The Analytics Revolution 2011: Optimizing Reporting and Analytics to Make Actionable Intelligence Pervasive

Prepared for IBM by:

David Loshin
Knowledge Integrity, Inc.
March, 2011

Contents

1 Executive Summary

The emergent recognition of the value of analytics clashes with the rampant growth of the volume of both structured and unstructured data. Competitive organizations are evolving by adopting strategies and methods for integrating business intelligence and analysis in a way that supplements the spectrum of decisions that are made on a day-to-day and sometimes even moment-to-moment basis. Individuals overwhelmed with data may succumb to analysis paralysis, but delivering trustworthy actionable intelligence to the right people when they need it short-circuits analysis paralysis and encourages rational and confident decisions.

In this paper, we examine the business value drivers that catalyze the introduction of pervasive analytics, and review some common examples where actionable knowledge can be used to develop competitive advantage. We consider the concept of pervasive analytics, in which operational and transactional data sources are streamed into an enterprise analytical engine, with the results used to enhance and hopefully measurably improve business processes.

The paper then provides an overview of the technical components needed to enable pervasive analytics. A deeper dive provides details about these data management fundamentals:

- Continuous data synchronization ;
- Cohesive information integration;
- A robust architecture for analysis;
- Performance tuning to handle large structured and unstructured data sets;
- Large-Scale computing;
- Comprehensive analytics services; and a
- Framework for reporting, ad hoc queries, dimensional analysis, and visualization and communication of actionable knowledge.

We then look at some of the key challenges to success, and summarize with some recommendations that can support a successful enterprise-wide business intelligence, reporting, and analytics program with a complete suite of tools that support the organization's end-to-end reporting and analytics needs.

2 Growing Data Volumes and Smarter Decisions

For years, technical analysts have been speculating about the monumental growth of data. A 2010 article suggests that data volume will continue to expand at a healthy rate, noting that "the size of the largest data warehouse … *triples approximately every two years.*"[1] The data explosion is globally recognized as a key Information Technology (IT) concern. According to a Gartner study in October of 2010, "47% of the respondents to a survey conducted over the summer ranked data growth in their top three challenges."[2] As an example of rampant data growth, retailer Wal-Mart executes more than 1 million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes. Persistence is not a requirement either – a 2010 report suggests that by 2013, the amount of traffic flowing over the internet annually will reach 667 exabytes.[3]

Structured information in databases is just the tip of the iceberg; some important milestones document the explosive growth of unstructured data as well: by the end of 2009, the amount of digital information was estimated to have grown to almost 800,000 petabytes (or 800,000,000,000 gigabytes), with an expectation that by the end of 2010 that amount would total approximately 1.2 million petabytes! At this rate, the amount of digital data could grow to 35 zettabytes (1 zettabyte=1 trillion gigabytes) by 2020.[4]

Complex business processes are increasingly expected to be executed through a variety of interconnected systems. Integrated sensors and probes not only enable continuous measurement of operational performance, the interconnectedness of many systems allows for rapid communication and persistence of those measures. Every day, it is estimated that 15 petabytes of new information is being generated, 80% of which is unstructured[5].

We can conclude from these few examples of rapid expansion of the amount of digital information that exciting new vistas can be opened up as never before through combined analysis of structured and unstructured data. New and improved means of data analysis allow organizations to identify new business trends, assess the spread of disease, or even combat crime, among many other opportunities explored in this paper. It appears that we are reaching the point where information is becoming the most significant focus of the business, where "statisticians mine the information output of the business

[1]Merv Adrian, "Exploring the Extremes of Database Growth" IBM Data Management, Issue 1 2010

[2] Lucas Mearian, "Data growth remains IT's biggest challenge, Gartner says," Computerworld Nov 2, 2010, (downloaded from http://www.computerworld.com/s/article/9194283/Data_growth_remains_IT_s_biggest_challenge_Gartner_says)

[3] The Economist, "Data, data everywhere," Feb 21, 2010, (downloaded from http://www.economist.com/node/15557443?story_id=15557443)

[4] Gantz, John and Reinsel, David, "The Digital Universe Decade – Are You Ready? The IDC 2010 Digital Universe Study," May 2010 (downloaded from http://idcdocserv.com/925)

[5] Bates, Pat, Biere, Mike, Weideranders, Rex , Meyer, Alan, and Wong, Bill, "New Intelligence for a Smarter Planet," ftp://ftp.software.ibm.com/common/ssi/pm/bk/n/imm14055usen/IMM14055USEN.PDF

for new ideas."[6] And these new ideas are not just lurking in structured databases. Rather the analysis must encompass unstructured data artifacts as well.

But as the volume of data grows, so does the complexity of finding those critical pieces of information necessary to make those business processes run at their optimized level. The issue is no longer the need to capture, store, and manage that data. Rather, the challenge is the need for distilling out and delivering the relevant pieces of knowledge to the right people at the right to time to enhance the millions of opportunities for decision-making that occur on a daily basis.

In essence, this can be summarized as the desire to integrate actionable intelligence in a pervasive manner into both the strategic and the operational processes across all functions and levels of the organization. And whether this means notifying senior management of emerging revenue opportunities, providing real-time insight into corporate performance indicators, or hourly realignment of field repair team schedules to best address customer service outages, the ability to accumulate, transform, and analyze information to provide rapid, trustworthy analyses to the right people at the right time can enhance growth opportunities and competitiveness.

3 Delivering Actionable Intelligence

Within almost every business process, there are situations in which information is accumulated and employed to help individuals make decisions. Some decisions are far-reaching, driving global corporate strategies, while others are operational and narrower in scope, such as selecting the least costly, right-sized boxes to best accommodate a customer's order. And for each business process, there are measures of performance, such as return on investment, time to value, or cost of materials.

In a perfect world, each decision would be the *optimal* one – the decision whose results lead to the best overall performance. However, decision-makers are often not provided with all the information they need to make the optimal decision. Providing more information is not always the answer – when massive amounts of unfiltered data are channeled across the organization, overwhelmed individuals are stunned into "analysis paralysis" – the compulsion to delay decision-making while waiting for just a *little bit more data* that can simplify (and perhaps justify) that impending decision. This paralysis can be alleviated when the information overload is throttled back to filter out the specific information necessary to optimally drive the decision process and allow specific actions to be taken. Delivering trustworthy **actionable intelligence** to the right people when they need it short-circuits analysis paralysis and encourages rational and confident decisions.

At one end of the continuum, overall company performance is reviewed and alternatives are considered to help adjust corporate strategy for long-term value generation. At the other end, operational activities are improved with specific pieces of intelligence that can adjust and optimize activities in real time. In the best scenario, the business process *itself* incorporates the acquisition and presentation of actionable

[6] Op. cit., The Economist

intelligence, streaming the results of analyses directly into the process, tracking performance measures, and indicating when better decisions are being made.

In essence, there is a “closed loop” relationship between operations, analysis of operations, modifications to business processes driven by the analysis, and then changes to the operational business processes and corresponding applications. Actionable intelligence informs both strategic and operational processes, and its pervasive delivery to staff members up and down the organizational chart can facilitate a transition from reacting to what has happened in the past to streamline making the optimal decisions moving forward.

3.1 Strategies that Drive Organizational Optimization

As more strategists within the organization are becoming aware of the capabilities of reporting and analysis, there is a growing desire to employ data analysis to proactively manage and improve business strategies. Observing the environment provides insight into how the organization works, as well as allows the analyst to understand areas of failure, success, and differentiate those tactics that can be replicated and thereby lead to greater success.

