普通高等院校数据科学与大数据技术专业“十三五”规划教材

大数据采集与预处理技术

DASHUJU CAIJI YU YUCHULI JISHU

刘丽敏 廖志芳 周 韵 ⊙ 编著

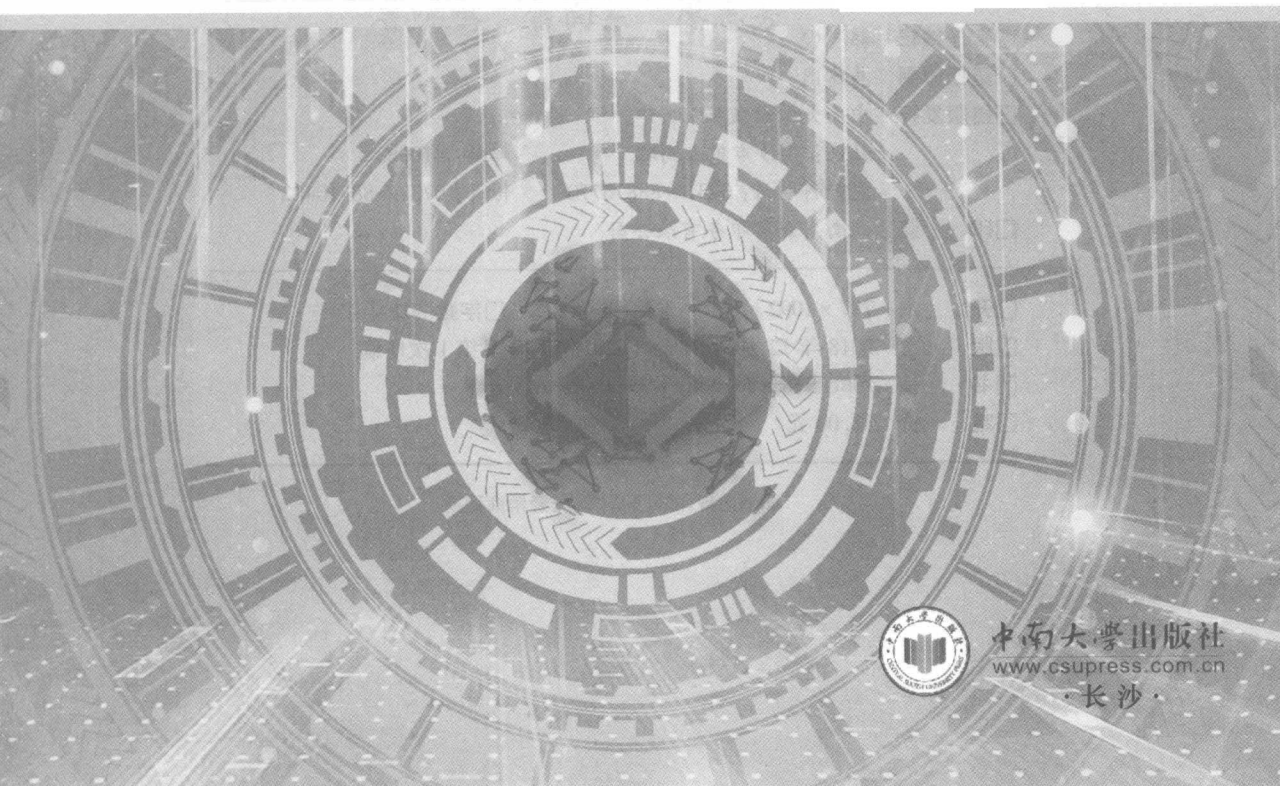
中南大学出版社
www.csupress.com.cn

普通高等院校数据科学与大数据技术专业“十三五”规划教材

大数据采集与预处理技术

DASHUJU CAIJI YU YUCHULI JISHU

刘丽敏 廖志芳 周 韵 ⊙ 编著



中南大学出版社
www.csupress.com.cn
·长沙·

图书在版编目（CIP）数据

大数据采集与预处理技术 / 刘丽敏，廖志芳，周韵编著. --长沙：中南大学出版社，2018.12

ISBN 978-7-5487-3411-6

Ⅰ.①大… Ⅱ.①刘… ②廖… ③周… Ⅲ.①数据采集 ②数据处理 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2018)第 213376 号

