The 8th International Scientific Conference
eLearning and software for Education
Bucharest, April 26-27, 2012
10.5682/2066-026X-12-092

CLOUD-BASED E-LEARNING TOOLS FOR DATA ANALYSIS

Grigore ALBEANU
Faculty of Mathematics and Informatics, ″Spiru Haret″ University Ion Ghica Str., Bucharest, Romania
E-mail: galbeanu@gmail.com

Florin Popențiu-VLĂDICESCU
″UNESCO″ Department, University of Oradea, Universității Str., Oradea, Romania
E-mail: popentiu@imm.dtu.dk

***Abstract**: This paper describes the most important characteristics of cloud computing paradigm in order to be used as powerful arrows towards efficient approaches in teaching data analysis. Also High Performance Computing aspects are considered for solving large complexity data analysis problems. The usage of both Cloud Computing and High Performance Computing will assure not only an excellent framework for research but also a powerful and flexible environment for e-Learning. The presentation is structured in three sections, being closed with the list of references. The first section considers the resources of cloud computing: SaaS (Software as a Service), IaaS (Infrastructure as a Service), PaaS (Platform as a Service - predefined integrated platform), and Managed Services. For every mentioned resource the most important features which are valuable for e-learning are outlined. High Performance Computing for data analysis is described in the second section. Also, software tools useful to analyse collections of data having small, medium or large size are presented and compared in order to identify the best tool to be used in a virtual learning framework. The third section describes both data manipulation techniques and results of experiments. Finally, concluding remarks are presented related to an online course on data analysis based on cloud computing paradigm.*

***Keywords:** cloud computing, high performance computing, data analysis, e-learning*

I. INTRODUCTION

From visual point of view, the cloud icon is used in network diagrams to depict the Internet environment, but in technical sense a cloud computing solution is based on clients (the mobile systems like PDAs or smart phones, computers without internal hard drives connected to servers, regular computers), a data centre (a collection of applications usable by subscription which are running on physical servers, placed in one large location, allowing multiple instances of virtual servers), and distributed servers (physical servers allowing virtual servers instantiation, being installed in geographically different locations). Cloud computing allows multiple small applications running in the same time and is different from grid computing which assume the participation of computers (with their resources) in a network to work for solving one and only scientific or technical problem in the same time (a large project was divided among multiple computers).

Empowering a cloud is possible by full virtualization (when a complete installation of a machine is emulated and run on another) and paravirtualization (allowing multiple operating systems to run on a single machine). The other concept used in modern computing paradigms is the service (usage of reusable resources across an environment).

The next section considers the resources of cloud computing: SaaS (Software as a Service), IaaS (Infrastructure as a Service), PaaS (Platform as a Service - predefined integrated platform), and Managed Services. For every mentioned resource the most important features which are valuable for e-learning are outlined.

High Performance Computing for data analysis is described in the third section. Also, software tools useful to analyse collections of data having small, medium or large size are presented and compared in order to identify the best tool to be used in a virtual learning framework.

II. THE RESOURCES OF CLOUD COMPUTING

According to Vaquero et al. (2009), "Clouds are a large pool of easily usable and accessible virtualized resources (such as hardware, development platforms and/or services) […] dynamically reconfigured to adjust to a variable load (scale), allowing also for an optimum resource utilization. This pool of resources is typically exploited by a pay-per-use model in which guarantees are offered by the Infrastructure Provider by means of customised SLAs [service-level agreements] […] The set of features that most closely resemble this minimum definition would be scalability, pay-per-use utility model and virtualization".

For any kind of clouds (public – offered by third-party providers, hybrid – offered by both in-house and third party providers, private – in-house cloud, community cloud – services offered only to community members) the offered resources are modelled by: SaaS (Software as a Service), IaaS (Infrastructure as a Service), PaaS (Platform as a Service - predefined integrated platform), and Managed Services.

The SaaS model is a software distribution model in which applications are hosted by a vendor or service provider in order to be accessible by clients over a computer network. In Internet, this approach uses web services under the service-oriented architecture paradigm. The SaaS model is related to, but different in many aspects from, the "*application service provider*" and "*on demand computing"* models.

The ideal candidates for SaaS model are that software performing a single task without interaction with other systems. The set of applications cover, but is not limited to, the following: CRM – Customer Resource Management, Video conferencing, Web content management, Web analytics, Accounting, IT service management.

The IaaS (sometimes called HaaS – Hardware as a Service) model supports the outsourcing of the equipment used to support operations, including storage, hardware, servers and networking components. According to Lee et al (2011), the IaaS model covers four categories of services offered under virtualization approach: Resource Management Service, Virtual Machine Service, Clone Service, and Security Service. The infrastructure can be scaled up and down depending on the client needs.

