、快速的数据流转（Velocity）、多样的数据类型（Variety）和低的价值密度（Value）四大特征。

其实，大数据的真正意义在于如何高效地从这些价值密度较低的海量数据中寻找价值，为各行各业的商业决策服务。

任务 2　初识 Hadoop

1.“三驾马车”与 Hadoop

谈到 Hadoop，不得不提谷歌公司的“三驾马车”：Google FS（GFS）、MapReduce、BigTable。

谷歌的“三驾马车”主要为谷歌的核心搜索业务服务。谷歌搜索要存储整个互联网的内容，同时要基于内容构建倒排索引。为了能够大幅提升计算效率并降低硬件成本，谷歌开发了三项技术：

- 谷歌文件系统（GFS），基于大量的、廉价的个人计算机构建分布式的海量的数据存储系统，可轻松存储整个互联网内容。
- 海量数据计算引擎（MapRecude），用来大规模地处理整个互联网的所有文档，建立倒排索引。虽然其有天然的缺陷，但是大幅提升了倒排索引的计算效率。
- 键值存储系统（BigTable），可以存储一个主键为不同时期的多个版本的值，通过使用互联网地址作为主键的方式，可以实现增量更新索引。

虽然谷歌公司没有公布这三个产品的源代码，但是发布了这三个产品的详细设计论文，从而奠定了 Hadoop 的基础。

2004 年，受到 Google 发布的 GFS 和 MapReduce 思想的启发，Doug Cutting 等人用两年的时间实现了 DFS 和 MapReduce 机制，使得 Nutch 性能大幅提升。

2005 年，Hadoop 作为 Lucene 的子项目 Nutch 的一部分正式引入 Apache 基金会。

2006 年，Hadoop 作为一套完整、独立的软件框架被分离，正式命名为 Hadoop，成为 Apache 顶级项目。

2008 年，Hadoop 赢得世界最快的 1 TB 数据排序记录（在 900 个节点上用时 209 s）。

2009 年，雅虎团队使用 Hadoop 对 1 TB 数据进行排序，只花费了 62 s。自此，Hadoop 作为大数据离线分析的开源框架在大数据领域显露峥嵘。

2. 认识 Hadoop

Hadoop 是一套分布式系统基础框架。用户可在不了解底层细节的前提下基于 Hadoop 框架运行分布式程序，而 Hadoop 框架将为分布式程序提供可靠性和数据处理能力。

Hadoop 实现了 MapReduce 编程模型，它能够将分布式应用程序自动地分成小的工作片段，每一个工作片段都可以在集群的任何节点执行。

Hadoop 还提供了一个分布式文件系统（HDFS），用户可以在计算节点存储数据。HDFS 容错性高，能够提供高吞吐量来访问应用程序的数据。

（1）Hadoop MapReduce

MapReduce 是一种编程模型，它通过使用键值对数据集的分布式操作序列实现大型分布式计算。在 Map 阶段，框架将输入数据集分拆为大量的片段，将每个片段分配给各 Map 任务。为运行在 Map 任务的所有集群节点分发大量的 Map 任务，每个 Map 任务都从框架为其分配的数据集片段中获取键值对数据，经过计算后生成新的键值对数据。Map 任务将调用用户定义的 Map 函数完成每一个键值对数据的转化，从而转化为新的键值对数据。

Map 阶段结束后，框架将键值对元组集拆分为与 Reduce 任务数量相同的片段。

在 Reduce 阶段，每个 Reduce 任务都会从框架为其分配的数据集片段中获取键值对元组片段作为输入数据，调用用户定义的 Reduce 函数进行计算，完成将元组转化为输出键值对数据。与 Map 阶段机制相同，Reduce 阶段框架也会为运行 Reduce 任务的所有集群节点分发大量的 Reduce 任务，并将键值对元组数据片段发送给每个 Reduce 任务。

（2）Hadoop DFS

Hadoop DFS（HDFS）组件是一个硬件设备集群，用于存储海量数据文件，并且从机制上能够保证数据存储的可靠性。其设计灵感来自 GFS。HDFS 将大文件以数据块的方式分割存储，除了最后一个数据块外，所有的数据块具有相同的大小。为了实现高容错，同一数据文件的数据块被创建为多个副本进行分布式存储，框架提供了配置文件，便于用户根据实际业务需求配置数据块的大小及副本数量。HDFS 中的文件采用“WORM”（write once read many）机制，数据是只读的，不允许对数据进行修改。

3. Hadoop 技术生态组件

以 MapReduce 和 HDFS 为核心的 Hadoop 软件架构，由于其高可靠性、高效及可伸缩的特点，使 Hadoop 开源社区活跃。围绕着 Hadoop 用于解决特定问题的一系列开源组件不断涌现，构建了一套以 Hadoop 为核心的技术生态。其中，使用最为广泛、最重要的几个组件如下。

