WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Blockchain in Data Science

Javid Hussain A[1], Kishoth G [2]

[1]Student, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu-641008
[2]Student, Department of Computer Science, Sri Krishna Arts and Science College, Coimbatore, Tamil Nadu-641008
[1]Email:javidhussaina21bds018@skasc.ac.in, Phone:9787586878 [2]Email:kishothg21bds020@skasc.ac.in,Phone:8248064889

ABSTRACT

Today, most data scientists are using blockchain technology to confirm the authenticity and track data at all points on the chain.Blockchain is Immutable Security. So, Most Data scientists use Blockchain in Data Science for its large-scale adoption. Ina blockchain, data is protected at all steps through different signatures.Blockchain technology is a hot topic current day, especially with the recent need in finance, the rapid growth of Bitcoin and cryptocurrencies, and the current NFT. From a Data Scientist's point of view, blockchains are also an interesting source of important data that can be used to tackle a wide range of interesting problems using Statistics and Machine Learning.

1.Introduction

1.1 Blockchain

Blockchain is a digital ledger that records every transaction that takes place. As there is no single authority, meaning that no one can change the transactions that take place in the ledger.

The information that is registered in the blockchain data structure cannot be done as altering one block means changing all other blocks that follow it. In case one past block is changed, all the following blocks are changed as a result. Thus, it is not possible for the change in even one block to escape being noticed.

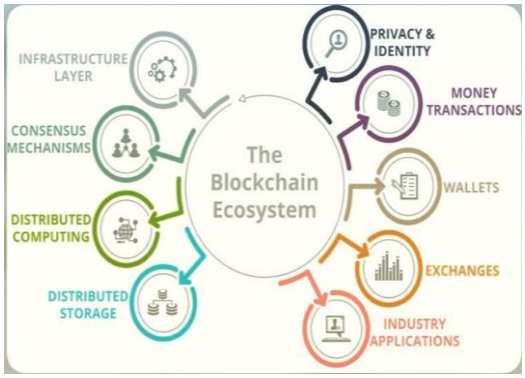

Fig.1

1.2 History of blockchain

Blockchain technology was invented in 1991 by the research scientist Stuart Haber and W. Scott Stornetta. They wanted to introduce a computer-related practical solution for time-stamping digital documents so that they cannot be backdated. Using Cryptography, they developed a system of the secured chain of blocks to store the time-stamped documents.
Structures that are called blocks store data as records. If Multiple Structures are connected in a chain, Then it is called a block in the chain. Then the storage unit is called Digital Ledger.

1.3 Data Science

Data Science is one of the trending technology nowadays. The field includes a lot of innovations such as Predictive analytics, Diagnostic analytics, and Descriptive Analytics.
Data Science aims to extract insights and also other information from data, both structured and unstructured data. The field of data science includes machine learning, data analysis, statistics, and other advanced concepts that are assigned to gain knowledge of the actual processes that use data.
Corporate kings such as Facebook, Google, Apple, and Amazon are mining volumes of data every day. The vast field of data science has spread like fire for the demand for data scientists who are tasked with deriving meaning from data and assisting in solving real-world problems.

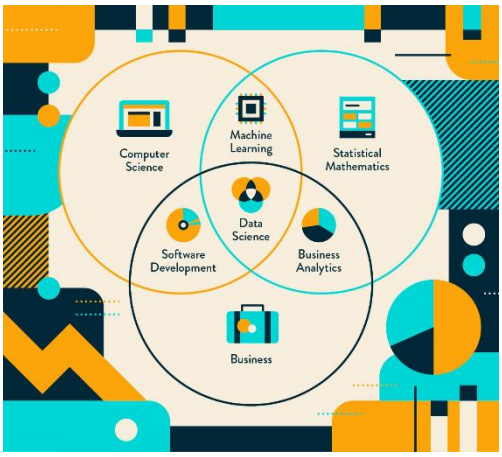

Fig.2

This demand is also fed by the area of Big Data, an advanced area of data wisdom that deals with extremely huge volumes of data that cannot be handled by the conventional data handling ways.

Blockchain related to Data Science

The relationship between blockchain and data wisdom, if there's any, has not been delved into much. Looking at it simplistically, both these technologies have data at the center. While blockchain validates and records data, data wisdom focuses on inferring meaningful perceptivity from data for the problem- working.
Both of these technologies employ algorithms to control relations with different data parts. In crux, data wisdom is for prognosticating and blockchain is for validating data.

- **High data quality**: All new records go through a blockchain-specific validation process done by one of the many consensus mechanisms. Once validated, these records areunchangeable.No one can modify them for any purpose, good or malicious. Blockchain data are typically well structured and are well documented. This makes the work of a Data Scientist who works with such data easier and more predictable.

- **Traceability**:Blockchain records contain all the information needed to track their origin and content, e.g., what the transaction started, when it happened, the amount of the asset transferred, and what address they received that asset. In addition, most social blockchains have "testers" - websites where anyone can check any record that has been created on the appropriate blockchain (see, for example, Bitcoin, Ethereum, and Ripple testers).

