Survey on BigData, MapReduce and Blockchain Technologies

[1]Dr Sneha K, [2]Vijaya Durga K

[1]Professor, [2]Student
[1, 2] M.Tech – CSE,
[1, 2]BNM Institute of Technology, Karnataka, India.

Abstract: Blockchain is a growing improvement for decentralized and function-based statistics distribution over a big gadget of suspicious participants. These dissertation talks are about generation of information of the blockchain generation in big data and how it is different from the currently used centralized transactions structures. Moreover there are talks about how blockchain novelty could be applied as a part of several industry forums within the retail location to profit the customers and the outlets. The paper also describes the advantages of blockchain in big data techniques used in different projects.

***IndexTerms* - Mapreduce, Cloud Computing, Cluster, Blockchain, Big Data.**

I. Introduction

The importance of Blockchain generation has been rising, for the reason, that concept was coined in 2008. A blockchain is a decentralized ledger of all dealings in a network. In the blockchain era, members within the community can verify transactions without they want to a relied on a 0.33 birthday party intermediary. Influential programs include finance transfers, balloting, cloud computing, and a lot of other makes use of. Fig -1 shows how blockchain blocks are generated. A blockchain is defined as an open area consisting of all Bitcoin dealings that have been prepared until the modern deal or the closing deal. As completed blocks are enclosed to it as and when the transactions are whole, the blockchain is turning larger and larger. These blocks are approaching the blockchain following a chronological order, in a linear manner. The computer systems which are part of the Bitcoin community are referred to as nodes. All of those nodes get hold of a duplicate of the blockchain, this taking vicinity automatically when a customer joins the Bitcoin community. There are lots of statistics blanketed within the blockchain, for an instance the addresses and their balances from the beginning till the latest finished block.

II. Big Data:

Blockchain involves Local Tax Big Data case in its operations and applications. The case can be advanced on different criteria like veracity, interoperability, security and so on. Local Tax for data heavy apps is an advanced baggage that uses blockchain framework. Construction of blockchain involves intense security that assures integrity, confidentiality, accessibility and also interoperability of information which are demonstrated on the cases of supply chain. There are many techniques which could be used to apply for the improvement of veracity of records by using token-primarily based value crowd sourcing, identification of supply and token primarily based trade and achievement on value basis. One Data Policy of Indonesia could have benefited by the utilization of Blockchain [1].

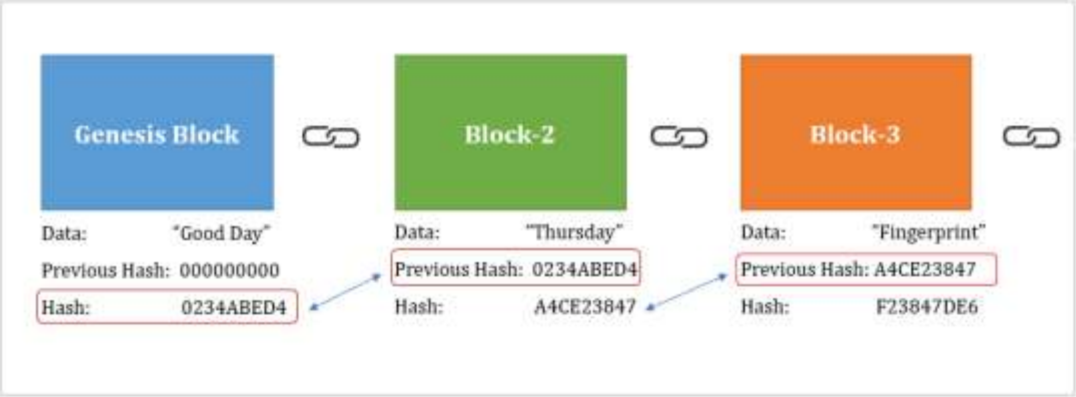

Figure 1 Blockchain

Production environment involves obligations allocation which has to be managed in a wide range. This led to the need of Internet of Responsibilities (IoR). The technique applies IoR system that is pushed using more number of records and generation of blockchain. The realistic developments of the technique proposed showed a great growth in rating of responsibility and cognizance of protection for each and every company and employee working in company [2].

