International Journal of Engineering Research & Management Technology ISSN: 2348-4039
November-2016 Volume 3, Issue-6
Email: editor@ijermt.org www.ijermt.org

A STUDY ON THE BIG DATA TOOLS AND TECHNIQUES AND LOG ANALYSIS FOR SECURITY

Alok Kumar
Department of Computer Science and Engineering,
SRM University NCR Campus
INDIA

ABSTRACT: Big Data is the large amount of data that cannot be processed by making use of traditional methods of data processing. Due to widespread usage of many computing devices such as Smartphone's, laptops, wearable computing devices; the data processing over the internet has exceeded more than the modern computers can handle. Due to this high growth rate, the term Big Data is envisaged. However, the fast growth rate of such large data generates numerous challenges, such as data inconsistency and incompleteness, scalability, timeliness, and security. This paper provides a brief introduction to the Big data technology and its importance in the contemporary world. This paper addresses various challenges and issues that need to be emphasized to present the full influence of big data. The tools used in Big data technology are also discussed in detail. This paper also discusses the characteristics of Big data and the platform used in Big Data i.e. Hadoop.

Recently, cyber-attack has become the serious national treat such as shut down industry control system, and an act of war. Therefore, the issue is suggested about the necessity of Enterprise Security Management (ESM) that is for integrated management of network system such as firewall, IPS, VPN, and etc. However, current ESM has the limit of blocking only cyber-attack from outside due to using the networking attack detection method that monitoring the traffic inflows from outside to inside

KEYWORDS: Big Data, Hadoop, Map Reduce, Log, Security, Cyber-attack, analysis.

I. INTRODUCTION

Big Data has gained much attention from the last few years in the IT industry. As we can witness billions of people are connected to internet worldwide, generating large amount of data at the rapid rate. The generation of this large amount of engenders various challenges. Along with Big Data's huge benefits too many organizations, the challenges and issues should also be brought into light. A forecast from International Data Corporation (IDC), the Big Data technology and services market represents a fast-growing -multibillion-dollar worldwide opportunity. In fact, a recent IDC forecast shows that the Big Data technology and services market will grow at a 26.4% compound annual growth rate to $41.5 billion through 2018, or about six times the growth rate of the overall information technology market. Additionally, by 2020 IDC believes that line of business buyers will help drive analytics beyond its historical sweet spot of relational (performance management) to the double-digit growth rates of real-time intelligence and exploration/discovery of the unstructured worlds. [1].

The A.P.T attack, which recently raises a significant social chaos by new malignant code, is one of the types of targeting attack in the past. It is difficult to detect and block because it achieves the goal by using whole possible methods after securing information of the target. In the past, cyber-attack was mainly a random attack type. However, recently hackers choose systematically and long-term attack type by cooperating to find weakness of the target. Therefore the company damaged by APT attack has been increased such as Hyundai Capital, Auction, SK communications, and Nonghyup. It is crucial to have ESM (Enterprise Security Management) system that manages integrated in-house network secure systems such as firewall, IPS, and VPN to prevent outflow of company's assets by the treats of intelligent A.P.T attack. Current ESM collects logs and saves it in database system (RDBMS, same as DB), and then

shows present condition on a dash board after analyzing the saved data. At the end it alarms to a manager when there is a problem. Unfortunately, current secure system only blocks cyber-attack from outside because it uses network-based attack detective method which only monitors traffic inflow from outside to inside. It is the reason why it shows the weakness to the method of direct attack to in-house

II. PHASES IN BIG DATA PROCESSING

Before processing big data it must be recorded from various data generating sources. After recording, it must be filtered and compressed. Only the relevant data should be recorded by means of filters that discard useless information. In order to facilitate this work specialized tools are used such as ETL. ETL tools represent the means in which data actually deployment and the organizational context, whereas the latter two are concerned with nature and applicable use of Big Data. Other challenges of Big Data are heterogeneity and incompleteness, scale, timeliness; another closely related concern is data security [4].Processing of BigData using existing technologies and methods is not possible. According to data analytics standard tools have not been designed to search and analyse large datasets. As a result, organizations encounter early challenges in creating, managing, and manipulating large

