Big Data and Analytics in Block Chain

Ashok Kuppuraj
Indiana University
Bloomington, Indiana 43017-6221
akuppura@iu.edu

ABSTRACT

This paper describes how Big data and its technologies helps in augmenting or improving the current Block chain technology and overcome the problems around it.

KEYWORDS

Big Data, Block Chain i523

Images are not in figure environment, se original template or other peoples papers

1 INTRODUCTION

The objective is to concur the abilities of the two broad topics in the current technology world, the Big Data, and Block Chain. Block chain and Big data are still evolving technologies, which gives us enough opportunity to explore and invent new concepts for its own good. As these are still evolving, we can leverage one's solution on the other. To leverage each one's problems and solutions, we must first identify the similarities in two frameworks and how these similarities are related and what solution we are going to adopt.

2 WHAT IS BIG DATA

Big data can be described as any type of data with large volume, velocity, and variety [3]. The history of Big data starts the moment we started using the computer back in the 90s, however, we choose not to use all the generated data due to constraints in the processing system and storage systems. Later, people understood that they are missing lot of information useful to the business due to these constraints, and started leveraging Data warehouse to processes the data in Batch after data generation. At a certain point in time, even the data warehouse systems are not capable to handle the volume and velocity of data we are generating. This exponential growth in the data generation due to wide adoption of computers by humans in the form of mobile, PCs and introduction of IoT sensors, resulted in the need for technology to process these data and it is termed as "Big Data" [4]

3 BLOCK CHAIN

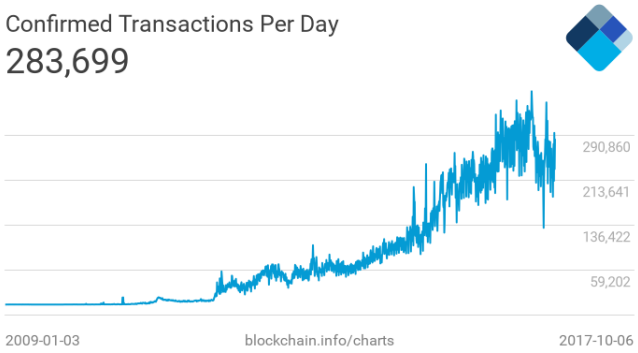
Blockchain can be defined as a decentralized, public ledger persisted in a connected set of immutable Blocks. The core idea is to perform any set of a transaction without a governing third-party avoiding Double spending by Distributed consensus. A transaction happens with an entity called tokens, tokens are the actual digital asset of a blockchain. The implementation begins with an entity A initiating the transaction, the initiated transaction request from A to B is broadcasted with Gossip protocol to most of the nodes, the transaction is validated by miners with the ledger available with them, the validation includes checking digital signatures and the previous input to that entity (i.e current withholding). Later the validated transactions are grouped with reference to its previous address and added as the current block. This block is then broadcasted to the network and the network peers validate the block and added them to their ledger, confirming the transaction. Hence, termed as "BlockChain" [7].

4 BIGDATA VS BLOCKCHAIN

As far as data is concerned, both Big data technologies and Blockchain go in parallel.Both involve processing data at volume, velocity, and variety which is the basic evaluation factor for defining big data. In the below section, analysis are made on how these three V's corresponds to Block Chain, with an example from Bitcoin, one of the front-runners in implementing Blockchain technologies.

4.1 Data Volume

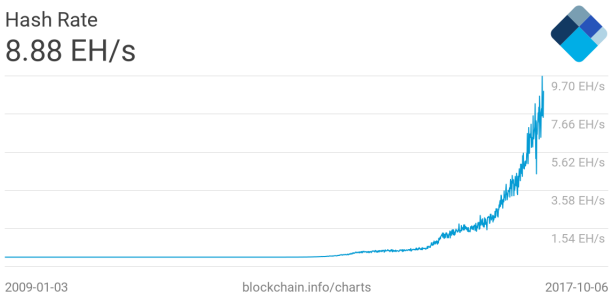
Though the data volume share of blockchain is considerably low compared with current Big data average, the volume it generates in an overall network perspective in terms of network I/O, logs, transaction data, it fits well with the terms of big data. For example, consider the transaction growth of Bitcoin [6], the volume of the transaction was averaging 5K in 2011, whereas in 2017 the average is 200K with an increase of 400 percent over 5 years and the volume is likely to grow in an exponential scale with the global acceptance of Blockchain technologies.



