Data Computing Appliance

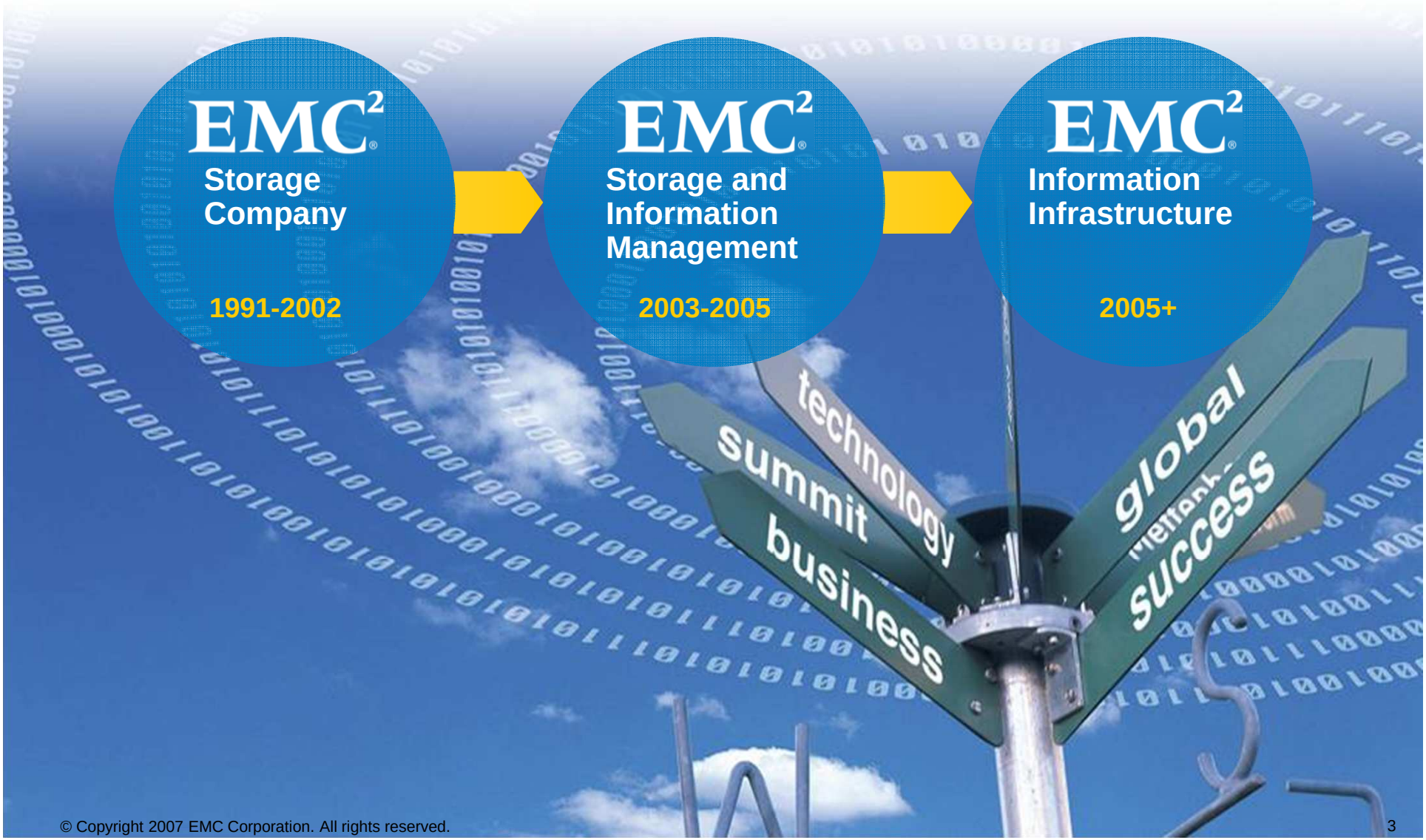
Presented by **Emin ÇALIKLI**
Sr. Technology Consultant
emin.calikli@emc.com

Greenplum

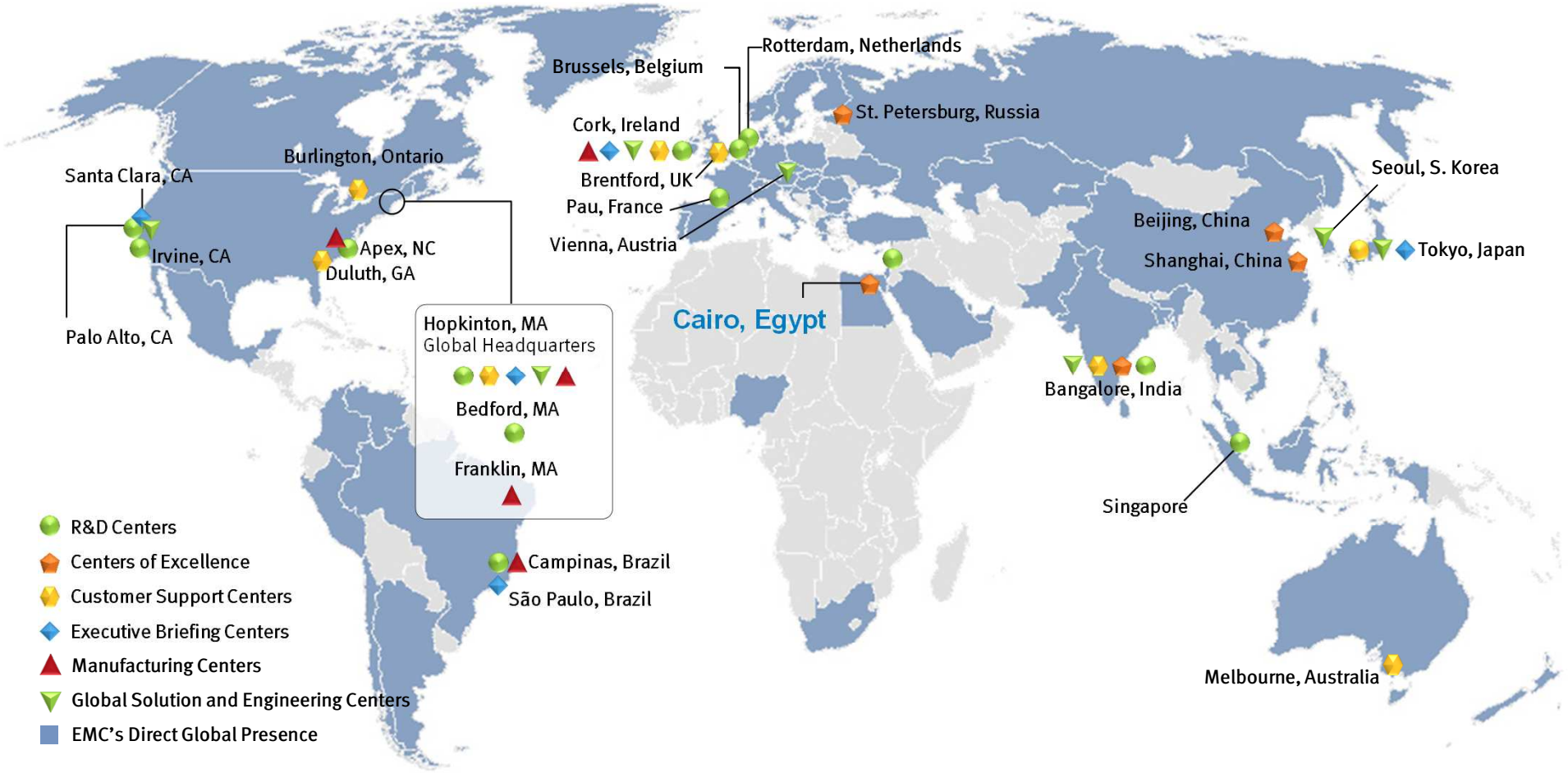
Agenda

- Who is EMC?
- Greenplum Innovative Technology
- Greenplum Data Computing Appliance

