

i0t0C101006t01000D¹010010100arororornmio70:0roD1010mroror0ro101Di0

0 1D100T000r0H01010101011010(0KD10j010101010rulor0101010101010101010TeiT 10T00701eI0r05100C101010102分1014106010701010C101010100I0toTare

M00101010010 L¹01011 SD1QT07070rercI015rg10L0L4040CFD401Q[0106TO10101011EGI040100TU40200109

**第** **1** **章** **数据科学概论**



**1.1** **数据科学的定义**

科学是对已经发现的、不断积累的、公认的普遍真理的总结，科学是系统化的知识体 系。科学包含自然科学和社会科学两大类别。

数据科学是2010年以来逐渐兴起的科学分支。人们普遍认为该门科学正在逐步形成， 其知识体系仍在创立之中。即便如此，人们也已经对其有一些基本的共识。科学的含义，一 部分在于对已经了解的事物的基本原理的整理，另一部分则在于探索的过程。数据科学家能 够利用系统化的、经过严格训练的方法，进行新的探索，为不断建立的理论体系做出贡献。

对于数据科学这个概念，有很多的解释。在本书中，我们侧重于从数据层面对其进行 解读，包括数据本身具有的内在价值，以及获取数据价值的技术手段等。

数据科学是关于对数据进行分析、抽取信息和知识的过程提供指导和支持的基本原则 和方法的科学。数据科学研究各种类型的数据、数据的不同状态、属性及其变化规律，它 研究各种方法和技术手段，对数据进行简单的以及复杂的分析，从而揭示自然界和人类行 为等不同现象背后的规律。

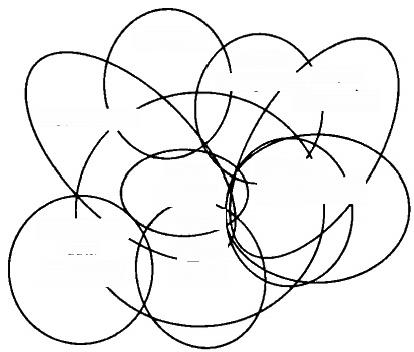
数据科学的核心任务是从数据中抽取信息、发现知识。它的研究对象是各种各样的数 据及其特性。数据科学包含一组概念 (Concept) 、 原 则 (Principle) 、 过程 (Process) 、 技 术/方法 (Technique/Method) 以及工具 (Tool), 为其核心任务服务。其中的概念和基 本原则 (Fundamental Principle) 给予我们观察问题、解决问题的一套完整的思想框架 (Structural Framework to Systematically Treat Problems)。大量的数据分析技术/方法和 工具，则帮助我们把思想落地，实现数据科学的目标。

简而言之，数据科学是以各类数据作为研究对象，建立在应对数据分析挑战的众多关 键技术基础上的通识意义上的科学。为了建立数据科学，人们需要从深层次上梳理关键的 数据分析处理技术，解析它们的定位和相互关联关系，在理论层面把这些技术联系起来， 也就是对基本概念、理论和技术加以系统化的整理。



**1.2** **数据科学的定位**

我们通过解读数据科学和统计学、人工智能/机器学习、数据挖掘、数据库与数据处 理、大数据分析、基于数据的决策的关系，来确定数据科学的定位(见图1-1)。

Statistics

Pattern

recognition

Neuro

computing

Visualization

Data

Machine Teaning

science

*AI*

KDD

Database & data

processing

Data mining

**图1** **-** **1** **数据科学是一门跨学科的新兴学科**

数据科学不是凭空发展起来的，它是一门新兴的交叉学科。它从数学/统计学、计算 机科学等传统学科领域，特别是从数据库、数据挖掘、大数据分析、人工智能/机器学习、 可视化等领域，借鉴了大量的理论和技术，吸收了有效的成分，逐步建立起自己的学科体 系。由于相关的理论和技术来自不同的研究领域，它们相互之间存在较大的差异，比如研 究的假设等。数据科学试图在此基础上构建和谐自洽的理论体系。

**1.2.1** **数据科学与数据库、大数据分析的关系**

业务系统的后台数据库，通过增加、删除、修改、查询等操作，对业务信息进行维 护，并且提供简单报表功能，支撑业务的运行。支撑业务运行的数据库，还不具备数据科 学的“范儿”,因为数据科学的核心任务是从数据中挖掘价值。但是，业务数据库的持续 运行，积累了大量的基础数据，为数据科学提供了重要的“原材料”。

信息技术的进步，大大降低了人们获取数据、存储数据和传输数据的成本，使得越来 越多的企业/机构有能力从自身的业务系统中或通过互联网等其他途径，获取并存储规模 日益庞大的数据。数据的价值对于很多现代企业而言越发重要，企业更加重视对历史数据 的积累。不断堆积的数据在规模和复杂程度上，逐渐超越了企业/机构所采用的已有技术 方案(或对其进行简单的升级改造)在执行数据管理和数据分析任务时所能达到的处理能 力，形成了大数据。

大数据具有三个主要的特点，其中最重要的特点是数据量大 (Big Volume), 其 规 模超出了已有工具的处理能力，需要研发新的工具来进行处理。当一个企业/机构意识 到不断堆积的数据具有非常大的价值，却无法通过对已有的信息系统进行传统意义上的

升级改造，来有效地从数据中提炼出足够的价值时，需要对承担数据管理和分析任务的 信息系统进行革命性的改造。大数据的第二个特点是数据类型多样，人们希望把不同来 源、不同类型的数据关联起来，进而分析其中隐藏的规律性。大数据的第三个特点是数 据生成速度快。比如在传感器网络中，传感设备生成的数据，数量大，速度快，需要及 时处理。

