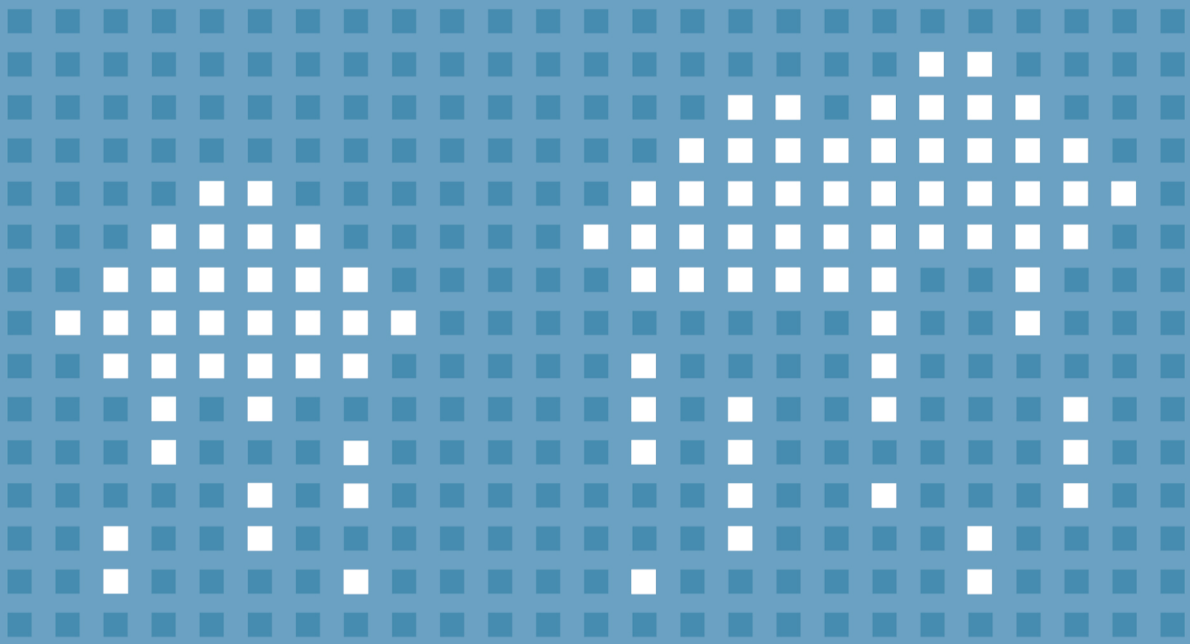


大数据架构师指南

朱进云　陈　坚　王德政　编著

A GUIDE FOR

BIG DATA ARCHITECTS

清华大学出版社

大数据架构师指南

朱进云　陈　坚　王德政　编著

A GUIDE FOR
BIG DATA ARCHITECTS

清华大学出版社
北　京

图书在版编目(CIP)数据

大数据架构师指南/朱进云，陈坚，王德政　编著. —北京：清华大学出版社，2016
ISBN 978-7-302-43516-7

Ⅰ. ①大… Ⅱ. ①朱… ②陈… ③王… Ⅲ. ①数据处理—指南 Ⅳ. ①TP274-62

中国版本图书馆 CIP 数据核字(2016)第 079520 号

责任编辑：陈　莉　高　屾
封面设计：周晓亮
版式设计：方加青
责任校对：曹　阳
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：180mm×250mm　　印　　张：18.5　　字　　数：320 千字
版　　次：2016 年 6 月第 1 版　　印　　次：2016 年 6 月第 1 次印刷
印　　数：1～3500
定　　价：45.00 元

产品编号：065794-01

本书编委会

序一 Preface One

这是一个数据爆发的时代，宽带化、移动互联网、物联网、智能终端的普及与人工智能的兴起，促使全球数据每两年翻一番，预计2020年全球数据规模将达到44ZB，较2013年将增长10倍。有资料报告，2013年全球数据的来源基本上是消费者、企业与政府各贡献1/3。按照用户数计算，在中国，无论是互联网用户还是移动互联网用户，无论是固网宽带用户还是移动宽带用户，其规模都已经是全球第一，中国的数据拥有量的潜力为全球之冠。IDC公司曾经指出，2013年中国在全球数据占比13%，预计2020年将上升到18%。

拥有数据并不意味着坐拥金矿，数据的产生与存储还要付出成本代价，大数据只有通过数据分析与挖掘，发现知识和生成智慧才能创造价值。大数据挖掘的应用将总结事物发展规律，提升人类生产与管理活动的准确性，减少传统方式下的“试错”成本，进而提升社会的总生产效率。

大数据的挖掘需要很多技术支持，反过来说也带动了海量存储、高效计算、深度学习、可视呈现等很多技术和学科的发展，它是当代信息技术的集中体现。大数据挖掘本身是产业，但其效益更多地反映在其应用到的社会管理和其他行业中，大数据之所以受重视正是因为其溢出效益明显，大数据将成为影响国家竞争力的重要因素。

美国、英国、欧盟、日本和韩国等国政府越来越重视大数据所产生的价值，鼓励使用大数据以推动社会进步，支持政府数据的公共资源化，并发布促进大数据技术发展的政策纲要。2015年中国国务院发布《促进大数据发展行动纲要》，提出了五大目标 、三大任务、十大工程以及七项政策，在国家层面推动大数据的应用与落地。大数据的挖掘应用正在引起各行各业的关注，成为“互联网+”行动的主要抓手，将发掘经济增长新动能。

大数据的挖掘不仅需要技术，更需要人才，麦肯锡公司预测，到2018年美国对大数据深度分析人才的需求与实际可供给之间相差一倍以上。我国与发达国家相比

更缺乏深度分析人才，尤其是大数据架构师。高校承担了培养人才的责任，但更需在实践中锻炼，为加速大数据架构师的成长过程，实用经验的传承十分重要。

中兴通讯对大数据的知识与工程经验进行系统性的概述，正好契合了当前大数据挖掘应用的浪潮，弥补了此类书籍的空白，为促进大数据技术的发展与应用提供了宝贵的经验。

中国工程院院士
中国互联网协会理事长

序二 Preface Two

数据并不是一个新概念，几千年来我们一直在利用数据。但数据的价值，特别是大数据的价值，最近几年才成为公众关注的焦点，是有其时代背景的。

就如同石油在几千年前就被发现了，但是其用途一直是作为日常生活或战争中的燃料，并不是特别重要的战略物资。只有内燃机被发明后，石油才成为最重要的动力能源，在最近的一百年才成为战略物资。

