Data Discovery Automation

Learning from the Warehouse Experience

April 2016

A White Paper by

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

As data discovery becomes increasingly important for democratizing decision making and draws data from ever more sources, business must face the issue of managing the population of these environments. This leads to the concept of governed data discovery to ensure clean, consistent data for the business and to avoid a spaghetti-like population architecture for IT.

This paper proposes that such governance demands the creation of a Discovery Hub. This is a data store where core business information can be cleansed, reconciled, and made available as a consistent resource to data discovery users, while allowing them the freedom to source other data they need elsewhere.

Beyond defining the Discovery Hub, we further suggest that the best approach to populating this Hub, as well as the downstream data discovery environments, is via Data Warehouse Automation (DWA) approaches and tools.

In practice, we look at two leading tools, Qlik in data discovery and TimeXtender TX DWA, to discover some of the key requirements and implementation aspects of the Discovery Hub.

Contents

Sponsored by TimeXtender

Alas, poor Yorick! I knew him…" Rather like Shakespeare's Danish prince, data warehousing has of late been contemplating the empty eye sockets of its own mortality. On one side stand the behemoths of big data, declaring that no more is needed today than a data lake, filled to the brim with raw data in all its original—and unmodeled—glory. On the other side gather the scrums of self-servicers, secure in their over-confidence that data meaning and consistency are, well, obvious. At its 2016 BI and Analytics Summit in London, even Gartner has declared BI dead.

It's not the first time that the death of business intelligence or data warehousing has been announced. Early arguments proclaimed that a data warehouse was too expensive or took too long to build. Today's rationale is that it is no longer needed, because today's business users are supported by modern tools and have access to every last byte of data. The improved tools and access are undeniable facts. However, they offer no help in three key areas: (1) the long-standing, pervasive issues of poor data quality and documentation, (2) the lack of data consistency across sources, and (3) the analysis by business users of data beyond the boundaries of their knowledge and competence.

The purpose of this paper is not to defend the data warehouse; I have no doubt of its continuing necessity. In a big data, self-service world the data warehouse becomes the undisputed repository of the core, legally-binding information about the business' position and history. The question addressed here is: how can the wisdom and experience of data warehousing be applied to address the three problems mentioned above? In particular, how does this apply in the area of *data discovery*, where these issues are most pressing as users explore data and create insights at the edge of their own experience?

The answer is—in retrospect, perhaps—obvious. Using similar, but simplified, ideas that have been developed in data warehousing, we here define a ***Discovery Hub*** where core business data can be cleansed, reconciled and documented prior to making it available to business users through data discovery tools. As this paper explains, the structure of a Discovery Hub is simpler and its contents narrower than a traditional data warehouse. Its design allows and encourages business users to make it the trusted foundation in their discovery processes. It can easily be complemented with data from other sources, both local and external. Rather than being a single version of the truth, the Discovery Hub is a ***single source of truth***, but only in those areas where such assurance is needed. Appreciating this answer begins with understanding data discovery.

The Whys and Wherefores of Data Discovery

To quote Wikipedia[1]: *"Data discovery is a business intelligence architecture aimed at interactive reports and explorable data from multiple sources… [It is] a user-driven process of searching for patterns or specific items in a data set. Data Discovery applications use visual tools… to make the process of finding patterns or specific items rapid and intuitive."* As Wikipedia definitions go, this is a good place to start to explore the core aspects of data discovery.

The vision of being user-driven is far from novel or unique to data discovery. From the earliest days of data warehousing, we have striven to empower business users. The first paper[2] to describe a data warehouse architecture in 1988 identified the problem: *"many potential end users, especially at management level, have become* requestors *of information rather than* accessors *of information"*. The paper introduced the data warehouse as the core data storage and management component to address this issue. It also described at length an "End-User Interface" and modeling techniques to make data more useful and usable for business users. Unfortunately, technology limitations and management preferences of the time meant that much of what was delivered to business users was in the form of fixed reports and, more recently, dashboards.

