International Journal of Research Studies in Computing
April 2013, Volume 2 Number 1, 11-20

International Journal of Research Studies in
Computing
Volume 1 Number 1 April 2012

The main challenges and issues of big data management

Almeida, Fernando ✉
Faculty of Engineering, University of Porto, Portugal (almd@fe.up.pt)
Calistru, Catalin
Innovation and Development Centre, ISPGaya, Portugal (cmc@ispgaya.pt)

ISSN: 2243-772X
Online ISSN: 2243-7797

Received: 21 August 2012 **Revised**: 6 September 2012 **Accepted**: 9 September 2012
Available Online: 11 September 2012 **DOI:** 10.5861/ijrsc.2012.209

OPEN ACCESS

Abstract

Big data is a disruptive force that will affect organizations across industries, sectors and economies. Through better analysis of the large volume of data that are becoming available, there is the potential for making faster advances in many scientific domains and improving the profitability of many enterprises. This paper examines the main challenges and issues that will have to be addressed to capture the full potential of big data. Additionally, we propose some initiatives and good practices that will help companies in the transition process for the big data analysis.

Keywords: big data; data management; data analysis; management policies; business intelligence; data warehousing

The main challenges and issues of big data management

1. Introduction

The term “big data” has recently grown in prominence as a way of describing the phenomenon of growth in data volume, complexity and disparity. The definition of big data is not totally consensual in literature and there may be some confusion around what it really means. Big data is not just an environment in which accumulated data has reached very large proportions. The word “big” does not just refer to size. If it was just a capacity issue the solution would be relatively simple. Instead, big data refers to environment in which data sets have grown too large to be handled, managed, stored and retrieved in an acceptable timeframe (Slack, 2012). According to Floyer (2012) big data has the following characteristics:

- Very large (petabytes/exabytes of data, millions/billions of people, and billions/trillions of records);
- Distributed aggregations of loosely structured data (often incomplete and inaccessible);
- Flat schemas with few complex interrelationships;
- Often involving time-stamped events and made up of incomplete data. Frequently connections between data elements must be probabilistically inferred.

Many citizens around the world regard this collection of information with deep suspicion, seeing the data flood as nothing more than an intrusion of their privacy. But there is strong evidence that big data can play a significant economic role to the benefit not only of private commerce but also of national economies and their citizens. In fact, the data can create significant value for the world economy, enhancing the productivity and competitiveness of companies and the public sector and creating substantial economic surplus for consumers (Lehdonvirta & Ernkvist, 2011). According to Manyika et al. (2011) estimate that government administration in Europe could save more than €100 billion in operational efficiency improvements alone by using big data. Furthermore, this estimate does not include big data levers that could reduce fraud, errors, and tax gaps (i.e., the gap between potential and actual tax revenue).

Digital data is currently in every economic sector, organization and user of digital technology. While this topic might once have initially concerned only a few data geeks and pioneers, big data is now relevant for leaders across every sector, and consumers of products and services stand to benefit from its application. According to IDC (2009), it is expected a 2500 exabytes of new information in 2012 with digital content as the primary driver. Moreover, digital universe grew by 62% in 2011. This scenario is illustrated in figure 1.

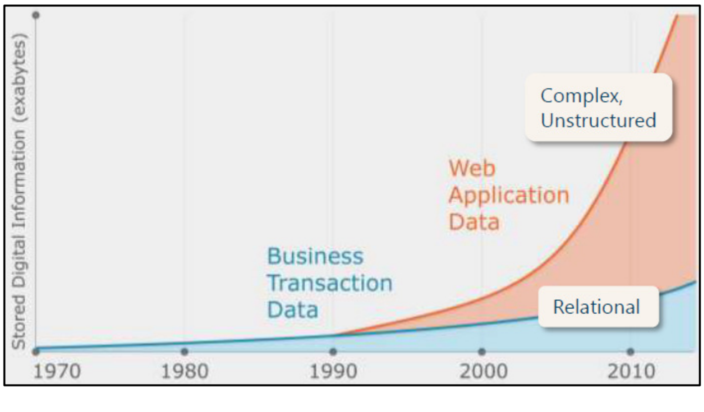

Figure 1. Data growth and expansion (IDC, 2009)

