

# Big Data Summit

**2010** Bay Area



## Making Advanced Analytics on Big Data Fast and Easy

Tasso Argyros, CTO, Aster Data

Molly Stamos, Director of Product Management, Aster Data

Westin SFO

February 18, 2010

**aster data**  
big data. fast insights.™

*The reality of big data has driven the emergence of a **new generation of analytics...***

# The Evolution of Analytics

## Advanced "Big Data" Analytics

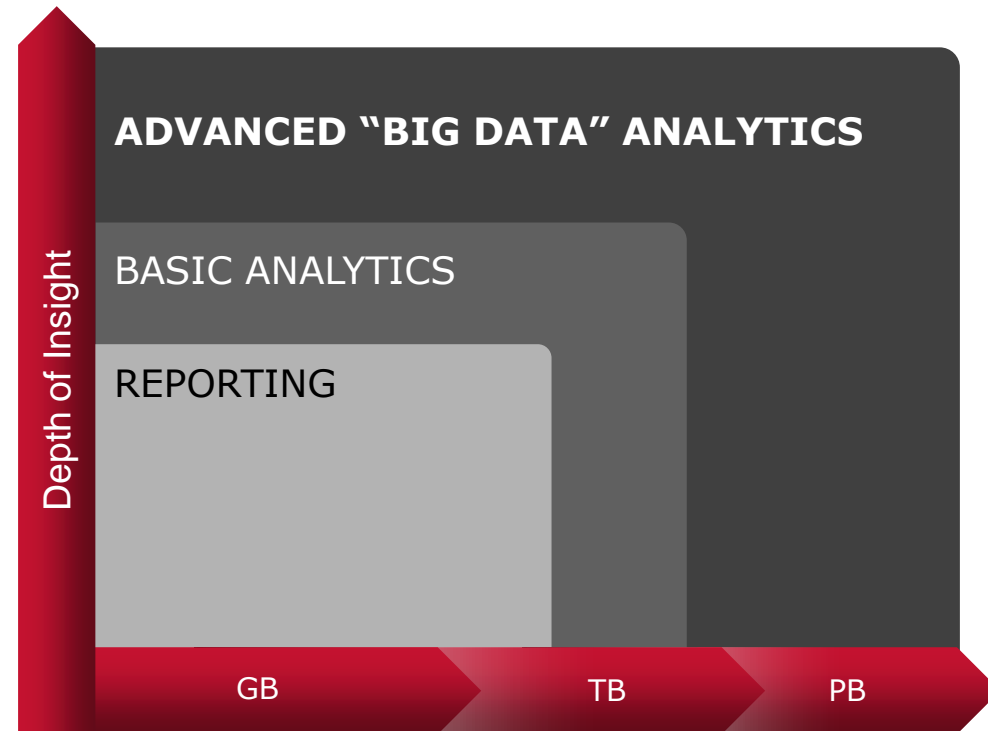
- **Rapid, rich** analysis of **big data sets**

## Basic Analytics

- Simple operations on limited amounts of data

## Reporting

- Summaries and dashboards of large volumes of data from limited numbers of sources



# Requirements for Big Data Analytics

## SCALE

- Analysis of **terabytes to petabytes** of data — not just summaries
- Including both **current and historical data**

## SPEED

- **Frequent analysis** – on-going, frequent analysis (e.g. daily, weekly)
- **Fast results** — insights in minutes/seconds

## RICHNESS

- **Deep data exploration** – complex analysis on massive data sets to find rare events, common patterns, outliers
- **Ad hoc, interactive analysis** – versus simple structured reports

# What is Big Data Analytics: Example 1

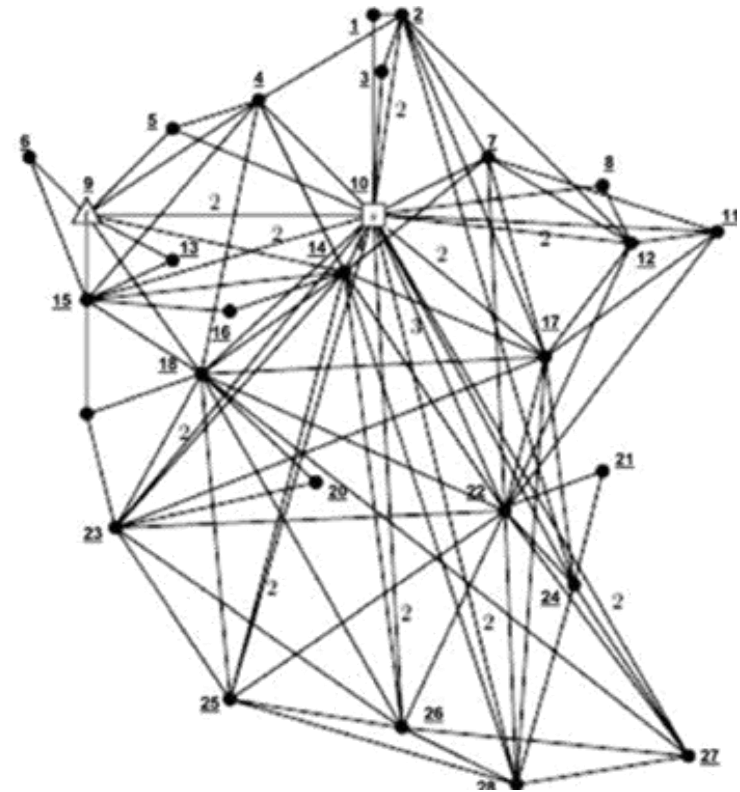
## Deep graph analysis

### Business Problem

- *Retailer*: How can we design marketing, packaging, and promotions to target key segments?
- *Telco*: What are the common calling patterns for a specific user group?

### Analytics Problem

- What are the most important clusters and interconnections?
- What are the patterns within a cluster or set of interconnections?