数据也一样。传统的数据库技术，在数据处理的能力上都有很大的局限性，超过100T这个量级，要么是处理效率急剧降低，要么是系统成本上升到难以接受的昂贵程度。所以，在大数据时代之前，数据在生产系统中的使用目的往往是单一的、即时的。大量的历史数据与过程数据，按照当时的IT技术，既无法存储，更无法处理。那些被备份到磁带机上的数据，大部分都成为死亡的数据化石。

当前大数据处理的技术，特别是云存储与云计算技术的成熟应用，为大数据的存储与处理提供了技术可能性。企业可以利用生产系统以及管理系统中产生的大量数据，对海量的数据进行存储、挖掘分析。一方面可以对生产活动进行更为准确的预测与指导，从而提高企业生产活动的准确性；另一方面还可以通过对数据价值的挖掘，产生新的业务，帮助企业充分开发数据的价值。政府也可以利用大数据来提高管理水平和效率。

2014年Gartner发布的HypeCycle曲线中，大数据技术已经越过炒作顶点。从HypeCycle曲线来看，越过炒作顶点的技术，往往是已经满足技术可行性的技术。技术进展并辅以商业模式创新，大数据在部分细分市场已经具备商业可行性，可以为企业的现在与未来带来收益。

2015年8月国务院发布了《促进大数据发展行动纲要》，将大数据的应用与落地提升到国家层面。在这种背景下，当前大数据系统建设出现一波高潮。商业级的大数据系统建设周期长，复杂度高，资金投入量大，所以需要合理的系统架构以应对未来业务需求的变化。由于业界大数据系统的建设刚起步，当前阶段急需对相关的系统架构知识以及实际项目建设经验进行共享，提升业界的整体建设水平。

纵观当前业界大数据相关的书籍，偏重于两大类型。其一是偏重于大数据理念，描绘大数据前景，说明大数据可以有哪些应用；其二是偏重于大数据基础知识，偏重于实际的编程与开发。

但在大数据项目的实际建设过程中，架构师在进行端到端方案设计时，需要对大数据庞大的知识体系进行总揽性把握，并辅以实际项目的经验，才有可能把握此类系统的关键需求与要点。而此类的知识与经验，业界分享较少，只能通过各类交流活动才能获取，不仅费时费力，而且还很难将这些知识系统化。

中兴通讯作为业界知名企业，在大数据研发上投入大量资源，并具备丰富的实际工程经验。本书不仅针对大数据知识进行系统化概述，并且将实际大型项目的经验进行总结。这种无私分享的宝贵经验，正是业界所亟需的，对大数据从业者具备较好的参考价值。相信本书分享的知识与经验，对推动大数据应用与落地起到积极的促进作用。

中兴通讯股份有限公司董事长兼总裁

赵先明

前言 Foreword

毫无疑问，这是属于大数据的时代。随着移动互联网的进步、自媒体的风行和物联网的兴起，信息传播技术和信息传播渠道得到极大发展，海量级甚至银河级的数据不断涌现，呈现出“信息爆炸”的态势。这种情况下，似乎我们获取信息变得更加容易和方便；而实际上，由于对个体有用的信息淹没在浩如烟海的无关信息中，获取“有用信息”反而变得更加困难。

大数据相关技术就是在这种情况下应运而生的。作为一门新兴技术，大数据技术被人熟知和掌握需要一个过程；同时，由于其始终处于一个高速发展的过程，对其认识也是不断修正提高的过程。

鉴于此，本书总结了中兴通讯大数据平台DAP团队对大数据技术的最新研究成果，结合中兴大数据平台在各行业的应用实践经验，旨在帮助读者建立系统化的大数据技术脉络，并针对业界一些似是而非的问题进行系统性的讲解与澄清。阅读完本书，读者就可以基本掌握大数据技术的系统架构和核心思想。

为何要写这本书

在大数据项目建设过程中，往往需要三个层次的知识。第一个层次是关于大数据是什么，能做什么等理念方面的知识；第二个层次是如果去端到端进行大数据方案设计，要厘清大数据方案所需的关注重点，并结合具体的实践案例进行说明；第三个层次是大数据相关的基础技术知识，例如，对HDFS、MR、SPARK等技术点的掌握。

第一个层次的书籍，业界已经有很多，其中以《大数据时代》为典型代表；第三个层次的书籍，业界也比较多，读者不难获得相关的学习材料。

但第二个层次的书籍，属于承上启下的层次。该层次的知识需要从实践中总结出经验与知识。由于大型项目的建设周期长，建设复杂度高，涉及面广，所以从大型项目的实践中总结出知识有较高的难度。鉴于此，市面上该层次的大数据书籍相对较少，大数据相关的从业者或建设者较难获得这方面的知识，往往只能通过各类交流活动获取这方面的知识，不仅费时费力，而且难以将这些知识系统化。

基于如上原因，我们感觉迫切需要将我们在大型项目中积累的经验总结出来，供业界同仁参考，同时，这也可以满足我们内部人员学习大数据相关知识的需求。

本书读者对象

如果您是IT市场营销人员，或者是企业IT主管，您可以直接阅读本书的第一部分与第三部分。通过对本书第一部分与第三部分的阅读，将帮助您建立起大数据技术概念和框架。如果您对具体的大数据技术不感兴趣，可以忽略掉第二部分纯技术的内容。

如果您是大数据技术人员，本书将会是一本较好的参考资料，有助于帮助您超越自己所从事的具体模块，将您的大数据知识体系系统化。

如果您是高校大数据相关课程的老师，由于本书较为系统，可以考虑将本书作为参考书或者教材。

如果您是大数据技术爱好者，也可以将本书作为泛读书籍，让您理解当前大数据的时代。当然，读者如果能具备一定的IT基础知识，将能够更好地汲取本书中的知识。这不仅有助于您快速理解大数据相关知识，也有助于启发您对特定专题的深入思考和独到分析。

本书特色

本书是首本系统化的方案实践方面书籍，系统化地阐述了大数据方案应该如何思考，以及大数据的技术基础知识，并辅以实际的案例进行说明。

以客户化的语言，描述大数据项目建设中应该重点考虑的问题。即使不是技术专家，也能很容易地理解本书第一部分的内容。

较为系统地阐述了大数据相关的体系，可以帮助读者迅速系统化大数据相关的知识。

结合实际的案例，总结在大数据建设实践中的经验与知识。

如何阅读本书

本书内容分为四大部分，不同的读者可以选择不同的内容进行阅读。

本书第一部分是“大数据架构师入门”，以虚构角色小明的视角，去理解大数据，理解客户的烦恼，并提出构建一个大数据系统时应该从哪些方面考虑。阅读完该部分后，读者将对大数据方案具备一定的“提问题”的能力。也就是说，如果您面前有一份大数据的建设方案，即使您以前对大数据了解甚少，也可以根据本书第3章的建议，去评判方案的完整性，评判方案的深度与广度。

