And now… a Production Analytic Platform

OCTOBER 2017

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

The data warehouse can and should evolve into a Production Analytic Platform to support operational implementation of predictive analytic models developed in the data lake and elsewhere. The Teradata Database embeds analytic function that supports this goal.

Business and the data warehouse evolve

> In an era when business needs access to all imaginable data and the ability to take near-instantaneous decisions and action based on it, the traditional data warehouse database is evolving to provide a foundational platform.

The traditional data warehouse architecture is based on five assumptions[1] that separate operational and informational processing, which suited the data types and business needs of the day. Modern data characteristics (size, structure and speed) and business needs for predictive analytics and immediate action have led to claims of the imminent death of the data warehouse. The future, according to this theory, lies with Hadoop, NoSQL, and more, implemented as a data lake.

The flaw in this thinking is that it forgets the technology underpinning the data warehouse architecture. This technology is a relational database management system (RDBMS) and has been growing in power and evolving in capability for over four decades.

On its own, RDBMS technology has proven its ability to handle the wildly diverse requirements of (1) operational systems, (2) enterprise data warehousing, (3) analytical data marts, (4) personal computers, and (5) open source environments as shown in the table overleaf. In addition, support for newer data types and technologies, such as XML, JSON, and advanced analytic functions such as pathing, sessionization, text analytics, and scoring functions, have been incorporated in RDBMSs, taking advantage of their existing scalability, reliability and performance characteristics.

This pattern of extending existing RDBMSs with complementary function suggests the evolution a new platform, capable of bridging today's hard operational-informational divide. We call this new approach a *Production Analytic Platform*.

SPONSORED BY
www.teradata.com TERADATA.

Requirement	Characteristics	Examples
Operational Systems	*Run the business:* high reliability, up-to-the-second read/write, ACID (Atomicity, Consistency, Isolation, and Durability)	Oracle, IBM DB2
Enterprise Data Warehousing	*Manage the business:* high reliability updating to build consistent history, read-only in daily use	Teradata Database
Analytical Data Marts	*Understand the business:* high speed read-only sorting, summarizing, querying, and reporting	Greenplum, Netezza, Vertica (now acquired by major S/W vendors)
Personal Computing	*Understand the business:* small scale read-only analytical data mart function for experimentation	Microsoft Access
Open Source	Extending all above strategies to the Hadoop world	Hive, HBase, Impala, Kudu, Parquet

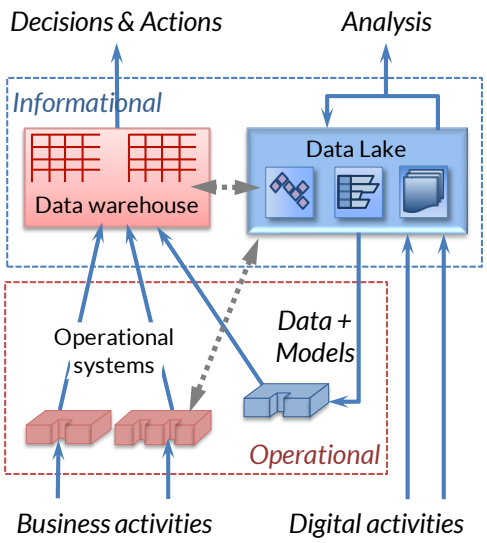
Understanding the operational-informational divide

Operational systems record and manage the parties, agreements and transactions that constitute the legal basis of business and must be fully current, consistent, reliable, performant, etc. Informational systems—data warehouses, marts, etc.—contain data extracted and derived only from operational systems to enable trusted analysis, querying, reporting, and decision-making, as shown on the left of Figure 1.

This convenient (for IT) division of labor has been disrupted by two factors. First, human-sourced information and machine-generated data[2] from customers' and partners' digital activities (from mainly external sources) is poorly suited to both traditional operational systems or data warehouses, because of quality, volume, velocity, and variety concerns. Second, large-scale predictive and operational analytics, based primarily on these new sources, demands functionality and scale that characterize the data lake, an environment that is informational in nature and where cheap data storage takes precedence over high performance analytics and predictable user SLAs.

Data scientists first cleanse raw data—an operational task—normalize and analyze it—an informational activity—in the data lake, then pass the data and derived models to the operational environment for ongoing production. This complex flow challenges the traditional operational-informational architecture. In addition, normalization and analysis of this new data often requires the use of reference and other data stored in the data warehouse and operational environments, shown by the gray dotted arrows in Figure 1. The current approach of building a stand-alone, data lake environment for digital activity data and analytics on a different, separate platform to that of traditional business computing thus also creates barriers to such data sharing and reuse.

