从系统角度审视大数据计算

郑纬民
清华大学计算机科学与技术系 北京 100084

摘要
大数据计算是实现大数据“巨大价值”的必要手段，而计算系统是大数据计算的有效载体。试着从系统角度审视大数据计算，透过大数据的体量巨大、速度极快、模态多样、真伪难辨等宏观特征，针对批量计算、流式计算、大图计算等计算形式，分别探讨大数据计算的典型特征，论述了这些特征给大数据计算系统的设计与实现带来的技术挑战，进而梳理了为了应对这些挑战所取得的研究成果，最后从系统角度指出未来大数据计算可能的一些研究方向。

关键词
大数据计算；批量计算；流式计算；大图计算；系统实例

Reviewing Big Data Computation from a System Perspective

Zheng Weimin
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract
Big data computing is a necessary way to acquire the “great value” behind the big data, and a computing system is an effective tool for big data computing. Big data computing from a system perspective was reviewed. Based on the fact that big data has the macro characteristics of huge volume, growing fast, complex structure, and quality disparity, the typical features of big data computing by analyzing batch computing, stream computing, and graph computing respectively, were discussed. These features may bring technical challenges to the design and implementation of big data computing system. The related works for overcoming these challenges were further categoried. In the end, some prospective research directions of big data computing from the system perspective were listed.

Key words
big data computing, batch computing, stream computing, graph computing, system instance

1 引言

大数据已成为当前社会各界关注的焦点[1~4]。从一般意义上讲，大数据是指在可容忍的时间内，无法用现有信息技术和软硬件工具对其进行感知、获取、管理、处理和服务的数据集合。大数据呈现出多种鲜明特征[3~8]，在数据量方面，体量巨大，当前全球所拥有的数据总量已经远远超过历史上的任何时期，更为重要的是，数据量的增加速度呈现出倍增趋势；在数据速率方面，速度极快，数据产生、传播的速度更快，在不同时空中流转，呈现出鲜明的流式特征，更为重要的是，数据价值的有效时间急剧减少，也要求越来越高的数据计算和使用能力；在数据复杂性方面，模态多样，种类繁多，在编码方式、存储格式、应用特征等多个方面也存在多层次、多方面的差异性，结构化、半结构化、非结构化数据并存；在数据价值方面，价值稀疏，真伪难辨，但价值总量巨大，随着数据规模的不断增大，隐含于大数据中的知识也随之增多，但这些知识隐含程度很深，对发现这些知识的方式、方法提出了更高的要求。此外，大数据还呈现出个性化、不完备化、交叉复用等诸多鲜明特征。

大数据蕴含大信息，大信息提炼大知识，大知识将在更高的层面、更广的视角、更大的范围帮助用户提高洞察力、提升决策力，将为人类社会创造前所未有的大价值。但与此同时，这些总量极大的价值往往隐藏在大数据中，表现出价值密度极低、分布极其不规律、信息隐藏程度极深、真伪信息交织混合发现有用价值极其困难的鲜明特性，这些特征必然为大数据的计算带来前所未有的挑战和机遇。

大数据计算是发现信息、挖掘知识、满足应用的必要途径，也是大数据从收集、传输、存储、计算到应用等整个生命周期中最关键、最核心的环节，只有有效的大数据计算，才能满足大数据的上层应用需要，才能挖掘出大数据的内在价值，才能使大数据具有意义。大数据计算系统是实现大数据科学计算的基础平台。对于规模巨大、价值稀疏、结构复杂、时效性强的大数据，其计算亦面临不同于传统数据计算的诸多新挑战，如计算复杂度高、任务周期长、数据实时性强、计算通用性差等。大数据及其计算的这些挑战对大数据计算系统的系统架构、计算框架、处理方法等提出了新的挑战。同时，大数据时代出现了很多新的应用需求，如面向社交媒体的大图关系分析与发现，需要结合具体的应用场景，开展针对性的关于计算模式的研究。