大数据采集与预处理技术

刘丽敏　廖志芳　周 韵　编著

□责任编辑　韩　雪
□责任印制　易建国
□出版发行　中南大学出版社
社址：长沙市麓山南路　　邮编：410083
发行科电话：0731-88876770　　传真：0731-88710482
□印　　装　湖南省众鑫印务有限公司

□开　　本　787×1092　1/16　□印张 11.75　□字数 301 千字
□版　　次　2018 年 12 月第 1 版　□2018 年 12 月第 1 次印刷
□书　　号　ISBN 978-7-5487-3411-6
□定　　价　32.00 元

普通高等院校数据科学与大数据技术专业“十三五”规划教材

编委会

内容简介

本书以大数据关键技术为主线，重点介绍了大数据采集技术和数据预处理技术。本书共7章。第1章为大数据概述，重点阐述了大数据的概念、大数据关键技术以及大数据采集和数据预处理的重要性，并对本书内容进行了概述；第2章在阐述传统数据采集相关技术基础上，从数据发展出发，剖析了大数据采集的特点和相关技术；第3章介绍了常用的大数据采集架构；第4章介绍了针对系统数据来源复杂、数据量大的企业数据的大数据迁移技术；第5章介绍了互联网数据抓取与处理技术；第6章介绍了数据预处理技术，包括数据清洗、数据集成、数据变换和数据归约等技术；第7章首先阐述了Hadoop相关理论基础，然后以淘宝网数据为例，介绍了大数据从数据采集、数据预处理、数据分析以及数据可视化的综合应用实例。本书可作为高等院校大数据相关专业的教学用书，也可以作为从事大数据相关工作的工程技术人员的参考用书。

总序

Preface

随着移动互联网的兴起，全球数据呈爆炸性增长，目前90%以上的数据是近年产生的，数据规模大约每两年翻一番；而随着人工智能下物联网生态圈的形成，数据的采集、存储及分析处理、融合共享等技术需求都能得到响应，各行各业都在体验大数据带来的革命，“大数据时代”真正来临。这是一个产生大数据的时代，更是需要大数据力量的时代。

大数据具有体量巨大、速度极快、类型众多、价值巨大的特点，对数据从产生、分析到利用提出了前所未有的新要求。高等教育只有转变观念，更新方法与手段，寻求变革与突破，才能在大数据与人工智能的信息大潮面前立于不败之地。据预测，中国近年来大数据相关人才缺口达200万人，全世界相关人才缺口更超过1000万人之多。我国教育部门为了响应社会发展需要，率先于2016年开始正式开设“数据科学与大数据技术”本科专业及“大数据技术与应用”专科专业，近几年，全国形成了申报与建设大数据相关专业的热潮。随着专业建设的深入，大家发现一个共同的难题：没有成系列的大数据相关教材。

中南大学作为首批申报大数据专业的学校，2015年在我校计算机科学与技术专业设立大数据方向时，信息科学与工程学院院领导便意识到系列教材缺失的严重问题，因此院领导规划由课程团队在教学的同时积累素材，形成面向大数据专业知识体系与能力体系、老师自己愿意用、同学觉得买得值、关联性强的系列教材。经过两年的准备，针对2017年《教育部办公厅关于推荐新工科研究与实践项目的通知》的精神，中南大学出版社组织对系列教材文稿进行相应的打磨，最终于2018年底出版“高等院校数据科学与大数据技术专业‘十三五’规划教材”。

该套系列教材具有如下特点：

1. 本套教材主要参照“数据科学与大数据技术”本科专业的培养方案，综合考虑专业的来源，如从计算机类专业、数学统计类专业以及经济类专业发展而来；同时适当兼顾了专科类偏向实际应用的特点。

2. 注重理论联系实际，注重能力培养。该系列教材中既有理论教材也有配套的实践教程。力图通过理论或原理教学、案例教学、课堂讨论、课程实验与实训实习等多个环节，训练学生掌握知识、运用知识分析并解决实际问题的能力，以满足学生今后就业或科研的需求；同时兼顾“全国工程教育专业认证”对学生基本能力的培养要求与复杂问题求解能力的