Traditional reporting within a longitudinal context allows the analyst to see what has happened within the organization and review how performance indicator scores trend over time. Organizing the data along different quantitative and qualitative hierarchical dimensions such as location, organization, customer profile, product categories, among many others, lets the analysts slice and dice the data to look for explicit business opportunities or process improvements. Interactive environments such as dashboards and mash-ups facilitate informed strategic decisions.

Upon making adjustments to the strategic direction of the organization, the incorporation of real-time delivery of key performance metrics allows senior stakeholders to review the impact and results of those strategic decisions within a short time window, thereby increasing agility, reducing risk, and allowing the organization to rapidly respond to emerging business opportunities. More pointedly, gaining a deeper insight into how the organization works (or doesn’t work!) can inform strategists to make significant changes to the way the company does business.

3.2 Integrating Analytics and Operations

Actually, the results of predictive analytics formulated as actionable intelligence can be integrated directly into operational processes even without the awareness of business consumers. For example, merging the results of predictive analytics with customer profiles empowers decision-makers at all levels of the organization to recognize and react rapidly to emerging opportunities, such as real-time adjustments to call-center scripts, or rerouting of product deliveries based on real-time warehouse and stock room inventory data, point-of-sales data, or even traffic or weather data. Customer preferences, profiles, and web statistics may drive dynamic content realignment on web sites to improve end-user response.

3.3 Common Value Drivers

Given the common desire to continuously improve business interactions with customers and partners, we can consider how business intelligence and analytics inform decision-making processes related to value drivers common across all industries. Today, most organizations use data in two ways: transactional/operational use ("running the business"), and analytic use ("improving the business"). When the results of analysis permeate the operational use, the organization can exploit discovered actionable knowledge to drive improvement.

A straightforward approach to managing business transformations using analytics involves classifying opportunities for positive business impacts within a simple classification scheme listing primary categories for potential business improvement, including the following areas:

- **Financial** opportunities, such as decreased operating costs, increased revenues, identifying new opportunities, speeding cash flow, or decreased penalties, fines, and other charges.
- **Risk and Compliance** opportunities associated with more precise credit assessment, reducing investment risks, becoming more competitive, wiser decisions regarding capital investment and/or development, reducing fraud and leakage, and auditable compliance with government regulations, industry expectations, or self-imposed policies (such as privacy policies).
- **Confidence and Satisfaction**-based opportunities, such as increased customer satisfaction, improved working conditions and improved employee retention, streamlining the supply chain, more precise and accurate forecasting, increased organizational trust, greater consistency in operational and management reporting, and better decisions.
- **Productivity** opportunities such as decreased workloads, increased throughput, decreased manufacturing defects, and improved end-product quality.

This categorization is intended to support the analytics process and help in clarifying general business objectives and corresponding performance metrics and indicators. Improving the business using analytics requires more than just installing and running the tools; the key stakeholders must define achievable targets and use the tools to both inform the decision-making processes and to assess, measure, and control the degree to which objectives are being met. Specific programs can be designed and developed around improvements within any of these key categories. Consider these examples:

- *Revenue Generation via Customer Profiling and Targeted Marketing* – Customer analytics can encompass a continuous refinement of individual customer profiles incorporating demographic, psychographic, and behavioral data about each individual to support customer community segmentation into a variety of clusters based on discriminating variables and corresponding value sets. Given these different clusters of customers and understanding the classification schemes allows one to develop micromarketing strategies to augment campaigns targeting small clusters of customers with similar profiles. "Laser-style" marketing can focus directly at individuals using the results of customer analytics.

- *Risk Management via Identification of Fraud, Abuse, and Leakage* – Fraud, which includes intentional acts of deception with knowledge that the action or representation could result in an inappropriate gain, is often perpetrated through the exploitation of systemic scenarios. Fraud detection is a type of analysis that looks for transaction patterns with relevant frequency within some of these identified scenarios. At the same time, comprehensive analysis of the products and services provided to customers within the contexts of their contracts/agreements may highlight leakage. Both of these risks can be analyzed and brought to the attention of the proper internal authorities for remediation.

- *Improved Customer Satisfaction via Profiling, Personalization, and Customer Lifetime Value Analysis* – Customer lifetime value analysis calculates the measure of a customer's profitability over the lifetime of the relationship, incorporating the costs associated with managing that relationship as well as the revenues expected from that customer. Employing the results of customer profiling can do more than just enhance that customer's experience by customizing the presentation of material or content. Customer profiles can be directly integrated into all customer interactions, especially at inbound call centers, where customer profiles can improve a customer service representative's ability to deal with the customer, expedite problem resolution, and perhaps even increase product and service sales.

- *Improved Procurement and Acquisition Productivity through Spend Analysis* – Spend analysis incorporates the collection, standardization, and categorization of product purchase and supplier data to select the most dependable vendors, streamline the RFP and procurement process, reduce costs, improve the predictability of high-value supply chains, and improve supply-chain predictability and efficiency.

These are just a few of the many examples where analytics can be used to optimize value drivers common across all industries.

3.4 Industry Examples

From one standpoint, opportunities for improvement manifest themselves differently depending on the industry, while from another standpoint there are common dimensions of operations that can be improved no matter which industry. Using the same value driver hierarchies, a company within a particular industry can benefit from reporting and analytics specifically tuned to that industry, such as these "vertical" examples:

- **Health Care** – Monitoring business process performance permeates all aspects of quality of care. For example, understanding why some practitioners are more successful at treating certain conditions can lead to improved quality of care. Analytics can help to discover the factors that contribute to success of one approach over others, and see whether those successes are dependent on variables within the control of the practitioner or factors outside their control. Improved diagnostic approaches can reduce the demand for high-cost diagnostic resources such as imaging machinery, and better treatments can reduce the duration of patient stays, freeing up beds, improving throughput, and enabling more efficient bed utilization.

- **Logistics/Supply Chain** – Integrated analysis for transportation and logistics management sheds insight into evaluation of many aspects of an efficient supply chain. For example, business intelligence is used to analyze usage patterns for particular products based on a series of geographic, demographic, and psychographic dimensions. Predictability becomes the magic word – knowing what types of individuals in which types of areas account for purchases of the range of products over particular time periods can help in more accurately predicting (and therefore meeting) demand. As a result, the manufacturer can route the right amounts of products to reduce or eliminate out-of-stocks. At the same time, understanding demand by region over different time periods leads to more accurate planning of delivery packaging, methods, and scheduling. One can map the sales of products in relation to distance from the origination point; if sales are lower in some locations than others, it may indicate a failure in the supply chain that can be reviewed and potentially remediated in real time.

- **Telecommunications** – In an industry continually battling customer attrition, increasing a customer's business commitment contributes to maintaining a long customer lifetime. For example, examining customer cell phone usage can help to identify each individual's core network. If a customer calls a small number of residential land lines or personal mobile phones, that customer may be better served by a "friends and family" service plan that lowers the cost for the most frequently called numbers. Identifying household relationships within the core network may enable service bundling, either by consolidating mobile accounts, or by cross-selling additional services such as landline service, internet, and other entertainment services. On the other hand, if the calls from the customer's individual mobile phone are largely to business telephone numbers and have durations between a half hour to an hour, that customer may be better served with a business telephony relationship that bundles calling with additional mobile connectivity services.