The ability to store, aggregate, and combine data and then use the results to perform deep analysis has become ever more accessible as trends such as Moore's Law in computing, its equivalent in digital storage, and cloud computing continue to lower costs and other technology barriers (Brill, 2007). Further, the ability to generate, communicate, share, and access data has been revolutionized by the increasing number of people, devices, and sensors that are now connected by digital networks. According to Manyika et al. (2011) more than 4 billion of people in 2010 were using mobile phones, and about 12 percent of those people had smartphones, whose penetration is growing at more than 20 percent a year. At the same time, more than 30 million networked sensor nodes are present in the transportation, automotive, industrial, utilities, and retail sectors. The number of these sensors is increasing at a rate of more than 30 percent a year (Manyika et al., 2011).

Aberdeen's research demonstrates that companies are seeing an average year growth of 38% in data volume (Lock, 2012). The average company confronts 2.5 times more data than it did three years ago, which is an increase that might seem particularly small for particularly data-driven companies. However, data is growing in complexity and variety as well as volume. Between data warehouses, data marts, enterprise applications, spreadsheets and external unstructured or social data, companies are drawing on an increasing number of unique data sources to drive their business analysis.

However, the influx of data presents many barriers to effective analytics, and to the creation of business insight for most decision makers (Adhikari, 2012). Whether their data is inaccessible, fragmented, or simply unwieldy from a volume perspective, companies are seeking formalized data management strategies in response. According to the Aberdeen's study (Lock, 2012), late delivery of information is the top pressure driving to develop their data management initiatives. This situation is depicted in figure 2.

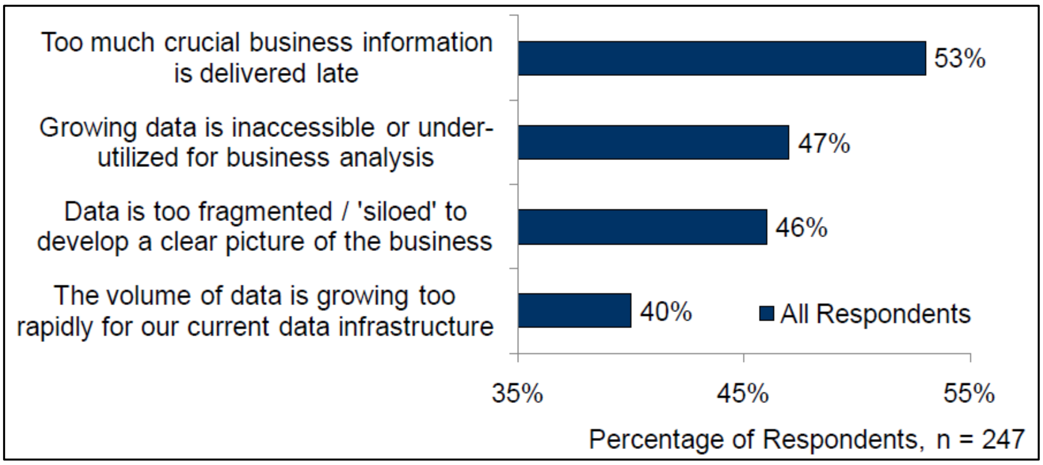

Figure 2. Top pressures driving data management initiatives (Lock, 2012)

There are many ways that big data can be used to create value across sectors of the global economy. Many pioneering companies are already using big data to create value, and others need to explore how they can do the same if they are to compete. Governments also have a significant opportunity to boost their efficiency and the value for money they offer citizens at a time when public finances are strongly constrained. According to Smith (2012) big data contributed an estimated £25.1 billion to the UK economy in 2011 but, as the adoption of analytics increases, it is forecasted to reach £40.7 billion by 2017. The Smith (2012) study suggests that at the same time big data analytics adoption will raise from 34 per cent in 2011 to 54 per cent by 2017, which is equivalent to 22 per cent of the UK net debt (c. £1 trillion) or more than the 2011/12 defense, healthcare and education budgets combined. The emergence of big data is expected to benefit economies is terms of business creation, efficiency gains and innovation. Additionally, the Centre for Economics and Business Research (CEBR) in UK predicts that the most beneficial sectors from big data analysis will be the financial services, public sector, retail and manufacturing (CEBR, 2012).