要求。

3. 在规范教材编写体例的同时，注重写作风格的灵活性。本套系列教材中每本书的内容都由教学目的、本章小结、思考题或练习题、实验要求等组成。每本教材都配有 PPT 电子教案及相关的电子资源，如实验要求及 DEMO、配套的实验资源管理与服务平台等。本套系列教材的文本层次分明、逻辑性强、概念清晰、图文并茂、表达准确、可读性强，同时相关配套电子资源与教材的相关性强，形成了新媒体式的立体型系列教材。

4. 响应了教育部“新工科”研究与实践项目的要求。本套教材从专业导论课开始设立相关的实验环节，作为知识主线与技术主线把相关课程串接起来，力争让学生尽早具有培养自己动手能力的意识、综合利用各种技术与平台的能力。同时为了避免新技术发展太快、教材纸质文字内容容易过时的问题，在相关技术及平台的叙述与实践中，融合了网络电子资源容易更新的特点，使新技术保持时效性。

5. 本套丛书配有丰富的多媒体教学资源，将扩展知识、习题解析思路等内容做成二维码放在书中，丰富了教材内容，增强了教学互动，增加了学生的学习积极性与主动性。

本套丛书吸纳了数据科学与大数据技术教育工作者多年的教学与科研成果，凝聚了作者们的辛勤劳动，同时也得到了中南大学等院校领导和专家的大力支持。我相信本套教材的出版，对我国数据科学与大数据技术专业本科、专科教学质量的提高将有很好的促进作用。

桂卫华

2018 年 11 月

前言

Foreword

目前全国各类高校、高职院校都已陆续开设了大数据相关的专业和课程。作为交叉型学科，大数据的相关专业强调培养具有多学科交叉能力的大数据人才。因此，作为承担大数据人才培养的高等院校，需要及时建立健全大数据专业的课程体系，培养一批具备大数据专业素养的高级人才，满足社会对大数据人才的需求。而高质量的教材是推动高等院校大数据专业课程体系建设的关键。

《大数据采集与预处理技术》一书侧重于介绍大数据关键技术中的大数据采集和数据预处理技术，该教材可作为入门级大数据基础教材，旨在为读者搭建起大数据的知识架构，阐述大数据采集和数据预处理的基本原理，开展相关的初级实践，为读者在大数据以及相关领域的学习奠定重要的基础。

教材系统论述了大数据的概念和关键技术、大数据采集基础知识、常用大数据采集架构、大数据迁移技术、互联网数据的抓取与处理技术、数据预处理等技术，最后给出了基于Hadoop的大数据综合分析案例。本书共7章，第1章介绍大数据的相关概念和大数据关键技术，帮助读者了解大数据的整体架构，形成对大数据关键技术的总体认识。第2章介绍传统数据采集技术和大数据采集技术，帮助读者形成大数据采集技术的初步认识，为后面章节的学习奠定基础。第3章介绍常用大数据采集架构，包括Chukwa、Flume、Scrible以及Apache Kafka等，帮助读者了解针对日志系统的采集方法。第4章介绍大数据迁移技术，包括基于存储的数据迁移、基于主机逻辑卷的迁移、基于数据库的迁移和服务器虚拟化数据迁移等，帮助读者了解针对系统数据来源复杂、数据量大的企业数据的采集方法。第5章介绍互联网数据抓取与处理技术，并开展了初步的实践，帮助读者了解互联网数据的爬虫技术以及分词技术。第6章介绍数据预处理技术，包括数据的描述以及数据清洗、数据集成、数据变换和数据归约等相关算法和技术。第7章首先介绍了Hadoop相关理论，然后开展了基于Hadoop的大数据分析案例，帮助读者了解大数据分析的全过程，建立起大数据从数据采集—数据预处理—数据分析—数据可视化的分析架构，加深对大数据相关知识的理解。

本书是中南大学大数据系列丛书之一，可以作为数据科学与大数据专业、大数据技术与应用专业、计算机科学等相关专业的教材，也可供从事大数据技术的相关技术人员参考使用。

本书由刘丽敏、廖志芳、周韵编写。其中，第1、2、3、6章由刘丽敏编写，第4章由廖志芳编写，第5章由廖志芳和刘丽敏合作编写，第7章由刘丽敏和周韵合作编写。全书由刘丽敏统稿。在撰写过程中，中南大学软件学院硕士研究生周杰、周亚辉等人做了大量辅助性工作，在此，向这些同学的辛勤工作表示衷心的感谢。

