廣東工業大學

课 程 设 计

课程名称_____高性能计算_____

学生学院_____计算机学院_____

专业班级_2013 级计算机科学与技术 2 班_

学　　号_____3113005816_____

学生姓名_____陈耿_____

联系电话_____18819471306_____

指导教师_____王卓薇_____

2016 年 12 月 10 日

利用云计算解决大量地理数据的挑战

Chaowei Yang *, Manzhu Yu, Fei Hu, Yongyao Jiang, Yun Li

美国，弗吉尼亚州，法尔法克斯市，乔治梅森大学，NSF 时空创新中心

Utilizing Cloud Computing to address big geospatial data challenges

Chaowei Yang *, Manzhu Yu, Fei Hu, Yongyao Jiang, Yun Li

NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA 22030, United States

Abstract Big Data has emerged with new opportunities for research, development, innovation and business. It is characterized by the so-called four Vs: volume, velocity, veracity and variety and may bring significant value through the processing of Big Data. The transformation of Big Data's 4 Vs into the 5th (value) is a grand challenge for processing capacity. Cloud Computing has emerged as a new paradigm to provide computing as a utility service for addressing different processing needs with a) on demand services, b) pooled resources, c) elasticity, d) broad band access and e) measured services. The utility of delivering computing capability fosters a potential solution for the transformation of Big Data's 4 Vs into the 5th (value). This paper investigates how Cloud Computing can be utilized to address Big Data challenges to enable such transformation. We introduce and review four geospatial scientific examples, including climate studies, geospatial knowledge mining, land cover simulation, and dust storm modelling. The method is presented in a tabular framework as a guidance to leverage Cloud Computing for Big Data solutions. It is demostrated throught the four examples that the framework method supports the life cycle of Big Data processing, including management, access, mining analytics, simulation and forecasting. This tabular framework can also be referred as a guidance to develop potential solutions for other big geospatial data challenges and initiatives, such as smart cities.

Key words Big Data;Cloud Computing;Spatiotemporal data;Geospatial science;Smart cities

摘要 大数据在研究、开发、创新和商业等方面已经出现了新的机会，它具有的四个特点可以用四个以英文字母 V 开头的词概括:数量大(volume)、速度快(velocity)、准确性高(veracity)和多样性（variety)。通过大数据的处理，可能会带来显着的价值。大数据从以上四个特点转化为第五个以 V 开头的词——“价值（value)”的过程，对计算机的处理能力来说是巨大的挑战。云计算已经形成一个新的模式，来提供大数据运算过程中的各种服务，以满足计算过程中不同需求，包括 a)服务的需求；b)整合集中的资源；c)灵活性；d)网络（宽带）的接入；e)可估量的服务，云计算可传递的计算能力促进了大数据向价值的转换。本文探讨了云计算如何被用于解决大数据向价值的挑战，我们列举了四个地理空间科学的例子，这四个例子包括了气候的研究、地理空间知识的挖掘、土地覆盖的模拟以及沙尘暴的建模。使用的方法是以表格的框架作为一个指导,并利用云计算的优势处理大数据的解决方案。通过这四个例子，说明了该方法在大数据处理的整个流程，包括管理、获取、挖掘分析、模拟和预测过程中的作用。这种表格的框架可以作为开发其他地理空间大数据的解决方案的引导,例如智慧城市。

关键词 大数据；云计算；时间与空间数据；地理空间科学；智慧城市

1. 引言

对地球的观测和模型的模拟每天产生了千兆比特的数据（Yang, Raskin, Goodchild, and Gahegan, 2010），非传统的地理空间数据获取方法，例如通过社交媒体（Romero, Galuba, Asur, and Huberman, 2011）、电话交

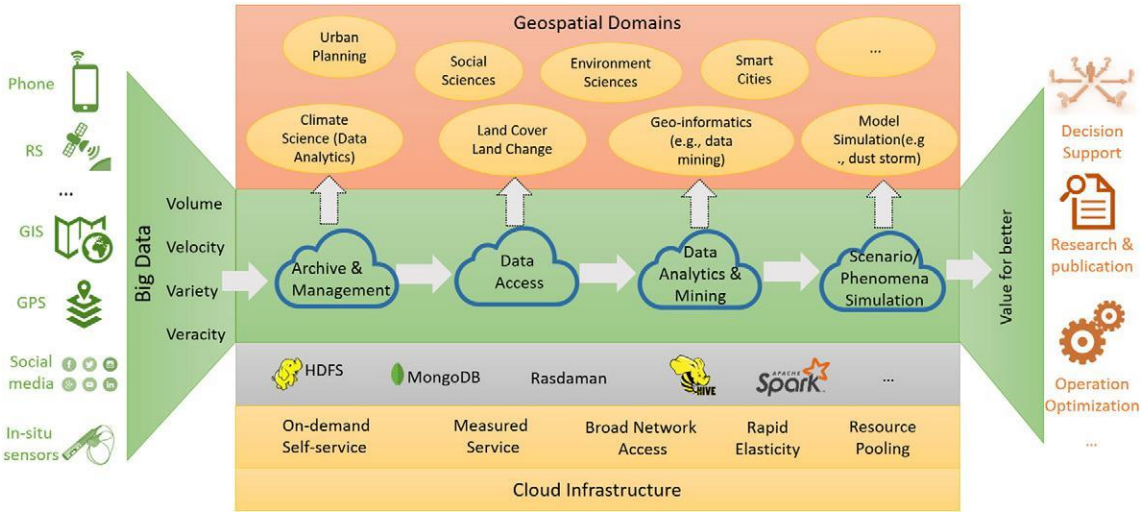


Fig. 1. Cloud Computing provides critical supports to the processing of Big Data to address the 4Vs to obtain Value for better decision support, research, and operations for various geospatial domains.

图 1.云计算为大数据的处理提供关键支持，以转化 4 个特点获得价值，从而为各种地理空间域提供更好的决策支持，研究和运营。

流（Frias-Martinez, Virseda, Rubio, and Frias-Martinez, 2010）和无人机（Einav and Levin, 2013）获取产生地理空间数据的速度更快，除此之外，地理空间数据（Marr,2015; Hsu, Slagter, and Chung, 2015）存在于各种各样的表格并以各种形式存在于不同的应用程序中，它们的精确度和不确定性被跨越了被真实性定义的范围。而且在实时的传感器中，数据高速产生（如图 1）。随着知识了解的深入和空前的信息，这些地理空间的大数据可以被处理以增加科学研究、工程开发和商业决策方面的价值（Lee and Kang, 2015）。他们设想通过完成大数据四个特点到价值的转换，来改善我们的生活和对地球的理解（(MayerSchönberger and Cukier, 2013）。

这种转变对数据管理和访问、分析与挖掘、系统架构和模拟等方面带来了巨大的挑战（(Yang, Huang, Li, Liu, and Hu, 2016)）。举例来说，面对的第一个挑战是如何处理大数据以获得能被用于单一决策支持系统的融合数据（Kim, Trimi, and Chung, 2014），另一个挑战是如何处理由于数据量波动带来的，大数据处理能力的可拓展性（Ammn and Irfanuddin,2013）。及时的数据分析功能也在创造价值方面带来了巨大的挑战（Fan and Liu, 2013; Chen and Zhang, 2014; Jagadish et al., 2014）。

云计算在作为服务提供计算能力方面已经形成一个新的模式，该模式具有以下五个方面的优势（如图 1 最底下两层）：a）快速且具有弹性地提供计算能力；b）聚合的计算能力以更好地利用和分享资源；c）为了快速沟通的宽带接入；d）作为服务的根据需求的运算接入；e）根据使用付费而不像传统运算资源一样预付费（Yang, Xu, and Nebert, 2013）。云计算采用了面向服务的架构，使“一切即服务”，包括基础设施即服务（IaaS）、平台即服务（PaaS）以及软件即服务（SaaS）（Mell and Grance, 2011）。云计算在重新开启地理科学和数字地球的可能性的同时，也启发了在多个地理科学及相关领域解决地理空间大数据问题的解决方案。

但是，利用云计算处理大数据现在仍处在初级阶段，而且如何从大数据从那四个特点转换为价值是一项艰巨的任务（如图 1）。本文通过四个科学的例子阐述了云计算是如何支持这个转变的过程，这四个例子包括了对气候的研究、知识的挖掘，土地的覆盖和利用的变化的分析以及沙尘暴的模拟。这四个例子都是极具代表性的，可以被吸纳到其他环境以及城市研究的领域，例如智慧城市方面（Batty, 2013; Mitton, Papavassiliou, Puliafito, and Trivedi, 2012; Odendaal, 2003）。地理空间大数据的生命周期（数据管理与获

取、分析/挖掘、现象/场景模拟）在这四个例子中得到充分的体现，并且细节都体现在每个部分中（如表格 1）。举例来说，例子 2.1 填充了按需自助式服务和表格 1 的交叉处，这以为 2.1 详细展示了多大量（的气候大数据）可以解决云计算的按需自助式服务。

2. 利用云计算支持气候分析

越来越大的水灾和洪涝等相关的气候变化，越来越多地影响城市的基础设施（ Rosenzweig, Solecki, Hammer,and Mehrotra, 2011），同时在过去 50 年里，人类的活动（如化石燃料的燃烧）越来越严重地影响了地球环境（Bulkeley and Betsill, 2005）。为了了解气候变化及其对环境和城市问题的影响，在过去观察到的大气候数据和未来的模拟应该得到很好的管理和分析。然而，观测和模拟这两方面都会产生大数据。例如，未来 IPCC 的报告将基于一百多 PB 字节的数据，美国宇航局将 2030 产生三百多 PB 的气候数据（skytland，2012），这些数据在格式、精确度以及研究对象方面各不相同（Schnase et al., 2014）。大数据有助于推进气候现象的理解，并且可以帮助确定如何补救气候变化对社会和生态系统的影响，如在人口稠密的地区全球温度异常检测和调查的极端天气气候事件的时空分布（如城市）（Das and Parthasarathy, 2009; Debbage and Shepherd, 2015）。

Table 1 The Big Data challenges as illustrated in the four examples are addressed by relevant cloud advantages to reach the Big Data Value and achieve the research, engineering and applica-tion objectives.

表格 1 如四个例子所示的大数据挑战通过相关云优势解决，以达到大数据价值并实现研究，工程和应用目标。

	On-demand Self-service	Broad network access	Resource pooling	Rapid elasticity	Measured service
Volume	2.1	4.1	2.1	2.1, 3.1, 3.2, 4.1, 4.2, 4.3,5.1	4.1
Veracity	2.1	3.1, 5.3			
Velocity			2.1	4.1	4.3
Variety		3.1, 5.2	2.1		
Value	2.1, 3.2			2.1	3.2

在利用大数据方面面对几个挑战:a）气候大数据的容量和产生速度已经远超许多独立计算机的存储和运算能力；b）气候大数据在格式和精度方面的多样性让人们很难找到一个简单易用的工具去分析它；c）对于很多气候科学家来说，建模的准确性设计到建模时的不确定性和混合模型的质量（Murphy et al., 2004）。这些由于数据量、产生速度、多样性和精度带来的问题可以被解决，解决方法是通过基于云的、高效的数据管理策略以及面向服务的数据分析架构，用于分析并挖掘这些气候数据。

2.1 气候大数据管理中的高效索引

成百上千 PB 的气候大数据只能被储存在分布式且可拓展的环境中，云计算可以从以下几个方面帮助管理：a）根据气候的数据量按需灵活地配置虚拟机（VM）；b）在虚拟机上自动部署并建立 HDFS、Hadoop 的分布式文件系统。气候数据可以以原生的格式保存，而无需为了节省存储空间二进行序列化。而且，逻辑数据的结构也被建立，以方便快速对数据进行识别、访问和分析（Li, Hu et al., 2016; Li, Yang et al., 2016），其中最核心的架构是一个对存储在 HDFS 上的多维的气候数据进行时空索引（Li, Hu et al., 2016; Li, Yang et al., 2016）。索引对 HDFS 上的数据内容进行字节级、文件级以及节点级的进行引导。九个特征被用于这些索引，其中包括大小、时间、状态信息用于描述数据的网状逻辑结构，这能与数据查询自己偏移量、字节长度、压缩编码、节点列表、文件路径等以识别在 HDFS 上的具体位置。这个索引让用户可以通过精确的时空信息和内容的描述，准确定位并访问这些数据。

从细节上看，时空索引的空间和时间属性能区分网格与时空包围盒重叠，节点列表的属性会被修正后出数到程序中的节点上，在那里这些数据将被保存下来。根据字节偏移量、自己长度和压缩编码等属性，计算机能以告诉的数据流的形式访问这些数据，数据的状态和属性将被用于将这些数据流重新转化为多维数组。

近 26 年来每个月的 MERRA 数据 MAIMNXINT1（大约 90GB）被用来评估基于云计算的时空索引的大数据管理效率。这

个实验分析了特定变量的每月平均值（通过改变他们的数字），这些数据来自特定气候范围内的以 1Gbps 连接的、基于 36 个虚拟机的 HDFS 集群，其中每个节点都配备了 8 个 CPU 核心（主频为 2.60GHz）、16GB 内存和 CentOS 6.5。结果显示了是否具有索引对处理时间造成的影响（如图 2），当不具有索引时，运行时间的倍数为 9.1，当具备索引时倍数仅为 1.8。由于受到时间的限制，在指定时间内完成任务准备的虚拟机数量不确定（Li, Hu et al., 2016; Li, Yang et al., 2016; Yang et al., 2015）。因此，按需的服务与弹性的计算性能相结合，具有高层次的管理效率，可以满足气候大数据的管理和分析需求。

2.2 一切皆服务以适应气候建模实验

当根据不同模型的输入分析模拟大数据或运行大量模型模拟时，气候模拟带来的挑战为科学实验获得了足够的计算资源，云计算满足了实验的一下方面：a）气候模型可以被当成服务（MaaS; Li et al., 2014），且足够多的虚拟机能被配置完备以满足特定模型的需求；b）应用程序可以被部署成为一种服务（Lushbough, Gnimpieba, and Dooley, 2015），且这种服务具有网络接口以便对模型进行操作和监控；c）具有不同分析的工作流拥有了直观的图形用户界面作为服务。云计算在基础设施层面上支持了气候大数据的分析工作。

