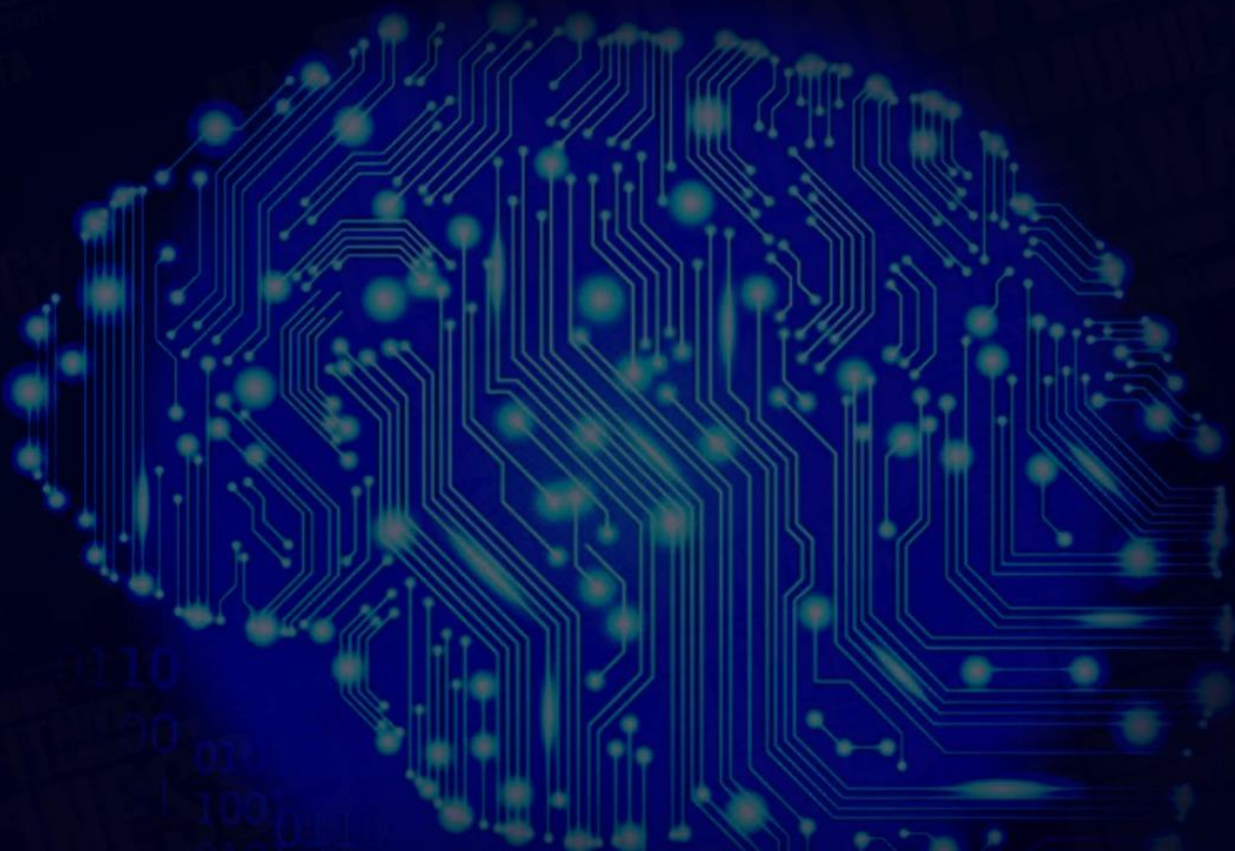# TekBusiness

Big Data Analytics

Four Key principals for extracting real-time business value from Big Data
3/5/17

Table of Contents

Introduction

Wal-Mart handles more than a million customer transactions each hour and imports those into databases estimated to contain more than 2.5 petabytes of data.

Radio frequency identification (RFID) systems used by retailers and others can generate 100 to 1,000 times the data of conventional bar code systems.

Facebook handles more than 250 million photo uploads and the interactions of 800 million active users with more than 900 million objects (pages, groups, etc.) – each day. More than 5 billion people are calling, texting, tweeting and browsing on mobile phones worldwide.

