Research Article

Performance Exploration of Hadoop in Fully-Distributed and Pseu-do-Distributed Method

Sushrut Naik [a], Shreyas Sail [b] Dr.Sanjeev Sannakki [c]

[a] Dept of Computer Science, Kls Gogte Insititue Of Technology, Belgaum,India
[b,c]Dept of Information Science, Kls Gogte Insititue Of Technology, Belgaum,India
[a]sushrutnaik1998@gmail.com, [b] shreyas.ajit.sail@gmail.com, [c] sannakkisanjeev@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: This electronic We propose to analyze the exhibition of Big Data applications with Hadoop and along these lines to make sense of the presentation in data security. Ha-doop is formed on MapReduce, one of the generally uti-lized programming models in Big Data. This paper aims to evaluate performance factors and compare the con-ventional method of providing security and distributed method of providing data security. This assessment will set up the fundamental abilities that ought to be consid-ered explicitly on a Distributed File System (HDFS: Ha-doop Distributed File System) for execution as far as speed. Large information has explicit attributes (Volume, Velocity, Value, Variety, Veracity - 5V) that make it hard to oversee from a security perspective. Consequently, we propose a model that gives enormous information secu-rity using Hadoop and Pig-Latin and analyze the afteref-fects of the traditional method and distributed method. Through this work, it was expected that Hadoop and Pig-Latin are skilled in Data Security

Keywords: Apache Hadoop, Big Data, Pig-Latin, Conventional Secu-rity, Fully-Distributed Security, HDFS (Hadoop Distributed File System), Data Security

1. Introduction

This paper tends to significant issues identified with security and insurance notwithstanding pointers that are related to the utilization of huge information examination in the medicinal services area. The Privacy and Security Workgroup (PSWG) of the Health Information Technology Policy Committee (HITPC) is by and by examining and giving significant clues to the National Coordinator for Health Information Technology on the U.S. Division of Health and Human Services (HHS) on inconveniences identified with protection and wellbeing to the electronic exchange of wellness insights. The execution on huge information in medicinal services impacts one of the PSWG's essential qualities, primarily, those patients' needs and desires must be mulled over, and that "patients have to never again be puzzled about or hurt with the guide of assortments, use or accord of their wellbeing subtleties."

With the growing quantity of daily information, it’s hard to exercise and analyze data on a single machine and additionally, it's tough to offer protection to huge information on a sole machine as a consequence there's a want of Multiple Node HDFS system for storing and exercising the information in a dispensed manner with security being the primary concern consideration. Once shifted to HDFS System pig scripting proves to be an enhanced tool to analyze data for colossal volumes in a secured method. Thus massive data may be without problems processed with high-end machines with the usage of the Hadoop distributed file system in a very systematic and protective manner.

With superior big data processing technologies, insights can be received to enable higher decision making for essential development regions which include fitness care, monetary productivity, energy, and natural catastrophe prediction. Big data refers to colossal quantities of digital data agencies and the government accumulate about us and our surroundings massive data are generated from a variety of customers and devices and are to be stored and processed ineffective information centers. As such, there's a sturdy call for building an unimpeded network infrastructure to acquire geologically dispensed and rapidly generated data, and flow them to data facilities for effective insights discovery. It’s just standard data that are usually allotted across more than one location, from a divergent array of sources, in various formats, and mostly unstructured. The issues encompass evaluation, search, curation, capture, storage, sharing, sources, visualization, and privacy violations.

Numerous associations as of now utilize Big Data for promoting and showcasing and research, yet might not have the nuts and bolts right – explicitly from a security position. Similarly, as with every single innovation, insurance seems, by all accounts, to be a bit of hindsight, best case scenario.

Large Data penetrates will be humungous as well, with the capacity for even extra basic reputational harm and licit repercussions than at present.

A creating assortment of organizations is practicing this innovation to hold and investigate petabytes of information which incorporate weblogs, clickstream subtleties, and online life substance to abuse upgraded bits of knowledge about their customers and their business endeavor.

As a final product, data classification transforms into even extra significant and information proprietorship should be routed to encourage any sensible class.

