www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 9 September 2017, Page No. 22463-22468
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i9.12

Review Paper on Big Data Tools and Techniques

Gurvinder Singh[1], Anurag Rana [2], Joyti[3]

[1]Arni University Dept. of Computer Science Engineering
Kangra Himachal Pradesh, India
[2]Arni University Dept. of Computer Science Engineering
Kangra Himachal Pradesh, India
[3]Arni University Dept. of Computer Science Engineering
Kangra Himachal Pradesh, India

Abstract:
In this paper we discussed about most emerging field of computer science engineering name as Big Data. This paper provides you detailed information about big data from basic to advance level. In this paper we describe about massive amount of structured, unstructured and semi structured because big data related to gathering and analysis of huge amount of data and analysis of data in this paper we also discussed about tools such as (Hadoop and HIVE) used in big data. This paper provides you deep insights about big data and various area of research available in this field.

Keywords: Big Data, Unstructured Data, Structured Data, Hadoop

1. Introduction

With the enhancement in devices and communication technologies, set of data are growing at a very fast manner. This increasing data is gathered by cheap and varied information sensing mobile devices, aerial (remote sensing), cameras, microphones and other sensor network. For handling such large amount of data. Relational database management systems often have difficulty. The huge storage requires "massively parallel software [1] running on tens, hundreds or even thousands of servers". This was achieved with the help of distributed computing [2]. In distributed computing system, various components are located on a network and they communicate with each other by passing message to each other. It is also referred as a system used to solve computational problems in a distributed manner. Generally, in this a problem is divided into many task, each of which is solved by several computers and these computers in a network communicate with each other by passing message. Also various software and hardware architecture are used for distributed computing, multiple CPU are connected at a lower level with some sort of network. At another higher level, it is important to do the interconnection of the process that run on the various CPUs with some specific communication system. Figure 1 show basic component big data. This figures sources of the big data from these are common sources of the big data. Now a days social media and online shopping website generate huge of data per day which require very good tools for the handling this huge amount of data for various purpose such as predictive analysis and decision support system.

Figure 1: Sources of Big Data

collection of large data sets [3] that cannot be processed by utilizing traditional computing techniques. Today many organizations are collecting, storing and analyzing massive amount of data, the amount of data produced by mankind is growing at a very fast manner. As per recent survey on big data, the amount of data produced by mankind from beginning of time till 2000, if we roughly count was just 5 billion Gigabytes, but the same amount was created in every two days in 2011, every minute in 2013, and every seconds in 2016, the rate of data is growing enormously day by day. The large amount of data sets is called as Big data, it involves bulk of data that is unstructured and such kind of data need much more real-time analysis [4]. Data stored in traditional ware houses is quite different from Big data, such data stored in warehouses need to be cleaned properly, documented completely, also it should be compatible with the basic structure of the particular warehouse [5], Big data involves not only handling of data that is needed to be stored in warehouse but it also deals with the data not suitable for storing in the ware houses. Thus, better data analysis and decisions are needed for handling of Big data. New opportunities in order to discover new values are also brought by Big data, it also help us to discover new challenges that is how to manage and organize data set in an effective manner. Big data include data generated by various sources.

2. Sources of Big Data

- **Media** be existent in-and-out of your association, may associate with APIs (think an API to gather images from Pinterest and surround them into a invention page or email handling) and is temperately structured.
- **Business apps** are organised, and consuming APIs you can tug data from mutually classified and outside your society. Internally, think assimilating your CRM or Web Gratified Running with your ecommerce organization. Externally, by means of Weather Co. or Climate Subversive data for native personalization is additional example. Community mesh is outdoor, but nearly very nonchalant and useful requests can be pounded up with it. For example, is your occupational exaggerated by the daily vacillation of currency (or Bitcoin value?), or exploration term capacity that can be dragged from Google Trends.
- **Social media** is high rapidity, high bulk data that you canister use to notice trends, analyses mawkishness about your variety, customer package and opponents, or target operations to social accounts that competition the email discourses in your client file (to name a few applications).You’re likely creation good use of mechanism log data finished your Web analytics, the subsequent step is using movable or third gathering facilities that help you healthier classify, board and change visitors.
- **Sensor data** is high rapidity, capacity, and variety and dare I add worth, when second-hand properly to comprehend user background and forecast performance. Sensors for geolocation, temperature, noise, consideration, assignation, biometrics, and more can assemble reams of data that is beneficial for better acquisitions and possession experiences in a variability of industries.
- **Website Analytics:** Your eLearning course website, online forums, and eLearning blogs are other Big Data sources. You can use analytics to track site traffic, engagement, and conversion rates. This data can even reveal where your online learners originate, how long they spend on the site, and which devices they utilize. For example, Google Analytics allow you to track the time online learners visit your site, and which keywords they used to find you. Furthermore, you have the ability to view all of your Big Data in the form of pie charts, graphs, and other visual representations. Big data includes huge volume, high velocity and extensible variety of data.

