International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 12 | Dec - 2021 ISSN: 2582-3930

AMALGAMATION OF BLOCKCHAIN AND BIG DATA

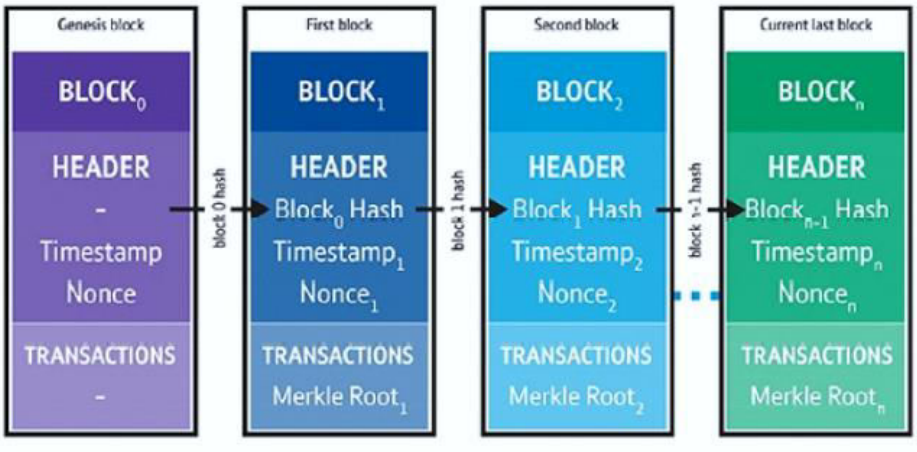
Rahul Dodke[1]

Abstract: Blockchain is a shared, distributed ledger that records transactions across business networks with the aim of helping businesses remove inefficiencies from trade. It does this through the use of cryptographic proofs, which help engender trust by ensuring facts are reported consistently to all those with a need to see them. Think of a database that instead of storing all the database entries on one computer it stores the data on multiple computers. Each block contains records and transactions; these blocks are shared across multiple computers and should not be altered absent an agreement (consensus) of the entire network. Big Data applies to information that can't be processed or analysed using traditional processes or tools. The term Big Data is quite vague and ill defined. It is not a precise term and does not carry a particular meaning other than the notion of its size. Big data task requires that large amount of computational space, to generate the terabytes of data for ensuring the successful data processing techniques. Three characteristics define Big Data: volume, variety, velocity, Veracity and Value. In blockchain technology covers the flaws of big data in fruitful relationship, with the factors of security, transparency, decentralization and flexibility, so that data to be analyze in different and efficient way for organizations all sizes in data analytics form.

Keywords: Big Data, Blockchain, Bitcoin, Cryptocurrency, Classification, Components, Use Cases

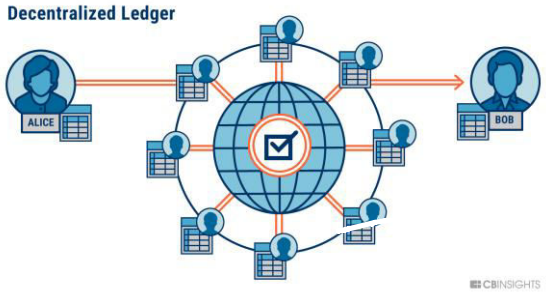
1. INTRODUCTION

Big Data is not Specific application type, but rather a trend –or even a collection of Trends- napping multiple application types. Data growing in multiple ways – More data (volume of data) – More Type of data (variety of data) – Faster Ingest of data (velocity of data) – More Accessibility of data (internet, instruments, …) – Data Growth and availability exceeds organization ability to make intelligent decision based on it. Indeed, we are dealing with a lot of complexity when it comes to data. Some data is structured and stored in a traditional relational database, while other data, including documents, customer service records, and even pictures and videos, is unstructured. Companies also have to consider new sources of data generated by machines such as sensors. A blockchain consists of a collection of data (a *block*) linked to the previous block. How are they linked? A block contains data, and each block references the block preceding it, so they are linked just as a chain link would be connected to the chain link before it. So, a blockchain contains blocks, which hold records of transactions. The private keys are held by the owner to show proof of ownership (this is the digital signature), so no one without the private key can decrypt the string and claim ownership.



