ISSN: 2754-6705

Journal of Mathematical & Computer Applications

SCIENTIFIC Research and Community

Review Article

Open Access

Performance Enhancing Analysis for Data in Motion in Big Data

Kartheek Pamarthi

USA

ABSTRACT

Big data systems are growing in size from terabytes to petabytes and zetabytes, and they have features that are structured, semi-structured, and unstructured. As a result, ensuring the safety of the data while it is being stored and moved is currently the most important requirement for businesses. Organizations like NIST are now working to define the architecture framework for big data security because security is still a relatively new concept in the big data business world. Within the context of a big data security framework, this paper explores the need, challenges, and potential solutions for protecting the privacy and integrity of data in transit. Protocols that are specialized to Big Data, such as Hadoop RPC and HDFS, as well as standard network protocols, such as Kerberos, SSL, and TLS, can work together to ensure the secure movement of large amounts of data. With the help of the AES cryptographic algorithm and its several variants, this study begins a collaborative system for the transfer of the contents of massive data. This study delves into the AES cryptographic system and its potential for performance optimization in Big Data settings, with the aim of achieving outstanding outcomes. It is proposed here that a structure consisting of five steps, which may be implemented in large data systems such as Hadoop, is a structure that focuses on efficiency and performance criteria.

***Corresponding author**
Kartheek Pamarthi, USA.

Received: May 16, 2024; **Accepted:** May 22, 2024, **Published:** May 29, 2024

Keywords: Hadoop, SAP HANA, Big Data, Data in Motion, Data at Rest

Introduction

Ability to make decisions based on data analysis is what we mean when we talk about decision making. Following the completion of the processing of data, the task of extracting meaningful information from the data is known as data analytics. Data analytics may be broken down into four distinct categories: descriptive, diagnostic, predictive, and prescriptive [1]. Each of these categories employs a unique set of algorithms and methods to conduct analysis. The goal of descriptive analytics is to make sense of what happened by combining the existing data. The goal of diagnostic analytics is to determine what caused an event to happen by providing a thorough understanding of the situation. Predictive analytics is the practice of looking at trends and patterns in past data to guess what might happen next. With the use of prescriptive analytics, we may narrow down a list of potential decisions to the one that will help us seize future chances while avoiding potential threats. All of this is carried out using the readily available data [2]. Data analysis would not be complete without processing the data. Data can be processed in two different ways: stream processing and batch processing. One way to think about it is that stream processing is processing data as it moves through the system in real time, as opposed to batch processing, which is processing data after it has already been collected. In Table 1 we can see a brief comparison of these two methods of data processing.

Table 1: Data Processing Methods Compared

Aspect	Stream Processing	Batch Processing
Data size	Unknown	Known
Performance	Limited time, it can be seconds or milliseconds	No limit, it can be hours or days
Dataset type	Unbounded	Bounded
Processing	It is processed only once	It can be processed many times
Example	E-commerce transactions	Payroll system

Data stream processing presents greater challenges than batch processing. Moreover, big data streaming analytics has certain challenges due to the characteristics of massive data. For instance, conventional approaches and data mining tools provide challenges when applied to big data stream analysis owing to the rapidity of data inflow and the diversity of data formats [3]. Another difficulty is figuring out how to handle them and get valuable insights from the data as it happens [4]. You can see the lifespan of streaming data in Figure 1.

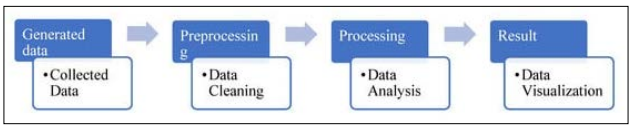

Figure 1: Streaming Data Lifecycle

Businesses and organizations can't react to changing circumstances in real time unless huge data streams are processed in real time [5]. Case in point: smart cities, oil and gas, transportation, and manufacturing can all benefit from real-time big data analysis from IoT devices; this allows for the early detection of anomalous data and the subsequent fixation of problematic issues before they escalate into larger and more expensive ones. Data cleaning and analysis go hand in hand; so, higher data quality will lead to more reliable and accurate results, which in turn will assist organizations in making more informed decisions. Therefore, improving data quality is mostly dependent on data cleaning.