为了满足和适应大数据计算的需要，随着大数据及相关技术的全面和深入发展，大数据计算模式也呈现出多样化、专业化特征，以满足不同领域大数据应用范式的要求。本文首先针对大数据计算的3种代表性模式进行了深入的分析，主要包括大数据批量计算、流式计算和交互计算，对其中各计算模式的基本概念、典型特征和技术挑战进行了系统的归纳和分类。其次，分别针对这3种计算模式中当前具有广泛代表性的系统进行了具体实例分析。再次，从系统的角度，对3种计算模式的未来研究方向和重点进行了初步分析。最后，对全文进行了总结。

2 大数据计算模式

大数据计算模式主要包括批量计算、流式计算、交互计算3种。其中，交互计算需要在计算过程中与用户进行互动，才能进行后续的计算动作，可以把交互计算看作批

量计算的一种特殊形式。本文不再对交互计算进行深入分析。大图计算本属于批量计算范畴，但随着互联网应用的发展，其重要性日益凸显，并且因其各个节点的关联紧密性而具有不同于其他普通批量计算的显著特征。本文对大图计算进行单独讨论。

2.1 批量计算的特征及挑战

大数据批量计算[9~13]（big data batch computing）是大数据计算的一种主要计算模式，当前阶段，大多数应用场景均通过批量计算模式实现。同时，批量计算也可以同其他计算模式进一步结合起来，以完成对数据的进一步处理。在大数据批量计算环境中，其计算架构如图1所示。数据通过多个数据源进行收集，按照与应用场景所需要的方式进行组织，在各种外存存储介质（如硬盘、磁带等）上静态地存储起来。当需要进行数据计算时，开启数据的计算过程，进行数据的集中处理，数据被处理完后，计算过程也随之结束。在数据的计算过程中，数据的计算顺序、计算速度等各种因素可以有效控制，也可以有选择地、重复地进行部分数据的重计算。数据的计算结果是确定、准确、全面、可重现的，但数据的计算时延往往较长，往往在数分钟到数小时之间。可见，对于先存储后计算的实时性要求不高，同时，对于数据的准确性、全面性更为重要的应用场景，批量计算模式更加适合。

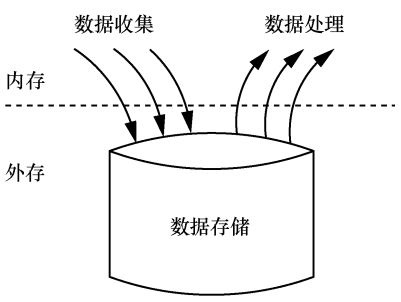

图1 大数据批量计算架构

大数据批量计算场景通常呈现出以下典型特征及挑战。

（1）数据体量巨大

数据量从TB级别跃升到PB级别，甚至更高。数据往往以静态的形式在硬盘等外部存储介质上永久存储，一次写入，很少再进行更新，存储时间长，可以重复多次利用，但很难对其进行移动和备份。面向如此体量的数据，需要在数据的组织方式、计算形式等方面根据具体的应用场景，构建一个高效、分布式的大数据计算系统，以满足对相关数据的并行、分布式处理要求。

（2）数据精确度高

批量数据通常是从应用中沉淀下来的，对于了解上次应用的各种内在关系、潜在逻辑以及预测未来发展都很关键。需要对其中所有数据进行全量式的计算，数据处理结果的精度要求较高。为了满足如此高的数据精度，需要在数据处理效率和数据处理结果精度等方面进行权衡，在数据的单次处理和再现方面进行权衡。

（3）数据价值稀疏

在数据的收集过程中，往往需要尽可能全面、密集地进行数据收集，避免任何有价值数据的遗失。随着数据收集工具和方法的不断进步，数据收集面和收集频率的不断增广和增加，数据价值的稀疏程度也急剧增强。因此，需要通过合理的计算架构和高效的数据处理算法才能从大量的数据中抽取少数有用的价值。此外，批量数据处理往往比较耗时，而且不提供用户与系统的交互手段，当发现处理结果和预期结果有很大差别时，会浪费很多时间。因此，批量数据处理适合大型的相对比较成熟的应用场景。数据价值稀疏性特征使得在大数据计算系统中，需要构建一个高效、精准、面向特定应用和领域的数据处理模

