ISSN:0975-9646

Big Data Analytics with Hadoop to analyze Targeted Attacks on Enterprise Data

Bhawna Gupta
Deptt. Of Computer Science
Baba Farid College
Bathinda, Punjab

Dr. Kiran Jyoti
Deptt. Of Computer Science
Guru Nanak Engineering College
Ludhiana, Punjab

***Abstract*: Big Data describes data sets that are too large, to unstructured or too fast changing for analysis. Big Data analytics is the process of analyzing and mining Big Data. Due to increase in number of sophisticated targeted threats and rapid growth in data, the analysis of data becomes too difficult. Today's Big Data security analytics systems rely, on untrustworthy data. As organizations open and extend their data networks- allowing partners, suppliers and customers to access corporate information in new and dynamic ways and this becomes more vulnerable to data misuse and theft. Attackers have become more adapt at highly targeted, complex attacks that overtake static threat detection measures. Today's attacks are prepared by advanced technologies are not detected until the damage has been occurred. Now the challenge is collecting and analyzing the Big Data fast enough to contain threats and perform last remediation. In this review paper, we are discussing about technique how Big Data is analyzed by using the technique of Hadoop and why the Big Data Security Analytics is important to mitigate the security threats to secure the enterprise data more efficiently.**

Keywords: Big Data Analytics, Hadoop, MapReduce, Big Data Security Analytics, Targeted Attacks.

1. INTRODUCTION

Due to rapid development of Internet and technology, all the machines are connected to each other either by networked system or via mobile communication. The users are producing more and more data through communication media in the unstructured form which is highly unmanageable and this management of data is the challenging job. The main focus is to gather the unstructured data from all the terminals, processed the data to convert into structured form so that accessing of the data would be easier [1]. For this, always a track is kept on data, that this data or event belongs to which category. Accordingly, data is analyzed and processed to convert it into meaningful and right information by using the concept of Big Data Analytics.

Big Data Analytics accepts the huge data sets and varied data types, both semi-structured and unstructured like videos, images, audio, web-pages, texts or e-mails etc. and convert it into reliable information. Big data analytics describes the simple algorithm for large amount of data without compromising performance. Analysis algorithms are provided directly to database which go beyond the pack and innovate newer more sophisticated statistical analysis. Big Data Analytics use number of tools to do the analysis and processing of data in meaningful way. Hadoop is one of the tools which is aimed to improve the performance of data processing.

With the huge amount of processed data available on internet, hackers also become so active with malicious attacks. Hackers target the analyzed data and create threats for information [2][3]. Big data security analytics is used for the growing practice of organization to gather and analyze security data to detect vulnerabilities and intrusions [4]. The aim is here to make use of Big Data techniques to analyze the data and apply same to implement enhanced data security mechanisms. To obtain data for such systems, organizations pick a variety of hosts with a range of Security Analytics Sources (SAS). It is a system that generates messages or alerts and transmits them to trusted server for analysis and action. It can be Host based Intrusion Detection System (HIDS), an antivirus engine that writes a syslog or interface that reports events to remote service e.g. Security and Information Event Monitoring (SIEM) system. The malicious and targeted attacks have become main subject for government, organization or industry [5]. A subset of threats is Advanced Persistent Threats (APT) which are well-resourced and trained adversaries that conduct multi-year intrusion campaigns targeting highly sensitive economic, proprietary or national security information [6]. Their aim to keep their persistency without getting detected inside their target environments.

II. BIG DATA ANALYTICS

Big data is now a big problem as the volume, variety and velocity of data coming into enterprises continue to reach extraordinary levels. This unexpected growth means that not only must you understand big data in order to process the information that truly counts, but you also must analyze the possibilities of what you can do with big data using big data analytics. Big data analytics is the process of analyzing big data to find hidden patterns, unknown correlations and other useful information that can be extracted to make better decisions. Fig. 1 demonstrates the potential use cases for Big Data Analytics. It shows the

relation between data variety i.e. unstructured, semi-structured and structured data with data velocity from batch system to real time system. It shows that what different analysis can be done by using Big Data Analytics and at which level.

