Introduction

This newsletter discusses the challenges faced by organizations attempting to deploy a Big Data strategy. We explain how the concept of the Logical Data Warehouse (LDW) can avoid the pitfalls associated with traditional data management techniques. We discuss how Big Data solutions can be delivered in a configurable and agile manner.

The immense reservoir of information across the Web and within corporate data stores can remain untapped because of the size of the data, its complexity and disparate nature, the short ‘relevance’ period of the data and the varied processing and connectivity requirements. How can Big Data be organized, analyzed and controlled so that the information it contains becomes explicit, available and actionable?

The traditional analytical approach to Information Management (IM), where physical repositories are used to consolidate and integrate the data for a preconceived purpose (in an industry model), relies on:

- The capabilities of the underlying data base and ETL technology,
- The strength of the solution design and, to a large extent,
- An understanding of the business questions that need to be resolved.

The problems faced by organizations building a Big Data analytics solution in the old way are that:

- The data is hard to prepare for analysis within the data base (due to unrealized rules and patterns),
- The large amounts of data needed in many different internal and external stores mean data base technology for consolidated analysis becomes cost prohibitive,
- The speed at which the analysis is required for use is usually real-time (or near real-time), making a data-mart or EDW-centric approach fail,
- The cost is at least three times greater than that of a LDW solution,
- The go-to-market speed of any new approach is very slow, reducing first-mover benefits, competitiveness and success rates in terms of customer acquisition and churn-reduction, and
- A single-view Enterprise Data Warehouse (EDW) or integrated data mart infrastructure, involving an unstructured integrated data set(s) and using a technology like HADOOP, cannot meet the demands of the Big Data analytics use case deploying a physical integration of the data.
In our experience, customers are moving towards mixed IM architectures where the data is brought together through metadata, data virtualization (processing data ‘in place’), distributed processing, where large data engines work on specific data preparation tasks, and sophisticated optimization and management capabilities.

This approach, described as a Logical Data Warehouse (LDW: a single view of the data, without moving it), needs a specific software solution for managing and deploying its development. ThoughtWeb’s Enterprise Analytics Studio™ (EAS) is a software solution for the central management, design and build of an enterprise LDW.

A truly enterprise solution, the EAS provides an organization-specific approach to Master Data Management, SLA management, an ability to handle taxonomy to ontology resolution, governance process management and data quality tools. The EAS ensures all aspects of the LDW are understood and managed.

This newsletter shows that the key to leveraging the power of Big Data is the Enterprise Analytics Studio™.

The ThoughtWeb EAS is a virtual development environment for building Big Data applications and the LDW. It constructs organization-specific virtual data marts, provides internal data transformation (ELT, not ETL), massively leverages structured, unstructured and tacit data in combination, enables automated organizational collaboration, and determines the relevance of every concept to the strategic goals of the organization, automatically assessing priority and risk.

The ThoughtWeb Enterprise Analytics Studio™ is the new world of Big Data analytics. I encourage you to explore it.

Vinay Samuel
President and Chief Executive Officer
ThoughtWeb Inc.
The hard disk invented by IBM in 1956 could store 5 megabytes of data. Today, a USB flash drive can store up to 256 gigabytes of data, while the IT infrastructure of a major corporation such as Walmart houses 2.5 petabytes of data. This represents growth in data storage capacity by a factor of more than five hundred and thirty-six thousand million over the past fifty years.

It is not a case of ‘too much information’ but of ‘too much data’, which is a more severe problem. Information can be used for decision-making, but data must be organized to become information: in its raw state it is useless.

For example, a powerful search engine can easily find as many as ten thousand documents in response to a particular query, although only five of those documents contain the gems of information you need. Similarly, every large organization faces the challenge of enabling its people to collaborate and ‘join the dots’ across the many workplaces and jurisdictional silos.

The only effective solution to these problems is intelligent collaboration, meaning the implementation of software that has the ability to discover automatically:

- Who needs to know what;
- Who and what can contribute to this knowledge; and
- To facilitate connection between the right people and the right information in real time.

Software can do this if it has access to the tacit knowledge that drives our need for information and if it can make sense of data in context. It is a lack of contextual intelligence that prevents organizations getting value from the data and from collaborating effectively.

The human brain has an amazing ability to build up layers of information, enabling us ‘intuitively’ to know when something is more important than something else. Despite this, we often become disconnected from the bigger picture and our decision-making becomes sub-optimized.

ThoughtWeb’s software is configured to build up as many layers of thoughts and ideas as are needed to connect the tiniest, most localized activity with the highest-level purpose drivers for example, the enterprise vision. ThoughtWeb’s software absorbs as nodes in this network all kinds of tacit knowledge, including goals, objectives, risks and opportunities, applicable to every unit or team or person in the organization and connects automatically all kinds of data, structured and unstructured, to appropriate entities or nodes in this same network.

The challenge is to unlock the resources of Big Data, to capture real-time data-feeds from both internal systems and external sources, including news feeds and social media, to allow people access to relevant knowledge and immediately useful guidance.

Big Data is a major contemporary example of a problem that is also an opportunity. Purposefully assembled and organized, Big Data is a source of enormous organizational power.

Source: ThoughtWeb
Practices around Data Warehousing (DW) are evolving rapidly, due partly to the rise of Big Data, coupled with the long delivery lead-times and costs of traditional DW, and the increasing push for self-service Business Intelligence (BI) and analytics. Organization-specific Knowledge Domains (KDs) within the ThoughtWeb Enterprise Analytics Studio™ (EAS) provide a cost-effective, highly adaptable implementation approach to Logical Data Warehousing (LDW), which is one of the keys to unlock the potential of Big Data.

