

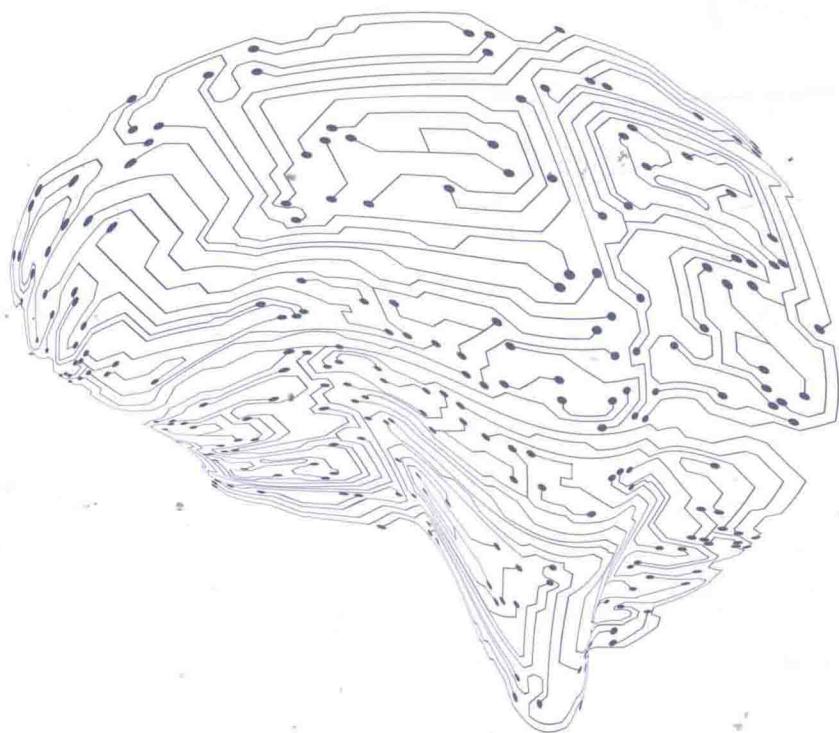
# 架构大数据

——大数据技术及算法解析

# BIG DATA

ARCHITECTING BIG DATA  
——TECHNOLOGIES AND ALGORITHMS EXPLAINED

赵 勇◎编著



# 架构大数据

## ——大数据技术及算法解析

赵 勇 编著

电子工业出版社

Publishing House of Electronics Industry

北京·BEIJING

## 内 容 简 介

本书从大数据架构的角度全面解析大数据技术及算法,探讨大数据的发展和趋势。不仅对大数据相关技术及算法做了系统性的分析和描述,梳理了大数据的技术分类,如基础架构支持、大数据采集、大数据存储、大数据处理、大数据展示及交互,还融合了大数据行业的最新技术进展和大型互联网公司的大数据架构实践,努力为读者提供一个大数据的全景画卷。

本书可作为大数据技术入门和进阶的专业书籍,同时也可作为高等院校大数据相关课程的教材和教学参考用书。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。  
版权所有,侵权必究。

### 图书在版编目(CIP)数据

架构大数据:大数据技术及算法解析/赵勇编著. —北京:电子工业出版社,2015.6  
ISBN 978-7-121-25978-4

I. ①架… II. ①赵… III. ①计算机网络—数据处理—算法分析 IV. ①TP274

中国版本图书馆CIP数据核字(2015)第089418号



责任编辑:董亚峰 特约编辑:王 纲

印 刷:北京京科印刷有限公司

装 订:北京京科印刷有限公司

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:787×1092 1/16 印张:27.75 字数:530千字

版 次:2015年6月第1次印刷

印 次:2015年6月第1次印刷

定 价:68.00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。

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

服务热线:(010)88258888。

## 编 委



---

李有福 陈 尧 张 迎  
刘春雷 杨凡超 袁双双  
孟庆祥 王心杰 查 敏

# 前言



大数据被称为新时代的黄金和石油，相关技术发展迅猛，所应用的行业也非常广泛，从传统行业如医疗、教育、金融、旅游，到新兴产业如电商、计算广告、可穿戴设备、机器人等。大数据技术更是国家科技发展和智慧城市建设的基石。当前“互联网+”新业态的发展，其核心也是大数据的采集、分析、价值挖潜和应用。

