模块 1　大数据生态系统

本模块从了解大数据(big data)入手,简明扼要地叙述大数据的产生、大数据的概念、大数据的“4V”特征、大数据应用案例,展示物联网(产生数据)、云计算(承载数据)、大数据(挖掘数据)和人工智能(学习数据)相辅相成、彼此依附和相互助力的关系,再通过对 Hadoop 的层层“揭秘”来认识 Hadoop 和它的核心组件及其常用的其他组件。

通过本模块的学习,学生将达到以下职业能力目标和要求。

- 了解大数据(产生、概念、特征)及“物、云、大、智”的关系;
- 了解 Hadoop 的应用案例;
- 了解及认识 Hadoop 和它的核心组件;
- 了解 MapReduce 作业的运行方式。

1.1　了解大数据

被誉为“大数据之父”的维克托·迈尔-舍恩伯格曾提到“世界的本质就是大数据”,当今社会生活中到处都是数据。不仅如此,在人人互联的廉价存储时代,我们收集的数据的性质也在发生变化。对于许多企业而言,它们的关键数据曾经仅限于其业务数据库和数据文件。在这些类型的系统中,数据被组织成有序的行和列,其中信息的每个字节在其性质和业务价值方面都很好管理和理解。当今这些数据和数据库仍然非常重要,但是应用的数据类型和数据处理方式发生了翻天覆地的变化,大数据悄然而至,让我们一起揭开大数据的神秘面纱。

1.1.1　大数据的产生

大数据的产生可追溯至 1887 年。1887—1890 年,美国统计学家赫尔曼·霍尔瑞斯为了统计 1890 年的人口普查数据,发明了一台电动器来读取卡片上的洞数,该设备让美国用 1 年时间就完成了原本耗时 8 年的人口普查活动,由此在全球范围内引发了数据处理的新纪元。

而现在,时钟的秒针每旋转一圈,超过 13 000 个 iPhone 应用被下载、Skype 上产生超过 37 万分钟的语音通话、新浪微博上会发布超过 98 000 条新微博、会有超过 6 600 张照片上传到微信朋友圈里、会有超过 1.68 亿封 E-mail 被发出、Facebook 上更新超过 69.5 万条新状态、抖音上上传超过 600 条新视频。

社交网络、电子商务等互联网应用成为新的数据来源,传感器、二维码、无线射频识别

(radio frequency identification,RFID)、位置信息等物联网应用成为新的数据采集方法;全时空数据的可采集性应用;智能算法的使用;非结构的数据形态飞速增加;数据获取成本、存储成本和处理成本的下降,所有因素都推动了数据量的爆炸式膨胀。

数据大爆炸和数据结构的变化,为数据处理带来新的挑战,人类正从信息技术(information technology,IT)时代走向数据技术(data technology,DT)时代"。

1.1.2 大数据的概念

大数据可一拆为二来看,就是"大"和"数据"。

何为大?数据最小的基本单位是 bit,可存储一个 '0' 或者 '1',8 bit 相当于 1 Byte。数据量等级单位有 Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB,它们依次按照进率 1 024(2^{10})来计算。常规 PC 的存储和处理数据的体量一般到达 GB 级别,而互联网、企业 IT、物联网、短信、电话、网络搜索、在线交易等,随时都在快速累积庞大的数据,数据量很容易达到 TB、PB 或 EB 等级,没有办法在可容忍的时间下使用常规软件方法完成存储、管理和处理任务。等级界值分分钟就临近了,"大数据"的概念延伸而出。

什么是数据?在计算机科学中,数据是指所有能输入计算机并被计算机程序处理的符号介质的总称,是用于输入电子计算机进行处理,具有一定意义的数字、字母、符号和模拟量等的统称。计算机存储和处理的对象十分广泛,表示这些对象的数据也随之变得越来越复杂。例如,应用下载记录、语音通话记录、淘宝"双 11"新订单记录等都为数据,有些是由二维表结构来逻辑表达和实现的数据,严格地遵循数据格式与长度规范,主要通过关系型数据库进行存储和管理,这种称为结构化数据。与之相对的是有些数据则"杂乱无章"或部分有序,不适于由数据库二维表来表现,包括所有格式的办公文档、XML、HTML、各类报表、图片和音频、视频信息等,这种称为非结构化或半结构化数据。

对于大数据,研究机构 Gartner 给出了定义,大数据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的海量、高增长率和多样化的信息资产。

