Beyond the Data Warehouse:

A Unified Information Store for Data and Content

May 2010

A White Paper by

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

The worlds of data and content are on a collision course! With ever-growing hordes of content gathering in the business and on the Internet, the old civilization of the data warehouse is under siege. But, never fear! A solution is emerging—the outcome will be integration, not annihilation.

Based on over twenty years of information architecture experience from data warehousing, this paper first shows data and content as two ends of a continuum of the same business information asset and explores the depth of integration required for full business value.

We then define a unified information store (UIS) architecture as the approach to unification. The heart of this store is a core set of business information, indexes and metadata, originating from up-front enterprise modeling and text analytics of information when loaded and at the point of use, which ensure both data quality and agility. The business outcome is analytics that combine the precision of data querying with the relevance of content search, independent of the information source and structure.

Software vendors from both viewpoints—data and content—are already delivering products that blend the two worlds. Businesses that begin to implement a unified information store stand to gain early adopter advantage in this rapidly growing market.

Contents

Sponsored by:

ATTIVIO

When worlds collide—Data meets Content

"Data (to Jean-Luc Picard): *'Since I do not require sleep… sir, I am content to stand.'"*
Star Trek, The Next Generation

Star Trek fans may recall that Lieutenant Commander Data's lack of emotions often led to misunderstandings with his human colleagues. In the world of business and IT, data and content have a similar, dysfunctional relationship. Data has long been the darling (and major investment) of IT—"data processing" was an early term for computing—while content was left to fend for itself. Now, thanks to Web 2.0 and the "Google experience", content is finally making the big time. IT is looking to improve management of this critical resource and to extend its search-like user environment. Because business users now expect an integrated view of all relevant information, an immediate challenge is to define and create a ***unified information store***, whether physical or virtual, of all types of information and common methods for retrieving and presenting information from this store.

Content and data are closely related. Data is what IT has made of content in order to control and process it in the structured world of computers. Content as simple as "I'll buy that red car" is transformed into a purchase transaction, with defined fields, allowed value ranges and keys normalized in a database. The use of two distinct words, "data" and "content" is unfortunate, since both are the same concept—information. Content is softer information, while data is harder[1]; two terms at opposite ends of a continuum. At the softer end, information exists as commonly used and interpreted by humans—documents, images, etc. Hard information is the structured records and fields suitable for logical and numerical computer processing, for example, in operational systems.

Conceptually, soft information is the original source of all hard information. In designing operational databases, hard information is defined up-front through a person modeling soft information for computer use. Simply put, modeling separates the meaning (what is an order) and relationships of terms (price and quantity as part of an order) from the values (ten items at $100 each) they may take in a particular instance. Meaning and relationships can also be distilled from soft information on the fly—during ingestion into a content store or even during use of the content—using text mining and analytic tooling that essentially automates the same modeling process.

Perhaps because of its more formal structure, data is often assumed to be more accurate and reliable than content. The aim of a "single version of the truth" is widespread in data warehousing. In reality, both assumptions are misleading. Reliability and accuracy of information depend solely on its source, and the format doesn't affect the quality of the information. Some sources are simply more or less dependable than others. Like accuracy and reliability, truth is also a relative term, as any reading of eye-witness reports can confirm. Resetting these erroneous beliefs is vital, especially for data warehouse experts, as we bring data and content together.

Harder information exists today in the regimented databases of operational systems, data warehouses, and so on. Softer information is found in a wide variety of content stores from the "world wild west" of the Web and social media to well-managed stores of e-mails, documents, call center logs, etc. in enterprises. To meet the demand to provide access to all relevant information—regardless of its source or form—technology leaders need to look for methods that unite information without losing either the relationships so valued in the database realm or the context and nuances so important in content.