当今全球大数据的竞争及战略布局，突出体现在大数据的技术创新和人才培养方面。技术创新能够保障在大数据发展的浪潮中始终处于引领地位，而大数据人才体系则是能最终实现技术创新和实践的根本。政府正在设立大数据局等管理和创新机构并开放政府数据，很多大型企业开始设立首席数据官（CDO）的职位，凸显对数据驱动的经济和业务模式的重视。未来5年全球将有数十万名的大数据人才缺口，以及数百万名大数据的管理和支持人员缺口，因此美国的哈佛商业评论把“数据科学家”称为21世纪最性感的职业，全球都在开始上演大数据的人才争夺战。

我们在《大数据革命——理论、模式与技术创新》一书中对数据科学理论、大数据创新模式以及大数据技术作了初步的探索 and 介绍，在成书的过程中我觉得如果有一本更深入、更全面的技术类的大数据书籍，能够更有效地帮助读者掌握和运用大数据，这也是写这本书的初衷。由于大数据技术是当今最前沿的技术，所涉及的领域及知识跨度都非常大，发展又日新月异，所以有些内容在开始写的同时，就面临老化的危险。过去一两年的时间里大数据技术的更新令人咋舌，老牌的 Hadoop 大数据处理平台已经成熟并广泛应用在诸多行业，而很多新兴的系统和算法，如加州大学伯克利分校研究开发的 Spark 大数据通用计算平台、用于图像和视频识别的深度学习大有后来居上的势头。因

此我们也是尽量跟上形势，努力提供给读者一个大数据技术的全景画卷。

本书对大数据相关的技术及算法做了系统性的分析和描述，梳理了大数据的技术分类如基础架构支持、大数据采集、大数据存储、大数据处理、大数据展示及交互，列举了资源管理调度、数据分析和挖掘、深度学习、精准营销、社会计算等大数据相关算法，还融合了大数据行业的最新技术进展和大型互联网公司的大数据架构实践，目的是为了帮助大数据产业及技术人才全面认识和了解大数据的相关技术及算法，掌握大数据行业的最新发展动态，学习互联网巨头的大数据架构实践，进而培养大数据的技术选型和系统架构能力，成为大数据时代的创新创业先锋。

最后感谢电子科技大学极限网络计算与服务实验室的老师和同学们为本书所付出的辛勤和努力，也感谢清华大学苏州研究院大数据处理中心的行业应用实践为本书提供的指导和建议。我之前在微软的同事沈寓实和李雨航两位专家在云计算体系及安全防护方面提供了详细的资料和建议。我想向本书的责任编辑董亚峰先生和电子工业出版社的编辑们致敬，他们是本书的幕后英雄。我的妻子昆和女儿 Sophie 给予我最大的理解和支持，我爱你们。

由于我们对大数据相关技术及算法的理解和专业知识水平都有局限，本书的错误和疏漏之处在所难免，敬请各位读者谅解和指正。请将您的意见和建议发送至 [dyf@phei.com.cn](mailto:dyf@phei.com.cn)，感谢您的支持。

赵 勇

2015年5月12日

# 目 录



第 1 章 大数据技术概述 .....	1
1.1 大数据的概念 .....	1
1.2 大数据的行业价值 .....	4
1.3 大数据问题的爆发 .....	9
1.4 大数据处理流程 .....	12
1.5 大数据技术 .....	13
1.5.1 基础架构支持 .....	14
1.5.2 数据采集 .....	14
1.5.3 数据存储 .....	15
1.5.4 数据计算 .....	16
1.5.5 展现与交互 .....	18
1.6 练习题 .....	19
参考文献 .....	19
第 2 章 大数据基础支撑——数据中心及云计算 .....	20
2.1 数据中心概述 .....	20
2.1.1 云计算时代数据中心面临的问题 .....	21
2.1.2 新一代数据中心关键技术 .....	22
2.1.3 业界发展动态 .....	24
2.1.4 小结 .....	25
2.2 云计算简介 .....	25
2.2.1 云计算定义 .....	26
2.2.2 云计算主要特征 .....	27