- Difficult to express in SQL
- Requires repeated iterations through data

# What is Big Data Analytics: Example 2

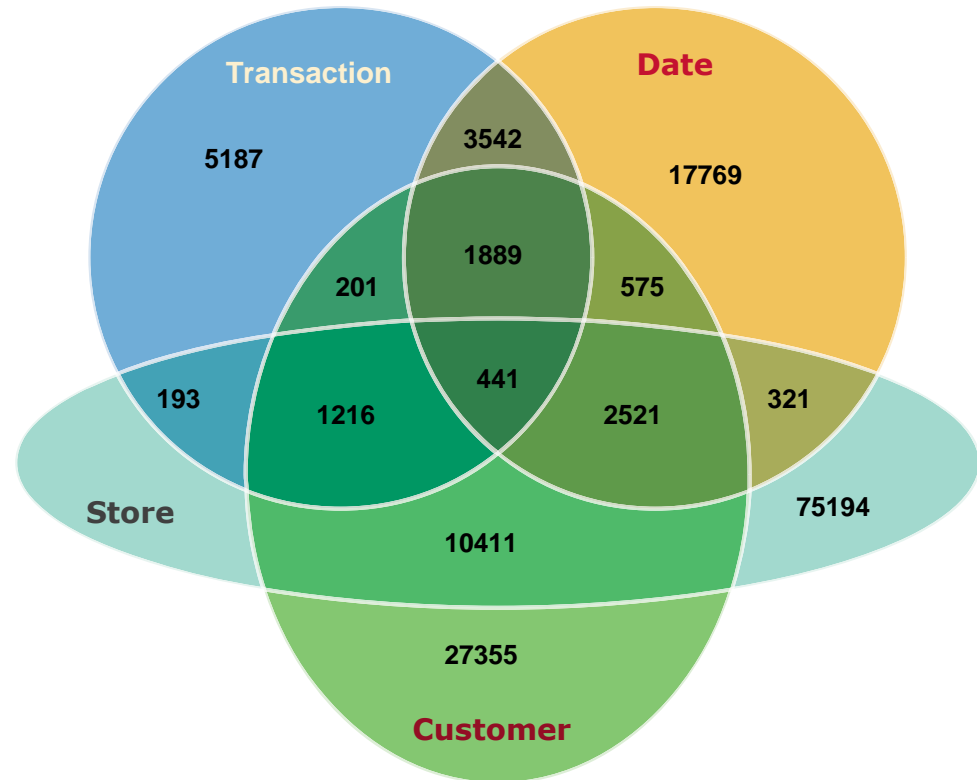
## Complex market basket analysis

### Business Question

- How do we optimize store stocking schedule, product placement, and promotions?

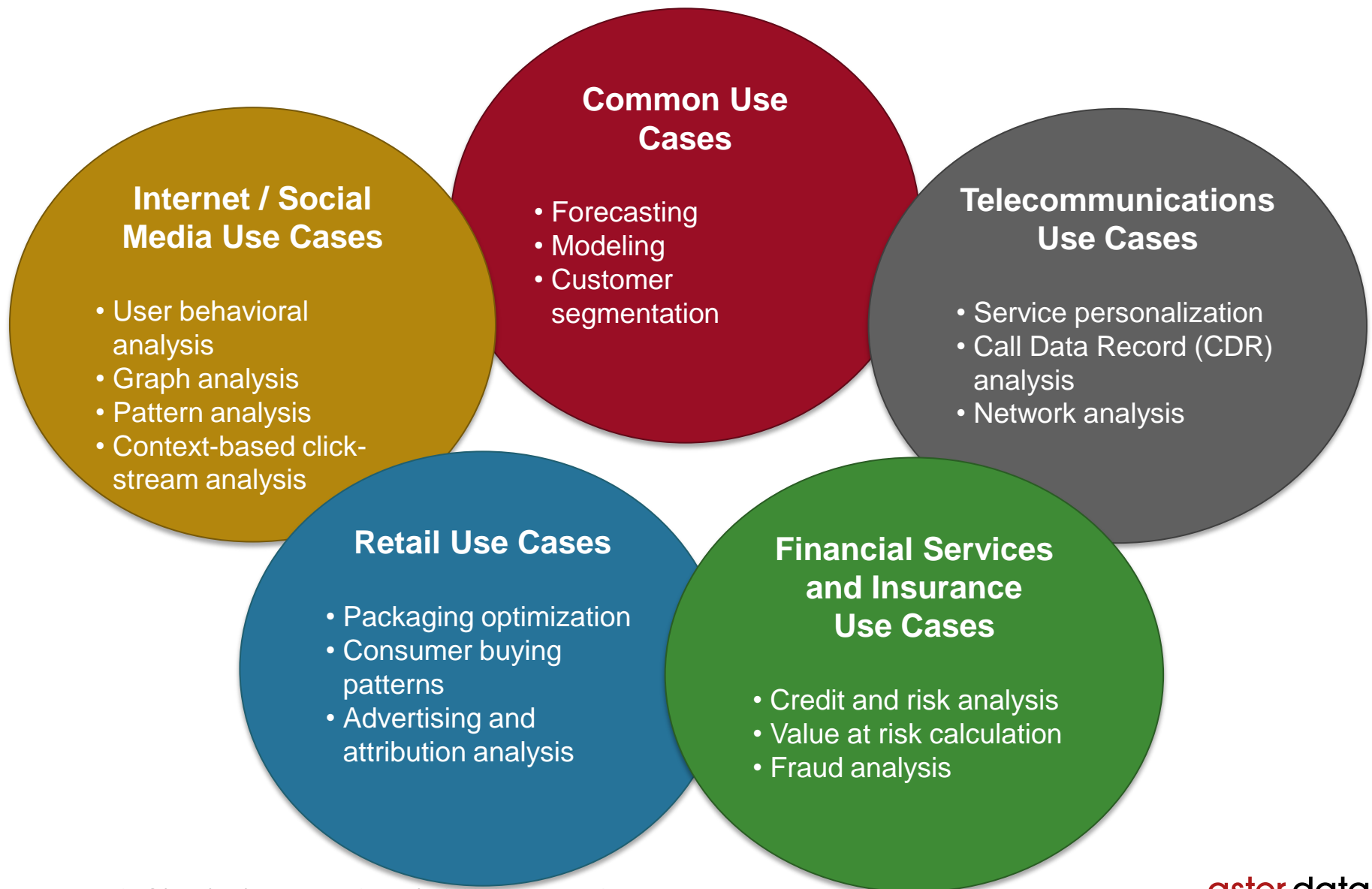
### Analytics Problem

- How to determine what items are commonly purchased together (by a customer, in a store, during a specific date range, etc.)?



