white paper

Growing Revenue with Frontline Data Warehouses

Contents

Executive Summary

The frontline data warehouse is the newest innovation in analytics and data warehousing for organizations building competitive advantage through analytics. The frontline data warehouse is designed to collect, manage and monetize the avalanche of data entering today's enterprises. Whether fed by data from hundreds of applications, thousands of ATMs, or millions of mobile phones, the frontline data warehouse provides fast access to a persistent, comprehensive, and up-to-the-minute record of the enterprise's frontline data. The frontline data warehouse provides agile in-database analysis that enables intelligence at the point of service delivery. With this, you can empower your applications and users to rapidly generate insights and rule updates that grow revenue even as data stays resident within the frontline data warehouse.

Architected to serve the needs of frontline applications and their users and analysts, the frontline data warehouse follows **three key design principles** that distinguish it from a **traditional data warehouse or analytic appliance**:

1. **Performance at scale**: Unlike a traditional data warehouse or analytic appliance, even as data scales to petabytes, the frontline data warehouse must continue to meet strict service levels for data loading, query performance, and speed of administration
2. **Continuous availability**: Unlike a traditional data warehouse or analytic appliance, an frontline data warehouse has to be available 24x7, even as planned and unplanned events occur due to component or network failures and system administration
3. **In-database analytics**: Unlike a traditional data warehouse or analytic appliance, an frontline data warehouse supports revenue-generating applications and their users, and needs to provide data analysis and transformations in-database.

Aster Data Systems now introduces Aster *n*Cluster 3.0, the market's most capable and cost-effective analytic database for frontline data warehousing. It provides high performance at massive data scales and solves key management challenges.

- The *n*Cluster architecture provides high performance at large data scales with production deployments managing hundreds of terabytes of data. We call this our **Always Parallel** architecture.
- *n*Cluster "Live Administration" functionality coupled with data replication provides unrivaled reliability and continuous availability. *n*Cluster is **Always On** – always ready to provide service.
- *n*Cluster provides the first enterprise-grade in-database analytics capability for the frontline data warehouse. We call our analytics framework **In-Database MapReduce** and it provides a powerfully-expressive analytic programming framework.

Just as important, Aster technology delivers business success at the lowest possible total cost. Aster *n*Cluster works on industry standard x86 servers and requires no special

components or operating systems. In addition, through careful engineering and attention to the automation of routine administration, Aster *n*Cluster minimizes training and ongoing staffing costs to provide an unmatched low administration cost.

The "Big Data" Era

Wired Magazine has memorably called our era "The PetaByte Age"[1] – an age of infinite storage, clouds of processors, and most importantly, exponential amounts of data generation and capture. We at Aster Data Systems prefer to call it the era of "Big Data" because the scale of the data involved is now well past the petabyte level. Google alone processes multiple exabytes of data per year; MySpace records 7 billion user events per day; and the latest physics super-collider produces more than a petabyte of data per month. Even a few years ago, these processing and data management levels would have been almost unthinkable. Added to the volumes of data is the relentless increase in the speed of competitor response in the marketplace. With the "flat world" of the Internet, the window that businesses have to attract, upsell, and retain customers is shrinking. Today's businesses face ever-shrinking response times to their prospects, customers and market. Almost every organization we talk to, from small services companies to large Fortune 500 enterprises, from global social networks to regional Internet retailers, is now busy devising and implementing strategies to digitize and monetize business data at scales and speeds never before contemplated.

Over the last decade, the diversity of data captured by the enterprise has exploded. This has been fueled by ubiquitous high speed networks, the ever decreasing cost of powerful microprocessors, and the infiltration of digital measuring devices into every part of the commercial world. Businesses routinely capture a huge diversity of information from an astonishing array of source points that include:

- Point of sale records
- Email/SMS
- Technical support records
- Call center speech-to-text conversions
- Mobile WWeb click-stream
- Meter/device events
- Blogs
- Content tags
- Call detail records
- ATM interactions
- Web click-streams
- Web offer trail
- Mobile GIS
- RFID trace records
- User-generated content
- Etc.

