

大数据

技术及应用探究

DASHUJU JISHU JI YINGYONG TANJIU

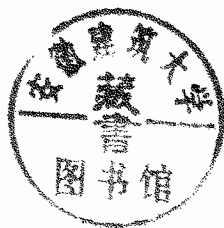
胡沛 韩璞 著



大数据

技术及应用探究

胡沛 韩璞 著



电子科技大学出版社

University of Electronic Science and Technology of China Press

图书在版编目(CIP)数据

大数据技术及应用探究/胡沛,韩璞著.--成都:
电子科技大学出版社,2018.1
ISBN 978-7-5647-5450-1

I.①大… II.①胡…②韩… III.①数据处理-研究 IV.①TP274

中国版本图书馆CIP数据核字(2018)第008984号

大数据技术及应用探究

胡 沛 韩 璞 著

策划编辑 李述娜

责任编辑 李述娜

出版发行 电子科技大学出版社

成都市一环路东一段159号电子信息产业大厦九楼 邮编 610051

主 页 www.uestp.com.cn

服务电话 028-83203399

邮购电话 028-83201495

印 刷 北京一鑫印务有限责任公司

成品尺寸 170mm×240mm

印 张 18.25

字 数 352千字

版 次 2018年8月第一版

印 次 2018年8月第一次印刷

书 号 ISBN 978-7-5647-5450-1

定 价 68.00元

版权所有,侵权必究

随着网络信息技术的不断发展和数据的爆炸式增长，人们进入到大数据时代。大数据已经从方方面面影响到人们的日常生活，但我们也要冷静地看到，大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是简单的将软硬件堆砌起来。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。

目前常见的大数据处理平台有很多，比如：Hadoop，Spark 和 Storm。由于篇幅限制，本书将会对 Hadoop 和 Spark 做出深入的研究和探讨。本书第一章深入探究了大数据的概念，产生的背景以及发展状况，应用案例，并针对大数据的历史机遇提出了自己的一些看法。

从第二章到第六章，通过系统研究 Hadoop 生态系统和基于 Hadoop2.X 版本架构，能够使读者对 Hadoop 有一个清晰直观的认识和了解；同时又探究 HDFS 和 MapReduce Java API 编程，使读者不仅能了解 Hadoop 相关原理还能掌握其编程思想。

Scala 是学习 Spark 的基础，因此在第七章介绍了 Scala 编程开发。Scala 作为一个面向对象函数式编程语言，如果读者具备 C/C++ 和 Java 编程基础的话，相信能够很快掌握；而如果没有，则只需要多花点时间即可掌握。Spark 是一个基于内存的分布式计算框架，目前被阿里巴巴，腾讯，亚马逊等很多公司使用。Spark 为什么能够后来居上，超越 MapReduce 成为最流行的分布式处理框架，通过第八章的学习，您将会知道其中的奥秘。

本书不仅适合计算机相关专业学生的大数据基础学习，也能够满足想继续深入了解 Hadoop 架构编程的学习者。本书在撰写过程中，参考了大量国内外的教材，专著和博客，作者再次表示感谢。本书的第二、三、四、五、六章由南阳理工学院的胡沛老师撰写，共计约 20 万字；第一、七、八章由南阳理工学院的韩璞老师撰写，共计约 15 万字。由于时间仓促，编者水平有限，本书难免存在不足之处，在此出版之际，笔者真诚地希望读者对本书提出宝贵的意见和建议。

1	大数据概论	001
1.1	大数据概述	001
1.2	Hadoop 生态系统	006
1.3	大数据处理工具	008
1.4	大数据与云计算的关系	013
1.5	分布式与集群	014
2	大数据平台 Hadoop	016
2.1	Hadoop 架构	016
2.2	Hadoop1.x 与 Hadoop2.x 的区别	018
2.3	Hadoop 平台搭建	018
3	分布式文件系统 HDFS	030
3.1	HDFS 架构	030
3.2	HDFS 读写过程解析	037
3.3	HDFS shell 命令	040
3.4	HDFS Java API 编程	042
4	分布式计算模型 MapReduce	057
4.1	MapReduce 编程模型	057
4.2	MapReduce 工作原理	060
4.3	Yarn	063
4.4	MapReduce 可编程组件	066
4.5	MapReduce Java API 编程	068