Wikipedia's definition emphasizes the use of visual tools in data discovery. While reasonable in the light of today's market trends, visual tools are but one approach to discovering data, albeit one that resonates with many business users. Some users still prefer tabular formats, perhaps in particular cases. Others may choose statistical tools. The fundamental challenge is not the way the results are presented to the user or even how the data is represented in the store. Rather, it is the way that users often misunderstand the data and ignore its context, and their inability to infer the meaning and use of the data.

> *The most challenging issues for data discovery are that users misunderstand the data and ignore its context, and are further unable to infer the meaning and use of the data.*

Understandably, perhaps, based on the phrase *data discovery*, the above definition also focuses on exploration and discovery within the data available to users. This misses three key aspects. First, the end business outcome is to take decisions and actions based on insights gained from the data. The validity of such decisions and actions depends strongly on how well the users have understood the meaning of the data they use, as mentioned above. It also depends on the second missed aspect: has the user included all relevant data in the analysis? Third, and closely related to the second point, is the data included of sufficient quality and consistency to trust the conclusions drawn?

Taken together, these considerations lead to one key conclusion: Comprehensive preparation of the data used in exploration and discovery processes is vital.

In the past, this conclusion led directly to the data warehouse architecture and the routing of *all* enterprise data through a "central" structure, usually called an enterprise data warehouse (EDW). In this EDW, data was cleansed, reconciled and prepared for use by business users, typically in data marts fed from this quality source.

While emphasizing data provenance and quality, this traditional approach has long proven costly and time-consuming to implement. These limitations led, at least in part, to the rise of data discovery as a market—and the vendors, such as Qlik and Tableau, most closely associated with it—as business users voted for speed of access and ease of anal-

ysis over slower and less sexy IT-led deployments. With the focus on visual analysis, such products usually offered simple data ingestion approaches, typically directly from single sources into the data discovery environment.

However, the market has, again, come full circle. The concept of *governed data discovery* has now emerged, emphasizing the need to ensure the completeness, quality and consistency of the data available to business users. IT departments have been re-engaged and are attempting, with support from data discovery vendors, to (re-)introduce consistency and quality in populating data discovery systems. The approach is to reduce the number of independent feeds from individual sources, to combine data from diverse sources up front, and to build an intermediate store of data where data quality and governance can be applied. This store becomes a preferred source of data for common or corporate data elements for business users of data discovery. Additional sources, both local, such as spreadsheets, and external to the enterprise are at the discretion of the individual business user.

Five key features of governed data discovery

1. *Business users have full, flexible access to and exploration / analysis of all the data contained in the data discovery system*
2. *Business users determine which data should be included in the system*
3. *Business users assume responsibility for all data included in the system, particularly in the governance of data coming from novel and/or external sources*
4. *Business users rely on expert help (usually from IT) to assure the quality, consistency and contextual validity of data included in the system, especially core business information*
5. *In contrast to "basic" data discovery, the need for governance means that business and IT must collaborate closely in designing and building data discovery systems*

In February, 2016, Gartner revamped its *Magic Quadrant for Business Intelligence and Analytics Platforms*[3] to focus heavily on self-service analytics and platforms centered on *governed data discovery* by business users. Of particular interest is their prediction that by 2018, *"smart, governed, Hadoop-based, search-based and visual-based data discovery will converge in a single form of next-generation data discovery that will include self-service data preparation…"* Their suggestion that data preparation also becomes self-service should, in my view, be seen only as one part of the preparation process. Beyond this, and its unreasonable emphasis on Hadoop-based platforms, this forecast aligns well with the points outlined above. In fact, the shift is already under way.

The parallel with early data warehousing is clear. As users draw more data from more sources into their data discovery systems, issues of quality and consistency of that data, and impacts on source system performance and security come to the fore. The obvious solution is to build a common intermediate store. Is this the same as an enterprise data warehouse? Although it has some similarities, the short answer is no. There are too many differences between today's business and data environments and those of the 1980s. The long answer—the functions and architecture of such an intermediate store—is the subject of the next section.