为气候模型研究设计的基于云计算的面向服务的工作流系统的组织架构如下（图 3）：a）模型服务对虚拟机上的运算和运行的模型负责，这提供基于系统中当前包含建模软件环境下运行的一个模型的快照；b）虚拟机监控服务想云平台提供了包含资源状态的虚拟机状态信息；c）并行计算解决了密集型计算的问题，数据分析服务供给模型的输出作为分析的输入。数据发布服务使用户能够通过互联网实时访问分析结果。所有的这些服务都能通过图形用户界面控制，它允许用户拖动并将服务连接到一起以构建一个复杂的工作流，然后系统就可以自动过渡到到工作流指定的应用程序并在云上自动配置和运行虚拟机。举个例子，Li（2015）建造了一个 ModellIE 的服务以研究这个模型的敏感性，实验结果表明，与传统方法相比，基于云计算的运算方法缩短了 10 倍的运行时间。

云计算所解决的气候研究中的挑战总结在表 1。首先，从观测和模拟获得的气候大数据存储在分布式和可扩展的云平台环境配置（2.1）；其次，气候数据方面面临的多种挑战是通过建立时空索引解决的，这可以同意他们的空间及时间；第三，在气候模型的各种挑战是通过建立面向服务的系统，以简化模型的配置、运行和输出分析（2.2）。这些方法可以被拓展到其他包含多维数据和复杂模型的地理空间领域，例如遥感、图像处理和基于代理的环境和城市建模事件。

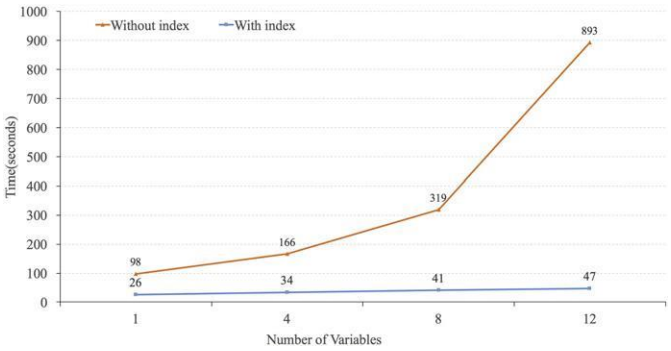


Fig. 2. Run time of the daily global mean calculation for different numbers of variables.

图 2　不同变量数的每日全局平均值计算的运行时间

3. 支持地理空间大数据的知识挖掘

我们已经收集了通过不同方法采集的拥有不同时空标志，用于环境和城市研究的地理空间的大数据，例如全球卫星定位（GPS）、遥感和网上的志愿者（Jiang and Thill, 2015; Yang et al., 2011）。在体积，获取速度和各种的时空数据的增量给研究人员带来了一个巨大的挑战，研究人员很难发现和访问用于研究和支持决策的正确数据（Yang et al., 2011）。其中的一个方法是通过从这些地理空间大数据和用处中发掘其中用于扩展查询、推荐和排序的知识。所挖掘

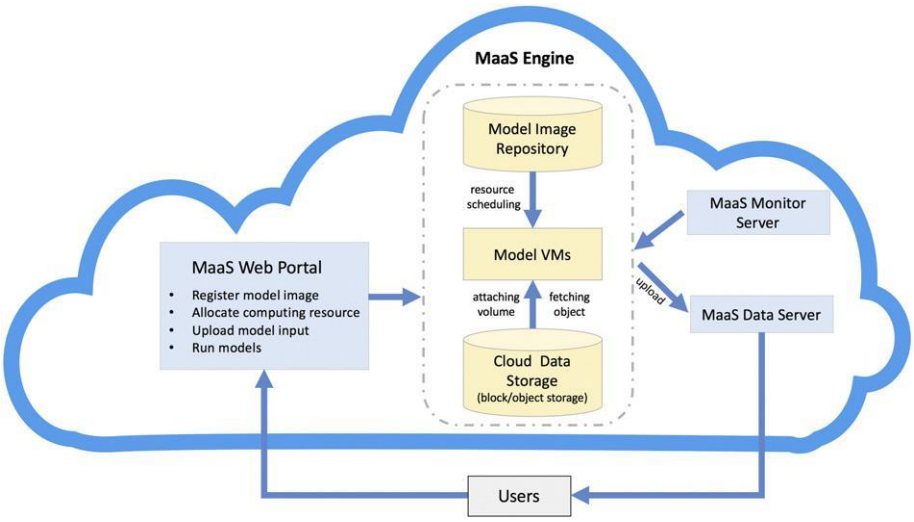


Fig. 3. The cloud-based service-oriented workflow system for climate model study.

图 3 基于云的面向服务的工作流系统气候模型研究

的知识包括，但不限于领域的热点话题，研究趋势，元数据链接和地理空间词汇的相似性。这个过程面临着巨大的数据量，产生速度和种类的挑战。这样一个挖掘的过程面临着两个挑战：a）如何将大数据转换成可被计算资源处理的并行数据块；b）如何利用可拓展的计算资源处理这些大数据。举 NASA 海洋物理分布式主动归档中心（PO.DAAC）的 MUDROD 项目为例子，2014 年的网络日志（包含了地理数据使用见闻）超过了 1 亿 5000 万条数据，挖掘任务在一台服务器上（6 核、12G 内存，使用 Windows 7 操作系统）运行了超过 5 小时。对于一些同时很多用户发送请求的高流量网站，日志以更高的速度生成，这个速度超过了单台服务器的数据处理能力，为了适应这种情况，日志以半结构化或非结构化的方式存储在不同的格式中（例如阿帕奇公司的 HTTP、FTP、NGINX、IIS 日志格式或用户定义的格式）。每一种格式都需要一个特定的处理协议以进行进一步的处理，这种不确定性以噪声数据（例如网络爬虫）的方式影响了所挖掘的数据，这要求了精确的爬虫探测算法以处理原始数据（Jiang et at.,2016）。

3.1 通过数据并行加速用户日志的挖掘工作

处理这些日志大数据的第一步是根据数据量和时间限制，将数据引导到进行相同操作的动态调整的虚拟机集群上（Gordon, Thies, and Amarasinghe, 2006）。为了将原始的日志分拆到相同数量的虚拟机集群上，一般来说可以通过两种方法加速这个过程，一个是基于时间的拆分，另一个是基于 IP 地址的拆分。在基于时间的拆分方法中（如图 4a），梁琳日期的日志被分到同一个文件中。一旦原始的日志被分到 k 个文件中（k 是虚拟机集群的数量），每个日志文件的综合被最小化了，这种拆分过程被以线性拆分问题解决了(Skiena, 1998)。在基于 IP 的拆分方法中（图 4b），日志文件中相同 IP 的日志采

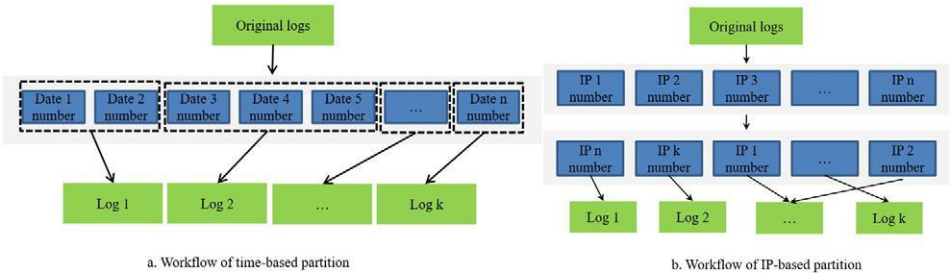


Fig. 4. The workflow of time-based partition (a) and IP-based partition (b).

图 4 基于时间的分区（a）和基于 IP 的分区（b）的工作流

用贪心算法归入同一个文件中，与基于时间的拆分方法不同的是，IP 地址种类的变化是允许的。

一旦这些原始的日志被拆分到一个虚拟集群中，繁多的日志文件在所有虚拟机中被并行处理，日志挖掘的速度显著提高。

3.2 在虚拟集群中提供按需计算资源

相比于手动建立一个集群，云计算促进了具有动态数量的虚拟机虚拟主机集群（Krämer and Senner, 2015），更多的计算机资源可以被部署用于处理历史数据，而动态数量的虚拟机可以被用于处理实时的数据流。按需分配的计算资源对于满足数据量变化的日志数据要求来说很有必要。举例来说，在 2014 年一月，PO. DAAC 的日志挖掘任务采用了具有更多虚拟机的集群，因此完成任务所花的时间更少了（如图 5a），基于时间的拆分和基于 IP 的拆分都加速了挖掘任务的进程。但是，基于时间的拆分改变了处理器产生的会话（图 5b）。

在整个 PO. DAAC 日志处理中，总共的处理时间从 190 分钟减少为 49 分钟，减少了 70%，与此同时，虚拟机的数量从 1 台提升为 4 台（图 6）。

就像地理空间数据使用日志，地理空间数据能被一个集群进行并行化处理。地理数据能根据不同的分类，例如经度、纬度、时间或文件大小，而被分成更小的部分，然后进入虚拟机中进行并行化处理。

如表 1 所体现的，广泛的网络接入和快速的弹性使使数据的并行化方法能够高效率地切分大数据，并使这些数据接下来能够被并行处理（3.1）。按需的自助服务、精确计算的服务以及能在短时间快速添加或溢出计算结点满足了动态匀速的要求（3.2）。

建议的网络日志知识挖掘方法能够与域名数据接口相结合，以帮助环境科学家和城市科学家快速发现有用的信息和知识。应该指出的是，针对用户特定的分析（知识）数据也可能构成隐私和安全问题。在城市的研究中，空间数据的挖掘和地理知识的发现已成为近年来的活跃的研究领域。GPS 数据，高分辨率的遥感数据和网络自发地理信息被收集用于提取未知或意外的信息（Mennis and Guo, 2009; Jiang and Thill, 2015）。这些数据集具有前所未有的巨大的数量，数据并行化算法可以被用于促进云计算进行高效地处理这些数据，例如用于在智慧城市中分析住房相关的工作（Long and Thill, 2015）。

4. 支持土地利用和土地覆盖变化的分析

土地利用和土地覆盖的变化（LULCC）已经成为环境变化和可持续发展研究的一个基本组成部分。Landsat 已经独立产生了 6PB 数据（Turner et al., 2003; Hansen and Loveland, 2012），土地变化的监控、评估和预测（LCMAP）对科研级别的实时和伪实时的地球观测土地变化产品产生了需求（Dwyer, 2014），但这面临了几个大数据方面的挑战：a）储存、访问和分享土地使用大数据的挑战；b）采用大量训练集和复杂算法的土地覆盖变化情况快速建模；c）快速的变化分析以及对土地变化数据的预测。

4.1 在云上储存、分享及分析土地覆盖

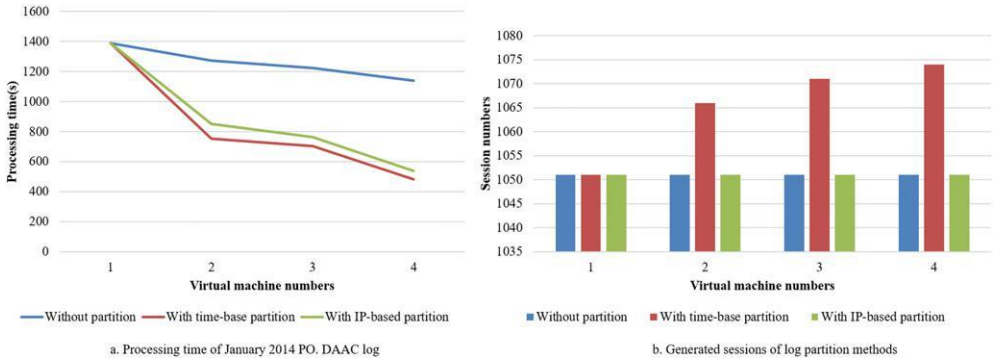


Fig. 5. Processing time of January 2014 PO. DAAC log (a) and generated sessions of log partition methods (b).

图 5 2014 年 1 月 PO. DAAC 日志(a)和生成的日志分区会话(b)的处理时间

大数据

PB 级的土地覆盖历史数据和 TB 级的土地覆盖流数据要求了要昂贵的、内部部署的硬件，然而这些是很难维护和管理的，而云储存是外包给第三方的云提供商进行升级及维护（USGS, 2016）。除此之外，云储存支持通过一个简单的网络服务接口即时访问，而且因为冗余设计和数据分发具有高可靠性，而且是根据使用付费（Calder et al., 2011）。作为持续时间最久的连续从太空中观察的地球土地数据，地球资源卫星数据在 2015 年后可以通过亚马逊 S3 访问。大多数从 2015 年起的地球资源卫星的影像是可用的，。所有新的 8 个地球资源卫星场景每天都是可用的，而且处理的时间经常不用几个小时（AWS， 2015）。另外，除了数据的可访问性增强，储存在云裳的陆地覆盖影像和发布在云上的覆盖模型结合在一起，以减少土地变化研究的工作流、结果分享和在线。举例来说，ARCGISt 在线允许快速在 AWS 的地球资源卫星数据的快速可视化和分析，利用地球资源卫星在 AWS MAPbox 功率的卫星直播，基于浏览器的地图，不断刷新着从 Landsat 8 卫星的最新图像（AWS，2015）。

4.2 利用大量训练集和复杂算法快速建模

在三种类型的土地变化建模，即图像分类、土地利用适宜性，与土地覆盖变化对环境的影响（Eastman, 2012），算法复杂且通常涉及大量的训练集建立一个健壮的模型。然而，大多数问题可以转换为通用的数据挖掘问题。例如，一个流行的 GIS 土地变化的建模工具，用于对开发的土地变化模型是基于

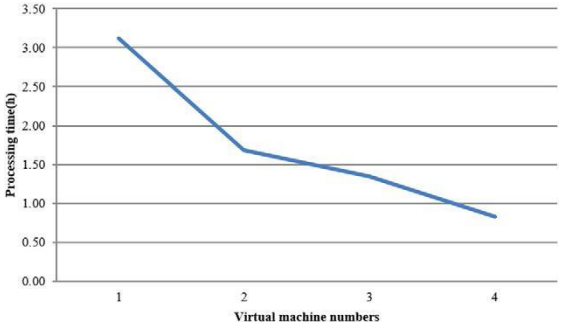


Fig. 6. Processing time of 2014 PO. DAAC log.

