THE INTERNATIONAL JOURNAL OF BUSINESS & MANAGEMENT

Data Management and the Efficacy of Big Data: An Overview

Rainer A. Sommer
Associate Professor, Department of Enterprise Engineering and Public Policy,
George Mason University, Founder's Hall, Arlington, USA

Abstract:
Enterprises exist to provide goods and services to a customer base. Whether in the private or public sectors, enterprises are composed of management hierarchies, as well as processes and technologies that serve the common purpose of developing, sustaining, and adding value to a customer. It is this latter concept (adding value) that has driven organizations to look ever deeper into their data in the hope of generating an added competitive advantage with which to enhance market impact or position through the use of Business Intelligence and Data Analytic techniques that primarily focus on the analysis of structured and unstructured data.

Keywords: *Business intelligence, data analytics, structured data, unstructured data, data lake*

1. Trusted Data and Data Type Definitions

Simply stated, business intelligence has many meanings, and they often depend on the domain in which are being described, and what type of data is being analyzed. Data can be organized into two distinct categories - Structured Data and Unstructured Data. In most enterprises structured data mostly represents traditional business data that is trusted and used to drive day-to-day operations; such as sales, accounting and human resource data. This data resides in very robust and secure relational databases and provides management the ability to operate on a continuing basis (Williams and Williams, 2003). General convention assumes that trusted data has the following characteristics[1].

- Authority: it is up-to-date and recognized as the reference copy of the relevant data,
- Authenticity: it is what it says it is and can be linked back to its source,
- Reliability: it can be trusted as a full and accurate representation of the relevant facts, transaction or business process,
- Integrity: it is complete, unaltered and preserves context and chain of custody, and
- Usability: it is accessible, and can be located, retrieved, presented and interpreted.

Unstructured data on the other hand is ancillary to running the enterprise. It is often captured as a by-product of normal business transactions, and stored in non-relational data repositories (often referred to as a "Data Lake"). This data is usually also untrusted and is generated from the tracking of web-site clicks, capturing user or customer sentiments from online sources or documents such as social media platforms, bulletin boards, telephone calls, blogs, or forums – just to name a few. Most important to remember is that much of this data is unstructured and quite often simply gets a time stamp along with a general categorical qualifying ID known as Meta-Data, before it gets deposited into data lake storage. Much of this data may never get used, but it is kept in the data lake for possible future analysis.

Within this context, enterprises strive to develop efficiencies within their Intra- and Inter- organizational business activities (Sommer, 2017)

Intra-organizational business activities are primarily focused on business processes that effectively support the unfettered flow of goods and information across internal organizational silos (or stovepipes). Hence, there is a great deal of emphasis placed on optimizing locally controlled business processes.

Inter-organizational business activities are more complex, since this model requires the enterprise to extend its internal processes beyond its traditional organizational boundaries to include other claimants. This model is often termed as an "Extended Enterprise" because it attempts to leverage integrated internal (Intra-organizational) efficiencies into a value chain that includes external suppliers and customers. This model would allow customer and suppliers to take advantage of the efficiencies created within the host organization to achieve an aggregate competitive advantage[2]. The extended enterprise is a model that expands the basic enterprise to include customers, suppliers, partners, and other corporate claimants. By definition, it also extends the traditional enterprise model to include new business models such as e-Commerce, Supply Chain Integration and Management, and Customer Relationship Management.

[1] International Organization for Standardization, ISO 15489-1:2001, *Information and documentation— Records management—Part 1: General*, 2001

[2] This is often referred to as the "Amazon Model" because the company has consistently included, customers, suppliers, and even competitors into its value chain.

Regardless of what types of business model an organization develops, the ultimate efficiency of the enterprise is governed by a complex balance of process flow, data access and integrity, regulatory and policy control, and business intelligence. Hence the proper management of data becomes a strategic competitive advantage (Schniederjans and Schniederjans, 2014)

2. The "Managing" of Big Data Management

Big Data Management is the organization, administration and governance of large volumes of both structured and unstructured data. Hence, the organization must address the management of data with the same zeal and caution as any other strategic asset that provides a long-term competitive advantage. Generally, a data management plan must at a minimum address the development, execution and supervision of various plans (i.e., backup/recovery, disaster, access, etc.), policies, programs and practices that control, protect, deliver and enhance the value of data and information assets. To aid in this process, the Data Management Association (DAMA) has provided the following framework to help guide the implementation of such a management structure[3].