由于编者能力有限，加之编写时间仓促，书中难免存在不足之处，敬请批评指正。

编者

2018年10月

目 录

Contents

第1章　大数据概述

随着云计算、透明计算和物联网等技术的兴起以及社会网络、移动支付、基于位置的服务(location based service, LBS)等应用的迅速发展，数据正以前所未有的速度在不断地增长和累积，大数据时代已经来到。

最早提出“大数据”时代到来的是全球知名咨询公司麦肯锡。麦肯锡曾说：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。”随着互联网和信息行业的发展，大数据在各个领域越来越彰显它的优势，各种利用大数据进行发展的领域正在协助企业不断地发展新业务。本章内容旨在让大家更好地认识和了解大数据。

1.1　大数据的概念

1965 年，英特尔创始人戈登摩尔提出了著名的“摩尔定律”，即集成电路上可容纳的晶体管数目，约每隔 18 个月便会增加一倍。1998 年，图灵奖获得者杰姆格雷提出了著名的“新摩尔定律”，即人类有史以来的数据总量，每过 18 个月就会翻一番。而根据国际数据调研机构 IDC 的估测，数据一直都在以每年 50% 的速度增长，也就是说每两年就增长一倍。这意味着人类在最近两年产生的数据量相当于之前产生的全部数据量。Intel 预测到 2020 年，全球数据量将会达到 35 ZB(1 ZB = 10 亿 TB = 1 万亿 GB)，而中国产生的数据量将会达到 8 ZB，也就是说 2020 年之后中国产生的数据量将约占到全球数据总量的的四分之一。

互联网技术的发展，尤其是社会网络的兴起，让每个人都是互联网信息的接受者，也是信息的产生者，每个人都成为数据源，几乎每个人都在用智能终端拍照、拍视频、发微博、发微信等。另外，物联网技术的发展，遍布地球各个角落的各种各样的传感器，无一不是数据来源或者承载的方式。此外，各行各业越来越依赖大数据手段来开展工作，例如，医疗行业有大量的病例、病理报告、治愈方案、药物报告等，通过对这些数据进行整理和分析，将会极大地辅助医生提出治疗方案，帮助病人早日康复。可以构建大数据平台来收集不同病理和治疗方案，以及病人的基本特征，建立针对疾病特点的数据库，帮助医生进行疾病诊断。所以随着科学技术的发展和人类活动的进一步扩展，数据规模会急剧膨胀，于是大数据这样一个在含义上趋近于无穷大的概念应运而生。

近些年来，大数据一直是学术界的研究热点。*Nature* 最早在 2008 年提出了大数据的概念，并推出了 *Big Data* 专刊。*Science* 在 2011 年推出了大数据专刊 *Dealing with Data*，主要围绕着科学研究中大数据的问题展开讨论，说明大数据对于科学研究的重要性。2012 年，美国

一些知名的数据管理领域的专家学者联合发布了一份白皮书“Challenges and Opportunities with Big Data”，该白皮书从学术的角度出发，介绍了大数据的产生，分析了大数据的处理流程，并提出大数据所面临的若干挑战。2012 年，达沃斯世界经济论坛上，大数据是会议铁主题之一，该次会议还特别针对大数据发布了报告“Big Data，Big Impact：New Possibilities for International Development”，探讨了在新的数据产生方式下，如何更好地利用数据来产生良好的社会效益。该报告重点关注了个人产生的移动数据与其他数据的融合与利用。2012 年 3 月，奥巴马政府在白宫网站发布了《大数据研究和发展倡议》，将其视为“未来的新石油”，提出通过大数据加速在科学、工程领域的创新步伐，强化美国国土安全，转变教育和学习模式。如何利用数据资源发掘知识、提升效益、促进创新，使其服务于国家治理、企业决策乃至个人生活服务，是大数据时代的重要战略课题。奥巴马政府的这一计划被视为美国政府继信息高速公路(information highway)计划之后在信息科学领域的又一重大举措。

中国信息通信研究院在 2016 年发布的《大数据白皮书(2016)》从大数据产业发展概述、大数据技术发展趋势、大数据资源开放与共享、重点行业大数据应用、大数据政策法规等方面分析了大数据行业的最新进展。