5	分布式数据库 HBase	095
5.1	HBase 介绍	095
5.2	HBase 数据模型	097
5.3	HBase 系统架构	099
5.4	HBase 数据读写流程	101
5.5	HRegion 的 split 和 compact	102
5.6	HBase 环境搭建	103
5.7	HBase shell	108
5.8	HBase Java API	112
6	分布式数据库仓库 Hive	136
6.1	Hive 简介	136
6.2	Hive 与 HBase, 关系型数据库的对比	138
6.3	Hive 的数据管理	139
6.4	Hive 的体系架构	140
6.5	Hive 的执行流程	142
6.6	HQL	142
6.7	Hive 环境搭建	145
6.8	Hive Java API	152
6.9	Hive 综合案例	157
7	Scala 编程基础	165
7.1	Scala 简介	165
7.2	Scala 安装	167
7.2	Scala 语法	168
7.3	Scala 程序控制结构	179
7.4	Scala 容器	183
7.5	cala 函数	198
7.6	面向对象编程	207

1 大数据概论



随着计算机技术全面融入人们的日常生活，数据已经积累到了一个由量变引起质变的程度，它不仅充斥在人们生活的方方面面，其累积速度也以指数级增加，最终出现了大数据这个用语。大数据几乎应用到了人们发展的所有领域中，不管是云计算、物联网，还是社交网络、移动互联网等都会与大数据扯上关系。大数据已经成为有特别含义的专用词汇，不再单指数据体量大。那么，什么是大数据，大数据发展的现状如何，大数据能给人们带来什么？本章将会详细介绍大数据产生的背景、发展历程、应用案例、特点和技术路线等。

1.1 大数据概述

随着网络和信息技术的不断普及，带动了互联网、物联网、电子商务、现代物流、网络金融等现代服务业的发展，也催生了车联网、新能源、智能城市、高端装备制造等新兴产业的发展。随着各行业信息化程度的提高，各种业务数据正以几何级数的形式爆发，其收集、储存、格式、检索、分析、应用等存在诸多问题，不能再以传统的信息处理技术加以解决。

资料显示，2011年全球数据规模为1.8ZB，预计到2020年全球数据将达到40ZB。我国2010年新存储的数据为250PB，而2012年中国的数据存储量达到364EB，约为日本的60%，北美的7%。大量新数据源的出现则导致了非结构化、半结构化数据爆炸式的增长，信息数据的单位也由TB跨越到ZB的级别。使用传统方法早已无法处理这些数据，如何管理和使用这些数据，逐渐成为一个新的领域，于是大数据的概念应运而生。

目前对大数据的准确定义尚有一些争论，这就导致大数据的定义有多种。维基百科给出的定义是：大数据是利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。美国国家科学基金会(NSF)则将大数据定义为“由科学仪器、传

感设备、互联网交易、电子邮件、音视频软件、网络点击流等多种数据源生成的大规模、多元化、复杂、长期的分布式数据集”。全球知名的咨询公司麦肯锡认为：大数据是指无法在一定时间内用传统数据库软件工具对其内容进行采集、存储、管理和分析的数据集合，但它同时指出“大数据”并非总是说有数百个TB才算得上。根据实际使用情况，有时候数百个GB的数据也可称为大数据，这主要要看它的第三个维度，也就是速度或者时间维度。IT权威研究与顾问咨询公司Gartner则将大数据定义为“在一个或多个维度上超出传统信息技术的处理能力的极端信息管理和处理问题”。

