Survey On Big Data Analytcis using Hadoop ETL

M.Saranya[1] A.Prema[2]

[1]MPhil Scholar, Department of Computer Science, Raja Doraisingam Govt Arts College, Sivagangai.
[2] Assistant Professor, Department of Computer Science ,Raja DoraiSingam Govt Arts College, Sivagangai

Abstract

The term big data refers to data sets whose volume, variability and speed of velocity make them difficult to capture, manage, procedure or analyzed. To examine this huge amount of data Hadoop is able to be used. Hadoop is an open source software project that enables the spread giving out of large data sets across a cluster of creation servers.ETL tools extract important information from various data sources, various transformation's of data are established out transformation phase and then load into the big data. HDFS (Hadoop Distributed File System), is a spread file system design to hold the very huge of data (petabytes or even zettabytes), and there high throughput admission to this information. Map Reduce method has been calculated in this paper which is required for implement Big Data Analysis using HDFS. In this paper the related topics of Big Data Analytics, and Hadoop, ETL, Map Reduce are reviewed.

Key Words: *Big Data, Hadoop, ETL, Map Reduce, HDFS*

I. INTRODUCTION :

Varsha B.Bobad describes "Big Data" is a collected works of large data sets that cannot be processed using usual computer techniques. Big Data is not just a data somewhat it has become a whole subject which involves an assortment of the tools, technique, and framework. the needed of big data generated by the large corporation like facebook, Yahoo, Google, Youtube etc for the purpose of analysis of huge amount of data also Google contains a large amount of information[1].

Shruti Tekadpande et al says Hadoop is an open source, Java-based encoding framework that supports the processing and storage of very large data sets in a distributed computer location HDFS stores three copies of each piece to three servers. Sizes of each building block have 64 MB. In this paper, we explore the new opportunity of utilizing Hadoop for performing business brain with a specifically ETL phase of Big Data [2].

Explicitly convey the multi-dimensionality of big Data when adding that "the data is too big, moves too fast, or doesn't fit the structure of your database architectures", this extract allows us to see that extra uniqueness should be added to large datasets to be careful as Big Data, or Big Size data as often found all through the literature[30]

big Data Analytics facilitate you to recognize the information contained within the data, but it will also help recognize the information that is most significant to the business and future business result. Big Data analytics basically want data that comes from examining the data [3]

Big Data Analytics the procedure of analyzing and mining Big Data-Can create ready and business knowledge at an unparalleled scale and specificity. The need to analyze and force trend data collected by industry is one of the main drivers for Big Data analysis tools. The technical advances in storage space, processing, and analysis of Big Data comprise the rapidly decreasing cost of storage and CPU control in recent years; the flexibility and cost effectiveness of data centers and cloud computing for elastic working out and storage and the progress of new frameworks such as Hadoop, which allows user to take gain of these spread computing systems storing large quantities of data through bendable parallel processing. These advances have created several differences between fixed analytics and Big Data analytics [4].

The technology and growing amount of data(Big Data), need is felt towards implement effective analytics technique (Big Data Analytics) to analyze this big volume of data for unknown and useful facts, patterns, Big data is the term for data sets so large and difficult that it becomes difficult to process using usual data management tools or allowance applications. This paper reveals most recent advancement on big data networking and big data [5].

The unstructured data that is so vast that it's difficult to process using usual database and software techniques. In most undertaking scenario the data is too large or it moves too fast or it exceeds current dispensation capacity. Big data has the likely to help a group to improve operations and make faster, smarter decisions [6]. The challenges include investigation, imprison, curation, search, distribution, storage, transfer, apparition, and privacy violation. The learning of larger data sets is due to the extra information derivable from analysis of single allowing correlation to be found to" spot business trends, prevent disease, combat crime and so on. "Vishal S Patil,Pravin et al [7] .Big Data can be characterized by well-known 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data have to be processed. yet though big data doesn't pass on to any complete quantity, the term is often used when speaking about petabytes and exabytes of data, a great deal of which cannot be integrated easily. [8].

The Complexity of today's data comes from manifold sources, and it is still a responsibility to link, match, rinse and change data across systems. However, it is necessary to connect and correlate relationships, hierarchies and many data linkage or your data can quickly out of control [9].

