



“云计算与大数据”及其对勘探地震技术发展的影响

唐杰^{1,2}, 魏嘉^{3*}, 武港山^{1,2}

1. 南京大学软件新技术国家重点实验室, 南京 210023;

2. 南京大学计算机科学与技术系, 南京 210023;

3. 中国石化石油勘探技术研究院, 南京 211103

* 联系人, E-mail: weijia.swty@sinopec.com

2017-02-13 收稿, 2017-06-05 修回, 2017-06-07 接受, 2017-07-28 网络版发表

摘要 互联网技术发展至今, 已进入了大数据和云计算时代, 并在传统商业领域取得了令人瞩目的成功, 产生了许多新的发展模式。但是, 在勘探地震专业领域, “云计算与大数据”技术的潜力尚没有充分发挥。本文介绍了“云计算与大数据”的方法论和基本特征, 分析了云计算、大数据和物联网、互联网+等关键技术之间的关系, 并结合地震采集、处理、解释的技术特点和发展需求, 探讨了“云计算与大数据”对地震技术发展的影响, 提出了建设地震专业软件生态系统和基于物联网改变地震采集现状的设想。

关键词 勘探地震技术, 大数据, 云计算, 物联网, 互联网+

油气地球物理, 尤其是勘探地震技术领域是计算机技术和软件技术应用最广泛的专业领域, 也是IT技术应用最早的专业领域, “云计算与大数据”已经彻底改变了传统产业的运作模式和人际交往的方式, 同时也在改变着油气地球物理领域的工作模式。然而, 油气地球物理领域过于“专业化”, 技术垄断程度远远高于其他领域, 其主要的经济收益不是来自规模效应, 无法通过互联网的平台化和便利性的特征快速扩大商业规模从而获得“额外的”收益, 因此, 无法照搬其他领域的成功经验, 需要从地震技术层面的特征出发。

近年来, 勘探地震技术面临着新的挑战, 一方面地震数据的采集量呈爆发式增长, 另一方面则是对实时处理来自井中各类传感器的数据流的需求也在不断提高。地震数据处理已成为计算密集型和数据密集型应用。虽然传统的高性能开发平台和模型

在处理计算密集型应用时有一定的优势, 但是随着数据规模和种类的不断增加, 传统的高性能平台已越来越力不从心。如何确保处理平台的易于使用、高效、易分享、具备较好的可伸缩性也是地震数据处理平台必须考虑的问题。此外, 地震处理解释专业软件现状是: 基于竖井式的全系列产品开发模式导致各厂商付出高昂的研发成本, 厂商之间形成了天然的技术壁垒, 资源不能共享。因此, 如何打造一个易于分享的地震数据处理生态系统也是目前的重要研究内容和发展方向。

在主题为“各向异性地球物理与矢量场技术”的香山科学会议第567次学术讨论会上, 与会人员围绕近年来基于互联网的云技术与大数据的发展, 不仅带来了地球物理观测方式的剧变, 还将可能给地球物理数据解译的理论架构、软硬件技术的发展带来颠覆性的冲击等一系列问题进行了深入的探讨, 达成

引用格式: 唐杰, 魏嘉, 武港山. “云计算与大数据”及其对勘探地震技术发展的影响. 科学通报, 2017, 62: 2630–2638

Tang J, Wei J, Wu G S. The impact of “Cloud Computing and Big Data” on the future trends of seismic exploration (in Chinese). Chin Sci Bull, 2017, 62: 2630–2638, doi: 10.1360/N972017-00158

了若干共识,并提出了建设地震专业软件生态系统和基于物联网改变地震采集现状的设想。

1 “云计算与大数据”概述

随着信息技术发展和信息社会需求的不断提高,可以说我们现在已经步入了“数据为中心”的时代,数据本身是资产,这一点已越来越为人们所共识。云计算^[1]技术(Cloud Computing)和大数据^[2](Big Data)技术的产生则是这一阶段的必然结果。总的来说,云计算通过互联网向用户提供计算力、存储服务,为互联网信息处理提供硬件基础。大数据技术则运用日趋成熟的云计算技术从浩瀚的数据信息海洋中通过对信息进行归纳、检索、整合,进而快速、高效地获得有价值的信息,为海量信息处理提供软件基础。

1.1 云计算技术

什么是云计算?目前业界公认的定义是:云计算是一种商业计算模型,它将计算任务分布在大量计算机构成的资源池上,使各种应用系统能够根据需要获取计算力、存储空间和信息服务^[1]。云计算是通过网络按需提供可动态伸缩的廉价计算服务。