（1）YARN

Apache Hadoop YARN（Yet Another Resource Negotiator）是一种全新的 Hadoop 资源管理器，是一个通用资源管理系统和调度平台，可以为层应用提供统一的资源管理和调度，它的引入为集群在利用率、资源统一管理和数据共享等方面带来了巨大好处。

第一代 MapReduce（称为 MRv1）是目前使用的标准大数据处理系统，但是在超大型集群上，其架构缺陷就被暴露出来了。当集群包含的节点超过 4 000 个时，就会表现出一定的不可预测性，其中最大的问题是级联故障，由于框架要尝试复制数据和重载活动的节点，所以任何一个故障都可能通过网络泛化导致整个集群的运行状况严重恶化。

YARN 分层结构的本质是 ResourceManager。它控制整个集群并管理应用程序向基础计算资源的分配。ResourceManager 将各个资源部分（计算、内存、带宽等）精心安排给基础 NodeManager（YARN 的每节点代理）。ResourceManager 还与 ApplicationMaster 一起分配资源，与 NodeManager 一起启动和监视它们的基础应用程序。

随着 YARN（称为 MRv2）的出现，开发者不再受 MapReduce 开发模式约束，而是可以创建更复杂的分布式应用程序。实际上，可以把 MapReduce 模型看成 YARN 架构可运行的一些应用程序中的一个子程序，只是为自定义开发公开了基础框架的更多功能。YARN 的使用模型几乎没有限制，不再需要与一个集群上可能存在的其他更复杂的分布式应用程序框架相隔离。随着 YARN 变得更加健全，它有能力取代其他分布式处理框架，从而完全消除专用于其他框架的资源开销，简化整个系统。

（2）Hive

The Apache Hive 是一个基于 Hadoop 的数据仓库基础架构，提供了类 SQL 查询语句（HQL）完成海量数据的汇总、查询和分析。

Hive 起源于 Facebook。Facebook 每天要产生大量的社交网络数据，通过 MapReduce 编程虽能实现对海量数据的分析，但业务逻辑变更导致的代码变更所产生的成本极大。因此，Facebook 研发并开源了 Hive，用于解决海量结构化日志数据的统计问题。

Hive 能够将存储在 HDFS 上的结构化数据文件映射为一张数据库表，通过简单的类 SQL 查询语句完成对海量数据的数据查询、统计的功能。Hive 可以将类 SQL 语句转化为 MapReduce 作业在分布式集群环境运行，学习成本低，适合基于数据仓库进行统计分析。

（3）ZooKeeper

Apache ZooKeeper 是一个分布式协调服务，用于维护配置信息、命名、提供分布式同步服务和组服务。

ZooKeeper 最早起源于雅虎研究院的一个研究小组。工作人员发现，雅虎内部的很多大型系统都需要依赖类似的系统来进行分布式协调，但这些系统往往都存在单点分布式问题。雅虎的研发人员试图开发一个通用的无单点问题的分布式协调框架，使开发人员能够将精力集中在业务处理逻辑上。

ZooKeeper 的设计目标是将那些杂且易错的分布式一致性服务封装起来，构建一个高效、可靠的原语集，并将一系列简单、易用的接口提供给用户使用。ZooKeeper 是一个典型的分布式数据一致性解决方案，分布式应用程序可以基于 ZooKeeper 实现诸如数据发布/订阅、负载均衡、命名服务、分布式协调/通知、集群管理、Master 选举、分布式锁和分布式队列等功能。

（4）Flume

Apache Flume 是一个高可用的、高可靠的、分布式的海量日志采集、聚合和传输的系统。其是基于流式数据的简单、灵活的框架。同时，它提供简单的可扩展数据模型来支持在线分析应用程序。

Flume 最早的版本是由 Cloudera 发布的，称为 Flume OG，其能够实时地将分布在不同节点、设备上的日志搜集到 HDFS。但随着 Flume 功能的扩展，Flume OG 代码工程臃肿、核心组件设计不合理、核心配置不标准等一系列缺陷逐渐暴露。

2011 年 10 月 22 日，Cloudera 完成了 Flume－782，对 Flume 进行了改动：重构核心组件、核心配置及代码架构，重构后的版本统称为 Flume NG。改版后的 Flume NG 正式纳入 Apache 旗下，改名为 Apache Flume。

Flume 的核心是把数据从数据源（source）收集过来，再将收集到的数据送到指定的目的地（sink）。为保证数据传输的成功率，在送到目的地（sink）之前，会将数据先缓存起来（channel），待数据真正到达目的地（sink）后，Flume 才会删除缓存的数据。

