Data Discovery Automation

Learning from the Warehouse Experience

VERSION 2, OCTOBER 2017

A White Paper by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

As data discovery becomes increasingly important for insightful and pervasive 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, agile 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.

Finally, we describe TimeXtender's Discovery Hub® software and how it addresses the requirements and implementation aspects of the Discovery Hub architecture in conjunction with a variety of data discovery tools.

Contents

IN COLLABORATION WITH TIMEXTENDER

www.TimeXtender.com

A*las, poor Yorick! I knew him…"* Rather like Shakespeare's Danish Prince Hamlet, data warehousing has of late been contemplating the empty eye sockets of its own mortality. Fear not, salvation is at hand.

On one side stand the behemoths of big data, declaring that no more is needed today than some vast, ill-defined 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 declared BI dead[1].

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: business users have modern tools and access to every last byte of data. The improved tools and agile 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 do not doubt 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 *data discovery*, where these issues conflict directly with users' need for agility in exploring data and creating insights at the edge of their own experience?

Finding answers to these questions begins with understanding data discovery.

The Whys and Wherefores of Data Discovery

So, what is *data discovery*? Where better to go for a definition than Wikipedia? As it happens, its definition there is rather complete and provides a good place to start to explore the core aspects of data discovery.

To quote Wikipedia[2]: *"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."*

The vision of being user-driven is not novel to data discovery. Since its inception, data warehousing has striven to empower business users. The first architecture paper[3] in 1988 identified the problem: *"many potential end users, especially at management level, have become* requestors *of information rather than* accessors*…"* The data warehouse was the core storage and management component to deliver consistent data. An "End-User Interface" and modeling techniques aimed to make data more usable by business users. Unfortunately, technology limitations and management preferences of the time meant that much of what business was offered was in the form of fixed reports.

Wikipedia's definition emphasizes the use of visual tools in data discovery. This, however, is not the fundamental challenge. Rather, it is the way that users often misunderstand the data and miss its context, as well as their inability to infer its meaning and use, as they explore and discover within the data *available* to them. This raises four key questions. First, has the user included all relevant data in the analysis? Second, is the data included of sufficient quality and consistency to trust the conclusions drawn? Third, is the data intelligible to the user? And, fourth, does all the available data have the same needs for governance and agility?

A challenging business issue for data discovery is that users misunderstand the data, miss its context, and are thus unable to infer its meaning and use.

All these considerations lead to one key conclusion: Comprehensive preparation of the data used must be combined with freedom and agility to discover and explore it.

The original solution, which emphasized governance, was the data warehouse architecture and the routing of *all* enterprise data through a "central" enterprise data warehouse (EDW). Here, data was cleansed and reconciled. Data marts, fed from this quality source, prepared data for use by business users. Varying implementation approaches imposed different structures on the main data store: Compare Inmon's normalized, subject areas with Kimball's query-driven dimensional, star schema.

However, these traditional approaches have long lacked agility. This led, at least in part, to the rise of data discovery—and the vendors and products, such as Qlik, Tableau and Microsoft PowerBI, most closely associated with it—as business users voted for speed of access and ease of analysis 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.

Today, the market has, again, come full circle. *Governed data discovery* emphasizes the need to ensure the completeness, quality and consistency of the data available to business users. Gartner's 2017 *Magic Quadrant for BI and Analytics Platforms* [4] says: *"influence is tipping back to include IT… to place greater emphasis on enterprise readiness, governance and price/value, in addition to the agility and ease of use demanded by business users."*

In short, we need agile, governed data discovery according to the following principles:

Four guiding principles of agile, governed data discovery

1. ***Comprehensive data availability:*** Decision makers must have full, flexible access to and exploration of all the data the business requires to address rapidly changing opportunities and emerging issues

2. ***Balanced governance and agility:*** Business must define and ensure the appropriate balance between governance and agility for all data, including both internal operational data as well as that coming from novel and/or external sources

3. ***Proportionate delivery infrastructure:*** IT must define and implement the necessary infrastructure to deliver data according to varying governance and agility needs, especially those of legally-binding operational sources

4. ***Business-IT collaboration:*** The above needs demand close collaboration between business and IT in implementing a pervasive, comprehensive data discovery system

These principles demand an IT-provided infrastructure to impose—or, at least, aid—governance before data arrives in the discovery tools, and allow agility in initial delivery and ongoing use. *Business must be directed to quality data where accountability requires and be allowed to improvise with additional data as innovation demands.* The architecture and functions of such an infrastructure is the subject of the next section.

