Blockchain Based Solutions for Big Data

[1] R. Swathi, [2]Dr.R. Seshadri
[1]Research Scholar, [2]DProfessor cum Director
[12]CSE
[12]SV University, Tirupati, India

Abstract: Big data era now and is producing different types of data daily from different sources. And this big data is important for all kinds of industries. Storing, processing and analysing of such Big data is possible. In any case, a standout amongst the most critical issues that obstructs the consistent appropriation of Big Data is the absence of security and privacy protection of data in the Big Data tools. While currently hunting down the most ideal approach to store, organize and process Big Data, the Blockchain innovation comes in giving significant information. Its proposed solutions about decentralized administration of private information, computerized property determination, IoT changes are having noteworthy effect on how Big Data may develop. This paper introduces the novel arrangements related with a portion of the Big Data areas that can be empowered by the Blockchain innovation.

IndexTerms **- Big Data, Block chain, Hadoop, MongoDB, Spark.**

I. INTRODUCTION

Blockchain is an underlying technology of Bitcoin that soon emerged as the real discovery of Satoshi Nakamoto [1], thus making Bitcoin just the first of the many future Blockchain implementations. Technically, Blockchain is a distributed public ledger that contains all the transactions that ever executed in the system. It exists on a P2P network where every full node stores a copy of the Blockchain ledger. There is no central authority that manages the Blockchain database. This concept of obtaining a database only between the actual and equal users of the system sets the base for building so called: "decentralized trust". For the transactions to be validated and authorized, a consensus of nodes that agree upon the issue is required. The concept of decentralized trust comes as opposite solution to almost every system that we have built using the client-server architecture. By removing the central authority out of the system, there is no longer a mediator processing the actions and the data. That results with lower transactional costs, non-reversible transactions and no need for trust in the governments or private corporations. In this solution, Blockchain users don't even need to trust the other party included in the transaction. They should trust only the system and the code.

Blockchain is a database that stores all the transactions grouped in blocks. When new transaction is created, the sender broadcasts it in the P2P network to all the other nodes. The transaction is still new and not confirmed. As the nodes receive the transaction, they validate it and keep it in their transactional pools. To validate transactions means to run predefined checks about the structure and the actions in the transaction. Special node types called miners create a new block and include all, or some of the available transactions from their transaction pool. Then the block is mined, which is a process of finding the proof of work using variable data from the new block's header [2]. Finding the proof of work is continuous calculation of a cryptographic hash that fits the defined difficulty target. Mining requires a lot of processing power and the miners use a dedicated mining hardware. The miner that first finds a solution for its block is the winner. His candidate block becomes the new block in the chain. Because transactions are added in the mining block as they arrive, we can say that the latest block in the Blockchain contains the latest transactions.

When a new block is created (mined) it is time-stamped and propagated to the network. Every node receives the block, validates it, validates the transactions in it, and adds the block to his local Blockchain copy. The transactions included in the block become authorized and non-reversible part of Blockchain in the moment the block is accepted by majority of the nodes. Blocks can also be inspected as a way of transactional and financial clearing. Valid transactions are approved in groups, at certain periods of time. This solution was done to avoid conflicts and solve the double spending problem. In addition to transactions, every block stores some metadata and the hash value of the previous block. So every block has a pointer to its parent block. That is how the blocks are linked, creating a chain of blocks called Blockchain[3].

Seemingly the most critical advancement in information technology during recent years, block chain can possibly change the way that the world methodologies big data, with improved security and information quality. The benefit of blockchain is that it is decentralized – no single individual or organization controls information passage or its integrity; in any case, the holiness of the blockchain is checked continuously by each computer on the network[4]. As all focuses hold a similar data, degenerate information at point "A" can't turn out to be a piece of the chain since it won't coordinate with the equal information at focuses "B" and "C".

Utilizing the case of a hospital or medicinal services supplier, inadequately oversaw quiet information builds the hazard that a patient will be misdiagnosed, treated mistakenly, or that test outcomes wind up plainly lost or debased. There's likewise a worry that two touchpoints on a patients' treatment journey may have distinctive datasets for a similar individual. Putting social insurance databases on the blockchain would make a solitary, unchangeable asset for experts to utilize while treating a patient. The most significant advantage the blockchain could offer healthcare is security [5-8].

A blockchain-based healthcare framework would likewise enable suppliers to impart records to equity divisions, safety net providers, managers and some other part with an enthusiasm for individuals' wellbeing without the exponential increment in chance factors that accompanies extending a system thin; all things considered, a multi-department framework is just as secure as the guards at its weakest point.

II. BIG DATA

Big data emerged in the early and mid-2000s to meet web scale computation needs: Zookeeper at Yahoo, Big Table and MapReduce at Google, Cassandra at Facebook; and so on. Then came open source projects like Hadoop File System (HDFS), Hadoop MapReduce, Cassandra, and more. By the late 2000s and mid-2010s, new companies like MongoDB, Cloudera, and DataStax had made organizations to change the open source successes into big business review offerings.