- Needs full data set – sampled data has too little detail
- Requires repeated iterations across data

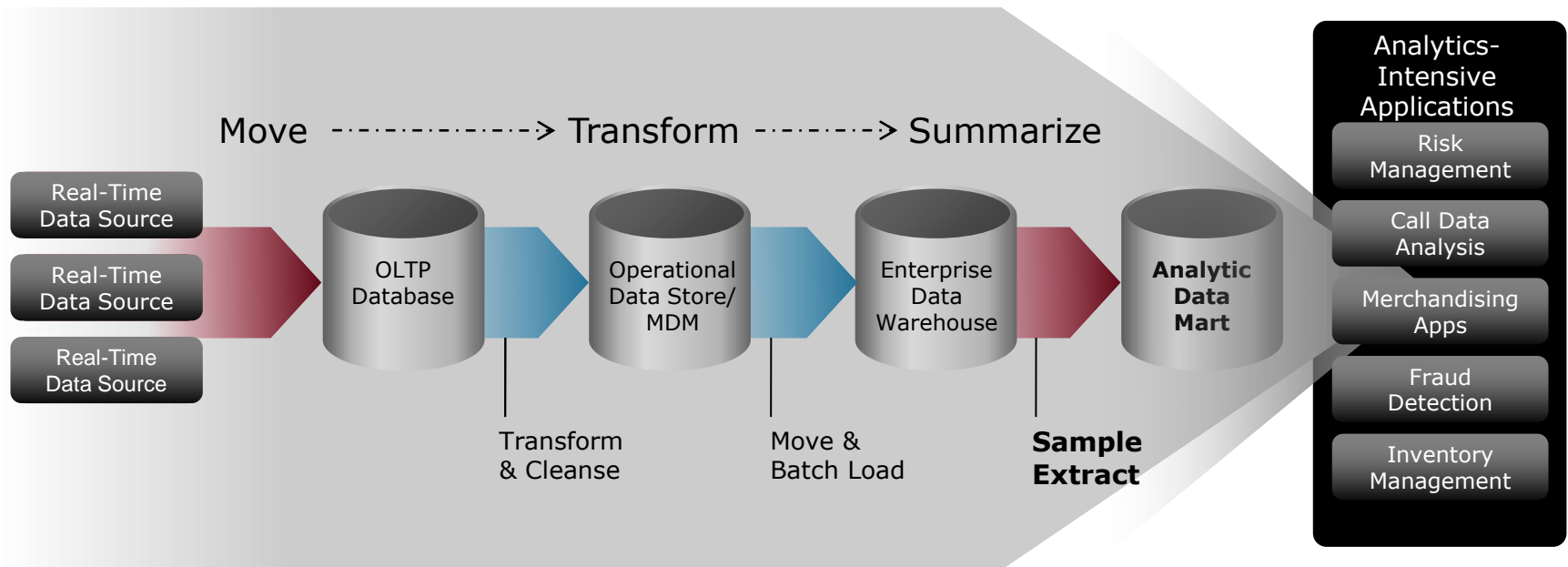
# More Examples of Big Data Analytics



*...but traditional systems and data architectures were **not designed for big data analytics...***

# 1. Traditional Data Pipelines Cannot Meet the Needs of Big Data Analytics

Rapidly growing volumes of data strain the traditional data pipeline to the breaking point



- Impossible to move TBs to PBs of data through the pipeline
- Forced to work with summaries and samples of data
- Cycle time from data to analysis = hours to weeks

## 2. Big Data Analytics Applications are Difficult to Build and Manage

The complexity and overhead of development, validation, and deployment limits implementation

### **Constrained by SQL limitations**

- SQL was designed for operational processing, not large-scale, complex analytics
- As a result, complex multi-step analytics logic is difficult or impossible to express in SQL
- Traditional UDFs are extremely hard to develop and not easily parallelized

### **Hindered by the difficulty and inefficiency of developing and managing applications**

- Complex to code and integrate
- Difficult to test and validate
- Costly to manage because of limited visibility and significant coordination required



*Solving these challenges requires a **new system and architecture** designed for big data analytics—*

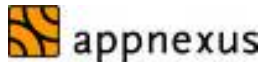
*a Data-Application Server*

# Introducing Aster Data

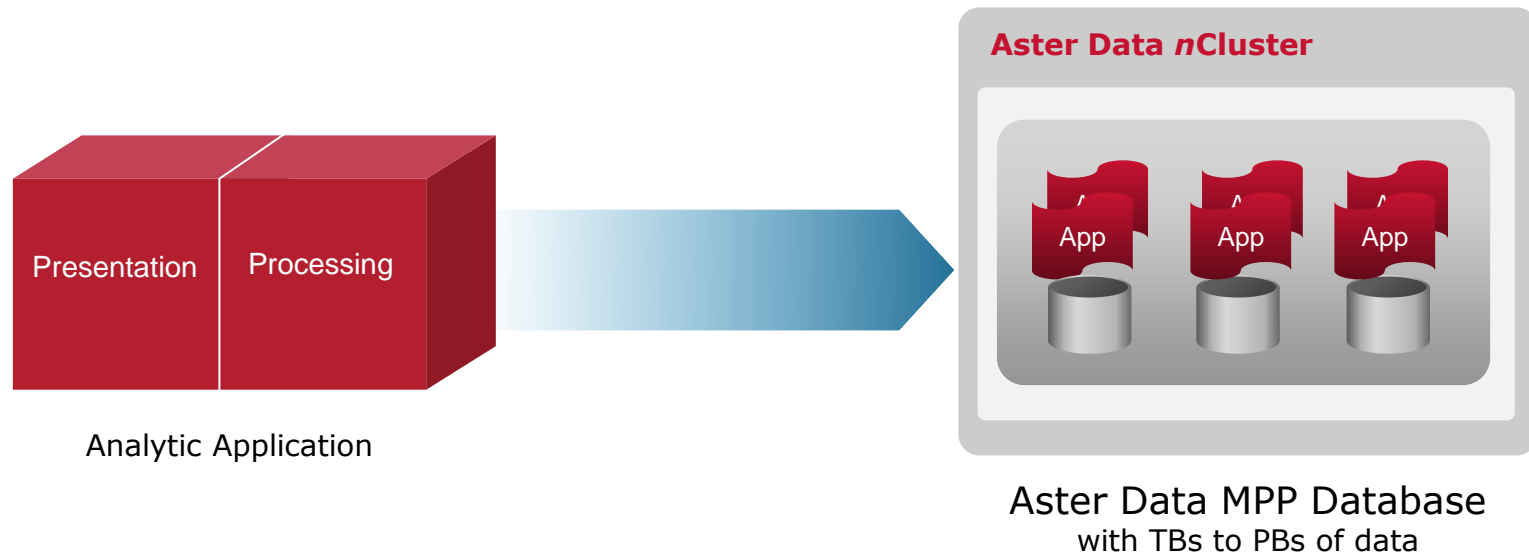
- Founded to create a platform to address the big data challenge:
  - Store and manage TBs to PBs of data AND
  - Enable ultra-fast, advanced analytics on large data volumes
- Rapidly growing customer list across industries and use cases