Introducing the Discovery Hub®

> A Discovery Hub® is an architectural structure that balances the governance and agility of data as it is provided to the data discovery tools that are used directly by business people in support of decision making.

The previous section indicates the need for a new architectural approach—the Discovery Hub—to facilitate balancing governance and agility in data discovery. We begin with defining and understanding the underlying principles and business requirements, and then construct the logical architecture.

Our starting point is to characterize different types of data/information* a modern business handles from the perspective of agility and governance. Whether data is internally or externally sourced is the first consideration. Internal data is usually better governed than external data, because the enterprise has control of its original, operational sources and defines approved processing and use patterns. However, when business users make and modify personal copies of internal data—in spreadsheets, for example—governance is compromised in favor of agility.

Strong governance is essential for *core business information* (CBI), the legally-binding record of the business, consisting of transactional and master data, as well as the context-setting information (or metadata) describing it[5]. It is created and well-governed in the operational systems and requires similar governance on its entire journey through data preparation and storage all the way to data discovery. Other types of internal data, especially derived data generated by business users, such as forecasts and summaries require lighter governance but demand substantial agility.

Similar agility vs. governance considerations apply to externally sourced data: Is it central to business processes or used in exploratory ways? However, with external data, one aspect is key. Because the business has little or no control over the sources or what happens to the data in transit, full governance can only apply *after* its arrival. Despite its business value, the quality of the data remains dubious and usage warnings should be considered. Furthermore, the quality of combined internally- and externally-sourced data requires detailed attention to determine and record valid and permitted use.

* The terms *information* and *data* are much abused. Information—fully contextualized and described—is what business needs to support decisions. In fact, *data discovery* might be better named *information discovery*. Data is simple facts—measurements, statistics, the output of sensors, etc.—gathered by IT and optimized for digital processing. Data is supported by a complementary, often separate, set of metadata. Metadata provides context for data, and would be better called *context-setting information*.

Simply speaking, when business wants information, IT must combine data with the comprehensive and appropriate metadata needed to (re-)create it. The fundamental purpose of the Discovery Hub—and, indeed, any data warehouse—is to do just that.

Principles underlying the Discovery Hub

These principles define the balance between governance and agility offered by the Discovery Hub:

1. *Single source of truth:* The Discovery Hub offers a consistent, managed location to store all information, both internally and externally sourced, where cross-enterprise consistency is important, balancing needs for governance and agility.

2. *Sources:* Internally sourced information in the Discovery Hub comes from the business' operational systems, either directly or through the EDW, if such exists and is deemed suitable. Other internal sources, such as spreadsheets, and external sources are included where managed to an acceptable level of quality. Where a data lake[†] exists, some content (of sufficient quality) from there may also be included.

3. *Consistency:* The Discovery Hub offers a single, consistent and managed source for all core business information used by business people in the data discovery environment. Where such CBI is offered, its use is mandatory unless specifically exempted by the data governance function.

4. *Single sourcing:* A single source is identified for each CBI 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.

5. *Deep sourcing:* For each identified source, *all* useful business information is brought into the Discovery Hub to encourage maximum possible future business use. (This contrasts with the traditional practice of loading only the subset of information that is needed for a known business use.)

A Discovery Hub is the single, consistent source of truth for business people working with data discovery tools.

Business requirements driving the Discovery Hub

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

1. *First stop shop:* The Discovery Hub is the first port of call for all data—both internally and externally sourced—required for business exploration and discovery.

2. *One stop shop:* The Discovery Hub is the only mandated source of core business information for use in data discovery. Other internally- and externally-sourced data should, in preference, flow through the Discovery Hub for convenience and quality.

3. *Neutrality of form:* The Discovery Hub provides data in a usage-neutral form for maximum agility in data discovery. The use and structuring of information and the tools used is at the discretion of the business users, with support and guidance from IT.

