IJCTA, 9(26) 2016, pp. 145-151

A Review Paper on Big Data & Cloud Computing Security Issues

V. Balaji[1] and P. Swarnalatha[2]

ABSTRACT

Big Data and cloud computing are two most essential topics in the current year. This permits computing resources to be provided as Information Technology services with high efficiency and usefulness. Now a day's big data is one of the major complications that researchers try to solve it and allows concentrating researches over it to get ride of the problem of how big data could be handling in the recent systems and managed with the cloud of computing. Also the most important issue is how to gain a perfect security for big data in cloud computing. This paper deals with a survey of big data with cloud computing security and the mechanisms that are used to protect and secure big data within the limitations.

Keywords: Big data, Cloud Computing, Cloud Security, Cloud Provider, Big Data Privacy.

1. INTRODUCTION

Big data is known as a datasets with size beyond the capability of the software tools that used today to accomplish and process the data within a dedicated time. With Variety, Volume, Velocity Big Data such military data or other unauthorized data need to be protected in a scalable and efficient way. Data confidentiality and security is one of most afraid problems for Cloud Computing due to its open environment with very limited user side control. It is also an essential task for Big Data. After few years later additional data globally would be affected with Cloud Computing which provides strong storage, computation and distributed ability in support of Big Data processing. Other thoughts are that information confidentiality and security challenges in both Cloud Computing and Big Data must be investigated. the confidentiality and security providing such forum for researchers, and developers to exchange the latest experience, research thoughts and development on fundamental topics and applications about security and confidentiality issues in cloud and big data environments . The cloud helps organizations and enables rapid on demand provisioning of server resources such as CPUs, manage, storage, bandwidth, and share and analyze their Big Data in a reasonable and simple to use [1]. The cloud infrastructure as a service platform which is supported by on demand analytics solution seller makes the large size of data analytics very affordable. As location independent cloud computing involves shared services providing resources, software and data to systems and the hardware on demand [1].

2. PRESENT OF BIG DATA AND FUTURE

The Big Data concept represents an in essence an "ocean of data", lots of information and the means to analyses them in the present days most of the humans can access more information than most of our descendants in a lifetime. Nowadays we double, in every year, the amount of data that we create

Forbes defines big data like thisBig data is a collection of data from traditional and digital sources inside and outside your company that represents a source of ongoing discovery and analysis" Big data is a

