All Is Well with Data Virtualization

APRIL 2019

A Blog Series by
Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

This series of four blog posts, published by Denodo between November 2018 and April 2019, explores data virtualization and its application in a variety of use cases.

- Virtualizing the Data Warehouse *introduces the concept of data virtualization and Denodo's product architecture*
- Gaining Real-time Insight *illustrates how data virtualization allows access to real-time data in combination with point-in-time warehouse data*
- A Warehouse in a Lake, Data Virtually *describes an architecture that supports both a data warehouse and data lake accessed through data virtualization*
- More Structured or Less, Data Virtualization Delivers *discusses how data virtualization allows access to and use of data of all levels of structuring within a typical application*

All these uses are described in the context of an imaginary bank and the efforts of its new CIO, Alice Well, to overcome years of legacy thinking and old systems.

SPONSORED BY
www.denodo.com

Virtualizing the Data Warehouse

November 28, 2018

http://bit.ly/2E0vUc5pr

Data virtualization allows builders of data warehouses (and a range of other data management systems) to offer integrated access to data distributed and stored over multiple physical platforms.

2018 is exactly thirty years since I published the first data warehouse architecture. The core concepts of the architecture remain strong and viable. However, the evolution of both business needs and technology solutions over the intervening decades has led to a whole new set of architectural considerations and concepts such as "logical data warehouse", "data lake", "enterprise data platform", and more.

These concepts are often presented as competing approaches or replacements for data warehousing, leading to much confusion among businesses trying to wring value from data. This misunderstanding is most easily eliminated by focusing on one key technology—data virtualization—which allows builders of warehouses (and other data management systems) to offer integrated access to data distributed over multiple physical platforms.

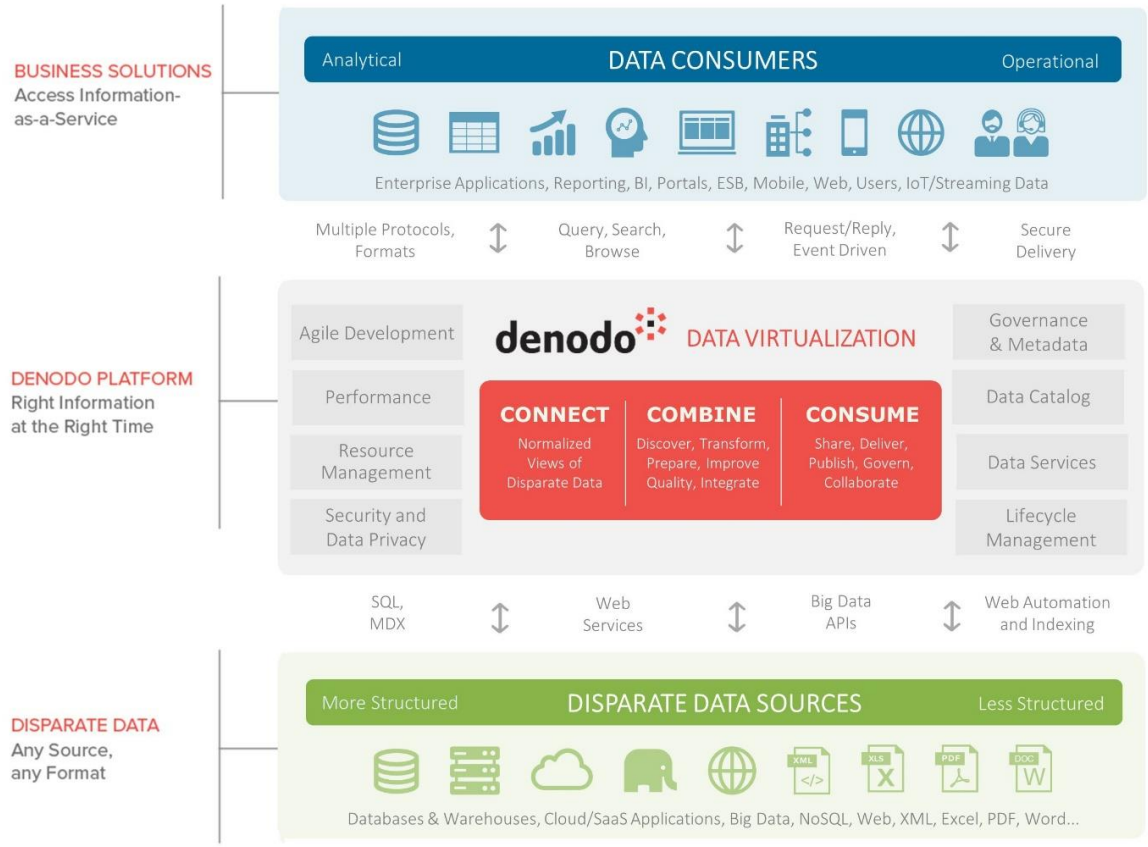
To understand the importance and central role of data virtualization, we need only condense the above-mentioned business and technology evolution into a single sentence: *Business requires and will continue to require ever more data, more quickly and more agilely, from more varied sources, both internal and external, than ever before.* Put like that, the idea of trying to force all the data the business may ever need into one store before business people can use it for gaining insight is complete nonsense.

