9sight Consulting

Collaborative Analytics

Sharing and Harvesting Analytic Insights across the Business

June 2009

A White Paper by

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

Business analysts are, by tradition, hunter-gatherers. Independently or in small, close-knit groups, they stalk the wild data resources of the business, seeking out new and unusual facts, and building from them deep insights into the meaning of business life and events. Armed with little more than their spreadsheets, they single-handedly recalculate cells and pivot tables in search of that "ah-ha" moment when innovation emerges from its lair.

Meanwhile, IT labors hard, and often at great cost, to provide a quality-assured, comprehensive and integrated warehouse of information as the basis for corporate reporting and planning. This valuable data resource often lies ignored or distrusted by pioneering business analysts. Today's business cannot afford the luxury of this disconnect.

This paper examines the sources of this unfortunate division and introduces the adaptive information cycle, a model that links the center-out approach of traditional data warehousing to the edge-based, emergent prototyping that characterizes today's analytic environment. By combining concepts from integrated development environments, social networking and collaborative working, the adaptive information cycle reunites Business Intelligence and Business Analytics.

Beyond the conceptual level, Lyza™ Commons shows the real functionality needed to allow business analysts to collaborate more closely and to enable IT to harvest the fruits of their innovation for the wider user community.

Contents

Working on the Information Edge

Emergent Prototyping and the Adaptive Information Cycle

Lyza Commons: First Steps in Collaborative Analytics

Conclusion

In collaboration with:

Lyzasoft, Inc.
www.lyzasoft.com

Joe Figger took one long, last look around his Dilbert-esque cubicle at Bravo Insurance Group Inc. His "data cube" (as he lovingly called it himself) for the last eight and a half years. He recalled with some considerable pride the analyses he had personally crafted here. The Case of the Shrinking Profit. The Mystery of the Missing Customers. Each analysis had been a personal triumph... and a significant business achievement.

Now, Joe had been asked to do something completely different. And sad though he was to give up the individual detective work, it made sense to him. The business had become too big to depend solely on one analytic wizard. Or even a small group of business analysts[1] working independently here in headquarters. The powers that be had given Joe his own office, unfortunately windowless, and had tasked him to come up with a way to share analyses, results and knowledge among a large corps of analysts distributed across various departments.

Joe had some good ideas about how this could be done. He knew how analysts worked and what internal processes they followed. What they would share and what they wouldn't. Those aspects would have to be incorporated in any collaborative approach if it was to have any hope of success. And as for the role of IT... Well, they would have to be followers rather than drivers. Joe smiled; Jill Joy, manager of the Business Intelligence team, would be unlikely to take that lying down. But, Joe had an ace up his sleeve that he had discovered in his last project[2]...

Working on the Information Edge

Traditional approaches to Business Intelligence start with information at the center and work out to the edge. The Data Warehouse was conceived as a centralized repository for information used in decision making, a "*single version of the truth*". Over time, data marts and similar components became the means of distributing this truth outwards, until it finally arrived in the hands of business analysts and other users who could use it with confidence in their deliberations. With the advent of PCs, users loaded this single version of the truth into spreadsheets, manipulating and changing it, and, depending on your point of view, innovating around it or invalidating its truth.

This ***center-out*** conceptual model of information usage has a long history and is often a valid and viable way of working. In some cases, and with particular types of information, it is well nigh mandatory. For example, information about financial results can and should be defined and controlled in the accounting department, and once posted should be immutable except under very limited conditions.

The center-out model posits three distinct and largely sequential steps in the information cycle:

1. **Record:** the basic *data* about business events and their status is documented

2. **Condition:** complete and consistent *information* is derived by cleansing, combining, filtering and enriching the basic data

3. **Utilize:** the information is used by business analysts and others to create real *knowledge* of the past and make informed decisions for the future

Implicit in this model is the assumption that there exists a central authority that knows users' needs and controls the information content and how it is manipulated in steps 1 and 2. Also implicit is the belief in a unidirectional progression from data, through information to knowledge. The model is directly supported and encouraged by the traditional Data Warehouse architecture, and the IT department often perceives it as key to achieving and maintaining data quality and auditability.