- Data Governance,
- Data Architecture Management,
- Data Development,
- Database Operations Management,
- Data Security Management,
- Reference & Master Data Management,
- Data Warehousing & Business Intelligence Management,
- Document & Content Management,
- Metadata Management, and
- Data Quality Management.

The DAMA framework divides implementation and execution responsibilities among both management and IT functions and whose ultimate goal is to allow companies to locate valuable information in large sets of structured, unstructured, and semi-structured data from a variety of sources, including traditional relational data as well as forum posts, images, blogs, texts, websites, clicks, cookies, e-mail, pop-ups, spam, and social media scans.

While each of these concepts merits its own investigation as to its overall contribution to organizational efficiency, increasingly the focus has turned to big data analysis for finding the next level of competitive advantage. Not surprisingly, due to the complexity of their vast business relationships, companies that have embraced inter-organizational business models have often taken the lead in leveraging large scale data analysis to support their rapidly changing value chains. Yet here-in lies the problem; even though "Big Data", "Data Mining", "Analytics", and "Business Intelligence" have been part of our professional and academic lexicon for many years, their underlying purpose and use is not very well understood.

3. Business Intelligence and Discovery Analytics: Two Distinct Views

Simply stated, business intelligence has many meanings, and they often depend on the domain in which are being described, and what type of data is being analyzed (Herschel and Jones, 2005), At its core there are two very distinct views about "Big Data" that must be explained (Figure 1).

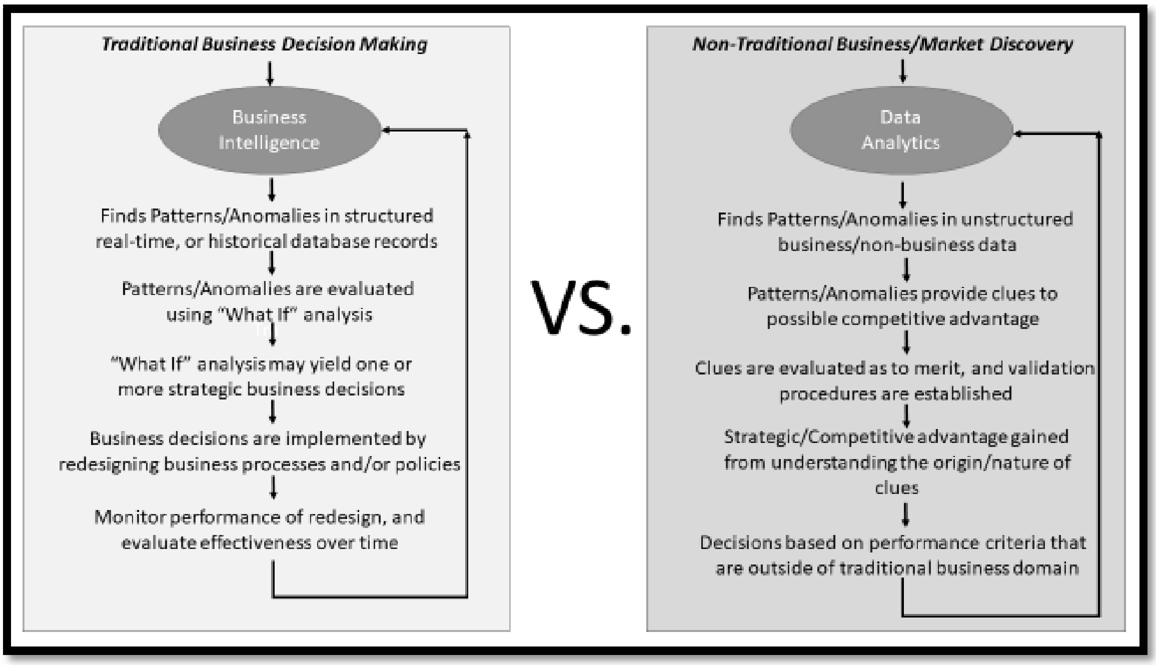
Traditional Business Decision Making
Business Intelligence
Finds Patterns/Anomalies in structured real-time, or historical database records
Patterns/Anomalies are evaluated using "What If" analysis
"What If" analysis may yield one or more strategic business decisions
Business decisions are implemented by redesigning business processes and/or policies
Monitor performance of redesign, and evaluate effectiveness over time
VS.
Non-Traditional Business/Market Discovery
Data Analytics
Finds Patterns/Anomalies in unstructured business/non-business data
Patterns/Anomalies provide clues to possible competitive advantage
Clues are evaluated as to merit, and validation procedures are established
Strategic/Competitive advantage gained from understanding the origin/nature of clues
Decisions based on performance criteria that are outside of traditional business domain