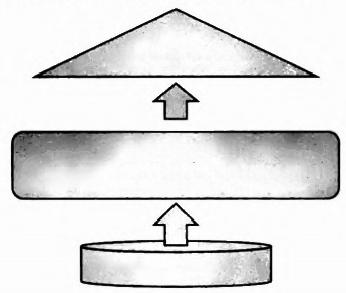
数据中蕴含着规律性，即数据中包含价值。数据的价值，在信息系统作为计算机科学 的一个重要的研究方向之初，就已经得到了人们的认可和重视。很多企业/机构对于收集 数据乐此不疲，究其原因是数据带来的价值(或者是潜在的价值)超出了它们收集数据和 管理数据的成本。在大数据时代，大数据特有的价值源于其规模效应，当数据量足够大 时，其价值能够产生从量变到质变的效应。

这一点在深度学习领域得到了很好的验证。2010年以来，在机器学习和人工智能领 域，研究者利用大数据更好地训练深度神经网络分类器，在语音识别和图像识别等应用领 域取得了巨大的进步。一直以来被认为人工智能难以攻破的围棋领域，新的人工智能技术 已经超越了人类。究其原因，是大规模数据，尤其是具有标注的大数据，使得人们构建更 精细化的分类器成为可能，让一度不被重视的神经网络分类器，甚至整个人工智能领域， 重新焕发了青春。这是对大数据进行深度分析，挖掘利用其价值的典型例子。

大数据及其分析(需要各种具体的机器学习与数据挖掘方法对数据进行分析)是数据 科学的有效组成部分。但是，我们不能把数据科学等同于大数据分析。从某种意义上讲， 大数据是一个相对的概念，今年的大数据，由于技术的进步，即计算机存储容量和计算能 力的增强，在不远的将来，不能再称为大数据了。数据科学研究的对象包括各种类型、各 种规模的数据，它除了研究具体的数据分析方法外，还要探讨更加根本的问题，即数据分 析处理的一些基本规律。

**1.2.2** **数据科学与基于数据的决策的关系**

基于数据的决策 (Data Driven Decision Making,DDD) 指的是人们基于数据分析的 结果进行决策，而不仅仅是基于直觉，拍脑袋进行决策。数据科学的目的是通过分析，让 我们理解数据并获得洞察力，它包含一系列的基本原则、过程、技术/方法和工具。由此 可见，数据科学是为基于数据的决策服务的，即我们从数据中挖掘其隐藏的模式，获得新 知，目的是指导我们新的行动。基于数据的决策见图1-2。



**基于数据的决策**

数据科学：

原则、过程、技术/方法和工具

数据

**图1-2** **基于数据的决策**



**1.3** **数据科学家**

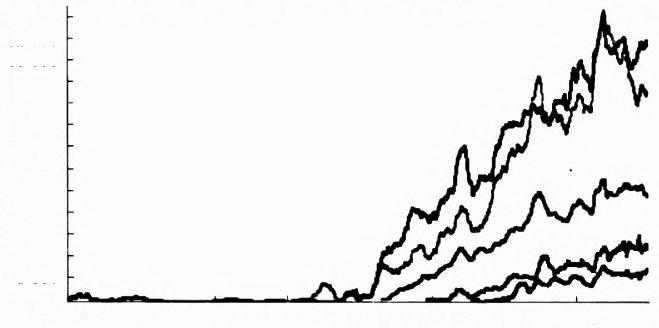
数据科学家是伴随着大数据技术的崛起和数据科学的兴起而出现的新的就业岗位。近年 来，对数据科学家的需求持续增长。数据科学家被誉为21世纪最性感的职业①,他们使用各 种技术，对不同来源的数据进行分析，帮助企业做出更加明智的决策。

图 1 - 3 展 示 的 是Indeed.com 发布的2006 — 2013年若干大数据开发人员岗位需求的 增长情况。图1 - 4展示的是Simple Hired 发布的2009—2011年若干大数据开发人员岗位

Job Trends from indeed.com

—mongodb—cassandra —hbase —couchbase —riak

Percentage Growth



13000

12000

11000

10000

9000

8000

7000

6000

5000

4000

3000

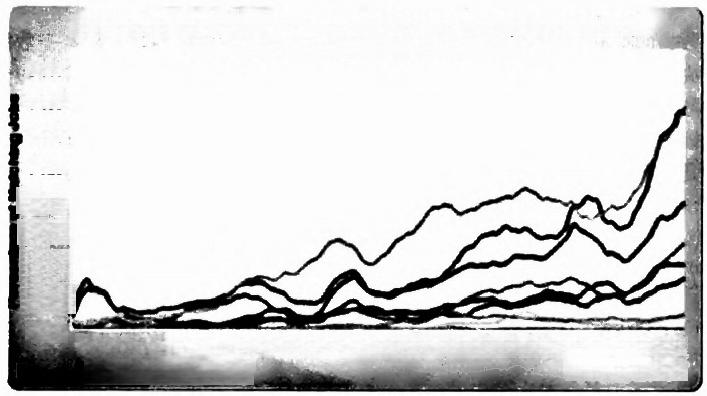
2000

1000

0

Jan'06 Jan'07 Jan'08 Jan'09 Jan'10 Jan'11 Jan'12 Jan'13

图 1 - 3 Indeed.com发布的2006—2013年若干大数据开发人员岗位需求的增长情况



JbTredo for -Cassandra -Redls -Voldemort 一Smpledb 一Couchdb

一 Mongodb 一 Hbase -Hypertable 一 Plak

**0040**

**0.035**

0.030

0.025

0020 D.01 0010 0005

D000 by0 10 Mer 10 10 M10 8p 10 oy 10 m11 11 y11

与 ， 0M Nc

图1-4 Simply Hired 发布的2009—2011年若干大数据开发人员岗位需求的增长情况



①[https://hbr.org/2012/10/data scientist-the sexiest-job-of-the-21st-century.](https://hbr.org/2012/10/datascientist-thesexiest-job-of-the-21st-century.)