从2009年开始，大数据成为互联网信息技术行业的流行词汇，大数据起初成熟应用多在互联网行业，互联网上的数据每年增长50%，每两年翻一番，全球互联网企业都意识到大数据时代的来临，数据对于企业有着重要意义。期刊《Nature》和《Science》针对大数据分别出版了专刊“Big Data”和“Dealing with Data”，从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多个方面讨论了大数据处理面临的各种问题。2011年5月，麦肯锡全球研究院发布题为《大数据：创新、竞争和生产力的下一个新领域》的报告。报告发布后，大数据迅速成为计算机行业的热门概念。2012年4月美国奥巴马政府推出大数据研究和发展倡议，并划拨2亿美元的专项资金之后，在全球掀起了一股大数据的热潮。

我国的大数据应用仍处于起步阶段。在工信部发布的物联网“十二五”规划中，把信息处理技术作为4项关键技术创新工程之一提出，其中包括了海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分。2013年，科技部正式启动863项目“向大数据的先进存储结构及关键技术”，启动5个大数据课题。

1.1.1 大数据产生的背景

近年来互联网、云计算、移动和物联网的迅猛发展和无所不在的移动设备、RFID、无线传感器，使得数据充斥着整个世界。与此同时，数据也成为一种新的资源，亟待人们对其加以合理、高效的利用，使之能够给人们带来更大的效益和价值。在科学研究（天文学、生物学、高能物理等）、物联网、互联网、移动设备等领域，数据量呈现快速增长的趋势。

1. 科学研究产生大数据

随着各种科研仪器的快速发展，人们对科学研究领域进行数据采集的能力和技术手段获得了极大提高，由此产生了海量的复杂数据。在这样的环境下，科研人员如何分析数据，发现新的科学规律，都需要科学研究方法随之发生变化。

2. 物联网的应用产生大数据

目前，物联网在智能工业、智能农业、智能交通、智能电网、节能建筑、安全监

控等行业都有应用，物联网设备通常每隔几秒就收集一次数据，从而催生了大数据。以深圳为例：平均每公里分布 510 辆机动车，全城拥有约 1.5 万个大社区、5 500 个大公司、几百个大商圈、日均产生约 7 亿条交通数据。

3. 互联网的发展产生大数据

为了进行有效的市场营销、推广和改善用户体验，用户在网上的每个点击及其事件都被记录下来；利用这些数据，服务提供商可以对用户存取模式进行仔细的分析，从而提供更加具有针对性的服务，这些日志数据通常极其巨大。淘宝网站每天有超过数千万笔交易，单日数据产生量超过 50TB，存储量 40PB。百度公司目前数据总量接近 1 000PB，存储网页数量接近 1 万亿页，每天大约要处理 60 亿次搜索请求，几十 PB 数据。Facebook 注册用户超过 8.5 亿，每天上传 3 亿张照片，每天生成 300TB 日志数据。新浪微博每天有数十亿的外部网页和 API 接口访问需求，每分钟都会发出数万条微博。

4. 移动设备产生大数据

移动互联时代，数以百亿计的机器、企业、个人随时随地都会获取和产生新的数据。比如智能汽车、智能电视、工业设备和手持设备等都将接入网络，由此产生的数据量及其增长速度比以往任何时期都要多。

1.1.2 大数据发展历程

大数据从产生到目前风靡全球，大致共经历了以下四个发展阶段：

1. 上世纪末

大数据的萌芽期。随着人工智能、数据挖掘理论和关系型数据库技术的成熟，一些商业智能工具和知识管理技术开始被应用。

2. 2003 年 -2005 年

大数据发展的突破期。微博和社交网络的兴起导致大量非结构化数据出现，传统数据处理模型难以应对，人们开始对数据处理系统、数据库架构进行重新设计。

3. 2005 年 -2009 年

大数据发展的成熟期。随着云计算相关技术的成熟，特别是 Hadoop 项目和相关大数据工具的诞生，大数据进入到新的发展阶段。

4. 2011 年至今

大数据时代来临。2012 年维克托·舍恩伯格《大数据时代：生活、工作与思维的大变革》宣传推广，大数据概念开始风靡全球；2012 年 3 月，美国奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，这一倡议标志着大数据已经成为重要的时代特征。

1.1.3 大数据的特征

目前关于大数据的特征还具有一定的争议，本书采用普遍被接受的 4V：规模性 (volume)，多样性 (variety)，价值密度 (value) 和高速度 (velocity) 进行描述。

1. 数据量大 (volume)

非结构化数据的超大规模和增长，导致数据集合的规模不断扩大，数据单位已从 GB 到 TB 再到 PB 级，甚至开始以 EB 和 ZB 来计数。

2. 类型繁多 (variety)

大数据的类型不仅包括网络日志、音频、视频、图片、地理位置信息等结构化数据，还包括半结构化数据甚至是非结构化数据，具有异构性和多样性的特点。

3. 价值密度低 (value)

大数据本身存在较大的潜在价值，但由于大数据的数据量过大，其价值往往呈现稀疏性的特点。虽然单位数据的价值密度在不断降低，但是数据的整体价值在提高。

4. 速度快时效高 (velocity)

要求大数据的处理速度快，时效性高，需要实时分析而非批量式分析，数据的输入、处理和分析连贯性地处理。

1.1.4 大数据应用案例

将大量的原始数据汇集在一起，通过数据挖掘等技术分析数据中潜在的规律，预测未来的发展趋势，有助于人们做出正确的决策，从而提高各个领域的运行效率，取得更大的收益。大数据冲击着许多行业，包括金融行业、互联网、医疗行业、社交网络、零售行业和电子商务等，大数据在彻底地改变着人们的生活。

1. 金融行业

大数据在金融行业应用范围较广，典型的案例有花旗银行利用 IBM 超级计算机系统 Watson 通过银行客户信息分析客户下一步需求，进而向消费者出售金融产品。阿里金融基于阿里巴巴在 B2C 多年来的建树，提出了基于互联网的大数据，通过数据分析，以自主服务模式为主，是纯互联网的小额信贷服务。

2. 互联网

大数据技术可以将用户在互联网上的行为记录下来，对用户的行为进行分析，打上标签并进行用户画像。特别是进入移动互联网时代之后，用户主要的访问方式转向了智能手机和平台电脑，移动互联网的数据包含了个人的位置信息，其 360 度用户画像更加接近真实人群。360 度用户画像可以帮助企业进行精准营销，提高广告的转化率。Google 等大型互联网搜索引擎通过对广告相关数据的批量处理，用来改善广告的

投放效果以提高用户的点击量。

3. 医疗行业

医疗行业拥有大量病例、病理报告、医疗方案、药物报告等。通过对患者以往的生活方式与医疗记录进行批量处理分析,提供语义分析服务,对病人的健康提供医生、护士及其他相关人士的回答,并协助医生更好地为患者进行诊断。

4. 社交网络

Facebook、Twitter、新浪微博、微信等以人为核心的社交网络产生了大量的文本、图片、音视频等不同形式的数据库。对这些数据库进行分析,发现人与人之间隐含的关系或者他们中存在的社区,推荐朋友或者相关的主题,提升用户的体验。

5. 零售行业

零售行业可以通过客户购买记录,了解客户关联产品购买喜好,将相关的产品放到一起增加来增加产品销售额。零售行业比较有名气的大数据案例是沃尔玛的啤酒和尿布的故事,以及 Target 通过向年轻女孩寄送尿布广告而告知其父亲女孩怀孕的故事。