Blockchain Decentralized trust control is proposed to solve the issues of IoT massive Statistics control. Relaxed Utilization manipulate scheme related to IoT massive statistics which involves processing of information operation, management inclusive of data amassing, garage, invoking and utilization over smart agreement of blockchain. Every procession for statistics and usage system is confirmed by maintaining security using cryptography signed and Merkle Tree based for all transactions including each and every block with great degree of protection in a wide range having disbursed ledger having the characteristic of tamper resistance in P2Ps (Peer to Peer). For statistical usage and intake, ease usage manage is proposed for digital rights management and token-based facts consumption method of excessive fee facts from being violated or unfold with no issue. In order to encourage IoT supplies of purchaser the information that is in excessive amount that is pleasant, blockchain based tokens are designed where tokens are awarded for proper supply contribution of records which are excessively pleasant. blockchain-primarily based decentralized accept as true with management platform of huge facts on the basis of common, consented blockchain, huge quantity assessment manifested as proposed scheme feasible, cozy and scalable for decentralized agree with control of IoT massive information[3].

Kratos is supplied which is gadget for statistics responsibility, auditability and transparency. The want for concrete control of records is proposed in order to answer it and the way Kratos would shape in along with a technical device for the interface between different disconnected statistics system across more than one educational stake holder. An idea is established upon a case having a look at school's device, Ex: Cambridge Public School., the challenges faced by them. The developing issues are defined facing facilities with regards to high school facts privacy, statistics use and the want for facts literacy and education for students, worries which direct consciousness on usability and carried out data gaining knowledge. Importance of Education Data Mining (EDM), Learning Analytics (LA) for a large number of stakeholders and a machine provided for allowing information share and exchange in transparent and immutable way. The objective is to run academic institutions and Schools to know the roadblocks to make sure about dealer compliance according to legal guideline and policies. Investigation is a must process to check how Kratos permits EDM, LA, scholar statistics studying and the business enterprises in greater shielding jurisdictions. Ex: Inside the context pertaining to Europe and its common Data Protection Regulation. The Kratos prototype is piloted in partnering schools each within the US and the European Union for benefiting remarks and preparing enhancements. The plan is to carry out studies by giving up customers including college students, parents and teachers for evaluating the system adoption in order to optimize functionality, architecture. The aim could be preserved to make Kratos as the open supply gadget to impose clean records governance to each player in community where the commitment remains to provide college students' corporation the better know-how of the position related to massive statistics of Big Data in the lives of school and beyond [4].

III. Map-reduce Technology:

The work in the paper supplied a better map-reduce technique, Figure-2. It is primarily based on K- approach clustering for massive facts analytics to make complete usage of K-method cluster algorithm, map-reducing approaches. K-method cluster approach with a set of rules has a lot of benefits and is more efficient. Every record's factor associated with the pivotal importance is assigned for each midpoint of the cluster randomly and choose the maximum likely corresponding statistics factors. Map reducing strategies are used to arrange a very clean computation procedure. K – Means clustering algorithm additionally improves the adaptability while keeping veracity of the result. These algorithms and techniques handle a massive amount of information and run effectively [5].

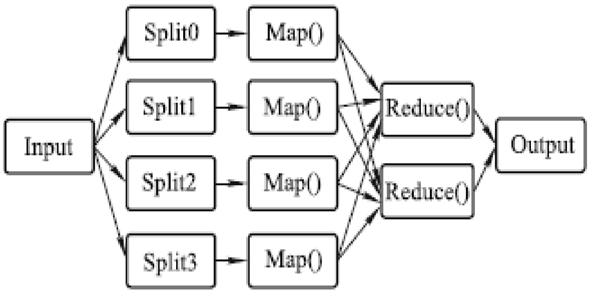

Figure 2 Map Reduce

A dynamic data auditing scheme is proposed for large data storage. This scheme supports dynamic auditing on Hadoop Distributed File System (HDFS). Based on Rivest-Shamir-Adleman (RSA) verifiable tag and DBIT, the scheme supports dynamic operations on HDFS. The scheme can face up to various attacks. Furthermore, a construction of the scheme is gifted primarily based on the Map Reduce framework shown fig in 2. As a part of future work, work could be increased to explore extra effective information auditing schemes. Our work could be extended to batch auditing for a couple of customers. Although RSA-based production is efficient without problems combined with the Map Reduce framework, it might cost a heavy overload on storage. To clear up this problem, bilinear mapping operation might be used; however, that is still a hard assignment. Exploring such a difficulty to reap efficiency and low storage consumption as a solution in future [6].

