Big data using cloud computing

Bernice M. Purcell
Holy Family University

ABSTRACT

Big Data is a data analysis methodology enabled by recent advances in technologies and architecture. However, big data entails a huge commitment of hardware and processing resources, making adoption costs of big data technology prohibitive to small and medium sized businesses. Cloud computing offers the promise of big data implementation to small and medium sized businesses.

Big Data processing is performed through a programming paradigm known as MapReduce. Typically, implementation of the MapReduce paradigm requires networked attached storage and parallel processing. The computing needs of MapReduce programming are often beyond what small and medium sized business are able to commit.

Cloud computing is on-demand network access to computing resources, provided by an outside entity. Common deployment models for cloud computing include platform as a service (PaaS), software as a service (SaaS), infrastructure as a service (IaaS), and hardware as a service (HaaS).

The three types of cloud computing are the public cloud, the private cloud, and the hybrid cloud. A public cloud is the pay- as-you-go services. A private cloud is internal data center of a business not available to the general public but based on cloud structure. The hybrid cloud is a combination of the public cloud and private cloud.

Three major reasons for small to medium sized businesses to use cloud computing for big data technology implementation are hardware cost reduction, processing cost reduction, and ability to test the value of big data. The major concerns regarding cloud computing are security and loss of control.

Keywords: Big data, cloud computing, private cloud, public cloud, hybrid cloud

INTRODUCTION

Big Data is a data analysis methodology enabled by a new generation of technologies and architecture which support high-velocity data capture, storage, and analysis (Villars, Olofson, & Eastwood, 2011). Data sources extend beyond the traditional corporate database to include e-mail, mobile device output, sensor-generated data, and social media output (Villars, Olofson, & Eastwood, 2011). Data are no longer restricted to structured database records but include unstructured data – data having no standard formatting (Coronel, Morris, & Rob, 2013).

Big Data requires huge amounts of storage space. While the price of storage continued to decline, the resources needed to leverage big data can still pose financial difficulties for small to medium sized businesses. A typical big data storage and analysis infrastructure will be based on clustered network-attached storage (NAS) (White, 2011). Clustered NAS infrastructure requires configuration of several NAS "pods" with each NAS "pod" comprised of several storage devices connected to an NAS device (White, 2011). The series of NAS devices are then interconnected to allow massive sharing and searching of data (White, 2011).

Data storage using cloud computing is a viable option for small to medium sized businesses considering the use of Big Data analytic techniques. Cloud computing is on-demand network access to computing resources which are often provided by an outside entity and require little management effort by the business (IOS Press, 2011). A number of architectures and deployment models exist for cloud computing, and these architectures and models are able to be used with other technologies and design approaches (IOS Press, 2011). Owners of small to medium sized businesses who are unable to afford adoption of clustered NAS technology can consider a number of cloud computing models to meet their big data needs. Small to medium sized business owners need to consider the correct cloud computing in order to remain both competitive and profitable.

BIG DATA AND THE CLOUD

The term big data is derived from the fact that the datasets are so large that typical database systems are not able to store and analyze the datasets (Manyika et al., 2011). The datasets are large because the data is no longer traditional structured data, but data from many new sources, including e-mail, social media, and Internet-accessible sensors (Manyika et al., 2011). The characteristics of big data present data storage and data analysis challenges to businesses.

A typical model for in-house storage of big data is clustered Network-Attached Storage (Sliwa, 2011). The configuration would begin with a network-attached storage (NAS) pod consisting of several computers attached to a computer used as the (NAS) device. Several NAS pods would be attached to each other through the computer used as the NAS device. Clustered NAS storage is an expensive prospect for a small to medium size business. A cloud services provider can furnish the necessary storage space for substantially lower costs.

Analyzing big data is done using a programming paradigm called MapReduce (Eaton, Deroos, Deutsch, Lapis, & Zikopoulos, 2012). In the MapReduce paradigm, a query is made and data are mapped to find key values considered to relate to the query; the results are then reduced to a dataset answering the query (Eaton, Deroos, Deutsch, Lapis, & Zikopoulos, 2012). The MapReduce paradigm requires that huge amounts of data be analyzed. The mapping is done concurrently by each separate NAS device; the mapping requires parallel processing. The parallel processing needs of MapReduce are costly, and require the configuration noted previously for storage. The processing needs can be met by cloud-service providers.

