iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VIII **Month of publication:** August 2020

DOI: https://doi.org/10.22214/ijraset.2020.31242

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VIII Aug 2020- Available at www.ijraset.com

Issues and Security Challenges in Bigdata: Hadoop

Vandana Malik
(Research Scholar), Dept. of Computer Science, BMU, Rohtak

Abstract: Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs. It is one of the tools designed to handle big data. Therefore, in recent years, with the popularization of the concept and application of cloud computing, it is focused on more and more by academia circles and industrial circles also. Hadoop is fundamentally meant to distribute storage & processing of Big Data sets on clusters of computer systems created on commodity hardware and it is an open source software framework. While designing the modules for Hadoop, one basic assumption is made that failures occurring in the hardware are common and framework would handle them all automatically. The biggest challenge in utilizing Hadoop to its full potential is the wisdom of knowing that where it can be used and where not. With the growing acceptance of Hadoop, there is an increasing trend to incorporate more and more enterprise security features. Therefore, the present research paper highlights the security challenges in Hadoop system.
Keywords: Hadoop, Cluster, Storage, Security Challenges, Processing Big Data.

I. INTRODUCTION

Hadoop projects treat Security as a top agenda item which in turn represents as a critical item. In present circumstances Big Data is referred to as large & complex structured or unstructured data set which the traditional data processing systems can not deal with. It's very difficult and nearly impossible for the normal data processing applications to process these large set of data which are being generated continuously by any entity around us. Big Data is one of the major areas of focus in today's digital world. As the World Wide Web grew, search engines and indexes were created to help locate relevant information amid the text-based content. It's being generated by any digital process; various social media produce it and exchange it in an exponential speed; every system, device, processor, sensor and mobile generates & transmits it.

A. Hadoop for Big Analytical Data

Hadoop is a distributed software solution. It is a scalable fault tolerant distributed system for data storage and processing. So Big Data is generated and transmitted through multiple sources and arriving in an alarming velocity. To decode its meaning, value and process it through proper channel we need a robust & powerful data processing system with ultimate analytical capabilities and skills. Processes include handling this huge data set are capture & analysis, search & process, data cure, querying, data visualization, storage, sharing & transferring, updating and the most importantly data privacy. Big Data is often related to trends, patterns & analytics of human behavior, interaction & association. Hadoop is a cost effective solution and can manage structured as well as unstructured data unlike traditional solutions such as RDBMS. The need to track and analyze consumer behavior, maintain inventory and space, target marketing offers on the basis of consumer preferences and attract and retain consumers, are some of the factors pushing the demand for Hadoop architecture solutions. Hadoop framework is written in java, it allows developers to deploy custom written programs coded in java or any other language to process the data in parallel time across hundreds and thousands of the servers. It includes various components, but two main components including a *MapReduce* set of functions and a *Hadoop Distributed File System* (HDFS) exists. It comprises a storage part known as Hadoop Distributed File System (HDFS) along with MapReduce, which is the processing part. The idea behind MapReduce is that Hadoop can first map a large data set, and then perform a reduction on that content for specific results. A reduce function can be thought of as a kind of filter for raw data. The HDFS system then acts to distribute data across a network or migrate it as necessary.

B. Why Hadoop is an essential requirement in current Scenario?

Risk management is an important part of every business and is extremely crucial when it comes to e-payments industry. The classical example would be companies like PayPal, Google and Amazon who are in a cut-throat competition to reach the consumer's pocket through online world. Thus risk analytics plays an import role. A typical flow of payment analytics consist of the transaction itself, the rules that are triggered as a result of transaction behavior and the alerts generated consequently.

Traditionally this would involves bringing the data and logs associated with the transaction to the ETL platform and finally loading it to some data warehouse. This has led to emergence of several technologies like Teradata, Oracle and Netezza. To track this various rules and modes are run that can differentiate between a normal and a fraudulent transaction. These rules are nothing but if-then conditions that generate an alert when something wrong happen. However to run these rules a lot of data needs to be analyzed like customer's present IP location, his past transactional history and many more factors. This would be a hectic and nearly impossible task when the past transactional details amounts to peta bytes of information and moreover all the processing has to be fast enough to keep the customer experience good. If the fraud models consume a lot of time then a legit customer may cancel the transaction which would impact the business and is certainly not acceptable. In such a scenario Hadoop would be quite impactful.
Hadoop can scale to analyze petabytes of information and due to its distributed nature it can significantly reduce the processing time as compared to the current system. Other use cases can be real time monitoring and analysis that can help prevent losses due to stolen financials and/or credentials.