1.1.3 大数据的特征

2001 年,即市场营销人员认识到"大数据"一词之前的几年,分析公司 META Group 发布了一份题为《3D 数据管理:控制数据量、速度和多样性》的报告。这篇报告涉及数据仓库挑战,以及使用关系技术克服挑战的方法。为了连续性和易理解性,接下来将进行大数据的特征描述。

大数据的"3V"特征,描述了数据挑战"大"的背后因素:

(1)数据体量巨大(volume):从 TB 级别跃升到 PB 级别。

(2)数据类型繁多(variety):以多种结构组织的数据,从原始文本(从计算机的角度来看,几乎没有或没有明显的结构,称为非结构化数据)到日志文件(通常称为半结构化数据)到有序数据在强类型的行和列(结构化数据)中。某些数据集甚至包括所有三种数据的一部

分(这被称为多结构数据)。

(3)处理速度快(velocity),1 秒定律,可从各种类型的数据中快速获得高价值的信息,这一点也和传统的数据挖掘技术有着本质的不同。进入组织并在有限的时间范围内具有某种价值的数据——通常在数据被转换并加载到数据仓库中以进行更深入的分析之前,该窗口会很好地关闭分析(如金融证券报价数据,它可能会显示购买机会,但仅显示一小段时间)。每秒进入组织的数据量越大,速度挑战就越大。

这些特征中的每一个显然对想要分析信息的人提出了自己独特的挑战。因此,这三个特征标准是评估大数据问题并提供清晰口号的简便方法。通常的经验法则是,如果数据存储和分析工作具有这三个特征中的任何一个,则很可能会带来巨大的数据挑战,须利用大数据技术来解决。另外大数据不仅仅是技术,关键是产生价值,前"3V"决定了挖掘大数据的价值类似沙里淘金,从海量数据中挖掘稀疏但珍贵的信息,价值密度低(value),是大数据的另一个典型特征。于是,业界将大数据特征归纳为"4V"——volume(数据体量巨大)、variety(数据类型繁多)、velocity(处理速度快)、value(价值密度低)。

大数据最核心的价值就是对于海量数据进行存储和分析。相比起现有的其他技术而言,大数据的"廉价、迅速、优化"这三方面的综合成本是最优的。

1.1.4　大数据应用案例

大数据应用颠覆人类思维,产生巨大成功的案例数不胜数,下面略举几例,一起来领略一下大数据的神奇应用。

1. 大数据应用案例之颠覆思维

Kaggle,一个为所有人提供数据挖掘竞赛的公司,在一次关于二手车的数据分析比赛中得到,橙色汽车有质量问题的可能性是其他颜色汽车的一半。为什么?

探寻事物的因果关系是人类的本性,但是大数据时代可以做某种程度的妥协,可以只需要关注"是什么",而忽略"为什么"。

2. 大数据应用案例之乔布斯抗癌

苹果公司的传奇总裁史蒂夫·乔布斯在与癌症斗争的过程中采用了不同的方式,对自身 DNA 和肿瘤 DNA 进行排序,他得到的不是一个只有一系列标记的样本,而是包括整个基因密码的数据文档。对于一个普通的癌症患者,医生只能期望他的 DNA 排列同试验中使用的样本足够相似。但是,史蒂夫·乔布斯的医生们能够基于乔布斯的特定基因组成,按所需效果用药。如果癌症病变导致药物失效,医生可以及时更换另一种药,也就是乔布斯所说的,"从一片睡莲叶跳到另一片上。"乔布斯开玩笑说:"我要么是第一个通过这种方式战胜癌症的人,要么就是最后一个因为这种方式死于癌症的人。"虽然他的愿望都没有实现,但是这种获得所有数据而不仅是样本的典型大数据思维方法还是将他的生命延长了好几年。大数据为人类的生命延续开启了一扇新的窗户。

3. 大数据应用案例之总统竞选

美国许多人通过 Facebook 更新个人状态、分享图片以及他们“喜欢”的内容。奥巴马的总统竞选活动也通过使用社交网络的各种数据功能完成了竞选，他们不仅通过社交网络寻找支持者，还通过社交网络召集了一批志愿军。

