Business Integrated Insight (BI^2)

Reinventing enterprise information management

August 2009

A White Paper by

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

"Le Roi est mort. Vive le Roi!" ***The message of this paper is at one with this ancient proclamation: the continuity of rule through the end of one era and the beginning of the next. The data warehouse has had a long and illustrious reign, but today a combination of business and technological change has laid the old king low. But, fear not! The young prince stands strong at his father's bedside, ready to lead the kingdom to new victories.***

This paper proposes a new architectural model for decision-making in all its guises throughout the enterprise. The new model, Business Integrated Insight (BI^2), emerges directly from re-evaluating decision-making in a 21st century business, and reviewing recent technological advances in data-bases, messaging and social computing. The message is one of technology evolution, rather than revolution—current data warehouse technologies, particularly dedicated implementations, will play a central role in the new order.

We begin with a review of the prevailing business and IT paradigms from which the original data warehouse architecture emerged and evolved in the 1980s and '90s and the problems it now faces. Section two makes the case for a new approach and proposes five new postulates for the future. In section three, we describe the BI^2 architecture, leading to a number of use cases and key considerations for implementation in section four. The final section summarizes the paper's main points.

Contents

The past is a foreign country: they do things differently there

Most of our assumptions have outlived their uselessness

A new architecture for Business Integrated Insight (BI^2)

From Data Warehouse to Business Integrated Insight

Conclusion

Sponsored by:

Teradata Corporation
www.teradata.com

Copyright © 2009, 9sight Consulting, all rights reserved

The past is a foreign country: they do things differently there[1]

Or do they? In business intelligence (BI), a review of the original business and technological drivers of the data warehouse architecture reveals that our current assumptions date back to the earliest days of decision support. Some are demonstrably no longer true, while others are, at best, questionable. So, is it time do something different; do we need to re-architect BI?

The concept of data warehousing emerged in the mid-1980s from the discipline of decision support (DSS), which dates back to the late '60s[2]. At that stage, both business management and technology were at an early stage of evolution. Business decision-makers operated on planning cycles of months and often ignored the fluctuating daily flow of business events. On the technology front, applications were hand-crafted and run in mainframes operating at the limits of their computing power and storage. These factors led to one of the longest-lived postulates in IT—the need to separate operational and informational computing and systems. From its earliest days, DSS envisaged extracting data from the operational applications into a separate system designed for decision makers. And, at the time, that made sense: business users liked it, and the technology of the time could not support decision-making systems that operated directly on operational databases.

> **Postulate 1('70s):**
> *Operational and informational environments should be separated for both business and technical reasons.*

The '70s and '80s saw the emergence of separate, often stand-alone, systems where data was extracted from operational systems and used for analysis and reporting. Relational databases (RDBMS), which were first commercialized then, were adopted as the primary platform for delivering DSS for a mix of practical and commercial reasons[3]. Some, such as IBM's DB2 (1983), evolved into general purpose databases supporting both sets of needs. Others, such as Teradata's DBC/1012 (1984), were optimized for decision support and became pioneers in the nascent field of data warehousing.

Enter the data warehouse

By 1986, the term "data warehouse" emerged in the industry, and Paul Murphy and I published the first formal architecture in 1988[4], shown in figure 1a. The architecture addressed a key issue in DSS: the multiplicity of data sets being created in large organizations, leading to inconsistent business decisions and operational problems for IT. We proposed a "*Business Data Warehouse… the single logical storehouse of all the information used to report on the business… In relational terms, a view / number of views that… may have been obtained from different tables*". The BDW was largely normalized, and its data reconciled and cleansed through an integrated interface to operational systems. A basic assumption of this architecture was that operational applications could not be trusted. The data they contained was often incomplete, inaccurate, and inconsistent across different applications. As a result, the

Figure 1: Evolution of the Data Warehouse Architecture

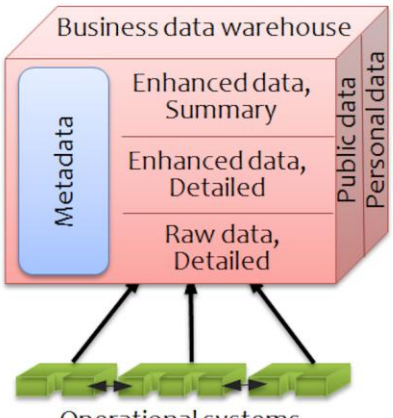

Fig. 1a *Adapted from Devlin & Murphy*[4]

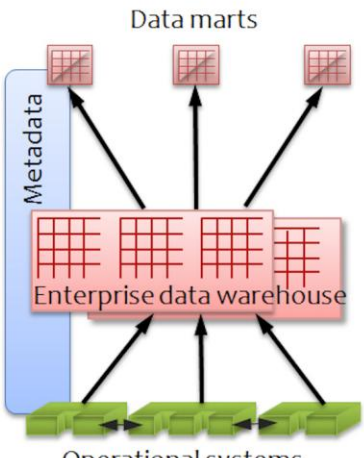

Fig. 1b *Adapted from Devlin*[5]

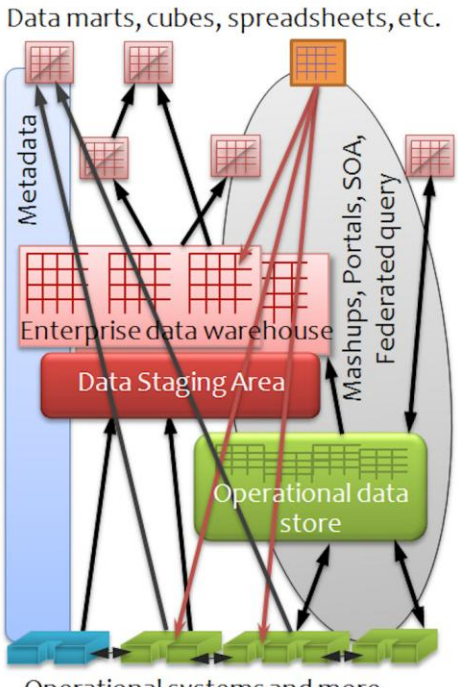

Fig. 1c

Copyright © 2009, 9sight Consulting, all rights reserved

Postulate 2 ('80s):
A data warehouse is the only way to obtain a dependable, integrated view of the business.

