大数据关键技术、主要特点及发展趋势

李纪舟 叶小新 丁云峰 朱党明

大数据是继物联网、云计算技术后世界又一热议的信息技术，发展迅速。大数据领域出现的许多新技术，是大数据采集、存储、处理和呈现的有力武器。本文从大数据的关键技术、特点分析、发展趋势三个方面进行阐述。

1 大数据关键技术

根据大数据的处理过程，大数据生命周期（图1）分为数据获取、数据预处理、数据存储、数据分析、数据检索、数据呈现、数据应用和数据安全等环节。由于大数据具有大规模、异构、多源等特点，大数据技术与传统的数据处理技术有所不同。在大数据处理的每个环节，都出现了许多针对大数据独特需求的新兴技术，这些技术大概分为11类（图2）。

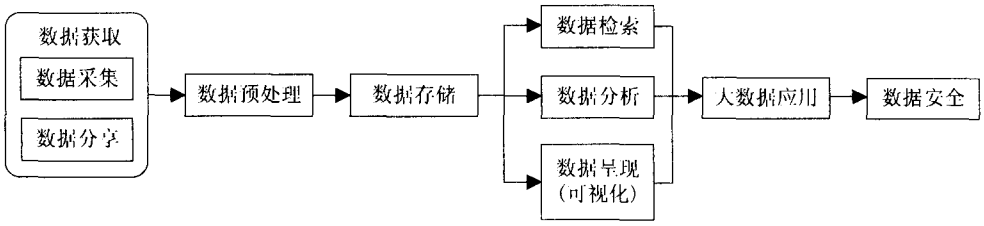

图1 数据生命周期图

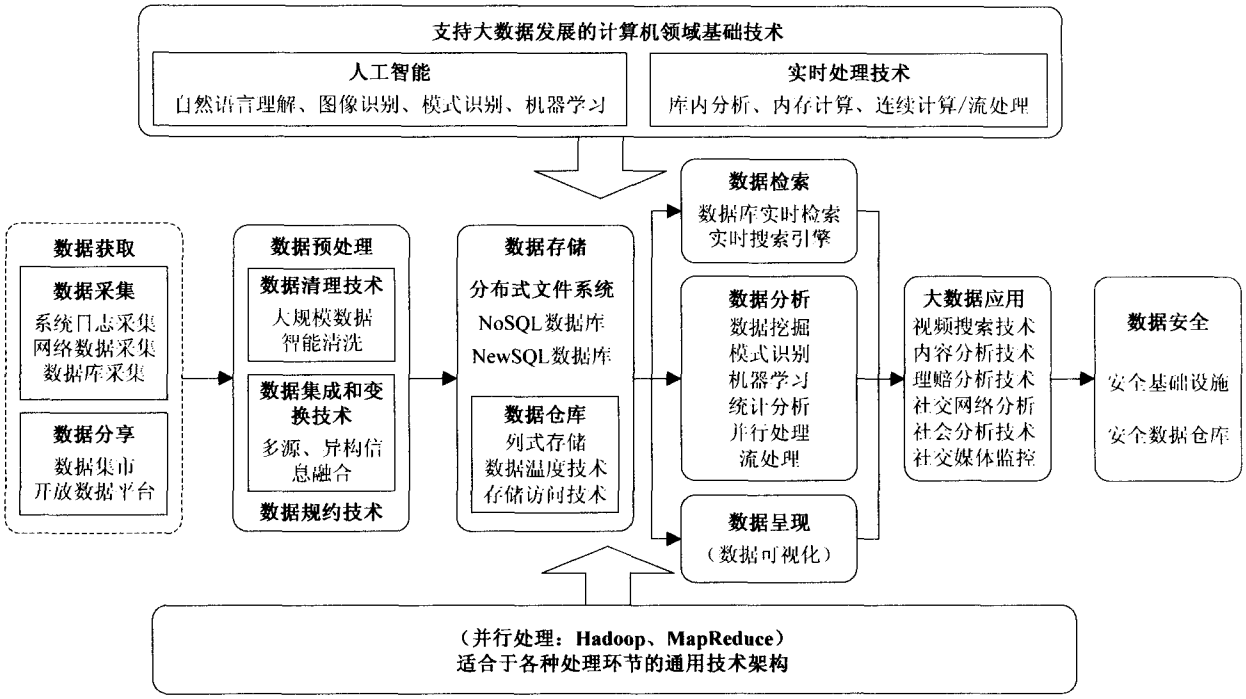

图2 大数据技术分类图

1.1 大数据处理通用技术架构

大数据的基本处理流程与传统数据处理流程的主要区别在于：由于大数据要处理大量、非结构化的数据，所以在各个处理环节中都可以采用并行处理。目前，MapReduce 等分布式处理方式已经成为大数据处理各环节的通用处理方法。

MapReduce 分布式方法最先由谷歌设计并实现，包括分布式文件系统 GFS (Google FileSystem)、MapReduce 分布式编程环境以及分布式大规模数据库管理系统 BigTable。

Hadoop 是谷歌分布式处理系统框架的开源实现，它在可伸缩性、健壮性、计算性能和成本上具有无可替代的优势。目前，Hadoop 已经成为大数据生态环境中不可或缺的一环，是拥有海量数据处理需求的公司的标准配置，许多商业创新和产品创新也围绕 Hadoop 展开。

（1）MapReduce 分布式处理技术

MapReduce 是一套软件框架，包括 Map 和 Reduce 两个阶段，可以进行海量数据分割、任务分解与结果汇总，从而完成海量数据的并行处理。

MapReduce 的工作原理是先分后合的数据处理方式。Map 即“分解”，把海量数据分割成若干部分，分给多台处理器并行处理；Reduce 即“合并”，把各台处理器处理后的结果进行汇总操作，以得到最终结果。用户只需要提供自己的 Map 函数以及 Reduce 函数就可以在集群上进行大规模的分布式数据处理。MapReduce 将处理任务分配到不同的处理节点，因此具有更强的并行处理能力。