- Partnered with industry leaders

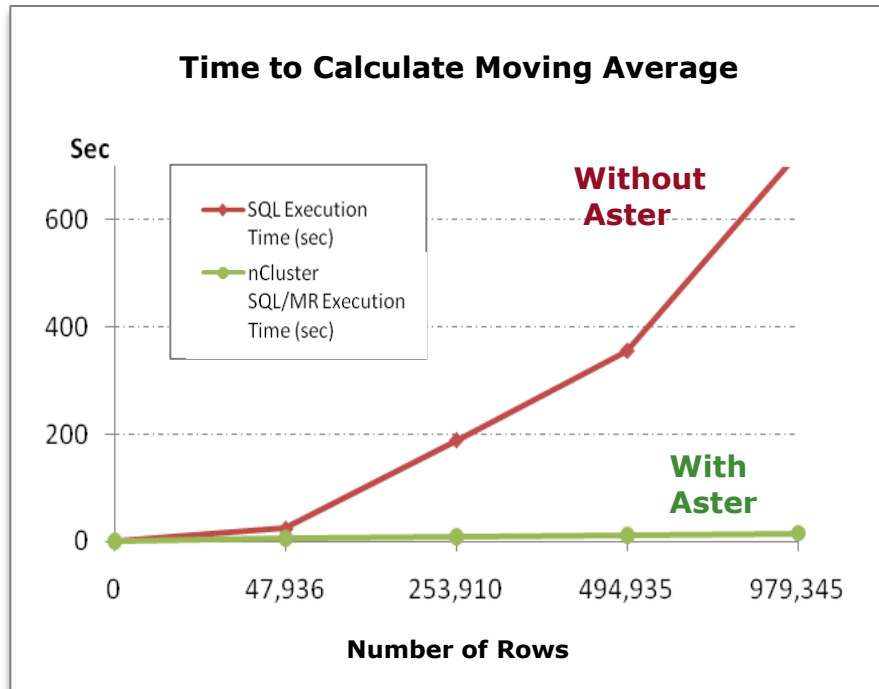


# Aster Data's Solution: Massively-Parallel Database Designed for Big Data Analytics



- Provide a high-performance, highly-scalable data management platform
  - **MPP architecture:** For unlimited scalability & high performance.
  - **In-database MapReduce:** For distributed processing on large data sets.
- Move applications to data to minimize costly data movement
  - **Application Push Down:** Move applications into the database next to the data
- Enable easier, faster development of analytic applications
  - **SQL-MapReduce:** Integrating SQL & MapReduce for speed and richness of analytics

# Result: Ultra-Fast, Scalable Analytics



- Much **shorter and less complex** to implement – half as many lines of code
- Execution is **significantly faster**, particularly over large numbers of rows
- Execution time **scales almost linearly** as amount of data grows
- and...**more flexible and repeatable**

# Real Results



## Analytic Application: “Media Metrics 360”

- Utilizes database with 320 billion rows
- Adding 2 billion rows of data per day, growing to 6-10 billion rows per day or 6TB of uncompressed data
- Analytics run daily with a few dozen business analysts concurrently running analysis

## Why Aster Data?

- Linear, one-click rapid scaling
- Query performance on commodity hardware exceeded their previous system
- Easily usable by business analysts (power of SQL-MR)
- 10x less costly than every other alternative considered



*Enables digital marketers and SEO professionals to dramatically improve organic search performance*

## Analytic Application: Collect web traffic and conduct **deep trend analysis** on how users search sites

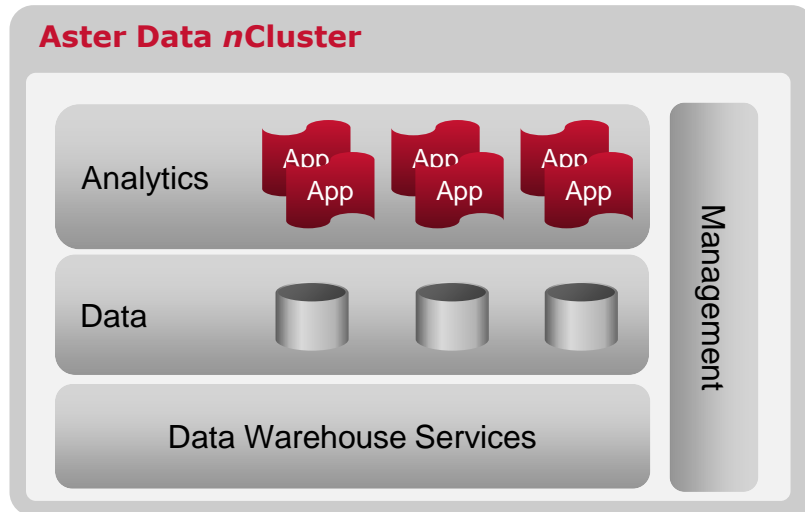
- 7 TBs of data with database being updated every 5 minutes with fresh data
- Need for ad hoc, real-time reporting for clients (e.g. REI, Overstock.com etc.)
- Complex mixed workload comprised of updates and read processing