This paper presents and analyzes the main challenges and issues that will have to be addressed to capture the full potential of big data. Section II details the three ETL phases in big data processing. Section III presents the mains challenges of big data analysis. Further, Section IV proposes relevant strategies and good practices for a big data analysis. Finally, Section V draws conclusions.

2. Phases in big data processing

Big data does not arise instantaneously, but it is recorded from some data generating sources, typically OLTP systems, spreadsheets, text files and web content. Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. However, relevant data must be collected and loaded in a target data warehouse and business intelligence system. One challenge in this process is to define these filters in such a way that they do not discard useful information. The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recoded and measured. Finally, another important issue here is data provenance. It is important to mention that, for instance, recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline.

Data warehouse operational processes normally compose a labor intensive workflow and constitute an integral part of the back-stage of data warehouse architectures, where the collection, extraction, cleaning, transformation, and transport of data takes place, in order to populate the warehouse. To deal with this workflow, and in order to facilitate and manage the data warehouse operational processes, specialized tools are already available in the market, under the general title Extraction-Transformation-Loading (ETL) tools.

ETL tools represent an important part of data warehousing, as they represent the mean in which data actually gets loaded into the warehouse. The figure 3 illustrates each individual stage in the process.

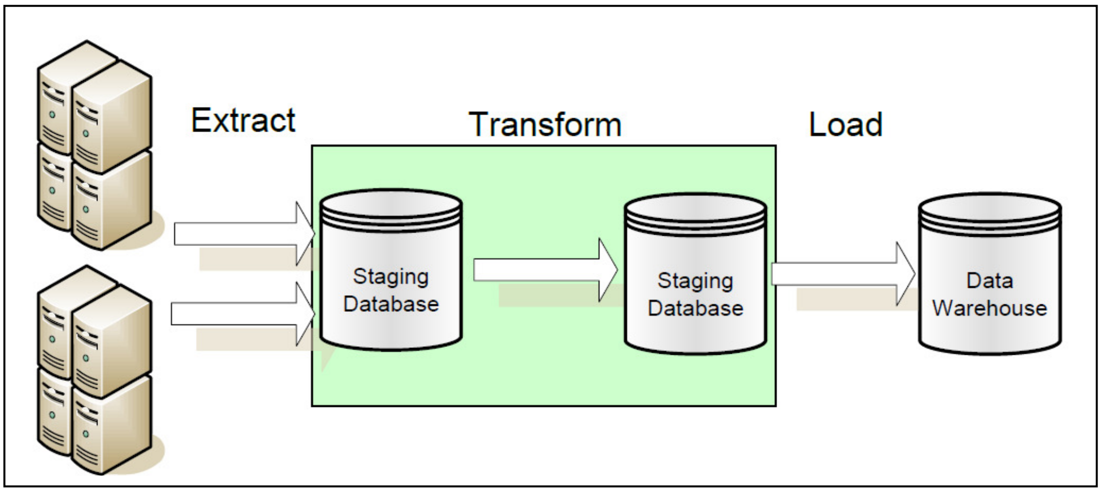

Figure 3. ETL process (Golfarelli & Rizzi, 2009)

Data is extracted from the data sources using a data extraction tool via whatever data connectivity is available. It is then transformed using a series of transformation routines. This transformation process is largely dictated by the data format of the output. Data quality and integrity checking is performed as part of the transformation process, and corrective actions are built into the process. Transformation and integrity checking are performed in the data staging area. Finally, once the data is in the target format, it is then loaded into the data warehouse ready for presentation.

The process is often designed from the end backwards, in that the required output is designed first. In so doing, this informs exactly what data is required from the source. The routines designed and developed to implement the process are written specifically for the purpose of achieving the desired output, and only the data required for the output is included in the extraction process. In addition, the output design must incorporate all facts and dimensions required to present both the aggregation levels required by the business intelligence solution and any possible future requirements.

Business rules that define how aggregations are achieved and the relationships between the various entities in both the source and target, are designed and therefore coded into the routines that implement the ETL process. This process leads to tight dependencies in the routines at each stage of the process.