早在 2006 年，Facebook 联合创始人克里斯·休斯就建议扎克伯格在网站上推出相关服务，帮助总统候选人在 Facebook 上建立个人主页，以便他们进行形象推广。2006 年 9 月，Facebook 全面开放，用户数量爆炸式增长，在当年年底达到 1 200 万，这一过程恰好有力地推升了奥巴马的知名度。此后，在克里斯的辅佐下，奥巴马掀起了一系列的网络活动，在 Facebook、MySpace 等社交网站上发表公开演讲、推广施政理念，赢得大量网民支持，募集到 5 亿多美元的竞选经费。

最终，“黑人平民”战胜了实力雄厚的对手，成为美国历史上第一位黑人总统，之后，在第二次的选举中更获得连任。此次选举被认为是美国民主的巨大进步，而互联网则提供了前所未有的实施手段，其中尤以 Facebook 代表的社交网站最为突出，以至于有人戏称之为“Facebook 之选”。

4. 大数据应用案例之流感预测

Google Flu Trends 通过对聚合搜索的结果进行分析，可以比疾控机构更快速地侦测到疾病的暴发。而且，尽管卫生报告每周都得到更新，但报告仅限于单个国家。Google Flu Trends 却有着几近涵盖全球的视角：它在任何人们使用 Google 搜索的地点收集数据。更重要的是，由于它是每日更新的，因而它可以向人们传递更即时的消息。

5. 大数据应用案例之音乐

10 多年前，音乐元数据公司 Gracenote 收到来自苹果公司的神秘忠告，建议其购买更多的服务器。Gracenote 照做了，而后苹果推出 iTunes 和 iPod，Gracenote 从而成为元数据的帝国。

在车内听的歌曲很可能反映你的真实喜好，Gracenote 就拥有此种技术。它采用智能手机和平板电脑内置的麦克风识别用户电视或音响中播放的歌曲，并可检测掌声或嘘声等反应，甚至还能检测用户是否调高了音量。这样，Gracenote 可以研究用户真正喜欢的歌曲、听歌的时间和地点。

Gracenote 拥有数百万首歌曲的音频和元数据，因而可以快速识别歌曲信息，并按音乐风格、歌手、地理位置等分类。

大数据应用案例枚不胜举，请读者结合自己生活学习中的应用，分析还有哪些大数据的应用。

1.1.5 “物、云、大、智”的关系

智能时代的今天，“物、云、大、智”存在于生活的每个角落，从产品营销至信息服务，从日常生活应用至高端科学研究。大数据的应用与研究总是与物联网、云计算、人工智能紧紧相

连，它们的关系如图 1-1 所示。

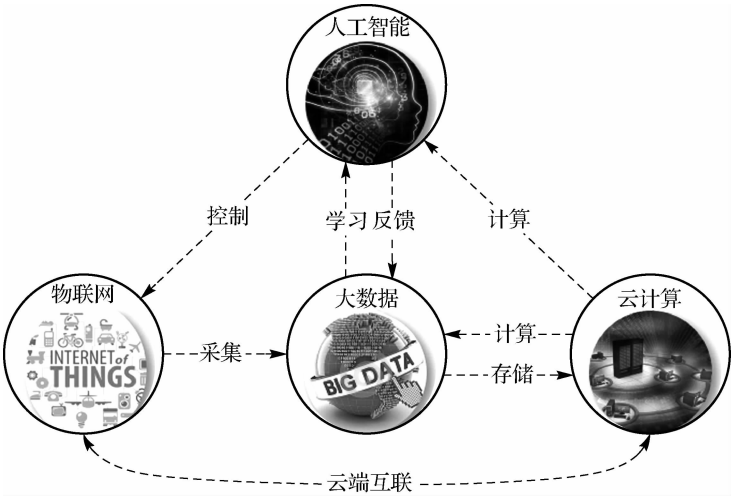

图 1-1　物联网、云计算、大数据和人工智能关系

1. 人工智能——智能时代的应用领域

人工智能的应用领域比较多，如机器人领域、语言识别领域、图像识别领域和专家系统等，而它的应用实例也是数不胜数，如指纹识别、人脸识别、视网膜识别、虹膜识别、智能搜索和博弈等。

2. 大数据——人工智能背后的基石

大数据是人工智能的基石，目前的深度学习主要建立在大数据的基础上，即对大数据进行训练，并从中归纳出可以被计算机运用在类似数据上的知识或规律。

3. 物联网 IOT——人工智能基础

IOT 全称是 Internet of things，可以简单地理解为物物相连的互联网，正是得益于大数据和云计算的支持，互联网才正在向物联网扩展，并进一步升级至体验更佳、解放生产力的人工智能时代。没有人工智能的物联网没大戏，而物流网又让人工智能更准确。