Postulate 3 ('80s):
The data warehouse is the only possible instantiation of the full enterprise data model.

data warehouse was the only place where a complete, accurate and consistent view of the business could be obtained. These two assumptions, probably well-founded in the 1980s, have solidified into immutable beliefs in the IT industry.

Another cornerstone of the data warehouse was that data is useless without a framework that describes what the data means, how it is derived and used, who is responsible for it, and so on. Thus arose the concept of metadata and one of its key manifestations—the enterprise data model. While this part of the architecture has proven to be somewhat problematic in practice, its necessity is widely agreed. A key tenet was that the enterprise data model should be physically instantiated as fully as possible in the data warehouse to establish agreed definitions for all information. It was also accepted that the operational environment is too restricted by performance limitations and too volatile to business change to allow instantiation of the enterprise data model there. The data models of operational applications were fragmented, incomplete, and disjoint so the data warehouse became the only reliable source of facts.

The data mart wars of the early 1990s

While envisaging a single logical storehouse accessed through relational views is straightforward, its physical implementation is another matter. Database, extract/transform/load (ETL) performance, data administration, data distribution, project size and other issues quickly arose. The full story of the data mart wars is simply too gory for this white paper. However, their outcome was that most data warehouse implementations became increasingly complex over the decade.

Postulate 4 ('90s):
A layered data warehouse is necessary for speedy and reliable query performance.

A data mart is often defined as a physically instantiated subset of a data warehouse, optimized for the needs of a particular department or set of users. However, in the'90s, data marts were promoted as independent DSS environments on a variety of technologies, sourced directly from the operational applications. Their attraction was largely based on the lengthy time and high cost of building an enterprise data warehouse (EDW). Architecturally, of course, this approach was a step backwards from the architecture. Such independent data marts are now considered, at best, to be complementary to an EDW strategy or, more negatively, as unnecessary political concessions that drain the IT budget.

During the '90s, the majority of data warehouses, limited by the performance of general purpose databases, moved to the hybrid or layered model[5], depicted in figure 1b, with dependent data marts sourced from an EDW and treated as an integral part of the warehouse environment. A few vendors, notably Teradata aided by their purpose-built parallel engine, achieved the goal of a single, integrated physical implementation of the original data warehouse model. Many vendors and consultants still promote a layered approach to ensure end-user query performance, or as a way to isolate data in a mart and/or shorten the development project timeline.

The vile virtual DW, the odious ODS and other abominations[6]

The simple and elegant layered architecture shown in figure 1b has proven its worth to vendors and implementers of data warehouses over the past two decades. However, additional business needs, technological advances and even marketing initiatives have added additional, and sometimes poorly characterized, components to the mix.

An early example of this trend can be seen in IBM's Information Warehouse Framework announcement in 1991, which included EDA/SQL[7] from Information Builders Inc. (IBI) to allow direct query access to operational data sources. More recently, mash-up technology has been promoted for similar "virtual data warehouse" approaches, extending access further to operational web-based and unstructured data sources. Such virtual or federated access eliminates the need to create additional copies of data and provides access to real-time information. However, it raises issues around consistency of results and performance impact on source data systems.

Copyright © 2009, 9sight Consulting, all rights reserved

In the mid-1990s, the operational data store (ODS) was introduced as an architectural concept[8] integrating operational data in a subject-oriented, volatile data store modeled along the lines of the EDW. While first positioned as part of the operational environment, it became an integral part of many data management architectures to support near real-time and non-historical reporting. Although it still appears regularly in architectural designs, the term ODS has been appropriated for so many different purposes that its original meaning is often lost. Regardless, the ODS supports a vital near real-time data warehouse workload, albeit implemented more like an independent data mart.

The architectural confusion doesn't end there. Local and personal data stores, the most problematic of which are spreadsheets (or "spreadmarts[9]") populate the edges of the warehouse, partially fed from the EDW but receiving other uncertified data from disparate sources. Stand-alone data mart appliances feed directly from the operational environment. The inclusion into the warehouse of unstructured information, especially Web-sourced data, has also stretched the original definition of the data warehouse to breaking point.

Well, here's another fine[10] mess you've gotten me into

All this harks back to the earliest days of decision support, where everyone made specialized copies of any data they needed irrespective of the resulting chaos. The resulting data warehouse "architecture" today, depicted in figure 1c, has lost its original simplicity and provides implementers with little guidance on how to structure DSS in a modern business. As the layers and silos increase, many problems become more pressing. Data duplication leads to ever growing levels of inconsistency, which has to be manually reconciled in the reporting process, reducing users' confidence in the data warehouse. Hardware, software and labor costs grow and maintenance becomes ever more complex, constraining the provision of new function and information to meet business demands. Customer and partner interactions suffer because of siloed and inconsistent information. And despite demands for more timely information, the added layers actually increase delays in getting data to the users.

However, the combination of recent business and technological changes makes it critical to address these issues urgently. From a business viewpoint, increased competition and higher customer expectations are driving demands that both structured and unstructured information from all sources is integrated and internally consistent across the organization and delivered at ever increasing speed. On the technology front, Service Oriented Architecture (SOA) approaches are dramatically changing the data and process structures of the operational environment, while Internet technologies are redefining how users expect to interact with all applications. These changes press in upon the analytic environment from above and below, challenging the fundamental assumptions upon which data warehousing was originally defined. Reports may exist forever, but they are less and less the *modus operandi* for delivering insights. Mashups, portals, and Blackberries demand insights too.

Most of our assumptions have outlived their uselessness[11]

The four postulates introduced earlier shaped the evolution of the data warehouse architecture. They are seldom mentioned, but they still underpin, to a greater or lesser extent, all thinking about how decision-making is supported to this day. Unfortunately, they are no longer true.

Postulate 1 established a fundamental boundary between the operational and informational worlds. Today, an increasing number of business processes span the operational to the informational worlds. Such behavior is most obviously seen in operational BI activities, where users switch back and forth between decision-making and action-taking within a single workflow. Even in more traditional or tactical BI, decisions need to be translated into action with ever greater alacrity. From a technical viewpoint, many of the old operational limitations have eased considerably. In the past, IT had to restrict data extraction to overnight batch windows to safeguard online performance and to take account of end-of-day reconciliation processes; today data is extracted 24 hours a day.

