Big Data & Virtualization: Concept familiarization and relation between them

accelerate the insight of Big Data and Virtualization with a laconic concept and significant overview

[1]Selina Sharmin, [2]Asoke Datta, [3]Md. Nurain Haider
[1]Senior Lecturer, [2]Graduate Student, [3]Student (final year at CSE)
[1]Dept. of Computer Science and Engineering,
[1]Leading University, Sylhet, Bangladesh

Abstract—One of the most popular catchphrases in the current world is Big data. It is used to describe the gargantuan bulk of both structured and unstructured data which is too large to process using traditional data-handling and software techniques. To solve this problem, virtualization has brought a greater range of quantifiable amenities in recent decades. The main purpose of the literature is to familiarize and find out the relationship between the two novel concepts - Big Data and Virtualization in an easily perceived manner. It begins by depicting the concept of both with their definitions and properties. Then, the main focus has been given to some of the applications of Big Data that use Virtualization in all diverse aspects of different fields have also been elucidated in a palpable fashion. This paper can be very handy for those who want to know the basics of the mentioned topics. The authors intend to decipher the insight in an intelligible manner embodying in smooth and to-the-point illustrations.

Index Terms—Big Data, Virtualization, Relation between Big Data and Virtualization.

What is Big Data

Typically, Big Data works with the data that are begotten from the content of digital systems such as - online transactions, emails, logs, posts, search queries, health records, social networking interactions, sensors, videos, audios, images, click streams, and cellular phones and their applications [3]. It encompasses the function of storing, operating, analyzing and visualizing potentially immense datasets in an agreeable timeframe that is unapproachable to standard IT technologies. In a broader range, the platform, equipment, and software used for this goal are collectively called "Big Data technologies". It is a relative term which deals with the volume, velocity, and variety of any data that can surpass the storage, computing capabilities or precise as well as timely decision making of a corporation. Nowadays likewise the business intelligence, business analytics, and data mining, Big Data has altered the business process from reporting and decision support to projection and further decision making [1][2].

One major reason why the latest technologies are being created to process Big Data is the incapability of typical database software tools to handle the massively growing data. Interesting to note that, until 2003, human-created only 5 exabytes (1018 bytes) of data, but this chunk of information is created in two days now. The Data used that time was extended to 2.72 zettabytes (1021 bytes) by 2012 and was predicted to double every two years which indeed, reached approximately 8 zettabytes of data by 2015 [4]. In the past, human genome decryption process used to take about 10 years, but at present within a week. Today's multimedia data have a major influence on the cornerstone of internet traffic and was expected to increase 70% by 2013. A few years ago, only Google had got more than one million servers around the worlds. There had been approximately 6 billion mobile subscriptions in the world and every day 10 billion text messages were transferred [3]. "**Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.**" [5] The introduction of some enormous systems such as - the mainframe computer, the PC, client-server computing, the Internet, cloud computing, mobile computing and social networking etc. have evolved the pace of technology. The emergence of Big Data as one of the most recent stages of the evolution aggregates technology in three trends - computing, data, and convergence. Here, computing storm comes as the output from the exponential growth of processing power mobile computing, social network, and cloud computing, the data storm from the availability of data with high volume, velocity and variety and lastly, the convergence storm comes as a resultant factor from the accessibility of open-source technology and commodity hardware. Datasets like Big Data represent a modern strategy of creating actionable business insights which can help organizations to sense and respond in a rapidly altering environment. [6] It is projected that by the year 2020, around 50 billion digital devices will be connected to networks and the internet [3].

Characteristics

Though there are a lot more discussions on the size of Big Data, there are other important characteristics of it to be taken care of - data variety and data velocity. These three components are represented as 3 Vs of Big Data. [7][9]

1. Volume: It represents the size of data which is at present bigger than terabytes, petabytes or even big. The majestic scope, as well as exponential growth of data, outrun traditional stock and analysis techniques. [8]

