IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 37-44

Compartive Analysis of Bigdata, Machine Learning, Block Chain: Technologies and Their Applications

[1]M.Vengateshwaran M.E., [2] A.Barkathunisha, [3]A.Vanisri,[4]R.Vijaya
[1]Assistant Professor, [234]UG Student
Department of Computer Science and Engineering
Arasu Engineering College
Kumbakonam
Tamilnadu, India

Abstract: *Now a days, large number of customers, on-line services and social media's are increasing day by day, so every second generating a huge amount of data in the format of unstructured. But it is not suitable for extracting particular information. Big Data technologies can be viewed as a new generation of technologies and architectures designed to extract value economically from very large volume of data by enabling high velocity capture, discovery and analysis. It surpasses the processing capacity of conventional DB systems. Big data might be petabytes (1024 terabytes) or Exabyte(1024 petabytes) of data consisting of billions to trillions of records from different sources such as web, sales, customer care, social media, mobile data etc., Big data refers to relatively large amounts of structured and unstructured data that is difficult to process using traditional database(RDBMS)and software techniques. Machine learning is an instance of these technologies. Various machine learning models that yield groundbreaking throughputs with high efficiency rates in predicting, detecting, classifying, discovering and acquiring in-depth knowledge about events that would otherwise be very difficult to ascertain have been made possible due to big data. The implementation of block chain in the proposed methodology results in low metadata access delay and thereby improves the execution time. This paper discusses the concept of big data, Machine Learning and Block chains. The aim of this paper is to encourage further research in incorporating the Block Chain Technology into Machine Learning.*
Keywords:- *BigData, Machine Learning, Block chain etc.,*

I. INTRODUCTION

Data can be defined as a collection of values of a specific variable either qualitative or quantitative. Whereas quantitative data highlights on quantity and numbers, qualitative data is more categorical and may be represented by categories such as height, color, race, gender, etc. Data is a very important resource in every research work. The type of data acquired coupled with the preprocessing techniques used contribute massively to great research achievements. Generally obtained through primary and secondary sources, data is primarily obtained by direct observations and through the conduction of surveys. Secondarily, data can also be acquired through rigorous market studies or information generated electronically or obtained from the worldwide web.

In Machine Learning, the bigger the data, the better the accuracy and greater the generalization ability of the model. i.e. Block chain implementation not only help save money but also helps in ensuring better machine learning models due to its decentralization ability cite100. In the next section, the concept of Big Data is broadly discussed, Section 4 discusses Machine learning and its associated technologies. In Section 5, the Block chain Technology is discussed

II. BIGDATA

BigData is a data that are enormous in size and exceeds the processing capacity of conventional database systems. It involves the data produced by different devices and applications. Now a day's used in agriculture, medical, marketing, social Medias and business informatics.

2.1 Characteristic of Big Data

BigData should have the following three Characteristics.

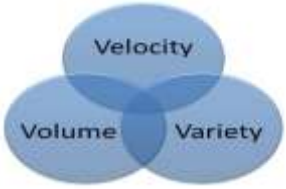

Fig.1 Characteristics of an BigData

✓ **Volume:** How much data(Size grows from terabyte to exabyte to zettabyte)
✓ **Velocity:** How fast that data is processed (Batch data, real-time data, streaming data)
✓ **Variety:** The various types of data (Structured, semi-structured, unstructured)
✓ **Veracity:**(Data inconsistency, ambiguities, deception)

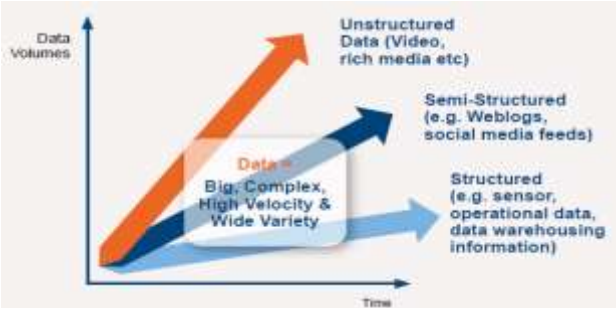

Fig.2 variety of data

Structured Data: (RDBMS[ERP,CRM], Data Warehouse, Microsoft Project Plan File) 10%
Defined as length and format. Examples of structured data include numbers, dates, and groups of words and numbers called *strings* (for example, a customer's name, address, and so on).likely SQL Example - Input data, Click-stream data, Gaming-related data etc..,
Semi Structured Data:(XML)10%
Unstructured Data:(Video, Audio, Text Message, Blogs, Weather Pattern, Location Coordinates, Web Logs & Dickstreams, Sensor Data/M2m,Email,SocialMedia,Geospatial Data) 80%
Unstructured data is data that **undefined as length and format.**
Now a day's 80% of data in the format of unstructured data and 20% of data in the format of structured data.