Table 2 lists the ten existing combinations of big data processing methodologies that this study intends to compare and contrast in order to assess how well the proposed framework performs. Data Storage module retrieval and batch processing within a specified time span are the hallmarks of Historical Data Processing, which #1–#3 of these combinations simulated. Apache Cassandra and Apache HDFS were used to construct the Data Storage module. Data was transferred from the Data Storage module to the Data Processing module, which contains tools such as Apache Spark and Apache Storm, at different throughputs.

Table 2: Integrated uses of Preexisting Big Data Processing Systems

Module	Scenario		Application Controller Module	Message Broker Module				Data Processing Module		Storage Module	
Experiments	Real-Time	Historical	Spring MVC	Kafka	RabbitMQ	Redis	NAT/ NATS	Spark	Storm	HDFS	Cassandra
#1		√	√					√		√	
#2		√	√						√	√	
#3		√	√						√		√
#4	√		√		√			√			
#5	√		√			√		√			
#6	√		√				√	√			
#7	√		√	√					√		
#8	√		√		√				√		
#9	√		√			√			√		
#10	√		√				√		√		

Literature Review

Big Data and Security-A Hadoop Approach

Before diving into Big Data, double-check that your architecture has all the necessary parts to analyze massive amounts of data from every angle. You will struggle to manage such a data deluge unless you have this setup right. The design of a Big Data system ought to be able to swiftly and cheaply absorb a multitude of data sources. Additionally, it should contain the following layers: data sources, ingestion, visualization, security, infrastructure, storage, and management of the Hadoop platform [6-8].

Visualization

This layer is helpful for data analysts and scientists because it allows them to grasp the data more effectively and more quickly, and as a result, this layer makes it possible for them to study various parts of the data using a variety of visual modalities [7].

Hadoop Platform Management

This layer provides the tools that are required for processing MapReduce as well as the query languages that are required in order to access NoSQL databases through the utilization of the distributed HDFS storage file system (PIG, HIVE, Sqoop, and so on) [8].

Hadoop Storage

This layer is specifically designed for the purpose of storing data through the utilization of massively dispersed storage and processing, which represents a shift in the manner in which an organization manages its data. Hadoop relies on HDFS, a distributed file system designed to store massive amounts of data (terabytes or petabytes) across many machines in a cluster [9-11]. It is able to store data in a dependable manner, it operates on fundamental hardware, it saves a file or a portion of a file using blocks, and so on. The Big Data architecture is shown in below figure 2.

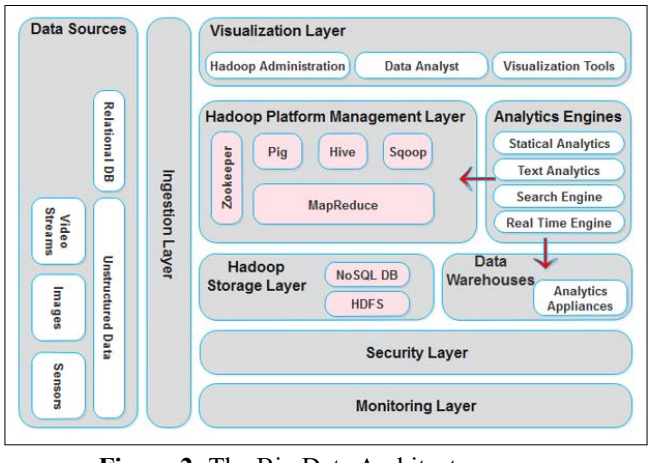

Figure 2: The Big Data Architecture

Security

Due to the fact that the safety of this data becomes a key concern, this layer was developed specifically for the purpose of protecting it. It is necessary to protect a wide variety of data kinds and uses, including but not limited to client purchasing habits, patient medical history, demographics of hereditary diseases, and many others. This is necessary not only to fulfill compliance requirements but also to safeguard the personal data of individuals. It is imperative that these security precautions are incorporated into any Big Data system from the very beginning [12].

Monitoring
Due to the extensive network of data source ingestion points and dispersed data storage clusters that make up the Big Data system, it is crucial to acquire a complete picture of the system by means of monitoring tools. Because of this, this layer is responsible for defining the concepts that are utilized by various monitoring systems in order to improve Hadoop's performance [13].

