第5章

大数据技术

本章将介绍大数据概述、大数据的关键技术和大数据主要应用领域。通过本章的学习，学生可了解大数据的产生背景及其基本概念、大数据发展历程、Hadoop 等大数据技术、爬虫、清洗等技术与工具、大数据分析、挖掘及可视化技术、大数据管理、大数据典型应用、数据中心及其智能化、大数据未来发展趋势等。

5.1 大数据概述

5.1.1 大数据产生背景及基本概念

1. 背景

从 1990 到 2003 年，由法国、德国、日本、中国、英国和美国 6 个国家的 20 个研究所 2800 多名科学家参加的人类基因组计划耗资 13 亿英镑，产生的 DNA 数据达 200TB。DNA 自动测序技术的快速发展使核酸序列数据量每天增长 106bp，生物信息呈现海量数据增长的趋势。生物信息学将其工作重点定位于对生物学数据的搜索（收集和筛选）、处理（编辑、整理、管理和显示）及利用（计算、模拟、分析和解释）。人类基因组计划研究开创了大数据或海量数据处理（数据密集计算）科学研究方法的先河。

2012 年 2 月，《纽约时报》的一篇专栏中称，大数据时代已经来临。在商业、经济及其他领域中，决策将日益基于数据和分析而做出，而并非基于经验和直觉。哈佛大学社会学教授加里·金说："这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。"此后，大数据一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展与创新。在国内，一些互联网主题的讲座沙龙中，甚至国金证券、国泰君安、银河证券等都将大数据写进了投资推荐报告。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然很多企业可能并没有意识到数据爆炸性增长带来问题的隐患，但是随着时间的推移，人们将越来越多地意识到数据对企业的重要性。

2. 定义

根据维基百科的定义，大数据指无法在可承受的时间范围内用常规软件进行捕捉、管理

和处理的数据集合。研究机构 Gartner 关于大数据的定义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。麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换而言之，如果把大数据比作一种产业，那么这种产业实现营利的关键，在于提高对数据的加工能力，通过加工实现数据的增值。

从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘。但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。随着云时代的来临，大数据也吸引了越来越多的关注。分析师团队认为，大数据通常用来形容一个公司创造的大量非结构化数据和半结构化数据，这些数据在下载到关系型数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要向数十、数百甚至数千台计算机分配工作。

5.1.2　大数据发展历程

1980 年，美国著名未来学家阿尔文·托夫勒就在其著作《第三次浪潮》中，将“大数据”称为“第三次浪潮的华彩乐章”。不过，他可能并没有在书中直接用到“大数据”这个词汇，因为公认的最早使用这个词汇的人是 20 世纪 90 年代在美国硅图公司担任首席科学家的 John Mashey。就像数据的概念从诞生到后来会发生意义转变一样，大数据的初始内涵与它现在的意义也肯定不甚相同。托夫勒也好，John Mashey 也罢，他们当时对大数据的理解更多地停留在表象层面，至于大数据的理论以及可能的应用范围等，还是后来在商用的刺激下被不断深化和放大的。

2008 年对“大数据”而言算得上是一个分水岭，因为国际知名杂志《自然》推出专刊，对其做了介绍。3 年后，美国的《科学》杂志也做了同样的事情。它们从互联网技术、互联网经济学、超级计算、环境科学、生物医药等多方面介绍了海量数据所带来的技术挑战，自此“大数据”一发不可收拾，成为学界研究的热点。鉴于《自然》《科学》等杂志在国际学术界中的权威及影响，推出专刊介绍大数据，无异于为其做了背书。如果说，大数据在此之前只是商人、学者零散的激情，那么此后则成为整个社会的共鸣。

2012 年数据科学家维克托·迈尔-舍恩伯格在《大数据时代》一书中，从理论的层面预言大数据将导致人类思维、商业以及管理领域的变革。大数据的出现已将“相关关系”推升到一个思维的高度。有学者甚至发出“理论的终结”之类的感叹。大数据作为一个时代的标签已经成型。这一判断非常容易得到确认，因为现代社会所有的设备和系统，如果没有数据，就无法进行智能推理和判断。云计算也好，人工智能也罢，从根本上讲，都是靠数据驱动的。19 世纪、20 世纪有很多标签，但其实都属于“石油时代”。同理，21 世纪也存在着诸多可能，但不妨碍我们称其为“大数据时代”。

5.2 大数据的关键技术

5.2.1 Hadoop 等大数据技术

1. Hadoop 概述

Hadoop 是一种处理大数据的分布式软件框架，具有可靠、高效、扩展、低成本和兼容性等特点。Hadoop 的可靠性表现在它有多个工作数据副本，以确保能够针对失败结点的重新分布处理。Hadoop 的高效性表现在其并行工作方式，能够在结点之间动态地移动数据，并保证各个结点的动态平衡，通过并行处理加快处理速度。Hadoop 的扩展性表现在可用的计算机集簇间分配数据，这些集簇可以方便地扩展到数以千计的结点中。Hadoop 低成本表现为它是开源的，依赖于社区服务。Hadoop 带有用 Java 语言编写的框架，可在 Linux 平台上运行，应用程序也可以使用其他语言编写。

Hadoop 框架的核心是 HDFS 和 MapReduce。HDFS 为海量的数据提供了存储，MapReduce 为海量的数据提供了计算。HDFS 是分布式文件系统，负责存储超大数据文件，运行在集群硬件上，具有容错、可伸缩和易扩展特性。MapReduce 功能实现了将单个任务打碎，并将碎片任务(Map)发送到多个结点上，之后再以单个数据集的形式加载(Reduce)到数据仓库里。

Hadoop 框架包括 Hadoop 内核、HDFS、MapReduce 和群集资源管理器 YARN。Hadoop 是一个生态系统，包括很多组件，除 HDFS、MapReduce 和 YARN 外，还有 NoSQL 数据库 HBase、数据仓库工具 Hive、工作流引擎语言 Pig、机器学习算法库 Mahout、数据库连接器 Sqoop、日志数据采集系统 Flume、流处理平台 Kafka、流数据计算框架 Storm、分布式协调服务 ZooKeeper、HBase SQL 搜索引擎 Phoenix、全文搜索引擎 Elasticsearch、安装部署配置管理器 Ambari、新分布式执行框架 Tez 等，见图 5-1。

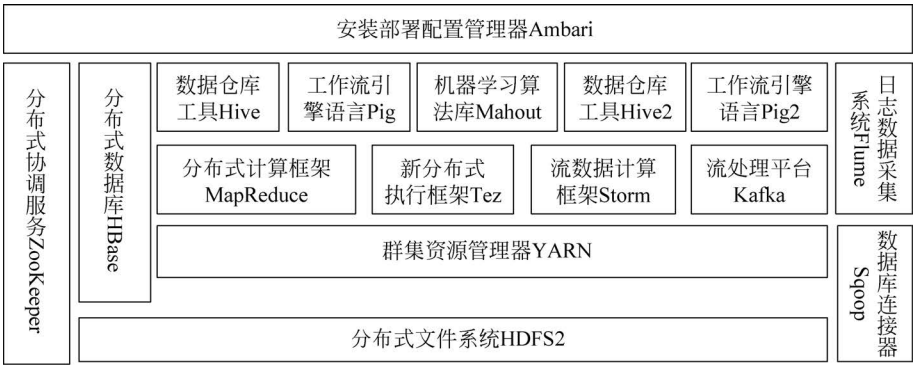

图 5-1 Hadoop 生态系统

Hadoop 由 Apache Software Foundation 公司于 2005 年作为 Lucene 的子项目 Nutch 的一部分正式引入。它受 Google Lab 开发的 MapReduce 编程模型包和 Google File System(GFS)的启发。2006 年 3 月，MapReduce 和 Nutch Distributed File System(NDFS)分别被纳入 Hadoop 项目中。Hadoop 最初只与网页索引有关，后成为分析大数据的平台。目前有很多公司开始提供基于 Hadoop 的商业软件、支持、服务和培训。Cloudera 公司于