In all phases of an ETL process, individual issues can arise, making data warehouse refreshment a very troublesome task. In next sections we briefly describe the most common issues, problems, and constraints that turn up in each phase separately.

2.1 Extraction

During the ETL process, one of the very first tasks that must be performed is the extraction of the relevant information that has to be further propagated to the warehouse. In order to minimize the overall processing time, this involves only a fraction of the source data that has changed since the previous execution of the ETL process, mainly concerning the newly inserted and possibly updated records. Usually, change detection is physically performed by the comparison of two snapshots (one corresponding to the previous extraction and the other to the current one). Efficient algorithms exist for this task, like the snapshot differential algorithms presented by Labio and Garcia-Molina (1996). Another popular technique is log sniffing, which consists in the scanning of the log file in order to reconstruct the changes performed since the last scan (Jorg & Dessloch, 2010).

2.2 Transformation

According to Rahm and Hai Do (2000), this phase can be divided in the following tasks: (a) data analysis; (b) definition of transformation workflow and mapping rules; (c) verification; (d) transformation; and (e) backflow of cleaned data.

In terms of transformation tasks, Lenzerini (2002) distinguishes two main classes of problems: (a) conflicts and problems at the schema level (e.g., naming and structural conflicts) and (b) data level transformations (i.e., at the instance level). According to Vassiliadis et al. (2005), the main problems with respect to the schema level are in terms of naming conflicts, where the same name is used for different objects (homonyms) or different names are used for the same object (synonyms). Furthermore, structural conflicts can also appear where one must deal with different representations of the same object in different sources.

The integration and transformation programs perform a wide variety of functions, such as reformatting, recalculating, modifying key structures, adding an element of time, identifying default values, supplying logic to choose between multiple sources, summarizing, merging data from multiple sources, etc.

2.3 Loading

The final phase of the ETL process has also its own technical challenges. A major problem is the ability to discriminate between new and existing data at loading time. This problem arises when a set of records has to be classified to the new rows that need to be appended to the warehouse, and rows that already exist in the data warehouse, but their value has changed and must be updated (Castellanos et al., 2009). Currently modern ETL tools already provide mechanisms towards this problem, mostly through language predicates.

An extra problem that can also appear is the simultaneous usage of the rollback segments and log files during the loading process. According to Reddy and Jena (2010), a technique that can be used that facilitate the loading task involve the creation of tables at the same time with the creation of the respective indexes, the minimization of inter-process wait states, and the maximization of concurrent CPU usage.

3. Challenges in Big Data Analysis

Applying big data analytics faces several challenges related with the characteristics of data, analysis process and social concerns.

The first challenge appears in terms of privacy. The privacy is the most sensitive issue, with conceptual, legal, and technological implications. This concern increases its importance in the context of big data. In its narrow sense, privacy is defined by the International Telecommunications Union (Gordon, 2005) as the "right of individuals to control or influence what information related to them may be disclosed". Privacy can also be understood in a broader sense as encompassing that of companies wishing to protect their competitiveness and consumers and stages eager to preserve their sovereignty and citizens. In both these interpretations, privacy is an overarching concern that has a wide range of implications for anyone wishing to explore the use of big data for development in terms of data acquisition, storage, retention, use and presentation.

Another challenge, indirectly related with the previous, is the access and sharing of information. It is common to expect reluctance of private companies and other institutions to share data about their clients and users, as well as about their own operations. Obstacles may include legal or reputational considerations, a need to protect their competitiveness, a culture of secrecy, and more broadly, the absence of the right incentive and information structures. There are also institutional and technical challenges, when data is stored in places and ways that make it difficult to be accessed and transferred.

Another very important direction is to rethink security for information sharing in big data use cases. Many online services today require us to share private information (i.e., facebook, linkedin, etc), but beyond record-level access control we do not understand what it means to share data, how the shared data can be linked, and how to give users fine-grained control over this sharing.

The size of big data structures is also a crucial point that cans constraint the performance of the system. Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. In the past, this challenge was mitigated by processors getting faster, which provide us with the resources needed to cope with increasing volumes of data. But there is a fundamental shift underway now considering that data volume is scaling faster than computer resources.