Introducing the Discovery Hub

The previous section clearly indicates the need for a new architectural component to facilitate governed data discovery. In common with all IT architectural thought, to define such a component, we need to define and understand the principles underlying the architecture, the business requirements driving it, and the technology constraints bounding it. We also need a name for this component. We introduce the Discovery Hub.

This Discovery Hub is envisaged as a collection and management hub for *core business information*[*] *(CBI)* that is used widely by the data discovery community and should be consistent and correct across this distributed use.

Principles underlying the Discovery Hub

These principles define the level and extent of information governance offered by the Discovery Hub component:

1. The Discovery Hub offers a single, consistent and managed source of all internally sourced core business information used by business people in the governed data discovery environment. Where such CBI is offered, its use is mandatory unless sanctioned by the data governance function

2. Internally sourced information in the Discovery Hub comes from the business' operational systems, either directly or from the Enterprise Data Warehouse, if such exists and is deemed suitable. Where a Data Lake exists, some content there may be used as a source, provided it is of sufficient quality.

3. A single source is identified for each core business information element to offer maximum consistency and quality. The process of reaching agreement on each source will require strong collaboration across multiple business areas as well as IT.

4. For each identified source, *all* business information is brought into the Discovery Hub to facilitate all possible future business uses of this information. (This contrasts with the traditional practice which loads only the known subset of information that is needed for a particular business use or data mart.)

Core business information (CBI)

The set of strictly defined, well-managed and continuously maintained information that defines the business—its legally-binding identity, activities and actions—from its time of inception (or first recording) to the present moment.

Core business information combines data traditionally found in operational, MDM and EDW systems with other legally recognized information, such as contracts, as well as context-setting information, to provide the most accurate, consistent, complete and reliable picture of the business possible.

[*] In this section, *information* and *data* are used as follows (excluding phrases such as "data discovery" and "data governance"):

Information: the recorded and stored symbols and signs we use to describe the world and our thoughts about it, and to communicate with each other. Most information is digital—loosely and strictly structured—but also includes paper, books and analogue recordings.

Data is simple facts—measurements, statistics, the output of sensors, etc.—in the form of numeric values and their related names or keys, distinct from those values. Data is optimized for processing by digital computers and is a formal subset of information.

5. The Discovery Hub also offers a consistent, managed location to store other information, both internally and externally sourced, where cross-enterprise consistency is important. Use of such information is optional.

6. The Discovery Hub stores ongoing historical snapshots of CBI at the maximum frequency needed by discovery users.

7. In addition, the Discovery Hub is an ideal location from which to provide context-setting information[4] (also known as metadata) to distributed data discovery environments.

Business requirements driving the Discovery Hub

These requirements drive the function of the Discovery Hub as seen and needed by business users of governed data discovery:

A Discovery Hub is a data store containing cleansed, consolidated core business information in a loosely normalized structure that supports the common data needs of diverse and distributed data discovery environments.

1. The data discovery environment offers unlimited access to all information required for business exploration and discovery, both internally and externally sourced.

2. The use and structuring of all information in the data discovery environment is fully at the discretion of the business users themselves.

3. With the exception of CBI from the Discovery Hub, the sourcing of information used in the data discovery environment is also at the discretion of the business users. Exceptions and additional constraints may be imposed at the discretion of the data governance organization.

Technology considerations bounding the Discovery Hub

These considerations arise from the current and foreseen characteristics of the technology used for both the governed data discovery environment and the Discovery Hub:

1. Data discovery tools, like all forms of business intelligence, provide a "quasi-relational" or "spreadsheet-like" view of the information available. The view is thus one of data in the form of rows and columns, with relationships often more implicit than explicitly stated. This provides substantial ease of use and flexibility to users.

2. To facilitate these aspects of the data discovery environment, the Discovery Hub manages information in a relational format.

3. This relational format is characterized as being *loosely normalized*. Data related to key business entities, such as customer, product and order are stored in "wide" tables—such that data that is often used together in data discovery is stored together in the Discovery Hub—rather than strict 3rd normal form.

4. Because information is being combined from multiple sources, an underlying layer of more "strictly normalized" data may be required for management and control purposes. Such a layer could be provided by an Enterprise Data Warehouse or within the Discovery Hub. In either case, it is invisible to and unused by business people.