Here we have some operation queries, model, and building algorithm to locate new insights. Mining requires included, cleaned, truthful data: at the same time, data mining itself can also be old to help recover the quality and reliability of the data, understand its semantics, and provide bright querying functions [29]

The quality of the data may be its high allowance level or its relevance according to the reality they stand for. In fact, since Big Data is big and messy, challenges can be classified into engineering tasks (managing data at an unimaginable height) and semantics (finding identified each relevant piece of data in big data:

1. The significant data addition challenge which can be seen as a five –step challenge: (1)define the problem to solve, (2) identify related pieces of data in Big Data,(3) ETL it into fitting format and store it for processing,(4) disambiguate it and (5) solve the problem.

II.RELATED WORKS

The second common objective of big skill and solutions is time reduction. Macy's merchandise price optimization request provides a typical example of reducing the cycle time for compound and large scale logical calculation from hours or even days to minutes or seconds [20]

The section store sequence has been talented to reduce the time to optimize pricing of its 73 million items for scale from over 27 hours to now over 1 hour. Described by some as "big data analytics,"this potential set hastily makes it probable for Macy's to re-price items much more frequently to adapt to change conditions in the retail marketplace. This big data analytics application takes data out of a Hadoop cluster and puts it into other parallel computing and in-memory software architectures [21].

This" big data structural propose and pattern" series presents a structured and pattern-based approach to simplify the task of defining overall big data architecture [34].

We can present the design and evaluation of a data ware cache framework that requires a minimum change to the original Map Reduce programming model for provisioning incremental processing for big data application using the Map Reduce model [20].

The author stated the importance of some of the technologies that handle big data like Hadoop, HDFS and Map Reduce. The author suggested about various schedulers used in Hadoop and technical aspects of Hadoop. The author also focuses on the importance of YARN which overcomes the limitations of Map Reduce [21]

The author continue with the bigdata definition and given in that includes the FiveV big data properties: Volume, Variety, Velocity, Value, Veracity and suggest that other dimensions for big data analysis and taxonomy, in particular comparing and contrasting big data technologies in e-science, industry, business, social media, healthcare [22].

The author is given some important emerging framework model design for big data analytics and a 3-tier architecture model is big data in data mining. In the proposed 3-tier architecture model is more scalable in working with different environment and also benefits to overcome the main issue in big data analytics for storing, analyzing, and visualization.

The framework model is given for Hadoop HDFS distributed data storage, real-time NoSQL database, and Map Reduce distributed data processing over a cluster of commodity servers [23].

The author state there is a need to maximize returns on BI investment and to overcome difficulties. Problems and new trends mentioned in this article and finding solutions by the combination of advanced tools, techniques and methods would help readers in BI projects and implementations. BI vendors are struggling and doing continuous effort to bring technical capabilities and to provide complete out of the box solutions with the set of tools and techniques. In 2014, due to rapid change in BI teams are facing a tough time to have the infrastructure with less skilled resources. Consolidations and convergence are going on; the market is coming up with wide range of new technologies. Still, the ground is immature and in a state of rapid evolution [24].

The author describes the concept of big data along with 3 Vs, Volume, Velocity and a variety of big data. The paper also focuses on big data processing problems and technical challenges must be addressed for efficient and fast processing of big data. The challenges include not just the clear issues of scale, but also heterogeneity, be deficient in of structure, error from data acquisition to result from interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. The paper describe Hadoop which is an open source software used for processing of big data [25]

With a long tradition of working with the constantly increasing volume of data, modern e-Science can offer industry the scientific analysis methods, while industry can bring advanced and fast developing big data technologies and tools to science and the wider public. [26]

Shvaiko and Euzenat mention the lack of evaluation of scalability as a challenge. Likewise, all these remarks could be made ours, after we have a present main aspect of big data semantic management. Surely, all the techniques and tools aforesaid can be improved by various parameters or heuristics, but in big data era, a significant place must be made to optimization. Tools must handle exabytes of data, streaming data, fast changing ones, very informal data etc. [27]

S.Vikram Phaneendra et al. illustrated that, in olden days the data was less and easily handled by RDBMS but recently it is difficult to handle huge data through RDBMS tools, which is preferred as ”Big data ”, In this paper they told that big data differs from other data in five dimensions such as volume, velocity, variety, value, and complexity. They illustrate the Hadoop architecture consisting of name node, data node, edge node, HDFS to handle big data system. Hadoop structural design switches large data sets, the scalable algorithm does log organization application of big data can be found out in a financial retail industry, healthcare, mobility,and insurance. The author also focused on the challenges that need to be faced by enterprises when handling big data: -data privacy, search analysis, etc [28].