Ingestion
This layer makes it possible to distinguish between the noise and the information that is pertinent. The enormous amount of data, the rapid speed, and the variety of data must all be within its capabilities. Beyond this, it needs data validation, cleaning, transformation, reduction, and integration capabilities so they may be used in the future by the Hadoop ecosystem [14].

Data Sources
Within a Big Data solution, this layer is responsible for defining the many kinds of data sources that are both internal and external to the organization and that need to be evaluated. Big data is characterized by a huge amount, variety, velocity, and value of data. Big data is also characterized by its velocity. As a result, it is a complicated data flow that must be handled in an impeccable manner in the intake layer [15].

Efficient Encryption Scheme for Big Data
The Right Algorithm for Data in Motion According to statistics and performance evaluations, the Advanced Encryption Standard (AES) is the most effective symmetric cypher in terms of strength and performance. The conventional AES algorithm has a memory complexity of 232 and a time complexity of 248 per second. Most new Big Data implementations now support standard AES algorithms and all of its important variants, modes, and other parameters. Along with this, there are other, less secure standards like 3DES and RC4. A conventional symmetric cryptosystem's encryption performance is illustrated in Figure 3, which describes the system.

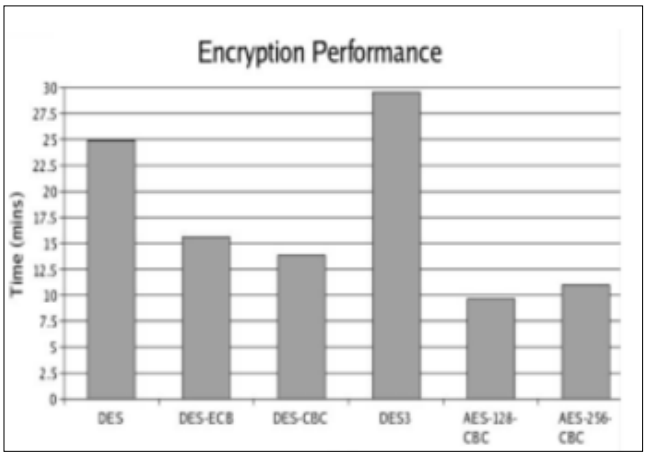

Figure 3: Efficiency of OpenSSL-Based Symmetric Algorithms with 8 GB of Data on an Intel 1.0 GHz Server

Big Data in Motion – Real Time Analytic Solutions for 21st Century Challenges
U.S. efforts to safeguard vital infrastructure from asymmetrical threats like car bombings and hijacked planes have received massive funding in the ten years following 9/11. Unfortunately, the remote locations of our most important infrastructure, like pipelines, ports, refineries, and power plants, make them just as susceptible to attacks and more difficult to safeguard. To make matters worse, the majority of these facilities are vulnerable to cyberattacks because they rely on Web-based command, control, and communications technologies. There are two ways that we have dealt with physical danger. When feasible, for starters, we should fortify these facilities, and secondly, we should set up many layers of sensors to sound the alarm anytime danger is in the area. In an effort to detect and localize security risks, a single facility may be equipped with a plethora of sensors, including magnetometers, video cameras, motion detectors, and listeners. The development of network flow sensors to detect cyber breaches into our networks is a collaborative effort between the U.S. government and private sector, however it has not yet seen widespread deployment. The massive amount of data coming in from these connected sensors is, of course, a difficulty to manage. The difficulties of Big Data have been widely publicized by this point. However, dealing with Big Data in Motion is the biggest problem for security professionals tasked with protecting infrastructure. If we wait two days—or even two minutes-to review the data from the sensors, it will probably be too late to avert an attack. Also, the challenges of big data were shown in figure 4.