## Why Aster Data?

- Queries that used to take 45 minutes now run in less than 7 seconds
- Previously impossible queries now execute in 20-30 seconds.
- Complete scan of database used to take 30 hours, now with Aster Data takes <1 minute
- Scaling the system takes <1 hour (other solutions considered took days)

*Introducing the first massively parallel  
database that runs applications inside to  
make next-generation analytics possible*

# Aster Data *n*Cluster: Massively Parallel Database for Big Data Analytics



*“The platform takes a different approach from traditional data warehouses, DBMS and data analytics solutions by housing data and applications together in one system, fully parallelizing both.”*

*“..a set of extensions to SQL called ‘SQL-MR’ has achieved growing, but not yet universal, adoption...”*

**Forrester Research, Nov. 2009**

## Advanced analytics via application embedding

- Run analytics applications inside the database to minimize costly data movement
- Support analytics applications without rewrite

## Advanced optimizations

- Dynamic Workload Management
- 10-1000x faster on standard SQL queries via intelligent query optimizers

## “Always on” availability

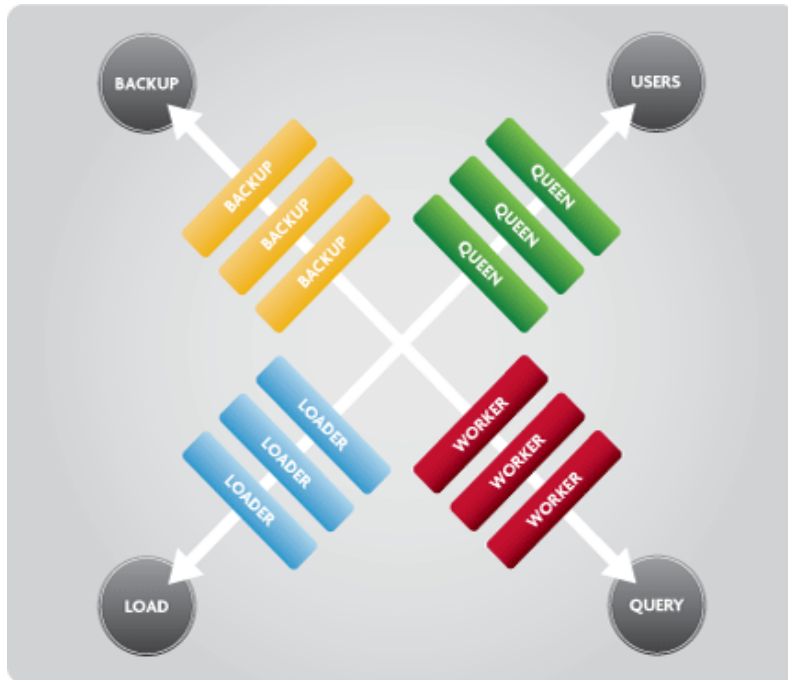
- 24x7 system availability through fault-tolerant design and non-disruptive administration

## “Always parallel” pervasive parallelism

- High user concurrency and predictable service levels
- Virtually unlimited scalability
- Patented SQL-MapReduce combines power of MapReduce with familiarity of SQL

# Massively-Parallel Processing Architecture

## Aster Data *n*Cluster

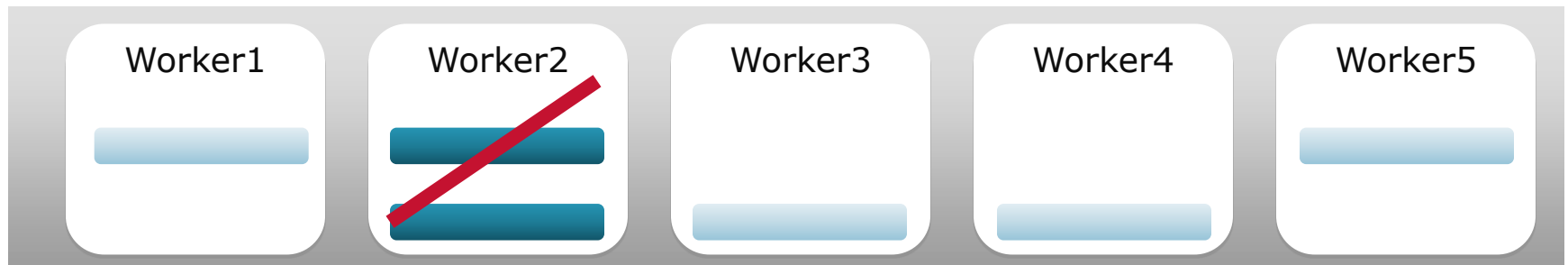


- Independent processing of backups, loads, and queries
- Reads, writes, backup, and data loading occur in parallel
- Each tier independently scalable