图 6 2014 年 PO. DAAC 日志的处理时间

逻辑回归和神经网络（Eastman, 2003）。这些数据的并行化挖掘算法是在云计算社区中被深入研究的，而且采用开源的、大型的处理框架，例如 Spark MLlib （Meng et al., 2016）。利用这些技术，一种中间件被开发出来用于将这些训练集转化成现有技术可以消化和转化的格式，这种格式是土地变化分析所要求的（如图 7）。

4.3 利用土地变化大数据进行快速分析和预测

一个分级或预测的模型可以通过传统的方法或在 4.2 中提到的方法进行提高，如果每个图片和像素都通过在土地变化模型中顺序计算，那这个过程会依旧是计算密集型的。通过以下几个步骤，可以通过虚拟集群加速这个过程：a）将研究区域并行化至小区域中；b）将土地变化数据分配到分析程序并行运行的虚拟机上；c）将运行结果正和岛一个结果集上。举个例子，一系列高分辨率的全球森林覆盖变化图在谷歌地球引擎（Hansen et al., 2013）通过其内在的并行访问谷歌云（Moore, 2015）表明，利用云计算平台，可以促进大型图像土地覆盖分类的可能性。

高速的网络，快速弹性和测量服务改进了大型 LULCC 数据的存储，访问和分析，以应对数据量和速度挑战（如表 1,4.1）。快速弹性允许大型数据处理框架中间件支持大型训练集和复杂算法的快速建模（4.2）。快速弹性和测量服务也使得所提出的并行计算框架可以提供近实时分类、土地覆盖变化和预测地图成为可能（4.3）。地球变化中提出的解决方案也可用于其他科学问题，如气候变化，生态系统服务和栖息地以及生物多样性模型。

5. 支持沙尘暴的预测

灰尘风暴对全球，特别是城市地区的健康，财产和环境造成严重危害（Knippertz 和 Stuut，2014; WMO，2011）。在沙尘暴期间和之后，由于能见度迅速下降，交通事故发生率增加；当灰尘颗粒保持悬浮在大气中时，空气质量和人体健康受到损害；当灰尘干扰时，

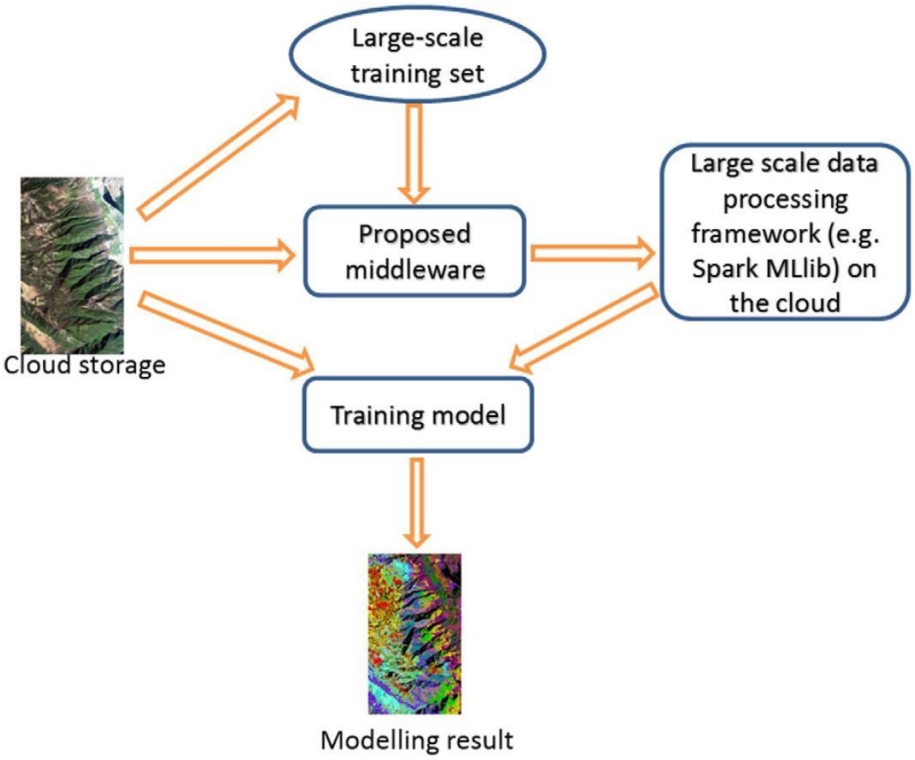


Fig. 7. The role the proposed middleware plays in the model building process

图 7 中间件在建模过程中的角色

可再生能源的效率降低（Wilkening, Barrie, and Engle, 2000）。因此，精确地预测即将发生的沙尘暴以减轻沙尘暴的环境、健康和其他财产的影响是至关重要的（Benedetti et al., 2014）。这种预测要求的标准要求是在两小时的计算时间内模拟一天天气情况（Xie, Yang, Zhou, and Huang, 2010），分辨率不高的（1/3 度）沙尘暴模型预测是很容易实现的，美国西南部使用单个 CPU，需要约 4.5 小时完成处理。对于高分辨率模拟（例如 3km×3km），模型输出数据的体积从 100GB 增加到 10TB。计算时间在三个维度（纬度，经度和时间）中的每一个中增加 4 倍。这导致整个系统增加 64（4×4×4 = 64）即 12 天以完成处理。将 12 天减少到 2 小时的这一挑战是如何处理大数据处理/计算，如何从地理，大气和生态系统数据中获取各种内容输入以及如何通过输入高精度的数据而提高预测模拟的真实性。

5.1 加速大数据运算和处理

为了应对将计算时间从 12 天减少到 2 小时的挑战，Huang、Yang、Benedict、Rezgui et al（2013）和 Huang、Yang、Benedict、Chen et al（2013）提出了一种自适应松散耦合模型策略，即将高分辨率/小尺度灰尘模型与低分辨率/大尺度模型进行链接。该策略运行低分辨率模型并识别具有预测的高灰尘浓度的子域作为敏感区域（AOIs）（如图 9a），这些敏感区域的较高分辨率模型被并行执行。在云计算的支持下，用于特定 AOI 的高分辨率模型运行的群集被并行地快速建立，并且比在整个域上执行高分辨率模型更高效地完成。当云计算并行处理所有敏感区域时，不同敏感区域所需的执行时间不到 2.7 小时（图 9b）。

5.2 输入大量种类的沙尘模型

随着灰尘预报模型的时空分辨率的增加，面临的挑战是获取具有不同格式、内容和不确定性的动态数据（Yang et al., 2011）。云计算广泛的网络访问的能力可以服务于具有高级网络带宽和可扩展性的、具有更多种类的模型输入数据的模型进行访问和预处理。 Huang、Yang、Benedict、Chen 等（2013）和 Huang、Yang、Benedict、Rezgui

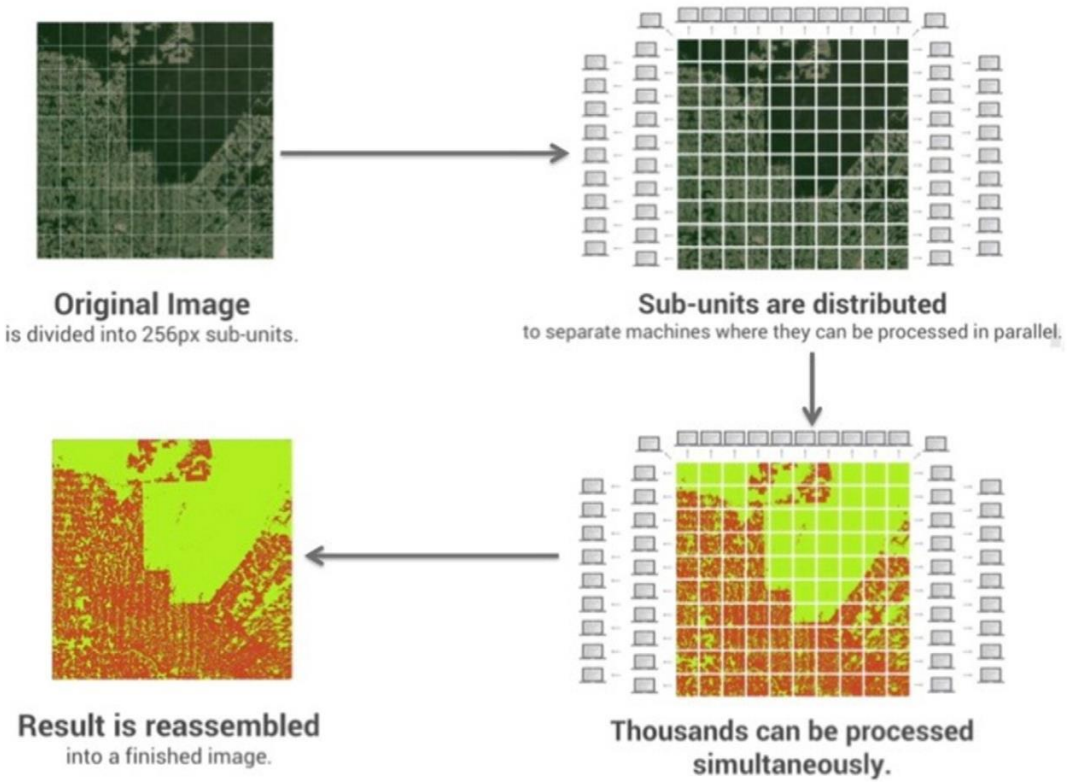


Fig. 8. Google Earth Engine divides Big Data to process in parallel using multiple computers (Moore, 2015).

图 8 谷歌地球引擎将大数据划分为使用多台计算机并行处理(Moore, 2015)

et al 等（2013）表明，与比 HPC 集群想比，亚马逊云实例可以在 更短的时间内完成大多数预测任务（图 10），表明云计算有潜力解决计算需求应用程序的并发强度。

5.3 提高沙尘预报的可靠性

影响模型输出真实性的最重要因素之一是模型初始条件的不确定性（Lin, Zhu, Wang,2008）。这些不确定性可以通过使用各

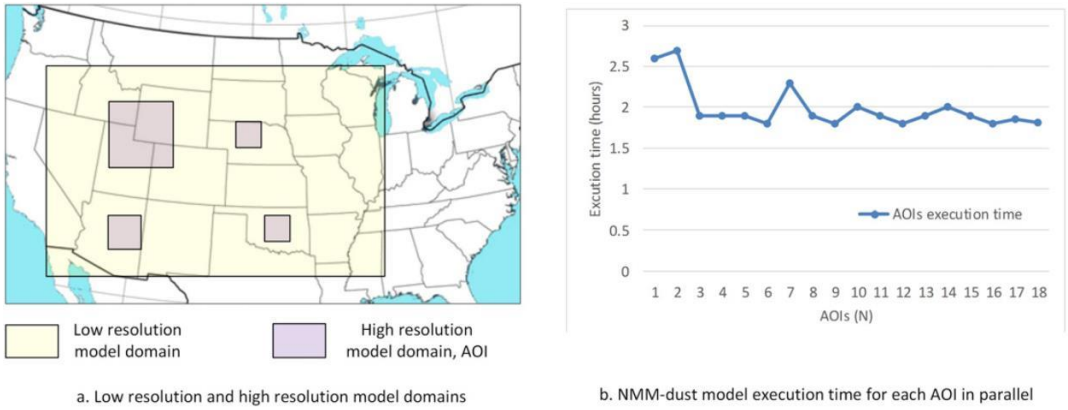


Fig. 9. Low-resolution model domain area and sub-regions (Area of Interests, AOIs) identified for high-resolution model execution (Huang, Yang, Benedict, Rezgui et al., 2013; Huang, Yang,
Benedict, Chen et al., 2013).

图 9 针对高分辨率模型执行确定的低分辨率模型域区域和子区域(Huang, Yang, Benedict, Rezgui et al., 2013; Huang, Yang,
Benedict, Chen et al., 2013)

种模型变量的敏感性测试来调查和表征（Zhao et al.,2010; Liu et al., 2012）。为了减少初始条件的不确定性，可以将数据同化技术应用于沙尘模型，通过将观测值同化到模型中以校正模型初始条件（Niu et al., 2008; Sekiyama, Tanaka, Shimizu, and Miyoshi,2010; Liu et al., 2011）。随着数据源的多样化，灵敏度测试和数据同化技术可以通过最小化的预处理和集成到模型中进行，从而使得能够努力提高模型精度，并最终减少模型不确定性（Lin et al., 2008; Darmenova, Sokolik, Shao, Marticorena, and Bergametti, 2009）。整个复杂的过程可以精确保存在虚拟机映像中，可以最小化工作量，并减少未来的手动错误。

因此，通过将云计算的特征与加速沙尘预报任务的净效应（表 1，5.1）相结合，可以解决大规模科学预测的挑战。 通过广泛的网络访问，实现了更多种类的输入数据的获取，并且在云上预处理而不消耗指定用于模型模拟的核心的计算资源（5.2）。 模型输入数据的选择更复杂，改进了模型的初始条件的表示和模型的模拟输出的可能的数据真实性（5.3）。 这些方法很容易适用于需要在短时间内得到结果的其他科学计算或模拟模型，包括洪水，飓风和空气污染的预测。

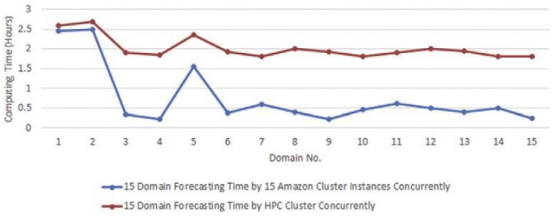


Fig. 10. NMM-dust execution time for 15 forecasting tasks on Amazon EC2 and HPC cluster
(Huang, Yang, Benedict, Chen et al., 2013; Huang, Yang, Benedict, Rezgui et al., 2013)

图 10 NMM-dust 在 Amazon EC2 和 HPC 群集上执行 15 个预测任务的时间(Huang, Yang, Benedict, Chen et al., 2013; Huang, Yang, Benedict, Rezgui et al., 2013)