2008年开始提供基于Hadoop的软件和服务。GoGrid于2012年开始与Cloudera合作，以加速企业的Hadoop应用推广。Dataguise公司于2012年推出了一款针对Hadoop的数据保护和风险评估。

2. 分布式计算框架 MapReduce

1）概念

MapReduce是一种编程模型，用于大规模数据集的并行运算。其软件实现是指定一个Map（映射）函数，用来把一组键值对映射成一组新的键值对，指定并发的Reduce（归约）函数，用来保证所有映射的键值对其中的每一个共享相同的键组。

MapReduce是面向大数据并行处理的计算模型、框架和平台，它包括三层含义：①MapReduce是一个基于集群的高性能并行计算平台，允许使用普通商用服务器构成一个包含数十、数百至数千个结点的分布和并行计算集群；②MapReduce是一种并行计算与运行软件框架，能自动完成计算任务的并行化处理，自动划分计算数据和计算任务，在集群结点上自动分配和执行任务以及收集计算结果，将数据分布存储、数据通信、容错处理等并行计算涉及很多系统底层的复杂细节交由系统处理，可减少软件开发人员的负担；③MapReduce是一种并行程序设计模型与方法，它借助于函数式程序设计语言Lisp的设计思想，提供简易并行程序设计方法，用Map和Reduce两个函数编程实现并行计算任务，提供抽象操作和并行编程接口，以简便大数据编程和计算处理。

2）主要功能

（1）大数据划分和计算任务调度。系统自动将大数据划分为很多个数据块，每个数据块对应于一个计算任务，并自动调度计算结点来处理相应的数据块。任务调度主要负责分配和调度计算结点（Map结点或Reduce结点），同时负责监控这些结点的执行状态，并负责Map结点执行的同步控制。

（2）数据/代码互定位。为减少数据通信，一个基本原则是本地化数据处理，即一个计算结点尽可能处理其本地磁盘上所分布存储的数据，这实现了代码向数据的迁移；当无法进行这种本地化数据处理时，再寻找其他可用结点并将数据从网络上传送给该结点（数据向代码迁移），但将尽可能从数据所在的本地机架上寻找可用结点以减少通信延迟。

（3）系统优化。为减少数据通信开销，中间结果数据进入Reduce结点前会进行合并处理；一个Reduce结点所处理的数据可能会来自多个Map结点，为了避免Reduce计算阶段发生数据相关性，Map结点输出的中间结果需使用一定的策略进行适当的划分，保证相关性数据发送到同一个Reduce结点；此外，系统还进行一些计算性能优化处理，如对最慢的计算任务采用多备份执行、选最快完成者作为结果。

（4）出错检测和恢复。以低端商用服务器构成的大规模MapReduce计算集群中，结点硬件（主机、磁盘、内存等）出错和软件出错是常态，因此MapReduce需要能检测并隔离出错结点，并调度分配新的结点接管出错结点的计算任务。同时，系统还将维护数据存储的可靠性，用多备份冗余存储机制提高数据存储的可靠性，并能及时检测和恢复出错的数据。

3. 群集资源管理器 YARN

Apache Hadoop YARN（Yet Another Resource Negotiator，另一种资源协调器）是一种Hadoop资源管理器，可为上层应用提供统一的资源管理和调度，它的引入为集群在利用

率、资源统一管理和数据共享等方面带来了巨大便利。

YARN 的基本思想是将 Job Tracker 的两个主要功能(资源管理和作业调度/监控)分离,创建一个全局 Resource Manager 和若干个针对应用程序的 Application Master。YARN 分层结构的本质是 Resource Manager。这个实体控制整个集群并管理应用程序向基础计算资源的分配。Resource Manager 将各种资源(计算、内存、带宽等)安排给基础 Node Manager(结点代理)。Resource Manager 与 Application Master 一起分配资源,与 Node Manager 一起监视其基础应用程序。Application Master 负责承担 Task Tracker 的一些角色,Resource Manager 负责承担 Job Tracker 的角色。Application Master 负责协调来自 Resource Manager 的资源,并通过 Node Manager 监视容器的执行和资源使用(CPU、内存的资源分配)。

4. 分布式协调服务 ZooKeeper

ZooKeeper 是一个分布式的,开放源码的应用程序协调服务器,是 Google 的 Chubby 开源实现,是 Hadoop 和 HBase 的重要组件。它是一个为分布式应用提供一致性服务的软件,其功能包括:配置维护、域名服务、分布式同步、组服务等。ZooKeeper 的目标就是封装好复杂易出错的关键服务,将简单易用的接口和性能高效、功能稳定的系统提供给用户。

5. 安装部署配置管理器 Ambari

Apache Ambari 是一种基于 Web 的工具,支持 Apache Hadoop 集群的安装、部署、管理和监控。Ambari 支持大多数 Hadoop 组件,包括 HDFS、MapReduce、Hive、Pig、HBase、ZooKeeper、Sqoop 和 Hcatalog 等。Ambari 是开源软件,是 Apache Software Foundation 中的一个项目。2017 年 11 月发布版本 2.6.0。

6. 新分布式执行框架 Tez

Apache Tez 是针对 Hadoop 数据处理应用程序的新分布式开源计算框架。它将多个有依赖的作业转换为一个作业,从而大幅提升 DAG(Database Availability Group,数据库可用性组)作业的性能。Tez 不直接面向最终用户,它帮助 Hadoop 批处理大数据。如果 Hive 和 Pig 使用 Tez 而不是 MapReduce 作为其数据处理工具,其响应时间会明显提升。Tez 构建在 YARN 之上。Tez 产生的主要原因是绕开 MapReduce 所施加的限制。除了必须要编写 Mapper 和 Reducer 的限制之外,还强制让所有类型的计算都满足这一范例,还有效率低下的问题。

7. 分布式文件系统 HDFS

HDFS 被设计成适合运行在通用硬件上的分布式文件系统,其容错性高,适合部署在廉价机器上。HDFS 能提供高吞吐量的数据访问,适合大规模数据集应用。HDFS 放宽了一部分 POSIX(Portable Operating System Interface of UNIX,可移植操作系统接口)约束,以实现流式读取文件系统数据。早年的 HDFS 是作为 Apache Nutch 搜索引擎项目的基础架构而开发的。HDFS 是 Apache Hadoop Core 项目的一部分。

HDFS 的特点包括:①硬件故障检测与恢复。HDFS 由数百或数千个存储着文件数据片段的服务器组成,每个组成部分都很可能出现故障,因此,它设计了故障检测和自动快速恢复功能。②数据访问。运行在 HDFS 上的应用程序必须流式访问其数据集,它不是运行在普通文件系统之上的普通程序。HDFS 被设计成适合批量处理的,而不是用户交互式的。

