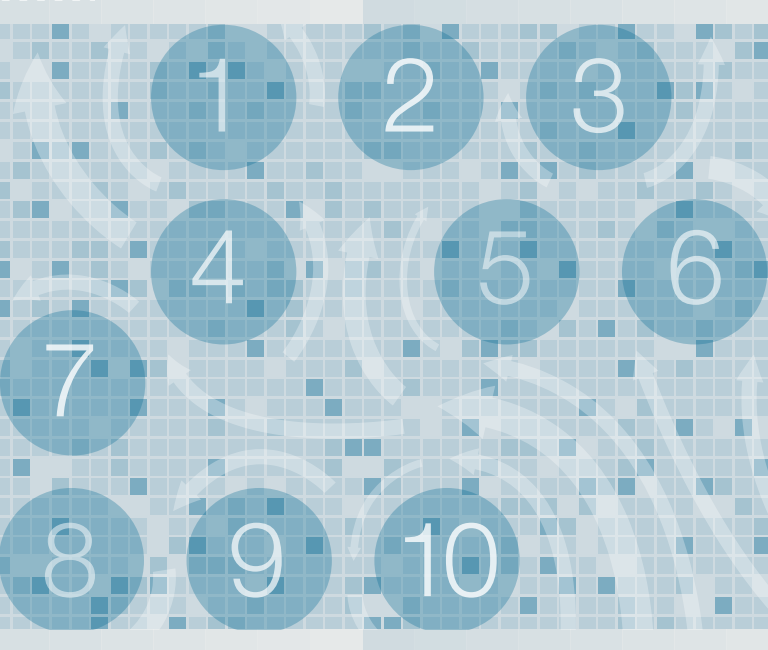
EXCLUSIVELY FOR
TDWI PREMIUM MEMBERS

FIRST QUARTER 2013

TEN MISTAKES TO AVOID

In Your Big Data Implementation

By Krish Krishnan



tdwi.org

tdwi

TEN MISTAKES TO AVOID

In Your Big Data Implementation

By Krish Krishnan

FOREWORD

Big data is the biggest buzzword in the industry today. Every organization—big or small—is looking into understanding and deploying a big data program. Big data doesn't just refer to having larger volumes of data. We must consider the source(s) of the data.

One purpose of a big data implementation is to incorporate additional data sets into the current data infrastructure to help the enterprise question anything from the data. Although the possibility of accomplishing this goal seems realistic with the evolution of technology and commoditization of an enterprise's infrastructure, there are several critical pitfalls to avoid. In this *Ten Mistakes to Avoid*, we will look at the most common mistakes that occur when implementing a big data program to help you enhance your analytical insights and the decision support processes in your enterprise.

ABOUT THE AUTHOR

Krish Krishnan is the founder and president of Sixth Sense Advisors, Inc. He is an expert in the strategy, architecture, and implementation of big data, text analytics, and high-performance data warehousing. He wrote *Building the Unstructured Data Warehouse* with Bill Inmon. His new book, *Data Warehousing in the Age of Big Data*, will be published soon. Krish teaches at TDWI and speaks at many conferences worldwide. You can follow him on Twitter: @datagenius.

MISTAKE ONE:
LACK OF A BUSINESS CASE

Big data is an extremely complex subject. To understand it, you will need a proper business case to determine the potential benefit of incorporating such data into your enterprise decision support platform. The business case must clearly identify the gaps in your data and platform.

Consider the following business case for incorporating social media data for brand monitoring:

XYZ Logistics is a leader in shipping and logistics management. The company's business has declined because of service and performance issues. In conducting research about the recent loss of customers and revenue, the company identified the need to monitor social media for trends and sentiments that directly and indirectly impact its brand; such data can provide valuable insights into the expectations of its customers. Among the critical aspects for this program:

Social media data: This data can provide geospatial information about the customer or prospect, sentiment, and processes that fail to engage the customer.

