Big Data and Its Challenges in Security

[1]Anna Scaria, [2]Anumol Jose, [3]Dr.Regunath K

[1]Student, [2]Student, [3]Asst.Professor
[1,2,3]Department of Computer Science, Santhigiri College of Computer Science, Vazhithala, Thodupuzha, India

Abstract: Because of our ability to create and gather digital data at an amazing rate, the concept of "Big Data" has become a reality. Despite the importance of it, the concept of big data is still often disregarded and undervalued. This particular study adds to the body of knowledge by effectively addressing the opportunities and limitations of Big Data, drawing on 7 case scientific studies of service providers and clients originating from various locations. The processing, analysis, and management of massive data are ineffective when using the present conventional resources, machine learning algorithms, and methodologies. Different scalable machine learning techniques, applications, and strategies (such as open source Hadoop and Apache Spark platforms) are widespread. In this particular research, we've identified the most important issues, including those related to big data, and we've compared numerous approaches to dealing with the issue of handling massive amounts of data.

Index Terms: Big Data, Opportunities, Applications, Security, Need of Security.

I. INTRODUCTION

Big data are gradually becoming ubiquitous. By all accounts, everyone is accumulating, dissecting, and making money from it. Big data are working on it, whether we're talking about dissecting zillions of Google searches to anticipate influenza flare-ups, zillions of phone calls to look for signs of terrorist activity, or zillions of aircraft details to find the optimal time to purchase tickets. It claims to resolve practically any issue, such as wrongdoing, general wellbeing, the growth of language structure, and so on, by combining the impact of modern computers with the vast data of the advanced era .

To maintain an extraordinary level of data quality and accessibility for corporate insight and big data investigation applications is the goal of big data organisation. Big data management strategies are used by businesses, governments, and other organisations to manage rapidly growing data pools, which typically contain many terabytes or even petabytes of information saved in a variety of file formats. Viable big data organisation enables firms to place important information in amazing arrangements of semi-organized[1] and formless data from a variety of sources, such as call detail records, system logs, and web-based social networking sites. In the last few years, there has been a tidal surge of data, and the web is the primary source of that data. Big data is too big, moves too quickly, and doesn't suit the layouts of the database models we've seen. Organizations can enter any reality and gain vital knowledge that was previously unthinkable by using Big Data solutions. Much like the term "cloud" refers to shifting technologies, the phrase "big data" can be confusingly unformulated. Big data usage necessitates changing the information structure to one that is more flexible, appropriate, and open.

Big data promises more in-depth knowledge, and data scientists are incredibly involved in examining this data in order to maximise profits for firms with complete customer approval. One of its vast new wildernesses is big data analysis. Rising innovations like the Hadoop framework and Guide Lessen present innovative and exciting ways to process and transform big data—defined as complex, unstructured, or large amounts of data—into meaningful experiences, but they also necessitate IT to reorganise the foundation in order to support the ongoing demands of big data analysis and the[2] circulated handling requirements. Big data is a broad term used to describe data volumes that are big or difficult to understand because there aren't enough applications for data preparation. Examination, capture, data term, find, part, storage space, transport, attention, questioning, and information division are among the challenges. The phrase repeatedly only refers to the application of explanatory or impacted new sophisticated tactics to eliminate hugeness from data, and inconsistently toward a careful size of data collection. Big data precision may directly lead to more assured decision-making, and better decisions may result in more organised efficacy, decreased cost, and dense plausibility. Big data is data that can handle more information than conventional database frameworks can.

II. CHARACTERISTIC OF BIG DATA

The big data can be characterized by 7 V's. These are listed below:

1. Volume
2. Velocity
3. Variety
4. Veracity
5. Value
6. Variability
7. Visualization

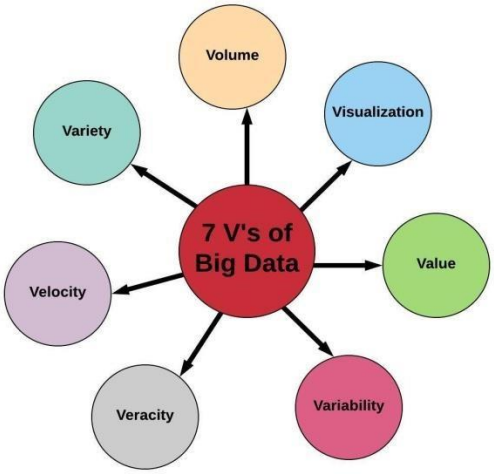

Figure 1: shows the 7 V's of big data III. OPPORTUNITIES & APPLICTIONS IN BIG DATA