EMC's Evolution



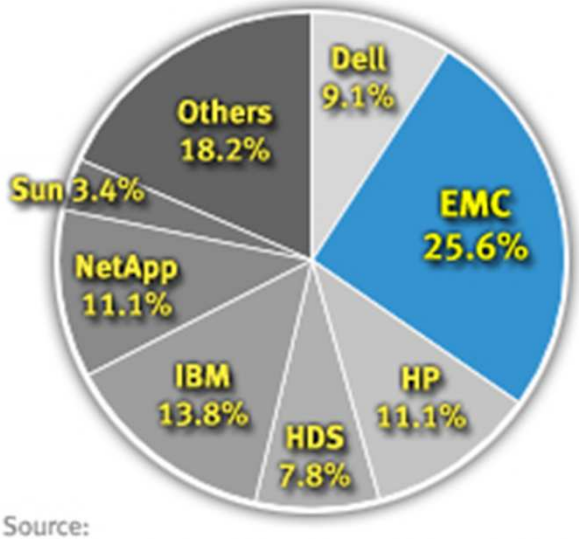
400+ Offices with Operations in 60+ Countries 43,000 Employees



Q4 '10Highlights

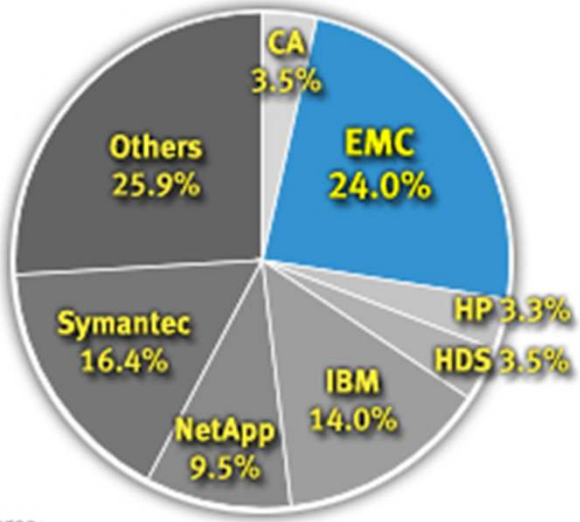
Thriving in a Challenging Global Economy

W/W External Disk Storage Systems, 2010



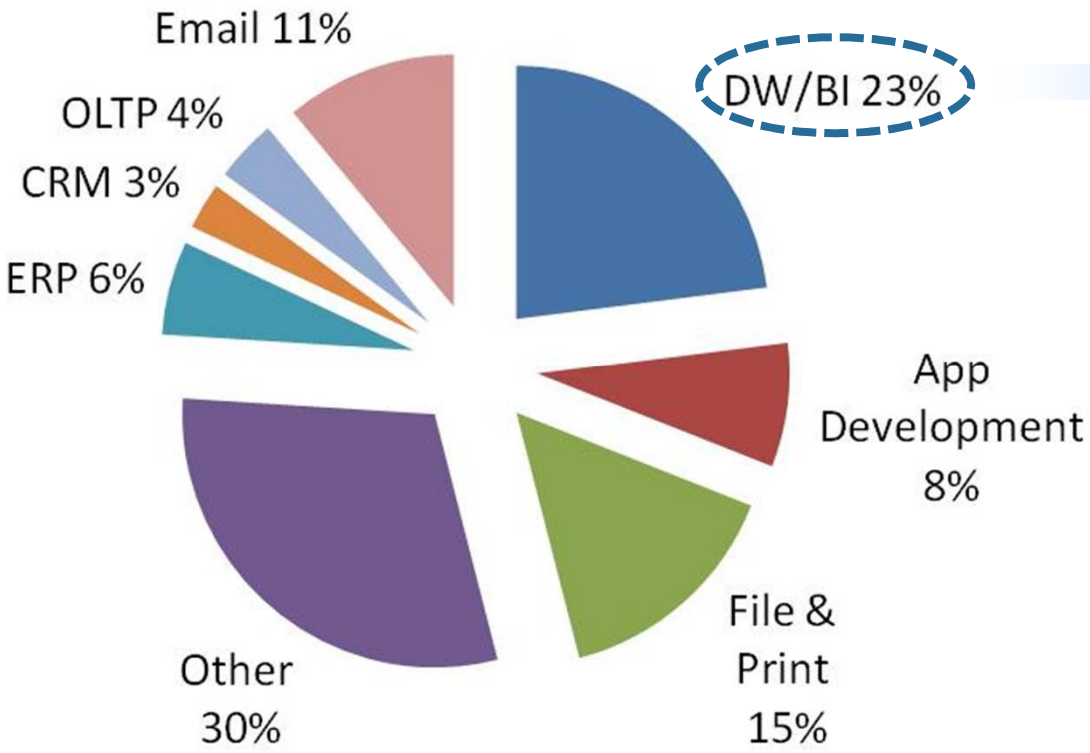
Source:
IDC W/W Quarterly Disk Storage Systems Tracker, 3/11.

W/W Storage Software, 2010



Source:
IDC W/W Quarterly Storage Software Tracker, 3/11.