Long gone is the era when a customer record was a small set of rows within an ERP database. Today, the complete customer record sprawls over multiple operational systems and databases, in varying formats, with inconsistent persistence and retrieval. Web click-streams are held in Web logs, call-detail records (CDRs) live in telecom Operational Support Systems, and product and service blogs are dispersed all over the Web. Data volumes have become so overwhelming that businesses are often forced to throw away much of the detail gathered because their infrastructure simply isn't capable of cost-effectively storing it all – even as they acknowledge that they're throwing away potentially

[1] "The Petabyte Age: Because More Isn't Just More. More is Different" Wired June 2008 Issue 16.07

valuable detail on their prospects, customers, partners, and employees. As data sizes continue to increase, this problem will only get worse. IBM has recently projected, for example, that the average footprint of a ***single business user*** will be 16 terabytes in just over a decade.[2]

The Revolution in Analytics

Along with the explosion in the size and diversity of business data, there is an accompanying increase in the pervasiveness of analytics and the sophistication of algorithms used. As described in books such as "Super Crunchers" and "Competing on Analytics," analytics is becoming the new competitive frontier for business.[3] As industries are reshaped by the Internet, and the increasing transparency and availability of pricing, channel, and product data, analytics becomes the key competitive weapon for business.

When analytic databases got their start in the 1980's, they were rarely asked to do anything more complex than sub-set selection and comparison on highly summarized data: in retail, for example, something as simple as "Compare the average spend per customer on camping gear in December vs. April." This kind of analysis was crucial for inventory, purchasing, and store space planning (and still is). Another common marketing example was "Tell me the coupon redemption rate for a 10%-off coupon vs. a 15%-off coupon in a test market." Or in telecom, an equivalent example: "What is the average weekly talk-time in the three months before a small business orders a second phone line."

Since then, however, there has been an increase in the sophistication of the algorithms used for analysis. Just a few of the more popular algorithms and methodologies adopted over the last decade include:

- Regression modeling for identifying and quantifying business performance drivers
- Bayesian probability analysis for fraud and spam detection
- Time-series analysis to detect repetitive behavioral patterns
- Graph analysis to understand social network contagion effects
- Conjoint analysis for optimizing offer attributes

Some of these techniques – like conjoint analysis – have remained in the back-office, employed by statisticians, product managers and marketers working with gigabytes of data. But many of the other techniques are migrating into frontline applications as their value is demonstrated and recognized by general managers and CFOs.

Coupled with a better menu of techniques, there has been growing awareness in the general business community that companies like Wal-mart, Capital One, and Google are successfully using advanced algorithms and analytics to grow their customer base, increase profit margins, and control risk by working at a level of detail far more granular than the "segment summary" level that most companies are used to. Google, in particular, has built

[2] http://www.eetimes.com/news/latest/showArticle.jhtml?articleID=210600317
[3] Super Crunchers Ian Ayres, Bantam 2007; Competing On Analytics Davenport & Harris, HBR Press 2007.

its business by using simple algorithms on huge data-sets with a wide variety of data-types to improve ad monetization and search relevance.[4]

Perhaps the most important driver of the use of more advanced algorithms is that there is far more data – from far more sources – that reveal the actions, motivations, and behaviors of prospects, customers, employees and partners. Web click-streams and offer histories alone are a goldmine of information about segments, promotion effectiveness, and uplift. Effectively analyzing that data, and using those insights to capture more customers, increase revenue and margins, optimize channels, and control business risk, is the goal of the revolution in analytics.

Just a few of the frontline applications and functions where analytics now drives the business include: credit scoring, risk modeling, ad-targeting, fraud detection, spam identification, cross-sell/up-sell bundling, and recommendations. In these applications, the race is on for businesses to use analytics to respond as fast as possible to the market and competitors.

For those companies that do not prepare for analytics-based competition, the cost can be high. For Yahoo! – who has been unable to match Google's capability to predict which ads should be presented to which customers – the consequence has been a plummeting stock price, massive brand erosion, and a crisis of employee and investor confidence. For Kmart, the consequence of being unable to predict what inventory should be in what store at what time was bankruptcy.[5]

Implementing Analytics for the Business Frontline

A successful information technology implementation is a successful combination of people, processes, and technologies. Seen this way, creating a market-leading analytics capability is no different than successfully implementing an email system, a new Web application, or a virtualization infrastructure. Having the right people means having skilled business analysts, application developers, and technical specialists. Analytic development can often be outsourced to statistical consulting and integration firms; and each vertical and application specialty has several specialist consulting firms skilled in the implementation of specific analytic applications.

