International Journal of Advance Research in Science and Engineering
Vol. No.5, Issue No. 03, March 2016
www.ijarse.com
IJARSE
ISSN 2319 - 8354

MINING BIG DATA: STUDY OF TOOLS OF OPEN SOURCE REVOLUTION

Ms. Palak Vaish[1], Dr. Saurabh Srivastava[2]

[1]*Research Scholar, Mewar University, Chittorgarh, Rajasthan, (India)*

[2]*Department of Mathematical Sciences and Computer Applications,*

Bundelkhand University, Jhansi, U.P (India)

ABSTRACT

Big data is a term that describes the large volume of structured and unstructured data that inundates a business on a day-to-day basis. It is a pool of large and complex data sets that are difficult to process using usual database management tools. Big Data mining is the ability of extracting useful information from huge streams of data or datasets that can be analyzed for insights that lead to better decisions and strategic business moves. The challenges include data capture, storage, search, sharing, analysis, and visualization. With this difficulty, a new platform of "big data" tools has arisen to handle sense making over large quantities of data, as in the Apache Hadoop Big Data Platform. This paper argues on Big Data, modern databases supporting it and Big Data Mining. The challenges, architecture and analytical tools of Big data and Big data mining are also taken into account.

Keyword: Architecture, Big Data, Big Data Mining, Hadoop

I. INTRODUCTION

The use of the data is rapidly changing the nature of communication, shopping, advertising, entertainment, and relationship management. Data sets grow in size in part because they increasingly gathered by widespread radio frequency identification readers, remote sensing, information-sensing mobile, software logs, microphones, cameras, and wireless sensor networks. For some organizations, facing hundreds of gigabytes of data for the first time may generate a need to reconsider data management alternatives. Gartner [1] summarizes this in their definition of Big Data in 2012 as "high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". It is described by majorly following Vs:

- **Volume**- data from business transactions, social media and information from sensor or machine-to-machine data [2].
- **Velocity**- Sensors RFID tags, and smart metering is driving the need to deal with torrents of data in near-real time.
- **Variety**- from structured, numeric data in traditional databases to unstructured video, text documents, email, audio, and financial transactions.

- **Variability**- Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data.
- **Value**- Resulting insights that for trends and patterns, difficult analysis based on graph algorithms, machine learning and statistical modeling. These analytics overtake the results of querying, reporting and business intelligence [3].

An analyzed Big Data can open new markets, deliver new business insights, and create competitive advantages. The real value of data is observed after analyzing i.e. after finding patterns, deriving meanings and making decisions. The importance of big data doesn't revolve around how much data you have, but what you do with it. Its analysis can lead to new product development, cost reductions, smart decision making, time reductions and optimized offerings.

There are many **applications of Big Data** that allow people to have better customer experiences, better services, and also be healthier, as personal data will permit to prevent and detect illness much earlier than before [4].

- Technology: reducing process time from hours to seconds
- Business: churn detection and customer personalization
- Smart cities: cities focused on high quality of life and sustainable economic development with proper management of natural resources
- Health: mining DNA of each person, to discover, examine, study, monitor and improve health aspects of every one

II. ARCHITECTURE OF BIG DATA

The big data landscape is comprised of four layers:

Infrastructure as a Service (IaaS): This includes the servers, storage, network and distributed file systems.

Platform as a Service (PaaS): This layer provides the logical model for the raw, unstructured data stored in the files.

Data as a Service (DaaS): This includes entire array of tools available for integrating with the PaaS layer using integration adapters, search engines, batch programs etc.

Big Data Business Functions as a Service (BFaaS): This layer includes packaged applications that serve a specific business need and leverage the DaaS layer for cross-cutting data functions like ecommerce, health, energy, retail, and banking.