4.2 Data velocity

In data terms, even 10 MB of data is considered huge when its getting generated within a span of minute, hence we must consider the velocity as an important metric in analyzing the data, here in Bock chain, though the transactions are not of high volume, but other non=token transactions like Gossip calls, smart contract transfers, block transfers, acceptance protocol were happening

every second, which in turns generate huge amount of data within 10 minutes of this interval with respect to Bitcoin.



The snippet describes the number of hashes resolved per second, in Bitcoin's network [1].

4.3 Data Variety

In a wide perspective, Blockchain deals with multiple varieties of structured data like Token data, Smart contracts, consensus data, logging data and unstructured data like videos [5], audio depending upon the use-case of the Blockchain. And the popularity it has now and based on the current growth trend in the acceptance of decentralization, Blockchain technology will tend to generate more data in a wide range of varieties.

5 IMPLEMENTATION - BIG DATA TECHNOLOGIES IN BLOCKCHAIN

Before we start on the implementation of Big data technologies, its required to identify the problems Blockchain faces now in scope with Big data's solutions, one of the main problems are Slow transactions and visualization through complex analytical calculation. Though we have other problems, we consider the above two as the most important to enhance the success rate of this technology. For example, consider Bitcoin, the overall transaction timing is very fewer in-terms of interbank transaction however for the end user it takes a minimum of 20 mins to complete a transaction whereas, in Visa, for instance, it can perform up to 1700 transaction per seconds.

5.1 Transaction processing

To deal with improving the transaction speed of a peer-to-peer network, first it is required to streamline the asynchronous process of gossip protocols, handshake between peers to increase the transaction processing speed, also by removing the block size limits[2] we can increase the frequency of the block building, these optimizations can be easily achieved with the help of Big data queuing utilities like Apache Kafka by creating individual topics for each set of Broadcasting, Block Acknowledgement, and consensus sharing between the peers, and second is to increase the Hash processing capacity by horizontal scaling with the help of Apache Spark or Apache Flink. By using these open source tools for hash processing, it is not required to invest on high-value GPU's to process data.

For an instance, Kafka can handle up to 200,000 messages/second (220MB/second)[**?**], which is way more than any other existing banking can provide.

5.2 Visualization

The current visualization options available with blockchain is based on the shared ledger available in the network, to fetch a real-time reporting or visualizing the happenings in the network, one must have to take part in the network and share all the interactions and the ledger details for any sort of analytical needs. As discussed in the previous section, if we start using Kafka for other peer-to-peer interactions via topics, we can seamlessly provide real-time reporting to users.

5.3 Smart-BlockChain

The next big leap in the Blockchain would be the implementation of Machine learning in the Blockchain. The current versions of blockchain don't give any machine learning modules or algorithm attached built along with. By including the machine learning modules in the blockchain network, Blockchain can be made smart by predicting malicious activities, optimizing transactions and evaluation of its sources.

6 CONCLUSION

Although Blockchain provides the solution for Real life problems, it would be nearly impossible without its implementations leaning towards Big data's solutions. Big data and the technologies is a front-runner in the handling of data of different volume, velocity, and variety which Blockchain is yet to reach. With the current acceptance rate of Blockchain, Bigdata, and Machine learning technologies, maybe in the future, countries don't need a leader to take a decision on their's behalf, people can collectively take a state's decision and election process will be so simple that it can happen everyday.

ACKNOWLEDGMENTS

The author would like to thank Dr. Gregor von Laszewski for his support and suggestions to write this paper.

REFERENCES

[1] 2015. Blockchain, velocity, Big Data. (2015). https://blockchain.info/charts/hash-rate?timespan=all&showDataPoints=true
[2] 2015. Optimizing Bitcoin. (2015). https://medium.com/@hudon/dear-bitcoin-this-is-how-you-can-beat-visa-b5ee857cf193
[3] 2015. What is Big data. (2015). https://www.sas.com/enus/insights/big-data/what-is-big-data.html/
[4] 2017. The Exponential Growth of Data. (2017). https://insidebigdata.com/2017/02/16/the-exponential-growth-of-data/
[5] 2017. Livepeer. (2017). https://medium.com/@petkanics/introducing-livepeer-a-decentralized-live-video-broadcast-platform-and-crypto-token-protocol-7
[6] 2017. n-transactions. (2017). https://blockchain.info/charts/n-transactions?timespan=all
[7] Satoshi Nakamoto (Ed.). 2008. *Bitcoin: A Peer-to-Peer Electronic Cash System.* 5.