Does this negate the idea of a data warehouse? Of course not! There remains a subset of data for which a single, reconciled store makes sense. That subset is called core business information: the set of strictly defined, well-managed and continuously maintained information that defines the business—its identity, activities and legally relied-upon actions—from its time of inception to the present moment. This is a smaller volume of data than stored in most modern data warehouses. And even this set of data may be divided and stored over more than a single platform. Older, historical data seldom needs the same level of access as more recent data and may well be moved to cheaper or slower storage.

No matter how you look at it, the data needed by a modern digital business already exists on multiple platforms and will increasingly do so. Business people need to access these multiple platforms transparently from their one favored tool. They also need to be able to combine or join data from these platforms into a single result. The only technique that can do this is data virtualization.

Simply defined, data virtualization takes a single query—be it from a business user or an app—and splits it into its constituent parts, sending the individual parts to different data platforms, receiving and combining the individual results into the required answer. And it does all this in real-time.

The concept is not new. Indeed, it dates to the early 1990s, when it was often limited to relational data on distributed platforms and called data federation. Since then, data vir-



tualization has been extended in scope, function and power until it has become the mature and comprehensive platform offered today by Denodo as seen in the accompanying diagram, which shows the scope and extensive functionality of the Denodo platform.

However, rather than work my way through this diagram, I believe it will be far more informative for readers to examine data virtualization from the point of view of the three most common use cases that apply to data warehousing and related approaches to delivering insight to business people. I'll dive deeper into these cases over the next three posts but first: a sampler here of where we are going.

One of the longest standing use cases for data virtualization in data warehousing relates to data freshness. If you need up-to-the-minute data as part of a query, copying it from an operational source via ETL (extract-transform-load) technology into the warehouse is unlikely to offer the timeliness needed by the business. In my next post, I show how data virtualization goes straight to the source for real-time results.

Another timeliness-related use case refers to development and maintenance time. Data warehouse projects traditionally have a poor reputation for rapid delivery of results, whether for the initial build, extensions of scope, or even ongoing maintenance. Solutions have ranged from narrowly scoped data marts to the now ubiquitous data lakes. In the third post of this series, I look at how data virtualization offers new levels of agility in design, development and maintenance of systems consisting of a mix of warehouse, marts and data lakes.

Over the past decade, businesses have faced a huge influx of data in a wide variety of structures—ranging from audio and video through free-form text to hierarchical and graphical data. What these structures have in common is their limited suitability to relational databases and their consequent management in many different stores. In the final post of the series, I examine how data virtualization allows us to mix and match best-of-breed storage technologies with business people's needs to see and manipulate all these stores as if they were one.

Gaining Real-time Insight

January 10, 2019

http://bit.ly/2USNmIk

Data virtualization offers an elegant solution to the challenge of allowing users to access real-time data from operational systems combined with reconciled data stored in a traditional data warehouse.

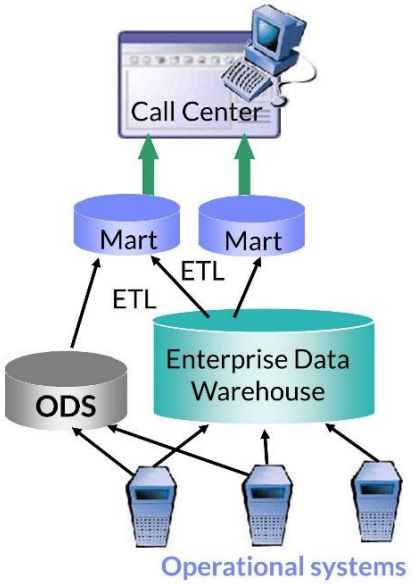
Allow me to introduce the (fictitious) Advanced Banking Corporation, or ABC for short. It may be advanced in name, but its IT systems are anything but. ABC will be our constant companion as we explore all three scenarios for data virtualization.

