Toward frictionless data discovery

IBM Fluid Query eases the way

August 2015

A White Paper by

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

In today's highly distributed, multi-platform world, the data needed to solve any particular decision making need is increasingly likely to be found across a wide variety of sources. As a result, traditional manual approaches requiring prior collection, storage and integration of extensive sets of data in the analyst's preferred exploration environment are becoming less useful. Data virtualization, which offers transparent access to distributed, diverse data sources, offers a valuable alternative approach in these circumstances.

This paper describes the problems of multi-source data discovery through the eyes of a typical business analyst, highlighting the difficulties encountered when using traditional manual methods of prior data integration. A modern, high-level architecture—the integrated information platform—is introduced that positions and explains data virtualization. The value of this technology in delivering business insight from data is then presented.

Finally, we explore a new offering, IBM Fluid Query, and the set of data virtualization functionality it offers through IBM PureData System for Analytics, IBM BigInsights and other IBM data management products. This initial offering provides valuable functions and also shows the direction of IBM thinking in this emerging area of data management function and value enablement.

Contents

Sponsored by:
International Business Machines

www.ibm.com

Data scientist is the sexiest job[1]…", muttered Raymond as he laboriously set up yet another extraction of updated and anonymized client data from the data warehouse for transfer to his sandbox environment for analysis in combination with the latest social media data. "Really? Feels more like data plumber." The SQL error code, row limit exceeded, popped up on the screen five minutes later. Raymond swore: "Or flipping sanitation engineer", recalling the cleverly named Drain Brain truck he'd seen that morning. And copying data from Hadoop was no easier; and the error messages were even more obscure. "Why does it all have to be so hard?"

Increasingly, it doesn't. But avoiding the difficulties Raymond encountered, and more, requires rethinking some old paradigms. In a world where data volumes are enormous, moving the data to the query—as Raymond has been doing for years—makes less sense. Many times we need to move the query to the data. In a world where data varieties are prolific, thinking about a data warehouse as a single, physical store of *all* the business information is becoming untenable. As is any strategy that suggests that one and only one platform can store and manage all data. We now need to envisage an architecture with multiple platforms (or pillars, as I will describe later). And we must have the means to completely hide the existence of such multiple locations from the user, or seamlessly combine data results from multiple sources. Furthermore, we need an infrastructure that manages data and creates information context across these platforms, allowing partial data duplicates—temporarily or permanently—to be populated in the background as required.

We will consider the shift from an architectural perspective. But first, let's understand what Raymond is attempting at BigLocal Underwriting Experts, or BLUE for short.

The world of multi-source data discovery

Information has long been central to insurance companies; they were among the earliest adopters of data warehouses to collect, manage and utilize their production data for reporting, risk analysis, prediction and more. BLUE's data came from multiple sources and had to be cleansed and consolidated before use. Sometimes it was a bit slow in arriving. But at least it originated from their own internal business processes, hence the term *process-mediated data**, as I refer to it. When Raymond first began analyzing risk (before data scientists were even dreamt of), he worked exclusively with this type of data, mostly in the warehouse or in specialized data marts, and often in spreadsheets. He became expert at *ad hoc* data integration—preparing, cleansing and copying the data he needed into his tools of choice. His close relationships with the owners and builders of BLUE's many operational systems allowed him to easily—relatively speaking—understand and gain access to the data he needed.