However, when we observe users in a real business, a very different behavior pattern can be seen. Rather than acting like civilized people who come to the supermarket to obtain pre-processed information, many business analysts behave more like hunter gatherers who forage for data throughout the business and create their own information stores and personal analyses. There is no concept of central control in this ***edge-based*** model, and the three steps above are tightly linked, often in a cyclical process performed by individual analysts. Aided only by their spreadsheets and similar tools, these users often provide the most innovative insights into business trends and answers.

At this information edge of the organization, analysts work closer to the real business and to its customers and suppliers, harvesting a rich and diverse bounty of timely and sometimes transient data to inform innovative decisions and solutions. Despite their clear value to the business, such users and their spreadsheets have long been ostracized by the IT organization and BI and data quality departments in particular. While the dangers to data quality and consistency of such foraging are clear, business today needs all the innovation it can get, and forward-looking companies recognize that this behavior needs to be encouraged and harnessed within the broad BI environment. In reality, for some classes of information, there are multiple versions of the truth, and the BI environment must support this too.

Characteristic	Center-out	Edge-based
Information provenance	Single, correct and centrally controlled version of truth exists	Multiple and possibly conflicting versions of truth can exist
Information flow	From central store to users	Directly from user to user
Information manipulation by users	Basic data is read-only; users control derived information	Users have full control over all information
Process focus	Reporting and *ad-hoc* performance analysis	Creative exploration of information and business scenarios
Typical tools	BI reporting and query tools	Spreadsheets and similar tools
Data quality & auditability	Can be closely controlled and managed	Open to rapid degradation
Work approach	Hierarchical and standardized	Emergent prototyping and innovation

Table 1: Center-out and edge-based models compared

Perhaps more importantly, the early steps of any new way of analyzing and understanding emerging patterns is a process of ***emergent prototyping***, where multiple visions of the truth are explored. Hypotheses are built and discarded iteratively and repeatedly. Information from diverse sources is combined and filtered. Analyses build upon and modify existing work. This exploratory process is the key innovative stage in the evolution of new solutions. Without emergent prototyping, analysis degenerates to static reporting. But the process appears disruptive to BI professionals who have focused long and hard on the quality and consistency inherent in the center-out approach. Let's look at an example of edge-based working.

Bob works in the Financial Planning department of a Pharmaceutical company, and he uses shipment data from SAP and ledger data for the last few years to develop Average Costs per Unit by Product, including R&D, marketing, distribution, and sales. Sue is responsible for commissions within the Sales department. She integrates sales responsibility assignments from Siebel with invoicing data to track Sales Rep revenue, and to calculate commission payments. But, are the reps rewarded for selling the most profitable lines? Hearing about Bob's analysis, she integrates his standard costs by product (excluding the sales cost) with her invoicing data, which allows her to compare commissions to margin by product, region, and rep. A new model for Sales Commissions begins to emerge.

The center-out and edge-based models are clearly based on diverse assumptions and thus have rather different characteristics as shown in Table 1. From a business point of view, both approaches are needed. But the question that arises is: can they be combined in a single technological framework?

Emergent Prototyping and the Adaptive Information Cycle

The astute reader will have already observed that each of these two models has its own particular strengths. Center-out favors control, consistency and stability; edge-based supports innovation and change. IT prefers the former; business analysts the latter. Given the past emphasis by IT on control and consistency, how can emergent prototyping be fully integrated into formal processes for decision making that support both innovation and control?

