# 第3章

# 大数据分析概论

## 导学

### 内容与要求

本章主要介绍大数据分析的基本方法和流程、大数据分析的主要技术及分析系统，以及实际应用情况，使读者对大数据分析有概括性的了解和掌握。

“大数据分析简介”一节要求理解大数据分析；掌握大数据分析的基本方法及流程。

“大数据分析的主要技术”一节要求熟悉主要的大数据分析技术，并对它们的作用有所了解。

“大数据分析处理系统”一节要求掌握4种类型大数据的特点，了解典型分析处理系统。

“大数据分析的应用”一节要求对网络与医学大数据的分析有所了解。

### 重点、难点

本章重点是大数据分析的方法、流程、主要技术和典型分析系统，难点是理解大数据分析的主要技术。

大数据分析就是研究包含各种数据类型的大型数据集的过程。大数据技术可以发现隐藏的数据模式、未知数据的相关性、市场趋势、客户喜好和其他有用的商业信息，其分析结果可以带来更有效的市场营销、新的收入机会、更好的客户服务、提高运营效率、获得竞争优势和其他商业利益。

# 3.1 大数据分析简介

在方兴未艾的大数据时代，人们要掌握大数据分析的基本方法和分析流程，从而探索出大数据中蕴含的规律与关系，解决实际业务问题。

## 3.1.1 大数据分析

大数据分析是指对规模巨大的数据进行分析。通过多个学科技术的融合，实现数据的采集、管理和分析，从而发现新的知识和规律。大数据时代的数据分析首先要解决的是海量、结构多变、动态实时的数据存储与计算问题，这些问题在大数据解决方案中至关重要，决定大数据分析的最终结果。

通过美国福特公司利用大数据分析促进汽车销售的案例，可以初步认识大数据分析。分析过程如图 3-1 所示。


图 3-1 福特促进汽车销售的大数据分析流程

**1. 提出问题**

用大数据分析技术来提升汽车销售业绩。一般汽车销售商的普通做法是投放广告，动辄就是几百万，而且很难分清广告促销的作用到底有多大。大数据技术不一样，它可以通过对某个地区可能会影响购买汽车意愿的源数据进行收集和分析，从而获得促进销售的解决方案。

**2. 大数据采集**

分析团队搜索采集数据，如这个地区的房屋市场、新建住宅、库存和销售数据、就业率等；还可利用与汽车相关的网站上的数据，如客户搜索了哪些汽车、哪一种款式、汽车的价格、车型配置、汽车功能、汽车颜色等；再有获取第三方合同网站、区域经济数据等。

**3. 大数据分析**

对采集的数据进行分析挖掘，为销售提供精准可靠的分析结果，即提供多种可能的促销分析方案。

**4. 大数据可视化**

根据数据分析结果实施有针对性的促销计划，如在需求量旺盛的地方有专门的促销计划，哪个地区的消费者对某款汽车感兴趣，相应广告就送到其电子邮箱和地区的报纸上，非常精准，只需要较少费用。

**5. 效果评估**

跟传统的广告促销相比，通过大数据的创新营销，福特公司花了很少的钱，做了大数据分析产品，也可叫大数据促销模型，大幅度地提高了汽车的销售业绩。

### 3.1.2　大数据分析的基本方法

大数据分析可以分为以下 5 种基本方法。

**1. 预测性分析**

大数据分析最普遍的应用就是预测性分析，从大数据中挖掘出有价值的知识和规则，通过科学建模的手段呈现出结果，然后可以将新的数据带入模型，从而预测未来的情况。

例如，麻省理工学院的研究者创建了一个计算机预测模型来分析心脏病患者丢弃的心电图数据。他们利用数据挖掘和机器学习在海量的数据中筛选，发现心电图中出现三类异常者一年内死于第二次心脏病发作的机率比未出现者高 1～2 倍。这种新方法能够预测出更多的、无法通过现有的风险筛查被探查出的高危病人，如图 3-2 所示。