Today, Raymond's sources of data are far broader and richer than he ever expected. Behavioral and relationship *human-sourced information* comes in large and growing volumes from social media and data aggregator sources. It is already piling up in the new Hadoop environment. Of course, analysis of this external data alone already offers valuable insights into social behavior and economic trends. BLUE has already used this data to revamp the risk evaluation process for younger prospects. The business has achieved high value from the work, although, in Raymond's view, the tools he had to use could have

* Descriptions of this term, as well as human-sourced information and machine-generated data can be found in my 2012 white paper for IBM *"The Big Data Zoo—Taming the Beasts"*, http://bit.ly/Big_Data_Zoo

been a bit more user-friendly. And, it is already clear that the real wins will come when external data is used with internal process-mediated data. Raymond has begun the initial tests of how customer data can be safely and securely merged with the incoming human-sourced information. He has moved data in both directions between the warehouse and Hadoop. But already, he is struggling to cope. His old, manual methods of managing the multiple copies he is creating are inadequate to the challenge of combining large volumes of high-velocity, but low quality external data with well-governed but highly sensitive internal data. And as for understanding the meaning of the new data… well, that's another story again.

In analytics, the real wins come when external data is used in conjunction with internal process-mediated data.

In the coming few years, Raymond is already anticipating vast swathes of *machine-generated data* from the Internet of Things as the business of automobile insurance is reinvented. What little he has seen of the expected speed and size of this data looks very daunting indeed. He has heard that yet another technology platform may be needed to store and manage it.

It is becoming increasingly clear to Raymond—and, indeed, most data scientists and experts in the world of analytics—that the tried and trusted ways of discovering insights from this ever-growing set of data sources, data types, and expanding volume of data is going to be inadequate. Data discovery on this scale cannot be a cottage industry. New architectural thinking and new technological solutions are required to reap the rich rewards promised by analytics across multiple data sources.

Beyond classical data warehousing—an integrated information platform

The original data warehouse architecture[2] dating back to the mid-1980s was largely driven by the needs of management and supervisory users for a consistent and accurate view of the business, including both performance reporting and problem investigation through query and analysis. Technology constraints of the time, and for many years after, dictated a largely centralized and layered architecture. All required data is funneled from multiple operational systems through an enterprise data warehouse (EDW) for cleansing and integration and, in most cases, delivered to data marts that are optimized for the query and reporting needs of business users. This approach has remained a staple of the business intelligence (BI) community for decades despite ongoing shifts in business needs toward timeliness and breadth of information rather than consistency.

However, recent years have seen an enormous explosion of so-called big data from social media and, now, the Internet of Things. This has driven dramatic shifts in the business perception of how value can be obtained from information through data discovery and analytical approaches. Add to this some impressive advances in technological capabilities—networking, hardware and software—and the foundations of the traditional data warehouse architecture must be reexamined. We have seen this trend clearly in the promotion of concepts such as logical data warehouses, data lakes and data reservoirs in the past five years or so. Unfortunately, some of these terms are more marketing concepts than well-defined architectures and fail to fully consider all emerging business needs, technological possibilities or vital aspects such as information quality, context, meaning, and emerging thinking about how decisions are really made by people in organizations.

The emergence of big data, together with impressive advances in technology, necessitate a reexamination of the traditional data warehouse architecture.

A full architectural exploration of all these considerations leads to a move from the layered data, single platform approach of the data warehouse to the new, pillared data, multi-platform approach described in depth in my recent book *"Business unIntelligence"*[3]. This seemingly simple—but technically complex—

shift in perspective is shown in Figure 1, which shows three pillars of data with different management and processing needs, supported by context-setting information (or metadata) that spans the pillars. Depending on business and technical circumstances, there may be more than three pillars; there is unlikely to be less.

The first and central core business data and reporting pillar is the consistent, quality-assured data found in the EDW and data marts. In terms of technology, this pillar is ideally based on relational database technology. Depending on business needs for higher query performance, in-memory, massively parallel processing (MPP), columnar databases or other specialized technologies, may apply. Example systems include IBM DB2 with BLU Acceleration, IBM PureData System for Analytics, dashDB and Spark. Deep analytic information requires highly flexible, large scale processing such as the predictive analytics and text mining often performed in the Hadoop or IBM BigInsights environment. Fast analytic data requires high-speed analytic processing that must be done in-flight, such as with IBM InfoSphere Streams, for example. At the intersection of speed and flexibility, we may have specialty analytic data, using specialized processing such as NoSQL, XML, graph and other databases and data stores.