- **Retail** – The large volume of point of sales data makes it a ripe resource for analysis, and retail establishments are always looking for ways to optimize their product placement to increase sales while reducing overhead to increase their margins, especially when market baskets can be directly tied to individuals via affinity cards. Understanding the relationship between a brick-and-mortar store location and the types of people that live within the surrounding area helps the store managers with their selection of products for store assortment. Strategic product placement (such as middle shelf or end-cap) can be reserved for those items that drive profitability, and this can be based on a combination of product sales by customer segment coupled with maps of customer travel patterns through the store. Product placement is not limited to physical locations; massive web logs can be analyzed for customer behavior to help dynamically rearrange offer placement on a web site, as well as encourage product upselling based on abandoned cart analysis, through collaborative filtering, or based on the customer's own preferences.

- **Financial Services/Insurance** – In both insurance and banking, identifying risks and managing exposure are critical to improved profitability. Banks providing a collection of financial services develop precise models associated with customer activities and profiles that identify additional risk variables. For example, analyzing large populations of credit card purchases in relation to mortgage failures may show increased default risk for individuals shopping at particular shopping malls or eating at certain types of fast food restaurants. In turn, recognizing behaviors that are indicative of default risk may help the bank anticipate default events and reach out to those individuals with alternate products that keep them in their homes, reduce the risk of default, and improve predictability of the loan's cash flow over long periods of time.

- **Manufacturing -**Plant performance analysis is critical to maintaining predictable and reliable productivity; tracking production line performance, machinery downtime, production quality, work in progress, safety incidents, and delivering measurements of operational performance indicators along the management escalation chain so that adverse events can be addressed within the proper context within a reasonable timeframe.

- **Hospitality** – Hotel chains assess customer profiles and related travel patterns ,and know that certain customers may be dividing their annual "night allocation" among the competitors. By analyzing customer travel preferences and preferred locations, the company may present incentive offers through the loyalty program to capture more of that customer's night allocation.

The examples for these industries are similar in that the analysis ranges from straightforward reporting of key business performance indicators to exploring opportunities for optimizing the way the organization is run or improving interactions with customers and other business partners. Investigation of the business processes and performance measures from any industry will yield suggestions for ways to specifically benefit from reporting and analytics.

4 Why Pervasive Analytics?

The concept of "business intelligence" and "analytics" include tools and techniques supporting a collection of user communities across an organization, as a result of collecting and organizing numerous (and diverse) data sets to support both management and decision making at operational, tactical, and strategic levels. Through data collection, aggregation, analysis, and presentation, actionable intelligence can be delivered to best serve a wide range of target users. Organizations that have matured their data warehousing programs allow those users to extract actionable knowledge from the corporate information asset and rapidly realize business value.

But while traditional data warehouse infrastructures support business analyst querying and canned reporting or senior management dashboards, a comprehensive program for information insight and intelligence can enhance decision-making process for all types of staff members in numerous strategic, tactical, and operational roles. Even better, integrating the relevant information within the immediate operational context becomes the differentiating factor. Offline customer analysis providing general sales strategies is one thing, but real-time actionable intelligence can provide specific alternatives to the sales person talking to a specific customer based on that customer's interaction history in ways that best serve the customer while simultaneously optimizing corporate profitability as well as the salesperson's commission. Maximizing overall benefit to all of the parties involved ultimately improves sales, increases customer and employee satisfaction, and improves response rate while reducing the cost of goods sold – a true win-win-win for everyone.

The wide ranges of analytical capabilities all help suggest answers to a series of increasingly valuable questions:

- **What?** – Predefined reports will provide the answer to the operational managers, detailing what has happened within the organization and various ways of slicing and dicing the results of those queries to understand basic characteristics of business activity (e.g., counts, sums, frequencies, locations, etc.).Traditional BI reporting provides 20/20 hindsight – it tells you what has happened, it may provide aggregate data about what has happened, and it may even direct individuals with specific actions in reaction to what has happened.
- **Why?** – More comprehensive ad hoc querying coupled with review of measurements and metrics within a time series enables more focused review. Drilling down through reported dimensions lets the business client get answers to more pointed questions, such as the sources of any reported issues, or comparing specific performance across relevant dimensions.
- **What if?** – More advanced statistical analysis, data mining models, and forecasting models allow business analysts to consider how different actions and decisions might have impacted the results, enabling new ideas for improving the business.
- **What next?** – By evaluating the different options within forecasting, planning, and predictive models, senior strategists can weigh the possibilities and make strategic decisions.
- **How?** – By considering approaches to organizational performance optimization, the senior managers can adapt business strategies that change the way the organization does business.

Information analysis makes it possible to answer these questions. Improved decision-making processes depend on supporting business intelligence and analytic capabilities that increase in complexity and value across a broad spectrum for delivering actionable knowledge (as is shown in Figure 1). As the analytical functionality increases in sophistication, the business client can gain more insight into the mechanics of optimization. Statistical analysis will help in isolating the root causes of any reported issues as well as provide some forecasting capabilities should existing patterns and trends continue without adjustment. Predictive models that capture past patterns help in projecting "what-if" scenarios that guide tactics and strategy towards organizational high performance.

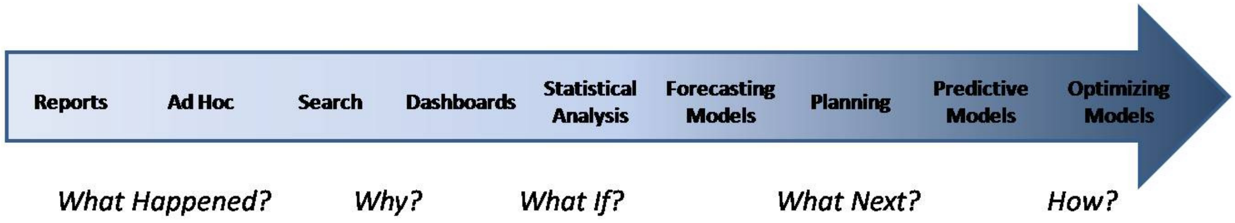

Figure 1: A range of techniques benefits a variety of consumers for analytics.

Intelligent analytics and business intelligence are maturing into tools that can help optimize the business. That is true whether those tools are used to help

- C-level executives review options to meet strategic objectives,
- Senior managers seeking to streamline their lines of business, or
- Operational decision-making in ways never thought possible.

These analytics incorporate data warehousing, data mining, multi dimensional analysis, streams, and mash-ups to provide a penetrating vision that can enable immediate reactions to emerging opportunities while simultaneously allowing one to evaluate the environment over time to discover ways to improve and expand the business.

5 Understanding the Technology Considerations

Early data warehousing architects struggled with designing and executing effective approaches for the extraction of data from source systems, transforming that data into a format suitable for analysis, and loading the data into the data warehouse. Once the data has been organized within the data warehouse, reporting and analysis tools such as Online Analytical Processing (OLAP) and query & reporting tools were employed to deliver data to the knowledge consumer.

Although some of the core aspects of those approaches remain unchanged, the details of that technical infrastructure have evolved since the mid 1990's in order to meet the increasing demand for orders of magnitude more data available from many more sources. In turn, both the variety and expectations of the pool of knowledge consumers has increased the demands on the types, scalability, and delivery mechanisms for actionable knowledge. Today, delivering the right information to the right people at the right time is the culmination of best practices in data management and organization combined with a technical infrastructure designed to direct many steady channels of data into a high performance analysis platform that can deliver trustworthy results within real time constraints. This depends on these areas of technology:

- Continuous synchronization of data from multiple sources to provide a coherent and consistent view;

- Cohesive information integration allowing for massive amounts of high quality data to be harmonized and aligned to enable effective analysis;

- A robust architecture for collecting, ordering, and managing very large sets of data for analysis;
- Performance tuning characteristics ranging from hardware improvements to adjusted programming and execution models to handle large unstructured data sets;
- Large-Scale computing, involving the processing and analysis of large data sets accumulated from many data sources, often with expectations of results provided in real time;
- Comprehensive set of analytics services (including undirected analysis, predictive models, and text analytics) that reduce or even eliminate the need to be a power user to derive benefit from actionable intelligence; and a
- Framework for reporting, ad hoc queries, dimensional analysis, and visualization and communication of actionable knowledge.