2.2.3	Web 服务、网格和云计算 .....	28
2.2.4	云计算应用分类 .....	29
2.2.5	小结 .....	31
2.3	大数据与云计算的关系 .....	32
2.3.1	大数据是信息技术发展的必然阶段 .....	33
2.3.2	云计算等新兴信息技术正在真正地落地和实施 .....	34
2.3.3	云计算等新兴技术是解决大数据问题的核心关键 .....	34
2.4	云资源调度与管理 .....	35
2.4.1	云资源管理 .....	36
2.4.2	云资源调度策略 .....	38
2.4.3	云计算数据中心负载均衡调度 .....	40
2.5	开源云管理平台 OpenStack .....	44
2.5.1	OpenStack 的构成 .....	45
2.5.2	OpenStack 各组件之间的关系 .....	46
2.5.3	OpenStack 的逻辑架构 .....	47
2.5.4	小结 .....	49
2.6	软件定义网络 .....	49
2.6.1	起源与发展 .....	50
2.6.2	OpenFlow 标准和规范 .....	50
2.6.3	OpenFlow 的应用 .....	53
2.7	虚拟机与容器 .....	55
2.7.1	VM 虚拟化与 Container 虚拟化 .....	55
2.7.2	Docker .....	55
2.8	练习题 .....	57
	参考文献 .....	57
<b>第 3 章</b>	<b>云计算先行者——Google 的三驾马车 .....</b>	<b>59</b>
3.1	Google 的三驾马车 .....	59
3.1.1	GFS——一个可扩展的分布式文件系统 .....	59
3.1.2	MapReduce——一种并行计算的编程模型 .....	64
3.1.3	BigTable——一个分布式数据存储系统 .....	69
3.2	Google 新“三驾马车” .....	77
3.2.1	Caffeine——基于 Percolator 的搜索索引系统 .....	77
3.2.2	Pregel——高效的分布式图计算的计算框架 .....	80
3.2.3	Dremel——大规模数据的交互式数据分析系统 .....	85
3.3	练习题 .....	89
	参考文献 .....	89
<b>第 4 章</b>	<b>云存储系统 .....</b>	<b>91</b>
4.1	云存储的基本概念 .....	91
4.1.1	云存储结构模型 .....	91



4.1.2	云存储与传统存储系统的区别	94
4.2	云存储关键技术	95
4.2.1	存储虚拟化技术	95
4.2.2	分布式存储技术	97
4.3	云存储系统分类	98
4.3.1	分布式文件存储	99
4.3.2	分布式块存储	105
4.3.3	分布式对象存储	109
4.3.4	统一存储	117
4.4	其他相关技术	124
4.5	练习题	127
	参考文献	127
<b>第5章</b>	<b>数据采集系统</b>	<b>129</b>
5.1	Flume	130
5.1.1	Flume 架构	130
5.1.2	Flume 核心组件	133
5.1.3	Flume 环境搭建与部署	134
5.2	Scribe	139
5.2.1	Scribe 架构	139
5.2.2	Scribe 中的 Store	140
5.2.3	Scribe 环境搭建与部署	141
5.3	Chukwa	144
5.3.1	Chukwa 的设计目标	144
5.3.2	Chukwa 架构	145
5.3.3	Chukwa 环境搭建与部署	147
5.4	Kafka	150
5.4.1	Kafka 架构	150
5.4.2	Kafka 存储	152
5.4.3	Kafka 的特点	152
5.4.4	Kafka 环境搭建与部署	154
5.5	练习题	155
	参考文献	155
<b>第6章</b>	<b>Hadoop 与 MapReduce</b>	<b>156</b>
6.1	Hadoop 平台	156
6.1.1	Hadoop 概述	156
6.1.2	Hadoop 的发展简史	157
6.1.3	Hadoop 的功能和作用	158
6.1.4	HDFS	159
6.1.5	HBase	162