The PaaS model is dedicated to application developers, testers, distributors, and administrators, and includes: the development environment, programming languages, compilers, testing tools and deployment mechanism. According to Natis et al. (2011), different PaaS models exist: aPaaS - *application PaaS* - the platform for hosting and managing individual application services and data: only a specific component or subset is offered, iPaaS - *integration PaaS* - the platform for intermediation and integration of the application services hosted and point-managed by aPaaS, kPaaS - *knowledge PaaS* - the platform for access and analysis of broad data resources in context, uxPaaS - *user experience PaaS* - the platform for multichannel, multidevice user-facing applications, dPaaS - *Data PaaS* - the platform for hosting and serving data.

According to OpSource, the Managed Services "enable small businesses, software developers, web design agencies and large enterprises to migrate and scale mission-critical applications in the cloud with the hands-on support of an experienced operator".

Some players in Cloud Computing world are (links included in references list): Amazon (Elastic Compute Cloud, Simple Storage Service, Simple Queue Service, Simple DB, etc.), Google (Google Apps Engine, Google Docs, etc.), Microsoft (Azure, SQL Services, .NET Services, Live Services, Share Point Services, Dynamics CRM Services, etc.), Yahoo, IBM, Intel, and Oracle.

Following Laisheng & Zhengxia (2011), cloud based E-Learning is a subfield of cloud computing on educational field for e-learning systems based on a five-layers architecture: HRL – the Hardware Resource layer, SRL – the Software Resource Layer, RML – the Resource Management Layer, SL – Service Layer (based on IaaS, PaaS, SaaS), and BAL – the Business Application Layer (supporting: content creation, content delivery, education platform, teaching evaluation and education management).

According to Ekanayake et al. (2010) there is a difference among "Cloud" and "cloud technologies" which associate the term "Cloud" with Xservices: SaaS, PaaS, IaaS. "Cloud technologies" is associated to cloud runtimes: Hadoop, Dryad, MapReduce, DFS, etc.

Mobile Cloud Computing can be used in mobile learning approach. The lowest cost in education can be obtained using small and cheap terminals (like Smart phones) connected to a cloud with large storage capacity and powerful processing ability. It is not necessary to use a powerful configuration because the computations will be assured by the cloud.

The open source JavaME UI framework and Jabber can be used for clients. Also Android provides powerful resources for a successful player on mobile e-learning market. However, the cost on Mobile Internet is large for the moment.

III. HIGH PERFORMANCE COMPUTING FOR DATA ANALYSIS

Firstly, we are interested in data oriented cloud computing (DOCC) architectures. Every DOCC architecture should consists of a basic layer solving concurrency issues for the distributed data service Computation Service. A data abstraction component is required to provide various interfaces to data. High level languages will support the computing layer.

For coordination reasons Google uses Chubby - a fault-tolerant system that provides a distributed locking mechanism and stores small files, while GFS – the Google File System is a distributed file system. Data abstraction, in Google, is supported by BigTable, which is a multidimensional non-relational data base suitable for non structured data. The Computation Layer is based on MapReduce programming model. As a high level, parallel data processing scripting language built on top of MapReduce is Sawzall.

The Cloud Computing approach is useful for data analysis if the workload is parallelizable. Another important concept when dealing with large databases processed in distributed environments is the "share nothing architecture" - each node being independent and self-sufficient. ACID characteristics (Atomicity, Consistency, Isolation, and Durability) are difficult to be maintained when replication over network is necessary. Google's Bigtable implements a replicated shared-nothing database suitable for analytical data management – that queries a data base in order to solve problems and support decisions.

MapReduce, Hadoop, Dryad, and other projects are all dedicated to automate the parallelization of large scale data analysis workloads. The most fault tolerant proposal is MapReduce. A Data Analysis Job is divided into many small tasks and in the case of failure, the task associated initially to a failing machine are reassigned to an available machine.

Limitations and opportunities were discussed by Abadi (2009). However, the most suitable for data analysis in the cloud are the following shared-nothing parallel databases: Teradata, Netezza, IBM DB2, Greenplum, and Vertica (see the references).

According to Law (2011/2012), "HPC has always been a tool for big companies, with high-level requirements – but it really can be used by small companies." As Farber et al. (2011) considers: "cloud computing provides new capabilities for performing analysis across all data in an organization." Intelligent data analysis assumes that when data is collected, the first step is to transform data, normalize it, and insert in the database. Sorting, data mining, image manipulation, social network analysis, inverted index construction, and machine learning are solved by MapReduce.