式，在极其稀疏甚至稀疏程度不断增加的应用场景下，能快速发现并挖掘出其中所存在的数据价值。

2.2 流式计算的特征及挑战

大数据流式计算[14~18]（big data stream computing）是大数据计算的另一种重要计算模式，特别是在数据时效性、实时性需要不断增加的应用场景不断增多的情况下，其重要性日益凸显。在大数据流式计算环境中，其计算架构如图2所示。数据以数据流的形式，通过多个不同的数据源实时到达大数据流式计算平台，然后，利用数据流图所描述的处理过程被在线处理，并实时产生结果，满足相关上层应用系统的需要。整个数据的处理过程往往在毫秒级的时间范围内完成，原始数据、中间状态、处理结果等数据根据具体应用场景的需要，不全部保存，只是选择性地存储。描述用户特定应用的数据流图一旦提交到系统中，将会永远在线运行，实时对输入的数据流进行处理，除非整个处理平台意外中断或显示终止。由于整个数据流的处理时间极短，判读的依据也往往集中在当前时间点附近（时间窗口）的数据，加上数据流中各数据项的不断变化，留给大数据流式计算平台进行调整和应对的时间也很少，因此，流式数据处理的结果往往不够精确、不够全面，只能给出一个实时性很强的、相对准确的、基于当前局部数据判断的结果。可见，对于无需先存储、可以直接进行数据计算、实时性要求很严格但数据的精确度往往不太重要的应用场景，流式计算具有明显优势。

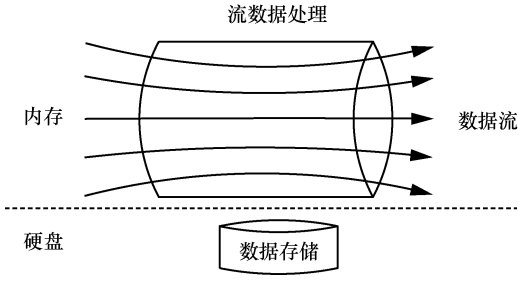

图2 大数据流式计算架构

大数据流式计算场景通常呈现出以下典型特征及挑战。

大数据流呈现出鲜明的实时性、易失性、突发性、无序性、无限性等特征。流式大数据是实时产生、实时计算，结果反馈往往也需要保证及时性。数据的使用往往是一次性的、易失的，即使重放，得到的数据流和之前的数据流也不同。数据的产生完全由数据源确定，由于不同的数据源，在不同时空范围内的状态不统一且动态变化，导致数据流的速率呈现出突发性的特征。各数据流之间、同一数据流内部各数据元素之间是无序的。数据是实时产生、动态增加的，只要数据源处于活动状态，数据就会一直产生和持续增加，可以说，潜在的数据量是无限的。

大数据流式环境中的数据计算在系统的可伸缩性、系统容错、状态一致性等方面均面临着前所未有的新的挑战。在系统的可伸缩性上，一方面，需要大数据流式系统具有很好的“可伸”特征，可以实时适应数据增长的需求，实现对系统资源的动态调整和快速部署；另一方面，当流式数据的产生速率持续减少时，需要及时回收在高峰时期所分配的目前已处于闲置或低效利用的资源，实现整个系统“可缩”的友好特征。在系统容错上，一方面，数据流实时、持续地到来，呈现出同时间相识的一维特征，一旦数据流流过，再次重放数据流的成本很大，甚至是不现实的；另一方面，在流式大数据的计算过程中，大部分“无用”的数据将被直接丢弃，所被永久保存下来的数据量是极少的，当需要进行系统容错时，其中不可避免地会出现一个

时间段内数据的不完整；再则，需要针对不同类型的应用，从系统层面上设计符合其应用特征的数据容错级别和容错策略。在各节点间状态的一致性上，一方面，如何从高速、海量的数据流中识别并维护一致性状态的数据是一个巨大的挑战；另一方面，在大规模分布式环境中，如何组织和管理实现系统状态一致性的相关数据，满足系统对数据的高效组织和精准管理的要求也是一个巨大的挑战。