There are several[3] opportunities that come with big data some of them are mentioned here. The opportunities with big data.

a.Opportunities of Big Data

- **Data-driven decision making:** Big data technologies have caused a paradigm shift in the way that designs are created. The majority of decisions made today are data-driven, which means that they are based on the data that the system generated, captured, saved, and analysed. This encourages a greater comprehension of the system and how it functions as a whole, which is essential for making wise judgements.
- **In-depth and better insights about the data and the system:** Effective big data mining can lead to the discovery of cuttingedge, previously unidentified patterns, which can be incredibly beneficial for businesses and organisations .
- **Unlocking new horizons of Information:** Big data's worth and visualisation clarity have the potential to revolutionise the basic foundation of information analysis and processing systems. The economic and strategic development of a person, a community, or a nation can all benefit from this information**.**
- **Better training of the systems and individuals:** We can now offer cutting edge training to the systems and people thanks to new tools and technologies like Deep learning, Artificial Intelligence, and Edge computing.
- **State-of-the-art SWOT Analysis:** We can design methods to create the best systems if we are able to examine the strengths, flaws, opportunities, and threats of the system**.**
- **Finding new relationships among data:** If there are any unexpected (yet familiar) relationships between the data, we can use those connections to provide our customers with improved products and services. For instance, on the off chance that we are dissecting a patient's data and we run over two data tests (which were already random) and somehow we process those individual data to reason a different relationship among them, at that point this new connection can be valuable in giving the patient a better diagnosis and course of treatment.
- **Improve Operational Efficiency:** By spotting and examining trends and connections between different operational divisions, big data can be used to enhance an organization's overall operational effectiveness.
- **Identifying new market**: The initiatives can use big data to identify potential customers and new markets for their goods and services. One example of these methods is currently being used by online shopping behemoths like Flip-kart and Amazon, where, in the event that a user selects one item, the website then displays additional items that are related to the item that the user has selected with the slogan "Users who buy this also buys this" or "items frequently brought together."
- **Improve customer satisfaction:** The right handling of big data enables businesses to monitor how their clients use their products (in the form of taking feedbacks, promotional giveaways etc). The ability to return an item or receive a refund if a consumer is dissatisfied with the product or service does wonders to win their business and their trust**.**
- **Informed strategic decision making:** The ability to identify the fundamental requirements and demands of the customers or residents will enable organisations and approach developers to bring about significant systemic improvements that will better serve the clients. This is made possible by examining customer purchasing patterns, posts on social media, blogs, and other online content.

b.Applictions of Big data

There are several applications of big data technology. Some of them are presented here. The[5] various applications of big data technology.

- **Predictive analysis and forecasting systems:** Based on the collection and processing of large data, these systems are able to forecast events that will occur soon. These include technologies for predicting seismic activity and healthcare systems.
- **Pattern Recognition System:** These are systems that are capable of locating previously unknown examples in the data. Finding new connections between the study subjects or items can be aided by this. These systems, for example, can be applied to DNA analysis and forensic sciences to produce cutting-edge findings.
- **Smart Education:** The institutions and regulatory authorities can design policies and regulations that are student-centric and provide a holistic development of the students and instructors alike with adequate analysis and processing of the big data related to the educational sector.

- **Learning Analytics:** It is described as a process of evaluating and keeping track of students' performance in order to spot their weak spots and problem areas and offer solutions to strengthen their focus, retention, and achievement in subsequent endeavours.
- **Industrial Management:** Companies and enterprises are currently using big data to identify fundamental system components both inside and outside the system in order to rebuild and adapt the working model in order to increase production with minimal and practical maintenance while also improving employee and customer satisfaction.
- **Translations Systems:** Big data and sophisticated learning algorithms can be combined to create cutting-edge tools, apps, and systems that continuously understand images, texts, sounds, and recordings. Fundamentally, these methods help advance holistic development by reducing semantic boundaries in training, research, and culture. For instance, Tiny Eye, Google Translate.
- **Smart Transport:** Applications for smart transportation, including real-time traffic analysis, alerts for alternate routes in the event of a traffic congestion, and the identification of the best and quickest routes to a destination, are now available and in use around the world.
- **Smart Healthcare:** Using big data technologies can enable innovative and cheap medical services. It is possible to develop widespread, financially astute, and adaptable human services models with increased reach and distant access. Some of these, such as Open APS, CGM systems, activity trackers, connected inhalers, ingestible sensors, etc., are already in use.
- **Smart Buildings:** Buildings and spaces equipped with IoT sensors are no longer a pipe dream. By lowering electricity use while maintaining a gorgeous environment for workplaces and homes in accordance with the desires of the clients, they are making great strides toward protecting nature.
- **Smart Manufacturing and Production:** Utilizing the hidden information in big data, innovative product development life cycles are being created, allowing manufacturers and industries to offer consumers high-quality services.
- **Behavioural Systems:** The best examples of behavioural systems that learn from their past actions to respond to current or upcoming circumstances are self-driving cars and drones.