Big Data analysis must occur in real-time while it is rapidly coming in from sensors, rather than after the fact, when the data has become static. Data that is not actively being processed is not as valuable. Volume, velocity, and variety of data all pose problems for processing and analysis in Big Data in Motion. Security experts, particularly those in the military and intelligence sectors, have discovered that by integrating the analysis of numerous sensor data streams in motion-and occasionally comparing this real-time data with historical data stored in a database-threats can be detected and even predicted earlier. Due to the large number of raw data types that need to be analyzed-often while considering non-linear correlations between and within the data sets-data variety exacerbates processing issues. We need this done in near-real time if at all feasible. There is no way to overstate the significance of doing holistic analyses of numerous streams of sensor data. It may not be necessary to be concerned about a single unusual occurrence, such as an unusual car captured on surveillance footage in the parking lot. On the other hand, it may signal problems if it occurs concurrently with the swiping of a card key at an off-hours side entrance. Without correlative analysis, which involves studying several cyber and physical data sets integratedly, it becomes more complicated to prioritize the level of threat and appropriate response. A novel concept in technology, Real Time Analytical Processing (RTAP) claims to have the answer to the Big Data in Motion issue. The foundations of RTAP are already utilized by models that aim to predict future events by continuously examining data from many sources, such as weather reports, crop status, and even Twitter activity, regardless of the time horizon (one year or one split second).

In order to trigger an appropriate response, these models continuously calculate and recalculate the chance of certain occurrences happening until a pre-determined threshold is exceeded.
While RTAP has achieved a lot, it still needs a lot of breakthroughs from the top names in IT. Up until now, the majority of the developments have concentrated on revising the RTAP code to speed up the analysis of one or more data streams to a fraction of a second. Developers of software, hardware, databases, and algorithms must collaborate on supplementary breakthroughs if this field is to continue to thrive. When it comes to security, RTAP upgrades will center on making it easier for sensors and detectors to incorporate complicated analytics, rather than just collecting

data and sending it somewhere for processing and analysis. This is true for cyber sensors as well as physical ones. Reducing or eliminating the time it takes to go from capture to analysis is crucial for RTAP to work. The current focus of RTAP research is on developing solutions that integrate analytic modeling capabilities into sensors at the point of data gathering. The security industry must think about the shifts it has to make in order to integrate RTAP into critical infrastructure protection, in addition to technology advancements. Since RTAP is fundamentally an IT solution, the first stage, according to previous industries' experiences, is to integrate security features with IT. The second requirement is a readiness to eliminate silos between the cyber and physical parts of the security network in order to integrate and analyze data streams as a whole. Keep in mind that infrastructure is now linked to the network, which means cyber-attacks can happen anywhere. There is no longer a distinction between cyber and physical components.

Money will be the deciding factor in whether or not RTAP technology is implemented in the critical infrastructure security industry. Although early adoption will certainly come with a price, it will be worth it in the long run to prevent the devastating effects of a security breach that compromises a vital facility or resource like a water supply, hydrocarbon pipeline, or refinery.

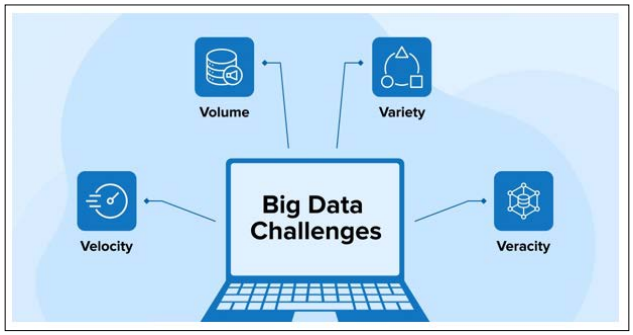

Figure 4: Big Data Challenges

The quantity of data produced has increased at an exponential rate due to the digitization of our world. Social media, sensors, and financial systems are just a few examples of the many potential sources of this type of data. Volume, velocity, variety, and veracity, or the "4 Vs" of big data, prove that this "big data" presents its own unique challenges.

The sheer volume of data generated can be too much for conventional data storage and processing infrastructures to handle. Data generation and processing speeds are measured in terms of velocity. Structured, semi-structured, and unstructured data are all examples of the many forms that data can take. When we talk about data's veracity, we're talking about its accuracy, consistency, and lack of prejudice or error. Big data presents both potential for innovation and progress and obstacles that make it hard to draw useful conclusions. Distributed computing, machine learning, and data visualization are some of the specialized tools and techniques needed for big data management and analysis. To keep up with the ever-shifting digital landscape and take advantage of big data's increasing pace, diversity, volume, and accuracy, enterprises need to constantly refine and update their strategies.

