

Aster Data Advanced In-Database Analytics

Access the Power of Massively Parallel, Advanced Analytics Inside the Database

Enterprises are moving beyond business intelligence and embracing advanced analytic applications. Advanced analytics are high-performance, complex queries that uncover rich insights within terabyte to petabyte size data sets. Examples include computing statistical measures, identifying behavioral patterns, processing graph analysis, or simply performing time-series analysis on rich, granular data sources like clickstreams, transactions, or system logs. Business intelligence and reporting can be performed with standard SQL queries, but advanced analytics requires deep, interactive analysis on fine grained data—requirements that go beyond the usability of standard SQL.

While there have always been capabilities within the database for running application logic beyond SQL—such as stored procedures or user-defined functions (UDFs)—these are largely being abandoned in favor of new data platforms that can host entire analytic applications inside the database. Leveraging innovations in SQL-MapReduce technology, Aster Data enables Advanced In-Database Analytics by pushing analytic applications into the database, to run where the data lies. This new approach to analytic application processing delivers the highest levels of performance, scalability, and deep, ad-hoc analysis for advanced analytics.

The Aster Data Solution

Advanced In-Database Analytics is a unique capability of Aster Data *nCluster* that moves beyond the constraints of running analytics in a traditional RDBMS database or on external application servers. By pushing complete analytic applications, not just analytic logic, into the data tier, Aster Data marries the power of standard SQL development with existing programming languages and new frameworks for analytic processing. This includes languages like Java, C/C#/C++, .Net, and Python, as well as new big data analytic processing techniques like MapReduce.

One of the first to face the challenge of analyzing petabyte-scale data, Google pioneered MapReduce to process large structured and unstructured data sets distributed across thousands of commodity hardware nodes. Until now, MapReduce has required specialized programming skills to take advantage of its powerful analytic capabilities. With Advanced In-Database Analytics, Aster Data brings the power of MapReduce to standard SQL with the patent-pending SQL-MapReduce.

All Advanced In-Database Analytics execute within Aster Data *nCluster*, a massively parallel database that runs on commodity hardware. Uniquely architected to support Advanced In-Database Analytics, *nCluster* maintains isolation between data management and analytic application processes in the database while at the same time ensuring that both data and application processes are treated as first-class citizens. This means that standard database services like fault-tolerance, workload management, and monitoring apply equally to data management and analytic application processes. In *nCluster* any existing analytic applications, including pre-packaged applications like R and SAS, run in-database, fully parallelized, and achieve significantly faster response times.

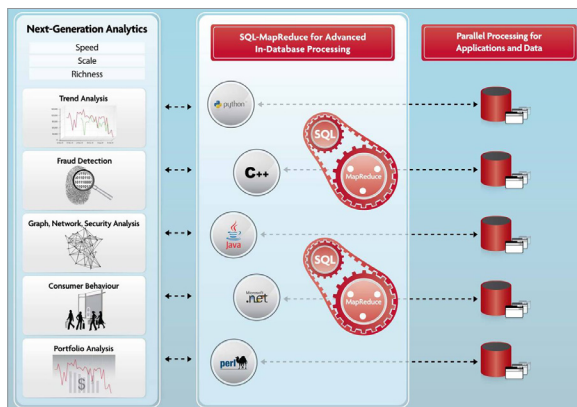


Figure 1: Architectural overview of advanced in-database analytics

Quick Overview

Aster Data Advanced In-Database Analytics is an integrated analytic engine delivered within Aster Data *nCluster* which parallelizes any analytic application.

Highlights

- Increases execution speed of advanced analytic applications with high performance, parallel processing
- Delivers advanced analysis on terabyte and beyond data sets—no sampling required
- Ensures that database processes are not impacted by analytic processing with in-database application isolation
- Rapid developer adoption with support for familiar languages like Java, C/C#/C++, Python, and .Net
- Avoid vendor lock-in typical when using proprietary UDFs

*“With Aster *nCluster* ... our data load time has decreased by over 95%, and our most important queries complete in seconds or less.”*

Tim Schigel, CEO
ShareThis



Scale Analysis to Large Data Sets

Terabyte to petabyte data stores are stressing traditional databases and data warehouses which are not designed to handle these massive loads. As a result analytics are delivered through sub-optimal workarounds that introduce data sampling and data latency.