IV. BIG DATA SECURITY ISSUES

A Big Data key is a distinct information system in and of itself, including applications, handling components, networks, and data storage, but with the unique feature that it requires extensive use of data from a variety of sources, as well as dispersed preparation and storage resources. Not unexpectedly, the[6] security precautions should be significant for this kind of circumstance, just as they are for every information system.

Typically, they fall into three categories:

- **Security Issues:** Big data is concerned with the storage, processing, and retrieval of data. These uses include networking, virtualization, memory management, transaction management, and many more technologies. Therefore, large data security challenges also apply to these technologies. Big data security concerns at the following four different levels: network, data, and generic[10].
- **Authentication level issues:** There are several nodes and clusters present. Each hub has distinct requirements or rights. Any node with access credentials can access any data. Nevertheless, there are times when a malicious hub will take or control basic client information if it has a need for authority. Numerous nodes join clusters for faster parallel processing and execution. In the absence of evidence, the cluster may be bothered by any vengeful hub. In big data, logging plays a key role. In the unlikely event that logging is not provided, no action that modifies or deletes data is logged. In the unlikely event that a new hub joins the cluster, this won't be seen because of logging nonappearance. Clients occasionally might use harmful data if a log isn't provided. However, it will be difficult to recover that data if any copy or data from another hub is deleted or controlled by a programmer.
- **Network level issues:** Clusters contain a large number of nodes, and these nodes are where calculations and data processing take place. Anywhere among the cluster's nodes should be able to process data in this manner. Therefore, it is challenging to determine which hub is processing data. This issue makes it difficult to determine which hub needs security. At least two nodes can communicate with one another and share data and resources through a network. RPC (Remote Procedure Call) is frequently used for network-based communication. In any event, unless and unless it is scrambled, RPC is not verifying.

  Many different technologies are used in the big data environment to handle the data, as well as some traditional security tools for security objectives. Traditional tools have been developed over time. Therefore, it's possible that these tools won't work well with the newly popularised type of big data. Big data uses a variety of technologies for data storage, processing, and recovery, hence there may be some complications due to these many technologies.

V. NEED OF SECURITY IN BIG DATA

Many firms use big data for marketing and research, but they might not have the resources they need, particularly in terms of security. In the event that big data suffers a security breach, there would be a lot more real legal implications and reputational loss than there are currently. Many businesses today are using the technology to store and analyse petabytes of data on their company, industry, and clients.

As a result, information organisation becomes considerably simpler. Systems like encryption, logging, and nectar pot recognition are crucial for making massive data secure. Big data organisation for extortion location is desirable and beneficial in many enterprises. Big data analysis is required to appreciate the difficulty of identifying and foreseeing advanced threats and vengeful intruders. These systems assist in identifying threats in the early stages by dissecting many data sources and using more sophisticated example analysis. Existing businesses and government entities have issues related to security and data privacy. Numerous firms are struggling with privacy challenges as big data usage in business increases. Companies must be cautious about privacy since data privacy is a risk. However, unlike security, privacy should be viewed as a benefit; as a result, it becomes a selling point for both customers and other partners. National security and data privacy should coexist in harmony.

a.Importance of Infrastructure-Level Optimizations

Cloud computing can be thought of as a type of network-based computing that refers to the on-demand distribution of remote computing resources. Instead of needing to create and maintain local infrastructure, it allows Cloud users to use computing resources as a utility. Users of the Cloud can flexibly allocate computing resources from the Cloud and release them freely as needed within this shared resource pool. Cloud computing can dramatically lower costs and free resource users from onerous maintenance compared to local infrastructures. This makes cloud computing a better option for people who just require the resources temporarily or who lack the skills to develop local infrastructure. The winners come in all shapes and sizes, from established businesses to start-ups. In this situation, cloud computing becomes a popular subject in the scholarly and industrial circles. Numerous businesses and academic institutions have developed a number of implementations, including public cloud and private cloud platforms (e.g., Open Stack). The Apache License governs the release of the open source project known as Open Stack. As a cooperative venture by Rack space Hosting and NASA, it started in 2010. Not only does Open Stack make it simple and quick for customers to construct private clouds, but it also enables on-demand cloud customization.[7]This enables several researchers to go further into the Cloud's architecture in order to pinpoint issues, find novel solutions, integrate them, and test them on a private Cloud test bed. Therefore, Open Stack is crucial for the advancement of cloud computing because it is an open source initiative.