DW/BI Storage Volumes



DW/BI is 23% of external storage workloads

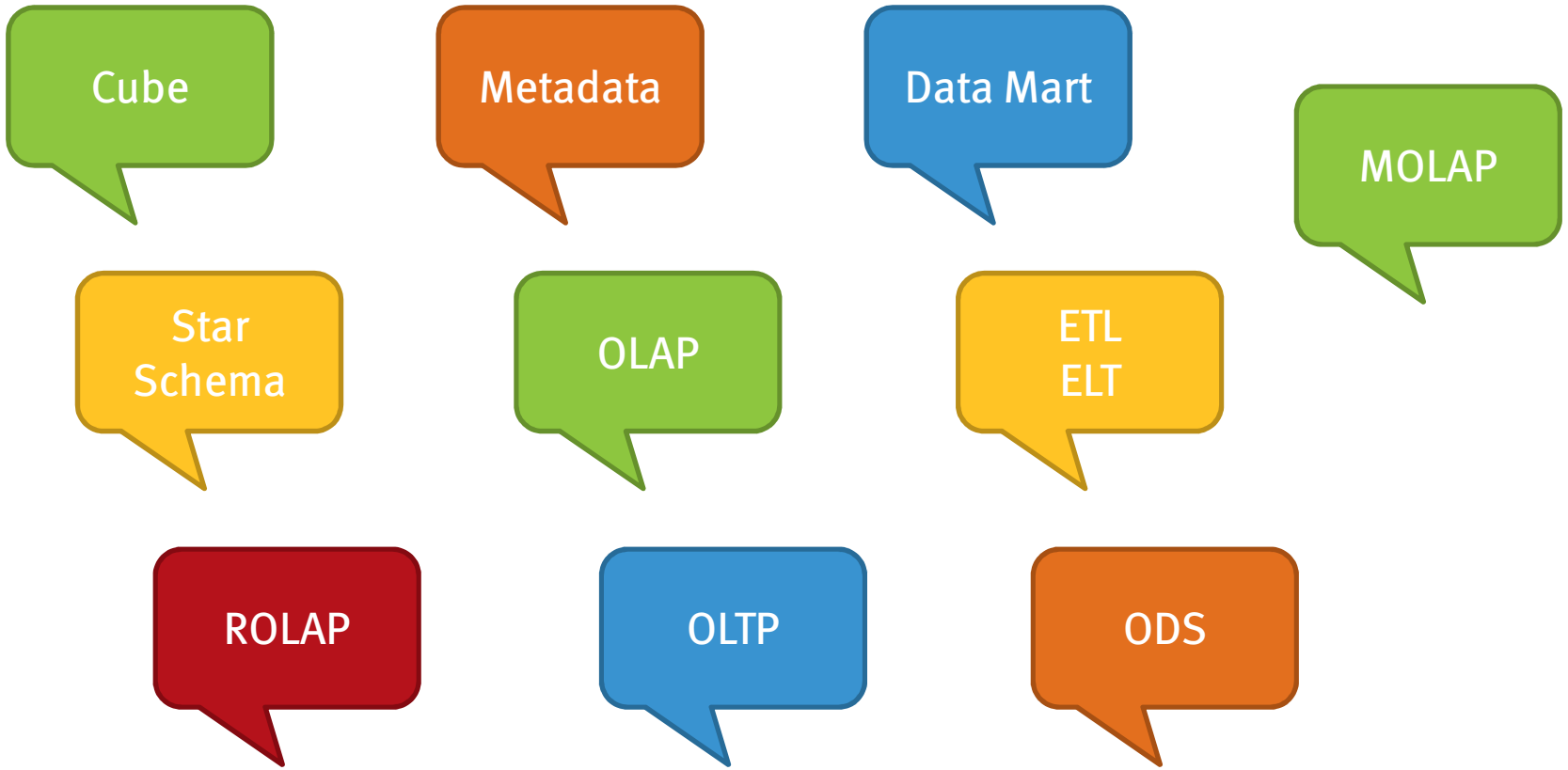
Source: IDC, Sep 2009

THE ANSWER MACHINE

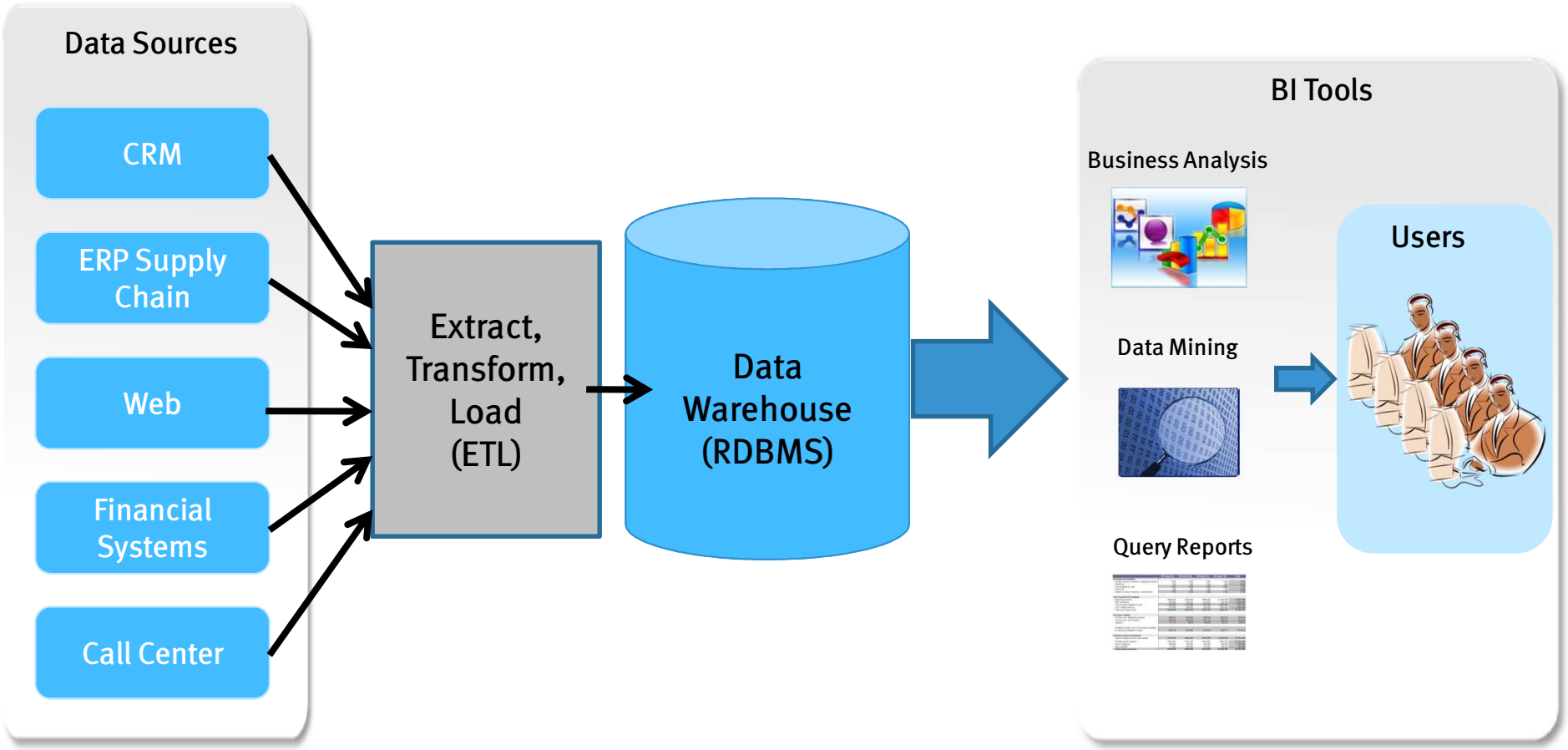
DATA IN. DECISIONS OUT.

Introducing the
Greenplum Data Computing Appliance

Data Warehousing?