[1] Research Scholar, SCOPE, VIT University

[2] Associate Professor, SCOPE, VIT University

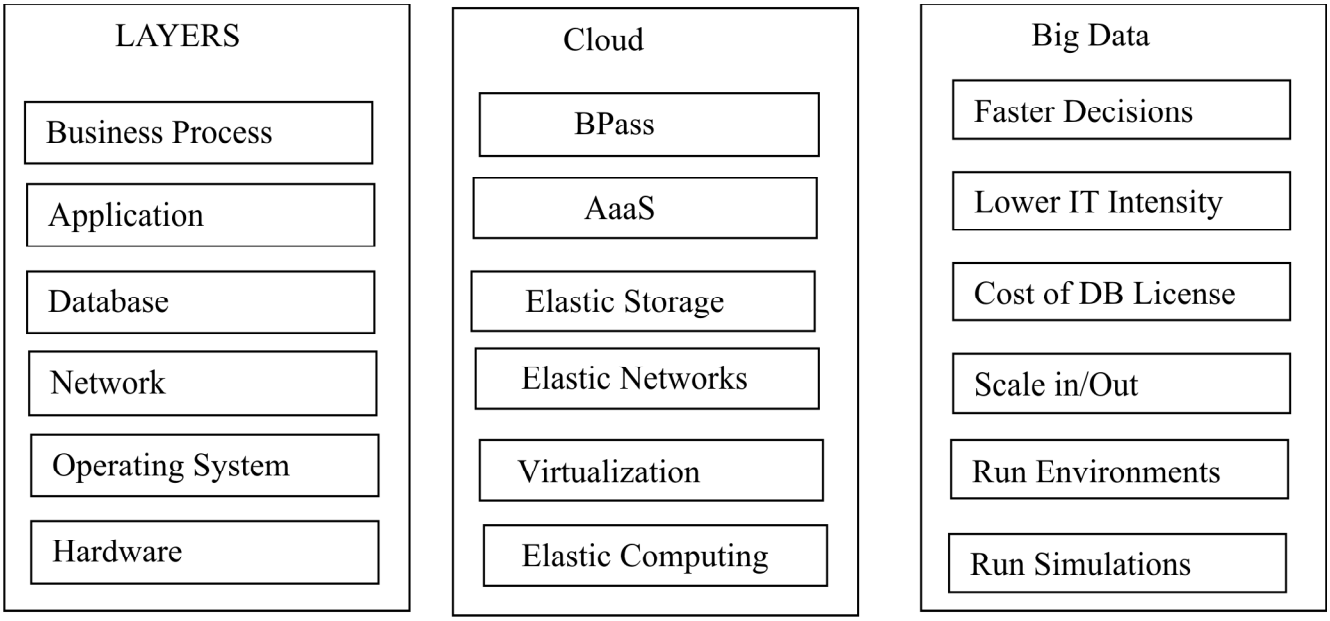

Figure 1: Big Data and Clouds

mix of unstructured and multi structured data, those types of data are analyzed together to get more knowledge and information to company than could be get using the usual methods and infrastructure.

Unstructured data is information that is not organized or easily interpreted by traditional data models or databases, and usually is text-heavy.

Good examples are posts from twitter, LinkedIn and other social media services.

Multi- structured data is represented by a variety of data formats that came from interaction between peoples and machines, such as web applications and social services. Those include text and multimedia formats, like photos and videos, with structured data.

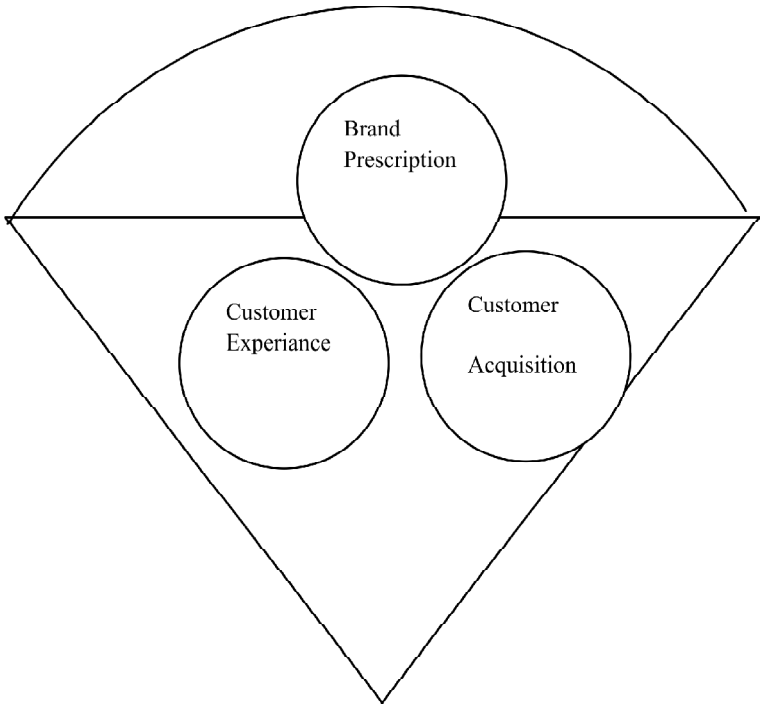

Figure : Social Media Business intelligence

This days the data produce from many sources such as social networks, website and sensor network. Also the total of data volume is expanding Continuity .however; big data refers essentially to the following data types such as traditional enterprise data such as Customer ýnformation in Data Base, the transactions websites companies. Machine generated and Sensors data such as smart meter, manufacturing sensors etc. and Social data such as social network and application platforms like Facebook, LinkedIn, what's app, Twitter and YouTube. According to a recent report the most of data unstructured or semi structured and the size of data exists now is doubling in every two years. So between 2013 and 2020 it will go to 44 trillion GB from 4.4 trillion GB. Moreover the huge amount of data recorded mostly in nonstandard forms which cannot be analyzed using traditional data models and methods. Big Data today have a wide range of challenges but the opportunities are also exists the right decision making, marketing strategies and improved customer relations, better public services and so on.

