# THE DARWINISM OF BIG DATA SECURITY THROUGH HADOOP AUGMENTATION SECURITY MODEL

**M.S.Mohan Sivam[1], R.Mahesh Muthulakshmi[2], D.Anitha[3]**

[1]*Assistant Professor, Electronics and Communication,* [2,3]*Assistant Professor, Computer Science*
*Indira Gandhi College of Engineering and Technology, Chengalpet, Tamil Nadu, (India)*

## ABSTRACT

*Data pours in millions of computers and millions of process every moment of every day so today is the era of Big Data where data interrelate to the volume, velocity, and variety of data interrelate. Huge volume, various varieties and high velocity create lots of other challenges and issues regarding its management and processing. Big Data enable any organization to collect, manage, analyze and making decision incredibly from large data sets. Big data is growing at an exponential rate but security feature not growing at an same rate. So it becomes important to develop new technologies to deal with it securities. So require latest technology and moderate theory about data, other than the traditional tools and technique to manage it due its nature. This paper introduces the big data technology along with its importance in the modern world and existing projects like hadoop which are effective and important in changing the concept of science into big science. Hadoop, Map Reduce and No SQL are the major big data technology. This paper also throws some light on other challenges and issues. The various challenges and issues in adapting and accepting Big data security and suggest some more security standards and concept that make robust hadoop ecosystem without any processing overhead.*

***Keywords: Bigdata, Hadoop, Mapreduce, ABAC, RBAC***

## I. INTRODUCTION

Data pours in millions of computers and millions of process every moment of every day so today is the era of Big Data. Big data refers to technologies that involve data that is too divers, fast changing or massive for conventional technologies, skill and infrastructure to address efficiently. Said differently the volume, velocity, and variety of data interrelation is too great. Big Data enable any organization to data creation, collection, retrieval, manage, analyze and making decision that is remarkable in terms of volume, velocity, and variety.

***a. Big Data 3 V's are [2].***

1. Volume: At present the data existing is in petabytes and is supposed to increase to zettabytes in nearby future. The social media, financial institution, medical institution, government, Sensors, Logs producing data in order of terabytes every day and this amount of data is definitely difficult to be handled using the existing traditional systems..

2. Velocity: At present data change rapidly through the archived data, legacy collections and from streamed data that comes from multiple resources sensors, traditional file records, cellular technology, social media and many more.
3. Variety: At present data comes in different forms including data-streams, text, picture, audio, video, structured, semi structured, unstructured. Unstructured data is difficult to handle with traditional tools and techniques. Thus our traditional systems are not capable enough on performing the analytics on the data which is constantly in motion.
4. There are volume, velocity and variety are main concern in big data technology. Some other issues are also considerable such as veracity, variability, complexity, Value.

The efflux of Big Data and the need to move this information throughout an organization has created a massive new target for hackers and other cybercriminal activity. Now this data is highly valuable, is subject to privacy laws and compliance regulation, and must be protected. Today the biggest concerns in our present age resolves around the security, privacy with audit access control, robustness, reliability, availability and protection of sensitive information such as financial data, sensors information, medical records, and social information on the social networking.

Big Data’s security in this process is becoming increasingly more important and same time organizations required to enforce access control and privacy restrictions on these data sets to meet regulatory requirements such information privacy laws. Most of Network security breaches from internal and external attackers are on the rise, often taking months to be detected, and those affected are paying the price. Organizations that have not properly controlled access to their data sets are facing lawsuits, negative publicity, and regulatory fines.

***b. Big Data Analytics***

Big Data analytics – the process of analyzing and mining Big Data – can produce operational and business knowledge at an unprecedented scale and specificity. The need to analyze and leverage trend data collected by businesses is one of the main drivers for Big Data analysis tools.

The technological advances in storage, processing, and analysis of Big Data include (a) the rapidly decreasing cost of storage and CPU power in recent years; (b) the flexibility and cost-effectiveness of datacenters and cloud computing for elastic computation and storage; and (c) the development of new frameworks such as Hadoop, which allow users to take advantage of these distributed computing systems storing large quantities of data through flexible parallel processing. These advances have created several differences between traditional analytics and Big Data analytics (Figure 1).