Information Flow in a DW/BI Environment



Data Input → ETL → Data Warehouse → Data Analysis → Presentation & Delivery

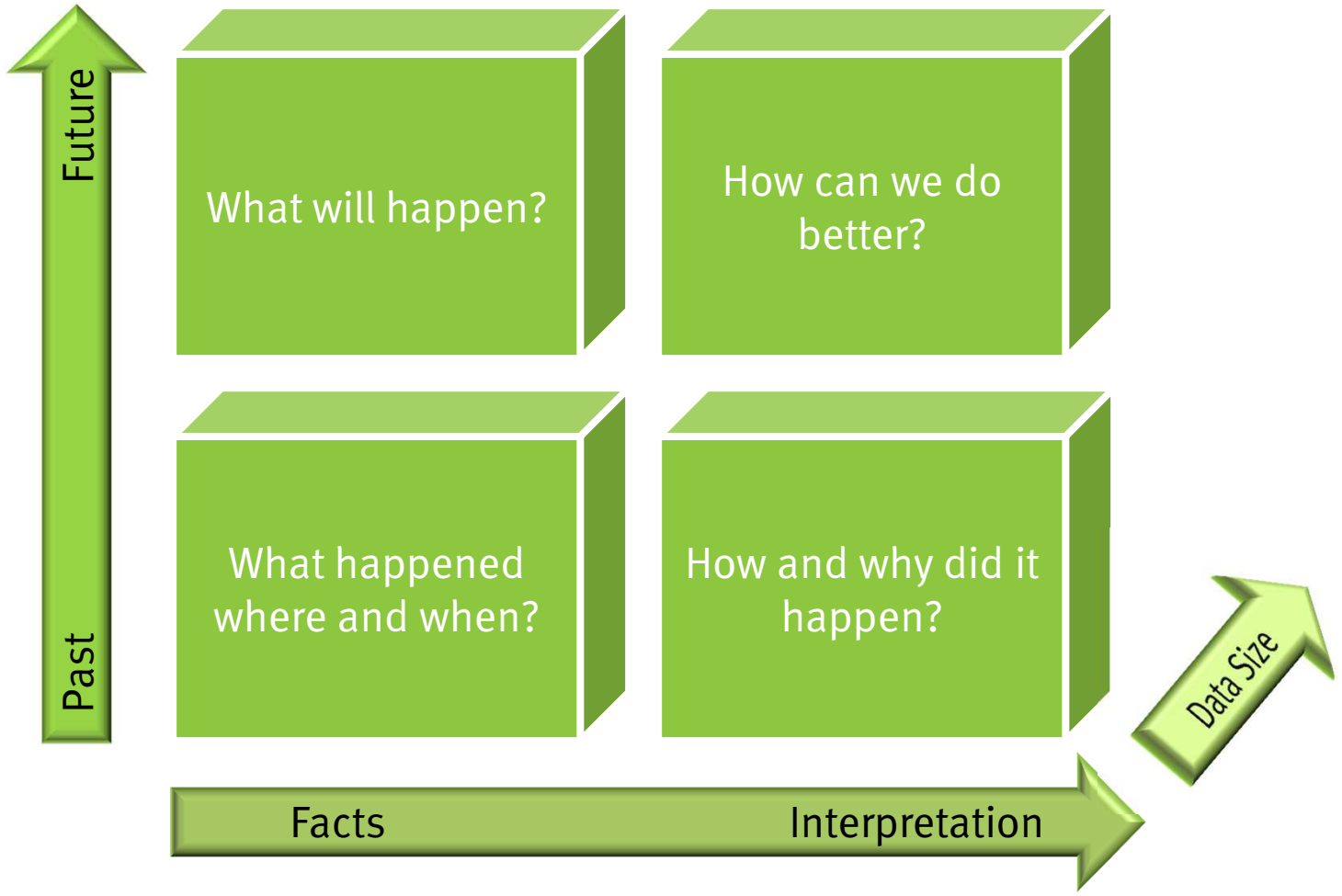
Greenplum Architecture

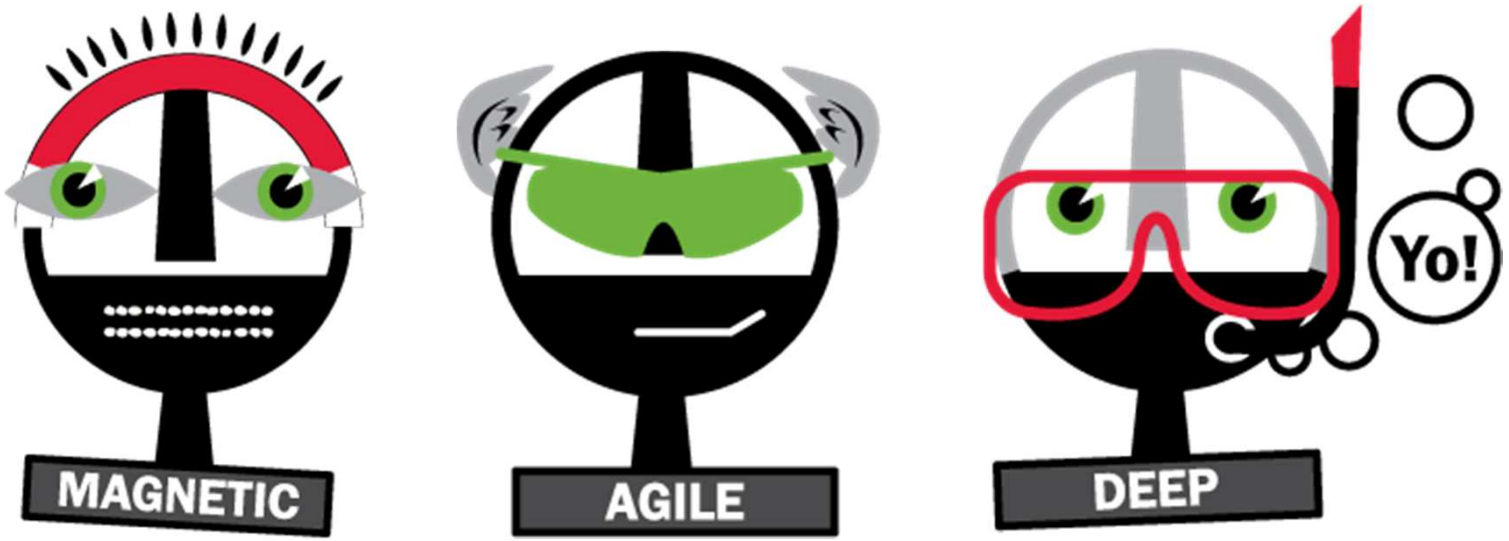
- MPP = Massively Parallel Processing
  - Two or more Servers (with own CPU/RAM/Disk) working on the same task
  - Multiple units of parallelism working together
    - Parallel Database Operations
    - Parallel CPU Processing
  - **Segments** = Greenplum Units of Parallelism (one Postgres database)
- 'Shared Nothing' Architecture
  - Each Segment is a separate Postgres Database
  - Segments only operate on their portion of the data
  - Segments are self-sufficient
    - Dedicated CPU Processes
    - Dedicated storage that is only accessible by the Segment

Things are changing

Yesterday's Data Warehouse and Analytic Infrastructure	The Greenplum Future
Proprietary	Commodity
Expensive	Cost-Effective
Centralized, Monolithic	Distributed
Process-Heavy	Self-Service
Batch	Real-Time
Summarized	Deep
Slow	Agile

Deeper Analysis – Bigger Business Value





MAD Skills

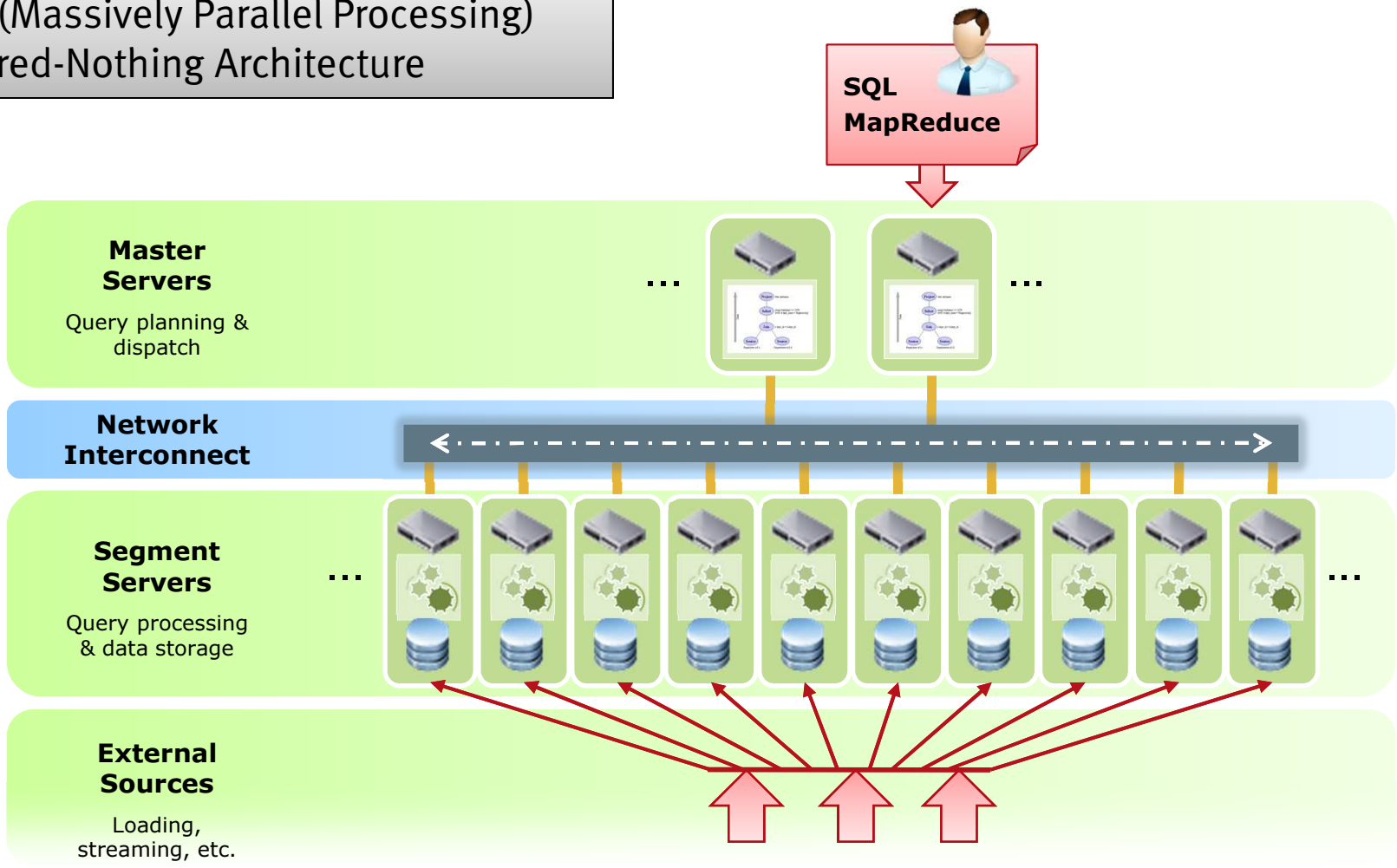
- **Business Intelligence**
  - MicroStrategy
  - Business Objects
  - Cognos
  - SAS
  - Informatica
  - …
- **The Analytics Library**
  - Logistic regression
  - Mann-Whitney U Test
  - Chi-Square Test
  - Sparse vectors, matrices
  - NLTK
  - …
- **Built-in Analytics**
  - Multiple linear regression
  - Naïve Bayes
  - Matrix operations
  - Window functions, OLAP
  - PL/R
  - …
- **Methods**
  - Log likelihood
  - Conjugate gradient
  - Re-sampling
  - K-means clustering
  - Association rules
  - …