This pillared approach answers modern business requirements for timeliness of data and access to a wide array of data and information types. However, it also presents technological challenges in ensuring data consistency across the pillars, as well as how users can seamlessly access data dispersed over the pillars. Let's now turn our attention to the data integration and virtualization components.

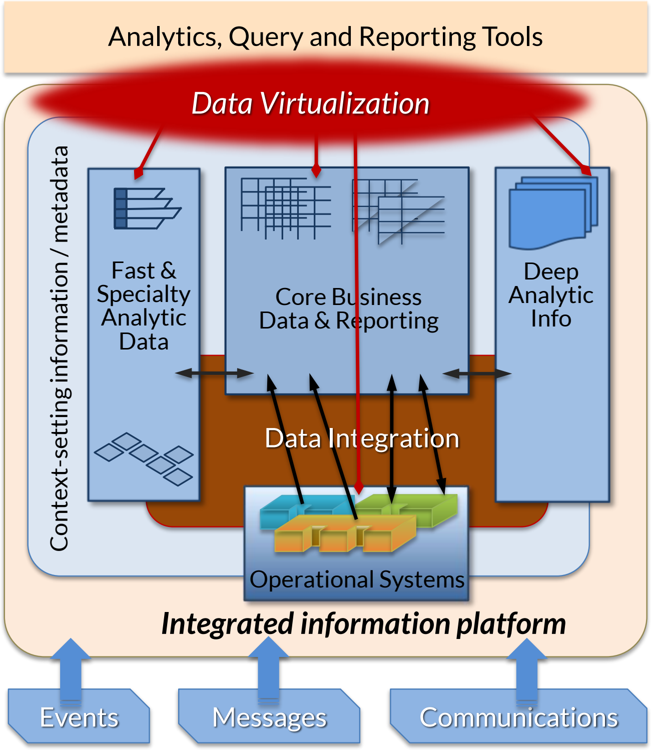
Analytics, Query and Reporting Tools
Data Virtualization
Context-setting information / metadata
Fast & Specialty Analytic Data
Core Business Data & Reporting
Deep Analytic Info
Data Integration
Operational Systems
Integrated information platform
Events
Messages
Communications

***Figure 1:** The modern "pillared" architecture*

From data to information and thence to insight

With data now dispersed across platforms for data volume, type and performance considerations, users are faced with the issue of how to physically access the data, combine it in a variety of information artifacts and derive useful, valuable business insight from it. These considerations are at the heart of data integration and data virtualization. These technologies have, of course, been in existence in various forms and levels of maturity since the earliest days of data warehousing. Nonetheless, the integrated information platform demands much greater sophistication in both areas.

Data virtualization—offering a single point of access

The basic premise of data virtualization is that business users need never be aware where or in which form the underlying data resides. Simply put, hiding that complexity is the responsibility of data virtualization. The required function can be characterized in three levels of increasing business value:

1. Functionality to route all or part of a query from one platform to another in a form usable by the second platform, to receive results back, combining with local results if necessary

2. The ability to use cross-platform technical metadata to build and optimize the performance of queries in a multi-platform environment through views

3. Provide linkage from a logical, business view or model of business information to the technical view through the metadata that describes where and how data resides on the different platforms

Data virtualization is a maturing technology, coming from two directions. In one approach, specialized data virtualization vendors start from the technical or business model functionality and build a platform that spans and sits above individual databases and data stores maximizing the number and variety of databases and stores accessible. This offers good generality of expression and avoids lock-in to one database vendor. In the other approach, database vendors create cross-platform query routing and execution function within their own environment, usually focused on and optimized for a subset of platforms of interest. This approach reuses much of the vendors' database function, and eliminates the need for an additional server layer with its overhead for all queries. It is particularly attractive when the majority of queries come through this platform.