Why Big Data Need:

- Increase of storage capacities
- Increase of processing power
- Availability of data

2.2 BIG DATA TECHNOLOGIES
Apaches's Hadoop File System

Hadoop is a java based free software framework. Hadoop is written in the Java, operating system is cross platform and speculators are APACHE software foundation. Hadoop consists of the Hadoop Common package(contains the necessary Java Archive (JAR) files and scripts needed to start Hadoop)which provides file system and OS level abstractions, a Map Reduce engine (either MapReduce or YARN) and the Hadoop Distributed File System (HDFS).The Hadoop framework transparently provides reliability to applications. Hadoop has two major components. They are Hadoop File System (HDFS) and Map-Reduce *Hadoop Distributed File System (HDFS)*The Hadoop File System (HDFS) are highly scalable, distributed, and portable which is written in Java for the Hadoop framework. TCP/IP protocols are used to Hadoop distributed file systems. User to communicate between each other via Remote Procedure Call (RPC).

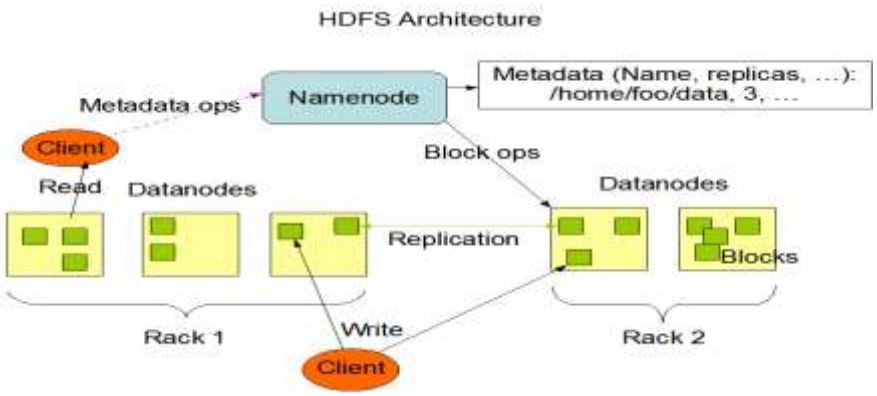

Fig.3 HDFS Architecture

HDFS has master/slave architecture. An HDFS cluster consists of a single master node and number of slave node, usually one per node in the cluster. Initially, a document is split and distributes big data across many nodes in the cluster. The Name Node is responsible for opening a file, closing a file, rename a file. The Data Nodes is providing a user request from read /write operation. Data nodes store all the files as replicated blocks and retrieve them whenever required. HDFS stores files in a replicated manner after breaking the file into fixed size blocks. Block size is 64MB and data node maintains three copies of a data.

The Name Node and Data Node are typically run a *limitations Of HDFS (Hadoop Distributed File System)* HDFS was designed for mostly endurable files and may not be suitable for concurrent write-operations. It is inconvenient to executing a job in HDFS file system. GNU/Linux operating system (OS). HDFS stores large files (gigabytes to peta bytes) across multiple machines.

2.3 MAPREDUCE

Map Reduce framework architecture provides a parallel processing to the huge amount of data. With Map Reduce, queries are fragmented and distributed across parallel nodes and processed (the Map step). Syndicates them to form an output (the Reduce step). Map Reduce programs run on Hadoop in languages like java, python etc., Map Reduce can take advantage of processing data paralleling.

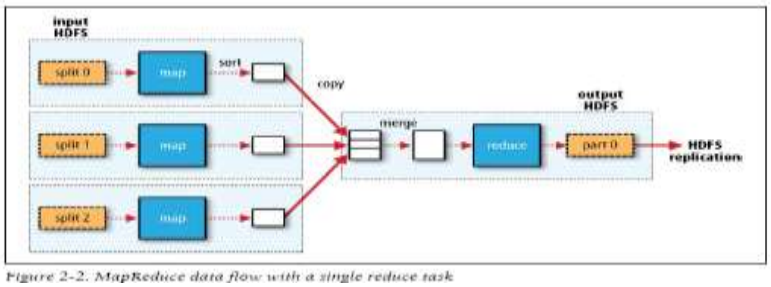

Figure 2-2. MapReduce data flow with a single reduce task

Fig.4 Map reduce

"Map" step

Mapper is applied in parallel on input data. User given the input (k1,v1) pairs from HDFS and produces a list of intermediate (k2,v2) pairs. Mapper output is partitioned per reducer.i.e. the number of reduce tasks for that job.