Volume

The sheer quantity of data produced is known as "volume," and it is one of the four key obstacles to big data. Data is being generated at an unprecedented rate by enterprises due to the proliferation of digital technology. Transactional systems, social media, and sensors are among the many places this data comes from. Such massive data sets may be too much for traditional data storage and processing systems, calling for fresh approaches to data administration and analysis. The sheer amount of big data presents a number of significant obstacles, including the following:

Storage

The massive amount of data generated necessitates large-scale storage systems that are both affordable and scalable. When dealing with massive amounts of data, standard relational databases may not be the ideal option due to their high maintenance costs and inability to scale. Cloud computing, distributed file systems, and object storage are some of the modern storage technologies that enterprises need to manage their large data loads.

Processing

The more data there is, the harder it is to process. Conventional batch processing systems are infamously inefficient and slow, and big data adds new dimensions to these problems. Companies often use distributed computing platforms like Hadoop, Spark, and Flink to quickly handle big data sets by distributing the processing load across several nodes.

Cost

Due to the need for specialized infrastructure and tools, managing massive amounts of data can be quite costly. Companies that generate enormous volumes of data on a daily basis may find that the expenses associated with big data processing and storage add up quickly. Businesses can get around this problem by taking stock of their processing and storage needs and then putting in place solutions that meet those needs while staying under budget.

Data Quality

The reliability and precision of the data may also be affected by the amount of data produced. It can be challenging to guarantee that data is comprehensive, accurate, and current with the amount of data being generated. This might cause data analysis biases or inaccuracies, which in turn can cause incomplete or erroneous findings. Organizations should put data cleansing and validation procedures in place to detect and fix data mistakes to guarantee data quality.

Data Governance

Governance issues can arise when dealing with massive amounts of data, including the need to manage rights and access, comply with data regulations, and guarantee data privacy and security. The appropriate and ethical management and use of data depends on organizations having robust data governance policies and processes.

Distributed computing systems enable data to be handled in parallel over numerous nodes, which is a common solution for organizations dealing with large amounts of big data. You may store and retrieve massive volumes of data with ease and scalability using cloud-based storage solutions. Data aggregation and compression methods can also lessen the storage demands of massive datasets.

In order for enterprises to gain valuable insights from their data, it is crucial to manage the large amounts of big data. Better consumer insights, data-driven decision-making, and innovation/growth-driving capabilities are all outcomes of efficient data management and processing.

Velocity
The speed of data creation and processing, or velocity, is another challenge with big data. The enormous data generation rates produced by the broad usage of real-time technology and the Internet of Things (IoT) have made real-time processing and analysis more difficult. Here are a few of the challenges that come with massive data and speed:

Real-Time Processing
Data processing in real-time is essential due to the ever-increasing volume and velocity of data. For conventional data processing systems, this means handling massive datasets in what seems like real-time. The following are obstacles

Data Ingestion
Efficient data ingestion techniques capable of handling massive volumes of data are necessary due to the rapidity with which data is being generated. Data intake pipelines that are robust and flexible enough to process data from many sources and formats are so essential.

Data Integration
Big data can be in many different formats and originate from many different sources. It could be difficult to integrate this data without first standardizing it and translating it to a common format.

Data Quality
High data velocity increases the likelihood of data errors and inconsistencies. The reliability of insights drawn from data could be compromised if data quality is an issue.

Data Storage
Data is moving at a rapid pace, necessitating storage systems that can process massive volumes quickly. The rapid rate of data growth may be too much for legacy data storage systems to handle.

Security
The ever-increasing data velocity highlights the need for robust security solutions to protect sensitive data. Encryption, access control, and secure data transport are all part of this suite of security measures.

Organizations frequently employ technology like complex event processing (CEP) systems and stream processing to handle the rapidity of large data. Organizations can make better, more timely decisions with the help of real-time data processing made possible by these platforms. With the help of caching technologies and in-memory databases, data access and processing speeds can be even improved.

Rapid analysis of big data is essential for companies that must respond quickly to changing market conditions and make quick decisions. Data handled in real-time allows organizations to quickly respond to new opportunities and threats, identify mistakes and outliers, and uncover trends and patterns.

Variety
One other difficulty with big data is the variety of data kinds and formats that are produced. Unstructured data includes things like text, images, and videos, whereas structured data includes things like database records and semi-structured data like XML and JSON.