Our products

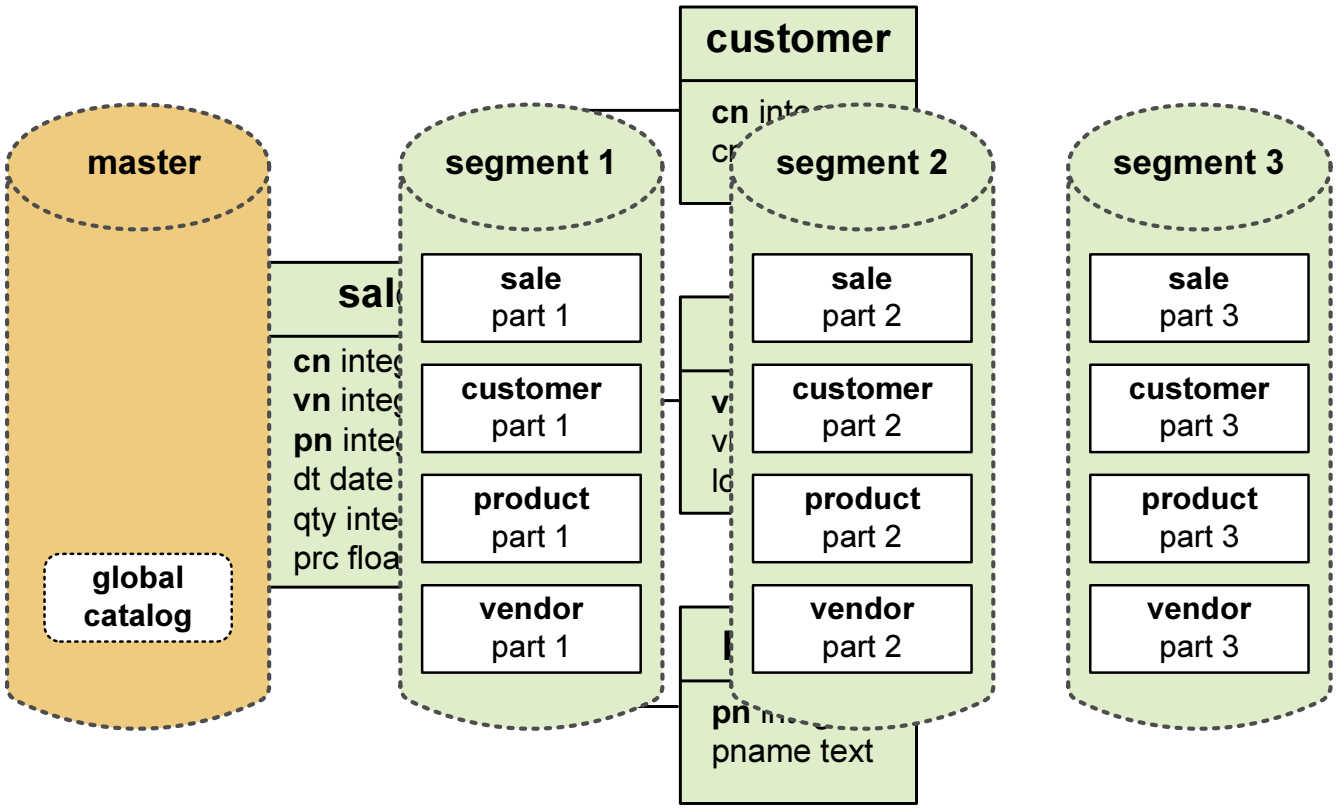
- **Greenplum Enterprise Data Cloud - Chorus**
  - Breakthrough data virtualization and collaboration software platform
- **Greenplum MPP Database**
  - Industry's leading MPP database software for large-scale data warehousing and analytics
- **Greenplum Single Node Edition**
  - Free analytic database software, highly optimized for single server deployments
- **Greenplum DCA – Data Computing Appliance**
  - DCA10, DCA100, DCA1000, DCA10000

Greenplum Database Architecture

MPP (Massively Parallel Processing)
Shared-Nothing Architecture

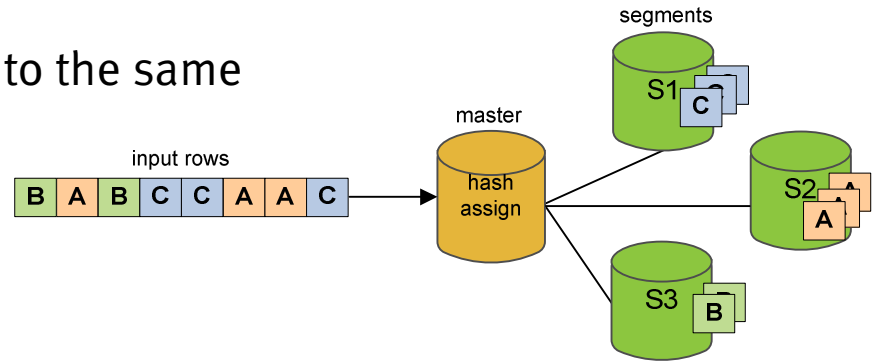


Distributed Tables

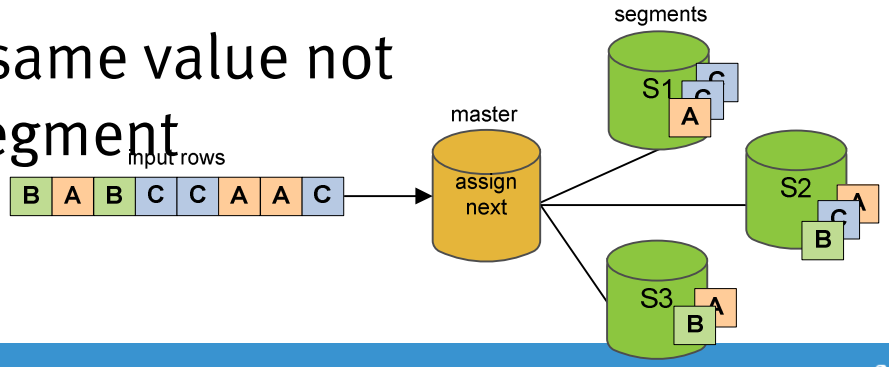


Distribution Policies

- **Hash Distribution**
  - `CREATE TABLE … DISTRIBUTED BY (column [,…])`
  - Keys of the same value always sent to the same segments



- **Round-Robin Distribution**
  - `CREATE TABLE … DISTRIBUTED RANDOMLY`
  - Rows with columns of the same value not necessarily on the same segment

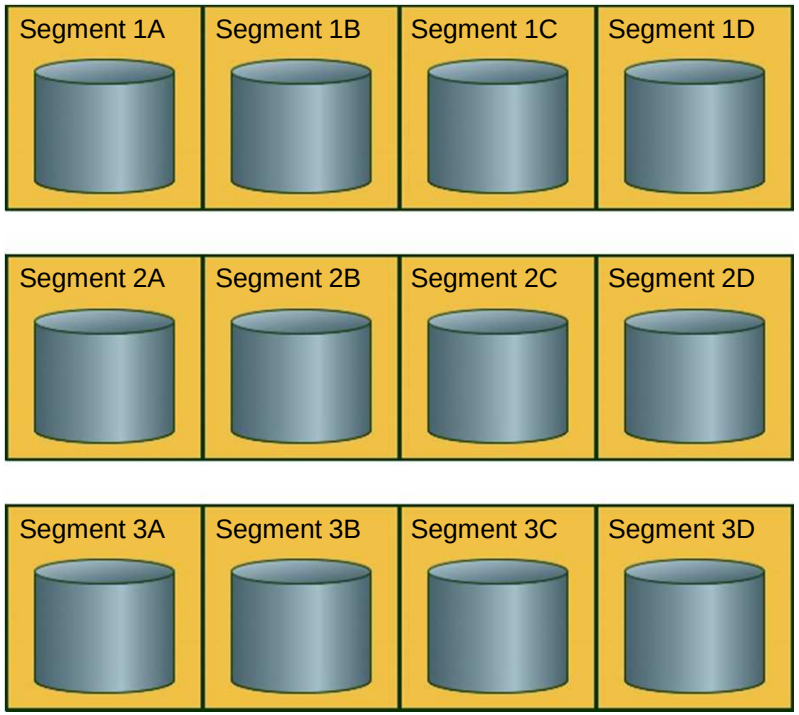


Parallel Data Scans

```
SELECT COUNT(*)
  FROM orders
 WHERE order_date >= 'Oct 20 2005'
   AND order_date <  'Oct 27 2005'
```