Data virtualization is now an accepted and maturing technology that is vital in a multi-platform world.

Data integration—building a cohesive information platform

In a classical data warehouse environment, the primary focus of data integration is on gathering, preparing and loading data from diverse sources into the EDW, and from there to data marts. ETL (extract, transform and load) and a variety of specialized tools, operating from real-time to batch, are seen as the primary focus in data integration. However, as in virtualization, we can distinguish three levels of increasing business value:

1. Specific source-target data movement, including multiple sources and targets, handling localized cleansing and combination tasks

2. Data integration platforms that offer a shared ETL environment including common ETL metadata, user interface, etc., often running on dedicated servers or on Hadoop

3. Fully automated development and management environments that bridge from user requirements and data models all the way to ETL function and use

Data integration is a mature technology in many respects, with different vendors offering a wide variety of solutions at all three levels.

Joining virtualization and integration into insight—context is the key

A little lateral thinking shows that data integration and virtualization are two sides of the one coin. They both aim to provide data from remote and often multiple sources to the user needing to query or otherwise use it. Data integration, the traditional way of doing this, copies all the data that may be required to a common destination, in anticipation and in advance of the users' needs. Originally done via bulk transfers, trickle feeding is becoming more common. Data virtualization, on the other hand, waits until the user requests some data and then tries to transfer the minimum amount needed during the course of the query. Virtualization is thus more circumscribed by users' immediate needs for timeliness of results, although integration is increasingly subject to the same constraints as most businesses move toward real-time operations. There is a functional symmetry to be seen here. It implies a need for common and shared infrastructure in order to ensure that users see compatible results irrespective of which technology is used for a particular business need, in an environment where both technologies coexist.

Of particular importance is that the underlying metadata is shared by both tool sets. This metadata covers both business and technical aspects of the data stored and the transformations it underwent. At its broadest extent, it describes the entire context of the data, including description, meaning, lineage, quality, valid values and uses, etc., enabling its use as true business information. In reality, metadata is too limited in perception to cover all these needs. This material is actually a key component of the business information itself and, as a result, I prefer the term *context-setting information (CSI)*. Only with such comprehensive context can users gain real business insights. And only when CSI is fully shared between integration and virtualization, can these insights be guaranteed to be valid and consistent.

Context-setting information provides the foundation for a consolidated virtualization / integration environment that delivers true insight to business users.

The integrated information platform of Figure 1 requires sophisticated tooling in virtualization, integration and context-setting information, as well as significant levels of integration between these functional aspects. We may already hear Raymond, in his inimitable style, responding: "Nice architecture, but what about some real help here? And I need it right now…"

IBM Fluid Query—modern virtualization and integration

Beginning with the announcement of IBM Fluid Query in March 2015 and a July refresh to version 1.5[†], IBM began to answer Raymond's plea for help. Fluid Query offers two distinct vantage points from which to address a multi-platform world. The first vantage point is that of classical power users, such as Raymond, whose history and skills lead them to start their data work from the relational environment—in this case, IBM PureData System for Analytics—and link from there to other data sources. The second vantage point is that of more recent data scientists, who work mainly from the Hadoop environment—here, IBM BigInsights—and look from there to the relational data warehouse and other platforms. We examine these two approaches separately, because they address different use cases and offer slightly different functionality. But first, let's look at some typical use cases.

Business use cases for modern virtualization and integration

1. Light-speed business: The business analyst, working in the data warehouse environment, is examining the downward trend in sales. Data in the warehouse (in this case) is a snapshot of the business as of end of yesterday, which is how most users want to see it or, maybe due to the time it takes to consolidate and cleanse it. The analysis (using yesterday's data) leads to a problem determination and remedial action is taken. Ideally, the analyst would now like to instantly see the results of the action. Querying real-time data from the operational environment and combining the result with the standard end-of-day data warehouse data allows the analyst to see the up-to-the-minute trend of sales when needed, without adding real-time loading capabilities and costs to the data warehouse.