重点是在数据吞吐量，而不是数据访问的反应时间，POSIX 的很多硬性需求对于 HDFS 应用都是非必需的，去掉 POSIX 一小部分关键语义可以获得更好的数据吞吐率。③大数据集。运行在 HDFS 之上的程序有大量数据集。HDFS 文件大小由吉字节(GB)到太字节(TB)级别不等，所以，HDFS 必须支持大文件，它需提供高聚合数据带宽，一个集群需支持数百个结点，一个集群还应支持千万级别的文件。④迁移计算。在靠近计算数据所存储的位置来进行计算是最理想的状态，尤其是在数据集特别巨大的时候。这样消除了网络的拥堵，提高了系统的整体吞吐量。HDFS 提供了接口，以便让程序将自己移动到离数据存储更近的位置。HDFS 被设计成可以简便实现平台间迁移的工具。⑤名字结点和数据结点。HDFS 是主从结构，一个 HDFS 集群是一个名字结点，它是一个管理文件命名空间和调节客户端访问文件的主服务器，用来管理对应结点的存储。HDFS 对外开放文件命名空间并允许用户数据以文件形式存储。内部机制是将一个文件分割成一个或多个块，这些块被存储在一组数据结点中。名字结点用来操作文件命名空间的文件或目录操作。它同时确定块与数据结点的映射。数据结点负责文件系统客户的读写请求，执行块的创建，删除和结点块复制指令。

8. 分布式数据库 HBase

HBase 是一个分布式的、面向列的、可伸缩的分布式开源数据库，是 Apache Hadoop 子项目。HBase 不是传统关系数据库，它是非结构化数据存储的数据库，是基于列的而不是基于行的模式。利用 HBase 技术可在廉价 PC Server 上搭建大规模结构化存储集群。HBase 利用 Hadoop HDFS 作为其文件存储系统。HBase 利用 Hadoop MapReduce 来处理 HBase 中的海量数据。

HBase 位于结构化存储层，Hadoop HDFS 为 HBase 提供了高可靠性的底层存储支持，Hadoop MapReduce 为 HBase 提供了高性能的计算能力，ZooKeeper 为 HBase 提供了稳定服务和 failover 机制。此外，Pig 和 Hive 还为 HBase 提供语言支持，使得在 HBase 上进行数据统计处理非常简单。Sqoop 可为 HBase 提供 RDBMS 数据导入功能，使传统数据库数据向 HBase 迁移非常方便。

9. 数据仓库工具 Hive

Hive 是基于 Hadoop 的数据仓库工具，可将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能，可以将 SQL 语句转换为 MapReduce 任务运行。Hive 是建立在 Hadoop 上的数据仓库基础构架。它提供了一系列的工具，可用来进行数据提取转换加载(ETL)，这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。Hive 定义了简单的类 SQL，称为 HQL，它允许熟悉 SQL 的用户查询数据。同时，这个语言也允许熟悉 MapReduce 的开发者开发自定义的 Mapper 和 Reducer 来处理内建的 Mapper 和 Reducer 无法完成的复杂的分析工作。

10. 工作流引擎语言 Pig

Pig 是一种数据流语言和运行环境，用于检索非常大的数据集，为大型数据集的处理提供了一个更高层次的抽象。Pig 包括两部分：一是用于描述数据流的语言，称为 Pig Latin；二是用于运行 Pig Latin 程序的执行环境。Apache Pig 是一种高级语言，适合 Hadoop 和 MapReduce 平台查询大型半结构化数据集。通过允许对分布式数据集进行类似 SQL 的查

询，Pig 可以简化 Hadoop 使用。

11. 日志数据采集系统 Flume

Flume 是 Cloudera 的一种分布式海量日志采集、聚合和传输系统。Flume 支持在日志系统中定制各类数据发送方，用于数据收集、简单数据处理，并将结果写到数据接收方。Flume 提供从 console（控制台）、RPC（Thrift-RPC）、text（文件）、tail（UNIX tail）、syslog（syslog 日志系统）、支持 TCP 和 UDP 两种模式，到 exec（命令执行）等在数据源上收集数据的功能。

Flume 有 Flume-og 和 Flume-ng 两个版本。Flume-og 采用多 Master 方式。为保证配置数据的一致性，Flume 引入了 ZooKeeper，Flume Master 间使用 Gossip 数据传输协议同步数据。与 Flume-og 相比，Flume-ng 取消了集中管理配置的 Master 和 ZooKeeper 而演变为一种数据传输工具。Flume-ng 另一个不同是读入数据和写出数据时由不同工作线程处理（称为 Runner）。

12. 流处理平台 Kafka

Kafka 是 Apache 基金开发的开源流处理平台，由 Scala 和 Java 编写。Kafka 是一种高吞吐量的分布式发布订阅消息系统，可处理消费者规模的网站中的所有动作流数据。Kafka 通过 Hadoop 并行加载机制统一线上和线下的消息，以便为集群提供实时数据。Kafka 特性包括：①通过磁盘数据结构提供消息的持久化，这种结构对于太字节（TB）级数据存储也能够保持长时间稳定性；②高吞吐量，即使是非常普通的硬件，Kafka 也可以支持每秒数百万的消息；③支持通过 Kafka 服务器和消费机集群来分区消息；④支持 Hadoop 并行数据加载。

13. 数据库连接器 Sqoop

Sqoop 是一种开源数据库连接工具，用于 Hadoop 与传统数据库间的数据传递和互转。Sqoop 项目开始于 2009 年，是 Apache 项目。对于 NoSQL 数据库，它提供了一种连接器，能够分割数据集并创建 Hadoop 任务来处理每个区块。随着云计算的普及，Hadoop 和传统数据库之间的数据传输和转移变得越来越重要。

14. 流数据计算框架 Storm

Apache Storm 是一种分布式实时大数据处理系统。是一种流数据计算框架，具有高摄取率、容错性和扩展性。Storm 最初由 Nathan Marz 和 BackType 的团队创建。BackType 是一家社交分析公司。后来 Storm 被收购，并通过 Twitter 开源。Apache Storm 已成为分布式实时处理系统的标准，允许大量数据处理。Apache Storm 用 Java 和 Clojure 编写，是实时分析的计算框架。

5.2.2 爬虫、清洗等技术与工具

1. 大数据采集

1）大数据采集的特点

（1）数据结构的多样性。大数据具有多样性的特点，即包含多种数据结构，有结构化的、半结构化的，也有非结构化的。而传统数据则是结构化的。有统计显示，大数据中，非结

构化数据占比高达85%。

(2) 大数据来源的多元化。大数据,除传统数据源外,还增加了一些新的数据源,例如,卫星影像数据、社交网络微博、微信、论坛、QQ聊天数据、股市数据、监控探头的视频数据、网络舆情监测数据等。

(3) 采集方式的智能化。与传统数据人工采集相比,大数据采集以智能化为主、人工为辅。一般情况下,大数据由智能设备、仪器、传感器、摄像头等产生,其过程无须人工干预。

2) 大数据资源的采集方法

(1) 按需求采集数据。大数据资源的应用需要经过采集、存储、分析、处理等过程,而大数据的应用决定了其初始的目标性。大数据的采集需要根据目标需求采集。采集是源头,也是关键。没有采集就没有后面的过程。

(2) 使用新的技术。大数据是普通软件和技术所不能处理的数据。因此,必须采用新的技术。无论是采集、存储、分析、处理,还是具体应用,大数据新技术的应用是前提条件。已经有的新技术包括云计算技术、海量异构数据存储、分布式集群计算、深度学习和挖掘技术、个性化推荐技术等。

(3) 可采取"众包"模式。大数据资源的采集可采用"众包"模式。"众包"是一种新的生产组织模式,通过"众包",将大数据资源的采集工作分配出去,一方面可降低成本;另一方面可以集思广益,并组织大规模生产。