In fact, the two models described above are but segments of a larger information lifecycle that is seldom explicitly described and supported in business decision-making today. The edge-based segment where business analysts work is where all new information of business relevance first appears and where decisions are made in the context of specific decision needs and drivers. The center-out segment is directed by the accumulated and consolidated wisdom of the business. Their combination is an ***adaptive information cycle*** that reunites the edge and center. Fresh data and specific analyses in context flow from the edge to the center. Controlled and consolidated information flows from the center as the basis for decision consistency.

Figure 1 shows the adaptive information cycle (AIC). When an event of interest occurs in the outer world, data is recorded about it. Conditioning brings together other existing and relevant data in the business to create useful information and ensure a complete and consistent picture of the overall relevance of the event. In the utilization step, the information is digested and decisions taken. So far, this is the same process that occurs at a corporate level in the center-out model.

It's also the same process an individual business analyst uses as the first steps in any investigation. However, in innovative analytic work, it is often clear that a good decision may require *new* information to delve deeper into the insights initially gained. New data needs to be gathered, recorded, conditioned and reused. This leads to the introduction of the ***assimilate*** step, which closes the AIC and returns the analyst to the record step.

The result is a "real-time", iterative approach termed ***emergent prototyping***. This approach is characteristic of innovative problem solving where analysis of the initially chosen information leads to a recognition that further variables may be of influence. The solution model thus emerges iteratively as a series of gradually converging prototypes including and excluding different parameters over the course of the analysis. Assimilation is the key step that takes insight gained in the first half of the cycle and posits that there may be more to discover about the problem by gathering new data from the real world. The outcome is that the intelligence assets of the business evolve in step with the evolution of its markets, technologies, competitors, and customers.

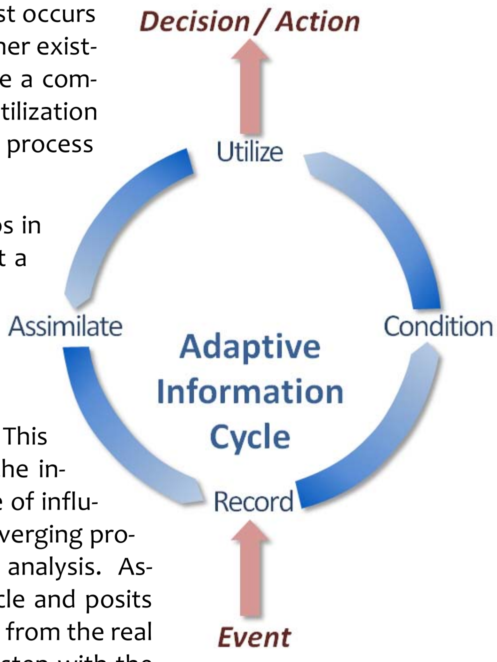


Figure 1: The Adaptive Information Cycle

The center-out model, most commonly implemented in Business Intelligence, actually does also support the assimilation step. However, the step is activated only occasionally. It occurs in the design and development phase of the warehouse, as business users contribute their knowledge of the data and conditioning needed to solve a business issue. It is also embedded in maintenance and upgrade projects of the warehouse. Closing the loop in the traditional BI environment occurs irregularly; information and analytic procedures thus fail to evolve with the business. Business analysts, driven by ever increasing time pressure, find that the data warehouse doesn't meet their needs, and go elsewhere for some or all of their data needs with predictable consequences for information quality.

In contrast, the edge-based approach includes assimilation, and the other three steps, on demand. An analysis may spin the cycle a number of times in a single day before coming to a conclusion. However, the entire loop is most often contained within the head of the individual analyst, without explicit support from either the tools used or IT.

In summary, what we see is information freshness and innovation at the edge, with an information cycle that is largely tacit, competing with information quality and consistency at the center with an explicit but incomplete information cycle. The requirement, then, is twofold. First, the tacit AIC that exists in analysts' heads must be made explicit with tools that document and structure the process and information used at the edge. Second, the largely broken feedback loop from the edge to the center must be reconnected so that the fresh information and collaborative solutions of analysts on the edge can be operationalized in a timely and regular manner. The outcome will be increased use of the central information sources, driving higher analytic consistency and quality at higher speeds.