2.3 大图计算的特征及挑战

大数据图计算[19~21]（big data graph computing）是大数据计算的一种计算模式，随着社交媒体、移动互联网的不断发展，在大数据计算中的重要性日益凸显。大数据图计算主要用来分析数据节点之间的关系和相似度，该计算范式已经广泛应用于用户分析、欺诈检测、社交媒体、移动互联网、生命科学等诸多领域，其巨大的商业价值已经凸显。

大数据图计算中的大图数据往往以图中的节点以及连接节点的边呈现，其中节点数目往往是数以万计的，边的数量更大，通常具有如下3个特征。

（1）节点之间的关联性

大图中各节点之间的关系是通过边来展现的。通常情况下，大图中边的数量是节点数量的指数倍。因此，节点和关系信息同等重要，图结构的差异也是由于对边做了限制，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、语义图以及特征图等。如何针对节点和边的不同作用和特征，进行节点和边的存储方式、组织模式以及计算途径等挑战的研究，结合具体应用，提供一种高效的存储方式、可扩展的组织模式以及有效的计算途径，满足具体应用场景的需要，是研究的关键点。

（2）图计算的数据耦合性强

在大图中，数据之间是相互关联的，对图数据的计算也是相互关联的。这种数据耦合的特性对图的规模日益增大达到上百万甚至上亿节点的大图数据计算提出了巨大的挑战。大图数据是无法使用单台机器进行处理的，但如果对大图数据进行并行处理，对于每一个顶点之间都是连通的图来讲，难以分割成若干完全独立的子图进行独立的并行处理。即使可以分割，也会面临并行机器的协同处理以及将最后的处理结果进行合并等一系列问题。这需要图数据处理系统选取合适的图分割以及图计算模型来迎接挑战并解决问题。

在大数据时代，大图的分割是大数据图计算最为突出的问题。由于对整个图的访问是随机进行的，在图划分时需要考虑3个方面：通信代价，访问跨机器的各边通信量；负载均衡，让每一台机器的问题规模基本接近；存储冗余，为了减少通信量，需要在机器上复制其他机器的存储信息（存在数据一致性问题）。通过考虑存储的冗余度，使综合开销达到最优。

此外，大数据图计算还存在以下问题：图数据的局部性差，由于节点众多，两个相连接的点（连接的点对也是随机的、无法预知的）可能存储的位置相隔很远，即不在同一个存储块，这使得系统需要随机访问这些节点及边，而访问磁盘的效率又极低，从而严重影响了计算效率；数据及图结构驱动，不同的图形结构会使用不同的计算方法，需要设计一个通用的方法；存储和效率，大图处理的规模（点的数量）基本上是10亿量级，依靠单台PC进行存储似乎不太可能，所以大多数图计算系统是分布式系统。由于这种系统是把存储容量和计算分摊给每一个机器，因此需要考虑如何划分才能使各机器负载均衡以及如何减少各个划分之间通信等问题。

3 典型计算系统

3.1 批量大数据计算系统

当前典型的大数据批量计算的应用系统有Hadoop[11]、Spark[13]。在Hadoop系统中，其体系结构如图3所示，由名字节点、数据节点、客户端节点组成。其中，名字节点负责管理文件系统的命名空间、集群配置以及数据块的备份、容错等内容；数据节点负责管理数据的存储位置、副本数目等内容，并以数据块的形式存储原始数据与校验信息；客户端节点通过与名字节点、数据节点进行通信，访问HDFS，实现文件操作。数据通过HDFS的方式进行组织，可以将各类数据存储在各种外部存储介质上，并通过MapReduce模式将计算逻辑分配到各数据节点进行数据计算和知识发现。

在Spark系统中，数据被转换成弹性分布式数据集（resilient distributed dataset，RDD），并以RDD为单位实现有效的数据处理。每个RDD都是一个不可变的分布式可重算的数据集，其记录着确定性的操作继承关系。如图4所示，每一个椭圆形表示一个RDD，椭圆形中的每个圆形代表一个RDD中的一个分区。通过对RDD的操作继承关系进行跟踪，当任意一个

