ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 6, Issue 7, January 2017

Presentation of a model for big data services based on cloud

Shahab Shakib, Pardis Shoumali

Abstract— ***Constant detailed data flux in organizations such as the emergence of social media, IoT, and multimedia has produced a slew of organized and disorganized data and management paradigms. Big data management may alter the definition of business, financial, engineering, science and health related products thus effecting the entire society. The advancement of data storage and technology results in correct analysis of real data which maximizes service user satisfaction and overall social welfare. Managing such big data entails vast investments and utilization of human resources. Cloud computing technology has enabled modern technology usage in the form of services. In this paper, a comprehensive assessment of big data aspects in cloud computing environments is presented which includes big data definition, characteristics, and categorization along with cloud computing topics. The relationship between big data, cloud computing and big data storage are also assessed.***

I. INTRODUCTION

Today, business firms are more capable compared to previous decades in terms of aggregating and extracting value from data. Corporations aim to make the most effective use of data to create value within their organizations. Data utilization is not limited to corporations since relevant and accessible data may also be utilized. Big data is promising a revolution in all industries. Our abilities in extracting economic and social value from new and current data depends on various factors. Working with such high volume of data requires modern tools and skills. Information is so vast that it cannot be entered into computers for analyzing with traditional statistics and database tools which are then presented by standard graphics software. Lack of expert human resources is one of the challenges that exist in utilizing opportunities stemming from big data. A traditional and limited approach will not be enough in creating value from data in the competitive modern age. Resolving new problems with a traditional perspective and utilizing relevant ineffective tools which were useful at a time, will eliminate advantageous opportunities. Such opportunities may not be compensated once missed since competitors are constantly advancing.

Cloud computing is one of the most prominent changes in modern ICT and corporate strategies which has turned into a powerful structure for large scale computations. Cloud computing does not only decrease costs relevant to automation limitations in companies, but also decreases maintenance and repair costs, provides efficient management, and optimizes accessibility for users. A major challenge that arises here is the growth rate for designing suitable substrates for big data analysis and results utilization.

II. BIG DATA DEFINITION AND ATTRIBUTES

Big data refers to a magnitude of data that exceeds the typical amount of data that can be analyzed, processed and managed by everyday software in a reasonable time. The “magnitude” concept in big data is continuously changing and increasing in size. Corporate and company capacities and capabilities in data management are also involved in this definition. A few terabytes may be considered big data for one company but for other companies, big data may be referred to as tens or hundreds of Exabyte. Regarding big data, data should be managed efficiently for information extraction, knowledge discovery and decision making for various practical issues. Data management commonly consists of 5 main activities; collection, storage, search, sharing and analysis.

Big data consists of a set of techniques and technologies to discover hidden values from a set of vast, varied and complex data on a large scale. So far, many challenges have risen in the big data field and are presented in the form of theories and assessed in various dimensions. These challenges first arise in three dimensions; data volume, growth velocity and variety as the three Vs’, although more challenges arise further on:

- Volume: the quantity of continuously produced data from various sources. The advantage of collecting vast data is to create hidden information and discover patterns by data analysis. Data volume is increasing exponentially. Various sources such as social networks, web server logs, traffic information, satellite images, audio information, bank transactions, web page content, government documents etc. exist which create high volumes of data.
- Variety: various data types exist that stem from various data structures. Various data types may be information collected from sensors, smart phones or social networks. Various data include video, image, text, audio and log data in the form of structure or unstructured data. For example, regarding the internet, users use various software and browsers to send information. A lot of the information is received form humans therefore mistakes are inevitable. This variety effects data integrity because the more varied the data, the higher chances of mistakes occurring.
- Velocity: this refers to the speed of data transfer. Data content changes continuously due to the addition of

supplementary data, the introduction of previously archived data and received data from multiple sources. Data is produced rapidly in real time via applications and sensors. On many occasions, users expect responses upon data entry. In other words, a response must be generated for users as soon as data is entered.

