International Journal of Information Management 34 (2014) 221–225

Contents lists available at ScienceDirect

International Journal of Information Management

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

Viewpoint

Discovering the potential of cloud computing in accelerating the search for curing serious illnesses

Nabil Sultan*

University Campus Suffolk, School of Business, leadership and Enterprise, Ipswich, Suffolk IP4 1QJ, United Kingdom

ARTICLE INFO

Article history:
Available online 21 January 2014

Keywords:
Cloud computing
High performance computing
DNA
Life science

ABSTRACT

Derided, when it emerged in 2007 as a fad, cloud computing has proved to be a viable business model for remotely delivering IT services through the Web (and other media) on a pay-as-you-go basis. The flexibility of this emerging computing service has opened many possibilities for organizations. Drug companies and medical research centers are among those organizations that are likely to benefit from this new IT service model. As well as providing massive cost-savings, cloud computing could offer these organizations the opportunity to greatly enhance the efficiency of their operations. For cloud providers, this is a new field to expand their reach. The aim of this article is to explore this new development and the potential of cloud computing in contributing to the advancement of research in life science and explain why this IT service model (despite many of its problems) could be game-changer for companies engaged in this business.

1. Introduction

Cloud computing is an IT solution and a business model that uses advances in ICT technologies such as virtualization and grid computing to remotely deliver (on demand) a range of ICT services (e.g., business and development software, processing power, storage) through the Web and other media such as a network infrastructure. There are mainly three types of services that are provided by cloud computing: Software as a Service (SaaS), Infrastructure as a Service (IaaS) and Platform as a Service (PaaS). SaaS involves the provision of software functionality (Google Apps is a good and popular example). IaaS provides a range of infrastructural services (e.g., processing power, storage, virtual servers). PaaS is a platform that enables access to development software and hosting options to developers of Web applications. Cloud computing represents a new business servitization model that is different from those described by servitization authors who saw a service as either a supplement to an existing physical product and/or a service that is based on a supplier using its skills and knowledge (i.e., its competences) to provide clients with a solution (see Baines, Lightfoot, Benedettini, & Kay, 2009; Vandermerwe & Rada, 1988; Vargo & Lusch, 2006, 2008). The cloud paradigm is different in the sense that a physical product (e.g., software and hardware) is transformed into a service.

Cloud computing is underpinned by two main technologies: virtualization and grid computing. Virtualization can be described as an approach for pooling and sharing technology resources to ensure greater efficiency of resources utilization. For example, virtualization can be used to take a single physical asset (e.g., server, storage device or network) and make it operate as if it were many separate, smaller assets. This process improves asset utilization and efficiency, and decreases costs by reducing the need for physical assets. Moreover, virtualization can also be used to combine multiple assets (e.g., storage devices and networks) and present them to servers and applications as if they were a single, larger asset; which simplifies server and application architecture and reduces costs (Sultan & Salim, 2010).

Grid computing is the technology that involves the use of software to combine the computational power of many different (and possibly geographically dispersed) computers, connected in a grid (hence the name "grid computing") in order to provide enhanced computer processing power. Grid computing also uses software that can divide and farm out pieces of a program to as many as several thousand computers. Grid technology, therefore, can be thought of as the technology that enables the establishment of network-distributed parallel processing and distributed and large-scale cluster computing.

2. The cloud and high performance computing

It is interesting to note that cloud computing for high performance computing (HPC) has not been on the priority list of many cloud providers. Until recently, HPC has not been a good candidate for cloud computing due to a number of factors such as its requirement for tight integration between server nodes via

* Tel.: +44(0)7 810377005.
E-mail address: nabil.sultan@yahoo.co.uk

0268-4012/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijinfomgt.2013.12.012

low-latency interconnects and high-speed networking (Shainer et al., 2010). For example, the performance overhead associated with host virtualization, a prerequisite technology for migrating local applications to the cloud, quickly erodes application scalability and efficiency in an HPC context which often involves sending messages back and forth many times per second, a process that is likely to increase the possibility of latency (Niccolai, 2009). However, new virtualization solutions that use KVM (kernel-based virtual machine) and XEN hypervisors[1] have managed to solve the performance issue by reducing the virtualization management overhead through enabling native performance capabilities from the virtual machines (VMs) and by allowing direct access from the VMs to the network. High-speed networking is also an important factor in HPC which requires fast communication between clusters of servers and storage (Shainer et al., 2010).

With the ability of more cloud vendors to provider faster speed and networking connections, cloud providers began to realize the business potential of providing reliable and cost-effective services for HPC-oriented organizations. Several application areas in this category appear particularly suitable for cloud computing. Among those are drug discovery, personalized medicine, translational medicine and genomics (Rubenstein, 2010). Currently, there are few cloud providers that can declare themselves as having cloud solutions for these types of applications. Amazon is no doubt the dominant player in this field. However, it is likely that more cloud providers will be moving into this computing area as more people from the scientific community begin to look at this computing service for solutions to their problems.

