

Machine Learning Approaches for Big Data Security in Hollerith Tabulation Systems

Ganesh Sahadevan, Padma Joseph

Assistant Professor, SRM University, India

Abstract— This paper gives complete guidelines on BigData, Different Views of BigData, etc. How the BigData is useful to us and what are the factors affecting BigData all the things are covered under this paper. The paper also contains the BigData Machine learning techniques and how the Hadoop comes into the picture. It also contains the what is importance of BigData security. The paper mostly covers all the main point that affect Big Data and Machine Learning.

Corresponding Author:

Ganesh Sahadevan

Keywords: *Dynamic latch comparator, speed, power consumption, high speed analog to digital converter.*

Copyright © 2024: Ganesh Sahadevan, Padma Joseph, This is an open access distribution, and reproduction in any medium, provided Access article distributed under the Creative Commons Attribution License the original work is properly cited License, which permits unrestricted use.

Citation: Ganesh Sahadevan, Padma Joseph, “Hollerith Tabulating Bigdata Security Using Machine Learning Techniques”, Journal of Science, Computing and Engineering Research, 9(3), March 2026.

I. INTRODUCTION

“Big Data” well-known word sounds around us. But only few are aware about it, we only know that Big Data means Large amount of data but it is not just amounts of data only, it is more than that. The term big data has come into use recently to refer to the ever-increasing amount of information that organizations are storing, processing and analyzing, owing to the growing number of information sources in use.

According to research conducted by IDC ([10].IDC Analyze the Future n.d.) there were 1.8 zettabytes (1.8 trillion gigabytes) of information created and replicated in 2011 alone and that amount is doubling every two years. Within the next decade, the amount of information managed by enterprise datacenters will grow by 50 times, whereas the number of IT professionals will expand by just 1.5 times.

The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics. These useful information's for company's helps of gaining richer and deeper insights and getting an advantage over the competition. For this reason, big data implementations need to be analyzed and executed as accurately as possible.

II. RELATED WORKS

Data mainly expanding on Three Views (fronts/views) at an increasing rate.

3.1. Data Volume

3.2. Data Velocity

3.3. Data Variety

3.1. Data Volume General meaning of volume is the amount of space that a substance or object occupies. As we relate with data volume the unstructured data streaming in from social media increasing rapidly required a large volume space.

Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

3.2. Data Velocity In general term velocity is the speed of something in a given direction. Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags (Radio-frequency identification), sensors and smart metering are driving the need to deal with torrents of data in near-real time. Dealing with data velocity is a challenge for most organizations.

3.3. Data Variety Data today comes in all types of formats. Structured, numeric data in traditional databases. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data are vast scope to gather information from, many organizations still struggle with. Now days there are two more additional dimensions has come when thinking about big data

3.4. Data Variability Variability means inconsistency in addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks.

3.5. Data Complexity Today's data comes from multiple sources. It is still an undertaking to link, match, cleanse and

transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

4. Big Data Security

Data volumes continue to expand as they take in an ever-wider range of sources, much of which is in unstructured form. Many organizations want to extract value from that data to uncover the opportunities for the business that it contains.

But the centralized nature of big data stores creates new security challenges. Traditional tools are not up to the task of processing the information the data contains, let alone ensuring it's secure in the process. Colin Tankard of Digital Pathways (Enterprise level security n.d.) explains that controls need to be placed around the data itself, rather than the applications and systems that store the data.

III. TYPE BASED KEYWORD SEARCH FOR SECURITY OF BIG DATA.

5.1.1 System Model We will design a secure big data storage system that supports multiple users. In this system, authorized users are able to store encrypted data and carry out keyword queries on encrypted data without decrypting all the files. Moreover, data owners could delegate certain type of files to other users.

5.1.2 Security Analysis We discuss our type based keyword search for encrypted data from the following security requirements: data confidentiality and query privacy. We assume that users' private keys are kept secret.

IV. TOP MOST CHALLENGES:

- 6.1. Secure computations in distributed programming frameworks
- 6.2. Security best practices for non-relational data stores
- 6.3. Secure data storage and transactions logs
- 6.4. End-point input validation/filtering
- 6.5. Real-Time Security Monitoring
- 6.6. Scalable and composable privacy-preserving data mining and analytics
- 6.7. Cryptographically enforced data centric security
- 6.8. Granular access control
- 6.9. Granular audits

V. DATA SET FOR BIG DATA

A data set is a named collection of data that contains individual data units formatted in a specific, there are many way and accessed by a specific access method that is based

on the data set organization. Types of data set organization include

- 7.1. Sequential
- 7.2. Relative sequential
- 7.3. Indexed sequential
- 7.4. Partitioned Access methods include:
- 7.5. Virtual Sequential Access Method (VSAM)
- 7.6. Indexed Sequential Access Method (ISAM).

