ISSN: 2321-7782 (Online)
Volume 3, Issue 6, June 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Study and Analysis of Big Data in Cloud Computing

V Akhila Reddy[1]
CSE Department
Jaya Prakash Narayan College of Engineering
Mahabub Nagar - India

G Rakesh Reddy[2]
CSE Department
Jaya Prakash Narayan College of Engineering
Mahabub Nagar - India

Abstract: ***With the rapid growth of emerging applications like social network analysis, semantic Web analysis and bioinformatics network analysis, a variety of data to be processed continues to witness a quick increase. Big Data and cloud computing are two important issues in the recent years, enables computing resources to be provided as Information Technology services with high efficiency and effectiveness. Cloud computing eliminates the need to maintain expensive computing hardware, dedicated space, and software. Big data is an emerging paradigm applied to datasets whose size or complexity is beyond the ability of commonly used computer software and hardware tools. From the view of cloud data management and big data processing mechanisms, we present the key issues of big data processing, including cloud computing platform, Cloud Computing key characteristics, Service models, deployment models, cloud architecture and the rise of big data in cloud computing and necessity of Security in Big Data is reviewed in this study. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology and Map Reduce are also discussed.***

Keywords: Big Data, Cloud Computing, Hadoop Technology, Map Reduce, Security.

I. INTRODUCTION

Big data is known as a datasets with size beyond the ability of the software tools that used today to manage 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. 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 .The cloud infrastructure as a service platform, supported by on demand analytics solution seller that makes the large size of data analytics very affordable. As location independent cloud computing Involving shared services providing resources , software and data to systems and The hardware on demand, actually the storage networking in cloud is a very strong because use driver for high performance.

Fig 1. Big Data and Cloud Computing

II. CLOUD COMPUTING

When we store our photos online instead of on our home computer, or use webmail or a social networking site, then we are using a "cloud computing" service. If we are an organization, and we want to use, for example, an online invoicing service instead of updating the in-house one we have been using for many years, that online invoicing service is a "cloud computing" service. Cloud computing refers to the delivery of computing resources over the Internet. Instead of keeping data on our own hard drive or updating applications for our needs, we use a service over the Internet, at another location, to store your information or use its applications.

Cloud computing is the delivery of computing services over the Internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Examples of cloud services include online file storage, social networking sites, webmail, and online business applications. The cloud computing model allows access to information and computer resources from anywhere that a network connection is available. Cloud computing provides a shared pool of resources, including data storage space, networks, computer processing power, and specialized corporate and user applications.

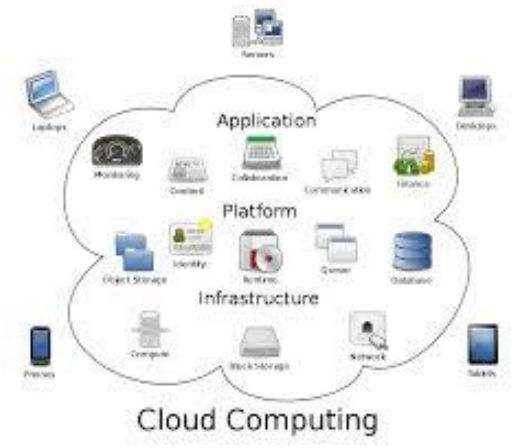

Fig 2. Cloud Computing

A. *Key Characteristics of Cloud Computing*

1) *Flexibility/Elasticity:* users can rapidly provision computing resources,as needed, without human interaction. Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out or up.

2) *Scalability of infrastructure:* new nodes can be added or dropped from the network as can physical servers, with limited modifications to infrastructure set up and software. Cloud architecture can scale horizontally or vertically, according to demand.

3) *Broad network access:* Capabilities are available over the networkand accessed through standard mechanisms that promote use by heterogeneous platforms (e.g., mobile phones, laptops, and PDAs).

4) *Location independence:* There is a sense of location independence, in that the customer generally has no control or knowledge over the exact location of the provided resources, but may be able to specify location at a higher level of abstraction (e.g., country, state, or data center).