本书第二部分是“大数据架构师基础”，本部分将较为系统地介绍大数据相关

的基础知识。如图Ⅰ-1所示，逐个介绍基础支撑层、计算存储层、中间件层、挖掘分析/应用层、展现层各部分内容，同时，对贯穿各层的安全和管理两大模块的相关内容做介绍，力图为读者呈现一个相对完整的大数据知识架构。

图Ⅰ-1 大数据技术框架

其中，计算存储层包括Hadoop架构、Spark架构、分析挖掘组件等内容；中间件层包括中间件的作用与意义，以及业界常用中间件及应用场景；展现层包括可视化相关的知识与内容；安全模块包括物理安全、主机安全、网络安全、数据安全等内容；管理模块包括自动部署、自动升级、自动巡检、自动维护等内容。

本书第三部分是“大数据架构师实践”，主要包括大数据开发实践中积累的一些经验，并结合案例进行阐述。这些实战中积累的知识与智慧，将帮助理论联系实践，更好地理解大数据技术。

本书第四部分是“大数据架构师拓展”，主要包括与大数据相关的其他技术。这些技术通常来说，并不属于大数据的技术范畴，但由于这些技术与大数据关系紧密，作为一名架构师，也需要系统地了解与思考这些相关的技术，才能对整个方案进行全局把握。该部分将试图对这些技术进行简单介绍，并试图说明这些技术与大数据之间的关系。

对于不需要关注具体技术的读者，则可以仅阅读第一部分“大数据架构师入门”；如果对具体的案例感兴趣，则可以阅读第三部分“大数据架构师实践”；如果是对技术感兴趣的读者，则可以阅读第二部分“大数据架构师基础”与第四部分“大数据架构师拓展”。

本书编写团队

大数据的知识非常广泛，不同层面的知识，以及不同技术模块的知识，很难由一个人完全掌握，所以本书是编写团队共同努力的成果。编写团队的成员都是在大数据领域担当重要工作岗位的技术骨干，大家在共同的理想与爱好下，聚集成一个团队，并为大数据架构师们完成了业界首本全面实践指导类的书籍。在此，请允许我列举参与编写的团队成员，并向他们致以诚挚的谢意。感谢他们牺牲周末与节假日的休息时间，为大家做的无私贡献。

团队成员包括：申山宏、梁平、薛清华、李敏、郭海生、杨荣康、牛家浩、刘

少麟、管云、洪科、简明、张强、艾红芳、关涛、刘淑霞、郭进良、汪绍飞、周治中、王利学、黄增建。

勘误与支持

尽管我们尽了各种努力来保证文章不出错误，但由于编者水平有限，加上编写时间仓促，难免会有错讹之处。如果你在书中发现了错误，例如错别字、书写错误等，请告诉我们，我们将整理成勘误表。通过勘误表，可以帮助其他读者节省阅读时间，提高阅读体验，并可以帮助我们提供更高质量的下一版。

错误反馈请发送至邮箱zhou.zhizhong@zte.com.cn，或者关注“中兴大数据”微信公众号(微信号ZTE_BigData)并留言，我们将在第一时间确认反馈。勘误表可以在“中兴大数据”微信公众号上获取。

致谢

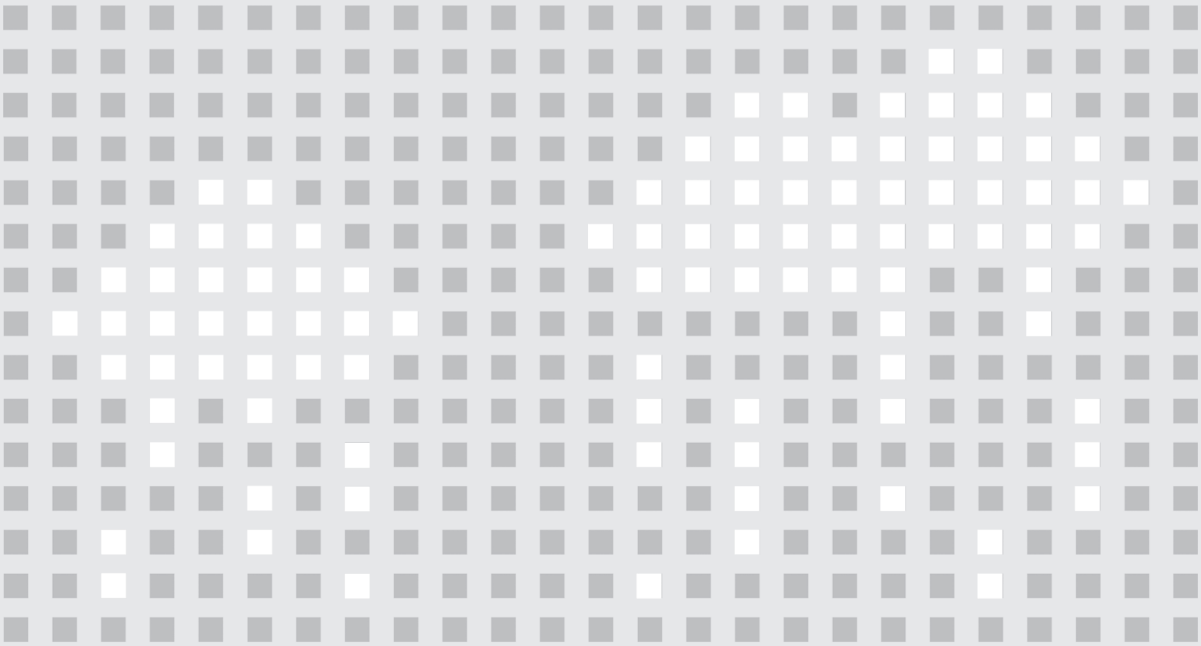
感谢中兴大数据平台DAP团队的所有成员，你们多年的潜心研究和积累是本书的基石。

感谢所有评审本书，并对本书提出过建议的朋友，你们的帮助对我们非常重要。

感谢关心本书的各界朋友，你们的关心与期望是我们的动力，更是对我们全心全意写好这本书的鞭策。

目录 Contents

第四部分 大数据架构师拓展



第一部分

大数据架构师入门

A GUIDE FOR

BIG DATA ARCHITECTS

第 1 章

大数据概述

故事是这样的，在英语课本中伴随我们成长的小明，中学毕业后考上了大学名校，“day day up”地苦修7年计算机、IT以及大数据知识后，终于成长为大数据咨询师。

记得那是明媚的春天，小明愉快地遨游在大数据一望无际的知识海洋里，春风十里不如大数据。忽然电话铃响了，电话那头传来Boss低沉的声音：“小明，请到我办公室来一趟。”