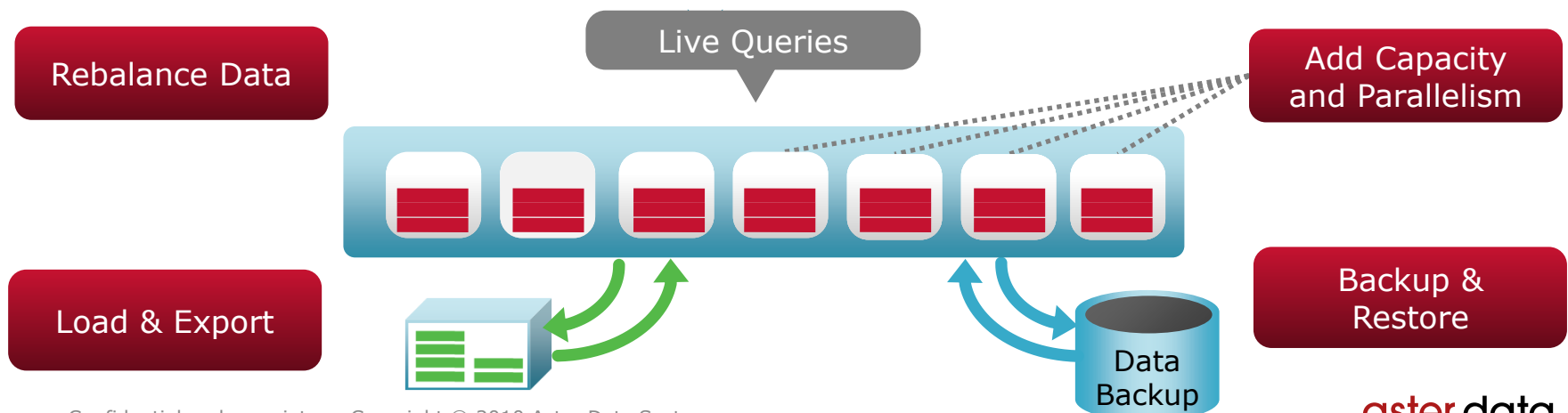
# "Always-On" Availability

Fault tolerant architecture and online management to minimize planned and unplanned downtime