According to a Gartner report 4.4 million new big data associated jobs will be created globally and only one third of these will be occupied. So, employment chances are huge in the big data job markets but there are very few training and education offerings focusing on this market. Big data has opened the increasing interest to new tools production Commencement with the introduction of Apache Hadoop and Map Reduce and also many open source have been implemented and developed by companies IBM, Oracle, Cloudera, SAP, Teradata, SAS Amazon and many others. Most big data products are generally based on open-source technologies [7]. Therefore, standards are especially important and needed for interoperability of the hardware and software components of commercial solutions. Lack of official standards also aggravates confidentiality and security problems

3. CLOUD COMPUTING IN BIG DATA

Cloud computing could be a powerful technology to accomplish massive-scale and complicated computing. It removes the requirement to keep up valuable computing hardware, ardent house, and software system. Huge growth within the scale information of knowledge of information} or huge data generated through cloud computing has been experimental. Addressing huge knowledge could be a difficult and time-demanding task that needs an oversized process infrastructure to make sure booming processing and analysis. The increase of huge knowledge in cloud computing is reviewed during this study. The definition, features, and classification of huge knowledge together with some concerns on cloud computing are introduced [6]. The connection between huge knowledge and cloud computing, huge knowledge storage systems, and Hadoop technology also are mentioned. Moreover, analysis challenges are investigated, with concentrate on quantifiability, accessibility, knowledge integrity, knowledge transformation, knowledge quality, knowledge no uniformity, privacy, legal and restrictive problems, and governance. Lastly, open analysis issues that need vital analysis efforts are summarized

The rise of cloud computing and cloud data stores has been a predecessor and implementer to the emergence of big data. Cloud computing is the commodification of computing time and information storage by means of standardized technologies. It has significant advantages over traditional physical deployments. However, cloud platforms come in several forms and sometimes have to be integrated with traditional architectures. This leads to confuse for decision makers in charge of big data projects, leads to a question of how and which cloud computing is the optimal choice for their computing needs, especially if it is a big information project? These projects frequently exhibit unpredictable, bursting, or immense computing power and storage requirements.

4. BIG DATA CLOUD PROVIDERS

Cloud providers come in all shapes and scopes and offer many different products for big data. Some are household names while others are recently emerging. Some of the cloud providers that offer IaaS services

that can be used for big data include Amazon.com, AT&T, GoGrid, Joyent, Rackspace, IBM, and Verizon/ Terremark.

Amazon's Public Elastic Compute Cloud for big data: Presently, one amongst the best high-profile IaaS service suppliers is Amazon net Services with its Elastic reason Cloud (Amazon EC2). Amazon didn't begin out with a plan to make an enormous infrastructure services business.

Instead, the corporate designed a vast structure to support its own retail business and discovered that its resources were underused. Rather than permitting this quality to sit down idle, it set to leverage this resource whereas adding to all-time low line. Amazon's EC2 service was thrown in 2006 and continues to evolve.

Amazon EC2 offers measurability underneath the user's management, with the user paying for resources by the hour. The utilization of the term elastic within the naming of Amazon's EC2 is critical. Here, physical property refers to the aptitude that the EC2 users ought to increase or decrease the infrastructure resources allotted to fulfill their needs [3] Amazon additionally offers different huge knowledge services to customers of its Amazon net Services portfolio. These embody the following

- **Amazon Elastic MapReduce:** Targeted for processing huge volumes of data. Elastic MapReduce utilizes a hosted Hadoop framework running on EC2 and Amazon Simple Storage Service (Amazon S3). Users can now run HBase.
- **Amazon DynamoDB:** A fully managed not only SQL (NoSQL) database service. DynamoDB is a fault tolerant, highly available data storage service offering self-provisioning, transparent scalability, and simple administration. It is implemented on SSDs (solid state disks) for greater reliability and high performance [3].
- **Amazon Simple Storage Service (S3):** A web-scale service designed to store any amount of data. The asset of its design center is presentation and scalability, so it is not as feature laden as other data stores. Data is stored in "buckets" and you can select one or more global regions for physical storage to address latency or regulatory needs [3].
- **Amazon High Concert Computing:** Tuned for specialized tasks, this service provides low-latency tuned high enactment computing clusters. Most often used by scientists and academics, HPC is entering the mainstream because of the offering of Amazon and other HPC benefactors. Amazon HPC clusters are purpose built for specific workloads and can be reconfigured easily for new tasks [2].
- **Amazon RedShift:** Available in limited preview, RedShift is a petabyte-scale data warehousing service built on a scalable MPP architecture. Managed by Amazon, it offers a secure, reliable alternative to in-house data warehouses and is compatible with several popular business intelligence tools [3].