The author stated learning from the application studies, we explore the design space for supporting data-intensive application on a large data-center-scale computer system.Traditional data processing and storage approaches are facing many challenges in meeting the continuously increasing computing demands of big data. This work focused on map reduce, one of the keys enabling approaches for meeting Big Data demands by means of highly parallel processing on a large number of commodity nodes.[33]

Dhole Poonam B et al Hadoop is open source software that enables reliable, scalable, distributed computing on clusters of low-priced servers [10]. Hadoop shows MAD characteristics. The 'M' stands for magnetic i.e. it can store all kind of data sources and attracts them towards itself. The "A" refer to the quickness as various operations on big data easily can be easily performed on it. The 'D' stands for Deep. It is capable of performing ad-hoc and complex analytics over the big data, Hadoop provides desired result song.Y.[11]

The Hadoop Distributed File System (HDFS) is a distributed file system planned to run on creation hardware [12] HDFS is designed to be deployed on low-cost hardware. It is highly faulted tolerant. HDFS is suitable for application that has large data sets T.K.Das [13]. HDFS is planned and optimized to store data in excess of a large amount of low-cost hardware in a distributed manner.

- Hadoop Common – contains libraries and utilities need by other Hadoop modules
- Hadoop Distributed File System(HDFS)- a distributed File system with the intention of provisions data on product machines, providing very high combined bandwidth across the cluster
- Hadoop Map Reduce – a programming model for large scale data dealing out.

Cost Effective: Hadoop saves cost as it employs a cheaper low-end cluster of commodity of machines instead of the costlier high-end server. Also, distributed storage of data and transfer of computing code rather than data saves high transfer costs for large datasets [35].

- Efficiency due to its inability to switch to the next stage before completing the previous stage tasks causing Hadoop unsuitable for pipeline parallelism, runtime, scheduling that causes degraded efficiency per node. Unlike RDBMS, it has no specific optimization of execution plans that could minimize the transfer of data among various nodes.
- Optimization [36], it does not random reads on small files [37].

III. Extract-Transforms-Load (ETL)

Expansion of big data involves the ETL process. It is a complex combination of process and technology. This system consists of three functional entities: Extract, Transforms, and Load.

Extract function extracts relevant information from sources data for decision making which then needed to be altered into the different schema to much the big data schema. The final function loads the data into big data [14]. Kuldeep Deshpande, et al The big data platform is categorized on the basis of their operating system, hardware server, and storage system. By allowing for the over cloud and Hadoop base data warehousing platform are the, most reliable to meet today's requirement.ETL process also needs to be developed on the same platform to gain the maximum profit [15] organizers have listed out their difficulty in order to process petabyte of data daily. They have demand the requirement developed the well-organized ETL, capable of handling all kinds of data. Yongqiang He et al [16]

Gregory S. Nelson et al. explained the methodology used to design the target database structure and transformations, create a mapping worksheet used to implement the ETL code, load the metadata, and create the process flows in Data Integration (DI) Studio. The paper further connects the dots for those interested in getting started with DI Studio not only as a tool but also how practitioners think about the DI Studio process [32]

IV. MAP REDUCE FRAMEWORK

Map reduce is the key algorithm that the Hadoop Map reduce engine uses to allocate work around a cluster. Typical Hadoop cluster integrates Map reduce and HFDS layer. In Map reduce layer job tracker assigns tasks to the task tracker. Master node job tracker also assigns tasks to the slave node task tracker output to a temporary storage. A master node orchestrates that for redundant copies of input data, only one is processed Vibhavari Chavan. [17]

To process a large amount of data one relevant programming pattern is Map Reduce. In this model, Map function process key/value pair to generate intermediate key/value pairs. Reduce function later get collectively these values to produce the output file J.Deen et al [18,19].

Chris Jermaine et al. proposes an online aggregation for large –scale computing. Given the potential for OLA to be newly relevant, and given the current interest and very large –scale, data-oriented computing; In this paper, they considered the problem of providing OLA in a shared-nothing environment. While they reflect on implementing OLA on top of the Map Reduce engine, many of author's most basic project contributions were not specific to Map Reduce and should apply broadly. Consider how online aggregation can be built into a Map Reduce system for large- scale data processing. Given the Map Reduce paradigm's close association with cloud online aggregation is a very attractive technology. Since large-scale cloud computations are typically pay-as-you-go, a user can monitor the accuracy obtained in an online fashion, and then save money by killing the computation early once sufficient accuracy has been obtained [31].