Organizations are inundated with data – terabytes and petabytes of it. To put it in context, 1 terabyte contains 2,000 hours of CD-quality music and 10 terabytes could store the entire US Library of Congress print collection. Exabytes, zettabytes and yottabytes definitely are on the horizon. Data is pouring in from every conceivable direction: from operational and transactional systems, from scanning and facilities management systems, from inbound and outbound customer contact points, from mobile media and the Web.

The explosion of data isn't new. It continues a trend that started in the 1970s. What has changed is the velocity of growth, the diversity of the data and the imperative to make better use of information to transform the business.

The hopeful vision of big data is that organizations will be able to harvest and harness every byte of relevant data and use it to make the best decisions. Big data technologies not only support the ability to collect large amounts, but more importantly, the ability to understand and take advantage of its full value.

What Is Big Data?

Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making.

Some of this data is held in transactional data stores – the byproduct of fast-growing online activity. Machine-to-machine interactions, such as metering, call detail records, environmental sensing and RFID systems, generate their own tidal waves of data. All these forms of data are expanding, and that is coupled with fast-growing streams of unstructured and semi structured data from social media.

5 V's of Big Data

Volume
The main characteristic that makes data "big" is the sheer volume. It makes no sense to focus on minimum storage units because the total amount of information is growing exponentially every year. That's a lot of data, but it is the reality for many organizations. By some estimates, organizations in all sectors have at least 100 terabytes of data, many with more than a petabyte.

Variability
Up to 85 percent of an organization's data is unstructured – not numeric – but it still must be folded into quantitative analysis and decision making. Text, video, audio and other unstructured data require different architecture and technologies for analysis.

Velocity
Thornton May says, "Initiatives such as the use of RFID tags and smart metering are driving an ever greater need to deal with the torrent of data in near real time. This, coupled with the need and drive to be more agile and deliver insight quicker, is putting tremendous pressure on organizations to build the necessary infrastructure and skill base to react quickly enough."

Veracity
Veracity refers to the trustworthiness of the data. Can the manager rely on the fact that the data is representative? Every good manager knows that there are inherent discrepancies in all the data collected.

Value
It may seem painfully obvious to some, but a real objective is critical to this mashup of the four V's. Will the insights you gather from analysis create a new product line, a cross-sell opportunity, or a cost-cutting measure? Or will your data analysis lead to the discovery of a critical causal effect that results in a cure to a disease?

Complexity
Difficulties dealing with data increase with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems. Organizations need to understand relationships, such as complex hierarchies and data linkages, among all data.

A data environment can become extreme along any of the above dimensions or with a combination of two or all of them at once. However, it is important to understand that not all of your data will be relevant or useful. Organizations must be able to separate the wheat from the chaff and focus on the information that counts – not on the information overload.

Rethinking Data Management

From Standalone Disciplines to Integrated Processes

Organizations are moving away from viewing data integration as a standalone discipline to a mindset where data integration, data quality, metadata management and data governance are designed and used together. The traditional extract-transform-load (ETL) data approach has been augmented with one that minimizes data movement and improves processing power. Organizations are also embracing a holistic, enterprise view that treats data as a core enterprise asset. Finally, many organizations are retreating from reactive data management in favor of a managed and ultimately more proactive and predictive approach to managing information.

Data Sampling

Data sampling is a statistical analysis technique used to select, manipulate and analyze a representative subset of data points in order to identify patterns and trends in the larger data set being examined. Sampling allows data scientist, predictive modelers and other data analysts to work with a small, manageable amount of data in order to build and run analytical models more quickly, while still producing accurate findings. Sampling can be particularly useful with data sets that are too large to efficiently analyze in full -- for example, in big data analytics applications. An important consideration, though, is the size of the required data sample. In some cases, a very small sample can tell all of the most important information about a data set. In others, using a larger sample can increase the likelihood of accurately representing the data as a whole, even though the increased size of the sample may impede ease of manipulation and interpretation. Either way, samples are best drawn from data sets that are as large and close to complete as possible.