在整个数据传输的过程中，流动的是 event，保证事务是在 event 的级别。event 是 Flume 传输数据的基本单元，如果是文本文件，通常是一条记录，event 也是事务的基本单位。event 从 source 流向 channel，再到 sink，其本身是一个字节数组，可携带头信息（headers）。event 代表着一个数据的最小完整单元，从外部数据源来，向外部目的地去。

Flume 的数据源是可定制的，可以用于传输大量事件数据，包括但不限于网络流量数据、社交媒体生成的数据、电子邮件消息和几乎所有可能的数据源。

（5）Sqoop

Apache Sqoop 是用于在 Hadoop 和结构化存储（例如关系型数据库）之间高效、批量传输数据的工具。用户可使用 Sqoop 将存储在关系型数据库（如 MySQL、MS SQL Server、Oracle 等）中的数据导入到 HDFS、Hive 或 HBase 中，也可以将存储在 HDFS、Hive 或 HBase 中的数据提取并导出到结构化数据库中。

4. Hadoop 发行版

由于 Hadoop 遵从 Apache 开源协议，任何人可以对其进行修改，并作为开源或商业产品发布/销售，因此，市面上出现了很多 Hadoop 版本，有 Intel 发行版、华为发行版、Cloudera 发行版（CDH）、Hortonworks 版本、MapR 的 MapR 产品等。

（1）Apache Hadoop

Apache Hadoop 是 Apache 官方的社区版本，是最原始的版本，所有发行版均是以此为基础进行改进的。

其发行版本主要有三代：

- Hadoop 1. x

  第一代包含三个版本，分别是 0. 20. x、0. 21. x 和 0. 22. x。其中，0. 20. x 最后演化成 1. 0. x，变成了稳定版；而 0. 21. x 和 0. 22. x 则支持 NameNode HA 等新的重大特性。

- Hadoop 2. x

  第二代 Hadoop 包含两个版本，分别是 0. 23. x 和 2. x，它们是一套全新的架构，均包

含 HDFS Federation 和 YARN 两个系统。

- Hadoop 3. x

  Hadoop 2. 0 是基于 JDK 1. 7 开发的，而 JDK 1. 7 在 2015 年 4 月已停止更新，这直接迫使 Hadoop 社区基于 JDK 1. 8 重新发布一个新的 Hadoop 版本，即 Hadoop3. x。Hadoop 3. 0 中引入了一些重要的功能和优化，包括 HDFS 可擦除编码、多 NameNode 支持、MR Native Task 优化、YARN 基于 cgroup 的内存和磁盘 I/O 隔离、YARN container resizing 等。

（2）Cloudera Hadoop

Cloudera Hadoop 是由 Cloudera 维护的 Hadoop 发行版，通常将该发行版称为 CDH（Cloudera Distribution Hadoop）。CDH Hadoop 版本的划分非常清晰，并且经过了严格的测试环节。CDH 包括收费的企业版及完全开源的社区版。

截至目前，CDH 共发行 5 个版本，其中，前三个版本已不再更新，最新的两个版本分别是 CDH4 和 CDH5。CDH4 基于 Hadoop 2. 0，CDH5 则基于 Hadoop 2. 2、Hadoop 2. 3、Hadoop 2. 5、Hadoop 2. 6，并且还在不断地更新。相比于 Apache Hadoop，CDH 发行版在兼容性、安全性和稳定性上有较大的增强。

（3）Hortonworks Data Platform

Hortonworks Data Platform 是由美国大数据公司 Hortonworks 开发的企业级 Hadoop 平台，通常将该发行版称为 HDP。HDP 完全是在开源的环境下设计、开发和构建的，它以 YARN 作为其架构中心，该平台支持一系列处理方法——批处理、交互式处理、实时处理。HDP 的功能包括数据管理、数据访问、数据管制与集成、运营、安全性。

（4）Hadoop 版本的选择方法

由于 Hadoop 版本比较多，很多用户不知怎样选择。实际上，当前 Hadoop 共有三个版本：Hadoop 1. x、Hadoop 2. x 和 Hadoop 3. x。其中，Hadoop 1. 0 由一个分布式文件系统 HDFS 和一个离线计算框架 MapReduce 组成，而 Hadoop 2. 0 则包含一个支持 NameNode 横向扩展的 HDFS、一个资源管理系统 YARN 和一个运行在 YARN 上的离线计算框架MapReduce。Hadoop 3. x 尚且不够成熟和稳定。

