10 Top Considerations for Big Data Management and Advanced In-Database Analytics
EXECUTIVE OVERVIEW

The era of big data management and deep analysis of massive amounts of data is opening new opportunities for significant competitive differentiation using advanced analytics. There are ten strong reasons why competitive organizations are turning to new data management solutions to handle their growing data volumes and evolving analytic needs. This new platform—a ‘data-analytics server’—merges data storage and data processing into one system to conquer the big data challenge. Big data storage is handled by a massively parallel database architecture; big data processing is handled by an integrated analytics engine, so that analytics run fully in-database. The results are cost-effective data storage, high-speed performance, and richer data analysis.

COST-EFFECTIVE DATA MANAGEMENT – BIG DATA ON COMMODITY HARDWARE

Every 18 months, an organization's data doubles. Most organizations are now faced with terabytes of available data for analysis. Terabyte to petabyte size data stores are stressing traditional database and data warehouse models, which cannot cost-effectively handle these massive loads. The hardware costs associated with traditional systems as they are pushed beyond terabytes become amazingly difficult to justify. Worse, business-critical analytic queries that should run in seconds to minutes instead take hours. As a result, costly hardware-dependent architectures force organizations to discard data that should be retained for a more thorough analysis of the business.

A new technology approach is required. What is needed is a massive parallel processing (MPP) data management architecture that parallelizes queries across a cluster of hardware devices rather than limits performance to a single box. This MPP architecture should parallelize all data management tasks, including queries, loads, back-ups, and exports. Not all MPP systems are alike. Those that parallelize all tasks provide 10x-1000x better performance than traditional relational database systems.

BEYOND HARDWARE COST – EASY AND LINEAR SCALABILITY

On average 70-80% of any organization’s data lives outside the traditional data warehouse. This data is generated rapidly and stored in new, rapidly-growing data stores and includes Internet, micro-transaction, and machine generated data—often in the form of log files. Traditional relational database systems have not kept up with either the growth of data sources or the changing profile of data processing today.

To accommodate unpredictable increases in data volumes and processing, you must embrace an unlimited yet cost-effective scaling model that does not depend on big hardware. Scaling not only needs to be cost-effective but should be easy and incremental.

- **On cost:** The ability to use commodity hardware rather than customized appliances or expensive servers is a key part of making that affordable.

- **On ease:** You should be able plug a new bare-metal server into the local network and perform one-click incorporation through a web interface. You should not have to purchase and provision a completely new analytic appliance.

- **On linearity:** Your platform should be flexible enough to support independent scaling by function (e.g. loading, backup, query processing) so your scaling model meets your specific needs.
STORAGE AND QUERYING – OPTIMIZE FOR BOTH

Big data is only one of two critical trends in the market. The other is a shift toward advanced analytics. Leading organizations differentiate from laggards in today’s data-driven economy by the ability to go beyond standard reporting and business intelligence, to identify patterns and behavioral trends, and to perform predictive analysis.

Some systems and architectures are very efficient at storing data (e.g. high compression, disk-heavy configurations) but trade that off against performance on the full range of analytic queries—reporting queries and advanced analytic queries. A strong solution needs to provide both.

ADVANCED ANALYTIC QUERIES – MAKE IT EASY

Delivering deeper business insight mandates systems that offer advanced analytic techniques. Standard reporting and business intelligence provide benchmark visibility into the past performance of your organization. These queries are predictable, predefined, and well-known. But more advanced analytics—including pattern analysis, churn prediction, trend analysis, behavioral and predictive analysis, etc.—provide insights into the present and future and are more exploratory in nature. This analysis is iterative and ad hoc, such that the discovery of one insight leads to another question, iteratively, until a model is honed. Traditional systems were not built for this type of analysis.

To meet the requirement for ad hoc, iterative analysis, look for a system architecture that can expand as your analytic needs grow beyond simple reporting and business intelligence. Key features should optimize both predictable query workloads as well as iterative, multi-pass query workloads. A data warehousing and analytics platform should provide the highest levels of performance across the continuum of analysis that your organization needs both today and in the future.

CONCURRENCY AND PERFORMANCE – MANAGING DIVERSE WORKLOADS

Traditional data warehouses and analytic platforms are being taxed as never before to keep up with the growth of new users. In a recent IDC survey, over 50% of respondents said that as they added more users to their data warehouse, performance degraded substantially.