5. BIG DATA CLOUD STORAGE

The cloud storage tasks in massive information analytics represent 2 categories: capability and performance. Scaling capability, from a platform perspective, are some things all cloud suppliers have to be compelled to watch closely. Information retention continues to double and triple year-over-year as a result of customers are keeping additional of it. Certainly, that impacts North American nation as a result of we want to supply capability. In skilled cloud storage has to be extremely obtainable, extremely sturdy, and should scale from many bytes to petabytes. Amazon's S3 cloud storage is that the most outstanding answer within the house. S3 guarantees a ninety nine. 9% monthly accessibility and ninety nine.999999999% sturdiness p.a. This can be under associate degree hour outage per month. The sturdiness are often illustrated with associate degree example. If a client stores ten 1 000 objects he will expect to lose one object each ten 1, 000,000

years on the average. S3 achieves this by storing information in multiple facilities with error checking and self-healing processes to observe and repair errors and device failures. This can be fully clear to the user and needs no actions or data.

A company may build and come through a likewise trustworthy storage answer however it'd need tremendous capital expenditures and operational challenges. International information focused firms like Google or Facebook have the experience and scale to try and do this economically [7]. Massive information comes and start-ups, however, enjoy employing a cloud storage service. They will trade cost for associate degree operational one that is great since it needs no capital outlay or risk. It provides from the primary computer memory unit reliable and ascendible storage solutions of a high quality otherwise impossible. This permits new product and comes with a viable choice to begin on a little scale with low prices. Once a product proves successful these storage solutions scale nearly indefinitely. Cloud storage is effectively a infinite information sink. significantly for computing performances is that several solutions additionally scale horizontally, i.e. once information is traced in parallel by cluster or parallel computing processes the output scales linear with the amount of nodes reading or writing.

6. BIG DATA PRIVACY AND SECURITY PROBLEMS

Big data analytics are being used more widely every day for an even wider number of reasons. These new methods of applying analytics certainly can bring innovative improvements for business. For example, retail businesses are successfully using big data analytics to predict the hot items each season, and to predict geographic areas where demand will be greatest, just to name a couple of uses [2].

The power of big data analytics is so great that in addition to all the positive business possibilities, there are just as many new privacy concerns being created. Here are ten of the most significant privacy risks [2].

1. **Confidentiality breaches and embarrassments:** The actions taken by businesses and other organizations as a result of big data analytics may breach the privacy of those involved, and lead to discomfiture and even lost jobs. Consider that some retailers have used big data analysis to predict such intimate private details such as the due dates of pregnant shoppers [5]. In such cases subsequent marketing activities resulted in having members of the household discover a family member was pregnant before she had told anyone, resulting in an uncomfortable and damaging family situation. Retailers, and other types of businesses, should not take actions that result in such situations.
2. **Anonymization could become difficult:** With so much data, and with powerful analytics, it could become impossible to completely remove the ability to identify an individual if there are no rules established for the use of anonymized data files. For example, if one anonymized data set was collective with another completely separate data base, without first determining if any other data items should be removed prior to combining to protect anonymity, it is possible individuals could be re-identified. The important and necessary key that is usually missing is establishing the rules and policies for how anonymized data files can be combined and used together [3].
3. **Data masking could be conquered to reveal personal information:** If data masking is not used appropriately, big data analysis could easily reveal the actual individuals who data has been masked. Organizations must establish effective policies, procedures and processes for using data masking to ensure privacy is preserved. Since big data analytics is so new, most organizations don't realize there are risks, so they use data masking in ways that could breach privacy. Many resources are available, such as those from IBM, to provide guidance in data masking for big data analytics.
4. **Unethical actions based on interpretations:** Big data analytics can be used to try and influence behaviors. There are my ethical issues with driving behavior. Just because you CAN do something doesn't mean you should. For example, in the movie The Fight Club, Ed Norton's character's job

was to determine if an automobile manufacturer should do a recall based strictly on financial consideration, [3] without taking into account the associated health risks. Otherwise, in other words, if it is cheaper for people to be killed or injured instead of fixing the faulty equipment in the vehicles. Big data analytics can be used by organizations to make a much wider variety of business decisions that do not take into account the human lives that are involved. The potential to reveal personal information because it is not illegal, but can damage the lives of individuals, must be considered [2].