Figure 1

[3] Data Management Association (DAMA) – www.dama.org

- Traditional Business Intelligence requires an organization to "know more about what they currently know".
- With Data Discovery organizations are looking for "things" they don't know, and may not have the questions "yet;" e.g., the discovery exercises generate "clues"

Since data can be organized into two distinct categories - Structured Data and Unstructured Data, in most enterprises structured data mostly represents traditional business data that is trusted and used to drive day-to-day operations; such as sales, accounting and human resource data. This data resides in very robust and secure relational databases and provides management the ability to operate on a continuing basis. As mentioned previously, unstructured data is ancillary to running the enterprise. It is a by-product of normal business transactions, and stored in the data-lake. Being unstructured, the data is generated from ancillary sources such as web-site clicks, bulletin boards, telephone calls, blogs, or forums – just to name a few. Since the data is not necessarily critical to daily business operations, much of it may never get used, but it is kept in the data lake for future use, or for specialized projects (Figure 2).

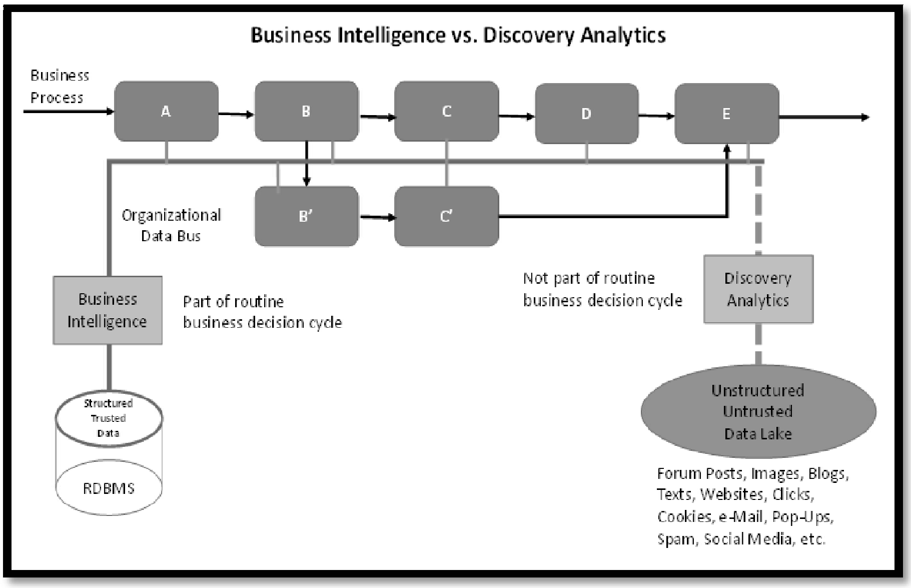
Business Intelligence vs. Discovery Analytics
Business Process
A
B
C
D
E
B'
C'
Organizational Data Bus
Business Intelligence
Part of routine business decision cycle
Not part of routine business decision cycle
Discovery Analytics
Structured Trusted Data
RDBMS
Unstructured Untrusted Data Lake
Forum Posts, Images, Blogs, Texts, Websites, Clicks, Cookies, e-Mail, Pop-Ups, Spam, Social Media, etc.

Figure 2

4. Finding a Competitive Advantage

4.1. Business Intelligence (BI)

From a historical business perspective, the analysis of structured data in making business decisions is a well-researched and known quantity. Thousands of tools and numerous methodologies exist to help businesses extract, transform and load (ETL) structured relational data into formats and data packages that can be distributed among any number of systems to develop an enterprise wide business solution. This is also the domain that is highly integrated around the concept of Business Intelligence (BI). Business Intelligence leverages the formal and well-defined ETL process to allow managers to pose and run "what-if" decision-making scenarios by extrapolating current trusted structured data forward into proposed future business cycles (Ranjan, 2009). Regardless of whether the future direction has to do with demand, profitability, outsourcing, mergers and acquisition, or any combination of other variables; the most relevant input data for the scenario usually comes from a trusted structured source. Hence BI concepts (and related tools, optimizers, rendering application platforms, etc.) are designed and optimized around a traditional and well understood relational database warehouse model (Chen, et al, 2012).