CLOUD COMPUTNG SERVICE MODELS

Common deployment models for cloud computing include platform as a service (PaaS), software as a service (SaaS), infrastructure as a service (IaaS), and hardware as a service (HaaS). Cloud deployment solutions can provide services that businesses would otherwise not be able to afford. Businesses can also use cloud deployment solutions as a test measure before adopting a new application or technology company-wide.

There are a wide number of alternatives for businesses using the cloud for PaaS (Géczy, Izumi, & Hasida, 2012). Platform as a Service is the use of cloud computing to provide platforms for the development and use of custom applications (Salesforce.com, 2012). The PaaS solutions include application design and development tools, application testing, versioning, integration, deployment, and hosting, state management, and other related development tools (Géczy, Izumi, & Hasida, 2012). Businesses attain cost savings using PaaS through standardization and high utilization of the cloud-based platform across a number of applications (Oracle, 2012). Other advantages of using PaaS include lowering risks by using pretested technologies, promoting shared services, improving software security, and lowering skill requirements needed for new systems development (Jackson, 2012). As related to big data, PaaS provides companies a platform for developing and using custom applications needed to analyze large quantities of unstructured data at a low cost and low risk in a secure environment.

Software as a service provides businesses with applications that are stored and run on virtual servers – in the cloud (Cole, 2012). The business is not charged for hardware, only for the bandwidth for the time and number of users necessary (Cole, 2012). The main advantage of SaaS is that the solution allows businesses to shift the risks associated with software acquisition while moving IT from being reactive to proactive (Carraro & Chong, 2006). Benefits of using SaaS are easier software administration, automatic updates and patch management, software compatibility across the business, easier collaboration, and global accessibility (Rouse, 2010a). Software as a Service provides companies analyzing big data proven software solutions for data analysis. The difference between SaaS and PaaS in this case is that SaaS is not going to provide a customized solution whereas PaaS will allow the company to develop a solution tailored to the company's needs.

In the IaaS model, a client business will pay on a per-use basis for use of equipment to support computing operations including storage, hardware, servers, and networking equipment (Rouse, 2010b). Infrastructure as a service is the cloud computing model receiving the most attention from the market, with an expectation of 25% of enterprises planning to adopt a service provider for IaaS (Cisco, 2009). Services available to businesses through the IaaS model include disaster recovery, compute as a service, storage as a service, data center as a service, virtual desktop infrastructure, and cloud bursting, which is providing peak load capacity for variable processes (Cisco, 2009). Benefits of IaaS include increased financial flexibility, choice of services, business agility, cost-effective scalability, and increased security (Cisco, 2009).

While not as yet being used as extensively as PaaS, SaaS, or IaaS, HaaS is a cloud service based upon the model of time sharing on minicomputers and mainframes from the 1960s and 1970s (ComputerWeekly.com, 2009). Time sharing developed into the practice of managed services (ComputerWeekly.com, 2009). In a managed services situation, the managed service provider (MSP) would remotely monitor and administer hardware located at a client's site as contracted (Rouse, 2007). A problem with managed services was the necessity for some MSPs to provide hardware on-site for clients, the cost of which needed to be built into the MSP's cost (Rouse, 2007). The HaaS model allows the customer to license the hardware directly from the service provider which alleviates the associated costs (Rouse, 2007). Vendors in the HaaS arena include Google with its Chromebooks for Business, CharTec, and Equus (Panettieri, 2011).

TYPES OF CLOUDS

Three types of clouds exist – the public cloud, the private cloud, and the hybrid cloud. A public cloud is the pay- as-you-go services previously discussed available to the general public (Armbrust et al., 2010). In a public cloud configuration, a business does not own the core technology resources and services but outsources these (Géczy, Izumi, & Hasida, 2012). A public cloud is considered to be an external cloud (Aslam, Ullah, & Ansara, 2010).

A private cloud is internal data center of a business that is not available to the general public but uses cloud structure (Armbrust et al., 2010). In a private cloud configuration, resources and services are owned by the business, with the services accessible within the business through the intranet (Géczy, Izumi, & Hasida, 2012). Since the technology is owned and operated by the business, this type of cloud is more expensive than a public cloud, but is also more secure (Géczy, Izumi, & Hasida, 2012). A private cloud is an internal cloud, residing inside the company's firewall and managed by the company (Aslam, Ullah, & Ansara, 2010).