The data produced from these various sources will be basically of three types

- **Unstructured data:** Such kind of data includes word file, pdf files, text, media logs. Structured Data: Structured data is very ordinary. It apprehensions all information which can be stored in database in the form of rows and columns we can say SQL in table. They have relational key and can be easily mapped into pre-designed fields. Today, individuals data’s are the most administered in progress and the unpretentious way to manage evidence’s. But designed data’s denotes only 5 to 10% of all informatics data’s.

- **Semi structured data**: Semi-structured data is information that doesn't reside in the form of the rows and tables means it is not having any pre-defined structure information is recognized by help of the Meta data. The in a relational database but that does have some organizational properties that make it easier to analyses. With some process you can store them in relation database (it could be very hard for some kind of semi structured data), but the semi structure exist to ease space, clarity or compute.
- **Unstructured data:** Unstructured data represent around 80% of data. This category of data includes multimedia and text content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents.

3. Literature Review

Assunção, Marcos D., et al [16] The amount of data currently generated by the various activities of the society has never been so big, and is being generated in an ever-increasing speed. This Big Data trend is being seen by industries as a way of obtaining advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more costumers, increase the revenue per customer, optimize its operation, and reduce its costs. Nevertheless, Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure, and effort. Cloud computing helps in alleviating these problems by providing resources on demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand.

Buyya, Rajkumar, et al. [17] in this world of technology the emergence of cloud computing has made dynamic provisioning of elastic capacity to applications on demand. Cloud data centers contain thousands of physical servers hosting orders of magnitude more virtual machines that can be allocated on demand to users in a pay-as-you-go model. However, not all systems are able to scale up by just adding more virtual machines. Therefore, it is essential, even for scalable systems, to project workloads in advance rather than using a purely reactive approach. Given the scale of modern cloud infrastructures generating real time monitoring information, along with all the information generated by operating systems and applications, this data poses the issues of volume, velocity, and variety that are addressed by Big Data approaches. In this paper, we investigate how utilization of Big Data analytics helps in enhancing the operation of cloud computing environments. We discuss diverse applications of Big Data analytics in clouds, open issues for enhancing cloud operations via Big Data analytics, and architecture for anomaly detection and prevention in clouds along with future research directions.

Dilpreet Singh et al. [1], Provides very deep study of the different hardware platforms available for big data analytics and assesses the advantages and drawbacks of each of these platforms based on various metrics such as scalability, data I/O rate, fault tolerance, real-time processing, data size supported and iterative task support. In addition to the hardware, a detailed description of the software frameworks used within each of these platforms is also discussed along with their strengths and drawbacks. Some of the critical characteristics described here can potentially aid the readers in making an informed decision about the right choice of platforms depending on their computational needs. In order to provide more insights into the effectiveness of each of the platform in the context of big data analytics, specific implementation level details of the widely used k-means clustering algorithm on various platforms are also described in the form pseudo code. A thorough comparison between different plat-forms based on some of the important characteristics (such as scalability and real-time processing) has also been made through star based ratings.

Matei Zaharia et al [2] discussed about the Map Reduce implementation success rate for data intensive application and provide deep insight of focuses on one such class of applications: those that reuse a working set of data across multiple parallel operations. This includes many iterative machine learning algorithms, as well as interactive data analysis tools. The proposed a new framework in his research finding which is called Spark that supports these applications while retaining the scalability and fault tolerance of Map Reduce. To achieve these goals, Spark introduces an abstraction called resilient distributed datasets (RDDs). An RDD is a read-only collection of objects partitioned across a set of machines that can be rebuilt if a partition is lost. Spark provides three simple data abstractions for programming clusters: resilient distributed datasets (RDDs).