4. 云计算——人工智能背后强大的助推器

云计算是将人们传统的 IT 工作转为以网络为依托的云平台运行，NIST（美国国家标准与技术研究院）在 2011 年下半年公布了云计算定义的最终稿，给出了云计算模式所具备的 5 个基本特征（按需自助服务、广泛的网络访问、资源共享、快速的可伸缩性和可度量的服务）、3 种服务模式[SaaS（软件即服务）、PaaS（平台即服务）和 IaaS（基础设施即服务）]和 4 种部署方式（私有云、社区云、公有云和混合云）。云计算发展较早，经过 10 年的发展，国内已经拥有超百亿规模，云计算也不再只是充当存储与计算的工具而已。未来可以预见的是，云计算将在助力人工智能发展层面意义深远。而同时，人工智能的迅猛发展、海量数据的积累，也将会为云计算带来未知性和可能性。

物联网（产生数据）、云计算（承载数据）、大数据（挖掘数据）和人工智能（学习数据）相辅

相成，彼此依附，相互助力，合力搭档在一起才更有力量：给未来多一些可能，给未知多一些可能性。

1.2 Hadoop 简介

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。换言之，如果把大数据比作一种产业，那么这种产业实现盈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”。大数据需要特殊的技术，以有效地处理大量的容忍经过时间内的数据。适用于大数据的技术包括大规模并行处理（MPP）数据库、数据挖掘技术、分布式文件系统、分布式数据库、云计算平台、互联网和可扩展的存储系统。

现在，可供组织使用的数据种类繁多，令人难以置信。在内部，企业拥有网站点击流数据，电子邮件和即时消息存储库等；在外部，公共和私人实体的开放数据计划，使大量原始数据可用于分析。这里的挑战是传统工具无法很好地处理许多此类数据的规模和复杂性，而量身定制的 Hadoop 就可以处理各种“混乱”情况。行业内的企业用户都已注意到，Hadoop 在 IT 部门中已迅速成为成熟的平台。

即使读者不熟悉大数据，也可以通过学习快速了解 Hadoop 的使用场景和技术。

1.2.1 认识 Hadoop

1. 大数据对 Hadoop 有什么样的需求

Hadoop 绝不是传统的信息技术工具，它非常适合应对许多大数据挑战，尤其是海量数据和各种数据结构，但是也有不太适合 Hadoop 的就是即时的分析高速数据。尽管 Hadoop 是进行大数据分析的重要工具，但它也不是能解决所有的大数据问题，整个大数据域并不是 Hadoop 的同义词。

我们正处在信息时代的高级状态——数据时代，数据通过物联网的传感器不断地生成和捕获，其增加的速度和变化是难以置信的。移动电话、照相机、摄像头、汽车、电视以及工业和医疗保健中的机器等设备都对我们今天看到的爆炸式增长的数据量做出了贡献。可以浏览、存储和共享该数据，但其最大价值仍未得到开发，该价值在于其提供洞察力的潜力，可以解决棘手的商业问题、应用问题，开拓新市场，降低成本并改善人类社会的整体健康状况。

庞大的数据量在许多情况下对传统的数据挖掘技术提出了几乎无法克服的挑战，即使条件良好，它也只能处理一部分可用的宝贵数据。Google 于是努力寻找一种新方法来分析其搜索引擎收集的大量数据，Hadoop 正是这种努力的结果，它代表了一种有效且经济高效的方式，可将大型分析挑战减少为可管理的小型任务。

2. Hadoop 的起源和有趣的名字

Doug Cutting 在 2003—2004 年发表了两个学术论文来描述 Google 的技术：Google File System（GFS）和 MapReduce。Doug Cutting 开发的 Hadoop 是一个开源平台，提供 MapReduce 和 GFS 技术的实现。Yahoo 公司在 2006 年雇用了 Doug Cutting，并很快成为 Hadoop 项目的坚定支持者。

回顾历史可以发现 Hadoop 最初旨在用作 2002 年开始的 Apache Nutch 项目的基础结构。Nutch 是一个开源 Web 搜索引擎，是 Lucene 项目的一部分。Apache 项目的创建是为了开发开源软件，并得到 Apache 软件基金会（ASF）的支持，该基金会是一个由分散的开发人员社区组成的非营利性公司。开源软件通常以公共和协作的方式开发，是一种源代码可供任何人免费研究、修改和分发的软件。Nutch 需要一种可以扩展到数十亿个网页的架构，并且所需的架构受到 Google 文件系统（GFS）的启发，并最终成为 HDFS。