- Value: the most important aspect of big data. Magnitude is referred to as the discovery of a vast set of varied hidden data produced at a rapid rate. This is a prominent issue when it comes to decision making for valuable data. In other words, whether or not data processing and maintenance is worth the decision making process. Typically, data may be relocated on different layers. Higher layers mean higher value. Therefore, some corporations may accept costs related to the maintenance of higher level data.
- Veracity: considering the fact that data is received from various sources, they may not be reliable. For example, regarding social networks, many comments may be presented for a given issue but whether or not all these comments are true or false is an issue that cannot be easily ignored when it comes to high volume data. Although, some research address this challenge by maintaining the characteristics of the main data in order to guarantee veracity. However, the second definition is credible for vast data generation to present characteristics of the main data.
- Validity: assuming that the data is correct, they may not be suitable for some applications or have the necessary credibility for utilizing in some applications.
- Volatility: the rate of various data value changes with time. In a typical e-business system, volatility speed is low and data value may be preserved for as long as a year but in other applications such as stock market applications, data is highly volatile and loses value while being replaced by new values. Although information maintenance during long periods of time for data volatility analysis is of great importance, increases in information maintenance periods entails higher implementation costs which must be taken into account.
- Visualization: one of the most challenging tasks in the big data field is information presentation. Information from vast complex data should be clearly presented. This is achieved by using suitable visual effects and analysis techniques.

One of big data's most prominent characteristics is that there is no structure and organization. Hence, it is necessary to run dozens of software and applications concurrently on hundreds or thousands of installed servers. This issue has led researchers and scientists to pursue new structures, methodologies, techniques and approaches to manage, control, and process such high volume of data. These efforts have been developed under the "large data" category.

Cloud computing platforms are one of the latest technologies to aid in large scale data processing and management in a reasonable time period. This is because the analysis of large scale data requires running hundreds or thousands of computers simultaneously which in turn, requires extensive investments by companies.

III. CLOUD COMPUTING

Cloud computing is still a growing concept. Its definition, application, fundamental technologies, and its advantages are discussed and improved by private and general sectors. These definition, attributes and properties are improving and changing with time. The expansion of this industry is so rapid that every year many companies are added to list of companies providing such services. Companies that provide cloud computing services have their own implementation methods and platforms. Therefore, it can state that the cloud computing industry is a large ecosystem of many models, service providers and markets.

Cloud computing is a model for ease of access to a set of flexible configured resources based on network order (such as networks, servers, storage environments, applications and services) in order to provide (or release) services in the most efficient manner possible. One of the advantages of cloud computing is that virtual resources, parallel processing, data security and service integration are scalable according to data storage.

IV. THE RELATIONSHIP BETWEEN CLOUD COMPUTING AND VIRTUALIZATION

Virtualization and cloud computing are two expressions used to optimize information technology infrastructures. In most cases, when there is talk of cloud computing, virtualization usually follows. Virtualization utilizes server hardware to create multiple virtual servers to satisfy user requirements. If these concepts were to be presented in a multi-layer structure, the first layer would consist of a SAN storage, the second layer would include server hardware and the third layer would consist of a virtual host server.

Virtual software such as Citrix, VSPhere, VMware, Microsoft Hyper-V and Sun xVM are some of the software active at the highest virtualization layer and play the role of host servers. Host servers may include any operating system based on requirements. Virtualization is in fact a technique for optimized use of hardware resources and a means to decrease relevant inconveniences and costs. Virtual servers provide the same services as specialized servers installed in real environments. This technique is known as soft virtualization. Another technique called hard virtualization includes the formation of a specialized supplementary server for the

DOI:10.17605/OSF.IO/DT5C3

virtual server. The concept of cloud computing is the utilization of virtual servers via virtualization techniques with resources such as operating systems, application software, and various services for network users in a way that users do not realize which servers at various physical locations are providing these services. In a cloud environment a user has no idea of the number of core operating systems, storage environments, data and processing capabilities provided by servers. Users only request their services from the cloud environment and receive them accordingly. The most prominent factor is that cloud computing uses virtualization techniques for achieving its objectives. The business aspect of this process is that services are provided for users without mandatory physical servers or software located at specific locations with relevant communication methods. All these services are provided virtually in the form of cloud computing and costs are decreased by renting required services. In other words, users can utilize off site hardware and there are no maintenance costs. Corporations can estimate their software and hardware requirements and order their services from cloud computing providers.