2. Variety: It is the type and nature of the data which assists people who want to analyze it for effective usage. The data gathered in an enterprise has become more intricate due to the fulmination of numerous sensors, smart devices, and social networking, and because of its inclusion of structured, semi-structured and unstructured data. Structured data depicts data which is aggregated into a relational scheme such as - rows and columns of a typical database. Semi-structured data is inherently self-describing and consists tags or other markers for strengthening hierarchies of records and fields within the data such as - weblogs, social media news feeds etc. Un-structured data consists of that kind of data which require extra effort to analyze or query such as - images, audio and video files etc. [1] [10] [7]
3. Velocity: It typically considers the speed at which the data can be arrived, stored and retrieved. Velocity also includes the study on various information streams and the rise in sensor network deployment having constant flows of which has made it complicated for traditional systems to handle. [1]

Applications of Big-data in current world

Due to the concurrent amelioration of big data, its significance has been so massive in recent years that big companies like - Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell etc. have been investing billion dollars on software firms in order to specialize data management and analytics [23]. To handle large data-sets areas like Internet search, fintech, urban informatics, and business informatics and difficulties encountered by scientists, business executives, practitioners of medicine, advertisers and governments, the practice of Big Data analytics has become ubiquitous [24]. Some of its recent applications are noted below:

1. **Real time applications**: At present, there are many domains for example – Transportation (selecting route and estimate time for destination, using sensor and other communication routes), Stock Market (predicting share prices, Automated trading shares), Clinical Care (clinical research) etc. where real-time big data analytics applications are needed. [25] [10]
2. **Data Visualization:** To help the business experts, big data analytics work as a better solution. Data are now handled efficiently in a comprehensive manner by the data scientists to ensure business stability and easiness of data complexion and so, have been adopted by the giant companies like – Google, EBay, Facebook and Twitter etc. [10].
3. **Finance sector**: Apart from stock market, there are other areas that have implemented Big Data analytics. For example - Quantitative Investing is a process data scientist incorporates computing technologies to predict security prices. Here, ideas are taken from newswires, Facebook, Twitter, magazines, earning reports, weather bulletins etc. [26].
4. **Sentiment analysis**: It is the procedure to detect and categorize data computationally, generally represented as text, in order to know the intention of writing the content. At present, it is the most used application of big data which has been implemented by many giant companies like IBM and SAP. It also uses techniques to know customer behavior and thus, increase customer experience [26].
5. **Food industry**: The applications of Big Data on the food industry is quite impactful to track the quality of products, recommending food items to customers or to ameliorate marketing strategies. IBM is currently providing techniques to the Cheesecake Factory to decompose structured data for example – find out location of restaurants, and unstructured data such as adding flavors to taste better and to communicate critical supply chain data faster in order to improve customer satisfaction. There are also technologies such as the Food Genius [28] which implements big data to speculate specific recommendations to foodies. Besides, the giant companies such as – the Starbucks, Dominos and Subway take utilize Big Data analytics in order to track individual customer preferences as well as offer customers with customized recommendations to improve more satisfaction. [27]

Besides, there are tremendous usage of Big Data in various fields among of which important fields' name are categorized below:

- Telecom Industry
- Fraud detection
- Defense sector
- Natural disasters
- Medical science
- Cloud computing
- Banking and Securities
- Communications, media and entertainment
- Education sector
- Manufacturing and natural resources
- Retail and whole sale trade
- Transportation and so on. [29] [10] [25]

Now this is the list of the current industries apart from the mentioned companies above in discussion which are dependent on Big Data for their business:

- Amazon – the online retail giant: has massive database, use algorithms to improve engagements.
- American express – use Big Data to predict consumer behavior and loyalty.
- BDO - Online Banking Service – to identify risks and fraud during audits
- Capital One – to ensure succession on customer offerings.
- General Electric – use data sensors to ameliorate working procedures and reliability.
- Miniclip – to improve user experience
- Netflix – big streaming service, has extreme data. [30]

Virtualization

By definition, virtualization describes the task of abstracting of computer resources as a form of a virtual version of the platforms like - virtual computer hardware platforms, storage devices, and computer network resources etc. Though at present it has become one of the most popular concepts in ICT, the concept is not brand new. [13] It is related to the technologies that give a logical view instead of providing a physical view of computing resources which assist to trick operating systems to work like a group of servers as a single chunk of computing resources. Besides, the system allows to run multiple operating systems concurrently and simultaneously on a single machine. [11] [14] [15]

It can also be termed as an IT asset optimization because it is being successfully used as Infrastructure as a Service (IAAS) for data centers, SMB, and other larger organizations through offering a solution for resource management. By using virtualization technique, it has been possible to achieve massive system utilization, abating cost as well as the efficiency of management. [12] [16]

Here, some terminologies related to virtualization technologies can help to deepen the knowledge of this concept:

Host Machine: It is a physical machine which operates the virtualization software. There are the physical resources in host machines for example - memory, hard disk space, and CPU, and also other resources like - network access which are utilized by virtual machines.