When a company uses a hybrid cloud, it uses a public cloud for some tasks and a private cloud for other tasks. When using a hybrid cloud model, a company will use the public cloud to expedite extra tasks that are not able to be easily run in the company's data center or on its private cloud (Armbrust et al., 2010). A hybrid cloud allows a company to maintain critical, confidential data and information within it firewall while leveraging the public cloud for non-confidential data (Aslam, Ullah, & Ansara, 2010). Figure 1 illustrates a hybrid cloud. The private cloud portion of the hybrid cloud is accessed by company employees, both in the company and on the road, and is maintained by the internal technology group. The private cloud part of the hybrid cloud is also accessed by the company employees but is maintained by external service providers. Each portion of the hybrid cloud can connect to the other portion.

WHICH CLOUD FOR YOUR DATA?

The type of cloud a company uses depends upon the company's needs and resources. The public cloud is considered the least secure of the three types, with services and resources able to be accessed over the Internet through protocols adopted by the provider (Géczy, Izumi, & Hasida, 2012). The communications protocols adopted by the provider are not necessarily secure; the choice of using secure or non-secure protocols is up to the provides (Géczy, Izumi, & Hasida, 2012). The public cloud is also the least costly of the cloud types, with cost savings in the areas of information technology deployment, management, and maintenance (Géczy, Izumi, & Hasida, 2012).

The private cloud provides services to company employees through an intranet (Géczy, Izumi, & Hasida, 2012). If mobile employees are able to access the private cloud, the access is typically through secure communication protocols (Géczy, Izumi, & Hasida, 2012). All services and resources provided are tailored to the needs of the business, and the business has total control over the services and resources (Géczy, Izumi, & Hasida, 2012). Due to the financial and human resources needed to deploy, manage, and maintain the information technology resources and services provided, the private cloud is the most expensive type of cloud (Géczy, Izumi, & Hasida, 2012).

When a business uses a hybrid cloud, the business owns its core information technology resources and services and will host and provide the resources and services in-house (Géczy, Izumi, & Hasida, 2012). Non-critical services are outsourced and maintained on a public cloud (Géczy, Izumi, & Hasida, 2012). Typically, core information technology resources and services are mission-critical and are often confidential (Géczy, Izumi, & Hasida, 2012). Therefore, resources and services that need to be secure are hosted and maintained on the private cloud, with the public cloud used for other services as a cost saving measure (Géczy, Izumi, & Hasida, 2012).

CLOUD COMPUTING FOR BIG DATA IN A SMALL TO MEDIUM SIZED BUSINESS

Cloud computing provides an environment for small to medium sized businesses to implement big data technology. Benefits that businesses can realize from big data include performance improvement, decision making support, and innovation in business models, products, and services (Manyika et al., 2011). Three major reasons for small to medium sized businesses to use cloud computing for big data technology implementation are the ability to reduce hardware costs, reduce processing costs, and to test the value of big data before committing significant company resources. The major concerns regarding cloud computing are security and loss of control (Géczy, Izumi, & Hasida, 2012).

Platform as a Service is a cloud computing model that provides hardware cost savings. Hardware cost savings are accrued using PaaS through standardization and high utilization of the cloud-based platform across a number of applications (Oracle, 2012). Businesses can also realize hardware cost savings from the SaaS model since the business incurs no additional hardware costs for implementation; the only costs are for bandwidth based on the time and number of users (Cole, 2012). Hardware as a Service is not currently used as often as other models, but businesses can derive hardware cost savings through the model since HaaS allows customers to license the hardware directly from the service provider (Rouse, 2007).

In-house processing of big data typically requires use of the MapReduce programming paradigm (Eaton et al., 2012). The parallel processing needs of MapReduce entails a huge commitment of processing power. Use of cloud computing for big data implementation lowers the in-house processing power commitment by shifting the data processing to the cloud.

The use of big data could provide sufficient benefit to a small to medium sized company to the extent that the business would be willing to commit resources to implement big data technology in-house. However, the level of benefit is difficult to determine without some experience. Cloud computing implementation of big data processing could provide the business with justification to adopt the technology in-house. If the benefit accrued from big data use on the cloud is significant, the business has established a reason to adopt the technology in house. Otherwise, the business can continue cloud computing use of big data or rely on its current data processing environment.