Most organizations as of now war with forcing these measures, making this an incredible test. We should select owners for the yields of Big Data procedures, notwithstanding the crude information. In this way, data ownership will be marvelous from data ownership – perhaps with IT claiming the crude information and business gadgets assuming liability for the yields.

Not very many organizations are likely to build a Big Data condition in-house, so cloud and Big Data could be inseparably associated. The same number of associations are perceptive, putting away information inside the cloud does now not dispose of their commitment to shielding it - from both an administrative and modern point of view.

Procedures including trademark-based complete encryption might be basic to secure delicate information and apply consent controls (being properties of the information itself, instead of nature in which it is put away). A significant number of those standards are unfamiliar to enterprises today.

2. Literature Survey

Sl No.	*Author*	*Outcome*
1	Elisa Bertino	This paper talks about examination difficulties and bearings concerning data confidentiality, protection, and dependability in the setting of big data. Key exploration issues talked about in the paper incorporate how to reconcile security with protection, the thought of data ownership, and how to implement get to control in big data stores
2	Pryasto M, Alamsyah	This paper discusses NIST Risk Assessment framework described in NIST SP800-30 [4] can be use for big data. The methodology in obtaining the data for risk assessment is still the same, although we may have to deal with larger data

3	Sung-Hwan Kim	This paper focused on two things. property importance in huge data is a key segment for removing information. In this perspective, we focused on the most ideal approach to ensure significant data by making sure about noteworthy information inside. Additionally, it is hard to guarantee every enormous datum and its qualities. We consider tremendous data as a singular article which has its attributes. We acknowledge that a trademark which has a higher hugeness is a higher need than various characteristic
4	Rao, S	This paper focused on two things. characteristic relevance in huge data is a key part of removing information. In this perspective, we focused on the most capable strategy to ensure significant data by guaranteeing significant information inside. Additionally, it is hard to guarantee every colossal datum and its qualities. We consider colossal data as a lone article which has its properties. We expect that a property which has a higher essentialness is a higher need than various properties.
5	Matturdi, B.	In his paper, his focal point is the troubles of security and security in Big Data. By then, we present some likely techniques and strategies to ensure Big Data security and assurance
6	Parmar, Raj R.Et al	This paper proposed for encryption of records very still in Hadoop proposed Encryption Zone plot which is utilized to Encrypt and unscramble the measurements straightforwardly yet it has trouble that MapReduce commitments execution debases. Along these lines, the extra productive methodology of encryption Kuber structure has been executed utilizing a form of Salsa20 known as ChaCha20.

3. System Design

- **System Perspective**

Our framework is a trade for the current Bigdata security framework for the accompanying reasons.

1. We can load any kind of bigdata as we are having hadoop’s file system which is flexible enough to store any kind of data.

2. As we are working in a distributed or clustered environment we can perform the major operations like encryption and decryption parallel and independently.

3. Parallelism helps in achieving better performance

4. Our system is time efficient and space efficient due to distributed storage and distributed processing.

Existing System

With regards to existing framework we follow the old traditional technique for encoding information so security breaks don't happen. Cryptography is a regularly utilized procedure [1] to ensure information against illicit perusing and undetected mutilation of information. Approximately, two sorts of cryptosystems exist: symmetric and asymmetric (or open key [2, 3]) cryptosystems. Encryption and decoding of information is done a lot quicker with a symmetric cryptosystem than with an unbalanced cryptosystem. In this way, generally, a symmetric cryptosystem is utilized to scramble or unscramble mass information in a framework. No doubt symmetric encryption is faster than asymmetric encryption but when it comes to encrypting huge amount of data this method lacks in performance factor, as the data becomes huge the time needed to encrypt the data will be more. The existing system can be viewed in the below Figure 1.1

