Review on-Exploring the Limitations and Challenges of Large Scale Cloud Computing

K. Sivakumar
Assistant Professor
Department of Computer Science and Engineering
JCT College of Engineering and Technology,
Coimbatore, India.

M. Rupa
Assistant Professor
Department of Computer Science and Engineering
JCT College of Engineering and Technology,
Coimbatore, India.

Abstract: **Cloud Computing is a creative platform and popular system in big data. Big Data is accomplished through the principle of virtualization. This paper deals with the time consumption in large-scale data storage. Security, data format and data processing problems occur while sharing large scale of data. To overcome all these issues Map Reduce and Hadoop technology is used. Map Reduce is based on the Modified Hilbert Curve (MHC) Algorithm, which helps to reduce the execution time. Addressing of Big Data is a challenging task and requires large computational infrastructure. Modules like Replication, Fault Tolerance and Data Encryption are used. This paper also discusses about the definition, characteristics, and classification of Big Data storage using large-scale Computing and research challenges.**

Keywords: Cloud Computing, Big Data, Hadoop Technology, Modified Hilbert Curve (MHC) Algorithm, Map Reduce, Replication, Fault Tolerance, Data Encryption.

I. INTRODUCTION

Big Data is a new paradigm for next-generation analytics development, enabling large-scale data computing, sharing and exploration of large volumes. Data using Cloud Computing technologies like large-scale data and service-oriented computational infrastructure facility. Large scale Data Computing is another worldview which consolidates large-scale computing with new data-intensive techniques and scientific models to construct data investigation for intrinsic data extraction. Large scale data computing is developed as service-oriented computing model to convey infrastructure platform and applications as administrations from the suppliers to the consumers meeting the quality of services (QOS) parameters, by empowering the reported and processing of huge volumes of rapidly developing data at a faster scale. Big Data demands large data computing and data resources and clouds offer large-scale infrastructure, hence both these technologies could be integrated.

The proposed research work deals with the challenges in integration of both these technologies. Big Data is a powerful metaphor for the administration of large-scale data computing in adaptable computing and store infrastructures. The proposed work examines an architectural system for Big Data computing in clouds that support large-scale distributed data-intensive applications. Date Aware Scheduling model for effectively scheduling the jobs gets the data from remote distributed storage utilizing transformative genetic approach, composed by Hadoop Distributed File System (HDFS) and Map Reduce. The proposed research work will demonstrate their sufficiency by performing scheduling experiments in both simulation and real-time environments utilizing Hadoop clusters.

1.1 BIG DATA

Big Data refers to the extension of the volume of data that are difficult to store, process, and analyze. The difficulty can be identified with data capture, storage, sharing and visualization [1].

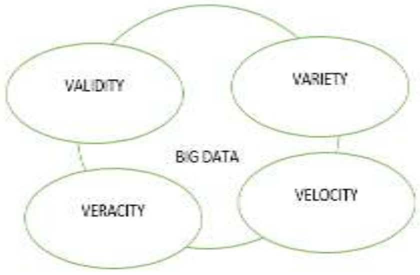

Fig 1 Big Data Characteristics

Big Data is characterized by four dimensions:

a. Data Variety
b. Data Validity
c. Data Veracity
d. Data Velocity

Data Variety: Variety refers to the different kinds of data accumulated by sensors and smart phones. Such types of data include video, image, text, and audio [2]. It reaches beyond the organized data and unorganized data [1].

Data Validity: It alludes to data authenticity. Description due to correctness or accuracy of data used to extract result in the form of information [2].

Data Veracity: Different types of data arrived from different sources by means of different platforms [2].

Data Velocity: The transfer rate of data between source and destination [2]. It ought to be utilized when streaming into the enterprise in order to maximize its value to the business and the part of the time is very critical here.

