10 Mistakes Enterprises Make in Big Data Projects

Avoid common pitfalls when planning, creating, and implementing big data initiatives

By Krish Krishnan | Published February 28, 2014

Big data and big data analytics are ubiquitous in the industry today. Organizations of every stripe are looking to understand and deploy a big data program to improve business outcomes. But big data is not just about large volumes of data. The sources of all this data should be considered. One purpose of a big data implementation is typically to incorporate additional data sets into the current data infrastructure, which gives organizations the capability to question anything from their data set.

Editor's note: *This article by Krish Krishnan, president and CEO at Sixth Sense Advisors, Inc., is offered for publication in association with The Big Data Seminar 2014 in New York City, March 13–14, 2014, featuring the author, and the Big Data and Enterprise Architecture Conference, June 11–13, 2014, in New York City featuring Bill Inmon. Data Management Forum sponsors both conferences.*

While accomplishing this goal seems realistic given the progression of technology and the commoditization of infrastructure, there are 10 common pitfalls that enterprises, in particular, need to avoid when planning and implementing a big data program. By avoiding these drawbacks, outcomes can enhance an organization's analytical insights and decision support processes.

1 Lacking a business case

Big data is an extremely complex subject area to understand without using a proper business case associated with the value that organizations can derive when incorporating big data into the enterprise decision support platform. The appropriate business case should include a clearly developed requirement for the gaps. Consider the following business case for incorporating social media data for brand monitoring:

Background: YYY Logistics is a leading-edge shipping and logistics organization where management has recently seen a slump in its business because of service and performance problems.

Business case: In conducting research about the recent loss of customers and business, the organization discovered it needed to monitor social media for trends and sentiments that directly and indirectly impact its brand. Monitoring social media would provide valuable insights into its customers' expectations. The following aspects were deemed critical for this program:

- **Social media data** can provide geospatial information about Twitter or Facebook users, their sentiments, and processes that caused the organization to fail them in some way.
- **Metrics** offer direct influence (friends), indirect influence (reach and amplification), impacted geographies, brand and competitive analysis, and number of fans.
- **Value** enables measuring the value delivered when the social media data analytics are integrated with existing enterprise analytics including customer analytics, marketing analytics, sales analytics, and campaign analytics.

In summary, the YYY Logistics organization believes this exercise is mandatory to help regain customer confidence and market share.

2 Minimizing data relevance

Big data is available all around us in various shapes and sizes. Understanding the relevance of each of these data sets to business needs is a key aspect to succeed with big data initiatives. The following categories of big data are available today:

- **Unstructured data** includes text, videos, audio, and images.
- **Semistructured data** includes email, earnings reports, spreadsheets, and software modules.
- **Structured data** includes sensor data, machine data, actuarial models, financial models, risk models, and other mathematical model outputs.

Enterprises need to have access to all these data sets, but do they understand their relevance to their analytics? Consider the following example from a call center's recordings and customer service. Every call center has a mandatory message that states, "This call may be recorded for quality or training purposes." In any conversation a customer engages in with a call-center

representative, there are tons of sentiments, competitive analyses, service-level issues, and cost-related discussions.

When converted from audio to text and processed, this data can be extremely useful in augmenting sentiment analytics, competitive research, and performance analytics. It also offers rich context to provide holistic insight into any particular area based on the needs of the enterprise. Without the appropriate context and relevance, the analytics can be skewed heavily by the additional data.

3 Underestimating data quality

Data quality is a highly significant consideration. Poor quality can ruin analytics in any organization. For big data, overall data quality can degrade as unstructured and semistructured data are integrated into data sets. While understanding the impact of data quality and taking the appropriate steps to resolve problems prior to processing big data are extremely important, organizations need to know how to improve data quality for data that it may not own or have generated.

For unstructured data, quality of text data can be improved by using language correction libraries prior to processing. If there is language translation involved, then end-user inputs are needed to provide the appropriate contextualization rules required for each linguistic connotation in speech or text. Image and video data quality are acquired from the source. If the data is sourced from Internet sites or from third-party sources, organizations can use semantic libraries, taxonomies, and ontologies with end-user inputs to enhance the quality of this data.

Semistructured data that has text or numeric values can be processed like textual data for correction. There are a lot of end-user inputs needed to ensure the validity of the data and its context.