2. 爬虫技术

1) 概念

网络爬虫是一种按照一定的规则,自动地抓取万维网信息的程序或者脚本。其可以从万维网上下载网页,是搜索引擎的重要组成。爬虫在抓取网页的过程中,不断从当前页面上抽取新的URL放入队列,直到满足系统的一定停止条件。所有被爬虫抓取的网页将会被系统存储,进行一定的分析、过滤,并建立索引,以便之后的查询和检索。

2) 结构技术分类

网络爬虫按照系统结构和实现技术,大致可以分为以下几种类型:通用网络爬虫、聚焦网络爬虫、增量式网络爬虫、深层页面爬虫等。实际中的网络爬虫通常是几种爬虫技术的结合。

(1) 通用网络爬虫。通用网络爬虫又称全网爬虫,爬行对象从一些种子URL扩充到整个Web,主要为门户站点、搜索引擎和大型Web服务提供商采集数据。

(2) 聚焦网络爬虫。聚焦网络爬虫是指选择性地爬行那些与预先定义好的主题相关页面的网络爬虫。与通用网络爬虫相比,聚焦网络爬虫只需要爬行与主题相关的页面,极大地节省了硬件和网络资源,保存的页面也因数量少而更新快,还可以很好地满足一些特定人群对特定领域信息的需求。聚焦网络爬虫是需要关注的重点爬虫类型。聚焦爬虫(又被称为网页蜘蛛、网络机器人,在FOAF社区中间,更经常地被称为网页追逐者),是一种按照一定的规则,自动地抓取万维网信息的程序或者脚本。

(3) 增量式网络爬虫。增量式网络爬虫是指对已下载网页采取增量式更新和只爬行新产生的或者已经发生变化的网页的爬虫,它能够在一定程度上保证所爬行的页面是尽可能新的页面。与周期性爬行和刷新页面的网络爬虫相比,增量式爬虫只会在需要的时候爬行新产生或发生更新的页面,并不重新下载没有发生变化的页面,可有效减少数据下载量,及

时更新已爬行的网页，减小时间和空间上的耗费，但是增加了爬行算法的复杂度和实现难度。

(4) 深层页面爬虫。Web 页面按存在方式分为表层网页和深层网页。表层网页是传统搜索引擎可以索引的页面，是以超链接可以到达的静态网页为主构成的 Web 页面。深层网页是大部分内容不能通过静态链接获取的，隐藏在搜索表单后的，只有用户提交一些关键词才能获得的 Web 页面。例如，那些用户注册后内容才可见的网页就属于深层页面。深层页面爬虫包含六个基本功能模块(爬行控制器、解析器、表单分析器、表单处理器、响应分析器、LVS 控制器)和两个爬虫内部数据结构(URL 列表、LVS 表)，其中，LVS(Label Value Set)表示标签/数值集合，用来表示填充表单的数据源。深层页面爬虫爬行过程中最重要部分就是表单填写，包含两种类型：基于领域知识的表单填写和基于网页结构分析的表单填写。

3) 应用分类

抓取目标的描述和定义是决定网页分析算法与 URL 搜索策略如何制定的基础。而网页分析算法和候选 URL 排序算法是决定搜索引擎所提供的服务形式和爬虫网页抓取行为的关键所在。这两部分的算法又是紧密相关的。现有聚焦爬虫对抓取目标的描述可分为基于目标网页特征、基于目标数据模式和基于领域概念 3 种。

(1) 基于目标网页特征的爬虫，其所抓取、存储并索引的对象一般为网站或网页。根据种子样本获取方式可分为：预先给定的初始抓取种子样本；预先给定的网页分类目录和与分类目录对应的种子样本，如 Yahoo! 分类结构等；通过用户行为确定的抓取目标样例，分为：①用户浏览过程中显示标注的抓取样本；②通过用户日志挖掘得到访问模式及相关样本。其中，网页特征可以是网页的内容特征，也可以是网页的链接结构特征等。

(2) 基于目标数据模式的爬虫，其所针对的是网页上的数据，所抓取的数据一般要符合一定的模式，或者可以转换或映射为目标数据模式。

(3) 基于领域概念的爬虫。其需要建立目标领域的本体或词典，用于从语义角度分析不同特征在某一主题中的重要程度。

4) 爬虫工具

(1) Web Scraper。它是一个独立的 Chrome 扩展，安装数目已经到了 20 万。它支持点选式的数据抓取，另外支持动态页面渲染，并且专门为 JavaScript、AJAX、下拉拖动、分页功能做了优化，并且带有完整的选择器系统，另外支持数据导出到 CSV 等格式。它们还有自己的 Cloud Scraper，支持定时任务、API 式管理、代理切换功能。

(2) Data Scraper。它是 Chrome 的扩展，可以将单个页面的数据通过单击的方式爬取到 CSV、XSL 文件中。在这个扩展中已经预定义了 5 万多条规则，可以用来爬取将近 1.5 万个热门网站。

(3) Listly。它是 Chrome 的插件，可以快速地将网页中的数据进行提取，并将其转换为 Excel 表格导出，操作非常便捷。例如，获取一个电商商品数据、文章列表数据等，使用它就可以快速完成。另外，它也支持单页面和多页面以及父子页面的采集，值得一试。

3. 清洗技术

1) 数据清洗的概念

数据清洗(Data Cleaning)指发现并纠正数据文件中可识别错误的过程，是对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误、处理无效值和缺失值等，

并提供数据一致性。数据清洗是一个反反复复的过程，不可能一次完成。

数据清洗从名词上解释就是把“脏数据”“洗掉”。因为数据仓库中的数据是面向某一主题的数据集合，这些数据从多个业务系统中抽取出来而且包含历史数据，这就难免有的数据是错误数据、有的数据相互之间有冲突。这些错误的或有冲突的数据显然就是“脏数据”。按照一定的规则把“脏数据”“洗掉”就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是过滤掉还是由业务单位修正之后再进行抽取。

2）待清洗的数据

不符合要求的数据主要是不完整数据、错误数据和重复数据三大类。

（1）不完整数据。这类数据主要是一些应该有的，但实际缺失的信息，如供应商的名称、分公司的名称、客户信息缺失。应将这类数据过滤出来，按缺失的内容向客户提交，要求在规定的时间内补全。

（2）错误数据。这类错误数据产生的原因是业务系统不够健全，在接收输入后没有进行判断直接写入后台数据库，例如，数值数据输入成全角数字字符、字符串数据后面有一个回车操作、日期格式不正确、日期越界等。

（3）重复数据。这类数据是因为重复操作导致的结果。待确认是重复的数据后，可将其去除，以保证数据的唯一性。

3）数据清洗方法

一般来说，数据清洗是将数据精简以除去重复记录，并使剩余部分转换成标准可接收格式的过程。数据清洗标准模型是将数据输入到数据清洗处理器，通过一系列步骤“清洗”数据，然后以期望的格式输出清洗过的数据。数据清洗从数据的准确性、完整性、一致性、唯一性、适时性、有效性几方面来处理数据的丢失值、越界值、不一致代码、重复数据等问题。数据清洗一般针对具体应用，因而难以归纳统一的方法和步骤，但是根据数据不同可以给出相应的数据清洗方法。

（1）解决不完整数据的方法。大多数情况下，缺失的值必须手工填入（即手工清洗）。当然，某些缺失值可以从本数据源或其他数据源推导出来，这就可以用平均值、最大值、最小值或更为复杂的概率估计代替缺失的值，从而达到清洗的目的。

