From Layers to Pillars—A Logical Architecture for BI and Beyond

Barry Devlin

Barry Devlin, Ph.D., is a founder of the data warehousing industry and among the foremost worldwide authorities on business intelligence and the emerging field of business insight. He is a widely respected consultant, lecturer, and author of the seminal book *Data Warehouse: From Architecture to Implementation*. He is founder and principal of 9sight Consulting.
barry@9sight.com

Abstract

The traditional BI architecture approach consists of a single stack of layers with data managed and moved from layer to layer. There were (and still are) good reasons for this design, but modern business needs drive another approach. Today's speed of response and breadth of data types dictate an architecture composed of pillars of data across multiple technologies and a new approach to integrating metadata as context-setting information across these pillars.

This article outlines the conceptual- and logical-level architectures that emerge from the data and processing needs of modern business operating in a world of abundant information, high connectivity, and powerful technology. Recognizing three distinct types of data, the architecture supports shared context across these types and the key role of traditional, modeled data in creating consistency and enabling governance. By defining pillars of data as a logical design, this approach supports optimization of technology choices and eases migration from current implementations, in contrast to the data lake approach favored by some in the industry.

Introduction

As far back as 2008, I began posing the question: Is the data warehouse architecture first described in the 1980s (Devlin and Murphy, 1988) becoming obsolete in the light of changing business needs and advances in technology? In a previous *Business Intelligence Journal* article, "Beyond Business Intelligence" (Devlin, 2010), I described at a conceptual level the characteristics of a new architecture that could support this new world. Bringing that architecture to the logical level—where IT can plan and design the infrastructure required—was slowed by the rise of big data as an industry priority and the more recent emergence of the Internet of things. These factors and others have now been incorporated into the conceptual architecture and detailed to the logical architecture level.

This article will provide an overview of that logical architecture. However, our starting point is a brief review of the conceptual level, the terminology of which has evolved as big data has become prevalent. The major portion of the article deals with the logical architecture. Its defining characteristic is a focus on pillars of information, as opposed to layers of data.

Both conceptual and logical levels are described in greater detail in *Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data* (Devlin, 2013). The final part of this article explores several technological aspects and, in particular, compares this architecture with a popular, technologically driven concept—the data lake.

The Conceptual or IDEAL Architecture

The fundamental purpose of a conceptual architecture is to frame a problem set or opportunity in a way that is simple and comprehensive enough to enable people with very different backgrounds to have a meaningful discussion about the topic. At the same time, the architecture must be sufficiently nuanced to begin the process of defining, describing, and designing solutions.

The problem (or opportunity) we address here is both stunningly simple and exceedingly broad: How to support today's decision-making needs in business, spanning the entire spectrum from automated decision management through predictive analytics to strategic planning, within the organizational and personal context of the business. The IDEAL architecture, shown in Figure 1, provides three conceptual *thinking spaces* within which solutions may be found: information, process, and people.

The framework may be expressed as a simple sentence: To make decisions, people process information. Although apparently simple, this formulation is powerful. It causes us to think about three equally important aspects of the solution:

- **Information:** What information is needed? Indeed, what is information, and how does it relate to data? From whence does information come? What are its defining and classifying characteristics?

- **Process:** Which processes are needed to both create information and make it available, reliable, and meaningful? How should these processes behave in both automatic and manual decisions?

- **People:** What roles do people have in making decisions? How do people interact socially in decision making? What goes on inside their minds (and bodies) in terms of rational thinking, intention, intuition, and even emotion as they contemplate and reach decisions? Which processes do people use to gather information? Indeed, how and where do people find information relevant to the decisions they need to make?

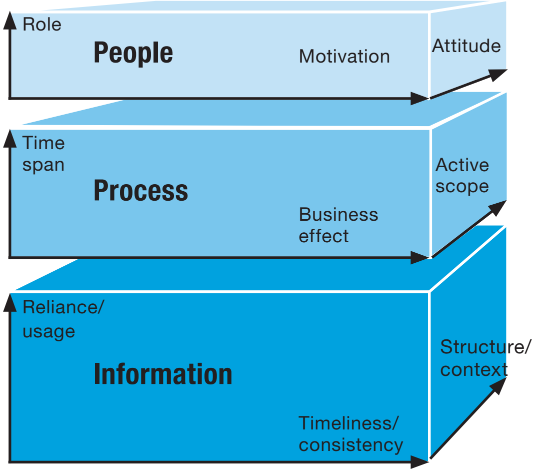

Figure 1: The IDEAL conceptual architecture.