6. 电子商务

电子商务中产生大量的购买历史记录、商品评论、商品网页的访问次数和驻留时间等数据,通过分析这些数据,商家可以精准地选择其热卖商品,从而提升商品销量;这些数据还能够分析出用户的消费行为,为用户推荐相关商品,提升优质用户数量。电商行业的巨头天猫和京东,已经通过用户的购买习惯,将用户日常需要的商品例如尿不湿,卫生纸,衣服等商品依据用户购买习惯事先进行准备。当用户刚刚下单,商品就会在 24 小时内或者 30 分钟内送到客户门口,提高了用户体验。

1.1.5 大数据的机遇与挑战

随着近年来大数据的不断升温,人们也逐渐意识到大数据中提到的数据是全部数据,而不是随机采样;预测是大体方向,而不是精确制导。随着对大数据研究的不断深入,人们越来越意识到对数据的利用可以为其生产生活带来巨大便利的同时,也带来了不小的挑战。

1. 大数据的安全与隐私问题

在互联网上浏览网页,就会留下一连串的浏览痕迹;注册登录网站需要输入个人的重要信息,例如用户名、登录密码、手机号,甚至还有身份证号、住址、银行卡等信息。通过相关的数据分析,就可以轻易挖掘出人们的行为习惯和个人重要信息。如果这些信息运用得当,可以帮助相关领域的企业随时了解客户的需求和习惯,便于企业调整相应的产品生产计划,取得更大的经济效益;但若是这些重要的信息被不良分子窃取,随之而来的就是个人信息、财产等安全性问题。

2. 大数据的计算模型

由于大数据中超过 85% 的数据都是半结构化和非结构化的数据，因此对数据存储、分析处理和计算提出了极高的要求。以 Hadoop 为代表的大数据平台因其具有良好的简单、易扩展等能力而在大数据分析领域得到了广泛应用，现已成为大数据处理的主流技术。尽管这样，Hadoop 在计算性能方面依然不能尽如人意，还需根据实际应用情况不断研发更高效、更实用的大数据分析技术。

3. 领域融合问题

数据的产生和应用领域已经不局限于某几个特殊的场合，几乎所有的领域如政治、经济、社会、科学、法律等都能看到大数据的身影，涉及这些领域的数据交叉问题就不可避免，而当前大数据的研究局面是各个学科都以自己为主处理本领域的海量数据。目前大数据研究多侧重于大数据的获取、存储、处理、挖掘和信息安全等方面，鲜有从宏观的角度探讨大数据对人们的生活习惯、行为规范和企业的生产管理、商务运营带来的冲击和变革。

1.2 Hadoop 生态系统

Hadoop 是一个能够对大量数据进行分布式处理的软件框架，具有可靠、高效、可伸缩的特点。Hadoop2.0 版本引入了 HA（高可用）与 Yarn（资源调度），这是与 Hadoop1.0 的最大区别。Hadoop2.0 主要由 3 部分组成：HDFS 分布式文件存储，MapReduce 编程模型和 Yarn 资源管理。

由图 1-1 可以看出，整个 Hadoop 家族由以下几个子项目组成：

1. HDFS

HDFS 是 Hadoop 平台的分布式文件储存系统，HDFS 集群包含了一个 NameNode（主节点），这个节点负责管理所有文件系统的元数据及存储了真实数据的 DataNode（数据节点，可以有多个）。HDFS 具有高容错性、适合大数据批处理、可构建在廉价机器上等优点，缺点是不支持低延迟数据访问、小文件存取、并发写入、文件随机修改。