**2. 可视化分析**

不管是对数据分析专家还是普通用户，他们二者对于大数据分析最基本的要求就是可视化分析，因为可视化分析能够直观地呈现大数据特点，同时能够非常容易地被地用户所接受。可视化可以直观地展示数据，让数据自己说话，让观众听到结果，数据可视化是数据分析工具最基本的要求。如图 3-3 所示是报纸发行量的可视化分析。图 3-4 所示是超市开业情况的地理位置可视化分析。

图 3-2　心电图大数据分析


图 3-3　北京日报发行量数据分析

**3. 大数据挖掘算法**

可视化分析结果是给用户看的，而数据挖掘算法是给计算机看的，通过让机器学习算法，按人的指令工作，从而呈现给用户隐藏在数据之中的有价值的结果。大数据分析的理论核心就是数据挖掘算法，算法不仅要考虑数据的量，也要考虑处理的速度，目前在许多领域的研究都是在分布式计算框架上对现有的数据挖掘理论加以改进，进行并行化、分布式处理。

常用的数据挖掘方法有分类、预测、关联规则、聚类、决策树、描述和可视化、复杂数据类型挖掘（Text、Web、图形图像、视频、音频）等，有很多学者对大数据挖掘算法进行了研究和


图 3-4 超市新店开业数据分析

文献发表。例如，有文献提出了对适合慢性病分类的 C4.5 决策树算法进行改进，对基于 MapReduce 编程框架进行算法的并行化改造；有文献提出对数据挖掘技术中的关联规则算法进行研究，并通过引入了兴趣度对经典 Apriori 算法进行改进，提出了一种基于 MapReduce 的改进的 Apriori 医疗数据挖掘算法。

**4. 语义引擎**

数据的含义就是语义。语义技术是从词语所表达的语义层次上来认识和处理用户的检索请求。

语义引擎通过对网络中的资源对象进行语义上的标注以及对用户的查询表达进行语义处理，使得自然语言具备语义上的逻辑关系，能够在网络环境下进行广泛有效的语义推理，从而更加准确、全面地实现用户的检索。大数据分析广泛应用于网络数据挖掘，可从用户的搜索关键词来分析和判断用户的需求，从而实现更好的用户体验。

例如，一个语义搜索引擎试图通过上下文来解读搜索结果，它可以自动识别文本的概念结构。如有人搜索“选举”，语义搜索引擎可能会获取包含“投票”、“竞选”和“选票”的文本信息，但是“选举”这个词可能根本没有出现在这些信息来源中，也就是说语义搜索可以对关键词的相关词和类似词进行解读，从而扩大搜索信息的准确性和相关性。

**5. 数据质量和数据管理**

数据质量和数据管理是指为了满足信息利用的需要，而对信息系统的各个信息采集点进行规范，包括建立模式化的操作规程、原始信息的校验、错误信息的反馈、矫正等一系列的过程。大数据分析离不开数据质量和数据管理，高质量的数据和有效的数据管理，无论是在学术研究还是在商业应用领域，都能够保证分析结果的真实和有价值。