那么，何谓大数据？《著云台》的分析师团队认为，大数据(big data)通常用来形容一个公司创造的大量非结构化和半结构化数据，这些数据下载到关系型数据库用于分析时会花费过多时间和金钱。刘鹏在《大数据》一书中给出了大数据的定义，大数据又称巨量数据，指的是无法在可承受的时间范围内用常规软件工具进行捕捉，管理和处理的数据集合，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来适应海量高增长率和多样化的信息资产。

由此可见，大数据是大小超出典型数据库软件采集、储存、管理和分析等能力的数据集合。因此大数据并非大量数据简单无意义的堆积，通常需要从大数据中获取更多有价值的“新”信息，所以必然要求这些大量的数据之间存在着或远或近、或直接或间接的关联性，才具有相当的分析挖掘价值。数据间是否具有结构性和关联性，是“大数据”与“大规模数据”的重要差别。大数据技术是从各种各样类型的大数据中，快速获得有价值信息的技术及其集成。“大数据”与“大规模数据”“海量数据”等类似概念间的最大区别，就在于“大数据”这一概念中包含着对数据对象的处理行为。为了能够完成这一行为，从大数据对象中快速挖掘更多有价值的信息，使大数据“活起来”，就需要综合运用灵活的、多学科的方法，包括数据采集、分布式处理、数据挖掘等，而这就需要拥有对各类技术、各类软硬件的集成应用能力。可见，大数据技术是使大数据中所蕴含的价值得以发掘和展现的重要工具。

目前工业界普遍认为大数据具有 5V + 1C 的特征：大量(volume)、多样(variety)、价值(value)、高速(velocity)、准确性(veracity)和复杂(complexity)。

(1)大量：数据量大，包括采集、存储和计算的量都非常大。大数据的起始计量单位至少是 P(1000 个 T)、E(100 万个 T)或 Z(10 亿个 T)。

(2)多样：数据种类和来源多样化。包括结构化、半结构化和非结构化数据，具体表现为网络日志、音频、视频、图片、地理位置信息等。多类型的数据对数据的处理能力提出了更高的要求。

(3)价值：数据价值密度相对较低。随着互联网以及物联网的广泛应用，信息感知无处不在，信息海量，但价值密度较低，如何结合业务逻辑并通过强大的机器算法来挖掘数据价

值，是大数据时代最需要解决的问题。

(4)高速：数据增长速度快，处理速度也快，时效性要求高。比如搜索引擎要求几分钟前的新闻能够被用户查询到，个性化推荐算法尽可能要求实时完成推荐。这是大数据区别于传统数据挖掘的显著特征。

(5)准确性：即数据处理结果要保证一定的准确性和可信赖度，即数据的质量。

(6)复杂：由于数据大量、多样、产生速度快，对数据的处理和分析的难度大。

从大数据的特征可以看出，大数据的产生和用户使用需求呈指数级增长，数量极其庞大；大量非结构化的数据使得数据复杂度提高，传统的数据处理方式已经无法来处理；数据处理的时效与数据分析结果的准确性难度非常大。可见，大数据技术涉及数据采集、数据存储、数据处理、数据分析等各个方面，运用传统的数据处理工具和技术，无法满足实时大数据的需求。

1.2　大数据关键技术

大数据正带来一场信息社会的变革，大量的结构化数据和非结构化数据的广泛应用，致使人们需要重新思考已有的数据模式，庞大的数据需要进行清洗、预处理、归类、建模、分析等操作，才能得到需要的服务和产品。因此，大数据关键技术涵盖了数据采集、数据存储、数据处理、数据分析、数据应用等多方面技术。如图 1-1 所示，根据大数据的处理过程，可将其分为数据采集、数据预处理、数据存储、数据分析与挖掘以及数据可视化等环节。由于大数据具有规模大、异构、多源等特点，大数据技术与传统的数据处理技术也有所不同，对于每个处理环节，都体现出大数据需求的新技术。