CIOs and architects are seeing the need to integrate an enormous variety of data types when planning information solutions for the future. They are focusing more than ever on delivering comprehensive solutions that retain enterprise knowledge, improve business efficiency, save people's time and drive business value.

These factors are driving the rise of Logical Data Warehousing (LDW) as a significant evolution in DW practices. We are seeing an evolutionary approach in which a significant number of leading architects are adopting a combination of data federation, multi-staged data caching and repositories.

The ThoughtWeb Enterprise Analytics Studio™ (EAS) provides an agile development environment in which the power of the LDW is realized through rapid configuration of a distributed network of Knowledge Domains (KDs). These KDs capture tacit knowledge and business processes in context, using purpose drivers to prioritize personal awareness and decisions. They provide intelligent personalized alerts and navigation, integrating knowledge management across the enterprise.

The KD is where Big Data applications and data management processes are designed and implemented. For every business subject, KDs are deployed, federated and integrated to construct a single, purpose-driven view of your business. In most cases, the physical data remains in place at its source.

Source: ThoughtWeb
Understanding the Logical Data Warehouse: The Emerging Practice

The logical data warehouse is a clear demarcation between centralized repository approaches and managed data services for analytics. Gartner presents the major components of the LDW.

Key Findings

- The top two issues reported during 2011 and 2012 by Gartner’s traditional data warehouse clients are the long time to delivery to add new information types and even new structured data, and the increasing performance management issues.
- A small percentage of practitioners have begun to pursue a “mixed” architecture by using virtualization, distributed processes and repositories together.
- Semantic tools, data virtualization layers and some database management systems are capable of initiating externally managed processes. They are demonstrating a consistent ability to offer mixed solutions for deploying new data – often using mixed architectures.

Recommendations

- Evaluate your existing data warehouse (DW) database platform to determine if new technology or software is to support data virtualization and distributed process management. Do not forget to include your current vendor’s road map. Many DW vendors have already anticipated at least portions of the logical data warehouse (LDW).
- Determine the extent to which you are capturing metadata for auditing performance, data quality metrics, and completeness of the datasets involved in your warehouses and sourcing strategies.
- Build one instance of a data virtualization end-user case with a strong preference for including a “big data” class of data. Build one distributed process operation, this can be single server size. Evaluate approaches for managing the conversion of one information management approach to another, such as converting distributed process output into an extract, transform and load job or converting it to a virtualized view.

Analysis

Complaints about the slow evolution of DWs continue to grow among end-user organizations. At the same time, big data is putting pressure on the warehouse to apply traditional principles of analytics on a grand scale as well as forcing the inclusion of less traditional information assets. The past 25 years of data warehousing practices have taught us that reliance upon faster hardware and network capability creates only temporary relief from the pressing demands to analyze more data for more users. Experienced DW practitioners utilize technology advances and innovative information management practices in a balanced approach to ensure DW longevity – as opposed to focusing on insular advances or practices as a “replacement.”

Traditional applications, including business intelligence (BI), follow the premise that the responsibility for service-level guarantees and information governance is shared by the application and the repository; DWs are effectively a consolidation semantic (see “Information Management in the 21st Century Is About All Kinds of Semantics”) in which multiple information governance models and information models from source systems, other warehouses were deployed on new information models and rationalized all sources to that model. However, the rationalization effort was focused on creating a repository for analytics and the chosen BI application frequently influenced the data model. In effect, a new silo of information was created by the same application and repository agreement regarding service level and governance. The LDW (see Note 1) breaks down this false dependency and supports a much wider range of semantic access to enterprise information.

Business Drivers

1. Big data (see “Big Data’ Is Only the Beginning of Extreme Information Management” and “The Importance of ‘Big Data: A Definition”). Organizations are actively seeking business benefits, from adding new information asset types and/or very large datasets enhancing situational awareness and discovery, improved decision-making and process automation. New asset types are not always best served by deploying a repository-only solution. Existing hardware and software platforms may be far more costly than is desirable for infrequently used data that is rarely, if ever, updated. Under these circumstances the business driver would be that the cost of using today’s technology for the new opportunities that big data offers is prohibitive.

2. End users are embracing self-service BI and analytics. The capability to manipulate existing data quickly is exposing gaps in the information in the traditional DW. It is already difficult to get agreement on the meaning of enterprise information. Self-service flexibility increases the demand for more data and this is accelerating cycle to delivery time for adding new data to the warehouse.

3. The cost-benefit ratio for traditional DWs is starting to reach the point of inversion – with diminishing returns on the funding/budgets in place. Many aspects of traditional DW delivery utilize the same solution for every problem. At the same time, budgeting allocations and cost projections based on existing technology are mistakenly being interpreted as instructions from finance to continue with the current technology options instead of exploring new choices.

Defining the Logical Data Warehouse

The LDW consists of seven major components. These components are graphically depicted in Figure 1 and can be summarized as follows:
• Significantly enhanced management and utilization of information management metadata.

• Three information management approaches – data virtualization and distributed processing (usually on a cluster) join the repository as equal partners.

• Three primary classes of services – audit, SLA management, and taxonomy to ontology resolution.

• Additionally, the LDW provides an information management platform to support data quality, master data management (MDM) and data governance, and receives auditing and statistics from those practices.

• All of this is accessed by use cases such as analytics, BI, corporate performance management and others.

1 **Repository management.** The LDW will continue to utilize enterprise data warehouse (EDW) data marts and operational data stores. However, the repository will only serve specific classes of service-level expectations, which we describe below. There are many types of repositories that will be accessed by the LDW, but here the approach refers specifically to building a physically consolidated solution.