3. Life science research and development (R and D)

Advances in personalized medicine hold immense potential for human health. The variation in human genetic make-up causes different responses to drugs in many people. This is often why drugs fail to treat many people as they are only suitable for a small proportion of the entire human population (i.e., those with the "right" genome for a particular drug). The genome is the entire hereditary information of an organism (such as a human). Deciphering an individual's genome is a process known as DNA sequencing which involves determining the order of the nucleotide bases[2] – adenine, guanine, cytosine, and thymine – in a molecule of DNA. This is a complex and time-consuming process; which explains why less than 20 human genomes have been sequenced so far in the entire scientific community (Reid, 2008). However, with the advent of next-generation DNA sequencing machines there is potential for creating individual genetic "maps" that can be used to provide medicines tailored according to each person's genome. However, there is great potential for this process to be advanced through the unlimited processing power and storage of cloud computing.

Next-generation DNA sequencing machines create massive amounts of data and require massive amounts of processing power. Most of the IT infrastructures of pharmaceutical and biotechnology companies are unable to provide this capability. For example, the client-server architectural approaches of these companies are often based on relational databases that are too expensive to develop, deploy and maintain, thus causing these companies to increasingly seek commercial solutions to avoid unforeseen scaling risks and hidden costs. Fig. 1 provides an illustration of how life science researchers can interact with an IaaS cloud infrastructure for processing and storing the massive amounts of data generated by this type of research.

There is no doubt that Amazon is capturing a great deal of the HPC market, thanks to its Amazon Web Services (AWS), its cloud service infrastructure, which includes Elastic Compute (EC2), its main IaaS platform. Google (with its new Google Compute Engine) has recently entered the IaaS market and analysts are already seeing it as a potential and major threat to Amazon in this field (Darrow, 2013). Pharmaceutical and medical research companies and institutes, especially those engaged in genomics research, count among Amazon's cloud customers. For example, in the US, the Broad Institute of MIT and Harvard along with Harvard Medical School are users. Life Technologies, an instrument maker, and Seattle-based Geospiza, a bioinformatics software company, have a partnership to use Amazon's servers to store genomic data. DNAnexus (a Californian bioinformatics startup), has built its business model around using Amazon Web Services (Timmerman, 2010).

However, an increasing number of small ICT companies (especially those servicing the life sciences sector) are entering the cloud market. Among those are GenoLogics, GenomeQuest and Geospiza. For example, GenoLogics focuses its collaborative data-management software platform on biomedical and drug discovery/development applications in the cloud, with emphasis on translational medicine and systems biology in pharmaceutical, biotechnology companies, and academic organizations. GenomeQuest provides a SaaS solution that allows researchers to perform sequence alignment and data mining on next-generation sequencing data. Sequence alignment is a method of comparing two or more nucleotide or protein sequences in order to determine the degree of similarity between them that might be interpreted as the result of functional or evolutionary relationships between them. The solution provides centralized sequence data management resources and applications for biological research and other tasks such as patent research. Geospiza offers a solution called "GeneSifter" for next-generation DNA sequencing in the cloud through Amazon Web Services.

Given the potential of cloud-based genomic research, acquisitions of small innovative companies in this area by large pharmaceutical, biotechnology or ICT companies is likely to accelerate. Recently, Geospiza was bought by Perkin Elmer, the giant maker of tools for life scientists, and CloudSwitch was bought by Verizon, a major global provider of communications services and an emerging cloud vendor.

The highly data-driven and integrative nature of research in genomic medicine presents significant challenges in formulating and testing important translational hypotheses. Advances in high throughput experimental technologies continue to add to the exponential growth in publicly available genomic data. The integration and interpretation of these immense volumes of data toward direct, measureable improvements in patient health and clinical outcomes is a great challenge in genomic medicine due to the scale of the computation and the amount of processing power required (Dudley, Pouliot, Chen, Morgan, & Butte, 2010). On that basis, according to Dudley et al. (2010), cloud computing can potentially offer an efficient and economic means to facilitate this task. In a case study designed to use statistical analysis to discover cancer-associated eQTLs[3] through integration of two high-dimensional genomic data

[1] A hypervisor is a virtual machine monitor (VMM) that allows multiple operating systems to run concurrently on a host computer.

[2] These are molecules that, when joined together, make up the structural units of RNA (Ribonucleic acid) and DNA (Deoxyribonucleic acid). RNA is one of the three major macromolecules (along with DNA and proteins) that are essential for all known forms of life. DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of most living organisms.

[3] eQTLs stands for 'expression quantitative trait loci'. They are genomic loci (i.e., specific locations of a gene or DNA sequence on a chromosome) that regulate expression levels of mRNAs or proteins. A gene expression is the most fundamental level at which the genotype (the genetic makeup of a cell, an organism, or an individual) gives rise to the phenotype (an organism's observable characteristics or traits