8. Dataset freely available Cross-disciplinary data repositories, data collections and data search engines which are freely available most popular dataset are as follow.

8.1. USGOV XML

Dataset US Government Dataset which is freely available ([17].us Government n.d.)

8.2. Amaxon Amazon Dataset ([18].Amazon dataset n.d.)

9. Big Data Attributes There are mainly 5 Attributes of a Big Data Strategy

9.1. Business Case To improve real-time reporting and predictive analysis.

9.2. Architecture It covers Ingestion/extraction job control, data storage area, refinery & data prediction, security, metadata, analytical, data discovery, BI, model execution tool, HW Platform, Hadoop distribution/targeted release.

9.3. Skill Development We must have our own patterns, technique and algorithms to handle the big data.

9.4. Governance Data governance must be exception base.

9.5. Big Data POC Work through architect details, provide a plan based on real-world experience, Test BI/Data Discovery, and provide Sizing information, business use-case validation.

VI. THREATS OF BIG DATA SECURITY INTELLIGENCE

with Big Data solution will empower an organization to address the needs of a changing security landscape. The following are categories of use cases where it can prove at least beneficial if not essential

10.1. Establish a Baseline Organization gains an understanding of its ecosystem, what needs to be defended or observed as well as formulating a risk profile enabling it to detect abnormalities. Common Use Case Questions: Who are the attractive targets within any enterprise? Which applications and what data do someone need to defend due

to their sensitivity? What is the normal behavior profile for users, assets, and applications.

10.2. Recognize Advanced Persistent Threats: Organization gains awareness of a motivated or incentivized attacker who attempts to hide or disguise the attack as innocuous interactions, potentially over a long period of time (months, years). Common Use Case Questions: Which external domains may be the source of attacks? Are there any low profile network traffic elements that might signal an ongoing or imminent attack?

10.3. Qualify Insider Threats Organization gains evidence or is warned of users within the organization's network who may be inclined to steal intellectual property, compromise enterprise systems or perform other actions that are detrimental to the organization's operations. Common Use Case Questions: What data is being leaked or lost and by whom? Who internally has the motivation and skills to compromise the cyber operations of the company? Who is exhibiting abnormal usage behavior?

10.4. Predict Hactivism Organization is alerted to attack from groups or entities that sympathize with causes that are contrary to the business interests of an enterprise. Common Use Case Questions: Which controversial issues may trigger a negative sentiment about the organization triggering an increased risk of attack? How to identify and monitor intentions of entities antagonistic to the organization's business practices? How does publicity of the company in the media impact risk?

10.5. Counter Cyber Attacks Organization is informed of an impending or on-going attack by criminal enterprises or government funded or government sponsored groups. Common Use Case Questions: What is the origin of an attack? Which hacking tools may be used and who is gaining access to them?

Are their symptoms of an attack underway or being planned manifesting themselves as support issues? 10.6. Mitigate Fraud Organization is appraised of new or existing fraud methods that may compromise its compliance with regulations or cause significant losses to its financial operations. Common Use Case Questions: How can the organization identify a fraudulent activity? Which users have compromised identities that may lead to fraudulent activity? Can well known fraud attempts have patterns can either be detected or even anticipated?

11. Solution Combating Advanced Threats with Security Intelligence A good security intelligence solution enables complex problem-solving capabilities, uniquely equipping them to defend against advanced threats. Let's look at critical capabilities of effective security intelligence solutions.

11.1. Consolidation of data silos for 360-degree view Connect the dots between seemingly unrelated or benign activities and ultimately deliver better insight for advanced threat detection.

11.2. Pre- and post-exploit insights Gather and prioritize information about existing security gaps to prevent breaches, as well as suspicious behavior to detect breaches.

11.3. Forensic capabilities Exhaustively research the impact of the breach using captured packet data, easing the burden on the security and network staff who have to build a remediation plan.

11.4. Anomaly detection capabilities Baseline current activity and identify meaningful deviations — a core and vital aspect of detecting advanced threats in progress.