2. WHAT DO YOU MEAN BY BLOCKCHAIN?

Blockchain is the core technology, or the heart behind bitcoin and in fact behind all cryptocurrency platforms. Cryptocurrency is a digital asset designed so that electronic cash is able to be exchanged using strong cryptography (encryption and decryption) to ensure the security of funds, transactions, and the creation of new funds. An example of cryptocurrency is bitcoin. Although bitcoin was not the first cryptocurrency invented, it's generally considered the first successful cryptocurrency. Blockchain aims to solve the problems of ledgers and contracts by sharing them in an unambiguous form between the participants of the business networks. Blockchain can be referred to as a *shared, distributed*
ledger with smart contracts.



SMART CONTRACT: Smart contracts refer to computer code that is shared between participants of the business network, and these implement the business rules associated with each transaction. As the code is shared, it can be executed by all relevant participants and they can agree on the output.

The blockchain implements several related qualities of service including consensus, provenance, immutability, and finality.

- *Consensus* is the process by which transactions are agreed upon by participants on the network. This means agreeing which transactions occurred, in what order, and what the result of running each transaction was.
- *Provenance* means that it should be possible to review prior transactions to determine the history associated with assets.
- *Immutability* is the fact that the shared transaction history cannot be tampered with. Once a transaction has been agreed to through consensus by the network and stored on the blockchain, it cannot then be edited, deleted, or have new transactions inserted before it.
- *Finality* is the property that the transaction cannot be modified once it is agreed upon.

It is a digital decentralized (no financial institutions involved) and distributed ledger. In layperson's terms, it is a database that stores records and transactions on multiple computers without one controlling party and according to an agreed policy. The data that is stored is a block, and the blocks are linked (chained) together to form a blockchain. The blockchain is not just one master computer and aims to work globally. It achieves integrity with a consensus of the data by all the computers connected on the network. A distributed consensus means that a pool of peers, geographically apart, agree in a decentralized manner, instead of one master computer (centralized). Instead of regulations, there are rules that are usually set in an open-source environment instead of being set by a government entity.

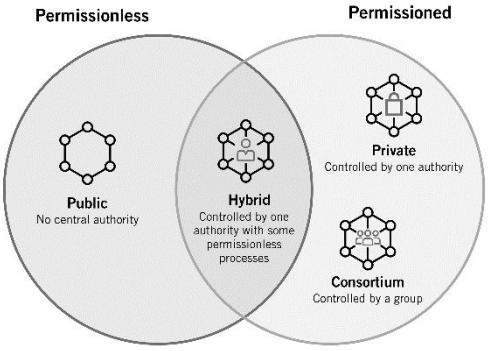
TYPES OF BLOCKCHAIN: There are four types of Blockchain,

Public Blockchain: Public blockchains are permissionless in nature, allow anyone to join, and are completely decentralized. Public blockchains allow all nodes of the blockchain to have equal rights to access the blockchain, create new blocks of data, and validate blocks of data.

Private Blockchain: Private blockchains, which may also be referred to as managed blockchains, are permissioned blockchains controlled by a single organization. In a private blockchain, the central authority determines who can be a node. The central authority also does not necessarily grant each node with equal rights to perform functions. Private blockchains are only partially decentralized because public access to these blockchains is restricted.

Consortium Blockchain: Consortium blockchains are permissioned blockchains governed by a group of organizations, rather than one entity, as in the case of the private blockchain. Consortium blockchains, therefore, enjoy more decentralization than private blockchains, resulting in higher levels of security.

Hybrid blockchain: Hybrid blockchains are blockchains that are controlled by a single organization, but with a level of oversight performed by the public blockchain, which is required to perform certain transaction validations.



3. WHAT DO YOU MEAN BY BIG DATA?

Although the term "Big Data" has become popular, there is no general consensus about what it really means. Often, many professional data analysts would imply the process of extraction, transformation, and load (ETL) for large datasets as the connotation of Big Data.