Improving data quality is an important consideration for processing big data. Without taking this step, the output often results in skewed results and can negatively impact the analytical systems in the enterprise.

4 Overlooking data granularity

Because big data is ambiguous by nature, there is no clear definition for the grain of data that is present within the data. Organizations discover and learn the granularity as they process the data sets. A significant weakness that can confront them in this scenario is the inability to process metrics and associate levels of hierarchies with the metrics. This situation is particularly true when handling text and semistructured data. These two types of data definitely make up a large portion of big data in any enterprise.

If the data processing does not identify the appropriate level of granularity, then the chances are high an erroneous result set will occur that skews the analytical outputs. Processing unstructured data requires that hierarchy definitions are available and elastic in nature. The elasticity of

hierarchies arises when organizations may 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-based forum. A customer considering a recently released lens for a specific camera, for example, may write the following post in a forum catering to enthusiasts of the camera:

> *I will not purchase this lens because a big advantage of this camera is its flash and lighting system, and yet the manufacturer has diminished its value by basically offering a proprietary accessory. I own other lenses and flashes that I can't use with the camera. The ability to purchase nonproprietary components would allow for more creative, compatible options.*

When processing the text of this post, discerning what the person has written about is tough based on different camera components and associated features, plus some standards in camera technology. This ambiguity is where the hidden layers of granularity of big data come into the picture. If processing this grain of data with the hierarchy of cameras versus the hierarchy of accessories, different results with varying degrees of accuracy are likely to be obtained because the hierarchy is not defined clearly.

5 Improperly contextualizing data

The fundamental logic behind processing textual data and executing text analytics lies with contextualization of the data. Without proper contextualization, the data can be processed with a lot of inaccuracy and produce skewed analytics. Consider a doctor's notes on hospital charts, for example. Doctors tend to use common shorthand notations such as "HA" and "EP." Processing the data without an expanded notation is not useful for metrics when looking at patients with a particular pattern of disease states, treatment options, or drug-to-drug interactions. When a cardiologist uses "HA," it stands for "heart attack," but when a nephrologist uses the term, it means "hyperactive bladder." Without contextualizing the business rules for processing each specialist's notations, garbage data sets result.

For example, the result might reflect that the patient, who has both a heart condition and kidney problems, had a heart attack in his or her kidney or an excessively active bladder in the heart, both of which are nonsensical. There are several additional steps with text analytics that need to be processed beyond contextualization such as homographs, alternate spellings, and categorization to create the accuracy of the data and to derive value from its processing. However, the key business rule for processing data is its contextualization.

6 Not grasping data complexity

Big data has multiple layers of hidden complexity that are not visible by simply inspecting it from an end-user perspective. The complexities are present in the data itself because of its structure and formats, content, and metadata. Without understanding the complexity, modeling a

solution for the data set—whether statistical, mathematical, or text mining—can create erroneous results.

Complexity is compounded by metadata being sparse with the data itself, and multiple formats can cause problems when analyzing the data, not when storing it. Consider consumer sentiment from Twitter, Facebook, and web forums. They all have attributes that help define a consumer and identify the brand and sentiments expressed by the consumer. The issue, however, is that Twitter data can be very cryptic and of course is limited to 140 characters, and data from web forums can be very verbose, spanning multiple lines.

When analyzing this varying degree of information, it requires multiple cycles of processing for one data set—Twitter—and a different set of processing cycles for another data set—web forums. A tweet to notate failure of service may be "@brand service or process #fail," and a web forum posting likely contains a fully formed sentence: "very disappointed with the brand (name) and service (name). They have no regard for the consumer and have failed to deliver any value." These hidden complexities of data format and language are key areas to understand when processing textual-, semistructured-, and semantic layer–dependent big data. Failure to analyze and define the appropriate business rules may result in skewed output results.

7 Ignoring data preparation

Big data processing requires organizations to prepare the data prior to processing and during the processing cycles, and provide additional inputs as needed for taxonomies and metadata. Organizations that ignore the preparation steps can skew the results. Weblogs or machine logs, for example, have a fixed format and field layout that can be useful in processing the data for analyzing the behavior of products, machines, and human-to-machine interactions. This format can also be used for associating these analyses with enterprise analytical platforms for visualization.