2 **Data virtualization.** The LDW must support the concept of rendering memory or cursor-only types of data resources, which directly read source systems. Data virtualization reads the data in place. The virtualization engine must keep audit statistics which report not only the virtualized processing performance, but the level of disruption or interruption occurring against the sources.

3 **Distributed processes.** A distributed process is a managed service call to an external processing cluster or engine, that obtains a result and reports that result to an interface (as data virtualization does) or submits it for data integration tasks that may or may not write the result to a repository in the LDW. Current examples of this approach include Hadoop MapReduce processing, content management and analysis, or even object analysis with image or video.

4 **Auditing statistics and performance-evaluation services.** The LDW keeps statistics regarding the performance of any of the information management approaches (repository, virtualization, distributed processes). It also keeps statistics on end-user access preferences, connected application access or named connections.

5 **SLA management.** The LDW has a quantifiable and qualitative metadataset which describes, for example, the expectations of all connected applications, end uses and named connections. The LDW constantly monitors the SLA performance relative to the audit statistics it keeps, and makes recommendations or dynamically switches between different information management architectures based on the circumstances of current operations and how they compare to the service-level expectation.

6 **Taxonomy/ontology resolution.** The LDW has a layer of metadata which describes how each information asset taxonomy relates to each use-case ontology in place. This permits the location of data assets across all of the available information management solutions that are registered with the LDW. Additionally, the combination of this metadata with the metadata from the audit and SLA services permits the identification of commonly used information assets and enables

---

**FIGURE 1**
Logical Data Warehouse Reference Framework

DQ = data quality; Gov. = governance; MDM = master data management
Source: Gartner (June 2012)
7 Metadata management. The LDW consists primarily of services and metadata. The metadata must be reusable across all classes of services operating. For example, if a distributed process emerges as commonly used over time, the same metadata should be usable to convert the process into a data integration job and move the results to tables or other file support in a repository to reduce the processing demand on distributed clusters and make those resources available for new requirements. Or, as data virtualization jobs begin to specify recurring relationships in data, moving the virtual data toward a high-performance repository rendering. Similarly, data governance, MDM and data quality rules that emerge from the services in the LDW should be easily shared with data quality, MDM and information governance applications and processes.

Repository or No Repository, That is the Question

A traditional DW is a physical implementation of a data consolidation process – it is a repository. Less traditional approaches have existed in the past, such as the all-federation-driven virtual data warehouse and the distributed warehouse (see “Emerging Trends: Introducing the Distributed Data Warehouse”). But the predominant implementation style, including more than 90% of all DWs deployed in the world, is to build some form of a repository. Despite this fact, a DW has always had the primary goal of being a data consolidation engine or service – not specifically a consolidated repository. For many years, however, engineering that was available in terms of storage access rates, processing capacity and even the software used to deploy the warehouse, relied upon a physical deployment to meet many of the performance and consistency requirements in the warehouse – the industry built repositories.

With the availability of both new hardware and new processing solutions, DW architects and designers must ask the question, “When is a repository still appropriate and when is it not?” The information provider SLA is derived from end-user requirements, and guidelines have emerged.

When to Build a Repository

- Pervasive use of any representations of consolidated information is expected above, and over, any other consideration. This assumes consolidated information is commonly understood and agreed as being primarily consistent among most end-user requirements to access the data. Similar to the 80/20 rule, if 80% of users expect to see the same data appear together in the same relationship in multiple use cases.
- Persistence of a representation is required beyond first access. This addresses compliance and reporting use cases, but also consistency over an established period of time.
- Latency tolerance allows for data to experience any delay in arrival, and with the proper management of that latency interval the information still meets the use-case requirement.
- Information quality standards are agreed and the rationalization of the data requires data quality and enrichment to be infused into any latent, persisted and pervasive data use case.

When to Consider Data Virtualization

- Representations of consolidated information are not agreed and ongoing experimentation or discovery of how information assets relate to each other is under way.
- Transient consolidations that experience only a temporary end-use-need are required.
- Native tools which authored/created a given information asset are capable of executing all other forms of processing the information (such as storage management, read/access and analysis) without a significant variance from the write speeds. When the native application is processing all other information utilization at rates 10 times slower than accumulating the data, consider a repository.
- Information quality and determination of anomalies can be delegated to the sources or can simply be resolved at runtime.

When to Consider Distributed Processing

- Use case is compatible with batch processing, that is, the data is cold but pattern analysis over it is valuable – the data is not archived.
- Data collected in the distributed file system does not follow predefined relationships, but rather requires the programmers knowledge to infer these relationships and deduce the patterns supporting the use cases.
- Data collection does not need to follow strict governance or data quality expectations, but rather requires adapting the governance principles to the use case. For example, clickstream analysis used to discover recommendation patterns only requires that the click data collected is the result of real user interaction rather than of Web robots.

Does the Traditional EDW Go Away?

In a word, no. The traditional EDW is an enterprise repository implementation. However the repository implementation by itself will not meet the potential range of SLAs. The role of the traditional EDW will be determined by the architectural preferences of each organization and should include evaluation of the use cases.

- The EDW can become the repository solution for certain use cases, and can be accessed beneath a semantic manager that directs processes to the repository when it best serves the expected service levels of a given use case.
- The EDW can serve as the semantic manager and call out to distributed processes or even run the data virtualization tasks required. Most relational databases can perform both of these tasks.
• The EDW can jointly control the semantic management with a semantic layer, sharing auditing metrics and coordinating which tasks will run in each portion of the LDW.