### 3.1.3 大数据处理流程

整个处理流程可以分解为定义问题、数据理解、数据采集、数据预处理、数据分析、分析结果解析等，具体如图 3-5 所示。

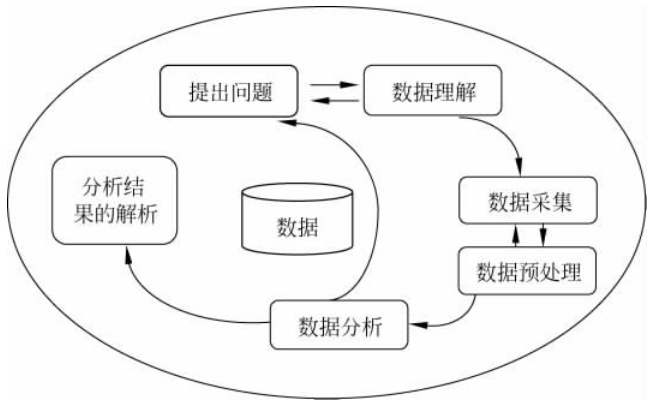


图 3-5　大数据分析处理流程图

**1. 提出问题**

大数据分析就是解决具体业务问题的处理过程，这需要在具体业务中提炼出准确的实现目标，也就是首先要制定具体需要解决的问题，如图 3-6 所示。


图 3-6　提出问题的过程

**2. 大数据理解**

大数据分析是为了解决业务问题，理解问题要基于业务知识，数据理解就是利用业务知识来认识数据。如大数据分析“饮食与疾病的关系”、“糖尿病与高血压的发病关系”，这些分析都需要对相关医学知识有足够的了解才能理解数据并进行分析。只有对业务知识有深入的理解，才能在大数据中找准分析指标和进一步衍生出来的指标，从而抓住问题的本质，挖掘出有价值的结果，如图 3-7 所示。

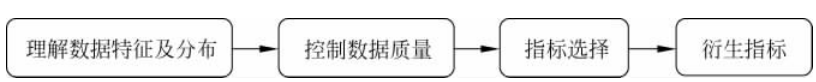


图 3-7　理解数据的过程

**3. 大数据的采集**

传统的数据采集来源单一，且存储、管理和分析数据量也相对较小，大多采用关系型数据库和并行数据仓库即可处理。大数据的采集可以通过系统日志采集方法、对非结构化数据采集方法、企业特定系统接口等相关方式采集，如用户利用多个数据库来接收来自客户端（Web、App 或者传感器等）的数据。

**4. 大数据的预处理**

如果要对海量数据进行有效的分析，应该将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，并且可以在导入基础上做一些简单的清洗和预处理工作。也有一些用户会在导入时对数据进行流式计算，来满足部分业务的实时计算需求。导入与预处理过程的特点和挑战主要是导入的数据量大，每秒钟的导入量经常会达到百兆，甚至千兆级别。

**5. 大数据分析**

大数据分析包括对结构化、半结构化及非结构化数据的分析，主要利用分布式数据库，或者分布式计算集群来对海量数据进行分析，如分类汇总、基于各种算法的高级别计算等，涉及的数据量和计算量都很大。

**6. 大数据分析结果的解析**

对用户来讲，最关心的是数据分析结果与解析，对结果的理解可以通过合适的展示方式，如可视化和人机交互等技术来实现。

## 3.2 大数据分析的主要技术

大数据分析的主要技术有深度学习、知识计算及可视化等，深度学习和知识计算是大数据分析的基础，而可视化在数据分析和结果呈现的过程中均起作用(关于可视化的具体处理方法见第 4 章)。

### 3.2.1 深度学习

**1. 深度学习的概念**

深度学习是一种能够模拟出人脑的神经结构的机器学习方式，从而能够让计算机具有人一样的智慧。其利用层次化的架构学习出对象在不同层次上的表达，这种层次化的表达可以帮助解决更加复杂抽象的问题。在层次化中，高层的概念通常是通过低层的概念来定义的，深度学习可以对人类难以理解的底层数据特征进行层层抽象，从而提高数据学习的精度。让计算机模仿人脑的机制来分析数据，建立类似人脑的神经网络进行机器学习，从而实现对数据有效的表达、解释和学习，这种技术在将来无疑是前景无限的。

**2. 深度学习的应用**

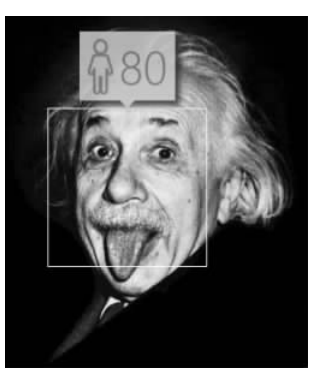


图 3-8 人脸识别判断年龄

近几年，深度学习在语音、图像以及自然语言理解等应用领域取得一系列重大进展。在自然语言处理等领域主要应用于机器翻译以及语义挖掘等方面，国外的 IBM、Google 等公司都快速地进行了语音识别的研究；国内的阿里巴巴、科大讯飞、百度、中科院自动化所等公司或研究单位，也在进行深度学习在语音识别上的研究。