需求的增长情况。来自O'Reilly 出版集团研究部门的报告显示，2009—2010年对 Hadoop 和Cassandra 的开发人员的需求呈现稳定增长的态势。图1-5展示了这一时期企业发布的 Cassandra 开发人员招聘岗位数量，以及发布这些招聘信息的企业数量的变化情况。这些 信息可以作为数据科学人才需求热度的冷暖表。

Cassandra jobs/companies

**一** **jobs**

**一** **companies**

6

2009-02009-03

2

2009-

心

2010- 010

**图1-5** **O’Reilly 发布的2009—2010年** **Cassandra开发人员岗位需求的增长情况**

著名咨询机构McKinsey Global Institute 预计，到2018年，美国面临14万～19万具 有数据深度分析技能 (Deep Analytical Skills) 的专业人员的缺口，此外，具有利用数据 科学提高决策效率技能的经理人的劳动力短缺则达到150万左右。 McKinsey 认为，是否 具备收集并挖掘数据、从中提取价值的能力，构成一个公司竞争力的关键因素。

**数据科学家的技能**

数据是新的石油，它正成为一种生产资料、稀有资产，它是重要的战略资源，全面融 入社会、生产、生活各个方面，深刻改变着世界的经济格局、利益格局、安全格局。数据 包含信息，可以为我们的决策服务。为了发挥数据的潜在价值，需要大量的数据科学家。 他们的工作是结合相关领域的背景知识，对数据进行建模、分析、展现等。

数据科学家需要什么样的具体技能呢?这是我们关心的问题，这也为数据科学概论这 门课程应该提供什么样的内容，提供一个指引。为了满足社会对数据科学人才的需求，许 多高校设立了专门的数据科学类专业，或在相关专业开设了数据科学类课程。数据科学专 业或课程，在高校中越来越受到学生的欢迎和重视。当然，数据科学不能落入讲授一些技 能的陷阱。

为了完成所赋予的数据分析任务，即利用实际数据解决实际问题，数据科学家需要拥 有一系列知识和技能，包括一定的数学基础，统计分析、机器学习、数据挖掘、数据可视 化方面的知识，编程能力，以及对具体应用领域(具体行业问题)的深入了解。此外，数 据科学家需要具有良好的沟通能力，能够和业务部门沟通，抽象其问题，并且把分析结果 清清楚楚地展示给业务部门，帮助他们做出决策。

**在这里，我们强调数据科学家需要具备跨学科的知识和技能，包括数学和统计分析、** **人工智能与机器学习、数据库与数据挖掘等。他们要具备数据处理全流程的能力，包括理** **解业务数据** **(Understand Data)、收集数据** **(Collect Data)、对数据进行集成** **(Integrate**

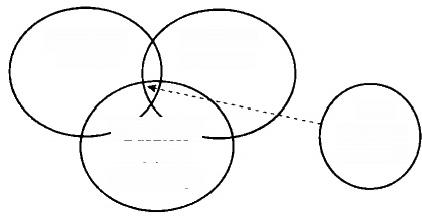


Data) 、 对数据进行分析挖掘 (Analyze 及把结果表达给目标听众 (Communicate



Data)、对结果进行可视化 (Visualize Result), 以 Result) 等。他们要善于讲故事，是一个好的讲

者 (a Good Story Teller)。他们能够设计(Design)、 实 现 (Implement, 即用具体的编程 语言实现)、评估 (Evaluate) 数据处理的工作流 (Data Science Workflow)。数据科学家 的技能见图1- 6。



Computer Science

Business

Domain

Knowledge

Data Scientist Skill Set

Math and Statistics

**图1-6** **数据科学家的技能**

数据科学家不能仅仅止于熟悉各种数据分析工具，他们对具体的业务领域要有深入的 了解，具备良好的沟通能力，能够问出正确的问题 (Ask the Right Question on Data), 并且运用其储备的数据分析知识和技能分析数据，并且评估这些分析的结果能否为提高业 务效率真正发挥作用，这正是发挥数据价值的关键所在。

**1.4** **数据科学的基本原则**

在本书中，我们试图总结数据科学的若干基本原则。这些原则在数据科学的实践中不 断重现，反过来可以用于指导新的数据科学实践。

**1.4.1** **原则1:数据分析可以划分成一系列明确的阶段**

分析数据，获得对数据和具体业务的理解，甚至获得知识，从而进一步指导我们解决 具体的业务问题，是数据科学的核心任务。数据分析可以划分成一系列的阶段，包括理解 业务数据、收集数据、对数据进行集成、对数据进行分析挖掘、对结果进行可视化，以及 把结果表达给目标听众等阶段。

把数据分析任务看作一个工作流 (Data Science Workflow), 划分成一系列明确的阶 段，是结构化地分析问题、解决问题的思想方法 (Structured Approach),使得我们不易 忽略重要的步骤。

**1.4.2** **原则2:描述性分析与预测性分析**

对数据进行分析有两个目的，即了解过去和预见未来。由此，数据分析分为两类，分 别是描述性分析和预测性分析。描述性分析侧重过去，预测性分析侧重未来。

(1)面向过去，发现隐藏在数据表面之下的历史规律或模式，这类分析称为描述性分

析 (Descriptive Analysis)。这些隐藏的模式可以帮助人们更好地进行决策。描述性分析 使用的技术包括统计分析方法和数据挖掘方法。

诊断型分析 (Diagnostic Analysis) 主要用来揭示一些现象背后的成因。可以看出， 它比描述性分析更为深入。由于我们在这里把分析仅仅分为两类，所以把诊断型分析纳入 描述性分析类别来把握。