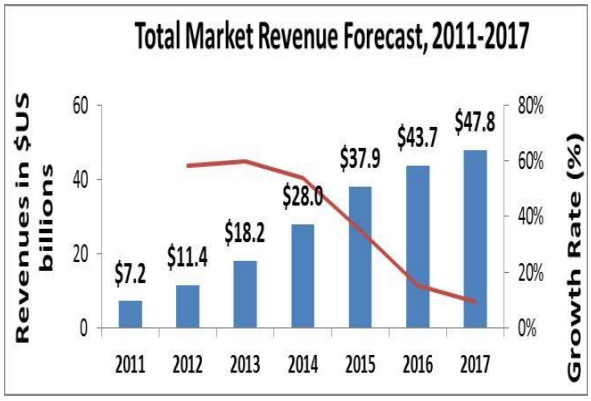

Figure 1: Growth rate of Big Data from 2011-2017[2]

Sr. No.	Properties	Description
1.	Volume	Many factors contribute towards increasing Volume streaming data , live streaming data and data collected from sensors etc.,
2.	Variety	Data comes in all types of formats-from traditional databases ,text documents, emails, video, audio, transactions etc.,
3.	Velocity	This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.
4.	Variability	Along with the Velocity, the data flows can be highly inconsistent with periodic peaks.
5.	Complexity	Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Figure 2. Properties of Big Data

III. BIG DATA OVERVIEW

Big Data is a compendium of big datasets that cannot be processed using traditional computing techniques. It is not a technique that can be worked on its own or in isolation; rather it involves many areas of business and technology. The properties of signify Big Data are volume, Variety, Velocity, Variability and Complexity as shown in figure 2[3]

Big data involves the data produced by different devices and applications. Some of the sources of Big Data are shown in the figure.

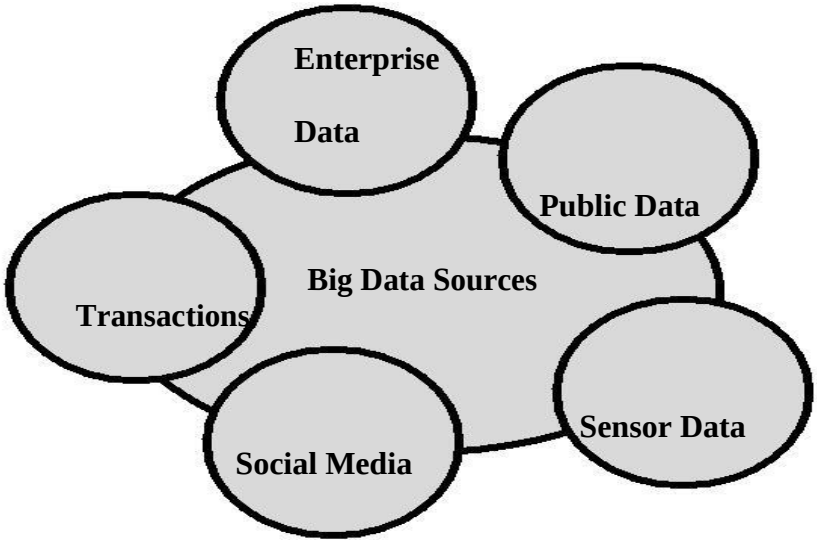

Figure 3. Some Sources of Big Data

IV. BIG DATA CHALLENGES

In computational sciences, Big Data is a critical issue that requires serious attention. There are only two or three main issues appear capable of making or breaking the promise of Big Data, and these are related to: solution approach, personal privacy and intellectual priority (IP). The first issue deals with technology, datasets. Systems of data replication have also displayed some security weaknesses with respect to the generation of multiple copies.

1.Heterogeneity

Machine analysis algorithms expect homogenous data, and cannot understand nuance. In consequence, data must be carefully structured as a first step to (or prior to) data analysis.

2. Scale

As we have examined above that volume or scale is the second major challenge of Big Data. Actually Size is directly related with term BIG and this data is grown rapidly. Managing large and rapidly increasing volumes of data has a challenging issue for many decades. Data volume is scaling faster that computer resources and CPU speeds are static. These unprecedented changes require us to rethink how we design, build and operate data processing components [5].

3. Timeliness

Personal Privacy

In the era of Big Data our personal information that is stored and transmitted through ISPs, mobile network operators, supermarkets, local councils, medical and financial service organizations (e.g hospitals, banks, insurance and credit card agencies). Also information shared and stored on social networking sites like facebook, twitter etc. Privacy is an important issue for everyone. All wants to hide their personal information in order to avoid the misuse of that information. But as the Big Data is grown it is very difficult to achieve. Timeliness is directly related with size, larger the size of data more time is required to process and analyze data. The best system is that which gives user data in correct form on

right time [5].