For a clearer understanding of the difference between virtualization and cloud computing, it is necessary to emphasize the fact that virtualization is a technique and cloud computing is a concept that utilizes virtualization techniques. Virtualization is normally used within organizations but cloud computing is used online via the internet.

V. THE RELATIONSHIP BETWEEN BIG DATA AND VIRTUALIZATION

The management of high volume distributed data is required alongside data computing applications to face big data challenges. Virtualization provides additional levels of productivity to turn big data platforms into a reality. Even though technically, virtualization is not a requirement for big data analysis, software frameworks in virtual environments are more efficient. Virtualization has three characteristics that support big data environments in terms of scalability and operational efficiency:

- Classification: in virtualization environments, many applications and operating systems are supported by classifying resources in physical systems.
- Isolation: each virtual machine is isolated from its physical host and other virtual machines. Due to the isolation feature, if one virtual machine fails, other machines and hosting systems will not be effected. In addition, data is not shared between virtual machines.
- Packaging: a virtual machine may be shown as a single file therefore it can be identified based on its providing services.

Hereon, various types of big data virtualization will be presented in regard to the relationships between cloud computing, virtualization and big data.

A. Big data server virtualization

In server virtualization, a physical server is divided into multiple virtual servers. Machine resources and hardware consisting of a RAM, CPU, hard drive, and network controller can be virtualized by a set of virtual devices with their own application programs and operating systems. A virtual machine is the presentation of a physical machine with the same responsibilities. A thin layer of software is added to hardware consisting of a virtual machine supervisor and hypervisor. Server virtualization utilizes a hypervisor for efficient use of physical resources. However, installation, configuration and management responsibilities are coordinated with virtual machine implementation. Server virtualization provides reliability for addressing high volume varied big data analysis which entails enhanced scalability. The volume of data to be analyzed may not be clear. This lack of clarity increases the need for server virtualization such that unpredicted requests for large scale data processing can be addressed.

In addition, server virtualization provides a foundation for cloud services as data resources in analyzing big data. Virtualization enhances cloud efficiency which simplifies the optimization of complex systems.

B. Big data applications virtualization

Infrastructure application virtualization is an efficient method to manage applications based on customer demand. Applications are packaged such that its dependencies on physical computer systems are hidden. This aids in the improvement of transferrable application management. In addition, infrastructure applications virtualization software enables technical and business policies and applications to ensure reliability concerning predictable virtual and physical resources. The required output is achieved because users may distribute IT resources based on relative business values of applications.

Infrastructure application virtualization combined with server virtualization may aid in increasing reliability in regard to service level agreements. Server virtualization supervises processor and memory utilization but does not determine changes in business prioritizations while allocating resources.

C. Big data network virtualization

Network virtualization is an efficient method to utilize networking as a pool of connected resources. Instead of relying on a physical network to manage traffic, users may utilize various virtual networks that use physical implementations. If there is a need for a network to collect data with specific operating characteristics and another network for applications with different operations and capacities, this method will be useful. Network virtualization eliminates bottlenecks and enhances distributed big data management for big data analysis.

D. Big data and storage virtualization

Data virtualization may be used to create a platform for dynamically linked data services. This will provide users with a search option via the unified reference source. Therefore, data virtualization provides an abstract service where integrated data is provided without considering its fundamental physical base. In addition, data virtualization reveals cached data for applications to improve performance.

Storage virtualization combines physical storage resources to enable more efficient sharing. This will decrease storage costs and simplifies big data analysis.

