“大数据”背景下的信息处理技术分析与研究*

□ 于薇 / 中国科学技术信息研究所 北京 100038

摘要：“大数据”是信息科技领域出现的一个研究焦点。文章从“大数据”的概念进入，对其特征进行梳理总结，对“大数据”处理的核心技术进行分析，比较分析各个“大数据”处理商业解决方案的特点，最后结合“大数据”特征分析科技文献信息，对“大数据”在科技文献信息处理领域的应用进行探讨性分析与研究。

关键词：大数据，Hadoop，信息处理技术
DOI：10.3772/j.issn.1673—2286.2012.11.003

引言

“大数据”是继“云计算”之后，在信息科技领域出现的一个研究焦点。无处不在的传感器和微处理器，形成了庞大的数据来源。科学研究领域，气象数据、地理数据、生物信息数据等是传统的海量数据集；制造业领域，很多机器上都安装了一个或多个微处理器来采集生产数据；商业消费领域，网上购买记录、消费评价等等数据都成为大数据问题；各国政府的海量统计数据和文件也因计算机技术的发展而成为亟待分析处理的大数据问题。国际数据公司（IDC）的数字宇宙研究报告称，2011年全球被创建和被复制的数据总量为1.8ZB，并预测到2020年，全球将拥有35ZB的数据量。可以说，世界已经进入到以数据为中心的时代——“大数据”时代。

1 关于“大数据”

目前为止，“大数据”还没有一个统一的定义，信息处理领域的大企业和机构都提出了“大数据”定义，总结起来，主要有以下几个代表性定义：

● 麦肯锡在其报告中指出，“大数据”是指那些大小已经超出了传统意义上的尺度，常规的数据库技术难以完成捕捉、存储、管理和分析的数据集合[1]。

● IBM把大数据概括成了3个V，即规模性（Volume）、多样性（Variety）、高速性（Velocity）。规模性指大数据的规模很大，突破了PB级数据量；多样性指大数据来源广泛，包括结构化和非结构化数据，如文本、传感器数据、视频、音频、网络日志等；高速性指大数据对数据实时处理能力要求极高[2]。

● 按EMC的界定，“大数据”其中的“大”是指大型数据集，一般在10TB规模左右；多用户把多个数据集放在一起，形成PB级的数据量；同时这些数据来自多种数据源，以实时、迭代的方式来实现[3]。

● 在Forrester分析师布赖恩•霍普金斯（Brian Hopkins）和鲍里斯•埃韦尔松（Boris Evelson）撰写的《首席信息官，请用大数据扩展数字视野》报告中，他们提出大数据的4项典型特征——海量（Volume）、多样性（Variety）、高速（Velocity）和易变性（Variability）[4]。

● IDC提出的“4V”原则，即容量、类型、速度和价值（volume、variety、velocity和value）。

可以看出，大数据必须同时具备海量（Volume）、多样性（Variety）、高速（Velocity）等特征，才能称为“大数据”。除此之外，数据的真实性（Veracity）在未

* 中国科学技术信息研究所预研基金项目“基于数字资源的异构语义元数据融合技术及服务研究”（编号：YY201219）。

来的大数据应用中会越来越重要[2]。同时，“大数据”不仅仅局限于符合上述特征的数据集合，还包括捕获、存储、管理和分析这些数据的技术。

2 Hadoop

大数据技术涵盖了硬软件多个方面的技术，目前各种技术基本都独立存在于存储、开发、平台架构、数据分析挖掘的各个相对独立的领域。这一部分主要介绍和分析大数据处理的核心技术——Hadoop。

2.1 Hadoop的组成

大数据不同于传统类型的数据，它可能由TB甚至PB级信息组成，既包括结构化数据，也包括文本、多媒体等非结构化数据。这些数据类型缺乏一致性，使得标准存储技术无法对大数据进行有效存储，而且我们也难以使用传统的服务器和SAN方法来有效地存储和处理庞大的数据量。这些都决定了“大数据”需要不同的处理方法，而Hadoop目前正是广泛应用的大数据处理技术。