Each of these conceptual thinking spaces is three dimensional, as shown by the axes drawn in Figure 1. Each axis describes a key characteristic (or interdependent combination of characteristics) of that space that must be considered in the context of decision making. For example, attitude in the people space shows the different mental or emotional approaches that people may take in decision making. “Timeliness and consistency” in the information space describes the interaction between these two information characteristics that is familiar to every data warehouse designer. These nine axes provide the conceptual framework within which business and IT can explore together what the business requires and what the technology enables in support of decision making.

Although it may be simpler to think in terms of three independent spaces, these three conceptual thinking spaces are deeply and intimately interconnected, as seen even in the questions used above to frame them. It is difficult to think of information divorced from the processes that generate it or the people who use it. For people and for business, process is the natural bridge between information and people. People are the *raison d'être* of both information and process, although they are rarely considered in detail in any prior IT architecture. Therefore, the IDEAL architecture is deliberately drawn with information as a foundational block and people in the topmost space.

This image emphasizes that thinking about decision making should begin and end with people as opposed to data or information, which has long been the primary focus of business intelligence and which is being reemphasized today in the phrase “data-driven business.” In truth, business is (or should be) people driven. It is a social enterprise to facilitate the exchange of goods, energy, and information between people. The decisions made in the course of business are, at least for the foreseeable future, made solely and exclusively by people, whether directly (for example, in strategic decision making) or indirectly (through automated, operational procedures based on principles and patterns conceived originally by people).

The framework constrains the conversation to an appropriate balance of generalization and detail so that IT can understand the business issues and business can see the strengths and weaknesses of technology. Business does not need to understand the difference between MapReduce programming and a SQL optimizer; nor must IT explore the details of collateralized debt obligations.

The term *IDEAL architecture* contains an acronym that captures the five key characteristics of the conceptual architecture:

- **Integrated:** Within and across all three layers, a unity of thought and purpose drives the approach; all aspects of this environment must link seamlessly together

- **Distributed:** Each layer of this architecture consists of a concept space with diverse attributes of equal importance, individual independence, and mutual dependence; there is no single, central control point

- **Emergent:** Ours is a mathematically and socially chaotic and complex environment, the characteristics of which cannot all be predicted or calculated in advance; order materializes from disorder to drive coherent structure and behavior

- **Adaptive:** As business needs and technological possibilities change, the architecture is sufficiently agile to adjust to and take advantage of them without re-architecting

- **Latent:** Being latent (hidden), the conceptual architecture is not in a form that can be directly implemented; it is a guide for business and IT thought and conversation about what is desired and possible

The latent nature of the conceptual architecture leads us now to discuss the logical REAL architecture.

The Logical or REAL Architecture

Perhaps the most obvious feature of the REAL architecture, shown in Figure 2, is the absence of people given the importance we attributed to them in the previous section. However, the rationale for the omission is rather obvious. A logical architecture provides a basis for implementation of technology by IT. At least for now, IT has proven incapable of (or unwilling to) build people!

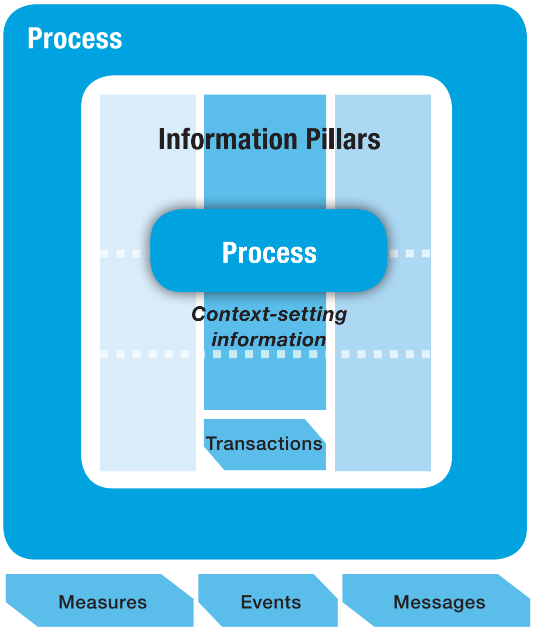

Figure 2: The REAL logical architecture.