The whys and wherefores of reuniting Data and Content

Why reunite data and content? Simply put, because the business neither understands nor accepts the difference—nor ever did! In the past, users accepted hard information, but they were never too pleased. To use it, they had to think and behave more like computers than people. Informa-

[1] In IT parlance, the terms "structured" and "unstructured information" are used. The latter is an oxymoron, because information, by definition, has structure and without it would simply be noise. But "soft" and "hard" have their pitfalls too—there is nothing soft about an email offered as hard legal evidence!

tion was placed rigorously in defined fields, with only certain values allowed. Searching for data required knowledge of where it was stored, the query had to be structured very precisely and the answer (if found) restructured into something meaningful. As often as not, IT had to be involved and usually took a long time to deliver an answer as it struggled with its long backlog. In short, a disjoint data/content environment lacks agility and accessibility.

The maturation of the Web changed users' expectations radically. Search engines like Google deliver answers instantly and often with surprising relevance, offering user experiences that are far more "human" and accessible. Information can be easily stored and shared in Facebook and other social networking systems. But how do they guarantee the integrity of the content? How can predetermined, *a priori* models of established relationships describe this rapidly morphing world?

Despite differing storage structures for hard and soft information, business increasingly needs a combined view where both precise and relevant answers are dictated by the context of the question, not by the source or structure of the information. Call centers, messaging systems from e-mail to Twitter, social networking tools and even compliance practices routinely collect vast quantities of soft information about customer desires, product problems, etc. Interpreting and linking such soft information to the hard data of sales, returns, and more is vital for quick and appropriate reaction to emerging trends. Amazon, for example, is renowned for its recommendation system that combines hard data from purchases and page visits with the softer information in users' reviews to influence buying behaviors. The five-star rating system quantifying reviewer opinion allows Amazon to provide the soft content of reviewer opinion as a valued data point for buyers.

The search for meaning

All business use of information depends on understanding its real meaning in the context of the people and activities involved. Such meaning is implicit; it must be made explicit to be useful in IT systems. However, extracting meaning from the putatively separate classes of data and content has long been a tale of two cities. One set of vendors comes from the hard information space, starting with relational and other database management systems. The second set starts from soft information, with search tools and content/document management systems as their technologies of choice. Until recently, both sides have generally focused on their core markets, sporadically issuing largely unfulfilled promises to try to cross the data / content divide from one side or the other.

Databases, relational and otherwise, emphasize hard information, from storage to querying and processing. However, they also provide a place (CLOBs and BLOBs—character/binary large objects) for free text and other softer information. Most databases provide search and manipulation function, albeit limited, within these fields. A further refinement sees the use of text analytics to create and populate indexes and other metadata in the database itself. The metadata may be stored in relational or XML formats, with the text remaining in its original form. While these enhancements do support content in databases, there are pitfalls: retrieval is largely dependent on pre-defined models and IT-generated SQL queries, reducing openness to changing content and decreasing agility to respond to unpredicted user exploration needs. Database systems were never designed to maintain rich interaction with content that dynamically responds to users' needs.

Predictably enough, vendors from the content end of the spectrum take an index-centric approach to addressing hard data. A typical index used for text search can be expanded to simply treat a row of relational data as a rather specialized "document", ingesting it as such in the indexing mechanism. Both the existing metadata, such as table and column names, and the actual data values in the database are included into the index. Text analytics is also used to understand meaning and identify textual relationships and patterns. But the cardinal relationship—the *raison d'être* of the relational system—is lost because with the search index, everything is flattened.

Conceptually, the database and index approaches are rather similar. The common function is the analysis that models meaning and relationships within information, whether hard or soft. For hard information, modeling is performed at design time, and permanently stored in the database structure and metadata. While this provides data quality and consistency as well as efficiency in use, it lacks agility to respond to unexpected queries. Softer information doesn't require a formal design-

time model; the "modeling" in this case occurs as a byproduct of ingestion. When a document arrives in the environment, its content is analyzed and indexed, often deploying various text analytics to add meaning, such as entity extraction, clustering, sentiment analysis, or classification.