Similarly, postulate 2 is becoming obsolete. In many companies, operational applications have migrated from a plethora of in-house developments to a minimal standardized set of commercial-off-the-shelf offerings supplemented by specialized applications. Such offerings as SAP, Oracle Applica-

Copyright © 2009, 9sight Consulting, all rights reserved

tions and others provide highly integrated operational data. Furthermore, SOA approaches depend implicitly on consistency and integration in the operational environment, both within and across company boundaries. While often depending on integrated data stores such as Master Data Management today, in the longer term, SOA requires and will drive greater levels of real-time, message-based integration between operational systems as well as the data warehouse environment.

Such advances are made possible by pervasive enterprise data models, which underpin all moves towards an integrated environment. As a result, postulate 3 tying enterprise model instantiation solely to the data warehouse is also being invalidated. This is good news for the warehouse! Much of the processing that takes place in feeding the warehouse through extract-transform-load (ETL) tools derives from mismatches between the source and target data models. A more integrated, enterprise-scope model in the entire environment simplifies transformation and reduces ETL processing.

Postulate 4 derived from the performance limitations of vendor products which drove the proliferation of data marts. From a business point of view, the layered data structure is under increasing pressure from the needs for lower data latency and increased integration. From a technology viewpoint, ever more powerful hardware, specialized databases and other tools present the opportunity to reduce or eliminate layered architectures.

Resetting assumptions

Postulate 1(2009):
Modern business processes seamlessly combine action-taking and decision-making, and require an integrated continuum of consistent information.

Having parted company with our original postulates, it's clear we need a new set of foundational principles that support modern decision-making. Foremost, we know that today's business demands close integration of operational and informational support. While some personnel with purely operational responsibilities will remain, the range of processes that combine operational and informational tasks and the number of business people who use them continue to grow. A recent example, which would have seemed ridiculous even a few years ago, is a data warehouse ingesting website clicks in real time so a call center agent can see a customer's activity today and over the last year for someone who has just now called. Such seamless processes absolutely require a consistent and integrated information base spanning the operational and informational environments.

A second widespread trend is that all information types are vital inputs to many decision-making tasks. This has enormous implications for information management and storage. Any modern architecture must explicitly include:

- Real-time data from transactional databases and message queues
- Structured data from less trusted sources, such as personal storage and external companies
- Unstructured information of all types, such as text, image, multimedia and more
- Unstructured information from collaborative tools such as e-mail, instant messaging and beyond
- Structured data and unstructured information from the Internet

Postulate 2(2009):
The new information architecture must be based on a comprehensive enterprise information model, spanning all types of information used in the business.

The characteristics of such information differ significantly from 1990s data warehousing. Nonetheless, this information must be part of an expanded enterprise-wide information model that both provides high level consistency across all information types and allows multiple definitions and subsets of enterprise information to support diverse and perhaps contradictory uses. Furthermore, the model and extended metadata that describes the information must be explicitly considered as part of the same information resource in order to ensure deep integration and maximum flexibility.

The volumes of information involved here are far in excess of those previously encountered in BI. From the earliest days of decision support, when summary reporting was sufficient, the data warehouse has expanded to include most all structured operational data. With the inclusion of unstruc-

Copyright © 2009, 9sight Consulting, all rights reserved

tured information, both external and internal, as well as external tured data, the volume of information needed by the business continues to explode. Most sources put the ratio of unstructured to structured information volumes in the order of 5:1 to 10:1. Add web-clicks, GPS, RFID, sensor networks and more, and every company will see another wave of data expansion.

Postulate 3 (2009):
The business information resource is best maintained as a single copy of each data item, with only minimal resort to transient layers or copies of specific subsets of data for specialized needs.

Beyond the exploding volumes of information, we have the need for ever increasing timeliness of information in decision-making tasks. Maintaining increasingly large volumes of rapidly changing information poses a substantial challenge for IT. Storage and processing costs may creep up but more worrying are the rapidly growing administrative and management costs of ensuring integrity and consistency when data is duplicated in many places. Under these circumstances, the old approach of making multiple copies of data for different purposes becomes increasingly untenable. While having only a single copy of every piece of data may be the vision, the aim in reality must be to minimize the number of copies of data, especially those data stores with the largest volumes and the most rapid rates of change.

Today's business also demands increasingly faster and more flexible reaction to market changes. This need is seen in sometimes challenging requests for near real-time data. More taxing by far is the concurrent demand for highly integrated yet completely flexile processes. In the 1980s' data warehouse architecture, the only process considered was batch population. A new architecture, on the other hand, must support an integrated process approach based on an enterprise model that defines all the processes the business needs.

Postulate 4 (2009):
An integrated, model-based and closed-loop process environment is needed to create, maintain and use both the business information and activities.

Such processes extend from creation and maintenance of business transactions to real-time population of, and user access to and use of information for analysis and decision-making. They further include activities in the IT domain that create and manage the information and function used by business users, because today's demands for flexibility extend to the actual application and process structure and components. An example of this can be seen in the insurance industry, where novel products that require both new data and process innovations must be introduced at short notice.

In this environment, closed-loop processes that explicitly link business changes and events to actions taken and back to evaluation of the business impact of these actions are required. The sense-and-respond model in Steve Haeckel's "Adaptive Enterprise"[12] is a particularly useful description that helps explore how BI can adapt to SOA[13].

A new architecture must also explicitly consider the expectations of today's business users. In the 1980s' data warehouse, usage was confined to a small, data-savvy subset of the population who ran reports and analyses. While such users continue to be important, BI tasks have now expanded to a much wider audience. These users are typically front-line workers rather than data-focused, and are linked directly to operational processes and deadline-driven. These employees use decision-making as discrete steps in a larger workflow. They also use collaborative tools, from office automation and e-mail to social networking and instant collaboration, often as part of formal business processes.

Postulate 5 (2009):
An integrated, flexible and role-based user interface provides access to the entire business information.

Business user interaction also demands a holistic approach irrespective of what type of task or information is involved. Sometimes called pervasive BI, there is an increasing trend to insert analytic information in workflows, transactions, and web sites—i.e. nearly every operational process. Conversely, operational and analytic processes are increasingly embedded in collaborative systems such as e-mail or messaging-based workflows where unstructured information dominates. The outcome is a convergence for all computing activities on a common, role-based user interface that supports all information processing needs. Furthermore, this interface must be flexible enough to allow tasks and workflows to be constructed and reconstructed on the fly with minimal technical expertise and based firmly on metadata from the enterprise-wide information and process models.