6.2	MapReduce	171
6.2.1	第一代 MapReduce (MRv1)	172
6.2.2	MapReduce 2.0——Yarn	180
6.3	Hadoop 相关生态系统	184
6.3.1	交互式数据查询分析	184
6.3.2	数据收集、转换工具	187
6.3.3	机器学习工具	188
6.3.4	集群管理与监控	188
6.3.5	其他工具	189
6.4	Hadoop 应用案例	191
6.5	练习题	192
	参考文献	192
<b>第 7 章 Spark——大数据统一计算平台</b>		<b>193</b>
7.1	Spark 简介	193
7.1.1	Spark	193
7.1.2	BDAS	195
7.2	RDD	197
7.2.1	RDD 基本概念	197
7.2.2	RDD 示例	199
7.2.3	RDD 与分布式共享内存	200
7.3	Spark SQL	201
7.4	MLlib	203
7.5	GraphX	206
7.6	Spark Streaming	206
7.6.1	基本概念	207
7.6.2	编程模型	208
7.7	Spark 的安装	210
7.7.1	单机运行 Spark	210
7.7.2	使用 Spark Shell 与 Spark 交互	213
7.8	Shark、Impala、Hive 对比	214
7.9	练习题	216
	参考文献	216
<b>第 8 章 Storm 流计算系统</b>		<b>218</b>
8.1	流计算系统	218
8.1.1	流计算系统的特点	218
8.1.2	流计算处理基本流程	219
8.2	Storm 流计算框架	220
8.2.1	Storm 简介	220
8.2.2	Storm 关键术语	221

8.2.3 Storm 架构设计 .....	223
8.3 Storm 编程实例 .....	225
8.4 Storm 应用 .....	228
8.4.1 Storm 应用场景 .....	228
8.4.2 Storm 应用实例 .....	228
8.5 其他流计算框架 .....	229
8.6 练习题 .....	231
参考文献 .....	231
<b>第 9 章 SQL、NoSQL 与 NewSQL .....</b>	<b>232</b>
9.1 传统 SQL 数据库 .....	232
9.1.1 关系模型 .....	232
9.1.2 关系型数据库的优点 .....	233
9.1.3 关系型数据库面临的问题 .....	234
9.2 NoSQL .....	234
9.2.1 NoSQL 与大数据 .....	235
9.2.2 NoSQL 理论基础 .....	235
9.2.3 分布式模型 .....	238
9.2.4 NoSQL 数据库分类 .....	241
9.3 NewSQL .....	255
9.3.1 系统分类 .....	255
9.3.2 Google Spanner .....	256
9.3.3 MemSQL .....	258
9.3.4 VoltDB .....	260
9.4 练习题 .....	263
参考文献 .....	263
<b>第 10 章 大数据与数据挖掘 .....</b>	<b>264</b>
10.1 数据挖掘的主要功能和常用算法 .....	264
10.1.1 数据挖掘的主要功能 .....	264
10.1.2 常用算法 .....	265
10.2 大数据时代的数据挖掘 .....	280
10.2.1 传统数据挖掘解决方案 .....	280
10.2.2 分布式数据挖掘解决方案 .....	280
10.3 数据挖掘相关工具 .....	282
10.3.1 Mahout .....	282
10.3.2 语言工具——Python .....	288
10.4 数据挖掘与 R 语言 .....	289
10.4.1 R 语言简介 .....	289
10.4.2 R 语言在数据挖掘上的应用 .....	290
10.5 练习题 .....	294