很多数据挖掘方法为诊断型分析服务，比如相关性分析、因果关系分析等，都是要通 过对数据的深度分析，揭示某些现象背后的成因。比如， 一家互联网金融公司使用描述性 分析技术，发现其某类产品在过去一个月的销售出现明显下滑，原因是什么呢?分析人员 需要借助一些诊断型数据分析方法，了解造成销售下滑的主要原因。比如，到底是不是产 品定价影响了销售。一般来讲，分析人员在某些假设前提下，对不同维度的数据进行相关 性分析和因果性分析。更为复杂的，可能需要引入一些图分析技术。比如，通过对用户行 为图谱的分析，辨别是否存在群体性恶意抵制的行为，影响商品的销售。

(2)面向未来，对现有的(大)数据进行深度分析，构建分类/回归模型，对未来趋 势进行预测，称为预测性分析 (Predictive Analysis)。预测性分析使用的技术包括统计分 析的回归分析技术以及机器学习方法。特别需要注意的是，机器学习方法特别是深度学习 技术，适用于预测性的数据科学任务，但是不太适合描述性分析的任务，因为有些机器学 习模型缺乏可解释性。对于描述性分析任务，统计分析方法更为合适。把数据分析的范 围，从已知拓展到未知，从过去引申到将来，是数据科学的真正魅力所在。预测性模型的 运用，赋予我们预见的能力，我们可以预测非常有可能发生的事件，可以大幅度降低企业 运营成本，规避风险，以及提高客户的体验。

规范性分析 (Prescriptive Analysis) 是对描述性分析和预测性分析的整合，是数据分 析的高级阶段。规范性分析不仅要预测将要发生的事情，以及什么时候发生，而且给出事 情发生的原因。此外，规范性分析还要给出若干决策选项 (Decision Option),以及每个 决策选项的可能后果。为决策者提出建议，如何充分利用未来的机会，减轻未来的风险。 规范性分析把通过分析获得的知识提升到智慧的层次。规范性分析需要考虑很多因素，包 括期望的结果、所处的环境、资源条件等。

为了实现规范性分析，需要结合历史数据和新数据，持续分析各种各样的新数据，包 括结构化数据、非结构化数据，根据新老数据的分析结果以及商业规则，重新进行预测， 提高预测的准确性，以及提出更好的决策选项。

在很多社会治理问题上，规范性分析能够发挥巨大的作用。比如，交通拥堵是一个困 扰很多大城市的难题。规范性分析通过对城市交通数据、气象数据、就业数据、地理信息 等数据进行综合的分析和建模，从宏观层面到微观层面，制定合理有效的交通疏导策略， 以缓解城市的交通拥堵问题。

规范性分析使得数据分析系统所发现的策略在执行后，各项指标按事先预定的趋势发 展。当然，这是一个愿景，目前面临的研究挑战还非常大。

**1.4.3** **原则3:实体的相似度**

在数据科学中，我们经常要计算实体间的相似度。比如在推荐系统中，我们要计算用 户之间的相似度，或者计算商品之间的相似度。计算相似度是数据科学的基本方法。

在实际工作中.我们虽然为特定实体建立了高维的刻画模型，但还是有可能遗漏某些

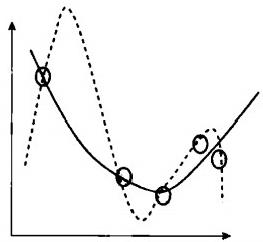
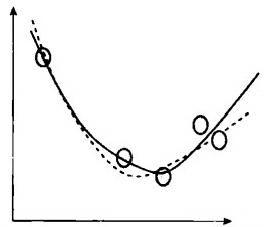
信息，没有完整地刻画客观对象。即便这样，我们还是有信心使用已有的属性信息，计算 实体之间的相似度。因为一般来讲，在一些属性上相似的实体，在其他属性上也是相似 的，这些属性可能是未知的属性，我们没有进行采集和数字化。

**1.4.4** **原则4:模型的泛化能力**

在机器学习中，我们使用历史数据训练出一个模型，这个模型有可能对历史数据适配 (Fit to Historical Data) 得很好，但是当我们把这个模型运用到新数据时，就要特别小心 了。因为这个模型没有见过新的数据，它的分类或者预测结果有可能很不理想，或者说它 的泛化能力欠佳。

在机器学习中，我们一般要避免模型对历史数据的过度匹配，这种现象称为过拟合 ((ver Fit)。过拟合导致模型的泛化能力差，也就是模型在新数据上的分类或者预测的效 果不好。

拟合与过拟合见图1-7。



**图1-7** **拟合与过拟合**

注：圆圈是数据点，实线为数据本来遵循的模型，虚线为学习到的模型。

**1.4.5** **原则5:分析结果的评估与特定应用场景有关**

对数据进行深入分析以后所获得的结果，是否具有实际应用价值，能否帮助我们做出 更好的决策?需要我们结合具体的应用场景来进行评估。数据科学家要和领域专家合作， 对数据分析结果进行评估。数据科学家自身也应该通过不断学习，对具体的业务领域有深 刻的理解，这样才能对数据分析结果有很好的把握。

数据科学家和用户一起工作，发现问题，对问题进行分析，建立模型和算法来解决问 题，并且负责把模型的结果解释给用户。沟通能力和数学、编程等能力一样重要。

**1.4.6** **原则6:相关性不同于因果关系**

从大量的基础数据中，我们可能分析出变量之间的相关性。相关性很有用，在一定程度 上可以帮助我们进行预测，但是相关性和因果关系有重大区别，相关性不意味着因果性。

因果关系的特征是原因在先，结果在后，前者的出现，必然导致后者。因果性是逻辑 上的概念，A 发生导致B 发生。相关性是统计上的概念，数据多了，A 发生时B 发生的概