Virtual Machine: Virtualized representation of a physical machine, it is operated and maintained by the virtualization software. Every virtual machine works as like as an individual, physical, non-virtualized system at is specially implemented as a single file or a small collection of files in a single folder on the host system.

Virtualization Software: Allows a user to operate virtual machines on a host machine. [12]

Characteristics

Basically, virtualization provides an abstraction to the underlying hardware and software platforms by concealing the intricacy and heterogeneity of them and creates a single virtual machine which operates a single virtual data source. By using many sophisticated methodologies as well as techniques such as hardware and software partition or aggregation, partial or complete machine simulation, emulation and time sharing, virtualization either agglomerates or divides the computing resources of a server-based environment for providing various operating environments. As the evolution and amelioration of virtualization is being continued, a big number of various fields are adopting it. [17]

Virtualization has three main characteristics in terms of cloud computing:

1. Partitioning: In virtualization, the available resources are partitioned in a way that many software applications, as well as operating systems, can be supported in a single physical system.

2. Isolation: The host physical system and other virtualized machines isolate every corresponding virtual machine so that in if any crash occurs, it cannot affect the other virtual machines. Besides, one virtual container cannot share the data of the other.

3. Encapsulation: An encapsulated virtual machine might act as a complete entity in an application. So, each application can be protected by encapsulation as it does not interrogate with other application. [18]

Why do we need Virtualization?

There are many immense benefits of using virtualization for which a company has to consider it as a better assistor. Some important reasons are:

- Reduces investment and maintenance expense
- Optimize resources
- Diminishes power consumptions and consolidate the Data centers
- Saves the workload of an administrator
- Simplifies the management
- Maximizes uptime
- Allows easy installation process of software
- Improves CPU utilization
- Allows virtual servers to run on almost any server
- Ameliorates productivity, agility and efficacy of Big Data analytics
- Builds a true software-oriented data center. [31] [12]

Recent giant industries that uses virtualization:

- VMware
- Microsoft
- Citrix
- Red Hat
- Oracle
- Amazon
- Google
- Parallels / Odin
- Huawei

- VERDE VDI and so on. [33]

Relation between Big Data and Virtualization:
Efficient management of massively distributed Data and data integrated applications require to solve the challenges of Big Data. Virtualization has made it possible to manage the data by providing an extra phase of effectiveness to establish big data platforms a reality. It's not the only way to optimize the process of Big Data management, but by using virtualization the organization of data becomes more generic. Almost all of the traditional storage systems cannot meet the requirements for dealing with Big Data, whereas virtualization allows more flexibility to use and control the Big Data from any location. [19] [21]
Big Data and virtualization have different applications and they are interconnected in some aspects. Important sides are discussed below:

Big Data and Server virtualization: A server, physical server, of course, is divided into multiple virtual servers in a server virtualization. Virtualization provides a way in which all the hardware and computer resources such as - the random-access memory (RAM), CPU, hard drive, and network controller can be virtualized into a series of virtual machines so that every component runs individual applications and operating system. It is an authenticated software technology that gives access to run multiple operating systems as well as applications on the same server concurrently at the similar time. To provide effectivity in the use of physical resources, hypervisor or Virtual Machine Monitor (VMM) is used. Server virtualization helps to manage Big Data analysis in many ways for example it -

- Can improve agility, suppleness, and scalability of IT. Deployment of workloads occur faster, output and accessibility ameliorate and main operations occur automatically.
- Saves significant costs.
- helps to confirm that any platform can scale as needed to handle the large volumes and varied types of data including Big Data analysis.
- provides the base which enables many of the cloud services used as data sources in a big data analysis.
- increases the efficiency of the cloud that makes many complex systems easier to optimize. [22] [19]