4.2. Discovery Analytics (DA)

Conversely, unstructured untrusted data sitting in a data lake can't necessarily be extracted, transformed, and loaded (ETL) into a relational table format. That doesn't mean it's impossible, but considering how much unstructured data can be generated on an hourly, daily, and monthly basis, it would be quite expensive to attempt such a project on a large scale. The better solution (mostly from a time and money perspective) would be to take manageable portions of the data lake and analyze that data for anomalies that could help support a business-related hypothesis, or point to a pattern of behavior that could possibly provide a competitive advantage in current or future business cycles. This then is the domain of Discovery Analytics (DA).

With Discovery Analytics, an organization is essentially placing an informed bet that there is a high probability that the expense associated with collecting and storing vast amounts of unstructured and untrusted data will yield a competitive advantage at some point in time. Essentially DA is used to mine vast portions of the data lake for randomly occurring patterns. The odds of finding a pattern that can actually provide a competitive advantage are relatively low – UNLESS, the data lake has the capacity to capture the smallest and most trivial of data. The reasoning is simple. The more data there is in the lake, the better the odds of finding a random pattern that could pay off and take the organization to the next level of efficiency and profitability (Agiu, et al, 2014)

Unlike structured data management paradigms that rely on traditional ETL methods to align critical elements of the business process, unstructured data is rather difficult to align with traditional data management architectures. To make the domain even more complicated, many of the terms and constructs used to define structured data management may have a different meaning when describing the dynamics of managing unstructured data. Therefore, business functions that rely on some form of unstructured data may first have to be mapped to a procedural model that defines the goal of the exercise (Figure 3).

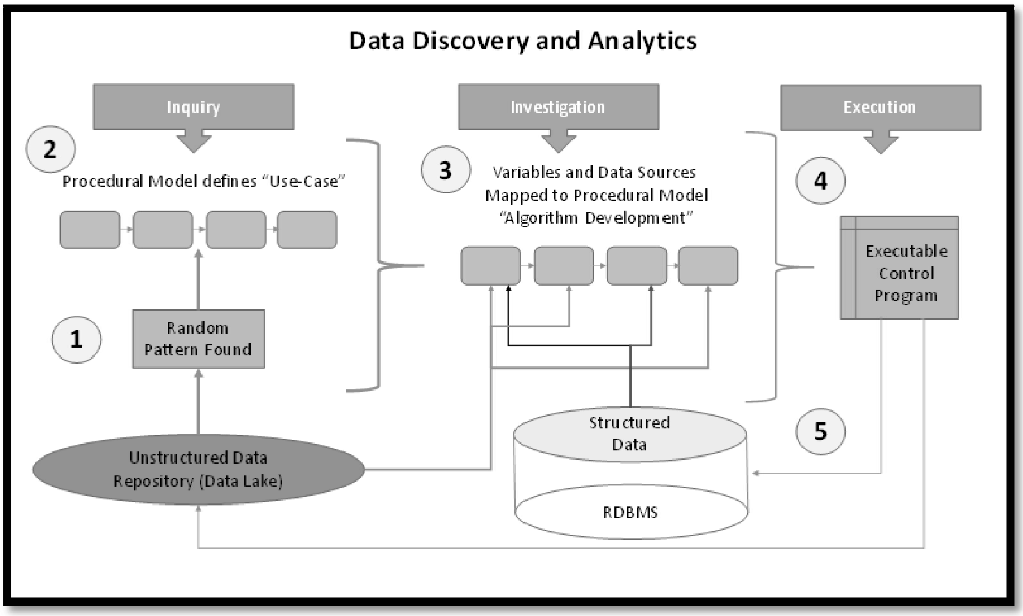
Data Discovery and Analytics
Inquiry
Investigation
Execution
2
Procedural Model defines "Use-Case"
3
Variables and Data Sources Mapped to Procedural Model "Algorithm Development"
4
Executable Control Program
1
Random Pattern Found
Structured Data
5
Unstructured Data Repository (Data Lake)
RDBMS