Copyright © 2009, 9sight Consulting, all rights reserved

A broadly-based review of current business needs and technological possibilities thus leads us in 2009 to a new set of postulates that more accurately reflect the reality of the early 21st century. These new postulates are much broader in scope than those that drove the original data warehouse architecture. In a number of cases, they diametrically oppose those earlier assumptions.

A new architecture for Business Integrated Insight (BI^2)

No architecture is so haughty as that which is simple[14].

The radically new set of postulates introduced above leads inevitably to the need for a new architecture. This architecture should be simple, especially in the light of the complexity that has crept into the old model as seen in figure 1c. And yet, the breadth required of the new architecture is breath-taking. It must literally cover the entire IT support for the business. It must be fully integrated. And it must stretch beyond "simple" intelligence. Hence, "*Business Integrated Insight*".

Figure 2 illustrates this architecture at a conceptual level. Unlike the 1980s architecture, the BI^2 architecture places all information—operational, informational and collaborative information, as well as metadata—in a single layer called the ***Business Information Resource (BIR)***. This is the foundational layer of the architecture: without it, the two upper layers cannot exist. If the BIR is incomplete or inaccurate, the two upper layers are seriously degraded from an ideal state.

The middle layer consists of processes, workflows, tasks, applications, tools and so on: in short, all functionality and processing that runs on the business' computers. This ***Business Function Assembly (BFA)*** depends on the underlying BIR layer for all information, from creation to ongoing usage.

Finally, the top layer, the ***Personal Action Domain (PAD)*** represents the intent and behaviors of all users of the system, from executives to front line staff to IT, within the company or external to it. The PAD is beyond the physical computing resources of the company; it is actually the people who run and interact with the business. The IT-based information and processes in the lower two layers are meaningless without this top layer. Understanding the scope of needs, goals and behaviors of users in this domain is necessary to design the lower two layers.

Before delving deeper, there is an obvious question to answer: what is the value of such a simplified, conceptual view? Business today is so interlinked, its information needs so interdependent, its processes so entwined and its reaction times so pressured that only a new architected approach can support it going forward. Business and IT, in particular, need to take a fully integrated view of what is required from its IT environment. This will entail novel forms of cooperation across the business and IT. There will be new IT roles and responsibilities and a significant level of empowerment for business

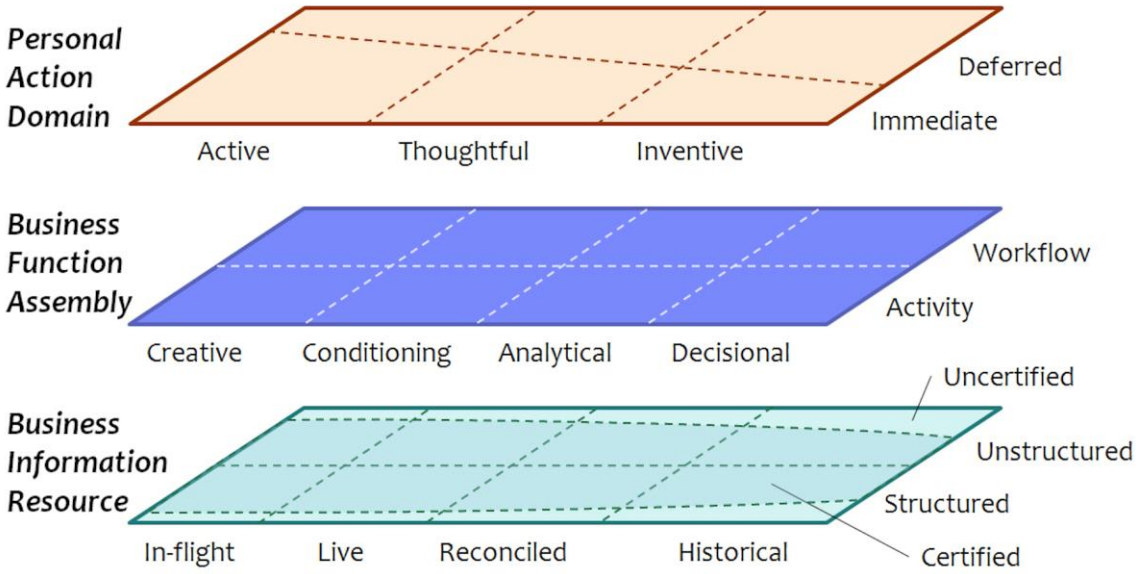

Figure 2: The Business Integrated Insight (BI^2) Architecture

Copyright © 2009, 9sight Consulting, all rights reserved

users. Such changes are facilitated by the common vision that this architecture provides. The alternative is to multiply the existing chaos that has formed in the absence of a modern planned architecture—high costs, complexity, and congestion.

Business Information Resource—the Data

Within the BIR, information is defined along three principal axes: ***timeliness***, ***structuredness*** and ***reliability***. The placement of any ***information nugget*** on these three axes determines to a large extent the level of management it requires, what processes can interact with it and how people can use it.

Metadata is also a part of the BIR—it is clearly information of business value. Traditionally, metadata was placed in a separate repository, leading to debates on where a particular field (e.g., a timestamp) should be placed. Furthermore, responsibility for managing metadata could be assigned to "somebody else". Bringing metadata into the BIR avoids these problems and further allows full consideration of metadata along the above three axes. The following sections describe these axes in terms of classical business information; however, the same principles apply to metadata.

The information nugget

An information nugget is the smallest set of related information that is of value to a business user in a particular context. It is the information equivalent of an SOA service, which is also defined in terms of the smallest piece of business function *from a user viewpoint*. An information nugget can thus be as small as a single record when dealing with individual transactions, or as large as a complete dataset describing the business status at some time. As with SOA services, information nuggets may be composed of smaller nuggets and as well as being part of many larger nuggets. They are thus granular, reusable, modular, composable and interoperable and may span traditional information types.

As modeled, an information nugget exists only once in the BIR, although it may be widely dispersed along the three axes. At a physical level, it ideally maps to a single data instantiation, although the usual technology performance and access constraints may require some duplication. However, the purpose of this new modeling concept is to minimize the level of physical data redundancy, while ensuring that information, as seen by business users, is uniquely and directly related to its use.

