

Aster Data® Embedded Analytics

Unleash the Power of Big Data Analytics

Enterprises have a critical need to rapidly extract the value locked in the exploding volumes of multi-structured big data available to them from new and emerging sources. Unlocking this value requires advanced analytic applications that can iteratively explore and investigate multi-structured data at scale even when the relationships in that data are not yet known. These analytic applications need to process high-performance, complex queries on terabyte to petabyte size data sets to perform analyses including pattern, time-series, and graph analysis on rich, granular data sources such as raw web logs, mobile devices, social networks, text, and sensor data. Big data analytics such as these require more than just standard SQL queries, they also require analysis that is impractically difficult and complex to perform in standard SQL.

Delivering on the needs of big data analytics requires new technology capable of processing diverse analytics on multi-structured data at massive scale. Achieving this requires bringing together data and analytic processing to enable analytics on full data sets, enabling a broad array of SQL and non-SQL analytics, and providing a solution that addresses both structured data and multi-structured data from new and emerging sources that is not typically stored and processed in traditional systems.

The Aster Data Solution

Embedded Analytics is a unique capability of the Aster Data *n*Cluster analytic platform that delivers powerful new analytics by bringing together massively-parallel processing of full analytic applications with massively-parallel data stores for multi-structured data from new and emerging sources. Complementing the capabilities that existing technologies for in-database analytics bring to structured data residing in traditional systems, Aster Data embedded analytics makes it possible to incorporate large-scale multi-structured data that does not typically reside in traditional systems into analytics and leverage new technologies like MapReduce for techniques such as pattern, graph, and time series analysis that are difficult or complex to express in SQL.

Aster Data delivers embedded analytics by bringing together its massively-parallel data stores with an embedded analytics processing engine based on MapReduce, a framework first pioneered by Google for processing petabyte-scale data sets distributed across thousands of commodity hardware nodes. Other implementations of MapReduce require specialized programming skills to take advantage of its powerful analytic capabilities, but Aster Data embedded analytics brings the power of MapReduce to standard SQL through Aster Data's patented SQL-MapReduce®. Aster Data's SQL-MapReduce marries the power of standard SQL development with broad support for existing programming languages and frameworks including Java, C/C#/C++, .NET, and Python, offering a broad array of tools for analytics.

Aster Data embedded analytics makes it possible to process complete analytic applications, not just portions of analytic logic, alongside multi-structured data to eliminate the limitations imposed by moving data to separate systems for analytics. It allows any existing analytic application, including pre-packaged applications like R and SAS, to run in-database inside *n*Cluster, fully parallelized, and achieve significantly faster response times.

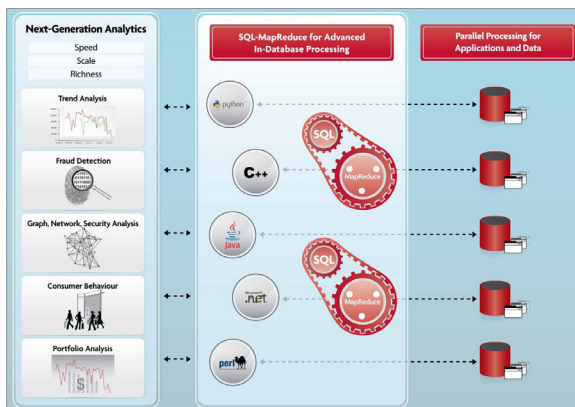


Figure 1: Architectural overview of Aster Data embedded analytics.

Overview

The Aster Data analytic platform provides an integrated analytic engine that processes complete analytic applications in parallel together with data to deliver high performance and scalability for big data analytics on multi-structured data from new and emerging sources.

Highlights

- Provides high-performance processing of advanced analytic applications at scale
- Allows full analytic applications to be embedded and processed together with data
- Brings MapReduce processing to big data analytics with an embedded analytics engine
- Makes it possible to process analytics on full data sets, eliminating the need to limit analytics to samples or windows of data
- Supports a broad variety of common development languages including Java, C, C#, C++, Perl, Python, and .NET

“Aster Data’s calling card is in-database analytic processing, an increasingly popular approach that speeds processing by running applications next to the data rather than extracting data and processing in the application environment. Teradata and Netezza have led the way in making in-database processing a reality, but Gartner says Aster’s four-tier architecture is particularly well-suited to the approach.”

