



# NAVIGATING THE IN-DATABASE LANDSCAPE

FIVE BEST PRACTICES FOR DEPLOYING  
ADVANCED ANALYTICS IN TODAY'S ENTERPRISE

TERADATA®

## EXECUTIVE SUMMARY

Since the mid-1990s, in-database analytics has helped companies manage massive amounts of data while maintaining high performance. Yet even as open source tools push the boundaries of innovation, many of these tools remain memory-bound and lack parallel capabilities, thus limiting their ability to process large volumes of data.

Companies can overcome these limitations by applying some key best practices for broadening and optimizing the capabilities of advanced analytics tools. These best practices include embracing parallelism, gaining access to the largest possible in-database library, selecting a platform that supports a variety of analytics, and building a solid foundation on a unified data architecture.

This paper reviews the current landscape of in-database technology and advises today's analysts on how best to exploit emerging innovations for maximum competitive advantage.

## BRIEF HISTORY

In-database analytics is not a new development. In fact, by the standards of enterprise technology, it's quite established, with a history stretching back some 15 years.

In 1998, Teradata was the first to offer in-database analytics to address a number of emerging challenges. With information scattered in siloed data marts across the enterprise, analysts struggled to pull together the right data, manipulate it, and analyze it. Most of these jobs were performed on a desktop or a server, and the iterative nature of analytics required many representative data samples to be pulled from source systems and transferred over the network, requiring hours of cycle time.

With the introduction of in-database technology, users could push analytics into the database instead of pulling it out. Further efficiencies arose from integrating siloes of data into a single data warehouse, eliminating the need to search for data across platforms and reducing analytic cycle time from hours to minutes.

Today, open source technology plays an increasingly prominent role in pushing the boundaries of innovation in advanced analytics. The programming language known simply as *R* is now the basis for many of the leading tools on the market, and a host of vendors are designing solutions that expand upon *R*'s advanced capabilities.

Simply put, the creation of in-database analytics was a major milestone in the field, and data scientists continue to

“Experienced analytics practitioners are incorporating cutting-edge approaches such as in-database analytics, text mining, and sentiment analysis to outdo their competitors.”

—2013 Analytics & Info Management Trends, InformationWeek, November 2012.

rely heavily on this technology for high-performance data management across every major industry. The question, then, is how to capitalize on these latest innovations to achieve even deeper insight across the enterprise.

## INNOVATIONS AND LIMITATIONS

Currently, open source *R* analytics is available on the Comprehensive *R* Archive Network (CRAN), the technology of choice for data scientists everywhere. *R* language and software includes the most innovative analytics technology available. This open source solution enables analysts to progress from a research paper to a callable algorithm within days.

However, there are still inherent limitations within the *R* language itself. For example, *R* is memory-bound, which means it is unable to process large volumes of data. It also lacks parallel capabilities. It can be used in a parallel framework such as Apache™ Hadoop® but only effectively by parallel programmers. Because *R* is not inherently parallel and is also limited to running in memory, running it across multiple nodes will execute analytics only on the data within each node. This can often result in multiple or aggregated answers, which can obscure or complicate one's view of the available data and even produce wrong results if not programmed with parallelism in mind. There are, however, solutions to overcome these limitations.

## FIVE BEST PRACTICES FOR DEPLOYING IN-DATABASE ANALYTICS

New developments offer many opportunities to broaden and optimize analytics implementations and R's considerable capabilities. Companies can apply some key best practices to combine optimal insight with the greatest possible efficiency.

### 1. Embrace parallelism.

R is not a parallel language. A number of solutions simply run R concurrently, which is not true parallelism. In practice, this approach often means taking the easy way out—and saving the hard parallel programming work for later.

If, for example, an analyst calculates an average while running R concurrently, every linked node will perform the same calculation, return the results, and calculate the average of the average, which can often be the wrong answer. The responsibility is placed on the programmer to bypass the functions and write the calculation correctly. The programmer will need to submit a sum and a count to each node, then perform the aggregation and division at the system level. The whole process becomes far more complicated than necessary, with frequently off-target results.