一般认为,云计算的核心特征有5点,即按需自服务、泛在可接入、资源全池化、快捷可伸缩、服务可量化。在云计算中,提供资源的网络被称为“云”。“云”中的资源在使用者看来是可以无限扩展的,并且可以随时获取,按需使用,随时扩展,按使用付费。这种特性经常被称为像水电一样使用IT基础设施。它意味着计算能力也可以作为一种商品进行流通,就像煤气、水电一样,取用方便,费用低廉。最大的不同在于,它是通过互联网进行传输的。

美国国家标准和技术研究院的“云计算”定义中明确了3种服务模式^[3]。

软件即服务(SaaS):消费者使用应用程序,但并不掌控操作系统、硬件或运作的网络基础架构。这是一种服务观念的基础,软件服务供应商,以租赁的概念提供客户服务,而非购买,比较常见的模式是提供一组账号密码。

平台即服务(PaaS):消费者使用主机操作应用程序。消费者掌控运作应用程序的环境(也拥有主机部分掌控权),但并不掌控操作系统、硬件或运作的网络基础架构。平台通常是应用程序基础架构。

基础设施即服务(IaaS):消费者使用“基础计算

资源”,如处理能力、存储空间、网络组件或中间件。消费者能掌控操作系统、存储空间、已部署的应用程序及网络组件(如防火墙、负载均衡器等),但并不掌控云基础架构。

此外,常见的服务模式还有公用计算(utility computing):即建立虚拟的数据中心使得其能够把内存、I/O设备、存储和计算能力集中起来成为一个虚拟的资源池来为整个网络提供服务。

从发展角度来看,云计算是并行计算(parallel computing)、分布式计算(distributed computing)和网格计算(grid computing)的演化,或者说是这些计算机科学概念的商业实现,同时也是虚拟化(virtualization)、公用计算(utility computing)、基础设施即服务(IaaS)、平台即服务(PaaS)、软件即服务(SaaS)等概念混合与演化的结果。

与传统模式相比,云计算具有很多优点。首先,云计算提供了可靠、安全的数据存储中心,用户不用再担心数据丢失、病毒入侵等麻烦;其次,云计算对用户端的设备要求较低,使用起来也最方便;此外,云计算可以轻松实现不同设备间的数据与应用共享;最后,云计算为我们使用网络提供了几乎无限多的可能。

现阶段,云计算在很多领域都有了成功的应用,如云政务、云教育、云物联、云安全、云存储……这些日益涌现的云技术应用融合了并行处理、智能感知、普适计算、网络互连、海量存储等新兴技术,通过分布在各领域的客户端对互联网中的事件进行监测、获取、处理、反馈,为各应用部门提供最为便捷和成本低廉的解决方案。

1.2 大数据技术

大数据是指所涉及的数据量规模巨大,无法通过人工或者计算机在合理的时间内截取、管理、处理并整理成为人类所能解读的形式信息^[4]。大数据具有四大特征:(1)数据体量巨大,从TB级快速升到PB甚至ZB级别;(2)数据类型繁多,如网络日志、音频、图像、视频等等;(3)价值大但密度低,以视频为例,连续不间断监控过程,可能有用的数据只有1, 2 s;(4)处理速度要求要快。因此,业界将“大数据”特征归纳为5V,即体量大(volume)、速度快(velocity)、模式多(variety)、难辨识(veracity)和价值大密度低(value)。

大数据技术则是指从各种类型的数据中快速获得有价值信息的技术、手段和方法。一般来讲,大数据技术的总体架构包括3层:数据存储、数据分析和数据分析。数据首先通过存储层存储下来,数据处理层提供强大的并行计算和分布式计算能力来保障分析的时效性,数据分析层则是根据需求和目标,建立相应的数据模型和指标体系,对数据进行分析,从而充分挖掘数据的价值。三层相互配合,让大数据最终产生价值。

(1) 数据存储层。数据有许多不同的分类方法,可以划分为结构化、半结构化或非结构化数据,也可以划分为元数据、主数据、业务数据,还可以分为GIS、视频、文件、语音、业务交易类的等各种数据。传统的结构化数据库已经无法满足数据多样性的存储要求,因此又产生了HDFS, NoSQL等存储系统,可以用于非结构化或结构化和半结构化的数据文件存储。

(2) 数据处理层。数据处理层需要解决的核心问题在于随着数据规模增大,存储出现分布式后带来的数据处理上的复杂度和时效性问题。通常是在传统的云计算相关技术架构上,整合Hive, Hadoop-MapReduce等框架的相关技术形成大数据处理层的能力。

(3) 数据分析层。分析层的重点是挖掘大数据的价值所在。大数据的出现使得使用更加复杂的模型来更有效地表征数据、解释数据成为可能。深度学习^[5,6]利用层次化的架构学习出对象在不同层次上的表达,解决更加复杂抽象的问题。知识计算^[7,8]通过对大数据进行高端分析,抽取有价值的知识,并把它构建成支持查询、分析和计算的知识库。可视化^[9,10]技术通过交互式、动态化以及大规模的展示可以让用户更好地理解分析的结果,帮助人们分析大规模、高维度、多来源、动态演化的信息,并辅助做出实时的决策。