（2）错误值的检测及解决方法。用统计分析的方法识别可能的错误值或异常值，如偏差分析、识别不遵守分布或回归方程的值，也可以用简单规则库检查数据值，或使用不同属性间的约束、外部的数据来检测和清洗数据。

（3）重复记录的检测及消除方法。数据库中属性值相同的记录被认为是重复记录，通过判断记录间的属性值是否相等来检测记录是否相等，相等的记录合并为一条记录。

（4）不一致性的检测及解决方法。从多数据源集成的数据可能有语义冲突，可定义完整性约束用于检测不一致性，也可通过分析数据发现联系，从而使得数据保持一致。

4）ETL 技术

ETL（Extract-Transform-Load）是用来描述将数据从来源端经过抽取（Extract）、转换（Transform）、加载（Load）至目的端的过程。ETL 是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，经过数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。

ETL 结果的质量表现为正确性、完整性、一致性、完备性、有效性、时效性和可获取性等

几个特性。而影响其质量的原因有很多，主要取决于系统集成和历史数据，包括以下几方面：业务系统不同时期系统之间数据模型不一致，业务系统不同时期业务过程有变化，旧系统模块在运营、人事、财务、办公系统等相关信息中不一致，遗留系统和新业务、管理系统数据集成不完备带来的不一致性。

ETL 的核心是 ETL 转换过程，主要包括以下几方面。①空值处理：可捕获字段空值，进行加载或替换为其他含义数据，并可根据字段空值实现分流加载到不同目标库。②规范化数据格式：可实现字段格式约束定义，对于数据源中的时间、数值、字符等数据，可自定义加载格式。③拆分数据：依据业务需求对字段可进行分解，如主叫号为 861082585313-8148，可进行区域码和电话号码分解。④验证数据正确性：可利用 Lookup 及拆分功能进行数据验证。针对该主叫号，进行区域码和电话号码分解后，可利用 Lookup 返回主叫网关或交换机记载的主叫地区，进行数据验证。⑤数据替换：对于业务因素，可实现无效数据、缺失数据的替换。⑥Lookup：查获丢失数据并返回用其他手段获取的缺失字段，保证字段完整性。⑦建立 ETL 过程的主外键约束：对无依赖性的非法数据，可替换或导出到错误数据文件中，保证主键唯一记录的加载。

很多情况下，数据清洗工具与 ETL 工具是合二为一的平台。常见的工具包括：①Datastage，是一种数据集成软件平台，能够帮助企业从分散在各个系统中的复杂异构信息获得更多价值。②Informatica，数据集成平台，可以在改进数据质量的同时，访问、发现、清洗、集成并交付数据。③Kettle，开源 ETL 工具，用 Java 编写，可以在 Windows、Linux、UNIX 上运行，数据抽取高效稳定。④ODI(Oracle Data Integrator)，是 Oracle 数据库厂商提供的数据集成类工具，开放数据链路接口，受 Oracle 数据库的影响，有局限性。⑤OWB(Oracle Warehouse Builder)，是 Oracle 的一个综合工具，提供对 ETL(提取、转换和加载)、完全集成的关系和维度建模、数据质量、数据审计，及数据和元数据的整个生命周期的管理。⑥Cognos，是 BI 核心平台之上，以服务为导向进行架构的一种数据模型，是唯一可以通过单一产品和在单一可靠架构上提供完整业务智能功能的解决方案。⑦Beeload，是国产 ETL 工具。Beeload 集数据抽取、清洗、转换及装载于一体，通过标准化企业各个业务系统产生的数据，向数据仓库提供高质量的数据。

5) 数据清洗工具

(1) 思迈特软件 Smartbi。其数据清洗功能非常强大，Smartbi 具有轻量级 ETL 功能，可视化流程配置，简单易用，业务人员可以参与。采用分布式计算架构，单结点支持多线程，可处理大量数量，提高数据处理性能。强大的数据处理功能不仅支持异构数据，还支持内置排序、去重、映射、行列合并、行列转换聚合、去空值等数据预处理功能。

(2) Python 语言。它是一种面向对象的动态语言，最初被设计用来编写自动化脚本。它越来越多地被用来开发独立的大型项目，因为版本不断更新，语言新功能也在增加。

(3) PyCharm。它是一种 Python 集成开发环境，有一整套工具，可以帮助用户在使用 Python 语言开发时提高效率，如调试、语法亮点、Project 管理、代码跳转、智能提示、自动完成、单元测试、版本控制等。

5.2.3 大数据分析、挖掘及可视化技术

大数据分析与挖掘的目的是获得更有价值的数据或结果，它也是知识发现的过程。

1. 大数据分析与挖掘概述

数据分析，是指对类型多样、增长快速、内容真实的数据进行分析，从中找出可以帮助决策的隐藏模式、未知的相关关系以及其他有用信息的过程。在这一过程中，有两个关键技术：一个是文本分析，另一个是机器学习。而大数据分析则是根据数据生成机制，进行数据采集与存储，并对数据进行格式化清洗，以大数据分析模型为依据，在集成化大数据分析平台的支撑下，运用云计算技术调度计算分析资源，最终挖掘出大数据背后的模式或规律的数据分析过程。

数据挖掘是知识发现的一个步骤，指从大数据中通过算法挖掘隐藏于其中有价值信息的过程。通过统计、在线分析处理、情报检索、机器学习、专家系统和模式识别等诸多方法和手段来实现数据挖掘的目标。数据挖掘是多学科和技术的应用，涉及的学科方法包括：统计学的抽样、估计和假设检验；人工智能、模式识别和机器学习的搜索算法、建模技术和学习理论。涉及的学科包括运筹学、进化计算、信息论、信号处理、可视化和信息检索等。近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是从大数据获取的信息和知识可以广泛用于商务管理、生产控制、市场分析、工程设计和科学探索等领域。

2. 大数据分析与挖掘方法

(1) 事物分类。分类是指按照种类、等级或性质分别归类。把无规律的事物分为有规律的，按照不同的特点划分事物，使事物更有规律。建立分类模型，对于没有分类的数据进行分类。

(2) 变量估计。估计与分类类似，不同之处在于，分类描述的是离散型变量的输出，而估值处理连续值的输出；分类的类别是确定数目的，估值的量是不确定的。一般来说，估值可以作为分类的前一步工作。给定一些输入数据，通过估值，得到未知的连续变量的值，然后根据预先设定的阈值，进行分类。

(3) 未知预测。预测指在掌握现有信息的基础上，依照一定的方法和规律对未来的事情进行测算，以预先了解事情发展的过程与结果。目的是对未来未知变量的预测，这种预测是需要时间来验证的，即必须经过一定时间后，才知道预测的准确性是多少。

(4) 关联分析。关联分析，又称关联挖掘，是一种简单、实用的分析技术，其目的是发现数据集中数据的关联性，从而描述一个事物中某些属性和某种规律和模式。关联分析是从大量数据中发现项集之间的关联。关联分析的一个典型例子是购物篮分析。该过程通过发现顾客放入其购物篮中的不同商品之间的联系，分析顾客的购买习惯。通过了解哪些商品频繁地被顾客同时购买，帮助零售商制定营销策略。其他的应用还包括价目表设计、商品促销、商品的排放和基于购买模式的顾客划分。

(5) 对象聚类。聚类是对记录分组，把相似的记录放在一个聚集里。将物理或抽象对象的集合分成由类似的对象组成的多个类的过程被称为聚类。由聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。聚类分析又称群分析，它是研究(样品或指标)分类问题的一种统计分析方法。聚类分析有很多方法，如系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。