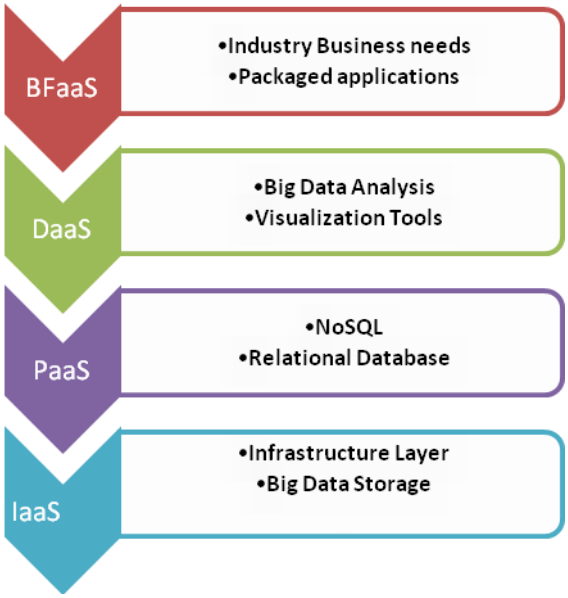

Fig. 1 Layered Architecture of Big Data

III. BIG DATA ANALYTICS IN MODERN DATABASE LANDSCAPE

With the evolving volume of data, new techniques are needed to analyze it for which new data-warehousing and technologies are a solution [5]. The Big Data phenomenon is intrinsically related to the open source software revolution. Large companies as Yahoo!, Facebook, LinkedIn, and Twitter benefit and contribute working on open source projects. Working with volumes of structured data has always been simple by using a relational database but if the volume or velocity of the data denies using a relational database, then the problem is called "Big Data Problem". Following table provides a broad perspective of this modern database landscape with a selection of the main systems in each family [6].

TABLE 1: Tools for Big Data Mining and Analytics

Relational	MySQL, Oracle, BB2, MS Access, SQL Server etc.	works with large amounts of structured data
NoSQL (Not Only SQL)	Used by web companies like Facebook, Google, Amazon, Twitter, etc.	Enable high levels of horizontal scale and high throughput by simplifying the database model, query language and safety guarantees offered. Its performance is far better than traditional relational databases.
NewSQL	It has recently emerged to strike a compromise between the features of traditional relational databases and the performance of NoSQL stores.	Suitable for executing live queries NewSQL supports SQL and uphold ACID (Atomicity, Consistency, Isolation, Durabilty) properties of Database.

Distributed Analytical Frameworks	**Hadoop, MapReduce,** They offer an alternative to databases for analytics	MapReduce framework was introduced by Google in 2004 as an abstraction for executing batch-processing tasks over large-scale at data (unindexed/raw files) in a distributed setting [4]. A MapReduce job divides the input dataset into independent subsets that are processed by map tasks in parallel. MapReduce is proprietary and kept closed by Google. It is considered as one of the core Big Data technologies.
		Apache Hadoop [7] is an open source project which offers a mature implementation of the MapReduce framework and a distributed file system called Hadoop Distributed Filesystem (HDFS). It is used for data intensive distributed applications and allows writing applications that rapidly process large amounts of data in parallel on large clusters of compute nodes. Apache Hadoop related projects [8]: Apache Pig, Apache Hive, Apache HBase, Apache ZooKeeper, Apache Cassandra, Cascading, Scribe and many others.
	Pregel/ Graph These are both MapReduce style frameworks designed specifically for abstraction of graph-based analytics as clustering, shortest	Pregel was first introduced by Google in 2009 suitable for distributed computation on very large graphs [9].
		Apache Graph is an open source implementation of MapReduce style frameworks built on top of Hadoop

	path, computing connected components, centrality measures etc., Both are a message passing system between vertices in a graph	
Object databases	Allow for storing data in the form of objects as per object oriented programming languages.	support a query language and other features, including versioning, constraints, triggers, etc.
XML databases	data are stored in an XML-like Data structure.	Store XHTML, RSS, ATOM feeds and SOAP messages etc. in a native format that enables them to be queried directly through languages such as XQuery, XPath or XSLT.
RDF databases	Provides query functionality over data represented in the core Semantic Web data model.	Use SPARQL query language
Big Data Mining Tools	R	It was designed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand beginning in 1993 and is used for statistical analysis of very large data sets [10]. It is an open source software environment and programming language designed for statistical computing and visualization.
	MOA	It is a stream data mining open source software to perform data mining in real time [11]. It started as a project of the Machine Learning group of University of Waikato, New Zealand, famous for the WEKA software [12].

	Vowpal Wabbit	It is an open source project started at Yahoo! Research and now used at Microsoft Research [13]. It is used to design a fast, scalable, useful learning algorithm from tera feature datasets. When doing linear learning, it can exceed the throughput of any single machine network interface through parallel learning.
	Apache Mahout	It is scalable data mining and machine learning open source software [14] based mainly in Hadoop. It is widely used for classification, clustering, frequent pattern mining. and collaborative filtering.
Big Graph mining	GraphLab	high-level graph-parallel system built without using MapReduce. It computes over dependent records which are stored as vertices in a large distributed data-graph [15].
	Pegasus	It is a big graph mining system built on top of MapReduce [16]. It allows finding patterns and anomalies in massive real-world graphs [17].

Big Data technologies primarily focus on the challenges of velocity and data volume. NoSQL scales by by supporting query languages more lightweight than SQL, distributing data management over multiple machines and relaxing traditional ACID guarantees. However, NewSQL aims to seek a balance by supporting ACID/SQL as per traditional relational databases while trying to compete with the performance and scale of NoSQL systems.