Syntactic search tool within cloud provides way for getting information required, desired over specific words usage which can be furnished by the person in line with his search. It additionally gives statistics or the consequences of the keywords entered on the premise of preceding searches making the word "Big Data" mild thus enabling making of the algorithms that help within the search phase of the records just like the hunt. Semantic Search tool of Cloud provides method to get the precise facts with some similar alternative results at once with no heaps pertaining to internet page for getting the facts from. Such search could be finished through those big records algorithms and makes the searching and collection of facts enormously smooth. This allows data for filtering to help the clients [7].

The study involves work proposing risk worries and hazards mitigation techniques in numerous superior computing environments like massive data, the webpage factors and cloud computing. Every superior measurable model requires large variable information procedures at high record costs agreeable, prevalent environments, so danger reduction required in gadget exceptional development. Cloud computing risks are also dealt while the subsequent segment describes approximately net of factors. Exploring threat control in huge facts surroundings is also dealt. The automation tools for risks primarily based on checking out might be utilized in future in all-pervasive environments to discover faults within the structures [8].

IV. BLOCKCHAIN AUTHENTICATION:

The paper involves design of the Implementation of Distributed File Storage (IPFS) based garage model for blockchain to mitigate the downside of garage and transaction get right of entry to each block in the block-chain grid. The IPFS repository prototype result affords efficient repository area due to content material labeled hash of the transactions. Moreover, the content addressed scheme is implemented to get entry to the transactions of blocking the usage of the hash fee supplied with the aid of the IPFS allotted garage. The prototype brought up here involves saving hash figures pertaining to transaction in preference to storing the complete transaction or original transaction to make certain the efficient storage scheme for the blockchain network. The proposed structure of repository prototype could be made use along with a lot of varieties of transaction files which include video-audio in blockchain. Prevailing repository models such as bitcoin, ethereum, hyper ledger suffer from the repository of cumbersome statistics within the disbursed ledger of the blockchain network. For this reason, the prototype could be operated at present storage version to reduce each block dimension. [9].

Authentication and authorization processes may be taken into consideration as an essential direction in many allotted systems, in particular blockchain mechanism.. As consistent with these tactics, it can be figuring out whether or not someone's registration is authenticated for permitting the person to get admission to some resources in blockchain mechanism. The proposed AABchain machine guarantees about registered person, a right authenticated person can be authorized for accessing the essential aid based totally considering hash cost pertaining to block as authentication, authorization. Possible upcoming projects the suggested mechanism on blockchain community that gets hold on numerous guidelines along with; The AABchain system implementing in simulation surroundings so may be applied or execution on the digital or actual environment; Real servers/nodes might be delivered for network of blockchain; Adjusting AABchain tool for a big community consisting from massive numbers of nodes/servers to make the gadget greater reliable. [10].