当决定是否将某个软件用于开源环境时，通常需要考虑以下几个因素：

①是否为开源软件，即是否免费。

②是否有稳定版，一般官网会有说明。

③是否经实践验证，这个可以通过查看是否有企业应用案例体现。

④是否有强大的社区支持，当遇到问题时，是否能够通过社区、论坛等资源获取解决问题的办法。


学习笔记

项目二
Hadoop 平台的搭建与安装

项目描述

假设需要完成一个大数据离线分析项目，从现在开始，进入 Hadoop 离线分析平台的学习。

项目分析

以“大数据相关技术构建市场招聘需求监控分析系统”为例，对项目进行分析，组建团队，分析项目需求，进行项目设计、开发。这里从项目团队组建开始。

假设将自己的角色定义为一名系统运维组人员，需要通过不断的学习和训练来完成项目组分配给自己的工作。

通常情况下，企业在构建大数据分析项目团队时，会采取两种组织方式：

1. 项目型团队

采用该方式组织的团队包括项目组的全部职能岗位，需要项目经理、系统架构师、数据采集工程师、数据清洗工程师、数据分析工程师、数据可视化工程师及大数据测试工程师等诸多岗位，这些岗位通常统一由项目经理对其进行管理。

2. 职能型团队

采用该方式组织的团队不设项目组，通常会包括一位项目经理及一位系统设计师。由系统设计师根据项目需求对项目整体进行模块化设计与拆分，项目经理按照软件开发工期进行项目安排，而实际的工作内容都是由各职能部门或职能小组完成的。项目组里同样需要系统运维组、数据采集组、数据清洗组、数据分析组、数据可视化开发组及软件测试组分别承担各自职能范围内的工作任务，项目经理不能管理各职能组的内部人员。

本项目以职能型团队的组织方式为例。

经过需求分析工程师、软件设计工程师的工作，已经对系统进行了整体的设计，系统由数据采集子系统、数据存储与分析平台、数据可视化子系统三个子系统构成，其中数据采集子系统、MapReduce、数据分析、数据可视化子系统为定制开发内容，其他组件均采用主流大数据开源组件。系统结构如图 2－1 所示。

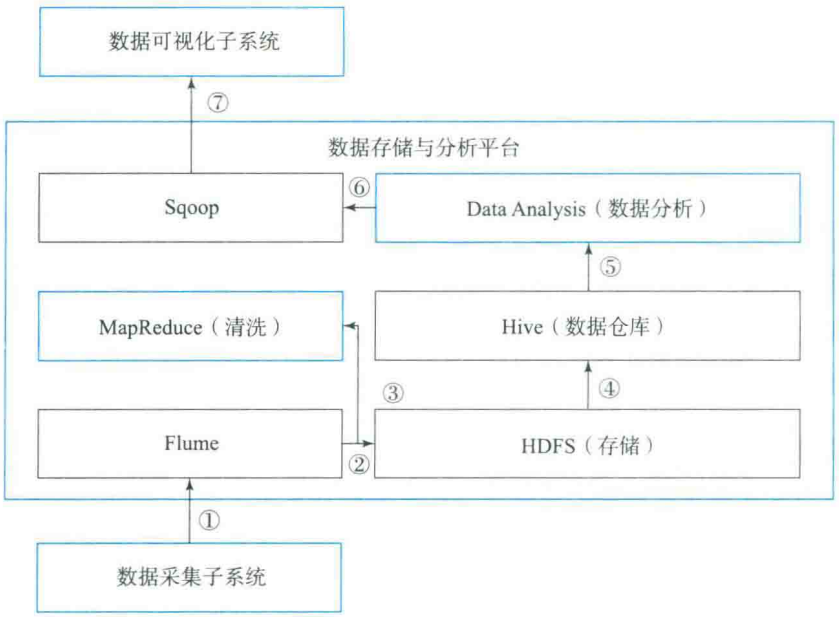

图 2－1　系统结构

作为系统运维组成员，项目经理分配给我们的工作有如下几项：

任务 1：搭建 Hadoop 2.6.0 基础开发环境，为数据清洗提供支撑平台。

任务 2：在开发环境中增加 Hive 1.1.0 组件，为数据分析提供支撑平台。

任务 3：在开发环境中增加 Flume 1.6.0 组件，实现与数据采集的集成。

任务 4：在开发环境中增加 Sqoop 1.4.7 组件，实现与数据可视化的集成。

任务 5：搭建工作流任务调度系统 Azkaban，实现自动化运维。

根据项目经理安排的工作，从第一项工作开始，搭建 Hadoop 2.6.0 基础开发环境，为数据清洗提供支撑平台。

任务 1　Hadoop 平台搭建基础

完整的 Apache Hadoop 2.x 发行版包含以下四个组件：

- Hadoop Common：为其他 Hadoop 组件提供基础配置。