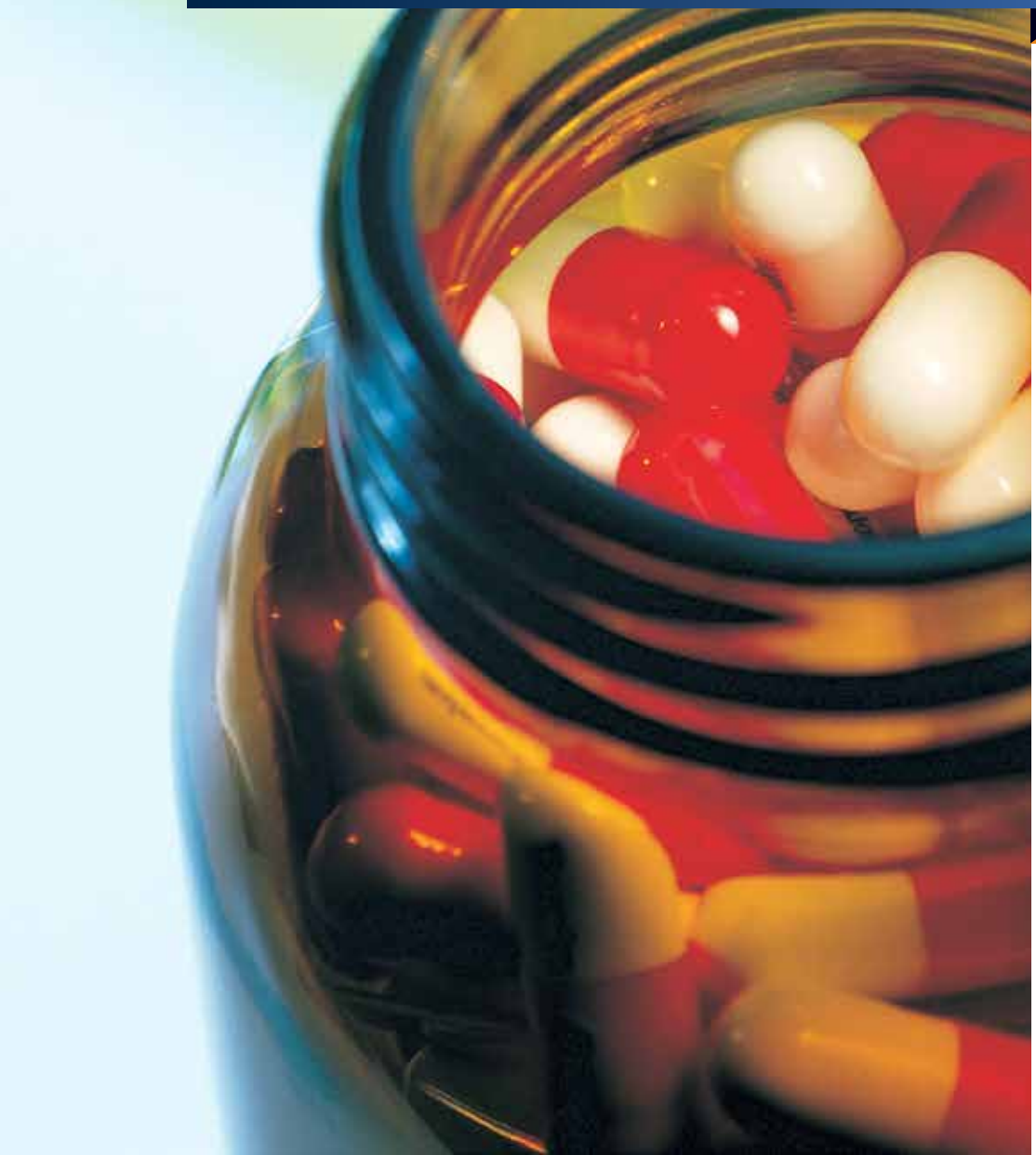
LIFE SCIENCES

06.13 | EB 7662

TRANSFORMING THE ECONOMICS OF REAL-WORLD EVIDENCE ANALYTICS

INTEGRATED ENVIRONMENT EXPEDITES DATA AVAILABILITY, REDUCES COSTS, AND EXTENDS BUSINESS INTELLIGENCE ACCESS THROUGHOUT THE ORGANIZATION

TERADATA



EXECUTIVE SUMMARY

Pharmaceutical companies are expanding their research and development (R&D) and commercial business intelligence (BI) capabilities to include real-world evidence (RWE) analytics that track, measure, and predict the clinical outcomes of drug treatments across various medical conditions, patient populations, and treatment regimes. These analyses deploy a diverse range of analytical techniques against very large datasets acquired from various sources with different data structures and semantics. First-generation platforms based on popular analytical software and shared-server environments have widely failed to deliver the data management and query processing performance necessary to make these analytics affordable or accessible to a meaningful number of users.