Decreasing cost of storage
Flexibility and cost-effectiveness of datacenters and cloud computng
Development of new frameworks (such as Hadoop)

**Figure 2. Technical factors driving Big Data adoption.**

1. Storage cost has dramatically decreased in the last few years. Therefore, while traditional data warehouse operations retained data for a specific time interval, Big Data applications retain data indefinitely to understand long historical trends.

2. Big Data tools such as the Hadoop ecosystem and NoSQL databases provide the technology to increase the processing speed of complex queries and analytics.
3. Extract, Transform, and Load (ETL) in traditional data warehouses is rigid because users have to define schemas ahead of time. As a result, after a data warehouse has been deployed, incorporating a new schema might be difficult. With Big Data tools, users do not have to use predefined formats. They can load structured and unstructured data in a variety of formats and can choose how best to use the data.

Big Data technologies can be divided into two groups: batch processing, which are analytics on data at rest, and stream processing, which are analytics on data in motion (Figure 2). Real-time processing does not always need to reside in memory, and new interactive analyses of large-scale data sets through new technologies like Drill and Dremel provide new paradigms for data analysis.

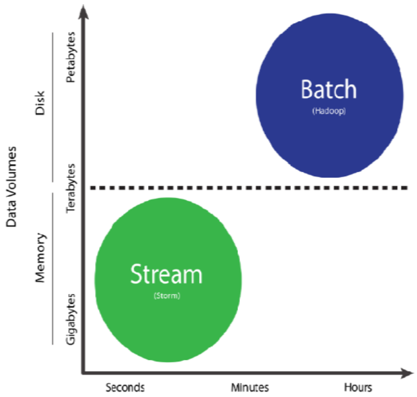

**Figure 2. Batch and stream processing**

***c. Big Data Analytics for Security***

This section explains how Big Data is changing the analytics landscape. In particular, Big Data analytics can be leveraged to improve information security and situational awareness. For example, Big Data analytics can be employed to analyze financial transactions, log files, and network traffic to identify anomalies and suspicious activities, and to correlate multiple sources of information into a coherent view.

Data-driven information security dates back to bank fraud detection and anomaly-based intrusion detection systems. Fraud detection is one of the most visible uses for Big Data analytics. Credit card companies have conducted fraud detection for decades. However, the custom-built infrastructure to mine Big Data for fraud detection was not economical to adapt for other fraud detection uses. Off-the-shelf Big Data tools and techniques are now bringing attention to analytics for fraud detection in healthcare, insurance, and other fields.

In the context of data analytics for intrusion detection, the following evolution is anticipated:

- 1st generation: Intrusion detection systems – Security architects realized the need for layered security (e.g., reactive security and breach response) because a system with 100% protective security is impossible.
- 2nd generation: Security information and event management (SIEM) – Managing alerts from different intrusion detection sensors and rules was a big challenge in enterprise settings. SIEM systems aggregate and filter alarms from many sources and present actionable information to security analysts.
- 3rd generation: Big Data analytics in security (2nd generation SIEM) – Big Data tools have the potential to provide a significant advance in actionable security intelligence by reducing the time for correlating,

consolidating, and contextualizing diverse security event information, and also for correlating long-term historical data for forensic purposes.

Analyzing logs, network packets, and system events for forensics and intrusion detection has traditionally been a significant problem; however, traditional technologies fail to provide the tools to support long-term, large-scale analytics for several reasons:

1. Storing and retaining a large quantity of data was not economically feasible. As a result, most event logs and other recorded computer activity were deleted after a fixed retention period (e.g., 60 days).

2. Performing analytics and complex queries on large, structured data sets was inefficient because traditional tools did not leverage Big Data technologies.

3. Traditional tools were not designed to analyze and manage unstructured data. As a result, traditional tools had rigid, defined schemas. Big Data tools (e.g., Piglatin scripts and regular expressions) can query data in flexible formats.