Doug Henschen, Editor-in-Chief
Intelligent Enterprise

intelligent
enterprise

Scale Analysis to Large Data Sets

Traditional analytics architectures were not designed to handle the demands of complex analytics on terabyte to petabyte volumes of data, particularly multi-structured data from new and emerging sources. As a result analytics are delivered through sub-optimal approaches that require data sampling and increase the time required to obtain updated analysis of fresh data.

Consider fraud detection as an example. Instead of running analytics on full data sets with near-real-time updates, analysts are required to rely on extracts from the original data, introducing data latency and limiting the scope of analysis. The result is that incidents of fraud go undiscovered or are detected only long after they occur. Other types of analytics face similar challenges with sample bias and inaccurate analysis.

With Aster Data embedded analytics, all data stored in the database is available to every analytic application for deeper, more accurate data analysis. With its massively-parallel architecture, Aster Data embedded analytics scale linearly to massive data volumes without data access or processing bottlenecks. This means that analytic applications can run directly on terabytes or petabytes of data, in near real-time—eliminating the limitations imposed by data sampling.

Achieve High Performance Data Analysis

Aster Data embedded analytics delivers not only for massive scale but also high performance. Using its embedded MapReduce analytic engine, the Aster Data analytic platform spreads analytic processing across an entire cluster of servers, parallelizing analytics processing without requiring complex parallel systems programming techniques.

To illustrate the performance benefits of Aster Data embedded analytics, we compared the performance of a use case originally written with standard SQL. In tests performed in 2009 a standard SQL query on Aster Data's *n*Cluster performed 9x faster when implemented using Aster Data's SQL-MapReduce framework than when running SQL alone.

As processing increases in complexity with packaged analytic applications like SAS and R, the Aster Data analytic platform continues to deliver high performance on large data sets. For example, on standard SAS analytics use cases, Aster Data has proven 8x–10x faster processing for data mining execution than a traditional database (see figure 2).

Ad-hoc Analytics and Deep Data Exploration

Moving beyond basic reporting to rich, ad-hoc analytics is challenging with standard SQL. Examples of rich ad-hoc analytics include pattern detection, time-series analysis, graph analysis, and behavioral pattern detection. These techniques require iterative, multi-pass processing, for which traditional SQL systems were not designed.

With SQL-MapReduce, queries that require multi-pass SQL are simplified to single-pass queries that perform at petabyte and terabyte scale. To the developer, MapReduce is presented as a simple set of SQL table functions. Although these functions may be highly sophisticated internally, the developer does not need to understand the internals of the function to use them. Furthermore, these functions can be automatically leveraged by any standard reporting or analytic tool that understands SQL—no toolset changes required.

All Aster Data SQL-MapReduce functions, whether pre-built or user-designed, are extremely flexible because they support the evaluation of a SQL-MapReduce analytic function at run-time rather than design time. Why is this important? Unlike traditional user-defined functions, which are rigid in their input and output data schema structure, SQL-MapReduce functions offer dynamic input/output data schema that can be determined at run-time. This allows SQL-MapReduce functions to be written once and then used many times without changes—there is no need to change their code or write additional declarations to change input or output schemas or parameters. As a result, Aster Data embedded analytics provide exceptional productivity for the data analyst, complete reusability, and expressive power for rich, ad-hoc analysis.

About Aster Data

Teradata's Aster Data analytic platform is the market-leading big data analytics solution. The Aster Data analytic platform embeds MapReduce analytic processing for deeper insights on new data sources and multi-structured data types to deliver analytic capabilities with breakthrough performance and scalability. Aster Data's solution utilizes Aster Data's patented SQL-MapReduce to parallelize the processing of data and applications and deliver rich analytic insights at scale. For more information visit www.asterdata.com.

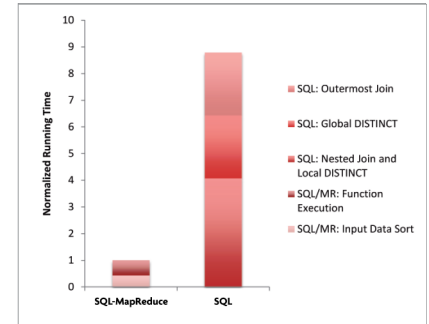


Figure 2: Runtime comparison of SQL and SQL-MapReduce clickstream analysis.

“With Aster Data, response times for large queries have dropped from 5 minutes to 5–10 seconds, and queries that previously were not possible now can be executed in 20–30 seconds.”

*Richard Zwicky, Founder and President
Eightfold Logic*