Metrics: XYZ Logistics needs to determine direct influence (friends), indirect influence (reach and amplification), geographies impacted, brand and competitive analysis, and the number of fans for each of its customers or prospects. This will help leverage better models for campaigns and other efforts.

Value: XYZ Logistics can measure the value when social media data analytics is integrated with existing enterprise analytics, including customer analytics, marketing analytics, sales analytics, and campaign analytics.

We can see from this little exercise the impact of establishing the business case and associated value from big data. Without the appropriate business case, it is nearly fatal to engage in big data programs. Such an exercise is mandatory to prove the short- and long-term value and return on investment.

MISTAKE TWO:
DATA THAT LACKS RELEVANCE

Big data is available in all shapes and sizes. Knowing the relevance of each of these data sets to your business needs is a key success factor with big data. Categories of big data available today include:

- **Unstructured data**, including text, video, audio, and images
- **Semi-structured data**, such as e-mail messages, earnings reports, spreadsheets, and software modules
- **Structured data**, including sensor data, machine data, actuarial models, financial models, risk models, and other mathematical model outputs

As an enterprise, you will have access to all of these data sets. The question is whether you know each data set's relevance to your enterprise analytics. Let's look at an example in a call center. Every call center has a mandatory message that states, "This call may be recorded for quality or training purposes." There are hundreds of sentiments, competitive analysis, service-level issues, and cost-related discussions that a customer can have in any conversation with a call center representative. This data, when converted from audio to text and incorporated into the data warehouse, can be extremely useful in augmenting your business intelligence with sentiment analytics, competitive research, and performance analytics—along with rich context—to provide holistic insight based on your enterprise needs. Without the appropriate context and relevance, the analytics can be skewed heavily by the additional data.

MISTAKE THREE:
LACK OF DATA QUALITY

Lack of data quality can ruin analytics in any organization. With big data, overall data quality can degrade as you integrate unstructured and semi-structured data. Although data quality is an important issue to understand and resolve prior to processing big data, you must determine how to improve the quality of data that may not be generated or owned by your organization.

In the case of unstructured data, text data quality can be improved by using language correction libraries prior to processing. If languages must be translated, then user inputs can provide the appropriate contextualization rules as needed for each linguistic connotation in speech or text. In the case of image and video files, data quality is determined at the source. If the data is sourced from Internet sites or third parties, you can use semantic libraries, taxonomies, and ontologies with user inputs to improve the quality of data.

For semi-structured data with text or numeric values, correct the data as you would textual data. User inputs are critical to ensure the validity of the data and its context.

Improving the overall data quality is an important consideration for processing big data. Although this is a tedious exercise in many cases, without this step the output produces skewed results and will negatively impact the analytical systems in the enterprise.

MISTAKE FOUR:
INSUFFICIENT DATA GRANULARITY

Big data is ambiguous by nature. There is no clear definition of the "grain" of data present within the data. You discover and learn the granularity as you process the data sets. The biggest weakness you may find is that you cannot process metrics and associated levels of hierarchies with the metrics you have, whether you are working with structured or semi-structured data. These two types of data constitute the largest portion of any organization's big data.

If the data processing does not identify the appropriate level of granularity, then you increase the chances of producing an erroneous result set and skewing the analytical outputs. Processing unstructured data requires that your hierarchy definitions are available and elastic. The reason for the elastic definition of hierarchies arises because you might encounter jagged and rolled-up data in the same data set, and associating the wrong grains of data into relationships can create several kinds of errors in the analysis and integration processes.

A classic example is processing user sentiments from a Web forum. Consider this post by a user about an SLR camera:

"As a Nikon user and owner, I don't know if I'd get this. One of the biggest pluses of being in the Nikon system is its flash and CLS system. But they have crippled this interchangeable lens camera with basically a proprietary hot shoe. I have an SU-800 and five different flashes, and can't use them on this camera. Would an ISO hot shoe have been so tough? There would be a lot more creative possibilities if the Nikon 1 system were CLS compatible."