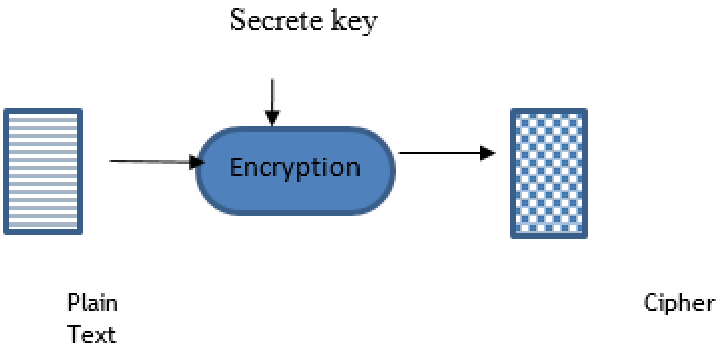
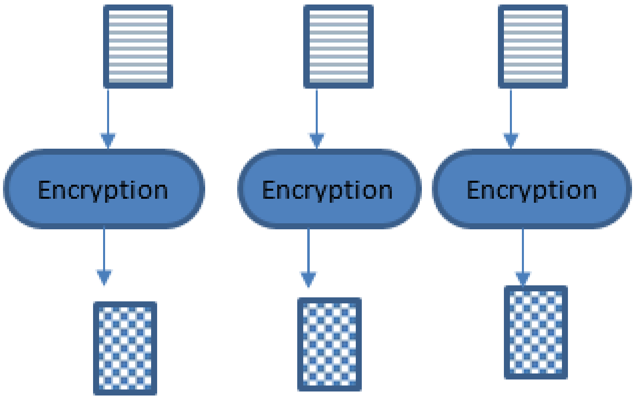


Figure 1.1

Proposed System

As compared to existing system we propose a new way of encrypting the bulk data / big data in a distributed manner. Yes the idea behind proposed system is to take the big data divide it into parts and later apply the encryption to the parts of the big data independently, the advantage of such method is time efficient and faster as compared to the existing system, as can be seen from the figure 1.2 that the data is broken into parts and individual parts are given as input to encryption algorithms independently. The process of encrypting the small blocks can be done on different nodes independently



Functional Requirements

Functional requirements are capabilities that a framework must show to take care of an issue. Functional requirements of Big data Security System**.**

Loading the BigData: This is one of the major functionality of the system where we are trying to load a very huge volume data into the HDFS[Hadoop Distributed File System]. We load the health data using Hadoop command as follows

Hadoop fs –put data.txt /Security

The above commands tell that we are loading a huge file "data.txt' into Hadoop's file system into a directory by name 'Security'.

There are other ways of loading data into the Hadoop file system, in pig Latin, we load the data as follows

HData = load 'data.txt' using PigStorage(',') as (Schema:datatypes);

In Hive, we load the data as follows

Load data local in path 'data.txt' into the table Health;

Slicing of data: As we are dealing with Bigdata and our system is supposed to encrypt the huge data, but encrypting such huge data is time-consuming. Hence the solution for this is to divide the huge data into blocks, the block size is 64 MB in Hadoop generation 1 and its size is 128 MB in generation 2. By dividing the huge data into blocks becomes time efficient in encryption as well as decryption.

Encryption: After the data being sliced or divided the data is scattered across the nodes that form the cluster, hence encryption is done on the partial data at block level to form blocks of ciphertext. We make use of AES encryption to encrypt the partial data and the encryption is done at the individual systems. This leads to take the advantage of parallel processing, in our application parallel processing is parallel encryption. This kind of encrypting of distributed data across the cluster is time-efficient.

Decryption: It's the reverse process of Encryption, where it takes blocks of cipher data decrypts it using the decryption mechanism and then decrypts the cipher to plain text. As Encryption is done in a distributed manner, the same process is applied for decryption, hence this is also time-efficient.

Non-Functional Requirements

Non-Functional requirements express the needs in terms of performance, design constraints, standards compliance, reliability, availability, security, maintainability, and portability.

Performance

The system should be designed with a small number of subsystems. Because of a lesser number of subsystems, lesser the communication between subsystems and better the performance. As we are going to test the system in two modes, one is called single machine pseudo-distributed mode and the other is fully distributed mode. Experiments specify that a fully distributed model of encrypting of data is time-efficient, and performance is better.

Reliability

The system will not fail at any given time and if it fails how fast the system can recover from failure is high reliability.

Availability

The system should deliver useful services at any given time. i,e, the system should be available to the customer 24*7 hours.

Security How far the system can resist or persist itself from external attacks or accidents. Security is required for all critical systems.

Maintainability

Bigdata which is huge in volume in the System shall be easily maintained in HDFS.