[4] http://anand.typepad.com/datawocky/2008/03/more-data-usual.html
[5] "Kmart files for Bankruptcy, Largest Ever for a Retailer" New York Times, January 23, 2002

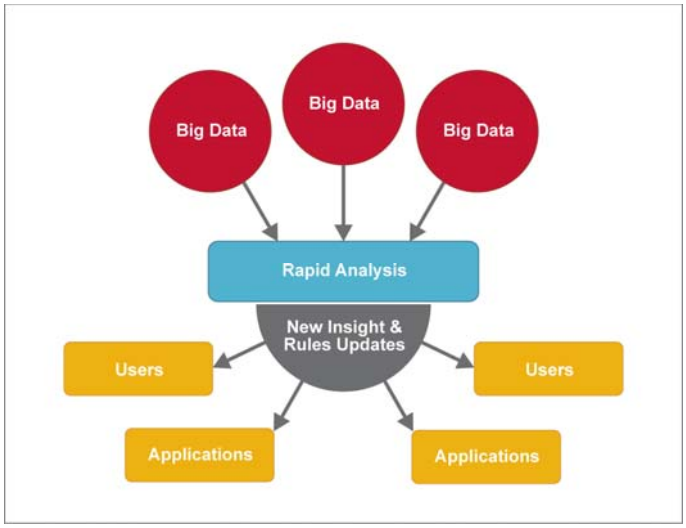

Figure 1: Essentials of the frontline data warehouse

Frontline Data Management

The new processes required to bring analytics to the business frontline are fundamentally about designing data management based on the need for fast, low-latency access to data.

In legacy data warehouses focused on back-office support, it typically takes a day or more for new business data to become available to analysts and applications. From the moment of data generation, data usually takes a multi-step journey through an OLTP database, an operational data-store or master-data manager, as well as an ETL (extract/transform/load) batch job. After all this happens, the data must wait for an available load-window before being loaded into the data warehouse. For the business frontline, 24 hours is an unacceptable amount of latency. Not only is the data too stale for frontline use, it has usually been over-summarized, sampled, or otherwise impoverished before it is made available to the data warehouse. To enable analytics for the business frontline, there are two alternative strategies. The traditional option is to attempt to scale the existing multi-step process (see Figure 2) and its infrastructure for the frontline requirements of large data volumes and low latency. This usually proves difficult, sometimes impossible and in any case exceedingly expensive in hardware costs. "Big Data" puts huge stresses on enterprise networks, and ETL tools and master data managers are simply not designed to cope with the volume and speed of transformations required.

Instead, a subtle paradigm shift is needed to create a new data management process (complementing the legacy data management process) that captures Big Data directly from source to eliminate summarization, sampling, and latency. In this data-flow, data transformations are performed either at the application source, while the data is in-flight to the frontline data warehouse, or in-database, once the data has been loaded into the FDW.

With this, frontline applications and users have access to a full source data-set that reflects the current state of the business without the burdens of heavyweight legacy warehouse and ETL data management.

