International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 6, June-2014, pp: (463-467), Impact Factor: 1.252, Available online at: www.erpublications.com

Big Data Management: Towards Managing & Analyzing Massive Datasets

Manoj Kumar Singh[1], Dr. Parveen Kumar[2]
[1]Research Scholar, Faculty of Engineering & Technology.Sri Venkateshwara University, Gajraula, U.P, India
[2]Professor, Department of Computer Science & Engg., Amity University, Haryana, India

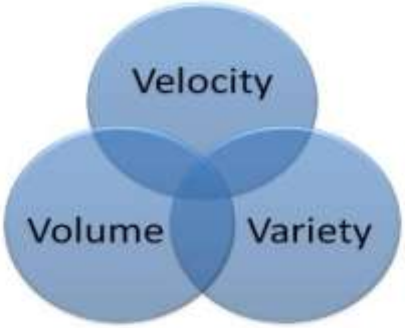
Abstract: An incredible "data deluge" is currently drowning the world. Data sources are everywhere, from Web 2.0 and user-generated content to large scientific experiments, from social networks to wireless sensor networks. This massive amount of data is a valuable asset in our information society. There is a trend that, virtually everyone, ranging from big Web companies to traditional enterprisers to physical science researchers to social scientists, is either already experiencing or anticipating unprecedented growth in the amount of data available in their world, as well as new opportunities and great untapped value but this is the big challenge to manage this massive & Complex datasets. This paper reviews big data management in the view of managing and analyzing large scale data.

Keywords: Big data, Massive, Large scale, datasets, databases.

Introduction

The explosive growth of big data, transactions and digitally aware devices is straining IT infrastructure and operations. At the same time, storage budgets are shrinking and user expectations continue to multiply. Managing file storage becomes more complicated when the user community grows, and are scattered across the globe. In the past, organizations addressed data management challenges by adding file servers or using network attached storage. Traditional network-attached storage solutions can be restricted in performance, security and scalability. A single file server doesn't scale, and even a roomful of file servers means complex storage management. Neither environment is well suited to continuous data access that a data-intensive computing environment requires. This is especially true when files are physically located in different buildings, towns or countries. To overcome these issues, you need to look at a new, more effective approach to managing data. However, the growth of the data volume in our digital world seems to out speed the advance of our computing infrastructure. Conventional data processing technologies, such as database and data warehouse, are becoming inadequate to the amount of data we want to deal with. This new challenge is known as big data. Due to its importance and commonness, it has gained enormous attention in recent years.

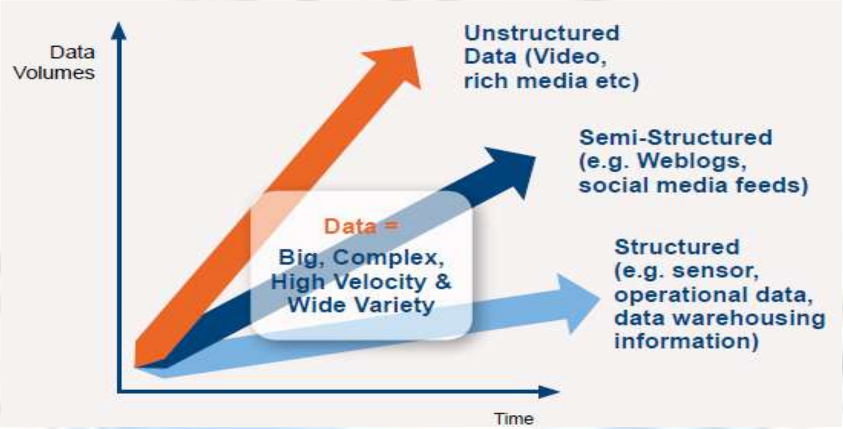
According to McKinsey[4], Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyse. There is no explicit definition of how big a dataset should be in order to be considered Big Data. New technology has to be in place to manage this Big Data phenomenon. IDC defines Big Data technologies as a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis.



Based on this concept, researchers have summarized three important aspects of big data that go beyond the ability of our current data processing technology. They are Volume, Velocity and Variety, also known as 3Vs.

Volume is synonymous with the “big” in the term, “Big Data”. Volume is a relative term – some smaller-sized organisations are likely to have mere gigabytes or terabytes of data storage as opposed to the petabytes or exabytes of data that big global enterprises have. Data volume will continue to grow, regardless of the organisation’s size. There is a natural tendency for companies to store data of all sorts: financial data, medical data, environmental data and so on. Many of these companies’ datasets are within the terabytes range today but, soon they could reach petabytes or even exabytes. Despite of the various big-volume issues, there is still no agreement on the quantification of big data. Such quantification depends various factors. First, the complexity of the data structure is an important factor. A relational dataset of several petabytes may not be called big data, since it can be readily handled by today’s DBMSs. In contrast, a graph dataset of several terabytes is commonly regarded as big data, as graph processing is very challenging to our technologies.

Data can come from a variety of sources and in a variety of types. With the explosion of sensors, smart devices as well as social networking, data in an enterprise has become complex because it includes not only structured traditional relational data, but also semi-structured and unstructured data.


Data Volumes
Unstructured Data (Video, rich media etc)
Semi-Structured (e.g. Weblogs, social media feeds)
Structured (e.g. sensor, operational data, data warehousing information)
Data =
Big, Complex, High Velocity & Wide Variety
Time

The velocity of data in terms of the frequency of its generation and delivery is also a characteristic of big data. Conventional understanding of velocity typically considers how quickly the data arrives and is stored, and how quickly it can be retrieved. In the context of Big Data, velocity should also be applied to data in motion: the speed at which the data is flowing. The various information streams and the increase in sensor network deployment have led to a constant flow of data at a pace that has made it impossible for traditional systems to handle. There have been sophisticated technologies to deal with each of these data types, such as those of database and information retrieval. However, a seamless integration of these technologies remains as a challenge.

Big Data Management

Big data management is about two things—big data and data management—plus how the two work together to achieve business and technology goals. For very large data sets, big data can also be an eclectic mix of structured data (relational data), unstructured data (human language text), semi-structured data (RFID, XML), and streaming data (from machines, sensors, Web applications, and social media). In this report, the term multi-structured data refers to data sets or data environments that include a mix of these data types and structures.

This is where data management disciplines, tools, and platforms (both old and new) are applied to the management of big data (in the base definition or the extended one). Traditional data and new big data can be quite different in terms of content, structure, and intended use, and each category has many variations within it. To accommodate this diversity, software solutions for BDM tend to include multiple types of data management tools and platforms, as well as diverse user skills and practices.

Big Data Management for Different Data Types & Structure

Structured data retains its hegemony, even with diverse big data. At 88%, structured data is by far the most managed data type today.In fact; we can safely assume that most of this structured data is actually relational, meaning that relational data is still very prominent. In turn, that means that DBMSs, SQL, and other tool types and technologies for relational data are important to managing big data.

Semi-structured data is the most prominent secondary data format. A number of data formats include a mixture of structured data, hierarchies, text, and so on. Common examples include documents that adhere to standards for XML, JSON, and RSS. Coincidentally, these documents are often used as formats for messages and events, so they may also be considered event data, which (along with semi structured data) also ranked highly in the survey as a prominent secondary data format for big data.

Web data ranked surprisingly low. Web servers and Web applications have been with us for almost 20 years, and Web data is a common source of big data today. So it's surprising that almost half of survey respondents (45%) don't manage Web logs and clickstreams at all.

Social media data ranked surprisingly high. Social media Web sites are only a few years old, and it's only in the last three years that user organizations have started to collect social data for study. So it's surprising that so many organizations surveyed are already managing social data.

Unstructured data still eludes many organizations. All forms of unstructured data require highly specialized technologies and skills, which may explain why approximately half of organizations surveyed still don't manage unstructured big data in the form of human language or audio/video (45%), personal productivity files (43%), or e-mail (53%).

Technology Drivers behind Big Data Management

Big data just gets bigger: It's important to beef up data management infrastructure and skills as early as possible. Otherwise, an organization can get so far behind from a technology viewpoint that it's difficult to catch up. From a business viewpoint, delaying the leverage of big data delays the business value. Similarly, capacity planning is more important than ever, and should be adjusted to accommodate the logarithmic increases typical of big data.

Resistance is futile: big data will be assimilated into enterprise data. You have to start somewhere, even if it's a data management silo devoted to one form of big data. Typical silos manage Web logs, sensor and machine data logs, and persisted data streams. Yet, it's also important to determine how each form of big data will eventually fit into an overall architecture for enterprise data.

Leverage big data, don't just manage it. It costs money to collect and store big data, so don't let it be a cost center. Look for ways to get business value from big data. As you select data platforms for managing big data, consider low-cost new ones and open source.

Advanced analytics is the primary path to business value from big data. This fact is so apparent that there's even a name for it: *big data analytics*. In many ways, the current uptick in advanced analytics among user organizations is driven by the availability of new big data, plus the new business facts and insights that can be learned from its study.

Joining big data with traditional data is another path to value. For example, so-called 360-degree views of customers and other business entities are more complete and bigger when based on both traditional enterprise data and big data. In fact, some sources of big data come from new customer touch points (mobile apps, social media) and so belong in your customer view.

Big data can enable new applications. For example, in recent years, a number of trucking companies and railroads have added multiple sensors to each of their fleet vehicles and train cars. The big data that streams from sensors enables companies to more efficiently manage mobile assets, deliver products to customers more predictably, identify noncompliant operations, and spot vehicles that need maintenance.

Big data can extend older applications. This includes any application that relies on a 360-degree view, as mentioned above. Big data can also beef up the data samples parsed by many analytic applications, especially those for fraud, risk, and customer segmentation.

ANALYZING BIG DATA

Equally relevant as the sources of data are the methodologies to analyze them and the standards of evidence that would be acceptable to management scholars for their publication. As with any nascent science, there is likely to be a trade-off between theoretical and empirical contribution, and the rigor with which data are analyzed. Perhaps, with big data, we are liable to initially be confounded by the standard of evidence that should be expected. The typical statistical approach of relying on *p* values to establish the significance of a finding is unlikely to be effective because the immense volume of data means that almost everything is significant. Using our typical statistical tools to analyze big data, it is very easy to get false correlations. However, this doesn't necessarily mean that we should be moving toward more and more complex and sophisticated econometric techniques to deal with this problem; indeed, such a response poses a substantial danger of over-fitting the data. Instead, basic Bayesian statistics and stepwise regression methods may well be appropriate approaches. Beyond these familiar approaches, there is a range of specialized techniques for analyzing big data, each of which is important for those entering this field to understand, though beyond the scope of this editorial. These techniques draw from several disciplines, including statistics, computer science, applied mathematics, and economics. They include (but are not limited to) testing, cluster analysis, data fusion and integration, data mining, genetic algorithms, machine learning, natural language processing, neural networks, network analysis, signal processing, spatial analysis, simulation, time series analysis, and visualization.

A mere Tweet from a trusted source can cause losses or profits of billions of dollars and a chain reaction in the press, social networks, and\ blogs. This situation makes information goods even more difficult to value, as they have a catalytic impact on real-time decision making. Meanwhile, entrepreneurs and innovators have taken aggregate open and public data as well as community, selfquantification and exhaust data to create new products and services that have the power to transform industries. In private and public spheres, big data sourced from mobile technologies and banking services, such as digital/mobile money, when combined with existing "low-tech" services, such as water or electricity, can transform societies and communities. There is little doubt that, over the next decade, big data will change the landscape of social and economic policy and research.

As big data became gradually available to web researchers, major methodological concerns emerged such as how to store and analyze such large quantities of data simultaneously—even samples carefully drawn from such large data pools. Specific technologies and software emerged to address these concerns. Big data technology includes big data files, database management, and big data analytics (Hopkins & Evelson, 2011).

One of most popular and widely available data management systems for dealing with hundreds of gigabytes or petabytes data simultaneously is the Hadoop programming model popularized by
Google. Its strengths include providing reliable shared online storage for large amounts of multiple sourced data through the Hadoop Distributed Filesystem (HDFS), analysis through Map Reduce (a batch query processer that abstracts the problem from disk reads and writes), and transforming data into a map-and-reduce computation over sets of keys and values (White, 2012). Map Reduce works well with unstructured or semi-structured data because it is designed to interpret the data at processing time (Verma, Cherkasova, & Campbell, 2012). While the Map Reduce system is able to analyze a whole "big data" set and large samples in batch fashion, the Relational Database Management System (RDBMS) shows more strength in processing point queries where the data is structured into entities with defined format (i.e., structured data) as may occur in key-word or key characteristic sampling (White, 2012). Different from Map Reduce's linear scalable programming that is not sensitive to the change of data size and cluster, RDBMS is a nonlinear programming which allows complex functions such as quadratic or cubic terms (Sumathi & Esakkirajan, 2007) in the model. RDBMS could be retrieved from http:// mysql-com.en.softonic.com/. Google's success in text processing and their embrace of statistical machine learning was decoded as an endorsement that facilitated Hadoop's wide-spread adoption. Hadoop, the open-source software can be downloaded from http://hadoop.apache.org/ releases.html. On the other hand, additional technologies and software are available for use with big data sets and samples. They represent reasonable alternatives to Hadoop, especially when data sets display unique characteristics that can be best addressed with specialized software.

CONCLUSION

The development of internet technology has made large comprehensive data sets readily available, including publically-available, textual, online data. Such data sets offer richness and potential insights into human behavior, but can be costly to harvest, store, and analyze as huge data sets can translate into big labor and computing costs for mining, screening, cleansing, and textual analysis. Incredibly large and complex data sets cry out for effective sampling techniques to manage the sheer size of the data set, its complexity, and perhaps most importantly, its on-going growth. In this research paper, we review managing and analyzing large scale datasets for Big Data Management.

References

[1]. Labrinidis A, Jagadish H. Challenges and opportunities with big data. Proceedings of the VLDB Endowment, 2012, 5(12): 2032–2033.

[2]. Lu J, Senellart P, Lin C, Du X, Wang S, Chen X. Optimal top-*k* generation of attribute combinations based on ranked lists. In: Proceedings of the 2012 International Conference on Management of Data. 2012, 409–420.

[3]. Doan A, Naughton J F, Baid A, Chai X, Chen F, Chen T, Chu E, DeRose P, Gao B J, Gokhale C, Huang J, Shen W, Vuong B Q. Th case for a structured approach to managing unstructured data. In: Proceedings of the 4th Biennial Conference on Innovative Data Systems Research. 2009.

[4]. Badke, W. (2012). Big search, big data. *Online*, *36*(3), 47–49.

[5]. Gartner, 2013. Gartner Big data Survey [Online] Available at <http://www.gart ner.com/newsroom/id/2593815> [Accessed 14.04.2014].

[6]. Haselden, K., Wolter, R., 2006. The What, Why, and How of Master data Management. Microsoft Corporation [Online] <http://msdn.microsoft.com/en us/library/bb190163.aspx> [Accessed 23.04.2014].

[7]. IBM, 2014. IBM Big Data [Online] Available at <http://www-01.ibm.com/software/data/bigdata/> [Accessed 15.04.2014]

[8]. Needham, J., 2013. Disruptive Possibilities: How Big Data Changes Everything. 1st Ed. O'Reilly Media.