• The EDW can become a processing platform to assist in tasks to support distributed processes and data virtualization tasks—especially in the case of scoring data records for use in analytics. This use case is particularly true of appliances.

• The EDW can become the place where the patterns discovered using distributed data processing are stored and used, along with other traditionally-collected data elements.

Source: Gartner Research, G00234996, M. Beyer, R. Edjlali, 21 June 2012

Note 1. Origin of the Logical Data Warehouse

The LDW is not a disruption, but a significant evolution in data warehousing practices. It is, however, a response to disruptive forces in the data and information management markets such as big data. In mid-2008, Gartner clients began (in small numbers) reporting a combination of various data management and integration technologies that was achieving success outside of traditional repository approaches. By April of 2009, a significant number of leading edge architects in end-user organizations had started to follow a pattern of combining data federation, multistaged data caching and repositories. The term “logical data warehouse” was first proposed in May 2009, and by October 2009 the “LDW” terminology was being used internally and intermittently at Gartner events. The term was published in official Gartner research in August 2011. As of October 2011, fewer than 3% of DW practitioners were addressing more than 50% of the overall concept. The interest of end-user organizations in finding the next evolution of analytics data management was approaching approximately 20% of DW architecture inquiries. Currently, Gartner inquiries indicate that slightly more than 5% of DW practitioners are now pursuing reference architectures and designs, and in any given month as many as 40% of DW inquiries include aspects of the LDW.
Telcos and Big Data

Understanding your customers and the level of real satisfaction they have with you is sometimes not evident until they tell you they are leaving. At this point it is normally too late for you to do anything about it.

Matt Carroll, CEO of Riteway Solutions Group, says: “We use ThoughtWeb to combine all aspects of our relationship with our customer into a single scorecard and use this scorecard to measure our Customer Satisfaction levels and give our teams specific and targeted messages to deliver.

A better understanding of my customers and how happy they are with me gives me the ammunition I need to serve them better. However, I need the information now. Not in two weeks’ time, when the data warehouse can provide it. ThoughtWeb brings me the data I need, including the sentiment of my customers, in real time, so I can address, improve and measure the service I am offering when it matters most.

“ThoughtWeb brings me the data I need, including the sentiment of my customers, in real time, so I can address, improve and measure the service I am offering when it matters most.”

Matt Carroll, CEO, Riteway Solutions Group

We take data from a range of systems and combine them into meaningful, actionable insight. For example, you can combine outage data from a single network cell, with details about the customers getting coverage from that cell, add information about those particular customers’ products and spending habits, match it with notes from recent contacts they have made with you, and determine whether the impact of the outage is going to be negative or neutral on their overall satisfaction level. Based on this assessment, you can trigger a call-out, SMS, cross-sell, or any other retention action you may have available to each specific customer.

You can do all this in near real time – without having to wait days for your data warehouse to give you a general feeling about the impact of the outage.”

Carroll is also impressed with the ability of ThoughtWeb to provide greater accuracy in the measurement of product performance.

“The telecommunications landscape is changing at an unprecedented pace,” he says, “and the ability of a telco operator to stay in front of this change and to drive revenue from a customer base that can shift loyalties to your competitors easily is getting harder each day.

One of the biggest challenges a Director of Marketing or Product Manager faces is to demonstrate the successful performance of your products and to show a direct contribution to the company’s top-line and bottom-line revenues. Understanding product performance goes beyond simply looking at sales trends, product revenue, product margin and profitability – these are traditional lagging indicators that often do too little too late to drive the performance of products, processes or people. In fact, teams that focus solely on these measurements usually struggle to establish clear value and contribution to the company’s objectives.

“Telcos use ThoughtWeb to combine analysis of the lagging indicators with current and predictive data to give a more pertinent view of how your product is performing and the contribution it is making to revenues.”

Telcos use ThoughtWeb to combine analysis of the lagging indicators with current and predictive data to give a more pertinent view of how your product is performing and the contribution it is making to revenues. With the ability to ingest data from both structured and unstructured sources, ThoughtWeb combines standard performance indicators with activity and operational data such as customer satisfaction measures, real-time sales data, competitor advertising, blogs, website comments, campaign management reports, customer surveys, and a range of other product and brand data to deliver a clear message on the performance of each product in each market over time.

With more detailed and comprehensive performance data available, companies can identify and focus on the high return improvement opportunities in product development to improve performance in revenue.”

Source: ThoughtWeb
Banking and Big Data

The financial crisis has demonstrated the need for banks to understand their business models together with the associated risks and to have confidence that performance indicators and executive incentives reinforce desired behaviours.¹

To improve marketing effectiveness, best-in-class banks are crafting ‘right here, right now’ banking sales and service models to enhance the customer experience and drive revenue. These banks are seeking to understand customer perceptions of the brand, detect significant events, and link these to customer lifecycle profiles so they can drive multi-channel, multi-step campaigns; i.e., which customers should we communicate with, through which communication channels, with what offers?

ThoughtWeb has an adaptable solution integrating structured customer data with unstructured reports and social media, able to deliver a real time, adaptable campaign management solution.

While it is true that many banks have made significant changes to their organizations and the way they run their businesses since the last crisis, the fragile confidence in the sector is further underlined by the presence of credit risk, liquidity and capital availability among the top four Banana Skins identified by this survey. It is also clear that much work has still to be done: banks remain both unpopular and under the spotlight.²

In dealing with Basel III compliance banks are focusing on capital requirements (higher minimums and conservation buffers), leverage ratios (tracking underlying components), and liquidity requirements (deriving LCR and NSFR). At a broader level banks must deal with a plethora of risk categories, including counter-party, trading, credit and operational risks. It is common for banks have more 20 risk management committees aggregating and analyzing data from many sources. How can every executive be confident that s/he is aware of the combined enterprise risk across all these sources of risk?

‘ThoughtWeb delivers exceptionally high value in this area,’ says Chris Murray, ThoughtWeb’s Founder. ‘A federated network of adaptable Knowledge Domains sitting across all the data sources supporting collaborative decision-making across all parts of the organization in real time is a practical and cost effective solution.’

In developing strategies to maximize share price, banks are dealing with the identification and prioritisation of global opportunities and threats, diverse stakeholders and counter-balancing goals. There is a need for predictive analytics and collaborative decision-making around opportunities (e.g., countries, industries, major clients), emerging competitors, and shifting macro-economic conditions.

ThoughtWeb contextual reasoning capabilities help facilitate the development of strategies supporting goals and perspectives. Projects are prioritized automatically based on contributions to strategies, goals and outcomes, and management at all levels have total visibility of relevant projects in all dimensions, e.g., by divisions, countries, programs, strategies, stakeholders. The adaptable solution provides optimized budget and resource allocation and performance tracking in support of collaborative decision-making.

When you consider the number of financial analysts operating in silos, and the amount of time they spend manually researching documents and conducting analytics before they can write their reports, you realize the enormous potential for banks to do more with less.

ThoughtWeb Insight automates the ingesting and reading of documents, machine-learning, and visualization, and the prioritization of relevant, paraphrased materials for each analyst. ThoughtWeb Enterprise Analytics Studio™ (EAS) integrates analyst comments and assessments, facilitating intelligent information-sharing and collaboration across the organization.

Source: ThoughtWeb

Government and Big Data

Tony Hindmarsh, Acting Head of DMO Reform at Australia’s Defence Materiel Organization, says: ‘We have embarked on a journey of disruptive innovation. We don’t have time to navel-gaze: we have a need to operate differently and we need to do it now. We need to understand each other’s perspective, how our respective plans, performance and risks impact each other. We need an integrated picture of how it all hangs together so that we can improve governance, by making it active and outcomes-focused, and improve functions where it makes most sense.’

The Defence Materiel Organization has found that ThoughtWeb can meet its need for speed and reduced costs.

Petar Bielovich, ThoughtWeb’s Vice President of Professional Services, comments: ‘We are working with a number of government agencies. In these tough budget times, our focus is on helping enable their vision of improving how they work together - at all levels within and across their organizations.

‘Each client has access to all the data they need - the decision makers know it’s there, but the organizations have struggled to pull it together so that they can make sense of it - so that they can avoid strategic surprise.

‘In response to a number of government agencies’ need to improve their ability to prioritize what they do to deliver on their required outputs and mitigate enterprise risks, we have implemented solutions for senior decision-makers and their supporting strategic management community that integrate strategic planning, enterprise risk management, business transformation frameworks, and information. These solutions have enormous business significance, resulting in improved shared understanding, clear accountabilities and improved ability to prioritize where they need to invest to achieve outcomes.’

Bielovich notes that ThoughtWeb’s products and services can be instrumental in developing organizational trust.

‘It’s truly satisfying for a client team to first acknowledge the thinking model and business rules that define their context and decision-making and then, when real data is aggregated and prioritized, how they react,’ he says. ‘They experience the “Ah-ha!” moment, when they say, “This is exactly what we need to do!”, or when one agency says to the other, “What exactly do you mean by that risk, and could you please explain why it’s important to you?”, or when the Department Head says, “We need this transparency because it builds the trust we need.”’

Source: ThoughtWeb

“\nIn response to a number of government agencies’ need to improve their ability to prioritize what they do to deliver on their required outputs and mitigate enterprise risks, we have implemented solutions for senior decision-makers and their supporting strategic management community. “

Petar Bielovich, Vice President Professional Services, ThoughtWeb
Effective Master Data Management (MDM) is obviously crucial to the success of a Logical Data Warehouse. However, as Wolter and Haselden showed in their seminal white paper, the apparent simplicity of master-data items can mask the complexity and organization-specific nature of a successful Master Data Management strategy.

Organizations can create competitive advantage by leveraging sources of Big Data, including social networks. There is an increasing need to extract knowledge buried in content and include it in business processes and analytics. Specifically, organizations need to implement Master Data Management (MDM) in the form of a MDM Hub and a distributed network of MDM Master Data.

ThoughtWeb’s Knowledge Domains (KDs) facilitate the dynamic discovery of Master Data-related content in the unstructured data and provide a framework for the integration of structured data and related content in context, which is a fundamental requirement for the implementation of MDM Master Data.

The key factor to bear in mind is that Master Data Management, Social CRM and the development of Knowledge Domains are not separate activities. Your organization will benefit from a unified strategy that addresses all of these key areas of opportunity; they are simply different paths up the same mountain!

Source: ThoughtWeb

---

To create competitive advantage, organizations will need to leverage sources of big data, especially social networks, in combination with their MDM strategy. But not all big data is the same and organizations need to understand the complexities and plan accordingly.

**Impacts**

- Enterprises that generate structured “big data” will use MDM processes to link or apply appropriate governance to it, leading to closer collaboration between the MDM team and teams working on big data initiatives.
- The desire to leverage social data generated outside the organization will lead organizations to link together MDM and social CRM systems by resolving identifiers and extending the customer profile and product record with social data.
- New parsing and analysis capabilities that form part of the big data advances will potentially allow organizations to link to or govern unstructured big data as part of their MDM capability.