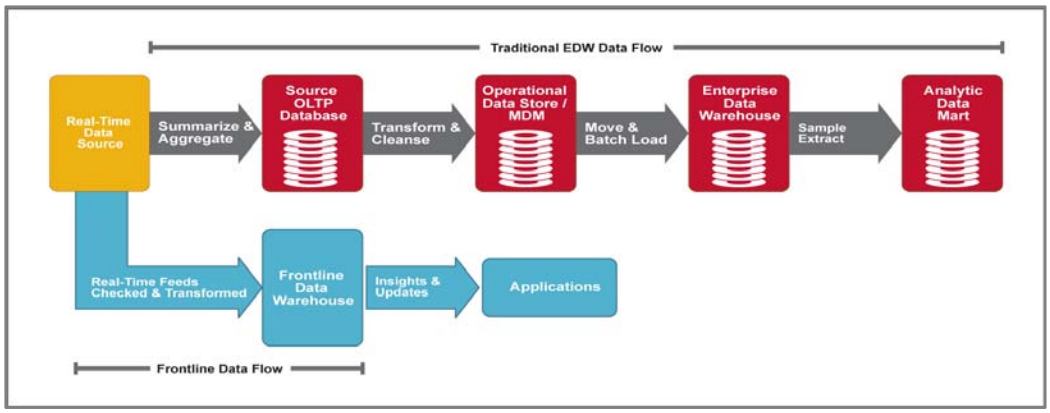

Figure 2: Data flow in traditional vs. frontline data warehousing

The Frontline Data Warehouse: A Unified Source for Frontline Truth

The final piece of implementation for a frontline analytics capability is to design and select a technology infrastructure that can capture, manage, store, and analyze the Big Data generated by the business frontline. We call this the frontline data warehouse – a database that can capture all the data generated by the business frontline in a single data repository accessible by applications and analysts. As the single source of truth for granular prospect, customer, device, and application data, the frontline data warehouse faces unique demands not usually encountered by traditional analytic appliances or data-marts, namely:

- Scaling to the hundreds of terabytes (soon to be petabytes) of data on the business frontline
- Providing continuous-service availability under component failure and routine administration
- Analyzing data in-place without moving data to an analytics engine

(The last point perhaps needs a brief clarification up front. When dealing with the new era of Big Data, it becomes increasingly expensive to transfer data over the network – for example, out of a database for processing in a separate analytics engine. In the era of Big Data, the application logic has to come to the data – the data can no longer come to the application logic. While a sample set (1-5% is typical) may work for some application and forecasting needs, in the era of Big Data, there is competitive advantage in analyzing a larger or full set of data. That's why it is vital that a frontline data warehouse have a significant in-place analytics capability.)

In addition to these unique criteria for the frontline data warehouse, the FDW also needs to support standard database capabilities such as a providing a SQL interface, workload management, security, and backup/restore.

The Frontline Data Warehouse in Context

It's perhaps worth taking a minute to compare and contrast the frontline data warehouse with more familiar analytic database deployments such as enterprise data warehouses, data marts and analytic appliances. Due to the unique demands placed on the frontline data

warehouse, both its features and its usage-profile are substantially different from these other types of analytic databases.