5. Data discovery environments are essentially general-purpose data marts that offer extreme flexibility in access and use through innovative data structures and processing approaches. The Discovery Hub matches these needs with loose normaliza-

tion rather than data structures, such as star schemas, that are optimized for specific types of analytic processes.

Architecture of a Discovery Hub

In today's data discovery environment, the users of each instance of the environment take responsibility for collecting data from any and all required sources. In some cases, BI departments create common and/or shared data sets of frequently used information. The Discovery Hub provides a more formal and generally applicable approach to creating, managing and providing such shared, consistent and quality information in a more governed environment.

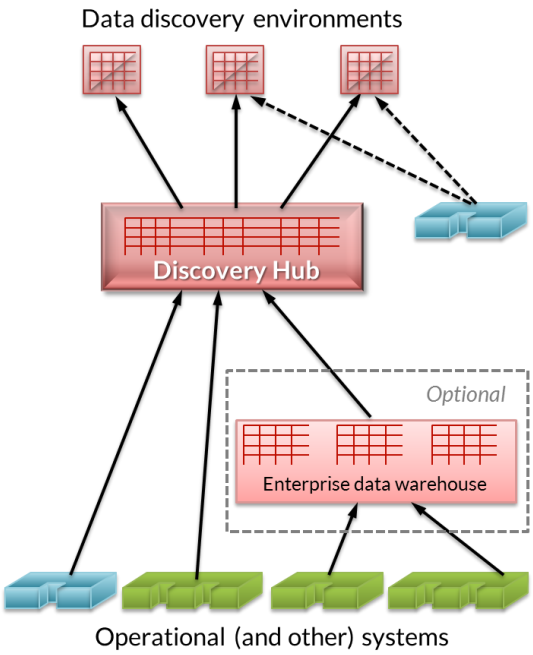

Figure 1: Discovery Hub architecture

Figure 1 illustrates the logical architecture of a governed data discovery environment supported by a Discovery Hub. Core business information originates from operational systems shown in green. Such CBI is consolidated and cleansed in the Discovery Hub and stored in a loosely normalized structure suitable for easy loading into data discovery environments. In an organization with good data governance all CBI is offered through the Discovery Hub. Where data governance approaches are still evolving, CBI may be migrated over time to a full Discovery Hub approach.

In organizations with mature data governance implementations, it is likely that an existing Enterprise Data Warehouse (EDW) can supply some or all of the core business information. To ease data management and limit conflict or duplication in such cases, information from the same source should flow either directly to the Discovery Hub or via the EDW. In implementation terms, both EDW and Discovery Hub components would likely reside on the same platform.

Information from other (non-CBI) systems shown in blue can be loaded into the Discovery Hub where a sufficient case for common information and shared use is established. More generally, such information flows directly to data discovery environments, as shown by the dashed arrows.

Longtime fans of data warehousing will surely feel comfortable with the architecture shown in Figure 1. Even excluding the optional EDW, the structure is reassuringly familiar: consolidating information from multiple sources into a common hub and then redistributing the cleansed and reconciled results to a distributed set of "data marts". While the final business purpose and internal storage structure of the Discovery Hub differ somewhat from the EDW, the key similarity lies in their population strategies. And, as in the case of the data warehouse, simplifying and automating this population is vital.

A Note on Data Lakes

While many organizations are adopting data lakes, an agreed definition is lacking. Most use the data lake as a first landing point for non-core data, such as social media, click logs and data from the Internet of Things. Others propose, in addition, to replace existing data warehouses with data lakes. In this case, the data lake may be a data source for the Discovery Hub. However, it is vital to ensure that the part of the data lake being used as a source contains core business information of high levels of quality and consistency.

Automated population of the Discovery Hub

Since the early days of data warehousing in the 1990s, experts have been extolling the virtues of automating the creation and management of data coming into and within that environment for informational use. As a consequence, extract, transform and load (ETL) tools emerged. Many of the original ETL tools invented then still exist and would certainly be appropriate for use in populating the Discovery Hub.