11.5. Real-time correlation and analysis Process massive data sets using advanced analytical methods and purpose-built data repositories, allowing for earlier and more accurate detection of advanced threats, and helping to distinguish the signal from the noise.

11.6. Helping reduce false positives De-prioritize unusual yet benign activity to reduce the time spent investigating anomalous but harmless activity, helping the organization focus on its top incidents.

11.7. Flexibility Constant environmental changes require constant product evolution to add data sources, create and tune analytics, create new user views and reports, and expand and evolve the overall deployment architecture.

11.8. Unified approach Prevention of complex, multi-pronged attacks requires a unified or integrated platform to help organizations intelligently wade through hundreds of security alerts and massive quantities of raw event and flow data.

12. Machine Learning “Big Data” and “Machine Learning” as connected activities. People have been talking about the need for more ‘analysis’ and insight in big data, which is obviously important, because we’ve been in the ‘collection’ phase with big data until now. But the innovation in the big data world that I’m most excited about is the ‘prediction’ phase — the ability to process the information we’ve collected, learn patterns, and predict unknowns based on what we’ve already seen. Machine learning is to big data as human learning is to life experience: We interpolate and extrapolate from past experiences to deal with unfamiliar situations. Machine learning with big data will duplicate this behaviour, at massive scales.

13. Big Data needs Big Compute: Where Hadoop and Spark fit in the picture Think of big data and machine learning as three steps:

13.1. Collect 13.2. Analyze 13.3. Predict These steps have been disconnected until now, because we've been building the ecosystem from the bottom up — experimenting with various architectural and tool choices — and building a set of practices around that. The early Hadoop stack is an example of collecting and storing big data. It allows easier data processing across a large cluster of cheap commodity servers. But Hadoop MapReduce is a batch oriented system, and doesn't lend itself well towards interactive applications; real-time operations like stream processing; and other, more sophisticated computations. 131 International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, March 2016 For predictive analytics, we need an infrastructure that's much more responsive to human-scale interactivity: What's happening today that may influence what happens tomorrow? A lot of iteration needs to occur on a continual basis for the system to get smart, for the machine to "learn" — explore the data, visualize it, build a model, ask a question, an answer comes back, bring in other data, and repeat the process. The more real-time and granular we can get, the more responsive, and more competitive, we can be. Compare this to the old world of "small-data" business intelligence, where it was sufficient to have a small application engine that sat on top of a database. Now, we're processing a thousand times more data, so to keep up the speed at that scale, we need a data engine that's in-memory and parallel. And for big data to unlock the value of machine learning, we're deploying it at the application layer. This means "big data" needs "big compute". This is where Apache Spark comes in. Because it's an in-memory, big-compute part of the stack, it's a hundred times faster than Hadoop MapReduce. It also offers interactivity since it's not limited to the batch model. Spark runs everywhere (including Hadoop), and turns the big data processing environment into a real-time data capture and analytics environment. 14. Five major advantages of Hadoop

14.1. Scalable Hadoop is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel. Unlike traditional relational database systems (RDBMS) that can't scale to process large amounts of data, Hadoop enables businesses to run applications on thousands of nodes involving thousands of terabytes of data.

14.2. Cost effective Hadoop also offers a cost effective storage solution for businesses' exploding data sets. The problem with traditional relational database management systems is that it is extremely cost prohibitive to scale to such a degree in order to process such massive volumes of data. In an effort to reduce costs, many companies in the past would have had to down-sample data and classify it based on certain assumptions as to which data was the most valuable. The raw data would be deleted, as it would be too cost-prohibitive to keep. While this approach may have

worked in the short term, this meant that when business priorities changed, the complete raw data set was not available, as it was too expensive to store. Hadoop, on the other hand, is designed as a scale-out architecture that can affordably store all of a company's data for later use (htt). The cost savings are staggering: instead of costing thousands to tens of thousands of pounds per terabyte, Hadoop offers computing and storage capabilities for hundreds of pounds per terabyte.

14.3. Flexible Hadoop enables businesses to easily access new data sources and tap into different types of data (both structured and unstructured) to generate value from that data. This means businesses can use Hadoop to derive valuable business insights from data sources such as social media, email conversations or click stream data. In addition, Hadoop can be used for a wide variety of purposes, such as log processing, recommendation systems, data warehousing, market campaign analysis and fraud detection.

