



## The Enterprise Data Cloud™

---

Redefining the Data Warehousing and Analytics Market

June 2009

### Contact

Greenplum

1900 S. Norfolk Drive,  
Suite 224

San Mateo, CA 94403

United States

+1 650 286 8012 (ph)

+1 650 286 8010 (fax)

# Table of Contents

- Executive Summary 3
- Data Management in the Enterprise 3
  - Challenge: Ever-Increasing Scale of Data 3
  - Challenge: Putting the Data to Work 4
- Fragmentation and Friction – How We Got Here 5
  - Unrealistic Practices 5
  - Dealing with Volatility and Change 6
  - Promise vs Reality – The ‘Enterprise Data Warehouse’ 6
  - Promise vs Reality – The ‘Data Warehouse Appliances’ 7
- The Way Forward 7
  - Business and Technology Trends 8
  - The Missing Piece – Self Service 8
- Greenplum’s EDC™ Initiative 9
  - EDC Platform 9
    - Self-Service Provisioning 10
    - Unified Data Access 11
    - EDC Technical Stack 12
    - At the Core – Greenplum’s MPP Database 13
  - EDC Methodology 14
  - EDC Ecosystem 15
- Use Cases 15
  - Use Case: Data Mart Consolidation 15
  - Use Case: Project Sandboxes 16
- Conclusion 16

## Executive Summary

Data warehousing technology and practices have evolved over the past 30 years, and yet have failed to reconcile the tension between business analysts (who want control and flexibility) and IT (who wants centralization and efficiency). The resulting technology and practices treat the warehouse like a precious mainframe, locking down processes and forcing a proliferation of shadow IT databases by the business.

There's a better way. Greenplum's Enterprise Data Cloud™ (EDC) Initiative is built around a bold new vision for bringing the power of Self-Service to data warehousing and analytics. Greenplum's EDC Initiative empowers both business and IT to do more and work together without friction.

### **It comprises 3 core elements:**

1. **The Platform** – Greenplum's EDC™ platform
  - a. Self-Service provisioning
  - b. Elastic scale
  - c. Massively parallel
2. **The Methodology** – New agile and pragmatic best practices
3. **The Ecosystem** – Working with customers and partners to bring to market new capabilities and standards

## Data Management in the Enterprise

Data, and the business intelligence that can be derived from it, are critical to any significant business. Data flows in trickles and torrents across these businesses – as transactions, event streams, logs, emails, and in countless other forms. Data volumes are exploding – to the tune of 1.5x to 2.5x a year at most companies, and new sources and uses for data appear every day.

Around the world, companies understand that there is value in the data that swirls within their walls, but almost universally they struggle to store, manage and empower the organization to derive value from this data. Their challenges are twofold – the ever-increasing scale of the data, and the difficulty in 'putting the data to work' – i.e. efficiently supporting the diverse uses and users of data across the company.

### Challenge: Ever-Increasing Scale of Data

Data growth is a given for almost every company. But the first challenge that many companies face is that their data volumes far exceed the capabilities of mainstream OLTP-oriented database systems (e.g. Oracle, Microsoft SQL Server). These systems can comfortably handle low numbers of TBs (Terabytes) and modestly complex analytics queries, but they start to break down or require extraordinary tuning in the low 10s of TBs. However many companies today need to store and analyze 10s or 100s of TBs, or even low numbers of PBs (Petabytes). Mainstream OLTP-oriented databases systems lag the scale required by today's largest databases by 3 orders

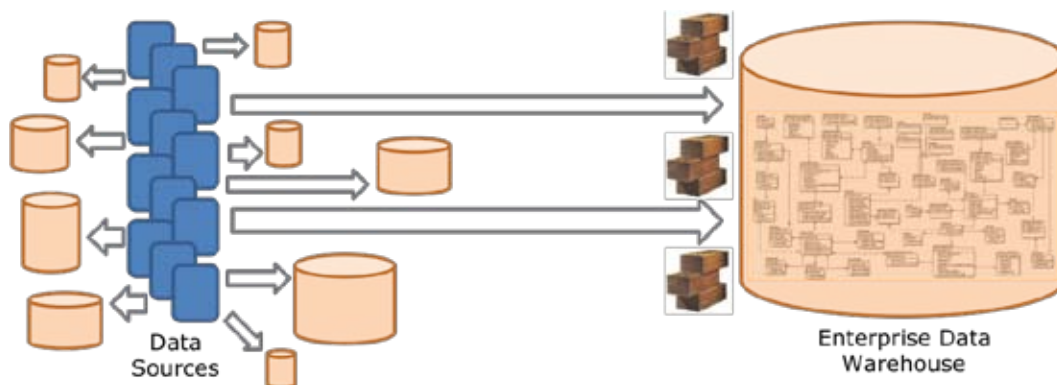
of magnitude (1000x). And companies often hit these limitations head-on before coming to understand that better alternatives exist.