5) *Reliability:* improves through the use of multiple redundant sites, which makes cloud computing suitable for business continuity and disaster recovery.

6) *Economies of scale and cost effectiveness:* Cloud implementations, regardless of the deployment model, tend to be as large as possible in order to take advantage of economies of scale. Large cloud deployments can often be located close to cheap power stations and in low-priced real estate, to lower costs.

7) *Sustainability:* comes through improved resource utilization, more efficient systems, and carbon neutrality.

B. *Cloud Computing Service models*

Cloud computing service models are classified as:

1) *Software as a Service (SaaS):* It is the delivery of application. In SaaS a complete application is provided to user which is running on cloud infrastructure. As software is hosted by provider, users do not need to buy, install or manage hardware for it. In SaaS instances of a software application are shared as a service. Examples of SaaS are Google Docs, Cloud Drive, and Salesforce.com CRM application.

2) *Platform as a Service (PaaS):* PaaS enables developers to deploy their application on the cloud. The consumer can control their application but do not have any control over underlying infrastructure. It provides user an integrated set of software through the internet. PaaS is a delivery of computing platform as a service. Examples of PaaS are Google App Engine, Amazon Web Services, and Microsoft Azure.

3) *Infrastructure as a Service (IaaS):* Using IaaS user get access to resources like storage, server, networks, data center space. It shares pool of computing resources. User can deploy and run both application and operating system on IaaS. It frees user from buying or managing underlying software and hardware. Example of IaaS is Amazon EC2.

C. *Deployment Models of cloud Computing*

Cloud services are typically made available via a private cloud, community cloud, public cloud or hybrid cloud.

1) *Public cloud:* Public clouds are offered over the Internet and are owned and operated by a cloud provider. Some examples include services aimed at the general public, such as online photo storage services, e-mail services, or social networking sites. However, services for enterprises can also be offered in a public cloud.

2) *Private cloud:* In private cloud, the cloud infrastructure is operated solely for a specific organization, and is managed by the organization or a third party.

3) *Community cloud:* In community cloud the service is shared by several organizations and made available only to those groups. The infrastructure may be owned and operated by the organizations or by a cloud service provider.

4) *Hybrid cloud:* It is a combination of different methods of resource pooling (for example, combining public and community clouds).

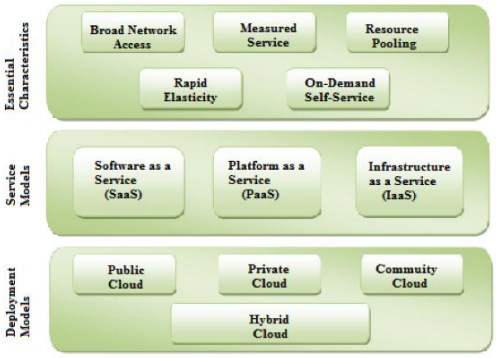

Fig 3. NIST Model of Cloud Computing

D. Architecture Of Cloud Storage

Cloud storage architectures are primarily about delivery of storage Data on demand in a highly scalable and multi-tenant way. Cloud storage architectures consist of a front end that exports an API to access the storage. In traditional storage systems, this API is the SCSI protocol; but in the cloud, these protocols are evolving. There, you can find Web service front ends, file-based front ends, and even more traditional front ends (such as Internet SCSI, or ISCSI). Behind the front end is a layer of middleware that may also called as storage logic. This layer implements a variety of features, such as replication and data reduction, over the traditional data-placement algorithms (with consideration for geographic placement). Finally, the back end implements the physical storage for data. This may be an internal protocol that implements specific features or a traditional back end to the physical disks.