**Recommendations**

- Get up to speed on big data concepts and technologies and start thinking about how your organization’s MDM capability may need to evolve to cope with the demands of big data, in the context of how it is likely to impact your organization and what business value could be derived from it.
- Work with the group leading your organization’s social CRM strategy and work out how you can link together the social CRM initiative and MDM initiative. Also, determine what private social networks your organization has or is planning and work out if there is value in linking the MDM initiative to them.
- Work with the data warehousing team to see if it is formulating a vision, which includes big data, around the concept of the logical data warehouse. Ensure that it involves the MDM team in the planning process, since master data will sit at the heart of the logical data warehouse.
- Investigate the new big data era parsing and analysis tools that focus on unstructured data to determine how they will enable increased leverage of unstructured data within the MDM program.

**Analysis**

Big data is a term used to describe high-volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Too often, discussion of big data challenges revolves around the volume dimension (and the very name suggests this), but it is often a combination of increased volume, velocity and variety that is creating challenges for enterprise architects and information management professionals, and opportunities for analytical specialists and data scientists. Gartner has written extensively about big data (and is continuing to write about it), outlining the business opportunity (see “CEO Advisory: ‘Big Data’ Equals Big Opportunity”), defining what big data is (see “The Importance of ‘Big Data’: A Definition”) and discussing the 12 dimensions of big data from an information management viewpoint (see “Big Data is Only the Beginning of Extreme Information Management”). In addition, discussion of big data technology tends to focus on Hadoop and MapReduce (see “How to Choose the Right Apache Hadoop Distribution”), but although they and accompanying technologies are rightly seen as breakthrough technologies allowing the storage and analysis of large volumes of data at new price points, they are not the only technologies that should be considered in combination with big data (see “Cool Vendors in Information Infrastructure and Big Data, 2012”).

Master data management (MDM) program managers need to come up to speed on big data concepts and, in particular, early adopter organizations should explore the potential business advantage that can be achieved by linking the MDM initiative with the social CRM initiative. In this research we explore the relevance of big data trends and concepts to MDM, discuss how a particular form of big data, social data, can be leveraged by and through an organization’s MDM initiative, and show how new big-data-related technologies will potentially enable greater inclusion of unstructured content in an organization’s MDM program, and hence its business processes and analytics.

**Enterprises that generate structured big data will use MDM processes to link or apply appropriate governance to it, leading to closer collaboration between the MDM team and teams working on big data initiatives**

Big data takes many forms. It may be structured data, unstructured data or content. It could be generated by IT systems, such as transactional business applications or it could be generated by operational technology (OT) devices such as smart meters, sensors or other parts of the “Internet of things.” Big data could also be generated internally within the organization, in concert with its business partners, or it could be generated externally, such as in social networks. It could often be data that has so far been underutilized. Something that we call “dark data.” In addition, the growth in adoption of mobile technology is leading to massive growth in the generation of location data and is also driving the usage of it in context-aware applications.

If the big data is generated internally, whether by IT or OT systems, then it potentially needs to be governed in some way (although there are likely to be different levels of governance applied), if there is business value in governing it. As with traditional data, big data is likely to contain direct and indirect references to the core business entities of the organization, namely its customers, products, suppliers, employees, assets and locations. There are different scenarios for MDM in the context of big data.
In most situations, the master data itself will not be “big,” rather the transaction data – such as clickstream data, blog files, email messages, telco call detail records (CDRs), Internet protocol data records (IPDRs) or sensor readings – is “big.” In this situation, the MDM team will need to work out a way to integrate the MDM system with the big data sources, via Hadoop adaptors for example, and provide entity resolution capabilities so that the big data transactions can be matched across to existing master data in the MDM system.

Sometimes there could be the need to govern the “party,” “thing” and “place” master data that is referenced in the big data transactions. In some situations this is not “big” and so existing methods could be used, but in a minority of situations the master data itself will be “big.” For example, if a government intelligence or law enforcement agency was attempting to master Internet Protocol addresses, mobile phone numbers, locations or very large populations, then the governance process will need to be fit for purpose. It is unlikely that a people-driven, workflow-based authoring process would be suitable, so we would be looking at a highly automated and scalable process whereby “big” master data is authored, probably in a distributed fashion, and matched, linked or merged in the MDM hub.

One of the impacts of big data will be a retreat from the concept of the physical enterprise data warehouse in a favor of a logical data warehouse (see “Does the 21st-Century “Big Data” Warehouse Mean the End of the Enterprise Data Warehouse?”) that can draw on different data sources that are based on different database technologies, such as relational (both row- and column-based), NoSQL and Hadoop. MDM will be key to linking this complex environment together, and this will require that the MDM team and the data warehousing team work together on the future vision.

It is also worth discussing the concept of a master data trust model (see “The Emergence of Information Stewardship Applications for Master Data”) in the context of big data. The trust model positions different kinds of data, including master data, on a continuum depending on the depth of trust that you seek and observe in the quality of a particular piece of data. At one end of the scale is the data generated by the organization’s core business systems, followed by that generated by more informal internal systems. Then, outside the firewall there is data supplied by trusted third parties, partners and industry groups, until finally you come to crowdsourced and social data. You need to potentially segregate these classes of data based on trust labels, such as assured, affirmed, proven, confirmed, asserted and unknown. And with the burgeoning rise of data and its new complexities, the model will need to be more explicit. Big data potentially falls into several of these segments, but the data relating to social media would need to be labeled as less trusted, unless and until actively assured.
The desire to leverage social data generated outside the organization will lead organizations to link together MDM and social CRM systems by resolving identifiers and extending the customer profile and product record with social data.