6. 结论

空间数据大数据在数据存储、访问、管理、分析、挖掘和建模的生命周期中构成了巨大的挑战。 这四个例子说明了云计算利用五个云计算的按需自助服务，广泛的网络访问，资源池，快速弹性和测量服务（见表 1）的优势解决四个“V”挑战达到价值的能力。包含部分编号的框表示这些部分利用云计算的功能来解决大型地理空间数据中的相关挑战。 表 1 作为评估其他大地理空间数据挑战的解决方案的指南也有价值。

虽然已经开展了利用云计算解决大地理空间数据挑战的研究，但仍有许多挑战需要解决：

地理空间大数据存储和管理仍然是优先的，包括如何在云环境中优化不同传统（例如 MySQL，PostgreSQL）和新兴的数据库管理系统（例如 NoSQL，HDFS，SPARK，HIVE）和分析（Agrawal，Das 和 El Abbadi，2011）

时空大数据挖掘需要实时数据处理、信息提取和自动化来提取信息和知识。应该开发更多可扩展的时空挖掘方法（Vatsavai et al., 2012），以利用云平台的弹性存储和计算资源（Triguero, Peralta, Bacardit, García, and Herrera, 2015）。

安全性是确保对敏感数据和用户隐私保护的挑战。科学家需要更多的研究来跟踪和维护被信任的信息，以识别和防止对云平台的攻击（Manuel, 2015）。

云平台上的使用行为（例如，何时，何地以及使用什么虚拟机）直接影响云计算资源的能效和可持续性。云计算需要更多工具来测量资源的使用，包括用于定价目的的计算资源和数据，以及指导云计算服务的使用（Yang et al., 2016）。

时空思维方法是至关重要的，并且更应该被开发和正式化，以优化云计算进行地理空间大数据处理（Yang et al., 2015; Yang et al., 2016）。

应该从主动上下文（Odendaal, 2003）调查云计算和大数据技术在智慧城市和智慧社区等新举措中的应用（Batty, 2013;Mitton et al., 2012），相关数据选择，融合，挖掘

（Jiang and Thill, 2015）和知识演示（Fox, 2015）。

致谢

本研究由 NSF Cyber Polar，创新中心，EarthCube 和计算机网络系统程序（PLR-1349259，IIP-1338925，CNS-1117300，ICER-1343759）和 NASA（NNG12PP37I）以及 Microsoft，Amazon，Northrop Grumman，和 Harris。我们由衷感谢匿名审稿人的有见地的评论。George Taylor 博士编写了本文的早期版本。

参考文献

Agrawal, D., Das, S., & El Abbadi, A. (2011). Big Data and Cloud Computing: Current state

and future opportunities. Proceedings of the 14th International Conference on Extending

Database Technology (pp. 530–533) ACM.

Ammn, N., & Irfanuddin, M. (2013). Big Data challenges. International Journal of Advanced

Trends in Computer Science and Engineering, 2(1), 613–615.

AWS (2015). Landsat on AWS. https://aws.amazon.com/public-data-sets/landsat/.

Batty, M. (2013). Big Data, smart cities and city planning. Dialogues in Human Geography,

3(3), 274–279.

Benedetti, A., Baldasano, J. M., Basart, S., Benincasa, F., Boucher, O., Brooks, M. E., et al.

(2014). Operational dust prediction. In P. Knippertz, & W. J. -B. Stuut (Eds.), Mineral

dust: A key player in the Earth system (pp. 223–265). Dordrecht: Springer Netherlands.

Bulkeley, H., & Betsill, M. M. (2005). Cities and climate change: Urban sustainability and

global environmental governance. 4. (pp. 1–2). Florence: Psychology Press, 1–2.

Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolsvold, A., McKelvie, S., ... Haridas, J.

(2011). Windows Azure Storage: A highly available cloud storage service with strong

consistency. Proceedings of the Twenty-Third ACM Symposium on Operating Systems

Principles (pp. 143–157) ACM.

Chen, C. P., & Zhang, C. -Y. (2014). Data-intensive applications, challenges, techniques and

technologies: A survey on Big Data. Information Sciences, 275, 314–347.

Darmenova, K., Sokolik, I. N., Shao, Y., Marticorena, B., & Bergametti, G. (2009). Development of a physically based dust emission module within the Weather Research and

Forecasting (WRF) model: Assessment of dust emission parameterizations and

input parameters for source regions in Central and East Asia. Journal of Geophysical

Research. Atmospheres, 114(D14).

Das, M., & Parthasarathy, S. (2009). Anomaly detection and spatio-temporal analysis of

global climate system. Proceedings of the third international workshop on knowledge

discovery from sensor data (pp. 142–150) ACM.

Debbage, N., & Shepherd, J. M. (2015). The urban heat island effect and city contiguity.

Computers, Environment and Urban Systems, 54, 181–194.

Dwyer, J. L. (2014). Development of Landsat information products to Support Land

Change Monitoring, Assessment, and Projection (LCMAP). AGU fall meeting abstracts.

1. (pp. 3725).

Eastman, J. R. (2003). IDRISI Kilimanjaro: Guide to GIS and image processing. Worcester:

Clark Labs, Clark University, 305.

Eastman, J. R. (2012). IDRISI Selva manual. Worcester, Massachusetts, USA: Clark

University.

Einav, L., & Levin, J. D. (2013). The data revolution and economic analysis (no. w19035). National Bureau of Economic Research.

Fan, J., & Liu, H. (2013). Statistical analysis of Big Data on pharmacogenomics. Advanced

Drug Delivery Reviews, 65(7), 987–1000.

Fox, M. S. (2015). The role of ontologies in publishing and analyzing city indicators.

Computers, Environment and Urban Systems, 54, 266–279.

Frias-Martinez, V., Virseda, J., Rubio, A., & Frias-Martinez, E. (2010). Towards large scale

technology impact analyses: Automatic residential localization from mobile phonecall data. Proceedings of the 4th

ACM/IEEE international conference on information and communication technologies and development (pp. 11) ACM.

Gordon, M. I., Thies, W., & Amarasinghe, S. (2006). Exploiting coarse-grained task, data, and pipeline parallelism in stream programs. ACM SIGOPS Operating Systems Review, 40(5), 151–162.

Hansen, M. C., & Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. Remote Sensing of Environment, 122, 66–74.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ...

Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. Science, 342(6160), 850–853.

Hsu, C. H., Slagter, K. D., & Chung, Y. C. (2015). Locality and loading aware virtual machine mapping techniques for optimizing communications in MapReduce applications. Future Generation Computer Systems, 53, 43–54.

Huang, Q., Yang, C., Benedict, K., Chen, S., Rezgui, A., & Xie, J. (2013a). Utilize Cloud Computing to support dust storm forecasting. International Journal of Digital Earth, 6(4), 338–355.

Huang, Q., Yang, C., Benedict, K., Rezgui, A., Xie, J., Xia, J., & Chen, S. (2013b). Using adaptively coupled models and high-performance computing for enabling the computability of dust storm forecasting. International Journal of Geographical Information Science, 27(4), 765–784.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and its technical challenges. Communications of the ACM, 57(7), 86–94.

Jiang, B., & Thill, J. C. (2015). Volunteered geographic information: Towards the establishment of a new paradigm. Computers, Environment and Urban Systems, 53, 1–3.

Jiang, Y., Li, Y., Yang, C., Armstrong, E. M., Huang, T., & Moroni, D. (2016). Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. ISPRS International Journal of Geo-Information, 5(5), 54.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. Communications of the ACM, 57(3), 78–85.

Knippertz, P., & Stuut, J. B. W. (2014). Mineral Dust. Dordrecht, Netherlands: Springer.

Korf, R. E. (2011). A hybrid recursive multi-way number partitioning algorithm. IJCAI proceedings-International Joint Conference on Artificial Intelligence, 22(1), 591.

Krämer, M., & Senner, I. (2015). A modular software architecture for processing of big geospatial data in the cloud. Computers & Graphics, 49, 69–81.

Lee, J. G., & Kang, M. (2015). Geospatial Big Data: Challenges and opportunities. Big Data Research, 2(2), 74–81.

Li, Z. (2015). Optimizing geospatial cyberinfrastructure to improve the computing capability for climate studies. (Ph.D. Dissertation, George Mason University. http://ebot.gmu.edu/handle/1920/9630).

Li, Z., Yang, C., Huang, Q., Liu, K., Sun, M., & Xia, J. (2014). Building model as a service to support geosciences. Computers, Environment and Urban Systems. http://dx.doi.org/10.1016/j.compenvurbsys.2014.06.004.

Li, Z., Yang, C., Liu, K., Hu, F., & Jin, B. (2016a). Automatic scaling Hadoop in the cloud for efficient process of big geospatial data. ISPRS International Journal of Geo-Information, 5(10), 173.

Li, Z., Hu, F., Schnase, J. L., Duffy, D. Q., Lee, T., Bowen, M. K., & Yang, C. (2016b). A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. International Journal of Geographical Information Science doi:/10.1080/13658816.2015.1131830.

Lin, C., Zhu, J., & Wang, Z. (2008). Model bias correction for dust storm forecast using ensemble Kalman filter. Journal of Geophysical Research. Atmospheres, 113(D14).

Liu, Z., Liu, Q., Lin, H. C., Schwartz, C. S., Lee, Y. H., & Wang, T. (2011). Three-dimensional variational assimilation of MODIS aerosol optical depth:

Implementation and application to a dust storm over East Asia. Journal of Geophysical Research. Atmospheres,

116(D23).

Liu, X., Shi, X., Zhang, K., Jensen, E. J., Gettelman, A., Barahona, D., ... Lawson, P. (2012). Sensitivity studies of dust ice nuclei effect on cirrus clouds with the Community Atmosphere Model CAM5. Atmospheric Chemistry and Physics, 12(24), 12061–12079.

Long, Y., & Thill, J. C. (2015). Combining smart card data and household travel survey to

analyze jobs–housing relationships in Beijing. Computers, Environment and Urban

Systems, 53, 19–35.

Lushbough, C. M., Gnimpieba, E. Z., & Dooley, R. (2015). Life science data analysis

workflow development using the bioextract server leveraging the iPlant collaborative

cyberinfrastructure. Concurrency and Computation: Practice and Experience, 27(2),

408–419.

Manuel, P. (2015). A trust model of Cloud Computing based on quality of service. Annals

of Operations Research, 233(1), 281–292.

Marr, B. (2015). Big Data: Using SMART Big Data. Analytics and metrics to make better decisions and improve performance. Wiley 258pp.

Mayer-Schönberger, V., & Cukier, K. (2013). Big Data: A revolution that will transform how

we live, work, and think. Houghton Mifflin Harcourt.

Mell, P., & Grance, T. (2011). The NIST definition of Cloud Computing.

Meng, X., Bradley, J., Yuvaz, B., Sparks, E., Venkataraman, S., Liu, D., ... Xin, D. (2016). Mllib:

Machine learning in apache spark. JMLR, 17(34), 1–7.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge

discovery—An introduction. Computers, Environment and Urban Systems, 33(6),

403–408.

Mitton, N., Papavassiliou, S., Puliafito, A., & Trivedi, K. S. (2012). Combining cloud and sensors in a smart city environment. EURASIP Journal on Wireless Communications and

Networking, 1, 1.

Moore, R. (2015). How a Google engineer, 66,000 computers, and a Brazilian tribe made a

difference in how we view the Earth. (http://earthzine.org/2015/01/27/how-a-google-engineer-66000-computers-and-a-brazilian-tribe-made-a-difference-in-howwe-view-the-earth/).

8 C. Yang et al. / Computers, Environment and Urban Systems xxx (2016) xxx–xxx

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, Computers, Environment and Urban

Systems (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., &

Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble

of climate change simulations. Nature, 430(7001), 768–772.

Niu, T., Gong, S. L., Zhu, G. F., Liu, H. L., Hu, X. Q., Zhou, C. H., & Wang, Y. Q. (2008). Data

assimilation of dust aerosol observations for the CUACE/dust forecasting system.

Atmospheric Chemistry and Physics, 8(13), 3473–3482.

Odendaal, N. (2003). Information and communication technology and local governance:

Understanding the difference between cities in developed and emerging economies.

Computers, Environment and Urban Systems, 27(6), 585–607.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in

social media. Joint European Conference on Machine Learning and Knowledge Discovery

in Databases (pp. 18–33). Berlin Heidelberg: Springer.

Rosenzweig, C., Solecki, W. D., Hammer, S. A., & Mehrotra, S. (Eds.). (2011). Climate change

and cities: First assessment report of the urban climate change research network

(pp. xvi). Cambridge: Cambridge University Press.

Schnase, J. L., Duffy, D. Q., Tamkin, G. S., Nadeau, D., Thompson, J. H., Grieg, C. M., ...

Webster, W. P. (2014). MERRA analytic services: Meeting the Big Data challenges of climate science through cloud-enabled climate analytics-as-a-service. Environment and

Urban Systems: Computers.

Sekiyama, T. T., Tanaka, T. Y., Shimizu, A., & Miyoshi, T. (2010). Data assimilation of

CALIPSO aerosol observations. Atmospheric Chemistry and Physics, 10(1), 39–49.

Skiena, S. S. (1998). The algorithm design manual: Text. 1.Springer Science & Business

Media.

Skytland, N. (2012). Big Data: What is NASA doing with Big Data today. (Open. Gov open

access article).

Triguero, I., Peralta, D., Bacardit, J., García, S., & Herrera, F. (2015). MRPR: A MapReduce

solution for prototype reduction in Big Data classification. Neurocomputing, 150,

331–345.

Turner, B. L., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., Eckley, N., ...

Martello, M. L. (2003). Illustrating the coupled human–Environment system for vulnerability analysis: Three case studies. Proceedings of the National Academy of Sciences,

100(14), 8080–8085.

USGS (2016). Landsat 8 missions. http://landsat.usgs.gov/landsat8.php.

Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012).

Spatiotemporal data mining in the era of big spatial data: algorithms and applications. Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for

big geospatial data (pp. 1–10) ACM.

Wilkening, K. E., Barrie, L. A., & Engle, M. (2000). Trans-Pacific air pollution. Science,

290(5489), 65.

World Meteorological Organization (WMO) (2011). WMO Sand and Dust Storm Warning

Advisory and Assessment System (SDSWAS)—Science and implementation plan

2011–2015. Geneva, Switzerland: WMO.

Xie, J., Yang, C., Zhou, B., & Huang, Q. (2010). High-performance computing for the simulation of dust storms. Computers, Environment and Urban Systems, 34(4), 278–290.