十里的春风，忽然变幻成浓郁的雾霾。小明走三步停一步，终于走到Boss面前。“国务院2015年8月31日已经印发了《促进大数据发展行动纲要》，你为啥到现在都没有向我报告？给你三天时间，给我说说，什么是大数据？大数据可以干啥？未来的技术方向是啥？”

小明熬了三天三夜，终于将业界关于大数据的科普知识整理出了一份报告，趁着早上Boss还没有来上班，悄悄地将报告放在Boss办公桌上。

1.1 什么是大数据

大数据，英文为Big Data。这个如今耳熟能详的名字，是《自然》(*Nature*)杂志于2008年9月4日的专辑“Big Data”中首次提出的。

Google在其推动世界范围内的信息整合过程中，极大地推动了大数据技术的创新和发展。

然而，到底什么是大数据？它的概念和外延包括哪些？由于大数据是最近新衍生出来的概念，它的内涵和外延也在不断地拓展和变化着，目前还没有一个业界广泛采纳的明确定义。

2011年6月，麦肯锡全球研究院(MGI)在它的报告《大数据：创新、竞争和生产力的下一个前沿领域》中这样描述：大数据是指无法用传统数据库软件工具对其内容进行抓取、管理和处理的大体量数据集合(“Big data” refers to datasets whose size

is beyond the ability of typical database software tools to capture，store，manage，and analyze)。

几乎同时，IDC(International Data Corporation)在它编制的年度数字宇宙研究报告《从混沌中提取价值》(*Extracting Value from Chaos*)中给大数据下了一个定义：大数据技术是新一代的技术与架构，它被设计用于在成本可承受(economically)的条件下，通过非常快速(velocity)的采集、发现和分析，从大体量(volumes)、多类别(variety)的数据中提取价值(value)(Big data technologies describe a new generation of technologies and architectures，designed to economically extract value from very large volumes of a wide variety of data，by enabling high-velocity capture，discovery，and/or analysis)。

IDC的定义描述了大数据时代的四大特征，即俗称的4V，而这4V(volumes、velocity、variety、value)也被广泛地认可为大数据的最基本内涵。

(1) 海量化(volumes)

数据体量巨大是大数据的首要特征，也是大家最容易发现的特征。全球数据正以前所未有的速度增长着，每天都有数以百万兆字节的数据在互联网上产生。据估计，全球可统计的数据存储量在2011年约为1.8ZB，2015年将超过8ZB。数据的爆炸式增长引发了数据存储和处理的危机。

(2) 多样化(variety)

数据类型的日趋繁多是大数据的另一个特征。传统的数据可以用二维表的形式存储在数据库中，我们称之为结构化数据。但随着互联网多媒体应用的兴起，图片、声音和视频等非结构化数据成为了数据的主要组成部分，统计显示，目前全世界非结构化数据已占数据总量的90%左右。如何有效地处理非结构化数据，并挖掘出其中蕴含的商业价值和经济社会价值，是大数据技术要解决的问题。

(3) 快速化(velocity)

快速处理是大数据必须满足的要求。经济全球化形势下，企业面临的竞争环境越来越严酷。在此情况下，如何及时把握市场动态，深入洞察行业、市场、消费者的需求，并快速、合理地制定经营策略，就成为企业生死存亡的关键。而对大数据的快速处理分析，是实现这一目标的前提。

(4) 价值化(value)

大数据蕴含的整体价值是巨大的，但是由于干扰信息多，导致其价值密度低，

这是大数据在价值维度的两个特征。挖掘出大数据的有用价值并加以利用，是数据拥有者的自然目标。但市场形势瞬息万变，因此，如何在海量的、多样化的、低价值密度的数据中快速挖掘出其蕴含的有用价值，是大数据技术的使命。

虽然后续不断有人增加对“V”的理解，如veracity(真实和准确)，强调真实而准确的数据才能让对数据的管控和治理真正有意义；如vitality(动态性)，强调数据体系的动态性等。这些对大数据的内涵都有一定的推动作用，但都不及开始的4V具有广泛性。

1.2 大数据的本质

所有技术的发展都是为社会进步服务的，大数据技术也不例外。但是，大数据技术对社会生产的促进作用是变革性甚至是颠覆性的。

“大数据商业应用第一人” Viktor Mayer-Schönberger在其著作《大数据时代》中，前瞻性地指出，大数据正在变革我们的生活、工作和思维。大数据开启了一次重大的时代转型，为我们带来了思维变革、商业变革和管理变革。其中最重要的三个思维变革颠覆了千百年来人类的思维惯例，对人类的认知和与世界交流的方式提出了全新的挑战。

(1) 全样本

我们将使用更多的数据甚至是全部数据来进行分析，而不再采用随机样本。从可能性角度，当前的技术能力已经可以支撑海量数据的处理；从必要性角度，有时候数据分析的目的就是要发现大量正常数据中的少数异常情况，例如跨境汇款中的异常交易，这无法通过采样分析获得。

(2) 概率化

我们将不再沉迷于精确性，而是允许劣质数据混杂其中。大数据时代不可能实现精确，反之用概率来表示事物发展的大方向，混杂性变成了一种标准途径。

(3) 相关性

我们将更关心相关关系，因果关系被放到次要的位置。在很多场景下，“是什么”比“为什么”对决策的帮助更大，可以在快速变化的环境中帮助你先发一步。

甚至，在一些不知道“为什么”的场景下，知道“是什么”反而有助于人们取得发现“为什么”的突破。

基于这种思维发展起来的大数据技术，具有以往的各种技术不具备的准确性和实时性优势，当它应用到社会各行业生产中时，对社会生产效率的提升是异常显著的。

很多人对于大数据应用的认识，都始于Google对于流行性疾病的成功预测。Google利用当前人们喜欢上网搜索解决方案(如搜索流感症状或者治疗药物)的习惯，找出了对应时段内某些特定字段的搜索频率与美国疾控中心历史记录中某些流行性疾病在空间和时间上的相关性，并据此而建立了一个数学模型。利用这个数学模型，Google成功预测了2009年H1N1流感的发展过程。

而这个成功应用带来的振奋远不止如此。首先，作为一家互联网公司，Google在与其毫无关联的医学专业领域获得了成功；更重要的是，它的预测在准确性特别是实时性方面，远远超过专业的美国疾控中心。

于是，更多的人在更多的行业开始了大数据应用尝试。

在零售业：梅西百货(Macys)已经实现对多达7300万种货品进行实时调价，以实现销量和利润的双重最大化；塔吉特(Target)公司通过对用户历史消费记录的大数据分析，实现对用户下一阶段消费行为的预测，从而实现精准投放。