BIG DATA MANAGEMENT

Hadoop Component	Functions
(1) HDFS	Storage and replication
(2) Map Reduce	Distributed processing and fault tolerance
(3) HBASE	Fast read/write access
(4) HCatalog	Metadata
(5) Pig	Scripting
(6) Hive	SQL
(7) Oozie	Workflow and scheduling
(8) Zookeeper	Coordination
(9) Kafka	Messaging and data integration
(10) Mahout	Machine learning

Tools include Hadoop, Map Reduce, and Big Table. Out of these, Hadoop is one of the most widely used technologies.

Hadoop:

Hadoop is an Apache open source framework which is written in java. High volumes of data, in any structure, are processed by Hadoop. Hadoop allows distributed storage and distributed processing for very large data sets. The main components of Hadoop are:

1. Hadoop distributed file system (HDFS)
2. Map Reduce

The architecture of Hadoop is shown in the figure 3. Hadoop has three layers. The two major layers are Map Reduce and HDFS.

HDFS (Storage layer): Hadoop has a distributed File System called HDFS, which stands for Hadoop Distributed File System. It is a File System used for storing very large files with streaming data access

patterns, running on clusters on commodity hardware. [8] There are two types of nodes in HDFS cluster, namely namenode and data nodes. The name node manages the file system namespace, maintains the file system tree and the metadata for all the files and directories in the tree. The data node stores and retrieve blocks as per the instructions of clients or the namenode. The data retrieved is reported back to the namenode with lists of blocks that they are storing. Without the namenode it is not possible to access the file. So it becomes very important to make name node resilient to failure. [11]
Table 2: Hadoop components and their functionalities.

Map Reduce [16]:-

Map Reduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate keys. Many real world tasks are expressible in this model [16]. The computation takes a set of *input* key/value pairs, and produces a set of *output* key/value pairs. The user of the Map Reduce library expresses the computation as two functions: *Map* and *Reduce*. *Map*, written by the user, takes an input pair and produces a set of *intermediate* key/value pairs. The Map Reduce library groups together all intermediate values associated with the same intermediate key I and passes them to the *Reduce* function. The *Reduce* function, also written by the user, accepts an intermediate key I and a set of values for that key. It merges together these values to form a possibly smaller

V. TECHNIQUES FOR BIG DATA HANDLING

The large complex data into small units and process them. There are many techniques available for data management. That includes Google BigTable, Simple DB, Not Only SQL (NoSQL), Data Stream Management System (DSMS), MemcacheDB, and Voldemort [7] .But these traditional approaches are only applicable to traditional data and not big data as it cannot be stored on a single machine. The Big Data handling techniques and it reads the data from HDFS in an optimal way. However, it can read the data from other places too; including mounted local file systems, the web, and databases. It divides the computations between different computers (servers, or nodes). It is also fault-tolerant. If some of nodes fail, Hadoop knows how to continue with the computation, by re-assigning the incomplete work to another node and cleaning up after the node that complete its task. It also knows how to combine the results of the computation in one place. [9]. the other core components in Hadoop architecture includes Hadoop YARN, it is a framework for job scheduling and cluster resource management. The other component is the cluster which is the set of host machines (nodes).

ARCHITECTURE OF SECURITY LOG SYSTEM USING BIG DATA

1. Intelligent Information Analyzing Platform

Intelligent information analysis platform is composed of collecting, saving, analyzing, and optimized element of data. Each function constructs different formats. They collect data stably from diverse data source, save the data equally by multiple parallel structures, offer the system structure that is able to analyze intelligently based on high-speed searching. There is the intelligent information analysis platform in Table 3.

2. Suggesting the Algorithm of Collecting Massive Data

Develop data collecting structure in considering all data collecting technique, massive data transmission, management stability, and high availability for collecting and engaging data. All of the information created in the security equipment is saved in real-time in collector through data transmitter such as sources, format data, structured/unstructured original log, and original log. The current data collecting process shows vulnerability in unusual traffic analysis while processing Web log in main homepages, WEB/WAS/DB management log, Net-Flow statistic information, detective event of DNS

sinkhole, local network communication record, and detective event of Web symptom. Suggested integrated collector can analyze unusual traffics, detect web symptom, block harmful web site through DNS sinkhole, detect web hacking early by using Web Shell. It is also suggested to use two methods, agent/agent less for collecting information, and add flexibility to selection of collection methods by

considering real-time and stability. Data transmitter automatically disperses error and load of data, and prevents loss of data by using automatic load distribution, detect error/repeat, and log forwarding technique.