In-database technology's latest advances enable analysts to bring R from the node level to the system level. That way, they can perform the hardest work up-front in the database, and not force parallelism later in the process.

### 2. Get access to the largest possible in-database library.

The simple fact is that in-database analytics will run most effectively when paired with the largest possible library.

Today's companies need to analyze massive amounts of data in many different ways, using a variety of algorithms and techniques. With a large library at their disposal—preferably a library based on a standard language such as SQL—they can gain fast, easy access to the broadest range of business intelligence tools on demand. Tasks that once took hours can be rewritten and run in a matter of minutes.

For example, Fuzzy Logix's comprehensive in-database analytics library is more complete than any other, with more than 600 algorithms that work inside Teradata® platforms. When combined with Teradata SQL-H™ capabilities, Fuzzy Logix functions can be run on Teradata Database and Teradata Aster Database with additional data from Hadoop.

Analysts can use this extensive library to run everything from simple statistical analyses to Monte Carlo simulations, all from deep within the Teradata and Teradata Aster Databases.

### 3. Select a platform that supports a variety of data types and analytics.

Analysts should be able to embrace a wide range of data for truly enhanced analytics. That means going far beyond structured data to integrate nontraditional information types for maximum insight.

For example, many businesses load XML data by bits and pieces, failing to gain a full picture of the information at their disposal. By combining XML documents with business data, analysts can ask new questions and solve new problems without jumping across tiers to process data. They can even retain all details in each XML document for further analysis and traceability.

Other nontraditional data types such as spatial data and temporal data (data that captures time) can provide unprecedented insight into customers' locations and paths of travel—an important capability, given the rise of GPS technology. By quickly combining customer profiles with location data, companies can make great strides toward optimizing location-based applications and mobile applications.

### 4. Make it easy to view your data.

More capabilities shouldn't lead to greater complexity. Instead, analysts should be able to take advantage of built-in automation to simplify a broad range of functions. For example, temporal data analytics has traditionally required a great deal of manual calculation. Now, however, ANSI-compliant data types allow for maximum application portability, and companies can deploy automated functions to easily track customers' loyalty across products and services over time.

Analysts can further streamline processes by running functions in parallel. And by using standard languages to simplify queries, they can integrate more information—including XML data—for comprehensive visibility across complex systems.

### 5. Build a solid foundation for your data.

To make the most of in-database analytics, companies need an architecture that can support analysis of data across the enterprise. Analysts must have access to all of the available data, all the time.

With a truly unified data architecture in place, companies will have the infrastructure necessary to capitalize on in-database technology's latest advances. The right architecture will be capable of capturing structured, multi-structured, and unstructured data so that analysts will never need to move data to a middle tier—enabling them to realize greater competitive advantage more quickly and easily than ever. A unified data architecture can also provide seamless interconnectivity so that analysts are empowered to use their tools of choice against all data across the enterprise.

## TERADATA: A SIMPLIFIED APPROACH

The analytics process is heuristic in nature. With each analysis, users travel further down the path toward true insight. And because they learn more with each cycle, they need a system responsive enough to continue their creative flow.

### SOLVING BUSINESS PROBLEMS: SELECTED USE CASES

By bringing together two industry leaders—Teradata, with its comprehensive Teradata Unified Data Architecture™ and Fuzzy Logix, with its scalable in-database analytics—companies can finally realize the potential that has long been just out of reach:

- ~ A global entertainment company was able to refine its new release forecast model, resulting in higher sales.
- ~ A major marketing company lowered churn by more than three percent, reduced the cost of Internet sales by 50 percent, and increased revenue 1,000 percent over four years.
- ~ A multi-channel outdoor lifestyle retailer was able to triple analytics output and extend predictive analytics to other lines of business.
- ~ A major financial service provider was able to increase the effectiveness of its model-building team by a factor of ten. With in-database analytics, tasks that previously took 175 hours can now be done in just 25 minutes.
- ~ A major retailer was able to eliminate a number of legacy data marts, reducing development cycles from four months to a matter of days.