Considering the size issue, we also know that the larger the data set to be processed, the longer it will take to analyze. The design of a system that effectively deals with size is likely also to result in a system that can process a given size of data set faster. However, it is not just this speed that is usually meant when we refer to speed in the context of big data. Rather, there is an acquisition rate challenge in the ETL process. Typically, given a large data set, it is often necessary to find elements in it that meet a specific criterion which likely occurs repeatedly. Scanning the entire data set to find suitable elements is obviously impractical. Rather, index structures are created in advance to permit finding qualifying elements quickly.

Finally, working with new data sources brings a significant number of analytical challenges. The relevance and harshness of those challenges will vary depending on the type of analysis being conducted, and on the type of decisions that the data might eventually inform. The big core challenge is to analyze what the data is really telling us in a fully transparent manner. The challenges are intertwined and difficult to consider in isolation, but according to King and Powell (2008), they can be split into three categories: (a) getting the picture right (i.e., summarizing the data), (b) interpreting or making sense of the data through inferences, and (c) defining and detecting anomalies.

4. Strategies and good practices

Strategies for dealing with big data challenges will differ depending on the data maturity of the organization. How efficiently and effectively can data be collected for analysis, and for that matter, is the organization aware of all of the different types and sources of data that should be included to maximize insight and answers? How well can the organization reconcile different data formats? What is the cost of collecting and analyzing data, and how is that cost weighed against the anticipated value of the outcome?

In a first step it is important to bring together line of business leaders and IT practitioners to identify which data pools have the greatest value. Additionally, it would be interesting to evaluate the data stores that have been prepared for analysis and consider how they could be expanded or improved, and look at unstructured data sets and prioritize which ones should be converted to more usable formats. Business leaders and IT staff must also work together to highlight use cases for the data based on existing business to determine which approaches will yield the most business value in the shortest window.

In a perfect world free from budgets, every piece of data that is collectable would be collected, and every byte would be analyzed in as many ways as the mind can consider. But in reality, collecting, storing, and analyzing data comes at a cost. Companies will need to make economic decisions about which data is worth collecting and analyzing. In fact, different parts of the business will have to make compromises. Business leaders are likely to lean towards collecting and analyzing more data, while IT leaders will aware if technology budget limitations and staff restrictions may lean in the other direction. Given the iterative nature of big data, these decisions will need to be revisited on a regular basis to ensure the organization is considering the right data to produce insight at any given point in time.

The more data collected, the bigger the economic problem become. It is more expensive to store and manipulate more data, and the more data there is to process the more computational power is required, layering on more cost. Yet more data produces better informed decisions. Approaching big data analytics with finite definitions of what data will be considered sounds counterproductive, but companies that are just starting out on big data projects, will need to set some parameters around just which data is involved and gauge expectations of results accordingly.

As an ever large amount of data is digitized and travels across organizational boundaries there is a set of policy issues that will become increasingly important in terms of privacy, security, intellectual property, and liability. Clearly, privacy is an issue whose importance particularly to consumers, is growing as the value of big data becomes more apparent. Personal data such as health and financial records are often those that can offer the most significant human benefits, however, consumers also view these categories of data as being the most sensitive.

Another closely related concern is data security (e.g., how to protect competitively sensitive data or other data that should be kept private). Recent examples have demonstrated that data breaches can expose not only personal consumer information and confidential corporate information but even national security secrets (Terence & Ludloff, 2011). With serious breaches on the rise, addressing data security through technological and policy tools will become essential.

At the same time that big data increases its importance also raises a number of legal issues, especially when coupled with the fact that data are fundamentally different from many other assets. Data can be copied perfectly and easily combined with other data. The same piece of data can be used simultaneously by more than one person. All of these are unique characteristics of data compared with physical assets. Questions about the intellectual property rights attached to data will have to be properly answered and managed.

In order to capture value from big data, organizations will have to deploy new technologies and techniques that include new types of analyzes. The range of technology challenges and the priorities set for tackling them will differ depending on the data maturity of the institution. Legacy systems and incompatible standards and formats too often prevent the integration of data and the more sophisticated analytics that create value from big data. New problems and growing computing power will spur the development of new analytical techniques. There is also a need for ongoing innovation in technologies and techniques that will help individuals and organizations to integrate, analyze, and visualize the growing amount of big data.