深度学习在图像领域也取得了一系列进展。如微软推出的网站 how-old，用户可以上传自己的照片“估龄”。系统根据照片会对瞳孔、眼角、鼻子等 27 个“面部地标点”展开分析，判断照片上人物的年龄，如图 3-8 所示。

**举例**：德国用深度学习算法让人工智能系统学习绘画。

2015年德国一个综合神经科学研究所用深度学习算法让人工智能系统学习梵高、莫奈等世界著名画家的画风绘制新的“人工智能世界名画”。他们在视觉感知的关键领域，如物体和人脸识别等方面有了新的解决方法，这就是深层神经网络。基于深层神经网络的人工智能系统提供了绘画模仿，提供了神经创造艺术形象的算法，用以理解和模拟人类去创建和感知艺术形象。该算法是卷积神经网络算法，模拟人类大脑处理视觉时的工作状态，在目标识别方面较其他可用算法甚至人类专家更好。

图3-9是德国一个小镇的原始照片，图3-10、图3-11和图3-12的左下角显示的是名画原作，右侧是人工智能学习后变形的图3-9图片效果。

图3-9　德国小镇一瞥

图3-10　特纳弥诺陶洛斯的沉船风格的小镇

图3-11　梵高的星夜风格的小镇

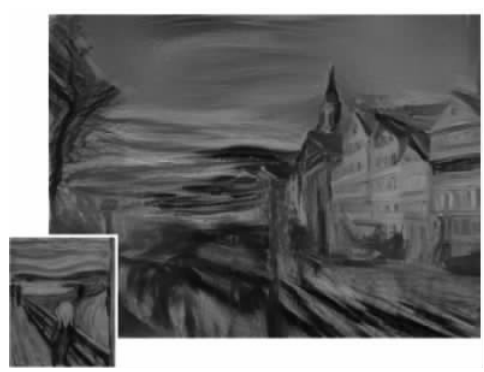

图3-12　爱德华·蒙克的呐喊风格的小镇

以上这些图像结合了一些著名的艺术绘画风格，这些图像被创建时，首先学习艺术品的内容表示和风格表示，然后应用在给定的图3-9中，并进行重新排列组合进行相似性视觉对比绘画，形成人工智能版的世界名画。

## 3.2.2　知识计算

### 1. 知识计算的概念

知识计算是从大数据中首先获得有价值的知识，并对其进行进一步深入的计算和分析

的过程。也就是要对数据进行高端的分析，需要从大数据中先抽取出有价值的知识，并把它构建成可支持查询、分析与计算的知识库。知识计算是目前国内外工业界开发和学术界研究的一个热点。知识计算的基础是构建知识库，知识库中的知识是显式的知识。通过利用显式的知识，人们可以进一步计算出隐式知识。知识计算包括属性计算、关系计算、实例计算等。

**2. 知识计算的应用**

目前，世界各个组织建立的知识库多达 50 余种，相关的应用系统更是达到了上百种。如维基百科等在线百科知识构建的知识库 DBpedia、YAG、Omega、WikiTaxonomy；Google 创建了至今世界最大的知识库，名为 Knowledge Vault，它通过算法自动搜集网上信息，通过机器学习把数据变成可用知识，目前，Knowledge Vault 已经收集了 16 亿件事件。知识库除了改善人机交互之外，也会推动现实增强技术的发展，Knowledge Vault 可以驱动一个现实增强系统，让人们从头戴显示屏上了解现实世界中的地标、建筑、商业网点等信息。

知识图谱泛指各种大型知识库，是把所有不同种类的信息连接在一起而得到的一个关系网络。这个概念最早由 Google 提出，提供了从“关系”的角度去分析问题的能力，知识图谱就是机器大脑中的知识库。