Some relevant projects in cloud computing for data analysis, implemented using Hadoop, CGL-MapReduce, and DryadLINQ, are (Ekanayake et al. (2010)):

- CAP3 - DNA Sequence Assembly Program for small-scale assembly of EST (expressed sequence tag) sequences with or without quality values.
- HEP - High Energy Physics Applications.
- Iterative MapReduce - Kmeans clustering and Matrix Multiplication Algorithms.

- ALU Sequencing Studies – Clustering, dissimilarities, etc.

The procedure used by DataMine Lab consists of the following steps:

1) Load data on Amazon Cloud and analyze information by simple techniques;
2) Use software such as Hadoop or HBase to mine and further analyze the data from as many perspectives as they can in order to generate the most useful set of results.
3) Use Pentaho for reporting.

In conclusion, distributed cloud computing, the recent multi-core architectures, and massive many-core parallelization through graphics processing units (GPUs) can provide several orders-of-magnitude performance in fast simulation, optimization and stochastic search methods, applied to fit and evaluate complex models from large collections of data. The next section discuses the benefits of using cloud computing for online teaching/learning, and experimental research on large collections by data analysis approaches.

IV. CLOUD COMPUTING BASED ELEARNING ENVIRONMENTS FOR DATA ANALYSIS

Cloud computing is an important paradigm of computing having various applications including education. This section describes the objectives and structure of a two weeks summer course in cloud-based data analysis addressed to master students in Applied Statistics, an international master program under TEMPUS project. Thought as in campus course, however the lectures and practice activities can be organized using modern Information and Communication Technologies commonly used in the E-Learning field, as Zaharescu (2012) described.

The course objectives are:

- Introduce Cloud computing paradigm for managing large data collections in distributed manner – 2 days.
- Introduce the R programming language by examples on small, medium and large data files – 2 days.
- Introduce data mining methods on distributed data bases – 2 days.
- Practice on case studies like: Predicting Algae Blooms, Predicting Stock Market Returns, Detecting Fraudulent Transactions, and Classifying Microarray Samples. – 3 days per project/(team composed by 3 to 5 people).
- Dissemination: Every team will present the experience gained during course and the project results – the tenth day.

The following resources will be used: A private Data server, Amazon Elastic Cloud, ROSTLAB Resources, Elastic-R experience, Printed books, articles, and on-line available lectures.

The preliminary study and experiments are positive giving the hope for a successful project.

Acknowledgements

During this research, the first author was supported by the Research Centre of the Faculty of Mathematics and Informatics according to the programs: “Applied informatics for software engineering” 107/2010, and “Research in theoretical informatics and computing methodologies” 106/2010. The second author acknowledges the support of the “UNESCO” Department of the University of Oradea.

References

[1] Abadi, D.J., 2009. Data Management in the Cloud: Limitations and Opportunities, *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*..

[2] Albeanu, G., Şerbănescu, L., Popenţiu-Vlădicescu, Fl., 2007. On teaching data analysis and optimisation using software tools. *In Grigore Albeanu, Dorin Mircea Popovici , Marin Vlada (eds.), Proceedings of the 2nd International Conference on Virtual Learning, Constanţa, 26-28 October, Bucharest*, Romania, I, pp. 255-260. 2007.

[3] Amazon Web Services. http://aws.amazon.com/

[4] Amazon Elastic Compute Cloud (Amazon EC2). http://aws.amazon.com/en/ec2/.

[5] Android – E-learning applications. http://www.androidzoom.com/android_applications/e+learning.

[6] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.A., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica, I., Zaharia, M., 2009. Above the Clouds: A Berkeley View of Cloud Computing, http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf.

[7] CAP3 – Sequence Assembly. http://deepc2.psi.iastate.edu/aat/cap/cap.html

[8] Chine, K., 2010. Learning Math and Statistics on the Cloud, Towards an EC2-Based Google Docs-like Portal for Teaching / Learning Collaboratively with R and Scilab, 10th International Conference on Advanced Learning Technologies (ICALT), pp. 752-753, http://dx.doi.org/10.1109/ICALT.2010.120.

[9] DataMine Lab. http://www.dataminelab.com/technology/.

[10] DB2 (IBM Redbooks). http://www.redbooks.ibm.com/abstracts/sg244695.html.

[11] Dinh, H.T., Lee, C., Niyato, D., Wang, P., 2011. A Survey of Mobile Cloud Computing: Architecture, Applications, and Approaches, Wireless Communications and Mobile Computing, John Wiley & Sons, Ltd., http://dx.doi.org/10.1002/wcm.1203.

[12] Dryad (The project). http://research.microsoft.com/en-us/projects/dryad/.