参考文献	294
<b>第 11 章 深度学习</b>	<b>298</b>
11.1 深度学习介绍	299
11.1.1 深度学习的概念	299
11.1.2 深度学习的结构	299
11.1.3 从机器学习到深度学习	301
11.2 深度学习基本方法	302
11.2.1 自动编码器	302
11.2.2 稀疏编码	304
11.3 深度学习模型	305
11.3.1 深度置信网络	306
11.3.2 卷积神经网络	308
11.4 深度学习的训练加速	310
11.4.1 GPU 加速	310
11.4.2 数据并行	311
11.4.3 模型并行	312
11.4.4 计算集群	313
11.5 深度学习应用	313
11.5.1 Google	314
11.5.2 百度	314
11.5.3 腾讯 Mariana	315
11.6 练习题	316
参考文献	316
<b>第 12 章 电子商务与社会化网络大数据分析</b>	<b>318</b>
12.1 推荐系统简介	318
12.1.1 推荐系统的评判标准	319
12.1.2 推荐系统的分类	319
12.1.3 在线推荐系统常用算法介绍	320
12.1.4 相关算法知识	323
12.2 计算广告	327
12.2.1 计算广告简介	327
12.2.2 计算广告发展阶段	327
12.2.3 计算广告相关算法	330
12.2.4 计算广告与大数据	332
12.2.5 大数据在计算广告中的应用案例	333
12.3 社交网络	333
12.3.1 社交网络中大数据挖掘的应用场景	334
12.3.2 社交网络大数据挖掘核心算法模型	334
12.3.3 图计算框架	335

12.3.4	大数据在社交网络中的应用案例	337
12.4	练习题	338
<b>第 13 章</b>	<b>大数据展示与交互技术</b>	<b>339</b>
13.1	数据可视化分类	339
13.1.1	按照展示内容进行划分	340
13.1.2	按照数据类型进行划分	341
13.2	可视化技术分类	351
13.2.1	2D 展示技术	351
13.2.2	3D 渲染技术	356
13.2.3	体感互动技术	360
13.2.4	虚拟现实技术	362
13.2.5	增强现实技术	364
13.2.6	可穿戴技术	365
13.2.7	可植入设备	368
13.3	练习题	369
	参考文献	369
<b>第 14 章</b>	<b>大数据安全与隐私</b>	<b>372</b>
14.1	云计算时代安全与隐私问题凸显	372
14.2	云计算与大数据时代的安全挑战	374
14.2.1	大数据时代的安全需求	374
14.2.2	信息安全的发展历程	375
14.2.3	新兴信息技术带来的安全挑战	376
14.3	如何解决安全问题	380
14.3.1	云计算安全防护框架	381
14.3.2	基础云安全防护关键技术	384
14.3.3	创立本质安全的新型 IT 体系	387
14.4	隐私问题	389
14.4.1	防不胜防的隐私泄露	389
14.4.2	隐私保护的政策法规	390
14.4.3	隐私保护技术	391
14.5	练习题	393
	参考文献	393
<b>第 15 章</b>	<b>大数据技术发展趋势</b>	<b>394</b>
15.1	实时化	394
15.2	内存计算	396
15.2.1	机遇与挑战	396
15.2.2	研究进展	397
15.2.3	发展展望	399

15.3	泛在化	399
15.3.1	发展现状	400
15.3.2	发展趋势	401
15.4	智能化	406
15.4.1	传统人工智能	406
15.4.2	基于大数据的人工智能	407
15.5	练习题	410
	参考文献	410
<b>第 16 章</b>	<b>知名企业大数据架构简介</b>	<b>411</b>
16.1	腾讯	411
16.1.1	背景介绍	411
16.1.2	整体架构	412
16.2	淘宝	416
16.2.1	背景介绍	416
16.2.2	整体架构	416
16.3	Facebook	417
16.3.1	背景介绍	417
16.3.2	整体架构	418
16.3.3	技术架构展望	420
16.4	Twitter	420
16.4.1	背景介绍	420
16.4.2	整体架构	420
16.4.3	技术架构展望	422
16.5	Netflix	422
16.5.1	背景介绍	422
16.5.2	整体架构	423
16.5.3	Netflix 个性化和推荐系统架构	426
16.6	练习题	430
	参考文献	430