在国内，中文知识图谱的构建与知识计算也有大量的研究和开发应用，图 3-13 是心房颤动知识图谱，图 3-14 是心肌炎知识图谱。具有代表性的有中国科学院计算技术研究所的 OpenKN、中国科学院数学研究院提出的知件（Knowware）、上海交通大学最早构建的中文知识图谱平台 zhishi. me、百度推出了中文知识图谱搜索、搜狗推出的知立方平台、复旦大学 GDM 实验室推出的中文知识图谱展示平台等，这些知识库必将使知识计算发挥更大的作用。

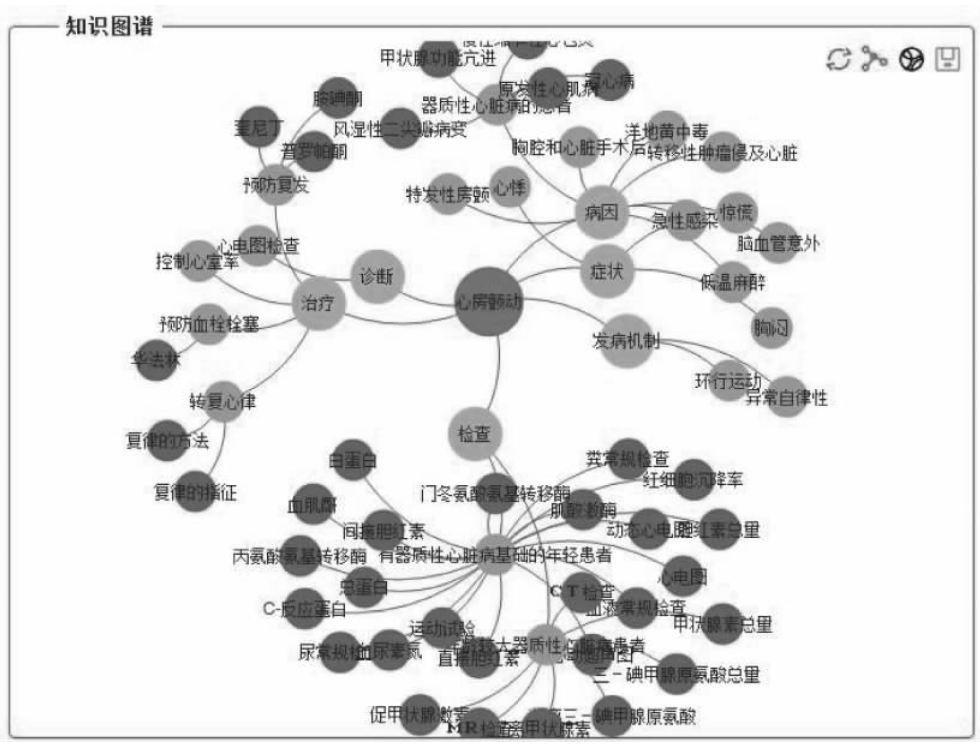


图 3-13　心房颤动知识图谱

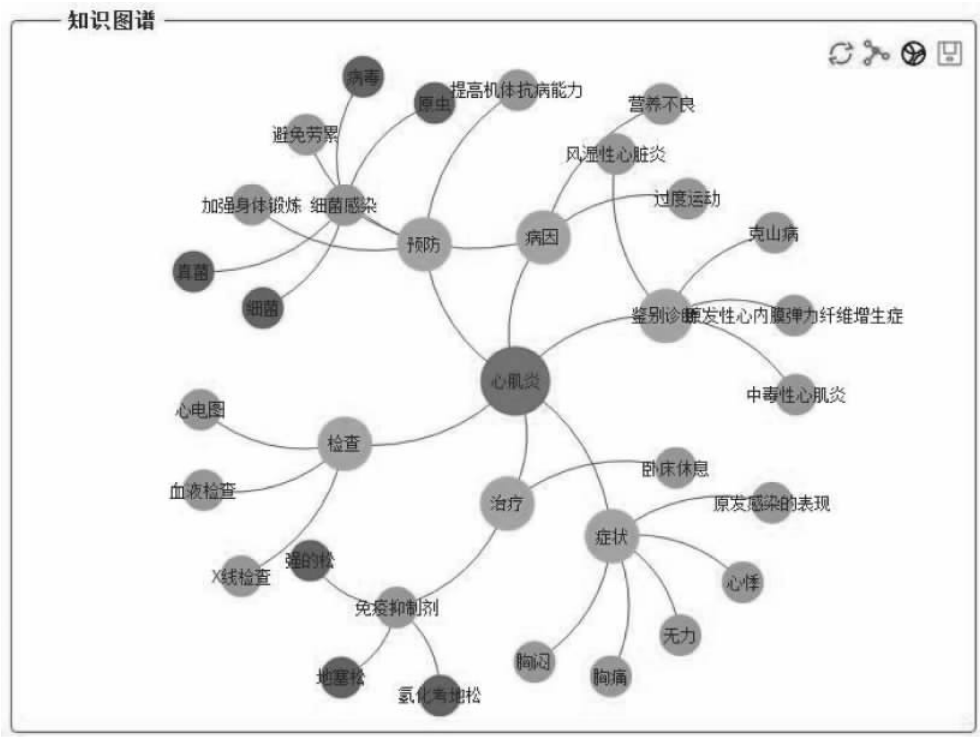


图 3-14　心肌炎知识图谱

## 3.3　大数据分析处理系统简介

针对不同业务需求的大数据，应采用不同的分析处理系统。国内外的互联网企业都在基于开源性面向典型应用的专用化系统进行开发。

### 3.3.1　批量数据及处理系统

**1. 批量数据**

批量数据通常是数据体量巨大，如数据从 TB 级别跃升到 PB 级别，且是以静态的形式存储，这种批量数据往往是从应用中沉淀下来的数据，如医院长期存储的电子病历等。对这样数据的分析通常使用合理的算法，才能进行数据计算和价值发现。大数据的批量处理系统适用于先存储后计算，实时性要求不高，但数据的准确性和全面性要求较高的场景。

**2. 批量数据分析处理系统**

Hadoop 是典型的大数据批量处理架构，由 HDFS（Hadoop Distributed File System，Hadoop 分布式文件系统）负责静态数据的存储，并通过 MapReduce 将计算逻辑、机器学习和数据挖掘算法实现。关于 Hadoop 与 MapReduce 的具体处理流程和方法见本书第 5 和 7 章。