1.3 云计算与大数据系统

云计算与大数据技术发展至今,已在电子政务、网络通信、医疗卫生、能源气象、金融零售等行业领域中得到了应用。然而,不同领域中的大数据需求各不相同,随之建立的大数据系统也各具特色。总体来说目前大数据系统主要分为批量处理、流式处理和实时交互处理等几种类型^[4]。

批量数据通常体量巨大、很少更新、数据精度高

但数据价值密度低。因此,需要合理的算法才能从批量数据中高效地抽取出有用的价值。典型的应用如能源领域对来自地表深处地震时产生的数据进行批量的排序和整理,可能发现海底石油的储量等。批量处理的代表性系统为Google公司的MapReduce^[11]编程模型和GFS^[12]文件系统。该系统具备模型简单、易于使用、可伸缩性好等优点^[13,14],一经推出就在学术界于工业界引起了很大反响,历经10余年的发展,形成了丰富的Hadoop软件生态系统(<http://hadoop.apache.org/>)。

流式数据通常带有时间标签,流速差异较大,数据格式可以是结构化、半结构化或非结构化,且垃圾信息较多。典型应用如环境监测、灾害预警、金融数据分析等。流式处理的典型系统为Twitter的Storm系统(<http://storm.incubator.apache.org/>)。Storm是一套分布式、可靠、可容错的用于处理流式数据的系统。其流式处理作业被分发至不同类型的组件,每个组件负责一项简单的、特定的处理任务,具有编程模型简单、系统容错性好、扩展能力强等优点。

交互式数据处理以人机对话的方式进行操作,同时输入的信息需要及时处理与反馈,以保证交互能够继续进行。交互式处理的典型应用为互联网中的交互式工具,如即时通讯、社交网络、电子商务、搜索引擎等。交互式处理的典型系统为Berkeley的Spark^[15]系统。Spark是一个基于内存计算的可扩展开源集群计算系统,配合Hive(<https://hive.apache.org/>), Pig(<https://pig.apache.org/Spark>)等为代表的分布式数据仓库能够支持上千台服务器的规模,使用内存进行数据计算以便快速处理查询,实时返回分析结果。此外,还可以利用Spark-Shell可以开启交互式Spark命令环境,提供交互式查询。这些优点使得Spark在分布数据处理领域有着很好的应用前景。

2 勘探地震技术的阶段性及其特点

地震采集是勘探地震技术的基础,地震采集技术的本质就是通过“网络化(网格化)”的传感器布设获取地下地质信息的技术过程。地震采集目前采用的“地震仪+检波器”的采集方式并不是地震采集技术的内在要求,而是技术发展阶段的产物,当通用计算机能力有限和部署方式受限制时,地震勘探不得不研制专用设备(地震仪)进行地震采集的控制,地震仪也是根据技术条件的限制,只能控制一定数量排列组

成的检波器“阵列”，从千道发展到万道的规模，再发展到现在百万道的规模(Sercel公司的508XT)。虽然地震仪控制的道数发生了巨大的进步，但仍没有达到“随心所欲”的地步。再者，这样的地震仪价格极其昂贵。通俗地讲，地震采集的每个接收点就是地震采集“网络”中的一个节点，实际上可以理解为局部“物联网”中的一个节点，整个地震采集工区内的所有地震采集接收点就组成一个局部“物联网”，正是由于这样的特点，近年来出现了AutoSeis和Wireless等公司的无线的节点式地震检波器。上述节点式检波器虽然没有完全采用“物联网”技术进行整体架构，相较于传统的地震仪而言，已经看到了技术变革的端倪。

随着地震采集技术的不断发展，地震数据处理面临的数据量日益增大，地震数据处理能力需要能够适应这种数据量的急剧增长。但是，地震数据并不严格等同于其他领域所说的“大数据”，大数据技术不在于掌握庞大的数据资源，而在对巨量数据进行专业化处理，能够获得额外价值的技术。通常认为，大数据具有“5V”特征，地震数据并不是价值大密度低的数据，而是关联性较强的数据，需要通过地震射线和波动理论进行数据处理；另外，其他领域的大数据往往具有记录数总量巨大，而单条数据记录数量不大的特征。但是，如果因为海量地震数据与通常意义上的大数据的差异，就将大数据技术与地震数据处理技术割裂开来显然也是不具备说服力的。地震数据首先是数据，其次才是地震专业领域的数据。在地震处理过程中，观测系统描述数据、近地表模型的数据、初至的数据、速度谱数据和速度模型等数据也表现出了多样性的特点。同时，作为大数据技术支撑技术的分布式存储技术，主要是解决巨量数据存储和高效访问的问题；高性能计算技术是解决数据分析时效性的问题，这些技术在处理巨量数据方面与以往的技术相比有明显的优势，虽然目前尚没有成功的商业应用，但是可以在分析地震数据特点和处理过程数据特点的基础上，在现有大数据技术基础之上发展针对性技术是可行的。

