Data Management and Big Data Text Analytics

Zaheeruddin Ahmed
School of Engineering and IT
Manipal University Dubai Campus, Dubai, U.A.E
zaheeruddin@manipaldubai.com

----------ABSTRACT----------

Big data is now one of the most important technology trends that have the potential for changing the way organizations transform massive amounts of data into knowledge. It is a combination of data-management technologies that have evolved over time. It enables organizations to store, manage, and manipulate vast amounts of data at the right speed and at the right time to gain the right insights. On the other hand big data analytics is the process of examining large amounts of data in an effort to uncover hidden patterns or unknown correlations [22]. Big data comprises of Structured and Unstructured data. In this paper we highlight some of the efficient way of storing unstructured data and appropriate approach of analyzing it. Unstructured data is largely produces from social network sites, email conversations which are in the form of text and graphics. This paper will discuss some of the ways to extract and retrieve efficient way of handling large amount of unstructured data using current technologies. We highlight some of the challenges and techniques related to big data analytics with respect to text analytics.

Keywords – Big data, Unstructured data, Text analytics, Hadoop

----

I. INTRODUCTION

The advent of contemporary computer systems in recent times has greatly increased the degree and intricacy of the decision process. Nowadays, these systems ensure that it requires the analysis of gigabytes or even terabytes of data. This clearly indicates how much data is stored on today's computer systems. While computers are exceptionally useful in accumulating and storing large volume of data, they are of limited help in searching and analyzing the same[3]. This is because computers at their core are still mathematical devices, and are better at analyzing structured data than unstructured data (documents, e-mails, or multimedia files) which makes up the huge volume of data as a whole. Furthermore the complicating aspect is analyzing different kinds of unstructured data, as unstructured data itself is divided into two broad group textual and non-textual data.

Textual data includes all the documents, e-mails, text files and chat files and non-textual data, on the other hand, comprises graphics, sound and movies. Not even this say if a presentation document contains textual or non-textual data, and sound or graphics are embedded in a word document or email, such complexity raises different problems for attempting to analyze them. The effective solution for problems associated with analysis of both structured and unstructured data is Big Data Analytics[5].

We have particularly identified the need of further work on practices and tools for dealing with text analytics in unstructured data To address these issues, this paper presents a model for collecting and combining heterogeneous data into integrated data models, in a fashion that allows us to work with various levels of data (structured, semi-structured, unstructured) with potentially restricted data access [6]. The data collection, refinement and integration model is based on an iterative process that is actively driven by an analyst. Instead of a purely data-driven approach, the process is rooted in the idea that, in the spirit of interactive computing, the analyst can actively participate in the data collection and refinement process, while the process can still retain repeatability and transparency [11].

II. BIG DATA

The new innovation in data management can be viewed as advancement in the technologies like hardware, storage, networking, and computing models such as virtualization and cloud computing. The convergence of such emerging technologies and reduction in costs for everything from storage to compute cycles have transformed the data landscape and made new opportunities possible. As all these technology factors are transforming the way we manage and leverage data. Big data is the latest trend to emerge because of these factors.

Big data has become important because it enables organizations to gather, store, manage, and manipulate vast amount of data at the right speed, at the right time, to gain the right insights. Organizations today are at a tipping point in data management. So far the technology was designed to support a specific business need, now that the organizations have more data from more sources than ever before. These organizations are facing some technology challenges like:

- How to deal with data when it is difficult to recognize the patterns that are the most meaningful for your business decisions.
- How to deal with massive amounts of data in a meaningful way.

So what is Big Data?

Big Data is information that can't be processed or analyzed using traditional processes or tools. Big Data is the capability to manage a huge volume of disparate data, at the right speed, and within the right time frame to allow real-time analysis and reaction. Big data is a term used to

describe the exponential growth and availability of data, both structured and unstructured.

Every day, 2.5 quintillion bytes of data is created, so much that 90% of the data in the world today has been created in the last two years alone.

This data comes from like: Sensors used to gather climate information, Posts to social media sites, Digital pictures and videos, Purchase transaction records, Cell phone GPS signals .

Big data is used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques. In most enterprise scenarios the data is too big or it moves too fast or it exceeds current processing capacity. Big data has the potential to help companies improve operations and make faster, more intelligent decisions [5].