In this section we look at some of the key technical considerations necessary for a modern analytics environment, and how the results can be fully integrated into hundreds, if not thousands, of daily business processes. Given an understanding of the technology components, we will begin to see some challenges emerge in environments with heterogeneous technologies cobbled together to support the analytics program.

5.1 Continuous Availability: Data Synchrony & Coherence

A recurring challenge of the traditional data warehousing framework involves the elongated turnaround time between the articulated need for reporting or analysis and its delivery. Delaying the delivery of actionable knowledge derived from current data makes it difficult to make timely decisions, and long data latencies will impact the ability to assess the results of those decisions within a reasonable time frame. Essentially, data synchronization has become a critical component within the infrastructure in order to integrate with the strategic and operational decision-making activities. The criticality of timely and current data cannot be understated. For example, increased sales by product category in particular regions may suggest a growing demand requiring immediate reallocation of logistics resources to prevent out-of-stocks and feed that demand. Relying on week-old order information is insufficient; instead, current shelf and warehouse stock information coupled with shipping resource allocation enables immediate decisions that maintain a steady flow of product to the customers.

Pervasive business intelligence and analytics require a high degree of data synchrony, meaning that the environment must

- Reduce or eliminate data latency;
- Maintain coherence of the data being used for analysis across the enterprise;
- Provide timely and current information; and
- Provide consistent, deterministic results to similar requests.

Maintaining a reasonable degree of synchrony and coherence among the multitude of data sources available within (as well as from outside) the organization requires technical strategies for continuous data availability that do not impose a strain on the environment. Some of those strategies include:

- Change data capture (CDC) for data replication, involving a managed, synchronized copying of data from a source to one or more target data systems. CDC is an event-driven mechanism for capturing changes from source datasets and propagating those changes through various channels, either directly to target databases or through a message queue for subsequent processing. *Synchronizing data modifications via CDC enables coherence between operational systems and analytical systems, enabling the discovery of actionable opportunities in real-time while maintaining consistency across reporting systems.*

- Data Federation, which enables transparent access to heterogeneous (and generally physically distributed) data types, platforms, and sources, and in numerous formats, without requiring a staging area or centralized repository (see Figure 2). Federation is an effective way to capture subsets of very large or very distributed data sets, and is frequently used, when data is offsite, is in an older format, or is infrequently used. For example, a data federation framework will allow an application to access databases, XML data, flat files, or even data services or data streams using a uniform access mechanism. In turn the federation server dynamically accesses the data sources and returns the synchronized results. *Federation simplifies consolidating data from multiple sources, enabling "cross-pollination" of information to better discover opportunities, and is very effective at joining dissimilar data before it arrives to the target.*

- Information Stream processing, which provides applications with continuous access to streaming data sources in real time. *Connecting streaming data with persistent data sources supports complex event processing and real-time discovery of opportunities based on emerging knowledge activities, such as weather-based commodity trading or immediate deliveries to prevent retail out-of-stocks.*

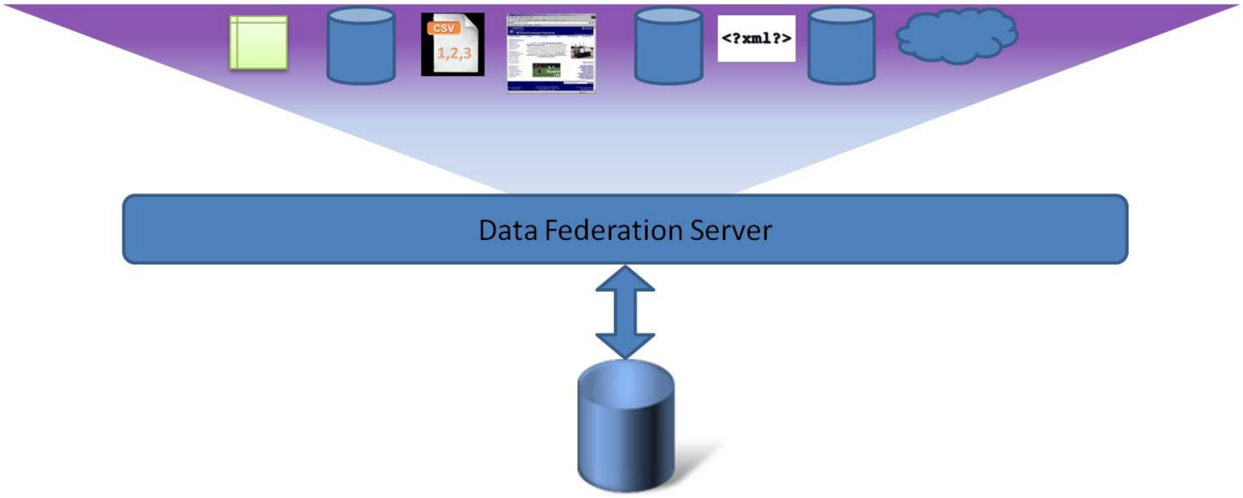

Figure 2: Data federation services provide dynamic, transparent access to numerous data sources.

5.2 Cohesive Information Integration

Because most of the data to be analyzed is created upstream from the point of analysis for specific transactional or operational purposes, it is necessary to provide a cohesive information integration framework to facilitate information acquisition and sharing. In other words, preparing enterprise data for analysis requires collaborative techniques for acquiring data from multiple sources, managing representations of similar data concepts in variant representations, and then sharing information with multiple consumers across different applications.

- Data Integration and ETL – Like data federation, data integration relies on the ability to seamlessly access data from many different sources and deliver that data to numerous targets, but it encompasses much more when it comes to parsing extracted data, normalizing it into standard formats, cleansing, and transforming the data into a representation that is suitable for loading into a data warehouse and any subsequent uses for reporting and analysis. *ETL is a key enabler of data warehousing, providing the fundamental information flows that enable business intelligence altogether.*

- Master Data Management – "Master data" objects are those core business concepts represented in different data silos and used in the different business applications across the organization, along with their associated metadata, attributes, definitions, roles, connections, and taxonomies and hierarchies. Some common examples include Customers, Suppliers, Parts, Products, Locations, Agreements, Contact mechanisms. Master Data Management (MDM) incorporates the business applications, information management methods, and data management tools to implement the policies, procedures, infrastructure that support the capture, integration, and subsequent shared use of accurate, timely, consistent and complete master data. MDM should provide the ability to manage access to uniquely identifiable representations for each master data entity across the application infrastructure. *Providing a* *"unified view" of core data concepts assembled from multiple sources helps to reduce duplication of customers, products, suppliers, or any other key data asset class.*

- Metadata Management – Just the right amount of metadata is necessary in order to support the data extraction, integration, and analysis requirements to ensure consistency of meaning. Metadata incorporates the business definitions (associated with data concepts, business terms, and associated definitions and semantics), conceptual reference data domains and their corresponding values, data element definitions, data element formats, structures, and data types, as well as entity models, entity relationships, and supporting metadata for data governance such as information usage maps, data quality rules, and access controls. *Managing data definitions, reference tables, and data mappings inspires information reuse and helps synchronize the semantics of data as it is integrated from across many systems.*