There are many different methods for drawing samples from data, and the ideal one depends on the data set and situation. Sampling can be based on probability, an approach that uses random numbers that correspond to points in the data set. This approach ensures that there is no correlation between points that are chosen for the sample. Further variations in probability sampling include simple, stratified and systematic random sampling and multi-stage cluster sampling. Once generated, a sample can be used for predictive analytics. For example, a retail business might use data sampling to uncover patterns about customer behavior and predictive modeling to create more effective sales strategies.

We look at a critical first step in ensuring that predictive analytics models yield valid outcomes. The proper preparation of data -- a process made more complex with the proliferation of big data and the expansion of predictive analytics applications to leverage it -- requires new tools and techniques.

Data Classification and Storage

The process of data classifications helps Business understand what data is actually available, it's location in the enterprise, how the data is being accessed, and how it must be protected to meet legal and regularity requirements. But other benefits are often overlooked. Data classification offers a cost savings by allowing less important data to be migrated from expensive primary storage, such as Fibre Channel drives, to less expensive secondary storage systems.

Our data classification process allows companies to organize their information in a way that corresponds to business requirements.

Data classification can help companies identify data that would benefit most from indexing, then move that data to the location that provides the best storage performance for indexing tasks. Backups can also benefit from data classification. Traditional backup processes will typically save everything to disk or tape, but companies rarely need every fragment of data to run their business. Data classification can help a business identify the mission-critical information needed for continuous operation, then focus backups on that essential data. This allows for faster backups and restores, reducing recovery time objective (RTO).

"The answer is not based on size, but based on the value of the information," says Michael Peterson, program director of the Storage Networking Industry Association's Data Management Forum. He points out that a small medical shop with an Insurance Portability and Accountability Act requirement can sustain just as much risk as a larger organization with more files or larger data volumes. The real issue is whether the risk or needs can justify the work involved.

TekBusiness helps the organization to classify the data by understanding what data is present, where it's located, how it's being used and how it's growing, etc. This can have a significant impact on storage planning in any organization. But larger and deeper initiatives are often best justified with significant data volumes.

From Sample Subsets to Full Relevance

The true value of big data lies not just in having it, but in harvesting it for fast, fact based decisions that lead to real business value. Financial institutions were essentially taking bundles of thousands of loans and looking at them as one. We now have the computing power to assess the probability of risk at the individual level. Every sector can benefit from this type of analysis.

"Big data provides gigantic statistical samples, which enhance analytic tool results," wrote Philip Russom, Director of Data Management Research for TDWI in the fourth quarter 2011 TDWI Best Practices Report, Big Data Analytics. "The general rule is that the larger the data sample, the more accurate are the statistics and other products of the analysis.

TekBusiness offers advice and examples of what makes a strong data integration strategy.

However, organizations have been limited to using subsets of their data, or they were constrained to simplistic analysis because the sheer volume of data overwhelmed their IT platforms. What's good? Is it to collect and store terabytes of data if you can't analyze it in full context, or if you have to wait hours or days to get results to urgent questions? On the other hand, not all business questions are better served by bigger data. Now, you have choices to suit both scenarios:

Incorporate massive data volumes in analysis
If the business question is one that will get better answers by analyzing all the data, go for it. The game-changing technologies that extract real business value from big data – all of it – are here today.
One approach is to apply high-performance analytics to analyze massive amounts of data using technologies such as grid computing, in-database processing and in-memory analytics. TekBusiness has adopted the concept of an analytical data warehouse that surfaces for analysis only the relevant data from the enterprise data warehouse, for simpler and faster processing.

Determine upfront which data is relevant
The traditional modus operandi has been to store everything; only when you query it do you discover what is relevant. TekBusiness provides the ability to apply analytics on the front end to determine data relevance based on enterprise context. This analysis can be used to determine which data should be included in analytical processes and which can be placed in low-cost storage for later availability if needed.

Information Management for Big Data

Many organizations already struggle to manage their existing data. Big data will only add complexity to the issue. What data should be stored, and how long should we keep it? What data should be included in analytical processing, and how do we properly prepare it for analysis? What is the proper mix of traditional and emerging technologies?
Big data will also intensify the need for data quality and governance, for embedding analytics into operational systems, and for issues of security, privacy and regulatory compliance. Everything that was problematic before will just grow larger.