“互联网+传统产业”的成功案例表明，所有的传统应用和服务都可以被互联网改变，没有被互联网改变的传统产业中一定意味着有商机，一旦引入了互联网就会产生新的格局。在寻找“互联网+地震技术”的过程中，首先地震数据量已经进入了海量数据

时代，地震数据处理需要性能更高的计算机系统，如此规模的计算机集群系统只有通过云计算技术才能实现随时随地的应用；其次在技术发展方面，就目前的情况而言，地震处理、解释软件作为地震数据处理技术的载体，由于技术难度和深度，目前存在明显的“寡头化”的趋势，而且这种趋势越来越明显。按照传统的技术研发思路，没有长期的技术积累，很难打破少数公司对技术和软件的垄断地位，只能采用基于互联网建立新的技术和软件研发模式才能打破现在的垄断局面；最后从目前的商业模式来看，地震处理解释软件均是采用购买软件许可，本地化部署方式进行应用。这种模式下，每个生产单位配备齐全的软件许可需要非常巨大的资金投入，这种模式可以基于互联网和云计算得到全面的颠覆。

勘探地震技术主要分为采集、处理、解释三个阶段。当然，这种划分并不是地震技术本身的内在要求，而是一定的技术发展阶段的产物，尤其是地震处理、解释的划分。由于前期技术发展的不平衡性和数据分析方面技术手段的不同要求，长期以来，地震数据处理和地震综合解释始终处于相互分离的状态。虽然地震处理解释一体化的设想已经提出了将近20年的时间，但是真正的一体化(或者说技术载体的一体化)进程是缓慢的。这既有软件技术方面的原因，也有当前业务流程方面的原因，更重要的是地震处理解释一体化过去只是一种方法论，或者说是一种理念，缺少支撑业务流程一体化的支撑体系。地震处理解释一体化本质上应该是根据地震技术在油气勘探开发中的作用，重新梳理处理、解释的技术流程，按照目标进行业务流程的重塑。在此方面，可以借鉴智能化数据分析技术的发展，形成新的数据分析技术系列。

地震采集、处理、解释的一体化是地震技术发展的必然趋势，就这个发展趋势而言，地震处理解释一体化是第一个阶段目标，只有在地震处理解释实现了一体化以后，才有可能实施地震采集处理解释的一体化。在地震处理解释一体化的过程中，重点是地震专业软件的一体化，通过地震处理解释一体化软件首先实现重塑地震处理解释业务和技术流程的目标，核心内容是地震处理解释一体化的软件平台。只有具备了地震处理解释一体化软件平台才能支撑地震处理解释业务和技术流程的一体化，以此为基础辅以“物联网”技术改变地震采集现状以全面提供地震采集、处理、解释一体化技术解决方案。

3 面向勘探地震技术的云计算与大数据

随着地震勘探和采集技术的发展,地震数据处理和其他行业一样进入到了“云计算和大数据”时代.研究人员纷纷投入到面向地震数据处理的“云计算和大数据技术”的研究中^[16~23].与此同时,相关技术也取得了长足的进展.例如,数据存储技术的发展以及存储成本的下降使得我们可以更加有效地存储各种类型的海量数据,云计算、高性能计算保证了地震数据处理的效率,数学领域的进展提供了更加丰富的处理手段,而类似于Google公司的Hadoop软件生态系统提供了一体化软件开发和应用平台.因此,尽管还处于起步阶段,但随着相关领域的发展以及不断飞涨的处理需求,云计算和大数据技术在地震技术中的应用已显现出蓬勃的生机和巨大的潜力.

目前云计算和大数据分析处理平台大多以Map Reduce架构和HDFS等分布式文件系统为基础,并辅之于各种数学、图像分析工具和可视化模块.典型的面向地震数据处理的云计算平台如图1所示^[24].在该架构中,底层是操作系统,可以通过虚拟机或硬件直接访问机制提供服务.操作系统之上是编译层和运行时系统,包括JRE, JDK, Scala和Python等几种选择.中间层是存储部件,如分布式文件系统HDFS、数据库Cassandra等,资源调度部件,如YARN, Mesos等,计算部件,如Spark, Hadoop等.当然现有的一些分析工具也不可或缺,如机器学习库Breeze, 信号分析工具FFTW, 图像处理工具OpenCV等.在这些基础层之上是定制的面向应用的云服务环境,可以提供分析处理、工作流、可视化等服务.最上层是用户开发的各种地震处理应用.