The advantages of cloud computing are tempered by two major concerns – security and loss of control (Géczy, Izumi, & Hasida, 2012). While the public cloud provides the greatest costs savings, it also incurs the greatest security risk and loss of control, since all of the company's big data is transferred to the cloud service provider (Géczy, Izumi, & Hasida, 2012). If the data being processed is considered mission critical to the company, the more expensive private cloud, implemented in-house, would provide a more secure environment with the company keeping the mission critical data in-house.

CONCLUSION

Cloud computing enables small to medium sized business to implement big data technology with a reduced commitment of company resources. The processing capabilities of the big data model could provide new insights to the business pertaining to performance improvement, decision making support, and innovation in business models, products, and services. Benefits of implementing big data technology through cloud computing are cost savings in hardware and processing, as well as the ability to experiment with big data technology before making a substantial commitment of company resources. Several models of cloud computing services are available to the businesses to consider, with each model having trade-offs between the benefit of cost savings and the concerns data security and loss of control.

REFERENCES

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G…Zaharia, M. (2010, April). A view of cloud computing. *Communications of the ACM*, 53(4), 50-58. DOI: 10.1145/1721654.1721672.

Aslam, U., Ullah, I, & Ansara, S. (2010, November). Open source private cloud computing. *Interdisciplinary Journal of Contemporary Research in Business. 2*(7), 399-407.

Carraro, G., & Chong, F. (2006, October). Software as a service: An enterprise perspective. Retrieved from http://msdn.microsoft.com/en-us/library/aa905332.aspx #enterprisertw_topic3

Cisco. (2009). Infrastructure as a Service: Accelerating time to profitable new revenue streams. Retrieved from http://www.cisco.com/en/US/solutions/collateral/ns341/ns991/ns995 /IaaS_BDM_WP.pdf

Cole, B. (2012). Looking at business size, budget when choosing between SaaS and hosted ERP. *E-guide: Evaluating SaaS vs. on premise for ERP systems.* Retrieved from http://docs.media.bitpipe.com/io_10x/io_104515/item_548729/SAP_sManERP_IO%23104515_EGuide_061212.pdf

ComputureWeekly.com. (2009, March). Hardware as a service. Retrieved from http://www. computerweekly.com/feature/Hardware-as-a-Service

Coronel, C., Morris, S., & Rob, P. (2013). *Database Systems: Design, Implementation, and Management*, (10th Ed.). Boston: Cengage Learning.

Eaton, Deroos, Deutsch, Lapis, & Zikopoulos. (2012). *Understanding big data: Analytics for enterprise class Hadoop and streaming data.* New York: McGraw-Hill.

Géczy, P., Izumi, N., & Hasida, K. (2012). Cloudsourcing: Managing cloud adoption. *Global Journal of Business Research, 6*(2), 57-70.

IOS Press. (2011). Guidelines on security and privacy in public cloud computing. *Journal of E-Governance, 34* 149-151. DOI: 10.3233/GOV-2011-0271

Jackson, K. L. (2012). Platform-as-a-service: The game changer. Retrieved from http://www.forbes.com/sites/kevinjackson/2012/01/25/platform-as-a-service-the-game-changer/

Manyika, J., Chui,. M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011, June). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Oracle. (2012). Oracle platform as a service. Retrieved from http://www.oracle.com/us/technologies/cloud/oracle-platform-as-a-service-408171.html

Panettieri, J. (2011, June 13). Can Google take hardware as a service (HaaS) mainstream? *MSPMentor*. Retrieved from http://www.mspmentor.net/2011/06/13/can-google-take-hardware-as-a-service-haas-mainstream/

Rouse, M. (2010a, August). Software as a service. Retrieved from http://searchcloudcomputing.techtarget.com/definition/Software-as-a-Service

Rouse, M. (2010b, August). Infrastructure as a Service. Retrieved from http://searchcloudcomputing.techtarget.com/definition /Infrastructure-as-a -Service-IaaS

Rouse, M. (2007, December). Hardware as a service. Retrieved from http://searchitchannel.techtarget.com/definition/Hardware-as-a-Service-in-managed-services

Salesforce.com. (2012). The end of software: Building and running applications in the cloud. Retrieved from http://www.salesforce.com/paas/

Sliwa, C. (2011, June 16). Scale-out NAS, object storage, cloud gateways replacing traditional NAS. Retrieved from http://searchstorage.techtarget.com/feature/Scale-out-NAS-object-storage-cloud-gateways-replacing-file-storage

Villars, R. L., Olofson, C. W., & Eastwood, M. (2011, June). Big data: What it is and why you should care. *IDC White Paper*. Framingham, MA: IDC.