Both approaches generate metadata describing the semantics and syntactics of the information. The resulting metadata is stored in an index or within the database to enable business use of the information. The differences lie in the timing of the analysis and the permanence of the resulting metadata. Hard information has its structure hardened when the schema is created. Because, in practice, schema change is cumbersome, all information must conform to the model (one part of the large cost of the 'T' of ETL). Soft information, in contrast, is defined on the fly. In other words, the metadata is generated as each piece of content is indexed, creating potentially unique metadata for each document. The downside is the difficulty in recognizing relationships spanning multiple documents and in creating and maintaining consistency of meaning across document stores.

The lure of the "mashup"

With two opposing sets of vendors and two different—superficially at least—solutions to the growing demand for a combined user view of hard and soft information, it was inevitable that an attempt to bridge the two worlds would emerge. Mashups of database and search technologies have attracted considerable attention in the market, promising users the possibility of combining hard and soft information from multiple sources, despite doubts about the levels of integration and agility they deliver. Users want the "Google experience" integrated with the precision and analytics of their business intelligence, enhanced with the awareness of context and personalization, and applied across the entire information landscape.

Application level mashups extend business intelligence to embrace search-like user capabilities such as natural language querying, spell correction, tag clouds, and more. If content is needed, it can still be added to the warehouse creating a convenient single-source information store. However, the limitations of the relational architecture remain; the soft information is essentially made hard, and much relevancy and agility lost in the process. More advanced mashups allow business to build dashboards and portals that request information from both database and content, using the appropriate querying model each time. Each portlet displays either hard or soft information, but never both together, with bridging logic that links information across the portlets.

The difficulty with mashups is twofold. At the level of the underlying data and content stores, the inherent limitations in capturing meaning and relevancy of both approaches still exist. At the level of the mashup, the bridge between the two worlds is weak and limited. Both approaches lack the insight and agility to dynamically interact with the information in response to the user's exploration of the information. Neither "*ad hoc* querying" promoted by business intelligence—the ability to ask new queries on demand based on the answers to the previous query—nor the Google-like intuitive search easily span the chasm to the opposite information class. The underlying meaning has not been sufficiently modeled or integrated to bridge the gap.

Refining the problem definition

Imagine a mashed up customer profile dashboard for a wealth management provider. A Google-like search for a customer's name can produce hard ERP data identifying the customer's investment history. Selecting one of the investments—a company name—can trigger a search of news articles mentioning the company, the most relevant articles first. But is this sufficient? Consider the queries: "What are our top selling products that get good or better reviews?" or "Give me all the people living in a college town where the news mentions the 'students are happiest.'" Mashups struggle because each question contains both hard and soft components that cannot be easily broken down into independent hard and soft queries. The problem is that there is no true integration of hard and soft information in the information layer, and thus no way to relate them.

Consider this common requirement: a seemingly simple scatter plot, wherein a non-technical user can intuitively plot any two or more key performance indicators against each other to unveil patterns that lead to better business decisions. Anticipating what measures the user will plot against one another is impossible. Furthermore, in many plots the data needs filtering. For example, a

CMO's dashboard includes a scatter plot for sales data and an area for listing news articles about the company and its products. The CMO plots point-of-sale location (state, province, etc.) against price for "toddler toys", and notices a cluster of "free" sales. In fact, these are free replacements for products recalled because of press reviews describing them as "hazardous to children", but the queries required in a mashed up system to discover this and filter these points from the plot are well nigh impossible to conceive. But imagine if the CMO could simply enter:

toddler toys –"hazardous to children"

Answering this requires a query that creates a scatter plot based on the sales data for the subset of toys classified as "toddler", filtering out results related to articles that mention the company's products in the context of "hazardous to children". The relationship is between the product data and news articles, the JOIN occurring between the sales data's product column and product names extracted from the news articles through entity extraction (a common text mining technique).

To execute such a query requires:

1. **Fully integrated metadata** covering and interlinking hard and soft information equally
2. **Pre-defined models** in key areas of the information (especially the hard information) to assure the quality and integrity of the data
3. **Post-defined models**, created at the time of document ingestion, of key concepts, phrases and relationships within the soft information and across to existing hard information
4. **Post-discovery[2] relationship creation** applied on demand at query time and defined by the context of the query

This is the "Google experience" combined with real business intelligence, with guaranteed consistency and integrity as well as true agility.