3.1 “大数据和云计算”技术在地震处理技术发展中的作用

大数据和云计算本身是一种技术发展战略,也是从技术到服务的新商业模式,可以看作“互联网+”商业运作模式的两个重要支柱.通过大数据基础技术的应用,可以支撑海量地震数据和复杂储层表征多学科的高效管理,同时为地震成像和复杂储层表征提供高效的数据吞吐能力;云计算技术,不仅可以实现超大规模计算集群资源的高效管理,而且可以随时随地提供高性能的计算资源以满足地震数据分析(地震成像和储层表征)的需要.

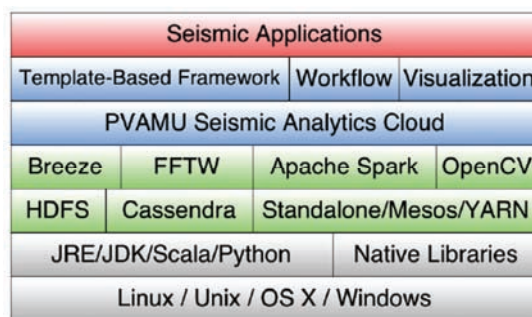


图1 面向地震数据分析与处理的云计算软件平台^[24]

Figure 1 The software stack of seismic analytics cloud platform^[24]

事实上,地震数据处理一直是高性能计算的重要应用领域之一.研究人员也一直利用MPI, Open MP和PGAS^[25]等实现高效的并行处理.然而,地震数据处理不仅仅是计算密集型任务,而是越来越朝着计算密集型和数据密集型的任务发展,同时对系统的扩展性、容错性甚至交互性都有了越来越高的要求.传统的高性能平台已不能满足要求. MapReduce架构的出现使得克服上述缺点成为可能,尤其是Hadoop生态系统的飞速发展吸引了大量的研究人员投入到相应的研究中. Seismic Hadoop (<https://github.com/cloudera/seismichadoop>)是其中的一个大胆的尝试,它基于Seismic Unix (<http://www.cwp.mines.edu/cwpcodes/index.html>),力图让用户能够更加方便地在Hadoop集群上进行偏移等地震数据的处理.还有研究人员在MapReduce架构上实现了传统的偏移等地震数据处理算法^[26,27].文献[28]利用近数据计算(near-data-processing)技术提出了一种新的逆时偏移算法.总体来说,地震数据处理是并行计算、高性能计算的重要应用领域.然而近年来由于基于Map Reduce的云计算框架具有编程模型简单、容错性强、扩展性好等优点,展现出很好的应用前景.随着Spark, Apache Flink (<http://flink.apache.org/index.html>)等内存计算模型的发展,以及HDFS分布式文件系统、NoSQL数据库、Redis (<http://www.redis.io>)数据结构的进步,基于云计算和大数据的地震数据处理技术将展现出更加灵活、高效的一面.

3.2 “大数据和云计算”技术在地震数据分析技术发展中的作用

地震数据分析是地震技术的重要组成部分,对计算能力、数据存储访问能力、分析算法等都有很高

的要求。近年来,大数据分析技术在互联网购物、多媒体处理等相关领域的应用中取得了很好的成果,其相关的并行计算(MapReduce, Spark和Storm)、人工智能、数据挖掘、深度学习等技术可以为重塑地震处理解释业务流程提供有力的技术支撑。文献[29]利用人工神经网络来预测沙体中的裂缝。文献[30]利用训练集的5种地震属性和储层厚度训练了一个支持向量机分类器,并用该分类器来预测储层厚度。文献[31]设计了一种后向传播神经网络来自动检测地震数据中的P波。

当然,如何将机器学习、神经网络等算法应用到云计算模式下,仍然是一个值得研究的课题,也具有相当的挑战性。Xing等人^[32]提出的Petuum是一个分布式的机器学习框架,可以方便人们实现自己的应用程序。Graphlab (<http://graphlab.org>)是另一个基于Python的分布式机器学习的应用框架。这些基于分布式的机器学习框架可以有效地帮助人们实现面向大数据的机器学习与数据挖掘算法,具有很好的应用价值。

3.3 通过大数据和云计算技术改变地震采集模式

利用“互联网+”时代的大数据技术和云计算技术,可以改变现有的地震技术研发和专业软件研发的模式。通过大数据技术的应用,可以建设能够适应未来发展的地震处理解释一体化软件平台,并以此为核心建立地震专业软件生态系统,以促进地震技术的可持续发展。另外,“互联网+”不仅仅局限在地震处理解释阶段,还可以依托云计算和物联网技术,实现地震采集技术的跨越发展。这种变革不是简单意义上的技术发展,而是以“消灭”地震仪为标志的全面技术变革。过去由于技术条件的限制,地震采集不得不采用专用设备(地震仪)进行地震数据采集的控制,目前通用计算机和云计算等技术的发展与互联网所具备的能力相结合,已经可以彻底改变(或替代)地震仪进行远程的地震采集过程控制。

