PakTech Today | Pakistan's Premier Technology Magazine
Vol. 01, Issue 13 • ISSN: 3136-2729

Cybersecurity Distributed Systems / Big Data Analytics / Enterprise Security /

Security Challenges in Distributed Big Data Systems

Everyday, distributed big data systems take trillions of records and events. These systems store the most sensitive records of government, bank and hospitals. Everyday those systems are attacked by threat actors Can we avoid this? Security of big data system is most critical challenge of modern era .

By **Hafiz Zubair Akram** | Cybersecurity & Distributed Systems Research | Published: July 2026

Focus Areas: *Distributed Security, Zero Trust, Hadoop/Spark Security, Encryption, Privacy Compliance, Insider Threats, Pakistan Digital Infrastructure*

Introduction : The Expanding Attack Surface of Distributed Data

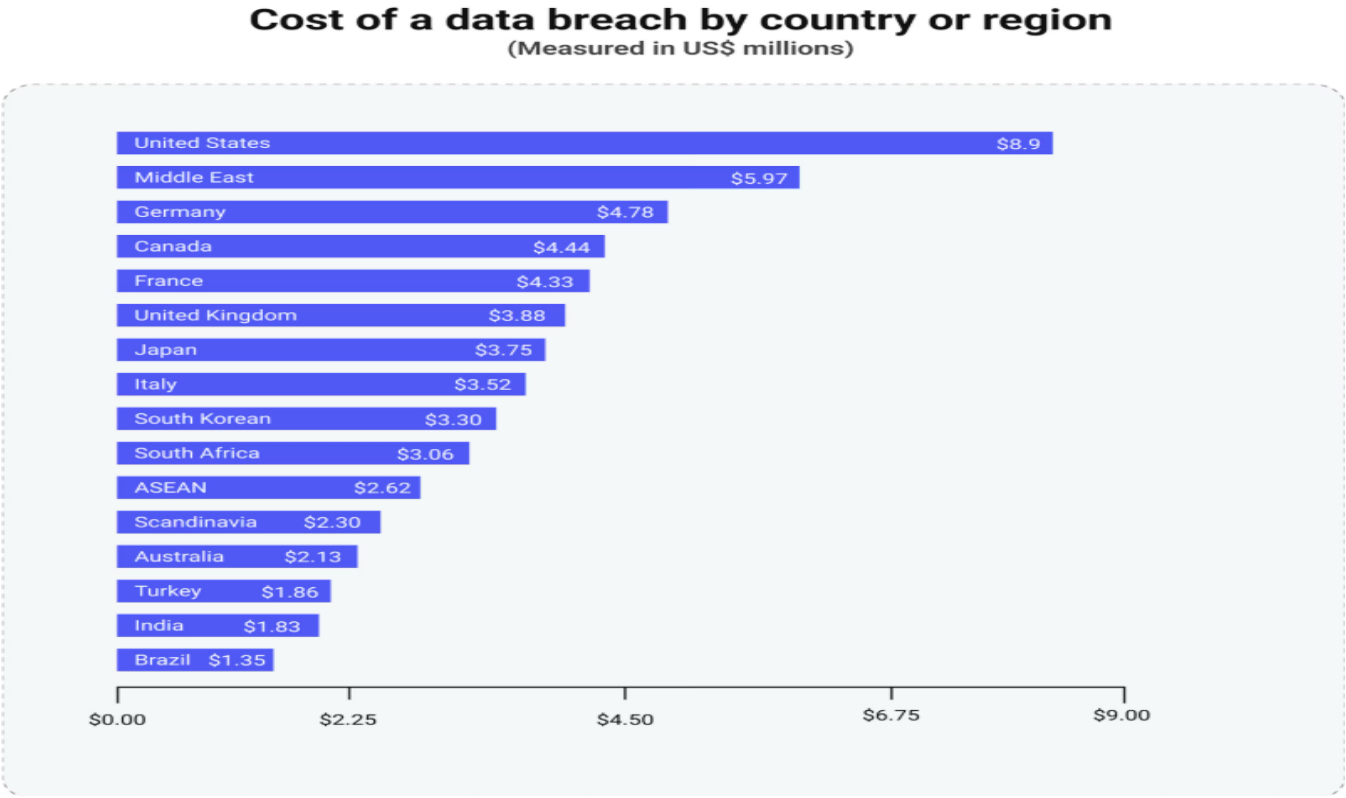
There is a harsh truth about modern big data systems good things that make a distributed system helpful are its scale and computational power of clusters, its decentralisation, and different data sources coming from different sources, like we know structured data, semi structured data, and unstructured data, which are extremely useful in some cases but challenges come with its security it is very difficult to secure them from threat actors. A single database of a company is easy to protect because it has a single access point and a smaller attack surface; we can use a firewall and other security measures. A distributed system with many nodes and many access points has a larger attack surface. A Kafka streaming pipeline takes data from thousands of different sources, and mostly the data is unfiltered and could be malicious.