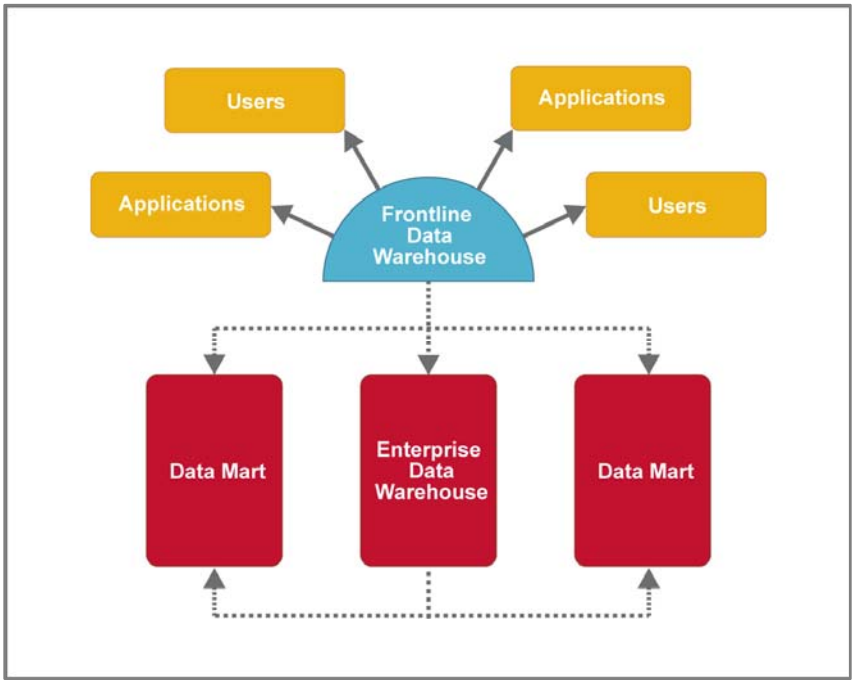

Figure 3: The frontline data warehouse in context

The Enterprise Data Warehouse

As the single source of truth for the entire enterprise, the enterprise data warehouse (EDW) faces unique organizational and technical stresses. Because the EDW serves the entire enterprise, it typically does not contain transient granular frontline data – some of which has a useful lifetime of less than a day, in some cases. In fact, the EDW tends to store summarized and historical persistent data – less useful for frontline applications trying to react quickly to market trends and events. In addition, once the hard work of establishing an enterprise data model for the EDW is accomplished, there is considerable reluctance to change it. As a result, the data model becomes difficult to evolve as new applications and analyses arise – an inflexibility that makes it hard for the business frontline to react quickly to opportunities in the market and against competitors. In short, the frontline is about speed, agility, and intimacy with data – criteria that the EDW was never intended to deliver against.

However, the frontline data warehouse is not intended to replace the enterprise data warehouse in a company's data and analytics strategy. EDW's serve valuable purposes for business reporting and cross-organizational consistency. A natural data management model is to have the frontline data warehouse feed the EDW with its familiar daily batch-loads in the normal course of business.

Analytic Appliances and Data Marts

Data marts and analytic appliances have proven useful for backline applications such as departmental analysis and reporting as well as business reporting over long time periods. However, data marts and appliances are not engineered for mission-critical environments like the business frontline, and instead focus on radical simplicity and ease of learning for less challenging environments. Significant system downtime for scaling, backup, loading and other routine administrative tasks is a common characteristic. And, they have notable deficiencies in their ability to support high volume data loads and rich queries without serious performance degradation. In addition, most appliances and data mart platforms are not designed for petabyte scalability on industry-standard hardware and networking.

As summarized in Figure 4, the frontline data warehouse takes EDW requirements for scale, fault tolerance, high volume loading, and rich queries - and today some vendors who claim EDW status for their products have trouble meeting even these criteria - adds data mart requirements for in-database analytics and cost-effectiveness, and meets new requirements for load isolation, and minimal planned downtime.

Key Requirements	Enterprise DW	Frontline DW	Datamart/Appliance
Petabyte scale	X	X	
Fault tolerance	X	X	
High volume load	X	X	
Rich queries	X	X	
Load isolation		X	
Minimal planned downtime		X	
In-database analytics		X	X
Cost effectiveness		X	X

Figure 4: Key requirements for the EDW vs. frontline data warehouse vs. data mart appliances

Key Elements of the Frontline Data Warehouse

The frontline data warehouse is designed to support frontline applications and their analysts, so it's perhaps worthwhile to talk about the common characteristics of these applications. In all these applications, whether we talk about fraud detection, ad targeting, or offer optimization, the key is that large amounts of data per customer needs to be captured and analyzed in a small time window, within the context of the historical information already captured for that prospect or product. The results of those analyses then create significant, order of magnitude differences in revenue generated for the business or in terms of business costs. The speed with which new data can be analyzed and incorporated into the business rules of the frontline applications is key to the competitive advantage generated by these applications. Finally, since these applications are generally

operating as close to 24x7 as possible, the supporting infrastructure has to meet those requirements as well.

As a result, the frontline data warehouse follows three key design principles:

1. **Scalability**: a frontline data warehouse can scale to petabyte data and peta-op analytics without encountering performance bottlenecks
2. **Continuous availability**: an frontline data warehouse avoids unplanned downtime with fault tolerant features and eliminates planned downtime and performance degradation for routine administration
3. **In-database analytics**: an frontline data warehouse provides programmable analytics that analyzes and transforms data in-place without off-loading to a separate processing platform

In the sections below, we describe how Aster *n*Cluster 3.0 meets these requirements with its "Always Parallel" architecture, "Always On" availability, and "In-Database MapReduce" as its analytics implementation.