As a result, the REAL architecture consists of two main components: information and process. As in the case of the conceptual architecture, the acronym *REAL* here is also an expression of the key characteristics of this level of the architecture:

Realistic: Implementing this architecture can begin today with existing technology and its full, foreseen extent is achievable with tools and techniques that can be expected within a few years

Extensible: Given the early stage of emergence of the fully integrated decision making described by the IDEAL architecture, the functions and features of the logical level are open to extension and expansion to allow technology evolution

Actionable: The actions and approaches required of the business and IT are clearly identified at a high level and can easily be extrapolated to lower levels of detail

Labile: The architecture is flexible enough to allow changes in business needs as the business-technology ecosystem evolves

In the interest of brevity, the remainder of this article will focus on the information components of the REAL architecture. As will become clear, this information architecture is realistic, given its starting point in existing data warehouse and/or Hadoop and other technologies. It has already proven extensible and labile, showing an early ability to incorporate the emerging Internet of things within its scope, and it is actionable, as seen in the final part of this article, when contrasted with the data lake approach.

The starting point for defining the new REAL information architecture is an examination of some key aspects of the IDEAL information space, illustrating how we move from conceptual to logical architecture. As shown in Figure 3, we focus on two axes of the information space: timeliness/consistency and structure/context. By focusing on the characteristics of information along these two axes, we begin to discern three relatively distinct (although somewhat overlapping) sets of information and data.

Process-mediated data is perhaps the set most familiar to BI practitioners, being the data that is gathered, stored, and managed in traditional operational and informational systems. The term *process-mediated* emphasizes the intimate relationship between this data and business processes. Data in this category is defined

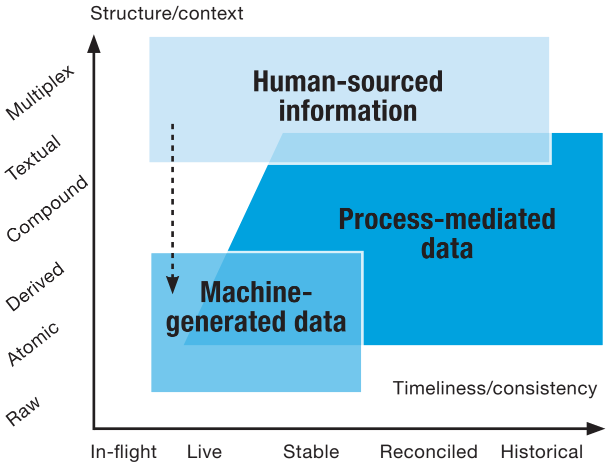

Figure 3: Three categories of data/information.

and modeled up front as business requirements are honed and applications designed. It reflects exactly—at least in the ideal case—the business needs expressed by the users, the logical constraints that emerge from the interrelation of data elements, and technological limitations of IT systems. Within these boundaries, this data is purposely constructed to reflect the legally binding reality of the business. Its role as a dependable record of business transactions demands the closest attention to considerations of reliance and consistency, clearly distinguishing such data from the other two sets.

Human-sourced information is the subjective and highly personal record of people's beliefs about and recollections of what has happened or may happen in the physical world. In a practical sense, it consists of tweets and Facebook posts, written e-mails and reports, even prose and poetry, images, and audio and video recordings, captured in social media or content management systems. It is loosely structured, textual, and multiplex information spanning the full spectrum of timeliness/consistency, from in-flight tweets to historical accounts and video documentaries, and everything in between.

At a fundamental level, human-sourced information provides insights into the inner, personal landscape of people's minds. In the old world of business computing, before the emergence of big data, this landscape was largely invisible. However, careful consideration reveals that such information is a foundation for process-mediated data. It is, in fact, human-sourced information that data modelers and application designers gather in JAD sessions and requirements documents in order to build operational and informational applications.

In daily life, human-sourced information precedes and pervades the formal transactions between people and businesses. Before buying a dress (a transaction recorded in process-mediated data), a potential customer creates a story (human-sourced information) about her needs, preferences, and expectations, either internally or externally via social media. The content of the story leads directly to the business transaction, or not, as the case may be. It is, of course, this linkage that imbues human-sourced information with its predictive power, provided it can be captured and interpreted in a timely fashion.

While mining sentiment and intention from tweets and Facebook posts is the state of the art today, analysis and interpretation of image and video is rapidly gathering speed. Known as affective computing, this technology bridges from human emotions to analytics, based on machine recognition and modeling of human emotional expression. A number of startups are already developing systems for market research, medicine, entertainment, and government initiatives (Dwoskin and Rusli, 2015). These developments give some indication of the extent to which human-sourced information is expanding in size and scope, and indicate that image and video rather than text may soon become the data formats that define significant future technological characteristics for the human-sourced information category.