A Discovery Hub is the first and—ideally—only stop for business people working with data discovery tools.

[†] 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, external data, such as social media, click logs and data from the Internet of Things. Others propose, in addition, to replace existing data warehouses. If a data lake is to be a source for the Discovery Hub, it is vital to ensure that the part of the data lake being used as the source contains data of high levels of quality and consistency.

4. *Freedom of information:* External information used in data discovery may be obtained from sources other than the Discovery Hub at the discretion of the business users. Exceptions and constraints may be imposed for quality assurance by the data governance organization where such information is already in the Discovery Hub.

Discovery Hub—a Logical Architecture

The above principles and requirements drive the logical Discovery Hub architecture.

Data discovery tools provide a "quasi-relational" or "spreadsheet-like" view of data: rows and columns, with relationships often more implicit than explicitly stated. This provides substantial ease of use and flexibility to users. To facilitate these aspects of the data discovery environment, the Discovery Hub manages information in a *relational format*. Furthermore, data discovery tools—through their innovative data structures and processing approaches—offer extreme flexibility in data access and use. Therefore, the Discovery Hub avoids data structures optimized for specific types of access or analysis, and the ideal relational format in the Discovery Hub is *loosely normalized*.

The *Modern Data Warehouse (MDW)*, shown in Figure 1, is the key component of the Discovery Hub responsible for providing a *single source of truth* for users of data discovery tools, where such assurance is needed. (Contrast this aim to the largely unachievable single version of the truth that has been the historical aim of data warehousing.) Within the MDW, data related to key business entities, such as customer, product and order are stored in "wide", loosely normalized tables—so that data that is often used together in data discovery is stored together—rather than the less flexible 3rd normal form (3NF). Other, more specific data structures, such as star schemas, that are explicitly optimized for predefined types of analytic processes, may be used in the MDW as starting points for eventual implementation of the preferred loosely normalized format.

A Discovery Hub is a data store containing cleansed, consolidated business information supporting the common data needs of diverse and distributed data discovery.

The Modern Data Warehouse stores a long-term historical view of business information at the maximum frequency needed by data discovery users. This view may consist of snapshots or be a true bitemporal structure.

Because information is being combined from multiple sources in the MDW, an underlying layer of more "strictly normalized" data aligned with the source systems is required for management and control purposes. This layer is the *Operational Data Exchange (ODX)* within the Discovery Hub. In normal use, it is invisible to and unused by business people, although Data Scientists may access it for analytic purposes. The ODX is an IT construct, and owned and managed by IT.

For every internal source accessed, shown in green in Figure 1, the ODX is populated with all data of likely business value found in that source in its raw, native (likely normalized) form. Cleansing and transformation are kept to the absolute minimum needed to load the data in a form acceptable to the target database. The ODX thus contains an auditable copy of the source systems and serves to reduce the impact of the Discovery Hub on primary source systems.

External data sources—such as social media or IoT— and colored blue in Figure 1 require individual consideration. Initially, they may be loaded directly to the data discovery tools by business users. This directly supports agility, but may pose problems for governance.

If such data becomes vital to key decision processes, business and IT together should consider routing such data through the Discovery Hub. Internal sources of lower quality data—such as user-created spreadsheets—may follow the same process for inclusion.

In situations where significant data transformation is required between the ODX and MDW, a technical *Data Staging Area* may be needed. This store is invisible to the business in almost all cases.

The Discovery Hub contains significant amounts of context-setting information / metadata to enable and simplify its use by data discovery tools. This *Semantic Layer* allows mapping from the MDW to a business context model that meets the needs of a specific business unit—organizational, geographical, etc.—or analytical purpose. Business context models can be automatically mapped to a variety of tools and formats, such as OLAP cubes or product-specific formats for Qlik, PowerBI, Tableau, and so on. This shared Semantic Layer is central to users' understanding of the data available in the MDW and to their well-governed and innovative use of it. It enables business to use the same language across the entire organization in every front-end fed through the Discovery Hub. Meanwhile, it provides different perspectives—grouping together related data for area-specific purposes. In short, models in the shared Semantic Layer are defined once and automatically used to deliver data in the right form and context to any supported data discovery tool.