率足够显著,那么A 和 B 就是相关的。

比如，我们看见闪电 (A) 和听见雷声 (B) 是高度相关的，但是两者并不具备因果

关系。闪电和雷声都有一个共同的原因，就是大气中的放电现象。

相关性不等于因果关系，重要的原因是一些额外因子的存在，而且我们先前没有考虑 进来。换句话说，当我们从数据的分析结果中，试图得出一些因果关系的结论时，必须考 虑到额外的因子。

额外因子或变量 (Confounding Factor/Variable) 是和目标变量相关的未知变量。由 于有额外因子的存在，相关性不等于因果关系。因为相关性有可能是由我们没有考虑到的 额外因子导致的，它有时候甚至违反我们的直觉。在数据建模和分析过程中，需要仔细检 查是否存在这样的额外因子。

比如.我们从一个人群的调查结果中发现，给他们使用某种治疗方案(服用特定药 物)之后，他们的症状减轻了。如果我们调查的人群是平均年龄为25岁的年轻人.对比 人群是平均年龄为75岁的老年人，老年人群服用安慰剂 (placebo), 那么年轻人症状的减 轻，有可能是因为他们年轻，恢复快，而不是因为服用了特定药物。这里，年龄差异是额 外因子，我们并没有考虑到。额外因子的作用见图1-8。

**What it look like**

**治疗方案** **(Treatment)**  **症状减轻** **(Less Disease)**

What is actually happening

治疗方案 (Treatment) 症状减轻 (Less Disease)

更年轻 (Younger)

图1-8 额外因子的作用

**1.4.7** **原则7:通过并行处理提高数据处理(分析)速度**

发明Pascal 语言的计算机科学家尼古拉斯 · 沃斯 (Niklaus Wirth) 提出了一个著名的公 式：程序=数据结构+算法。利用计算机进行数据处理，我们关心的不外乎两个方面，即数 据是什么,我们要对数据进行什么样的处理。在本书的后面部分.我们将围绕数据和数据分 析展开论述，这是数据科学的核心内容，可以说，数据科学=数据+数据上的计算。

通过并行处理，可以提高数据处理的速度。并行处理分为任务并行 (Task Parallel- ism) 和数据并行 (Data Parallelism) 两种类型。任务并行就是通过多个进程(正在运行 的应用程序),对数据进行处理，通过操作系统的多任务处理能力，提高数据处理的效率。 比如，不同储户对银行核心业务系统进行账户明细查询，各个储户对自己的账户进行查 询.互不干扰。我们在不同的进程中，运行用户的查询，提高处理速度，而不是一个储户 查询完成了另一个储户才能开始查询。需要注意的是，在某些场合，必须协调并发执行的 不同进程对数据的存取，才能保证数据的正确性。比如，有一个公共账户，储户A 查询现 有余额(数据库里)为10元钱，然后他把余额提取到内存(数据进入内存才能进行处 理),从余额10元钱里扣掉2元钱(在内存里进行操作),这时候储户B 从数据库里查询

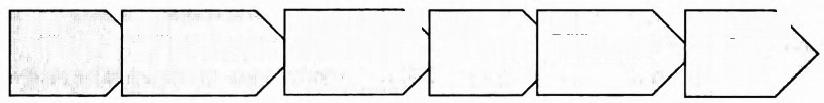
公共账户，得到的余额也是10元钱，他也从余额10元钱里扣掉2元钱，这时候，储户A 把扣掉2元钱以后的余额8元钱(还在内存里)写入数据库(数据库里),接着，储户B

把他扣掉2元钱以后的余额8元钱(还在内存里)也写入数据库(数据库里),数据库中 的公共账户余额最后是8元钱。显然，这个余额是错误的。数据库的并发控制 (Concur- rency Control) 方法，将用来协调不同进程对同一数据的读写操作，保证其正确性。数据 并行指的是把整个数据集(大规模),划分成一系列小的数据集，然后利用多个进程对这 些小的数据集进行并行操作，以达到提高数据处理速度的目的。比如，我们有一张储户 表，里面保存了10亿个储户，现在需要查找年龄为20～30岁的储户。我们可以从头到尾 扫描这张表，提取符合条件的储户。由于这张表实在太大，效率不高，于是我们把这张表 划分成100个分区 (Partition), 然后启动100个进程，每个进程负责扫描一个数据分区， 从中查找符合条件的储户信息，最后把100个进程的查询结果合并起来，就可以得到最终 结果。这些进程可以在一个节点(一台计算机)上运行，也可以分布到若干节点上运行， 当然数据也需要事先分布到若干节点上。

通过数据并行处理大规模的数据集，有时候并不像上述实例那么简单。比如，我们需 要在一个大数据集上进行统计模型的参数估计(或者训练一个机器学习模型)。如果我们 简单地把数据划分成若干分区，然后在这些分区上分别进行统计模型的参数估计(或者训 练机器学习模型),将得到若干组参数估计(若干机器学习模型),这不是我们想要的。对 估计出来的参数，简单进行平均不是办法，若干机器学习模型如何整合成一个统一的模 型，也不是那么容易。解决这个问题的一个办法是首先对数据进行洗牌 (Shuffle), 然 后 抽取一个具有代表性的采样 (Sample), 也就是保证这个数据抽样能够代表原有数据集的 分布特点。然后在数据抽样上运行参数估计算法，或者训练机器学习模型，我们将得到一 套参数和一个机器学习模型。但是这种办法有可能丢失数据的一些细节信息，参数估计不 准，或者训练出来的机器学习模型预测能力有限。另外一个办法是对串行的 (Serialized) 参数估计算法和机器学习算法进行改造，设计并行版本 (Parallelized Version),使之可以 在多个数据分区上运行。