"Always Parallel": Scale and Performance for Frontline Applications

Aster *n*Cluster provides maximum system performance and efficiency through a network-optimized massively parallel processing (MPP) architecture. The *n*Cluster architecture provides unlimited query scaling by parallelizing query and analytics execution across all the nodes in a cluster using a global, network–aware query optimizer. Adding more capacity is a simple matter of plugging a new bare-metal server into the local network and performing one-click incorporation through a Web-interface. *n*Cluster manages user data in locally-attached storage on individual servers within the cluster. Internally, the cluster consists of three separate classes of nodes: Queens, Workers, and Loaders.

The Queen servers provide the external single-system interface to the data warehouse. End-users and database administrators connect to a Queen through ODBC/JDBC, while systems administrators monitor *n*Cluster through the APIs provided by a Queen. The Queen servers are also responsible for coordinating the cluster servers in query processing, result aggregation, and failure handling.

Worker servers are responsible for storing partitions of data and replicas of data that reside at other Worker servers. Worker servers also participate in query processing and maintenance tasks (e.g., indexing, backups, load balancing) as invoked by Queen servers.

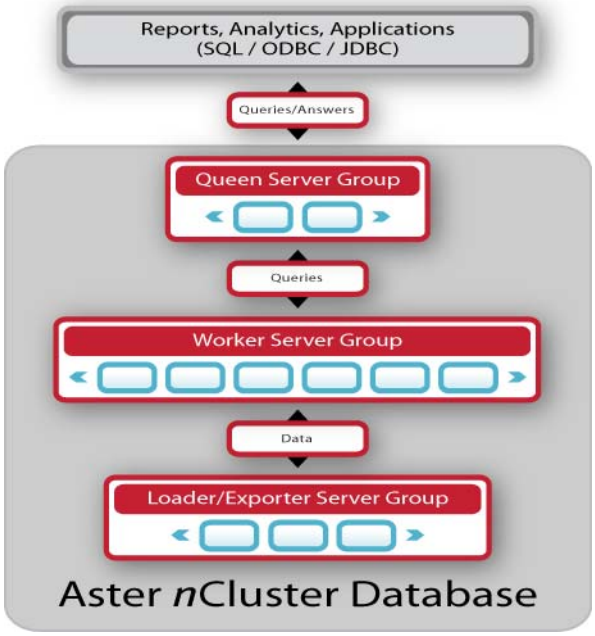

Figure 4: Aster nCluster architecture in use

The Loader servers are responsible for partitioning and loading of new data into the Worker servers. Additionally, the Loader servers can export data for use in other systems, disaster recovery, and data back-up.

The three-tier Aster *n*Cluster design encapsulates a task isolation philosophy. Each tier can be independently scaled in response to workload characteristics. For example, if the number of connections increases, the cluster can be populated with more Queen servers; if data volumes grow, the number of Worker servers can be increased; if faster loads are desired, more Loader servers can be provisioned.

Aster *n*Cluster also provides a high volume load/export capability. With proven capacity to load terabytes per day of fresh data, load execution is kept independent of query execution by *n*Cluster's unique task isolation. In addition, *n*Cluster provides scalable online backup and restore through fully parallelized, direct-to-server backup and restore. Backups run non-intrusively in the background ensuring fast backup and fast restore for even the largest frontline data warehouses.

"Always On": Availability Strategies for Mission Criticality

Aster Data Systems *n*Cluster provides hands-free 24x7 availability and resiliency for mission-critical frontline users and applications. According to Gartner research, the ability for data warehouses to support business critical applications is becoming a must-have requirement.[6]

[6] Gartner, Inc. "Operational Analytics and the Emerging Mission-Critical Data Warehouse". Beyer, Mark. 14 May 2007.

"More than 90% of Global 2000 companies plan to incorporate analytics into multiple operational applications by 2010," said Gartner vice president and distinguished analyst Mark Beyer, "but fewer than 15% of data warehouses have been designed for high availability, failover, disaster recovery and the remaining components of mission-critical systems." [7]

Aster *n*Cluster provides 24x7 planned uptime by distributing data replicas across the server cluster. These replicas activate immediately in the case of a server failure or even temporary unavailability. In the event of failures, the system can then recover lost servers without downtime. In addition, Aster *n*Cluster provides "Live Administration" – the ability to repartition and rebalance data across the cluster without requiring database downtime. Because of Aster's task isolation, there is no such thing as "load windows". By separating load and export tasks onto separate servers, query execution is protected from load/export workloads.