- Data Profiling – Like any asset, it is valuable to inventory the asset – determine what it is, who it belongs to, where it is used, and its serviceability. Data profiling does this – it supports a combination of artifact review and empirical analysis of source data sets to understand the characteristics of data element metadata that are critical for analysis. Data profiling can provide the "ground-truth" associated with the actual data values as well as provide evidence of consistency with metadata, and will provide insight into the suitability of candidate sources to satisfy the target analytical needs. *The result of a data quality assessment using data profiling in conjunction with a review of the consuming application expectations can uncover data quality rules that are managed within the metadata repository.*

- Data Cleansing – Data profiling will expose potential anomalies and errors in the data, and erred data used as input to analytical processing will impact the believability of its results. When the source data system owners are able to address design or process flaws that introduce data issues, then the root causes can be eliminated and so can the errors. However, when the source data is outside of the organization's administrative control, processes and technical infrastructure must be in place to parse, standardize, enhance/enrich, and cleanse the data to satisfy the downstream analytic needs. Data cleansing is generally required when the data is being repurposed, especially when the new purposes have stricter data quality expectations. *Data cleansing enhances the value of the data and high quality, consistent data makes the decision-making process trustworthy.*

- Data Validation – Alternatively, there are often opportunities for introducing new errors into the data, and identifying and eliminating potential data flaws early in the information production flow will reduce the variance and inconsistency downstream and improve overall operational efficiency. *But even more importantly, validating supplied data with defined data quality rules contributes to the delivery of trustworthy data.*

- Identity Resolution – Due to the organic growth of the myriad operational systems deployed across the enterprise, it is not unusual that multiple data instances in different systems represent the same real-world entities in a variety of ways. Alternatively, there are often situations in which two real-world entities share the same identifying attributes, making it difficult to distinguish between them. Both of these types of issues reflect the same core challenge: the ability to evaluate the similarity between pairs of records and determine whether they represent the same thing. Identity resolution addresses these issues by calculating and scoring the degree of similarity between any two records. When the score is above a specific threshold, the two records are presumed to match; below another threshold, they are deemed to not match. Identity resolution is used to match records in the presence of variations or incomplete attributes, or to determine that two records truly represent distinct entities. *Identity resolution is a key component of master data management, and increasing precision in entity matching reduces data duplication and supports high quality reporting and analysis.*

- Pervasive Delivery Mechanisms – The spreadsheet is no longer the sole means for delivering analytical results. Self-service configuration of reports and query results from within business applications eliminates the bottleneck caused by relying on IT staff for support, and web-based delivery simplifies the availability of results using a variety of intuitive, interactive visual presentation objects (such as graphs, heat maps). *By automating event-driven notifications that can be tailored to an assortment of interfaces, ranging from desktops to hand-held PDAs, actionable intelligence can be provided directly where it is needed.*

5.3 Managing Knowledge Using a Robust, yet Extensible Architecture

While the key stakeholders at most organizations are concerned with each of our high level value drivers, each company's areas of business coupled with external pressures specific to its lines of business drive a need for a framework that addresses improvements to the common value drivers while informing the business processes and the corresponding reporting and analytics needs that are unique to each industry. That means a data warehouse model integrated with those commonly-used master data domains (such as "customer," "product," or "agreement") that allows flexibility in enhanced models to solve industry-specific challenges. Consider these examples:

- **Banking**, where there is a need for models that focus on a balance of enterprise risk management and oversight of customer visibility and product profitability. From the risk perspective, this combines managing compliance with (and the statutory reporting associated with) regulatory directives at a variety of governmental levels ranging from Basel II, Sarbanes-Oxley, international financial reporting, as well as managing corporate financial, counterparty, and credit risk. At the same time, these businesses still have the need for a variety of reports and analyses of customer interactions that help in maintaining customer loyalty, increase retention, and measure channel effectiveness.
- **Insurance**, in which there is a need for models that address challenges associated with reviewing and managing claims, per member/per month reporting, intermediary performance, compliance (such as the Solvency II requirements), risk management, basic life and pension

actuarial, and corporate pensions compliance, yet still requiring visibility into customer management and retention, product performance, and marketing and sales effectiveness.

- **Retail**, where the need is for reporting and analytics for customer management, merchandising management, products & services management, store operations management, corporate finance management, supply chain, management of multiple sales channels, as well as regulatory compliance visibility. In addition, the results of real-time predictive analytics can be streamed back into operational marketing and sales processes to improve revenue generation and reduce customer attrition.
- **Telecommunications**, in which there is a desire to monitor marketing performance, advertising management, revenue assurance, provisioning, rating, billing, credit risk, customer device and plan analytics, as well as reporting usage for both voice and non-voice communications, customer retention and attrition, as well as system and network performance, and usage analysis (and dropped-calls!) to inform decision-making for capital improvements in the telecommunications infrastructure.

As you can see, all of these examples require a robust architecture that supports the collection, aggregation, ordering, and management of very large sets of data. And although all these industry examples share a number of common objectives surrounding marketing and sales effectiveness or customer relationship management, each has different mixes of workloads for reporting, analytics, or both that can benefit from robust industry data warehousing models.

5.4 Performance Enhancing Techniques

In light of the incredible rate of data volume growth, the question of maintaining high levels of system performance moves to the forefront, especially with a community of business intelligence and analytics users spanning a broad spectrum of consumer types, such as

- Standard business consumers, who use domain-specific reports coupled with their own ad hoc queries and direct interactive analyses;
- Casual consumers, who typically review functional or operational performance metrics summarized from pre-designed reports presented via scorecards or dashboards;
- Operational analytics consumers whose business processes are informed through embedded results from operational analytics;
- Extended enterprise consumers, comprising external parties, customers, regulators, external business analysts, partners, suppliers, or anyone with a need for reported information for tactical decision-making; and
- Power users and sophisticated analysts using a variety of tools and techniques to analyze massive amounts of data and whose results will inform decision-making processes.

In general, the different consumers and their prototypical uses reflect a general mix of demand for reporting and analysis. Consequently, because of the different types of needs for results of these different types of analyses, the business intelligence architecture must support a mixed workload of rolled-up and pre-designed reports, reporting and analytics prepared by power users, ad hoc queries, interactive analysis, as well as interactive drill down into the supporting data sets.

Preparing massive amounts of data, as well as supporting a mixed workload of reporting, interactive analysis, and complex analytics means more than just throwing “big iron” at the problem. Rather, one must consider how any aspects of the information production flow that could be impacted as a result of poor performance, from the point at which data flows into your analytics environment, to the point where massive amounts of data must be made available for reporting, end-user interaction, slicing, and dicing.

The need for performance has to be evaluated in light of four areas of the information production flow: at the point of data integration and loading into the data warehouse, as the data is subjected to queries and algorithmic analysis, delivering results in an interactive manner to the user community, and persistent storage. We can look at these specific techniques for enhancing performance:

- Analytic database appliances, which are systems designed with built-in architectural enhancement intended to improve processing speeds;
- Alternate programming models such as Hadoop to enable collaboration between algorithmic analysis and database analytics; Compression, which is intended to reduce storage to improve loading and query response times; and
- Memory hierarchy enhancements, such as using solid state storage devices in place of spinning disks.

5.4.1 Large-Data Analytics

End users should not be limited to only accessing filtered and reduced results. When massive amounts of unstructured data contribute to an analysis process, the end-user may need to drill back to the original source. However, the absence of context may lead to inconsistency and confusion. Maintaining historical information about the relevance and content of each collected data artifact (such as a web page, presentation, or video) within the context of its connections to other artifacts provides one set of dimensions for analysis. In addition, enriching the data artifacts with tagged meanings based on defined semantic hierarchies provides additional contexts and consequently additional dimensions for review.