The Adaptive Information Cycle—from Personal to Corporate Instantiation

How this can be achieved is shown in figure 2. We've already seen that the AIC exists in the analyst's head. This is the inner, ***personal*** cycle shown. As also mentioned, a good analyst can do this without tool support (although such support can help, as we'll see later). Now, imagine that the analyst has come up with a procedure that could be of general value. How does that procedure get incorporated in the production environment? Today, at best it depends on an entirely manual development process in IT reaching out to the analyst; at worst and most likely, it will never happen.

However, we know that even today such valuable analyses do get shared among analysts on an informal basis. The results of a worthwhile spreadsheet developed by one analyst are used as the basis for another spreadsheet solution by a second analyst. In some cases, chains of spreadsheet use and reuse may develop. Emergent prototyping crosses from the personal realm to the group, and benefits from the collective wisdom, knowledge and diversity of the group. Conceptually, this is what is shown in figure 2 as the ***group*** level. Although decision context and information relevancy are important determinants of what is shared among peers, personal relationships and trust tend to have even greater relevance in decisions to share knowledge[3]. This implies that the collaborative and social networking tools now emerging can be the basis for support of the group level in general and, in particular, the peer-to-peer interactions that lead to the promotion of personal analyses to group usage and later to the emergence of cooperative development within the group.

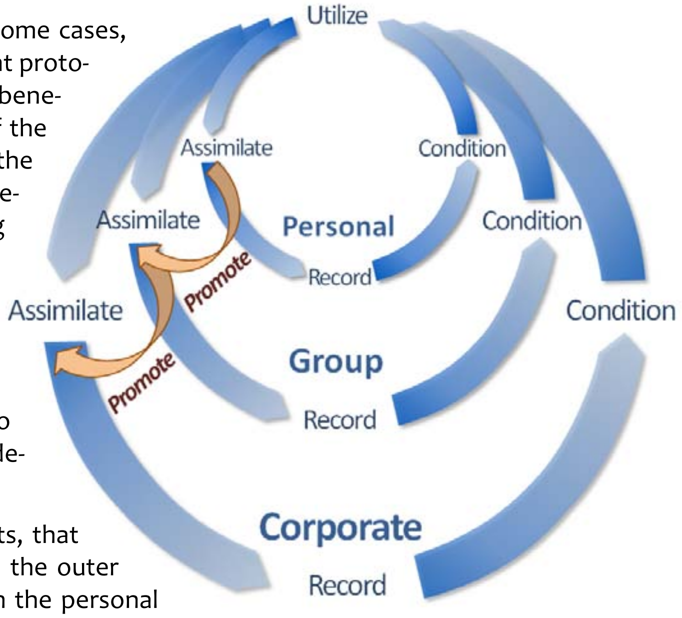


Figure 2: Levels of use of the AIC

Promoting a procedure, and potentially, associated datasets, that has been proven at the group level to the ***corporate*** level, the outer loop in figure 2, is more likely to happen than directly from the personal level. At least, they have been tested by more than one person and are more widely used, and so more likely to be accepted by IT as a valid solution to a real need. But again, the promotion process today is a manual development step. So, what support and tool function is needed to enable personal and group innovation to be promoted to the production level? Again, the answer comes in part from social networking and collaborative tools, especially those playing in the area known as *Enterprise 2.0*—the application of such tooling to the corporate environment.

In either case, it's in the assimilation step where linkage between the levels logically occurs. Between the personal and group levels, the mechanism is essentially a push, where the user who creates an analysis declares the value and availability of a particular function to his or her peers. Facilitating peer-to-peer interaction and sharing is key to easy and speedy movement of knowledge and analyses from the personal to the group level. Going from group to corporate level, the mechanism is more likely to be a pull, with the IT function proactively seeking suitable procedures on a regular basis from those shared at the group level.