## Online failover and replica restoration

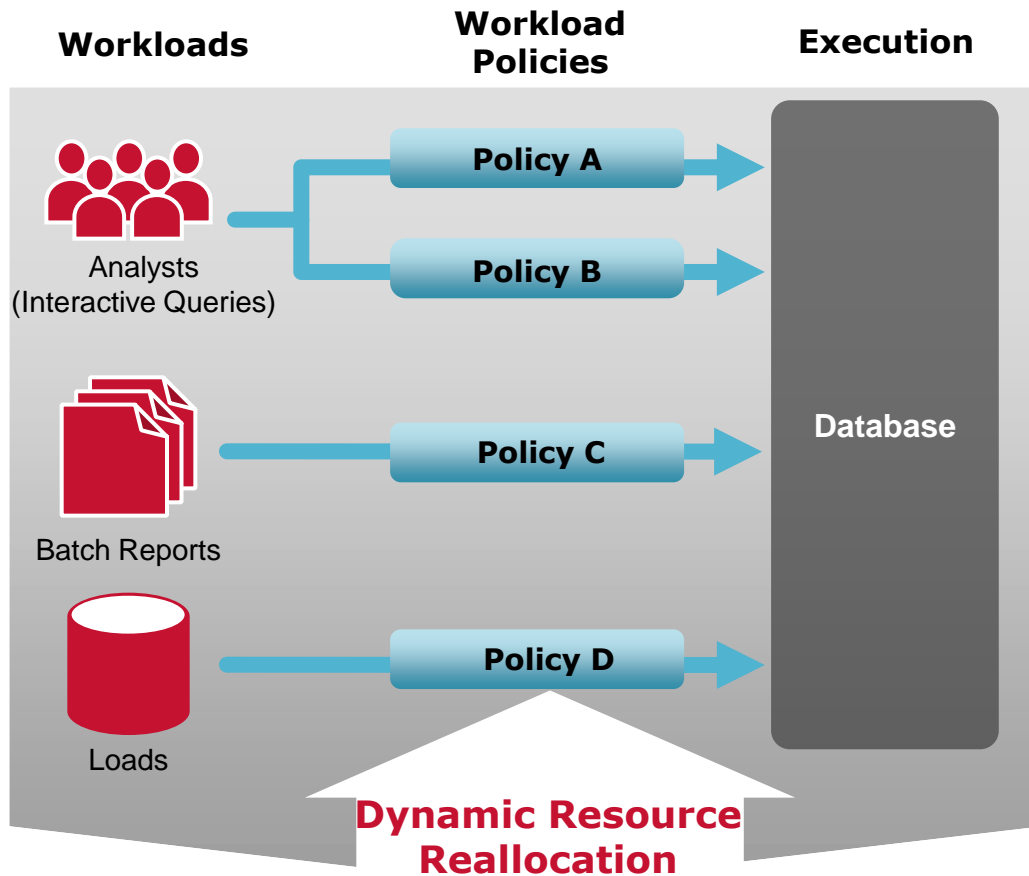


## Online management



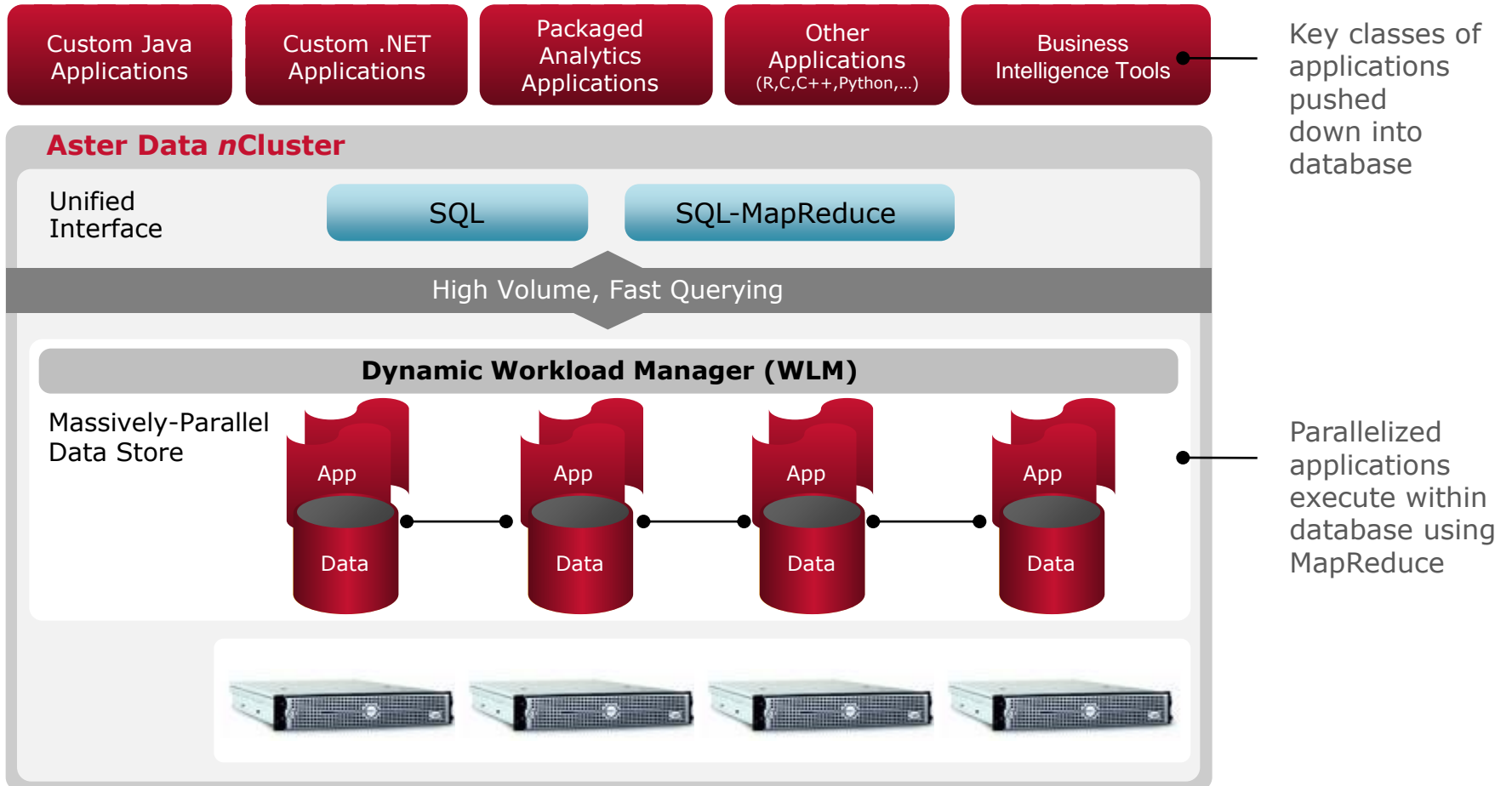
# Aster's Dynamic Workload Management

Optimize use of resources for consistent performance



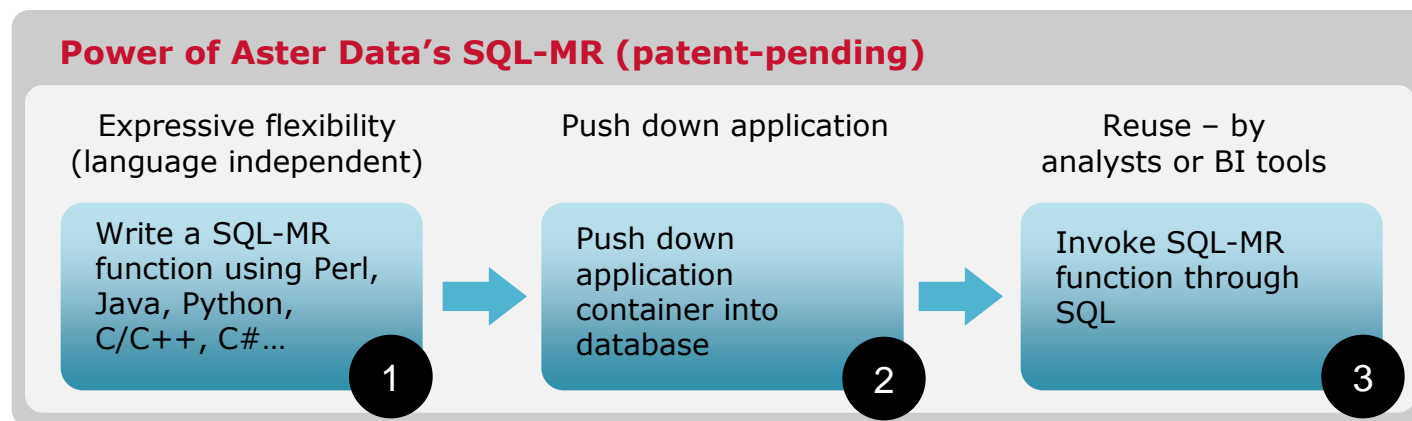
- Granular rules-based prioritization
- Dynamic resource allocation and re-allocation
- 300+ mixed concurrent Workloads

# Running Applications Inside the Database for Ultra-Fast Analytics



# Aster Data's SQL-MapReduce for Faster, Easier, Powerful Analytics

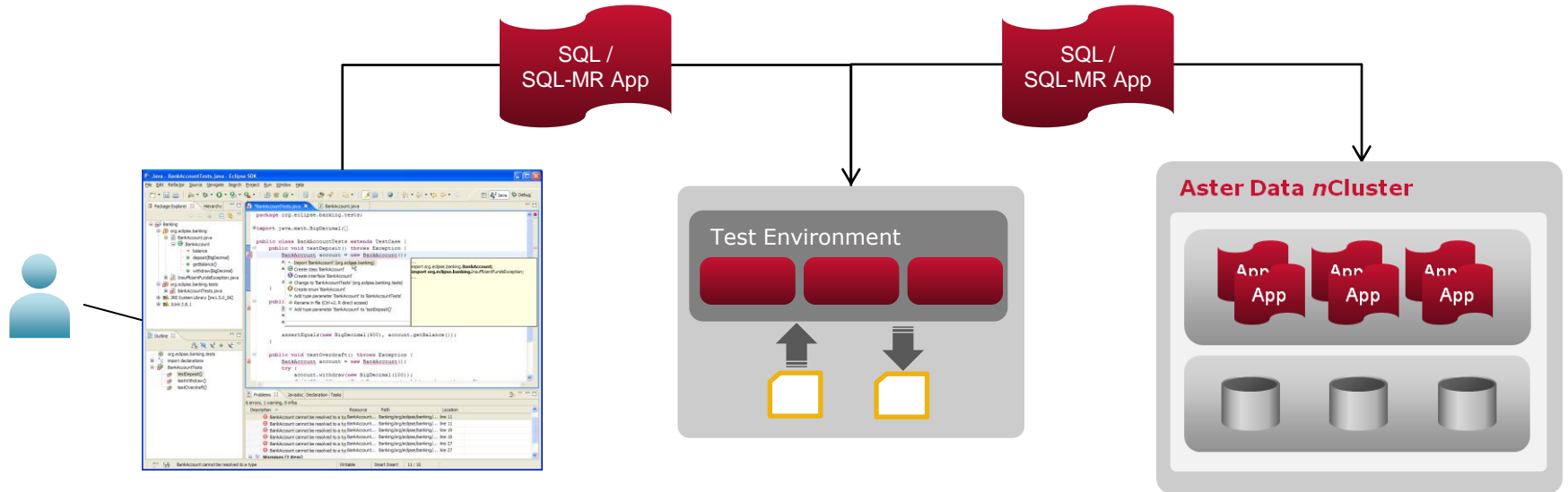
- Integrates SQL with MapReduce to deliver the first enterprise-ready MapReduce framework that executes inside the database
- SQL-MapReduce enables:
  - Ultra-fast and deep data analysis with seconds response time
  - Running rich applications in parallel – 10s to 1000s degrees of parallelism
- SQL-MapReduce is key enabler to running applications inside the database