V. MD5 HASHING:

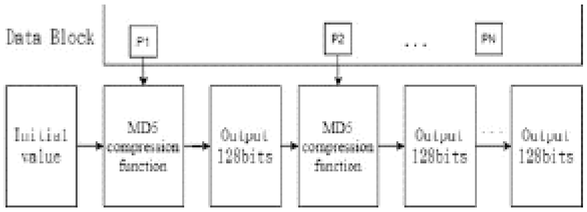

Figure 3 MD5 Hashing

The paper involves proposal of a brand new blockchain machine with first-rate-grained get entry to manipulate for IoT applications in which the membership and people's traits such as information owner, user, miner, maybe up to date securely in anti-tamper blockchains with the aid of integrating Chameleon Hash (CH) functionality among brand latest multilayer chain structure. This system conveyed direct statistics outflow as a result of revoked individuals and maintained exact compatibility with principal consensus protocols, go-chain protocols, encryption algorithms. The system can be in addition adjusted using some Trusted Authorities (TAs) over distributed way for assuring TA's great safety degree. The gadget scrunity, prototype shows the fact that proposed device could outperform current answers among overhead and complexity phrases. By utilizing the proposed machine, relaxed, possible, and decentralized statistics get admission to manage which enables huge-scale and high-precision records control maybe effortlessly installed for IoT web. Get entry to control could at initial stage stop cancel customers/miners of having access to no longer only the destiny records but the past information in a blockchain, thereby notably enhancing manageability not left out I adjusting the tamper resistance of the blockchain. This permits bendy, secure blockchain IoT offerings, control to be adopted through nearby/international leagues of IoT. The evaluation supplied managerial guidelines for threat evaluation, price assessment, and power, demands for storage in such a way where suitable technical answers may be designed to satisfy precise commercial enterprise requirements [11].

The use and benefits as well as demerits with respect to three-D-Playfair, MD5 and XOR offered via paper presentation. For conquering dangers along with saving statistics to avoid customizing, modification and verify the reliability of the supply, XOR computation of three-D-play fairs key is proposed and ciphers to confirm the validity related to supply, utilize MD5 for getting hash cost to confirm the integrity of the ctrecords. Practically, efficient confirmation about the usage related to XOR and MD5 combination that could correctly confirm if the information is affected with and guarantees the righteousness of the information. [12].

VI. SURVEY DETAILS

Paper no.	Algorithm and Techniques	Advantages	Disadvantages
[1]	Blockchain Implementation Assessment Framework	Using the Blockchain framework, Local Tax Big Data may be advanced especially on records safety, interoperability, and veracity.	Blockchains aren't scalable as they counterpart centralized device.
[2]	Internet of Responsibilities Model driven by blockchain and big data	The realistic deployments show an apparent boom within the obligation rankings and protection focus in both employees and businesses.	Compatibility issue as there is no widespread tagging and monitoring with sensors.
[3]	BlockBDM (Blockchain enabled decentralized trust management)	Resolves the issues of trust in IoT systems. Record transactions done based on blockchain based smart agreement.	Transactions of duplicate records cause huge storage and time consumption
[4]	Kratos system	Need for concrete records explored, to shape in with technical gadget among multiple disconnected statistics structures across a couple of educational stakeholders.	Suitable for business enterprises, educational institutions. As the data is transparent, it is not suitable for securing personal information.
[5]	K-Means cluster algorithm, DNA(Deoxyribonucleic acid) algorithm Map reduce algorithm,	Huge garage space of dataset achieved via map reduce technique	Difficult to expect K-value for extremely large clusters
[6]	Dynamic auditing on HDFS(Hadoop Distributed File System)	Secure to resist forge attack, replay attack, replace attack on big data platform	Hadoop doesn't support small records, it has slow processing speed
[7]	Semantics search engine	Efficiently deals with finding words from large amount of data.	Big data in cloud is automatically processed before it is presented to users. Hence, not suitable for Data security.
[8]	Mitigation strategies	Specializes identification of dangers, threats in advanced computing environments. Aimed at system high quality assurance and renovation	Consumes more time for massive computations.
[9]	IPFS based blockchain storage model	Reduces the size of block transactions suitable for data heavy applications.	IPFS consumes lot of bandwidth which is not always favored by metered net customers. This leads to huge data consumption in case of duplicate files.
[10]	Cryptographic Hash Function	Hash tables give quick lookup of transaction entries and hence saves time.	Hash collisions are unavoidable among a subset of large set of possible keys. Also, hash tables grow pretty longer for more number of collisions
[11]	Blockchain-IoT Systems	Improved security, suitable for getting more efficient supply chain	Not suitable for confidential files
[12]	3D-playfair Cipher	MD5 set of rules suitable for storing and comparing hashes. Integrity of the data is ensured.	Need for information availability that is subject to changes has to be reported to authorized users. Not suitable for storing confidential data.