II. HADOOP TRADITIONAL SECURITY

Originally Hadoop was developed without security in mind, no security model, no authentication of users and services and no data privacy, so anybody could submit arbitrary code to be executed. Although auditing and authorization controls (HDFS file permissions and ACLs) were used in earlier distributions, such access control was easily evaded because any user could impersonate any other user.
Because impersonation was frequent and done by most users, the security controls measures that did subsist were not very effective. Later authorization and authentication was added, but that to have some weakness in it. Because there were very few security control measures within Hadoop ecosystem.
All users and programmers had the same level of access privileges to all the data in the cluster, any job could access any of the data in the cluster, and any user could read any data set. Because MapReduce had no concept of authentication or authorization, an impish user could lower the priorities of other Hadoop jobs in order to make his job complete faster or to be executed first – or worse, he could kill the other jobs.
Even Hadoop security is not properly addressed by firewalls, once a firewall is breached; the cluster is wide-open for attack. Firewalls offer no protection for data at-rest or in-motion within the cluster. Firewalls also offer no protection from security failure which originates from within the firewall perimeter. An attacker who can enter the data centre either physically or electronically can steal the data they want, since the data is un-encrypted and there is no authentication enforced for access.

A. When to Use Hadoop (Hadoop Use Cases)

Hadoop can be used in various scenarios including some of the following:

1) Analytics Search
2) Data Retention
3) Log file processing
4) Analysis of Text, Image, Audio, & Video content
5) Recommendation systems like in E-Commerce Websites

B. When Not to Use Hadoop

There are few situations in which Hadoop is not the right fit. Following are some of them:

1) Low-latency or near real-time data access.
2) If you have a large number of small files to be processed. This is due to the way Hadoop works. Name node holds the file system metadata in memory and as the number of files increases, the amount of memory required to hold the metadata increases.
3) Multiple writes scenario or scenarios requiring arbitrary writes or writes between the files.

III. SECURITY THREATS

Just managing a complex application such as Hadoop can be challenging. Hadoop present some unique set of security issues for data centre managers and security professionals. Hadoop is also missing encryption at the storage and network levels, which is a major selling point for government agencies and others that prefer to keep their data under wraps. A few of the security issues are depicted below:

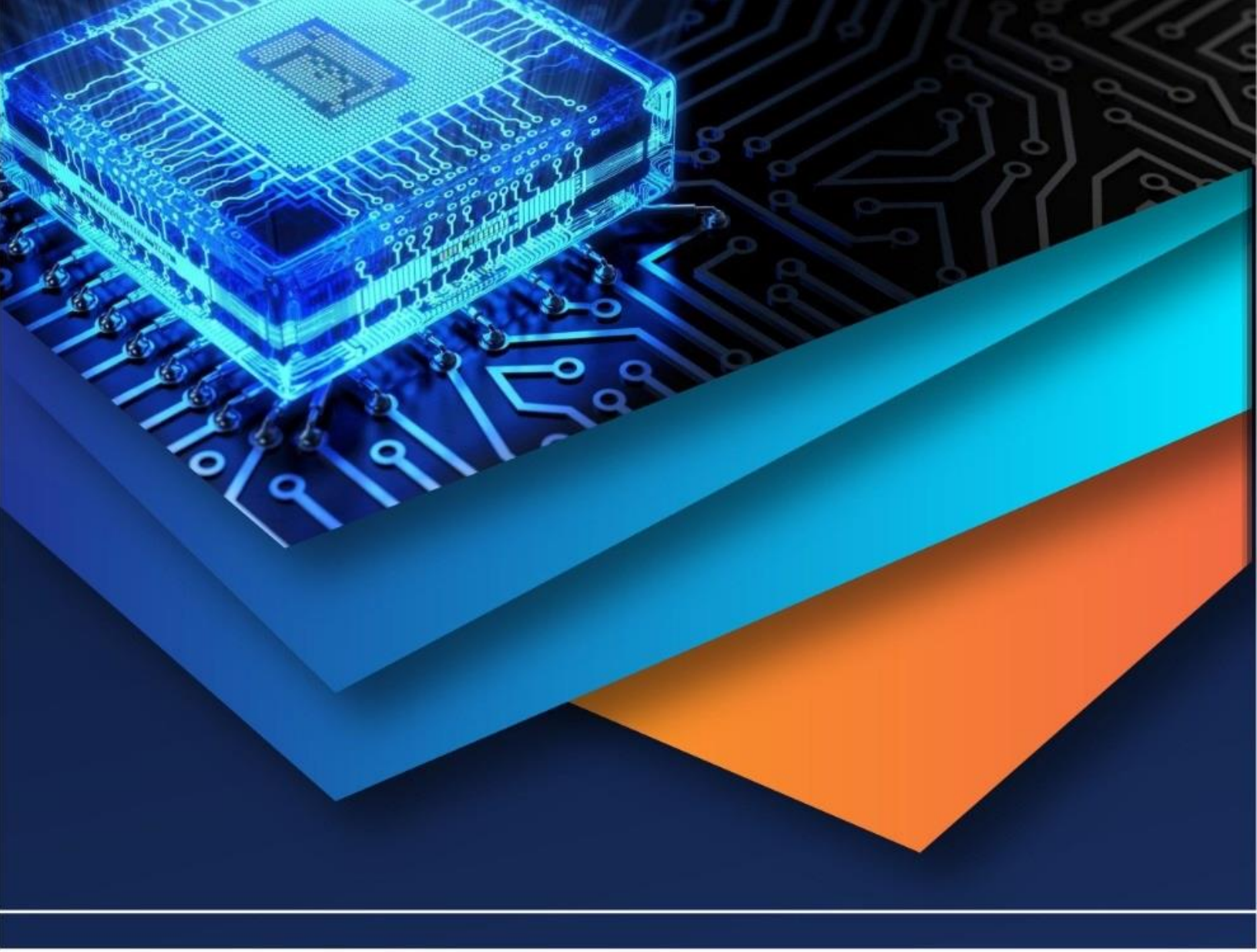
1) *Authentication:* There is no way to ensure users authentication by which one can say they are allowed to access it. An unauthorized user may access an HDFS file via the RPC or via HTTP protocols and could execute arbitrary code or carry out further attacks. An unauthorized client may read/write a data block of a file at a Data Node via the pipeline streaming Data-transfer protocol. Even an unauthorized client may gain access privileges and may submit a job to a queueor delete or change priority of the job. Further to it an unauthorized user may access intermediate data of Map job via its task trackers HTTP shuffle protocol.
2) *Controlling Data Access:* In Hadoop system there is no way to ensure users who access Hadoop can only access the data they are entitled to access, with the same policies applied consistently however they access the Hadoop system. Hence an unauthorized user may eavesdrop/sniff to data packets being sent by Data nodes to client. Commissioned data environments provision access at the schema level, devoid of finer granularity in addressing proposed users in terms of roles and access related scenarios. Many of the available database security schemas provide role based access.
3) *Distributed Computing:* Since, the availability of resources leads to virtual processing of data, at any instant or instance where it is available. This progresses to large levels of parallel computation. As a result, complicated environments are created that are at high risks of attacks than their counterparts of repositories that are centrally managed and monolithic, which enables easier security implications.
4) *Vulnerable Environment:* Speaking of security, the very makeup of Hadoop makes running it a risky proposition. The framework is written almost entirely in Java, one of the most widely used yet controversial programming languages in existence. Java has been heavily exploited by cybercriminals and as a result, implicated in numerous security breaches.
5) *Small Data Concerns:* There are a few big data platforms in the market that aren't fit for small data functions. Hadoop is one such platform wherein only large business that generates big data can utilize its functions. It cannot efficiently perform in small data environments. While big data is not exclusively made for big businesses, not all big data platforms are suited for small data needs. Unfortunately, Hadoop happens to be one of them. Due to its high capacity design, the Hadoop Distributed File System, lacks the ability to efficiently support the random reading of small files. As a result, it is not recommended for organizations with small quantities of data.
6) *Fragmented Data:* Big Data clusters contain data that portray the quality of fluidity, allowing multiple copies moving to-and-fro various nodes ensuring redundancy and resiliency. The data is available for fragmentation and can be shared across multiple servers. As a result, more complexity is added as a result of the fragmentation which poses a security issue due to the absence of a security model.
7) *Risky Functioning:* Java is one of the most widely used programming languages. It has also been connected to various controversies because cyber criminals can easily exploit the frameworks that are built on Java. Hadoop is one such framework that is built entirely on Java. Therefore, the platform is vulnerable and can cause unforeseen damages. Every platform used in the digital world comes with its own set of advantages and disadvantages. These platforms serve a purpose that it vital to the company. Hence, it is necessary to check if the pros outweigh the cons. If they do, then utilize the pros and take preventive measures to guard yourself against the cons. To know more about Hadoop and pursue a career in it, enroll for a big data Hadoop certification. You can also gain better with big data Hadoop training online courses.
8) *Client Interaction:* Communication of client takes place with resource manager, data nodes. However, there is a catch. Even though efficient communication is facilitated by this model, it makes cumbersome to shield nodes from clients and vice-versa and also name servers from nodes. Clients that have been compromised tend to propagate malicious data or links to either service.
9) *Virtually no Security:* Big data stacks were designed with little or no security in mind. Prevailing big data installations are built on the web services model, with few or no facilities for preventing common web threats making it highly susceptible.
10) *There's a Widely Acknowledged Talent Gap:* Hadoop does not have easy-to-use, full-feature tools for data management, data cleansing, governance and metadata. Especially lacking are tools for data quality and standardization. It can be difficult to find entry-level programmers who have sufficient Java skills to be productive with MapReduce. That's one reason distribution providers are racing to put relational (SQL) technology on top of Hadoop. It is much easier to find programmers with SQL skills than MapReduce skills. And, Hadoop administration seems part art and part science, requiring low-level knowledge of operating systems, hardware and Hadoop kernel settings.