在博彩业：Tipp24 AG公司用KXEN软件来分析数十亿计的交易以及客户的特性，然后通过预测模型对特定用户进行动态的营销活动。这项举措减少了90%的预测模型构建时间。

在通信业：中兴通讯创新性地提出了基于大数据技术的电信系统反馈环理念，让电信网络作为一个整体获得实时的系统反馈，从而使网络性能更加稳定，网络运维更加高效；而全球120家运营商中，已经有48%的企业正在实施大数据战略，通过提高数据分析能力，他们正试图打造着全新的商业生态圈，实现从电信网络运营商(Telecom)到信息运营商(Infocom)的华丽转身。

在金融业：阿里通过对用户消费习惯的大数据分析，已经可以将余额宝第二天的赎回规模的预测准确率保持在97%以上，连“双十一”等大促销造成的大规模资金流动也不例外；中信银行与中兴通讯大数据平台强强联合，打造一个全新的“数据银行”，利用金融大数据更科学地实现加强风险管控、精细化管理、业务创新等业务转型。

在公共管理行业：中兴通讯为2014南京青奥会打造的“环宁护城河”项目，将

各种警务数据在大数据平台上集中处理，从时间和空间两个维度进行实时统计和展现，为青奥安保工作部署提供科学的决策依据。

越来越多的实践证明，大数据运用可以为各个行业带来巨大的收益。

麦肯锡在它的报告中，根据各行业利用大数据技术获取利益的潜力，将各个行业分为5个组别。

(1) 计算机和电子产品及信息行业必然能够从大数据中获取巨大利益，该行业本身就有巨大的信息池且具有快速创新的特点，与大数据天然吻合。

(2) 社会公共管理及金融业则需要通过细分和自动化算法来克服技术障碍，从而大为受益。

(3) 建筑、教育服务、艺术和娱乐等行业则面临着获取海量数据价值的系统障碍。当然，如果这些障碍是可以克服的，则也可以从大数据中获益。

(4) 制造业、批发贸易等行业全球交易程度高，如果能够克服数据和技术上的障碍，则从行业普遍意义上讲获益巨大，但面临的困难同样不小。

(5) 零售、医疗、住宿和食物等本地服务行业全球交易程度低，则从行业普遍意义上讲，从大数据中获取价值的潜力相对较小。

1.3 大数据技术当前状态[①]

随着大数据在各个行业的广泛应用，各个行业在得到大数据带来的收益的同时，也在推动着大数据技术的飞速发展。

不同的行业有着不同的业务特征，进而也有不同的需求。如何满足这些不断涌现的需求，成为推动大数据技术发展的动力。

1. 零售行业

(1) 业务特征

零售行业同类产品的差异小，可替代性强，提高销售收入离不开出色的购物体

① 本节行业分析内容改写自《大数据生命周期全景与产业发展IADP模型研究(赛迪顾问)》，改写方式：缩写。

验和客户服务。同时，零售行业需要增强产品流转率，实现快速营销。

(2) 需求分析

提升客户购物体验的一个关键途径是精准营销，而精准营销的核心是用户消费行为分析，即用户识别。这个过程涉及消费历史记录、电话/WEB/电子邮件等数据中折射出的用户消费习惯识别。

快速营销的分析和决策基于对产品产、销、存及物流各个环节的大数据分析，涉及条码技术、标签技术、全息扫描技术、RF技术等技术。

2. 互联网行业

(1) 业务特征

互联网行业主要特征之一是数据量呈爆炸性增长，数据结构类型日趋复杂。各种类型的信息和数据都呈现爆炸式地增长。全球90%的数据都是在过去两年中生成的。在未来几年，数字信息会呈现更加惊人的增长，预计到2020年，信息和数据总量将增长44倍。

另一个特征是用户行为丰富，WEB社群关系复杂。互联网已经不再是单纯地浏览网页信息，互动已经成为主要方式。用户行为和网络中的社会群体变得更加多样化、复杂化。

(2) 需求分析

用户粘性对于互联网公司来说是至关重要的测评指标。而从爆炸性增长的数据和复杂的用户行为中，提取有价值的信息，分析用户行为，建立用户模型，来提高用户体验、增加用户粘性，是大数据技术发展的挑战和动力。

3. 电信行业

(1) 业务特征

数据量激增，保存时间长。近些年，由于无线上网和智能手机的推广，导致电信行业数据量呈现爆炸性增长。从全球移动网络中语音和数据流量的状况来看，2009年末，数据流量超过了语音流量，到2011年数据流量已经超过语音流量的两倍。根据研究预测，到2015年全球移动数据流量将比2010年上升26倍。电信行业不仅仅数据量大，而且保存时间长，一般电信行业要求数据保存2年6个月。

受众群体大，市场饱和度高。电信业务已经是人们生活中的必需品，用户数量

非常巨大，整体市场饱和度高。

(2) 需求分析

一方面，流量和用户的激增，给现有网络带来了巨大的压力。如何保持现有网络的稳定高效运转，成为各大运营商首先需要考虑的问题。而大数据技术能解决这一问题，例如中兴通讯提出的“基于大数据技术的电信系统反馈环理念”。

另一方面，运营商面临着从业务提供者到管道提供者的转变。如何在这个转变过程中，高效、合理地优化网络建设，同时能够发现潜在的信息应用需求并转变为商业价值，也需要大数据技术的支撑。

4. 金融行业

(1) 业务特征

金融业有着数据池积累巨大的天然优势，但同时如何挖掘数据价值也成为挑战。另外，金融业是高风险行业，有着其他行业不可比拟的安全性要求。

(2) 需求分析

从大量数据中挖掘有价值的信息，并将其作为判断的依据，及时准确地进行金融智能决策，是金融业迫切的需求。

金融业对安全的苛刻要求，成为大数据技术的挑战。

5. 交通行业

(1) 业务特征

1) 数据量大，数据类型多。随着车辆保有量的不断攀升，交通综合监控呈多维、立体化趋势，数据分析面对的是文本、语音、图片、视频等多种类型数据的飞速增长。

2) 实时性要求高。交通系统受很多因素的影响，时间、天气、路况、突发事件等都让交通状况产生突然并且累积性的变化。

(2) 需求分析

面对多种类型的海量数据加上极高的实时性要求，大数据技术需要在存储、计算、分析、处理等方面表现出超强的性能，才能满足对瞬息万变的交通状况进行及时调度和快速响应的要求。

1.4 大数据的技术发展趋势

随着大数据技术的发展，IT相关系统也正发生着变革。系统的硬件设计、软件设计，甚至商业部署都开始以数据为中心。也正是在这些实践和应用中，发现痛点并解决痛点的过程和探索，反过来推动大数据技术的发展。

从技术层面讲，以下几个方面将是大数据的热点。

(1) 硬件对架构的冲击

大数据对性能的要求非常高，而硬件的变化对性能会产生直接而巨大的影响，因此当硬件提升时，会推动大数据系统架构的变革，以达到充分利用硬件、大幅度提升性能的目的。

例如，下一代非易失内存(NVRAM) 的性能接近DRAM(最短延迟为DRAM的2～3倍)，这将对文件系统为主的存储架构产生巨大影响；同时，远程直接数据存取(RDMA)可将NVRAM连接成PB级(或更大)资源池，实现更简洁的内存计算，这将促进内存计算发展。

而针对数据的不同场景的专用硬件，将直接改变对应的系统架构。例如，对于很少使用的大容量数据，可以开发高密度/低IO/低功耗的低成本存储。

当大数据系统部署在云/虚拟化系统上时，系统架构需要考虑：存储部署在虚拟机上时，如何保证高IO需求；MR等计算框架，采用移动计算到数据侧的模式，其计算资源如何虚拟化，等等。

(2) 计算框架

随着大数据应用逐渐广泛，单一的计算框架已经无法满足需求。2014年图灵奖获得者Stonebraker认为：一刀切(one size fits all)的数据处理架构将寿终正寝，在流处理、数据仓库、数据库和科学数据库等方面会出现专用化引擎。

SPARK在持续走热，也揭示了从单一的MapReduce计算框架逐渐演变为多种计算框架并存的趋势。未来的计算框架将以通用计算框架为主(SPARK很可能成为主流)，在特殊场景下辅以较为专业的计算框架。

(3) 数据封装的中间件

实现数据的封装，是生态型平台必须具备的功能。大数据中间件层就是实现这一功能的组件。它位于应用层与底层数据库之间，屏蔽掉底层传统数据库、MPP、

Hadoop等数据存储的差异，同时为上层应用提供统一的开发接口，让应用层无须考虑底层的实现。

在从传统架构向大数据架构演进的过程中，多技术混搭是现实的需求，而大数据中间件层使得混搭方案成为可能。

(4) 非结构化数据处理

在今天的互联网数据中，结构化数据仅仅占到10%，非结构化数据成为最重要的源数据。非结构数据通常有音频/视频、文本、特定行业数据(如电信信令)等。对音频/视频数据的分析，已经有较为成熟的分析软件；对于特定行业数据，业内相关公司已经开始探索，如中兴通讯对电信信令的大数据分析；而文本分析也是最近在开源社区较为活跃的话题，通过和不同行业的结合，可以产生较多衍生应用。

(5) 智慧发现

学习可以分为数据、信息、知识、智慧4个层次，其中，智慧发现在未来很重要。在智慧发现领域，人工智能与大数据有较多的交叉重叠，其中深度学习是一个热点。深度学习是通过构建具有很多层的学习模型和海量的训练数据，来学习更有用的特征。

(6) 可视化

只有能被人类所理解的数据，才是有价值的数据；而可视化是最直观、最容易被理解的展示方式。

并不是只有传统的结构化数据可以可视化，操作、流程、信息等，一切皆可可视化。当前可视化技术呈现如下三个趋势。

1) 扁平化，即放弃一切装饰效果，所有界面元素的边界都干净利落，更加简单直接地将事物的工作方式展示出来，减少认知障碍的产生。同时，扁平化设计更简约，可以保证在所有的屏幕尺寸上都有相同的展示效果。

2) 动态化、可交互，即动态图形的表现力更丰富；通过界面的拖拽、点击、放大缩小，即可完成条件选择和切换。采用更少的菜单和更少的对话框，而不用复杂的条件选择对话框。

3) 多维度、多图联动，即通过多张图从不同维度展示同一个东西，即可在交互时，通过操作一张图引起其他相关图的联动，并且可以同时获得更多的信息。

第 2 章

大数据项目常见场景

又是一个风和日丽的下午，Boss将小明叫到办公室，办公桌上是一份A大学的某大数据咨询项目的case，Boss说：“上次的大数据报告写的不错，后生可畏啊！这里有个case，正好让你实战一下。”

2.1 实验型部署场景

2.1.1 背景介绍

A大学是国内知名大学，其计算机科学与应用数学都是国内学科的翘楚，该校毕业生很多都进入相关研究机构或国际著名IT企业任职。

随着大数据技术与应用的蓬勃发展以及国家大数据发展战略在学科建设上的深入落实，A大学决定抓住该历史机遇，充分发挥其学科优势，在大数据分析领域培养出新的优势学科。因此，A大学推动计算机科学系与应用数学系成立一个跨学科联合实验室，该实验室(后文统称大数据科学实验室)紧密结合社会需求，响应时代呼唤，定位于培养能够适应时代要求的大数据人才。

2.1.2 面临的问题

大数据科学实验室作为计算机科学系与应用数学系的联合实验室，承担了两个系在计算机科学与应用数学方向上的本科、研究生教学任务。与此同时，应用数学上的很多科学研究任务需要使用大量的计算资源进行数据分析，也需要使用大数据科学实验室的设备。

而这两个学科方向以及不同类型的教学、科研任务对大数据科学实验室的设备

有竞争关系，有时相互间甚至有冲突，影响了各项任务的顺利进行。如计算机科学的本科教学大纲中有实验课程，安排本科生动手搭建基本的Hadoop环境并在此基础上开发简单的分析应用，而应用数学研究需要大量的计算资源对海量数据进行机器学习模型训练，这两种任务之间就有明显的竞争关系。甚至经常因为机器分配和人员操作失误将运行了数百个CPU时间的模型训练任务意外关闭，这又进一步加剧了问题的严重性。

虽然有一笔资金可用于购买设备，但预算并不充分，对于如何配置软硬件产生了分歧。

一方面预算有限，另一方面又想以最高的性价比获得尽量多的计算和存储能力，甚至要有不受限制的节点数量，这真是愁坏了实验室主任，只能求助于专业的大数据咨询师。

2.1.3 需求分析

小明与几位实验室老师和学生坐下来聊了聊发现，大数据科学实验室开展的实验包括以下几部分。

(1) 教学实验：安装大数据环境，在此基础上设计并运行简单的分析应用，这部分应用对存储和计算资源的要求比较低。

(2) 新架构研究与实验：对大数据存储、计算架构进行实验研究，并通过大量的压力测试对架构性能进行评估和改进，这部分应用对存储和计算资源要求一般都比较高。

(3) 机器学习研究实验：训练机器学习模型，包括神经网络模型、统计模型、图模型等，主要是计算密集型的批处理应用。

小明还了解到由于实验室成立时间短，实验室设备虽然安排了专人管理，但缺乏管理工具的支持，仅通过机器密码实现简单的安全管理。

实验室设备构成复杂，既有高性能服务器，又有老式桌面机，而且各类实验对计算资源占用率不同，因此通过管理员人工对设备资源进行调度效果不佳。

因为各类实验对计算资源的消耗差异很大，有些实验严重浪费了宝贵的计算资源。

实验所用数据都是可公开获得的开放数据，对数据安全性要求不高，各数据集的规模大小不同，但均不超过GB。

经过上述摸底，小明对问题胸有成竹，这是计算密集型、存储规模小、数据安全性要求低又相对封闭的系统。

小明很快给出咨询建议：使用廉价的PC服务器+虚拟化解决方案+开源全栈式数据分析平台。PC服务器就能满足存储需求，同时也能获得不错的计算能力。采用虚拟化方案提高机器的利用率，同时减少实验间的干扰。开源全栈式数据分析平台更是能够将神经网络模型、统计模型、图模型的运算统一在一个计算框架下。

小明回到单位很快完成了咨询报告的撰写并提交到老板邮箱，第一个咨询项目大功告成！

2.2 中小型部署场景

第二天电话里又传来Boss低沉的声音：“小明马上到我办公室来一下！”小明原本欢乐的心脏顿时咯噔一声，难道报告Boss不满意？小明忐忑地迈进Boss办公室，没想到Boss拍着小明的肩膀说：“最近干得不错，上个case客户很满意，我这里还有一个case，我看好你哟！”

小明从Boss手中拿到case材料，看着眼熟，不是上周在网上大搞营销的那个B公司嘛，他们的产品设计新颖，用户口碑还挺不错的。小明明白，Boss这是在向他委以重任，于是小明立刻向Boss立下军令状，签下任务承诺书。

2.2.1 背景介绍

小明回到办公位立刻对B公司进行了全方位搜索，B公司是行业新玩家，但是凭借自己在用户体验方面的独门绝技，很快站稳了脚跟，并且拥有了忠实的客户群。B公司虽然员工规模不大，但业务发展异常迅猛，正向行业领头羊地位发起总攻。

小明对B公司做了360度调查后，拨通B公司电话说明来电意图，并约定当面拜访B公司进行现场调研。

B公司老总接待了小明，并向小明介绍了咨询目的。B公司是一家新创立的企业，其最初的企业定位就是以极致的用户体验与客户参与为差异化竞争点，通过微

信、论坛、问答网站等各种渠道建立起与用户的密切联系与紧密互动，直接将客户声音融入产品开发流程，提升客户的参与度与粘度，提高产品的用户体验。同时B公司特别重视市场分析与品牌战略，将量化的市场分析融入决策流程，所有重要决策都要有数据支撑，并且不遗余力地进行品牌建设，打造科技、时尚的企业形象。这种独特的竞争策略取得了巨大的成功，使得B公司业务规模持续保持高速增长。

老总还邀请小明参观了客服中心，安排小明参加了某产品开发团队、市场分析团队的例会，并与团队成员进行了面对面交谈，还邀请小明参加了一个产品的策划会。小明确实感受到这是一家朝气蓬勃的公司，虽然工作压力很大，但每位员工都清楚地知道自己的责任并为之奋斗。

2.2.2 面临的问题

在B公司开发团队例会和产品策划会上，小明发现各种各样的数据、图表是团队做决策的重要依据。而这些报表都是由市场分析团队综合网络、呼叫中心等各种用户沟通渠道反馈的用户需求，以及各类竞争厂商相关竞情信息，进行深入分析得到的深度洞察，B公司虽然年轻，但却有一个能准确把握客户需求与竞情事态的分析团队。B公司处于业务的快速增长期，对数据分析的需求也持续增长，但合格的数据分析师短缺的问题却很难在短期内通过招聘和内部培养解决。小明在与分析团队沟通中了解到，B公司信息化程度处于初级水平，信息的采集需要分析人员人工进行，耗时费力。而数据分析与可视化主要依赖Excel表格，而Excel模版开发的周期长，响应市场、开发团队的需求变化不灵活，随着产品线的扩大，新的分析需求不断涌现，而用户的增长也使得数据量急剧上升，传统的方式已经逐渐不能应对新的形势。因为采用人工采集的方式，因此原始数据格式不一，保存归档也没有工具支撑，导致数据的重用性差。分析团队迫切地**需要自动化数据采集、清洗与预处理流程，并且需要更加高效的数据分析与可视化工具**。

同样面临人手缺乏问题的还有客服团队，随着客户群的快速增长，客服团队虽然一再扩充，仍然难以满足实际需求，而场地、成本等其他因素也制约了客服团队进一步的扩大。小明参观客服中心时了解到客服系统虽然实现了统一通信，建立了客户资料数据库，但用户问题的解答全部依赖话务员经验，并且每次客户沟通都要客服人员手工录入客户资料数据库，进一步加大了话务员的工作压力。与此同时，

虽然花费巨大力气建立了客户资料数据库，但却没有充分地发挥出其作用。客服团队**迫切需要一个智能机器人**帮助自动回答一些常见问题，并能够自动地补充客户资料数据库，以减轻话务员的压力。

B公司老总是典型的精英人才，关于信息化和数据分析都有更深入的思考，考虑到公司未来几年可预期的高速增长，老总希望能够高起点地搭建一套大数据系统，将数据的采集、清洗、预处理、存储、分析自动化，重构现有的应用。同时**基于大数据平台和累积的用户数据、问答数据和各类实时数据，以构建全新的用户画像系统为核心，依此构建舆情监控、自动问答、客户关怀等上层应用**。投资预算相对宽松。

完成对客户的拜访，小明带着调研资料回到公司，马上投入到紧张的需求分析中。

2.2.3 需求分析

从调研可见，B公司现有业务系统比较简单，若将现有业务全部迁移到新系统中重新实现，则系统的设计受历史因素约束少，在预算宽裕的前提下，系统架构可以主要基于当前和未来的业务需求进行设计。

从调研结果看，B公司的需求涉及数据的采集、清洗、预处理、存储与分析计算几个方面，所需实现的业务都是BI、用户画像、知识体系、知识管理、舆情监控、问答系统等非关键型业务。数据规模中等，对计算能力、实时性、高可用性、冗余备份的要求都不太苛刻。但因为涉及的都是公司核心数据，因此对数据安全性要求很高。