Figure 1: The operational and informational worlds collide



Some data lake proponents favor moving everything into the lake. For enterprises with significant investment in traditional environments, this approach is neither feasible nor financially attractive. Furthermore, the maturity of data lake tools lags behind that of traditional systems, especially in terms of the reliability, availability and maintainability needs of mission-critical operational systems and the data warehouse.

Introducing the Production Analytic Platform

A Production Analytic Platform bridges the traditional operational and informational worlds. Its initial focus is to ease the task of putting models developed in a data lake into a high-performance, reliable environment, but can extend to a range of other borderline operational/informational activities.

As we've seen, the traditional operational-informational divide impedes modern analytics based on digital activity data. Bridging these two environments, the Production Analytic Platform provides a better balance of function and performance between them, based on the eight key characteristics listed here.

Key Characteristics

1. Embedded support for a wide range of reporting, query, and advanced analytics functions, as well as openness to inclusion of new, emerging features
2. Built-in storage and support of all common data formats and the ability to easily include user-defined formats with adequate processing performance
3. Scalability to data volumes necessary for production use of digital activity data (may be less than for initial analytic requirements and must be defined)
4. Ability to access data stored remotely (data virtualization) and to optimize use of such data by local caching and other means
5. Users' ability to access all data types via native languages as well as via SQL
6. Access to multiple analytics engines that provide a choice of tools and analytic methods to users, including commercial and open-source products
7. Support for a wide variety of user types and their preferred tools: business analysts, data scientists, application developers, executives, and business knowledge workers
8. Reliability, availability, maintainability, and performance levels compatible with production use for daily operational decisions, as well as tactical and strategic decision making

The first and central practical question about this platform is to identify its base technology. The most likely candidates are a classic RDBMS and open source Hadoop-related technology. The comparison is between a mature technology that has been designed and built in a controlled and managed approach over decades (RDBMS) and one that is still evolving and that is driven by the immediate and often short-term interests of a development community over less than a decade. The latter traits certainly contribute to rapid technological advances and, for characteristics (1) - (4), means that even where RDBMS is currently ahead, we may assume that open source will catch up and perhaps overtake in a few years. For characteristics (5) - (7), both approaches play fairly evenly.

Characteristic (8), therefore, becomes key. The open source environment presents extensive, well-known software maintenance issues. Rapid feature development often comes at the expense of reliability and operability. In such areas, RDBMS has a considerable advantage, a long history, and is already embedded in the operational and informational environments of most enterprises. And, as discussed above, RDBMSs continue to add support for a variety of novel data formats and analytical functions.

The Production Analytic Platform is thus positioned across the operational-informational boundary, as shown in Figure 2. It allows direct ingestion of digital activity data to support production analytics based on models generated previously in the data lake. With a shared store of core business data from the traditional operational systems and the data warehouse, the Production Analytic Platform is the ideal location for ancillary data and function used in analytics both in production and in the model-build phase in the data lake.

In terms of implementation, the Production Analytic Platform is an extension of the data warehouse (and, in particular, the enterprise data warehouse) to undertake more operational activities. These activities are firstly related to day-to-day operation of predictive analytic models, but can extend to other borderline operational/informational activities. Central to all of this is the analytical feature richness, robustness, and performance characteristics of the RDBMS underpinning the data warehouse, all of which points to the Teradata Database as a suitable starting point.

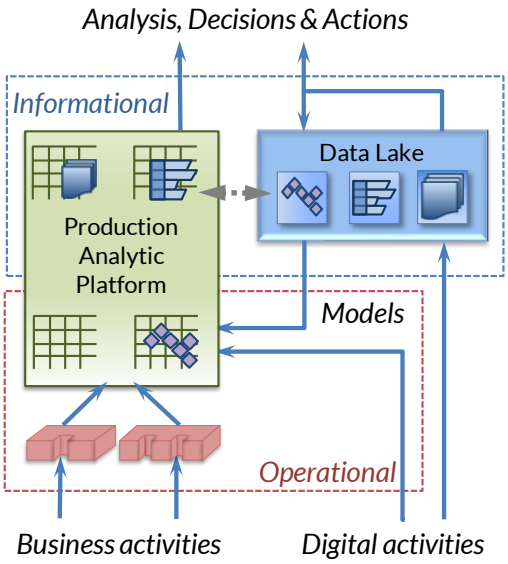


Figure 2: Production Analytic Platform

Further ThoughtPoints in this series explore specific aspects of the analytic extensions in the Teradata Database that support this direction.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His book, ***"Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data"*** *(http://bit.ly/BunI-TP2) was published in October 2013.*

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

[1] Devlin, B., *"Business Integrated Insight BI2—Reinventing enterprise information management"*, (2009), http://bit.ly/2xZc77b

[2] Devlin, B., *"Business unIntelligence—Insight and Innovation beyond Analytics and Big Data"*, (2013), Technics Publications LLC, NJ, http://bit.ly/BunI-TP2