[13] Ekanayake, J., Qiu, X., Gunarathne, T., Beason, S., Fox, G., 2010. Performance Parallel Computing with Cloud and Cloud Technologies, *Book chapter to Cloud Computing and Software Services: Theory and Techniques CRC Press (Taylor and Francis)*; 2010.

[14] ELASTIC-R Project. http://www.elasticr.net/.

[15] Farber, M., Cameron, M., Ellis, C., Sullivan J., 2011. Massive Data Analytics and the Cloud, 8 p., http://www.boozallen.com/media/file/MassiveData.pdf.

[16] Google Cloud Connect. http://www.google.com/apps/intl/en/business/officeconnect.html.

[17] Greenplum, Big Data Analytics. http://www.greenplum.com/.

[18] Grossman, R., Gu Y., 2008. Data Mining Using High Performance Data Clouds: Experimental Studies Using Sector and Sphere, *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 920-927, ACM, Las Vegas*, http://sector.sourceforge.net/pub/grossman-gu-ncdm-tr-08-04.pdf

[19] Hadoop (The project). http://hadoop.apache.org/.

[20] HBASE (The Hadoop Database). http://hbase.apache.org/.

[21] Hogan, M. Shared-Disk vs. Shared Nothing, Comparing Architectures for Clustered Databases, http://www.scaledb.com/pdfs/WP_SDvSN.pdf.

[22] IBM Cloud Computing. http://www.ibm.com/cloud-computing/us/en/index.html.

[23] Intel Cloud Computing. http://www.intel.com/content/www/us/en/cloud-computing/intel-s-cloud-computing-vision.html.

[24] Katz, R., Goldstein, P., Yanosky, R., 2010. Cloud Computing in Higher Education, Retrieved http://net.educause.edu/section_params/conf/CCW10/highered.pdf.

[25] Laisheng, X., Zhengxia, W., 2011. Cloud Computing: A New Business Paradigm for E-learning, *Proceedings of ICMTMA*, pp. 716-719.

[26] Lammel, R., 2008. Google's MapReduce Programming Model – Revisited, *Sci. Comput. Program*. 70(1), pp. 1-30.

[27] Law, G., 2011/2012. Cloud in HPC, Scientific Computing World, 121, pp. 20-22.

[28] Lee, B.S., Yan, S., Ma, D., Zhao, G., 2011. Aggregate IaaS service, *Annual SRII Global Conference*, http://www.hpl.hp.com/techreports/2011/HPL-2011-22.pdf.

[29] Microsoft Cloud Computing. http://www.microsoft.com/en-us/cloud/default.aspx?fbid=rRvRp1Ax8Bo

[30] Natis, Y.V., Lheureux B.J., Pezzini, M., Cearley D.W., Plummer, D.C., (2011). PaaS Road Map: A Continent Emerging, Gartner, http://www.gartner.com/resId=1521622.

[31] Netezza. http://www.netezza.com/data-warehouse-appliance-products/cloud.aspx.

[32] Nokia Developers, JavaME UI frameworks, http://www.developer.nokia.com/Community/Wiki/Java_ME_UI_Frameworks.

[33] OpSource, Managed Services. http://www.opsource.net/Services/Cloud-Hosting/Managed-Services.

[34] Oracle Cloud Computing. http://www.oracle.com/us/technologies/cloud/index.html.

[35] ROSTLAB. http://rostlab.org/cms/?id=185.

[36] Pentaho Data Analytics. Big Data Business Analytics. http://www.pentaho.com/big-data/.

[37] Sawzall Language Specification. http://szl.googlecode.com/svn/doc/sawzall-spec.html.

[38] Teradata. http://www.teradata.com/products-and-services/database/.

[39] Torgo, L. 2011. Data mining with R: Learning with case studies, Taylor and Francis Group.

[40] Vaquero, L.M., Rodero-Merino, L., Caceres, J., Lindner, M., 2009. A break in the clouds: Towards a cloud definition. *ACM SIGCOMM Computer Communication Review*, 39(1), pp. 50-55.

[41] Vecchiola, C., Pandey, S., Buyya, R., 2009. High Performance Cloud Computing: A view of Scientific Applications, The 10th International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN), pp. 4-16, http://dx.doi.org/10.1109/I-SPAN.2009.150.

[42] Vertica, Big Data Analytics. http://www.vertica.com/the-analytics-platform/.

[43] Yahoo Cloud Computing. http://labs.yahoo.com/Cloud_Computing

[44] Zaharescu, E., 2012. Enhanced Virtual E-Learning Environments Using Cloud Computing Architectures, IJCSRA, 2(1), pp. 31-41.