Timeliness

The timeliness axis defines the time period over which an information nugget exists and is considered to be valid. From left to right, timeliness moves from information that is essentially ephemeral to eternal. ***In-flight*** information consists of messages on the wire or the enterprise service bus; it is valid only at the instant it passes by. This data-in-motion might be processed, used, and discarded. However, in-flight information is normally recorded somewhere, at which stage it becomes live. ***Live*** information has a limited period of validity and is subject to continuous change. It also is not necessarily completely consistent with other live information. This is the characteristic of ***reconciled*** information, which is consistent and stable over the medium term. ***Historical*** information is the final category, where the period of validity and consistency is, in principle, forever. But, like real-world history, it contains much personal opinion and may be rewritten by the victors in any power struggle!

The timeliness axis broadly mirrors the lifecycle of information from creation through use and either disposal or archival. Some call this spectrum hot, warm, and cold data. In the diagram, dashed lines delineate boundaries between the different classes of timeliness described above. However, the reality is that these classes gradually merge from one into the next. It is therefore vital to apply critical judgment when deciding which technology is appropriate for any particular base information nugget. Note that all the dashed boundary lines shown in figure 2 are similarly vague in practice.

Structuredness

The structuredness axis reflects the ease with which meaning can be discerned in information. This characteristic is often described in a binary manner: ***structured*** vs. ***unstructured***. In some usage, data is described as structured, and information unstructured. In reality, there exists a continuum of structuredness that applies to all information. The highest degree of structuredness is found in nu-

Copyright © 2009, 9sight Consulting, all rights reserved

merical or codified information stored in labeled fields—classical business data. When textual information is stored in named fields (in XML, for example), the degree of structuring decreases because there may be additional meaning in the stored text beyond that explicitly defined by the field's metadata. Moving to e-mail and then more general documentation, increasing amounts of information exist inside the structured fields, and the information as a whole becomes increasingly unstructured. At the extreme, multimedia information is highly unstructured.

Placing information on this axis has become increasingly important in modern business as more and more unstructured information is used. It is widely asserted that unstructured data makes up 80% or more of all stored data. It makes sense therefore that much useful information can be found there, for example, by text mining tools. Just as we have transformed and moved information along the timeliness axis in data warehousing, we now face decisions about whether and how to transform and move data along the structuredness axis, usually from unstructured to structured.

Reliability

The final axis, reliability, has been largely ignored in traditional data warehousing, which confines itself to centrally managed and "dependable" data. However, the widespread use of personal data such as spreadsheets has always been problematic for data management. Similarly, data increasingly arrives from external sources, from trusted business partners all the way to the "world wild west" of the Internet. All of this unmanaged and undependable information plays an increasingly important role in running a business. It is thus becoming clear that centrally managed and certified information is only a fraction of the information resource of any business.

The reliability axis, therefore, classifies information according to how much faith can be placed in it. Highly certified information is strongly managed, often at an enterprise level. It adheres closely to the enterprise information model, is highly consistent and may be subject to audit. By definition, reconciled, and to a slightly lesser extent, historical information is highly reliable. Reliability of information also varies depending on its source. Internal operational systems, with their long history of management and auditability, are usually considered sources of very reliable data. Information produced and managed by a single individual, on the other hand, often has low reliability. A collaborative effort by a group of individuals produces information of higher reliability. Information from the Internet is highly unreliable and requires validation and verification before use. And information from other external sources, such as business partners, has varying levels of reliability. The placement of information on this axis, and the definition of rules and methods for handling different levels of reliability are topics that are still in their infancy, but they will become increasingly important as the volumes and value of less closely managed and controlled data grows.

Business Function Assembly[15]— Processing

The process layer of the BI^2 architecture—the Business Function Assembly (BFA)—can be characterized along two axes: ***effect*** and ***scope***. The BFA conforms to SOA principles. Functionality is structured as well-defined, callable and largely immutable services that perform meaningful ***activities*** at a business level. Services may be built upon lower level services recursively, and linked together into transactions or ***workflows.*** They can be inserted, replaced or removed with minimal technical expertise from a workflow in a plug-and-play manner. The scope axis reflects this characteristic structure. Activities and workflows, and their actions and interfaces, are all described by metadata in the BIR.

The effect axis describes how a particular function affects the business, and by extension, the BIR. ***Creative*** function produces new business entities or instances; for example, create a new purchase order or create an entirely new type of order. Such function operates on the left-most side of the BIR. ***Conditioning*** function modifies existing entities or instances; for example, update an order or archive an old order. The effect of conditioning extends over the entire BIR, as it is the mechanism by which information nuggets are transformed and moved along the various axes of the BIR. ***Analytical*** function uses existing information to monitor and understand what is occurring in and outside the business, forming hypotheses about causes and effects. This function does not modify existing information, but can create new information based solely on existing information. ***Decisional*** function applies

Copyright © 2009, 9sight Consulting, all rights reserved

insight to analytical hypotheses and existing information to take action. Such actions link back directly to invoke creative, conditioning or analytical functions, thus closing the sense and respond loop.

The BFA includes both business and IT processes, as they are currently called. Creating a data instance (a business process) and creating a data entity or replacing a service in a workflow (IT processes) are treated equivalently. This approach is necessitated by the expanding role of business users / developers who are today are more likely to create a spreadsheet or mashup a couple of web pages than to approach IT for a new application or database.

Personal Action Domain—People

The Personal Action Layer is both the most important and at the same time most unexpected layer in the BI^2 architecture. Most important because it characterizes how all the various users of the system behave and what they expect as a result of any action. Most unexpected because it actually doesn't represent anything that can be physically instantiated in a computer environment!

The PAD represents a very simplified model of human behavior along two axes: ***intent*** and ***gratification time***. Modeling what a particular user intends in a specific moment is a vital step to understanding what type of business function is required in the BFA and what is needed in the BIR to satisfy that intent. ***Active*** intent initiates an action, causing something to happen. It initiates a creative or conditioning function that creates or changes information mainly in the in-flight or live classes. ***Thoughtful*** intent gathers information of any appropriate class for use by analytical function. ***Inventive*** intent uses decisional functionality to innovate and recreate some aspect of the business cycle, linking back to active intent and thus close the sense and respond loop.

The gratification time axis simply reflects the normal reality of the world that some things you can have now, while others you have to wait for. In the modern business, there is a growing desire and need for ***immediate*** gratification, which places added emphasis on access to in-flight and live data. However, such data may not be internally consistent, as previously described, implying that ***deferred*** gratification may have to be endured.

From Data Warehouse to Business Integrated Insight

A journey of a thousand miles begins with a single step[16].