The figures tell the stakes clearly. According to IBM's Cost of a Data Breach Report, the global average cost of a data leak reached USD 4.88 million in 2024. Approximately 46% of data breaches are due to cloud hosted services or distributed systems that provide services. If distributed systems help us it also create distributed vulnerabilities in massive amount. For every additional computer or node in a distributed system, create an additional attack surface for the hacker or threat actor, and due to the increase in complexity of a distributed structure, it becomes very difficult for the security team to handle the security of the complex systems.

For Pakistan's economy, where state organisations, financial services firms, communication companies, and government platforms are building and investing in big data systems, the security measures are very necessary. The companies that are building the big data infrastructure that will drive Pakistan's digital economy for the next decade, and most importantly, the measures that they will implement will determine the future of Pakistan, whether this infrastructure becomes an asset or a risk.

This article provides every aspect of the security challenge technically inherent in a distributed big data architecture. In this article we will explore the threat and categories of vulnerabilities that are present in the system and in real-world deployments. We will also explore the defensive techniques, architecture, and cryptographic approaches that we can use in the system to mitigate the risk.

> ***"The average cost of a data breach in a distributed system is USD 4.88 million in 2024 which is the highest in seventeen years. Forty-six percent of all breaches come from hybrid distributed systems. It is the central critical security challenge of the current data age."***



The Distributed Architecture Problem: Why Scale Creates Vulnerability?

Some where in Okara a wheat farmer got message on his mobile that his fields need water. Down in Rahim yar Khan, drones spray the cotton fields, a job nobody wants to do in by hands in heat. In Sawabi someone takes picture of damaged maize leaf and gets pest ID without going anywhere.

It is important to understand why distributed big data systems present so many vulnerabilities and security challenges. For understanding this system, it is important to know the meaning of the “distributed” term, which actually means this at the architecture level, and why these security measures are different from traditional security measures of a centralised system without clustering.

The Multi-Node Trust Problem

In a traditional database which operates behind a single server. Authentication happen more securely than distributed system. Data resides only in one location. In a distributed system data is split across hundred of nodes. MapReduce jobs are executed on multiple nodes each has relationship with NameNode for coordination. Apache Kafka distribute message across multiple systems. The communication between the nodes can be intercepted which is very critical and is not secure and data could be intercepted. Each node could be compromise and attack surface increase for the hacker.

The distributed system specially Hadoop Architecture works in its default configuration which is vey insecure and comes with no authentication between nodes by default. This is flaw in design of Hadoop early development, because developers were not focused on security. They only focus on the simplicity and easiness. Now the Hadoop Architecture has moved to enterprise level but this is significant issue of security. This default open source architecture has become significant liability. Many authentication mechanism like Kerberos

authentication framework has been implemented in the ecosystem, issue is that misconfiguration in this framework is very common which lead to data breaches.

Data in Motion vs. Data at Rest

Distributed system has two categories of data to protect, one data is at rest and other data is in motion each comes with its own requirement. Data at rest stored across HDFS blocks, unencrypted data at rest is readable, although application layer is secure data at rest is not secure. Key Management Server is required but that is itself is target for hackers.