This means that the analytic environment must be capable of aspects of provisioning solutions to “big” problems. This involves a combination of managing rapid access to a massive number of data artifacts and integrating the results of computational analytics (such as those performed using Hadoop) with structured data for reporting and analysis within a data warehouse. Together, these capabilities contribute to the types of emerging advanced analytic services that drive the discovery of insight and actionable knowledge.

5.4.2 Architectural Enhancements: the Analytic Database Appliance

By virtue of the fact that data warehousing has extremely limited need to support transactions, data warehouse designers are willing to sacrifice transaction performance in return for increased performance of complex queries. This idea has led to the concept of the "analytic database appliance" – a system that integrates high performance computational resources, high speed I/O and networking, database management, and potentially other tools pre-installed, configured, and optimized for mixed-workload reporting and analysis.

Analytic database appliances are often comprised of specially optimized software layered on top of a massively parallel processing system, although the underlying storage architecture can be configured in different ways. Three common approaches are described based on their storage configuration:

- Shared nothing, where each independent processing unit is connected to its own memory and communicates directly with its own disk system.
- Shared disk, in which each independent processing unit is connected to its own memory but differ from the shared-nothing approach in that all processing units access a common disk system.
- Shared everything, where every processing unit has access to a shared memory systems as well as access to shared disk storage.

Analytic database appliances also are engineered for optimized data allocation. Instead of a traditional row-oriented data layout, some appliances organize and store their data using a columnar layout. Because each column can be stored separately, for any query, the system can evaluate which columns are to be accessed and therefore only retrieve the values requested from those columns. Indexing can be simplified also, since the values within each column can be ordered and used to form the index, which not only modulates the expansion of the database footprint, it also reduces the amount of data to be streamed in from secondary storage, which can dramatically improve query performance.

5.4.3 Integrated Complex Algorithmic Analysis Using Alternate Programming Models

Power users for advanced analytics focus on "big problems," such as distributed number crunching, complex statistical analyses, model development using data mining tools, and large-scale filtering and reduction, as a way of evaluating massive data sets, possibly a mixture of structured and unstructured data, often with the results to be integrated with operational systems. These users are likely to employ a variety of analytical techniques operating on embarrassingly parallel tasks, such as combining data warehousing with data mining algorithms or crafted parallel applications. In turn, the results of these applications may still need to be reported through traditional business intelligence approaches, so any BI environment should enable complex algorithmic analysis coupled with standard data warehouse models, especially in a massively parallel processing architecture.

Hadoop is a prime example of this type of alternate programming environment that is structured to exploit data distribution and massive parallelism. Hadoop is an open-source framework most importantly composed of two services. The first is a reliable distributed file system, and the second is a parallel programming model based on an approach called "MapReduce." The MapReduce programming model was introduced and described by Google researchers for parallel, distributed computation involving massive data sets (ranging from hundreds of terabytes to petabytes). As opposed to the familiar procedural/imperative languages like Java and C++, MapReduce's programming model mimics functional languages (notably Lisp and APL), mostly due to its dependence on two basic operations that are applied to sets or lists of data value pairs:

- *Map* which describes the computation or analysis applied to a set of input key/value pairs to produce a set of intermediate key/value pairs, and
- *Reduce,* in which the set of values associated with the intermediate key/value pairs output by the *Map* operation are combined to provide the results.

With some applications applied to massive data sets, the theory is that the computations applied during the *Map* phase to each input key/value pair are independent from one another. Combining both data and computational independence means that both the data and the computations can be distributed across multiple storage and processing units and automatically parallelized. This parallelizability allows the programmer to exploit scalable massively parallel processing resources for increased processing speed and performance. In turn, the results of the algorithms can be realigned with a data warehouse's dimensional model to integrate algorithmic analysis with traditional reporting.

5.4.4 Relieving Storage Demands Using Compression

Even though the average cost of disk storage continues to decrease, the explosive growth rate for all sorts of data outpaces the relative decrease in cost. To make matters worse, the storage needs of a business intelligence environment go beyond what is stored in the data warehouse. Because of the use of indexes, operational data stores, staging areas, test environments, and a development environment, there is a need for ways to reduce the storage footprint without a loss of information.

One common technique for achieving this objective is compression. Compression is a process of reducing the need for storage through recognition and replacement of common patterns with smaller-sized data items. Compression becomes particularly relevant when loading data into your data warehouse environment. As structured data sets grow by leaps and bounds, the restricted bandwidth of the common Input/Output (I/O) channels puts a stranglehold on timely movement of data into the analysis framework. Alternatively, user queries reliant on multi-table joins can also be hobbled as a result of limited I/O bandwidth. Reducing the storage footprint through compression and applying creative data layouts and filters can reduce the demand for data movement, thereby leading to increased performance, reduced costs, and increased sustainability.

There are different compression algorithms; some examples include run length encoding, in which repeating data items are represented once along with their corresponding counts, as well as applications of Lempel-Ziv compression along data values in rows. In the latter example, common patterns are extracted and represented using compressed numeric values, which then replace the pattern each time it occurs, thereby reducing the storage need. Good compression techniques can reduce storage by 50%, and frequently result in reducing storage by as much as 80%.

5.4.5 Increased Performance via Memory Hierarchy Improvements

Different computing architectures often vary in relation to the degree to which multiple system aspects such as CPUs, cache memory, core memory, temporary disk storage areas, and persistent disk storage contribute to maximizing system performance, especially in the context of a mixed business intelligence workload. These storage modules, among others, constitute what is referred to as the “memory hierarchy,” which hardware architects employ in varying configurations to find the right combination of memory devices with varying sizes, costs, and speed to provide optimal results by reducing the latency for responding to increasingly complex queries and analyses.

Performance-driven applications attempt to use the memory hierarchy to minimize latency through intelligent data access and caching strategies. The value of these improvements can be multiplied with strategic use of hardware. For example, responding to different types of queries requires writing to temporary memory space, most frequently using spinning disks as the medium. Using solid state flash memory for temporary storage provides faster access, reduces I/O latency, decreased query response time, and consequently results in increased query throughput.

5.4.6 Stream Computing

The decision-making process is increasingly informed with a growing variety of inputs, of which fewer and fewer are based on structured data. Alternatively, decision-makers consume a broad swath of different kinds of unstructured data from many different sources. Manually scanning all of these inputs would create a significant bottleneck to productivity. Instead, there is a need for methods to not only rapidly scan the multitude of data sources and continuously analyze streamed data in real time so that the right bits and pieces of information can be filtered, reduced, and transformed into actionable knowledge to be delivered to the right individual at the right time. As more sources of information become available, these methods must adapt to heterogeneous data in different sources, formats, and even types, including text, audio, pictorial, and video.

In essence, this requires efficient methods for analyzing data streams, allowing one to analyze every bit of data being sent so that immediate action can be taken when critical information can be inferred. Data can also be selectively saved for subsequent processing. This type of information stream analysis effectively provides a “data federation” capability layered on top of a query engine applied to analyze massive amounts of dynamic data from many heterogeneous sources. The analyses provided by stream computing enable rapid response to events and changing environments by updating query result sets in lock-step as the data sources are refreshed. In essence, multiple (variant) data sources can be reviewed, filtered, cleansed, aggregated, or subjected to additional transformations to feed numerous downstream data consumer needs.

