**面向大数据分析的可扩展系统：技术教学**

【摘要】最近技术的进步已经导致出现大量来自不同领域的数据（例如医疗保健、科学传感器、用户生成数据，互联网和金融公司以及供应链系统），超过了过去二十年。“大数据”这个术语在这个新兴趋势之下被创造了，除了数量之外，大数据跟传统数据相比还表现出其他独特的特点。例如，大数据通常是非结构化的、需要更多的实时分析。这门技术的发展需要具备新的数据采集，传输，存储和大规模数据处理机制的系统架构。在本文中，我们为大数据分析平台提供了一份文献调查和系统教程，目的是为非专业读者提供一幅具体全面的架构图，并为高级读者灌输一种落实动手定制他们自己大数据解决方案的精神。首先，我们介绍大数据的定义并讨论大数据的挑战。 接下来，我们提出了一个系统框架，将大数据系统分解为四个顺序模块，即数据生成，数据采集，数据存储和数据分析。这四个模块组成了大数据的价值链。接下来，我们对研究和工业界的众多方法机制进行了详细的调查。另外，我们还介绍了用于解决大数据挑战的流行Hadoop框架。最后，我们概述了大数据系统的几个评估基准和潜在的研究方向。

【索引】大数据分析，云计算，数据采集，数据存储，数据分析，Hadoop。

1. **介绍**

由于其更广泛的影响，新兴的大数据模式已经深刻地改变了我们的社会，并将继续吸引技术专家和普通公众的多方关注。很明显，我们正处于数据洪流时代，这体现在来自各种来源的庞大数据量以及不断增长的数据生成速度。例如，IDC报告[1]预测，从2005年到2020年，全球数据量将增长300倍，从130艾字节到40000艾字节，每两年增长一倍。“大数据”这个术语的出现是为了捕捉这一数据爆炸趋势的深刻含义，而且事实上大数据已被吹捧为可望改变我们社会的新石油。例如，Mckinsey报告[2]指出，全球个人位置数据的潜在价值估计在未来十年将为服务提供商带来1000亿美元的收入，并将为消费者和商业终端用户带来高达7000亿美元的价值。与大数据相关的巨大潜力导致了一个新兴的研究领域，这个领域迅速引起了来自不同行业的巨大兴趣，例如工业，政府和研究机构。广泛的兴趣首先体现在工业报道[2]和公共媒体（如“经济学人”[3]，[4]，纽约时报[5]和国家公共广播电台（NPR） [6]，[7]）范围。政府在制定新计划[8]方面也发挥了重要作用，以加速应对大数据挑战的进展。最后，“自然”和“科学杂志”发表了一些专题，以讨论大数据现象及其挑战，并将其影响扩大到超越技术领域。因此，人们对来自不同领域的大数据越来越感兴趣，需要对其定义，演化历史，构建技术和潜在挑战有一个清晰直观的理解。

本教程主要讨论可扩展的大数据系统，其中包括一组工具和机制，用于加载，提取和改进不同数据，同时利用大规模并行处理能力执行复杂的转换和分析。 由于大数据的独特性，设计可扩展的大数据系统面临一系列技术难题，其中包括：

首先，由于各种不同的数据源和庞大的数量，很难从分布式节点可扩展性地收集和整合数据。 例如，包含文字，图片，视频，社交关系的超过1亿5千5百万条推文是由全球数百万的帐户生成的[9]。

其次，大数据系统需要存储和管理收集到的大量异构数据集，同时在快速检索，可扩展性和隐私保护方面提供功能和性能保证。 例如，Facebook需要存储，访问和分析超过30皮秒的用户生成数据[9]。

第三，大数据分析必须有效地实时或近实时地在不同层次上挖掘海量数据集 - 包括建模，可视化，预测和优化 - 从而可以揭露固有承诺改善决策，并获得进一步的优势。

这些技术挑战要求对当前的数据管理系统进行重新审视，从架构原则到实施细节。事实上，许多领先的行业公司[10]都放弃了过渡性解决方案去采用新兴大数据平台。

但是，主要基于关系数据库管理系统（RDBMS）的传统数据管理和分析系统在解决上述大数据挑战列表方面存在不足。具体而言，传统RDBMS与新兴大数据模式之间的不匹配分为以下两个方面，其中包括：