This paper describes an integrated analytical environment that combines optimized solutions for interrogating structured and multistructured data, coupled with a high-performance data-staging solution and linked with high-throughput tools for data movement and loading. The result is a dramatic improvement in the performance, cost, and accessibility of real-world evidence analytics.

NEW DECISION LANDSCAPE FOR THE PHARMACEUTICAL INDUSTRY

Developing a new drug and shepherding it successfully through the clinical trial process has always been a data-intensive operation. Historically, post-launch analytics focused on initial uptake, follow-on prescription volume, and margin performance. With the availability of RWE data in unprecedented volume, the range and complexity of analytics have greatly increased.

Drug makers have entered a complex decision-making process where success requires an advanced portfolio of analytical capabilities. Converging changes in the competitive and regulatory environments are forcing drug companies to focus on each compound's observed clinical and financial performance throughout the entire product life cycle, encompassing the true diversity of real-world treatment circumstances. The market forces driving the adoption and evolution of RWE analytics include the following:

- **Fewer blockbusters in the development pipeline:** Major firms are responding to patent expiration, generic competition, and margin erosion by refocusing their development investments on new applications for existing drugs and on targeted treatments for smaller populations that are less attractive to high-volume generic specialists.
- **Falling ROI:** Longer and more expensive approval processes are now widely reported to exceed $1 billion in costs for each successful new treatment.
- **Vast new volumes of data:** Information related to medical conditions, clinical treatments, and medical outcomes in real-world environments is becoming available as a result of the industrywide transition to electronic health records.
- **Pricing and reimbursement pressure:** Many payers have developed advanced capabilities of their own to monitor drug efficacy and clinical outcomes in real-world treatment settings.
- **Rising awareness of drug safety issues:** More frequent product recalls have been made possible, in part, by the U.S. Food and Drug Administration's (FDA) own investment in improved post-market monitoring capabilities within the Center for Drug Evaluation and Research (CDER).

To support the broad interest in new outcomes analytics, the Pharmaceutical Research and Manufacturers of America, the FDA, and the Foundation for the National Institutes of Health launched a public-private partnership in 2009: the Observational Medical Outcomes Partnership (OMOP). This interdisciplinary research group has been charged with developing standard data models, vocabularies, methods, metrics, and algorithms for outcomes analysis of very large heterogeneous datasets.

BARRIER TO ENTRY: DATA MANAGEMENT CHALLENGES ON STEROIDS

Even with a nascent set of standards and best practices to follow, developing an internal RWE analytical capability remains a formidable undertaking. This is primarily due to the vast size and diversity of the datasets involved. These include clinical data from healthcare providers, claims data from insurers, electronic medical records, and even self-reported outcomes data from social media sources.

These datasets are structurally diverse because they originate in different systems within very different hardware and software environments. They include structured,

CASE STUDY: QUERY TIME CUT FROM DAYS TO MINUTES

Situation: The epidemiology department in a top-ten pharmaceutical company runs the OMOP standardized Univariate Self-Controlled Case Series (USCCS) macro to determine cause and effect relationships between drugs and conditions in large populations as well as precise surveillance windows. The department's original platform was SAS software running on shared UNIX servers.

Problem: Very large input tables caused very slow runtimes. Only one analysis could be run at a time, taking four hours to complete with only a limited data sample.

Solution: A hybrid approach runs the SAS USCCS macro in-database on Teradata.

Results: The processing time for one analysis was reduced from four hours to five minutes. Processing time for the longest-running query dropped from 4.3 days to 3.8 minutes, a 1,600X performance improvement. And all USCCS analytics now run on full datasets.

multistructured, and potentially unstructured data types, not all of them well suited to relational databases or to SQL-based analytics. They are semantically diverse because similar data is often encoded differently by different data providers. Bringing all this information into a single repository for analysis is inherently difficult; it requires a comprehensive set of transformation rules, very large storage volumes, and extremely efficient translation and loading engines.

These intrinsic data management challenges are only aggravated by the common practice of conducting RWE analyses using one of the popular statistical analysis software packages (SAS® software is certainly the most widely used) running on shared-server hardware. These implementations suffer from two significant shortcomings: They consume very large volumes of expensive storage, and their performance in complex analytical processing is limited by the parallel execution resources available in the hardware platform.

CONSTRAINTS AND THE COSTS OF COMPROMISE

One inevitable result is very long execution times that limit the number of analyses that can be run, reduce the number of potential users (especially business users), and increase the cost of each inquiry. The result? Users try to work around the performance constraints. Less data is brought onto the platform and inquiries are run on aggregates or data samples that limit the richness and utility of the results. Input datasets are often not retained on the platform, leading to wasteful redundancy in data management, preparation, and loading, as well as quality assurance and data lineage issues. Costs on these platforms tend to be high and productivity low. More tellingly, users' ability to stimulate the pace and extend the reach of scientific inquiry with robust analytics is distinctly limited by the computing environment.