如前所述,地震采集设备的布设具有其固有的“物联网”特征,基于地震一体化软件平台可以开发与地震处理解释一体化的地震采集控制软件,并且软件可以部署在云端(地震专业软件生态中),通过互联网与野外现场的地震检波器(可以支持无线和有线的连接方式)组成的局部“物联网”相连,就可以实现基于互联网的远程采集控制,再辅之于高效的地震

增量处理技术,能够实现地震资料的实时处理,并能够根据实时处理结果优化完善野外数据采集。这种模式将彻底改变地震采集的现状,并进一步实现地震采集、处理、解释的一体化,为油气勘探开发提供全面的地震技术支撑。Potapov等人^[33]提出了一种基于Google App引擎的地震数据采集与处理系统,该系统利用云计算技术可以对采集的信号进行有效的预处理,提高了计算精度和效率。

3.4 地震专业软件生态系统

生态系统的概念源于自然生态学,往往具有可持续健康发展的内在动力。软件生态系统是软件发展到“互联网+”时代的一个新生事物,是社会组织结构和软件产业链结合的产物。一个软件生态系统包括一套软件解决方案,同时也是一个“虚拟”的社会组织结构。在软件生态系统的组织结构中,既有生态系统维护者,又有生态系统参与者,其中维护者主要职责是维护系统稳定健康地运行,参与者能够支持和自动执行活动和交易。软件生态的优势在于迅速适应不断变化的业务需求并提供高度可靠性与伸缩性的应用服务。苹果AppStore就是一个成功的软件生态系统。软件生态系统相对于传统的软件产品线,最大的优势在于按需使用,可以避免传统的软件解决方案非常昂贵的费用。

地震专业软件生态将改变目前的地震专业软件需要购买许可本地化部署的应用模式。首先是基于云计算实现超大规模计算资源实现了集中化管理,其次是基于大数据技术的地震一体化软件平台,同时基于互联网展现软件生态系统的“无所不在(anyone, anywhere, anytime, anymore)”的特点。地震专业软件生态与通常的软件生态系统的共同点在于软件平台是开放的,任何人随时随地都可以进行技术和软件研发,并实现研发产品的在线交易和在线应用。不同点在于地震专业软件的App以功能插件(交互插件和批处理插件)为主,获得阶段性许可后,并不需要下载到终端设备,而是利用云端部署的超大规模计算和存储资源,在云端进行地震数据解析。基于地震专业软件生态系统,可以为技术和软件研发的参与者(开发用户)提供基于云计算的PaaS服务,能够支撑上述参与者高效的开发新技术和软件模块,同时基于类似于AppStore的模式,实现软件模块的快速应用。通过这种新的模式,能够聚合全球零散的物探研发

力量,而且可以吸引非物探领域(数学、图像等)的人员参与地震技术研发,快速实现新技术的软件集成,从而实现相对于当前技术垄断地震专业软件服务商的超越发展.基于地震专业软件生态系统,可以为从事地震资料处理和复杂储层表征的参与者(生产用户)提供基于云计算的SaaS服务.生产用户可以根据生产项目的实际需求,在项目运行阶段按需购买软件模块和计算资源的使用权,一旦项目完成,这些软硬件资源被收回,以达到有效降低成本的目的.

斯伦贝谢公司的Ocean应用商店是地震专业领域软件生态系统的一个尝试.Ocean应用商店采用了与苹果AppStore相似的策略,开发用户利用Petrel软件平台开发自己的应用模块,然后再在应用商店中进行交易.但是,Ocean应用商店对于地震专业领域而言,其专业覆盖面仅仅包括了地震解释和储层表征,尚未涉及地震处理,也不提供在线的云服务.此外,中国石化石油物探技术研究院也发布了自主开发的基于云计算和大数据技术的地震处理解释一体化平台“ π -Frame”.

4 结论

综上所述,“云计算和大数据”正在改变着传统产业的商业运作模式,地震专业技术领域籍此发生变革.通过大数据、云计算、移动互联网和物联网技术的应用,可以突破现有地球物理观测静态模式,推动动态观测;改变现有数据存储与共享模式,形成数据的实时交互、互联互通;改变现有地球物理数据孤立封闭的处理模式,向开放、融合、联合、实时在线的方向发展;改变现有地球物理装备与软件技术的小作坊共享模式,向基于云和大数据的生态系统过渡,实现以地震专业软件生态系统为核心内容的、以地震采集物联网架构为重要标志的“互联网+地震技术”,以彻底改变当前地震勘探的运作模式,不仅可以真正实现地震技术的一体化,提高地震采集的效率和投资效益,而且可以快速推进地震数据解析的智能化和自动化,提高地震资料的可信度,从而达到降低成本、缩短周期的目的,让地震数据资产在油气勘探开发中创造更大的效益.