"Nice concept," I hear you say, "but how do I get there?" Unlike the old Irish joke, however, the answer cannot be: "If I wanted to get there, I wouldn't start from here!" This section first describes some use cases and then focuses on the Business Information Resource, this being the foundational layer of the BI^2 architecture. Understanding the current business data and information environment in the context of the three axes and rationalizing existing data stores according to this model are the first steps towards BI^2. More importantly, these are also the first steps towards a more manageable, cost-effective and future-proof information resource.

BI^2—three diverse use cases

The business analyst

It is widely assumed that business analysts' information needs can be supported through a single data mart accessed through one particular BI tool. The falsity of this assumption can be seen in the widespread phenomenon of spreadmarts[9] mentioned earlier. Mapping the business analyst use case (A) on to the BI^2 architecture, shown in red in figure 3, illustrates how the architecture works and provides an insight into the diverse information needs of the analyst. In the PAD, the analyst's role is classed as thoughtful and spans immediate and deferred gratification. In the BFA, the function required is analytical and consists of stand-alone activities.

The most interesting aspect is in the BIR. In this case, we see information is structured and historical, but split between certified and uncertified. This immediately poses the very valid implementation question: how can these diverse data types be made available most effectively to the user? Today,

Copyright © 2009, 9sight Consulting, all rights reserved

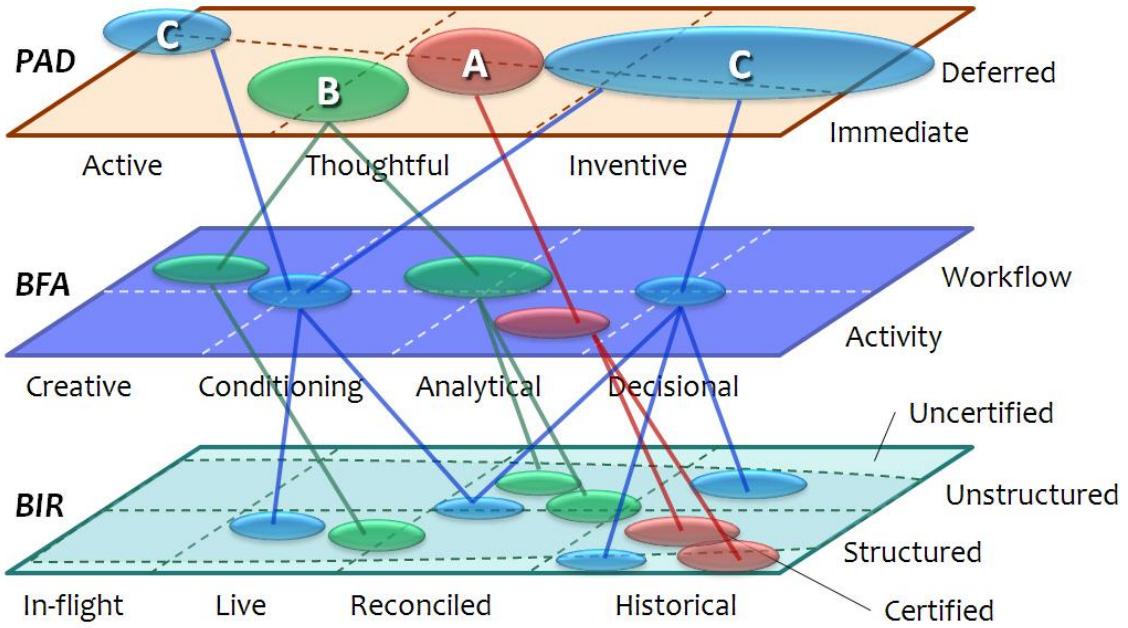

Figure 3: Business Integrated Insight (BI^2) Use Cases

the analyst probably copies data manually from the certified domain into uncertified personal spreadsheets. A number of more effective strategies could be considered, from federated access to a managed "playmart[17]" environment. We can also envisage analyst scenarios needing live information, structured information, etc. that can also be mapped onto and implemented with this model.

The call center

A basic call center use case (B) is shown in green in figure 3. The agent's role involves both active (create or cancel a subscription) and thoughtful (what could we offer to keep the customer?) behaviors in the PAD. Almost all of the agent's activities are constrained by a predefined process, so the function required resides in the workflow segment of the BFA. Two quite distinct types of function are required—the creative function that generates new information and the conditioning/analytical function that generates new insights for the agent about what to do to retain the customer. In the BIR, the creative function generates structured data, while the conditioning/analytical function requires a mix of structured and unstructured information. Recall that the three separate information segments in the BIR comprise a single information nugget for this user process and must thus be modeled together.

Marketing and sales support

The marketing and sales support use case (C), shown in blue in figure 3, is the most complex case we mention here, having behavior, function and information widely spread over many axes of the three layers. The large oval in the thoughtful/inventive segment of the PAD represents the insightful market analysis work that precedes and continues throughout a sales campaign. Actions that change the campaign in-flight are reflected by the smaller blue oval on the left. In fact, the two ovals are linked (indicated by their extension beyond the bounds of the layer) insofar as invention must lead to action, so that value can be realized.

Given the scope and complexity of the user behavior in this use case, the function and information mappings in the BFA and BIR become correspondingly broad and diverse. In this case, we show only a subset of the types of function and information that are required. Readers are invited to explore these mappings in more depth themselves.

Mapping current data stores onto the Business Information Resource

Figure 4 maps a number of today's common data stores onto the BIR. Comparing this with figure 3, we can see that discrete user processes (even the simplest ones) usually require data from multiple