14.4. Fast Hadoop's unique storage method is based on a distributed file system that basically 'maps' data wherever it is located on a cluster. The tools for data processing are often on the same servers where the data is located, resulting in much faster data processing. If you're dealing with large volumes of unstructured data, Hadoop is able to efficiently process terabytes of data in just minutes, and petabytes in hours.

14.5. Resilient to failure A key advantage of using Hadoop is its fault tolerance. When data is sent to an individual node, that data is also replicated to other nodes in the cluster, which means that in the event of failure, there is another copy available for use (Hadoop n.d.). The MapR distribution goes beyond that by eliminating the NameNode and replacing it with a distributed No NameNode architecture that provides true high availability. Our architecture provides protection from both single and multiple failures. When it comes to handling large data sets in a safe and cost-effective manner, Hadoop has the advantage over relational database management systems, and its value for any size business will continue to increase as unstructured data continues to grow. Michele Nemschoff is vice president of corporate marketing at MapR Technologies.

In comparator circuits to reduce power consumption the Power gating technique is proposed. In this technique, circuit operates in sleep mode by switching off the current in circuit. Power gating has the benefit that is it measures current (I_{dd}) in the quiescent state. In this paper the different architectures of double tail comparator is presented. The proposed comparator is designed by using power gating technique. Using this technique power and delay is reduced.

VII. CONCLUSION

The availability of Big Data, low-cost commodity hardware, and new information management and analytic software has produced a unique moment in the history of data analysis. The convergence of these trends means that we have the capabilities required to analyze astonishing data sets quickly and cost-effectively for the first time in history. These capabilities are neither theoretical nor trivial. They represent a genuine leap forward and a clear opportunity to realize enormous gains in terms of efficiency, productivity, revenue, and profitability. The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work together they can achieve a great benefit.

REFERENCES

- [1]. P. Nirmala, T. Manimegalai, J. R. Arunkumar, S. Vimala, G. Vinoth Rajkumar, Raja Raju, "A Mechanism for Detecting the Intruder in the Network through a Stacking Dilated CNN Model", *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 1955009, 13 pages, 2022. <https://doi.org/10.1155/2022/1955009>.
- [2]. D. Sathyanarayanan, T. S. Reddy, A. Sathish, P. Geetha, J. R. Arunkumar and S. P. K. Deepak, "American Sign Language Recognition System for Numerical and Alphabets," 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-6, doi: 10.1109/RMKMATE59243.2023.10369455.
- [3]. J. R. Arunkumar, Tägele berihun Mengist, 2020" Developing Ethiopian Yirgacheffe Coffee Grading Model using a Deep Learning Classifier" International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4, February 2020. DOI: 10.35940/ijitee.D1823.029420.
- [4]. Ashwini, S., Arunkumar, J.R., Prabu, R.T. et al. Diagnosis and multi-classification of lung diseases in CXR images using optimized deep convolutional neural network. *Soft Comput* (2023). <https://doi.org/10.1007/s00500-023-09480-3>
- [5]. J.R.Arunkumar, Dr.E.Muthukumar," A Novel Method to Improve AODV Protocol for WSN" in *Journal of Engineering Sciences*" ISSN NO: 0377-9254 Volume 3, Issue 1, Jul 2012.
- [6]. R. K, A. Shameem, P. Biswas, B. T. Geetha, J. R. Arunkumar and P. K. Lakineni, "Supply Chain Management Using Blockchain: Opportunities, Challenges, and Future Directions," 2023 Second International Conference on Informatics (ICI), Noida, India, 2023, pp. 1-6, doi: 10.1109/ICI60088.2023.10421633.
- [7]. Arunkumar, J. R. "Study Analysis of Cloud Security Challenges and Issues in Cloud Computing Technologies." *Journal of Science, Computing and Engineering Research* 6.8 (2023): 06-10.
- [8]. J. R. Arunkumar, R. Raman, S. Sivakumar and R. Pavithra, "Wearable Devices for Patient Monitoring System using IoT," 2023 8th International Conference on Communication and Electronics Systems (ICES), Coimbatore, India, 2023, pp. 381-385, doi: 10.1109/ICES57224.2023.10192741.
- [9]. S. Sugumaran, C. Geetha, S. S, P. C. Bharath Kumar, T. D. Subha and J. R. Arunkumar, "Energy Efficient Routing Algorithm with Mobile Sink Assistance in Wireless Sensor Networks," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ACCAI58221.2023.10201142.
- [10]. R. S. Vignesh, V. Chinnammal, Gururaj.D, A. K. Kumar, K. V. Karthikeyan and J. R. Arunkumar, "Secured Data Access and Control Abilities Management over Cloud Environment using Novel Cryptographic Principles," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ACCAI58221.2023.10199616.
- [11]. Syamala, M., Anusuya, R., Sonkar, S.K. et al. Big data analytics for dynamic network slicing in 5G and beyond with dynamic user preferences. *Opt Quant Electron* 56, 61 (2024). <https://doi.org/10.1007/s11082-023-05663-2>
- [12]. Krishna Veni, S. R., and R. Anusuya. "Design and Study Analysis Automated Recognition system of Fake Currency Notes." *Journal of Science, Computing and Engineering Research* 6.6 (2023): 16-20.
- [13]. V. RamKumar, S. Shanthi, K. S. Kumar, S. Kanageswari, S. Mahalakshmi and R. Anusuya, "Internet of Things Assisted Remote Health and Safety Monitoring Scheme Using Intelligent Sensors," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ACCAI58221.2023.10199766.
- [14]. R. S. Vignesh, R. Sankar, A. Balaji, K. S. Kumar, V. Sharmila Bhargavi and R. Anusuya, "IoT Assisted Drunk and Drive People Identification to Avoid Accidents and Ensure Road Safety Measures," 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ACCAI58221.2023.10200809.
- [15]. I. Chandra, G. Sowmiya, G. Charulatha, S. D, S. Gomathi and R. Anusuya, "An efficient Intelligent Systems for Low-Power Consumption Zigbee-Based Wearable Device for Voice Data Transmission," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10083856.
- [16]. G. Karthikeyan, D. T. G, R. Anusuya, K. K. G, J. T and R. T. Prabu, "Real-Time Sidewalk Crack Identification and Classification based on Convolutional Neural Network using Thermal Images," 2022 International Conference on Automation, Computing and Renewable Systems (ICACRS), Pudukkottai, India, 2022, pp. 1266-1274, doi: 10.1109/ICACRS55517.2022.10029202.
- [17]. R. Meena, T. Kavitha, A. K. S, D. M. Mathew, R. Anusuya and G. Karthik, "Extracting Behavioral Characteristics of College Students Using Data Mining on Big Data," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10084276.
- [18]. S. Bharathi, A. Balaji, D. Irene, J. C. Kalaivanan and R. Anusuya, "An Efficient Liver Disease Prediction based on Deep Convolutional Neural Network using Biopsy Images," 2022 3rd International Conference on Smart

- Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 1141-1147, doi: 10.1109/ICOSEC54921.2022.9951870.
- [19].I. Chandra, G. Sowmiya, G. Charulatha, S. D, S. Gomathi and R. Anusuya, "An efficient Intelligent Systems for Low-Power Consumption Zigbee-Based Wearable Device for Voice Data Transmission," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICECONF57129.2023.10083856.
- [20].Revathi, S., et al. "Developing an Infant Monitoring System using IoT (INMOS)." International Scientific Journal of Contemporary Research in Engineering Science and Management 6.1 (2021): 111-115.
- [21].J.R.Arunkumar, Dr.E.Muthukumar, A Novel Method to Improve AODV Protocol for WSN in Journal of Engineering Sciences ISSN NO: 0377-9254 Volume 3, Issue 1, Jul 2012.
- [22].R. S. Vignesh, A. Kumar S, T. M. Amirthalakshmi, P. Delphy, J. R. Arunkumar and S. Kamatchi, "An Efficient and Intelligent Systems for Internet of Things Based Health Observance System for Covid 19 Patients," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ICECONF57129.2023.10084066.
- [23].I. Chandra, K. V. Karthikeyan, R. V, S. K, M. Tamilselvi and J. R. Arunkumar, "A Robust and Efficient Computational Offloading and Task Scheduling Model in Mobile Cloud Computing," 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ICECONF57129.2023.10084293.
- [24].R. K, A. Shameem, P. Biswas, B. T. Geetha, J. R. Arunkumar and P. K. Lakineni, "Supply Chain Management Using Blockchain: Opportunities, Challenges, and Future Directions," 2023 Second International Conference on Informatics (ICI), Noida, India, 2023, pp. 1-6, doi: 10.1109/ICI60088.2023.10421633.
- [25].J. R. Arunkumar, and R. Anusuya, "OCHRE: A Methodology for the Deployment of Sensor Networks." American Journal of Computing Research Repository, vol. 3, no. 1 (2015): 5-8.