IV. CONCLUSION AND FUTURE PROSPECT

There is a lot of innovation around security in Hadoop today. But another challenge centers around the fragmented data security issues, though new tools and technologies are surfacing. The Kerberos authentication protocol is a great step toward making Hadoop environments secure. Therefore, there is a bunch of focus on making all these security frameworks work together and to make them simple to manage. In future prospects Hadoop is a secure system and offers key features for securely processing enterprise data. But the security work never ends. Several agencies are working on numerous projects to enhance Hadoop security from the inside, shore up defenses from the outside with Apache Knox and to keep up with evolving requirements by providing more flexible authentication and authorization and by improving data protection. We are also working to improve integration with enterprise Identity Management and security systems.

REFERENCES

[1] Vivek. P, John Leo. A. An Analysis of Big Data Analytics Techniques. IJEMR. October-2015,Volume-5, Issue-5, 2013.
[2] Kashyap G.H., Ahmed, Afzal, Hoque N,. Big Data Analytics in Bioinformatics: A Machine Learning Perspective. Hadoop for Data Science, Vol. 13, NO. 9. Sept.2014.
[3] L. Dai, Yan Guo, J. Xiao & K. Zhang. Hadoop for big data manipulation. Biology Direct, Dec. 2012.
[4] Welcome to apache Hadoop. hadoop.apache.org. retrieved 2016-11-27.
[5] Hadoop: Solution for Big data challenges in Bioinformatics and it's prospective in India International Conference On Recent Advances In Computer Science, Engineering And Technology, March 2013.
[6] http://www.marketresearchstore.com/report/hadoop-market-z59712.retrieved 2016/11/28
[7] E. Khan. Addressing Hadoop using Natural Language Processing. Modern Computer Applications in Science and Education. ISBN: 978-960, June 2014.
[8] Divya Kumari, Ravi Kumar. Impact of Hadoop in Bioinformatics, I.J.C.A, Sept., Vol.101, No.11, 2014.
[9] M. Herland, T.M Farooqui and R. Wald. A review of data mining using Hadoop. International Journal of Computer Science, June 2014.
[10] Hadoop release. Apache.org Apache software foundation. Retrieved 2016-11-27.
[11] Ronald C Taylor. An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. Taylor BMC Bioinformatics 2010, 11(Suppl 12).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY