ISSN: 2277-9655
Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

BIG Data – A Review

Anuradha Bhatia[*1], Gaurav Vaswani[2]
[*1]Senior Lecturer, Computer Department, Mumbai, India
[2]Student , Second Year Computer Technology, Khar, india
anubhatia31@rediffmail.com

Abstract

As more data becomes available from an abundance of sources both within and outside, organizations are seeking to use those abundant resources to increase innovation, retain customers, and increase operational efficiency. At the same time, organizations are challenged by their end users, who are demanding greater capability and integration to mine and analyze burgeoning new sources of information.

Big Data provides opportunities for business users to ask questions they never were able to ask before. How can a financial organization find better ways to detect fraud? How can an insurance company gain a deeper insight into its customers to see who may be the least economical to insure? How does a software company find its most at-risk customers those who are about to deploy a competitive product? They need to integrate Big Data techniques with their current enterprise data to gain that competitive advantage.

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. The problems start right away during data acquisition, when the data tsunami requires us to make decisions, currently in an ad hoc manner, about what data to keep and what to discard, and how to store what we keep reliably with the right metadata. Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modelling are other foundational challenges. Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Finally, presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge.

Keywords: Big data, Knowledgebase, Data environment, cloud

Introduction

We are awash in a flood of data today. In a broad range of application areas, data is being collected at unprecedented scale. Decisions that previously were based on guesswork, or on pain staking constructed models of reality, can now be made based on the data itself. Such Big Data analysis now drives nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. Scientific research has been revolutionized by Big Data .The Sloan Digital Sky Survey has today become a central resource for astronomers the world over. The field of Astronomy is being transformed from one where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database. In the biological sciences, there is now a well-established tradition of depositing scientific data into a public repository, and also of creating public databases for use by other scientists. In fact, there is an entire discipline of bioinformatics that is largely devoted to the duration and analysis of such data. As technology advances, particularly with the advent of Next Generation Sequencing, the size and number of experimental data sets available is increasing exponentially. Big Data has the potential to revolutionize not just research, but also education. A recent detailed quantitative comparison of different approaches taken by 35 charter schools in NYC has found that one of the top five policies correlated with

ISSN: 2277-9655
Impact Factor: 1.852

measurable academic effectiveness was the use of data to guide instruction .Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance.

One clear example of Big Data is the Square Kilometre Array (SKA) (www.skatelescope.org) planned to be constructed in South Africa and Australia. When the SKA is completed in 2024 it will produce in excess of one exabyte of raw data per day (1 exabyte = 1018 bytes), which is more than the entire daily internet traffic at present. The SKA is a 1.5 billion Euro project that will have more than 3000 receiving dishes to produce a combined information collecting area of one square kilometre, and will use enough optical fibre to wrap twice around the Earth. Another example of Big Data is the Large Hadron Collider, at the European Organisation for Nuclear Research (CERN), which has 150 million sensors and is creating 22 petabytes of data in 2012 (1 Petabyte = 1015 bytes, see Figure 1). In biomedicine the Human Genome Project is determining the sequences of the three billion chemical base pairs that make up human DNA. In Earth observation there are over 200 satellites in orbit continuously collecting data about the atmosphere and the land, ocean and ice surfaces of planet Earth with pixel sizes ranging from 50 cm to many tens of kilometres.

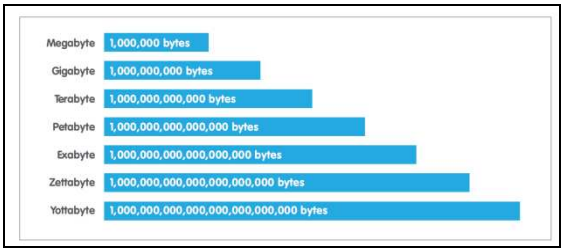

Fig 1: Overview of data scale from megabytes to yottabytes

The Big Data Workflow

The notion of exploring Wikipedia's view of history is a classic Big Data application: an open-ended exploration of "what's interesting" in a large data collection leveraging massive computing resources. While quite small in comparison to the hundreds-of-terabytes datasets that are becoming increasingly common in the Big Data realm of corporations and governments, the underlying question explored in this Wikipedia study is quite similar: finding overarching patterns in a large collection of unstructured text, to learn new things about the world from those patterns, and to do all of this rapidly, interactively, and with minimal human investment.

From Words to Connections: Transforming a Text Archive into a Knowledge Base

Documents are inherently large collections of words, but to a computer each word holds the same meaning and importance as every other word, limiting the types of patterns that can be explored in an archive to simply word frequencies. The creation of higher-order representations capturing specific dimensions of that information, recognizing words indicating space, time, and emotion, allow automated analyses to move closer towards studying patterns in the actual meaning and focus of those documents. The frst generation of Big Data analysis focused largely on examining such indicators in isolation, plotting the tone of discussion of a topic over time or mapping locations and making lists of persons mentioned in that coverage. Connections among indicators have largely been ignored, primarily because the incredible richness of human text leads to networks of interconnections that can easily reach hundreds of trillions of links from relatively small collections. Yet historical research tends to revolve around these very connections and the interplay they capture between people, places, and dates and the actions and events that relate them. Thus, the grand challenge questions driving the second generation of Big Data research tend to revolve around weaving together the myriad connections scattered across an archive into a single cohesive network capturing how every piece of information fits together into the global picture. This in turn is driving an increasing focus on connections and the enormous theoretic and computational challenges that accompany them. In the case of Wikipedia, mapping mentions of locations and creating timelines of date mentions and tone in isolation can be enlightening, but the real insight comes from coupling those dimensions, exploring how tone diffuses over space through time.

Data Environment

More than one out of 10 data managers now have in excess of a petabyte of data within their organizations, and a majority of respondents report their levels of unstructured data are growing. Less than one out of five feels their IT infrastructure will be ready to handle this incoming surge of data. Protecting data overall is important, but unstructured data gets low priority at this time. There are multiple ways to measure Big Data—which can be based on volume, variety, velocity, and value. For the purposes of this survey, we

ISSN: 2277-9655
Impact Factor: 1.852

looked at two of the key differentiators of Big Data versus traditional data stores—volume and variety. In terms of volume, the survey finds considerable amounts of data now being supported within today's Oracle enterprises. For instance, 11% of respondents have data stores within their enterprises that exceed the one-petabyte mark. Another 20% report they are managing data in the hundreds-of-terabytes range. Overall, 42% can be considered large data shops, supporting more than 50TB. (Fig 1) Of course, these levels vary greatly by company size. For example, 28% of the largest organizations in the survey (with more than 10,000 employees) report having more than a petabyte's worth of information in their shops, compared to only 1% of the smallest firms with fewer than 1,000 employees. Another measure of Big Data is variety, as seen in the degree of unstructured data (web logs, social media data, sensor data, documents, imagery, and audio) coursing through enterprise systems. At this time, close to one-fifth of the enterprises surveyed say a significant percentage of their data (25% or more) is unstructured. Even more telling, close to two-thirds of respondents indicate they expect the amount of unstructured data in their organizations to increase over the next three years, either significantly or moderately. The leading data types that comprise the growing Big Data stores include transactional data, email files, and office documents, the survey finds. (Fig 2) While many respondents are bracing for the Big Data deluge if they aren't already in the thick of it current systems, as they are configured, may not be ready for the onslaught. Most data managers in the survey, 65%, feel that for the most part, their current IT infrastructure and database systems are adequate for managing all their data—at this time. However when they consider their systems' adequacy to handle data management requirements three years from now, the percentage expressing such confidence drops to 54%. In addition, this confidence is lukewarm at best—only 19% indicate they feel completely assured that their IT and data infrastructure will be up to the task. The sense of inadequacy is even more intense among the Big Data organizations identified in this survey. Only 11% of respondents in Big Data sites—defined as those reporting more than 500TB and more than 25% unstructured data stores—are confident their systems will be up to the task in three years. While managing all this data is one thing, another key consideration with these growing volumes and the variety of data is its criticality to the business. Keeping data highly available and secure is an ongoing challenge for data managers. How much data presented to respondents' infrastructure and database systems can be lost without repercussions to the organization? For example, can the business afford the loss of a store of unstructured data, such as graphics files, such as videos or web logs? More than one-third, 37%, indicates that absolutely none of their data can be lost. In total, 65% indicate they can't sustain data losses exceeding 5% of their total information assets. While there is a drive to guard almost all data against loss, levels of protection vary significantly by data type—as indicated by half the respondents. And, accordingly, unstructured data receives low priority on the data protection Spectrum. A majority of respondents, 77%, consider transactional data to be most important. Just under half make every effort to protect the integrity of office documents stored on their premises, and about 45% consider their email to be too important to lose. One out of five say it is important to protect device-generated or location data, but only one out of 10 worry about web logs and audio or video files.

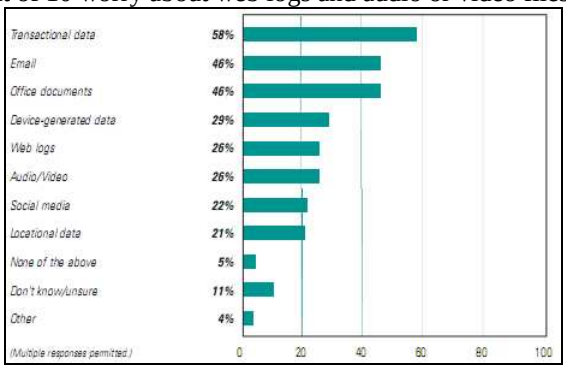

Fig 2: Data Types Driving Need for Big Data Technologies

The analysis of Big Data involves multiple distinct phases as shown in the (Fig 3) below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modelling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase.