VI. THE RELATIONSHIP BETWEEN CLOUD COMPUTING AND BIG DATA

Cloud computing and big data are intertwined. Big data enables commodity computing to process distributed queries on multiple data sets and returns results to users. Cloud computing provides this fundamental engine which is a class of distributed data processing platforms.

The results from big data extraction facilitates corporation personnel in the decision making process. This will also provide corporations with a competitive advantage. For example, many companies analyze collected big data and transfer products from warehouses to nearby client locations before orders are placed. When an order is placed, clients can expect to receive their orders within 5 minutes. In the transport field, the time it takes for taxis to arrive will be significantly decreased. Clients will not notice the process but will definitely notice the difference between such quick services compared to traditional services.

Today, many large companies utilize big data technology to create competitive advantages. By using big data analysis, crime can also be predicted and measures can take place to prevent crimes such as terrorist activities.

Some of the techniques used in big data management include extensive resources management, parallel computing, and distributed computing. In parallel computing, processing stages are completed at the same time. This usually takes place on common hardware platforms or across customized or conjunct networks. This method is also used in cloud computing implementation. In distributed computing, processing is divided on multiple computers. Each computer works individually and sends results for aggregation. This takes place in typical private networks but can also take place in general networks. Some of the data analysis takes place on thousands of distributed computers in order to take advantage of idle times.

In cloud computing, users utilize a distributed architecture over remote or virtual facilities. Emphasis is on remote maintenance and conformity with various APIs which vendors have access to (data security, authentication, payment, etc.) cloud and distributed computing have common capabilities but differ in the fact that users can rent, borrow or own the data center.

Distributed computing can be defined as the utilization of a distributed system to solve major issues by dividing roles. Each role is defined on separate computers from the distributed system. A distributed system consists of more than one independent computer which is connected to the system via a network. All computers placed in this network are connected to achieve a common goal by utilizing local memory. Users of a computer may have different requirements and the distributed systems maintain coordination between shared resources and aids them to achieve individually defined objectives.

Big data on cloud platforms and websites stored on distributed databases can process parallel big data in one cluster using a distributed algorithm and programming model. The main objective of big data management is to visualize and provide correct data analysis. Analysis results are presented in the form of various graphs which aid in making decisions.

Virtualization and parallel processing are fundamental technologies in implementing cloud computing. The basis of platform characteristics for providing access, storing and managing computing components in a macro environment stem from virtualization. Virtualization is the process of sharing resources and separating hardware to improve the utilization of computer resources, increase output and increase scalability.

Cloud computing is a style of computing where vast scalable and flexible capabilities relating to the IT field are provided as services via the internet. These services include: infrastructure, platforms, and storage programs. Users pay for these resources and services. They do not need to create the required infrastructure. In addition, it is necessary to state a few reasons for big data management on cloud computing platforms to satisfy the modern world’s needs.

When there is talk of big data, the first issue that emerges is investments to purchase data collection infrastructures and data storage facilities. Today, instead of investing in servers, purchasing information storage facilities such as SAN storage, procuring rack spaces, purchasing network equipment, renting bandwidth, maintenance costs and paying for consultancy services to analyze data, companies can order their required services from a cloud server via cloud computing. Here, there is no need for corporations to dedicate costs for infrastructure facilities and maintenance costs will decrease substantially. Although, it depends whether the corporations utilize public or private clouds which differ in terms of services costs. The business aspect of big data considers high costs relating to purchasing high standards tools to search and analyze data. Correct analysis is required to achieve valuable and significant results.

As previously mentioned, storage, classification, management and processing data are prerequisites for big data analysis and are requirements for expanding cloud computing services. Today, advanced cloud services are facing the growing demand for big data analysis. Big data infrastructure services, visual reports resulting from analysis and any XaaS service can be provided by sending requests to distributed platforms. Analysis results are promptly generated and various fields benefit from utilizing these results.