The third category, **machine-generated data**, has recently garnered attention with the Internet of things. However, it is not new by any means. The data generated by machines has long been an input to process-mediated data. The key switches and bill counters of ATMs have been generating the input data to business transactions since the 1980s. Network devices recording start and end times and destinations of phone calls have been the ultimate source of billing information for even longer.

Machine-generated data consists of events or measures gathered by particular devices at known times. Such data is rather simple in its structure. Conceptually, it consists of a device identifier, date and time, and a list of name-value pairs showing what has been measured and its value. When such data comes from internally owned or managed devices, the structure is usually well defined and relatively stable over time. In such cases, machine-generated and process-mediated data have so much in common that it has never been necessary to distinguish between them in the past.

This situation changes dramatically with the introduction of the Internet of things. Here, the vast majority of machine-generated data arrives from external sources over which the receiver has little or no control. Therefore, the structure or content of such data may be poorly understood or may vary over time. The data itself may be incomplete or in error. Different or upgraded devices may produce new or different measures and events. As a result, machine-generated data is often described as being semi-structured, reflecting the variability that may occur, especially within the list of name-value pairs.

A further characteristic of this data is the speed and volume at which it arrives. Jet engines, for example, produce 5,000 data points per second for maintenance or flight optimization analysis.

The very different characteristics of these three data and information types leads directly to a simple conclusion: We are likely to need different technologies to support these differing requirements. What works for the well-defined, pre-modeled, and stable process-mediated data may not work for loosely defined, variably structured, time variant, machine-generated data. The management, processing, and storage requirements of very high volume and very loosely structured human-sourced information differ considerably from those of process-mediated and machine-generated data. We are immediately led, therefore, to an architecture that allows multiple technologies with different strengths and weaknesses to store, manage, and manipulate the range of modern data and information.

Forging Layers into Pillars

The traditional data warehouse architecture can be summarized very simply in two thoughts. First, bring all the data you need together in a single place to ensure the consistency and cleanliness required for effective decision making and management reporting. This single place, or enterprise data warehouse (EDW), has been implemented in a general-purpose relational database since its inception. Second, in order to cater to different processing needs, pass this data through several storage layers, each optimized for different purposes. Thus, we have an operational systems layer, a staging layer, an EDW layer, and a data mart layer, to name but a few.

As seen in the previous section, the very different characteristics of the three types of data and information preclude bringing all data together in one place—typically, the relational EDW. Having evolved together, process-mediated data and relational databases are well matched. The relational model provides a firm theoretical foundation for the type of storage, processing, and analysis normally associated with such data. Years of research and development and, in particular, recent hardware and software advances, suggest that relational databases will likely remain the best platform for process-mediated data. For human-sourced information, enterprise content management systems have traditionally been favored.

However, Hadoop is making a strong case for its use in many aspects of storage, processing, and analysis of such data. For machine-generated data, a case can be made for relational databases, particularly at smaller volumes and lower speeds. Streaming systems are probably mandatory at the highest speeds and volumes. NoSQL data stores are growing in performance and favor for this type of data.

From this set of simple technological observations, the concept of pillars of data emerges. The REAL information architecture shown in Figure 4 chooses to show three pillars, corresponding to the three categories described earlier. Actual implementations might show more than three depending on the technical requirements and limitations encountered. It's seldom likely to be fewer. The REAL information architecture is most definitely heterogeneous.

Choosing pillars does not necessarily mean abandoning layers; a pillar might consist of a number of layers, for example. However, there are strong arguments for eliminating, or at least reducing, the number of layers. Layers, by definition, imply multiple copies of data, some identical and others extended or enhanced in some way. With big data volumes, a strategy involving multiple copies rapidly becomes prohibitively expensive, not only in terms of storage and processing, but particularly for management. Layering also implies transporting and processing data between those layers, which introduces complexity of design and maintenance as well as runtime challenges. These latter include increased latency of data available to end users and higher complexity in two-way synchronization of data across the layers.

These challenges have long been evident in the case of process-mediated data, in the operational BI and analytics use case. In a traditional data warehouse environment, operational data is ingested into the data warehouse for analysis and creation of predictive models. These models are transferred to and run in the operational environment, influencing the actions of user-facing applications in real time, such as when cross-selling or upselling on retail websites. The outcomes are fed back into the analytics environment to improve the predictive models. As business demands ever faster feedback loops, the limitations of a layered architecture become clear.