Naidila Sadashiv et al [3] discussed comparison between Cluster, Grid and cloud computing all are formed with the collection of parallel or distributed computer so it is important to understand difference between all these because they are similar and related to each other. This paper provides comparison between them and helps us to understand difference between them and also challenges in all the fields.

Cluster: A cluster is a collection of parallel or distributed computer which are interconnected among them using high speed networks such as gigabit Ethernet, SCI, and infinite band.

Dejan S. Milojicic et al [4] discussed about the P2P systems and applications by summarizing the key concepts and giving an overview of the most important systems. Design and implementation issues of P2Psystems are analyzed in general, and then revisited for eight case studies. This survey will help people in the research community and industries understand the potential benefits of P2P. For people, unfamiliar with the field it provides a general overview, as well as detailed case studies. Comparison of P2P solutions with alternative architectures is intended for users, developers, and system administrators (IT). P2P is an important technology that has already found its way into existing products and research projects. It will remain an important solution to certain inherent problems in distributed systems. P2P is not a solution to every problem in the future of computing.

M. Ali et al [5] discussed about RSA-Grid: A grid computing based framework for power system reliability and security analysis. In this paper, a grid computing based framework is proposed for the probabilistic based power system reliability and security analysis. In this paper introduce a framework for power system reliability and security analysis because high performance computing power is needed in case of power system reliability assessment.

S. Agarwal et al [6] In this paper author proposed a RoPE (Re- optimizing Data-Parallel Computing) which is used for the re-optimization of the data-parallel jobs. RoPE uses piggybacking for the collection of the certain data properties and code on job execution and then it adapts the plan of the execution by serving these properties to a query optimizer. Accurate estimation of the property of code and data is not an easy task in distributed environment and forecasting attributes by the collection of statistics of the stored raw data is not appropriate because the commonness of user operation. The knowledge of these attributes provides us a huge space of enhancements. The absolute numbers of jobs indicate attributes are estimated dynamically. The proposed solution RoPE gathers statistics from multiple location and uses innovative way to combine the entire attribute. The elasticity permitted for the user to state random code leads to a ample snugger connection between data and computation in data parallel clusters. Future scope of proposed work develops more advance technique that selects plans We defer to future work some advanced techniques that choose plans having defined level of the validity range which is specified over given statistics and perform substituting to these plans during runtime depending on the detected statistics.

4. Properties of Big Data

Generally Big data consist of three basic properties which is also represented by V3, V3 is nothing but three attributes of big data named as volume, velocity and variety [6].

- Volume: Organization collect data from different sources including various business transactions , social media , in past storing the data was a great problem .But new technologies have eased the burden .The Big word in big data itself defines the volume .At present the existence of data is in petabytes and is assumed to increase to more than zettabytes in coming era .The social networking sites are producing data in order of terabytes and this amount of data is definitely difficult to be handled using the existing traditional systems.
- Velocity: Velocity basically deals with the rate at which data arrives from various nodes to one server. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows. For example, the data from the sensor devices would be constantly moving to the database store and this amount won’t be small enough on performing the analytics on the data which is constantly in motion

- Variety: Data that is being produced does not include only a single category of data but it also includes semi structured data in addition to traditional data from various sources. All this data is totally different that consist of raw, structured, semi-structured and even unstructured data which is difficult to be handled by existing traditional analytics systems.

Further there are also two additional dimensions of Big data, these are variability, complexity, values. Variability indicates that data can flow in an inconsistent manner with periodic peaks. Moreover, complexity in big data makes it difficult to link, match, cleanse and transform data across systems.