Meeting the needs of user concurrency means predictable performance and guaranteed service levels for mixed, diverse workloads. A system that supports high concurrency on simple workloads does not necessarily perform at that same level as workloads get more complex. To ensure that time-critical queries are processed immediately, you should be able to dynamically reallocate CPU and storage resources based on in-progress transactions. This dynamic workload management should work on hundreds of mixed concurrent workloads. A measurement should not be based on logged-in users, but on active, concurrent workloads.
ADAPTABLE ANALYTICS –
REUSABLE PRE-DEFINED OR CUSTOM-BUILT
FUNCTIONS

Business intelligence and reporting can be performed with standard SQL queries, but advanced analytics requires techniques that stretch the limitations of standard SQL. Advanced analytics includes computing statistical measures, identifying behavioral patterns, processing graph analysis, or performing time-series analysis on deep, granular data sources like micro-transactions, click streams, or system logs.

The new analytic framework SQL-MapReduce enables a diverse set of advanced analytics to be performed easily. Pre-built and custom SQL-MapReduce analytics are extremely flexible, with support for dynamic input and output data schema. This allows SQL-MapReduce functions to be coded once and then used many times. Look for a system that runs SQL-MapReduce analytics completely in-database, co-located with your data for exceptional data analyst productivity with complete reusability and the necessary expressive power for rich, ad-hoc analysis.

IN-DATABASE ANALYTICS –
MOVE YOUR ANALYSIS TO THE DATA

While traditional databases can run application logic beyond SQL—such as stored procedures or user-defined functions (UDFs)—these approaches are often constrained. For the near real-time responsiveness leading organizations require, you must be able to run 100% of the analytics in-database. The underlying architecture for execution must maintain isolation between data management and analytic application processes and at the same time ensure that both data and application processes are treated as first-class citizens. Only through this architecture will you see the true performance benefit of in-database analysis, often 8x-10x improvement over out-of-database approaches.

This platform approach should also allow you to develop and execute a wealth of analytics written in standard languages such as Java, C/C++, C#, .Net, Python, and Pearl inside the database. New, innovative frameworks for analytic parallelization like MapReduce should also be embedded natively for in-database analysis to provide automatic parallelization of in-database analytics.

ALWAYS-ON AVAILABILITY –
END YOUR MAINTENANCE WINDOWS

As analytics become more critical to business operations, users will expect always-on database availability. This means your platform must allow simultaneous load/export during queries, online backup and recovery, online restoration, and online scale-out capabilities, eliminating the need for scheduled downtime.

Of course, not all downtime is scheduled. Unplanned outages can occur due to hardware failure or local and site disasters. To prevent these types of failures, you need massive-scale fault tolerance with replication, automatic failover, NIC bonding, failure heuristics, and clustered backup. If a system component goes down, intelligent load redistribution should minimize overhead on surviving nodes, with parallel replica restoration allowing immediate recovery to optimal system state.
EASY TO USE – FOR BOTH ANALYSTS AND DEVELOPERS

For widespread adoption of advanced analytics, you need a platform that is dead simple for both developers and business analysts. The developer requires an intuitive, fully-integrated development environment for coding, testing, and deployment of rich analytic applications. This should be complimented by an extensive suite of pre-built analytic functions to accelerate application development. For the business analysts to take advantage of the system, applications written by developers should be accessible through standard SQL or SQL-based tools. The analyst can then extend and modify application parameters without requiring lengthy development cycles.

DEPLOYMENT CHOICES – DON’T RESTRICT YOURSELF

There are many different types of analytic application projects and initiatives. Some are developed for short-term competitive or strategic insight; others are developed as long-term operational applications. Traditional database systems force you to make heavy upfront hardware investments, constraining the ability to support shorter-term but highly impactful projects.

Your organization should have a choice of deployment options, including on-premise software deployment, on-premise appliance deployment, and cloud deployment.

CONCLUSION

You probably have many questions about what big data management and advanced analytics can mean for your company and how to get started. This new platform called a “data-analytics server”—the merging of data storage and analytic processing in one platform—is a powerful tool for your organization to conquer the big data challenge. Big data storage must be handled by a new massively parallel database architecture; big data processing must be handled by a complete and integrated analytics engine, so analytics run fully in-database. The result is cost-effective data storage, high-speed data analysis, and richer analytics.

To learn more about how this new platform has delivered value to organizations like yours, visit our customer use cases at http://www.asterdata.com or contact us directly at 1.888.Aster.Data.

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 solution effectively utilizes Aster’s patent-pending SQL-MapReduce together with parallelized data processing and applications to address the big data challenge. Today companies with 10’s to 100’s of terabytes of data turn to Aster Data for their big data management and advanced data analysis needs.