IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 8, August 2016
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

A Study of Statistical and Analytical Approaches for Handling Big Data

K.Vishnu Vandana[1], S.Yunus Basha[2] and D.Tanuja[3]

[1] Asst.Professor, Department of CSE, Brindavan Institute of Technology and Sciences, Kurnool, Andhra Pradesh, India

[2]Asst.Professor, Department of CSE, Brindavan Institute of Technology and Sciences, Kurnool, Andhra Pradesh, India

[3]Asst.Professor, Department of CSE, Brindavan Institute of Technology and Sciences, Kurnool, Andhra Pradesh, India

Abstract

Inspite of any domain, big data is having a greater impact on the business. The amount of data that is generated every day is getting increasing. Most types of this information are being stored in digital formats for better accessibility. One of the challenges is to learn how to identify the information which is potentially useful to individual and for an enterprise. It is not just the issue of access to new data sources, but the method of implementing certain pattern recognition systems and their inter-relationships. We need analytics to discover the uncover insights that will help enterprises.

In this Paper, we study few techniques to discover hidden patterns and other useful information and identified how these techniques have greater impact in text mining.

Keywords: *Analytics, Big Data, Digital Formats, Pattern Recognition, Text Mining.*

1. Introduction

Big Data refers to datasets patterns whose storage size is beyond the ability of a typical database tools to capture, store, manage and analyze. There is no specific reason and explicit definition of how big a dataset should be in order to be considered as Big Data. IDC defines technologies of Big Data as a new generation of architectures and technologies designed to extract very large volumes of a wide variety of data by enabling high volume capture, discovery and analysis.

Wal-Mart, in 2004 has claimed to have the huge data warehouse with 500 terabytes storage which is equivalent to 50 printed collections. In 2009, eBay stored data up to 8 petabytes. After 2 years, the Yahoo data warehouse is estimating a total of 170 petabytes its business data. Since the demand of digitization, enterprises from various sectors have amassed burgeoning amounts of data in digital form, capturing trillion bytes of data about their customers, suppliers, internal and external operations. Data volume is also growing exponentially due to the explosion of data generated by a machine in the form of data records, web-log files, sensor data and from growing human engagement among the social networks.

It is highly impossible to stop the growth of data. According to an IDC study, in 2005, more than 130 exabytes of data were created and stored. The data amount moved to 1,227 exabytes in 2010 and in 2015 it was increased by 45.2% to 7,910 exabytes. The continuous growth of data constitutes the “Big Data” technological phenomenon brought about by the rapid data growth and parallel advancements in technology that have given a choice to rise a demand on ecosystem of hardware and software products that are enabling users to analyze data and produce new and more granular levels of insight.

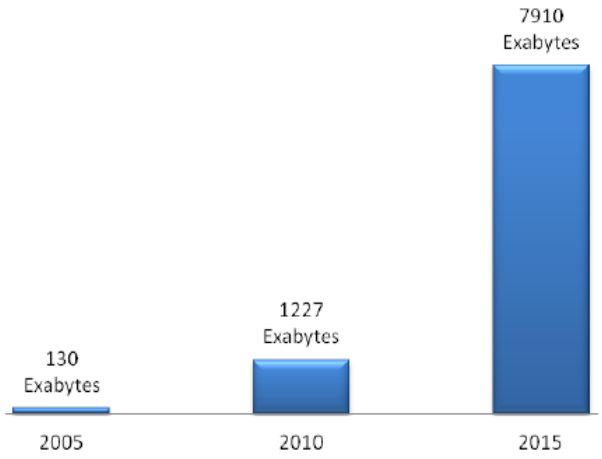


Fig.1 Data Storage in Exabytes in Universe

To obtain value from big data, you need a cohesive set of solutions for data capturing, processing, and analyzing, from acquiring and discovering new insights of data to

make decisions repeatedly and scaling the associated information systems.

The new Big Data era is bringing with it an immediate need of advanced data acquisition, storage management, and analysis mechanisms. The challenge of data integration from social media and other unstructured data into a traditional business intelligence environment became an urgent need today. Organizations are trying to develop analytic platforms that can synthesize both traditional structured and semi-structured data.

2. Preliminary Study

The convergence across enterprises has ushered in a new system that is redefining relationships among producers, distributors, goods and services and consumers. In this increasingly complex world, business domains are intertwined and what happens in one vertical means has a direct impact on others. Within a business organization, this complexity makes it difficult to rely solely on experience and to make decisions. They need to rely on good data services for their business decisions. By placing data for business operations and to provide access to new insights, they will then be able to compete effectively.