VII. CONCLUSION

This paper discussed how the blockchain era could be useful for the big data region and the way it could be used for MapReduce. Blockchain framework is a combination of at ease report storage. It helps in regulations of larger files. The framework proposes measures to make certain the combination of Blockchain and Mapreduce tackles the hassle of data garage because it utilizes the off-chain storage Blockchain generation also the interoperability demanding situations in the MapReduce atmosphere. Mapreduce IT systems exchange records with the Mapreduce blockchain. Blockchain technology along with the interoperability demanding situations within the Mapreduce ecosystem, it could be proved that Mapreduce plays a crucial role within the big data system and its application. Mapreduce in bigdata gadget consists of the need to enhance efficiency in healthcare provider shipping, affected person safety, boom gets right of entry to fitness care services especially they need to reduce the expenses of scientific research and applications.

REFERENCES

[1] Satriyo Wibowo, Tesar Sandikapura, “Improving Data Security, Interoperability, and Veracity using Blockchain for One Data Governance, Case Study of Local Tax Big Data”, 2019 International Conference on ICT for Smart Society (ICISS), **Date Added to IEEE Xplore:** 27 January 2020, Bandung, Indonesia.

[2] Xuejiao Tang; Jiong Qiu; Wenbin Zhang; Ibrahim Toure; Mingli Zhang; Enza Messina; Xueping Xie; Xuebing Wang, “The Internet of Responsibilities – Connecting Human Responsibilities using Big Data and Blockchain”, 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9-12 Dec. 2019.

[3] Ma Zhaofeng; Wang Lingyun; Wang Xiaochang; Wang Zhen; Zhao Weizhe, “Blockchain-Enabled Decentralized Trust Management and Secure Usage Control of IoT Big Data”, IEEE Internet of Things Journal (Volume: 7, Issue: 5, May 2020), 18 December 2019.

[4] Velislava Hillman; Varunram Ganesh, “Kratos: A secure, authenticated and publicly verifiable system for educational data using the blockchain”, 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9-12 Dec. 2019.

[5] S. Dhanasekaran; R. Sundarrajan; B. S. Murugan; S. Kalaivani; V. Vasudevan, “Enhanced Map Reduce Techniques for Big Data Analytics based on K-Means Clustering”, 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 11-13 April 2019.

[6] Xingyue Chen, Tao Shang, Feng Zhang, Jianwei Liu & Zhenyu Guan, “Dynamic data auditing scheme for big data storage”, Frontiers of Computer Science volume 14, pages219–229(2020). https://link.springer.com/article/10.1007/s11704-018-8117-6

[7] Nripendra Narayan Das, Mohit Chowdhary, Rahul Luthra, Maisera, Saurabh Garg, “Semantic Big Data Searching In Cloud Storage”, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019. **Date Added to IEEE Xplore:** 10 October 2019

[8] Vinita Malik, Sukhdip Singh, “Cloud, Big Data & IoT: Risk Management”, 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019. **Date Added to IEEE Xplore:** 10 October 2019

[9] Randhir Kumar, Rakesh Tripathi, “Implementation of Distributed File Storage and Access Framework using IPFS and Blockchain”, 2019 Fifth International Conference on Image Information Processing (ICIIP), **Date Added to IEEE Xplore:** 10 February 2020.

[10] Wasan Ahmed Ali, Naji Mutar Sahib, Jumana Waleed, “Preservation Authentication and Authorization on Blockchain”, 2nd International Conference on Engineering Technology and their Applications 2019-IICET2019-Islamic University, Alnajaf-Iraq. Date Added to IEEE Xplore: 27 February 2020

[11] Guangsheng Yu , Xuan Zha , Xu Wang , Wei Ni, “Enabling Attribute Revocation for Fine-Grained Access Control in Blockchain-IoT Systems”, IEEE Transactions on Engineering Management (Volume: 67, Issue: 4, Nov. 2020), Page(s): 1213 – 1230.

[12] Wen-Chung Kuo; Wan-Hsuan Kao; Chun-Cheng Wang; Yu-Chih Huang, “3D-Playfair Encrypted Message Verification Technology based on MD5”, 2020 15th Asia Joint Conference on Information Security (AsiaJCIS), 20-21 Aug. 2020. **Date Added to IEEE Xplore:** 10 September 2020