**1.5** **数据处理流程：时间维度的纵向视角**

数据有完整的生命周期 (life cycle), 数据的生命周期包括数据的产生、数据的表示 和保存、数据的销毁等各个阶段。需要注意的是，随着存储设备价格的下降，我们有可能 无须销毁数据而是永远保存它，暂时不用的数据及时归档。近期的数据有可能经常用到称 为热数据， 一般保存在高速设备上，比如固态硬盘或者普通硬盘。暂时不用的历史数据称 为冷数据，则可以保存到低速设备上，比如普通硬盘或者磁带等设备。数据分析处理流程 见图1 - 9。

Data

Collection

Data

Representation & Storage

Data Cleaning &Integration

Data

Analyzing

Data

Visualization

Insights & Decision Making

**图1-9** **数据分析处理的流程**

伴随着数据整个生命周期的，是人们对数据的分析处理流程。在数据存续的整个生命 周期内，我们有可能对数据进行多次分析。分析处理流程划分成数据采集、表示和存储、 清洗和集成、分析、展现等主要的阶段。从生命周期的各个阶段观察数据，以及从数据分 析流程的各个阶段观察数据，是对数据进行观察的一种基于时间的、纵向的视角。

(1)数据采集：是把相关的业务数据采集保存起来。比如，当一个风力发电企业要对 风力发电机的健康状况进行持续监控，就需要把风力发电机的各类传感器的数据采集起 来；当一个电商要对客户进行画像 (Profiling), 就需要把客户在电商网站上的浏览、点 击、购买等行为记录下来。

(2)数据表示和存储：采集的数据需要以某种格式保存起来。当我们对现实世界的活 动进行数字化的记录和保存时，需要设计合适的表示形式。比如，对电商用户的购买行为 进行记录，需要记录客户ID 、时间、商品ID 、数量、单价等基本信息。大量客户的购买 记录组织在一块儿，就形成了一张二维表，每行对应一个购买记录，这是一种结构化的信 息。当我们对社交网络上用户的好友、关注等关系进行记录时，这种二维表的形式就不太 合适了，表达社交网络关系最自然的数据表示形式是图 (Graph) 。 一张图由一系列的节点 以及节点之间的边构成，比如社交网络中的用户表示为图的节点，用户之间的好友关系表 示为边。有时候图中的节点类型多样，图中的边也有多种类型，这是一种异构的图。比 如，在论文出版网络中，图中的节点有作者、论文、期刊/会议等，作者和论文之间有 “写作”的关系.论文和期刊/会议之间有“出版于”的关系。

(3)数据清洗：数据中的错误，对数据分析产生重大影响。低质量的数据导致低质量 的分析结果。数据清洗是想办法把数据中的错误尽量剔除掉。

(4)数据集成：是把不同来源的、异构的数据整合在一起。这些数据有可能类型不一 样，需要建立数据之间的关联性。如何找到不同类型数据集之间的联系，比如建立起用户 基本信息和社交网络信息之间的联系尤其关键。集成是数据在数量上的变化导致质量上的 变化。数据聚集和融合以后，它的价值在于给我们提供一种可能性，可以在更大的语境下 (Context) 分析世界万物之间的关系、模式。

(5)数据分析：是数据分析处理流程的重要步骤，利用统计分析、数据挖掘和机器学 习方法，对数据进行分析，获得分析结果。

(6)数据可视化：是把数据分析结果以图形的方式展现出来，方便最终用户(管理 者)理解和把握。

(7)基于数据的决策：通过对数据分析结果的解读.我们获得对数据的深入理解，进 而帮助我们在新的业务里进行科学的决策。

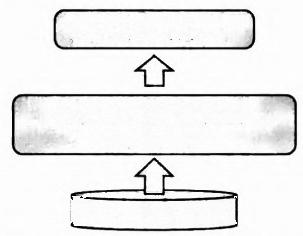
**1.6** **数据处理系统的架构：系统维度的计算视角**

**1.6.1** **数据处理系统的层次架构**

数据处理系统依赖于计算机系统的存储和计算能力建立。整个系统可以切分成数据 库、存储/检索与分析系统、应用系统(数据产品)等主要层次或者子系统(见图1-10)。



这是对数据处理系统进行观察的一种视角。



应用软件、数据产品

存储、检索、分析软件

(计算设备)

**数据库(存储设备)**

**图1-10** **数据处理系统的层次架构**

(1)数据库：数据保存在存储设备上。在这里，需要解决数据的表示和保存的问题， 即数据要以一定的逻辑结构和物理结构进行保存。比如，关于用户购买行为的信息，在逻 辑上是一张二维表，我们可以以行存储格式 (Row-wise Storage) 或者列存储格式 (Co- lumnar Storage) 进行物理存储。大量数据按照一定的结构保存在存储设备上，构成了一 个数据库。注意，不能把数据库等同于关系数据库，那是一种狭义的理解。

(2)数据的存储、检索以及分析：数据的存储和检索是对数据进行增加、删除、修 改、查询、简单汇总分析等操作。数据的分析指的是对数据进行深入的统计分析、挖掘， 甚至利用这些数据来训练一个机器学习模型，用于分类和预测任务。数据的存储、检索以 及分析，在软件形式上，表现为数据库管理系统(比如RDBMS) 、 大数据处理平台等(比 如 Hadoop) 形式。这些软件是系统级的软件，运行在一台计算机或者由多台计算机构成 的集群上。