2004 年，Google 发表了一篇介绍 MapReduce 的论文，到 2005 年，Nutch 同时使用 MapReduce 和 HDFS。2006 年初，MapReduce 和 HDFS 成为 Lucene 子项目的一部分，该子项目名为 Hadoop。2008 年 2 月，Hadoop 集群生成雅虎的搜索索引。2008 年初，Hadoop 已成为 Apache 的顶级项目，并被许多公司使用。2008 年 4 月，Hadoop 用时 209 秒，利用 910 个节点集群对 TB 级数据进行排序打破了世界纪录。2009 年 5 月，Yahoo 能够使用 Hadoop 在 62 秒内排序 1 TB 数据。

至于 Hadoop 这个名字，它只是 Doug Cutting 的儿子给他的毛绒大象取的名字。这个名称是唯一的且易于记忆的——其特性使其成为一个不错的选择。

3. Hadoop 究竟是什么

Hadoop 的核心是一个框架，用于将数据存储在大型商用硬件集群上——负担得起且易于使用的日常计算机硬件，并针对该数据运行应用程序。使用负担得起的计算资源网络来获得业务解决能力是 Hadoop 的关键价值主张，通俗地讲，就是把一堆 PC 通过网络连接起来能完成大型数据处理。

Hadoop 由两个主要组件组成：MapReduce 分布式处理框架和 Hadoop 分布式文件系统（HDFS）。在 Hadoop 上运行的应用程序将其工作分配给集群中的 node 节点（集群是一组互连的计算机，它们可以共同解决同一问题），HDFS 中存储将要处理的数据。Hadoop 集群可以跨越数千台计算机，HDFS 将数据存储在其中，并且 MapReduce 作业在数据附近节点进行处理，从而使 I/O 成本保持较低。同时，MapReduce 也极其灵活，可以开发各种应用程序。

Hadoop 集群是一种计算集群，也就是主要用于计算目的的集群，许多计算机（Node）可以共享计算工作负载，并利用集群中非常大的聚合带宽。Hadoop 集群通常由主节点和许多从属节点组成，主节点主要用于控制 Hadoop 中的存储和处理系统，而从节点存储集群中的所有数据并在其中处理数据。

1.2.2 Hadoop 核心组件

Hadoop 的三大核心组件是 Hadoop 的数据存储工具 HDFS(Hadoop distribute file system,分布式文件管理系统)、分布式计算框架 Hadoop MapReduce 和 Hadoop 的资源管理器 YARN(yet another resource negotiator,另一种资源协调者)。

Hadoop 使用 HDFS 进行数据存储。HDFS 具有主/从体系结构,主服务(NameNode)控制对数据文件的访问。从站服务(DataNodes)在集群中的每个节点上进行分布,DataNodes 管理与节点相关联的存储,为客户端读取和写入请求以及其他任务提供服务。

Hadoop 使用 MapReduce 进行分布式处理。MapReduce 涉及对分布式数据集的一系列操作的处理。数据由键-值对组成,并且计算只有映射阶段和归约阶段。用户定义的 MapReduce 作业在集中群的计算节点上运行。一般来说,MapReduce 作业的运行方式如下:

(1)在 Map 阶段,输入数据被分为大量的片段,每个片段都分配给一个 Map 任务。

(2)这些映射任务分布在整个集群中。

(3)每个映射任务都会从其分配的片段中处理键-值对,并生成一组中间键-值对。

(4)中间数据集按键排序,并将排序后的数据划分为多个与 Reduce 任务数量匹配的片段。

(5)在 Reduce 阶段,每个 Reduce 任务都会处理分配给它的数据片段,并生成一个输出键-值对。

(6)这些 Reduce 任务也分布在整个集群中,并在完成后将其输出写入 HDFS。

Hadoop 早期版本中的 Hadoop MapReduce 框架具有一个称为 JobTracker 的单一主服务和多个称为 TaskTrackers 的从属服务,集群中每个节点一个。将 MapReduce 作业提交给 JobTracker 时,该作业将放入队列中,然后根据管理员定义的调度规则运行。JobTracker 管理着 MapReduce 任务到 TaskTrackers 的分配。但 Hadoop 后期版本中,一个新的资源管理系统 YARN 提供通用的计划和资源管理服务,因此不仅可以在 Hadoop 集群上运行 MapReduce 应用程序,还可以进行资源的调度管理。