Characteristics of Big Data:
Three Characteristics defined big data. Volume, Velocity and Variety.

Volume: Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

Velocity: Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

Variety: Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

III. BIG DATA CHALLENGES

Big data offers incredible insight. But with terabytes and petabytes of data pouring in to organizations today, traditional architectures and infrastructures are not up to the challenge. IT teams are burdened with ever-growing requests for data. It takes hours or days to get answers to questions, if at all. More users are expecting self-service. This initiates the question: How do you present big data that users can quickly understand and use. Mining millions of rows of data creates difficult task for analysts with sorting and presenting data. Organizations often approach the problem in one of two ways: Build “samples” so that it is easier to both analyze and present the data, or create template charts and graphs that can accept certain types of information. Both approaches miss the potential for big data. Instead, consider pairing big data with visual analytics so that you use all the data and receive automated help in selecting the best ways to present the data. This frees staff to deploy insights from data. Think of your data as a great, but messy, story. Visual analytics is the master filmmaker and the gifted editor who bring the story to life.

To fully take advantage of data will need to address several challenges related to big data. Here are some outlined challenges.

Speed: Meeting the need for speed. Companies not only have to find and analyze the relevant data they need, they must find it quickly. The challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.

Understanding the data: It takes a lot of understanding to get data in the right shape. If the data comes from social media content, you need to know who the user is in a general sense –such as a customer using a particular set of products – and understand what it is you’re trying to visualize out of the data.

Quality: The value of data for decision-making purposes will be at risk if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data, it becomes even more pronounced.

Meaningful results: Graphical analysis becomes difficult when dealing with extremely large amounts of information. One way to resolve this is to cluster data into a higher-level view where smaller groups of data become visible.

IV. BIG DATA ANALYTICS

Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights [1]. It is the process of examining large data sets containing a variety of data types to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information.

The analytical findings can lead to more effective opportunities, better customer service, improved operational efficiency, competitive advantages and other business benefits for organizations [1].

The primary goal of big data analytics is to help make informed decisions by enabling analytics professionals to analyze large volumes of transaction data, as well as other forms of data.

The data that comes from Web server logs and Internet click stream data, social media content and social network

activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things.

Big data is associated with semi-structured and unstructured data. Big data can be analyzed with the software tools commonly used as part of advanced analytics disciplines such as predictive analytics, data mining, text analytics and statistical analysis. But the semi-structured and unstructured data may not fit well in traditional data warehouses based on relational databases. Furthermore, data warehouses may not be able to handle the processing demands posed by sets of big data that need to be updated frequently or even continually -- for example, real-time data on the performance of mobile applications or of oil and gas pipelines. As a result, many organizations looking to collect, process and analyze big data have turned to a newer class of technologies that includes Hadoop and related tools such as YARN, MapReduce, Spark, Hive and Pig as well as NoSQL databases. Those technologies form the core of an open source software framework that supports the processing of large and diverse data sets across clustered systems.

V. UNSTRUCTURED BIG DATA

Unstructured data usually refers to information that doesn't reside in a traditional row-column database. As you might expect, it's the opposite of structured data -- the data stored in fields in a database.

Unstructured data files often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, web pages and many other kinds of business documents. While these sorts of files may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database.

Experts estimate that 80 to 90 percent of the data in any organization is unstructured. And the amount of unstructured data in enterprises is growing significantly -- often many times faster than structured databases are growing.

Unstructured Data Management: Organizations use of variety of different software tools to help them organize and manage unstructured data. These can include the following:

Big data tools: Software like Hadoop can process stores of both unstructured and structured data that are extremely large, very complex and changing rapidly[9] .

Business intelligence software: Also known as BI, this is a broad category of analytics, data mining, dashboards and reporting tools that help companies make sense of their structured and unstructured data for the purpose of making better business decisions.

Data integration tools: These tools combine data from disparate sources so that they can be viewed or analyzed from a single application. They sometimes include the capability to unify structured and unstructured data.

Document management systems: Also called "enterprise content management systems," a DMS can track, store and share unstructured data that is saved in the form of document files.

Information management solutions: This type of software tracks structured and unstructured enterprise data throughout its lifecycle.

Search and indexing tools: These tools retrieve information from unstructured data files such as documents, Web pages and photos [11].