Alice Well, new CIO since last Monday, has heard nothing but bad news from or about IT. They are a roadblock to every new development and when they get started it takes twice as long and costs 50% more than planned. Existing systems are old and slow; decision makers must make do with yesterday's data—if they're lucky, and the nightly batch jobs all worked. If that wasn't enough, ABC has just announced the acquisition of a small, nimble rival, who use a Hadoop-based data lake for decision support, and which, Alice knows, is largely incompatible with ABC's legacy environment. She has been repeating her personal mantra "All is well, all will be well" almost non-stop since her she arrived.

Alice's first challenge has come from the very top. Unsurprisingly, it's labeled "Urgent".

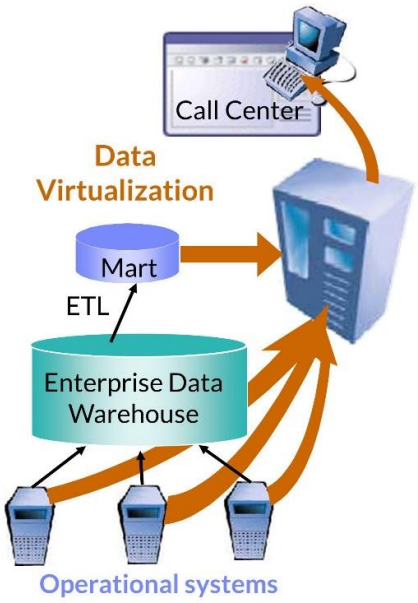
"Why," asks CEO, Bill Costly, "do agents in the new Call Center never have a customer's most recent transactions to hand when taking a call?"

The answer, as Alice easily discovers, is that the system used by ABC's Call Center was not designed from the ground up to deliver real-time data. As seen in the figure below, ABC's data warehouse is a very traditional design, fed by batch ETL (extract, transform and load) tools from the operational checking, credit, and client management systems. The checking and credit extracts run nightly, but data from the client management system—a nasty mix of ancient code supplemented by spreadsheets—is only updated weekly, after manual checking by IT.



In the mid-1990s, ABC adopted best practice of the time and built an operational data store (ODS) in an attempt to improve the timeliness of data provided to business users. This ODS is a largely unmodified store of data from the checking and credit systems, loaded hourly. Including the client management system in the ODS proved beyond the skills and budget of IT. Because of the mismatch of timings between these data sources to the warehouse and ODS, call center agents struggle to answer all but the simplest of client enquiries.

Fortunately, Alice has recently attended an event organized by Denodo and a well-known consultant who showed and discussed the very pictures shown here!



The new solution, based on data virtualization technology, is shown here.

Whenever an agent receives a client call, a query is sent from the call center app to the data virtualization (DV) server, which retrieves in real-time the historical, integrated data about the client from the data warehouse via a previously developed mart, today's client transactions up-to-the-second from the operational checking and credit systems, as well as the most recent client status from the "spreadsheet enhancement" to client management system. The DV server joins together the data from the various systems and returns it to the agent.

Despite the simplicity and elegance of the solution, Alice will have to work hard to sell it to the BI manager, Mitch Adoo, whose nickname, "About Nothing" reflects his strictly traditional approach to data warehousing. Among Mitch's objections are data consistency, network loading, and the impact on performance of the operational systems. Each argument has some validity, of course, but Alice has good answers, as well as holding a trump card, the CEO's memo demanding urgent action.

It is true that more care is required when joining data in real-time than in batch where potential problems can be anticipated, and solutions applied in retrospect. Alice is aware of these issues but is making a very deliberate trade-off between immediacy of insight against guaranteed data consistency. Here, the trade-off is justified by the improved customer satisfaction of the majority of customers. The quality of the spreadsheet-based client management data is a major concern, but Alice is gambling that errors here will be of less concern to clients. And maybe she can use the savings from virtualization to offset some investment in the base system problems.