Data in motion travel across the network between different nodes, clusters and external system. It should be protected from interception and man in the middle attack. Transport Layer Security (TLS) is the basic requirement for the communication between nodes. Large cluster is very complex and implementing TLS on this is vey complex. The operational complexity of this system is major hinder in implementing the encryption , that is the reason why many production clusters operate without proper security measures which make system exploitable.

The Data Store Challenge

Moder big data architecture are rarely built around single technology. There are different system for each task, like HDFS for data storage, Apache Kafka for streaming , Apache Hive or Spark SQL for analytics and for storge we also use Amazon S3 or Azure Blob Storage. Each of these has its own security mechanism, its own authentication and encryption techniques. For achieving consistent security policy across system we need unified policy management layer. There are many tools like Apache Ranger and Apache Sentry were designed for this purpose.

The Threat Landscape: Who Attacks Distributed Data and How?

Understanding the threat landscape is necessary for distributed Hadoop cluster. The threat actor are not single, a nation state threat actor targeting a telecommunication company's network may have different objective than other threat actor deploying ransomware against an unprotected Hadoop cluster.

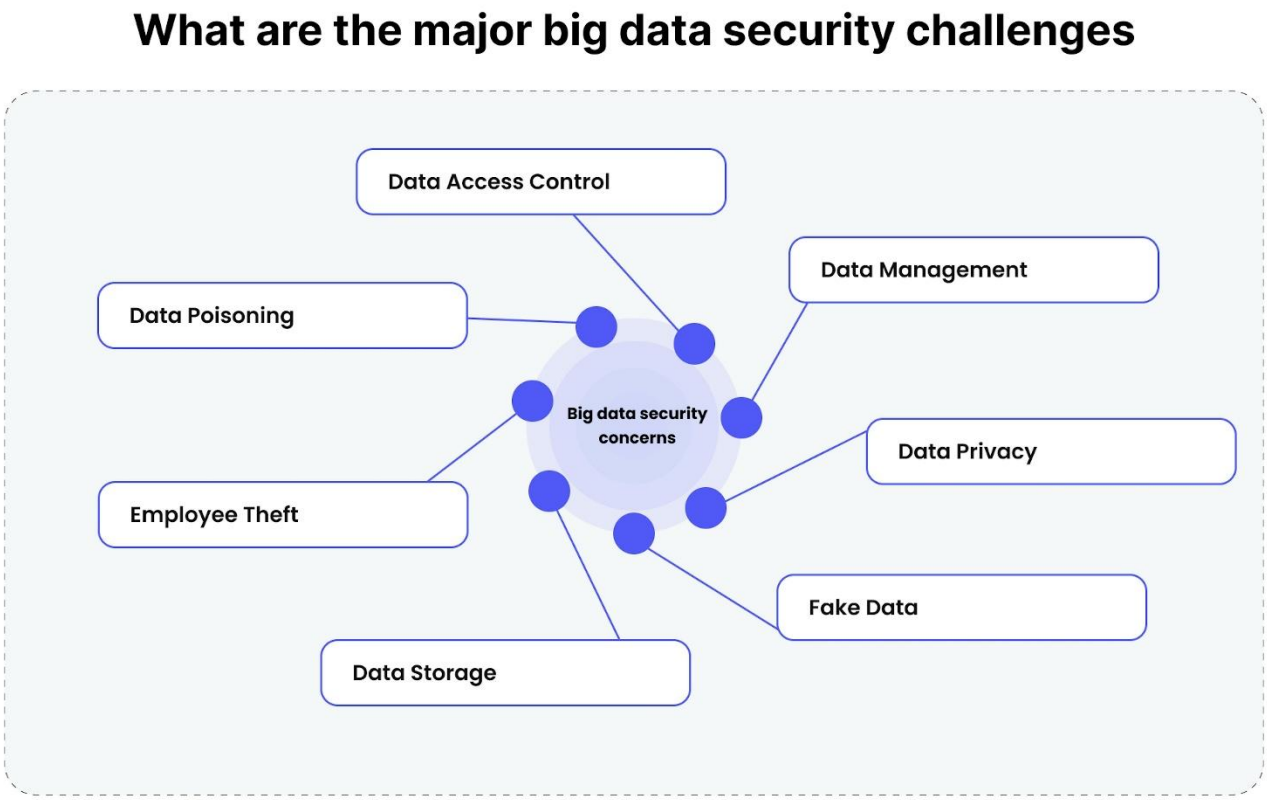
Insider Threats: The Most Difficult to Detect