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure:

Past, present and future. Computers, Environment and Urban Systems, 34(4), 264–277.

Yang, C., Wu, H., Huang, Q., Li, Z., & Li, J. (2011). Using spatial principles to optimize distributed computing for enabling the physical science discoveries. Proceedings of the

National Academy of Sciences, 108(14), 5498–5503.

Yang, C., Xu, Y., & Nebert, D. (2013). Redefining the possibility of digital Earth and

geosciences with spatial Cloud Computing. International Journal of Digital Earth,

6(4), 297–312.

Yang, C., Sun, M., Liu, K., Huang, Q., Li, Z., Gui, Z., ... Lostritto, P. (2015). Contemporary computing technologies for processing big spatiotemporal data. Space-time integration in

geography and GIScience (pp. 327–351). Netherlands.: Springer.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2016). Big Data and Cloud Computing: Innovation opportunities and challenges. International Journal of Digital Earth. http://dx.doi.

org/10.1080/17538947.2016.1239771.

Zhao, C., Liu, X., Leung, L. R., Johnson, B., McFarlane, S. A., Gustafson, W. I., Jr., ... Easter, R.

(2010). The spatial distribution of mineral dust and its shortwave radiative forcing

over North Africa: Modeling sensitivities to dust emissions and aerosol size treatments. Atmospheric Chemistry and Physics, 10(18), 8821–8838.

ARTICLE IN PRESS

CEUS-01097; No of Pages 9

Computers, Environment and Urban Systems xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Utilizing Cloud Computing to address big geospatial data challenges

Chaowei Yang *, Manzhu Yu, Fei Hu, Yongyao Jiang, Yun Li

NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA 22030, United States

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form 20 October 2016
Accepted 20 October 2016
Available online xxxx

Keywords:
Big Data
Cloud Computing
Spatiotemporal data
Geospatial science
Smart cities

ABSTRACT

Big Data has emerged with new opportunities for research, development, innovation and business. It is characterized by the so-called four Vs: volume, velocity, veracity and variety and may bring significant value through the processing of Big Data. The transformation of Big Data's 4 Vs into the 5th (value) is a grand challenge for processing capacity. Cloud Computing has emerged as a new paradigm to provide computing as a utility service for addressing different processing needs with a) on demand services, b) pooled resources, c) elasticity, d) broad band access and e) measured services. The utility of delivering computing capability fosters a potential solution for the transformation of Big Data's 4 Vs into the 5th (value). This paper investigates how Cloud Computing can be utilized to address Big Data challenges to enable such transformation. We introduce and review four geospatial scientific examples, including climate studies, geospatial knowledge mining, land cover simulation, and dust storm modelling. The method is presented in a tabular framework as a guidance to leverage Cloud Computing for Big Data solutions. It is demostrated throught the four examples that the framework method supports the life cycle of Big Data processing, including management, access, mining analytics, simulation and forecasting. This tabular framework can also be referred as a guidance to develop potential solutions for other big geospatial data challenges and initiatives, such as smart cities.

1. Introduction

Earth observation and model simulation produce tera- to peta- bytes of data daily (Yang, Raskin, Goodchild, and Gahegan, 2010). Non-traditional, geospatial data acquisition methods, such as social media (Romero, Galuba, Asur, and Huberman, 2011), phone conversations (Frias-Martinez, Virseda, Rubio, and Frias-Martinez, 2010) and unmanned aerial vehicles (Einav and Levin, 2013), produce geospatial data at even faster speeds. In addition to the large Volume (Marr, 2015; Hsu, Slagter, and Chung, 2015), geospatial data exist in a Variety of forms and formats for different applications, their accuracy and uncertainty span across a wide range as defined by Veracity, and data are produced in a fast Velocity through real time sensors (Fig. 1). With unprecedented information and knowledge embedded, these big geospatial data can be processed for adding Value to better scientific research, engineering development and business decisions (Lee and Kang, 2015). They are envisioned to provide innovation and advancements to improve our lives and understanding of the Earth systems (Mayer-Schönberger and Cukier, 2013) when transformed from the first four Vs to the last V (value) through advancements in a variety of geospatial domains (Fig. 1).

Such transformations pose grand challenges to data management and access, analytics, mining, system architecture and simulations (Yang, Huang, Li, Liu, and Hu, 2016). For example, the first challenge is how to deal with the Variety and Veracity of Big Data to produce a fused dataset that can be utilized in a single decision support system (Kim, Trimi, and Chung, 2014). Another issue is how to deal with the velocity of Big Data to have scalable and extensible processing power based on the fluctuation of the data feed (Ammn and Irfanuddin, 2013). Supporting on-demand or timely data analytical functionalities also pose significant challenges for creating the Value (Fan and Liu, 2013; Chen and Zhang, 2014; Jagadish et al., 2014).

Cloud Computing has emerged as a new paradigm to provide computing as a utility service with five advantageous characteristics (Fig. 1 bottom two layers): a) rapid and elastic provisioning computing power; b) pooled computing power to better utilize and share resources; c) broadband access for fast communication; d) on demand access for computing as utility services; and e) pay-as-you-go for the parts used without a significant upfront cost like that of traditional computing resources (Yang, Xu, and Nebert, 2013). Service-oriented architecture is adopted in Cloud Computing and enables "everything as a service", including Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) (Mell and Grance, 2011). While redefining the possibilities of geospatial science and Digital Earth (Yang et al., 2013), Cloud Computing engaging Big Data enlightens potential solutions for big geospatial data problems in various geosciences and relevant domains.

However, utilizing Cloud Computing to address Big Data issues is still in its infancy, and it is a daunting task on how the five advantageous

* Corresponding author.
E-mail address: cyang3@gmu.edu (C. Yang).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010

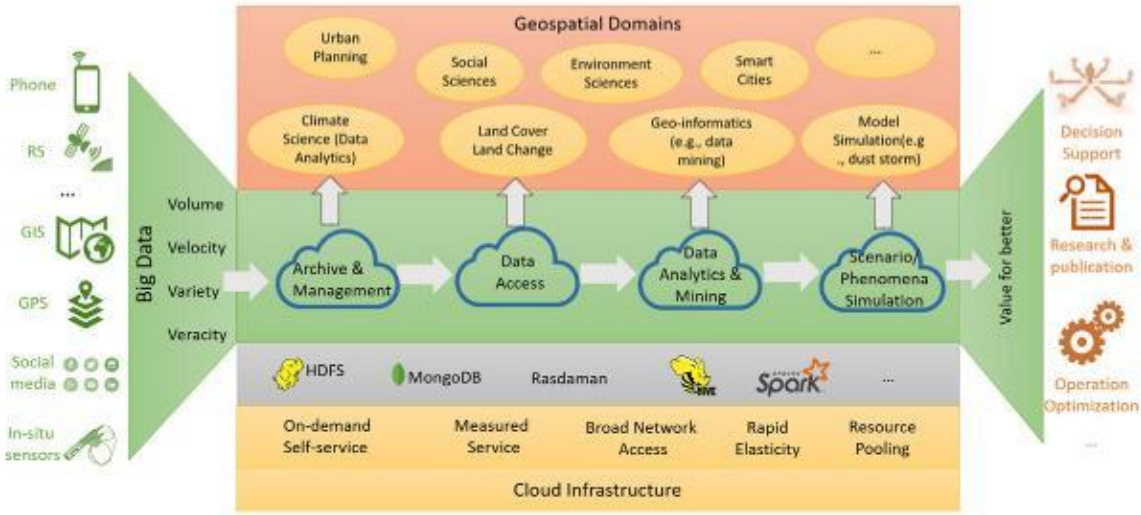


Fig. 1. Cloud Computing provides critical supports to the processing of Big Data to address the 4Vs to obtain Value for better decision support, research, and operations for various geospatial domains.

characteristics can address the first four Vs of Big Data to reach the 5th V (Fig. 1). This paper illustrates how Cloud Computing supports the transformation with four scientific examples including climate studies, knowledge mining, land-use and land cover change analysis, and dust storm simulation. These four examples are highly representative and can be easily adopted to other environmental and urban research fields, such as smart cities (Batty, 2013; Mitton, Papavassiliou, Puliafito, and Trivedi, 2012; Odendaal, 2003). The big geospatial data life cycle (data management and access, analyses/mining, phenomena/scenario simulation) are examined through the four examples and detailed in each example section (Table 1). For example, 2.1 is filled in the intersection cell of on-demand self-service and volume of Table 1. This means that 2.1 details how the volume (of big climate data) are addressed with the on-demand self-service of Cloud Computing.

2. Utilizing Cloud Computing to support climate analytics

The interrelated climate changes, such as greater incidence of heavy downpours and increased riverine flooding, are increasingly compromising urban infrastructure (Rosenzweig, Solecki, Hammer, and Mehrotra, 2011). Meanwhile human activities (e.g. the burning of fossil fuels) heavily impacted the global environment in the past 50 years (Bulkeley and Betsill, 2005). In order to understand climate change and its impacts to environmental and urban issues, the big climate data observed in the past and simulated for the future should be well managed and analyzed. However, both observation and simulation produce Big Data. For example, the next IPCC report will be based on 100 + petabytes of data, and NASA will produce 300 + petabytes of climate data by 2030 (Skytland, 2012). These data differ in format, spatiotemporal resolution, and study objective (Schnase et al., 2014). Big Data can help advance the understanding of climate phenomena and help identify how impacts of climate change on society and ecosystems can be remedied, such as detecting global temperature anomalies and investigating spatiotemporal distribution of extreme weather events, especially over highly populated regions (such as urban areas, Das and Parthasarathy, 2009; Debbage and Shepherd, 2015).

There are several challenges in the use of Big Data: a) the volume and velocity of big climate data have far exceeded the stand-alone computer's storage and computing ability; b) the variety of climate data in format and spatiotemporal resolution make it difficult to find an easy-to-use tool to analyze climate data; c) the veracity in model simulation is a concern for climate scientists of the uncertainties and mixed model qualities (Murphy et al., 2004). The combined complexities of volume, velocity, variety, and veracity can be addressed with cloud-based, advanced data management strategies and a service-oriented data analytical architecture to help process, analyze and mine climate data.

Table 1
The Big Data challenges as illustrated in the four examples are addressed by relevant cloud advantages to reach the Big Data Value and achieve the research, engineering and application objectives.

	On-demand Self-service	Broad network access	Resource pooling	Rapid elasticity	Measured service
Volume	2.1	4.1	2.1	2.1, 3.1, 3.2, 4.1, 4.2, 4.3, 5.1	4.1
Veracity	2.1	3.1, 5.3			
Velocity			2.1	4.1	4.3
Variety		3.1, 5.2	2.1		
Value	2.1, 3.2			2.1	3.2

2.1. Advanced spatiotemporal index for big climate data management

The hundreds of petabytes of climate data can only be managed in a distributed and scalable environment. Cloud Computing could help the management as follows: a) provisioning on-demand flexible virtual machines (VM) according to the volume of climate data; and b) automatically deploying HDFS, Hadoop Distributed File System, on the VMs to build a distributed filesystem. Data can be maintained in native format instead of sequenced text for saving storage space. A logical data architecture is also built to facilitate fast identification, access, and analyses (Li, Hu et al., 2016; Li, Yang et al., 2016). The core architecture is a spatiotemporal index (Li, Hu et al., 2016; Li, Yang et al., 2016) for the multi-dimensional climate data stored on HDFS. The index maps data content onto the byte, file and node levels within the HDFS. Nine components are used for the index and include: space, time and shape information describe the data grid's logical information which correlates to data query, byte offset, byte length, compression code, node list and file path identify specific location on the HDFS. This index enables users to directly locate and access data with exact spatiotemporal and content description.

In details, the space and time attributes in the spatiotemporal index will identify the grids overlapped with a spatiotemporal bounding box. The node list attribute is leveraged to deliver the computing programs to the node where the grids are stored. Then the computing programs can read the data as a data stream with high data locality, according to the byte offset, byte length, and compression code attributes. The

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010

shape and data type attributes can be used to reshape the data stream into a multiple-dimension array.

The monthly MERRA data for 26 years, MAIMNXINT[1] (about 90 Gb), are used to evaluate the cloud based and spatiotemporal indexed Big Data management efficiency. This experiment analyzes the monthly mean value of specific climate variables (by changing their numbers) in a specified spatiotemporal range from 36 VM-based HDFS cluster connected with 1 Gigabit (Gbps). Each node is configured with eight CPU cores (2.60 GHz), 16 GB RAM and CentOS 6.5. Results with and without using the index (Fig. 2) show when the number of processed variables increased the run time without the index increased by a factor of ~9.1, whereas the run time with the index only increased by a factor of 1.8. Based on the time constraints, a flexible number of VMs can be provisioned on demand to finish the tasks within a specific time frame (Li, Hu et al., 2016; Li, Yang et al., 2016; Yang et al., 2015). Therefore, on-demand service and elasticity in combination with a high level management effectively accommodate the big climate data management and analytical demands.

2.2. Anything as a service to ease the climate modelling experiments

Climate simulation poses challenges on obtaining enough computing resources for scientific experiments when analyzing big simulation data or running a large number of model simulations according to different model inputs. Cloud Computing addresses this experiment as follows: a) the climate models can be published as a service (MaaS; Li et al., 2014) and enough VMs can be provisioned with specific model configurations for each ensemble modelling run on demand; b) the application is deployed as a service (Lushbough, Gnimpieba, and Dooley, 2015) with a popular web portal to support model operation and monitoring; and c) the workflow involving different analytics is operated as a service (WaaS; Krämer and Senner, 2015) with intuitive GUIs. The big climate data analytics are supported by Cloud Computing at the computing infrastructure level.

The architecture of the cloud-based service-oriented workflow system for climate model study includes (Fig. 3): a) the model service is responsible for compiling and running models on VMs, which are provisioned based on the snapshot of the system containing the modelling software environment to run a model; b) the VM monitor service provides the cloud platform with VM status information for resource scheduling; c) the data analysis service feeds the model output as the input for analytics, while analyzing data in parallel to address data intensive issues. Data publishing service enables users to access the analysis results in real time via the Internet. All of these services are controllable through a GUI, which enables users to drag and connect services together to build a complex workflow so the system can automatically transition to the applications specified by the workflow and run on the cloud with automatically provisioned VMs. As an example, Li (2015) built ModelE as a service to study the sensitivity of ModelE, and the experiment showed that this cloud-based method reduced time consumption by 10 times over the traditional method.