5.5 Emerging Advanced Analytics Services

In addition to what might be called “mainstream” analytics, there are emerging techniques that are rapidly being incorporated into the environment to support pervasive operational intelligence, such as:

- Data mining and predictive modeling
- Embedded predictive analytic models
- Entity recognition and extraction
- Text analysis
- Sentiment analysis

5.5.1 Data Mining and Predictive Modeling

Data mining and other advance statistical techniques enable analysts to build models that, to some extent, replicate some of the thought processes performed to recognize patterns for success. Business analysts have applied data mining techniques with no preexisting expectations to identify patterns warranting further determination of business value.

This *undirected analysis* relies on no preconditions, and can be a good way to start developing predictive models. Analysts using data mining techniques develop predictive models that can be refined, trained, and then applied to very large data sets to identify patterns corresponding to opportunities or risks. For example, clustering customer data is an undirected process to group customers based on discovered similarities and differences. Evaluating the dependent variables resulting from clustering customer data enables directed classification of new customer records into the discovered clusters.

There are a number of data mining methods and techniques that can be combined to develop predictive models, such as:

- **Clustering**, a process to group items such that each group is clearly different from all others, and that the members of each group are recognizably similar.
- **Association**, in which a collection of data instances is analyzed to find rules that could be used to predict the occurrence of one set of values based on the occurrences of other values within each instance.
- **Regression**, which is a statistical method to analyze a data set to fit a collection of data points to a mathematical formula. In turn, that formula can be used to plug in values to enable prediction of dependent variable.
- **Market Basket Analysis**, which is a process that looks for relationships of objects that belong together within the business context, and its name is derived from the concept of analyzing the contents of a supermarket shopper’s cart to see if there are any “naturally occurring” affinities that can be used for business advantage.
- **Case-based reasoning**, in which known situations are used to form a model for analysis. New situations are compared against the model to find the closest matches, which can then be reviewed to inform decisions about classification or for prediction.
- **Decision tree analysis**, which looks at a collection of data instances and given outcomes, evaluates the frequency and distribution of values across the set of variables, and constructs a decision model in the form of a tree. The nodes at each level of this tree each represent a

question, and each possible answer to the question is represented as a branch that points to another node at the next level.

- **Neural networks**, which uses probabilistic and statistical methods to analyze training data to create a “black box” process that takes some number of inputs and produces some predictive output.

5.5.2 Embedded Predictive Analytic Models

The predictive models developed using a variety of data and text mining algorithms can be integrated into business processes to supplement both operational decision-making as well as strategic analysis, using the patterns that have been revealed to predict future events or help in achieving a particular goal. For example, customer profiles created though the application of a clustering analysis can be used for real-time classification based on specific demographics or other characteristics. These profiles can be used to recommend opportunities for cross-selling and up-selling, thereby leading to increased revenue. Embedded predictive models can be used to address all of our value drivers, and are used in many different scenarios, including customer retention, acquisitions and procurement, supply chain improvements, fraud modeling, improving forecasting accuracy, clinical decision-making, credit analysis, and automated underwriting.

5.5.3 Entity Recognition and Entity Extraction

A consequence of the growing flood of unstructured artifacts is the challenge in isolating key terms in text (such as people, places, and things) and establishing linkages and relationships among those concepts. “Real-time” entity identity recognition has traditionally been a batch process, but time-critical operations relating to online comments, customer service, call center operations, or more sensitive activities involving security, bank secrecy act/anti-money laundering, or other “persons of interest” applications become significantly more effective when individual identities can be recognized in real time.

The challenge goes beyond finding name patterns in sequential text, but builds on natural language processing concepts to expose relationships (such as an individual’s affinity for a particular charity), causal relationships (such as correlation of product issues within geographical regions), or multiple references to the same entity (such as the introduction of pronouns like “He” or “It” that refer to named entities with proper pronouns such as “George Washington”). Real-time identity recognition enables rapid linkage between individuals and their related attributes, characteristics, profiles, and transaction histories and can be used in real-time embedded predictive models to enhance operational decision-making.

5.5.4 Text Analysis

Text analysis can be used to isolate key words, phrases, and concepts within semi-structured and unstructured text, and these key text artifacts are analyzed semantically, modeled, and their source documents correlated based on recognized concepts. This implies the need for concept taxonomies in which like terms can be collected and aggregated at different levels of precision, such as car makes, models, as well as alternative versions resulting from the presence or absence of specific features.

Algorithms for entity recognition, entity extraction, and text analysis need to be more sophisticated in relation to the text's structure. While simple, pattern-based unstructured entities (such as telephone numbers) can be scanned using techniques such as regular expression parsing, more complex pattern and context sensitive techniques are increasingly used. A standard for content analytics called the Unstructured Information Management Architecture (UIMA) has been established by OASIS in 2009, and guides the development of a framework for the integration of text analysis components and rules to drive the development of unstructured analysis software.

These components allow you to perform frequent term analysis, determine sentinel or signal terms, build concept hierarchies, create dictionaries, and document rules for phrase recognition and for concept extraction, among other techniques. Once this analysis is completed, the information contained within the documents can be clustered, categorized, and organized to support intelligent searches, and filtering concepts from streaming text helps identify important text artifacts that can be routed directly to individuals with a particular interest in the supplied content. Once the concepts have been ordered, identified, and extracted, they can be subjected to data mining and other types of analytic analysis to help the knowledge worker draw conclusions from actionable information.

5.5.5 Sentiment Analysis

Social media sites, online networks, and blogs provide ample opportunity for a wide variety of individuals to post subjective reviews, product evaluations, service ratings, experiences, and other opinions, often ones that influence a broad spectrum of other individuals within the author's network. As opposed to idly reflecting on what has been presented, organizations seek to take advantage of these growing networks of influence by rapidly addressing negative sentiment or exploiting positive sentiments.

Sentiment analysis takes text analytics to the next level through the analysis of unstructured text to review and evaluate subjectivity in the attitude of the material's author. For example, a product's manufacturer may analyze call center reports for part names appearing frequently with negative interactions such as product failures. Doing so can help identify common failure patterns, thereby allowing for proactive actions to reach out to product owners before the parts fails. Sentiment analysis presents other opportunities as well, such as identifying emerging consumer trends, identifying customer preferences, or finding unhappy customers. This allows businesses to manage their online reputation by highlighting positive opinions while reducing impacts due to negative opinions.

Sentiment analysis is a culmination of a number of techniques discussed in this paper – analyzing term frequency, deducing taxonomies and hierarchies, tagging document artifacts with their corresponding concept tags, organizing concepts in relation to alternate structured data, and applying data mining analyses to look for patterns, associations, causality, and other types of relationships.

5.6 Analytics Delivery Services

The range of analytics services supports the variety of data consumers across the organization:

- Reporting and Ad Hoc Querying – Standard, static reports derived from user specifications provide a consistent view of particular aspects of the business, generated in batch and typically delivered on a scheduled basis through a standard (web) interface. The static nature of reports drives the need for alternative methods for additional insight. One approach is to extract the reported data into spreadsheets for additional data manipulation, while also allowing ad hoc queries to gather additional data for analysis. *Standard reports can provide knowledge to a broad spectrum of consumers, even if those consumers must have contextual knowledge to identify the key indicators and take action. However, given the growth of data into the petabytes, standard reporting is rapidly yielding to exception reporting.*

- Scorecards and Dashboards – If a trained eye is required to scan key performance metrics from canned reports, simplifying the presentation of key performance metrics may better enable the knowledge worker to transition from seeing what has already happened to understanding the changes necessary to improve the business process. Scorecards and dashboards customize an up-to-date presentation of summarized performance metrics, allowing continuous monitoring throughout the day. Pervasive delivery mechanisms can push dashboards to a large variety of channels, ranging from the traditional browser-based format to hand held mobile devices. *Through the interactive nature of the dashboard, the knowledge worker can drill down through the key indicators regarding any emerging opportunities, as well as take action through integrated process-flow and communication engines.*

- Mash-ups – The mash-up takes the dashboard to the next level, allowing the knowledge consumers themselves the ability to identify their own combination of analytics and reports with external data streams, news feeds, social networks, and other web 2.0 resources in a visualization framework that specifically suits their own business needs and objectives. *The mash-up framework provides the "glue" for integrating data streams and business intelligence with interactive business applications.*

- Multidimensional Analysis and Online Analytical Processing (OLAP) – The multidimensional analysis provided by OLAP tools helps analysts "slice and dice" relationships between different variables (within their own hierarchies), such as "what are corporate revenues by time?" or "What is the availability of products by supplier by location?" The use of the word "by" suggests a pivot around which the data can be viewed, allowing one to look at sales grouped by time period, then by region, or the other way around, grouped by region *then* by time period. *OLAP lets the analyst drill up and down along the hierarchies in the different dimensions to uncover dependent relationships that are hidden within the hierarchies*.