All three types of data are considered big data. Because of this variety of data, conventional methods of data management are inadequate for storing, processing, and analyzing it. Additional difficulties that businesses encounter when trying to manage the diversity of big data include the following:

Lack of Standardization
Given the diversity of data formats, it may be challenging to ensure data consistency and standardization across sources. Because of this, issues with data integration and analysis could arise.

Data Silos
Data silos occur when several parts of an organization consistently use distinct formats for generating and storing data. Consequently, gathering all the required data for analysis and decision-making could prove to be quite a challenge.

Data Quality
Unstructured data, including social media posts and customer evaluations, makes analysis results more prone to biases and errors.

Skill Gaps
Specific skills are required for organizing and evaluating diverse data sets; these include data modeling, data integration, and natural language processing, among others. A lot of businesses are having trouble finding employees that have the skills necessary to deal with the variety of big data.

Security and Privacy
Protecting sensitive information and following privacy standards are two things that organizations need to keep in mind as they combine and analyze data from various sources. This becomes more complicated when dealing with potentially sensitive unstructured data, like social media posts.

Technologies like Hadoop Distributed File Systems (HDFS) and NoSQL databases are commonly used by enterprises to handle the diversity of big data. When it comes to storing and retrieving data, NoSQL databases offer a schemaless approach that is both flexible and efficient, while HDFS is great for handling massive amounts of unstructured data. It is also possible to combine data kinds from different sources for analysis by using data virtualization and data integration techniques.

In order to acquire a comprehensive understanding of their operations, customers, and markets, businesses must master the art of managing big data's diversity. Organisations can have a better understanding of their data and make better decisions when they combine and analyze various kinds of data.

Veracity
The veracity of the data, which encompasses its quality and dependability, is yet another obstacle to big data. It includes things like making sure the data is accurate, thorough, consistent, and credible. Data might be noisy, incomplete, or contain mistakes, which can lead to erroneous analysis and decision-making, making veracity a critical difficulty to deal with. Some of the challenges in veracity include:

- Data Quality: Verifying the quality of the data is crucial to ensure that the data is accurate, consistent, and complete. This requires a thorough understanding of the data sources, the data collection methods, and data management processes.

Data Integration
Data may have various structures, formats, and standards, making it difficult to combine data from several sources. The data may become inaccurate, duplicated, or inconsistent as a result of this.

Data Cleaning
Preparing and cleaning the data is necessary to improve its quality. As part of this procedure, we will eliminate duplicate data, rectify errors, and fill in missing values. Nevertheless, a thorough familiarity with the material is necessary for this procedure, which can be lengthy.

Data Privacy
Protecting sensitive information from prying eyes and making sure everything is up to code can seem like an insurmountable task. A number of security measures, including encryption, anonymization, and access limits, are required to accomplish this.

Bias and Interpretation
Using biased data can result in drawing the wrong conclusions and making poor decisions. A thorough familiarity with the data sources and analysis methods is necessary to guarantee the objectivity and neutrality of the data analysis.

Data Governance
If you care about the reliability, correctness, and safety of your data, you must establish data governance policies and processes. Establishing duties and responsibilities for data governance, creating data management procedures, and defining data standards are all part of this.

Dealing with honesty is crucial if you want to make use of big data for insights and value. Organizations can improve their decision-making and outcomes by implementing suitable data governance rules, integrating data from various sources, and ensuring data quality.

The Bottom Line
Any business that needs to store, handle, or analyze massive volumes of data will find big data to be a formidable obstacle. Distributed computing systems and cloud-based storage solutions are examples of novel processing and storage options that enterprises can use to tackle these problems.

Also, they need to figure out how much processing and storage space they need, and then find cheap and efficient ways to get it. To get the most out of big data, you need to make sure it's secure, well-governed, and of good quality. With efficient management of big data volume, organizations may gain a better knowledge of their consumers, make decisions based on data, and drive innovation and growth.

Elevondata Your Helping Hand
Looking to leverage the power of big data for your business? Our company offers comprehensive Big Data Implementation Services to help you unlock the full potential of your data. Collaborating with you, our team of seasoned experts will assess your company's individual requirements and provide a tailored solution to fit those demands.

Whether you need help with data integration, data warehousing, data analysis, or any other aspect of big data implementation, we have the expertise and experience to help. With our proven track record of success, we can help you achieve your business goals and drive growth and profitability.