### 3.3.2 流式数据及处理系统

**1. 流式数据**

流式数据是一个无穷的数据序列，序列中的每一个元素来源不同，格式复杂，序列往往包含时序特性。在大数据背景下，流式数据处理常见于服务器日志的实时采集，将 PB 级数据的处理时间缩短到秒级。数据流中的数据格式可以是结构化的、半结构化的甚至是非结构化的，数据流中往往含有错误元素、垃圾信息等，因此流式数据的处理系统要有很好的容错性及不同结构的数据分析能力，还要完成数据的动态清洗、格式处理等。

**2. 流式数据分析处理系统**

流式数据处理有 Twitter 的 Storm、Facebook 的 Scribe、Linkedin 的 Samza 等。其中 Storm 是一套分布式、可靠、可容错的用于处理流式数据的系统，其流式处理作业被分发至不同类型的组件，每个组件负责一项简单的、特定的处理任务。

Storm 系统有其独特的特性。

(1) 简单的编程：类似于 MapReduce 的操作，降低了并行批处理与实时处理的复杂性。

(2) 容错性：如果出现异常，Storm 将以一致的状态重新启动处理以恢复正确状态。

(3) 水平扩展：其流式计算过程是在多个线程和服务器之间并行进行的。

(4) 快速可靠的消息处理：Storm 利用 ZeroMQ 作为消息队列，极大地提高了消息传递的速度，任务失败时，它会负责从消息源重试消息。

### 3.3.3 交互式数据及处理系统

**1. 交互式数据**

交互式数据是操作人员与计算机以人机对话的方式产生的数据，操作人员提出请求，数据以对话的方式输入，计算机系统便提供相应的数据或提示信息，引导操作人员逐步地完成所需的操作，直至获得最后处理结果。交互式数据处理灵活、直观、便于控制，采用这种方式，存储在系统中的数据文件能够被及时地处理修改，同时处理结果可以立刻被使用。