It is tough to discern what the user has written about when processing this text because so many cameras and associated features are mentioned; the user also discusses an accessory and some standards in camera technology. This ambiguity is where the hidden layers of granularity of big data are found. If you were to process this grain of data with the hierarchy of cameras versus hierarchy of accessories, you would generate different results with varying degrees of accuracy because the hierarchy is not clearly defined. It's also ambiguous when you look at the granularity as it pertains to different subject areas.

MISTAKE FIVE:
MISSING DATA CONTEXTUALIZATION

The fundamental logic behind processing textual data and executing text analytics lies with contextualization of the data. Without proper contextualization, we will inaccurately process the data and skew the analysis. Consider the following example of a doctor's note about a patient.

If you examine notes from hospital charts, you will find doctors using common shorthand notations, such as "HA" and "EP." Processing the data without expanded notation is not useful for metrics when looking at patients with a particular pattern of disease states, treatment options, or drug interactions. When a cardiologist uses HA, the acronym stands for "heart attack." When a nephrologist uses HA, the acronym means "hyperactive bladder."

Without contextualizing the business rules for processing each specialist's notations, we will end up with errors in the result set. For example, the result might reflect that the patient, who has both a heart condition *and* kidney issues, had a heart attack in the kidney or an excessively active bladder in the heart, both of which are nonsense. Although there are several additional steps with text analytics that need to be processed beyond contextualization (such as homographs, alternate spelling, and categorization) to improve the accuracy of the data and derive value from processing the data, the key business rule to process is the contextualization of the data.

MISTAKE SIX:
NOT UNDERSTANDING DATA COMPLEXITY

Big data has multiple layers of complexity that are not visible through simple inspection by an end user. These complexities are present in the data itself because of its structure, format, content, and associated metadata. Without understanding the complexity, a model of a solution for the data set (whether statistical, mathematical, or text mining) will create erroneous results. The complexity is compounded by the fact that metadata is sparse with the data itself and multiple formats can cause issues when analyzing the data (not when storing the data).

Consider consumer sentiment from Twitter, Facebook, and Web forums. The attributes of this text will help define a consumer and identify the brand and sentiments expressed by the consumer. The issue, however, is that Twitter data is cryptic and restricted to 140 characters; Web forum data is verbose, spanning multiple lines. This difference in the amount of information, when analyzed, needs multiple cycles of processing for one data set (i.e., a cycle for Twitter and a different set of processing cycles for a different data set such as a Web forum). A tweet to note a service failure will include "@brand/service" or "process #fail" whereas a Web forum will contain a fully formed sentence, such as "Very disappointed with the [brand name] and [service name]—they have no regard for the consumer."

These hidden complexities of data format and language are critical to understand when processing textual, semi-structured, and semantic-layer-dependent big data. Failure to analyze and define the appropriate business rules will result in skewed output results.

MISTAKE SEVEN:
POOR DATA PREPARATION

Big data processing requires you to prepare the data prior to and during the processing cycles and to provide additional inputs as needed for taxonomies and metadata. Failure to execute the preparation steps will skew the results of processing big data. For example, processing log files will help you understand the need for data preparation.

Weblogs or machine logs have a fixed format and field layout that can be useful in analyzing the behavior of products, machines, and human/machine interactions; the logs can also be useful in associating the behavior with enterprise analytical platforms for visualization. However, there are a few steps developed by every organization that address how the data needs to be named, enriched, associated with metadata, and parsed. These steps must be followed to ensure that data is ready for processing; the steps must be completed in the preparation stage of processing the data into the analytical systems and the operational data store. Pay special attention to the date/time format, relevance to master data or metadata, ambiguous data, and column values.

If you do not allow adequate time and appropriate processes for data preparation, you may end up with data issues in your program.

MISTAKE EIGHT:
ORGANIZATIONAL IMMATURITY

The success of any program related to data and analytics is aligned with the team that owns and drives the program, as well as with the maturity of the team in terms of domain and data knowledge. This is also true of big data and analytics programs. Business owners and data experts must be aware of what big data means to their line of business, how they will treat the new data sets, and what the results mean to the organization.