An architecture for a Unified Information Store (UIS)

Prior architectures dating back to the 1980s focused almost exclusively on hard information and its lifecycle from creation to archival or deletion. The original data warehouse architecture[3] is a good example. A more recent approach is the Business Integrated Insight (BI^2) architecture[4], driven by modern business needs for speed, agility and collaboration as well as emerging technologies such as service oriented architecture (SOA), Web and Enterprise 2.0 and new database approaches, clearly recognizes the need to architect all information and business processes more broadly. The unified information store defined here is a subset of the entire information resource of the business focused exclusively on the analytic use of combined hard and soft information. Figure 1 shows the two characteristic dimensions of the store: ***timeliness / consistency*** and ***structure / knowledge density***.

The timeliness / consistency dimension has long been recognized as a key characteristic of hard information, describing both the life cycle of data from creation through use to disposal and its journey from operational to data warehouse to data marts. These concepts apply also to soft information, but have seldom been explicitly stated. Along this dimension, ***live*** information is data in transit or in use to run the business. Such information is in constant flux. It may be an instant message or e-mail used in the negotiation of a new contract. It may be an order transaction in SAP or a

[2] Albala, M., *"Post-discovery intelligent applications: The next big thing,"* White Paper (2009), http://attivio.web101.hubspot.com/post-discovery-intelligent-applications-the-next-big-thing/

[3] Devlin, B. A. and Murphy, P. T., *"An architecture for a business and information system,"* IBM Systems Journal, Vol 27, No 1, Page 60 (1988) http://bit.ly/EBIS1988 and
Devlin, B., *"Data warehouse—From Architecture to Implementation,"* Addison-Wesley, (1997)

[4] Devlin, B., *"Business Integrated Insight (BI2)—Reinventing enterprise information management,"* White Paper (2009), http://bit.ly/BI2_White_Paper and
Devlin, B., *"Beyond Business Intelligence,"* Business Intelligence Journal, Vol 15, No 2, available 2nd Quarter 2010

status flag, such as "currently online". As time passes and we move to the right, information becomes more ***consistent***. Such consistency is achieved in data warehouses or content management systems and is key to dependable enterprise-wide information usage. Information finally becomes an ***historical*** record of the business.

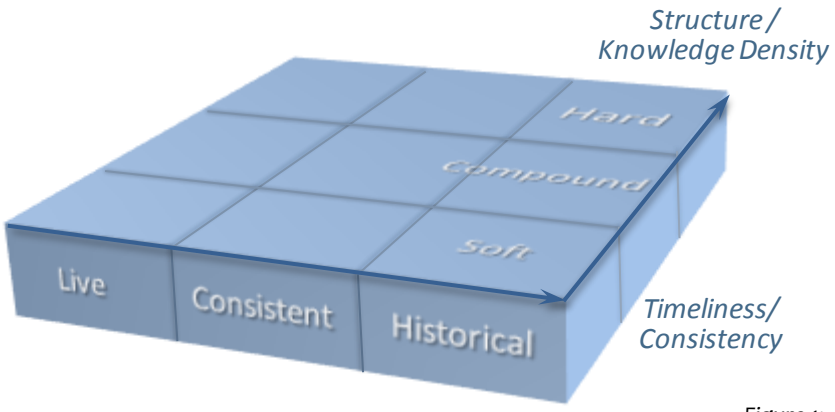


Figure 1:

The unified information store

On the structure / knowledge density dimension, we have already encountered ***hard*** and ***soft*** information in terms of its structure. Hard information is the highly structured data used to record business transactions, while softer information is text, image and so on. The center row—***compound*** information—is key to joining data and content for business use. Structurally, compound information is simply an admixture of hard and soft components, stored together. XML is a classic example, where tags provide the structure that defines hard data amongst the softer information components of text and images. But, to fully understand the importance of compound information, we need to understand why this dimension is also labeled knowledge density.