Functional enablers for an Adaptive Information Cycle

Although mostly performed by business users, developing an analytic procedure is as close as you can get to application development without calling yourself a programmer. So, it makes sense that the type of functionality provided in a modern integrated development environment (IDE) would likely be valuable here too. For analytic procedure development, the key **IDE** functions are:

1. **Authoring and editing, with extensive visual support:** function templates and wizards are important for creating analytic procedures, but providing a visual map of the relationships between the various components of the analysis is key to ensuring valid overall procedure logic

2. **Metadata access:** definitive definitions of the meaning, usage, scope and interrelationships of data used in the analysis enhance productivity and validity of results

3. **Tracing, debugging and testing:** analytic procedures often contain multiple steps using different data, emphasizing the need to be able to trace the source(s) of a particular result

4. **Cross-procedure tracing:** tracing information and logic provenance across procedures is very important as procedures are often build upon one another, especially when shared among users

5. **Management:** documenting and versioning of analytic procedures is key to reuse

In contrast to a standard application development IDE, deployment of an analytic procedure at the personal level is simple: just run the analysis! However, the process by which procedures are shared and finally made publically available requires substantial support, and in a very different manner to that found in a standard IDE. In fact, this process is characterized more by collaborative working ideas than software deployment. **Collaborative deployment** function includes:

6. **Sharing of analytical procedures and datasets:** enabling authors to share procedures and data and revoke access to them
7. **Catalogue of shared procedures and datasets:** allowing analysts to search for suitable base procedures and data of interest
8. **Feedback and broadcast mechanisms:** for ongoing development and improvement of analyses
9. **Rating mechanisms:** procedures with high popularity and ratings among the business community become prime candidates for harvesting
10. **Group Creation and Management:** based on security, organizational and other criteria to limit access to sensitive information and/or procedures

At the group level, collaborative deployment is usually based on trusted social networks and is often limited to the sharing of the outputs of analyses. Within the usually small group of power business analysts in any company, certain individuals gain a reputation for being expert with particular sets of data or analyses. These people then become the hubs of networks of shared analytic function.

While sharing at a group level is largely driven peer-to-peer within the business analyst community, at the corporate level, the mechanism is best described as harvesting by the IT support group responsible for Business Intelligence in the organization. As particular analyses become widely and repetitively used, it makes sense that a central support function examine them for possible wider use, validity for the business as a whole, performance impacts and so on. Suitable procedures can then be moved to the production environment, optimized and rapidly made part of the BI environment. In addition, shared procedures can be subject to auditing and other controls.

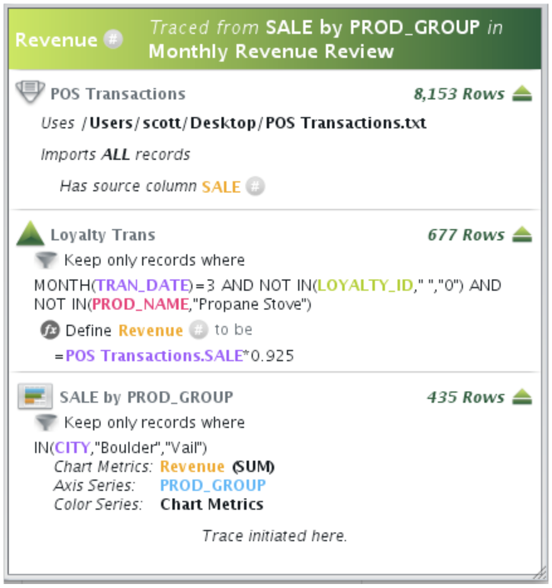
With such mechanisms in place, a more innovative and flexible approach to the ongoing development and delivery of Business Intelligence becomes possible. New procedures designed by the business analysts who know the data best can be shared with peers and the best analyses subsequently brought more rapidly into wider use. With better metadata and a more transparent development environment, as well as wider peer testing, analytic procedures are less likely to contain calculation or logic errors. Improved auditability is a further benefit.