(3)应用软件与数据产品：在数据库、数据存储/检索及分析系统软件上，可以针对 特定领域进行软件开发，方便用户使用，称为应用软件。 一些厂商在其掌握的数据上进行 分析，向客户销售分析结果，这种产品形式称为数据产品 (Data Product)。

由此可见，数据处理系统由硬件和软件构成，硬件包括存储设备和计算设备，软件则 包括数据库、数据存储/检索与分析软件，以及应用软件和数据产品。

**1.6.2 数据处理系统的Lambda架构**

在数据处理系统中，数据的处理模式分为三类，分别是批处理、流式数据处理以及交 互式处理。

(1)批处理：是把数据首先保存起来(把数据入库),然后进行分析。批处理模式处 理的数据量一般是全量数据(或者是绝大部分数据)。由于对大量数据进行分析，它的响 应时间比较慢， 一般以分钟和小时计。对于训练准确的机器学习模型，以及从数据中发现 隐藏的模式来讲，我们需要对大量的数据进行处理。

(2)流式数据处理：在这种处理模式下，新到达(新产生)的数据要及时进行处理， 无须首先入库，处理过后的数据一般不予保存，直接丢弃掉(当然也可以根据需要保存起 来，但这不属于流式数据处理的范畴)。流式数据处理对新到达的每个数据元素 (Data El- ement) 都立即处理，其响应时间以毫秒(甚至微秒)计。系统可以把最近的数据暂时保 存在内存中，在处理当前数据元素时做参考，但是参考的数据量一般较少。流式数据处理 系统一般用于监控、数据转换、数据装载等场合。比如，通过传感器采集到的设备运行参

数，可以持续监控设备的运行状态，这类应用要求比较强的实时性。由此可见，批处理模 式处理的数据是已经入库的数据 (Data at Rest), 流式数据处理模式处理的数据是运动中 的数据 (Data in Motion)。

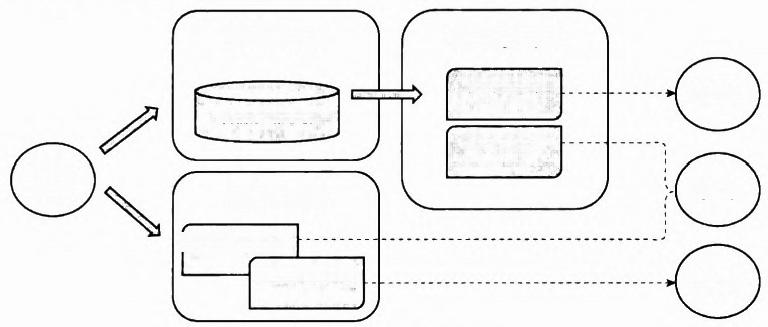
(3)交互式处理：交互式处理和批处理，都需要把数据首先保存起来，再进行处理。 交互式处理一般查询部分数据，进行简单的统计分析，它所存取的是整个数据集的一小部 分，因此，它的响应时间比批处理要快得多，一般达到秒级响应时间。

批处理、流式数据处理、交互式处理等处理模式各有侧重，满足不同应用场合的需 求，我们可以把这三种模式整合起来，这就是著名的Lambda 架构。

Lambda 架构分成三个层次 (Tier), 分别是批处理层 (Batch Layer)、实时处理层 (Speed Layer)、服务层 (Serving Layer)。从数据源不断到达的数据，经过适当的暂存 (Staging), 交给Batch Layer,一般要经过抽取、转换和装载 (Extraction Transformation and Loading,ETL) 的过程加载到数据库中，形成主数据集 (Master Dataset)。

此外，从数据源不断到达的数据，经过适当的暂存后，同时分发给Speed Layer。 Speed Layer 对数据的处理模式是一个一个数据元素单独处理，有可能参考近期少量的数 据元素 ， 老的历史数据就直接丢弃掉了 。

用户可以向系统发起查询(分析)请求，查询分为三类：(1)有些查询直接发送到 Speed Layer, 查询结果是最近数据的分析结果，比如传感器读数的最近变化情况。(2)有些 查询直接发送到Batch Iayer, 对历史数据进行分析，查询的响应时间要慢得多。(3)有些时 候，用户希望把历史数据的分析结果和最近数据的分析结果整合起来，获得更加完整的视 图。用户的查询需要分解成两个子查询，分别发送到Speed Layer和 Batch Layer, 查询处理 器把这两个查询的查询结果进行必要整合之后，呈现给用户，如图1 - 11所示。



Batch Layer

Batch view

query

Master dataset

Batch view

Speed Layer

Real-time view

Real-time view

query

Serving Layer

New Data

query

**图** **1** **-** **1** **1** **Lambda架构**

**1.7** **数据的多样性：数据类型维度的横向视角**

针对不同的应用，我们采集到的数据类型丰富多样，包括表格数据、 HTML 网页文 件、XML 文件、RDF(Resource Description Framework, 资源描述框架)数据、文本数 据、图(社交网络)数据、多媒体数据(音频/视频/图像)等。这些数据可以划分成结构

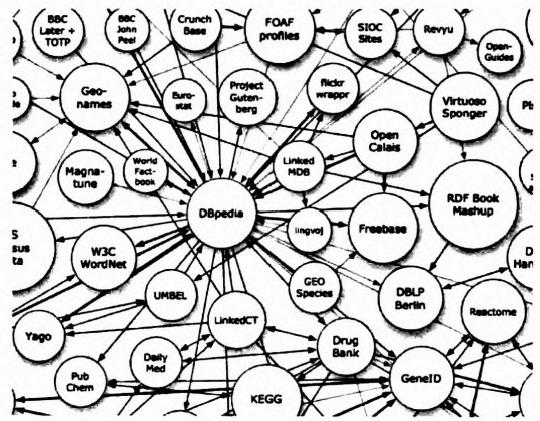
化数据、非结构化数据和半结构化数据等不同类型。结构化数据主要是指符合关系数据模 型的二维表数据，半结构化数据包括各种包含结构标记 (Tag) 的 HTML 网页、XML 文 档 、RDF 数据等，非结构化数据则包括文本数据、图数据以及各种多媒体数据。