(6) 数据的可视化。可视化是对数据挖掘结果进行直观表示的方式。在数据挖掘时，可以通过软件工具进行数据的展现、分析、钻取，将数据挖掘的分析结果形象地显示出来。

3. 可视化的基本概念

可视分析是大数据分析的重要方法。大数据可视分析旨在利用计算机自动化分析能力的同时，充分挖掘人对于可视化信息的认知能力优势，将人、机的各自强项进行有机融合，借助人机交互式分析方法和交互技术，辅助人们更为直观和高效地洞悉大数据背后的信息、知识与智慧。

可视化的概念于 1987 年 2 月在美国国家科学基金会(National Science Foundation)的一个专题研讨会上提出。1995 年前后，网络信息技术的发展使可视化技术有了比较大的突破。近年，大数据的出现使之有了更为广泛的用途。

可视化就是将科学计算的中间数据或结果数据，转换为人们容易理解的图形图像形式。有学者将信息可视化(Information Visualization)定义为：对抽象数据使用计算机支持的、交互的、可视化的表示形式以增强认知能力。与传统计算机图形学以及科学可视化研究不同，信息可视化的研究重点更加侧重于通过可视化图形呈现数据中隐含的信息和规律，所研究的创新性可视化表征旨在建立符合人的认知规律的心理映像。

经过 30 年的发展，信息可视化已成为分析复杂问题强有力的工具。随着计算机、图形图像技术的发展，现在已经可以用丰富的色彩、动画技术、三维立体显示及仿真(虚拟现实)等手段，表现各种现实和虚拟的场景。科学计算可视化已广泛应用于流体计算力学、有限元分析、医学图像处理、分子结构模型、天体物理、空间探测、地球科学、数学等领域。

4. 面向大数据应用的可视化技术

大数据的出现刺激了可视化技术的快速发展。大数据可视化技术涉及传统的科学可视化和信息可视化，其目的是掘取信息和洞悉知识。信息的可视化分为文本、网络或图、时空及多维数据等。

1) 文本可视化

文本信息是典型的非结构化数据，是互联网中最主要的信息类型，也是物联网各种传感器采集后生成的主要信息类型，人们日常工作和生活中接触最多的文档多以文本形式存在。文本可视化的意义在于，能够将文本中蕴含的语义特征(例如，词频与重要度、逻辑结构、主题聚类、动态演化规律等)直观展示出来。典型的文本可视化技术是标签云，将关键词根据词频或其他规则进行排序，按照一定规律进行布局排列，用大小、颜色、字体等图形属性对关键词进行可视化，如图 5-2 所示。

图 5-2　标签云

2) 网络图可视化

网络关联关系是大数据中最常见的关系，例如，互联网与社交网络。基于网络结点和连

接的拓扑关系，直观地展示网络中潜在的模式关系，例如，结点或边聚集性，是网络可视化的主要内容之一，如图 5-3 所示。

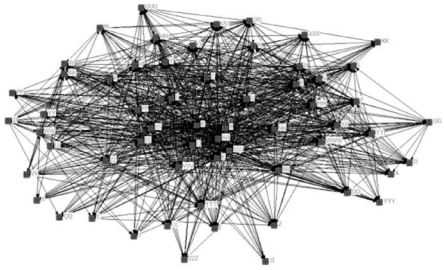

图 5-3　网络图可视化

3）多维数据可视化

多维数据分析的目标是探索多维数据项的分布规律和模式，并揭示小同维度属性之间的隐含关系。二维散点图将两个维度属性值集合映射至两条轴，在二维轴确定的平面内通过图形标记的小同视觉元素来反映其他维度的属性值，如图 5-4 所示。平行坐标是可视化高维几何和分析多元数据的常用方法。为了在 n 维空间中显示一组点，绘制由 n 条平行线组成的背景，通常是垂直且等距的。高维空间的一个点，可以表示为一条折线，这条折线在 n 条平行线之间，折线的拐点在平行线上，n 条平行线为平行坐标轴。第 i 个轴上顶点的位置对应于该点的第 i 个坐标。此可视化与时间序列可视化密切相关，如图 5-5 所示。

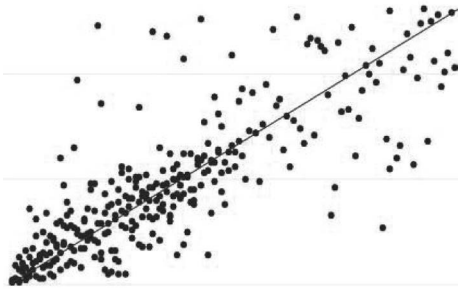

图 5-4　散点图

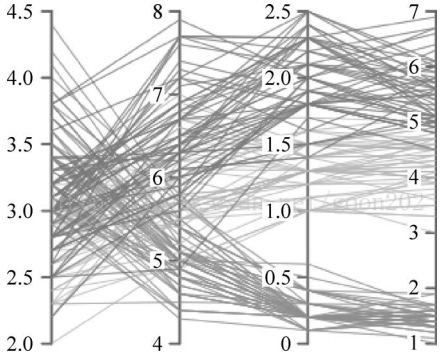

图 5-5　平行坐标

5.2.4　大数据管理

1. 大数据平台框架概述

国家大数据战略，不仅要将大数据作为战略资源，也要将其作为国家治理的创新手段。在统筹布局建设国家大数据平台的基础上，逐渐推动数据的统一、整合、开放和共享机制。

大数据管理架构应该是“一个机制、两个体系和三个平台”的结构。大数据管理工作机制应包括数据共享与开放、工作协同、大数据科学决策、精准监管和公共服务机制等。两个体系指大数据的交换、共享、一致、整合和应用的安全保障和标准化工作。三个平台担负大数据集约化基础设施、网络资源、计算资源、存储资源、安全资源、集中管理、系统运维、大数据采集、处理、分析和应用，见图 5-6。

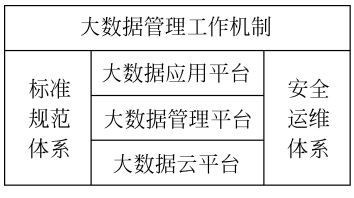

图 5-6　大数据平台框架

大数据云平台是国家大数据战略的基础设施。因为单台计算机无法处理大数据，因此必须采用分布式架构，对海量数据进行分布式数据挖掘，必须依托云计算的分布

式处理、分布式数据库、云存储和虚拟化技术。云计算的核心是将对被用网络连接的计算资源统一管理和调度，构成一个计算资源池向用户提供按需分配的服务。

大数据管理平台是云平台之上的数据交换、存储、共享和开放的平台，它为大数据应用提供统一的数据支持。其具体工作包括破除信息孤岛、整合与集中数据资源、建立数据资源目录、构建数据中心、分布式数据管理、数据互联互通等。

大数据应用平台侧重数据（关联、趋势和空间）分析模型的构建，利用可视化、仿真技术和数据挖掘工具，通过数学统计、在线分析、情报检索、机器学习、专家系统、知识推理和模式识别等，提升治国理政的能力，使决策过程科学化。