Without such maturity of thought and clarity of requirements, the overall success of the integration of big data into the analytical systems is highly questionable. Some of today's organizations have attained a high level of maturity in a particular line of business in terms of big data and analytics, and these teams will evolve to become the new leaders for enabling the same success across their organizations.

One word of advice: if you are embarking on a big data project, do not let IT drive the program. This program is *for* the business and *about* the business and needs to be owned and defined *by* the business. This is the first step of the organizational maturity that is required for big data success. Without business involvement and drive, analytics for your big data integration project will suffer, as will quality, dedication, and high-quality decisions and insights.

Before business users are ready to use the program, they must be mature in their thought process about what they expect from the data and how it will enhance the associated analytics. Another indication of organizational maturity is that IT teams recognize the need to let go of the program and become facilitators instead.

MISTAKE NINE:
LACK OF DATA GOVERNANCE

Data governance is the lynchpin for the success of enterprise data integration and management of data across its life cycle. Big data is no exception when it comes to data governance; it needs to be treated as another data set within the enterprise that requires stewardship and associated processes for managing the data to be designed, developed, and deployed.

Fundamentally, the processes associated with the governance of traditional data—stewardship, program governance, business rules, data quality, and master data management—can be extended to big data, with additional focus on metadata, business rules, and semantic library integration. The challenges in governing big data include the complexity associated with the processing of data, the business rules definition for processing the data, and the multi-owner-based stewardship of the data, where conflicting requirements will be the norm. We have already mentioned several areas of big data governance; one additional point is to remember that big data integration projects will need sponsorship and executive guidance.

Without the right type of data governance, your program will not only fail but may also damage the existing analytical programs that have been deployed by skewing results and decreasing user confidence about the value of integrating big data.

MISTAKE TEN:
BELIEVING TECHNOLOGY IS A SILVER BULLET

The biggest hype in the industry today is that Hadoop is the panacea for all issues related to data. There are already strong rumblings that Hadoop is a legacy technology for the big data companies and that more innovations are on the way.

Every time a technology tipping point helps solve a data problem, a new class of data problems evolves along with it. In the case of big data, the problem with open source platforms is the maturity of the technology to support enterprise-scale deployments as the platforms evolve as an ongoing ecosystem.

Do data warehouses have a future? If we define the data warehouse as simply the relational database management system (RDBMS), then we might assert there is a future for the RDBMS platform. However, the new data warehouse is a combination of the RDBMS, Hadoop, NoSQL, and other technologies. This heterogeneous approach is the "new normal" and is here to stay.

Most big data technology evolved between 2005 and 2012 into mainstream, though these technologies have been in incubation for more than a decade. This newness needs to be considered when understanding the architecture and placeholder for the technology in the enterprise framework.

Not realizing the maturity of the technology and its fit to the enterprise is a big mistake. The solutions from the big data stack can be effectively integrated into the enterprise for the right purpose; otherwise, the project will yield minimal benefits, and, more important, can result in misguided analytical processing, leading to more chaos in the enterprise. Technology is not a "silver bullet" for your big data program.

As you can see, due to the complexities of the data apart from its volume, velocity, and variety, there are several risks associated with implementing a big data program. Careful planning and learning can help every big data program be successful.

ABOUT TDWI

TDWI, a division of 1105 Media, Inc., is the premier provider of in-depth, high-quality education in the business intelligence and data warehousing industry. TDWI is dedicated to educating business and information technology professionals about the best practices, strategies, techniques, and tools required to successfully design, build, maintain, and enhance business intelligence and data warehousing solutions. TDWI also fosters the advancement of business intelligence and data warehousing research and contributes to knowledge transfer and the professional development of its members. TDWI offers a worldwide membership program, five major educational conferences, topical educational seminars, role-based training, on-site courses, certification, solution provider partnerships, an awards program for best practices, live Webinars, resourceful publications, an in-depth research program, and a comprehensive website, tdwi.org.

1201 Monster Road SW
Suite 250
Renton, WA 98057-2996

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org