Hadoop 不仅限于 MapReduce 和 HDFS,它还是一系列相关项目(实际上是一个生态系统),用于分布式计算和大规模数据处理。这些项目中的大多数由 Apache Software Foundation 托管,故形成了 Apache Hadoop 生态系统。Apache Hadoop 生态系统中其他部分开源组件及其具体描述见表 1-1。

表 1-1 Hadoop 生态系统中其他部分开源组件及其具体描述

组　　件	描　　述
Ambari	一套集成的 Hadoop 管理工具,用于安装、监视和维护 Hadoop 集群,还包括用于添加或删除从属节点的工具
Flume	一种数据流服务,用于将大量日志数据移动到 Hadoop 中
Hive	用于存储在 HDFS 中的数据的分布式数据仓库,还提供基于 SQL(HiveQL)的查询语言

（续表）

组　　件	描　　述
Pig	一个用于在 HDFS 上运行的超大型数据集的分析平台，其基础结构层由生成 MapReduce 程序序列的编译器组成，语言层包括名为 Pig Latin 的查询语言
HBase	使用 HDFS 作为基础存储的分布式列式数据库。使用 HBase，可以将数据存储在具有可变列结构的超大型表中
ZooKeeper	一个简单的界面，用于分布式应用程序使用的服务（如命名、配置和同步）的集中协调
Sqoop	Sqoop 用于将关系型数据库（如 MySQL）或者其他结构化的数据导入 Hadoop 的生态系统（HDFS、Hive、HBase）中，反过来也可以将 Hadoop 的数据导出为对应的结构形式
Spark	Spark 是个开源的数据分析集群计算框架，其采用基于内存的分布式数据集，优化了迭代式的工作负载及交互式查询。同时支持分布式数据集上的迭代式任务，可在 Hadoop 文件系统上与 Hadoop 一起运行

Hadoop 生态系统及其商业发行版还在继续发展中，新技术或新工具将不断出现，目前各种 Hadoop 生态系统项目及它们之间的关系如图 1-2 所示。

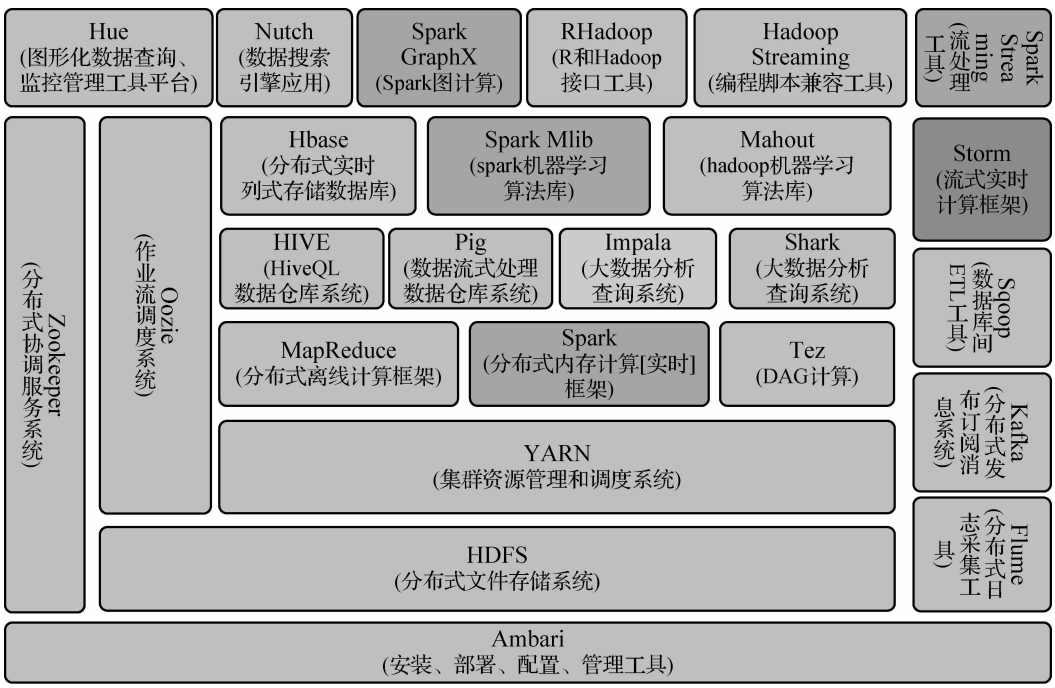

图 1-2　各种 Hadoop 生态系统项目及它们之间的关系

可以从 Apache 软件基金会或提供自己的 Hadoop 发行版的公司中获得 Hadoop。仅可直接从 Apache Software Foundation 获得的产品可以称为 Hadoop 版本。其他公司的产品可以包括官方的 Apache Hadoop 发行文件，但是 Apache 软件基金会不支持从 Apache Hadoop 源树中“派生”（并代表其修改或扩展版本）的产品。