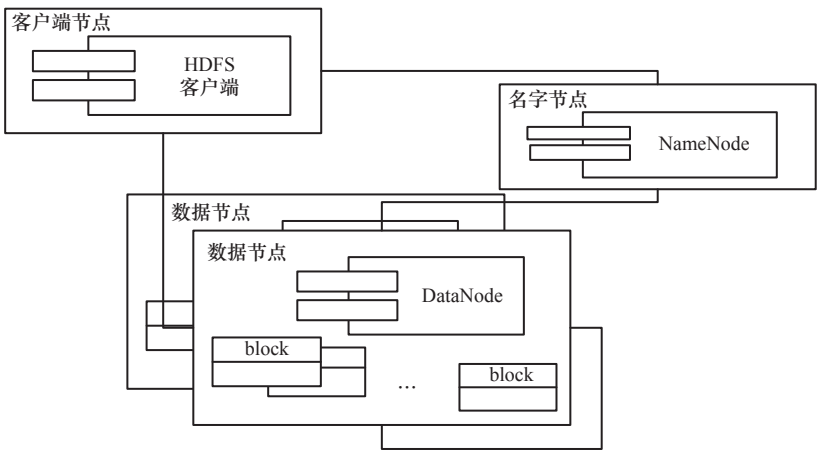

图3 Hadoop体系结构

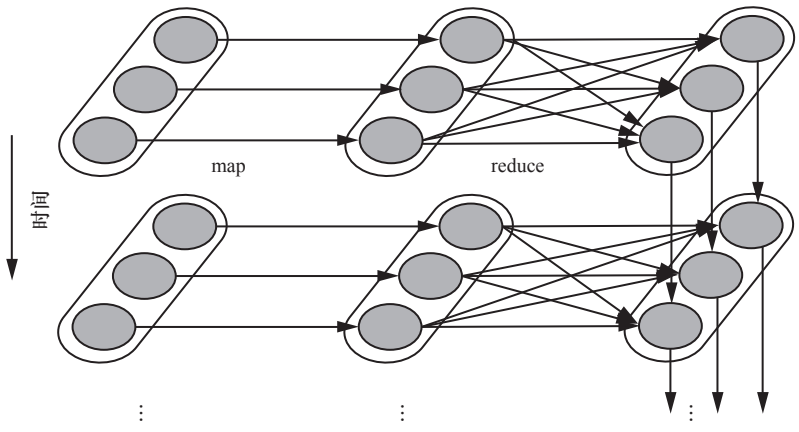

图4 RDD的操作继承关系

RDD的分区出错或不可用时，只要输入数据可重现，就可以利用原始输入数据通过转换操作而重新算出，实现系统的容错。

同时，Spark系统也可以在一定程度上支持大数据流式计算和交互计算的应用范式。

3.2 流式大数据计算系统

早期流式计算的研究往往集中在数据库环境中开展数据计算的流式化，数据规模较小，数据对象比较单一。大数据环境中的流式数据在实时性、易失性、突发性、无序性、无限性等方面提出了更高要求，现阶段关于大数据流式计算的研究则更多地从系统架构、数据传输、编程接口、高可用策略等方面开展和实施。当前典型的大数据流式计算的应用系统有Storm[17]、S4[18]。

在Storm系统中，采用主从式系统架构。如图5所示，一个Storm系统中有两类节点，即一个主节点Nimbus和多个从节点supervisor，有3种运行环境，即master、cluster和slaves。其中，主节点Nimbus运行在master环境中，是无状态的，负责全局的资源分配、任务调度、状态监控和故障检测；从节点supervisor运行在slaves环境中，也是无状态的，负责监听并接收来自于主节点Nimbus所分配的任务，并启动或停止自己所管理的工作进程worker，其中，工作进程worker负责具体任务的执行。zookeeper是一个针对大型分布式系统的可靠协调服务和元数据存储系统，通过配置zookeeper集群，可以使用zookeeper系统所提供的高可靠性的服务。Storm系统引入zookeeper，极大地简化了Nimbus、supervisor、worker之间的设计，保障了系统的稳定性。

