**Week 10 Assignment**

**Blockchain**

Monroe College

CS 675: Big Data Management and Analytics

**Group 1**

Gita Gurung

Madeline (Zhiyang) Lin

Priyanka Chaudhari

Sukriya Gurung

Professor Syed Farid

November 13th, 2022

*Use the provide reading to provide a summary of what do you understand about blockchain and how can you implement this technology within the data management system.*

*Upload your paragraph using this link.*

**ANS:**

Blockchain technology depicts a bag of lego or bricks. From the bag, you can take out different bricks and put them together in different ways to create different results. The blockchain is a decentralized ledger, or list, of all transactions across a peer-to-peer network.

Blockchain in Hadoop can impose security while making transactions, enables the possible inter-cluster communication, race between miners for validating and authorizing transactions, integrating the global economy with electronic currency for any kind of trading goods, materials, money, etc.

The BlockHDFS is the proposed solution which enhances the security of HDFS in a user transparent way by storing metadata of files from HDFS in a blockchain. BlockHDFS consists of three components: an HDFS cluster including the NameNodes and DataNodes, a permissioned blockchain network such as Hyperledger Fabric, and a NodeJS Client which acts as a bridge connecting the HDFS cluster and the blockchain network. In BlockHDFS, the blockchain is responsible for storing the metadata of the files. Integrating blockchain with distributed file systems such as HDFS can potentially improve security and traceability. In the context of this paper, the original design of Hadoop is more optimized for file processing instead of security-by-design. Hence, in this paper, we proposed a new approach to introduce blockchain (and more specifically, Hyperledger) to enhance the security of the HDFS ecosystem. In the current implementation, we have only added minimal metadata to the blockchain, but with BlockHDFS, one can easily add more features suited to their application needs. For future work, BlockHDFS can be extended to work in real-time with the file system and track all data between NameNode and DataNodes of the HDFS in the secure ledger with multiple nodes.

Details of how we implement Blockchain technology into data management system are elaborated in the following documentation of the reading materials.

**Documentation**

**A New Era Needed In Distributed Computing**

This paper highlights the security concerns of Hadoop and need for the enhancements along with electronic currency and blockchain functionality. The paper also emphasizes on why and how BitCoin and BlockChain can fit in Hadoop Eco-Systems and their possible pros and cons. Especially, in validating and authorizing business transactions with some mathematical cryptographic techniques like hashcode with the help of BlockChain Miners.

Hadoop is a parallel distributed computing framework by Apache for addressing BigData issues. It was introduced to reduce the time required to query the data of various data formats.

The Hadoop ecosystem is independent software that runs on top of HDFS and MapReduce. The components include MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, HBase, and Kafka.

Up recently, Apache's Spark and Flink got included into BigData systems to support structured data processing and distributed streaming. These are in-memory processing engines whereas Hadoop uses HDFS for read and write operations. Hence, these systems give much faster results

There are some common security concerns with respect to Hadoop to be addressed in the near future as well as existing security implementations in Hadoop.

Next, the article discusses Bitcoin and Blockchain.. Bitcoin is a digital currency to facilitate the exchange of goods and services by offering a commonly accepted good. Blockchain technology depicts a bag of lego or bricks. From the bag, you can take out different bricks and put them together in different ways to create different results. The blockchain is a decentralized ledger, or list, of all transactions across a peer-to-peer network. Bitcoin uses code proof of work algorithms.

The reasons why the blockchain is needed in Hadoop distributed computing systems include inter cluster communication, data storage, speculative execution, open ledger, loT security, cryptographic authentication and avoiding delays in completing the transactions.

Hadoop Blockchain can fit in Hadoop on top of basic core, MapReduce, Flume, Kafka, and Zookeeper. It can leverage existing frameworks for both processing and securing the transactions. Blockchain in Hadoop can impose security while making transactions, enables the possible inter-cluster communication, race between miners for validating and authorizing transactions, integrating the global economy with electronic currency for any kind of trading goods, materials, money, etc. However, blockchain in Hadoop will publish some confidential data with the open ledger, revamp Hadoop basic architecture for inter cluster communications, which would result in some cost and time.

This paper analyzes the Hadoop ecosystems regarding the security concerns and the needed technology of blockchain. Currently, the Hadoop framework does not have blockchain functionality but we hope that it can be implemented in the future releases.

**Big-Crypto Big Data, Blockchain and Cryptocurrency:**

This article introduces and summarizes the interactions between cryptocurrency and Big Data in the digitized world. It presents a comprehensive investigation on these fields and provides a systematic review of recent developments. It covers all the stakeholders whether they are in the field of academic or industries. It covers all the stakeholders whether they are in the field of academic or industries, give them a better understanding, and aim to explore its future potential.