4,423,323

Master



Develop Query Plan

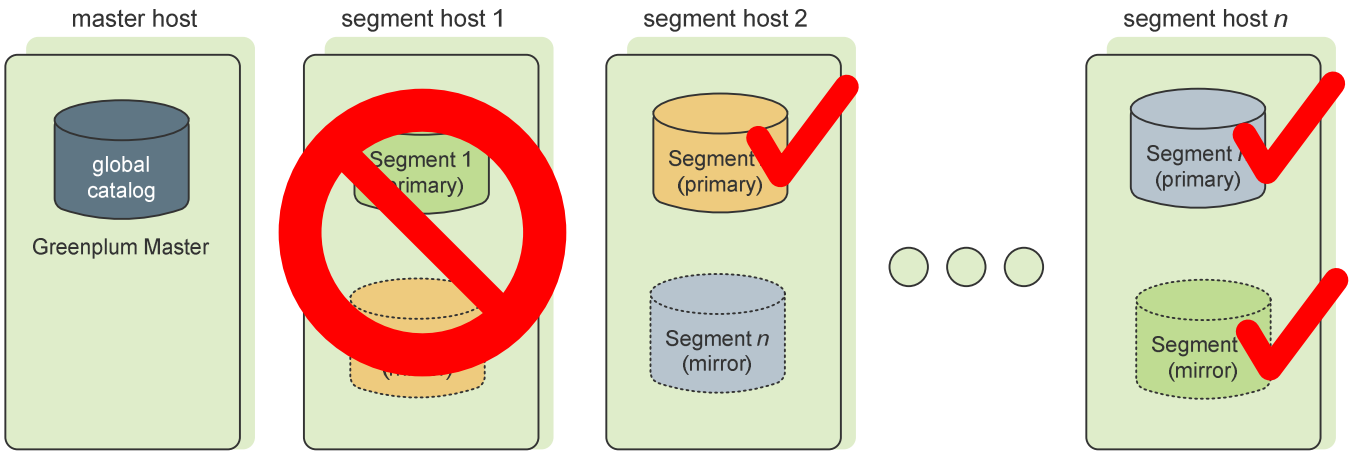
Send Plan to Segments

Segments Return Results

Return Results

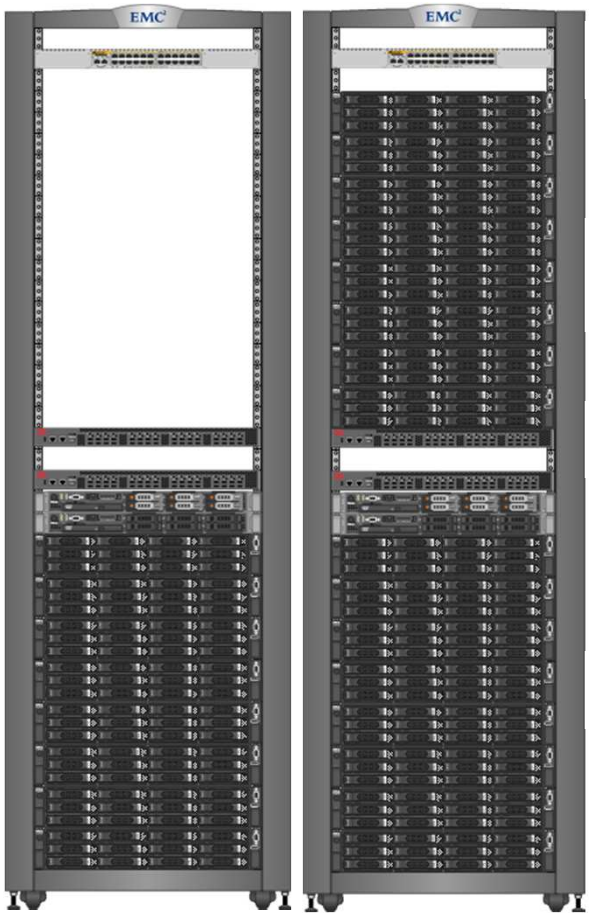
Each Segment Scans Data Simultaneously

Data Redundancy - Segment Mirroring



Configuration Specifications

GP100 GP1000



Expansion bus

Eight segment servers full rack

Interconnect bus

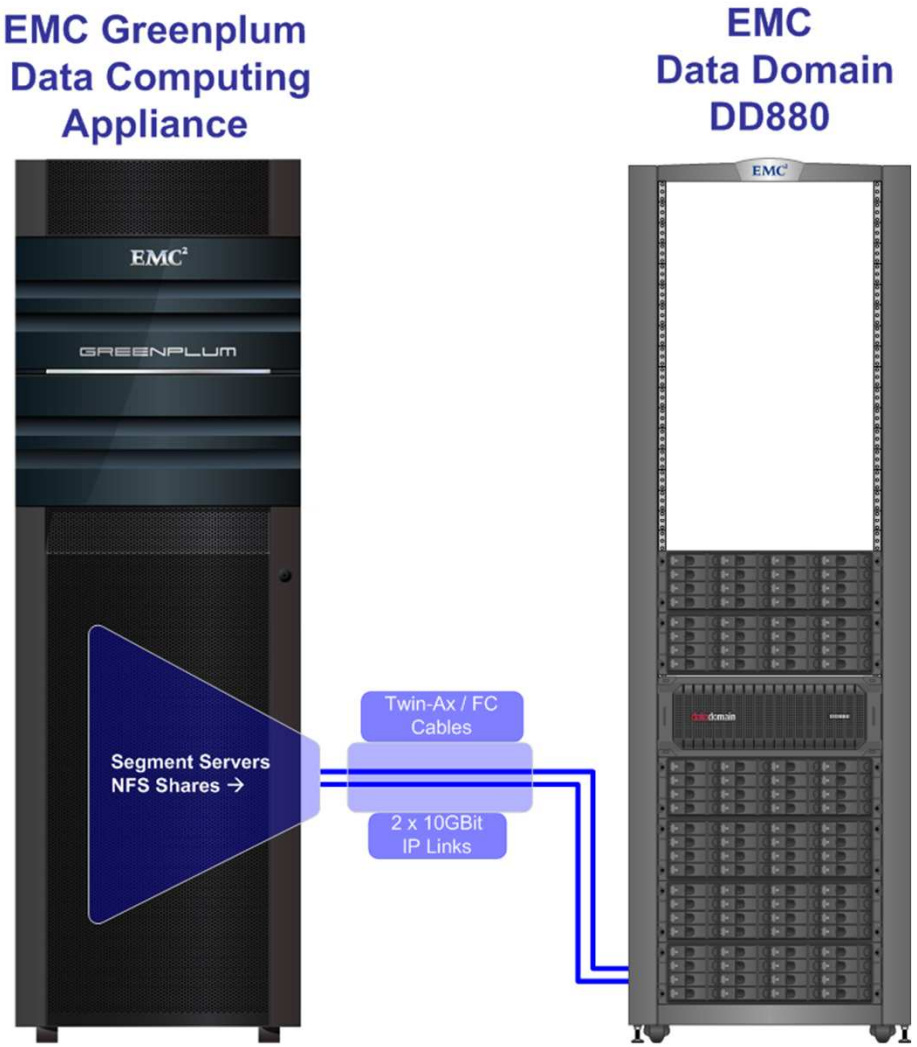
Two master servers

Eight segment servers

Specifications	GP100: Half Rack	GP1000: Full Rack
Master servers	2	2
Segment servers	8	16
Memory per server	48 GB	48 GB
Total memory	384 GB	768 GB
Segment HDDs (SAS)	96	192
Usable capacity (uncompressed)	18 TB	36 TB
Usable capacity (compressed)	72 TB	144 TB
Scan rate	12 GB/s	24 GB/s
Data load rate	5 TB/hour	10 TB/hour