1.1 Big Data Challenges

In any case, big data has its difficulties, which include control, data authenticity and adaptation.

1.To start with, who controls the infrastructure when there are numerous performing actors included? For example:

•If you're a multinational venture, how would you share information around the planet? If we have different duplicates, how would we know which one is the most up and coming? How would we accommodate an alternate framework manager part at each territorial office?

•In industry consortium, how to share control of the biological system framework among the organizations in your consortium? This is particularly hard if those organizations are contenders!

•Why can't there be information simply "out there" as a solitary shared wellspring of truth that nobody on the planet possesses or controls, in essence? Or maybe, information would be an open utility like power or the web itself.

Second, how well can you trust the data?

•If we create the information our self, how would we demonstrate we were the originator? On the off chance that we get information from others, how would we know it was genuinely them?

•What about crashes and malicious behaviour? Machines crash, glitches happen, bits flip. Zombie IoT toasters may include refuse. So after all your favour Spark calculations, is it still only junk out?

Finally, how do you monetize the data?

•How do we exchange the privileges of the information, or purchase rights from others?

•There's a long standing long for a widespread information commercial centre; how?

III. BIG DATA AND BLOCK CHAIN TECHNOLOGY

A new tool is introduced for processing big data called block chain technology.The current surge in blockchain innovation was started by Bitcoin. In fact, all blockchains are basically databases, however databases with "blue ocean" benefits: decentralized/shared control, immutability/review trails, and local resources/trades.By present day database standards, customary blockchains have awful adaptability and don't have query languages; in any case, the blue ocean benefits were sufficient to catch the creative ability of the globe.

Even better, later technology—the BigchainDB blockchain database—joins the advantages of appropriated databases (scale, queryability) and blockchains (decentralized, immutable/audit trails, resources/exchanges) [9-10].This new blockchain database innovation has the scalability required in huge information conditions, by expanding over best-in-class disseminated databases like MongoDB. This opens the potential for exceptionally fascinating applications in big data: shared control of framework, review trails on information, and the likelihood for a general information trade. How about we investigate both in detail.

3.1 Shared Control of Big Data Infrastructure

A big data blockchain database like BigchainDB is decentralized, which implies that its control can be shared [11-13].That sharing can occur in one of numerous unique circumstances:

- Across workplaces inside an undertaking. That is, you gain shared power of a big data database crosswise over geologically spread workplaces.
- Across organizations inside a biological system. That is, you gain shared power of a major information database among organizations (even contenders) in a biological system.

- On a planetary level. Shared control of an open, public big data database signifies "information as an utility" like air or the web. Such a database is getting taken off now: it's called IPDB. IPDB is both a system and a philanthropic establishment.

IV. BLOCK CHAIN TECHNOLOGY AS METHODOLOGY

Problem: If we're a multinational undertaking, how would we share information around the planet? If we have various duplicates, how would we know which one is the most up and coming? How would we accommodate an alternate framework director part at each territorial office?

•A: Each local office with its own sysadmin controls one hub of the general database. So, they control the database altogether. The decentralized nature additionally implies that if a sysadmin or two denounces all authority, or a territorial office is hacked, the information is as yet secured. (Expecting encryption is set up as well, obviously).

- Problem: If we're a multinational undertaking, how would we share information around the planet? If we have various duplicates, how would we know which one is the most up and coming? How would we accommodate an alternate framework chairman part at each territorial office?

•Problem: If we're an industry consortium, how to share control of the environment framework among the organizations in our consortium? This is particularly hard if those organizations are contenders!

•A: Like over, each organization controls one hub in the general database.

•Problem: Why can't there be information simply "out there" as a solitary shared wellspring of truth that nobody on the planet possesses or controls as such? Or maybe, information would be an open utility like power or the web itself.

•A: IPDB, the Interplanetary Database, is getting taken off at this point.

Blockchain innovation enables us to have review trails on information, to enhance the dependability of the information. You get verified information stories.

How 1. Here's how it works. Suppose that you have an information pipeline of six stages: IoT sensors → Kinesis/Event Hub + stream investigation → HDFS storage→ Spark information cleaning → Spark standardization → MongoDB stockpiling →Tableau examination.

Prior to every datum pipeline step begins, time-stamp the info information as takes after:

1.Create an exchange, formed as a JSON report, that incorporates a hash of the information, hashes of each line and segment on the off chance that you like, and any meta information you wish to incorporate (e.g. where you got the information from, exact hashing formula).

2.Cryptographically sign the exchange with your private key. This is an exemplary advanced mark.

3.Write the exchange to the blockchain database (BigchainDB). It will naturally time-stamp the exchange. Presently you have unchanging proof that you approached that information by then, which others can cryptographically check in view of your open key. After every datum pipeline step is done, time-stamp the progression's yield information in a similar three stages.