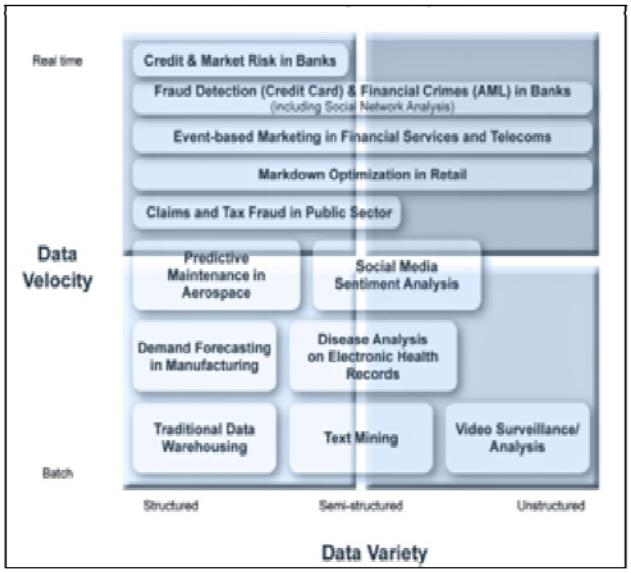

Fig.1. Potential use cases for Big Data Analytics

With big data analytics, scientists and others can analyze huge volumes of data that old analytics and business intelligence solutions can't find. Consider this; it's possible that enterprise could accumulate billions of rows of data with hundreds of millions of data combinations in multiple data stores and abundant formats. Fig. 2 is demonstrating the value of Big Data Analytics by drawing the graph between time and cumulative cash flow. Old analytics techniques like any data warehousing application, you have to wait hours or days to get information as compared to Big Data Analytics. Information has the timeliness value when it is processed at right time otherwise it would be of no use. It might not return its value at proper cost.

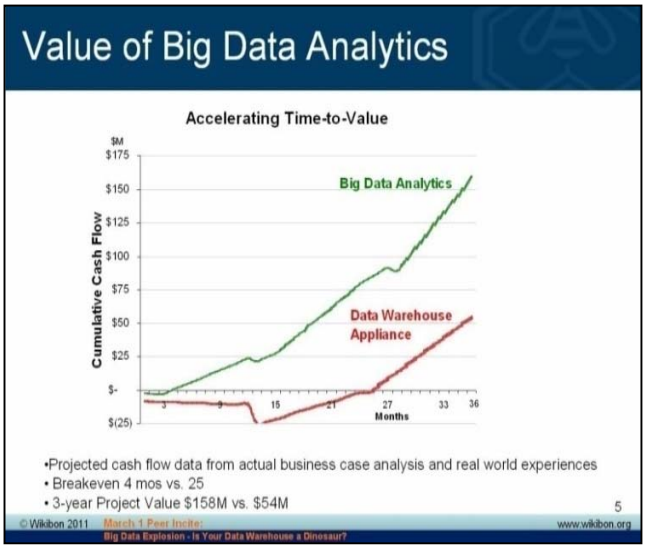

Fig.2. Value of Big Data Analytics

With new advancement in computing technology, you can tackle even the most difficult and challenging business problems. For simpler and faster processing of only relevant data, high-performance analytics can be used [11]. Using high-performance data mining, predictive analytics, text mining, forecasting and optimization on big data, perfect innovation can be extracted to make the best possible decisions. In addition, enterprises are discovering that the unique properties of machine learning are ideally suited to addressing their fast-paced big data needs in new ways [1].

Big data analytics is helpful in managing more or diverse data. It also helps to generalize new questions from observation, formulating new hypotheses, explore and discovery of new processed concepts, and making decisions from testing. The main efforts done by big data analytic is the use of new analytics techniques on either new data or data that has been mixed in new ways.

Big data Analytics used the tool Hadoop for processing the unstructured data. The main point is whether Hadoop will become as indispensable as database management systems. Hadoop has proven its advantages of use and cost where volume and variety are extreme [14]. Cloudera, Hortonworks, and MapR are doing work for Hadoop on high-scale storage and MapReduce processing data into the world of analytics. Data analysis is a do-or-die requirement for today's businesses. Hadoop upstarts to traditional database players by analysis done by vendors.

Today, only 8 percent of large global organizations are using big data analytics to identify patterns [13] and proactively execute attempts to weasel into their payment systems and compromise both their status of enterprise and their customers' financial information. But this 8 percent is expected to jump to at least 25 percent of companies by 2016 [13], according to a new Gartner report, as enterprises using technologies and protocols that can give them faster access to more contextualized information both within and outside their network area.

III. HADOOP - TOOL FOR ANALYSIS

Hadoop is a software framework for storing and processing Big Data and work under Big Data Analytics. It is an open-source tool build on java platform and aimed at to improve the performance in terms of data processing on clusters.

- Hadoop comprises of multiple concepts and modules like HDFS, Map-Reduce, HBASE, PIG, HIVE, SQOOP and ZOOKEEPER to perform the easy and fast processing of huge data [14].
- Hadoop is different from Relational databases and can process the high volume, high velocity and high variety of data to generate value [14].

Fig. 3 demonstrates that how hadoop picked the data in any form and analyze the whole data and take action instantaneously. Hadoop is using Nosql or mongodb languages for processing of the data.