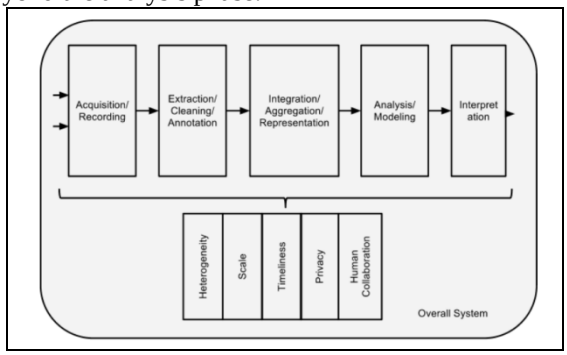

Fig 3: Phases of Big data

ISSN: 2277-9655
Impact Factor: 1.852

For example, Big Data has to be managed in context, which may be noisy, heterogeneous and not include an upfront model. Doing so raises the need to track provenance and to handle uncertainty and error: topics that are crucial to success, and yet rarely mentioned in the same breath as Big Data. Similarly, the questions to the data analysis pipeline will typically not all be laid out in advance. We may need to figure out good questions based on the data. Doing this will require smarter systems and also better support for user interaction with the analysis pipeline. In fact, we currently have a major bottleneck in the number of people empowered to ask questions of the data and analyze it. We can drastically increase this number by supporting many levels of engagement with the data, not all requiring deep database expertise. Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how we manage data analysis.

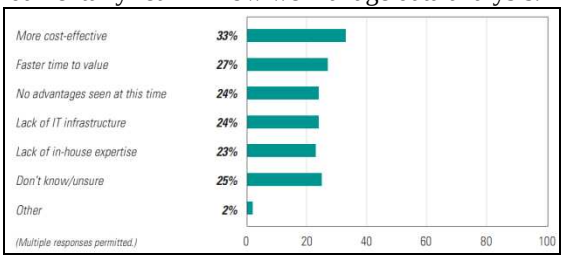

Fig 4: Advantage of Big Data Cloud

Fortunately, existing computational techniques can be applied, either as is or with some extensions, to at least some aspects of the Big Data problem. For example, relational databases rely on the notion of logical data independence: users can think about what they want to compute, while the system (with skilled engineers designing those systems) determines how to compute it efficiently. Similarly, the SQL standard and the relational data model provide a uniform, powerful language to express many query needs and, in principle, allows customers to choose between vendors, increasing competition. The challenge ahead of us is to combine these healthy features of prior systems as we devise novel solutions to the many new challenges of Big Data.(Fig 5)

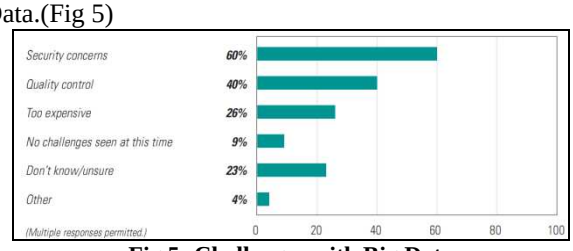

Fig 5: Challenges with Big Data

Conclusion

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result 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. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

References

[1] Hey, T., Tansley, S. & Tolle, K. (2009) The Fourth Paradigm. "Data-intensive scientifc discovery", Microsoft.
[2] Hilbert, M. & Lopez, P. (2011) "The world's technological capacity to store, communicate and compute information", Science 332, 1 April 2011, 60-65.
[3] IDC (2010) "IDC Digital Universe Study, sponsored by EMC", May 2010, available at http://www.emc.com/collateral/demos/microsites/idc-digital-univperse/iview.htm
[4] ICSU Strategic Plan 2006-2011, International Council for Science, Paris, 64pp
[5] All the reports are available at the ICSU website, www.icsu.org
[6] http://www.mysql.com/
[7] Leetaru, K. (2011) "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space", First Monday. 16(9).
[8] http://frstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040
[9] Bellomi, F. & Bonato, R. (2005) "Network Analysis for Wikipedia", Proceedings of Wikimania.
[10] Leetaru, K. (forthcoming). "Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia",

ISSN: 2277-9655
Impact Factor: 1.852

[11] D-Lib Magazine.
[12] http://www.tei-c.org/index.xml
[13] http://history.state.gov/historicaldocuments
[14] Leetaru, K. (forthcoming). "Fulltext Geocoding Versus Spatial Metadata For Large Text Archives: Towards a Geographically Enriched Wikipedia", D-Lib Magazine.
[15] http://en.wikipedia.org/wiki/Golden_Retriever
[16] Leetaru, K. (2011). "Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space", First Monday. 16(9). http://frstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040
[17] http://www.perl.org/
[18] http://www.graphviz.org/