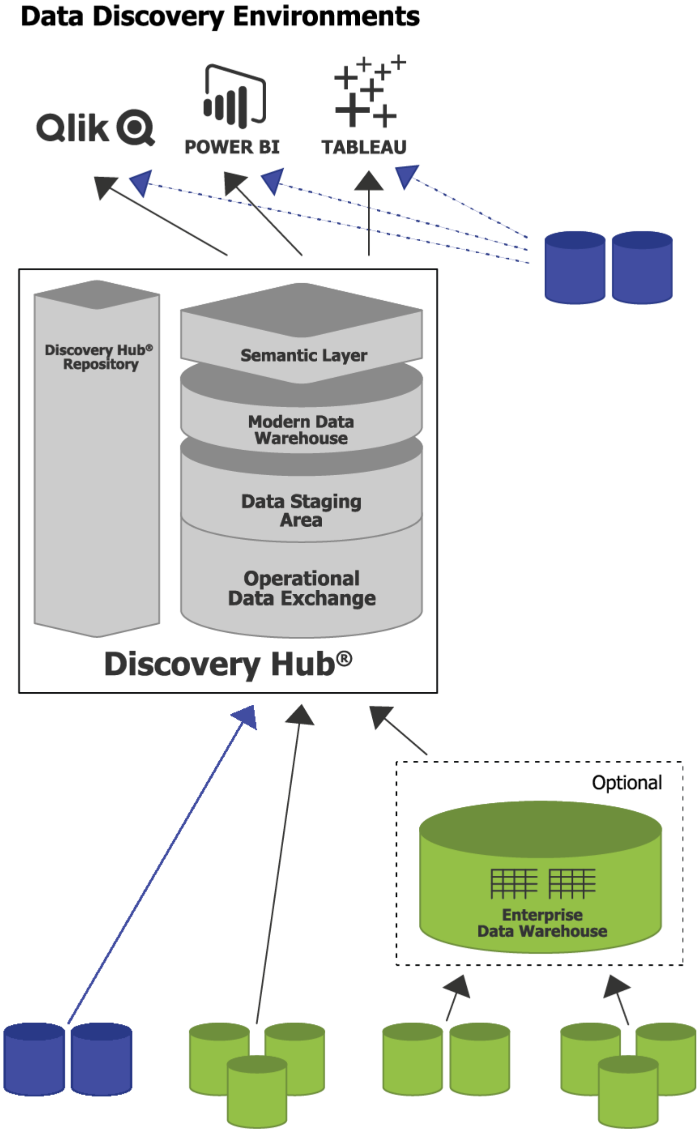

Data Discovery Environments
Qlik
POWER BI
TABLEAU
Discovery Hub® Repository
Semantic Layer
Modern Data Warehouse
Data Staging Area
Operational Data Exchange
Discovery Hub®
Optional
Enterprise Data Warehouse

Figure 1: Discovery Hub architecture

Much of the metadata in the Semantic Layer is created in the design and build process of the Discovery Hub. Automating this process is vital to delivering the required governance and agility of the environment. Such automation produces a *Discovery Hub Repository*, containing vital context-setting information, including versioning and lineage of data, usage documentation, scheduling, and so on. This repository is further updated during operation of the environment with information about data updates, failed loads, etc.

Longtime fans of data warehousing will surely feel comfortable with this architecture as shown in Figure 1. 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. The difference is that the Discovery Hub provides an integrated development and delivery environment from sourcing through to data discovery tools, enabling improved agility and governance for both business and IT.

Automating Discovery Hub development

Data Warehouse Automation provides an agile development approach and an integrated metadata store for joint use by business people and IT from initial inception to ongoing use and maintenance of the Discovery Hub and data discovery environment.

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. Therefore, extract, transform and load (ETL) tools first emerged.

More recently, Data Warehouse Automation (DWA) tools address the entire scope of the full process of information / data preparation with an emphasis on agile development and consistent, well-governed data stores. This process automated and largely hidden by DWA 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. Therefore, DWA tooling, rather than ETL, provides the ideal development platform of a Discovery Hub architecture.