在S4系统中，采用对等式系统架构。如图6所示，一个S4系统由用户空间、资源调度空间和S4处理节点空间组成。其中，在用户空间中，多个用户可以通过本地的客户端驱动实现服务的请求访问；在资源调度空间中，为用户提供了客户适配器，通过TCP/IP实现用户的客户端驱动与客户适配器间的连接和通信，多个用户可以并发地同多个客户适配器进行服务请求；在S4处理节点空间中，提供了多个处理节点Pnode，进行用户服务请求的计算，主要包括监听并分发接收到的事件计算请求，实现对事件流的路由选择、负载均衡、逻辑影射、故障恢复等功能。各个处理节点间保持相对的独立性、对等性和高并发性，极大

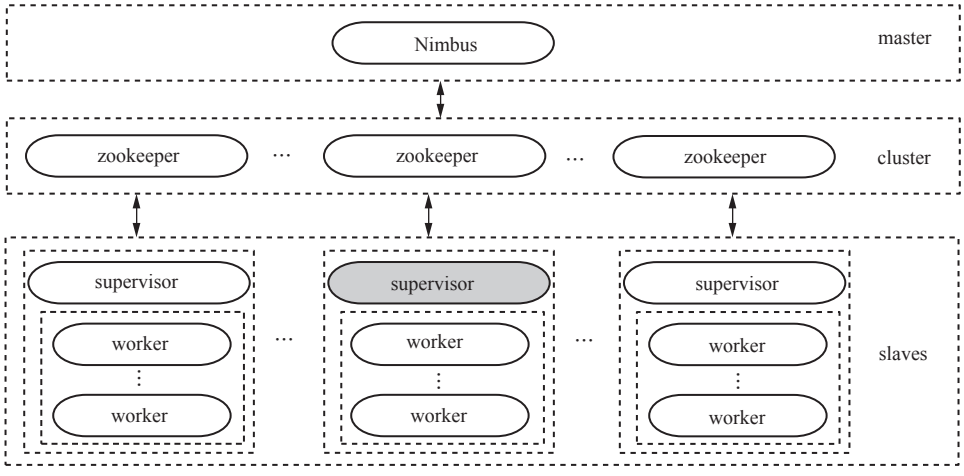

图5 Storm系统架构

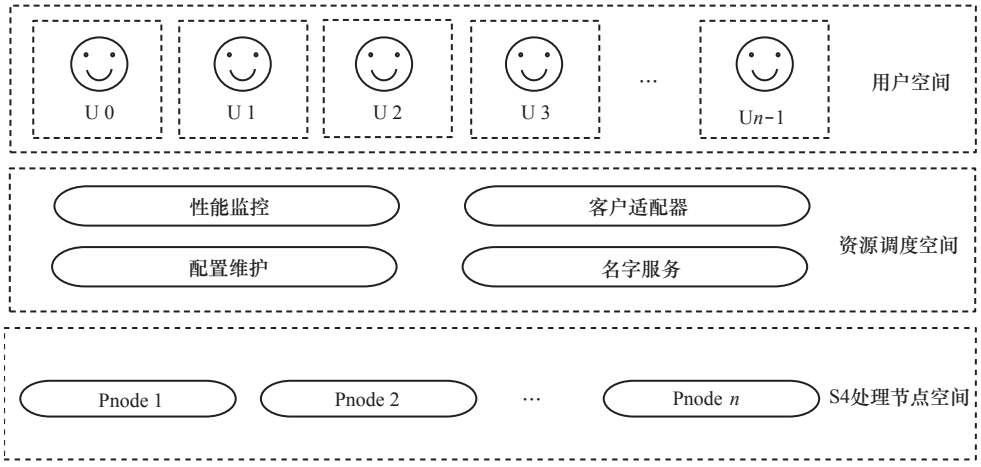

图6 S4系统结构

地提高了系统的性能，并通过散列方式将事件路由到一个或多个目标处理节点上。

3.3 大数据图计算系统

大数据图计算主要用来分析数据节点之间的关系和相似度，其巨大的商业价值已经凸显。例如，利用PageRank技术发现有影响力的用户，将GraphLab技术[20]用于社区、欺诈检测和推荐系统，还有一些分布式计算应用到Giraph、GraphX、Faunus和Grappa。GraphLab是美国卡耐基大学开发的一个并行的图挖掘分布式系统。该技术解决了传统MapReduce中有关机器学习处理频繁迭代计算和大量节点通信导致的计算效率低下的问题。具体来讲，在GraphLab中，以顶点为计算单元，将机器学习算法抽象为聚集、应用和分散3个步骤。在每一个迭代过程中，点的计算需要经过这3个步骤。并且，Graphlab是在共享内存的基础上，各机器异步、动态并行地执行计算任务，比BSP（bulk synchronous parallel，整体同步并行）计算效率更高，并且能够很好地保证数据的一致性。