"Reduce" step

It collects all answers from data node and form an output. Reducer takes (k2,list(v2)) values as input, make sum of the values in list(v2) and produce new pairs (k3,v3) as final result.

Map and Reduce function

The "*Reduce"* Map Reduce is useful in applications like distributed pattern-based searching, function is then functional in parallel to each group and delivers a collection of values in the same domain. Thus the Map Reduce framework transforms the list of values.

Uses of Map reduce

Map reduce usefull in application like distributed pattern-based searching document clustering, machine learning and statistical machine translation. Besides that, Map Reduce model has been revised to numerous computing environments like cloud environments and mobile environments. Map Reduces inputs and outputs are usually stored in a distributed file system.

III. MACHINE LEARNING

Machine Learning is the study of Computer algorithms that improve automatically through experience and make predictions on data. Machine Learning is a computer program that is said to learn from experience E with respect to any class of tasks T and performance measure P, if its performance at the task improves with the experiences (Mitchell 1997). Machine learning algorithms are either classified as supervised or un-supervised depending upon the labeling of data. Supervised learning is based on the training data with pre-defined label where the class label of each instance is known in advance. Classification is an example for supervised systems that learns the given example with the class label and assigns a correct class label for unknown instances. On the other hand, unsupervised learning is based on unlabeled training data. The patterns are grouped based on similarity and termed as clustering.

3.1Machine Learning Algorithms

Many supervised machine learning techniques were developed for document classification (Sebastiani 2002), some of the renowned classifiers are Regression model, k-Nearest Neighbor (k-NN), Decision Tree classifier, Naive Bayes (NB), Support Vector Machines (SVM) and Neural Networks. The un-supervised machine learning algorithm is a Clustering that generates groups or clusters of related documents and the similarity among them.

31.1 k-NearestNeighbor (k-NN)

k-NN is employed to perform tests on the degree of likeliness between test documents and k training data. Here certain classification data are stored to determine the category of test documents(Ko & Seo 2000). This method is designated as an instant-based learning algorithm which categorizes the data objects based on the proximity of the feature space in the training set. Here training data are represented in a multi-dimensional feature space. This feature space in turn divided into regions based on the class of the training data. A data point in the feature space is mapped to a particular class if it is the most frequent classes among the k nearest training data. The distance between the data points are often calculated using the Euclidean Distance measure.

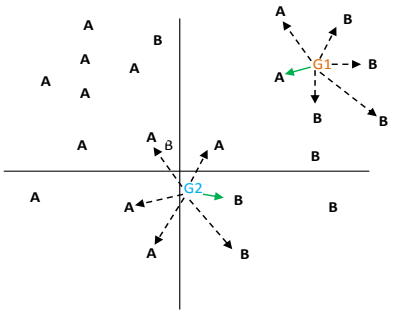

Fig.5 k-Nearest Neighbors of Classes G1 and G2

4.1.2 Naive Bayes (NB) Classifier

Naive Bayes classification is a probabilistic approach that does not require more instances for all possible combinations of attributes. Here, every attribute of interest assumed to be independent of each other. One of the important assumptions behind the technique is that the influence of an attribute is always independent of other attributes for a given class and this assumption is called as class conditional independence (McCallum& Nigam 1998). The joint probability of document features is calculated based on the probability that a new document fit in a specific class is defined in given below Equation

$$P(c_i|d') = \frac{P(d'|C_i).P(C_i)}{\sum_{cj \in C} P(d'|C_i).P(C_j)}$$

3.1.3 Genetic Algorithm (GA)

Genetic algorithm proposed by Holland (1992) performs better on hybrid problems dealing with both discrete as well as continuous data and on combinatorial problems. Here, three important processing elements such as selection operation, crossover operation and mutation operations are used to generate various learning models. GA is best suited for the problem domain where optimal solution is needed, but the possibilities of algorithm are restricted to stick at local optimal solutions. The computational requirements of learning system are not fulfilled by GA, hence it may not be concerned when there is strong computing power is demanded.