She is also injecting added agility into the IT department. We'll return to this aspect in greater depth in a later article. In this case, the ability of the IT department to deliver real-time data to business users based entirely on existing systems has the important side benefit of proving that the BI department is willing to try something "new"!

Oddly enough, novelty in this case is largely a market illusion. Data virtualization is now a mature technology, having been around in different forms since the early 1990s. However, like Mitch, the market has been largely unaware of the technical advances the products have made. In relation to network load problems, Denodo, for example, uses multiple techniques to minimize the problem, including local caching of data and query optimization that takes the data volumes that need to be transferred between servers into account.

Regarding impact on operational systems' performance, it certainly can be a problem if these systems are heavily loaded. However, it should be noted that the DV server is sending very specific queries to these sources, using fields that should normally be indexed, and getting results limited to a small number of rows. With such specific querying of operational sources, data virtualization in fact offers a far more efficient solution to joining current and historical data than the ODS approach.

And so, this is how Alice wins the first battle in her war to drag ABC's IT systems into the 21st century. In my next post, we'll see her set sail on that newly acquired data lake.

A Warehouse in a Lake, Data Virtually

February 6, 2019

http://bit.ly/2XFfNXd

Data lakes have in recent years become the preferred approach to storing data, especially that coming from voluminous external sources. However, there is a continuing need for a traditional data warehouse. Data virtualization supports both approaches and offers users consolidated access to both.

Fresh from her success in supplying real-time transaction data to the call center using the Denodo Data Virtualization Platform, Alice Well, recently appointed CIO of Advanced Banking Corporation (ABC), hears the three familiar, demanding raps on her office door. She glances toward the heavens, whispering "all is well, all will be well."

"Why are you waging war on *my* warehouse?" BI manager, Mitch Adoo (About Nothing) barges into the room. "Why are you promoting a *data lake* from that upstart acquisition over all careful work we've put into the warehouse? You know they just keep dumping data in there until it has become a stinking swamp."

"Slow down, Mitch," says Alice. "I'm giving you the opportunity to focus the warehouse on what it does best: the creation and management of a consistent set of core business information for the entire corporation. Meanwhile, the data lake can handle all the other, dirty, less critical data. The warehouse will be a key component in allowing business and IT folks to get to it in a well-managed and understandable way. Your work will be even more important for ABC."

Mitch is now smiling broadly. That's more like it!

"Read my architecture document," Alice adds. "It's based on our new data virtualization platform."

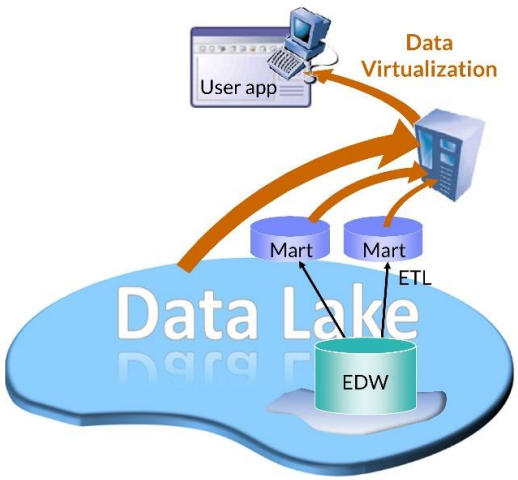
The Old Way: Lake Displaces Warehouse

Since its initial introduction in 2010 by James Dixon, CTO of Pentaho, the data lake has grown in popularity as the place to store all data for analytic use. Its open source technology foundation offers advantages in cost, agility in development, speed in loading, and ease of use. This is because of its "schema-on-read" approach, which allows data to be stored in its original form rather than having to be structured into a relational model on loading (schema-on-write).

The above advantages are of most benefit for externally sourced information. However, the cost argument, in particular, has led many organizations to plan to eliminate their warehouses, although they were built mainly for well-structured, internally sourced data. Issues arise, however, in this elimination plan. Existing investment must be written off and current staff skills are devalued. Data management in data lakes remains primitive in comparison to the relational environment, a real concern in the case of financially and legally binding data.

Alice, like many thoughtful and experienced CIOs, has concluded that a better approach is to keep the data warehouse with a more focused role while digging and filling a new data lake.