1.2 METHODOLOGIES:

1.2.1. Hadoop Distributed File System (HDFS):

Hadoop is an efficient, reliable distributed platform that gives versatile storage and computing on large data. It offers a system to process a large volume of data by running a large number of jobs in parallel. Because of its features of scalability, reliability, cross-platform support, it has been generally adopted by an industry. Hadoop comprises of two modules are, Map Reduce and Hadoop Distributed File System (HDFS) [3].

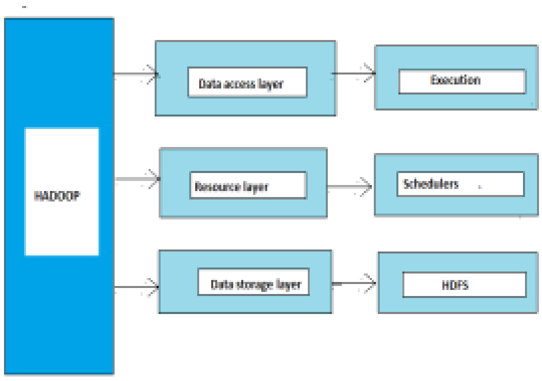

Fig 2 Hadoop Architecture

Figure 2: shows the fundamental parts of the Hadoop. The data storage layer is HDFS and it is the platform for distributed data storage for the Hadoop ecosystem. HDFS's essential obligation is to characterize how data is written to and read from a distributed system is made out of a few thousand servers. Each server has resources such as memory, IO, and CPU. The resource layer's obligation is to assign, schedule, and manage these resources on a per-task basis. The data access layer is important for executing client requests on the cluster. These requests could be in the form of SQL queries or languages such as scalability [4].

HDFS properties:

a) *High Availability*: Provides mission-critical work procedure and applications.

b) *Fault Tolerance*: Automatically and impeccably recuperate from failures.

c) *Tunable Replication*: Many duplicates of each document provide data protection and computational execution.

1.2.2 Map Reduce:

One of the most generally utilized distributed computing systems is Map Reduce. Map Reduce enables developers to perform difficult computations basically while stowing away the details of Data Distribution, parallelization, and Fault Tolerance; it examines both organized and unorganized data. Map Reduce observation can be processed individually. Map Reduce was first created by Google for large-scale data processing. Executes user jobs specified as two Functions:

- Map Function
- Reduce Function

Map Function:

A basic function is utilized to create key pairs in parallel similar to utilize primary keys in the relational database world [5].

Reduce Function:

The fundamental concept of Map-Reduce removes many conventional challenges in HPC to accomplish fault tolerance and availability. Therefore, it makes the way for the improvement of highly parallel, highly reliable and distributed applications on large [5].

Map Reduce properties:

a) *Resource Manager*: Employs data locality and server resources to decide ideal computing operations.

b*) Optimized Scheduling*: According to prioritization the jobs are completed.

c) *High Availability*: Process fails independently and restarts automatically.

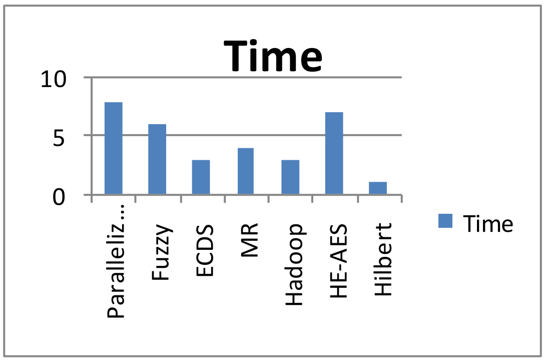

Fig 3 Performance Analysis

Fig 3 depicts a graphical representation of algorithms based on the performance analysis. Hilbert Curve Algorithm requires low computation time compared to all other algorithms. The worst case for computation is parallelization algorithm.

1.3 ISSUES WITH BIG DATA

The present large scale computing design has a few issues yet to be solved:

* Basic DBMS has not been tailored to Cloud Computing.

* Data Acts is a serious issue so it would be ideal to have Data Centers located closer the user than the provider.