参考文献

- Wang L Z, Laszewski G, Younge A, et al. Cloud computing: A perspective study. *New Gener Comput*, 2010, 28: 137–146
- Li G J, Cheng X Q. Research status and scientific thinking of Big Data (in Chinese). *Bull Chin Acad Sci*, 2012, 27: 647–657 [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. *中国科学院院刊*, 2012, 27: 647–657]
- Badger M L, Grance T, Patt-Corner R, et al. Cloud computing synopsis and recommendations. *Nist Spec Publ*, 2012, 800: 146
- Sun D W, Zhang G Y, Zheng W M. Big data stream computing: Technologies and instances (in Chinese). *J Softw*, 2014, 25: 839–862 [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: 839–862]
- Schmidhuber J. Deep learning in neural networks: An overview. *Neural Netw*, 2015, 61: 85–117
- LeCun Y, Bengio Y, Hinton G E. Deep learning. *Nature*, 2015, 521: 436–444
- Zhong X Q, Liu Z, Ding P P. Construction of knowledge base on hybrid reasoning and its application (in Chinese). *Chin J Comput*, 2012, 35: 761–766 [钟秀琴, 刘忠, 丁盘苹. 基于混合推理的知识库的构建及其应用研究. *计算机学报*, 2012, 35: 761–766]
- Chen L W, Feng Y S, Zhao D Y. Extracting relations from the Web via weakly supervised learning (in Chinese). *J Comput Res Dev*, 2013, 50: 1825–1835 [陈立玮, 冯岩松, 赵东岩. 基于弱监督学习的海量网络数据关系提取. *计算机研究与发展*, 2013, 50: 1825–1835]
- Keim D, Qu H, Ma K L. Big-Data visualization. *IEEE Comput Gr Appl*, 2013, 33: 20–21
- Olshannikova E, Ometov A, Koucheryavy Y, et al. Visualizing Big Data with augmented and virtual reality: Challenges and research agenda. *J Big Data*, 2015, 2: 1–27
- Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. *Commun ACM*, 2008, 51: 107–113
- Ghemawat S, Gobioff H, Leung S T. The Google file system. In: Michael L, ed. *Proceedings of the nineteenth ACM symposium on Operating systems principles (SOSP '03)*. New York: ACM, 2003, 37: 29–43
- Dean J, Ghemawat S. MapReduce: A flexible data processing tool. *Commun ACM*, 2010, 53: 72–77
- White T. *Hadoop: The Definitive Guide*. Sebastopol: O'Reilly Media, Inc., 2012
- Zaharia M, Chowdhury M, Franklin M, et al. Spark SI: Cluster computing with working sets. In: Nahum E, ed. *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud'10)*. Berkeley: USENIX Association, 2010. 10
- Anand P. Guest editorial: Big Data is a big deal. *J Pet Technol*, 2013, 65: 18–21
- Beckwith R. Managing Big Data: Cloud computing and co-location centers. *J Pet Technol*, 2011, 63: 42–45