Hadoop是一个基于Java的分布式密集数据处理和数据分析的软件框架。该框架在很大程度上受Google在2004年白皮书中阐述的MapReduce的技术启发。Hadoop主要组件包含[5]：

Hadoop Common：通用模块，支持其他Hadoop模块

Hadoop Distributed File System（HDFS）：分布式文件系统，用以提供高流量的应用数据访问

Hadoop YARN：支持工作调度和集群资源管理的框架

HadoopMapReduce：针对大数据的、灵活的并行数据处理框架

其他相关的模块还有：

ZooKeeper：高可靠性分布式协调系统

Oozie：负责MapReduce作业调度

HBase：可扩展的分布式数据库，可以将结构性数据存储为大表

Hive：构建在MapRudece之上的数据仓库软件包

Pig：架构在Hadoop之上的高级数据处理层

在Hadoop框架中，最底层的HDFS存储Hadoop集群中所有存储节点上的文件。HDFS的架构是基于一组特定的节点构建的（图2），这些节点包括一个NameNode和大量的DataNode。存储在HDFS中的文件被分成块，然后将这些块复制到多个计算机中（DataNode）。这与传统的RAID架构大不相同。块的大小（通常为64MB）和复制的块数量在创建文件时由客户机决定。NameNode可以控制所有文件操作。HDFS内部的所有通信都基于标准的TCP/IP协议。NameNode在

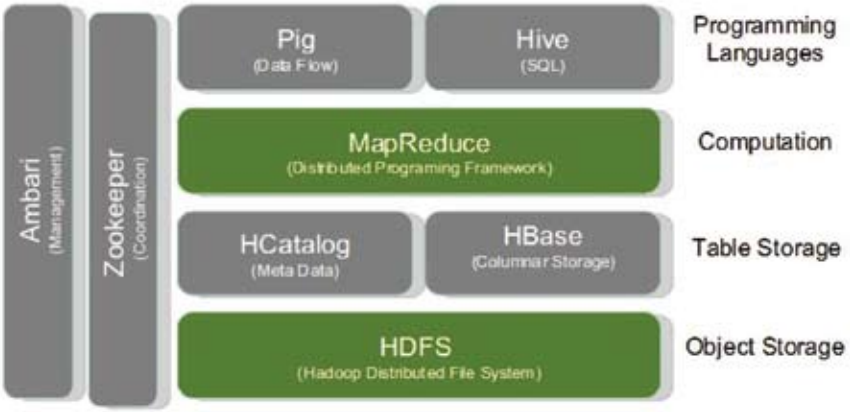

图1 Hadoop框架组成模块

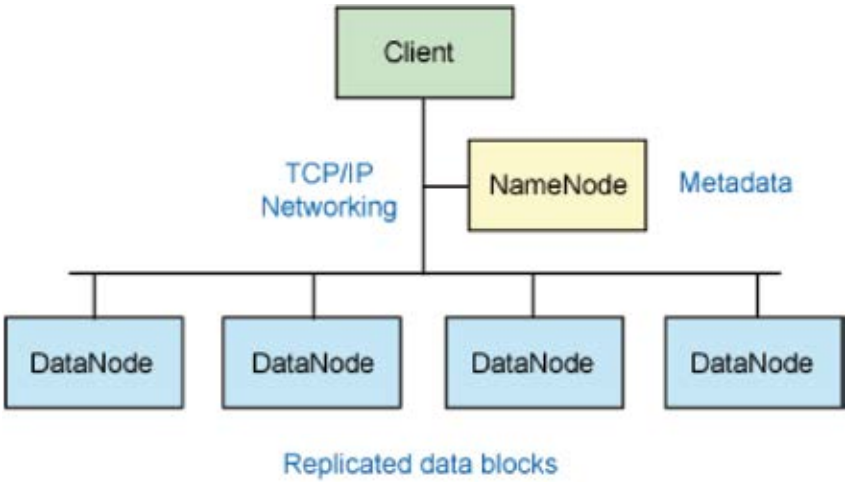

图2 HDFS框架节点

HDFS内部提供元数据服务，负责管理文件系统名称空间和控制外部客户机的访问。它决定是否将文件映射到DataNode上的复制块上。DataNode通常以机架的形式组织，机架通过一个交换机将所有系统连接起来。