The New Way: A Warehouse on an Island in a Lake



The phrase "warehouse on an island in a lake" suggests that the data warehouse is an island of stability, structure and management in an otherwise fluid and changeable data lake. The warehouse becomes both a store of core business information and a contextual reference point for all the rest of the data in the lake. The warehouse retains its relational structure, while the lake uses whatever set of technologies is appropriate.

However, business people must be able to access and use all this data, irrespective of its storage technology. They must be able to "join" data across the warehouse and lake with ease and elegance. This, of course, is where data virtualization technology is vital in implementing the approach, as shown below.

The warehouse, consisting of an enterprise data warehouse (EDW) and data marts, remains as it was. However, further growth is first reined in and its use for advanced analytics and externally sourced information may be restricted. Non-core data may be later removed. The EDW serves two purposes. First, it is the certified and reconciled source for data from the operational environment, provided to business people via data marts in the usual fashion. Second, for the data lake, it acts as reference data and metadata in support of the wide variety of more loosely structured and less well-defined data there.

Data virtualization mediates access by business people and apps to all data in the lake and data marts (and also the EDW, if required). In the Denodo platform, such mediation is built on an extended relational model and supported by a comprehensive and dynamic Data Catalog. The model of an existing EDW acts as a foundation for the extended relational model of the broader data lake.

The Denodo Data Catalog is a dynamic catalog of curated, reusable and timely information about the content and context all data available through data virtualization. Its dynamic nature arises from its direct linkage with the data delivery infrastructure, thus ensuring that its contents are always current. With the EDW data and metadata available from its island within the data lake, data stewards and business users alike can be assured that, not only is the catalog up-to-the-minute, but contains the reconciled and agreed definition of the business' core data.

Leveraging powerful searching and browsing functionality of the Denodo platform, business people can seamlessly move between metadata and data wherever they reside, with no need to understand or construct complex SQL queries. Data stewards and administrators also benefit from the Catalog with real-time metrics of data usage, timeliness and quality in the data lake, oriented around the key business concepts and structures found in the EDW.

As Alice says, all is well, and all will be well when IT can reuse and repurpose the high-quality design work that went into the data warehouse to support new platforms such as

the data lake. Such reuse is only possible with a powerful data virtualization platform built around a dynamic Data Catalog such as that offered by Denodo.

More Structured or Less, Data Virtualization Delivers

April 17, 2019

http://bit.ly/2KPTT1Y

Applications, such as campaign management, requiring access to a mix of both more and less structured data are ideally supported by data virtualization that offers consolidated access to data in multiple formats.

Alice Well's early successes as CIO at Advanced Banking Corporation (ABC) in solving the old problem of getting real-time data to the call center and the newer opportunity presented by the data lake have meant that she now has significant influence with the board on how to implement digital business at the bank. Modestly, she attributes a large part of that success to her early discovery and adoption of Denodo's Data Virtualization Platform.

It's also fair to say that another key contributor to her success is her understanding of the organizational implications of major changes in the IT infrastructure. Now, with the pressure building at board level for significant and visible advances in delivery of digital transformation, all of Alice's organizational skills and technological nous will be needed.

Never one to shirk a challenge, Alice has determined that a central precept for a digital business would be to provide a single, fully integrated, contextually sensitive interface to all ABC's multitudinous data sources for all the business. As she commits her new annual objectives to CEO, Bill Costly, she reflexively turns to her by now well-known mantra: "all is well, all will be well."

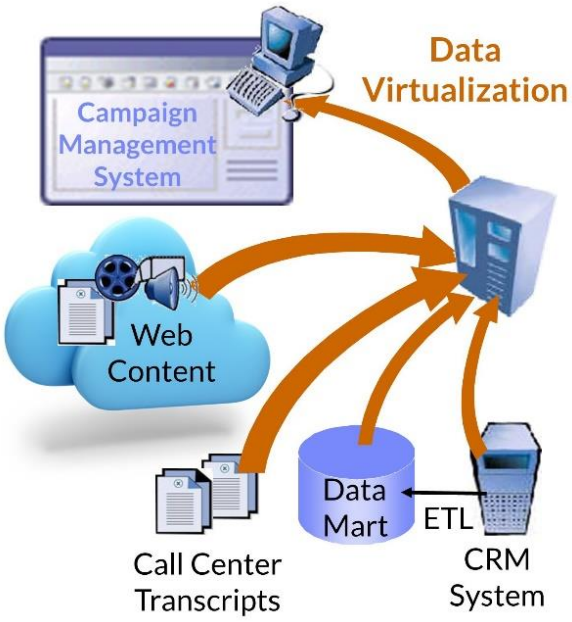
As with every large business, ABC's slate of data sources has grown exponentially over recent years. Not only has the variety of sources of numerical, well-structured data—SQL, NoSQL, and spreadsheets—grown rapidly. In addition, the business has developed a huge appetite for all types of more loosely structured data—everything from text files and e-mails in content management systems to web pages and sources, such as LinkedIn and Facebook.