从数据结构的角度来看，RDBMS只能支持结构化数据，但对半结构化或非结构化数据提供的支持很少。

从可扩展性的角度来看，RDBMS随着昂贵的硬件扩大规模，不能与商品硬件并行扩展，这不适合应对日益增长的数据量。

为了应对这些挑战，研究机构和行业已经以特别的方式为大数据系统提出了各种解决方案。 云计算可以作为大数据系统的基础设施层进行部署，以满足某些基础设施要求，如成本效益，弹性和扩展或缩减的能力。分布式文件系统[11]和NoSQL[12]数据库适用于持久存储和大规模免计划数据集的管理。MapReduce[13]是一个编程框架，在处理群组聚合任务（如网站排名）方面取得了巨大成功。Hadoop[14]集成了数据存储、数据处理、系统管理和其他模块，形成了强大的系统级解决方案，正成为处理大数据挑战的主流。基于这些创新的技术和平台，我们可以构建各种大数据应用程序。鉴于大数据技术的扩散，一个系统的框架应该是为了捕捉大数据研究和开发工作的快速发展，并基于不同角度发展。



图1。南洋理工大学建立了模块化数据中心，用于系统/试验台的研究。试验台主机270台服务器，分成10个机架。

在本文中，将从我们在私有模块化数据中心测试平台上构建大数据解决方案的第一手经验中学习（如图1所示），我们努力为可扩展大数据系统提供系统教程，重点关注使能技术和建筑原则。我们希望这篇论文能够成为领域专家，大数据用户和普通读者的第一站，以便根据他们对大数据解决方案的特定需求寻找信息和指南。例如，领域专家可以遵循我们的指导原则开发自己的大数据平台，并在大数据领域进行研究;大数据用户可以使用我们的框架来评估供应商提出的替代解决方案;普通观众可以了解大数据的基础及其对工作和生活的影响。为了达到这个目的，我们首先列出大数据的替代定义，并辅以大数据和大数据模式的历史。 随后，我们引入了一个通用框架，将大数据平台分解为四个组件，即数据生成，数据采集，数据存储和数据分析。对于每个阶段，我们都会调查当前的研究和开发工作，并为架构设计提供工程见解。迈向特定的解决方案之后，我们将深入研究Hadoop--大数据分析平台的事实选择，并为大数据平台提供基准测试结果。本文的其余部分安排如下。 在第二部分中，除了处理范例之外，我们还介绍大数据的定义及其简要历史。 然后，在第三节中，我们将介绍大数据价值链（由四个阶段组成），大数据技术图，分层系统架构和挑战。接下来的四节将描述与大数据价值链相关的不同大数据阶段。具体而言，第四部分着重于大数据生成，并介绍了具有代表性的大数据源。第五节讨论大数据采集并提出数据采集，数据传输和数据预处理技术。第六部分研究大数据存储方法和编程模型。第七节讨论大数据分析，第八节讨论了一些应用。第九部分介绍了Hadoop，它是大数据运动的当前支柱。第十节概述了评估大数据系统性能的几个基准。第十一节提出了一个关于未来研究建议的简短结论。

1. **大数据：定义，历史和范例**

在本节中，我们首先列出大众数据的流行定义，然后是其发展的简要历史。 本节还讨论了两种可选的范例，流处理和批处理。

1. **大数据定义**

鉴于目前的流行度，大数据的定义相当多样化，达成共识非常困难。 从根本上说，大数据不仅意味着大量的数据，而且意味着其他功能将其与“海量数据”和“非常大的数据”的概念区分开来。事实上，有关大数据的一些定义可以在文献中找到，而三种定义在塑造大数据的重要性方面起着重要作用：

归属定义：IDC是研究大数据及其影响的先驱。 它在2011年由EMC（云计算领导者）赞助的报告中定义了大数据[15]：“大数据技术描述了新一代技术和架构，旨在从大量数据中经济地提取价值，通过实现高速捕获、发现、和/或分析。“这种定义描述了大数据的四个显着特征，即体积，品种，速度和价值。结果，“4Vs”的定义被广泛用于描述大数据。在2001年的研究报告[2]中也有类似的描述，META集团（现为Gartner）分析师Doug Laney指出，数据增长的挑战和机遇是三维的，即增加的数量，速度和多样性。尽管这个描述本来并不是为了定义大数据，Gartner和包括IBM [16]和某些微软研究人员[17]在内的许多行业在后面的10年里仍然继续使用这种“3V”模型来描述大数据。