2. MapReduce

MapReduce 是一个计算模型，用于大规模数据集的并行运算。它极大地方便了编程人员在不会分布式并行编程的情况下，将自己的程序运行在分布式系统上。

3. ZooKeeper

ZooKeeper 是一个分布式应用程序协调服务，是 Hadoop 和 HBase 的重要组件。

它是一个为分布式应用提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务等。

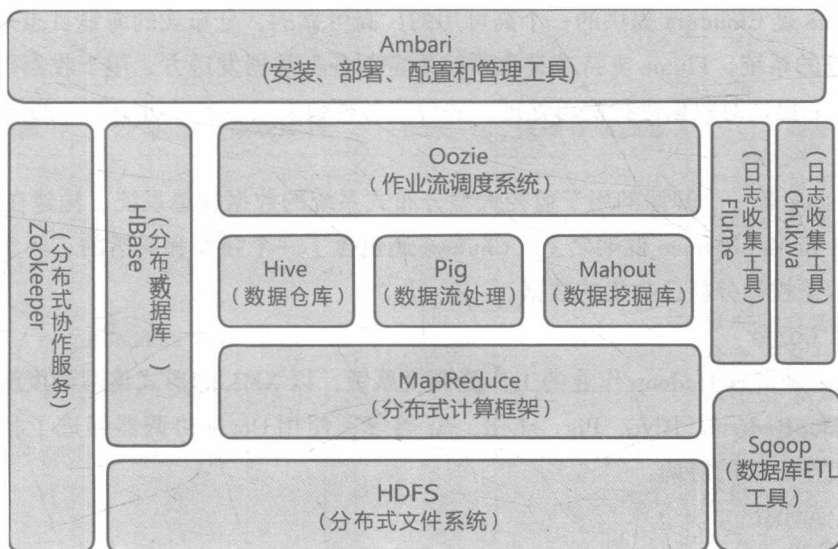


图 1-1 Hadoop 生态系统图

4. HBase

HBase 是一个高可靠性、高性能、面向列、可伸缩的分布式存数据库，利用 HBase 技术可在廉价机器上搭建起大规模结构化存储集群。

5. Hive

Hive 是基于 Hadoop 的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能，可以将 SQL 语句转换为 MapReduce 任务运行。

6. Pig

Pig 是一个高级过程语言，适合于使用 Hadoop 和 MapReduce 平台来查询大型半结构化数据集。通过允许对分布式数据集进行类似 SQL 的查询，Pig 可以简化 Hadoop 的使用，目前在实际的生产中已使用的较少。

7. Mahout

Mahout 是个可扩展的机器学习和数据挖掘库，旨在帮助开发人员更加方便快捷地创建智能应用程序。Mahout 包含许多实现，包括聚类、分类、推荐过滤、频繁子项挖掘。此外，通过使用 Apache Hadoop 库，Mahout 可以有效地扩展到云计算。

8. Sqoop

Sqoop 是一个用来进行数据相互转移的工具，可以将一个关系型数据库中的数据

导入到 Hadoop 的 HDFS 或 HBase，也可以将 HDFS 或 HBase 中的数据导入关系型数据库。

9. Flume

Flume 是 Cloudera 提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume 支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume 提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

10. Chukwa

Chukwa 是一个开源的用于监控大型分布式系统的数据收集系统，构建在 Hadoop 的 HDFS 和 MapReduce 框架之上。Chukwa 还包含了一个强大和灵活的工具集，可用于展示、监控和分析已收集的数据。

11. Oozie

Oozie 是管理 Hadoop 作业的工作流调度系统，以 XML 的形式编写调度流程，可以调度 MapReduce, Hive, Pig, shell, jar 等等，使用 Oozie 协调器促进了相互依赖的重复工作之间的协调。

12. Ambari

Ambari 是一种基于 Web 的工具，用于创建、管理、监视 Hadoop 的集群，支持 Hadoop HDFS、Hadoop MapReduce、Hive、HCatalog、HBase、ZooKeeper、Oozie、Pig 和 Sqoop 等的集中管理。

1.3 大数据处理工具

任何完整的大数据平台，一般包括以下的几个过程：数据采集、数据存储、数据计算和数据展现。其中，数据分析是整个大数据处理流程的核心。在这一层所处理的是从异构数据源抽取和集成的数据，也称为分析的原始数据。根据应用的不同需求，可以从中取舍，利用全部或部分数据进行分析，大数据的价值正是产生于此处。