Data quality concerns may still exist, given the more diverse development population and less centralized control than today's officially preferred center-out model. However, the fact is that almost every organization that has a data warehouse also has a parallel edge-bound environment of business analysts toiling away in totally uncontrolled and unaudited spreadsheets. Bringing such users into the fold as it is currently defined is impossible. But creating a new, unified model brings substantial innovative benefits to the business as a whole. It also enables IT to better support business analysts with timely information and business analysts to leverage common and consistent information as the base on which to build.

Lyza Commons: First Steps in Collaborative Analytics

The basic analytical approach and function of Lyza has been outlined in a previous paper[1]. As can be seen there and can be checked out via the free 30-day trial[4], Lyza provides a visual development environment for analytic procedures that are usually developed and run in spreadsheets. In contrast to development in a spreadsheet environment where the analysis is built from individual functions entered individually in specific cells and then copied manually back and forth, Lyza provides a column-based approach with the analysis built as a workflow of explicitly related functionality. This approach, with its comprehensive set of underlying metadata and the recently introduced tracing function shown in figure 3, amounts to a very usable and useful visual development environment for analytics as described in the previous section.

Lyza Commons extends the environment to support collaborative working. The concept is simple: a shared repository allows users who register with the Commons to share analyses in a controlled and managed fashion. When a user develops an interesting analytic procedure, s/he can share it with colleagues who are also members of the same group in the Commons. The distribution model employed here is publish / subscribe through a brokered peer-to-peer service, implemented through a physical copy of the procedure which is saved in the Commons and made available to the designated users.



Revenue # Traced from SALE by PROD_GROUP in Monthly Revenue Review

POS Transactions 8,153 Rows

Uses /Users/scott/Desktop/POS Transactions.txt

Imports ALL records

Has source column SALE #

Loyalty Trans 677 Rows

Keep only records where

MONTH(TRAN_DATE)=3 AND NOT IN(LOYALTY_ID," ","0") AND NOT IN(PROD_NAME,"Propane Stove")

Define Revenue # to be

=POS Transactions.SALE*0.925

SALE by PROD_GROUP 435 Rows

Keep only records where

IN(CITY,"Boulder","Vail")

Chart Metrics: Revenue (SUM)

Axis Series: PROD_GROUP

Color Series: Chart Metrics

Trace initiated here.

Figure 3: Tracing in Lyza

Note that it is the logic of the procedure and its information *result* set that are shared, rather than access to all of the underlying information set from which the results derived. This has important security consequences: sharing of a procedure doesn't give full access to the breadth of data that was available to the original developer of that procedure. This approach also takes into account the model of trust most commonly found in social networks (which is how groups of analysts are found to behave), where information is provided and accepted based largely on the identities of the people involved. Such a peer-to-peer model of interaction is particularly useful in promoting the collaborative emergent prototyping required for true analytic innovation. This function relates closely to the collaborative deployment enablers for the AIC listed above.

When a procedure has been shared, the receiver can use it directly as the base on which to construct his/her own analysis, joining in further data sources and adding analytic steps. And this new procedure can also be shared with another user who now has access the both lower levels of shared function and information. In such a case, it's even more important to be able to trace data provenance and manipulation back across all the levels of the procedure; and the tracing function in Lyza allows exactly this (function 4 above).

True collaboration between business people starts with sharing of results and analysis procedures between peers and extends to collaborative work on joint projects within the group. Such collaboration fosters reuse and builds networks of users who know who knows what—the tacit knowledge of the organization.