In a complicated computing environment like a hybrid or heterogeneous cloud, many academics work to assist users in reducing the cost of their consumptions while maximising performance. They offer a variety of techniques to analyse potential performance and unit costs, along with different combinations to identify the best resource allocation plans. Contrarily, cloud service providers aim to provide for more customers while using fewer resources, which results in cheaper prices (e.g., monetary expenditure, energy consumption). The ability of cloud computing to manage resources autonomously and dynamically is known as scalability or elasticity, and it is a hot topic in the field of cloud computing. These studies cover a wide range of topics, including self-adaptation, cybernetics, etc. This trait serves as a key differentiator between cloud computing and other cluster or grid computing. Security affects not only the user's privacy but also the regular use and upkeep of cloud computing. Beyond algorithm creation and[4] architecture optimization, resource sharing and multi-tenancy have a significant impact on these studies. The ability to withstand hazards is strengthened by reliability, which enhances cloud computing. This is crucial for cloud computing, especially when there is congestion or overload that might easily lead to user requests failing. Agility enables users of the Cloud to swiftly supply or de-provision resources, which obviously impacts the user experience and other factors. To the best of our knowledge, no methods have yet been put out for creating the infrastructure-level optimization we aim for, one that draws inspiration from trash collectors and is applied to the Cloud. The strategy that utilises Platform-as-a-Service (PaaS) technology is the most comparable.

VI. CONCLUSION

Big data refers to data that is so large, complex or fast that it's difficult or impossible to process using traditional data base methods used. The act of accessing and storing large amounts of information for analytics has been around for a long time. Big data allows companies to improve their products and create tailored marketing by gaining a 360 -degree view of their customers' behavior and motivations. Therefore, the platform- and infrastructure-level resource management in Big Data is the main focus of this article. Two unique arrangements are suggested to optimise resource usage in both layers, i.e., platform and infrastructure, based on the various resource leaks produced by misconfiguration and firm framework mechanisms. Elasticity is widely used in current research as the major technique for dynamically tuning framework resources to ensure the framework performance is maintained within a suitable range. In any case, beyond scaling the infrastructure, our recommendations in this thesis aim to achieve the best performance based on provisioned resources, or to maximise resource consumption to serve as many people as is practical. As a result, we think the study findings in this thesis can be used to augment existing elasticity research, both at the infrastructure and platform levels. In this thesis we have discussed about the Big Data and its opportunities, Applications, Security and Need of Security. Methods like Block chain and encryption can be used as the solution for the security issues of Big Data.

REFERENCES

1. Ren, K., Gibson, G., Kwon, Y., Balazinska, M., and Howe, B. Hadoop's Adolescence; A Comparative Workloads Analysis from Three Research Clusters In SC Companion: High Performance Computing, Networking Storage and Analysis (2012).
2. Romano, P. Elastic, scalable and self-tuning data replication in the cloud-tm platform. In Proceedings of the 1st European Workshop on Dependable Cloud Computing (New York, NY, USA, 2012), EWDCC '12, ACM, pp. 5:1–5:2
3. Pandey, S., Wu, L., Guru, S. M., and Buyya, R. A particle swarm optimization-based heuristic for scheduling workflow applications in cloud computing environments. In Advanced information networking and applications (AINA), 2010 24th IEEE international conference on (2010), IEEE, pp. 400–407
4. Polo, J., Becerra, M. Adaptive map reduce scheduling in shared environments. In Cluster, Cloud and Grid Computing, 14th IEEE/ACM International Symposium on (2014), pp.61– 70
5. H., Lent, R., Mahmoodi, T., Sannelli, D., Mezza, F., Telesca, L., and Dupont, C. Energy efficient resource allocation strategy for cloud data centres. In Computer and information sciences II. Springer, 2011, pp. 133–141.
6. Ren, K., Gibson, G., Kwon, Y., Balazinska, M., and Howe, B. Hadoop's Adolescence; A Comparative Workloads Analysis from Three Research Clusters In SC Companion: High Performance Computing, Networking Storage and Analysis (2012).
7. Romano, P. Elastic, scalable and self-tuning data replication in the cloud-tm platform. In Proceedings of the 1st European Workshop on Dependable Cloud Computing (New York, NY, USA, 2012), EWDCC '12, ACM, pp. 5:1–5:2