（2）开源分布式软件架构 Hadoop

Hadoop 是 Apache 软件基金会开发的分布式密集数据处理和数据分析的软件框架。目前大数据处理各个环节的工具普遍采用 Hadoop 架构，亚马逊、微软、IBM、甲骨文等大数据产品供应商也纷纷提供了基于 Hadoop 的大数据处理工具。

Hadoop 框架中包括以下主要项目：

- HDFS：Hadoop 分布式文件系统(Hadoop Distributed FileSvstem)。
- MapReduce：并行计算框架。
- HBase：类似 GoogleBigTable 的分布式 NoSQL 列数据库。
- Hive：数据仓库工具。
- Zookeeper：分布式锁设施，提供类似 Google Chubby 的功能。
- Avro：新的数据序列化格式与传输工具。
- Pig：大数据分析平台，为用户提供多种接口。

1.2 大数据采集

大数据的采集是指利用数据库等方式接收发自客户端（Web、App 或者传感器形式等）的数据。大数据采集的主要特点是并发访问量大，因为同时有可能会有成千上万的用户来进行访问和操作，比如火车票售票网站的并发访问量在峰值时达到上百万，这时传统的数据采集工具很容易失效。大数据采集方法主要包括：系统日志采集、网络数据采集、数据库采集、其他数据采集等四种。

1.3 大数据分享

目前数据分享主要通过数据集市和开放数据平台等方法实现。开放数据平台可以提供涵盖本地服务、娱乐、教育和医疗等方方面面的数据集合，用户不但可以通过 API 访问，还可以很方便地通过 SDK 集成到移动应用当中。在线数据集市除了提供下载数据的功能外，还为用户提供上传和交流数据的场所。数据平台和数据集市不但吸引有数据需求用户，还能够吸引很多数据开发者在平台上进行开发。

1.4 大数据预处理

数据预处理就是对采集的数据进行清洗、填补、平滑、合并、规格化以及检查一致性等

处理，并对数据的多种属性进行初步组织，从而为数据的存储、分析和挖掘做好准备。通常数据预处理包含三个部分。

（1）数据清理

数据清理包含遗漏值处理、噪音数据处理以及不一致数据处理。对于大型数据库，某个属性中的数据往往会有遗漏，对这种情况有忽略该值、人工填写、全局常量填充、平均值填充、使用最可能的值填充等方法，可能值可以通过回归分析、贝叶斯形式方法或判定树等得出。