*The next step: making big data analytics  
easy to build and manage*

# Preview: Accelerating Development of Analytic Applications

First integrated MapReduce & SQL development environment



1

**Develop**

Point-and-click dev environment and wizards for SQL-MapReduce applications *plus* new suite of analytics functions

2

**Test**

Test advanced analytics applications in local desktop test environment

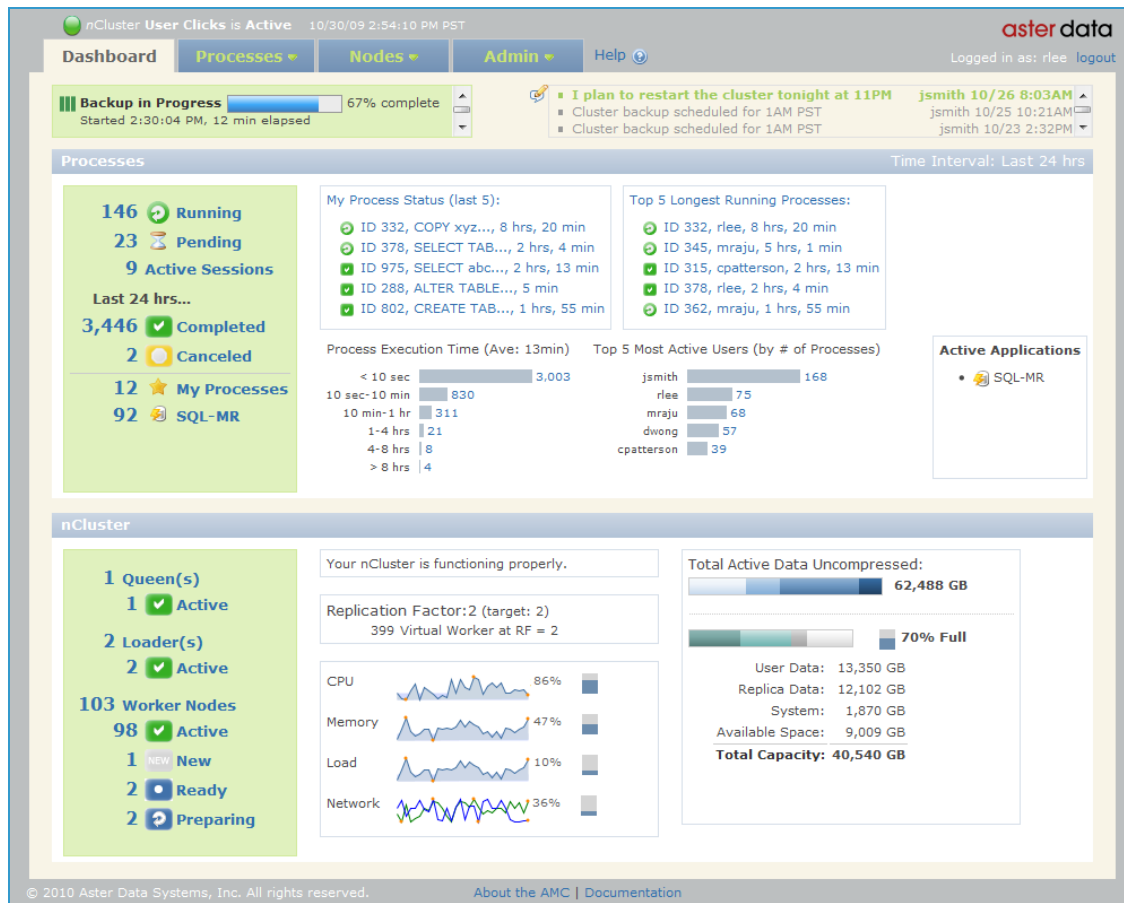
3

**Deploy**

*Push down* application into database with a single click

# Preview: Cutting-Edge Management Console

## Enhanced visibility and control of analytics processing



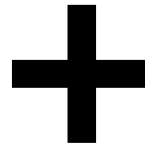
- **Dashboards** summarize cluster status and activity
- **Query & process statistics** provide detailed view of processing
- **Node drilldown** provides visibility into node health and status

# Why Companies Are Turning to Aster Data

Big data warehousing + big data analytics

## Big Data Warehousing

- **MPP shared-nothing architecture**
- **Pervasive parallelism** of data loads, queries, backup, exports, upgrades, etc.
- **High performance, linear scale out**
- **SQL-MapReduce** for easy parallelization of processing
- **Online, incremental scaling**
- **Dynamic Workload Management** for 100s of concurrent workloads
- **Optimized query planner**



## Big Data Analytics

- **Application push down** enables application execution inside the database for ultra-fast data processing on massive data scales
- **Applications are fully parallelized** to utilize TBs of memory and 1000s of CPU cores
- **Application Management Services** for failover, performance, resource management etc.
- **Dynamic Workload Management** spans traditional queries and app execution
- **Applications** include Java, C, C++, C#, .Net, Perl, Python, and packaged apps

The logo features the words "aster data" in a white, lowercase, sans-serif font. The background is a solid dark red color with several large, overlapping, semi-transparent circles in a lighter shade of red, creating a modern, abstract design.

aster data

big data. fast insights.™