* Data Replication must be carefully done else it affects data integrity and gives an error-prone analysis.

* Trust in the event of mission-critical data.
* Some deployment models are still in their formative stage.

1.4 ISSUES IDENTIFICATION

Searching for sequences in a large unorganized dataset can be both time-consuming and costly. Sequence alignment algorithms are frequently used to align many sequences. Because of memory limitation, aligning more than three to four sequences is frequently not allowed by conventional alignment tools. As expected, a Hadoop cluster with three nodes is able to search the sequence data much speedier than a single node. It is expected that search time will reduce as the number of Data Nodes are expanded in the cluster. In any case, when we execute a Map Reduce job in the similar cluster for more than one time, each time it takes the same amount of time. This study aims to present this problem and propose a solution that would enhance the time involved in the execution of Map Reduce jobs. Asymmetric encryption is one of the best techniques to secure data in storage (6), while data encryption requires more time for execution. Utilizing of novel approach we can reduce the time limitation in Big Data (8). Authentication is an intense method to finding access to storage data (7).

Table1 Discussion of Big Data Algorithm

Ref. No	Title	Author	Algorithms	Year of Publication	Merits	Demerits	Applications
1.	Cloud Burst: Highly Sensitive read mapping with Map Reduce [8]	Michael C.Schatz	Model for Parallelizing Algorithm	2009	Better Balance across the virtual machine	More Execution Time	Highly Scalable
2.	Secure Algorithm for Cloud Computing and its Application [`9]	Akshita Bhandari	Hybrid Encryption-AES	2016	Low-Cost	Security Issues	Minimize the Memory Size
3.	Authentication and Encryption in Cloud Computing [10]	Sunil Kumar	Elliptic Curve Digital Signature Algorithm	2015	Shorter key length.	Cannot be used for signatures with message recovery	Security purposes
4.	Challenges for Map Reduce In Big Data [11]	Michael Hayes	Map Reduce	2014	scalability	High Latency	Image Analysis
5.	Securing Big storage: Present and Feature [12]	Kavitha Ammayappan	Hadoop	2017	Inexpensive	Slow processing speed	Open source
6.	A new approach to improve load balancing for increasing fault tolerance and decreasing energy consumption in Cloud Computing [13]	Ali Moghtadaeipour	Fuzzy Method	2016	Reliability	complicated	Load Balancing
7.	Enhancing confidentiality and privacy of outsourced spatial data [14]	Ibrahim Kamel	Hilbert curve	2016	Privacy protection Low cost		Image processing

1.5 PROPOSED WORK ON BIG DATA

The proposed for this problem is Map Reduce based Modified Hilbert Curve Algorithm, which finds out the queried block and uses the same one for future jobs required for different users or nodes which are match with this job. This algorithm first searches for the availability of the previous job, if not available it takes from the direct block in the clusters. It avoids the extra time for search, and avoids the load to the same block of the cluster, and reduces the reachable time.