当类型多样的数据描述的是现实世界中同样的实体、事件时，便具有内在的联系，必 须建立它们之间的关联，以便实现跨媒体的数据分析。多种类型的数据，描述客观世界同 样的对象，是透过数据观察世界的一种多通道的视角，是一种360°的全面视角。

比如，社交网络的用户具有一些基本的属性，这些属性保存在关系数据库表中，同时 他们拥有一些好友关系，这些关系保存在图数据库中。当关系数据库表中的记录以及图数 据库中的节点表示同一个用户时，就有必要建立关联。

又比如，近年来，人们建立了大量的知识库，包括通用知识库和专用知识库，比如 DBpedia,Freebase,Geo-names 等。这些知识库之间的实体是互相覆盖的，即一些实体， 既在这个知识库中出现，又在另外一个知识库中出现。建立不同知识库中实体的关联性， 使我们正确地解析实体，利用各个知识库互为补充，丰富实体的信息。

Linked Open Data项 目 (<http://linkeddata.org/>) 就是这样的一种尝试，它试图通 过Web 把相关的知识库连接在一起。它提出了一整套办法，把各个独立的知识库在互联 网上分享出来，使用URI(Uniform Resource Identifier, 统一资源标识符)和RDF 技术， 把语义网 (Semantic Web) 上的数据、信息和知识连接起来。 Linking Open Data项目的 云图见图1-12。

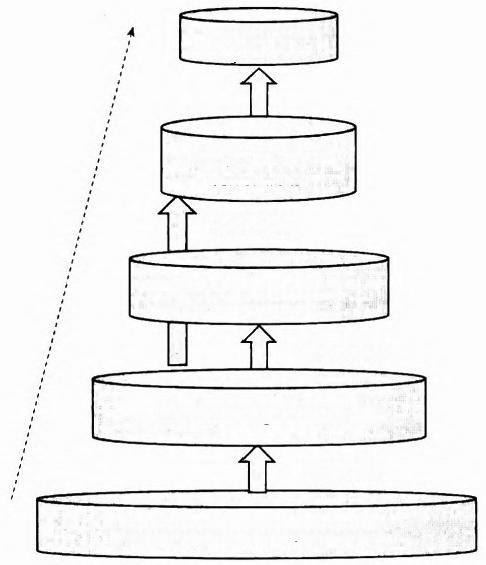


**图1-122009年7月的Linking** **Open** **Data(LOD)项目的云图(Cloud** **Diagram)局部**

注：完整版请参见<http://lod-cloud.net/versions/2009-07-14/Lod-cloud.png。>

**1.8数据价值的挖掘：价值维度的价值提升视角**

数据的价值发现过程和具体的应用场景密切相关。对于不同的应用来讲，价值提升的 过程具有共性，如图1-13所示。



Knowledge, high value

Insights :pattern, correlations

复杂分析

Metrics &report

简单分析

Cleaned,integrated data

清洗集成

Raw data:voluminous,low value



知识库

**分析软件**

**报表软件**

数据仓库

分布式文件系统/ 分布式数据库

**图1-** **13** **数据的价值提升过程**

首先，原始数据记录了现实世界的实体和事件， 一般数据量较大，但是数据的价值密 度低，有可能包含很多的噪声(即错误数据)。这些数据一般存放在具有高度扩展能力的 分布式文件系统或分布式数据库中。

这些数据必须经过清洗，以便剔除错误，提高数据的质量。此外，来自不同来源的数 据，需要集成起来，删除重复数据。多源异构数据还要建立数据之间的关联。我们掌握的 数据越全面、越多样，分析结果越有可能反映客观实际。经过清洗和集成的数据， 一般存 放到数据仓库中。

对数据进行分析的方法，根据分析的复杂度，分为简单分析和复杂分析。简单分析就 是对数据进行多维的汇总统计、生成报表等操作。 一般使用具有多维分析和可视化功能的 报表软件实现数据的分析，分析结果可以保存到数据库中，方便再次使用，无须从头 计算。

复杂分析则包括运用统计分析方法、数据挖掘方法、机器学习方法，对数据进行深入 分析。通过适当的分析，我们可以挖掘到数据中隐藏的模式、相关性等。数据的复杂分析 依赖于通用的分析软件，或者针对用户应用进行定制开发的分析软件实现，分析结果可以 保存到数据库中备查。

如果数据中反复出现一些模式，我们可以在此基础上抽象出知识。知识是比模式、相 关性等更加具有普遍性 (Generalized) 的规律。知识放在知识库中。

数据价值提升或者价值挖掘的过程，伴随着数据(信息)规模的缩小和数据(信息) 价值密度的提高。通过收集大量的数据并进行分析，原始的数据就变成了对人们有用的 信息。

**1.9** **思考题**

(1)数据科学的定义。

(2)数据科学和统计学、人工智能/机器学习、数据挖掘、数据库与数据处理、大数 据分析、基于数据的决策的关系。

(3)数据科学家的技能要求 (Skill Set)。

(4)数据科学的基本原则 (Principle)。

(5)数据的生命周期、数据分析处理的流程(时间维度的视角)。

(6)数据处理系统的层次架构、数据处理的不同模式以及Lambda 架构。

(7)数据的多样性(横向视角)。

(8)数据价值的挖掘(价值提升视角)。