A cryptocurrency is a digital currency. It circulates without centralized authority. Big data and cryptocurrency are convergent in complementary ways. The characteristics of the cryptocurrency network are useful resources for Big Data analytics. Because of the rapid evolution and reliance on artificial intelligence, crypto trading and transfers are becoming stable day by day. Blockchain is one of the emerging technologies that is paving the way for financial and industrial services. Cryptocurrency is inseparably connected to blockchain technology. The combination of varied multidisciplinary concepts is called crypto economics which includes software engineering, cryptography and distributed computing. It lets buyers and sellers make the transactions directly (peer-to-peer), maintaining privacy, the records are irreversible, applicable and efficient for double spending. Some of the most well-known cryptocurrencies are Bitcoin, Ethereum, Litecoin, Stellar, Ripple, Z-cash and Dash. Although there are lots of flaws, developers and researchers have been enhancing this technology through a variety of approaches, and the majority focused on improving security and privacy.

Tangle is a protocol that has the main feature of directing an acyclic graph. It performs a faster machine to machine micro-payment system without no transaction fees. It uses the web tangle instead of chain structure for consensus. Hashgraph is another technology that uses gossip for achieving consensus. A transaction is performed when the participant shares its information and every node gathers all received information along with a new transaction to pass on to multiple random nodes in the network. Artificial intelligence uses advanced computer science to analyze and make sense of complex data. It is trained to react intelligently in terms of reasoning and problem solving.

This paper directly focuses on the convergence of the above-mentioned concepts. With the rapid growth of interest and attention of an emerging, tremendous, and valuable market, it brings the target of cybercriminals. Researchers have been investigating limiting those drawbacks. Tangle and Hashgraph technologies can improve efficiency and reduce costs.

**Blockchain-integrated Hadoop distributed file system for secure provenance traceability**

Hadoop Distributed File System (HDFS) is one of the widely used distributed file systems in big data analysis for frameworks such as Hadoop. HDFS allows one to manage large volumes of data using low-cost commodity hardware. However, vulnerabilities in HDFS can be exploited for nefarious activities. This reinforces the importance of ensuring robust security to facilitate file sharing in Hadoop as well as having a trusted mechanism to check the authenticity of shared files. This paper focuses on improving the security of HDFS using a blockchain-enabled approach (BlockHDFS). Specifically, the proposed BlockHDFS uses the enterprise-level Hyperledger Fabric platform to capitalize on files' metadata for building trusted data security and traceability in HDFS.

The new approach called BlockHDFS incorporates blockchain functions into the HDFS and stores them inside Hyperledger as an asset. Hyperledger Fabric is a permissioned blockchain with a distributed ledger that supports smart contracts. Using Hyperledger Fabric also requires less computation power to run, as everything resides on a private business network suited for organizations and enterprise needs. Consequently, it eliminates the necessity of blockchain mining.

Many hackers have targeted HDFS mainly because of the values of the data it holds. The data in HDFS is mostly organizational data that contains much information. A malicious user who gains access to such a port can easily modify the data inside the DataNode, and all it takes is the name of the user who is using the HDFS. To tackle such attacks or deal with the threat model, we need to take the advantage of blockchain to create an immutable set of logs for HDFS, which can serve as a trusted source of information for an effective investigation when something has gone wrong in HDFS.

The BlockHDFS is the proposed solution which enhances the security of HDFS in a user transparent way by storing metadata of files from HDFS in a blockchain. BlockHDFS consists of three components: an HDFS cluster including the NameNodes and DataNodes, a permissioned blockchain network such as Hyperledger Fabric, and a NodeJS Client which acts as a bridge connecting the HDFS cluster and the blockchain network. In BlockHDFS, the blockchain is responsible for storing the metadata of the files. Overheads incurred by storing the HDFS file metadata into the blockchain are from two aspects. First, the WebHDFS REST API needs to read the metadata, such as the hash value, of a file from HDFS. Retrieving file metadata from HDFS can introduce some latency. However, as will be shown in the experiments, such latency is not significant. Second, it takes additional operations to store the metadata into the blockchain. BlockHDFS includes the components of User, WebHDFS REST API, File checksum, NodeJS client and Hyperledger Fabric blockchain.

Then, the paper compares the performance difference between and HDFS and BlockHDFS and also talks about the security limitations current limitations and of BlockHDFS.

Integrating blockchain with distributed file systems such as HDFS can potentially improve security and traceability. In the context of this paper, the original design of Hadoop is more optimized for file processing instead of security-by-design. Hence, in this paper, we proposed a new approach to introduce blockchain (and more specifically, Hyperledger) to enhance the security of the HDFS ecosystem. In the current implementation, we have only added minimal metadata to the blockchain, but with BlockHDFS, one can easily add more features suited to their application needs. For future work, BlockHDFS can be extended to work in real-time with the file system and track all data between NameNode and DataNodes of the HDFS in the secure ledger with mul tiple nodes.

**HDFS**

World was different before the internet took the world by storm. Big data concept was not introduced. The amount of data generated was relatively slow paced. All the data was mostly documents and in the form of rows and columns. Even a single storage and processor could handle both the storage and processing jobs without any issues.