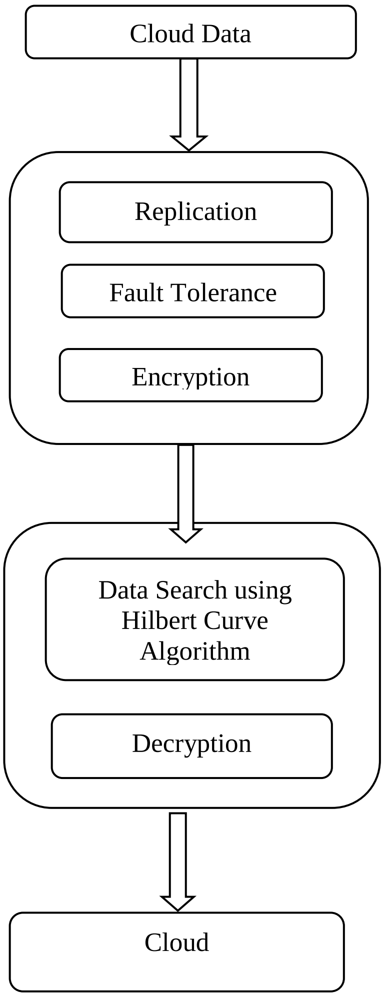

Fig 4 Process of large scale data computing

1.6 MODULES USED

1. Fault Tolerance
2. Replication
3. Data Encryption
4. Data Search using Hilbert Curve
5. Data Decryption

Fault Tolerance: The failure causes in large- scale data computing are Data Failures and Hardware Failure. Data Failures is Data Corruption. Hardware failures are storage access exception and different defects in the data. Proactive Fault Tolerance technique avoids recovery from failures and predicts the failures [9]. Reactive fault tolerance reduces the effect of failures on application execution when the failures effectively occurs [10].

Replication: Replication is the process of copying data changes from one database into another database. The two databases are the most part situated on a different physical server. With the objective of reducing this time, data replication is utilized to confer greater availability of the files required most generally. The control of the number of copies of a file is carried out by replication factor, and this informs the file system of quantity copies that must be maintained available for a given file. The larger the replication factor utilized, the greater the availability of the file [7]. Hadoop Distributed File System (HDFS) technology is utilized. This strategy reduces the execution time and satisfies the necessity for parallelism. Adaptive replication

algorithms such as Latest Access Largest Weight (LALW) and Pop Store are presented with simple static replication. Static Replication is the most simple and routinely utilized replication approach in Cloud Computing. The quantity of replicas is pre-configured before the system starts and the system replicates the static number of data whenever the data is stored. The popular file systems in an exhibit cloud environment are HDFS and GFS [8].

Data Encryption: Encryption process generates cipher-text that can only be viewed in its original form if decrypted with the correct key. Decrypted data with two fundamental kinds of encryption keys, Symmetric-key or asymmetric-key. Rivest- Shamir Adleman and Advanced Encryption Standard are public key view on cryptography implementation. AES merits are it requires less time and less memory.

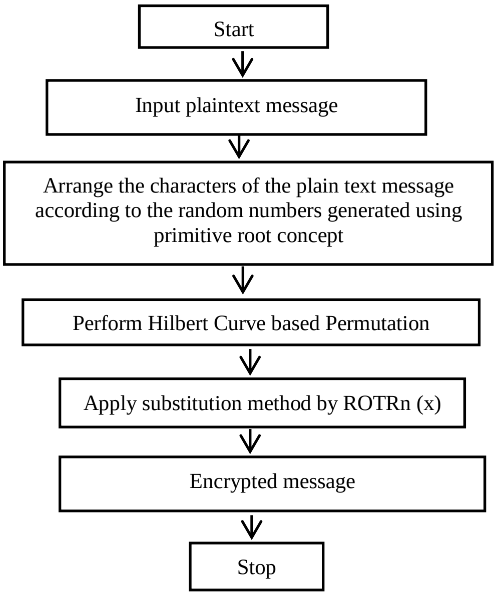
Start
Input plaintext message
Arrange the characters of the plain text message according to the random numbers generated using primitive root concept
Perform Hilbert Curve based Permutation
Apply substitution method by ROTRn (x)
Encrypted message
Stop

Fig 5 Overall working model for Encryption method

At sender side, the plaintexts are arranged column-wise is a N×N matrix, N value is chosen as 4 or 8. For arranging the plaintext characters the permutation key is from the primitive root concept. Then the matrix characters are given to Hilbert curve for further permutation to get the permuted text message. ROTRn(x) is a circular right shift of the argument 'x' by 'n'. If the plaintext message contains more than 64 characters then the entire process is repeated for the remaining characters.

Hilbert Curve: Hilbert curve is also known as Hilbert space-filling curve. Hilbert curve is easy to generate when applied over a digitized photography. Based on the current research in the field, the Hilbert curve does not take the distribution of spatial point into consideration when transforming the original data.