6 Challenges

A mature business intelligence and analytics program will have a full complement of these technology components to support knowledge consumers across the full analysis spectrum. In many environments the analytics program has grown organically, with a variety of acquired tools, internally-developed solutions, integrated with different choices of hardware, network, and software (such as database management systems), leading to a workable, if not the most efficient, solution.

And while the horde of vendors have endeavored to support interoperability (so that they all seem to play well together), the BI systems have been engineered over time with heterogeneous components provided by different vendors, with little thought of the complexity of component integration, let alone performance or optimization. In fact, as the need for speed grows, we are recognizing that there are some latent challenges that exist for organizations as they try to “home brew” their comprehensive analytics infrastructure, including these factors:

- **Organic Development and Heterogeneity** – The organic nature of the development means that the analytic applications have been incorporated on an “on-demand” basis, with neither a comprehensive program plan nor an assessment of business needs across the enterprise. This leads to technical dependencies based on development decisions that are not specifically related to addressing business needs, and in time, those dependencies may impede the maturation of a flexible end-to-end analytics solution.

- **Flexibility and Extensibility** – Despite the attempts by many vendors to enable interoperability, each is limited in its ability to work well with those (usually released) product versions for which there are published specifications. In reality, this imposes stringent integration constraints which may prevent the customer from enabling all available product capabilities. For example, if the selected data cleansing tool only works with version 5.7 of the selected ETL product, the customer must refrain from upgrading to version 6 of the ETL product until the data cleansing vendor enhances its product to support the ETL product upgrade. Also, as business needs, requirements, analytics expectations, or numbers of consumers change, the underlying analytics infrastructure will have to adapt to those changes. This suggests a need for an ability to easily add capabilities and functionality to the business intelligence infrastructure.

- **Data quality** – Even though there is an increased concentration on data governance and quality management, there are still opportunities for data flaws to be introduced. Intermittent data validation, different data quality tools (for parsing, standardization, and cleansing), and conflicting rule sets *still* contribute to data flaws and inconsistency.

- **Time to value** – Installing, testing, and validating a variety of components and ensuring that they operate well together requires a significant time and resource investment in planning, design, implementation, and deployment. The increased complexity of implementation and deployment increases the time until the systems can be used productively.

- **Performance and scalability** – Many BI systems become a victim of their own success; as the number of users increases, the query load grows, or as the amount of data to be analyzed grows, the system's ability to scale appropriately leads to performance degradation. This scalability challenge only increases when interoperability constraints artificially throttle back the performance potential of any of the integrated components.

- **Rapid solutions** – Having to reengineer the same (or similar) reports and analyses is a drain on productivity and resources, and extends the time to value. In retrospect, organizations would benefit if they could incorporate tools and technologies to support frequently-used applications as customer insight and profitability, market analysis, spend analysis, and other common value drivers.

7 Summary: Addressing the Challenges

When considering all these challenges together, a common thread emerges: inefficiencies introduced due to the piecemeal accumulation of technology components from a multitude of sources will ultimately degrade the organization's ability to deliver actionable intelligence to the appropriate individuals at the necessary times, especially as the demand for analytics increases, whether that is due to increased embedded operational intelligence, larger data volumes, increased numbers of users, or (more likely) a combination of these demands.

And if a root cause of these risk factors is the variety of technology components, then mitigating those risks requires considering the options for creating an end-to-end solution that is architected to best take advantage of complementary components. A complete suite of tools that support the entire organization's reporting and analytics needs addresses our challenges:

- **End-to-end solution** – With a well-designed architecture, the program team can articulate a strategy for meeting business needs with a comprehensive solution in which all the components are designed to fit together. Selecting a complete solution from a single vendor not only simplifies the implementation and deployment, it also simplifies the procurement process while reducing the risks of a heterogeneous environment.

- **Flexibility and Extensibility** – A single provider can provide greater flexibility, especially when upgrades and releases can be synchronized in a way that ensures that functional improvements are not artificially limited by product versioning. In addition, the customer can introduce functionality as needed by deploying new upgrades or modules in lock-step with the strategic program plan.

- **Data quality** – Standardizing the data validation, cleansing, and enhancement tools and the way those tools are used provides a predictable level of consistency of enterprise data from its initial entry (or creation) to the numerous downstream consumers.

- **Time to value** – Reducing systemic complexity by unifying the complete solution platform simplifies the acquisition process, reduces the resource requirements for implementation, training, and deployment, thereby accelerating the time to value.

- **Packaged solutions** – Vendors with experience in solving general client challenges are able to integrate best practices into their packaged solution leveraging their technical componentry to address common value drivers such as revenue generation through cross-selling and up-selling or cost reductions using spend analysis. Alternatively, years of working with clients in defined lines of business or industries allows vendors to customize their solution offerings to take advantage of subject matter expertise to compose solutions to meet business needs within specific industries, such as location placement for retail business, routing optimization for logistics, or hazard zone assessment for the insurance industry.

- **Performance and scalability** – When an end-to-end solution is designed to run on top of specific hardware, the developers are able to take advantage of a number of optimizations integrated directly into both the hardware and software platforms, such as workload management, task and process scheduling, load balancing, parallel I/O channels, or high availability. Optimized analytical database management services allow for high performance analytical data warehousing, supported by parallelized data integration plus high speed federation services. Increasing numbers of queries can be offloaded to alternate processing units or routed to in-memory databases, decreasing DBMS loads while increasing response rates and throughput.

By transitioning from an organic evolution of corporate business intelligence and analytics environment built on top of a myriad of technology components to a strategically-architected end-to-end solution, your organization can gain a rapid time to value through real-time, integrated analytics, resulting in advantageous intelligence delivered to the appropriate decision-makers at the right time.

About the Author

David Loshin, president of Knowledge Integrity, Inc, (www.knowledge-integrity.com), is a recognized thought leader and expert consultant in the areas of data quality, master data management, and business intelligence. David is a prolific author regarding BI best practices, via the expert channel at www.b-eye-network.com and numerous books and papers on BI and data quality. His book, "Business Intelligence: The Savvy Manager's Guide" (June 2003) has been hailed as a resource allowing readers to "gain an understanding of business intelligence, business management disciplines, data warehousing, and how all of the pieces work together." He is the author of "Master Data Management," which has been endorsed by data management industry leaders, and the recently-released "The Practitioner's Guide to Data Quality Improvement," focusing on practical processes for improving information utility. Visit http://dataqualitybook.com for more insights on data quality, and http://mdmbook.com for thoughts on master data management.

David can be reached at loshin@knowledge-integrity.com.