4. Big Data systems use cluster computing infrastructures. As a result, the systems are more reliable and available, and provide guarantees that queries on the systems are processed to completion.

New Big Data technologies, such as databases related to the Hadoop ecosystem and stream processing, are enabling the storage and analysis of large heterogeneous data sets at an unprecedented scale and speed. These technologies will transform security analytics by: (a) collecting data at a massive scale from many internal enterprise sources and external sources such as vulnerability databases; (b) performing deeper analytics on the data; (c) providing a consolidated view of security-related information; and (d) achieving real-time analysis of streaming data. It is important to note that Big Data tools still require system architects and analysts to have a deep knowledge of their system in order to properly configure the Big Data analysis tools.

## II. RELATED WORK

Currently Hadoop is in initial phase of development many of companies participating in it, our literature also based on companies reports. Some of Hortonworks [3] works with the Hadoop community to bring innovation to the platform, for the enterprise. Employees have collectively contributed more lines of code to Hadoop than any other company. Hortonworks have brought together a collection of resources that are of p articular interest of developers, analyst, and system administration. Also provide tools and training and hadoop solution for business users, java developers, data analyst, data scientist and administrators.

Security is a top agenda item and represents critical requirements for Hadoop projects. Over the years, Hadoop has evolved to address key concerns regarding authentication, authorization, accounting, and data protection natively within a cluster and there are many secure Hadoop clusters in production. H adoop is being used securely and successfully today in sensitive financial services applications, private healthcare initiatives and in a range of other security-sensitive environments. As enterprise adoption of Hadoop grows, so do the security concerns an d a roadmap to embrace and incorporate these enterprise security features has emerged. [4]

1. Securing a Hadoop cluster today according to Hortonworks

a. Authentication verifies the identity of a system or user accessing the system b. Authorization specifies access privileges for a user or system.

c. Accounting provides the ability to track resource use within a system.

d. Data Protection ensures privacy and confidentiality of information. Hadoop and HDP allow you to protect data in motion.

2. Securing a Hadoop cluster tomorrow according to Hortonworks

a. Perimeter level Security with Apache Knox

b. Improved Authentication

c. Granular Authorization

d. Accounting & Audit

e. Protecting data with Encryption

3. According to IBM, Security within the hadoop today [5]: Hadoop supports strong security at the file system level.

Recall that the Hadoop Distributed File System (HDFS) is implemented within another native file system (such as the third extensible file system [ext3]). Access controls for Hadoop are implemented by using file-based permissions that follow the UNIX® permissions model. Although this model provides file-level permissions within the HDFS, it lacks more fine-grained access controls

4. Right now IBM associated following project with Hadoop Ecosystem

a. Sentry with HDFS, Hive, and Impala

b. Project Rhino provide multicomponent security

c. Apache Knox Gateway

d. Delegations Tokens

As an example, consider a file within the HDFS that contains movie reviews for a set of users. This data consists of a user ID, zip code, gender, age, movie title, and review. In Hadoop, access is an all-or nothing model. If you can access the file using the permissions model, you can access all fields within the file. What's needed is a more fine- grained model of access. Where more secure access is granted to all data within the file, lower security access could be provided for individual fields of the data (such as all data except the user ID and zip code). Lower security access minimizes the possibility of leaking user information, and the role-based access of individual fields makes it possible to restrict access within files instead of all-or-nothing file access.[7]

The overall problem of data security within Hadoop becomes even more difficult when you consider its implementation. Hadoop, and its underlying file system, is a complex distributed system with many points of contact. Given its complexity and scale, the application of security to this system is a challenge by itself. Any security implementation must integrate with the overall architecture to ensure proper security coverage.

## III. SECURITY AND PRIVACY ISSUES

There we apply some security concept over hadoop ecosystem and mainly in data processing job. But first of all consider following cases and use of this incremental security process in following condition.

There are some following cases due to security breaches.