IV. CHALLENGES IN BIG DATA ANALYTICS AND MINING

Despite of so many advantages, there are many future important challenges in Big Data management, mining and analytics based on the nature of data and diversity [18]. There are some challenges to be met in the future [19] as-

- Big data is so big that it is very difficult to find user-friendly visualizations of photographs, infographics and essays [17].

- How the Analytics Architecture to deal with historic data and with real-time data at the same time. An interesting proposal is the Lambda architecture [20].
- Big Data mining techniques should be able to adapt with evolving time and in some cases to detect changes first.
- How to manage statistical significance accurately as it's always easy to go wrong with huge data sets and thousands of questions to answer simultaneously.
- Dealing with Big Data, the storage quantity is very relevant. There are two main approaches for it-compression we may take more time and less space but can be considered as a transformation from time to space. However, using using sampling, we are loosing information, but the gains in space may be in orders of magnitude.
- Not allow Distributed mining techniques to paralyze as to have distributed versions of some methods, a lot of research is needed with theoretical and practical analysis to provide new methods.
- Since new data is largely file based, untagged and unstructured, large quantities of useful data are getting lost. However, currently only 3% of the potentially useful data is tagged, and even less is analyzed as per The 2012 IDC study on Big Data [21].

V. CONCLUSION

Big data is going to be more diverse, larger, and faster in future as it's becoming the new Final Frontier for scientific data research and for business applications [22]. Driven by key industrial stakeholders and real-world applications, managing and mining Big Data is yet a challenging yet and compelling task. Big Data is autonomous with distributed and decentralized control, huge with heterogeneous and diverse data sources and complex and evolving in data and knowledge associations. High performance computing platforms are required to support Big Data mining, which impose systematic designs to use the full power of the Big Data. Different sources and ways of data collection often result in data with complicated conditions, such as missing/uncertain values. Privacy concerns, noise and errors can be introduced into the data, to produce altered data copies. So, developing a safe and sound information sharing protocol is a major challenge. Thus, Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains

REFERENCES

[1]. Gartner, http://www.gartner.com/it-glossary/bigdata, 2012.

[2]. Bloem, J. Doorn, M. V. Duivestein, S. Manen & Ommeren, "*Creating clarity with Big Data*", Sogeti, 2012.

[3]. Vinayak Borkar, Michael J. Carey, Chen Li, "Inside "*Big Data Management": Ogres, Onions, or Parfaits?*", EDBT/ICDT 2012 Joint Conference Berlin, Germany, 2012.

[4]. Intel. Big Thinkers on Big Data, http://www.intel.com/content/www/us/en/bigdata/ big-thinkers-on-big-data.html, 2012.

[5]. J. Dean and S. Ghemawat. MapReduce: Simpli_ed Data Processing on Large Clusters. In OSDI, pages 137{150, 2004.Russom, "Big Data Analytics", TDWI Research, 2011.

[6]. F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. ACM Trans. Comput. Syst., 26(2), 2008.

[7]. Apache Hadoop, http://hadoop.apache.org.

[8]. P. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, and G. Lapis. IBM Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Companies,Incorporated, 2011.

[9]. G. Malewicz, M. H. Austern, A. J. C. Bik, J. C. Dehnert, I. Horn, N. Leiser, and G. Czajkowski. Pregel: a system for large-scale graph processing. In SIGMOD Conference, pages 135-146, 2010.

[10]. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[11]. C. Bockermann and H. Blom. The streams Framework. Technical Report 5, TU Dortmund University, 12 2012.

[12]. A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer. MOA: Massive Online Analysis http://moa. cms.waikato.ac.nz/. Journal of Machine Learning Research (JMLR), 2010.

[13]. J. Langford. Vowpal Wabbit, http://hunch.net/~vw/, 2011.

[14]. Apache Mahout, http://mahout.apache.org.

[15]. Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. M. Hellerstein. Graphlab: A new parallel framework for machine learning. In Conferference on Uncertainty in Artificial Intelligence (UAI), Catalina Island, California, July 2010.

[16]. D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.

[17]. R. Smolan and J. Erwitt. The Human Face of Big Data. Sterling Publishing Company Incorporated, 2012.

[18]. C. Parker. Unexpected challenges in large scale machine learning. In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, BigMine '12, pages 1{6, New York, NY, USA, 2012. ACM.

[19]. V. Gopalkrishnan, D. Steier, H. Lewis, and J. Guszcza. Big data, big business: bridging the gap. In Proceedings of the 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Big- Mine '12, pages 7{11, New York, NY, USA, 2012. ACM.

[20]. N. Marz and J. Warren. Big Data: Principles and best practices of scalable realtime data systems. Manning Publications, 2013.

[21]. J. Gantz and D. Reinsel. IDC: The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. December 2012.

[22]. Kataria. M, Mittal. P, "Big Data: A Review", International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 7, July 2014, pg.106 – 110