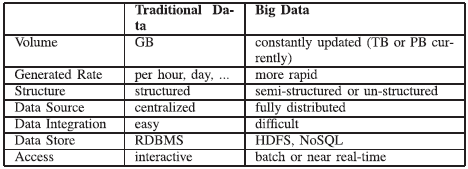
比较定义：2011年，Mckinsey的报告[2]将大数据定义为“数据集，其大小超出了典型数据库软件工具捕获，存储，管理和分析的能力”。这个定义是主观的，并不根据任何特定的度量来定义大数据。然而，它将数据集必须被视为大数据的定义纳入进化方面。

架构定义：美国国家标准与技术研究院（NIST）[19]认为，“大数据是指数据量，采集速度或数据表示限制了使用传统关系方法进行有效分析的能力，或者需要为有效处理使用显著的水平缩放。”特别是，大数据可以进一步分类为大数据科学和大数据框架。大数据科学是“对大数据的获取，调节和评估的技术的研究”，而大数据框架是“在计算节点上能够跨集群进行分布式处理和分析大数据问题的软件库及其相关算法”。 一个或多个大数据框架的实例化被称为大数据基础架构。

同时，各个行业和学术界对于大数据的实际含义有很多讨论[20]，[21]。

然而，就大数据的定义达成共识是困难的，即使不是不可能的。一个合乎逻辑的选择可能是包含所有备选定义，每个定义都关注大数据的特定方面。在本文中，我们采用这种方法并开始了解大数据科学和工程中的常见问题和方法。

表1.大数据和传统数据之间的比较。



上述关于大数据的定义提供了一组工具来比较新兴的大数据与传统的数据分析。 表1总结了这种比较，在“4Vs”的框架下。 首先，大量数据集是区分大数据和传统数据的关键因素。例如，Facebook报告称其用户在2012年2月注册了27亿“点赞”和每天的评论[22]。其次，大数据有三种结构：结构化，半结构化和非结构化。传统数据通常是结构化的，因此可以很容易地标记和存储。但是，来自Facebook，Twitter，YouTube和其他用户生成内容等资源的绝大多数数据都是非结构化的。第三，大数据的速度意味着数据集必须以与数据生产速度相匹配的速度进行分析。对于时间敏感的应用（如欺诈检测和RFID数据管理），大数据以流的形式注入企业，这要求系统尽可能快地处理数据流以最大化其价值。最后，通过利用各种挖掘方法来分析大数据集，重要价值可以以深度洞察力或商业利益的形式，从具有低密度值的大量数据中推导出来。

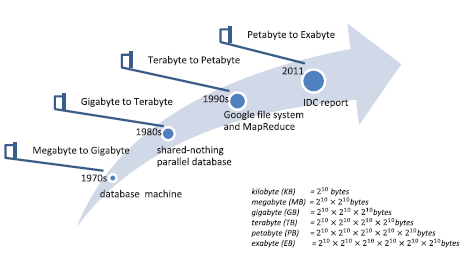


图2.主要里程碑的大数据简史。 根据数据大小的顺序增长

**C．大数据范例：流vs批**

大数据分析是使用在强大支持平台上运行的分析算法来发现隐藏在大数据中的潜力的过程，例如隐藏模式或未知相关性。根据处理时间的要求，大数据分析可以分为两种备选模式：

流处理：流处理范例[34]的起点是假设数据的潜在价值取决于数据的新鲜度。因此，流式处理范例尽快分析数据以得出其结果。 在这个范例中，数据到达一个流中。在持续到来的时候，由于流速快，体积庞大，只有一小部分流存储在有限的内存中。通过流的一个或几个通道可以找到近似结果。数据流处理理论和技术已经研究了数十年。代表性的开源系统包括Storm [35]，S4 [36]和Kafka [37]。流式处理范例用于在线应用程序，通常在秒级或甚至毫秒级。

批处理：在批处理范例中，数据首先被存储然后被分析。MapReduce[13]已经成为主要的批处理模型。MapReduce的核心思想是数据首先被分成小块。接下来，这些块被并行处理并以分布式方式处理以产生中间结果。最终结果是通过汇总所有中间结果得出的。该模型将计算资源调度到数据位置附近，避免了数据传输的通信开销。MapReduce模型简单，广泛应用于生物信息学，web挖掘和机器学习。