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 ideal for the Discovery Hub and data discovery environment.

The key characteristics of a DWA development environment for the Discovery Hub are:

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 captures the conversation between business and IT at definition time and is available for reuse in ongoing maintenance and change of the Discovery Hub. This metadata store is an integral part of the Discovery Hub and used by both business and IT.

3. *Automated creation and management of population procedures* 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.

With an architecture and development environment characterized, it's time to look at a product implementation.

The Discovery Hub in practice

TimeXtender’s Discovery Hub®, built upon their existing data warehouse automation platform, provides a complete development, operating and maintenance environment for implementation.

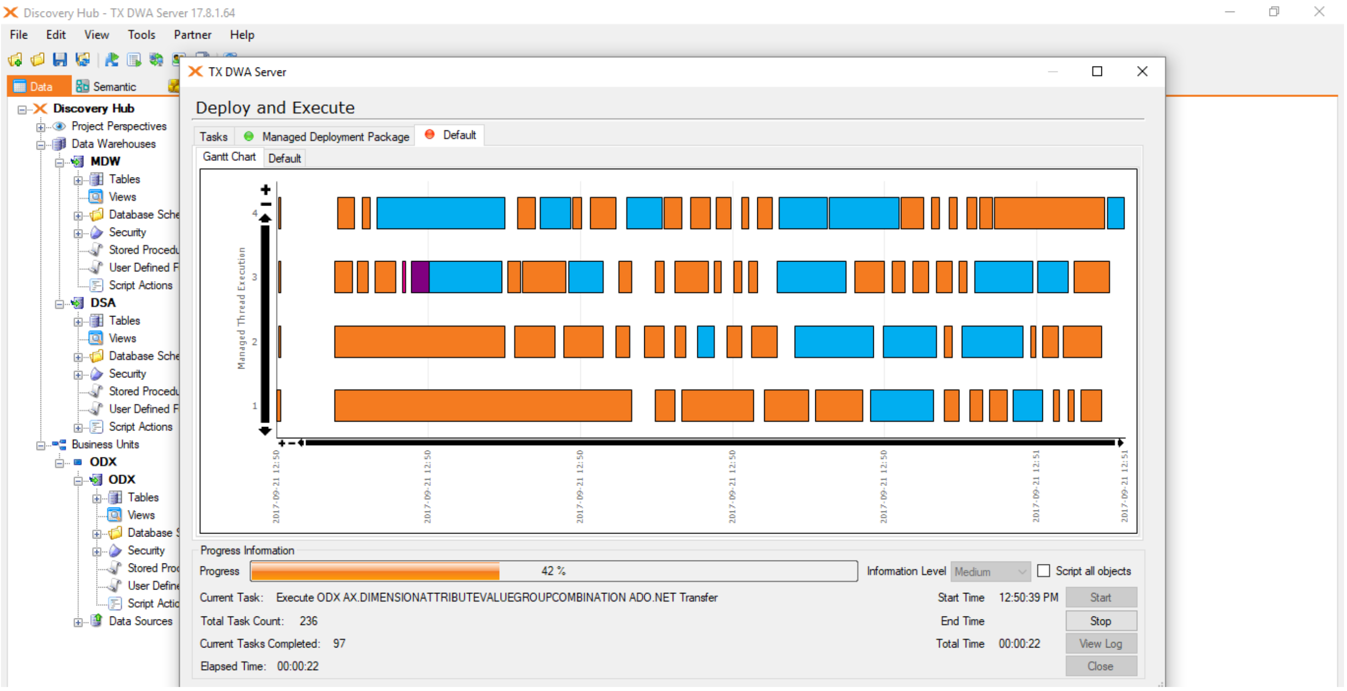
So, how would you go about building a Discovery Hub?

TimeXtender, a long-time developer of data warehouse automation software since 2005. The concept of a Discovery Hub emerged from their experience with DWA projects that were increasingly driven by the growing business need for an agile data discovery, self-service environment. As a result, they have been extending their DWA environment to meet the needs of the Discovery Hub architecture, and have now inaugurated their software with the same name.