White, C. (2011). *Data Communications and Computer Networks: A business user's approach*, (6th ed.). Boston: Cengage Learning.

Appendix A

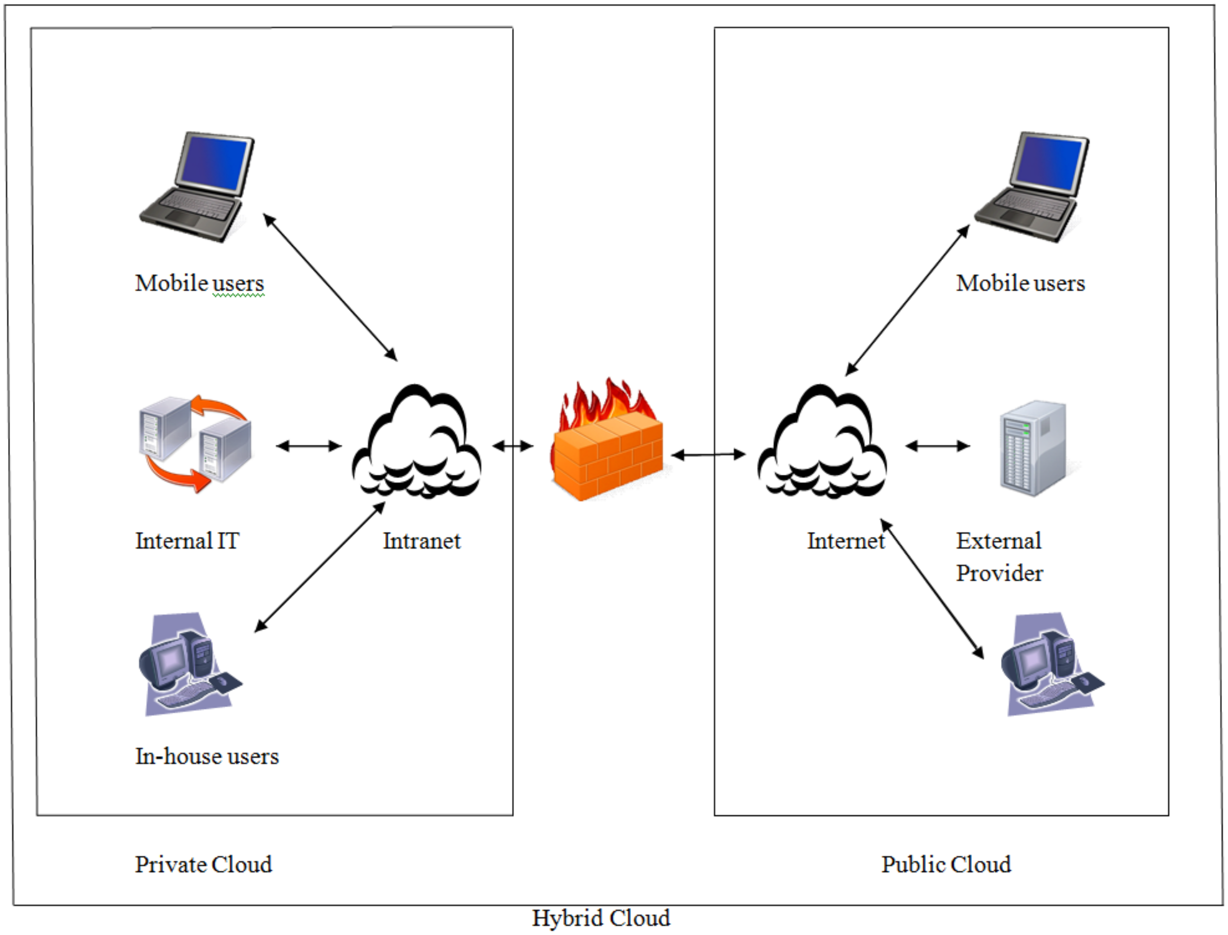

Figure 1 Hybrid Cloud