- 18 Binotto A P D, Sultanum N, Cerqueira R. Cloud-based remote visualization of Big Data to subsurface exploration. In: Proceedings of SPE Large Scale Computing and Big Data Challenges in Reservoir Simulation Conference and Exhibition. Istanbul: SPE, 2014
- 19 Coles D, Prange M, Djikpesse M. Optimal survey design for big geodata. In: Proceedings of the Society of Exploration Geophysicists Conference. Denver: SEG, 2014
- 20 Crockett B, Kurrey K. Smart decision making needs automated analysis: Making sense out of Big Data in real-time. In: Proceedings of the Society of Petroleum Engineers (SPE) Intelligent Energy Conference and Exhibition. Utrecht: SPE, 2014
- 21 Feblowitz J. Analytics in oil and gas: The big deal about Big Data. In: Proceedings of the Society of Petroleum Engineers (SPE) Digital Energy Conference and Exhibition. Woodlands: SPE, 2013
- 22 Perrons R K, Jensen J. The unfinished revolution: What is missing from the E&P industry's move to "Big Data". *J Pet Technol*, 2014, 66: 20–22
- 23 Perrons R K, Jensen J W. Data as an asset: What the oil and gas sector can learn from other industries about "Big Data". *Energy Policy*, 2015, 81: 117–121
- 24 Yan Y, Huang L, Yi L. Is apache spark scalable to seismic data analytics and computations? In: Ho H, Ooi B C, Zaki M J, et al, eds. 2015 IEEE International Conference on Big Data. Santa Clara: IEEE, 2015. 2036–2045
- 25 Numrich R W, Reid J. Co-array Fortran for parallel programming. *SIGPLAN Fortran Forum*, 1998, 17: 1–31
- 26 Rizvandi N, Boloori A, Kamyabpour N, et al. Mapreduce implementation of Prestack Kirchhoff time migration (PKTM) on seismic data. In: 12th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT). Gwangju: IEEE, 2011. 86–91
- 27 Addair T, Dodge D, Walter W, et al. Large-scale seismic signal analysis with Hadoop. *Comput Geosci*, 2014, 66: 145–154
- 28 Perrone M, Liu L K, Lu L, et al. Reducing data movement costs: Scalable seismic imaging on blue gene. In: Parallel Distributed Processing Symposium (IPDPS), 2012 IEEE 26th International. Shanghai: IEEE, 2012. 320–329
- 29 Chaki S, Routray A, Mohanty W. A novel preprocessing scheme to improve the prediction of sand fraction from seismic attributes using neural networks. *IEEE J Sel Top*, 2015, 8: 1808–1820
- 30 Deng Y, Wang H. The support vector machines for predicting the reservoir thickness. In: The Eighth International Conference on Natural Computation (ICNC). Chongqing: IEEE, 2012. 118–120
- 31 Kaur K, Wadhwa M, Park E. Detection and identification of seismic p-waves using artificial neural networks. In: The 2013 International Joint Conference on Neural Networks (IJCNN). Dallas: IEEE, 2013. 1–6
- 32 Xing E P, Ho Q, Dai W, et al. Petuum: A new platform for distributed machine learning on Big Data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney: ACM, 2015. 1335–1344
- 33 Potapov V P, Oparin V N, Giniyatullina O L, et al. Cloud computing in seismic data processing based on Voronoi diagrams using Google App Engine. *J Min Sci*, 2015, 51: 1041–1048

Summary for “‘云计算与大数据’及其对勘探地震技术发展的影响”

The impact of “Cloud Computing and Big Data” on the future trends of seismic exploration

TANG Jie^{1,2}, WEI Jia^{3*} & WU GangShan^{1,2}

¹ State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China;

² Department of Computer Science and Technology, Nanjing University, Nanjing 210023, China;

³ Geophysical Research Institute, SINOPEC, Nanjing 211103, China

* Corresponding author, E-mail: weijia.swty@sinopec.com

Oil and gas exploration is one of the riskiest and most expensive of all commercial endeavors. Exploring for these oil and gas reservoirs is done by generating enormous amounts of seismic data and running that data through the computational ringer to reveal the most promising locations to drill. Data sets in this domain are among the largest in the industry, reaching into multiple petabytes, and which must be funneled from the storage subsystem to the compute nodes at tens of gigabytes per second. “Cloud Computing and Big Data” technologies describe a new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data (structured and unstructured) by enabling high-velocity capture, discovery as well as analysis. These technologies have made remarkable success in some traditional business areas. However, in the field of seismic exploration, the potential of “Cloud Computing and Big Data” technology has not yet been fully emancipated. There are many challenges. Data are locked in applications and cannot be shared efficiently. Data sets of this size compel the storage resources and network bandwidth. The processing workloads are a great challenge to the computing infrastructure and costs. Jobs may be random or sequential or both in an unpredictable pattern. Automatic or semi-automatic extraction of useful information for interpretation and visualization of seismic images is still unpractical. Furthermore, the workflow of oil exploration has formed into a processing chain, i.e. the output from one job becomes the input for another, covering from acquisition, processing to interpretation and visualization, which means that a software ecosystem is needed so that the newly acquired seismic data could be processed in real time to cope with the time demands and the workflow flexibility. To address those topics, this paper introduces the methodology and basic features of “Cloud Computing and Big Data”, analyzes the relationship between cloud computing, big data and Internet of Things, Internet + as well as other key technologies. Several existing Cloud Computing and Big Data systems are introduced. Furthermore, the technical characteristics of seismic acquisition, processing and interpretation are analyzed. The presented evidence shows that the integration of acquisition, processing, interpretation and visualization has been the trend of the seismic exploration. The paper then discusses the influence of “Cloud Computing and Big Data” on the evolution of seismic technology in three aspects. MapReduce framework as well as some success applications is presented to verify that Cloud Computing could handle the processing of tremendous huge and complex seismic data. Machine learning based on cloud computing is becoming a prosperous solution for seismic analysis to automatically extract more value from data. And the combination of Internet Things and Big Data is a good choice for seismic acquisition. Finally, the paper points out that building a software ecosystem of seismic processing and improving the technologies of seismic acquisition based on internet of things is the right choice to fully realize digital seismic exploration.

seismic exploration, Big Data, Cloud Computing, internet of things, internet+

doi: 10.1360/N972017-00158