HadoopMapReduce是Google MapReduce的开源实现。MapReduce技术是一种简洁的并行计算模型，它在系统层面解决了扩展性、容错性等问题，通过接受用户编写的Map函数和Reduce函数，自动地在可伸缩的大规模集群上并行执行，从而可以处理和分析大规模的数据[6]。Hadoop提供了大量的接口和抽象类，从而为Hadoop应用程序开发人员提供许多工具，可用于调试和性能度量等。

在Hadoop应用实例中，一个代表客户机在单个主系统上启动MapReduce的应用程序称为JobTracker。类似于NameNode，它是Hadoop集群中唯一负责控制MapReduce应用程序的系统。在应用程序提交之后，将提供包含在HDFS中的输入和输出目录。JobTracker使用文件块信息（物理量和位置）确定如何创建其他TaskTracker从属任务。MapReduce应用程序被复制到每个出现输入文件块的节点，将为特定节点上的每个文件块创建一个唯一的从属任务。每个TaskTracker将状态和完成信息报告给JobTracker。图3显示一个示例集群中的工作分布。

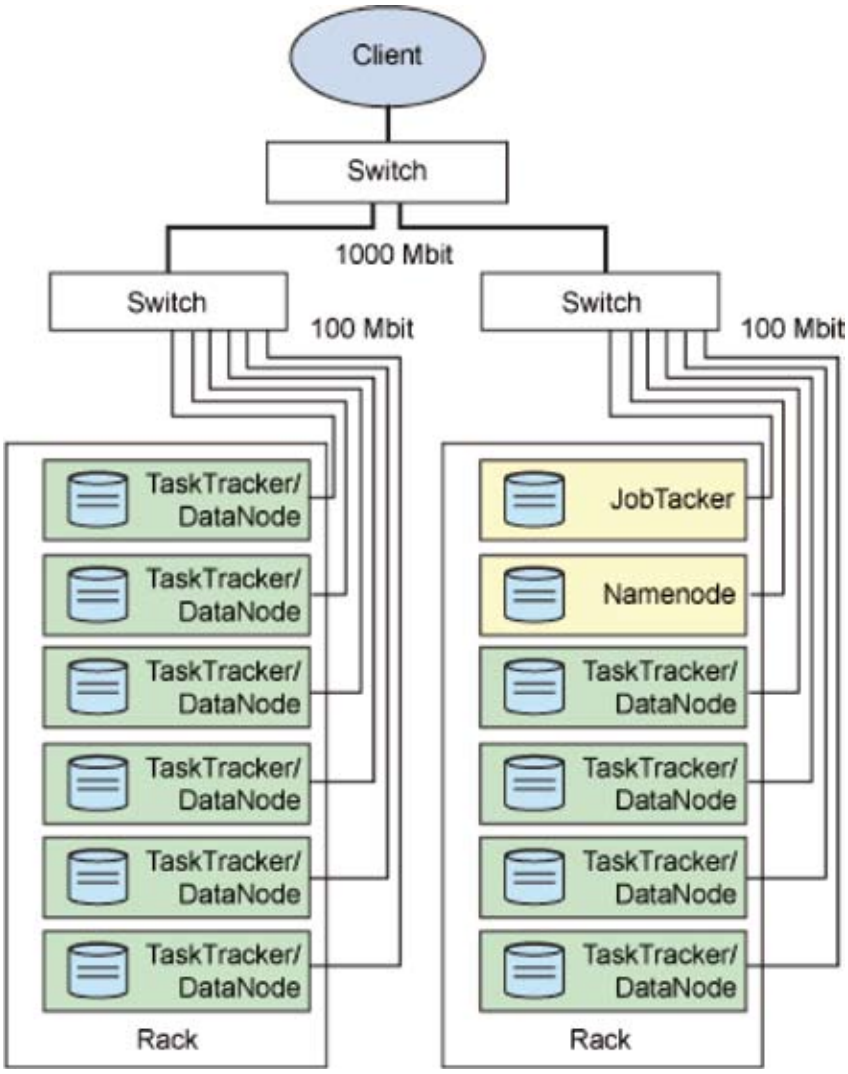

图3 HadoopMapReduce工作分布示例

2.2 Hadoop的优点[7]

Hadoop能够使用户轻松开发和运行处理大数据的应用程序。它主要有以下几个优点：

（1）高可靠性。Hadoop按位存储和处理数据的能力值得人们信赖。

（2）高扩展性。Hadoop是在可用的计算机集簇间分配数据并完成计算任务的，这些集簇可以方便地扩展到数以千计的节点中。

（3）高效性。Hadoop能够在节点之间动态地移动数据，并保证各个节点的动态平衡，因此处理速度非常快。

（4）高容错性。Hadoop能够自动保存数据的多个副本，并且能够自动将失败的任务重新分配。

Hadoop带有用Java语言编写的框架，因此运行在Linux生产平台上是非常理想的。Hadoop上的应用程序也可以使用其他语言编写，比如C++。

2.3 Hadoop的不足

Hadoop作为一个处理大数据的软件框架，虽然受到众多商业公司的青睐，但是其自身的技术特点也决定了它不能完全解决大数据问题。

在当前Hadoop的设计中，所有

的metadata操作都要通过集中式的NameNode来进行，NameNode有可能是性能的瓶颈[8]。