**Case 1:**

The 2006 incident, known as the Data Valdez [4], occurred when employees at AOL posted three months' worth of search queries from 650,000 members. AOL employees did so for research purposes, and took steps to "anonymize" the members. AOL made the data available for several weeks on the site research.aol.com. By the time the company realized the privacy implications and pulled the material, the data had already been downloaded by third parties and made available on mirror sites. It's not yet clear how many AOL members will submit claims -especially because many users don't know whether their search queries were publicly released. The settlement notice itself states there is no way for people to determine whether their data was published, based on their usernames.

Hadoop Incremental Security Model provide authorization, authentication and control with encryption using a policy that consider right user meet with its regulatory data.

**Case 2:**

In 2006, Netflix offered a $1 million prize for a 10 percent improvement in its movie recommendation system, and released an "anonymized" training data set of the movie viewing history of half a million subscribers so that developers participating in the contest would have some data to use for the contest. This data set had the ratings of movies that the Netflix subscribers had watched, with all personally identifying information removed.Netflix pays $9M to settle user-data misuse charges; aims to misuse more data with Facebook [6].Netflix accounts that the video-streaming company had kept copies of their personal information and rental history from accounts that had been closed long before.

Retaining data on individual users, as well as anonymized aggregations of data showing user behaviour, makes it easier to recreate recommendation lists for customers returning to the service after having closed previous accounts. Though restrictions on the type of data a service company can keep and the length of time it can retain personally identifiable records could cause endless trouble for non-video-rental companies such as Facebook, they are currently keeping Netflix itself away from Facebook.

Hadoop Incremental Security Model concern all data and its accessibility and use of sensitive information over the system with auditing of the track data provenance.

**Case 3:**

Two researchers, Dr.Arvind Narayanan and Dr.VitalyShmatikov from the University of Texas at Austin, linked together the Netflix data set with the Internet Movie Database (IMDB) review database, applying a new "de-anonymization algorithm."[7] They published a research paper showing that they could mathematically identify many of the users in the released Netflix data set. Based on a user's IMDB ratings of just a few movies, the researchers showed that their algorithm could personally identify the same individuals in the Netflix data set to find the Netflix subscriber's entire movie viewing history prior to 2005, resulting in potential revelations related to the subscriber's religious beliefs, sexuality, and political leanings.

As a result, a Netflix subscriber filed a lawsuit against Netflix, claiming that its release of their data violated the Video Protection Privacy Act (VPPA) and "outed" her as a lesbian. Netflix settled the lawsuit for $9 million in 2010.

*Hadoop Incremental Security Model revoke this types of activity that based on cross domain and retain all information from the server also check third party authentication from ABAC or RBAC.*

## IV. SUCCESSFUL PROTECTION IN THE AGE OF BIG DATA

### a. Use of Big Data to Manage Security Threats

Because of the scale of the Internet and the fact that the world's population is steadily coming online, protecting users from cybercrime can be viewed as a numbers game. The same forces that are driving big data are driving threats concurrently. New methods of addressing cyber threats are needed to process the enormous amount of data emerging from the world and to stay ahead of a sophisticated, aggressive, and ever-evolving threat landscape. No off-the-shelf solution can address a problem of this magnitude. The traditional rules of engagement no longer apply. Scaling up to manage the changes in the threat landscape is necessary, but it must be done intelligently. A brute force approach is not economically viable.

Successful protection relies on the right combination of methodologies, human insight, an expert understanding of the threat landscape, and the efficient processing of big data to create actionable intelligence

Complicating the issue further, security software companies need to not only stop malicious behavior that has already been initiated, but to predict future behavior as well. Predicting the next threat can mean preventing an attack that could potentially cause millions of dollars in damages. Accurate prediction requires knowledge of previous history. Successful security software companies examine past behaviors and model them to predict future behavior. This means employing effective mechanisms to archive historical information, access it, and provide instant reporting and details. Consumers rarely glimpse the enormous amount of effort conducted below the surface to protect them from cyber threats.