Fig.3. Hadoop converting data into value

Hadoop is designed to process large volumes of information by connecting many commodity computers together to work in parallel in efficient manner [8]. The 1000-CPU(or processor) machines would cost a very large amount of money, far better than 1,000 single-CPU or 250 quad-core machines. Hadoop have tied these smaller and more reasonably priced machines together into a single cost-effective computer cluster.

In a Hadoop cluster, data is distributed to all the nodes of the cluster present on which data can be loaded as shown in fig. 4. The Hadoop Distributed File System (HDFS) will do this distribution of large data files into chunks which are managed by different nodes in the cluster [9]. An active monitoring system then re-replicates the data in response to system failures (if occurs) which can provide partial storage. Even though the file chunks are replicated and distributed across number of machines, they form a single namespace, so their contents are universally accessible.

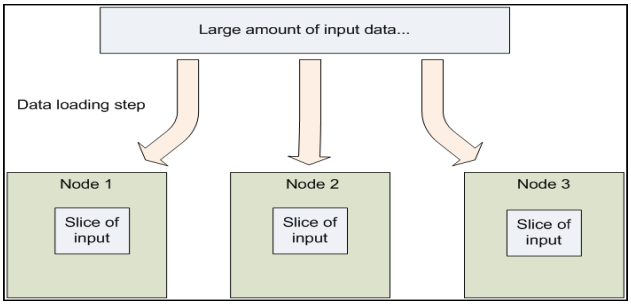

Fig.4. Hadoop cluster

Data is conceptually record-oriented in the Hadoop programming framework. Since files are spread across the distributed file system, each compute process which is running on a node operates on a subset of the data. This strategy of moving computation to the data allows Hadoop to achieve high data locality which in turn results in high performance.

Hadoop limits the amount of communication done by the individual processes, as each record is processed by a isolated task which are different from one another. Programs must be written to a particular programming model, named "MapReduce." In MapReduce, records are processed in separately by isolated tasks called *Mappers*. The output from the Mappers is then moved together into a next set of tasks called *Reducers*, where results from different mappers can be merged together as shown in fig. 5.

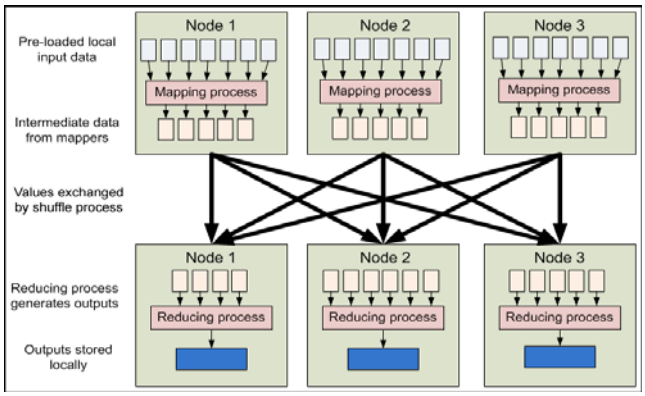

Fig.5. Mapping of records using Hadoop Distributed File System

The Hadoop Distributed File System allows individual servers in a cluster to fail without aborting the computation process by ensuring data is replicated with redundancy across the cluster [10]. There are no limits on the data that HDFS stores as it can be unstructured and schema-less.

In contrast, relational databases require that data should be structured and schemas should be defined before storing the data. With HDFS, making sense of the data is the responsibility of the developer's code.

IV. BIG DATA SECURITY ANALYTICS

There should be number of opportunities for big data security analytics to enter the enterprise security mainstream because of:

1. **Continuing problems with detection and response of threats.** Existing security analytics tools cannot handle advanced virus, malware, stealthy attack techniques, and the growing army of well-organized global cyber attacks [7]. When enterprises get finished buying advanced virus, malware tools from Bit9, Damballa, FireEye, and Invincea, they often think that they still need to replace new layers of defense with real-time and asymmetric big data security analytics [15]. This would generate RFIs/RFPs, evaluations, and actual sales.
2. **Moore's law and open source.** Multi-core 64-bit Intel servers with 10 Gbps network interfaces are small, fast and cheap comparable to older system. These servers have the required horsepower for massive data crunching for stream and batch processing. On the software side, security vendors are increasing development cycles by customizing open source tools like Cassandra, Hadoop, MapReduce, and Mahout for security analytics purposes [15]. This should help to accelerate innovation to protect systems from threats.

3. **Tons of activity on the supply side.** Aside from the usual suspects like HP, IBM, McAfee, and RSA Security, enterprises want security alerts from newcomers. Some, like 21CT, ISC8, Hexis Cyber Solutions, Leidos, Narus, and Palantir will move beyond government business alone and push into the private sector [11][15]. Others, like Click Security, Fortscale, and Netskope have intelligence backgrounds and understand this to a greater extent.