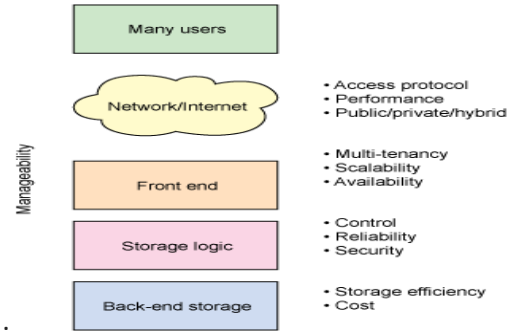

Fig 4. Generic cloud storage architecture

III. BIG DATA

Big data is defined as large amount of data which requires new technologies and architectures to make possible to extract value from it by capturing and analysis process. New sources of big data include location specific data arising from traffic management, and from the tracking of personal devices such as Smart phones. Big Data has emerged because we are living in a society which makes increasing use of data intensive technologies. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are need to be understood. Big Data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concepts & tools. The difficulties can be related to data capture, storage, search, sharing, analytics and visualization etc.

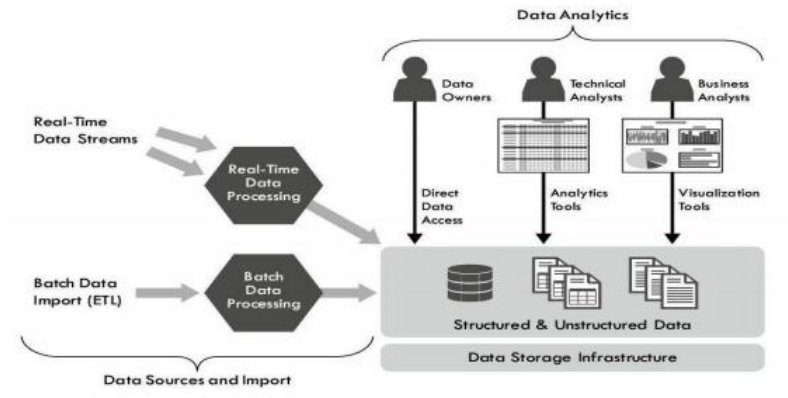

Fig 5. Example of Big Data Architecture (Aveksa Inc., 2013)

Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. The various challenges faced in large data management include – scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more. In addition to variations in the amount of data stored in different sectors, the

types of data generated and stored—i.e., encoded video, images, audio, or text/numeric information; also differ markedly from industry to industry.

A. BIG DATA CHARACTERISTICS

1) Data Volume: The Big word in Big data itself defines the volume. At present the data existing is in peta bytes and is supposed to increase to zetta bytes in nearby future. Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it.

2) Data Velocity: Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows and aggregated.

3) Data Variety: Data variety is a measure of the richness of the data representation – text, images video, audio, etc. Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web Pages, Web Log Files, social media sites, e-mail, documents.

B. Big Data Technologies

When making an attempt to understand the concept of Big Data, the words "Hadoop" and "Map Reduce" can not be avoided

1) *Hadoop:* Hadoop, which is a free, Java-based programming framework, supports the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, flexible and fault tolerant. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data. Hadoop has two main sub parts – Map Reduce and Hadoop Distributed File System (HDFS).

2) *Map Reduce: Hadoop* Map Reduce is a framework used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file-system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework.

3) *Hadoop Distributed File System (HDFS):* HDFS is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources to overcome node failures.

C. BIG DATA MANAGEMENT

The needs of the big data are not being satisfied by the current technologies and the speed of increasing storage capacity is much less compared to the data. Thus a revolution reconstruction of information framework is needed very much. For this we need to design a hierarchical architecture for storage. The heterogeneous data are not efficiently handled by the efficient Algorithms that exist now and thus we need to even design a very efficient algorithm for the effective handling of the heterogeneous data.

Necessity of security in big data:

The big data is used by many of the business but they may not have assets from perspective of the security. If any security threat occurs to big data, it may come out with even more serious issue. Nowadays, companies use this technology to store data of petabyte range regarding to the company, business and customers. This result in severe criticality for classification of information to secures the data we either need to encrypt, log or use honey pot techniques. The challenge of detecting threats and malicious intruders, must be solved using big data style analysis.

Analysis and computation of big data:

Speed is the main thing when we look up for querying in the big data. However the process may be time consuming only because of the reason that it cannot traverse all related data in the whole database in a short time. While the big data is getting complicated, the indices in the big data are aiming at the simple type of the data. The traditional serial algorithm is inefficient for this big data.

IV. BIG DATA CLOUD STORAGE

The cloud storage challenges in big data analytics fall into two categories: capacity and performance. Scaling capacity, from a platform perspective, is something all cloud providers need to watch closely. Data retention continues to double and triple year-over-year because customers are keeping more of it. Certainly, that impacts us because we need to provide capacity.

In Professional cloud storage needs to be highly available, highly durable, and has to scale from a few bytes to petabytes. The durability can be achieved this by storing data in multiple facilities with error checking and self-healing processes to detect and repair errors and device failures. This is completely transparent to the user and requires no actions or knowledge. A company could build and achieve a similarly reliable storage solution but it would require tremendous capital expenditures and operational challenges. Global data centered companies like Google or Face book have the expertise and scale to do this economically. Big data projects and start-ups, however, benefit from using a cloud storage service. They can trade capital expenditure for an operational one, which is excellent since it requires no capital outlay or risk. It provides from the first byte reliable and scalable storage solutions of a quality otherwise unachievable. This enables new products and projects with a viable option to start on a small scale with low costs. When a product proves successful these storage solutions scale virtually indefinitely. Cloud storage is effectively a boundless data sink. Importantly for computing performances is that many solutions also scale horizontally, i.e. when data is copied in parallel by cluster or parallel computing processes the throughput scales linear with the number of nodes reading or writing.

V. FUTURE SCOPE

Recently , researchers focusing their efforts in how to manage , handling and also processing the huge amount of data as known a Big data deals with three concepts volume , Variety and velocity which requires a new mechanisms to manage , processing , storing , analyzing and securing the big data . As managing and processing of big data have many problems and required more efforts to handle these requirements when deal with big data , security is one of the challenges that arise when systems try to handle the concept of big data. More researches required to increase computation speed and to overcome the security of big data instead of current security algorithms and methods.

VI. CONCLUSION

This paper gave a description of use of big data in the cloud computing. We discussed about the key issues of big data processing, including cloud computing platform, and the rise of big data in cloud computing and necessity of Security in Big Data and Big data Technologies. Big Data is not a new concept but very challenging. It is very much cooperate and make a successful and long term use of cloud computing and explore new ideas for the usage of the big data over cloud environment.

References

1. 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), Vol.6, No.3, May 2014
2. K. R. C. Wang, Q. Wang and W. Lou, "Ensuring data storage security in cloud computing," in Proc.17th International Workshop on Quality of Service (IWQoS '09), pp. 1–9, 2009.
3. L. M. Kaufman, "Data security in the world of cloud computing," Security & Privacy, IEEE, vol. 7, no. 4, pp. 61–64, 2009.
4. S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," Journal of Network and Computer Applications, vol. 34, no. 1, pp. 1–11, 2011.
5. 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.
6. F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Busines on big data applications." Big Data, 2013 IEEE International Conference, Silicon Valley, CA, Oct 6-9, 2013, pp.32 - 37.
7. "Big data: science in the petabyte era," Nature 455 (7209):1, 2008.
8. Douglas and Laney, "The importance of 'big data': A definition,"2008.
9. Hao, Chen, and Ying Qiao. "Research of Cloud Computing based on the Hadoop platform.".Chengdu, China: 2011, pp. 181 – 184, 21-23 Oct 2011.
10. 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.
11. Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications.". Hangzhou, China: 2012, pp. 59 – 63, 21-24 Oct. 2012.

AUTHOR(S) PROFILE

V Akhila Reddy, received the M.Tech degree in Computer Science and Engineering from Jaya Prakash Narayan College of Engineering,JNTUH,India . Presently Working as Assistant Professor in Jaya Prakash Narayan College of Engineering, JNTUH, India.

G Rakesh Reddy, Research Scholar in Computer Science and Engineering at JNTUH and currently working as Assistant Professor in Jaya Prakash Narayan College of Engineering, JNTUH, India.