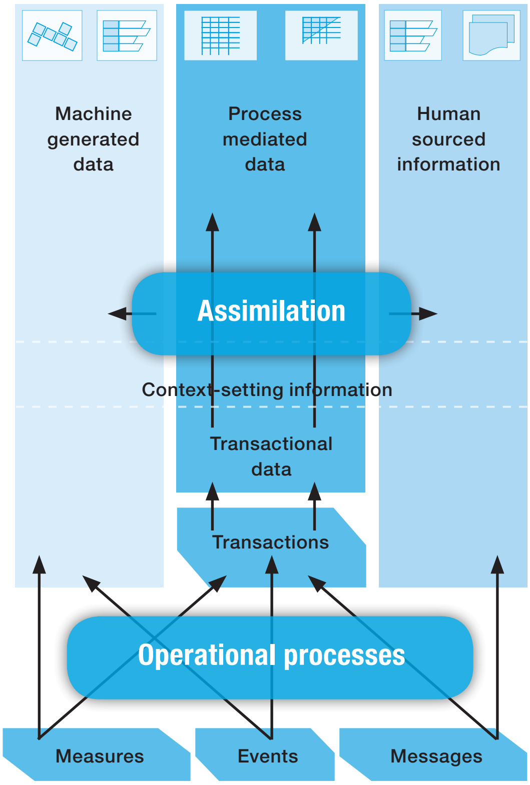

Figure 4: Information pillars in the REAL architecture.

In practice, the advantages of layering (such as design simplicity and separation of operational/informational processing concerns) are outweighed by the needs of the business for rapid response and agility. Advances in in-memory processing and relational database design (Devlin, 2014) are enabling the creation of combined operational/informational systems, of which SAP HANA is an example. Such systems are sometimes labeled (Gartner, 2014) as hybrid transaction/analytical processing (HTAP). For these reasons, layering within the process-oriented data pillar is likely to diminish over time. As machine-generated data and human-sourced information pillars are introduced, layering is best avoided in order to control data volumes.

The question that arises with multiple pillars, of course, is that old chestnut of consistency. The traditional data warehousing approach of data consolidation is no longer an option. The new answer is threefold. First, cleansing and consolidation can occur within each pillar. This is particularly true of process-mediated data, which is modeled prior to implementation, making such data a key resource for ensuring wider consistency and cleanliness. Second, the process of assimilation is responsible for comparing and consolidating data and information across the pillars. This process depends on data virtualization and data integration technologies. Third, we must depend on the stuff currently known as metadata.

Metadata, I contend, is a misnomer. The simple definition—data about data—reveals the name's severe limitations. I propose to call it *context-setting information* (CSI seems an appropriate acronym) because (i) its true purpose is to record the context in which data or information can be used, and (ii) in structural terms, it is much closer to information than data. If we think simply in relational database terms, the data is the values in the rows; context spans from column and table names through descriptions, all the way up to business definitions and usage scenarios, and down to the information about when the values were last updated.

Together, all this CSI amounts to a collection of meaning, at least insofar as information can be said to be truly meaningful outside of the mind and experience of each individual. Such context and meaning is, of course, necessary all three types of information. However, collecting and/or creating CSI for these newer categories of data is much more difficult than for process-mediated data, as is clear to many a data scientist who spends a majority of time and effort in trying to do so.

Context-setting information is thus shown spanning the three information pillars in Figure 4. New tools being developed for big data governance will begin to build out CSI in machine-generated data and human-sourced information. However, the CSI that forms part of process-mediated data should always take precedence where available because it is normally developed though formal modeling processes, as opposed to being harvested from externally sourced and thus potentially unreliable data.

Diving into the Data Lake

Having explored this new architecture, it is appropriate to compare it to other architectural thinking. The relationship to the traditional data warehouse architecture should already be clear. Current operational and informational systems consist, for the most part, of process-mediated data. That pillar of the REAL information architecture depicts the future of the data warehouse and other components, suggesting that, in some or perhaps many cases, "de-layering" will occur over time.

However, comparison of the REAL information architecture with emerging architectural thinking, known as the data lake, raises some concerns. Proposed by several vendors in the big data market, Wiktionary defines a data lake as "A massive, easily accessible data repository built on (relatively) inexpensive computer hardware for storing 'big data.' Unlike data marts, which are optimized for data analysis by storing only some attributes and dropping data below the level aggregation, a data lake is designed to retain all attributes, especially so when you do not yet know what the scope of data or its use will be" (Wiktionary, 2015).