The high volume of required resources for big data analysis along with relevant high equipment installation costs, severely decrease incentives of utilizing complex and rapid processing methods. By requesting big data based services, users can satisfy requirements for even a short period of time since the utilization of resources increases or decreases according to requirements.

VII. PRESENTATION METHODS

In this section, a model for presenting big data management services on a cloud computing platform will be presented.

A. Cloud software as a service SaaS

An application program is provided for clients which is run on a cloud infrastructure and is accessible by various client machines via a connector for weaker clients such as web browsers (such as web emails). This service is visually displayed on the data layer along with various reports. Clients do not control or manage high volume cloud data infrastructure, servers, operating systems, underlying storage space or application software. Only user level configuration settings can be modified by users.

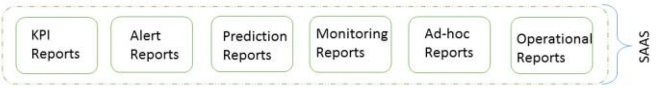

Fig 1-SaaS layer

B. Cloud platform as a service PaaS

In this layer clients are able to place their created applications on the cloud infrastructure. This program is created by using programming languages and tools that are supported by service providers (such as java, python, and .NET). Cloud infrastructure clients do not control the network, servers or underlying storage space but may configure the application program and host environment.

In this layer suitable tools are used for data display and utilization. Stored data clouds cannot delay in displaying information otherwise some of the data may become useless. Some of the applications may require prompt information for sudden decisions. Query generating systems for data should be able to manage results in the form of tasks and sub-tasks and send computed information to the storage facility while maintaining a performance standard. Tools and platforms used in this layer may include monitoring analytic, data mining engine, statistical analytic etc.

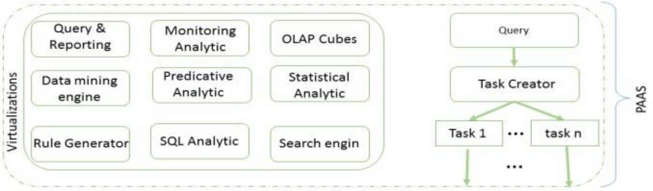

Fig 2-PaaS layer

C. Cloud infrastructure as a service IaaS

As previously mentioned, due to the rapid data increase, storage software must be scalable and flexible so there is no need to deactivate the system as a whole. In this layer, due to business requirements, the selection of any type of database is possible. Correct database selection enhances efficiency. For example, there may be a need for a business to determine the correlation between the number of "likes" on a social network and product sales. It may be more useful to utilize a graph database, or in other cases a column database to display information in columns instead of relational databases. This type of database has high information entry and extraction speed. When geographical or locational data is used it is best to use spatial or locational databases. These databases store data based on location and geographical relations.

Cloud environments provide processing power, storage space, networks and other fundamental computing resources. Distributed resources may also be used (such as signifying data to decrease data volume) and data is finally stored in the data warehouse. Platforms for big data may be used to transfer computations to the storage facility without transferring data such as virtual servers or parallel equipment. Results are then sent for displaying purposes. In this layer, clients do not manage or control the underlying cloud infrastructure but have access to operating system, storage space, application and network component configurations (such as firewalls, load balancing). Measures are taken in this layer for physical tools such as various servers which must be redundant. This means that these tools should tolerate errors and failures.