2. Bridging relational data islands: Most enterprises run multiple relational databases, sometimes dictated by application needs, other times through historical developments. That should not mean that data in one system is inaccessible from another. Data virtualization offers the ability to bridge to another relational database system to access and combine the data required. By pulling more

[†] Functions described in this paper are available in IBM Fluid Query version 1.5.

remote data into the PDA (or BigInsights or DB2) environment, it is also possible to bridge to multiple RDBMS in the same query, provided possible performance impacts are understood.

3. Creating deep context: Numbers from the data warehouse alone don't always tell the full story. Consider the query: "What are our top selling products that get good or better reviews?" This query seems reasonable to a business user, but it contains both process-mediated data and human-sourced information: *top selling* comes from sorting sales volumes in the data warehouse or mart, while *good / better reviews* requires analysis of social media and other data landed in Hadoop. In many cases, from customer relationship management to call center support, textual (and sometimes image) information creates a much deeper context around the numbers for business. This type of analysis allows SQL queries originating in the relational environment to access the big data landed in Hadoop stores pushing processing down to MapReduce and taking advantage of the power of the applications that reside there, such as pattern recognition, predictive analytics, etc.

4. Agile activities: Who has time to build a new data mart every time the business needs expand or change? Sometimes you just need an extra detail or two on an existing data mart. Or maybe you simply want to check if adding a new column would be valuable. The ability to join data from one mart to the data that resides in another in a simple query can give quick answers, either to discover if the result is useful or to quickly satisfy the new need without requiring an IT project to create a new mart. Agility to react to new needs or market changes is vital in today's business.

Business needs for timeliness, contextualization, agility and historical support drive the adoption of virtualization and integration functionality.

5. Historical support: In the big data era, it makes sense to regularly move old, less frequently used data to a cheaper environment. But just because it's less popular, it doesn't mean the business will be happy if takes hours to retrieve if needed. Moving older data regularly to Hadoop, but allowing access through a virtualized query from the data warehouse offers the best of both worlds—lower storage costs with seamless user access to the data when needed.

Vantage point I: Operate and Oversee

There's hardly a business left in the world that doesn't use a data warehouse to oversee their operations with reporting and query tools. In truth, many organizations use more than one—despite my and others' best advice. The outcome is that many decision-making needs require data from multiple sources. In addition, many traditional data warehouses contain data that is a snapshot of a point in time, typically close of business yesterday. So, what happens when a user of one data mart needs data that only resides in another warehouse or mart, or when more timely operational data is needed to answer a business question?

Two solutions have been available until now: (1) download data from the multiple sources and proceed in Excel, or (2) go to IT and ask them to build a new mart. Like many business analysts, Raymond preferred the former approach, having experienced some heartbreak with IT delivery delays. With new data appearing in Hadoop and NoSQL systems, the challenge is only increasing. IBM Fluid Query offers a third option: accessing the required remote data directly as part of a SQL query.

Business operation and oversight occurs in a traditional, relational environment that is becoming increasingly diverse, demanding virtualization solutions.

With this option, the user's view of the world is centered on the data residing on IBM's relational database systems such as IBM PureData System for Analytics (PDA), DB2, etc. Using standard SQL syntax, the scope of the query can be extended to include data on other platforms, both relational and non-relational. In the relational world, this includes other PureData System for Analytics implementations, IBM DB2, dashDB, PureData System for Operational Analytics and Oracle. Non-relational targets include IBM BigInsights, Hortonworks, Cloudera and Spark.