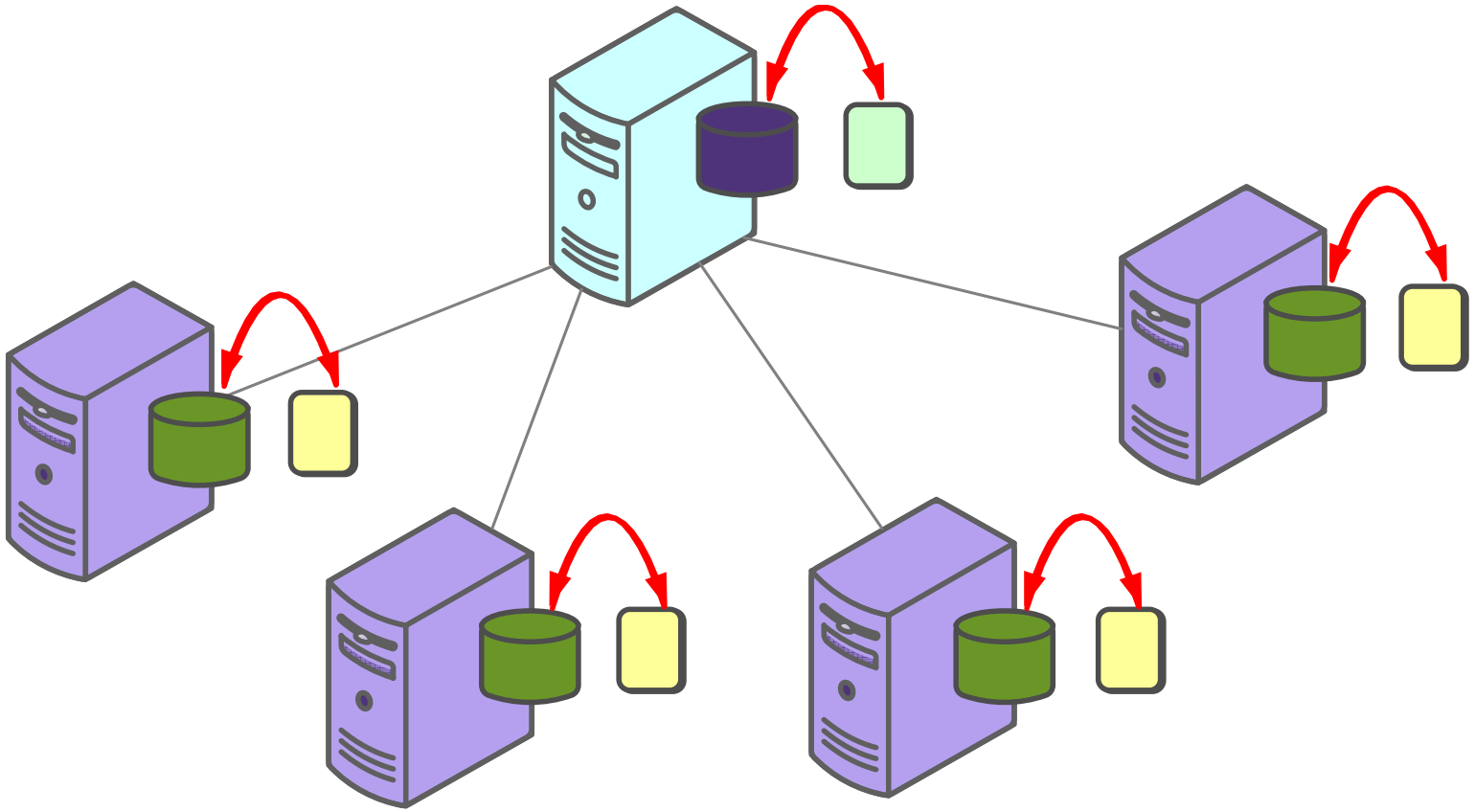
THE FASTEST

Integrated Data Domain Backup



- Backup and Recovery
  - With EMC Data Domain / Greenplum native utility
- Reduces storage backup requirements
  - Deduplicates data
- Fast, reliable data recovery
  - Reduced recovery time
- Flexible and efficient
  - Designate intervals to backup
  - Point-in-time copies

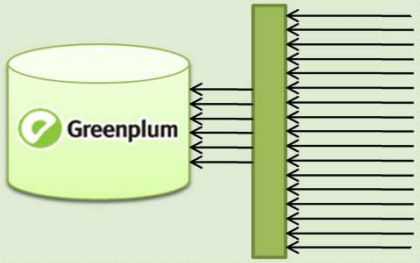
Parallel *Everything*, including Backups and Restores

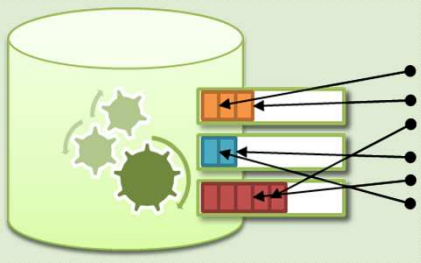


Greenplum Polymorphic Storage™
Flexible Row or Column Oriented Processing

- **Rather than take a side, we give customers the flexibility of both**
  - Results consistent with industry/academic findings of row vs column benefits
- **Row-orientation typically better for general purpose DW**
  - Avoid reassembly overhead that dominates in typical workloads
- **Column-orientation typically better for an important set of use cases**
  - Accessing small # of cols from a wide table – e.g. certain Data Mining use cases
- **Table Orientation**
  - Just specify 'orientation=row' or 'orientation=column' when creating a table
  - Gzip and LZ compression algorithms available with either orientation
- **Gives customers the choice of processing model for any table**
- **Efficient pre-projection, parallel execution in either case**

Workload Management Overview





Connection Management

- Control over how many users can be connected.
- Provides pooling (to allow large numbers) and caps (to restrict numbers if desired)
- Intelligently frees and reacquires temporarily idle session resources