**2. 交互式数据分析处理系统**

交互式数据处理系统有 Berkeley 的 Spark 和 Google 的 Dremel 等。Spark 是一个基于内存计算的可扩展的开源集群计算系统。关于 Spark 的详细介绍见本书第 9 章。

### 3.3.4 图数据及处理系统

**1. 图数据**

图数据是通过图形表达出来的信息含义。图自身的结构特点可以很好地表示事物之间的关系。图数据中主要包括图中的节点以及连接节点的边。在图中，顶点和边实例化构成各种类型的图，如标签图、属性图、特征图以及语义图等，如图 3-15、图 3-16、图 3-17 和图 3-18 所示。


图 3-15　价格标签图


图 3-16　服装颜色属性图


图 3-17　自然特征图

图 3-18　人脑语义地图

**2. 图数据分析处理系统**

图数据处理有一些典型的系统，如 Google 的 Pregel 系统、Neo4j 系统和微软的 Trinity 系统。Trinity 是 Microsoft 推出的一款建立在分布式云存储上的计算平台，可以提供高度并行查询处理、事务记录、一致性控制等功能。Trinity 主要使用内存存储，磁盘仅作为备份存储。

Trinity 有以下特点。

（1）数据模型是超图：超图中，一条边可以连接任意数目的图顶点，此模型中图的边称为超边，超图比简单图的适用性更强，保留的信息更多。

（2）并发性：可以配置在一台或上百台计算机上，提供了一个图分割机制。

（3）具有数据库的一些特点：是一个基于内存的图数据库，有丰富的数据库特点。

（4）支持批处理：支持大型在线查询和离线批处理，并且支持同步和不同步批处理计算。

## 3.4　大数据分析的应用

大数据分析有广泛的应用，以下从互联网和医疗领域为例，介绍大数据的应用。

**1. 互联网领域大数据分析的典型应用**

（1）用户行为数据分析。如精准广告投放、行为习惯和喜好分析、产品优化等。

(2) 用户消费数据分析。如精准营销、信用记录分析、活动促销、理财等。

(3) 用户地理位置数据分析。如 O2O(Online To Offline,在线离线/线上到线下)推广、商家推荐、交友推荐等。

(4) 互联网金融数据分析。如 P2P(Peer-To-Peer)、小额贷款、支付、信用、供应链金融等。

(5) 用户社交等数据分析。如流行元素分析、舆论监控分析、社会问题分析等。

**2. 医疗领域大数据分析的典型应用**

(1) 公共卫生:分析疾病模式和追踪疾病暴发及传播方式途径,提高公共卫生监测和反应速度。更快更准确地研制靶向疫苗,如开发每年的流感疫苗。

(2) 循证医学:分析各种结构化和非结构化数据,如电子病历、财务和运营数据、临床资料和基因组数据,从而寻找与病症信息相匹配的治疗方案、预测疾病的高危患者或提供更多高效的医疗服务。

(3) 基因组分析:更有效和低成本的执行基因测序,使基因组分析成为正规医疗保健决策的必要信息并纳入病人病历记录。

(4) 设备远程监控:从住院和家庭医疗装置采集和分析实时大容量的快速移动数据,用于安全监控和不良反应的预测。

(5) 病人资料分析:全面分析病人个人信息,找到能从特定保健措施中获益的个人。

(6) 疾病预测:如预测特定病人的住院时间,哪些病人会选择非急需性手术,哪些病人不会从手术治疗中受益,哪些病人会更容易出现并发症等。

(7) 临床操作:相对更有效的医学研究,发展出临床相关性更强和成本效益更高的方法用来诊断和治疗病人。

**3. 应用案例:某互联网公司对用户行为数据进行实时分析**

分析步骤如下。

(1) 首先提出分析方案:制定测试分析策略,数据来源于网站用户行为数据,数据量是90 天细节数据约 50 亿条。

(2) 简单测试:先通过 5 台 PC Server,导入 1～2 天的数据,演示如何 ETL(见注释),如何做简单应用。

(3) 实际数据导入:按照制定的测试方案,开始导入 90 天的数据,在导入数据中解决如下问题:①解决步长问题(每次导入记录条数),有效访问次数。②解决 HBase 数据和 SQLServer 数据的关联问题等。