对于噪音数据，可以用分箱、聚类、计算机人工检查和回归等方法去除噪音。分箱技术是把数据分类然后用合理的数值替换原先数据，除去原数据中的噪音；聚类技术是通过“距离”等判别把数据进行概念分层，过渡到更高一级的层次；回归技术则是利用回归模型，用模型预测值代替原有数据。对于不一致数据，可以进行手动更正。

（2）数据集成和变换

数据集成是把多个原数据中的数据结合、存放到一个数据库中存储，如数据仓库。这一过程中主要考虑三个问题:实体识别、数据冗余和数据值冲突检测与处理。数据变换是数据处理的必然结果，主要过程有平滑、聚集、数据泛化、规范化以及属性构造。

（3）数据规约

分析大型数据库中的海量数据是个很庞大的工程，如果对所有数据进行分析和挖掘，将要耗费很长的时间。数据规约能够把握主要数据，加快分析速度。主要包括:数据方聚集、维规约、数据压缩、数值规约和概念分层等。

1.5 大数据存储及管理

大数据需要行之有效的存储和管理，否则人们不能处理和利用数据，更不能从数据中得到有用的信息。目前，大数据的存储和管理技术主要分三类。

（1）分布式文件系统

分布式文件系统将大规模海量数据用文件的形式保存在不同的存储节点中，并用分布式系统进行管理。其目的是解决复杂问题，将大的任务分解为多个小的任务，通过让多个处理器或多个计算机节点参与计算来解决问题。分布式文件系统能够支持多台主机通过网络同时访问共享文件和存储目录，这使多台计算机上的多个用户能够共享文件和存储资源。典型的分布式文件系统产品有GFS (Google File System 文件系统)、HDFS (Hadoop 分布式文件系统)以及分布式数据库 HBase。

（2）数据仓库

数据仓库采用更适于数据查询的技术，以列式存储或 MPP（大规模并行处理)两大成熟技术为代表。数据仓库往往适合于存储关系复杂的数据模型（例如企业核心业务数据)，并且需要限制为基于二维表的关系模型。同时，数据仓库适合进行一致性与事务性要求高的计算，以及复杂的 BI（商业智能)计算。在数据仓库中，经常使用数据温度技术、存储访问技术来提高性能。

- 列式存储

列式存储将数据按行排序，按列存储，将相同字段的数据作为一个列族来聚合存储。不同的列族对应数据的不同属性，这些属性可以根据需求动态增加，通过这样的分布式实时列式数据库对数据统一进行结构化存储和管理，避免了传统数据存储方式下的关联查询。

当只查询少数列族数据时，列式数据库可以减少读取数据量，减少数据装载和读入读出的时间，提高数据处理效率。按列存储还可以承载更大的数据量，获得高效的垂直数据压缩能力，降低数据存储开销。

- 数据温度技术

数据温度技术可以区分经常被访问和很少被访问的数据。经常访问的就是高温数据，这类数据将存储在高速存储区，访问路径会非常

直接，而低温数据则可以放在非高速存储区，访问路径也可相对复杂一些。

● 存储访问技术

近两年，存储访问的技术也在变化着，比如 Teradata 的固态硬盘数据仓库，用接近闪存的性能访问数据，比原来在磁盘上顺序读取数据快很多。后来的内存数据库产品在数据库管理系统软件上进行优化，规避传统数据库（数据仓库)读取数据时的磁盘 IO 操作，再次大大节省访问时间。

目前典型的数据仓库产品有数据仓库一体机 IBM Netezza、自服务数据仓库 EMC Greenplum Choru,Teradata, SybaseIQ 等等。

（3）非关系型数据库（NoSQL）

NoSQL 是区别于传统关系型数据库的数据库管理系统的统称。与关系型数据库相比，NoSQL 最大的不同是不使用 SQL(结构化查询语言)作为查询语言，其数据存储可以不依照固定的表格模式，通常具备水平可扩展的特征。NoSQL 越来越普及，几乎所有的大型互联网公司都在这个领域进行着实践和探索，比如搜索、准实时统计分析、简单事务等。典型的 NoSQL 产品有 MongoDB、CouchDB、DynamoDB、Neo4j 等等。