Three major things have come together to drive attention on Big Data:

1. Combine and interrogate Big Data technologies have matured to a point where their deployments are fully practical.
2. The underlying infrastructure cost to power the analysis has dramatically fallen, making it economic to mine the data.
3. The competitive pressure on business has increased to the point where most traditional strategies are offering only few marginal benefits.

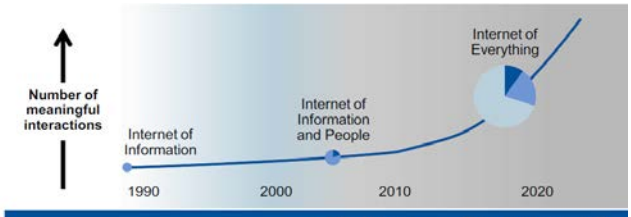
For years, business organizations have captured transactional structured data. The analysis of such data is retrospective and the datasets investigated are on patterns of business operations. In recent years, new big data technologies with lower costs have shown improvements in capturing data, storage and analysis. Organizations can capture more data from many more sources like social media and from blogs, feeds, audio and video file types. The options to store and process the data optimally have expanded dramatically and effective technologies such as MapReduce and in-memory computing provide optimized capabilities at a higher rate for different purposes. The data analysis can be done in real time on full datasets rather than summarized one. Meanwhile, the number of options to interpret and data analysis has increased, with the use of various visualization technologies also.

2.1 Proliferation of IoT and Big Data

According to Cisco's IBSG, and expectation of 50 billion devices will be connected to the Web by 2020. Meanwhile, Gartner reported that by 2010 more than 65 billion devices were connected to the internet. By 2020 this number will go up to more than 230 billion. Irrespective of the difference in an estimation, these connected devices, ranging from smart meters to a wide variety range of sensors and actuators which continually send out data in a huge amount that need to be stored and analyzed. Business organizations that deploy sensor networks will need to adopt relevant Big Data technologies to process the data in large amount sent by these networks.

2.2 Open Source Initiatives

Many of the Big Data technologies within the ecosystem have an open source origin due to participation and sharing by providers in an open source development projects. The Hadoop framework is the core of many Big Data issues today. The viability of these open source tools had driven vendors a chance to launch their own versions and integrate them with their products.



Key milestones	> 50% of Internet connections are things	Things create more traffic than information and people
Number of permanently Internet-connected devices	> 15 billion	> 30 billion
Number of intermittently Internet-connected devices	> 50 billion	> 200 billion
Volume of traffic from things	Minority	Majority

Fig.2 Exponential growth in increased number of connected computers

3. Big Data Analysis in Research and Development

Research and Development that involves data-intensive workloads, find value in Big Data technologies. Big Data technologies have to support such R&D efforts the growth of digital content and enable more efficient analysis outputs. Traditional technologies such as symmetric multiprocessing which enable system

scalability can be expensive for many granular use case scenarios. Hence, cost-efficient scalable hardware and software resources are needed to process business logics and data volumes becomes more apparent.

In this section, we study few analytical methods which can be implemented to produce needed patterns from the huge data sets

3.1 Text Analytics

Text analytics, a process to derive information from various text sources. These sources are forms of semi-structured data. The text analytics technology comes from basic fields like linguistics, statistics and machine learning. In general, modern text analytics uses statistical models, to capture human language patterns such that machines can "understand" the text meanings and perform various text analytics tasks. These tasks can be simple as entity extraction for fact extraction or concept analysis.

Entity Extraction: It identifies a product, a person or any individual piece of information.
Fact Extraction: A fact is a statement about something that exists, has happened and known in generally. It is defined by an entities collection and extraction of facts, to identify a role, or a relationship.
Concept Extraction: A functions which is used to identify an event, process, or behavior.

3.2 In-Memory Analytics

In-Memory Analytics is a layer of analytics in which data is loaded into the system memory collected from different varieties of data sources, directly for effective querying and to perform calculation. This approach partly removes the need to build metadata in systematic relational aggregations and pre-calculated cubes forms.

The use of IM processing as a back-end resource for business can improve analytics performance. On the traditional disk-based analytics, the metadata need to be created before performing the actual analytics process. The way of modeling the metadata is dependent on the analytics requirements. The way of change to model the metadata and to fulfill the new requirements requires a higher level of technical knowledge. IM Analytics removes the need to pre-model this metadata for every end user. Meanwhile, the developers need not to consider every possible issue of analysis. The analytics content relevance is also improved as data can be analyzed the moment when it is stored in the memory. The delivered speed by IM analytics makes it possible to improve interactive visualization of data sets, and makes data access more exploratory experience.