1.3 实训 1 收集 Hadoop 相关案例

1. 实训目的

- 了解大数据及“物、云、大、智”的关系；
- 了解 Hadoop 的应用案例；
- 了解及认识 Hadoop 与组件；
- 了解 MapReduce 作业的运行方式。

2. 实训内容

- 搜索 Hadoop 使用案例，制作案例 Hadoop 体系结构图；
- 查找国内尤其是互联网公司使用 Hadoop 的实际案例；
- 参照 Hadoop 体系中各组件功效，结合案例查看组件选用状态；
- 绘制案例体系架构图，示例如图 1-3 所示。

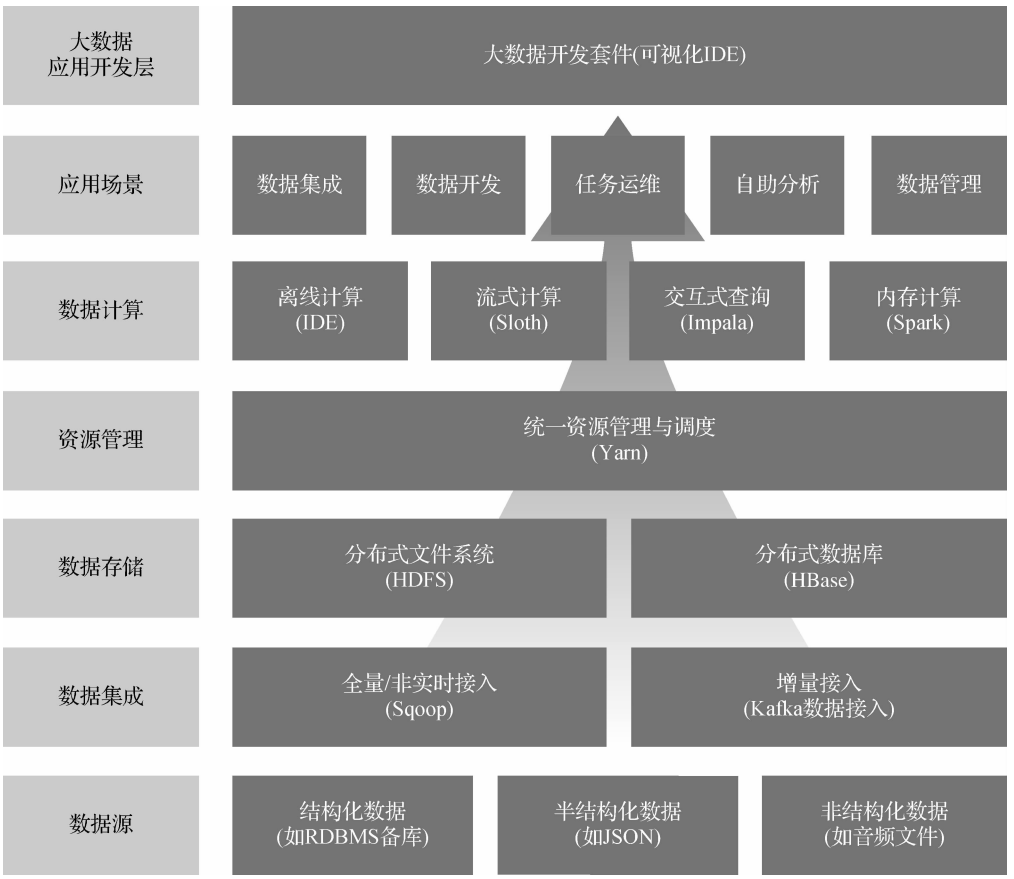

图 1-3 网易猛犸大数据架构平台

3. 实训要求

- 按题目要求查找案例并绘制案例体系架构图(“文字＋截图”方式)；
- 总结实训心得与体会。