表2.流处理和批处理之间的比较

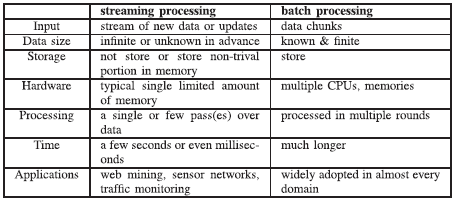


表2总结了这两种处理范例之间的差异。一般来说，流处理范例适用于以流的形式生成数据并需要快速处理以获得近似结果的应用程序。 因此，流处理范例的应用领域相对较窄。最近，大多数应用程序都采用了批处理范例;即使某些实时处理应用程序使用批处理范例来实现更快的响应。此外，已经做出了一些研究努力来整合这两种范式的优点。

大数据平台可以使用其他处理范例;然而，这两种范式的差异将导致相关平台的架构差异。例如，基于批处理的平台通常涵盖复杂的数据存储和管理系统，而基于流处理的平台则不然。实际上，我们可以根据数据特征和应用需求来定制平台。由于批处理范式被广泛采用，本文只考虑基于批处理的大数据平台。

1. **大数据系统架构**

在本节中，我们将重点放在大数据分析的价值链上。具体而言，我们描述了一个由四个阶段（生成，采集，存储和处理）组成的大数据价值链。接下来，我们展示一个大数据技术图，将该领域的领先技术与大数据价值链中的特定阶段和时间戳相关联。

1. **大数据系统：价值链视图**

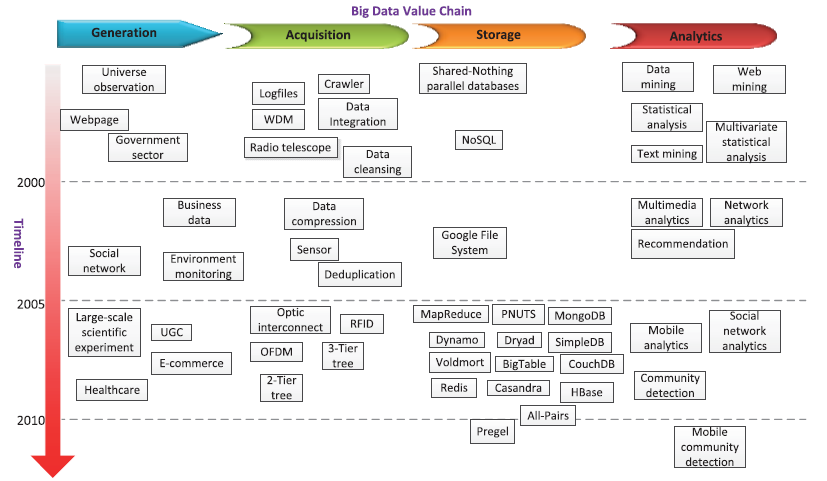


图3.大数据技术图。它在两个轴上转动，即数据值链和时间轴。数据价值链将数据生命周期分为四个阶段，包括数据生成，数据采集，数据存储和数据分析。在每个阶段，我们都会重点介绍过去10年的示范技术。

大数据系统非常复杂，提供了处理数字数据生命周期从诞生到销毁不同阶段的功能。同时，该系统通常涉及不同应用的多个不同阶段[38,39]。在这种情况下，我们采用系统工程方法，在业界广为接受[40,41]，将典型的大数据系统分解为四个连续阶段，包括数据生成，数据采集，数据存储和数据分析，如图3的横轴所示。注意数据可视化是数据分析的辅助方法。一般来说，应该将首先进行数据可视化以发现一些粗略模式，然后再采用特定的数据挖掘方法，我在数据分析部分提到了这一点。每个阶段的细节解释如下。

数据生成涉及如何生成数据。在这种情况下，术语“大数据”被指定为意味着从各种纵向和/或分布式数据源（包括传感器，视频，点击流和其他可用数字源）生成的大型，多样和复杂的数据集。通常，这些数据集与不同级别的域特定值相关联[2]。在本文中，我们将重点放在三个主要领域的数据集上，这些领域包括商业，互联网和科学研究，其价值相对容易理解。然而，在收集，处理和分析这些数据集方面存在巨大的技术挑战，这些数据集需要新的解决方案来迎接信息和通信技术（ICT）领域的最新进展。