4 未来研究方向

从系统角度看大数据计算，未来可能的研究方向包括以下几个方面。

（1）批量计算

大数据批量计算需要读写大量数据，而目前的存储系统主要针对计算密集型应用设计，从存储系统读出原始数据进行批量计算，计算结束后将计算结果写入存储系统。相应存储系统强调数据吞吐量，数据一致性保证程度高，数据读写时延相对较高。一个有潜力的研究方向是利用大数据批量计算的特征，解决大数据计算中的存储瓶颈问题。

另一类研究工作是针对典型应用进行定制化的性能优化，一个代表性例子是深度学习算法的并行加速技术研究。以深度学习中的卷积神经元网络为例，一个研究方向是使用GPU对卷积神经元网络进行加速，另一个方向是使用多台机器对卷积神经元网络进行并行化加速。

（2）大数据流式计算

需要构建一个高效、可扩展的计算平台，一方面需要具有很好的通用性，满足对

流式数据计算的需要，提供一系列公用的流式计算工具和属性；同时要在在线资源管理、状态一次性维护、用户级容错策略等方面具有良好的性能。

大数据流式计算中，数据流具有多流混合、流速波动等特性，一个研究方向是如何设计并优化流式计算中的资源调度策略，同时实现数据流速高时处理速度快和数据流速低时能耗低两个目标。大数据流式计算需要提供7×24 h的连续计算能力，对于系统可靠性方面的要求很高。另一个研究方向是如何利用流式计算的特征，同时实现数据流计算高可靠和可靠性维护开销低两个目标。

（3）大数据图计算

图计算系统的构建有两个思路：一种是为了避免数据关联性带来的机间通信而采用单机图处理。往往采用图数据分区的方法，每次加载一个分区，循环多次处理一张大图。网络大数据的多维关联性，导致大数据计算对网络图空间的访问发散性。由于缓存机制和介质特性，整个存储栈都对数据局部性表现出更好的性能。一个重要的研究方向是如何解决网络图空间的访问发散性与高效存储所需的数据局部性之间的矛盾。

另一种思路是充分发挥多台机器并行计算的优势而采用多机图计算。这种大数据图计算方式面临的最为突出的问题就是大图分割问题。由于对整个图的访问是随机进行的，一个研究方向是如何在图划分时实现通信代价低、计算及传输负载均衡、存储冗余度合理3个目标。

5 结束语

在大数据时代，大数据计算是大数据整个生命周期中的核心，是大数据中知识发现的关键。大数据计算模式主要包括大数据批量计算、流式计算、图计算、交互计算等，这些不同的计算模式分别满足不同的应用范式对数据计算结果在处理精度、实时性等方面的不同要求。这些计算模式并不是相互独立的，可以相互配合，满足同一应用范式在不同阶段对数据计算结果的要求。当前，批量计算是大数据计算的最主要模式。随着用户应用需求和技术的不断变化，所需要的计算模式也会不断变化，亟待根据最新应用范式的发展和要求，针对具体场景，开展对相关计算模式中出现的新情况、新问题的研究。

参考文献

[1] Chen C L, Zhang C Y. Data-intensive applications, challenges, techniques and technologies: a survey on big data. Information Sciences, 2014(275): 314~347

[2] Chang R M, Kauffman R J, Kwon Y. Understanding the paradigm shift to computational social science in the presence of big data. Decision Support Systems, 2014(63): 67~80

[3] Kambatla K, Kollias G, Kumar V, *et al*. Trends in big data analytics. Journal of Parallel and Distributed Computing, 2014(74): 2561~2573