The challenges in climate research addressed by Cloud Computing are summarized in Table 1. First, the large volume of climate data from observation and simulation are stored in the distributed and scalable environment provisioned by the cloud platform (2.1). Second, the variety challenge in climate data is addressed using the spatiotemporal index to unify them from the aspects of space and time (2.1). Third, the variety challenge in climate models is relieved by building the service-oriented system to simplify the model setup, running and output analysis (2.2). These methods can be extended to other geospatial domains which involve high dimensional data and complex models, such as remote sensing, image processing and agent-based modelling of environmental and urban events.

[1] http://disc.sci.gsfc.nasa.gov/mdisc/data-holdings/merra/merra_products_nonjs.shtml.

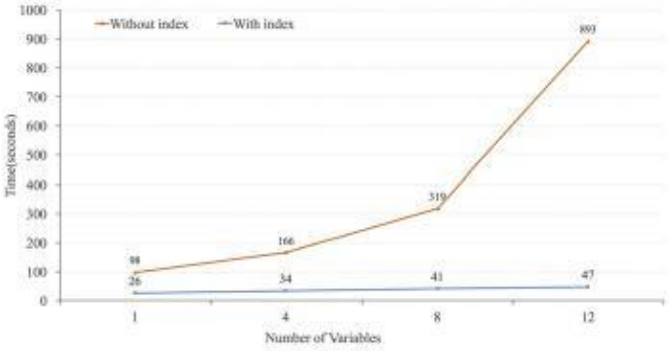


Fig. 2. Run time of the daily global mean calculation for different numbers of variables.

3. Supporting knowledge mining from big geospatial data

We have collected big geospatial data with different spatiotemporal stamps and resolutions for environment and urban studies using various methods, e.g., Global Positioning System (GPS), remote sensing, and Internet-based volunteer (Jiang and Thill, 2015; Yang et al., 2011). The increment in volume, velocity, and variety of the spatiotemporal data poses a grand challenge for researchers to discover and access the right data for research and decision support (Yang et al., 2011). One method of addressing this Big Data discovery challenge is to mine knowledge from the big geospatial data and their usages (Vatsavai et al., 2012) for query expansion, recommendation and ranking. The mined knowledge includes but is not limited to domain hot topics, research trends, metadata linkage and geospatial vocabularies similarity. This process is challenged with Big Data volume, velocity and variety. Such a mining process poses two challenges: a) how to divide Big Data into parallelizable chunks for processing with scalable computing resources; and b) how to utilize an adaptable number of computing resources for processing the divided Big Data. Take the MUDROD project for NASA Physical Oceanography Distributed Active Archive Center (PO. DAAC) as an example, the 2014 web log (contains geospatial data usage knowledge) was over 150 million records and the mining task takes >5 h to complete using a single server (6 cores, 12G memory and Win 7 OS). For high traffic websites with a large number of users sending requests concurrently, logs are produced at a much higher velocity, exceeding a single server's data-processing capability. In addition, logs are semi-structured or unstructured data stored in various formats (e.g. Apache HTTP, FTP, NGINX, IIS log format or user-defined format). Each format requires a specific processing protocol complicating the integration of different formats for further processing. The uncertainty affects the quality of mined knowledge with common noise (e.g., from web crawlers) requiring computational intensive crawler detection algorithms to preprocess original logs (Jiang et al., 2016).

3.1. Accelerating user log mining through data parallelism

The first step to processing big log files is to proceed in parallel by conducting the same operations on a dynamic number of VMs based on data volume and time constraints (Gordon, Thies, and Amarasinghe, 2006). To divide the original logs into the same number of VMs of a cluster, two data parallelism methods are applied to efficiently split logs, including time-based and IP-based log partition. In the time-based log partition (Fig. 4a), logs of consecutive dates are grouped into the same file. Once the original logs are split into k files (i.e., k = number of VMs in the cluster), the difference of the sum of logs in each file is minimized. This partitioning is solved as a linear partition problem (Skiena, 1998). In IP-based log partition (Fig. 4b), logs of the same IP are grouped into the same file using the greedy algorithm (Korf, 2011). Different from the time-based partition, the alteration of arrangement of IP is allowed.

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010

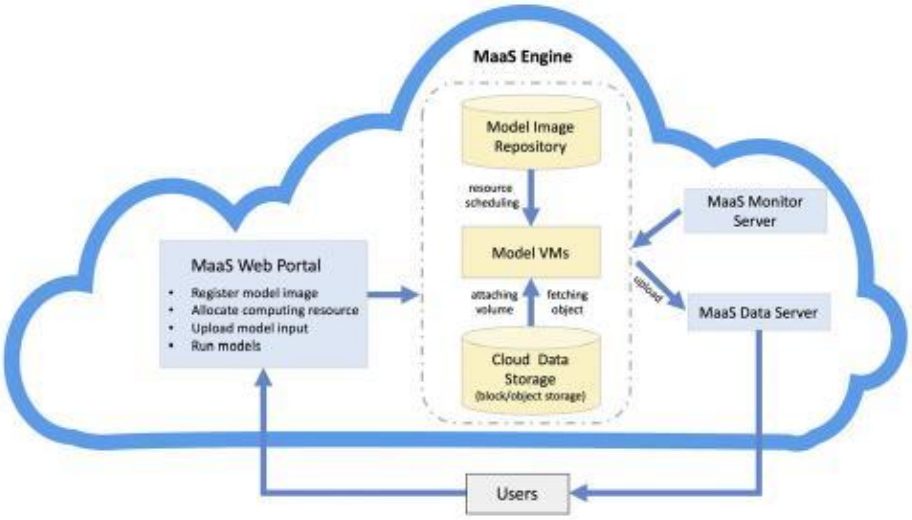


Fig. 3. The cloud-based service-oriented workflow system for climate model study.

Once the original logs are distributed in a virtual cluster, log pieces are processed in parallel in all VMs so log mining efficiency is notably increased.

3.2. Provisioning on-demand computation resources within a virtual cluster

In contrast to setting up a cluster manually, Cloud Computing facilitates the provisioning of a virtual cluster automatically with a dynamic number of VMs (Krämer and Senner, 2015). More computing resources can be deployed to process big historic data while a dynamic number of VMs can be provisioned to handle real-time data streams. On-demand computing resources are necessary to meet the requirement of dynamical log data volumes. For example, in the January 2014 PO. DAAC log mining task with more VMs in the cluster, less processing time was spent on finishing the task (Fig. 5a). Both the time-based partition and the IP-based partition dramatically accelerated the mining processes. However, the time-based partition changed sessions generated by log processer (Fig. 5b).

For the entire 2014 PO. DAAC logs, the total processing time was reduced 70%, from 190 to 49 min, as the VMs increased from 1 to 4 (Fig. 6).

Like geospatial data usage log, geospatial data can also be efficiently processed by a cluster leveraging data parallelism paradigm. Geospatial data can be partitioned into smaller parts based on different aspects, such as latitude, longitude, time or file size, and then distributed among VMs for parallel processing.

As summarized in Table 1, the broad network access and rapid elasticity enables data parallelism methods to efficiently segment Big Data and preparing the data for processing in parallel (3.1). The on-demand self-service, measured service and rapid elasticity add or remove computing nodes in a short time to meet the dynamic computing requirement (3.2).

The proposed knowledge discovery method of mining web log can be integrated with domain data portals to help environmental or urban scientists quickly discover useful information and knowledge. Though it should be pointed out that the user specific profiling (knowledge) data may also be of privacy and security concerns. In urban studies, spatial data mining and geographic knowledge discovery has emerged as an active research field in recent years. GPS data, high-resolution remote sensing data and internet-based volunteered geographic information are collected to extract unknown or unexpected information (Mennis and Guo, 2009; Jiang and Thill, 2015). These data sets are of unprecedentedly large size and the data parallelism paradigm can be leveraged to utilize Cloud Computing for efficient processing, e.g., for analyzing jobs-housing correlations (Long and Thill, 2015) in a smart cities context.

4. Supporting land-use and land-cover change analysis

Land-Use and Land-Cover Change (LULCC) has emerged as a fundamental component of environmental change and sustainability research. Landsat alone has produced 6 petabytes of data (Turner et al., 2003; Hansen and Loveland, 2012). The Land Change Monitoring, Assessment and Projection (LCMAP) pressed the need to generate science-quality land change products from current and near-real time Earth observations (Dwyer, 2014). However, several Big Data challenges exist as follows: a) storing, accessing and sharing big land use data; b) rapidly modelling LULCC with large-scale training set and complex algorithms; and c) rapidly changing analyses and predictions with LULCC data.

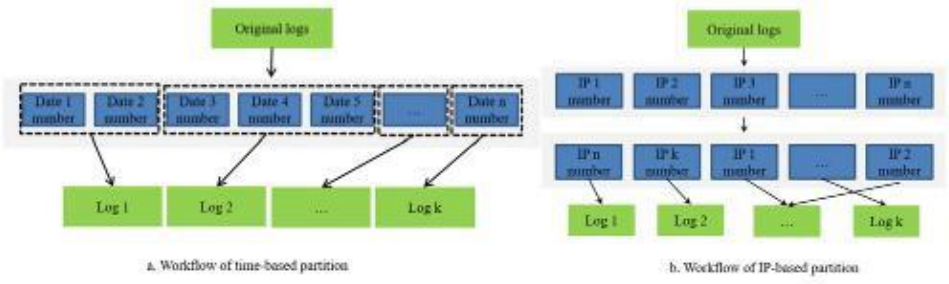
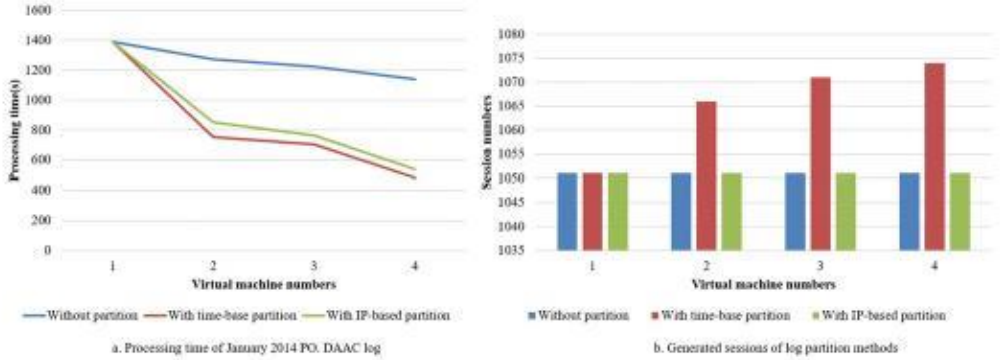


Fig. 4. The workflow of time-based partition (a) and IP-based partition (b).

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010



a. Processing time of January 2014 PO. DAAC log

b. Generated sessions of log partition methods

Fig. 5. Processing time of January 2014 PO. DAAC log (a) and generated sessions of log partition methods (b).

4.1. Storing, sharing and analyzing big land cover data on the cloud

Petabytes of historical LULCC and terabytes of streaming LULCC data require expensive and on-premise hardware that is hard to maintain and administrate, while cloud storage is outsourced to third-party cloud providers performing updates and maintenance (USGS, 2016). Additionally, cloud storage supports immediate access and exposes them through one simple web service interface, high reliability through redundancy and distribution of data and pay-as-you-go pricing (Calder et al., 2011). As the longest, continuous record of Earth's land surface observed from space, Landsat data is available via Amazon S3 since 2015. Most Landsat scenes from 2015 are available along with a selection of cloud-free scenes from 2013 and 2014. All new Landsat 8 scenes are available daily and often within hours of production (AWS, 2015). In addition to the improvement of data access, land cover imagery stored on the cloud is combined with land cover modelling published on the cloud to ease the LULCC research workflow, result sharing, and reproducing. For example, ArcGIS Online allows quick visualization and analysis of Landsat data on AWS. Mapbox uses Landsat on AWS to power Landsat-live, a browser-based map that is constantly refreshed with the latest imagery from the Landsat 8 satellite (AWS, 2015).

4.2. Rapid modelling with big training set and complex algorithms

Among the three types of LULCC models on image classification, land use suitability, and environmental impact of land cover change (Eastman, 2012), algorithms are complex and usually involve large training sets to build a robustic model. However, most can be converted to generic data mining problems. For example, the land change modeler of IDRISI, a popular GIS land change modelling tool, is based on logistic regression and neural networks (Eastman, 2003). The parallelization of these data mining algorithms is well studied in Cloud Computing communities and is supported by open source, large-scale processing frameworks, such as Spark MLlib (Meng et al., 2016). To leverage these technologies, a middleware was developed to convert the training set of the land cover images into the format that existing technologies can digest and convert the results back into ones that the LULCC requires (Fig. 7).

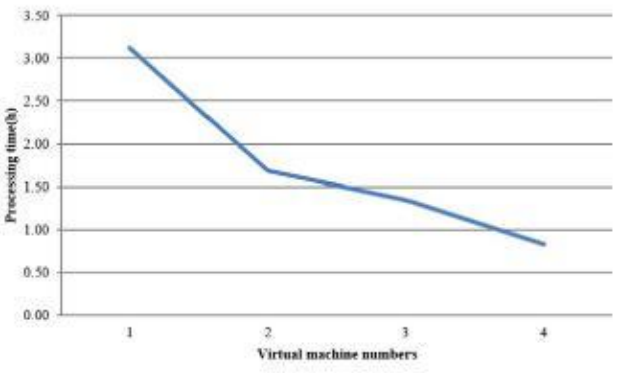


Fig. 6. Processing time of 2014 PO. DAAC log.

4.3. Rapid change analysis and prediction with big LULCC data

A classification or prediction model can be generated either through traditional approach or the one proposed in section 4.2. It would still be computationally intensive if each image and pixel is processed sequentially in LULCC models. This is handled with a virtual cluster to accelerate the processing through the following steps: a) parallelization of the study area into sub-areas; b) distribution of the LULCC data to the VMs where analyses run simultaneously; and c) aggregation of the results into a result dataset. As an example, a series of high-resolution global forest cover change maps from Google Earth Engine (Hansen et al., 2013) through its intrinsically-parallel computational access to Google cloud (Moore, 2015) demonstrates the possibility of utilizing a Cloud Computing platform to accelerate large land cover image classification (Fig. 8).