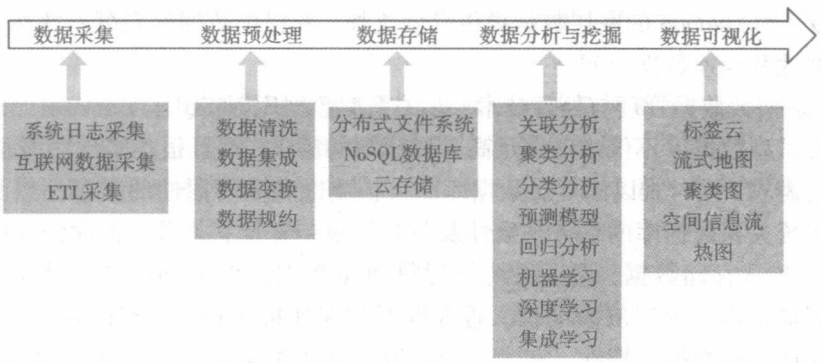


图 1-1　大数据处理过程

1. 大数据采集

对于大数据分析来说，获取大数据是重要的基础。数据采集，又称数据获取，是处于大数据生命周期的第一个环节，它是通过 RFID 射频数据、传感器数据、社交网络、移动互联网等方式获取各种类型的结构化、半结构化和非结构化的海量数据。由于可能存在成千上万的用户并发的访问和操作，因此，必须采用专门针对大数据的采集方法。

2. 大数据预处理

虽然数据采集端本身有很多数据库，但是如果要对这些海量数据进行有效的分析，还是应该将这些数据导入到一个集中的大型分布式数据库或者分布式存储集群当中，同时在导入的基础上完成数据清洗和预处理工作。

现实世界中，数据通常存在不完整、不一致的“脏”数据，无法直接进行数据挖掘，或挖掘结果差强人意，因此，为了提高数据挖掘的质量产生了数据预处理技术。

3. 大数据存储

大数据存储与管理要用存储器把采集到的大规模数据存储起来，建立相应的数据库，并进行管理和调用。针对大数据时代的复杂结构化数据，特别是半结构化数据和非结构化数据的海量存储和分布式存储的需求，大数据存储通常采用分布式文件系统、NoSQL 数据库以及云存储等技术。

(1)高效低成本的大数据文件存储技术：分布式文件系统(distributed file system，DFS)。

DFS 是指文件系统管理的物理存储资源不一定直接连接在本地节点上，而通过计算机网络与节点相连。使用分布式文件系统可以轻松定位和管理网络中共享资源、使用统一的命名路径完成对所需资源的访问。例如，Google 文件系统是一个可扩展的分布式文件系统，用于大型的、分布式的、对大量数据进行访问的应用。它运行于廉价的普通硬件上，将服务器故障视为正常现象，通过软件的方式自动容错，在保证系统可靠性和可用性的同时，大大降低了系统成本。除了 Google 的文件系统，Hadoop 也是一种常用的分布式系统架构，它可以让用户在不了解分布式底层细节的情况下开发分布式程序，充分利用集群的高速运算和存储。Hadoop 的分布式文件系统称为 HDFS(Hadoop distributed file system)。HDFS 不仅有着高容错性的特点，而且为 Hadoop 的底层数据提供存储支撑，并提供数据的高可靠性和容错能力，拥有良好的扩展性和高速数据访问性。

(2)非关系型大数据管理与处理技术：非关系型数据库 NoSQL。

传统的关系型数据库不仅难以满足高并发读写的需求，而且很难实现对海量数据高效率存储和访问的需求，同时难以满足对数据库高可扩展性和高可用性的需求。然而，NoSQL 数据库打破了传统关系数据库的事务一致性及范式约束，放弃了关系数据库强大的 SQL，采用 < key，value > 格式存储数据，保证系统能提供海量数据存储的同时具备优良的查询性能。NoSQL 数据存储不需要固定的表结构，通常也不存在连接操作，具有模式自由、备份简易、接口简单和支持海量数据等特性。在大数据存取上具备关系型数据库无法比拟的性能优势。例如，Google 设计的分布式数据库 BigTable，它为应用程序提供了比单纯的文件系统更方便、更高层的数据操作能力。BigTable 提供了一定粒度的结构化数据操作能力，解决一些大型媒体数据(Web 文档、图片等)的结构化存储问题。BigTable 的设计目的是可靠地处理 PB 级别的数据，并且能够部署到上千台机器。

DynamoDB 是 Amazon 提供的共享式数据库云服务，这种方法可用性和扩展性都很好，读写访问中 99.9% 的响应时间都在 30 ms 内。DynamoDB 通过服务器把所有的数据存储在固态硬盘上的三个不同区域。如果有更高的传输需求，DynamoDB 也可以在后台添加更多的服务器。

Hbase(Hadoop database)是 Hadoop 项目的子项目，它是一个分布式的、面向列的开源数