Methodology
Our experiments and literature review informed our recommendation of a simpler Five-Step structure for Big Data systems to achieve high performance in data transfer (Figure 5). Since we're interested in the data in motion, we'll be employing symmetric encryptions like AES to transport much of it.

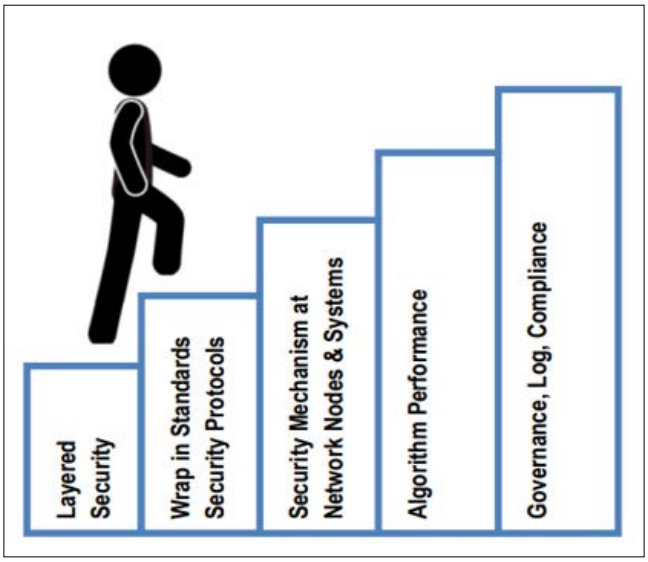

Figure 5: Data migration in Big Data Systems: A Simplified Five-Step Framework for Maximum Performance.

A. Layered Security Security and privacy mechanisms are provided at many stages by Big Data implementations like Hadoop, in addition to the network security layers already mentioned. To further ensure the delivery of secret data after encryption, these levels can be fine-tuned. The data that needs to be encrypted will be identified and protected by the layered structure. Key management servers (KMS) and similar products save all keys, including those for temporal symmetric AES, on file. A Big Data framework also places a premium on the integrity of the network, servers, and clients, as well as the enforcement of security regulations and their application at different points throughout the network. Layered security architectures can only be effective if security-related procedures like policy enforcement, security awareness training, and policy creation are also implemented.

B. Wrap Moving Data in Standard Security Protocols Data flow, control flow, message passing, secure file transfer (like FTP), and query processing are all made possible by Hadoop's several layers. An essential part of protecting big data communications is enclosing these data in a tried-and-true security protocol. For instance, Apache Knox's REST APIs for Hadoop will make it possible to transfer encrypted data over the web. To ensure the security of data exchanged using AES algorithms, SSL and HTTPS offer a secure key exchange mechanism. These algorithms generate a temporary session key, which is then destroyed after the communication is complete. The enforcement of perimeter security can be aided by existing standards like Kerberos and Java's Simple Authentication & Security Layer (SASL), which offer authentication and improve authorization procedures. Even though AES provides unbreakable security, current commercial solutions still rely on RC4 and 3DES encryption techniques, which are no longer sufficient for safe communication.

C. Security Mechanism at Network Nodes & Systems Intel Advanced Encryption Standard New Instructions (AES-NI) and its accompanying hardware framework offer support for Big Data Hadoop File System (HDFS), among numerous other accelerators and solutions that contribute to Big Data security. While the encryption process is taking place between memory and the HDFS file system, data in transit performance can be gained. Intel AES-NI is a set of seven instructions designed to speed up the most computationally demanding and complicated parts of the AES algorithms. Routers, nodes, and access points can implement such solutions by utilizing protocols like Network Security Services (NSS) that enable SSL/TLS or comparable standard protocols for machine-based encryption and decryption.
Another option is to use the graphics card's resources in conjunction with a gKrypt Engine; this will significantly outperform running the AES in a regular processor in terms of processing speed.

D. Algorithm Performance Many different versions and enhancements to the AES algorithm have been proposed, including the following: Dual Key, Single Lookup Table (LUT), AES Hybrid Approach, AES Lightweight models, and software solutions tailored for multi-core systems. Using the Single lookup AES, the GF lookup table's processing steps are reduced. To calculate the new state, this approach uses only 16 Load operations and 16 XOR operations, and it uses approximately one hundredth of the cycles that GF (28) uses. As mentioned in the section under "AES performance," it is crucial to choose an appropriate implementation and vendor that is compatible with the chosen platform.