The broad network process, rapid elasticity and measured service improves the storage, access and analytics of big LULCC data for the data volume and velocity challenges (Tables 1, 4.1). The rapid elasticity allows large-scale data processing framework middleware to support the rapid modelling of large training sets and complex algorithms (4.2). The rapid elasticity and measured service also make it possible for the proposed parallel computing framework to provide near real-time classification, land cover change and prediction maps (4.3). The solutions proposed in LULCC can also be adopted in other scientific issues such as climate change, ecosystem service and habitat and biodiversity modelling.

5. Supporting dust storm forecasting

Dust storms are serious hazards to health, property, and the environment worldwide, especially urban areas (Knippertz and Stuut, 2014; WMO, 2011). During and after a dust storm, traffic accidents increase because of the rapidly decreasing visibility; air quality and human health are compromised when dust particles remain suspended in the atmosphere; efficiency of renewable energy sources is reduced when dust interferes with the energy capture mechanics (Wilkening, Barrie, and Engle, 2000). Therefore, it is crucial to predict an upcoming dust event with high spatiotemporal resolution to mitigate the environmental, health, and other asset impacts of dust storms (Benedetti et al., 2014). A standard requirement for such prediction requirement is to simulate one day phenomena within a two-hour computational time (Xie, Yang, Zhou, and Huang, 2010). This is easy to achieve with a coarse-resolution (1/3 degree) dust model forecast for the U.S. Southwest using a single CPU that takes ~4.5 h to complete processing. For

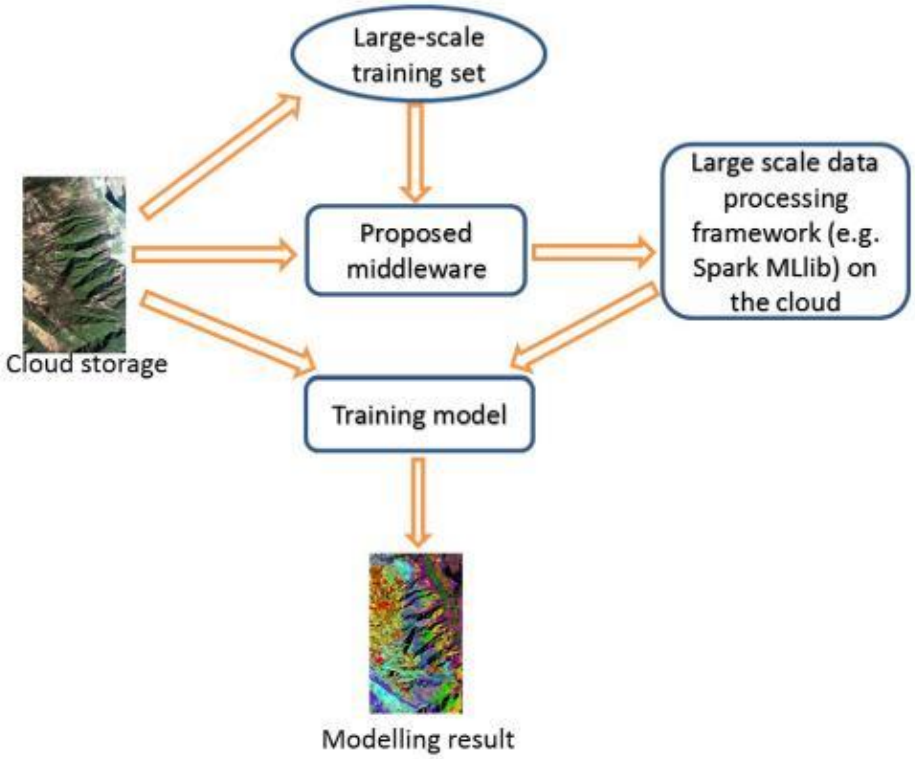


Fig. 7. The role the proposed middleware plays in the model building process.

high-resolution simulation (e.g. 3 km by 3 km), the volume of the model output data increases from 100 Gb to 10 Tb. The computational time increases by a factor of 4 in each of three dimensions (latitude, longitude and time steps). This results in an overall increase of a factor of 64 ($4 \times 4 \times 4 = 64$) or 12 days to complete the processing. This challenge of reducing from 12 days to 2 h is a Big Data problem in how to deal with the large volume of data processing/computing, how to ingest the variety of content input from geographic, atmospheric and ecosystem data and how to improve the veracity of model forecast data by ingesting high quality model input data.

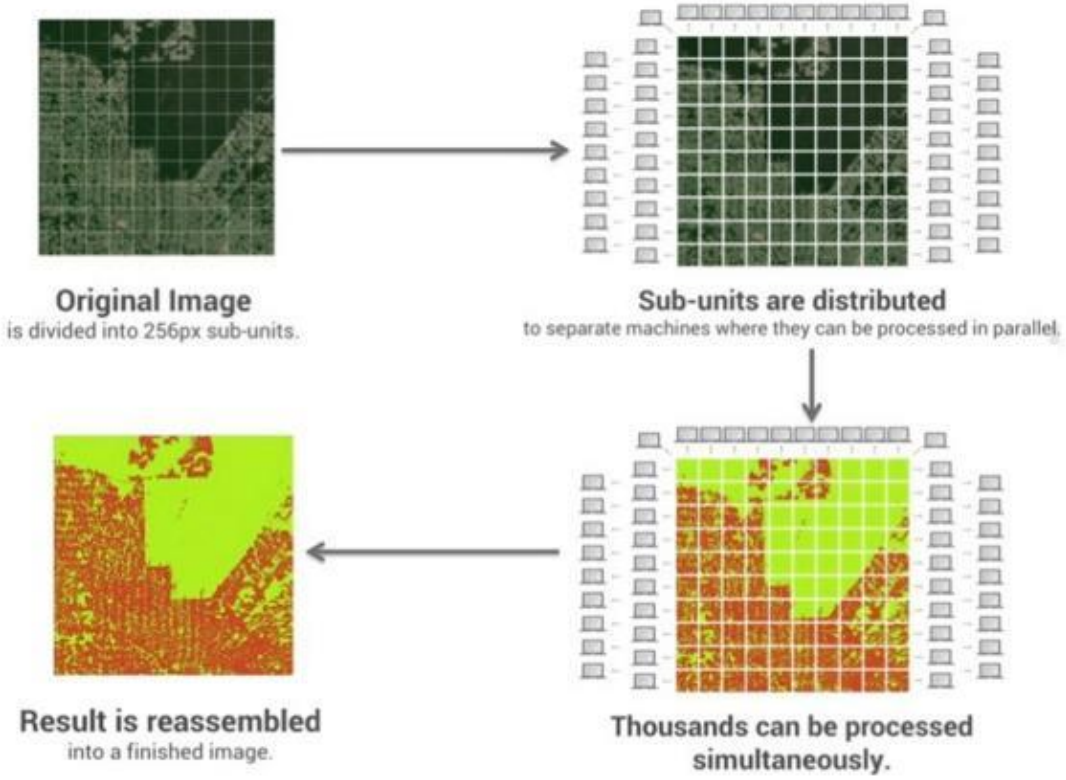


Fig. 8. Google Earth Engine divides Big Data to process in parallel using multiple computers (Moore, 2015).

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010

5.1. Accelerating large volume computation and processing

To deal with the challenge of reducing computing time from 12 days to 2 h, Huang, Yang, Benedict, Rezgui et al. (2013) and Huang, Yang, Benedict, Chen et al. (2013) proposed an adaptive, loosely-coupled model strategy, linking a high resolution/small scale dust model with a coarse resolution/large scale model. This strategy runs the low-resolution model and identifies subdomains, Areas of Interest (AOIs) with predicted high dust concentrations (Fig. 9a). A higher-resolution model for these AOIs is executed in parallel. With the support of Cloud Computing, clusters for high-resolution model runs for specific AOIs are established rapidly in parallel and are completed more efficiently than an execution of a high-resolution model over the entire domain. The execution time required for different AOIs when Cloud Computing handles all AOIs in parallel is <2.7 h (Fig. 9b).

5.2. Ingesting a big variety of dust model input

With the increase of spatiotemporal resolution of a dust forecast model, the challenge is to access dynamic data with different formats, content and uncertainties (Yang et al., 2011). The capability of broad network access of Cloud Computing can serve the access and preprocessing of a larger variety of the model input data with advanced network bandwidth and scalability. Huang, Yang, Benedict, Chen et al. (2013) and Huang, Yang, Benedict, Rezgui et al. (2013) showed that Amazon cloud instances can complete most of the forecasting tasks in less time than HPC clusters (Fig. 10), indicating that Cloud Computing has potential to resolve the concurrent intensity of the computing demanding applications.

5.3. Improving data veracity of dust forecasts

One of the most significant factors affecting the veracity of model output is the uncertainty of model initial condition (Lin, Zhu, and Wang, 2008). These uncertainties can be investigated and characterized through sensitivity tests using various model variables (Zhao et al., 2010; Liu et al., 2012). To reduce the uncertainty of the initial conditions, data assimilation techniques have been applied to dust models by assimilating the observations into the model to correct model initial conditions (Niu et al., 2008; Sekiyama, Tanaka, Shimizu, and Miyoshi, 2010; Liu et al., 2011). With the increasing variety of data sources, sensitivity tests and data assimilation can be conducted with minimum effort of preprocessing and integration into the model, thus enabling the efforts to improve model accuracy, and eventually reduce model uncertainty (Lin et al., 2008; Darmenova, Sokolik, Shao, Marticorena, and Bergametti, 2009). The entire complex process can be precisely preserved in a VM image that can be reused with minimum effort & reducing future manual errors.

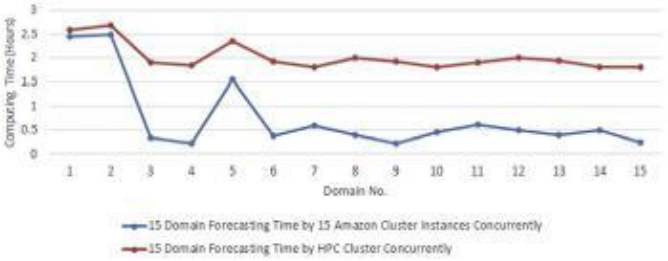


Fig. 10. NMM-dust execution time for 15 forecasting tasks on Amazon EC2 and HPC cluster (Huang, Yang, Benedict, Chen et al., 2013; Huang, Yang, Benedict, Rezgui et al., 2013).

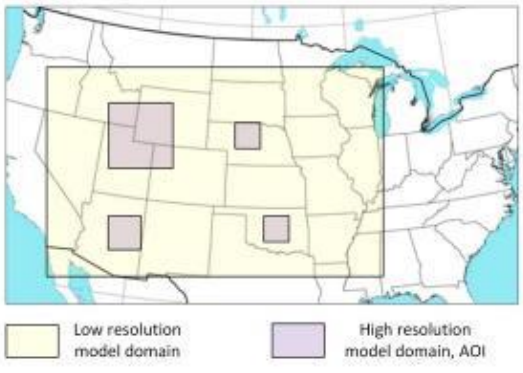
Therefore, the challenges for large-scale scientific prediction can be addressed by engaging the features of Cloud Computing with a net effect of accelerating a dust forecast task (Tables 1, 5.1). With broad network access, the ingestion of a larger variety of input data is achieved, and the ingestion is preprocessed on the cloud without consuming the computing resources designated for the core of model simulations (5.2). The selection of model input data is more sophisticated, improving the representation of the model's initial conditions and potentially data veracity of model's simulation output (5.3). These approaches are easily adaptable in other scientific computation or simulation models that require results within a short period of time, including the prediction of floods, hurricanes, and air pollution.

6. Conclusion

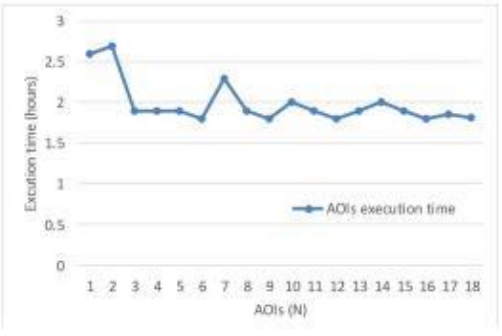
Big geospatial data pose grand challenges during the lifecycle of data storage, access, manage, analysis, mining, and modelling. The four examples illustrate the capability of Cloud Computing to address the 4 V challenges to reach Value with the five Cloud Computing advantages of on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service (Table 1). The boxes filled with section numbers indicate that these sections leverage the features of Cloud Computing to address the relevant challenges within big geospatial data. Table 1 is also of value as a guide to evaluate solutions for other big geospatial data challenges.

While research have been conducted to utilizing Cloud Computing to address big geospatial data challenges, many challenges remain to be addressed:

- Big geospatial data storage and management remains a high priority, including how to optimize different traditional (e.g., MySQL, PostgreSQL) and emerging database Management systems (e.g., NoSQL, HDFS, SPARK, HIVE) on the cloud environment for distributed storage, access and analytics (Agrawal, Das, and El Abbadi, 2011)



a. Low resolution and high resolution model domains



b. NMM-dust model execution time for each AOI in parallel

Fig. 9. Low-resolution model domain area and sub-regions (Area of Interests, AOIs) identified for high-resolution model execution (Huang, Yang, Benedict, Rezgui et al., 2013; Huang, Yang, Benedict, Chen et al., 2013).

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010

- Spatiotemporal Big Data mining requires real-time data processing, information extraction and automation to extract information and knowledge. More scalable spatiotemporal mining methods (Vatsavai et al., 2012) should be developed to take advantage of the elastic storage and computing resources of cloud platforms (Triguero, Peralta, Bacardit, García, and Herrera, 2015).
- Security is a challenge to assure protection for both sensitive data and the users' privacy. More research is needed to tracking and maintaining trust information to identify and prevent attacks on the cloud platform (Manuel, 2015).
- The usage behavior (e.g., when, where, and what VMs are used) on the cloud platform directly affects the energy efficiency and sustainability of the Cloud Computing resources. More tools are necessary to measure usage of resources, including computing resources and data for pricing purposes and to guide use of Cloud Computing services (Yang et al., 2016).
- Spatiotemporal thinking methodologies are critical, and more should be developed and formalized to optimize Cloud Computing for big geospatial data processing (Yang et al., 2015; Yang et al., 2016).
- Utilizing Cloud Computing and Big Data technologies in new initiatives, such as smart cities and smart communities (Batty, 2013; Mitton et al., 2012), should be investigated from the initiative context (Odendaal, 2003), application complexities (Long and Thill, 2015), relevant data selection, fusion, mining (Jiang and Thill, 2015), and knowledge presentation (Fox, 2015).