- **Built-in anonymity**:Blockchains do not require their users to provide any personal, important information in a world where maintaining personal privacy is a real problem. According to Data Scientist, this helps to overcome headaches associated with some of the rules (e.g., GDPR in Europe) that require personal data to be anonymous before processing.

- **Large Data Volumes**:Many machine learning algorithms require big amounts of data to train. This is not a problem for mature blockchains, which provide more data.Collecting data related to the problem is the first step that Data Scientists are likely to in their blockchain-related projects. Using the aforementioned explorer websitesindividual blockchain records can be easily detected. However, programmatically collecting larger datasets suitable for Data Science purposes can be a difficult task that may require special skills, software, and financial resources. There are three main options one could consider. They are,

- Use Datasets Already prepared
  by someone else

- API or ETL Tool

- Use Commercial Solutions

BlockChain will Improve Data Science

Allows data traceability: Peer-to-peer cooperation is facilitated using the blockchain. If a published account, for example, fails to adequately describe any method, any peer can analyze the whole process and determine how the results were achieved.Anyone can learn whether data is accurate to use, how to store it, how to update it, where it comes from, and how to use it properly due to open channels of damage. In conclusion, blockchain technology will allow users to track data from start to finish.

Allows for real-time analysis: Analyzing data in real-time is very difficult. The most effective way to locate fraudsters is to be able to monitor progress in real-time. In the long run, however, real-time analysis was not possible. Companies can now detect any data flaws from the beginning, due to the blockchain nature of the split.

Ensures the accuracy of data: Data on the blockchain digital log is stored in a variety of locations, both private and public. Data is checked and tested at the entrance before any additional blocks are added. This process itself is a way of verifying data.

Fig.3

3.**Big Data**

Big Data is a collection of data that is greater in volume but growing rapidly with time. Big data is so large in size and complexity that none of the traditional data management tools can store it. Big data is also known as Huge Data.

3.1Types of Big Data

Following are the types of Big Data:

1. Structured
2. Unstructured
3. Semi-structured

3.1.1 Structured

Any data that can be saved,accessed, and processed in the form of fixed-format is known as ‘structured’ data. Over the period, knowledge in computer science has achieved greatsuccess in developing techniques for working with such kind of dataand also deriving value out of it. However, we are foreseeing issues when the size of such data grows to a huge extent,sizes are being in the range of multiple zettabytes.

3.1.2 **Unstructured**

Any data with an unknown form or structure is classified as unstructured data. According to its huge size, unstructured data poses multiple challenges in its processing for deriving value from it. An example of unstructured data is a heterogeneous data source containing a combination oftext files, images, videos, etc. Nowadays organizations have wealth of data available to them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.

3.1.3 **Semi-structured**

Semi-structured data can contain both forms of data. We can see semi-structured data as a structured form but it is not defined with eg. a table definition in relational DBMS. An example of semi-structured data is data represented in an XML file.

3.2 **Characteristics of Big Data**

Big data can be described by the following characteristics:

•Volume
•Variety
•Velocity
•Variability

3.2.1 **Volume**

The name Big Data itself is related to a very large size. Data size plays a very important role in determining the amount of data. Also, whether certain data can be considered Big Data or not, depends on the amount of data. Therefore, 'Volume' is one factor that needs to be considered when working with Big Data solutions.

3.2.2**Variety**

The coming aspect of Big Data is its diversity.
Variety refers to the colorful sources and nature of data, both formal and informal. In earlier days, spreadsheets and websites were the only data sources considered for utmost operations. Moment, data in the form of emails, prints, vids, covering tools, PDFs, audio,etc. they're also considered in the analysis programs. This variety of randomized data raises specific problems for data storehouse, digging, and analysis.

3.2.3 **Velocity**

The term 'speed' refers to the speed of data processing. How quickly data is produced and processed to meet needs, determines the actual power of data.
Big Data Velocity is about the speed at which data enters from sources such as business processes, application logs, networks, social networking sites, sensors, Mobile Devices, etc. The data flow is huge and continuous.

3.2.4 **Variability**

This means inconsistencies that data can sometimes show, thus disrupting the process of managing and managing data effectively

4.Conclusion

Data Science and Blockchain Technology can be integrated to transform the way we process and analyze data. High-Level Level, Variety, and High Data Growth and the rapid development of the data app have put a tremendous need for user value, compliance, and efficient use of the power of privacy protection application. In this paper, we first provided an overview of blockchain, data science, blockchain-related data science, blockchain will improve data science and Big Data and its variants, as well as other key features of Block Chain.

References

[1] Vibhuthi Viswanathan, "Implication of Blockchain in Data Science".https://www.itprpportal.com/, March 15, 2022

[2] Sergey Mastitsky on Data Science for Blockchain: Understanding the Current Landscape, https://towardsdatascience.com/data-science-for-blockchain-understanding-the-current-landscape-c13615c367e, Apr 26, 2021

[3] Shaoliang Peng, "Blockchain for data Science", in Research Gate, March 2020