While data and information are vital to running a business, knowledge is indispensible to its understanding and management; knowledge is information with context and relevance. The journey from soft to hard information via modeling concentrates knowledge in information. In hard data, we know precisely what a particular piece of information stands for—because that is what it has been defined to mean. Its relationships are also predefined, so we understand a very specific context. However, this same predefinition of meaning and context creates its own problems: because it is based on the needs and understanding that motivated the original modeling, other meanings and relationships—its broader context and relevance, especially at the time when the information is explored and analyzed—may be lost. We cannot know everything beforehand. Soft information, while having lower knowledge density, often contains tacit knowledge about context and relevance of information that *a priori* modeling was not looking for and therefore overlooked. In terms of knowledge density, compound information has the best of both worlds.

As shown in figure 2, the level of structure and knowledge density of compound information precisely aligns with metadata, thus forming a conceptual and highly practical link between hard and soft information, between data and content.

Analytics based on compound information and metadata

Metadata is at the heart of the UIS architecture. Whether we approach the convergence of data and content from the hard or soft information side, we immediately encounter the need to create extensive indexes, pointers, descriptors and so on—metadata—to enable later use of the information. Whether it's created in the database design phase for hard information or at document indexing or use in soft information, this is still metadata. And whether derived from predefined structural elements in documents or folksonomies built on data warehouses, it is all still part of the compound information class of the UIS.

Figure 2:

Metadata and core business information

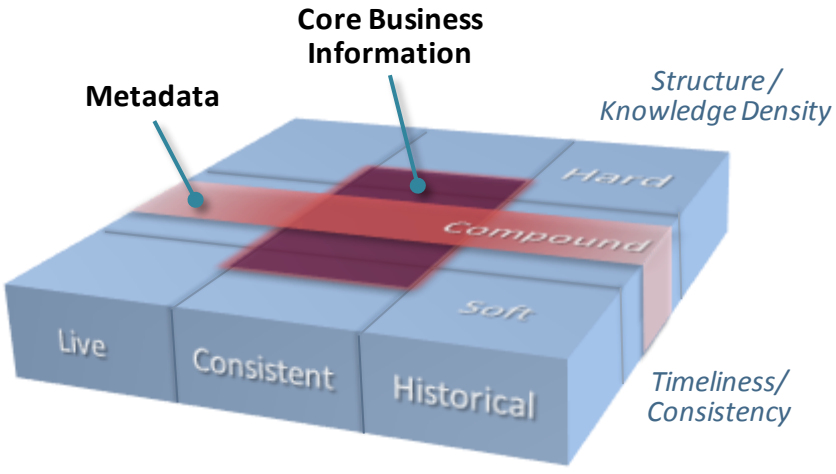


Figure 2 highlights metadata as part of the compound information class, and thus, as part of the information resource of the business, in must be considered and implemented together with the rest of the data and content. From left to right along the compound row of the grid, the live cell con-

tains the active, changing metadata needed for the agile, on-the-fly analysis of information characteristic of post-discovery intelligence. Shorter-term metadata is key to agility in the UIS. The consistent and historical cells represent more stable metadata, predefined via traditional modeling, distilled by text mining or gathered through usage analysis. Longer-term metadata is key to data quality; it defines basic meanings and relationships, which, if changed or ignored can render business information incorrect or meaningless. In its entirety, metadata underpins context and relevance of all information—hard and soft—in the unified information store.

All analytic work begins from the metadata layer. In content, from basic search to advanced post-discovery work, almost the entire analytic process occurs in this metadata component, using value-based or inverted indexes, document vector maps and so on. Traditional business intelligence (BI) pays lip-service to metadata; beyond using it to identify tables and columns, it is largely ignored. Analysis is presumed to happen in the hard data, but a large part of the analysis actually occurs in the indexes, as evidenced by the increase in query speed as indexes are added. These indexes are key-based and predefined at database creation, and lack the agility, context and relevance of the content approach. Extending inverted indexes to the hard data, a relatively simple technical step, is at the heart of the UIS. When this is done, all analytic work—for both hard and soft information—occurs almost exclusively in the metadata. In analytic work, data and content are re-united and users can immediately benefit from the combined strengths of the search and query paradigms—the agility and context of soft information and the accuracy and relationships of hard information.