There is, of course, the need to have some expertise to deal with big data analytics technology. The big data architect/engineer must design the big data environment, which should include:

- Store and management – here the company can take advantage of commodity technologies, possibly implemented in the cloud, which can accept and accommodate massive feeds or relatively unstructured data;
- Mapping and understanding – some providers of big data solutions expects that the company completes this arduous task on his own. However, it is no longer necessary, because some solutions include tools that facilitate and at least partially automate this essential discovery and mapping process;
- Analytics – looking to the market we can easily realize that not all business solutions are suitable for big data. The right choice will be highly capable of both ad hoc, discovery-focused analysis and efficient, ongoing measuring and monitoring. High performance is critical, and solutions that are at least partially cloud-based can provide data integration advantages along with the obvious scaling pluses.

To enable transformative opportunities, companies will increasingly need to integrate information from multiples sources. In some cases, organizations will be able to purchase access to the data. In other cases, however, gaining access to third-party data is often not straightforward. The sources of third-party data might not have considered sharing it. Sometimes, economic incentives are not aligned to encourage stakeholders to share data. A stakeholder that holds a certain dataset might consider it to be the source of a key competitive advantage and thus would be reluctant to share it with other stakeholders.

Invest in data quality and metadata is also a key issue in any system, and big data systems are no exception. However, big data systems require even much more automation and advance planning. In a first step, the company should ensure that data quality is not treated as a project or initiative, but as a fundamental layer of the data stack that receives adequate resourcing and management attention. Second, the system should be built considering multiple lines of defense: from data mastering (e.g., creation of customer accounts) to data collection (e.g., recording all customers' interactions with the company) to metadata (e.g., organizing and dimensionalizing the data to aid in future reporting and analysis). Third, the company should automate both the processes that identify and elevate data quality issues and the measurement and reporting of data quality progress.

The company should also ask for regular feedback. Big data is a learning process, both in terms of managing the data and in driving business value from its contents. The internal user base is a valuable source of feedback and integral to the company's learning and development process. Areas such as usability, data quality, and data latency are all categories within which users will give important feedback. In addition, the company should ask for ad hoc feedback from every level of the stakeholder organizations so they see the company commitment to making their business better. Regular feedback ensures that the big data system is tightly integrated into business decision making, so it can play a lead role in business improvement. This regular feedback process will help companies to build more effective big data capabilities, saving time and money, and driving maximum Return On Investment (ROI) for their businesses.

Finally, it is important to consider that sectors with a relative lack of competitive intensity and performance transparency, along with industries where profit pools are highly concentrated, are likely to be slow to fully leverage the benefits. According to Manyika (2011), the public sector tends to be a lack of competitive pressure that limits efficiency and productivity. As a result, the sector faces more difficult barriers than other sectors (such as telecom, manufacturing or retail) in the way of capturing the potential value from using big data.

5. Conclusion

The effective use of big data has the potential to transform economies, delivering a new wave of productivity growth and consumer surplus. Using big data will become a key basis of competition for existing companies, and will create new competitors who are able to attract employees that have the critical skills for a big data world. Leaders of organizations need to recognize the potential opportunity as well as the strategic threats that big data represent and should assess and then close any gap between their current IT capabilities and their data strategy and what is necessary to capture big data opportunities relevant to their enterprise. In this task, they will need to be creative and proactive in determining which pools of data they can combine to create value and how to gain access to those pools.

However, many technical and organizational 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 privacy, security, heterogeneity, integration, lack of structure, data quality and regular feedback. These challenges will require transformative solutions, and will not be addressed naturally simply by the evolution of business intelligence systems. Not only enterprise IT architectures will need to change to accommodate it, but almost every department within a company will undergo adjustments to allow big data to inform and reveal. Data analysis will change, becoming part of a business process instead of a distinct function performed only by trained specialists. Big data productivity will come as a result of giving users across the organization the power to work with diverse data sets through self-services tools.