The relational platforms and IBM BigInsights can return the results of queries sent to them and these results are joined in the originating PDA database. A further extension could also allow caching (or snapshotting) of commonly used data locally to avoid transferring data and rebuilding the required data set multiple times. Hortonworks, Cloudera and Spark, on the other hand, may return unqualified and potentially larger data sets if they are unable to process the appropriate SQL locally. This leaves the central PDA system to do the qualification function before the join. This can result in larger data transfer volumes across the network and additional load on the originating machine. However, as the SQL functionality of Hadoop systems improves, increasing amounts of the work can be pushed down to them, improving overall system performance.

Vantage point 2: Optimize and Outlook

Listening to the IT press and analyst chorus, you might be forgiven for thinking that the entire industry is focused entirely on predictive and other analytics to outlook future markets and behaviors and optimize operational systems accordingly. You might imagine that the entire universe is moving lock, stock and barrel to Hadoop and related platforms. This, of course, is untrue. In fact, it will never be true—migration costs would be prohibitive and, anyway, given a few years, another software messiah will likely appear. Nonetheless, significant new, mainly external, data sources are being implemented on Hadoop. And some emerging companies with no legacy systems are experimenting with running their businesses there. Undeniably, the Hadoop platform is today the home ground of data scientists (and unicorns). In such cases, analysis and reporting begins here in the Hadoop environment, and tools to do this are improving and expanding apace.

The vast majority of businesses continue to operate and oversee their operations in traditional relational environment, while the data scientists focus their work in the Hadoop environment. Nonetheless, they also certainly need to combine data from existing relational systems with the results of queries in the big data environment. Thus, a data scientist working in IBM BigInsights can access data in tables on a PureData System for Analytics (or DB2, PDOA, dashDB, etc.) system directly from the analysis run on Hadoop. With BigInsights, Fluid Query offers the ability to return data subsets as qualified by the query. For a data scientist working on Hortonworks, Cloudera and Spark, intermediate processing may be required to further qualify the query results. The value in both cases comes from the possibility of linking production information, such as customer numbers or validated data warehouse information, for example, to data from external, lower quality data, such as Twitter handles.

Predictive and other analytics are best performed on combined traditional and big data, demanding unified virtualization and integration solutions.

In either case, IBM Fluid Query can move full data sets from one platform to the other, enabling the creation and management of historical data stores on Hadoop or data caches on the PureData System for Analytics platform. And given the progress seen in data sources/targets and functionality seen in the short time between initial release and version 1.5, it is reasonable to expect that IBM will continue to expand both aspects going forward.

Conclusions

Although still in the early stages of evolution, IBM Fluid Query is setting a direction that is already beginning to meet Raymond's needs for a better way of working with data from multiple sources, both traditional internal data and the burgeoning wealth of external data. Fluid Query is, in this sense, both a program of functionality required to support an integrated information platform and the products that deliver that function in stages. And not only the needs of more traditional business analysts like Raymond, but also of the emerging breed of data scientists who do much of their work on Hadoop platforms.

Decision making support is today in transition from an environment where all the data was brought to and used in a relational data warehouse to a new, multi-platform world, sometimes called a logical data warehouse. In this modern world, data volumes and technology strengths/weaknesses mean that function must increasingly be moved to the data, wherever it resides, rather than *vice versa*. Such data virtualization and integration will increase in importance and popularity as the complexity and size of the data environment grows. IBM Fluid Query offers a starting point on this journey and provides the signpost toward an integrated information platform that drives extensive and valuable insights from all data managed and collected by the business.

As we move to a multi-platform, data rich world, IBM Fluid Query offers a good starting point and useful direction in the adoption of virtualization and integration in BI.

*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, writer, blogger and contributor to the development of the information industry, Barry is based in Cape Town, South Africa and operates worldwide.

[1] Davenport, T.H. and Patil D.J., *"Data Scientist: The Sexiest Job of the 21st Century"*, Harvard Business Review, (October 2012), https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[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/EBIS88

[3] Devlin, Barry, *"Business unIntelligence—Insight and Intuition Beyond Analytics and Big Data"*, Technics Publications, New Jersey, (2013), http://bit.ly/BunI-TP1