### *b. Best Practices in Achieving End User Results*

Addressing today's threat landscape requires a synergistic relationship with customers and other third parties that are constantly exposed to ever-evolving malicious content. A licensing agreement that allows customers to anonymously donate suspicious data for analysis and reverse engineering can provide valuable access to real data on real machines operating in the real world. Based on data gathered from this community network, specialized search algorithms, machine learning, and analytics can then be brought to bear on this data to identify abnormal patterns that can signal a threat.

For example, many computer users follow a typical daily pattern. That pattern may consist of visiting a news site, encountering several ad servers, and logging on to Facebook. If that pattern suddenly changes, perhaps moving the user to a domain never previously visited, this incident can be immediately prioritized for further analysis. These types of complex correlations can be identified only by a system that can perform a very large number of database searches per second.

A feedback loop for process improvement is another critical component. Keen observation and curation of key data that is fed back into the process allows for continual process improvement. Over time, the process can predict malicious behavior long before it occurs

## V. AUGMENTATION SECURITY MODEL

Hadoop Augmentation Security considers following

1. Access Control by Attribute Based Access Control (ABAC) [1] or Role Based Access Control (RBAC) [9] for access, modify and control jobs or precise data access.

2. Encryption of data in transit and rest state.

3. Accountable Audit of the events and track of data provenance.

4. Compliance Assurance for storing sensitive and non-sensitive without replication.

5. Broad usage that cover foundation of concurrency, authentication and authorization.

6. Easier Administration that based on functional role with appropriate access control.

7. Cleansing/Sanitization/Destruction.

8. Data ingest: Data ingestion is the process of importing, extracting and processing data for later use or storage in a database. This process often involves altering individual files by editing their content and/or formatting them to fit into a larger document that begins by validating the individual files, then prioritize, the source for optimal processing and validate results.

## VI. CONCLUSION

This paper described the new concept of big data, its importance and the existing projects. To accept and adapt to this new technology many challenges and security issues exist which need to be brought up right in the beginning before it is too late. All those issues and challenges have been described in this paper. These challenges and issues will help the business organizations which are moving towards this technology for increasing the value of the business to consider them right in the beginning and to find the ways to protest them. Hadoop, the system and its usage grew over the last decade. The early experiment use did not require security. Now security became critical issue in current scenario. As a result, security was recently added to Hadoop in spite of the axiom that states it is best to design and implement security in from the beginning.

## REFERENCES

[1] Computer Security Division Computer Security Resource Center (CSRC), "Attribute Based Access Control (ABAC)",http://csrc.nist.gov/projects/abac/,2015.

[2] Bermen, Jules J. “Principle of Big Data”, Morgan Kaufmann, Waltham, 2013

[3] Das D., & O'Malley O., Security for Enterprise Hadoop Webpage, http://hortonworks.com/labs/security/, 2011

[4] Davis W., AOL Settles Data Valdez Lawsuit For $5 Million Page, http://www.mediapost.com/publications/article/193831/aol -settles-data- valdez-lawsuit-for-5-million.html , 2013

[5] Jones M. T., Hadoop data security and Sentry, http://www.ibm.com/developerworks/security/library/se-hadoop/index.html 2014

[6] LOHR S., A $1 Million Research Bargain for Netflix, and Maybe a Model for Others Webpage, http://www.nytimes.com/2009/09/22/technology/internet/22netflix.html 2009

[7] Lemos R., Researchers reverse Netflix anonymization Webpage, http://www.securityfocus.com/news/11497/1 2007

[8] Roy, I., Setty, S. T. V. S. T. V, Kilzer, A., Shmatikov, V., &Witchel, E., Airavat: Security and privacy for MapReduce. In Proceedings of the 7th USENIX conference on Networked systems design and implementation (pp. 20–20). http://doi.org/10.1.1.149.25332010

[9] Understanding Role Based Access Control, http://technet.microsoft.com/enus/library/dd298183%28v=exchg.150%29.aspx

[10] Zettaset, The Big Data Security Gap : Protecting the Hadoop Cluster. (n.d.).,2014

[11] J. Singh, “Big Data : Tools and Technologies in Big Data,” vol. 112, no. 15, pp. 6 –10, 2015