Meaningful security monitoring data has an expiration date i.e. it should follow timeliness property. If it's not used immediately and effectively, hackers can continue to hack your processed information. If it is used effectively and at the same time, hackers can leave their targets forever.

"A year or two ago, hackers would look around, conduct extensive cyberespionage on their targets and then go in for the theft – whether it was for money or information, said Avivah Litan". Now hackers are aware of more effective and latest security and fraud prevention measures configured by their target victim enterprises, simply go directly to the theft without a drawn-out reconnaissance phase."

Gartner says this tripling of big data analytics for security could result in reducing false alarms in existing monitoring systems, correlate high-priority alerts to detect patterns of abuse and fraud and speed up their response by tuning their rules and models against data streaming in near real time [12].

V. ISSUES

Although a lot of research is going on big data but still many concepts are still to be researched. Researchers would try to enhance security platform to improve ability of software to find advanced threats, react accordingly and would develop preventive measures for future. Researchers would try to improve quality and reliability of security system. Some researchers are planning to taken up data collection, pretreatment, integration, MapReduce and analysis using machine learning techniques. They would use the results for securing and implementing preventive measures from threats to enterprise data.

Researchers would try to design the meet the production needs of enterprises for developing high quality product by applying security measures with the help of Big Data Analytics with Hadoop. Some researchers are using networking monitoring tools lie Packet pig, Mahout etc. to enhance the security levels. Researchers are planning to implement more typical targeted threats scenarios and will analyze them using Hadoop cluster. They would calculate efficiency of system to implement preventive measures using campaign analysis.

VI. CONCLUSION

In this paper, we propose the use of Big Data Analytics for analyzing the enterprise data. We discussed a framework based on Hadoop for dealing the targeted attacks using Big Data Security Analytics. We can manage the Big Data characteristics of large volumes of enterprise data. If enterprise has an unmet business need for strategic decision making with a high degree of processing, a Revolution Analytics and Hadoop combination offers significant opportunity to gain advantage.

VII. FUTURE PLANS

Enterprise data security is challenging task to implement and calls for strong support in terms of security policy formulation and mechanisms. We plan to take up data collection, pretreatment, integration, map reduce and prediction using machine learning techniques. We are developing security alerts which will provide employees with the ability to view the activity. Events will be filtered down and summarized view will be available to each individual employee.

REFERENCES

[1] Wang Cheng, Zeng Min, Liu qiong-mei. Practices of Agile Manufacturing Enterprise Data Security and Software protection. 2nd International Conference on Industrial Mechatronics and Automation, 2010.

[2] Wenguang Chai. Analyzes and solves the Top Enterprise Network Data Security Issues with the Web Data Mining Technology. 2009 First International Workshop on Database Technology and Applications, 2009.

[3] Li Xuemei, Li Yan2, Ding Lixing. Study on Information Security of Industry Management. Asia-Pacific Conference on Information Processing, 2009.

[4] J. Oltsik. Defining the big data security analytics. Networkworld, 1 April 2013.

[5] A. K. Sood, R.J. Enbody "Targeted cyber attacks: A Superset of advanced persistent threats' Security & Privacy, IEEE Volume 11, Issue 1, pages 54-61, Jan-feb. 2013

[6] Eric M. Hutchins, Michael J. Cloppert , Rohan M. Amin, "Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains" ,6th International Conference on Information Warfare and Security(ICIW2011) http://www.lockheedmartin.com/content/dam/lockheed/data/corporate/documents/LM-White-Paper-Intel-Driven-Defense.pdf

[7] A.K.Sood, R.J. Enbody "Targeted Cyber attack: A superset of advanced persistent threats" Security & Privacy, IEEE Volume 11 Issue 1, pages 54-61, Jan-Feb, 2013.

[8] Apache Hadoop Project http://hadoop.apache.org/

[9] "Hadoop Tutorial from Yahoo!", Module 7: Managing a Hadoop Cluster http://developer.yahoo.com/hadoop/tutorial/module7.html#machines

[10] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop distributed file system", in poc. The 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), 2010

[11] M.M. Anwar, M.F. Zafar, Z. Ahmed. A proposed Preventive Information Security System. IEEE International Conference on Electrical Engineering, April, 2007.

[12] MacDonald, Neil, 2012, Information Security is Becoming a Big Data Analytic Problem, Gartner, (23 March 2012), DOI= http://www.gartner.com/id=1960615

[13] Larry Barrett, "Big data analytics: the enterprise's next great security weapon?" February 2014.

[14] http://www.edupristine.com/courses/big-data-hadoop-program/big-data-hadoop-course/

[15] Jon Oltsik, Dec 13, 2013. "Strong opportunities and some challenges for big data security analytics in 2014" DOI= http://www.esg-global.com/blogs/strong-opportunities-and-some-challenges-for-big-data-security-analytics-in-2014/