1.3.1 数据采集

数据采集是大数据价值挖掘重要的一环，其后的数据分析和挖掘都建立在采集的基础上。大数据技术的意义不在于其掌握的规模庞大的数据信息，而在于对这些数据进行智能处理，从中分析和挖掘出有价值的信息。

数据的采集有基于网络信息的数据采集，比如交通摄像头的视频图像采集；也有基于互联网的数据采集，如对各类网络媒介的各种页面信息和用户访问信息进行采

集。在互联网数据采集中每天会产生大量的日志（一般为流式数据，如，搜索引擎的pv，查询等），处理这些日志需要特定的日志系统，目前比较有名的开源日志采集系统有 Flume、Scribe 和 Kafka 等，表 1-1 是对其简单的描述。

表 1-1 开源日志采集系统

采集系统	公司	实现语言	框架	数据存储位置	特点
Flume	Cloudera	Java	push/push	HDFS	支持在日志系统中定制各类数据发送方，用于收集数据。当节点出现故障时，日志能够被传送到其他节点上而不会丢失。采用 ZooKeeper 管理，可以有多个 maser。可扩展性好。
Scribe	Facebook	C/C++	push/push	HDFS 或本地文件	分布式收集，统一处理。当节点出现故障时，日志能够被传送到其他节点上而不会丢失。可扩展性好。
Kafka	LinkedIn /Apache	Scala	push/pull	HDFS	高吞吐量的分布式发布订阅消息系统。支持多订阅者，当失败时能自动平衡消费者。支持 online 和 offline 的场景。可扩展性好。

1.3.2 数据存储

由于大数据的存储量极大，因此其存储设备需要具有高扩展性、高可用、自动容错和低成本等特点。常见的存储形式有分布式文件系统和分布式数据库。分布式文件系统采用大规模的分布式存储节点来满足存储大量文件的需求，而分布式的非关系型数据库则为大规模非结构化数据的处理和分析提供支持。

常见的分布式文件系统有 GFS、HDFS、Lustre、Ceph、FastDFS 和 MogileFS 等, 具体描述见表 1-2, 而非关系型数据库主要有 Redis、Tokyo Cabinet、MongoDB、CouchDB、Cassandra、Voldemort 和 HBase, 表 1-3 是对其简单介绍。

表 1-2 分布式文件系统介绍

文件系统	公司	实现语言	特点
GFS	Google	C++	基于 Linux 的专有分布式文件系统, 尚未开源, HDFS、Ceph 和 Lustre 都是基于 GFS 的文件系统。
HDFS	Apache	Java	HDFS 是目前最流行的一个分布式文件系统, 具有高容错, 适合部署在廉价的机器上。
Lustre	Oracle	C	Lustre 主要用于大型计算机集群和超级电脑, 不支持副本。
Ceph	Sage Weil	C++	Ceph 具有高扩展性, 数据分布均衡, 并行化程度高, 没有单点故障的特点; 目前是 Linux 未来版本的备选文件系统, 但是并没有商业部署的经历。
FastDFS	阿里巴巴	C	轻量级分布式文件系统, 解决了大容量存储和负载均衡的问题, 特别适合以文件为载体的在线服务, 如相册网站、视频网站等。
MogileFS	Six Apart	C/C++	一套高效的文件自动备份组件, 通过 fuse 支持了标准的 posix。

表 1-3 非关系型数据库介绍

	公司	实现语言	特点
Redis	VMware/Pivotal	C	数据都是缓存在内存中, 但会周期性的把更新的数据写入磁盘或者把修改操作写入追加的记录文件, 并且在此基础上实现了 master-slave(主从)同步。
Tokyo Cabinet	Mikio Hirabayashi	C	数据库由一系列 key-value 对的记录构成, 其数据文件只有一个, 所有操作都是依据 key 做主键操作, 支持 master-slave 架构。