# 第 1 章

## 大数据技术概述

---

近年来，以物联网、移动互联网、云计算和大数据（Big Data）为代表的新一代信息技术发展迅猛，而大数据则风头最劲。无所不在的移动终端、智能设备、无线传感器等每秒都在产生数据，拥有数以亿计用户的互联网服务时时刻刻在产生巨量的交互，比如，百度每天大约要处理几十 PB 的数据，Twitter 每天会产生 7TB 的数据，Facebook 每天生成 300TB 以上的日志数据等。数据产生的速度太快，要处理的数据量也太大。据 IDC 预测，到 2020 年全球将拥有 35ZB（1ZB =  $10^{21}$  字节）的数据。与此同时，数据的价值也在不断凸显，数据被类比为新时代的黄金和石油，现代企业快节奏的业务需求和竞争压力对数据处理的实时性和有效性提出了更高的要求，传统的数据处理技术已经完全不能满足大量数据的实时处理的需求，大数据就全面爆发了。大数据涉及国家战略、区域及企业发展、社会民生的方方面面，掌握大数据的核心理念、模式和技术，就把握了新时代的脉搏。2013 年被称为大数据元年，在 2014 年，大数据不管在技术还是应用方面都取得了很大的发展。本书主要介绍大数据技术及算法的相关知识及最新进展，帮助大数据从业人士了解、掌握和架构大数据。

在本章中，我们将重点介绍大数据的相关概念、历史发展、大数据的价值、大数据问题、大数据处理流程以及大数据技术分类等几个方面。

### 1.1 大数据的概念

大数据指的是无法在规定时间内用现有的常规软件工具对其内容进行抓取、管理和处

理的数据集合。大数据技术则特指新一代的创新型的技术，能够突破常规软件的限制，是对大数据进行采集、存储和处理的技术的统称。

大数据（Big Data）一词正式出现是在 2011 年麦肯锡全球研究院发布的《大数据：下一个创新、竞争和生产力的前沿》研究报告中。但其实从 20 世纪 90 年代至 21 世纪初，大数据已经开始萌芽。那时对于大数据的研究主要集中在算法、模型、模式、识别等几个方面，数据挖掘理论和数据库技术也在趋于成熟，一批数据仓库、知识管理系统等商业智能工具逐步出现并被应用。这些方面的研究为大数据时代的到来奠定了良好的理论基础。

从 2003 年开始，业内学者开始围绕着半结构化数据或非结构化数据的处理进行探索，大数据的发展也取得了一定的突破。传统的数据处理技术能够较好地处理结构化的数据，而对非结构化的数据则难以应对。对非结构化数据的处理带动了数据处理技术的发展。2005 年 Hadoop 项目诞生，由多个软件产品组成的 Hadoop 生态系统为结构化和复杂数据的处理提供了一个快速、可靠分析的平台。其中分布式文件系统 HDFS 的可靠数据存储服务和高性能的 MapReduce 并行数据处理服务为大数据的处理提供了良好的技术基础。

2006—2009 年，并行计算和分布式系统成为主流，大数据技术的发展进入成熟期。从 2009 年开始，各国政府对于大数据技术的重视和应用已经初现端倪。在 2009 年，联合国启动了“全球脉动”（Global Pulse）计划，旨在推动数字数据快速收集和分析方式的创新；美国启动 Data.gov 网站，向公众开放各种政府数据；欧洲一些领先的研究型图书馆和科技信息研究机构开始合作，从而改善在互联网上获取科学数据的简易性。

2010 年以来，智能手机的应用日益广泛，随之而来的数据的碎片化、分布式、流媒体等特征更加明显，移动数据量急速增长。2011 年 5 月，麦肯锡全球研究院发布的《大数据：下一个创新、竞争和生产力的前沿》研究报告中指出，大数据已经渗透到当今的每一个行业和业务职能领域成为重要的生产因素。报告中提出了可能改变世界格局的 12 项技术，包括云计算、物联网、移动互联网、知识工作自动化、先进机器人等，而大数据则是这些技术的基础，每项技术都离不开大数据。同年 12 月，信息处理技术作为 4 项关键技术之一在我国的十二五规划上被提出来，其中包括海量数据存储、数据挖掘、图像视频智能分析等技术，而这些都是大数据的重要组成部分。另外 3 项关键技术，包括信息感知技术、信息传输技术、信息安全技术，也都与大数据密切相关。