2. 数据整合

数据整合就是把在不同数据源收集、整理、清洗，转换后的数据加载到一个新的数据源，为数据消费者提供一种统一数据视图的数据集成方式。数据整合有如下多种方法。

（1）多个数据库整合。通过对各个数据源的数据交换格式进行一一映射，从而实现数据的流通与共享。对于有全局统一模式的多数据库系统，用户可以通过局部外模式访问本地库，通过建立局部概念模式、全局概念模式、全局外模式，用户可以访问集成系统中的其他数据库；对于联邦数据库系统，各局部数据库通过定义输入/输出模式，进行各联邦数据库系统之间的数据访问。基于异构数据源系统的数据整合有多种方式，所采用的体系结构也各不相同，但其最终目的是相同的，即实现数据的流通共享。

（2）数据仓库整合。数据仓库是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。从数据仓库的建立过程来看，数据仓库是一种面向主题的整合方案，因此首先应该根据具体的主题进行建模，然后根据数据模型和需求从多个数据源加载数据。由于不同数据源的数据结构可能不同，因而在加载数据之前要进行数据转换和数据整合，使得加载的数据统一到需要的数据模型下，即根据匹配、留存等规则，实现多种数据类型的整合。

（3）中间件整合。中间件是位于用户与服务器之间的中介接口软件，是异构系统集成所需的黏结剂。现有的数据库中间件允许用户在异构数据库上调用 SQL 服务，解决异构数据库的互操作性问题。功能完善的数据库中间件，可以对用户屏蔽数据的分布地点、数据库管理平台、特殊的本地应用程序编程接口等差异。

（4）Web 服务整合。Web 服务可理解为自包含的、模块化的应用程序，它可以在网络中被描述、发布、查找及调用；也可以把 Web 服务理解为是基于网络的、分布式的模块化组件，它执行特定的任务，遵守具体的技术规范，这些规范使得 Web 服务能与其他兼容的组件进行互操作。

（5）主数据管理整合。主数据管理通过一组规则、流程、技术和解决方案，实现对企业数据一致性、完整性、相关性和精确性的有效管理，从而为所有企业相关用户提供准确一致的数据。主数据管理提供了一种方法，通过该方法可以从现有系统中获取最新信息，并结合各类先进的技术和流程，使得用户可以准确、及时地分发和分析整个企业中的数据，并对数据进行有效性验证。

3. 大数据共享

随着信息时代的不断发展，不同部门、不同地区间的信息交流逐步增加，计算机网络技

术的发展为信息传输提供了保障。当大数据出现时，数据共享问题被提上了议事日程。

数据共享就是让在不同地方使用不同计算机、不同软件的用户能够读取他人数据并进行各种操作运算和分析。数据共享可使更多人充分地使用已有的数据资源，以减少资料收集、数据采集等重复劳动和相应费用，把精力放在开发新的应用程序及系统上。由于数据来自不同的途径，其内容、格式和质量千差万别，因而给数据共享带来了很大困难，有时甚至会遇到数据格式不能转换或数据格式转换后信息丢失的问题，这阻碍了数据在各部门和各系统中的流动与共享。

数据共享的程度反映了一个地区、一个国家的信息发展水平，数据共享程度越高，信息发展水平越高。要实现数据共享，首先应建立一套统一的、法定的数据交换标准，规范数据格式，使用户尽可能采用规定的数据标准。如美国、加拿大等国家都有自己的空间数据交换标准，目前我国正在抓紧研究制定国家的空间数据交换标准，包括矢量数据交换格式、栅格影像数据交换格式、数字高程模型的数据交换格式及元数据格式，该标准建立后，将对我国大数据产业的发展产生积极影响。其次，要建立相应的数据使用管理办法，制定相应的数据版权保护、产权保护规定，各部门间签订数据使用协议，这样才能打破部门、地区间的信息保护，做到真正的信息共享。

4. 大数据开放

数据开放没有统一的定义，一般指把个体、部门和单位掌握的数据提供给社会公众或他人使用。政府数据开放就是要创造一个可持续发展机制发挥数据的社会、经济和政治价值，通过开放数据推动社会和经济发展。政府作为最大的数据拥有者，应当成为开放数据和鼓励其合理使用的主体。

1）数据开放存在的主要问题

（1）数据公开制度不完善。在实际工作中，很难确定数据有没有涉及个人隐私、商业机密等问题。开放是相互的，很多部门没有真正意识到数据开放的重要性和作用，往往以“保密”或者“不宜公开”为理由不愿开放数据。海量数据分散在各个部门或者层级，潜在的价值被忽略。数据的开放要经过存储、清洗、分析、挖掘、处理、利用等多个环节才能形成有价值的数据集，在每一个实施环节都需要有相应的制度法规和技术标准作为依据。

（2）数据开放程度不高。首先，各地平台提供的数据总量较小，无法满足经济发展与社会创新领域的需求。很多利用价值较高的数据只在部门内部共享，并未对社会开放。其次，数据质量参差不齐。由于对数据的质量没有严格的要求，而容易造成数据失真。另外，开放数据平台门槛较高，很多数据只开放不更新，不提供下载服务，无法形成有价值的数据源。

（3）数据安全隐患。大数据开放是双刃剑，在给人们生活带来便利的同时，构成了巨大隐患。传统方法是采用划分边界、隔离内外网等来控制风险。但是随着移动互联网、云计算、5G、Wi-Fi 技术的广泛应用，网络边界已经消失，木马、漏洞和攻击都可能威胁数据安全。

2）数据开放保障机制的建议

（1）法律保障机制。完善法律体系是促进政府数据开放的必经之路，加快制定大数据管理制度、法规和标准规范是当务之急。数据开放原则、使用权限、开放领域、分级标准及安全隐私等问题都需要细化。通过制度保障保证数据安全。

（2）数据共享机制。首先，要加快国家数据库的建设，消除部门信息壁垒；其次，统筹数据管理，引导各部门发布社会公众所需的相关数据；最后，制定统一的数据开放标准和格

式，方便数据上传和下载，满足不同群体的数据需求。

（3）技术保障机制。数据的有效性和正确性直接影响到数据汇聚和处理的成果，因此必须要保障数据的质量。一旦数据来源不纯、不可信或无法使用，就会影响科学决策。针对数据体量大、种类多的数据集，需要先进技术和人才的支撑，因此，既懂统计学也懂计算机的分析型和复合型人才要加强培养。

5.3 大数据主要应用领域

5.3.1 大数据典型应用

（1）精准医学。在2011年美国医学界首次提出“精准医学”的概念后，2015年1月奥巴马在其国情咨文中提出“精准医学计划”。精准医疗（Precision Medicine）是一种将个人基因、环境与生活习惯差异考虑在内的疾病预防与处置的新兴方法。中国也随之启动了“精准医学计划”。科学技术部决定在2030年前政府将在精准医疗领域投入600亿元，其中，中央财政支付200亿元，企业和地方财政配套400亿元。

（2）精准农业。精准农业又称为精确农业或精细农作，发源于美国，是20世纪80年代末，经济发达国家继LISA（低投入可持续农业）后，为适应信息化社会发展要求对农业发展提出的一个新的课题，是信息技术与农业生产全面结合的一种新型农业。精准农业是采用3S（GPS、GIS和RS）等高新技术与现代农业技术相结合，对农资、农作实施精确定时、定位、定量控制的现代化农业生产技术，可最大限度地提高农业生产力，是实现优质、高产、低耗和环保的可持续发展农业的有效途径。精准农业将农业带入数字和信息时代，是21世纪农业的重要发展方向。