TekBusiness provides the management and governance capabilities that enable organizations to effectively manage the entire life cycle of big data analytics, from data to decision.

TekBusiness provides a variety of these solutions, including data governance, metadata management, analytical model management, run-time management and deployment management.

TekBusiness Information Management technology and implementation services enable organizations to fully exploit and govern their information assets to achieve competitive differentiation and sustained business success.

Data Management and Analytics

TekBusiness follows the below 4 key principals in managing data.
1. **What's your goal:** We will focus on the business goal and carefully take every step to achieve your business goal. TekBusiness is unique for incorporating high-performance analytics and analytical intelligence into the data management process for highly efficient modeling and faster results.
2. **Data Management:** Choose the data carefully based on the above Data classification analysis and filter the data for storage and analytics
3. **Analytics Management:** Our High-performance analytics also makes it possible to analyze the above classified data to get precise answers for hard-to-solve problems and uncover new growth opportunities and manage unknown risks.
4. **Effective Decision Management:** Our Effective Decision Management capabilities to easily embed information and analytical results directly into business processes while managing the necessary business rules, workflow and event logic.

High-performance, scalable solutions slash the time and effort required to filter, aggregate and structure big data. By combining data integration, data quality and master data management in a unified development and delivery environment, organizations can maximize each stage of the data management process.

Stream it, score it, and store it

TekBusiness is unique for incorporating high-performance analytics and analytical intelligence into the data management process for highly efficient modeling and faster results.
For instance, you can analyze all the information within an organization – such as email, product catalogs, wiki articles and blogs – extract important concepts from that information, and look at the links among them to identify and assign weights to millions of terms and concepts. This organizational context is then used to assess data as it streams into the organization, churns out of internal systems, or sits in offline data stores. This up-front analysis identifies the relevant data that should be pushed to the enterprise data warehouse or to high-performance analytics.

To ensure that you have the right combination of high-performance technologies to meet the demands of your business, we offer several processing options. These options enable you to make the best use of your IT

resources while achieving performance gains you never would have thought possible.

Because it's not just the fact that you have big data, it's what you can do with the data to improve decision making that will result in organizational gains. TekBusiness can cut through the complexities of big data and identify the most valuable insights so decision makers can solve complex problems faster than ever before.

High-performance analytics from TekBusiness is optimized to address new business requirements and overcome technical constraints. In addition, TekBusiness is leading the way in empowering organizations to transform their structured and unstructured data assets into business value using multiple deployment options.

Conclusion

"One-third of organizations (34 percent) do big data analytics today, although it's new," wrote Russom of TDWI. "In other words, they practice some form of advanced analytics, and they apply it to big data. This is a respectable presence for big data analytics, given the newness of the combination of advanced analytics and big data." Given that more than one-third of organizations in Russom's research reported having already broken the 10-terabyte barrier, big data analytics will see more widespread adoption. Organizations that succeed with big data analytics will be those that understand the possibilities, see through the vendor hype and choose the right deployment model.

Big Data and Big Data Analytics

Big data is not just for large organizations. If we define big data as the data volume, variety, velocity, Veracity, and complexity that exceed an organization's ability to manage and analyze it in a timely fashion, then there are candidates in any industry. It doesn't matter if the breaking point is reached at hundreds of gigabytes or tens or hundreds of terabytes. The principles that apply to big data and big data analytics are similar and can help the smaller organization extract more value from its data assets and IT resources.

Closing Thought

Big data is not about the technologies to store massive amounts of data. It is about creating a flexible infrastructure with high-performance computing, high-performance analytics and governance – in a deployment model that makes sense for the organization.

TekBusiness provides comprehensive big data analytics capabilities, from robust information management support (data, analytics and decision management) to high-performance analytics infrastructure support, big data visualization and exploration capabilities, solutions that integrate structured and unstructured data, and prepackaged business solutions.

TekBusiness can come to your premise and discuss your business goals and demo the key principals followed by TekBusiness which can help assess, provide guidance and deliver solutions that support the best approach for any organization. Organizations can choose the approach that meets their needs today and scales for the future.