（4）NewSQL

NewSQL 是改进后的 SQL（结构化查询语言)系统，是对各种新的可扩展/高性能的 SQL 数据库的简称，可提供 SQL 独有的一些特性，同时还具备 NoSQL 的扩展性。NewSQL 把关系模型的优势发挥到分布式体系结构中，或者提高关系数据库的性能到一个不必进行横向扩展的程度，以满足可扩展性需求和无模式数据管理需求。

1.6 大数据分析及挖掘

大数据的分析和挖掘是一种决策支持过程，它主要基于人工智能、机器学习、模式识别、数据挖掘、统计学、数据库等技术，高度自动化地分析大数据，做出归纳性的推理，从中挖掘出潜在的模式，从而在大数据中提取有用信息。

大数据的分析和挖掘与传统的数据挖掘比较有两个特点：一是通常采用并行处理的方式；二是大数据分析对实时处理的要求很高，流处理等实时处理技术受到人们欢迎。

（1）机器学习

机器学习（Machine Learning）是研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能，是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。

（2）数据挖掘

数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。其含义包括:数据源必须是真实的、大量的、含噪声的；发现的是用户感兴趣的知识；发现的知识要可接受、可理解、可运用；并不要求发现放之四海皆准的知识，仅支持特定的发现问题。

（3）模式识别

模式识别（Pattern Recognition）是指对表征事物或现象的各种形式的（数值的、文字的和逻辑关系的）信息进行处理和分析，以对事物或现象进行描述、辨认、分类和解释的过程，是信息科学和人工智能的重要组成部分。

（4）统计分析

对于大数据的统计分析主要利用分布式数据库，或者分布式计算集群来对存储于其内的海量数据进行普通的分析和分类汇总等，以满足大多数常见的分析需求。一些实时性需求会用到 EMC 的 GreenPlum、Oracle 的 Exadata，以及基于 MySQL 的列式存储 Infobright 等，而一些批处理，或者基于半结构化数据的需求可以使用 Hadoop。

（5）并行处理

大数据分析的三大挑战是数据量的膨胀、数据深度分析需求的增长和数据类型不断多样化。面对这些挑战，传统的关系型、结构化处理方式已经力不从心，需要采用MapReduce等并行处理方式，将海量数据进行分解并分布存储，由数据挖掘系统并行处理，然后将多个局部处理结构合成最终的输出模式，实现海量数据挖掘。

1.7 大数据检索

（1）数据库实时检索

数据库实时检索是指在数据仓库或者NoSQL等大数据存储平台上，或者多个不同结构的数据存储平台之间快速、实时地查询和检索不同结构的数据。

（2）实时搜索引擎

实时搜索是对互联网上的大量数据和信息进行即时、快速搜索，实现即搜即得的效果。目前各大搜索引擎都在致力于实时搜索的实现。

1.8 大数据可视化

数据可视化可以提供更为清晰直观的数据感官，将错综复杂的数据和数据之间的关系，通过图片、映射关系或表格，以简单、友好、易用的图形化、智能化的形式呈现给用户供其分析使用，可通过数据访问接口或商业智能门户实现，通过直观的方式表达出来。可视化与可视分析通过交互可视界面来进行分析、推理和决策；从海量、动态、不确定甚至相互冲突的数据中整合信息，获取对复杂情景的更深层的理解；可供人们检验已有预测，探索未知信息，同时提供快速、可检验、易理解的评估和更有效的交流手段。可视化是人们理解复杂现象，诊释复杂数据的重要手段和途径。

1.9 大数据应用

（1）视频搜索

视频搜索是指对收集的视频进行搜索，它把社交网络、社交标签、元数据提取和应用程序、音频转录和传统搜索技术的相关元素合并。

（2）内容分析

内容分析定义为处理内容以及用户消费内容的行为，以获得特定问题答案的一系列技术。内容的类型包括各种类型的文本（包括文档、博客、新闻网站、客户对话）和社交网站上的交流互动。分析方法包括文本分析、流媒体和语音分析、以及情绪、情感意图和行为分析。