数据采集​​是指获取信息的过程，细分为数据采集，数据传输和数据预处理。首先，由于数据可能来自多种来源，承载格式化文本，图像和/或视频的网站 - 数据收集是指可从特定数据生产环境中获取原始数据的专用数据收集技术。其次，在收集原始数据后，我们需要高速传输机制将数据传输到适用于各种类型分析应用的存储维持系统。最后，收集的数据集可能包含许多无意义的数据，这会不必要地增加存储空间量并影响后续的数据分析。例如，从用于监测环境的传感器收集的大多数数据集中，冗余是很常见的，我们可以使用数据压缩技术来解决这个问题。因此，我们必须为高效的存储和挖掘执行数据预处理操作。

数据存储涉及持续存储和管理大规模数据集。数据存储系统可以分为两部分：硬件基础设施和数据管理。硬件基础设施由一系列共享ICT资源组成，以弹性方式组织各种任务以响应其即时需求。硬件基础架构应该能够扩展，并且能够动态地重新配置以解决不同类型的应用程序环境。数据管理软件部署在硬件基础架构之上以维护大规模数据集。另外，为了分析存储的数据或与存储的数据交互，存储系统必须提供多种接口功能，快速查询和其他编程模型。

数据分析利用分析方法或工具来检查，转换和建模数据以提取价值。许多应用领域利用丰富的数据和领域特定分析方法提供的机会来推导预期的影响。尽管各种领域提出了不同的应用要求和数据特征，但其中一些领域可能会利用类似的基础技术。新兴分析研究可分为六个关键技术领域：结构化数据分析，文本分析，多媒体分析，web分析，网络分析和移动分析。此分类旨在强调每个区域的关键数据特征。

**C. 大数据系统：分层视图**

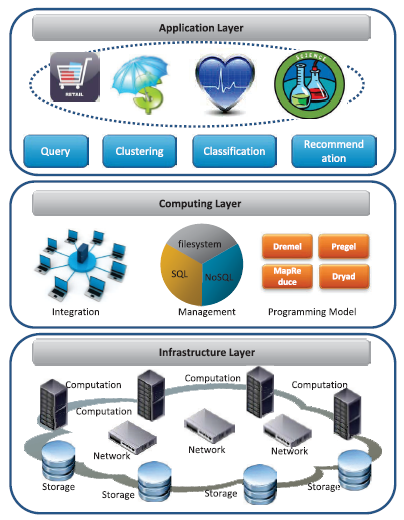


图4.大数据系统的分层架构，它可以从底层向上分解为三层，包括基础设施层，计算层和应用层。

或者，大数据系统可以分解为分层结构，如图4所示。分层结构可以从底层到顶层分为三层，即基础设施层，计算层和应用层。这种分层的视图仅仅提供了一个概念层次来强调大数据系统的复杂性。各层的功能如下。

基础设施层由一系列ICT资源组成，可以通过云计算基础架构进行组织，并通过虚拟化技术实现。这些资源将通过特定的服务级别协议（SLA）以细粒度的方式暴露给上层系统。 在此模型中，必须分配资源以满足大数据需求，同时通过最大化系统利用率，能源意识，操作简化等来实现资源效率。

计算层将各种数据工具封装在运行于原始ICT资源的中间件层中。 在大数据的背景下，典型的工具包括数据集成，数据管理和编程模型。数据集成意味着从不同的数据源获取数据，并将数据集合到统一的表格中，并进行必要的数据预处理操作。数据管理是指提供持久数据存储和高效管理的机制和工具，如分布式文件系统和SQL或NoSQL数据存储。该编程模型实现了抽象的应用程序逻辑并有助于数据分析应用程序。MapReduce[13]，Dryad[42]，Pregel [43]和Dremel[44]说明了这种编程模型。

应用层利用编程模型提供的接口来实现各种数据分析功能，包括查询，统计分析，聚类和分类; 然后，它结合了基本的分析方法来开发各种领先的相关应用程序。McKinsey提供了五大潜在的大数据应用领域：医疗保健，公共部门管理，零售，全球制造和个人位置数据。