Across the organization, the clamor is nearly deafening as every function competes for access to their favorite unique and usually urgently needed source. With the demand for an enterprise-wide data virtualization infrastructure growing, Alice takes a stand. She well knows that infrastructure projects claiming to "immediately benefit everybody" seldom deliver on their promise. What is needed is a focused project with well-bounded goals that delivers early value to one part of the business and can be rolled out widely afterwards. When CMO, Mark Ed Price, comes looking to invest in a new campaign management solution, Alice knows that she has found her perfect project to prove the value of data virtualization across the widest variety of sources.

Curing Campaign Management Misery

ABC has had an on-premises Customer Relationship Management system supported by a reporting data mart for some years now. The CMO would like to have better Campaign Management, with access to customer conversations from the ABC call center, as well

as sentiment from Facebook and Twitter feeds. Buying a new, integrated Campaign Management System (CMS) has been costed as far too expensive, especially when all the data consolidation and transfer work from existing systems is included. What is an embattled CMO to do?



Data virtualization to the rescue! As shown here, ABC's existing Denodo Data Virtualization Platform allows data from all the required platforms to be accessed and consolidated into a bespoke CMS application.

One of the long-standing strengths of Denodo Data Virtualization has been the wide range of different types of sources and systems from which it can obtain data.

Traditionally, data virtualization products focus on well-structured data sources. Relational databases (RDBs) are the most obvious example, and Denodo supports a comprehensive set of traditional RDBs, as well as cloud-based systems (such as Amazon Redshift and Snowflake), appliances (e.g. HP Vertica and IBM Netezza), and open sources systems (like PostgreSQL and Amazon Derby). With support for NoSQL stores, hierarchical databases (such as IMS and Adabas), CSV (comma separated variable) files and spreadsheets, Denodo clearly covers all the bases for well-structured data.

Turning to text sources, Denodo supports standard file formats, such as pdf, Word documents, text, and e-mails from file systems and content management systems. Content can also be accessed from search systems, web pages, web apps (such as Amazon, Facebook, LinkedIn, Twitter, etc), and even semantic repositories. And there's more sources to be found in the Denodo documentation.

Build vs. Buy

Alice's solution for Campaign Management builds on the power of data virtualization to access data from all the existing data sources in ABC. However, it clearly also involves the building of a new application, with all its attendant costs. Why not just buy an off-the-shelf CMS?

When building a new business, or when a business has been built around a single (often cloud-based) system such as Salesforce, an off-the-shelf CMS is often a good choice. However, in a case like ABC, with a plethora of existing systems that would need to feed the new CMS or be replaced by it, data virtualization can be a very cost-effective approach.

A further advantage is that data virtualization offers an agile development approach, allowing data sources to be added in stages and tested in realistic scenarios. With the Denodo's Dynamic Data Catalog, business people can locate additional data sources not previously included in the CMS scope, and with Dynamic Query Optimization, can be assured of well-performing access.

And They All Lived Happily Ever After…

With this, we reach the end of Alice's story. With Denodo Data Virtualization Platform, she has gifted ABC with an agile, performant infrastructure for direct access to data on a wide variety of platforms. The infrastructure can be used to support general business intelligence use for real-time data and external data sources, as well as enabling specific applications. These systems can be built in stages and modified with ease in an agile development environment. Business people can find what they need and discover data they might never have known would be useful to them.

As Alice Well is well known for saying: all is well, and all will be well* with data virtualization.

** Alice's mantra is based on the passage "All shall be well, and all shall be well, and all manner of thing shall be well" from "Revelations of Divine Love" by 14th century anchorite, Julian of Norwich. This is first book in the English language known to have been written by a woman.*

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 Upside, and more, Barry is based in Bristol, UK, and operates worldwide.