MORE PRODUCTIVE AND AFFORDABLE RWE ANALYTICS

To reduce the costs of RWE analytics, extend the benefits to a wider range of users, and fully realize the potential throughout the business, pharmaceutical companies need analytical environments that support several essential capabilities:

- Collect, manage, and access large heterogeneous data volumes with a high degree of efficiency, including the ability to centralize and retain large amounts of detail and history.
- Integrate and transform heterogeneous datasets, rendering them accessible and consumable by a diverse set of analytical processes and techniques.
- Execute any type of analytical process or technique in an optimized environment, with access to any type of data.
- Use large-scale parallelism to accelerate all data management and analytical processing tasks.
- Accelerate analytical execution and minimize data movement by performing all analytical computation on the data repository platform.
- Conduct rapid, agile, low-cost exploration of data and new analytical approaches.
- Extend access to the analytical environment beyond IT to all scientists and business users with an opportunity to improve operational performance through analytics and data-driven insights.

CASE STUDY: IMPROVED RESPONSE TIMES BY 450 TIMES

Situation: The outcomes research unit at a top-ten pharmaceutical company needed to improve its ability to use external data sources in answering key scientific questions.

Problem: Slow processing performance on the group's shared SAS platform required eight hours to build and save a patient summary for static analysis. Some queries ran for four days, severely limiting the group's ability to support additional workloads.

Solution: A hybrid approach allows both internal and external datasets to be loaded into a Teradata integrated data warehouse, and existing SAS queries are converted to run in-database.

Results: Response time improvements ranged from 20 times on a sample of the data to 450 times on the entire dataset. And processing time for a diagnosis code listing of patients dropped from five hours to approximately 40 seconds.

BUSINESS BENEFITS OF AGILE, AFFORDABLE RWE ANALYTICS

If life sciences companies go to the trouble of implementing an integrated analytical environment, do the benefits really extend beyond a small group of analysts and the IT staff charged with their support? To answer that question in the context of your own organization, consider the impact on operational performance and competitive advantage if your firm could do the following:

- Predict adverse-outcome events in time to respond proactively, reduce consumer impacts, and limit economic repercussions.
- Identify and use positive outcomes with targeted marketing that precisely aligns treatments and high-responding population segments and provides insight and focus for future clinical trials.
- Identify new indications for existing treatments and use prior R&D investments to capture new revenue with reduced development costs.
- Closely align development, marketing, and distribution operations with well-understood market and patient segments to maximize efficiency and ROI.

SOLUTION: INTEGRATED ANALYTICAL ENVIRONMENT

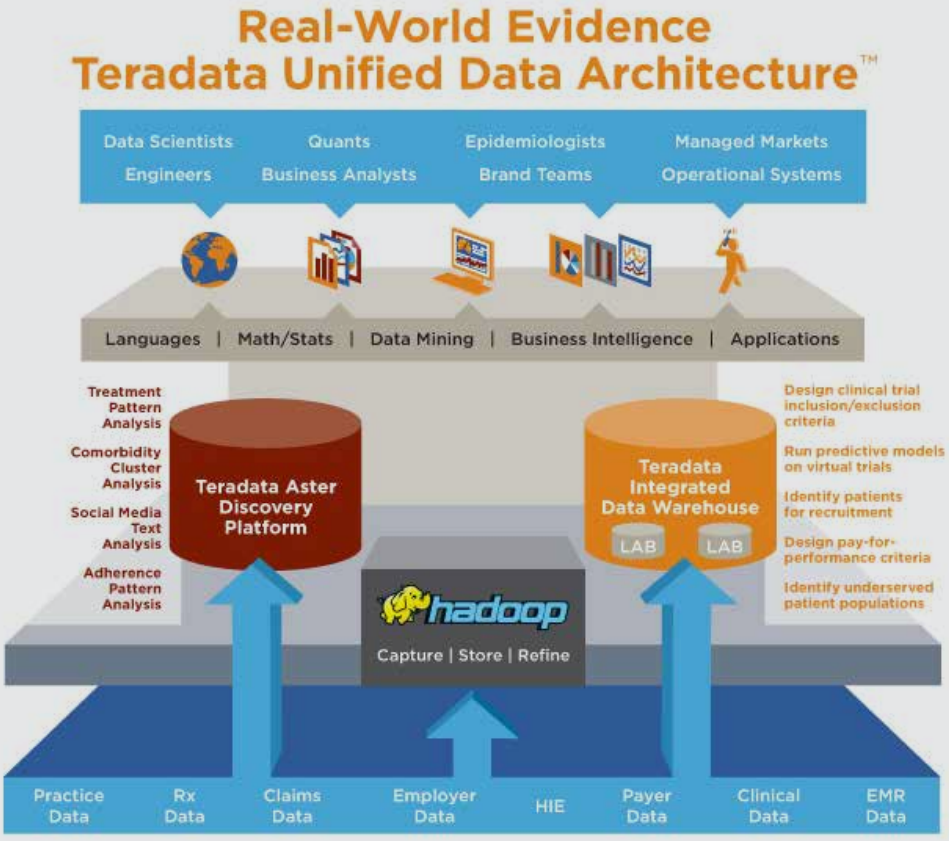
No single BI platform or solution provides all the necessary capabilities, so the logical approach is to carefully select and tightly integrate best-in-class solutions for the core functional requirements along with the applications, services, and tools necessary to support interoperability. Three core components are essential:

- **Integrated data warehouse (IDW):** This provides a shared environment for strategic and operational analytics and a single centralized source of data for reuse. For life sciences companies, this will include clinical and clinical trial data, purchased claim data, electronic health records (EHRs), and genetic markers. The IDW should combine a high level of software-based parallelism as well as hardware-based scalability and should support in-database analytical execution for leading third-party analytics providers. It should also support self-provisioning data labs that allow users to easily combine data from outside the IDW with read-only views from the database, enabling rapid, low-cost data exploration and ad-hoc analysis.
- **Data discovery platform:** This enables rapid investigation of big multistructured datasets using a variety of techniques and prepackaged analytics and applications. This platform should provide database extensions that bring the analytics to the data rather than the less efficient opposite. It should use the built-in parallelism of the MapReduce software framework and, like the IDW, should support in-database analytical processing. Ideally, it should simplify access by scientists and business users by providing a SQL interface to all types of data and integrating SQL into the MapReduce framework to mask the complexities of parallel programming. In a life sciences analytical environment, this platform will typically house data from social networks and EHRs and will be used to identify early safety concerns and measure the efficacy of various treatment pathways as well as other similar applications.
- **Data-staging platform:** This provides a cost-effective environment for loading, storing, and refining services to prepare all incoming data—structured, unstructured, or multistructured—for analysis based on the value of

the data. Apache™ Hadoop® is highly optimized for precisely these functions. In life sciences analytics, the data-staging platform receives and prepares diverse data streams and types, including sensor, social network, or genomic data, for iterative analytics within the discovery platform.

In addition to these three core platform components, a viable and effective integrated analytical environment for drug companies requires the following:

- **Life-sciences-specific logical data model:** This formally defines and documents the relationships among data extracted from disparate operational sources. A logical data model is the key to extending analytics and business intelligence into every area of an organization. It is the logical framework that makes it possible to ask any question of any data at any time.
- **Life-sciences-specific physical common data model:** This is based on a life sciences logical data model that normalizes the data to provide data model persistence, access paths, and extensibility.
- **Semantic interface:** This normalizes and standardizes the analytics interface.
- **Data access and movement tools:** These simplify user access to any data regardless of where it is stored and accelerates transfer and loading between platforms. It is essential that data visibility and movement throughout the environment is transparent, fast, and friction free.



Integrated data discovery and business intelligence

THE TERADATA UNIFIED DATA ARCHITECTURE

Teradata is helping to remove the data management barriers that constrain inquiry and analysis in the life sciences industry. Teradata has created and is continuing to develop a unified data management and analytical environment, the Teradata® Unified Data Architecture™. It dramatically increases the efficiency, productivity, and cost-effectiveness of life sciences analytics, including real-world outcomes research. For more information, download the business value white paper, "Turning Value into Profit with UDA": **www.teradata.com/products-and-services/unified-data-architecture/**.

CONCLUSION

Operational capabilities rely on unrestricted access to data and analytics for business users throughout the organization—a level of access and investigative freedom that simply isn't possible when the costs of analysis are too high and query response times are measured in days or weeks.

Achieving the necessary performance—both computational and economic—requires the ability to take advantage of an optimized analytical platform; to use scalable, parallel execution in both the software and hardware layers; to manage large, diverse datasets with speed and efficiency; and to move data effortlessly between analytic platforms. The inescapable solution is an integrated analytical environment.

ABOUT THE AUTHOR

Ed Acker is an R&D life sciences industry consultant at Teradata. He has more than 15 years of life sciences experience, primarily focused on health economics and outcomes research (HEOR), electronic submission, clinical trial transparency, and translational medicine. At Teradata, he is focusing on a common data model for translational research and the application of the unified data architecture in life sciences R&D.

10000 Innovation, Drive Dayton, OH 45342 teradata.com

TERADATA. THE BEST DECISION POSSIBLE™