Copyright © 2009, 9sight Consulting, all rights reserved

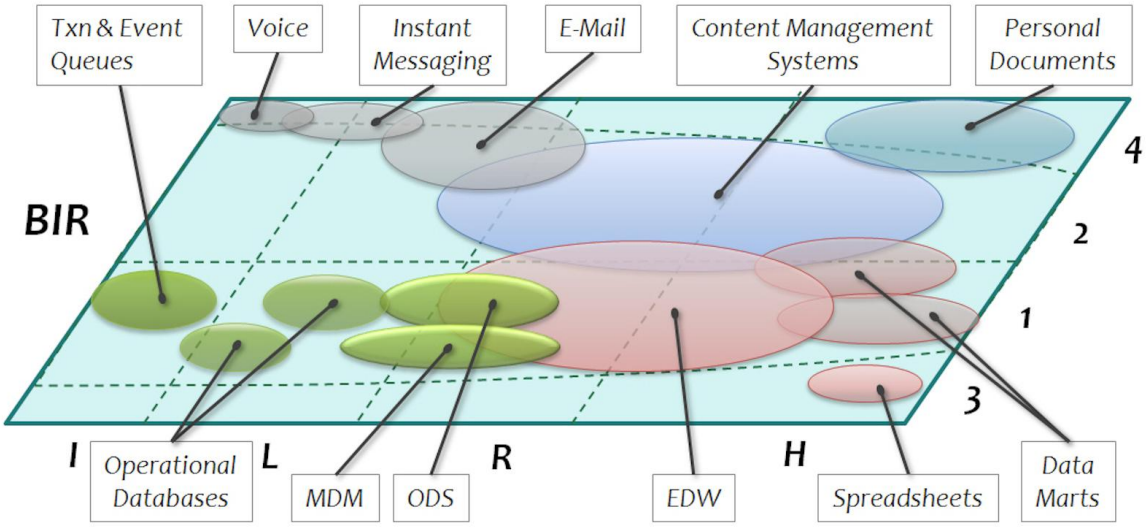

Figure 3: Mapping the BI² Architecture to current systems

stores. Given that these stores may be competing and inconsistent with one another, the most immediate and obvious conclusion is that today we have created an information management nightmare! However, the historical role of the warehouse is a point of consolidation for enterprise data, and its implementers have experience of in tackling such issues, so it makes sense to begin there.

Data warehouses and data marts

Today, after many years of scope creep on both sides, data warehouses and data marts play overlapping and often conflicting roles in the areas of reconciled/historical, structured and certified information (grid references R1 and H1) of the BIR. Companies have built multiple overlapping data warehouses and marts serving communities often with only slightly differing business requirements. This approach is architecturally unsound. And while it may be politically easier and financially cheaper in the development phase, the longer term maintenance and management costs are untenable. Addressing this issue now would provide significant immediate and long-term savings in running costs.

Applying the BI² concepts outlined earlier leads to an immediate conclusion—data marts, both dependent and independent, should be rolled into a single, comprehensive data warehouse component. As hardware has become more powerful and database systems better tuned, the technical case for data marts has weakened considerably. The remaining political and organizational issues should now be put aside to simplify and streamline this portion of the Business Information Resource.

Operational Data Stores

The boundary between live and reconciled information in the structured, certified sectors (L1 and R1) of the BIR has always been highly fluid. As a result, multiple solutions have emerged from the operational and informational worlds, of which master data management (MDM) and operational data stores (ODS) are but two examples. Confusion and multiple solutions in this area can be very costly, because such data requires near real-time management and inconsistencies can feed into instant errors in dealing with customers or suppliers.

To propose an optimal approach, we need to understand the differing business needs that can help define the boundaries of the relevant information nuggets. While this analysis varies widely by business, we see two general classes of consolidating in-flight or live data from independent sources to:

1. **Create a fully consistent transaction base for the business operations.** An example is the creation of a reconciled product catalog so that sales in different regions are recorded consistently.
2. **Create a fully consistent information base for decision-making.** An example is the creation of one consumer behavior profile to be fed to sales, call centers, and web sites for cross selling.

In the above examples, we can identify two distinct high-level information nuggets that, on deeper analysis, will be seen to share lower-level nuggets. For the non-shared nuggets, responsibility can

Copyright © 2009, 9sight Consulting, all rights reserved

clearly be assigned to either the data warehouse environment or the operational environment. In the case of shared nuggets, the general principle is to follow the strongest need. If timeliness is most important, the operational environment would take responsibility; if consistency is more important, the data warehouse would take charge.

The above analysis allows minimization of data duplication across the L1-R1 boundary, which is vital for reducing reconciliation, certification, data administration and management costs. However, having assigned the shared information nuggets to one or other side of the boundary, access from the other side is, by definition, remote. For example, let's assume (as is likely) that responsibility for maintaining a consistent set of purchase order transactions is assigned to operational applications, while a consistent view of customers is the responsibility of the data warehouse. In this case, access to the order transactions from the warehouse requires a remote (or federated) data access while the same applies to access to the customer information from the operational environment.

With increasing network bandwidths and speeds, as well as advances in SOA and database query optimization, remote or federated query and/or access has become more necessary, and especially with portals and mashups, more acceptable and reliable in recent years. At present, the programmatic approach to federation is favored by many vendors and implementers. However, as is the case for joining tables within a single database, SQL queries offer similar data design and management advantages for dispersed databases. However, it is almost mandatory that federated queries be predefined, as the dangers of *ad hoc* usage are significantly greater than in the case of localized databases.

The only other approach to reducing duplication is to consolidate *all* structured data into a single database, an option that is simply not realistic. And beyond this is unstructured information…

Unstructured information

Recent years have seen an explosion in interest in obtaining operational and especially analytical value from the huge stores of unstructured information, particularly in sectors R2 and H2 of the BIR. Given the very different processing approaches to structured and unstructured information, vendors typically propose solutions that favor their lineage on one side of the divide or the other. Relational database vendors favor solutions that copy unstructured information into the warehouse and use SQL to access it. Search vendors tend to prefer copying in the opposite direction. Neither approach aligns with the BI^2 architecture—they both potentially create duplication of information nuggets. Furthermore, both approaches lead to the possibility of duplicating potentially enormous volumes of information, with attendant storage and management costs.

Applying BI2 principles leads to a more practical hybrid solution. Recognize first that the real difference between structured and unstructured data lies in the explicit metadata that is created to describe structured information. For example, an order can be an unstructured sentence—"Jane Doe ordered 3 table lamps at $40 each last Tuesday"—or a record with well-defined (within the metadata) fields and contents—Customer = "Jane Doe", Product = "Table Lamp", etc. Value, either operational or informational, is obtained from unstructured information by extracting identifiable facts and by making explicit the metadata that is implicit in the information. In the example above, the word "ordered" implies both the identity of the customer and the number and type of product.

Text extraction is thus a process that creates semi-structured data and metadata from unstructured information, and is analogous to the reconciliation process that occurs on the structuredness axis. Conceptually, the approach to dealing with unstructured information is the same as that described in the previous section: understand the business needs, identify information nuggets and minimize duplication. Given the volume of unstructured information, it makes sense to leave the "raw" unstructured information where it is under the control of the content management environment. Extracting the semi-structured data and metadata into the data warehouse environment provides access to the value of the information through SQL. And federated access (in this case, often called "mashups") provides a way to use the raw unstructured information as required.