The central visible feature of TimeXtender’s Discovery Hub® is a comprehensive drag and drop user interface, shown in Figure 2, 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.

Through this interface, the software creates and populates the metadata that is the foundation of the Discovery Hub and auto-generates the code required for the whole lifecycle. This combination of metadata and code accelerates data sourcing, avoids human errors, keeps track of all data in the different levels of the Discovery Hub, and auto-generates the documentation required for everything from design and maintenance, through operation and auditing, all the way to regulatory compliance.

For business people, the rapid prototyping of target tables and ongoing delivery of requested changes by IT allows the users to focus on analyzing the data they require and



Progress Information

Progress 42 %

Information Level Medium ☐ Script all objects

Current Task: Execute ODX AX.DIMENSIONATTRIBUTEVALUEGROUPCOMBINATION ADO.NET Transfer

Total Task Count: 236

Current Tasks Completed: 97

Elapsed Time: 00:00:22

Start Time 12:50:39 PM

End Time

Total Time 00:00:22

Figure 2: TimeXtender’s Discovery Hub® user interface

know that it will be delivered when needed. With IT and business focusing on the same definition of requirements in the target tables, the Discovery Hub ODX/MDW structure eases collaboration and allows both parties to focus on their areas of expertise. Speedy, iterative analysis and development reduce the chances of misunderstandings and rework in an agile environment.

Team work and collaboration is an important element in building an integrated but distributed system of population function into and out of the Discovery Hub. TimeXtender's Discovery Hub software supports multiple concurrent developers at varying levels of collaboration based on its 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, handwritten code to support unique transformation needs.

TimeXtender's Discovery Hub optimizes system performance by using a code generation approach that benefits directly from Microsoft SQL Server features as they evolve and leverages the newest functions as appropriate, 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. Load failures are handled gracefully and with minimal rework.

TimeXtender's Discovery Hub is optimized for operation on SQL Server and Azure SQL database and can access data from a wide variety of data sources, data already stored in the current data warehouse, as well as new sources of external data.

Furthermore, TimeXtender's Discovery Hub simplifies the creation and management of populating diverse data discovery environments, including Qlik, PowerBI and Tableau, by replacing the normal hand-scripting approach with a drag and drop modeling interface that generates and optimizes scripts according to the needs of the specific tools. 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.

Four varied, real-life examples of the practical implementation of a Discovery Hub can be found in the associated white paper[6] "Data Discovery—Right on Time".

Conclusions

Data discovery is an increasingly important form of BI. Business appreciates its flexibility and usability. However, data quality can suffer as users explore and experiment. A Discovery Hub, driving both governance and agility, is key.

Business users delight in 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 is certain.

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.

Balancing governance and agility, the Discovery Hub drives faster, deeper business insight than previously possible.

Data warehouse automation tools, such as TimeXtender's Discovery Hub®, are a necessity when considering the population needs —both inbound and outbound—of the Discovery Hub. DWA tools simplify, speed and automate the design, development and maintenance of these feeds. Only using 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, support the rapid growth in data volumes, and drive faster and deeper business insight than previously possible.

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 book, ***"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"*** *(http://bit.ly/BunI-TP2) was published in October 2013.*

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and technological implications of deep business insight solutions combining all aspects of internally and externally sourced information, analytics, and artificial intelligence. A regular contributor to Twitter (@BarryDevlin), TDWI, BACollaborative, and more, Barry is based in Cape Town, South Africa and operates worldwide.

Brand and product names mentioned in this paper are trademarks or registered trademarks of TimeXtender, Qlik, Microsoft, Tableau, and other companies.

[1] Timo Elliott on Gartner's 2016 BI Summits. "BI is Dead", April 1, 2016, http://timoelliott.com/blog/2016/04/bi-is-dead.html

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

[3] 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

[4] Gartner, "Magic Quadrant for Business Intelligence and Analytics Platforms", 16 February 2017, ID:G00301340

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

[6] Devlin, Barry, "Data Discovery—Right on Time, How Automation Underpins Data Discovery", August 2017, http://go.timextender.com/hubfs/Analyst%20Reports/Discovery%20Hub%20Use%20Cases%20White%20Paper.pdf