The proposed content of the data lake varies by proponent, although there is general agreement that all externally sourced, machine-generated data and human-sourced information should be included. At its most extreme, even traditional operational data could become part of the data lake. Others suggest that the data warehouse should be a prime candidate for inclusion *and* elimination in its current form.

The definition above leads directly to an implementation approach for the data lake based on Hadoop. Although this is a rapidly evolving and expanding technology base, it would appear highly ambitious to suggest that it can be easily optimized for the many different types and uses of data and information required by a modern business. The pillar architecture, in contrast, recognizes this need for "horses for courses" and positions Hadoop and, indeed, other technologies to play to their strengths.

Although potentially attractive on the basis of potential hardware and software cost savings, the migration of existing data warehouses and, particularly, operational systems to a new platform seems both expensive and risky. The REAL architecture, while allowing or encouraging migrations where needed (for example, to in-memory database systems to support operational analytics), takes a much more realistic approach to the introduction of new platforms. In fact, one of the foundations of the architectural thinking was to be conservative in the level and speed of evolution required of existing systems.

Finally, the REAL architecture affords a central role to process-mediated data as the legally binding, historically dependable record of the business. Such data demands special care and treatment, characteristic of traditional relational databases (and, indeed, older hierarchical or network varieties). The Hadoop environment, on the other hand, has emerged from a very different set of needs: as "a solution to an economics problem faced by Google and other Internet giants a decade ago" (Hunt, 2014) involving enormous data quantities and relatively limited needs for consistency, reliability, or even speed. Adding such attributes to the Hadoop environment is, at best, a long and arduous process, which remains far from complete after eight years of development.

Conclusions

The data warehouse architecture, first conceived in the mid-1980s, was based on the business needs and technical limitations of its era. A key aspect of that architecture was layering of data. Business priorities have changed, although the underlying need for useful, consistent information has not. Although the old technological limitations have eased, big data has highlighted other restrictions. Taken together, these changes demand a rethinking of the original data warehouse architecture. The outcome presented here emphasizes pillars of information and data rather than layers.

The concept of pillars provides a welcome opportunity to use multiple technologies according to the characteristics of the information used and processing required. Pillars offer more agility in both design and use of information. However, without the single point of consolidation in an EDW, creating and maintaining consistency becomes more challenging. Metadata, or context-setting information, takes on an expanded role and an ongoing consistency-checking process across pillars is required. Traditional operational and informational data, or process-mediated data, plays a central governance role in the pillared REAL architecture, as well as providing a starting point for its evolution.

This multi-platform, evolutionary approach stands in contrast to the data lake concept. This concept begins with the needs of the "new" big data types—human-sourced information and machine-generated data—and attempts to retrofit the technology to the process-oriented world of legally binding transactions handled by traditional operational and informational systems. This one-size-fits-all approach will most likely lead to significant implementation challenges in terms of data consistency, governance, and migration.

The IDEAL conceptual and REAL logical architectures outlined here recognize that modern business uses a diverse set of data and information, some in large volumes and at high speeds and others with high reliability and consistency. The resulting pillared architecture offers the best approach to balance these competing needs, while simultaneously providing a clear evolutionary path from current implementations. ■

References

Devlin, Barry [2010]. "Beyond Business Intelligence," *Business Intelligence Journal*, Vol. 15, No. 2. pp. 7–16. http://tdwi.org/research/2010/06/business-intelligence-journal-vol-15-no-2.aspx

Devlin, Barry [2013]. *Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data*, Technics Publications.

Devlin, Barry [2014]. "The Emergent Operational/Informational World." http://bit.ly/wp-htap

Devlin, Barry, and Paul Murphy [1988]. "An architecture for a business and information System," *IBM Systems Journal*, Vol. 27, No. 1.

Dwoskin, Elizabeth, and Evelyn M. Rusli [2015]. "The Technology that Unmasks Your Hidden Emotions," *The Wall Street Journal*, January 28. http://www.wsj.com/articles/startups-see-your-face-unmask-your-emotions-1422472398

Gartner press release [2014]. "Hybrid Transaction/Analytical Processing Will Foster Opportunities for Dramatic Business Innovation," January 28. https://www.gartner.com/doc/2657815/hybrid-transactionanalytical-processing-foster-opportunities

Hunt, Matt [2014]. "The Big Problem Is Medium Data," High Scalability blog, December 17. http://highscalability.com/blog/2014/12/17/the-big-problem-is-medium-data.html

Wiktionary, The Free Dictionary, http://en.wiktionary.org/w/index.php?title=data_lake&oldid=27849084 (accessed January 29, 2015).