However, ETL tools have tended to focus on the individual population jobs (the arrows shown in Figure 1) rather than the overall process and have somewhat of a reputation for being overly IT oriented. In the case of Discovery Hub population, it would be more appropriate to focus on the overall population environment with improved ease of use for mixed business and IT audiences. These aspects are central to a more recent class of tools known as Data Warehouse Automation (DWA). Such tools address the entire scope of the full process of information (or data) preparation. This process starts with understanding specific and enterprise business needs, moves through finding and understanding data sources, deciding what the data must look like in the target to meet those needs, and finally, knowing how to keep all this going in rapid iteration as business needs and technologies change rapidly and often unexpectedly.

Data Warehouse Automation is the use of an integrated set of tools and techniques to automate the design, delivery and maintenance of data warehouses and marts. It is equally useful for other data stores, such as the Discovery Hub and data discovery tools.

In "Data Warehouse Automation Tools: Product Categories and Positioning"[5], Wayne Eckerson defines two broad approaches to DWA: *model-driven*, starting from a logical model of the business needs, and *data-driven*, which emphasizes the importance of first understanding the data available in the enterprise. At the heart of both approaches is the idea that business users must be involved from the beginning of the process, right through to the end. These same considerations, as well as the two approaches, apply to the implementation of Discovery Hubs.

While seemingly tangential to the discussion of populating the Discovery Hub, we must also consider the question of self-service population of data discovery environments—which will be of most interest to the business users—if we are to provide a useful and usable solution to true governance of this environment. Interest in this area is growing rapidly, driven by big data and the population challenges faced by data scientists, and goes under the moniker of data wrangling. Realistically, this market is simply another aspect of the bigger picture of populating informational data stores, which demands the creation of a valid, reusable and bidirectional bridge between business meaning and the data stored in computers. Further details and discussion of this topic can be found in my three-part series "Data Wrangling, Information Juggling and Contextual Meaning"[6].

With these considerations in mind, we define the key characteristics of a system providing automated population of the Discovery Hub:

1. *An interactive and inclusive system* supporting both business users who describe their needs for information (both core business information and other sources, where known) and IT support who understand data sourcing opportunities and constraints, and define the optimal data structures of the Discovery Hub.

2. *A store of context-setting information* (metadata) populated as per point (1) that provides the basis for conversation between business and IT at definition time and is

available for reuse in ongoing maintenance and change of the Discovery Hub. This store should form an integral part of the Discovery Hub infrastructure and be available to both business users and IT on an ongoing basis.

3. *Automated creation and management of population procedures*—either code generation or engine-based—to handle initial population of the Discovery Hub, regular updates with history, and changes due to evolving business needs.

4. *Self-service use by business people* to create and manage population of data discovery environments directly from novel or less-strictly governed information sources, and to enable requests for direct access to more strictly governed core business information as required.

The Discovery Hub in practice

How would you go about building a Discovery Hub? Let's take the case of a fictitious company, Amber Bank Corp (or ABC for ease of use) looking to democratize its decision making environment and a handful of real products—TimeXtender's TX DWA, Qlik and Microsoft SQL Server—to understand how to proceed.

ABC has separate operational systems on Oracle for its various lines of business, with SAP R/3 for payroll and general accounting. They have a data warehouse on SQL Server mainly supporting the Finance and Audit functions. Other departments do use some of that data but, truth to tell, much of the management information of the bank still comes from a variety of bespoke data marts used by a few business analysts, not to mention a plethora of spreadsheets used by all levels of management.

With a new strategic direction called "Democratizing Decisions", ABC plans to govern existing data properly and to expand their data sources to include transactional Oracle data, as well as ingesting social media and other external data. For decision making support at all levels, they have adopted a data discovery approach and the users themselves have chosen Qlik as their platform.

However, they face a challenge. With a history of data foraging for their spreadsheets, users pull data into Qlik from wherever they can find it and demand many special data feeds from IT. A spaghetti architecture is emerging with escalating costs and diminishing quality. The breaking point comes when the VPs of Finance and Operations come to blows in front of the CEO over the profit/loss figures for underperforming loans. The CIO receives an ultimatum: get governed in data discovery or get going on LinkedIn!