Figure 2 also highlights "core business information". This set of information is of particular importance in ensuring the long-term quality and consistency of the unified information store. This information needs to be modeled and defined at an early stage of the design and its content and structure subject to rigorous change management. While other information may undergo changes in definition or relationships over time, the core business information must remain very stable. In harder information, this data is often termed "master data" and it is widely understood that such data requires special controls, because it is at the heart of much of the data relationships that make up the business. Among softer information, there also exist key documents, such as legal agreements, that also require special control to ensure that the business is what it says it is.

Reducing unnecessary information duplication

Another key driver for BI^2 and the UIS is to reduce unnecessary information duplication, a cause of many business problems. Data quality deteriorates as duplicates diverge and users base decisions on different copies. Attempts to reconcile disparate data and models increase IT costs. Decision speed is impacted as copies of copies of copies are created and maintained. As we've seen, the UIS removes duplication of soft data into the data warehouse. However, the UIS creates a more subtle but significantly more beneficial opportunity to reduce duplication of hard information. A vast amount of data is duplicated in data marts, either directly from operational systems or via the data warehouse. Many of these data marts are modeled and structured in advance to support various putative types of analytics by end users. Where such analyses are mainly standard reports, such predefined structures can provide strong ease-of-use or performance benefits. But, in many cases, these marts are meant to support considerable levels of *ad hoc* queries. In such cases, IT finds itself restructuring databases, rebuilding indexes and creating further data copies as users' needs evolve.

Applying the more flexible post-design indexing methods described above presents an opportunity to eliminate a significant amount of duplication in the data mart layer of current warehouses. This is made possible because all potential relationships in the data are either represented in the indexes created during data loading to the warehouse or can be generated on demand at query time. Eliminating or avoiding creation of data marts, without any reference to the value of joining data and content, could probably, on its own, justify implementation of a UIS.

And the architecture offers one further intriguing possibility: a novel approach to providing access to operational data using these same post-design indexing methods. This could remove much of the duplication of near real-time data inherent in today's approaches to operational BI. But, that's a topic for another day!

Where are we now?

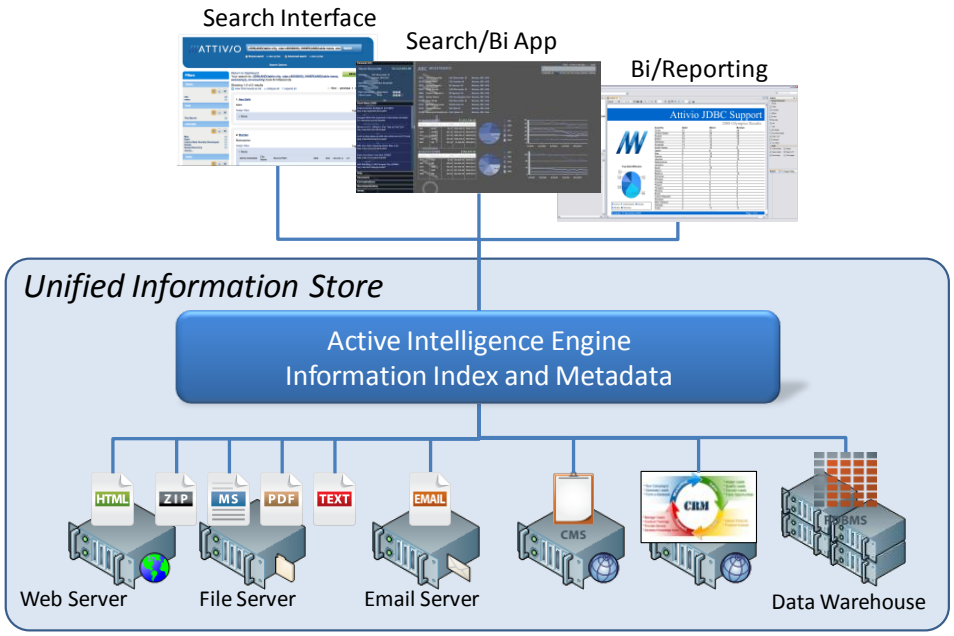
The UIS described above provides a logical information architecture for the convergence of data and content for analytical use. The architecture depicted is far from futuristic; in fact, many of the required components already exist in a variety of products. The approach builds on existing database and content management technologies. It strongly suggests that hard information (data) should remain in the data warehouse and marts where currently found and that soft information (content) should also be left where it is. However, the vital difference is the recognition that compound information and metadata form the basis for integrated analytic use of both classes. When fully created and populated, via prior modeling, during content or data loading and even during the analytic process, this metadata becomes the single, combined resource for agile, relevant and accurate access to and use of combined data and content. Even in this rather novel area, software developments are emerging in text analytics, XML and post-relational databases, as well as from vendors in the area of Unified Information Access[5].