Big Data & Storage virtualization – Storage virtualization becomes the most important factor for Big Data when it is needed to store data in multiple places or to access data from multiple places. Storage virtualization aggregates all data from many physical storages to one storage so that it looks like a single storage device. The process constitutes of abstraction and hiding the internal functionalities of a storage device from the host application, host servers or a general network for enhancing the process and network-independent management of storage. By doing this it has been greatly possible to reduce the expenditure of storage and thus become easier to manage and share all the data. [21] [19]

Big data processor and memory virtualization – Processor virtualization is used for processor performance maximization and the memory virtualization unlinks memory from the servers. There are many repeated queries of big data sets in big data analysis most pattern of which are hard to compute and require more processing power to meet an optimized solution. These analytics always search for efficient trends and patterns that are still to be understood. Besides, lots of memory (RAM) are needed to help CPU to generate all the data as fast as possible, as taking too much time can be the worst scenario for big data. To optimize and maximize performance of processor, processor virtualization is useful. [21] [19] [20]

Big data and network virtualization - Without a good network, it would be an arduous task to maintain all the storage and data from one location. If anyone can create multiple virtual networks and all networks can be utilized efficiently, it is better not to trust on physical network and manage the traffic of the physical network. Instead, a virtual network can reduce unaccepted interruption and improve the capability to manage the large distributed data required for big data analysis. [21] [19] [20]

Big data & Cloud Computing – A major capability for which the virtualization has brought enormous effect on Big Data is Cloud Computing. Every day new data is being created and lots of storage are required to handle these data. As the cloud computing provides storage expansion through a virtual machine and Big data is concerned with storage capacity, so cloud computing is very effective for Big Data analysis and management for providing huge computing power and storage resources. It has five essential characteristics: on-demand capabilities, broad network access, resource pooling, rapid elasticity and measured service. Firstly, via on-demand capabilities, a consumer can unilaterally provision computing capabilities for example - server time and network storage. Next, with broad network access, useful capabilities are available over the network which is accessed through mechanisms promoting usage by heterogeneous thin or thick client platforms. Then, a multi-tenant model with various physical and virtual resources is used to pool the provider's computing resources to serve multiple consumers via resource pooling.
With rapid elasticity, capabilities can be elastically provisioned and released to scale and version rapidly growing Big Data even automatically in some cases. Lastly, cloud systems can automatically manage the usage of available resources and to do this, they leverage a metering capability as an abstraction at some level that might be suitable to the type of service for example – storage, processing, and bandwidth etc. [21] [19] [20]

Conclusion
Some brief explanation and important characteristics of Big data and virtualization along with their relationship have been studied from various aspects, and it can be recapitulated that the relationship between them is complementary. Big data and virtualization form an integrated model in the world of distributed network technology. The evolution of big data, progression

and the requirements stimulate those service providers who work for virtualization optimization because the relationship between them is based on the service, the storage, and processing as a common factor. Simply, Big data represents the product or a service and the virtualization represent the container with massive storage capacity. Virtualization provides an environment of flexible distributed resources that uses high techniques in the processing and management of Big Data and yet abates the cost. To use an advanced-analytic resource of elastic and fluid topology, to efficiently originate data in any source, format, and schema, to have a latency-agile resource that persists, aggregates and processes data, an organization must combine the use of virtualization for effective management of Big Data.

REFERENCES

[1]. Dr.M. Moorthy, R. Baby and S. Senthamaraiselvi "An Analysis for Big Data and its Technologies", Dec 2014 | Vol 4, Issue 12,412-41.

[2]. Gang-Hoon Kim, Silv ana Trimi, and Ji-Hyong Chung, white paper March, 2014," Big-Data Applications in the Government Sector".

[3]. Seref SAGIROGLU and Duygu SINANC, "Big Data: A Review", 978-1-4673-6404-1/13/2013 IEEE.

[4]. Intel IT Center, "Planning Guide: Getting Started with Hadoop", Steps IT Managers Can Take to Move Forward with Big Data Analytics,

June 2012

http://www.intel.com/content/dam/www/public/us/en/documents/guides/
getting-started-with-hadoop-planning-guide.pdf

[5]. Apache Hive. Available at http://hive.apache.org

[6]. Joseph O. Chan, "An Architecture for Big Data Analytics", 2013 Volume 13 Issue 2.