[4] 李国杰，程学旗. 大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考. 中国科学院院刊，2012，27(6)：647~657
Li G J, Cheng X Q. Big data research: the major strategic areas of technology and economic development——research status and scientific thinking of big data. Bulletin of the Chinese Academy of Sciences, 2012, 27(6): 647~657

[5] 孙大为，张广艳，郑纬民. 大数据流式计算：关键技术及系统实例. 软件学报，2014，25(4)：839~862
Sun D W, Zhang G Y, Zheng W M. Big data stream computing: technologies and instances. Journal of Software, 2014, 25(4): 839~862

[6] 程学旗，靳小龙，王元卓等. 大数据系统和分析技术综述. 软件学报，2014，25(9)：

1889~1908
Cheng X Q, Jin X L, Wang Y Z, *et al*. Survey on big data system and analytic technology. Journal of Software, 2014, 25(9): 1889~1908

[7] 王元卓, 靳小龙, 程学旗. 网络大数据：现状与展望. 计算机学报, 2013, 36(6): 1125~1138
Wang Y Z, Jin X L, Cheng X Q. Network big data: present and future. Chinese Journal of Computers, 2013, 36(6): 1125~1138

[8] 李学龙, 龚海刚. 大数据系统综述. 中国科学: 信息科学, 2015, 45(1): 1~44
Li X L, Gong H G. Survey on big data system. Scientia Sinica Informationis, 2015, 45(1): 1~44

[9] Dobre C, Xhafa F. Intelligent services for big data science. Future Generation Computer Systems, 2014(37): 267~281

[10] Aisling O D, Jurate D, Roy D S. Big data, Hadoop and cloud computing in genomics. Journal of Biomedical Informatics, 2013(46): 774~781

[11] Hadoop. http://hadoop.apache.org/,2005

[12] Zaharia M, Das T, Li H, *et al*. Discretized streams: fault-tolerant streaming computation at scale. Proceedings of the SOSP 2013, Pennsylvania, USA, 2013

[13] Spark. http://spark-project.org,2013

[14] Cugola G, Margara A. Processing flows of information: from data stream to complex event processing. ACM Computing Surveys, 2012, 44(3): 51~62

[15] Zhang Z, Gu Y, Ye F, *et al*. A hybrid approach to high availability in stream processing systems. Proceedings of the 30th IEEE International Conference on Distributed Computing Systems, Genova, Italy, Jun 2010: 138~148

[16] Liu X F, Lftikhar N, Xie X. Survey of real-time processing systems for big data. Proceedings of IDEAS 2014, Porto Portugal, 2014: 356~361

[17] Storm. http://storm-project.net/,2015

[18] Chauhan J, Chowdhury S A, Makaroff D. Performance evaluation of Yahoo! S4: a first look. Proceedings of 7th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Victoria, BC, Canada, 2012: 58~65

[19] Chatziantoniou D, Pramatari K, Sotiropoulos Y. Supporting real-time supply chain decisions based on RFID data streams. Journal of Systems and Software, 2011, 84(4): 700~710

[20] GraphLab. http://graphlab.org/projects/index.html,2015

[21] Furedi Z, Kostochka A, Kumbhat M. Choosability with separation of complete multipartite graphs and hypergraphs. Journal of Graph Theory, 2014, 76(2): 129~137

作者简介

郑纬民，男，清华大学教授、博士生导师，中国计算机学会理事长，目前主要从事并行与分布式计算、存储系统的研究工作，主持和参与多项国家“973”计划、“863”计划、国家自然科学基金项目。近年来在IEEE TC/IEEE TPDS/ACM TOS/FAST等本领域顶级期刊与国际会议发表论文40余篇。

收稿日期： 2015-05-03；**修回日期：** 2015-05-06
基金项目： 国家重点基础研究发展计划（“973”计划）基金资助项目（No.2014CB340402），国家自然科学基金资助项目（No.61170008，No.61272055）
Foundation Items: The National Basic Research Program of China(973 Program)(No.2014CB340402), The National Natural Science Foundation of China(No.61170008, No.61272055)
论文引用格式： 郑纬民. 从系统角度审视大数据计算. 大数据, 2015002
Zheng W M. Reviewing big data computation from a system perspective. Big Data Research, 2015002