ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization *Volume 4, Special Issue 11, September 2015*

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India

Big Data on Cloud Computing- Security Issues

K Subashini, K Srivaishnavi

UG Student, Department of CSE, University College of Engineering, Kanchipuram, Tamilnadu, India

ABSTRACT: Cloud computing is now a new technology which includes sharing of computer resources. The user can access applications in the cloud network from anywhere by connecting to the cloud using the Internet. Some of the real time applications which use Cloud Computing are Gmail, Google Calendar, Google Docs and Dropbox etc. *Also we discuss security issues for cloud computing*,Big data, Map Reduce and Hadoopenvironment. The main focus is on security issues in cloud computing that are associated with bigdata. Big data applications are a great benefit to organizations, business, companies and manylarge scale and small scale industries. Cloud computing plays a very vital role in protecting data, applications and the related infrastructure with the help of policies, technologies, controls, and big data tools. Big Data is a heterogeneous mix of data ie.,both structured (traditional datasets in rows and columns like DBMS tables, CSV's and XLS's) and unstructured data like e-mail attachments, images, PDF documents, medical records. Businesses are primarily concerned with managing unstructured data, because over 80 percent of enterprise data is unstructured. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc.,

KEYWORDS: Hadoop; Big Data; Map Reduce; Distributed file system; Data mining.

I. CLOUD COMPUTING

Cloud Computing is a technology which depends on sharing of computing resources than havinglocal servers or personal devices to handle the applications[2]. Computing means a type of computing in which services are delivered through the Internet. The services are Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS). The goal of Cloud Computing is to make use of increasing computing power to execute millions of instructions per second. Cloud Computing uses networks of a large group of servers with specialized connections to distribute data processing among the servers. Instead of installing a software suite for each computer, this technology requires to install a single software in each computer that allows users to log into a Web-based service and which also hosts all the programs required by the user. Cloud computing technology is being used to minimize the usage cost of computing resources. The cloud network, consisting of a network of computers also the costof software and hardware on the user decreases.

II. CLOUD ARCHITECTURE

Cloud computing architecture refers to the components and subcomponents required for cloud computing. These components typically consists of a front end platform (fat client, thin client, mobile device), back end platforms (servers, storage), a cloud based delivery, and a network (Internet, Intranet, Inter-cloud). Combined, these components make up cloud computing architecture.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India

III. CLOUD STORAGE

An online network storage where data is stored and accessible to multiple clients. Cloud storage is generally deployed in the following configurations: public cloud, private loud, community cloud, or some combination of these three also known as hybrid cloud.

IV. CLOUD NETWORKING

1. High bandwidth: Allowing users to have uninterrupted access to their data and applications.
2. Agile network: On-demand access to resources requires the ability to move quickly and efficiently between servers and possibly even clouds.
3. Network Security: Security is always important, but when you are dealing with multi-latency, it becomes much more important because you are dealing with segregating multiple customers.

CLOUD COMPUTING SECURITY:

It is simply cloud security. It is evolving sub-domain of computer security, network security and information security[1]. It refers to a broad set of policies, technologies, and controls deployed to protect data, applications, and the associated infrastructure of cloud computing.

V. BIG DATA

Big Data is the word used to describe massive volumes of structured and unstructured data thatare so large that it is very difficult to process this data using traditional databases and softwaretechnologies[3]. Relational Database Management Systems(RDBMS) and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers".

BIG DATA CHARACTERISTICS:

The three main terms that signify Big Data[3] have the following properties:

a) Volume: Many factors contribute towards increasing Volume streaming data and data collected from sensors etc.,
b) Variety: Today data comes in all types of formats emails, video, audio, transactions etc.,
c) Velocity: This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.
d) The other two dimensions that need to consider with respect to Big Data are Variability and Complexity.
e) Variability: Along with the Velocity, the data flow can be highly inconsistent with periodic peaks.
f) Complexity: Complexity of the data also needs to be considered when the data is coming from multiple sources.

Technologies today not only support the collection of large amounts of utilizing such data effectively. Also transactions made all over the world with respect to a Bank, customer transactions, and Facebook users generating social interaction data. When making an attempt to understand the concept of Big Data, the words 'Map Reduce' and 'Hadoop' concepts are used.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India

BIG DATA ANALYTICS:

It is really about two things, Big data and Analytics. A problem with Big Data is that they use NoSQL and has no Data Description Language (DDL) and it supports transaction processing. Also, web-scale data is not universal and it is heterogeneous. For analysis of Big Data, database integration and cleaning is much harder than the traditional mining approaches. Parallel processing and distributed computing is becoming a standard procedure which are nearly non-existent in RDBMS[1].Google's technical response to the challenges of Web-scale data management and analysis was simple, by database standards, but kicked off what has become the modern Big Data revolution in the systems world. To handle the challenge of Web-scale storage, the Google File System (GFS) was created. GFS provides clients with the familiar OS-level byte-stream abstraction, but it does so for extremely large files whose content can span hundreds of machines in shared-nothing clusters created using inexpensive commodity hardware. To handle the challenge of processing the data in such large files, Google pioneered its Map Reduce programming model. This model, characterized by some as "parallel programming for dummies", enabled Google's developers to process large collections of data by writing two user-defined functions, map and reduce, that the Map Reduce framework[4] applies to the instances (map) and sorted groups of instances that share a common key (reduce) as similar to the sort of partitioned parallelism utilized in shared-nothing parallel query processing.

VI. 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 projects – Map Reduce and Hadoop Distributed File System (HDFS)[5].

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[1]. Map Reduce has following characteristics ; it supports Parallel and distributed processing, it is simple and its architecture is shared-nothing which has commodity diverse hardware (big cluster).Its functions are programmed in a high-level programming language (e.g. Java, Python) and it is flexible. Query processing is done through NoSQL integrated in HDFS as Hive tool. A Map Reduce[1][4] 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. Map Reduce jobs use efficient data processing techniques which can be applied in each of the phases of Map Reduce: namely Mapping, Combining, Shuffling, Indexing, Grouping and Reducing.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India

HADOOP DISTRIBUTED FILE SYSTEM(HDFS)

HDFS is a file system that spans all the nodes in a Hadoop cluster for data storage[1][5]. 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. HDFS is a distributed file system architecture which encompasses the original Google File System.

BIG DATA APPLICATIONS

The big data application refers to the large scale distributed applications which usually work with large data sets. Data exploration and analysis turned into a difficult problem in many sectors in the span of big data. With large and complex data, computation becomes difficult to be handled by the traditional data processing applications which triggers the development of big data applications.

- Google's map reduce framework and apache Hadoop are the software systems for big data applications, in which these applications generates a huge amount of intermediate data. Manufacturing and Bioinformatics are the two major areas of big data applications.
- Big data provide an infrastructure for transparency in manufacturing industry, which has the ability to unravel uncertainties such as inconsistent component performance and availability.
- In these big data applications, a conceptual framework of predictive manufacturing begins with data acquisition where there is a possibility to acquire different types of sensory data such as pressure, vibration, acoustics, voltage, current, and controller data.
- The combination of sensory data and historical data constructs the big data in manufacturing. This generated big data from the above combination acts as the input into predictive tools and preventive strategies like health management.
- Another important application for Hadoop is Bioinformatics which covers the next generation sequencing and other biological domains. Bioinformatics which requires a large scale data analysis, uses Hadoop.Cloud computing gets the parallel distributed computing framework together with computer clusters and web interfaces.

BIG DATA ADVANTAGES

In Big data, the software packages provide a rich set of tools and options where an individual could map the entire data landscape across the company, thus allowing the individual to analyze the threats he/she faces internally. This is considered as one of the main advantages as big data keeps the data safe. With this an individual can be able to detect the potentially sensitiveinformation that is not protected in an appropriate manner and makes sure it is stored according to the regulatory requirements[2]. There are some common characteristics of big data, such as

a) Big data integrates both structured and unstructured data.

b) Addresses speed and scalability, mobility and security, flexibility and stability.

c) In big data the realization time to information is critical to extract value from various data sources, including mobile devices, radio frequency identification, the web and a growing list of automated sensory technologies.

All the organizations and business would benefit from speed, capacity, and scalability of cloud storage. Moreover, end users can visualize the data and companies can find new business opportunities. Another notable advantage with big-data is, data analytics, which allow the individual to personalize the content or look and feel of the website in real time so that it suits the each customer entering the website. If big data are combined with predictive analytics, it produces a challenge for many industries. *The combination results in the exploration of these four areas:*

a) Calculate the risks on large portfolios
b) Detect, prevent, and re-audit financial fraud

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization Volume 4, Special Issue 11, September 2015

National Conference on Emerging Trends in Computing, Communication & Control Engineering [NCET 2K15]

On 4th September, 2015

Organized by

Department of CSE, ECE & EEE, Magna College of Engineering, Chennai-600055, India

c) Improve delinquent collections
d) Execute high value marketing campaigns

VII. NEED FOR SECURITY IN BIG DATA

For marketing and research, many of the businesses uses big data, but may not have the fundamental assets particularly from a security perspective[2]. If a security breach occurs to big data, it would result in even more serious legal problems and reputational damage. In this new era, many companies are using the technology to store and analyze petabytes of data about their company, business and their customers. As a result, information classification becomes even more critical. For making big data secure, techniques such as encryption, logging, honeypot detection must be necessary. In many organizations, the deployment of big data for fraud detection is very attractive and useful. The challenge of detecting and preventing advanced threats and malicious intruders, must be solved using big data style analysis. These techniques help in detecting the threats in the early stages using more sophisticated pattern analysis and analyzing multiple data sources.

Not only security but also data privacy challenges existing industries and federal organizations. With the increase in the use of big data in business, many companies are wrestling with privacy issues. Data privacy is a liability, thus companies must be on privacy defensive. But unlike security, privacy should be considered as an asset, therefore it becomes a selling point for both customers and other stakeholders. There should be a balance between data privacy and national security.
The challenges of security in cloud computing environments can be categorized into network level, user authentication level, data level, and generic issues[2][3].

Network level:The challenges that can be categorized under a network level deal with network protocols and network security, such as distributed nodes, distributed data, Internode communication.
***Authentication level*:** The challenges that can be categorized under user authentication level deals with encryption and decryption techniques, authentication methods such as administrative rights for nodes, authentication of applications and nodes, and logging.
Data level:The challenges that can be categorized under data level deals with data integrity and availability such as data protection and distributed data.
***Generic types*:** The challenges that can be categorized under general level are traditional security tools, and use of different technologies.

REFERENCES

[1] Ren, Yulong, and Wen Tang. "A Service Integrity Assurance Framework For CloudComputing Based On Mapreduce."*Proceedings OfIeee Ccis2012*. Hangzhou: 2012, Pp 240 –244, Oct. 30 2012-Nov. 1 2012
[2] Hao, Chen, And Ying Qiao. "Research Of Cloud Computing Based On The HadoopPlatform.".Chengdu, China: 2011, Pp. 181 – 184, 21-23 Oct 2011.
[3] 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.
[4] Wie, Jiang , Ravi V.T, And Agrawal G. "A Map-Reduce System With An Alternate Api For Multi-CoreEnvironments.". Melbourne, Vic: 2010, Pp. 84-93, 17-20 May. 2010.International Journal Of Network Security & Its Applications (Ijnsa), Vol.6, No.3, May 2014
[5] K, Chitharanjan, And Kala Karun A. "A Review On Hadoop — Hdfs Infrastructure Extensions.".JejuIsland: 2013, Pp. 132-137, 11-12 Apr. 2013.