Copyright © 2009, 9sight Consulting, all rights reserved

Uncertified information—personal and external

Uncertified information—rows 3 and 4 in figure 4—has long posed challenges for data warehousing as centralized control confronts autonomous behavior. BI^2 can point to solutions here too.

By its very nature, very little of any certainty can be said about uncertified information. Its provenance may be unknown, its meaning uncertain and its quality suspect. Unlike other information, it is likely to be largely unmodelled and widely duplicated. As a result, it is only when it reaches a certain acceptable level of certification that such information can be treated as a fully functioning component of the BIR. Uncertified information is included in the BIR not because it meets the architectural demands of the BIR, but because it is of growing business value and has the potential to contribute to the certified information therein.

Put simply, external information must be subject to a formal validation and importation process before it is made broadly available in the BIR. Particular care is needed in the case of external information that enters under the guise of personal information, gleaned from the Internet by business users and carried in together with information these users have derived from previously certified information in the BIR. *Caveat emptor!*

In the past, the promotion of personal information to the certified level, in the data warehouse, for example, was either expressly forbidden or allowed only through very centralized and formal process. With the acceptance of the value of collaborative working and tools that manage that process, we can see the emergence of an Adaptive Information Cycle described in detail elsewhere[18]. The key to this cycle, however, is the explicit inclusion of the business users' community in the process and the recognition that the collective knowledge and wisdom of these groups play a key role in certifying information for wider business usage.

Conclusions

In this paper we've journeyed through the end of one era into the beginning of the next. The 1980s data warehouse architecture has stood us in good stead, but the time has come to look at decision-making—as it spans the entire business process spectrum—in a new light. We have seen the demands for integrated behavior and near instantaneous access to all information across all business processes that characterize modern business. The ability of IT to meet these demands will make the difference between success and survival for every business. We thus require a new architecture that explicitly covers all aspects of IT support for the business as a whole.

Business Integrated Insight (BI^2) meets these demands and proposes a comprehensive, but conceptually simple, architecture through which a new vision can be created. Despite a radically different approach to architectural layering and a very broad scope, BI^2 nonetheless provides a direct evolutionary path from current data warehouse implementations.

Rationalizing existing data warehouse implementations is a key first step towards this new, comprehensive architecture. Rarely can the CIO actually do more with less. BI^2 not only prescribes this but illustrates the principles of how to get it done. This simplification also makes perfect sense from the viewpoint of reducing current ongoing costs of the enterprise-wide data infrastructure and diverting investment into initiatives that will enhance information usage and drive better decision-making throughout the organization.

Your journey has just begun. *Bon voyage!*

Copyright © 2009, 9sight Consulting, all rights reserved

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing. He is a widely respected consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation". Barry's current interest extends to a fully integrated business, covering informational, operational and collaborative environments to offer an holistic experience of the business through IT. He is founder and principal of 9sight Consulting, specializing in the human, organizational and IT implications and design of deep business insight solutions.

About Teradata Corporation

Teradata (NYSE: TDC) is the world's largest company solely focused on raising intelligence through data warehousing and business analytics. It's our passion and it's all we do. We deliver award-winning, integrated, purpose built platforms based on the most powerful, scalable, and reliable technology platform in the industry. Our assets include:

- Approximately 6,000 associates in more than 60 countries
- Strong diversified client base of over 900 customers worldwide and companies of all sizes
- 2,000+ implementations worldwide

Every day Teradata pushes analytical intelligence deeper into operational execution, enhancing efficiency and transforming corporate culture. Teradata… Smarter. Faster. Wins.™

Teradata Corporation
2835 Miami Village Drive
Miaisburg, OH 45342 www.teradata.com

Brand and product names mentioned in this paper may be the trademarks or registered trademarks of their respective owners.

[1] The oft-quoted opening sentence from "*The Go-Between*" (1953), L. P. Hartley

[2] Power, D.J. *A Brief History of Decision Support Systems,* v 4.0, (2007), http://bit.ly/1IZxHF

[3] White, C.J. "*In the beginning: an RDBMS history*", Teradata Magazine, (September 2004), http://bit.ly/ZPNtw

[4] Devlin, B. A. and Murphy, P. T., "*An architecture for a business and information system*", IBM Systems Journal, Volume 27, Number 1, Page 60 (1988) http://bit.ly/EBIS1988

[5] Devlin, B., "*Data warehouse—From Architecture to Implementation*", Addison-Wesley, (1997)

[6] With apologies to friends and colleagues who invented or promoted these concepts ☺

[7] Since 2001, this technology has become part of iWay Software, a subsidiary of IBI.

[8] Inmon, W.H., Imhoff, C. & Battas, G., "*Building the Operational Data Store*", John Wiley & Sons, (1996)

[9] Eckerson, W. W. and Sherman, R. P., "*Strategies for Managing Spreadmarts: Migrating to a Managed BI Environment*", TDWI Best Practice Report, First Quarter 2008, http://bit.ly/ZVx86

[10] From http://bit.ly/kARK1, Ollie's catchphrase in the Laurel and Hardy films was "…another nice mess…"!

[11] Marshall McLuhan

[12] Haeckel, S. H., "*Adaptive Enterprise*", Harvard Business Press, (1999)

[13] Devlin, B., "Business Intelligence is Dead—Long Live the Highly Evolved Business", http://bit.ly/R8EkM

[14] John Ruskin, British art critic and social thinker, 1819-1900 (In the 1800s, "haughty" meant "high" or "lofty" vs. its more modern meaning of "arrogant" or "proud")

[15] The Business Function Assembly and Personal Action Domain layers of the architecture are less fully developed at this time. Influence how this architecture evolves at BeyeConnect: http://bit.ly/yLl7c

[16] Lao Tzu, Chinese Philosopher and founder of Taoism, 600 BC-531 BC

[17] Devlin, B. "Playmarts: Agility with Control, Reconnecting Business Analysts to the data warehouse", (2008), http://bit.ly/lN5UV

[18] Devlin, B. "Collaborative Analytics, Sharing and Harvesting Analytic Insights", (2009), http://bit.ly/otVWP

Copyright © 2009, 9sight Consulting, all rights reserved