Insider threats include malicious employees. They mishandle the data or may leave the sensitive information, this accounts for approximately 17% of data leak. They are difficult to detect because they have legitimate access to the internal system and data. Their behaviour could not be easily detected rather than a sophisticated malicious actor. They are more dangerous than other actors because they have privilege to the system, For example a data engineer with full HDFS access and a DBA with full permission and privilege is dangerous and can be misused. A single compromised account with full HDFS access can leak petabytes of data in very short time like in minutes.

Ransomware Against Data Infrastructure

Ransomware attacks against big data system are increasing rapidly in both numbers and impact. Ransomware actors typically criminal work for money or financial gai. They often target these systems because big data systems
dramatically hold large amount of data and this data typically has high business value, and those business are ready to pay ransom money because data is very important for them.

Attacking big data system is typically complex and time taking rather than traditional file server, because petabyte of data is computationally intensive and time consuming, So mostly ransomware actors attack the infrastructure of the Big data system like HDFS NameNode, the Kafka controller rather than encrypting the data. When metadata of NameNode is compromised without backup then whole cluster fail and could not communicate even though data remains intact. This type of attack is very common (metadata ransom)



Core Vulnerability Categories: A Technical Taxonomy

Understanding the threat landscape and vulnerabilities present in the system is very important because these vulnerabilities create exploitable weakness in distributed big data system.

Authentication and Authorisation Gaps

Authentication weakness failure is one of the most common vulnerability present in the big data system, Mostly big system failed to authenticate before granting access to user. We know the system is complex and for each node it is difficult to handle the authentication. The challenge is not that authentication is present. We have many mechanisms like Kerberos that provides authentication for both human users and service principles in cluster.

Authorization weakness is failure enforce access control once authentication is done. There are also other tools which are present for this like Apache Ranger and Apache Sentry that provide centralised policy management for big data system

Unencrypted Data Pipelines

The major problem in big data system is that data encryption is performed at rest but data transfer between nodes is unencrypted or in cleartext. This is the major attack surface for attacks like man in the middle attack which is very common to intercept the data between two systems. The other problem is that internal network communication is often trusted and TLS is disabled for inter node communication. This assumption is not correct and extremely dangerous and could give backdoor access to malicious user.

Injection Attacks Against Big Data Query Layers

SQL injection is the injection of malicious query inside the database, if validation is not performed on input parameters and input is not filtered then attacker could gain the access to the database. It is one of the most common vulnerability presents in the system today.

About the Author
This article is a research piece by Hafiz Zubair Akram Cybersecurity and Distributed System team and intended for data engineer, security architects, cloud platform engineers.

Keywords: *Distributed Big Data Security; Hadoop Security; Apache Ranger; Kerberos Authentication; Zero Trust Architecture; Data Encryption; Insider Threats; Ransomware;*

References

[1] IBM Security (2025). Cost of a Data Breach Report 2025. IBM Corporation. www.ibm.com/security/data-breach
[2] Bhathal, G.S. & Singh, A. (2019). Big Data: Hadoop Framework Vulnerabilities, Security Issues and Attacks. Array (Elsevier), 1–2. https://doi.org/10.1016/j.array.2019.100002
[3] NIST (2024). Post-Quantum Cryptography Standards. National Institute of Standards and Technology. FIPS 203, 204, 205. www.nist.gov/pqcrypto
[4] Apache Software Foundation (2024). Apache Ranger Security Framework Documentation. ranger.apache.org
[5] Gartner Research (2024). Zero Trust Architecture Adoption Survey. Gartner, Inc. www.gartner.com