Portability

The system is implemented using JAVA hence should be easily transferable from one computer to another, along with the storage space if necessary.

Robustness

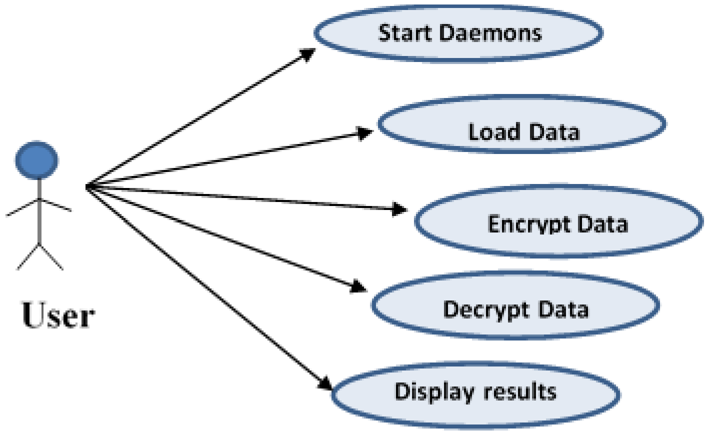
The Developed system is robust and fault-tolerant because of the block redundant nature of the data that is available as part of HDFS, hence the system can operate despite any node fails or abnormalities in input, calculations, etc.

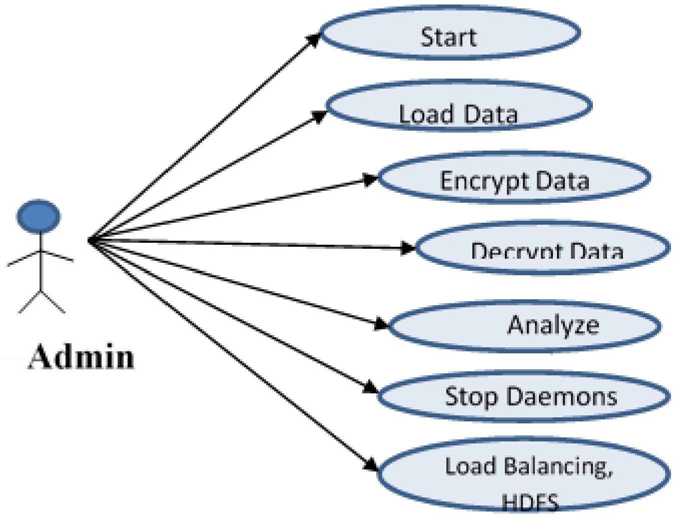
Testability

The Developed framework is tried with test objectives, test strategies utilized, and test assets. The outcome is put away with legitimate qualities so it very well may be alluded for future testing conditions.

4. Detailed Design

Use Case Diagram For User



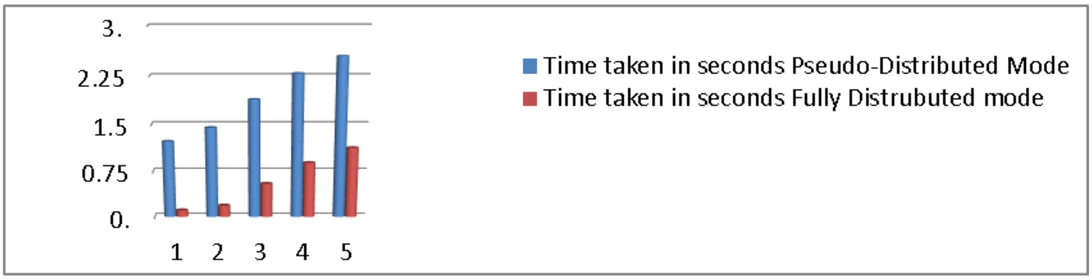


Outcomes

Experiment no	Amount of data in MB's	Time taken in seconds	
		Pseudo-Distributed Mode	Fully Distributed mode
1	5	1.23	0.11
2	10	1.45	0.19
3	15	1.89	0.55
4	20	2.29	0.89