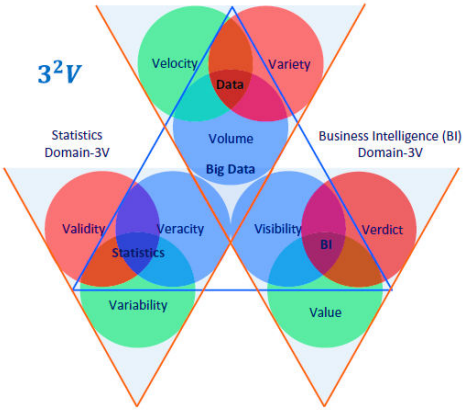
Table 1 Typical Size of Different Data Files

Media	Average Size of Data File	Notes (2014)
Web page	1.6–2 MB	Average 100 objects
eBook	1–5 MB	200–350 pages
Song	3.5–5.8 MB	Average 1.9 MB/per minute (MP3) 256 Kbps rate (3 mins)
Movie	100–120 GB	60 frames per second (MPEG-4 format, Full High Definition, 2 hours)

"Big Data will be a source of new economic value and innovation"

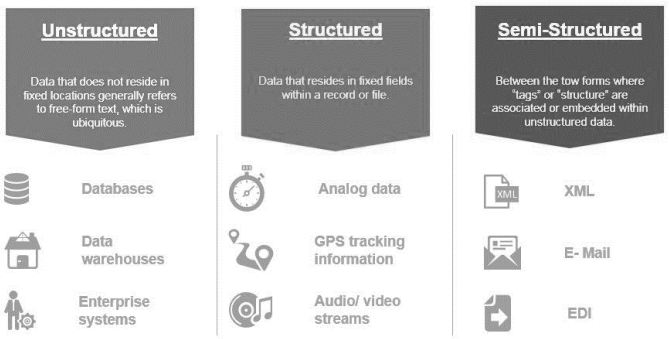
Big data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction. As we note earlier in this chapter, big data is typically broken down by three characteristics:

- **Volume**: How much data
- **Velocity**: How fast that data is processed
- **Variety**: The various types of data
- **Veracity:** How data is inconsistent and uncertain
- **Value:** The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.



Types of Big Data:

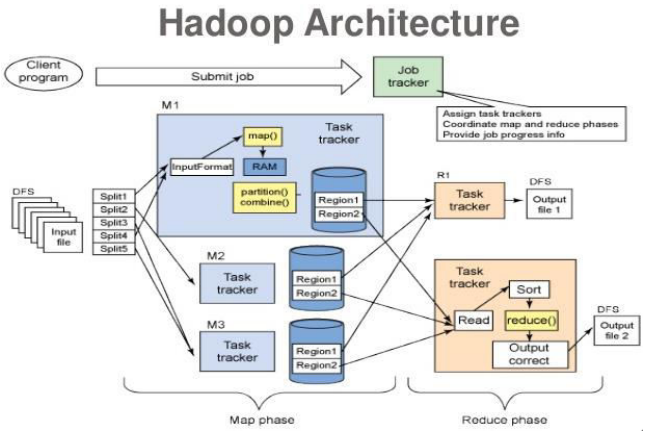
- Structured Data: Structured is one of the types of big data and By structured data, we mean data that can be processed, stored, and retrieved in a fixed format. It refers to highly organized information that can be readily and seamlessly stored and accessed from a database by simple search engine algorithms.
- Unstructured data: Unstructured data refers to the data that lacks any specific form or structure whatsoever. This makes it very difficult and time-consuming to process and analyze unstructured data.
- Semi structured Data: Semi structured is the third type of big data. Semi-structured data pertains to the data containing both the formats mentioned above, that is, structured and unstructured data.



Big Data Enabling Technologies:

It's a collection of data and process using different open-source tools. The open-source tools are; Apache Hadoop, Hadoop HDFS, Hadoop Yarn, Hadoop Map Reduce, Hive, Cassandra, Apache Zookeeper, Apache HBase, Apache Spark, No SQL, Kafka, Spark Streaming Ecosystem, Spark MLib, Spark GraphX

- Apache Hadoop – Apache Hadoop systems are open-source framework in big data.
- MapReduce function, In MapReduce function, Hadoop distributed file system move to the map function with the input of (key, value) pairs, the pairs are shuffle & sort; its produce the intermediate values, the values merged into reduce function, after reduce function final key and value can be produced
- YARN – YARN is a Yet Another Resource Manager. The resources can be allocated and executed on different cluster nodes.
- Hive – To access big data, Hive systems support for HSQL.
- Apache Spark – its otherwise big data analytics framework, academic & industry gained lot off attraction in Apache spark, its designed for fast computation
- Zookeeper - Zookeeper is a highly reliable distributed coordination kernel, which can be used for distributed locking, configuration management, leadership election, work queues... Zookeeper is a replicated service that holds the metadata of distributed applications. Key attributed of such data; small size, performance sensitive, dynamic and critical.
- NoSQL - NoSQL technology is new and with wide variety of databases that can stored unstructured data.
- Cassandra – Data is placed on different machines with more than one replication factor that provides high availability and no single point of failure.