How 2. There's a significantly less difficult path for a few stages, in case you're utilizing a disseminated database that BigchainDB as of now wraps (e.g. MongoDB, RethinkDB). You at that point just swap out that database (e.g. MongoDB) with its blockchain-ified form (e.g. MongoDB wrapped by BigchainDB).

There's no requirement for hashing or anything, since it's all understood. Note that BigchainDB does not uncover the entire interface of the wrapped database, however after some time it will uncover more in view of client criticism.

Advantages:

•Problem: If we produce the information ourrself, how would we demonstrate we were the originator?

•A: People who have your public key can see that you

•Problem: If you get information from others, how would you know it was really from them?

•A: You can check the exchange against that individual's public key

•Problem: What about accidents and noxious conduct? Machines crash, glitches happen, bits flip.

•A: You can run intermittent procedures to re-hash the information put away in the pipeline. If the new hash doesn't coordinate the past hash, something's incorrectly.

•Problem: Zombie IoT toasters may enter rubbish. So after all our favour Spark figuring's, is it still only rubbish out?

•A: First, utilize IoT gadgets with legitimate security, no compelling reason to bring down the DNS once more:) Those IoT gadgets ought to have an approach to sign the information where their private key isn't traded off. At that point, as some time recently, you can check the IoT gadget's information input exchange against its public key.

We can manufacture a widespread information commercial centre, to help vanish dividers of information storehouses. An adaptable blockchain database talking the convention of IP rights exchange empowers information to be purchased and sold as a benefit. It would be by and large controlled by an open biological system. Individuals can fabricate information trades on top to suit their wants.

Here's the way it works. We require a worldwide open blockchain database, which exists as IPDB. There could even be different systems, where resources stream among them with the Interledger convention for interoperability. The benefit is the information rights, upheld by copyright law. The advantage "lives" on the blockchain database. You possess the advantage on the off chance that you control the private key. The benefit can be cut and diced, and exchanged to others, utilizing a cutting edge, adaptable, blockchain-accommodating IP convention. This is likewise a current development, called Coala IP.

Thus, the stack is BigchainDB programming + IPDB arrange + Coala IP convention. With this, we have the substrate on which imaginative programmers and business visionaries can construct information trades of different shapes and sizes.

Advantages. How about we perceive how this tends to the issues I'd depicted before.

•Problem: How would you exchange the privileges of the information, or purchase rights from others?

•A: Create an exchange to exchange rights to someone else, talking the dialect of the Coala IP convention. Sign it. Compose it to the database.

V. CONCLUSION

Blockchain presents numerous guarantees for the eventual fate of Big Data. The first is that in numerous regions, clients could be responsible for every one of their information and exchanges. They can assume that exchanges will be executed precisely as the convention summons expelling the requirement for a trusted outsider. This idea can impact Big Data to discover an answer for putting away and overseeing information in a circulated way on a P2P arrange. Blockchain innovation can be another piece of the encompassing biological system of instruments that Big Data employments. As a matter of fact it can assume a significant part in security for client confirmation, limiting access in light of a client's need, recording information get to histories and legitimate utilization of encryption on information

REFERENCES

[1] Satoshi Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System", 2008.

[2] Andreas M. Antonopoulos, "Mastering Bitcoin", O'Reilly Media, Inc, 2015

[3] Elena Karafiloski, AnastasMishev"Blockchain Solutions for Big Data Challenges " IEEE EUROCON 2017, 6–8 JULY 2017, OHRID, R. MACEDONIA

[4] Dr. Gavin Wood "Ethereum: A Secure DecentralisedGeneralised Transaction Ledger",2014.

[5] Tal Rapke, MD "Blockchain Technology & the Potential for Its Use in Healthcare",2016.

[6] Gupta, Anand Jha, and Purna Roy, "Adopting Blockchain Technology for Electronic Health Record Interoperability", 2016.

[7] Laure A., Linn Martha B., Koo, M.D, "Blockchain For Health Data and Its Potential Use in Health IT and Health Care Related Research",2016.

[8] Ariel Ekblaw, Asaph Azaria, John D. Halamka, MD, Andrew Lippman, "A Case Study for Blockchain in Healthcare: "MedRec" prototype for electronic health records and medical research data",2016.

[9] Thanh Bui and Tuomas Aura, "Application of Public Ledgers to Revocation in Distributed Access Control",2016.

[10] Ahmed Kosba, Andrew Miller, Elaine Shi, Zikai Wen, CharalamposPapamanthou, "Hawk: The Blockchain Model of Cryptography and Privacy-Preserving Smart Contracts", Security and Privacy, IEEE,2016.

[11] "Internet of Things: Privacy & Security in a Connected World", FTC Staff Report,2015.

[12] Marco Conoscenti, Antonio Vetro, Juan Carlos De Martin, "Blockchain for the Internet of Things: a Systematic Literature Review",2016.

[13] Ali Dorri, Salil S. Kanhere, and Raja Jurdak, "Blockchain in Internet of Things: Challenges and Solutions",2016.

Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).