[7]. Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review". martinhilbert.net. Retrieved 7 October 2015.

[8] S. Singh and N. Singh, "Big Data Analytics", 2012 International Conference on Communication, Information & Computing Technology Mumbai India, IEEE, October 2011

[9]. A, Katal, Wazid M, and Goudar R.H. "Big data: Issues, challenges, tools and Good practices.". Noida: 2013, pp. 404 – 409, 8-10 Aug. 2013.

[10]. Samiddha Mukherjee, Ravi Shaw. "Big Data – Concepts, Applications, Challenges and Future Scope", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2016.

[11]. Richard Scroggins. "Virtualization Technology Literature Review", Global Journal of Computer Science and Technology Interdisciplinary, Volume 13 Issue 1 Version 1.0 Year 2013.

[12]. Rabi Prasad Padhy, Manas Ranjan Patra, Suresh Chandra Satapathy. "VIRTUALIZATION TECHNIQUES & TECHNOLOGIES: STATE-OF-THE-ART", Journal of Global Research in Computer Science, Volume 2, No. 12, December 2011.

[13] Ivan Pogarcic, David Krnjak and Davor Ozanic, "Business Benefits from the Virtualization

of an ICT Infrastructure", 1 August 2012.

[14] Nabeel Zanoon, Abdullah Al-Haj and Sufian M Khwaldeh, "Cloud Computing and Big Data is there a Relation between the Two: A Study", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017).

[15] Susanta Nanda Tzi-cker Chiueh, "A Survey on Virtualization Technologies".

[16] Heradon Douglas, Christian Gehrmann, "Secure Virtualization and Multicore Platforms", State-of-the-Art report, SICS Technical Report, Swedish Institute of Compute Sceince (December 2009).

[17] Davide Adami, Stefano Giordano, "Multidomain layer 1 Infrastructure Virtualization as a Feature Internet Servicesenabling Paradigm", Journal of Internet engineering 4(1), (December 2010).

[18] Judith Hurwitz, Robin Bloor, Marcia Kaufman and Fern Halper, "Cloud Computing For Dummies**",** November 2009.

[19] Judith Hurwitz, Robin Bloor, Marcia Kaufman and Fern Halper, "THE IMPORTANCE OF VIRTUALIZATION TO BIG DATA**",** November 2009.

[20] James Kobielus, "Big data needs data virtualization", JUN 20, 2013

[21] Bala M. Balachandran and Shivika Prasad, "Challenges and Benefits of Deploying Big Data Analytics in the Cloud for Business Intelligence", Procedia Computer Science, 112 (2017) 1112–1122.

[22] vSphere and vSphere with Operations Management, article - Virtualized Big Data, source - **https://www.vmware.com/products/vsphere/virtualize-big-data.html**

[23] **"Data, data everywhere"**. The Economist. 25 February 2010. Retrieved 20th March 2018.

[24] "Community cleverness required". Nature. 455 (7209): 1. 4 September 2008. doi:10.1038/455001a. PMID 18769385.

[25] Akinul Islam Jony, "Applications of Real-Time Big Data Analytics", International Journal of Computer Applications (0975 - 8887), Volume 144 - No.5, June 2016.

[26] Accessed from - **http://www.wsj.com/articles/how-computers-trawl-a-sea-of-datafor-stock-picks**

[27] Accessed from - **https://datafloq.com/read/big-datas-impact-food-industry/96**

[28] Accessed from - **http://www.forbes.com/sites/daniellegould/2012/09/24/foodindustry-understand-trends-big-data-tools/**

[29] Accessed from - **https://www.simplilearn.com/big-data-applications-in-industries-article**.

[30] Accessed from - **https://www.icas.com/ca-today-news/10-companies-using-big-data**.

[31] Zartasha Gul, Irfan Ahmed, Yasir Hafeez and Saiqa Bibi. "Virtualization benefits in High Performance Computing Applications", Journal of Computer Science and Information Technology, June 2014, Vol. 2, No. 2, pp. 101-109.

[33] Accessed from - **https://www.serverwatch.com/server-trends/slideshows/top-10-virtualization-technology-companies-for-2016.html**