5. **Big data analytics are not 100% accurate:** While big data analytics are powerful, the predictions and conclusions that result are not always accurate. The data files used for big data analysis can often contain inaccurate data about individuals, use data models that are incorrect as they relate to particular individuals, or simply be flawed algorithms (the results of big data analytics are only as good, or bad, as the computations used to get those results)[9]. These risks increase as more data is added to data sets, and as more complex data analysis models are used without including rigorous validation within the analysis process. As a result, organizations could make bad decisions and take inappropriate and damaging actions. When decisions involving individuals are made based upon inaccurate data or flawed models, as a result individuals can suffer harm by being denied services, being falsely accused or misdiagnosed, or otherwise be treated inappropriately.

10 major security and confidentiality challenges facing infrastructure suppliers and shoppers. By demarcation the problems concerned, in conjunction with investigation of internal and external threats and summaries of current ways to mitigating those risks, the alliance's members hope to prod technology vendors, educational students and practitioners to collaborate on computing techniques and business observes that decrease the risks related to analyzing large information sets mistreatment innovative data analytics. Existing cryptography technologies that don't scale well to massive datasets. Time period system observance ways that works well on smaller volumes of knowledge however not terribly massive datasets.

The growing range of devices, from smartphones to sensors, manufacturing information for analysis. General confusion "surrounding the varied legal and procedure restrictions that cause accidental approaches for making certain security and confidentiality [1].

Given the terribly massive information sets that contribute to an enormous information implementations, there's a virtual certainty that either protected information or important belongings (IP) are gift. This data is distributed throughout the massive information implementation as needed with the result that the complete information storage layer wants security protection [1].

7. CONCLUSION

Recently, researchers focusing their hard work in how to achieve, handle and also processing the enormous amount of information(Big Data) which deals with three concepts volume , Variety and velocity. This requires new mechanisms to accomplish, processing, storing, analysing and securing the big data which will be discussed in further papers. Also the managing and processing of big data have many problems wherein the required efforts have been made to handle these requirements. This finally deals with big data, security which has been mentioned as one of the challenges that stand up when systems try to handle the idea of big data. The effective research techniques required to overcome the security of big data which have been carried out by the researcher as a future work.

REFERENCES

[1] Elmustafa Sayed Ali Ahmed1 and Rashid A.Saeed "A Survey of Big Data Cloud Computing Security ",*International Journal of Computer Science and Software Engineering (IJCSSE)*, **3, 1**,138-145, 2014.

[2] Vahid Ashktorab1, Seyed Reza Taghizadeh2 and Dr.Kamran Zamanifar3, "A Survey on Cloud Computing and Current Solution Providers", *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, **3 2**, 123-131, 2012.

[3] Venkata Narasimha Inukollu, Sailaja Arsi ,and SrinivasaRao Ravuri, Security Issues Associated With Big Data In Cloud Computing, *InternationalJournal of Network Security & Its Applications (IJNSA)*,**6,3**,336-345 2014.

[4] Addison Snell, Solving Big Data Problems with Private Cloud Storage, Intersect360 Research, Cloud Computing in HPC: Usage and Types, **6, 1**, 235-244 2011.

[5] D. Borthakur, "The hadoop distributed file system: Architectureand design," *Hadoop Project Website*, **11**,57-65, 2007.

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

[7] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." *Open Cirrus Summit (OCS), 2012 Seventh*, Beijing, Jun 19-20, 2012, pp. **48 – 52**.

[8] Venkata Narasimha Inukollu , Sailaja Arsi and Srinivasa Rao Ravuri "Security issues associated with big data in cloud computing "*International Journal of Network Security & Its Applications(IJNSA*), **6, 3**, 457-465, 2014.

[9] Addison Snell, Solving Big Data Problems with Private Cloud Storage, Intersect360 Research, Cloud Computing in *HPC: Usage and Types*, **6, 1**, 235-244 2011.