With shared procedures and metadata residing in the Commons, the stage is set to move from group level collaboration in Lyza to harvesting for the corporate level of the AIC. Today, administrative function in Lyza focuses on peer-to-peer sharing and simple group management. Support for feedback and rating mechanisms (functions 8-10) enabling IT to begin automating the harvesting process is slated for future releases. Given the fact that Lyza's primary market today is business analysts rather than IT, this lower prioritization of corporate-level function is reasonable. Furthermore, it will take time for group-level analytic work to be fully socialized before harvesting can begin in earnest.

It's also worth noting from an IT point of view that one of the concerns raised in the prior paper over the use of Lyza as a Playmart environment—the ability to export from Lyza to spreadsheets and other external tools—has been addressed. An installation option is now available to restrict export functionality for members of specific groups. Together with the introduction of the Commons for sharing of procedures in a controlled manner, this facility points towards an environment where analysts can freely explore data in a network of peers while the business retains the ability to manage and audit this activity.

Conclusion

So, did Joe Figger succeed in his new mission? Did Jill Joy join him? Will they live happily ever after?

The architectural model described above does provide the basis for a new cooperative approach between the edge-based business analysts and the center-focused IT department. Joe's long experience and personal understanding of how analysts actually work suggests that the IT department change their approach. Criticizing spreadsheet usage is not enough. IT must provide tools to enable analysts to innovate with data in a managed environment and to allow socialization and collaboration within that community. In addition, IT must speed up the inward flow of new information and analyses from the edge to the center and back again, so that analysts can use accurate and consistent of centrally-sourced information.

For data quality, auditability and Business Intelligence in general, this all makes perfect sense. Closing the adaptive information cycle has the potential to enable innovative use of information among those best equipped to do so—the business analysts themselves. It allows new information and analytics to be easily harvested for reuse across the organization, reducing IT effort and bottlenecks in producing new reports and queries. And IT can focus on what it does best—creating a common, high-quality and consistent base of information in the areas where such attributes are needed. As the loop is closed, the business analysts an see that their strengths are not in endless foraging for information that already exists.

Lyza clearly provides a powerful, well-controlled and cooperative development environment for analytic procedures, in line with the adaptive information cycle model elucidated above. However, for all its benefits, adoption of this model depends on some relatively substantial behavior changes for business analysts in particular. The Lyza environment requires analysts to move from cell-based to column-based thinking. Lyza has made this move, I believe, because the consistency and control offered by the latter approach provides the basis for the process metadata Lyza uses to capture and reproduce analytic logic. From the point of view of the analytical procedures themselves and in terms of bridging between the data warehouse and analytic environments, the rationale for this shift is undeniable. The challenge will be in bringing the spreadsheet-minded analysts on the journey.

Dr. Barry Devlin is among the foremost authorities on business insight and data warehousing. He is a widely respected consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation". Barry's current interest extends to a fully integrated business, covering informational, operational and collaborative environments to offer an holistic experience of the business through IT. He is founder and principal of 9sight Consulting, specializing in the human, organizational and IT implications and design of deep business insight solutions.

About Lyzasoft Inc.

Founded in 2008, Lyzasoft is spin-off of a successful business intelligence consulting practice – chartered to create desktop and workgroup software for powerful analytics without the reliance on lengthy IT development cycles. The Lyza product suite has been designed by analysts and for analysts.

Lyzasoft, Inc.
1675 Broadway, Suite 1300
Denver, CO 80202
www.lyzasoft.com

Brand and product names mentioned in this paper may be the trademarks or registered trademarks of their respective owners.

[1] The title "business analyst" has many diverse meanings. In this paper, we use it to refer to business users who perform relatively complex collection, analysis and interpretation of data from a wide variety of sources.

[2] "Playmarts: Agility with Control—Reconnecting Business Analysts to the Data Warehouse", Barry Devlin, 2008

[3] "Information Foraging Theory: Adaptive Interaction with Information", Peter L. T. Pirolli, 2007

[4] See www.lyzasoft.com/try.php