1.7 ADVANTAGES OF BIG DATA

- Improving privacy
- Improving security
- Optimizing Business Process
- Improving science and research

1.8 LIMITATION OF BIG DATA

- Insecure data
- Service Reliability issues
- Cost

1.9 CHALLENGES OF BIG DATA

- Data storage and Quality
- Security of the data
- Privacy of the data
- Data access and sharing information
- Technical Challenges
  Fault Tolerance
  Replication

3. CONCLUSION

In this paper, the various big data characteristics and process in large scale data computing are analyzed. The failures in big data are hardware and software failure. This paper depicts different methodologies to handle large datasets. In this review, Fault could be analyzed by various modules such as Fault Tolerance, Replication, and Encryption. This paper tends that the execution time is reduced while exchanging large-scale data. Besides, the key issues in Big Data were discussed.

ACKNOWLEDEMENT

I express my sincere thanks to Dr.N.S.Nithya, Research guide, associate professor, Department of Computer Science and Engineering for her valuable advice and guidance. I also express my heartful thanks to Dr.G.Rajiv Suresh Kumar, Head of the Department Dept. of Computer Science and Engineering, for his cooperation. I would also like to thank staff of our library- reference section, and my friends for their support. Finally, and most of all, I would like to thank my parents for their eternal support and encouragement.

REFERENCES:

[1] Sachchidanand Singh and Nirmala Singh, "Big Data Analytics," *International Conference on Communication, Information & Computing Technology (ICCICT)*, pp. 1-4, Oct. 2012.

[2] GayatriKapil, Alka Agrawal, R. A. Khan, "A study of Big-Data Characteristics,"*International Conference on Communication and Electronics Systems (ICCES)*, pp. 1-4, Oct. 2016.

[3] Madhvaraj M Shetty and Manjaiah D.H "Data Security in Hadoop Distributed File System," *IEEE International Conference on Emerging Technological Trends (ICETT)*, pp. 1-5, Oct. 2016.

[4] Scott Shaw, "Hadoop Technology," *Published onInternet of Things and Data Analytics*, Handbook by HwaiyuGeng, pp. 383- 397, 2017.

[5] Khalid Adam Ismail Hammad and Jasni Mohamed Zain, "Big-Data Analysis and Storage," *Proc. of International Conference on Operations Excellence.*

[6] Gabriel Heleno, MaristelaHolandae and AleteiaAraujo, "Data Replication Policy in a Cloud Computing-Environment," *IEEE 11th Iberian Conference on Information System and Technologies*, pp. 1-6, Jun. 2016.

[7] Julia Myint and Axel Hunger, "Comparative Analysis File Replication Algorithms for Cloud Data Storage," Future Internet of Things and Cloud (FiCloud), *IEEE International Conference on Future Internet of Things and Cloud (FiCloud)*, pp. 115-123, Aug. 2014.

[8] Jiankun Hu, Athanasios, "Energy Big Data Analytics and Security Challenges and Opportunities*," IEEE Trans. on Smart Grid*, Vol. 7, no.5, pp. 2423-2436, Sept. 2016.

[9] Salma M.A. Ataallah, Salwa M. Nassar, Elsayed E. Hemayed,"Fault Tolerance in Cloud Computing–Survey,"*IEEE 11th International Computer Engineering Conference (ICENCO)*, pp. 241-245, Dec. 2015

[10] Pankaj Deep Kaur and KanuPriya, "Fault Tolerance Techniques and Architectures in Cloud Computing-A Comparative Analysis,"*IEEE International Conference on Green Computing and Internet of Things (ICGCIOT)*, pp. 1090-1095, Oct. 2015.

AUTHORS :

1. K.Sivakumar, Assistant Professor, Department of Computer Science and Engineering, JCT College of Engineering and Technology, Coimbatore, India.

2. M.Rupa, Assistant Professor, Department of Computer Science and Engineering, JCT College of Engineering and Technology, Coimbatore, India.