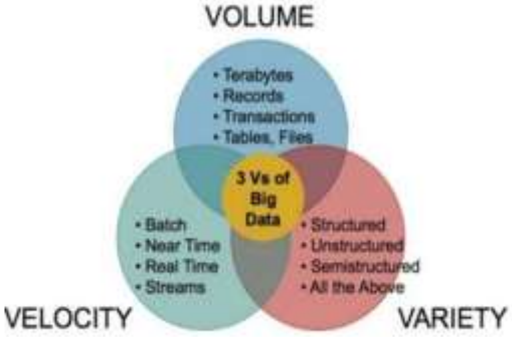

Figure 2: Properties of Big data

5. Big Data tools

- **Apache Hadoop:** Hadoop is open source framework which is based on java developed and maintained by Apache foundation. Hadoop is framework which is used for massive analysis of data and storage of data Hadoop is a java based free software framework that can effectively store large amount of data in a cluster. This framework also provides functionality of parallel processing and distributed processing with the help of HDFS and MapReduce programming model [11].
- **Microsoft HDInsight:** This is tool provide by Microsoft for Big Data solution and this is also powered by Apache Hadoop which is available as a service in the cloud. HDInsight practices Windows Azure Blob stowing as the nonappearance file system. This also delivers high obtainability with stumpy cost.
- **NoSQL:** This tool is used to handle unstructured data this is not follow any particular schema and in this each row consist on own values of the NoSQL provides improved performance in storing huge amount of data. There are many open-source NoSQL DBs available to analyse big Data.
- **HIVE**: is associated library of Hadoop which is used to distributed data management on Hadoop. HIVE supports query language which is called as HiveSQL to provide query solution on the big data. This runs on top of Hadoop.
- **Sqoop:** This is a tool that connects Hadoop with various relational databases to transfer data. This can be effectively used to transfer structured data to Hadoop or Hive.
- **PolyBase:** This the whole thing works on upper of SQL Server 2012 Parallel Data Warehouse (PDW) and is cast-off to admittance information stored in PDW. PDW is a data warehousing machine constructed for dispensation any capacity of interpersonal data and delivers incorporation with Hadoop permitting us to admittance non-relational data as well.

6. Conclusion

The biggest advantage is developer productivity, though this can come at the expense of execution speed (mostly latency) and efficiency (high throughput via brute force). First I'll point out that HiveQL is SQL, or at least a variant of SQL. And since no database vendor follows the SQL standard perfectly, they're all variants as far as I'm concerned. The Tableau (product) connector for Hive supports all of the same important functionality as any of the other connectors we offer for SQL databases. Hive does have benefits over other SQL systems implemented in databases. Hive has several interesting UDF packages and makes it easy to contribute new UDFs.

References

1. R. Chaiken, B. Jenkins, P.-A. Larson, B. Ramsey,D. Shakib, S.Weaver, and J. Zhou. Scope: easy and efficient parallel processing of

massive data sets.Proc. VLDB Endow., 1(2):1265–1276, Aug. 2008.

2. Remzi H. Arpaci Dusseau, Eric Anderson, Noah Treuhaft, David E. Culler, Joseph M. Hellerstein David Patterson, and Kathy Yelick. Cluster I/O with River: Making the fast case common.In Proceedings of the Sixth Workshop on Input/Output in Parallel and Distributed Systems(IOPADS '99), pages 1022,Atlanta, Georgia, May 1999.
3. M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014.
4. Martin Courtney, “The Larging-up of Big Data”, IEEE, Engineering &Technology,September 2012.
5. Sam Madden, “From Databases to Big Data”, IEEE, Internet Computing, May-June 2012.
6. Y. Demchenko, C. De Laat, and P. Membrey, “Defining architecture components of the big data ecosystem,” in Collaboration Technologies and Systems (CTS), 2014 International Conference on. IEEE, 2014, pp. 104–112.
7. S. L. Garfinkel, A. Juels, and R. Pappu, “Rfid privacy: An overview of problems and proposed solutions,” IEEE Security & Privacy, no. 3, pp.34–43, 2005.
8. Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money,“Big Data: Issues and Challenges Moving Forward”, IEEE, 46th Hawaii International Conference on System Sciences, 2013.
9. Yuri Demchenko, Zhiming Zhao, Paola Grosso, Adianto Wibisono, Ceesde Laat, “Addressing Big Data Challenges for Scientific Data Infrastructure”, IEEE, 4th International Conference on Cloud Computing Technology and Science, 2012.
10. F. Chang et al. Bigtable: A distributed storage system for structured data. In Proc. OSDI, pages 205{218.USENIX Association, 2006.
11. http://bigdata-madesimple.com/top-big-data-tools-used-to-store-and-analyse-data/