Organizations increasingly want to engage with their customers and prospects through public social networks like Twitter, Facebook and LinkedIn, as well as through private social networks, such those based on tools like Jive and salesforce.com’s Chatter. These social networks are seen as additional channels for organizations to market to, sell to and service their customers, but also as opportunities for the organization to understand its customers better, to get closer to them and to provide them with a better customer experience. This is increasingly seen as a key competitive differentiator.

The data generated by individuals in these social networks is one form of big data, but the public and the private social networks have different characteristics and different implications for MDM. Private social networks are maintained by the organization to encourage its employees to collaborate or to engage with its customers. The usual governance rules of MDM can be applied to both of these situations, because the organization is setting the rules and hosting the social network. The situation with external public networks like Twitter, Facebook and LinkedIn is different, because they are beyond the organization’s control and the data is being created external to its systems. The challenge is to connect these two worlds – the internal enterprise business applications and the external public social networks.

Social media monitoring platforms are part of the set of functionality that forms social CRM. These monitoring platforms can analyze sentiment toward companies, brands and products on public social networks, such as Twitter and blogs. They typically provide insight into positive or negative sentiment at an aggregate level, enabling a better understanding of what different customer segments (for example “people in this geographical area” or “18 to 24 year olds”) think and are saying about something. Going a step further, using identity resolution technology (essentially matching algorithms) and accessing the social networks through their APIs, it may be possible to determine the identity of that Twitter, Facebook or other social network user, to understand more about their network, to determine how influential they are and then when and how (via the appropriate trust model) to link that back to the enterprise’s master customer profile and potentially the product record in the MDM system. Social network identifiers and other attributes should now be seen as a “social customer profile” extension to the existing customer profile. This will enable the organization to understand and interact with its customers or prospects via social networks in addition to traditional channels. For example, this capability will help retail organizations understand and influence the customer’s buying journey (with respect to their interest in particular products, services and brands), help them to provide more relevant and timely marketing offers, protect their brand, and provide better customer service.

Public social networks differ in terms of how “public” they are. If someone tweets on Twitter it is usually public to the world and they can expect others to tweet on what they say. However, on Facebook and Google+ you have to be “friend”ed or “added” in order to join a community. There is a level of privacy associated with these networks and unwelcome approaches from big corporations would be seen as invasive or deemed “stalking.” From the organization’s point of view it could be technically possible to resolve the identities of members of these kind of communities, but it would be unwelcome if you approached them. It needs the individual to proactively “like” the company or product on Facebook and then to enter the company’s Facebook application which requests them to assert their identity and give marketing rights and permissions. Once that is achieved, the Facebook ID and profile can form part of the organization’s social customer profile and this channel can be used to develop a relationship.

In addition to the focus on social data relating to customers and prospects, some organizations will also focus on social data and sentiment relating to product and brand.

Recommendations:

- **Get up to speed on big data concepts and technologies** and start thinking about how your organization’s MDM capability may need to evolve to cope with the demands of big data, in the context of how it is likely to impact your organization and what business value could be derived from it.

- **Talk with the data warehousing team to see if it is formulating a vision, which includes big data, around the concept of the logical data warehouse.** Ensure that it involves the MDM team in the planning process.

- **Recommendations:**
  - Work with the group leading your organization’s social CRM strategy and work out how you can link together the social CRM initiative and MDM initiative. Also, determine what private social networks your organization has or is planning and work out if there is value in linking the MDM initiative to them.
  - Leverage identity resolution technologies and techniques in order to identify key customers and prospects that are active in social networks, and extend your MDM initiative to include master data identifiers and attributes that are authored in external social networks, so as to build out the social customer profile.

New parsing and analysis capabilities that form part of the big data advances will potentially allow organizations to link to or govern unstructured big data as part of their MDM capability.

As well as structured data, big data includes unstructured data and content, including clickstream analysis, blogs, application logs, emails, documents, Web pages, social data, audio and video. It has traditionally been difficult to include this content in the MDM initiative, although there are packaged solutions that (a) provide the ability to link to a
metadata dictionary that indexes the content (thus providing structure to the data), allowing a customer service application to not just retrieve the structured part of the customer profile, but also the unstructured part; for example, the photo, scan of driving license, passport of signature, or (b) provide storage and linkages for images of products. Master content management (MCM) includes both of these scenarios, but is still at an emerging stage (see “Hype Cycle for Master Data Management, 2011”).

The big data trend provides new opportunities. The variety and complexity of data formats is increasing, but the competitive need to include knowledge of this content in business processes and analytics is also increasing. To respond to this need, some of the vendors in the big data market, such as IBM, are creating technology that can parse and analyze text and more complex formats to produce an understanding of the content and the derivation of identifiers and attributes that could be linked to by an MDM system. This will potentially allow a more dynamic discovery of master-data-related content in this unstructured data, so that it can be leveraged more widely via the MDM system in business processes and analytics.

Recommendations:
- Investigate the new big data era parsing and analysis tools that focus on unstructured data to determine how they will enable increased leverage of unstructured data within the MDM program.

Evidence
This research is based on analysis of discussions with vendors and end-user organizations.

Source: Gartner Research, G00233199, J. Radcliffe, 23 June 2012
There is a global trend towards the democratization of analytics. The increased demand for Big Data analytics and the increased rate of organizational change coupled with a growing shortage of analysts are necessitating the move towards do-it-yourself analytics. The Enterprise Analytics Studio™ (EAS) of the future, with democratized access for all kinds of users, delivers intelligent collaboration and cost effective solutions extracting maximum value from the data, while simultaneously addressing the challenges of Big Data, Logical Data Warehousing, Master Data Management and organizational agility.