2012 年 1 月份，在瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响》宣称，数据像货币或黄金一样，已经成为一种新的经济资产类别。2012 年 3 月，美国政府发布了《大数据研究和发​​展倡议》，并划拨 2 亿美元的专项支持资金，标志着大数据已经成为重要的时代特征。美国政府对数据领域的投资，也使得大数据的应用从一开始单纯的商业行为上升到国家科技战略，引发了全球对大数据的关注和追捧。

2012 年 4 月，美国的大数据处理公司 Splunk 成功上市，这也是第一家上市的大数据处理公司。该公司提供大数据监测和分析服务的软件，其成功上市也促使其他 IT 公司加快对大数据技术和应用的布局。7 月，阿里巴巴推出数据分享平台“聚石塔”，为淘宝、天猫



上的电商和电商服务商提供数据云服务，希望通过分享和挖掘海量数据，为国家和中小企业提供价值。阿里巴巴集团也成为国内最早提出企业数据化运营的企业。

2012年7月，联合国在纽约发布了一份关于大数据政务的白皮书，总结了各国政府如何利用大数据更好地服务和保护人民。这份白皮书举例说明在一个数据生态系统中，个人、公共部门和私人部门各自的角色、动机和需求。例如通过对价格关注和更好服务的渴望，个人提供数据和众包信息，并对隐私和退出权力提出需求；公共部门出于改善服务、提升效益的目的，提供统计数据、设备信息、健康指标及税务和消费信息等，并对隐私和退出权力提出需求；私人部门出于提升客户认知和预测趋势目的，提供汇总数据、消费和使用信息，并对敏感数据所有权和商业模式更加关注。联合国还以爱尔兰和美国的社交网络活跃度增长可以作为失业率上升的早期征兆为例，表明政府如果能合理分析所掌握的数据资源，将能“与数俱进”，快速应变。

2013年称为大数据元年，几乎所有的大型互联网企业都将业务范围延伸至大数据产业。电子商务平台、社交平台、门户网站等，都存在着大数据的影子。大数据也由技术热词逐渐演变成社会浪潮，影响着国家、社会和生活的各个方面。

全球大数据产业的日趋活跃，技术演进和应用创新的加速发展，使各国政府逐渐认识到大数据在推动经济发展，改善公共服务，增进人民福祉，乃至保障国家安全方面的重大意义。2014年5月，美国白宫发布了2014年全球“大数据”白皮书的研究报告《大数据：抓住机遇、守护价值》。报告鼓励使用数据以推动社会进步，特别是在市场与现有的机构并未以其他方式来支持这种进步的领域；同时，也需要相应的框架、结构与研究，来帮助保护美国人对于保护个人隐私、确保公平或防止歧视的坚定信仰。

近年来大数据不断地向社会各行各业渗透，已经开始广泛应用于教育、金融、医疗等各个行业，使得大数据的技术领域和行业边界愈来愈模糊不定，应用创新已超越技术本身更受到青睐。大数据技术可以为每一个领域带来变革性影响，并且正在成为各行各业颠覆性创新的原动力和助推器。

大数据需要满足数据量足够大（Volume）、数据的种类多样（Variety）、数据的增长及处理速度快（Velocity）、数据蕴藏价值大（Value）这4个根本特征，才能称之为大数据。

数据量大（Volume）指的是数据的采集、存储和计算的量都非常大，大数据通常指10TB以上规模的数据量。造成数据量增大的原因有很多，例如，很多监控和传感设备的使用，使我们感知到更多的事务，这些事务的数据将被部分或者完全存储；（移动）通信设备的使用，使得交流的数据量成倍增长；基于互联网和社会化网络的应用的发展，数以亿计的用户每天产生大量的数据。

数据种类多（Variety）是指数据的种类和来源较多，例如多种传感器、智能设备、社交网络等。数据的种类包括结构化、半结构化和非结构化数据，包括图片、音频、视频、地理位置等多类型的数据。

数据的增长及处理速度快（Velocity）指数据每分每秒都在爆炸性地增长，而对数据的