当前Hadoop单一NameNode、单一Jobtracker的设计严重制约了整个Hadoop可扩展性和可靠性。首先，NameNode和JobTracker是整个系统中明显的单点故障源。再次，单一NameNode的内存容量有限，使得Hadoop集群的节点数量被限制到2000个左右，能支持的文件系统大小被限制在10-50PB，最多能支持的文件数量大约为1.5亿左右。实际上，有用户抱怨其集群的NameNode重启需要数小时，这大大降低了系统的可用性[9]。

随着Hadoop被广泛使用，面对各式各样的需求，人们期望Hadoop能提供更多特性，比如完全可读写的文件系统、Snapshot、Mirror等等。这些都是当前版本的Hadoop不支持，但是用户又有强烈需求的。

3 主要商业性“大数据”处理方案

“大数据”被科技企业看作是云计算之后的另一个巨大商机，包括IBM、谷歌、亚马逊和微软在内的一大批知名企业纷纷掘金这一市场。此外，很多初创企业也开始加入到大数据的淘金队伍中。Hadoop是非结构数据库的代表，低成本、高扩展性和灵活性等优势使其成为各种面向大数据处理分析的商业服务方案的首选。Oracle、IBM、Microsoft三大商业数据提供商是Hadoop的主要支持者。很多知名企业都以Hadoop技术为基础提供自己的商业性大数据解决方案。这一部分主要介绍以Hadoop为基础的典型商业性大数据解决方案。

3.1 IBM InfoSphere大数据分析平台

IBM于2011年5月推出的InfoSphere大数据分析平台是一款定位为企业级的大数据分析产品。该产品包括BigInsights和Streams，二者互补，Biglnsights基于Hadoop，对大规模的静态数据进行分析，它提供多节点的分布式计算，可以随时增加节点，提升数据处理能力。Streams采用内存计算方式分析实时数据。它们将包括HadoopMapReduce在内的开源技术紧密地与IBM系统集成起来。研究Hadoop这样开源技术的人很多，但是IBM这次是真正将其变成了企业级的应用，针对不同的人员增加不同的价值[10]。

InfoSphereBigInsights 1.3的存储和运算框架采用了开源的HadoopMapReduce，同时针对Hadoop框架进行了改造，采用了IBM特有的通用并行文件系统——GPFS。利用GPFS的目的是为了避免单点故障，保证可用性。BigInsights中还有两个分析产品——Cognos和SPSS，这两个分析产品在传统功能上加强了文本分析的功能，提供了一系列文本分析工具，并使用高级语言进行自定义规则，如文本格式转换等。目前BigInsights提供两种版本，一种是企业版（Enterprise Edition），用于企业级的大数据分析解决方案。另一种是基础版（Basic Edition），去掉了企业版中的大部分功能，用户可以免费下载，主要提供给开发人员和合作伙伴试用。