Figure 3

This is often referred to as building the "use-case" for the problem. For example,in the "Inquiry" phase a random heat-map pattern of cell phone usage generated from unstructured data (such a specific type ofimage upload /download activity) may prompta law enforcement team to seek further investigation into human traffickingor gambling activities. Based on this evidence an "Investigation" phase may be called for, and theteam will build a use-case procedural model to help bound the project. The model will outline the specific steps and procedures the project team thinks will be required to successfully complete this part of the investigation. This use-casewill then serve as the basis for developing the "Execution" phase where a control program (an executable algorithm) is developed, and which defines the variables that must be included in supporting the data requirements of the procedural model. The procedural model may define the problem in any number of dimensions (i.e., cultural, ethnic, socio-economic, and local consumer preference, etc.).The variables to support these dimensions may consist of both structured and unstructured data, and their relationship(s) to each other are defined within the structure of the control program. The output of the execution phase control program may help to pin-point human trafficking and organized crime activities within a very precisely defined geographic boundary.

5. Big Data Technical Challenges

Needless to say, the amount of data generated in today's global business environments are staggering. Inter- and intra-organizational processes generate vast amounts of structured and unstructured data. While structured trusted data can be culled and managed in relational databases, unstructured data is by definition "unruly" in its inability to be easily classified and categorized, and in the sheer volume in which it is generated. When collecting clicks, cookies, phrases, documents, images, texts, emails, spam, posts, etc. that are related to an organization' s dealings with the world at large, the store and maintenance costs of data are enormous. That cost is often justified with the premise that the data "may" at some point in time reveal a game changing competitive advantage. Until that time comes, storage and management costs are steadily climbing. Not surprisingly, many senior managers are questioning the continued use of data analytics because they see relatively few returns from the investment. Research suggests that they are correct in their assessments, and points to the fact that many companies don't know how to exploit the data already embedded in their core business systems.

6. Conclusion

The biggest problem with Business Intelligence and Data Analytics/Discovery is that there are no absolutes – especially in the case of the later technology (Data Analytics/Discovery). It remains up to the experience and insight of managers to determine which DA generated patterns show promise (hence necessitating further investment and investigation), and which patterns are to be ignored. As is often the case in businesses, "the art" of successful management is not something can be taught or readily formulated. When presented with a series of DA outputs, Warren Buffet would most likely key in immediately on a pattern that makes perfect sense to his superior investor instincts, while I on the other hand would look at the same pattern and say "so what!". Therefore, the true competitive advantage lies in the ability to interpret the DA output and proactively develop an effective business solution as a response. It can be argued that very few people have the insight to capitalize on DA generated patterns. Even when presented with a "goldmine" solution from a DA platform, it remains elusive unless someone can recognize it for its potential.

7. References

i. Agiu, D., Mateescu, V., Muntean, J. (2014) Business Intelligence overview. Database Systems Journal, 3, 23-36.

ii. CGMA® REPORT BUSINESS ANALYTICS AND DECISION MAKING: The Human Dimension (2016). Chartered Institute of Management Accountants.

iii. Elliott, T. (2003) Implementing a Business Intelligence Strategy! A Practical Guide to Business IntelligenceStandardization. White Paper, businessobjects.com

iv. Herschel, R.Y, Jones, N.E., (2005) Knowledge management and businessintelligence: the importance of integration. Journal of Knowledge Management, 9, 45-55.

v. Hsinchun, C., Chiang, R., & Storey, V.C. (2012) BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. MIS Quarterly, 36, 1165-1188.

vi. Ranjan, J. (2009) Business Intelligence: Concepts, Components, Techniques, and Benefits. Journal of Theoretical and Applied Information Technology, 9, 60-70.

vii. Schniederjans, M. J., Schniederjans, D. G. & Starkey C. M. (2014) Business Analytics Principles, Concepts, and Applications: What, Why, and How. Upper Saddle River, NJ; Pearson Education, Inc.

viii. Sommer, R.A., (2017) Public Sector Change: Understanding Vertical and Horizontal Integration, International Journal of Business & Management (IJBM), 5, 82-86.

ix. Williams, S., & Williams, N. (2003) The Business Value of Business Intelligence. Business Intelligence Journal. Decision Path Consulting, 1-11.