Take fraud detection for example. Instead of running analytics on full data sets with near real-time updates, fraud analysts are required to prepare extracts of the original data, introducing data latency and limiting the scope of analysis. The result is undiscovered fraudulent transactions and late detection of fraud incidents. Other advanced analytics face similar challenges with sample bias and inaccurate analysis.

With Advanced In-Database Analytics all data stored in the database is available to every defined analytic function for deeper, more accurate data analysis. With *nCluster*'s massively-parallel architecture, Advanced In-Database Analytics scale linearly to many hundreds of database instances 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 constraints posed by data sampling.

Achieve High Performance Data Analysis

Advanced In-Database Analytics make sense not only for big data analytic effectiveness but also for analytic efficiency. What is the standard *nCluster* performance gain over more traditional approaches? Let's first look at a standard SQL based use case. In tests performed in 2009 a standard SQL query on Aster Data's *nCluster* performed 9x faster when using the SQL-MapReduce framework than when running SQL alone.

As processing increases in complexity with packaged analytic applications like SAS and R, Advanced In-Database Analytics continue to deliver greater 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).

Advanced In-Database Analytics also take care of high availability and fault-tolerance of analytic applications by isolating SQL-MapReduce execution from core database processes. SQL-MapReduce functions are executed in their own process, meaning that bad analytic queries result only in an aborted query, leaving other database jobs uncompromised.

Ad-hoc Analytics and Deep Data Exploration

A business intelligence key performance indicator (KPI) is typically a simple running counter or sum that can be computed incrementally from summary tables. This is why SQL has excellent support for business intelligence. Moving beyond KPI-style reporting to rich, ad-hoc analytics is challenging with standard SQL. Examples of rich ad-hoc analytics include computing statistical measures, identifying behavioral patterns, processing graph analysis, or simply performing time-series analysis. These techniques require iterative, multi-pass processing, which is resource heavy in a traditional, SQL-only data architecture.

With SQL-MapReduce, queries that require multi-pass SQL are simplified to single-pass queries that perform at petabyte and terabyte scale. To the *nCluster* developer, MapReduce is presented as a simple set of SQL table functions. So 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 they are pre-built or custom designed, are delivered in an extremely adaptive way that supports the evaluation of a SQL-MapReduce analytic function at run-time rather than design time. Why is this important? Unlike UDFs, which are rigid in their input and output data schema structure, SQL-MapReduce functions offer dynamic input/output data schema. This allows SQL-MapReduce functions to be coded once and then used many times. Once they've been implemented there is zero need to change their code or write additional declarations. Aster Data In-Database Analytics provide exceptional productivity for the data analyst, complete reusability, as well as expressive power for rich, ad-hoc analysis.

About Aster Data

Aster Data is a proven leader in big data management and big data analysis for data-driven applications. Aster Data's *nCluster* is the first MPP data warehouse architecture that allows applications to be fully embedded within the database engine to enable ultra-fast, deep analysis of massive data sets. Aster Data's unique "applications-within™" approach allows application logic to exist and execute with the data itself. Termed a "Data-Analytics Server," Aster Data's solution effectively utilizes Aster Data's patent-pending SQL-MapReduce together with parallelized data processing and applications to address the big data challenge. Companies using Aster Data include Coremetrics, MySpace, comScore, Akamai, Full Tilt Poker, and ShareThis. Aster Data is headquartered in San Carlos, California and is backed by Sequoia Capital, JAFCO Ventures, IVP, and Cambrian Ventures, as well as industry visionaries including David Cheriton, Ron Conway, and Rajeev Motwani. For more information please visit www.asterdata.com, or call 1.888.Aster.Data.

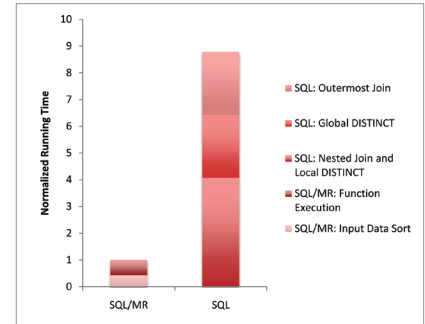


Figure 2: Runtime breakdown of SQL and SQL-MapReduce clickstream analysis.

“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