But as years passed by, everything digitized, induced the dramatic increase of data generation in a multitude forms and formats. Every millisecond raw data ( semi -structured and unstructured data) was available in the form of emails, images, audios and videos which is collectively known as big data. It became nearly impossible to handle this size of data. The combination of a single storage unit and processor was obviously not a solution. The ultimate solution is to have multiple storage units and processors. This was when the Hadoop framework was introduced. The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware that could store and process huge amounts of data efficiently. It is part of the Apache Hadoop Core project and was originally built as infrastructure for the Apache Nutch web search engine project.

HDFS consists of three components that were specifically designed to work on big data. The first component of Hadoop is its storage units, the HDFS storage. Storing large data on one computer is unfeasible hence data is distributed amongst many computers and stored in blocks for example if you have 600 megabytes of data to be stored, HDFS splits the data into multiple blocks of data that are then stored on several data nodes in the cluster.

128 megabytes is the default size of each block hence 600 megabytes will be split into four blocks a, b, c, and d of 128 megabytes each and the remaining 88 megabytes in the last block e. Even if one data node crashes we do not lose that specific piece of data. HDFS makes copies of the data and stores it across multiple systems for example when block ‘a’ is created it is replicated with a replication factor of 3 and stored on different data nodes this is termed the replicated method. It makes HDFS fault tolerant.

Once data is stored successfully it needs to be processed; this is where the second component of hadoop mapreduce comes into play. In the traditional data processing method, entire data would be processed on a single machine having a single processor. This consumed time and was inefficient especially when processing large volumes of raw data. To overcome this problem, mapreduce splits data into smaller parts and processes, each of them separately on different data nodes. The individual results are then aggregated to give the final output.

The next step is the mapper phase where occurrence of each word is counted and allocated a number depending on the words. Similar words are shuffled, sorted and grouped in the reducer phase where all the grouped words are given a count and finally, the output is displayed by aggregating the results. MapReduce processes each part of big data individually and then sums the results at the end, this improves load balancing and saves a considerable amount of time.

Next step is to run it on the Hadoop cluster. This is done with the help of a set of resources such as RAM, Network Bandwidth and CPU. Multiple jobs are run on Hadoop simultaneously and each of them needs some resources to complete the task successfully.

To efficiently manage these resources we have the third component of Hadoop which is a yarn that consists of a resource manager node, manager application master, and containers. The resource manager assigns resources. Node manager handles the nodes and monitors the resource usage. In the node, the container holds a collection of physical resources.

In order for us to process the MapReduce job we created, a few steps are required. First, the application master has to request the container from the node manager once the node manager gets the resources it sends them to the resource manager this way yarn processes job requests and manages cluster resources in Hadoop. In addition to these components, Hadoop also has various big data tools and frameworks dedicated to managing, processing, and analyzing data. The Hadoop ecosystem comprises several other components like hive, pig, apache-spark, flume, and scoop.

The blockchain is an elegant architecture for systems that track ownership of things, thanks to its democratic distributed nature, the absence of a central authority, and the immutability of data. It also has many other uses, “smart contracts” being perhaps the most hyped of all. However, the ability to program complex contracts into computer-trackable code is proving to be a significant challenge. There are several blockchains with different ways to tackle this problem. Theoretically, they all verify the source and history transactions with technology developed in the 90′s that made torrents & Napster. The ledger they hold does not trust anyone except for a consensus, i.e. the transaction is held in many places globally and verified by many machines, so they believe the consensus. This is the ‘work’ which helps push the supply side market called mining. The coin cap has nothing to do with blockchain; it is merely a method companies use to control valuation and devaluation. In a real life scenario there are many problems with the model that have to be addressed, particularly in Bitcoin. For example, bitcoin mining is extremely heavy on energy consumption, which effectively devalues the network liquidity. There are also extensive cyber security costs which protect the vault and operations, again devaluing the market. Also, with regards to coin caps, usually those can usually be changed in the future based on consensus of miners, or a small group of miners that control the pot.

**Blockchain can be Used At:**

* Online voting that cannot be hacked.
* Auditability of every single vote (only the voter can know who he/she voted for, but the vote itself can be audited)
* Political party registration cannot be hacked or changed except by the voter
* Voter registrations cannot be hacked or changed
* A direct auditable link between voter, voter registration, party registration, primary vote, and election vote
* Easy online voting convenience for all legally registered voters
* Voting machines at polling locations would be locked down using blockchain to prevent hacking
* Blockstack technology that allows the voter to control every aspect of their vote within the constraints of state-controlled election laws

The best thing about Blockchain technology is decentralizing the systems so that they are not controlled or moderated by central agencies. This eliminates the middleman or brokers. The systems become more Democratic and can avoid lots of scams. Secure, immutable, transparent and global transactions in a trustless environment is the main advantage of blockchain technology.

**Data Management on Blockchain:** How data gets managed on a blockchain - All transactions are grouped in an individual unit, and this unit is called a block.

Standard properties for every block which is used for managing data on blockchain technology :

* Index: Index is nothing but a sequential block number
* Timestamp: Time on which block the data is added
* Hash: this is a unique hash value for the data. It is generated using mathematical functions. Every block has a different hash value, which directly communicates with the data (i.e. if the data gets changed, then the hash value also gets changed)
* Previous Hash: It contains a hash value of the previous block, which is to get a backward reference.