As shown in figure 3, Attivio's Active Intelligence Engine (AIE) provides a relatively integrated solution[6], enabling well-balanced query and search access to a mixture of data and content (SQL and search syntax in one API). AIE ingests information from a wide variety of sources, content and data, into a single index. Text mining and analytics functions include entity extraction, content clustering and auto-classification. In hard data, existing relationships between tables and columns are retained, while new relationships can be generated on the fly based on the inverted index of all relational data. This enables arbitrary Joins within the hard information as well as across hard and soft information and provides the ability to deliver analytic solutions that seamlessly integrate data and content in a single user view.

Figure 3:

The UIS in Attivio's Active Intelligence Engine

Of course, in an emerging area, not all required function is yet available. No vendor yet provides support of the full SQL language set across data and content such that all existing BI implementations are guaranteed to work simply by redirecting them to the UIS. Also more traditional pre-canned reporting capabilities will still work most efficiently against existing relational databases. And there are still those who just can't pivot enough around a multidimensional cube (you know who you are)! Indeed, the aim at this stage is not to rip and replace existing content or BI applications or information stores. Rather, the goal for now is to begin the journey. A next step may be to build new dashboards and portals using a more UIS approach for applications where users need a truly integrated reporting and search experience.



[5] Evelson, B. and Brown M., *"Search + BI = Unified Information Access"*, Forrester White Paper, (2008)
[6] Meyer, S., *"Unified Information Access"*, White Paper, (2009), http://attivio.web101.hubspot.com/unified-information-access-part-two/

Conclusion

Content, or soft information, has always been of interest to the business in a wide range of processes, from marketing to executive decision-making. The explosion in volume and variety of soft information driven, in particular, by the Internet has sharpened that interest. However, with years of experience in business intelligence and data warehousing behind them, many users are clear that what they really need is an integrated view of soft information with the harder data already available in the warehouse. While soft information on its own does have value, the real business advantage will come from exploring the entire set of hard and soft information free from the limitations of the pervasive, predefined data structures of hard information.

This goal is made possible by the adoption of a unified information store architecture that explicitly (1) includes *all* classes of information—from soft to hard and from live to historical—and (2) creates a core set of metadata and business information at the heart of the architecture to ensure the overall quality and consistency of the information asset. The UIS integrates the precision of the database relationship with the richness of the search relevancy model without compromising the integrity of either. These principles of integration, quality and agility are at the heart of a true enterprise data warehouse and are thus extended to include both data and content as equal players in a unified information store.

Vendors from both the database and content sides of the industry are converging on this unified information space, bringing a variety of tooling and obvious technology preferences. The key to succeeding in truly unifying data and content will be in choosing tools that create, maintain and use an enhanced, common index and metadata store for both hard and soft information. For businesses that want early adopter advantage in this area, sufficient tools already exist to get started immediately.

Dr. Barry Devlin is among the foremost authorities on business insight and 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.