Streams最大的特点就是内存分析，利用多节点PC服务器的内存来处理大批量的数据分析请求。Streams的特点就是“小快灵”，数据是实时流动的，其分析反应速度可以控制在毫秒级别，而BigInsights的分析是批处理，反应速度无法同Streams相比。总体来说，二者的设计架构不同，也用于处理不同的大数据分析需求，并可以形成良好的互补。

InfoSphere平台仅仅是IBM大数据解决方案中的一部分。IBM大数据平台包括4大部分：信息整合与治理组件、基于开源Apache Hadoop的框架而实现的BigInsights平台、加速器，以及包含可视化与发现、应用程序开发、系统管理的上层应用。通过IBM的解决方案可以看出，解决大数据问题不能仅仅依靠Hadoop。

3.2 Oracle Big Data Appliance[11]

Oracle Big Data Appliance准确地说是一款硬件产品，添加了Hadoop装载器、应用适配器以及Oracle新的NoSQL数据库，主要目的是为了将非结构化数据加载到关系型数据库中去，并对软硬件的集成做了一些优化。Oracle Big Data机包括开源Apache Hadoop、Oracle NoSQL数据库、Oracle数据集成Hadoop应用适配器、Oracle Hadoop装载器、Open Source Distribution of R、Oracle Linux和Oracle Java HotSpot虚拟机。它能够快速、便捷地与Oracle数据库11g、Oracle Exadata数据库云服务器和Oracle Exalytics商务智能云服务器集成。分析师和统计人员可以运行现有的R应用，并利用R客户端直接处理存储在Oracle数据库11g中的数据，从而极大地提高可扩展

性、性能和安全性。

3.3 Microsoft SQL Server[12]

微软已经发布Hadoop Connector for SQL Server Parallel Data Warehouse和Hadoop Connector for SQL Server社区技术预览版本的连接器。该连接器是双向的，用户可以在Hadoop和微软数据库服务器之间向前或者向后迁移数据。微软的SQL Server 2012将并入Hadoop分布式计算平台，微软还将把Hadoop引入Windows Server和Azure（微软的云服务）。

3.4 Sybase IQ

Sybase IQ是Sybase公司推出的特别为数据仓库设计的关系型数据库，添加了Hadoop的集成，并提供了MapReduce的API。相比于传统的“行式存储”的关系型数据库，Sybase IQ使用了独特的列式存储方式，在进行分析查询时，仅需读取查询所需的列，其垂直分区策略不仅能够支持大量的用户、大规模数据，还可以提交对商业信息的高速访问，其速度可达到传统的关系型数据库的百倍甚至千倍。

4 其他“大数据”解决方案

“大数据”解决方案并非只有Hadoop一种，许多知名企业还提供了其他的解决方案。

4.1 EMC

EMC提供了两种大数据存储方案，即Isilon和Atmos。Isilon能够提供无限的横向扩展能力，Atmos是一款云存储基础架构，在内容服务方面，Atmos是很好的解决方案。在数据分析方面，EMC提供的解决方案、提供的产品是Greenplum，Greenplum有两个产品，第一是GreenplumDatabase，GreenplumDatabase是大规模的并行成立的数据库，它可以管理、存储、分析PB量级的一些结构性数据，它下载的速度非常高，最高可以达到每小时10TB，速度非常惊人。这是EMC可以提供给企业、政府，用来分析海量的数据。但是GreenplumDatabase面对的是结构化数据。很多数据超过90%是非结构化数据，EMC有另外一个产品是GreenplumHD，GreenplumHD可以把非结构化的数据或者是半结构化的数据转换成结构化数据，然后让GreenplumDatabase去处理。

4.2 BigQuery

BigQuery是Google推出的一项Web服务，用来在云端处理大数据。该服务让开发者可以使用Google的架构来运行SQL语句对超级大的数据库进行操作。BigQuery允许用户上传他们的超大量数据并通过其直接进行交互式分析，从而不必投资建立自己的数据中心。Google曾表示BigQuery引擎可以快速扫描高达70TB未经压缩处理的数据，并且可马上得到分析结果。大数据在云端模型具备很多优势，BigQuery服务无需组织提供或建立数据仓库。而BigQuery在安全性和数据备份服务方面也相当完善。免费帐号可以让用户每月访问高达100GB的数据，用户也可以付费使用额外查询和存储空间。