VI. TEXT ANALYTICS

A major segment of the unstructured data collected by any organization will be in textual format, from e-mail communication and corporate documents to web pages and social media content. So far many of the text processing techniques have been deployed in text-based enterprise search and document management systems but actually text analytics has its roots in information retrieval[21].

In information retrieval, document representation and query processing are the foundations for developing various models which in turn, became the basis for the digital libraries and search engines. In addition to this statistical natural language processing (NLP) techniques have also become important for representing text. In addition to document and query representations, user models are also important in enhancing search performance. Many of these foundational text processing techniques discussed have been deployed in text-based enterprise search and document management systems.

Text analytics also offers significant research opportunities and challenges in several focused areas, including web stylometric analysis for authorship attribution, multilingual analysis for web documents, and large-scale text visualization. Multimedia information retrieval and mobile information retrieval are two other related areas that require support of text analytics techniques, in addition to the core multimedia and mobile technologies. Similar to big data analytics, text analytics will continue to foster active research in industry[8].

Extraction Techniques:

In general, text analytics solutions use a combination of statistical and Natural Language Processing (NLP) techniques to extract information from unstructured data. NLP is a broad and complex field that has developed over the last 20 years. A primary goal of NLP is to derive meaning from text. Natural Language Processing generally makes use of linguistic concepts such as grammatical structures and parts of speech. Often, the idea behind this type of analytics is to determine who did what to whom, when, where, how, and why.

NLP performs analysis on text at different levels like :
Lexical/morphological analysis examines the characteristics of an individual word including prefixes, suffixes, roots, and parts of speech (noun, verb, adjective, and information that will contribute to understanding what the word means in the context of the text provided.

Lexical analysis depends on a dictionary, thesaurus, or any list of words that provides information about those words. In the case of a wireless communication company's sales promotion, a dictionary might provide the information that promotion is a noun that can mean an advancement in position, an advertising or publicity effort, or an effort to encourage someone's growth. Lexical analysis would also enable an application to recognize that promotion, promotions, and promoting are all versions of the same word and idea[16].

Syntactic analysis uses grammatical structure to dissect the text and put individual words into context. Here you are widening your gaze from a single word to the phrase or the full sentence. This step might diagram the relationship between words (the grammar) or look for sequences of words that form correct sentences or for sequences of numbers that represent dates or monetary values. For example, the wireless communication company's call center records included this complaint: "The customer thought it was ridiculous that roll-over minutes were not in the plan." Syntactic analysis would tag the noun phrases in addition to providing the part-of-speech tags.

Semantic analysis determines the possible meanings of a sentence. This can include examining word order and sentence structure and disambiguating words by relating the syntax found in the phrases, sentences, and paragraphs.

Discourse-level analysis attempts to determine the meaning of text beyond the sentence level.

VII. HADOOP FRAMEWORK

Hadoop is framework of tools, open source software that enables the distributed processing of large data sets across clusters of servers. It is designed to scale up from a single server to thousands of machines, with a very high degree of fault tolerance. Hadoop supports running of applications on Big Data[9].

Hadoop framework includes the common utilities that support the other Hadoop modules. It includes Distributed File System HDFS A distributed file system that provides high-throughput access to application data. A framework for job scheduling and cluster resource management and system for parallel processing of large data sets called MapReduce.

Hadoop enables a computing solution that is:
Scalable: New nodes can be added as needed and added without needing to change data formats, how data is loaded, how jobs are written, or the applications on top.

Cost effective: Hadoop brings massively parallel computing to commodity servers. The result is a sizeable decrease in the cost per terabyte of storage, which in turn makes it affordable to model all your data.

Flexible: Hadoop is schema-less, and can absorb any type of data, structured or not, from any number of sources. Data from multiple sources can be joined and aggregated in arbitrary ways enabling deeper analyses than any one system can provide.

Fault tolerant: When you lose a node, the system redirects work to another location of the data and continues processing without missing a fright beat.

VIII. CONCLUSION

In this paper we presented some of the issues related to big data analytics for analyzing unstructured data. This is a ongoing project. We have explained the away unstructured data can be handled and especially text analytics. This is my ongoing project I will be exploring more in my next phase of research

ACKNOWLEDGMENT

The author likes to acknowledge the valuable discussions with Dr. M.P. Singh Department of Computer Science Dr B.R. Ambedkar University Agra, Prof. Vijay Kumar (Curators' Professor), Computer Sc. Electrical Engineering, University of Missouri-Kansas City and Prof. S. K. Mattu Head, Department of Computer Science Delhi University and all my fellow colleagues of PhD scholars at Jaipur National University. The discussions with all have been valuable in preparing this paper.