4. BLOCKCHAIN USECASES IN BIG DATA & RESULTS

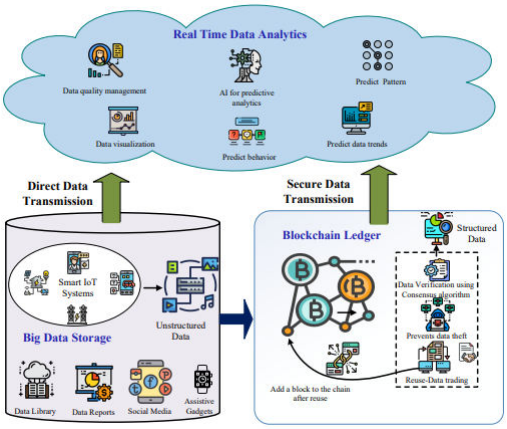
Big data analytics process provided in the form of both industry and academia. It operating and managing with own data in Blockchain technology. The blockchain systems ensure that privacy and integrity of data, it predicts large amount of data in big data techniques but focus on validating data, the data to be brought together in decentralized manner and the origin of the data linked in chained manner. The motivations of integrating blockchain with big data are discussed as follows.

- Improving Big Data Security and Privacy: As the number of devices connected to the Internet is growing day by day, the quantity of the data stored at third party locations like cloud is increasing rapidly.

- Improving Data Integrity: There exists a likelihood of people tampering the records in big data to influence the prediction of big data analytics in their favor. The immutability property of the blockchain ensures that it is next to impossible to tamper with the data stored in the blockchain network.
- Fraud Prevention: The existing big data solutions rely on the analysis of patterns in the historical data to detect fraudulent transactions. Hence big data cannot solve the problem of fraudulent transactions in the financial sector.

Real-Time Data Analytics: Since the blockchain stores every transaction, it makes the real-time analytics of big data achievable. The banks and financial institutes can settle the cross-border transactions including large amounts in near real-time as the blockchain integrated big data analytics enables the financial institutes to settle the transactions quickly.

- Enhancement of Data Sharing: The integration of blockchain with big data helps service providers to share the data to other stakeholders with minimal risk of data leakage.
- Enhancement of the Quality of Big Data: Data scientists spend most of their time on data integration as different sources follow different formats in data collection. By using blockchain for data storage, the quality of the data can be improved as it is structured and complete.
- Streamlining the Data Access: The use of blockchain would simplify the life cycle of big data analytics by online streamlining the data access.



5. BLOCKCHAIN SERVICES FOR BIG DATA

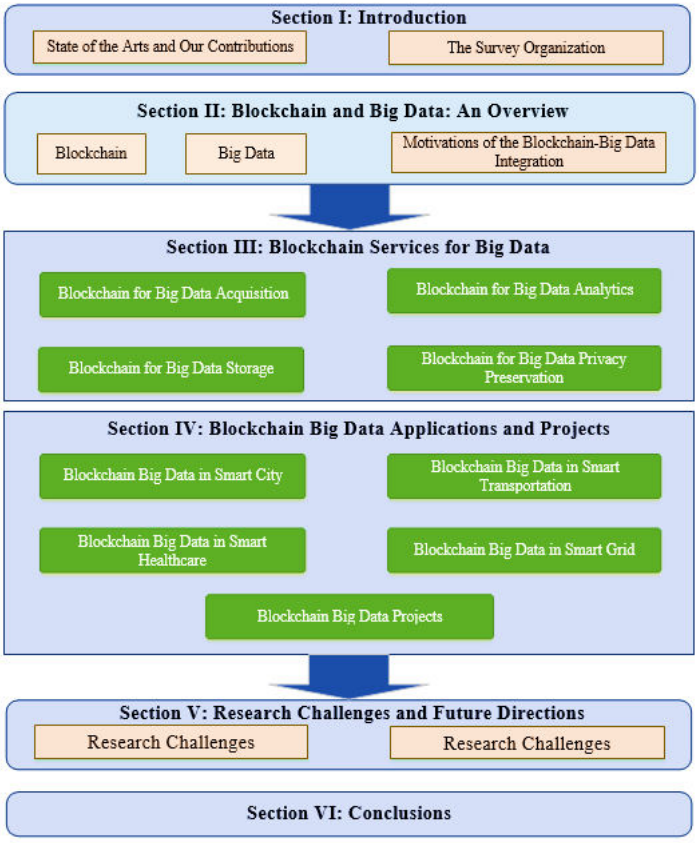
A. **Blockchain for Big Data Acquisition:** In general, big data applications acquire data from di- versified sources in a different format (unstructured data). These data cannot be processed in the native form. Therefore, the data must be converted to a structured format from which various predictions on the application domain can be made.