（3）精准营销。精准营销是在精准定位的基础上，依托现代信息技术手段建立个性化的顾客沟通服务体系，实现企业可度量的低成本扩张之路，是有态度的网络营销理念中的核心观点之一。公司需要更精准、可衡量和高投资回报的营销沟通，需要更注重结果和行动的营销传播计划，还有越来越注重对直接销售沟通的投资。在互联网里，人们面对的、可获取的信息（如商品、资讯等）成指数式增长，如何在这些巨大的信息数据中快速挖掘出对人们有用的信息已成为当前急需解决的问题，所以网络精准营销的概念应运而生。精准营销有三个层面的含义：第一，精准的营销思想，营销的终极追求就是无营销的营销，到达终极思想的过渡就是逐步精准；第二，是实施精准的体系保证和手段，而这种手段是可衡量的；第三，就是达到低成本可持续发展的企业目标。

（4）精准广告。也叫精准推送，指广告主按照广告接受对象的需求，精准、及时、有效地将广告呈现在广告对象面前，以获得预期转化效果。它是一种革命性的网络推广方式，即点对点推送，精准而高效。个人计算机中的Cookie文件（一种记录用户上网信息的文件）可以实现跟踪分析，利用特征关键词对用户进行分类，同时与广告主产品的特征进行关联、匹配和排序。

5.3.2 数据中心及其智能化

1. 数据中心

数据中心是企业用来在Internet基础设施上传递、加速、展示、计算、存储数据信息的基

础设施。

随着数据中心行业在全球的蓬勃发展，以及经济的快速增长，数据中心的发展建设将处于高速时期，再加上各地政府部门给予新兴产业的大力扶持，都为数据中心行业的发展带来了很大的优势。随着数据中心行业的大力发展，将来在很多城市中都会有很大的发展空间，一些大型的数据中心也会越来越多。数据中心将会成为企业竞争的资产，企业的商业模式也会因此而发生改变。

数据中心是与人力资源、自然资源一样重要的战略资源，在信息时代下的数据中心行业中，只有对数据进行大规模和灵活性的运用，才能更好地去理解数据、运用数据，才能促使我国数据中心行业快速高效发展，体现出国家发展的大智慧。海量数据的产生，也促使信息数据的收集与处理发生了重要的转变，企业也从实体服务走向了数据服务。产业界需求与关注点也发生了转变，企业关注的重点转向了数据，计算机行业从追求计算能力转变为数据处理能力，软件业将从编程为主向数据为主转变，云计算的主导权也将从分析向服务转变。在信息时代下，数据中心的产生，使更多的网络内容将不再由专业网站或者特定人群所产生，而是由全体网民共同参与。随着数据中心行业的兴起，网民参与互联网、贡献内容也更加便捷，呈现出多元化。巨量网络数据都能够存储在数据中心，数据价值也会越来越高，可靠性能也在进一步加强。

2. 数据中心智能化

数据中心被智能化后被称为数智中心，其主要包含以下几方面。

(1) 统一的数据收集及管理智能化。核心功能包括数据收集、数据持久化、数据管道等方面的智能化。

(2) 用户管理智能化。主要实现 NIS、LDAP 等多种方式的用户统一认证管理等方面的智能化。

(3) 报警管理智能化。主要实现根据设备的运行情况、发生的故障、日常的巡检结果，系统通过预设模板定时生成运维、巡检报告，并自动推送给指定人员。

(4) 可视化展示智能化。主要实现全面性能、链路监控及可视化展示等方面的智能化。对计算中心资源的服务器、存储、CPU、内存、网络(流入/流出速率、丢包率和错包率) 以及风火水电等场地设备等各项指标的实时监测提供可视化展示功能。

(5) 配置管理数据库(Configuration Management Database，CMDB)智能化。主要实现设备资产的导入/导出、变更管理、变更对比、自动添加以及资产检索等方面的智能化。

(6) 故障定位及修复管理智能化。主要实现设备及系统服务的故障自动发现、追溯、定位、分析、恢复，且通过以往解决的经验进行解决方案的自动推送，提升整体运维效率。

(7) 业务拓扑智能化。主要实现资产网络拓扑图的自动生成、资产拖曳、资产关联、告警联动、业务收/放。

(8) 知识库建设。主要实现运维单自动创建、故障评级、故障知识分类、故障检索、故障智能推送、知识统计。

5.3.3　大数据未来发展趋势

1. 未来推动大数据行业市场需求增长的主要因素

(1) 物联网的兴起。得益于物联网的发展，智能手机已经应用于控制家用电器上，

Google Assistant、小米、小爱相关智能设备在家庭中实现特定任务自动化。

（2）人工智能的广泛应用。AI 执行任务比人类更快、更精确，从而能够减少错误并改善整体流程，这使得人们可以更加专注于更关键的任务，同时提升服务质量。

（3）暗数据迁移到云。尚未转换为数字格式的数据称为暗数据，它是尚未开发的巨大存储库，未来这些模拟数据库将被数字化并在迁移到云中，它们的利用，有利于企业进行预测分析决策。

（4）量子计算。尽管量子计算尚处于起步阶段，但相关研究实验从未停止，量子计算将能够极大提升计算机数据处理能力，缩短处理时间。很快，诸如 Google、IBM 和 Microsoft 之类的大型科技公司将开始测试量子计算机，并将其集成到他们的业务流程中。

（5）边缘计算。随着物联网发展，企业收集数据方式逐渐转向设备端，对数据传输处理延迟提出更高要求，由于边缘计算相对云计算更加靠近数据源头，可以有效降低数据传输处理到反馈的迟延，同时具有显著效率成本优势和安全隐私保护优势。

（6）开源解决方案。越来越多的免费数据和软件工具被公开可用，例如开源软件，在加速数据处理方面取得了较大进步，同时具有实时访问和响应数据功能，小型组织和初创企业将从中受益。毫无疑问，开源软件更加便宜，可以帮助企业降低运营成本，未来将会蓬勃发展。

2. 大数据行业发展趋势

（1）数据安全与隐私保护需求日渐高涨。安全与隐私保护意识日益增强，敏感信息约束和数据安全检查成为互联网、移动端的用户数据管控的难点。

（2）单一大数据平台向大数据、人工智能、云计算融合的一体化平台发展。传统的单一大数据平台已经无法满足用户的应用需求，而将大数据、人工智能、云计算融于一体的大数据分析平台，将成为一个联系 IT 系统与从员工、客户、合作伙伴、社会，到设备的一个全生态的核心系统。

（3）发挥数据价值，企业逐步推进数字化转型。越来越多的企业将“数字”视为核心资源、资产和财富，纷纷制定数字化转型战略，以抢占数字经济的新的制高点。大部分企业将制定数字化战略，积极推进数字化转型，达到敏捷业务创新和业务智能的发展目标。

（4）中小大数据企业聚焦细分市场领域。随着大数据产业链逐渐完善，在大数据各个细分市场，包括硬件支持、数据源、交易层、技术层、应用层以及衍生层等，都出现了大量不同类型的中小独立公司，带动各自领域的技术与应用创新。随着大数据应用层次深入，产业分工将更加明确，中小企业需聚焦于细分领域，推出针对性产品，寻找自己的定位与价值。

思 考 题

1. 什么是大数据？大数据产生的背景是什么？
2. 大数据发展有几个阶段？
3. 大数据涉及哪些关键技术？
4. Hadoop 是什么？
5. 有哪几种爬虫？什么是数据清洗？什么是 ETL？

6. 什么是数据可视化？有几哪种可视化技术？
7. 什么是数据整合？什么是数据共享？什么是数据开放？
8. 大数据主要有哪几方面的应用？
9. 什么是数据中心？什么是数据中心智能化？
10. 大数据未来发展趋势是什么？