However, there are a few steps developed by every organization that govern how the data needs to be named, enriched, associated with metadata, and parsed. These steps must be followed to ensure data is ready for processing, and they are completed in the preparation stage of the data processing into the analytical systems and the operational data store (ODS). Special attention needs to be paid to the date and time format, relevance to master data or metadata, ambiguous data, or column values. If organizations do not allow adequate time for preparation of data for downstream processing, they may end up with downstream problems in the program.

8 Delaying organizational maturity

The success of any program related to big data and analytics is aligned with the team that owns and drives the program. The maturity of the team in terms of both domain knowledge and data knowledge to leverage the outcomes from the initiative can also impact success. This situation is the same with big data and analytics programs within any organization.

Business owners and data experts need to be aware of what the value of big data means to their line of business, how they will treat these new data sets, and what the results mean to the organization. Without this kind of maturity around thought processes and clarity of requirements, the overall success of integrating big data into analytical systems is highly questionable. There are organizations today that have attained a high level of maturity in a particular line of business in terms of big data and analytics, and those teams will evolve to become the new leaders for enabling the same success across the organization.

When embarking on a big data project, one piece of advice is to not let IT drive the program. This program is for the business, and it's all about the business. So it needs to be owned and defined by the business. This first step in organizational maturity is important for the success of big data initiatives. Without business involvement and drive, analytics for big data integration may suffer from quality malaise. Moreover, adoption as a result of the decisions made, insights gleaned, and the associated successes for the business may all be minimal.

Before business users are ready to use the program, they need to ensure that they know what to expect from the data and how it will enhance the analytics associated with the data. Another indication of organizational maturity is when IT teams recognize the need to let go of the program and instead be a facilitator for the business.

9 Forgoing data governance

Data governance is the lynchpin for the success of enterprise data integration and management of data across its lifecycle. Big data is no exception when it comes to data governance; it needs to be treated as another data set within the enterprise that needs stewardship and associated processes for managing the data to be designed, developed, and deployed. Fundamentally, the processes associated with the governance of traditional data from stewardship to program governance, business rules, data quality, and master data management (MDM) can be extended to big data. And these processes can also be extended with additional focus on metadata, business rules, and semantic libraries as integration areas of the process.

The challenges in governing big data include the complexity associated with the processing of data, which is a task in itself. Other issues include business rules definition for processing the data and the multi-owner–based stewardship of the data, in which conflicting requirements may be the norm.

There are several subtopics in big data governance that have been discussed previously, but the one additional point to remember is that big data needs sponsorship and executive guidance to gain acceptance within the enterprise. Without the right type of governance, any program may not only fail, but also may damage the existing analytical programs that have been deployed in the enterprise. In addition, the results can be skewed and create a lack of confidence in the minds of line-of-business users about the value of integrating big data.

10 Deploying technology as a silver bullet

A tremendous hype cycle in the industry today is about the Apache Hadoop framework being the panacea for all problems that are related to data. And yet there are already strong rumblings that Hadoop is a legacy framework for the big data companies, and more innovations are on the way. Every time a technology tipping point occurred to solve a data problem, a new class of data problems arose that evolved along with it. In the case of big data, the problem accompanying open source platforms is the maturity of the technology to support enterprise-scale deployments as the platforms evolve to an ecosystem on a continued basis.

Do data warehouses have a future? If a data warehouse is defined as just the relational database management system (RDBMS), there is a future and use for the RDBMS platform. But advanced data warehouses are a combination of the RDBMS, Hadoop, NoSQL, and other technologies. This heterogeneous approach is the new normal and is here to stay.

Much of today's big data technology has evolved into mainstream platforms, although they have been incubating for more than a decade. Consider these technology advances when understanding the architecture and placeholders for the technology in the enterprise framework.

The mistake in this area is not realizing the maturity of the technology and its fit within the enterprise. The solutions from the big data stack can be effectively integrated into the enterprise for the right purpose; otherwise, the exercise may result in minimal benefits. More importantly, it can result in misguided analytical processing that leads to more chaos.

Because of the complexities of big data—apart from its volume, velocity, and variety—there are several risks associated with implementing a big data program. But careful planning and learning can help every big data program become successful.