- Blockchain for Secure Big Data Collection
- Blockchain for Secure Big Data Transmission/Sharing

B. **Blockchain for Big Data Storage:** There are several cloud-based services available to store and access files from anywhere on any machine. Users, particularly organizations are hesitant to store sensitive information on the system managed by a third party.

- Blockchain for Secure File Systems
- Blockchain for Secure Database Management

- Blockchain for Big Data Storage Infrastructure



C. **Blockchain for Big Data Analytics:** The development of edge and cloud computing has increased the amount of data in various scenarios. The extensive construction and generation of data from sensors, social media, web and IoT devices resulted in the growth of Artificial Intelligence (AI) techniques.

- Blockchain for Secure Data Learning in AI Algorithms
- Blockchain for Secure Data Training

REFERENCES

1. E. Karafiloski and A. Mishev, "Blockchain solutions for big data challenges: A literature review," in IEEE EUROCON 2017-17th International Conference on Smart Technologies, Ohrid, Macedonia, 2017, pp. 763–768.
2. N. Tariq, M. Asim, F. Al-Obeidat, M. Zubair Farooqi, T. Baker, M. Hammoudeh, and I. Ghafir, "The security of big data in fog-enabled IoT applications including blockchain: a survey," Sensors, vol. 19, no. 8, p. 1788, Apr. 2019.

3. D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, “Integration of blockchain and cloud of things: Architecture, applications and challenges,” IEEE Communications Surveys & Tutorials, 2020.
4. Q. Zhou, H. Huang, Z. Zheng, and J. Bian, “Solutions to scalability of blockchain: A survey,” IEEE Access, vol. 8, pp. 16 440–16 455, 2020.
5. S. K. Singh, S. Rathore, and J. H. Park, “Blockiotintelligence: A blockchain-enabled intelligent IoT architecture with artificial intelligence,” Future Generation Computer Systems, vol. 110, pp. 721–743, 2020.
6. W. Viriyasitavat and D. Hoonsopon, “Blockchain characteristics and consensus in modern business processes,” Journal of Industrial Information Integration, vol. 13, pp. 32–39, Mar. 2019.
7. L. Da Xu and W. Viriyasitavat, “Application of blockchain in collaborative internet-of-things services,” IEEE Transactions on Computational Social Systems, vol. 6, no. 6, pp. 1295–1305, 2019.
8. C. Berg, S. Davidson, and J. Potts, Understanding the blockchain economy: An introduction to institutional cryptoeconomics. Edward Elgar Publishing, 2019.
9. Y. Yuan and F.-Y. Wang, “Blockchain and cryptocurrencies: Model, techniques, and applications,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 48, no. 9, pp. 1421–1428, 2018.
10. J. Abou Jaoude and R. G. Saade, “Blockchain applications–usage in different domains,” IEEE Access, vol. 7, pp. 45 360–45 381, 2019.
11. S. Shalini and H. Santhi, “A survey on various attacks in bitcoin and cryptocurrency,” in International Conference on Communication and Signal Processing (ICCSP), 2019, pp. 0220–0224.
12. G. T. Reddy, M. P. K. Reddy, K. Lakshmanna, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker, “Analysis of dimensionality reduction techniques on big data,” IEEE Access, vol. 8, pp. 54 776–54 788, 2020.
13. D. B. Rawat, R. Doku, and M. Garuba, “Cybersecurity in big data era: From securing big data to data-driven security,” IEEE Transactions on Services Computing, 2019, in press.
14. M. Tang, M. Alazab, and Y. Luo, “Big data for cybersecurity: Vulnerability disclosure trends and dependencies,” IEEE Transactions on Big Data, vol. 5, no. 3, pp. 317–329, Sep. 2019.
15. P. Sharma, R. Jindal, and M. D. Borah, “Blockchain technology for cloud storage: A systematic literature review,” ACM Computing Surveys (CSUR), vol. 53, no. 4, pp. 1–32, 2020.