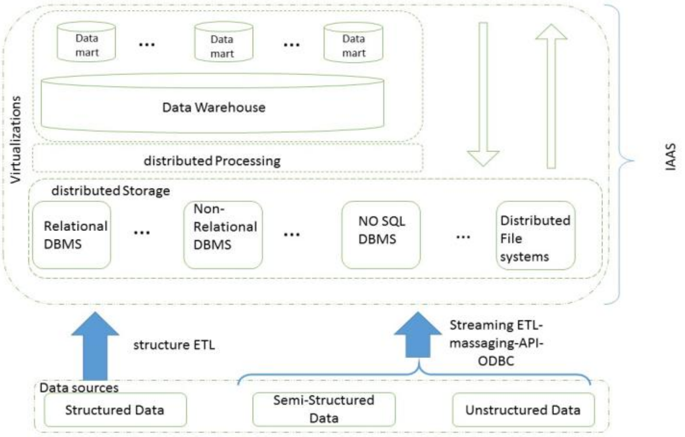

Fig 3-IaaS layer

VIII. CONCLUSION

Big data utilize distributed storage technology based on cloud computing instead of local storage connected to a computer or electronic device. Big data evaluation is derived using cloud based application programs with high growth rate via virtual technologies. Therefore, cloud computing not only provides equipment for big data processing, but also provides services as a service model. In this paper, the dependencies of these technologies is assessed upon explanations provided for cloud computing, big data and virtualization techniques. Finally, a model is presented for providing big data management services on a cloud platform.

APPENDIX

Appendixes, if needed, appear before the acknowledgment.

REFERENCES

[1] R. Cumbley, P.Church, Is Big Data creepy? Comput. Law Secur. Rev. 29 (2013) 601–609.

[2] R. Chow, P. Golle, M. Jacobson, E. Shi, J. Staddon, R. Masouka, and J. Molina, "Controlling Data in the Cloud: Outsourcing Computation without Outsourcing Control," presented at the ACM Cloud Computing Security Workshop, Chicago, Illinois, USA., 2009.

[3] J.J. Berman, Introduction, in: Principles of Big Data, Morgan Kaufmann, Boston, 2013, xix–xxvi (pp).

[4] G. Rowel, "Virtualization: The next generation of application delivery challenges," 2009.

[5] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, Big table: a distributed storage system for structured data, ACM Trans. Compute. Syst. (TOCS) 26 (2008) 4.

[6] W. Itani, A. Kayssi, A. Chehab, Privacy as a Service: Privacy-Aware Data Storage and Processing in Cloud Computing Architectures, Dependable, Autonomic and Secure Computing, 2009. DASC '09, in: Proceedings of the Eighth IEEE International Conference on, 2009, pp. 711–716.

[7] Khan, Abdul Nasir, et al. BSS: block-based sharing scheme for secure data storage services in mobile cloud environment. The Journal of Supercomputing (2014) 1–31.

[8] R. Sravan Kumar, A. Saxena, Data integrity proofs in cloud storage, in: Proceedings of the Third International Conference on Communication Systems and Networks (COMSNETS), 2011, pp. 1–4.

[9] L. Chang, R. Ranjan, Z. Xuyun, Y. Chi, D. Georgakopoulos, C. Jinjun, Public Auditing for Big Data Storage in Cloud Computing - a Survey, Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on, 2013, pp.1128–1135.

[10] D. Talia, Clouds for scalable big data analytics, Computer 46 (2013) 98–101.

[11] D. E. Y. Sarna, Implementing and Developing Cloud Computing Applications: Taylor and Francis Group, LLC, 2011.

[12] L. Hao, D. Han, IEEE Conference on The study and design on secure-cloud storage system, Electrical and Control Engineering (ICECE), 2011 International 2011, pp. 5126–5129.

AUTHOR BIOGRAPHY

Shahab Shakib born in 1986 and graduated with a B.Sc. degree in Computer Software engineering from Islamic Azad University, Karaj Branch, Iran in 2011. He has more than 6 years of experience as a Network Expert and Network manager in several companies. He is also a member of Researches Group in IT companies. He currently manages data center projects and teaches IT engineer management and expert courses.

Pardis Shoumali born in 1986 and graduated with a B.Sc. degree in IT from the Iran University of Science and Technology and Software Engineering from Islamic Azad University, South Tehran Branch, Iran in 2010. Upon completing her studies, she worked for AFACO (Aria Fan Abzar) as a Software Developer, MABNA Consultants Co as a BI/DW manager and then TIDM (TOSAN Intelligent Data Miners) as a Head of DW Development Team.

DOI:10.17605/OSF.IO/DT5C3