E. Control, Record, Lawfulness Installing a data protection solution, keeping an eye on emerging threats, enabling frequent system security updates, and monitoring network traffic for intrusion and flows are all crucial. Systems like Hadoop still necessitate a newly developed intrusion prevention system (IPS) that resembles conventional data flow.

Conclusion

This architecture ensures efficient and scalable real-time data processing for big data applications. By leveraging modern tools and techniques, it enhances performance, minimizes latency, and ensures data security and compliance. This setup is suitable for various industries, including finance, healthcare, and IoT, providing a robust framework for handling data in motion. The protection of data while it is in transit is of utmost significance in Big Data applications. It has been determined, after reviewing the existing literature, that there is no global standard or formula that can be utilized to ensure the safety of the data that is being sent. In this article, we made an effort to present a simplified framework and formula for ensuring the safety of data that is in motion within the context of Big Data, with a particular emphasis on the AES structure. The authors' affiliation with NIST's Big Data Working Group (NBD-WG) ensures that the subject-specific global standard will integrate more research and framework portions. By mid-2022, work is expected to commence on the next iteration of the NIST Big Data interoperability Framework (NBDIF), which may incorporate similar features.

References

1. Abdul Rahman S, Tout H, Ould-Slimane H, Mourad A, Talhi C, et al. (2020) A survey on federated learning: The journey from centralized to distributed on-site learning and beyond. IEEE Internet of Things Journal 8: 5476-5497.
2. Aho B, Duffield R (2020) Beyond surveillance capitalism: Privacy, regulation and big data in Europe and China. Economy and Society 49: 187-212.
3. Alharbi A, Zamzami H, Samkri E (2020) Survey on homomorphic encryption and address of new trend. International Journal of Advanced Computer Science and Applications 11.
4. Almutairi NM (2020) Privacy Preserving Third Party Data Mining Using Cryptography. The University of Liverpool https://cgi.csc.liv.ac.uk/~frans/CurrentResearch/Thessi/almutairiNawal_Feb2020.pdf.
5. Beltrán ETM, Pérez MQ, Sánchez PMS, Bernal SL, Bovet G, et al. (2023) Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. IEEE Communications Surveys & Tutorials 25: 2983-3013.
6. Sukhpreet Singh, Ashwani Kumar (2024) A Review Paper on Big Data and Hadoop. All content following this page was uploaded by Sukhpreet Singh 1.
7. Omar AR, Ishak S, Jusoh MA (2020) The impact of COVID-19 Movement Control Order on SMEs' businesses and survival strategies. Malays. J. Soc. Space 16: 139-150.
8. Akanbi A, Masinde MA (2020) Distributed stream processing middleware framework for real-time analysis of heterogeneous data on big data platform: Case of environmental monitoring. Sensors 20: 3166.
9. Chan TC, Leow MC, Ong LYA (2022) Practical usability study framework using the SUS and the affinity diagram: A case study on the online roadshow website. Pertanika J. Sci. Technol 30: 1439-1455.
10. Al-Sai ZA, Husin MH, Syed Mohamad SM, Abdin RM, Damer N, et al. (2022) Explore big data analytics applications and opportunities: A Review. Big Data Cogn. Comput 6: 157.
11. González Serrano L, Talón Ballestero P, Muñoz Romero S, Soguero Ruiz C, Rojo Álvarez JL (2020) A big data approach to customer relationship management strategy in hospitality using multiple correspondence domain description. Appl. Sci 11: 256.
12. Minatogawa VL, Franco MM, Rampasso IS, Anholon R, Quadros R, et al. (2019) Operationalizing business model innovation through Big Data Analytics for Sustainable Organizations. Sustainability 12: 277.
13. Mathrani S, Lai X (2021) Big data analytic framework for organizational leverage. Appl. Sci 11: 2340.
14. Chung SS, Aring D (2018) Integrated real-time Big Data Stream Sentiment Analysis Service. J. Data Anal. Inf. Process 6: 46-66.
15. Tsai CW, Lai CF, Chao HC, Vasilakos AV (2015) Big data analytics: A survey. J. Big Data 2: 21.