(4) 数据源及数据特征分析。

90 天的数据量:Web 数据 7 亿条,App 数据 37 亿条,总估计在 50 亿条。

每个表有 20 多个字段,一半字符串类型,一半数值类型,一行数据估计 2000B。

每天导入 5000 万行,约 100G 存储空间,100 天是 10T 的数据量。

50 亿条数据若全部导入需要 900G 的存储量(压缩比在 11∶1)。

假设同时装载到内存中分析的量在 1/3,那总共需要 300G 的内存。

(5) 硬件设计方案。

总共配制需要 300G 的内存。5 台 PC Server,每台内存:64G,4CPU 4Core。

机器角色:一台 Naming、Map,一台 Client、Reduce、Map,其余三台都是 Map。

(6) ETL(Extract Transform Load)过程(将数据从来源端经过抽取、转换、加载至目的端的过程)。

历史数据集中导：每天的细节数据和 SQL Server 关联后，打上标签，再导入集市。

增量数据自动导：每天导入数据，生成汇总数据。

维度数据被缓存：细节数据按照日期打上标签，跟缓存的维度数据关联后入集市。

(7) 系统配置：系统内部管理、内存参数等配置。

(8) 互联网用户行为分析结果。

浏览器分析：运行时间、有效时间、启动次数、覆盖人数等。

主流网络电视：浏览总时长、有效流量时长、浏览次(PV)数覆盖占有率、1 天内相同访客多次访问网站、只计算为 1 个独立访客(UV)占有率等。

主流电商网站：在线总时长、有效在线总时长、独立访问量、网站覆盖量等。

主流财经网站：在线总时长、有效总浏览时长、独立访问量、总覆盖量等。

(9) 技术上分析测试结果。

90 天数据，近 10T 的原始数据，大部分的分析查询都是被秒级响应。

实现了 Hbase 数据与 SQLServer 中维度表关联分析的需求。

(10) 分析测试的经验总结。

由于事先做了预算限制，投入并不大，并且解决了 Hive 不够实时的问题。有关 Hive，请参考 5.2 节。

## 本章小结

大数据分析为处理结构化与非结构化的数据提供了新的途径，这些分析在具体应用上还有很长的路要走，在未来的日子里将会看到更多的产品和应用系统在生活中出现。通过本章内容的学习，学生应该学会大数据分析的方法，掌握大数据分析的一般流程与主要技术，为大数据的分析应用奠定基础。

## 习题 3

### 一、填空题

1. 大数据分析是指_____。

2. 大数据分析的基本方法有预测性分析、可视化分析、_____、语义引擎、数据质量和数据管理。

3. 大数据处理流程可以分解为定义问题、数据理解、数据采集、_____、数据分析、分析结果解析等。

4. 深度学习和_____是大数据分析的基础。

5. 知识图谱泛指各种大型_____，是把所有不同种类的信息连接在一起而得到的一个关系网络。

6. 图数据中主要包括图中的节点以及连接节点的边。在图中，顶点和边实例化构成各

种类型的图，如标签图、属性图、语义图以及_____等。

7. 人们对大数据的处理形式主要是对静态数据的批量处理，_____，以及对图数据的综合处理等。

8. _____是典型的大数据批量处理架构。

9. 交互式数据处理系统的典型代表是 Berkeley 的_____系统等。

10. 图数据处理有一些典型的系统，如微软的_____系统。

**二、简答题**

1. 简述大数据的分析流程。

2. 简述深度学习。

3. 简述知识计算。

4. 简述批量数据。

5. 简述流式数据。