Acknowledgements

This research is supported by NSF Cyber Polar, Innovation Center, EarthCube and Computer Network System Programs (PLR-1349259, IIP-1338925, CNS-1117300, ICER-1343759) and NASA (NNG12PP37I) as well as Microsoft, Amazon, Northrop Grumman, and Harris. We thank the anonymous reviewers for their insightful comments and reviews. Dr. George Taylor edited an earlier version of the paper.

References

Agrawal, D., Das, S., & El Abbadi, A. (2011). Big Data and Cloud Computing: Current state and future opportunities. *Proceedings of the 14th International Conference on Extending Database Technology* (pp. 530–533) ACM.

Ammn, N., & Irfanuddin, M. (2013). Big Data challenges. *International Journal of Advanced Trends in Computer Science and Engineering, 2*(1), 613–615.

AWS (2015). *Landsat on AWS*. https://aws.amazon.com/public-data-sets/landsat/.

Batty, M. (2013). Big Data, smart cities and city planning. *Dialogues in Human Geography, 3*(3), 274–279.

Benedetti, A., Baldasano, J. M., Basart, S., Benincasa, F., Boucher, O., Brooks, M. E., et al. (2014). Operational dust prediction. In P. Knippertz, & W. J. -B. Stuut (Eds.), *Mineral dust: A key player in the Earth system* (pp. 223–265). Dordrecht: Springer Netherlands.

Bulkeley, H., & Betsill, M. M. (2005). *Cities and climate change: Urban sustainability and global environmental governance. 4.* (pp. 1–2). Florence: Psychology Press, 1–2.

Calder, B., Wang, J., Ogus, A., Nilakantan, N., Skjolsvold, A., McKelvie, S., ... Haridas, J. (2011). Windows Azure Storage: A highly available cloud storage service with strong consistency. *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles* (pp. 143–157) ACM.

Chen, C. P., & Zhang, C. -Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314–347.

Darmenova, K., Sokolik, I. N., Shao, Y., Marticorena, B., & Bergametti, G. (2009). Development of a physically based dust emission module within the Weather Research and Forecasting (WRF) model: Assessment of dust emission parameterizations and input parameters for source regions in Central and East Asia. *Journal of Geophysical Research. Atmospheres, 114*(D14).

Das, M., & Parthasarathy, S. (2009). Anomaly detection and spatio-temporal analysis of global climate system. *Proceedings of the third international workshop on knowledge discovery from sensor data* (pp. 142–150) ACM.

Debbage, N., & Shepherd, J. M. (2015). The urban heat island effect and city contiguity. *Computers, Environment and Urban Systems, 54*, 181–194.

Dwyer, J. L. (2014). Development of Landsat information products to Support Land Change Monitoring, Assessment, and Projection (LCMAP). *AGU fall meeting abstracts. 1.* (pp. 3725).

Eastman, J. R. (2003). *IDRISI Kilimanjaro: Guide to GIS and image processing*. Worcester: Clark Labs, Clark University, 305.

Eastman, J. R. (2012). *IDRISI Selva manual*. Worcester, Massachusetts, USA: Clark University.

Einav, L., & Levin, J. D. (2013). *The data revolution and economic analysis (no. w19035)*. National Bureau of Economic Research.

Fan, J., & Liu, H. (2013). Statistical analysis of Big Data on pharmacogenomics. *Advanced Drug Delivery Reviews, 65*(7), 987–1000.

Fox, M. S. (2015). The role of ontologies in publishing and analyzing city indicators. *Computers, Environment and Urban Systems, 54*, 266–279.

Frías-Martínez, V., Virseda, J., Rubio, A., & Frías-Martínez, E. (2010). Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development* (pp. 11) ACM.

Gordon, M. I., Thies, W., & Amarasinghe, S. (2006). Exploiting coarse-grained task, data, and pipeline parallelism in stream programs. *ACM SIGOPS Operating Systems Review, 40*(5), 151–162.

Hansen, M. C., & Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment, 122*, 66–74.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ... Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. *Science, 342*(6160), 850–853.

Hsu, C. H., Slagter, K. D., & Chung, Y. C. (2015). Locality and loading aware virtual machine mapping techniques for optimizing communications in MapReduce applications. *Future Generation Computer Systems, 53*, 43–54.

Huang, Q., Yang, C., Benedict, K., Chen, S., Rezgui, A., & Xie, J. (2013a). Utilize Cloud Computing to support dust storm forecasting. *International Journal of Digital Earth, 6*(4), 338–355.

Huang, Q., Yang, C., Benedict, K., Rezgui, A., Xie, J., Xia, J., & Chen, S. (2013b). Using adaptively coupled models and high-performance computing for enabling the computability of dust storm forecasting. *International Journal of Geographical Information Science, 27*(4), 765–784.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big Data and its technical challenges. *Communications of the ACM, 57*(7), 86–94.

Jiang, B., & Thill, J. C. (2015). Volunteered geographic information: Towards the establishment of a new paradigm. *Computers, Environment and Urban Systems, 53*, 1–3.

Jiang, Y., Li, Y., Yang, C., Armstrong, E. M., Huang, T., & Moroni, D. (2016). Reconstructing sessions from data discovery and access logs to build a semantic knowledge base for improving data discovery. *ISPRS International Journal of Geo-Information, 5*(5), 54.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big-data applications in the government sector. *Communications of the ACM, 57*(3), 78–85.

Knippertz, P., & Stuut, J. B. W. (2014). *Mineral Dust*. Dordrecht, Netherlands: Springer.

Korf, R. E. (2011). A hybrid recursive multi-way number partitioning algorithm. *IJCAI proceedings-International Joint Conference on Artificial Intelligence, 22*(1), 591.

Krämer, M., & Senner, I. (2015). A modular software architecture for processing of big geospatial data in the cloud. *Computers & Graphics, 49*, 69–81.

Lee, J. G., & Kang, M. (2015). Geospatial Big Data: Challenges and opportunities. *Big Data Research, 2*(2), 74–81.

Li, Z. (2015). *Optimizing geospatial cyberinfrastructure to improve the computing capability for climate studies*. (Ph.D. Dissertation, George Mason University. http://ebot.gmu.edu/handle/1920/9630).

Li, Z., Yang, C., Huang, Q., Liu, K., Sun, M., & Xia, J. (2014). Building model as a service to support geosciences. *Computers, Environment and Urban Systems*. http://dx.doi.org/10.1016/j.compenvurbsys.2014.06.004.

Li, Z., Yang, C., Liu, K., Hu, F., & Jin, B. (2016a). Automatic scaling Hadoop in the cloud for efficient process of big geospatial data. *ISPRS International Journal of Geo-Information, 5*(10), 173.

Li, Z., Hu, F., Schnase, J. L., Duffy, D. Q., Lee, T., Bowen, M. K., & Yang, C. (2016b). A spatiotemporal indexing approach for efficient processing of big array-based climate data with MapReduce. *International Journal of Geographical Information Science* doi:/10.1080/13658816.2015.1131830.

Lin, C., Zhu, J., & Wang, Z. (2008). Model bias correction for dust storm forecast using ensemble Kalman filter. *Journal of Geophysical Research. Atmospheres, 113*(D14).

Liu, Z., Liu, Q., Lin, H. C., Schwartz, C. S., Lee, Y. H., & Wang, T. (2011). Three-dimensional variational assimilation of MODIS aerosol optical depth: Implementation and application to a dust storm over East Asia. *Journal of Geophysical Research. Atmospheres, 116*(D23).

Liu, X., Shi, X., Zhang, K., Jensen, E. J., Gettelman, A., Barahona, D., ... Lawson, P. (2012). Sensitivity studies of dust ice nuclei effect on cirrus clouds with the Community Atmosphere Model CAM5. *Atmospheric Chemistry and Physics, 12*(24), 12061–12079.

Long, Y., & Thill, J. C. (2015). Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems, 53*, 19–35.

Lushbough, C. M., Gnimpieba, E. Z., & Dooley, R. (2015). Life science data analysis workflow development using the bioextract server leveraging the iPlant collaborative cyberinfrastructure. *Concurrency and Computation: Practice and Experience, 27*(2), 408–419.

Manuel, P. (2015). A trust model of Cloud Computing based on quality of service. *Annals of Operations Research, 233*(1), 281–292.

Marr, B. (2015). Big Data: Using SMART Big Data. *Analytics and metrics to make better decisions and improve performance*. Wiley 258pp.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Mell, P., & Grance, T. (2011). *The NIST definition of Cloud Computing*.

Meng, X., Bradley, J., Yuvaz, B., Sparks, E., Venkataraman, S., Liu, D., ... Xin, D. (2016). Mllib: Machine learning in apache spark. *JMLR, 17*(34), 1–7.

Mennis, J., & Guo, D. (2009). Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems, 33*(6), 403–408.

Mitton, N., Papavassiliou, S., Puliafito, A., & Trivedi, K. S. (2012). Combining cloud and sensors in a smart city environment. *EURASIP Journal on Wireless Communications and Networking, 1*, 1.

Moore, R. (2015). *How a Google engineer, 66,000 computers, and a Brazilian tribe made a difference in how we view the Earth*. (http://earthzine.org/2015/01/27/how-a-google-engineer-66000-computers-and-a-brazilian-tribe-made-a-difference-in-how-we-view-the-earth/).

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature, 430*(7001), 768–772.

Niu, T., Gong, S. L., Zhu, G. F., Liu, H. L., Hu, X. Q., Zhou, C. H., & Wang, Y. Q. (2008). Data assimilation of dust aerosol observations for the CUACE/dust forecasting system. *Atmospheric Chemistry and Physics, 8*(13), 3473–3482.

Odendaal, N. (2003). Information and communication technology and local governance: Understanding the difference between cities in developed and emerging economies. *Computers, Environment and Urban Systems, 27*(6), 585–607.

Romero, D. M., Galuba, W., Asur, S., & Huberman, B. A. (2011). Influence and passivity in social media. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 18–33). Berlin Heidelberg: Springer.

Rosenzweig, C., Solecki, W. D., Hammer, S. A., & Mehrotra, S. (Eds.). (2011). *Climate change and cities: First assessment report of the urban climate change research network* (pp. xvi). Cambridge: Cambridge University Press.

Schnase, J. L., Duffy, D. Q., Tamkin, G. S., Nadeau, D., Thompson, J. H., Grieg, C. M., ... Webster, W. P. (2014). *MERRA analytic services: Meeting the Big Data challenges of climate science through cloud-enabled climate analytics-as-a-service*. Environment and Urban Systems: Computers.

Sekiyama, T. T., Tanaka, T. Y., Shimizu, A., & Miyoshi, T. (2010). Data assimilation of CALIPSO aerosol observations. *Atmospheric Chemistry and Physics, 10*(1), 39–49.

Skiena, S. S. (1998). *The algorithm design manual: Text. 1*.Springer Science & Business Media.

Skytland, N. (2012). *Big Data: What is NASA doing with Big Data today.* (Open. Gov open access article).

Triguero, I., Peralta, D., Bacardit, J., García, S., & Herrera, F. (2015). MRPR: A MapReduce solution for prototype reduction in Big Data classification. *Neurocomputing, 150*, 331–345.

Turner, B. L., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., Eckley, N., ... Martello, M. L. (2003). Illustrating the coupled human–Environment system for vulnerability analysis: Three case studies. *Proceedings of the National Academy of Sciences, 100*(14), 8080–8085.

USGS (2016). *Landsat 8 missions*. http://landsat.usgs.gov/landsat8.php.

Vatsavai, R. R., Ganguly, A., Chandola, V., Stefanidis, A., Klasky, S., & Shekhar, S. (2012). Spatiotemporal data mining in the era of big spatial data: algorithms and applications. *Proceedings of the 1st ACM SIGSPATIAL international workshop on analytics for big geospatial data* (pp. 1–10) ACM.

Wilkening, K. E., Barrie, L. A., & Engle, M. (2000). Trans-Pacific air pollution. *Science, 290*(5489), 65.

World Meteorological Organization (WMO) (2011). *WMO Sand and Dust Storm Warning Advisory and Assessment System (SDSWAS)—Science and implementation plan 2011–2015*. Geneva, Switzerland: WMO.

Xie, J., Yang, C., Zhou, B., & Huang, Q. (2010). High-performance computing for the simulation of dust storms. *Computers, Environment and Urban Systems, 34*(4), 278–290.

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems, 34*(4), 264–277.

Yang, C., Wu, H., Huang, Q., Li, Z., & Li, J. (2011). Using spatial principles to optimize distributed computing for enabling the physical science discoveries. *Proceedings of the National Academy of Sciences, 108*(14), 5498–5503.

Yang, C., Xu, Y., & Nebert, D. (2013). Redefining the possibility of digital Earth and geosciences with spatial Cloud Computing. *International Journal of Digital Earth, 6*(4), 297–312.

Yang, C., Sun, M., Liu, K., Huang, Q., Li, Z., Gui, Z., ... Lostritto, P. (2015). Contemporary computing technologies for processing big spatiotemporal data. *Space-time integration in geography and GIScience* (pp. 327–351). Netherlands.: Springer.

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2016). Big Data and Cloud Computing: Innovation opportunities and challenges. *International Journal of Digital Earth*. http://dx.doi.org/10.1080/17538947.2016.1239771.

Zhao, C., Liu, X., Leung, L. R., Johnson, B., McFarlane, S. A., Gustafson, W. I., Jr., ... Easter, R. (2010). The spatial distribution of mineral dust and its shortwave radiative forcing over North Africa: Modeling sensitivities to dust emissions and aerosol size treatments. *Atmospheric Chemistry and Physics, 10*(18), 8821–8838.

Please cite this article as: Yang, C., et al., Utilizing Cloud Computing to address big geospatial data challenges, *Computers, Environment and Urban Systems* (2016), http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.010