To summarize, careful Aster engineering has resulted in a platform that delivers a uniquely compelling value proposition around continuous availability. Aster *n*Cluster can answer queries at peak performance even when it is:

- Experiencing a hardware failure
- Recovering from a hardware failure
- Adding capacity
- Performing backup
- Loading data
- Exporting data

Availability characteristics such as these are generally only found in OLTP databases. With Aster *n*Cluster, these capabilities now come to the world of analytic databases.

In-Database Analytics

While there have always been capabilities within the database for performing application logic – such as stored procedures or user-defined functions (UDFs) - these have largely been abandoned as a programming strategy over the last decade. With the rise in application servers in the late 1990's, there have been few significant advances for in-database application execution. But now, Aster *n*Cluster provides a first in the database world: Aster *n*Cluster In-Database MapReduce. MapReduce is a programming model invented at Google in 2003 to process large unstructured data-sets distributed across thousands of nodes. Now Aster *n*Cluster brings that capability to the world of structured data. In-Database MapReduce functions are simple to write and are seamlessly integrated within SQL statements. They rely on SQL queries to provide input and manipulate data. (For more details, please see our technical whitepaper on MapReduce.)

Delivering Cost-Effectiveness

Perhaps equally as important as the technology features of Aster products are its deployment and pricing model. Aster *n*Cluster has been designed from the ground up to run on industry standard x86 servers and networking, providing a cost-effective

[7] Gartner, Inc. "Operational Analytics and the Emerging Mission-Critical Data Warehouse". Beyer, Mark. 14 May 2007.

commodity platform for customers to implement their frontline data warehouse. It's important to emphasize the advantages derived from this:

1. System administrators don't have a new hardware platform to learn how to install, monitor and manage, including minutiae such as new cooling and rack requirements.
2. Enterprise purchase discounts from major server providers can be leveraged for the data warehouse deployment.
3. If there are urgent requirements to add capacity or performance, existing hardware in the organization can be re-deployed. The business is not hostage to the procurement speed of a vendor of proprietary hardware

The second aspect of cost-effectiveness is delivering a low cost of training and ongoing maintenance. Aster *n*Cluster has been engineered so that almost all routine administrative tasks are automated within the database. Skew adjustment, partitioning, and failure recovery are just some of the routine tasks automatically handled by the database without the need for administrative intervention. Unlike traditional data warehouses, Aster *n*Cluster doesn't make you wade through 10,000 pages of documentation just to deploy a data warehouse. Training is easy, and management is intuitively presented through a simple Web interface.

In addition, the Aster pricing model is based only on the size of the user data stored in the warehouse, meaning that expanding the frontline warehouse as additional applications and users leverage the data won't result in additional license expense.

Frontline Data Warehouses: Infrastructure for Data Monetization

Companies are building mission-critical applications that depend on detailed analysis of large data volumes to drive operations and profitability. Traditional back-office data warehouses can't cost-effectively meet the availability, performance, and analytic depth required of this new generation of data-driven applications.

For these enterprises, Aster *n*Cluster provides a frontline data warehouse that provides an always-on, always-parallel MPP architecture with the first-ever In-Database MapReduce programming framework; all in a cost-effective software platform that provides an easy-to-use, easy-to-manage model that's familiar to DBAs and system administrators.

For further information, visit our Website at www.asterdata.com , call us at 1-888-Aster-Data, or email us at info@asterdata.com. We'd love to start a conversation with you about your frontline data warehousing plans.

About Aster Data Systems

Aster Data Systems is a proven innovator in high-performance analytic databases for frontline data warehousing – bringing deep insights on massive data analyzed on clusters of commodity hardware. Co-founded in 2005 by three colleagues in the Stanford Computer Science Ph.D. program, the Aster *n*Cluster database provides patent-pending innovations in performance, availability, and in-database analytics. Aster is headquartered in Redwood City, California and is backed by Sequoia Capital, Cambrian Ventures, and First-Round Capital.

For more information please visit us at http://www.asterdata.com, write to info@asterdata.com, or call 1-888-ASTER-DATA.