（3）理赔分析

理赔分析为使用商业智能、报告解决方案、仪表盘、数据挖掘和预测建模技术对理赔数据进行的更高的性能管理和分析。总体而言，理赔分析工具支持需求分析、报告和预测建模三个过程。

（4）社交网络分析

社会网络分析（SNA）工具被用来分析人群之间的关系。可以用来研究个人或组织的社会结构和相互关系（或工作方式）。SNA包括收集多个数据源、分析数据、确定关系并挖掘关系的质量或有效性等步骤。

（5）社会分析

社会分析描述了对交互和关联结果进行收集、测量、分析和解释的过程。这种交互可能发生工作场合、内部或者外部社区以及社交网络中。社会分析本身是个涵盖性的专业术语，包括一系列专业的分析技术，如社会过滤、社会网络分析、情绪分析以及社交媒体分析等。

（6）社交媒体监控

社交媒体监控器（SMMs）是一种软件服务，用来跟踪社交媒体中被提到和讨论的兴趣话题，被用于对互联网上的大量评论进行搜索、监控和轻度的筛选、过滤。通过对社交媒体中大量分散的评论进行分析，提供对人们兴趣热点的频繁更新和隐含趋势跟踪。

1.10 大数据安全

大数据技术的发展，使得人们能够从这些

数据中观察和分析社会动态、人群的动作和行为、人群活动规律以及企业的商业秘密。海量数据本身，以及数据中蕴藏的信息涉及到国家、社会、企业和人们的隐私，这对大数据时代的信息安全提出巨大挑战。因此，大数据时代需要发展信息安全技术，确保关系到人们生活方方面面的数据和信息不会被泄漏。

目前除了传统的信息安全方法外，大数据领域还有安全基础设施、安全数据仓库等。此外，一些数据库安全管理软件能够对不同操作系统上运行的异构关系型数据库进行实时监控，一些大型安全数据库能够对与商务数据结合在一起的数据进行预防性的分析，以便识别钓鱼攻击，防止诈骗和阻止黑客入侵。

1.11 支持大数据发展的计算机领域基础技术

（1）人工智能技术

随着各种形式的非结构化数据不断增多，旨在从庞大的非结构化数据中获得知识和洞察力的计算机工具也在迅速发展中。这些工具的发展多依赖于不断进步的人工智能技术。人工智能是研究、开发用于模拟、延伸和扩展人的智能的理论、方法、技术及应用系统的一门新的技术科学，是计算机科学的一个分支。它企图了解智能的实质，并生产出一种新的能以与人类智能相似的方式做出反应的智能机器，该领域的研究包括语音识别、图像识别、自然语言处理、机器学习、模式识别、数据挖掘和专家系统等。

（2）实时处理技术

大数据时代最明显的特征之一就是数据处理速度的不断增加，用户要求对数据要有更深的洞察力，并能够实时访问最新的数据。

- 库内分析

库内分析是指数据在被过滤和处理之前不会离开数据库，也就是说，数据分析在数据库内可以即时完成，这就节省了数据移动所占用的大量时间。同时，通过将数据保持在数据库内，还可提高数据的安全性。

- 内存计算

内存计算技术支持在服务器的主内存中处理超大量的实时数据，从分析和交易中提供及时的结果。

- 连续计算/实时流处理

连续计算是指对数据流进行连续查询和分析，在计算时就将结果以流的形式输出给用户。

2 大数据技术的主要特点

（1）开源软件受到广泛欢迎

开源项目和产品正在主导新兴的大数据市场。分布式处理的软件框架 Hadoop、用来进行数据挖掘和可视化的软件环境、非关系型数据库 HBase, MongoDB 和 CouchDB 等开源软件都在大数据技术领域占据重要地位，2012 年排名前 5 位的数据挖掘工具中，有 4 个是开源软件。

（2）人工智能技术不断融入

“大数据”可以看作是对大规模数据集合的智能分析处理，能够帮助人们从似乎无穷多的数据中发现信息、发现规则、发现知识、发掘智慧，进而对未来态势发展做出预测。要想对大数据做出智能处理，就必须要用人工智能技术。大数据的管理与分析和可视化等技术无不与人工智能相关联，目前机器学习、数据挖掘、自然语言理解、模式识别等人工智能技术已经深深融入到大数据各流程的处理技术之中。