4.1.4 Neural Network Based Classi □ er

A Neural Network (NN) text classifier is a network of processing units called neurons. Each input units represent words of the document, the output unit(s) represent the class or classes of interest, and the weights on the link that connects the processing units denotes dependence relations (Miguel Ruiz & Padmini Srinivasan 1998). To classify a given testing document d_j, the word weights wk_j are assigned to the visible units, where wk_j denotes the k^{th} term in document d_j. The activations of these input units are propagated forward through the network and the value of the output unit(s) determines the prediction decision.

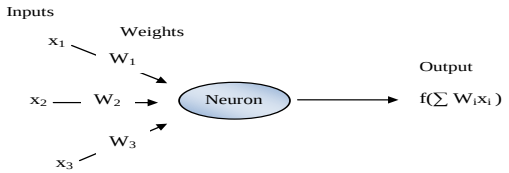

Fig.6 Structures of Artificial Neural Networks

3.1.5 Deep Learning Approach

According to Hinton(2007), Deep Learning is an advanced class of machine learning techniques that utilizes several layers of non-linear information processing for both supervised and unsupervised methods of feature extraction and transformation. It is extensively used for applications such as pattern analysis and classification.

Table 1. Comparison of Different Learning Algorithms

Approaches	Advantages	Limitations
Naive Bayes	Simple and quick classification	low accuracy
k- Nearest Neighbor	Cost of learning process is zero. Complex concepts can be learned by local approximation with simple procedure	Increase in the training data causes a decrease in efficiency, curse of dimensionality
Genetic Algorithm	Provides an optimal solution	Expects large amount of data and the algorithm often stick at local optimal solutions.
Neural Networks	Approaches an Expert's classification results	It require sufficient training data, Learning is to slow across multiple hidden layers
Deep Learner	Highly flexible to specify prior knowledge, handle large family of function parameterized with many individual parameters	Involves multiple layers with complex structures

IV. BLOCK CHAIN TECHNOLOGY

Block chain is a collection of records called *blocks*, which are connected using security. Block chains which are readable by the public are widely used by crypto currencies.

Each and every record includes plaintext and cypertext hash value of before records and intervals of records, transaction between the one block to another block. It is represented as a tree structure. By schema of a Block chain to changes of the data. It is "an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way".

Block chain is the interconnection of decentralized blocks of information. The technology thrives on peer to peer networks in order to achieve its decentralization ability. In Block chains, entries are written into a record by each peer. A number of records of information from a particular peer form a block. Each peer within the network has their own block. These blocks are interconnected to form a chain of blocks containing information. Information flows freely within these chained blocks. However, entries written into a record by each peer within the network of users has to be consented to by group. In Block chain technology, information is made readily available to all peers within a group or network.

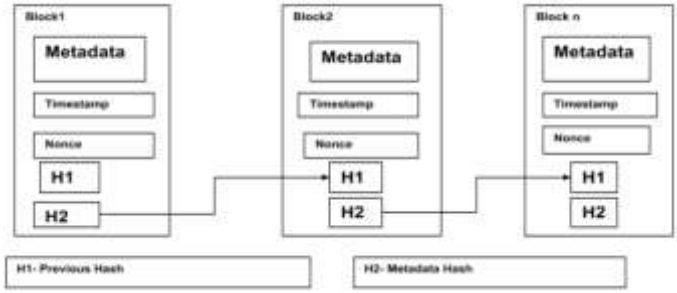

Fig.7 Blocks in the block chain

They then use specific protocols to determine whether an information amendment or update should or not occur. The technology derives its strength from 3 other technologies. They are Peer to Peer Network, Public Key Cryptography and the Block chain Protocol.

Peer to Peer Network: Peer to Peer Technology drives the authorization and decentralization ability of the Block chain Technology. Peers reach a consensus and decide on particular data updates or amendments. No one peer can effect change to information without the approval of others.

Public Key Cryptography (PuKC): The involvement of PuKC in the block chain technology ensures a secure digital identity. Using the associated private and public keys, a digital signature depicting strong sense of ownership could be created and hence a secure digital identity. In Public Key cryptography, a user that wishes to communicate sends a message along with its public key to a peer.

The receiving peer receives the message and uses their private key to decrypt and retrieve the message. This form of securing information provides high authentication access. A feature embedded in Block chain. The authorization and authentication process involved in Block chain makes it a force to reckon with in recent times.

Block chain Protocol: This protocol determines the underlying rules within which block chain operates i.e. broadcasting digitally signed information to all nodes/peers in a network at a given time. The nodes involved agree on the information update and each node/block gets a copy of the updated information hence no single point of failure. The major property of block chain ensuring security and overall effectiveness of the technology lies with decentralization /shared controls.