ThoughtWeb’s Enterprise Analytics Studio™ (EAS) is a virtual development environment for building Big Data applications. Solutions are constructed within Knowledge Domains (KDs) that capture tacit knowledge, metadata and business rules by point-and-click configuration. A multi-structured data layer enables Big Data storage. Analytical plug-ins that extract value are pre-packaged. A high-speed SQL interface to Hadoop, a Massively Parallel Processing (MPP) data layer and pre-configured industry applications accelerate solution development.

The main attributes of a ThoughtWeb Enterprise Analytics Studio™ (EAS) are:

- Structured and unstructured data ingestion;
- Structured, unstructured and tacit data storage and connector;
- Idea and role-based collaboration;
- Intelligent push and personalization;
- Object discovery, semantic and sentiment analytics, fuzzy logic and AI for automated reading and understanding data, networks and data frameworks; and
- Facilities for rapidly building Web-based enterprise applications, including collaborative Knowledge Domains and federated networks of Knowledge Domains.
ThoughtWeb Rapid Application Developer (RAD) is used to configure contextual frameworks which put structured, unstructured and tacit knowledge in context, giving meaning to the data and results, which are made available to the right people in near real time for appropriate action. It facilitates Big Data solution deployment within days or weeks.

“ThoughtWeb Rapid Application Developer (RAD) . . . facilitates Big Data solution deployment within days or weeks.”

ThoughtWeb Insight processes a wide variety of data-feeds, making the connections required for improved learning, situational awareness and faster, better decision-making. It is a powerful machine learning system with the ability to ingest, read and make sense of unstructured data. For people who are flooded with data of many types e.g., email, news services, social media, documents, ThoughtWeb Insight saves enormous amounts of time by reading, identifying, summarizing and prioritizing everything that is relevant to you.

ThoughtWeb Insight accumulates learning from everything it reads, building up enterprise knowledge. It uses this knowledge to identify people who have interests in common (even though they may be many degrees of separation apart) and facilitates intelligent collaboration by connecting people who need to collaborate and providing the bridge telling them what interests they have in common.

ThoughtWeb Intelligent BI (IBI) is the integration of ThoughtWeb’s contextual reasoning and tacit knowledge-capture capabilities and includes:

- Interactive dashboards;
- An intuitive Report Writer,
- Slice and dice;
- Drill features connecting multiple reports to form a chain of discovery;
- Proactive exception alerts and notifications;
- Discussion, sharing and collaboration with others;
- Many chart types;
- Formatting control;
- Reporting flexibility;
- Embedded interactive reports or dashboards anywhere in the Web;
- Mobile apps for phone or iPad connected to dashboards;
- View, analyse, share and collaborate in-memory data base;
- Direct connection to analytical databases;
- Drag and drop development (no scripting or coding);
- Open and standards-based architecture; and
- Role-, group- and data-level security.

ThoughtWeb Big Data Analytics (BDA) is a suite of analytical capabilities including statistical analysis and predictive analytics that consumes structured data and applies Massively Parallel Processing (MPP) computing power. It works with very large volumes of data and very wide varieties of data and attributes.

ThoughtWeb Intelligent Content Management (iCMS) is a modular, scalable content management system that operates as part of the EAS. The ThoughtWeb Knowledge Domains (KDs) provide context for the ICM content, with seamless aggregation of content around KD objects.

ThoughtWeb Rapid Data Integrator (RDI) enables organizations to respond rapidly and efficiently to business needs for holistic information from disparate sources of data. It is an approach to data integration based on both ELT (extract, load, transform) and CDC (change data capture).

ThoughtWeb EAS delivers a Logical Data Warehouse capability through a federated network of KDs.

The ThoughtWeb EAS transforms data into contextualized, actionable knowledge before it is presented to the user.

“The ThoughtWeb EAS transforms data into contextualized, actionable knowledge before it is presented to the user.”
The ThoughtWeb EAS provides situation awareness, the right information made meaningful in context. This allows an unmatched level of personalization of information for each individual in the organization, allowing him or her to make better decisions faster, guided by intelligent recommendations and prioritization.

The ThoughtWeb EAS:

- Provides internal data transformation (ELT, not ETL);
- Augments the power of existing data analysis tools;
- Massively leverages structured, unstructured and tacit data in combination;
- Provides unlimited views of the business;
- Enables automated organizational collaboration opportunities;
- Provides unlimited views of the market environment; and
- Automatically assimilates new data-feeds in real time.

The ThoughtWeb Enterprise Analytics Studio™ is the new world of Big Data analytics.

Source: ThoughtWeb
About ThoughtWeb

Visit our website www.thoughtweb.com or email us at info@thoughtweb.com

ThoughtWeb Pty Limited ABN 77 001 399 391

Australia
Sydney:
Australian Technology Park, Locomotive Workshop Suite 9001
2 Locomotive Street
Eveleigh NSW 2015 Australia
T: +61 2 9283 8689    F: +61 2 9209 4323

Melbourne:
Suite 2314, HWT Tower
40 City Road
Southgate, VIC 3006
T: +61 3 9674 7271

Canberra:
T: +61 2 6169 7953

ThoughtWeb Inc.

United States
Postal Address
2691A Hunter Mill Rd #806
Oakton VA 22124 – 1704
T: +1 (703) 657 7632    F: +1 (703) 434 3060