（3）非结构化数据处理技术受到重视

云计算时代的到来使得数据创造的主体由企业逐渐转向个体，而个体所产生的绝大部分数据为图片、文档、视频等非结构化数据。因此，对非结构化数据的处理需求越来越强烈，非结构化数据采集技术、NoSQL 数据库、流处理技术正取得快速发展。

（4）分布式处理架构成为大数据处理普遍模式

由于大数据要处理大规模、海量、异构的数据，传统的处理方法在存储空间、处理时间和效率上都难以满足人们对大数据处理的要求，

所以在各个处理环节中都普遍采用分布式方法进行并行处理。此外，云计算技术也是以分布式处理为核心的。目前，MapReduce等分布式处理方式已经成为大数据处理各环节的通用处理方法，分布式文件系统、大规模并行处理数据库、分布式编程环境等技术也普遍被使用。

3 大数据技术发展趋势

（1）技术趋向多样化

目前，大数据相关的技术和工具已经非常多，在未来还会继续出现新的技术和工具。在大数据生命周期的各个环节，不论是大数据的采集、存储、管理，还是分析、可视化以及应用都将出现创新。

（2）数据分析将成为大数据技术核心

数据分析是大数据的核心，数据分析技术是大数据技术的核心。“大数据”的价值体现在对大规模数据集合的智能处理，从而在无穷多的数据中发现信息、知识和智慧。要想实现这样的价值，最关键的步骤就是对数据的分析和挖掘。数据的采集、存储和管理都是数据分析步骤的基础，数据分析得到的智能结果可以应用到大数据相关的各个领域。未来大数据将充分利用机器学习、数据挖掘、模式识别、自然语言理解等人工智能基础技术，进一步实现数据分析的智能化。

（3）流线化、实时性的数据处理将广泛采用

目前大数据处理系统大多采用的是批量化的处理方式，这种方式适用于数据报告的频率不需要达到分钟级别的场合。传统的数据仓库系统、BI、链路挖掘等应用对数据处理的时间要求往往以小时或天为单位，但“大数据”应用突出强调数据处理的实时性。在线个性化推荐、股票交易处理、实时路况信息等数据处理时间要求在分甚至秒级。在一些大数据应用场合，由于没有足够的空间来存储接收到的所有数据，人们需要对数据实时处理并实时扔掉。因此，在未来几年中，内存计算、流处理、连续计算等实时计算技术将迅速发展，用于处理流线化和实时性分析的可伸缩框架和平台将会广泛应用。

（4）基于云的数据分析平台将更趋完善

云计算的发展为大数据提供了良好的处理平台和技术支持。云计算为大数据提供了分布式的计算方法与可以弹性扩展、相对便宜的存储空间和计算资源；此外，云计算IT资源庞大、分布较为广泛，是异构系统较多的企业及时准确处理数据的有力方式。未来，基于云平台的大数据分析工具和数据库将日趋成熟，推动大数据技术进一步发展。

（5）开源软件成为推动大数据技术发展新动力

开源软件的盛行不会抑制商业软件发展，相反开源软件将会给基础架构硬件、应用程序开发工具、应用、服务等各个方面的相关领域带来更多的机会。例如开源软件架构Hadoop推出后，成为大数据处理的通用架构，很多厂商基于开源Hadoop推出了自己的商业化产品。未来开源软件和商业化软件并存的局面还将持续，二者相互促进，共同发展。

（6）关键技术展望

未来大数据技术的发展将以数据分析和数据管理为中心，以人工智能、实时计算、分布式计算和数据库技术为基础。具体来看，以下几种技术将成为大数据发展的关键。

- 数据采集和预处理：多源、异构信息融合技术、大规模数据智能清洗技术。
- 数据管理：NoSQL数据库、NewSQL数据库。
- 数据分析和应用：知识计算（搜索）技术、大规模异构数据挖掘技术、大规模异构数据实时分析技术、面向行业的数据分析知识库、面向行业的大数据综合应用平台、大规模数据可视化分析技术。
- 数据安全：大数据异常检测技术、信息安全监测技术。