Initial analysis shows that business users are misusing and mis-combining data from some sources, leading to analysis errors. The same data is being worked multiple times. All this is immensely time consuming and diverts business users from real work. It is clear that basic sourcing and reconciliation should be performed once by data experts in collaboration with users and that a simplified but comprehensive view of core information—a Discovery Hub—should become its primary source. It is also clear that existing reconciliation work in the current data warehouse should be leveraged.

With the existing data warehouse on SQL Server and a Qlik-based front end, TX DWA is an obvious choice of data warehouse automation tool to orchestrate the complete

sourcing and reconciliation process. TX DWA is optimized for operation on SQL Server and can access data from all the original Oracle data sources, data already stored in the current data warehouse, as well as new sources of external data.

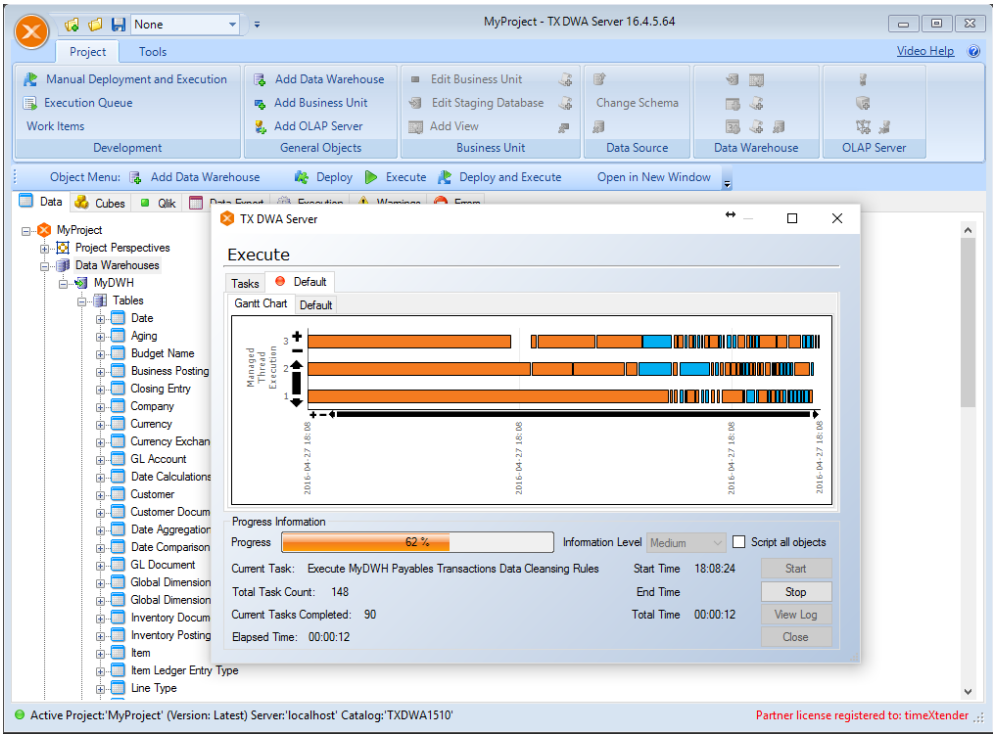

Figure 2: TX DWA user interface

Furthermore, TX TWA can simplify the creation and management aspects of populating the diverse and distributed Qlik environments, by replacing the normal hand-scripting approach with a drag and drop modeling interface that generates Qlik scripts and optimizes them by using surrogate keys and storage in native QVD files. Using the same tool inbound to and outbound from the Discovery Hub maximizes value from training and skills development, as well as providing opportunities for reuse. Indeed, ABC can consider revamping the existing, script-based sourcing of the current data warehouse as a follow-on project to gain further benefit from their investment.

TX DWA's central feature is a comprehensive drag and drop user interface that supports selection of sources and targets, creation of cleansing, transformation and reconciliation procedures, and the creation of a comprehensive store of metadata that documents the system and enables simple and rapid maintenance and upgrade.