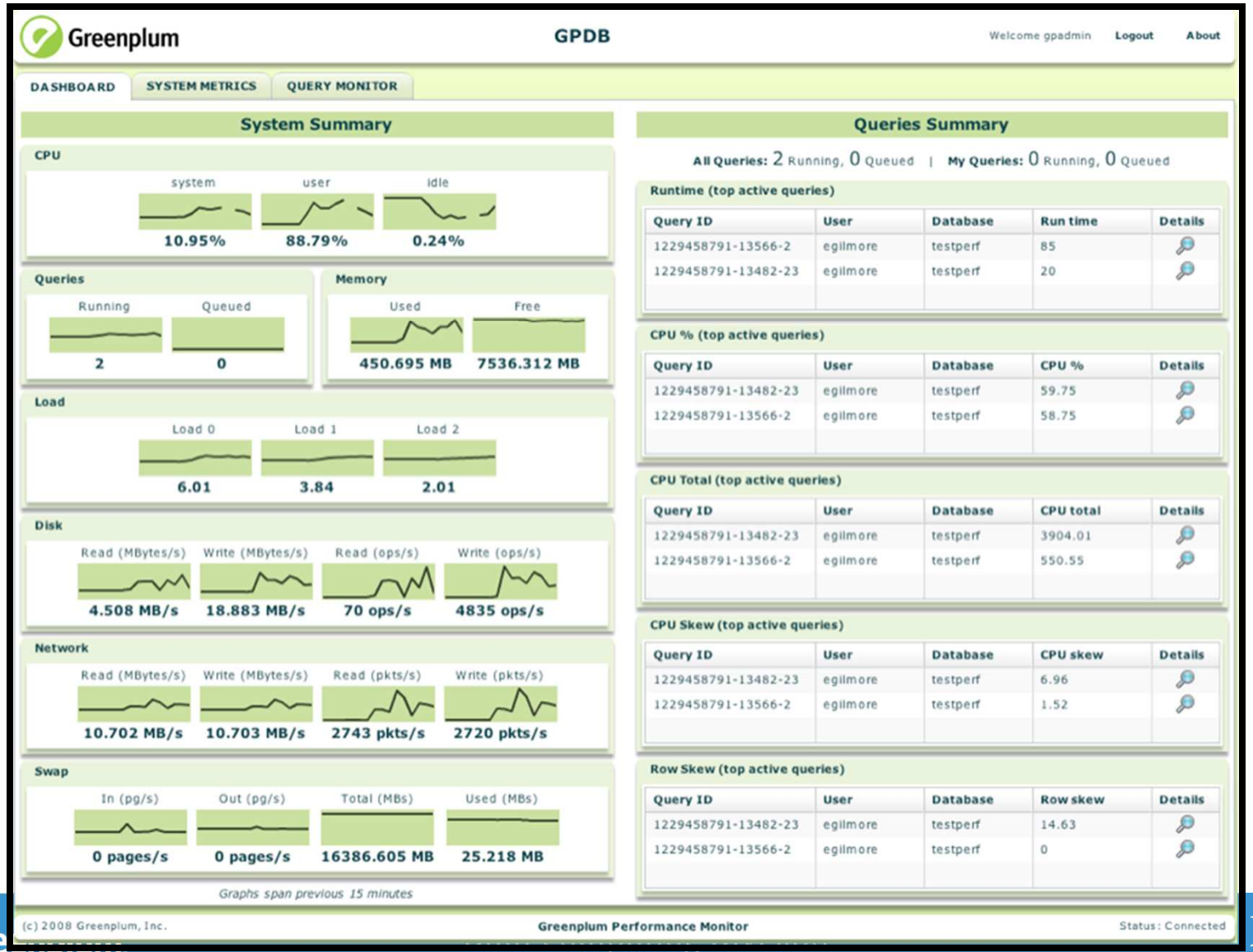
User-Based Resource Queues

- Each user is assigned to a resource queue that performs 'admission control' of queries into the database
- Allows DBAs to control the total number or total cost of queries allowed in at any point in time

Dynamic Query Prioritization [NEW]

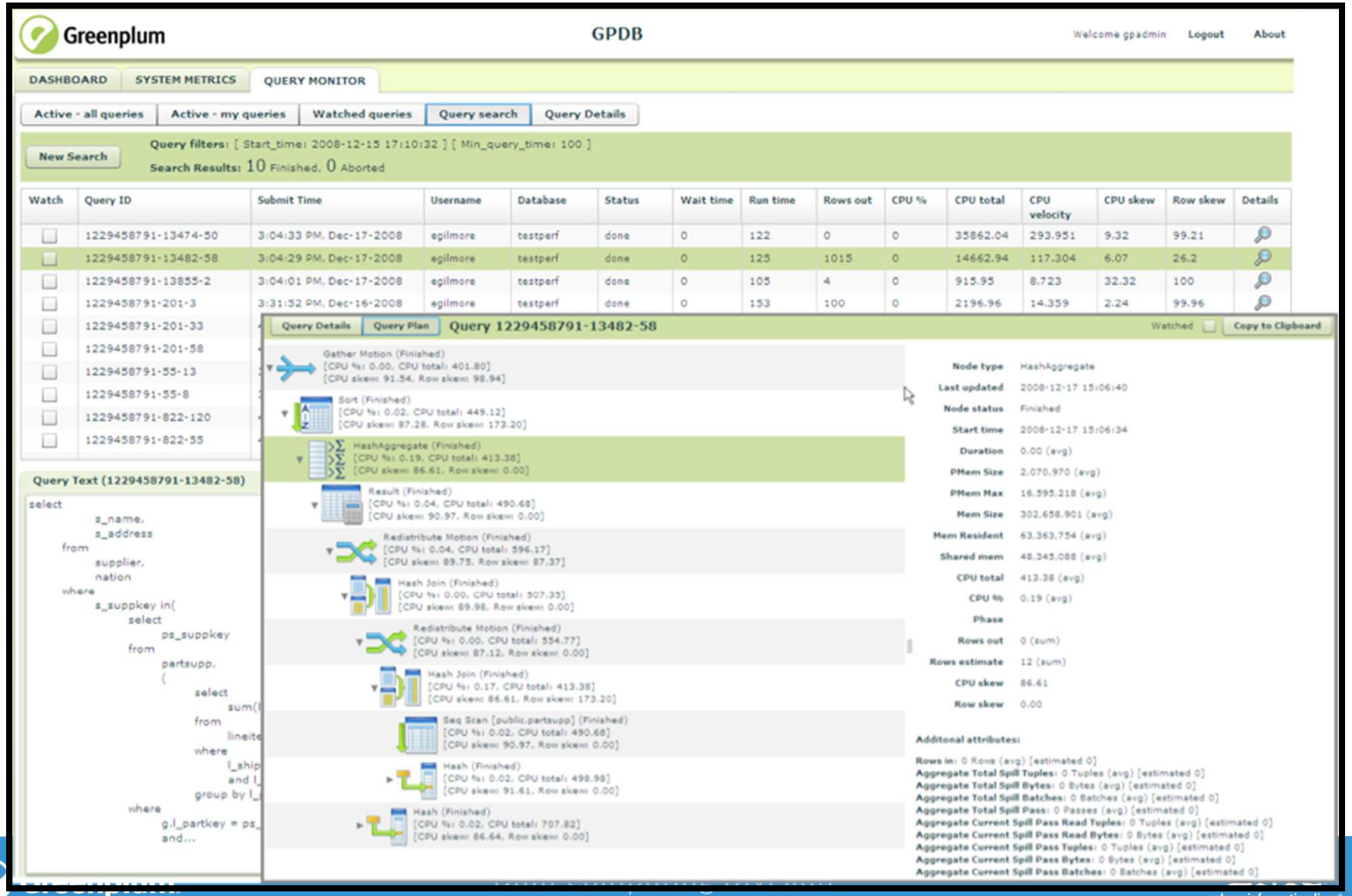
- Patent pending technique of dynamically balancing resources across running queries
- Allows DBAs to control query priorities in real-time, or determine default priorities by resource queue

Enhanced Database Monitoring - Dashboard



Enhanced Database Monitoring – System Metrics

Enhanced Database Monitoring – Query Monitor



EMC Consulting Services for DW/BI

BENCHMARKING AND PROOF OF CONCEPT

- Client use cases implemented on appliance and exercised at scale
- Standard performance metrics captured for current state and competitive comparison
- TCO and ROI calculations for decision makers
- Platform training for client IT staff

PLATFORM MIGRATION AND INTEGRATION

- Database Replatforming and Consolidation
- Application code porting, including Oracle, Teradata, DB2 and SQL Server
- High speed historic data migration, consolidation and harmonization
- Real/near real time and batch system integration via ETL and ELT

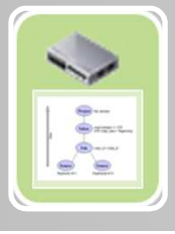
PERFORMACE DESIGN AND DATA MODELING

- Optimal Greenplum configuration, implementation and data model design aligned with business use cases
- MPP Distribution
- Workload Management
- Polymorphic Storage including:
  - Partitioning
  - Compression
  - Columnar

INFORMATION MANAGEMENT AND ANALYTICS

- Near real time predictive analytics across multi-Terabyte data sets
- Business Intelligence and Reporting
- Self provisioning analytic marts and business user collaboration
- Data Governance, Stewardship and Continuous Quality Improvement

Key Technical Innovations

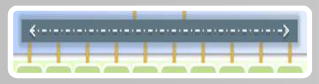


Scatter-Gather Data Streaming

- Industry leading data loading capabilities

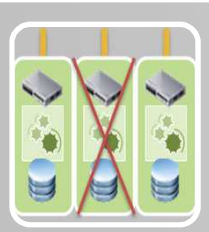
Online Expansion

- Dynamically provision new servers with no downtime



Map-Reduce Support

- Parallel programming on data for advanced analytics



Polymorphic Storage

- Support for both row and column-oriented storage

Greenplum Data Computing Appliance

Data In. Decisions Out.

- Industry leading price-performance
- Most powerful data loading machine on the planet (+10TB per hour)
- First best step to private-cloud, virtualized DW and analytic infrastructure