In-database solutions from Teradata do the heavy lifting up-front so that analysts can be more flexible, focused, and creative. Instead of asking a question and waiting hours for an answer that might be unclear or simply incorrect, analysts can sustain an ongoing conversation with their data.

The newest in-database technology from Teradata continues a long tradition of industry-leading innovation. Teradata now offers the world's first out-of-the-box parallel in-database R, along with the largest library of in-database advanced analytics anywhere. Additionally, Teradata solutions enable integrated XML data, optimized geospatial features, and enhanced temporal analytics. Taken together, these developments signal a major step forward in the industry.

Built on the Teradata Unified Data Architecture™, the Teradata solution for in-database analytics delivers the following benefits:

- ~ Easy access to all available data—All production data is consolidated in a single warehouse, enabling users to easily self-provision a workspace within the data warehouse and add experimental data for analysis, then focus on the analysis rather than the mechanics of extracting the right data.
- ~ Access to a variety of data types—Analysts are no longer limited to structured data. They can expand their analytics capability by incorporating XML, location data, and time-sensitive temporal data to enrich analytics.
- ~ Simple movement from development to deployment—Since R can be run directly in the warehouse in parallel, a company's R program can be easily ported for optimal performance as it processes massive volumes of information.
- ~ No more parallel programming—Analysts can simply write a program in R with Revolution R Enterprise with ScaleR, and Teradata will take care of the parallel processing and return more accurate answers in less time.
- ~ Full interoperability with analysts' favorite tools—By using the most popular and standard languages and interfaces, Teradata can integrate in-database capabilities with all the major tools. With XQuery support for XML, built-in functions for geospatial and temporal data, open source R for analytics tools, and hundreds of new Fuzzy Logix functions, analysts have more ways to make the most of their favorite tools.
- ~ Secure, enterprise-ready technology at a low total cost of ownership—The Teradata solution is a secure environment built on an enterprise scale, delivered at a cost significantly lower than traditional enterprise analytics technologies.

- ~ Comprehensive set of analytics functions—Applying Fuzzy Logix analytics at the source, users can eliminate uncorrelated variables or statistically insignificant observations, greatly reducing the response time necessary for standard modeling and reporting tools.
- ~ Solid foundation built on a unified data architecture—As businesses extend their data architecture to incorporate a wider variety of platforms, analysts have the flexibility to analyze any data with transparency and ease.

## THE TERADATA ADVANTAGE

The acceleration of industry change is relentless, putting tremendous pressure on every business to understand and discover new advantages—and exploit those advantages before anyone else.

Teradata offers truly groundbreaking solutions for in-database analytics, making it easier for business users to analyze all of their data and operationalize analytics for greater business value. With the world's first out-of-the-box parallel R, the largest library of in-database advanced analytics, and full integration across all types of data for fast, simple analysis, Teradata simply provides the most innovative solution on the market today.

For more information about how Teradata can help you extend and operationalize your analytics across the enterprise, **contact a Teradata solutions specialist**, or visit the **data mining and analytics solution page** on [Teradata.com](http://Teradata.com).



10000 Innovation Drive Dayton, OH 45342 [teradata.com](http://teradata.com)

# TERADATA®

Unified Data Architecture and SQL-H are trademarks, and Teradata, Aster, and the Teradata logo are registered trademarks of Teradata Corporation and/or its affiliates in the U.S. and worldwide. Apache is a trademark, and Hadoop is a registered trademark of Apache Software Foundation. Teradata continually improves products as new technologies and components become available. Teradata, therefore, reserves the right to change specifications without prior notice. All features, functions, and operations described herein may not be marketed in all parts of the world. Consult your Teradata representative or [Teradata.com](http://Teradata.com) for more information.