REFERENCES

[1] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, "Big Data: Issues and Challenges Moving Forward", 2013 46th Hawaii International Conference on System Sciences.

[2] Jaakko Salonen, Jukka Huhtamäki, Ossi Nykänen, "Challenges in Heterogeneous Web Data Analytics - Case Finnish Growth Companies in Social Media" 2013, AcademicMindTrek, Tampere, Finland. ACM 978-1-4503-1992-8/13/1.

[3] Bipin C. Desai, "The State of Data", Proceeding IDEAS '14 Proceedings of the 18th International Database Engineering & Applications Symposium Pages 77-86 ACM ISBN: 978-1-4503-2627-8.

[4] Claudia Loebbecke, Joerg Bienert, Ali Sunyaev, "A Parallel Platform for Big Data Analytics : A Design Science Approach" IJCSET |May 2013 | Vol 3, Issue 5, 152-156, ISSN:2231-0711.

[5] Zhonglin He, Xiaohong Xiao, Yuhua He "A Software Design Model Based on Big Data", Applied Mechanics and Materials Vols. 644-650 (2014), pp 2821-2825 Submitted: 18.07.2014 © (2014) Trans Tech Publications, Switzerland Accepted: 19.07.2014 doi:10.4028/www.scientific.net/AMM.644-650.2821.

[6] "Big Data and Predictive Analytics: What's New?", Seth Earley, Earley & Associates, 1520-9202/14/

$31.00 © 2014 IEEE Published by the IEEE Computer Society computer.org/ITPro.

[7] "Big Data and IS Research", Paulo B. Goes, MIS Quarterly Vol. 38 No. 3 pp. iii-viii/September 2014.

[8] Weiyi Shangy, Zhen Ming Jiangy, Hadi Hemmatiy, Bram Adamsz, Ahmed E. Hassany, Patrick Martinx, "Assisting Developers of Big Data Analytics Applications When Deploying on Hadoop Clouds", 978-1-4673-3076-3/13/c 2013 IEEE , ICSE 2013, San Francisco, CA, USA.

[9] T.K.Das, P.Mohan Kumar, "BIG Data Analytics: A Framework for Unstructured Data Analysis", International Journal of Engineering and Technology (IJET) 2013.

[10] Richard Branch, Heather Tjeerdsma, Cody Wilson, Richard Hurley, Sabine McConnell, "Cloud Computing and Big Data: A Review of Current Service Models and Hardware Perspectives", Journal of Software Engineering and Applications, 2014, 7, 686-693.

[11] Jeff Morris, "Big Data Needs Big Analytics", Software World Vol.43 No.6 5, 2013.

[12] Mark Anawis, "Big Data Analytics Continues to Evolve", 2014, Scientific Computing. com.

[13] Damianos Chatziantoniou, grFlorents Tselai, "Introducing Data Connectivity in a Big Data Web", DanaC'14, June 22 2014, Snowbird, UT, USA. Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-2997-2/14/06 http://dx.doi.org/10.1145/2627770.2627773.

[14] Alekh Jindal, Jorge Quiané-Ruiz, Samuel Madden "Adding Flexibility to the Hadoop Skeleton", SIGMOD'13, June 22–27, 2013, New York, New York, USA. Copyright 2013 ACM 978-1-4503-2037-5/13/06.

[15] Umut A. Acar, Yan Chen "Streaming Big Data with Self-Adjusting Computation", DDFP'13, January 22, 2013, Rome, Italy.Copyright © 2013 ACM 978-1-4503-1871-6/13/01.

[16] "Big Data Now" by O'Reilly Media, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472. ISBN: 978-1-449-31518-4.

[17] Paul Zikopoulos, Chris Eaton, "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data" McGraw-Hill 2012 Page:3.

[18] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

[19] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[20] http://www-01.ibm.com/software/data/bigdata/

[21] http://www.webopedia.com/TERM/B/big_data.html

[22] Judith Herwtz, Allen Lugent, "big data" John Wiley & Sons, Inc 2013, p13.