However, incorporating Block chain databases in Machine learning means having a

- shared data
- bigger and safer data
- much better machine learning models

Uses:

Smart contracts, Banks, sales, tracking digital use and payments to content creators, such as wireless users or musicians

V. CONCLUSION

Big Data technologies can be viewed as a new generation of technologies and architectures designed to extract value economically from very large volume of data by enabling high velocity capture, discovery and analysis. Various machine learning models that yield groundbreaking throughputs with high efficiency rates in predicting, detecting, classifying, discovering and acquiring in-depth knowledge. BigData and machine learning are produced in application such as search engines, business informatics, social networks, social media, metrology and weather forecast. The paper summarizes briefly big data, machine learning and block chain technology. The relevance of these technologies and how closely they relate with one another is further discussed citing major applications which makes use of these technologies together. The aim of this paper is to encourage further research in incorporating Block Chain Technology into Machine Learning.

REFERENCES

[1]. "IBM What is big data? —Bringing big data to the enterprise". www.ibm.com. Retrieved 2013-08-26.
[2]. George, Lars (September 20, 2011) "HBase: The Definitive Guide (1st ed.)", O'Reilly Media. p. 556. ISBN 978-1449396107
[3]. Kon stantin Shvachko et.al. "The Hadoop Distributed File System, Yahoo! "Sunnyvale, California USA, 978-1-4244-7153-9, 2010.
[4]. T. White, Hadoop: The Definitive Guide. O'Reilly Media, Yahoo! Press, June 5, 2009.
[5]. http://hadoop.apache.org/releases.pdf.
[6]. Bijesh Dhyani et.al. "Big Data Analytics using Hadoop", International Journal of Computer Applications (0975 – 8887), 108(12):1-5, 2014.

[7]. Sebastiani,F 2002, 'Machine Learning in Automated Text Categorization', in ACM Computing Surveys Archive vol. 34 , issue 1, pp. 1 – 47.
[8]. Sebastiani, F 2005, 'Text Categorization, in A. Zanasi ed., Text Mining and its Applications to Intelligence', CRM and Knowledge Management, pp. 109-129, WIT Press, Southampton, UK.
[9]. Selamat, A& Omatu, S 2004,'Web Page Feature Selection and Classification using Neural Networks', Information Sciences, vol.158, pp.69–88.
[10]. Sofia & Bulgaria Nyberg, K 2011, 'Document Classification Using Machine Learning and Ontologies'.
[11]. S. Athmaja, M. Hanumanthappa, and V. Kavitha. A survey of machine learning algorithms for big data analytics. In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pages 1–4,March 2017.
[12]. Nolan Bauerle. How does blockchain technology work? Available at:[url = https://www.coindesk.com/information/how-doesblockchain- technology-work/,, 2018. Accessed Feb 2018].
[13]. Carla E Brodley and Mark A Friedl. Identifying mislabeled training data. Journal of artificial intelligence research,11:131–167, 1999.
[14]. C. Cachin. Blockchains and consensus protocols: Snake oil warning. In 2017 13th European Dependable Computing Conference (EDCC), pages 1–2, Sept 2017.
[15]. Michael Crosby, Pradan Pattanayak, Sanjeev Verma, and Vignesh Kalyanaraman. Blockchain technology: Beyond bitcoin. Applied Innovation, 2:6–10, 2016.
[16]. W. Meng, E. Tischhauser, Q. Wang, Y. Wang, and J. Han. When intrusion detection meets blockchain technology: A review. IEEE Access, PP(99):1–1, 2018.
[17]. James Nechvatal. Public-key cryptography. Technical report,NATIONAL COMPUTER SYSTEMS LAB GAITHERSBURG MD, 1991.

Author's Profile

Name : Mr.M.Vengateshwaran

Designation : Assistant Professor, Department of CSE
Arasu Engineering College, Kumbakonam

Qualification: B.E., M.E.,

Specialization: Bigdata, Datamining, IR, Database, SE etc.,

Name : Ms. A. Barkathunisha **(B.E.-CSE)**

College : Arasu Engineering College, Kumbakonam

Specialization: BigData

Name : Ms. A.Vanisri **(B.E.-CSE)**

College : Arasu Engineering College, Kumbakonam

Specialization: BigData

Name **:** Ms. R.Vijaya **(B.E.-CSE)**

College **:** Arasu Engineering College, Kumbakonam

Specialization: BigData