Fig.3 In-Memory Analytics overview

3.3 Predictive Analytics

PA is a set of analytical and statistical techniques which are used to identify patterns and their relationships within larger data volumes that is used to predict events. PA analytics may mine information and patterns in structured and unstructured data sets and data streams to anticipate future outcomes. The real value of this analytics is to provide predictive support that moves beyond traditional reactive break and fix assistance towards proactive support system by preventing service related impacting events from occurring.

Gartner, in his research evaluated three methods within the marketplace for predicting technical issues internal product support space. Gartner's research believes that the mature predictive support services will fully use all the combinations of those three approaches.

Pattern-based technological approach used to compare real-time performance of the system and configuration of data with unstructured data sources which includes known profiles, historic failure records and data regarding customer configuration data. Most effective correlation engines use statistical patterns with huge, multifaceted repositories to determine customer's present configuration and performance data failures.

Rules-based statistical analysis approach of historic data performance, previously identified failure modes and testing the results of stress or load and is used to define a series of rules that is compared with. Each rule

sometimes interrogates multiple data points and other external factors against defined thresholds. These rules may then be collated and can be triggered, depending on the severity and resultant issues impact.

Control-based models chart theory has been a quality of manufacturing processes has proven an invaluable aid of managing systems which are complex process-driven. The advent of real-time telemetry, retrofit-capable, data improvements in acquisition solutions and network capacity supports large volumes of data. Now statistical techniques within the manufacturing space can now be used to industrialize delivery of information technology service.

The statistical anomalies can be identified readily and are used to initiate preventive action and ensures that performance of service is unaffected and the business can be functionally continued as normal.

3.4 Cloud Based Data Analytics

Software-as-a-Service owned, delivered and managed remotely by one or more service providers. A common code with a single set is provided in an application and can be used by customers at any time. Software service based business analytics makes user to deploy one or more of the prime components of business analytics quickly without significant technology involvement need to deploy and maintain an on-premise solution.

Analytic applications support performance management with pre existing functionality for specific solutions. Business analytics platforms provide an environment to develop, integrate, delivery, and analysis of information. Information management infrastructure provides the data architecture and its integration infrastructure.

4. Conclusion

Organizations are trying to make sense of the massive flow of big data, and to develop analytical platforms that can synthesize structured data with semi-structured and unstructured sources traditional information. Big Data can provide insights which are unique into market trends and enabling more targeted business decisions at a lower cost. It is possible when such data is properly captured and analyzed.

References

[1] Chaiken R. et. al.: SCOPE: easy and efficient parallel processing of massive data sets. PVLDB 1(2), 2008.

[2] Dean, J., Ghemawat, S.: MapReduce: a flexible data processing tool. Communications of the ACM 53(1): 72-77 (2010).

[3] The Apache Hadoop Project.http://hadoop.apache.org/core/, 2009.

[4] S. Das, Y. Sismanis, K. Beyer, R. Gemulla, P. Haas, and J. McPherson. Ricardo: Integrating R and Hadoop. In *SIGMOD*, 2010.

[5] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, and C. Welton.Mad skills: New analysis practices for big data. *PVLDB*,2(2):1481– 1492, 2009.

[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. In *OSDI*, pages 205–218, 2006.

[7] K.Vishnu Vandana, S.Yunus Basha, G.Pratiba Priyadarshini – "Exploiting and Gaining New Insights for Big Data Analysis", International Journal of Engineering Research & Techology, Vol 4, Issue 01, January 2015.

[8] Savitha K, Vijaya MS – "Mining of Web Servers Logs in a Distributed Cluster Using Big Data Technologies", International Journal of Advanced Computer Science and Applications, Vol. 5, No. 1, 2014.

[9] Sherin A, Dr S Uma, Saranya K, Saranya Vanim – "Survey On Big Data Mining Platforms, Algorithms and Challenges", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 5 No. 09 Sep 2014,p.p.no854–861.

[10] Han Hu, Yonggang Wen, Tat-Seng Chua, Xuelong Li – "Toward Scalable Systems for Big Data Analytics: A Technology Tutorial", `Vol 2, May 12,2014,p.p.no652-681.

[11] Chanchal Yadav, Shuliang Wang , Manoj Kumar – Algorithm and approaches to handle large Data - A Survey", International Journal of Computer Science and Network, Vol 2, Issue 3, 2013.

[12] T.K.Das1 , P.Mohan Kumar – "BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013, p.p.no 153-157.

[13] Zheng Zhao, Russell Albright, James Cox, and Alicia Bieringer – "Big Data Meets Text Mining", Vol 4, 2013.