5	25	2.55	1.13

Graph



5. Conclusion

With the expanding measure of day by day information, it's hard to process and break down information on a solitary framework and it's hard to give security to enormous information on a solitary framework in this way there's a need of Multiple Node HDFS framework for putting away and preparing the information in a dispersed way with security thought about. Once moved to HDFS System pig scripting ends up being a superior apparatus to break down information for tremendous volumes in a made sure about way. Hence colossal information can be effectively handled with top of the line frameworks utilizing Hadoop disseminated record framework in an exceptionally productive and secure way. The representations and results show that scrambling an enormous measure of information on a pseudo-disseminated mode will devour additional time when contrasted with the bunched method of encryption. The inquiry instruments make the investigation a lot simpler by giving arbitrary access to Big Data.

6. Future Scope

As of now, our framework is performing dispersed encryption on the information of volume in approximately hardly any MB's, as a major aspect of future improvements we would need to take a shot at the information of volume as far as TB and PB. The test outcomes that we have been on a bunch of 4 to 5 machines. In the future, we would need to chip away at a bunch of 10 to 20 machines with the goal that we can get persuading results.

Likewise, we need to upgrade the handling some portion of our framework, at present we are utilizing Hadoop's MapReduce for preparing. Hence as a feature of an upgrade, we need to utilize apache's top-level task 'Flash' for handling i.e encryption and decoding.

As far as handling SPARK is multiple times quicker than Hadoop's MapReduce

References

1. Simmons, G.J., Contemporary Cryptology, The Science of Infor-mation Integrity, IEEE Press, New York, 1992.
2. Diffie, W., and Hellman, M.E., New directions in cryptography, in: IEEE Transactions on Information Theory, IT 22(6), November 1976, pp. 644-650.
3. Rivest, R., Shamir, A., and Adleman, L., A method for obtaining digital signatures and public-key cryptosystems, in: Communications of the ACM, Vol.21, 1978, pp.120-126.
4. D. Ritchey, "Big data, big security," Security, Vol. 49, no. 7, pp. 28-30 2012.
5. C. Tankard, "Big data security," Network Security, Vol. 2012, no. 7, pp. 5-8 2012.
6. Bertino, E. paper Big Data - Security and Privacy, ISBN: 978-1-4673-7277-0, Accession Number: 15411749, Publisher: IEEE, New York, 2015
7. Paryasto, M, Alamsyah, A, Paper Big-data security management issues, Accession Number: 14649758, DOI: 10.1109/ICoICT.2014.6914040 Publisher: IEEE, 2014
8. Sung-Hwan Kim, Paper Attribute Relationship Evaluation Meth-odology for Big Data security, Page(s): 1 – 4 INSPEC Accession Number: 14047549, DOI: 10.1109/ICITCS.2013.6717808, Publish-er: IEEE, 2013
9. Rao, S., paper Security Solutions for Big Data Analytics in Healthcare, Page(s): 510 - 514
10. Print ISBN: 978-1-4799-1733-4 INSPEC Accession Number: 15557107, Conference DOI: 10.1109/ICACCE.2015.83 Publisher: IEEE, 2015
11. Matturdi, B, Paper Big Data security and privacy: A review, (Volume:11 , Issue: 14) Page(s): 135 – 145 ISSN : 1673-5447 IN-SPEC Accession Number: 15057605 DOI: 10.1109/CC.2014.7085614, Publisher: IEEE, 2014

12. Parmar, R., Roy, S., Bhattacharaya, D., Bandyopadhyay, S., & Kim, T. H. (2017). “Large Scale Encryption in Hadoop Environment: Challenges and Solutions.” IEEE Access.
13. K. Abouelmehdi, A. Beni-Hessane and H. Khaloufi, "Big healthcare data: preserving security and privacy", Journal of Big Data, vol. 5, no. 1, 2018. Available: 10.1186/s40537-017-0110-7
14. S. D M, D. G.K and M. Y.M, "A Technique For Big Statistics Securi-ty Based on Hadoop Distributed File System", SSRN Electronic Jour-nal, 2019. Available: 10.2139/ssrn.3508526.
15. "Privacy and Security in Big Data Management", International Journal of Recent Trends in Engineering and Research, pp. 251-256, 2018. Available: 10.23883/ijrter.conf.20171201.051.rsxao.