3. Data Saving

Collector is constructed by distribution-based log servers. There is the saving method of collector server. The data coming through the collecting system is made for clients to find the information initially through receiving and normalizing process, compared with normalized data, and obtain index value interacted with Index DB. The data of security log, system log, and application log is received and it is normalized through normalizing engine, normalizing file, and data tagging. Use distributed architecture to save the massive security log file. The dispersed architecture is processed in parallel processing to store massive data, and runs saving and real-time indexing work by distribution-based multi indexer. So the Tera byte (TB) data per day can be processed by the distributed architecture and each collector shows 200,000 EPS process performance. Especially that each collector automatically checks integrity when saving data, and saved data in compressed and encoded folder. The collectors automatically backup and restore by constructing data backup/hot spare collector to protect the original data automatically from possible defect of multi system. Theoretically, this management structure can store infinite data, and have expendability and stability. It also makes faster result than a serial process way by arranging collectors in parallel form causing proportion of number of collector and processing performance. The technique can make significantly huge effect to process the massive security log big data when data size is small.

4. Data Analyzing System

The speed of massive data cannot be guaranteed by checking in real-time. However, it can be operated by finding indexing data by entering keywords or conditions of the indexing data saved in collector. The searched massive security log data analysis makes multi-scanning easier by data drilldown that analyzing data in subdividing problems into small pieces. Also the data from security equipment is guaranteed real-time analysis performance by two types of distribution-based multi-scanning. One of them is to detect rapid changes of data based on baseline and threshold value. Another one is using trending analysis that is to predict data based on statistics.

Analyzing the correlation of all the events and intuitionally display it in diagram form by real-time monitoring in equipment/log type develop diverse dashboard of movement of users for visible analysis of data. Turn an alarm to show threat in visualized form when error is found in real-time monitoring process. Maximum of 2 billion cases of single scanning is ran, and scanning in a minute under the condition of simple scanning condition of 200G~400G per day.

VI. CONCLUSION

As there are huge volumes of data that are produced every day, so such large size of data it becomes very challenging to achieve effective processing using the existing traditional techniques Big data is data that exceeds the processing capacity of conventional database systems. In this paper fundamental concepts about Big Data are presented. These concepts include Big Data characteristics, challenges and techniques for handling big data .and I applied collecting, saving, processing, and analyzing techniques based on intelligent information analysis platform for system construction of security log analysis using big data. Saving massive data secures availability and expandability of collected security logs by using an Agent/Agentless way. Also, constructing the system becomes possible

to analyze which was impossible in the past by supporting fast searching and showing visualized method which is now possible to analyze. Moreover, it is expected to enhance customer service satisfaction by inflow of harmful code in-house, and real-time monitoring becomes possible.

In this research, there is a limitation of selecting a part of security field among diverse big data methodology. The extra study of analysis techniques of Big data analysis area can now be applied to diverse fields such as manufacturing, service, and finance as well as security.

VII. REFERENCES

1. https://www.idc.com/prodserv/4Pillars/bigdata
2. www.Wikibon.org
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. Golfarelli, M., & Rizzi, S. (2009). Data warehouse design: modern principles and methodologies. Columbus: McGraw-Hill
5. Almeida, F., and Calistru, C, "The Main Challenges and Issues of Big Data Management", International Journal of Research Studies in Computing, 2(1), 2013, pp. 11-20.
6. https://www.progress.com M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014
7. Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop. apache.org/docs/r1.2.1/hdfs_design.ht
8. Amrit pal, Pinki Aggrawal, Kunal Jain, Sanjay Aggrawal *"A Performance Analysis of Map Reduce Task with Large Number of Files*
9. *Dataset in Big Data using Hadoop"* Forth International Conference on Communication Systems and Network Technologies, 2014.
10. Rahm, E., & Hai Do, H. (2000). Data cleaning: problems and current approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), 3-13.), Apache Hadoop (2013). HDFS Architecture Guide [Online]