7 BIBTEX ISSUES

Warning–I didn't find a database entry for "kafka-performance"

Warning–no key, author, or editor in Bigdataintro

Warning–no author, editor, organization, or key in Bigdataintro

Warning–to sort, need author, editor, or key in Bigdataintro

Warning–no key, author, or editor in datagrowth

Warning–no author, editor, organization, or key in datagrowth

Warning–to sort, need author, editor, or key in datagrowth

Warning–no key, author, or editor in bitcointrans

Warning–no author, editor, organization, or key in bitcointrans

Warning–to sort, need author, editor, or key in bitcointrans

Warning–no key, author, or editor in hastratepersec

Warning–no author, editor, organization, or key in hastratepersec

Warning–to sort, need author, editor, or key in hastratepersec

Warning–no key, author, or editor in livepeer-BC-stream

Warning–no author, editor, organization, or key in livepeer-BC-stream

Warning–to sort, need author, editor, or key in livepeer-BC-stream

Warning–no key, author, or editor in Optimize-bitcoin

Warning–no author, editor, organization, or key in Optimize-bitcoin

Warning–to sort, need author, editor, or key in Optimize-bitcoin

Warning–no key, author, or editor in Bigdataintro

Warning–no key, author, or editor in Bigdataintro

Warning–no key, author, or editor in bitcointrans

Warning–no key, author, or editor in bitcointrans

Warning–no key, author, or editor in datagrowth

Warning–no key, author, or editor in datagrowth

Warning–no key, author, or editor in hastratepersec

Warning–no key, author, or editor in hastratepersec

Warning–no key, author, or editor in livepeer-BC-stream

Warning–no key, author, or editor in livepeer-BC-stream

Warning–no key, author, or editor in Optimize-bitcoin

Warning–no key, author, or editor in Optimize-bitcoin

Warning–no key, author, or editor in hastratepersec

Warning–no author, editor, organization, or key in hastratepersec

Warning–neither author and editor supplied for hastratepersec

Warning–no journal in hastratepersec

Warning–no number and no volume in hastratepersec

Warning–page numbers missing in both pages and numpages fields in hastratepersec

Warning–no key, author, or editor in Optimize-bitcoin

Warning–no author, editor, organization, or key in Optimize-bitcoin

Warning–neither author and editor supplied for Optimize-bitcoin

Warning–no journal in Optimize-bitcoin

Warning–no number and no volume in Optimize-bitcoin

Warning–page numbers missing in both pages and numpages fields in Optimize-bitcoin

Warning–no key, author, or editor in Bigdataintro

Warning–no author, editor, organization, or key in Bigdataintro

Warning–neither author and editor supplied for Bigdataintro

Warning–no journal in Bigdataintro

Warning–no number and no volume in Bigdataintro

Warning–page numbers missing in both pages and numpages fields in Bigdataintro

Warning–no key, author, or editor in datagrowth

Warning–no author, editor, organization, or key in datagrowth

Warning–neither author and editor supplied for datagrowth

Warning–no journal in datagrowth

Warning–no number and no volume in datagrowth

Warning–page numbers missing in both pages and numpages fields in datagrowth

Warning–no key, author, or editor in livepeer-BC-stream

Warning–no author, editor, organization, or key in livepeer-BC-stream

Warning–neither author and editor supplied for livepeer-BC-stream

Warning–no journal in livepeer-BC-stream

Warning–no number and no volume in livepeer-BC-stream

Warning–page numbers missing in both pages and numpages fields in livepeer-BC-stream

Warning–no key, author, or editor in bitcointrans

Warning–no author, editor, organization, or key in bitcointrans

Warning–neither author and editor supplied for bitcointrans

Warning–no journal in bitcointrans

Warning–no number and no volume in bitcointrans

Warning–page numbers missing in both pages and numpages fields in bitcointrans

Warning–empty publisher in Bitcoin

Warning–empty address in Bitcoin

(There were 69 warnings)

8 ISSUES

DONE:

Example of done item: Once you fix an item, change TODO to DONE

8.1 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

8.2 Writing Errors

Errors in title, spelling - Blockchain

Spelling errors

Are you using *a* and *the* properly?

Do not use the phrase *In this paper/report we show* instead use *We show.* It is not important if this is a paper or a report and does not need to be mentioned

8.3 Citation Issues and Plagiarism

Claims made without citations provided

8.4 Structural Issues

The paper has less than 2 pages of text, i.e. excluding images, tables and figures

8.5 Details about the Figures and Tables

Missing captions

put figures at end of paper