5“大数据”与科技文献信息处理

“大数据”目前主要指医学、天文、地理、Web日志、多媒体信息等数据，鲜有提及文献信息。事实上，现在的科技文献信息日益凸显出“大数据”的特征，主要表现在以下几个方面：更新周期缩短；数量庞大；文献的类型多样；文献载体数字化；文献语种多样化；文献内容交叉；文献信息密度大。

科技文献中所含的信息类型多样，既有结构性数据，也有非结构性文本和公式，如何利用“大数据”技术对文献内容进行分析，挖掘用户访问日志、评价反馈等数据的价值，为用户提供服务成为科技信息服务业急需思考和解决的问题。在科技文献信息处理中，文本分析技术、语义计算技术、数据安全需要与“大数据”解决方案结合起来考虑实施，这样才能更有效地提供知识服务。

6 结语

“大数据”技术还处于起步阶段，主要服务产品大都围绕Hadoop架构发展而来，但是“大数据”不等同于Hadoop，云存储与云计算、传统关系型数据库技术在“大数据”时代仍有其不可替代的优势。传统的信息组织方式与“大数据”技术的结合，是文献信息处理领域新的研究课题。

参考文献

[1] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity [OL]. [2012-08-21]. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.
[2] IBM. What is big data? [OL]. [2012-08-20]. http://www-01.ibm.com/software/data/bigdata/.
[3] 姜奇平.大数据时代到来[OL]. (2012-02-02) [2012-08-15]. http://www.ciweek.com/article/2012/0118/A20120118554491.shtml.
[4] 什么是大数据?[OL]. (2012-09-07) [2012-09-18]. http://www.enet.com.cn/article/2012/0907/A20120907159678.shtml.
[5] [OL]. (2012-02-10) [2012-09-02]. http://hadoop.apache.org/.
[6] 覃雄派,王会举,杜小勇,等.大数据分析——RDBMS与MapReduce的竞争与共生[J].软件学报,2012,23(1):32-45.
[7] Hadoop [OL]. [2012-09-22]. http://baike.baidu.com/view/908354.htm.
[8] PAVLO A, PAULSON E, RASIN A, et al. A comparison of approaches to large-scale data analysis [C]// SIGMOD '09: Proceedings of the 35th SIGMOD international conference on Management of data, New York, NY, USA, 2009: 165-178.
[9] HDFS-273 [OL]. [2012-09-02]. https://issues.apache.org/jira/browse/HDFS-273.
[10] [OL]. [2012-09-02]. http://www.searchdatabase.com.cn/showcontent_56707.htm.
[11] [OL]. [2012-09-02]. http://www.searchdatabase.com.cn/showcontent_53541.htm.
[12] 布局大数据 全球十四大数据厂商论战[OL]. (2012-08-01) [2012-08-17]. http://cloud.zol.com.cn/310/3108708.html.

作者简介

于薇，中国科学技术信息研究所信息技术支持中心助理研究员，中国人民大学信息资源管理学院在读博士。主要研究方向：自然语言处理，电子文件管理。近期关注方向是元数据互操作的方法研究。E-mail: yuwei@istic.ac.cn

A Research of Information Processing Technology for Big Data

Yu Wei / Information Technology Support Center, Institute of Scientific and Technical Information of China, Beijing, 100038

Abstract: Big Data is new research focused in the field of Information Science, which includes hardware and software. First, this paper introduces the concept of Big Data and summarizes the features of Big Data. Secondly, the paper focuses on Hadoop, the main technology of Big data. On the analysis of the advantages and shortages of Big Data, the paper compares and analyzes the typical services to resolve Big Data, which use Hadoop, cloud storage and cloud computing and traditional database technology. Finaly, the paper summarizes the features of science and technology information and proposes the application of Big Data technology in the field of science and technology information service.

Keywords: Big data, Hadoop, Information processing technology, Science and technology information service

（收稿日期：2012-09-22）