V. CONCLUSION

Big Data is a set of the great dataset that can't be processed using conventional figure techniques. Big Data is not merely a data quite it has become a total subject which engages various tools, techniques, and framework. In this paper, we reviewed some related papers of Hadoop, ETL, and Map Reduce techniques.

REFERENCES

[1] Varsha B.Bobad , "International Research Journal of Engineering and Technology (IRJET)", Volume: 03 Issue: 01

[2] Inmon, William "Data Mart Does Not Equal Data Warehouse". DM Review.com. (2000-07-18).

[3] Jeffrey R. Bocarsly, "The Data Warehouse Toolkit." Complex ETL Testing-A Strategic Approach

[4] R. Kimball and M. Ross. WileyPublishing, Inc., 2002.

[5] "Survey of Recent Research Progress and Issues in Big Data" www.cse.wustl.edu/~jain/cse570-13/ftp/bigdata2/index.html 1/13

[6] Dhole Poonam B, Gunjal Baisa L, "Survey Paper on Traditional Hadoop and Pipelined Map Reduce" International Journal of Computational Engineering Research‖Vol, 03 Issue, 12

[7] Varsha B.Bobade" Survey Paper on Big Data and Hadoop" International Research Journal of Engineering and Technology

[8] V. Bhanumurthy*, G Behera "Deliverable from space Data sets for Disaster Management-present and true trends"

[9] Ms. Vibhavari Chavan, Prof. Rajesh. N. Phursule, ―"Survey Paper on Big Data"‖ International Journal of Computer Science and Information Technologies, Vol. 5 (6), 2014.

[10] Amogh Pramod Kulkarni, Mahesh Khandewal, ―"Survey on Hadoop and Introduction to YARN"‖, International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 5, May 2014)

[11] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), ―"Big Data Framework‖" 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499

[12] Inmon, William (2000-07-18). "Data Mart Does Not Equal Data Warehouse". DMReview.com

[13] Katarina Grolinger, Miriam A.M. Capretz." Knowledge as a Service Framework for Disaster Data Management"

[14] V. Hristidis, S. Chen, T. Li, S. Luis, and Y. Deng, "Survey of Data Management and Analysis in Disaster Situations," Journal of Systems and Software, vol. 83, no. 10, pp. 1701-1714, 2010.

[15] Song .Y, Davis Karen C," Analytics over large scale Multidimensional Data: The Big Data Revolution, Communications of ACM," 2011

[16] Merinela Mircea," Business Intelligence--Solution for Business Development", Intech Publisher, 2012

[17] Kuldeep deshpande, and dr. Bhimappa desai,"limitations of dataware house platforms and Assessment of hadoop as an alternative," Volume 5, Issue 2, pp. 51-58, IJITMIS ,2014

[18] Yongqiang He et al RCFile: "A Fast and Space-efficient Data Placement Structure in MapReduce-based Warehouse Systems," ICDE, 2011

[19] Vibhavari Chavan et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939

[20] Franklin, M.,Halevy, A., Maier, D., 2005. From databases to data spaces:"A new Abstraction For Information Management. ACM SIGMOD Record" 34 (4), 27–33.

[21] Saleem, K., Luis, S., Deng, Y., Chen, S.-C., Hristidis, V., Li, T., 2008. "Towards a business Continuity information network for rapid disaster recovery." In: Proceedings of the 9th Annual International Conference on Digital Government Research, Montreal, Canada, May 18–21, pp. 107–116

[22] Senthi Vadivel Bhupatthi Rav" Disaster Management: A Global Issue" International journal of civil and structural engineering Volume 1, No 1, 2010

[23] Sagiroglu, S.Sinanc, D.,‖Big Data: A Review‖,2013, 20-24.

[24] Tekiner F. and Keane J.A., Systems, Man and Cybernetics (SMC), ―"Big Data Framework"‖ 2013 IEEE International Conference on 13–16 Oct. 2013, 1494–1499

[25] S.Vikram Phaneendra & E.Madhusudhan Reddy "Big Data-solutions for RDBMS problems- A survey" In 12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010) (Osaka, Japan, Apr 19{23 2013).