数据来自外部的互联网、社交网络和内部的呼叫中心等多个渠道，除历史数据统一迁移外，数据主要以增量的方式积累，需要相应的数据采集接口，且由于数据来源的多样性导致数据形式与质量不一，需要一套完善的ETL系统管理数据的接入、清洗与预处理。

原始数据很大一部分是语音、文字等非结构化类型的数据，需要采用相应的自然语言处理技术进行处理和分析，这类应用主要是以流式应用为主。结构化的数据主要用来做决策支持，需要搭建数据仓库和相应BI系统，这类应用主要以批处理和交互式应用为主。

B公司前期没有专门的机房和专业IT管理员，机房工程设计与施工能力缺失，在

需要部署和运维中等规模大数据集群的前提下，需要采用turn-key交付方式。在预算充足的情况下，为减轻对IT管理的压力，应尽量选择成熟、功能完善的大数据平台管理系统。

B公司对本次上大数据非常重视，成立了以老总为第一负责人的领导小组，但B公司整体的IT能力较欠缺，需要抽取骨干人员尽早接受专业培训，并且直接参与大数据建设工作。随着数据分析工具的变化，分析团队也应抽取骨干人员尽早接受新工具的培训和使用。

中兴通讯大数据平台DAP是经过大量实践检验的、成熟的大数据平台，能够提供完善的ETL、存储、流分析、批处理分析、管理、安全和技术支持能力，并且有强大而富有经验的工程服务团队，同时能够提供IT运维管理、大数据分析工具等全方位的培训服务。因此，在预算充分的情况下，小明觉得硬件采用商用刀片服务器，软件采用DAP大数据平台的方案是一个不错的选择。

连续奋战了一个昼夜，小明终于制定出来一套基于DAP的详细技术方案并交到Boss案头。Boss看了小明的方案，大加赞赏，将小明提升为团队主管。

2.3 大型部署场景

Boss将团队交给小明带领的同时，又给了小明另一个更大的挑战。

这次拿到的任务让小明格外兴奋，这是一家著名的国际化大公司C公司，希望采用大数据技术重构整合自己的业务系统，当前阶段虽然项目目标并不算明晰，但公司已经准备了过亿元人民币的预算规模。

能够为C公司量身定制一套大数据方案无疑是业内所有架构师的梦想，当然其中的挑战也是毋庸置疑的。如此难得的机遇，像鞭子一样鞭策着小明，促使小明立马向Boss表态：保证带领团队完成任务！

小明立即召集团队人员展开专题研究。首先与客户领导建立联合工作机制，收集和理解客户的需求，反馈需要求助的问题，协调项目整体进展；其次与客户各部门IT运维人员联系，摸清企业的数据视图与业务流程；最后，在前面工作的基础上，与客户各部门的业务人员一起制订业务的开发与交割方案。

2.3.1 背景介绍

C公司拥有长期的信息化建设历史和富有经验的IT管理团队，其系统信息化水平较高，各种类型的生产系统已经在公司运行多年，并积累了海量的历史数据。在大数据不断重塑互联网行业，并不断向各种传统行业渗透的浪潮下，C公司也希望引入大数据相关的技术，为公司开拓新的价值增长点。

为保障项目成功实施，C公司也组建以首席技术官(CTO)为总负责人、以各业务部门主管为成员的专项小组。小明带领团队入驻C公司后立即进入专项小组，并将各团队派驻到相关部门开展摸底调研。

2.3.2 面临的问题

作为公司高层领导，C公司CTO也深知当前存在的诸多问题，如业务系统新老并存，部门墙导致数据分散无法充分利用。C公司对于大数据虽然有总体的目标，但如何让项目满足经济可行性并最终落地，尚有较多的困惑。其中主要集中在如下两个方面：

(1) 大数据系统与现有生产系统之间是怎样的关系？采用何种方式获取数据？

(2) 利用大数据技术构建何种业务应用？如何证明这些业务应用产生了经济效益，而不仅仅是在消耗公司宝贵的资源？

2.3.3 需求分析

小明深知对于这类项目，**数据集成**是其成功的必要条件，而将原本分散在不同的业务系统中的数据清洗、集成到统一的大数据仓库，再借助大数据平台强大的分析能力，无疑可以极大地提升数据的可用性和价值，把原本沉睡在各部门数据库中孤立的、无法利用的包袱转化为有价值产出的金矿。

基于大数据仓库海量的存储能力能够提供的企业数据视角，结合大数据分析能力，企业可以做到对外部市场和内部运营状况的深度洞察，通过数据分析与指标的量化，提高企业运行的透明度，进而提高支撑决策的精确性与科学性，是未来提高企业竞争力的核心。

企业现有的应用系统基于的技术与平台复杂多样，甚至存在很多早已过时的技术与架构，且各种应用间数据交互与共享方式异常复杂，导致整个系统的维护成本很高。而且由于各个应用系统都是独立建设的，独享硬件资源而无法实现设备资源共享，也带来了大量的资源浪费，使用成本居高不下。如果能够基于大数据平台逐渐地将这些应用迁移过来，统一在一套大框架下，借助大数据平台先进的技术架构和能力，无疑可以极大降低后续使用与维护成本。

通过与团队成员的大量讨论，小明很快制订出了以数据集成和应用迁移为核心的宏大的大数据实施方案。当小明满怀信心地将方案提交给客户后，该方案却得到C公司CTO与各部门主管的一致反对。

经过短暂的震惊后，小明马上做出调整，分别与CTO和部门主管们进行深度交流。认识到老的业务系统虽然存在一些问题，但已经稳定运行了多年，可靠性和功能都是可以保证的。而新方案对公司当前IT系统的改变太大，存在很多不确定因素。该方案不仅实施周期长、见效慢，而且对公司的运作冲击太大，造成很大的成本压力。

经过深入调研，小明对客户的需求理解又得到了进一步的加深。小明意识到架构师在设计时不能仅考虑技术因素，还要考虑更多的现实约束，包括投资成本、建设周期，以及抗冲击与风险。**大数据项目的建设应尽快带来效益亮点**，通过滚动的规划，快速上线可实现经济效益的业务应用，给客户以信心，并进一步推动大数据项目的持续、深入开展。

经过深入的分析，小明对方案进行了大幅修改。在大数据平台层面，**强调大数据架构设计的灵活性和可演进性**，为系统未来的滚动重构预留了设计余量；在大数据业务应用层面，**优先规划可以提升当前生产系统效率的应用**，以满足经济可行性的要求与压力。

同时，通过对C公司原有业务流程的分析，小明在新规划的大数据仓库基础上，借助海量的数据及庞大的计算能力，构建了一组可反映企业运营状态的KPI指标，并每天以报表的形式输出。

新的方案提交客户后，得到了CTO和部门主管们的一致认可。