Team work and collaboration is a vital element in building an integrated but distributed system of population function into and out of the Discovery Hub. TX DWA supports multiple concurrent developers at varying levels of collaboration based on the central metadata store. Version control and support for development, test, and live environments are included. Reuse of SQL code enables higher levels of productivity for developers, as does the ability to fully integrate customized, hand-written code to support unique transformation needs.

For business users, the rapid prototyping of target tables and ongoing delivery of requested changes by IT allows the users to focus on the data they require and know that it will be delivered when needed. With IT and business focusing on the same definition of requirements in the target tables, collaboration is eased and the chances of misunderstandings reduced in a rapid turn-around, agile development environment.

Under the covers, TX DWA optimizes system performance by using a code generation approach that takes advantage of all current and upcoming features of SQL Server, as well as automatically managing the order of execution of different procedures. This fully utilizes available hardware and processing cycles. If required, certain sources can be

prioritized. Incremental load is supported from sources that can identify changed records or by using the structure and contents of the target database. Load failures are handled gracefully and with minimal rework.

All the above features are, of course, demanded by data warehouse development. The Discovery Hub also benefits from them. In this, we see the fundamental similarities between the two components. The key difference, however, is in their purpose. Because the Discovery Hub focuses only on supporting data discovery (as opposed to the entire spectrum of management information and decision making support targeted by a data warehouse), the process of building and maintaining the population environment can be more effectively simplified and speeded up by DWA tooling such as TX DWA.

Conclusions

Data discovery is becoming an increasingly important aspect of business intelligence. Business users appreciate the flexibility they gain by being able to collect the data they need, when they need it, and manipulate it easily, particularly in visual environments. The business value gained by these approaches is undisputed. Further growth in the use of such tools can be anticipated.

Nonetheless, this growth brings its own problems. Business users gather multiple copies of the same core information, creating challenges for the management and performance of the source systems. Users transform incoming data according to their own best—but often flawed—understanding of the source systems, leading to erroneous and conflicting interpretations. A new and improved approach to the governance of data arriving in the diverse and distributed data discovery environments is required. The concept defined here of a Discovery Hub, containing cleansed, consolidated core business information in a loosely normalized structure offers a solution.

Data warehouse automation tools are a necessity when considering the population—both inbound and outbound—needs of the Discovery Hub.

Data warehouse automation tools, such as TimeXtender TX DWA, are a necessity when considering the population—both inbound and outbound—needs of the Discovery Hub. DWA tools simplify, speed and automate the design, development and maintenance of these feeds. Only through the use of such tools can the business properly govern the information used in data discovery and can IT feasibly provide a usable and maintainable data provision environment. Together, business and IT are thus able to claim a truly governed data discovery environment.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His new book, ***"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"*** *(http://bit.ly/BunI-Technics) was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and IT implications of deep business insight solutions that combine operational, informational and collaborative environments. A regular tweeter, @BarryDevlin, and contributor to ITKnowledgeExchange and TDWI, Barry is based in Cape Town, South Africa and operates worldwide.

---

[1] Wikipedia contributors, "Data discovery", *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=Data_discovery&oldid=696373559 (accessed April 4, 2016).

[2] 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

[3] Gartner, "Magic Quadrant for Business Intelligence and Analytics Platforms", 4 February 2016, ID:G002758, http://www.gartner.com/doc/reprints?id=1-2WQY2ZG&ct=160121&st=sb

[4] For a definition and discussion of context-setting information, see Devlin, Barry, *"Business unIntelligence"*, (2013), Technics Publications LLC, NJ, http://bit.ly/BunI_Book

[5] Eckerson, Wayne, "Data Warehouse Automation Tools: Product Categories and Positioning", September 2015, http://eckerson.com/register?content=which-data-warehouse-automation-tool-is-right-for-you

[6] Devlin, Barry, "Data Wrangling, Information Juggling and Contextual Meaning", July-August 2015, http://bit.ly/1S6UJBr