Achieving the vast potential of big data demands a thoughtful, holistic approach to data management, analysis and information intelligence. Across industries, organizations that get ahead of big data will create new operational efficiencies, new revenue streams, differentiated competitive advantage and entirely new business models. Business leaders should start thinking strategically about how to prepare the organizations for big data.

6. References:

Adhikari, S. (2012). *Time for a big data diet*. Retrieved August 12, 2012, from http://technologyspectator.com.au/emerging-tech/big-data/need-big-data-speed?

Brill, K. (2007). The invisible crisis in the data center: the economic meltdown of Moore’s law. *Uptime Institute White Paper*, *7*, 1-8.

Castellanos, M., Simitsis, A., Wilkinson, K. & Dayal, U. (2009). Automating the loading of business process warehouses. In *International Conference on Extending Database Technology (EDBT)* (pp. 612-623), Saint-Petersburg, Russian Federation. <http://dx.doi.org/10.1145/1516360.1516431>

CEBR (2012). Data equity: unlocking the value of big data. *Centre for Economics and Business Research White Paper*, 4, 7-26.

Floyer, D. (2012). *Enterprise Big-data*. Retrieved July 12, 2012, from http://wikibon.org/wiki/v/Enterprise_Big-data

Golfarelli, M., & Rizzi, S. (2009). *Data warehouse design: modern principles and methodologies*. Columbus: McGraw-Hill.

Gordon, A. (2005). Privacy and ubiquitous network societies. *Workshop on ITU Ubiquitous Network Societies*, 6-15.

IDC. (2009). As the economy contracts, the digital universe expands. *IDC White Paper*, 5, 12-18.

Jorg, T., & Dessloch, S. (2010). Near real-time data warehousing using state-of-the-art ETL tools. *In enabling real-time for businessintelligence* (pp. 100-117), Heidelberg: Springer-Verlag, 100-117. <http://dx.doi.org/10.1007/978-3-642-14559-9_7>

King, G., & Powell, E. (2008). *How not to lie without statistics*. Working paper: Harvard university. Retrieved August 12, 2012, from http://gking.harvard.edu/gking/files/nolie.pdf

Labio, W., & Garcia-Molina, H. (1996). Efficient snapshot differential algorithms for dará warehousing. In *Proceedings of the 22nd International Conference on Very Large Data Bases* (pp. 63-74), Bombay, India.

Lehdonvirta, V., & Ernkvist, M. (2011). Converting the virtual economy into development potential: knowledge map of the virtual economy. *InfoDev/World Bank White Paper*, *1*, 5-17.

Lenzerini, M. (2002). Data integration: a theoretical perspective. In *Proceedings of the 21st Symposium on Principles of Database Systems (PODS)*, Wisconsin, USA, 233-246.

Lock, M. (2012). Data management for BI: big data, bigger insight, superior performance. *Aberdeen Group White Paper*, *1*, 4-20.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. (2011). Big data: the next frontier for innovation, competition, and productivity. *McKinsey Global Institute Reports*, *5*, 15-36.

Rahm, E., & Hai Do, H. (2000). Data cleaning: problems and current approaches. *Bulletin of the Technical Committee on Data Engineering*, *23*(4), 3-13.

Reddy, V., & Jena, S. (2010). Active datawarehouse loading by tool based ETL procedure. In *International Conference on Information and Knowledge Engineering (IKE'10)*, Las Vegas, USA, 196-201.

Slack, E. (2012). *What is big data?* Retrieved August 5, 2012, from http://www.storage-switzerland.com/Articles/Entries/2012/8/3_What_is_Big_Data.html

Smith, D. (2012). *Big data to add £216 billion to the UK Economy and 58,000 new jobs by 2017*. Retrieved August 3, 2012, from http://www.sas.com/offices/europe/uk/press_office/press_releases/BigDataCebr.html

Terence, C., & Ludloff, M. (2011). *Privacy and big data*. Sebastopol: O'Reilly Media.

Vassiliadis, P., Simitsis, A., Georgantas, P., Terrovitis, M., & Skiadopoulos, S. (2005). A generic and customizable framework for the design of ETL scenarios. *Information Systems*, *30*(7), 492-525. <http://dx.doi.org/10.1016/j.is.2004.11.002>