The technology to support these larger databases exists today, in the form of MPP (Massively Parallel Processing) Share-Nothing DBMS systems. These systems are optimized for the read-mostly analytical processing required by data warehousing and business intelligence, and utilize 10s or 100s of nodes working in parallel with data automatically partitioned across the nodes to support fully parallel query processing. Traditionally these solutions were built using proprietary hardware technology – as is the case with Teradata and Netezza. By contrast, Greenplum is the leading pure software vendor of this technology, with a DBMS solution that runs on commodity hardware from a wide variety of tier-1 hardware vendors.

As a result, it is reasonable to consider this challenge solved. There are a number of MPP database vendors that have credible solutions for storing and analyzing 100s of TBs. At Greenplum, we provide these solutions to over 70 of the largest and most innovative customers in the world (as of June 2009) – providing extreme price-performance, massive scalability and parallelism, and mission-critical reliability for companies with 10s of TBs, all the way up to the largest known database in the world – eBay’s 6.5PB (6500TB) data warehouse running on a single Greenplum Database cluster.

### Challenge: Putting the Data to Work

While this first challenge has been solved, there is an equally important challenge that has remained largely unaddressed until now. Look at any significant business, and you’ll find hundreds or thousands of data silos. A company might have one or more ‘enterprise data warehouses’, hundreds of departmental silos, data spread out in Excel spreadsheets and Access databases, and countless custom applications. In addition to this, most companies have a multitude of data that isn’t captured at all and just falls on the floor – event streams that tell the tale of customer interactions, logs of important systems and operational processes, and emerging business data that doesn’t yet have a business case to be captured.



This all-too-common story reflects the organic evolution of these systems, and inherent tensions between those responsible for the operation of these systems (usually, but not always, IT) and those that want to use these systems and the data within them to get work done (usually analysts within business units).

IT looks at this chaos and sees cost and complexity. They want to bring order, consolidate and reduce costs. They want to reduce the messiness of all this data and bring greater control and repeatability to the process.

Business analysts look at this chaos and see barriers to getting their jobs done. The data they need is in different systems, or isn't being captured at all. Getting the data they need into a place where they have free reign to analyze it could take months of lobbying for machines, data feeds, and support from overstretched IT staff.

Each group has perfectly reasonable demands that are inherently in conflict. This is the heart of the problem, and this problem must be addressed to truly put data to work in support of the business.

## Fragmentation and Friction – How We Got Here

### Unrealistic Practices

Today's data management practices have evolved over the past 30 years – in response to business needs and the capabilities of the technology of the day. Throughout this time, companies have had an aspiration to unify their data – i.e. bring it together in one place where it can be centrally managed and organized. The hope is that they can build a single data model that will reflect their business and act as the framework for everything going on in the company.

For a long time it was believed that this was an attainable goal. i.e. If enough smart business modelers spend enough time scouring the organization, then out would pop the grand unifying data model for the business. Companies spent months or years on business analysis and mapping exercises to figure out how all the data should be organized. Experts touted the importance of getting all of the modeling right before letting any data into the warehouse – and the importance of keeping out any data that hadn't yet been properly vetted and reconciled.

Fast forward to today, and most organizations have a nagging sense that this approach doesn't work. They've seen failed 'big-bang' modeling efforts that took months or years to produce long documents that were out of date the moment they were published. They've seen business units venting frustration at their inability to get work done because of a 6-12 month backlog to incorporate new data into the central data model. They've seen political squabbling and endless meetings to try to forge agreement on 'the model' instead of focusing on the pragmatic needs of the business. Amongst practitioners there is now a growing appreciation that the blind pursuit of one stable and unifying data model was a mistake.

However this approach still reflects the orthodoxy of data modeling and warehousing practice. Books and articles tout these ideas and discuss the purity of the warehouse and the primacy of the model above all else.

## Dealing with Volatility and Change

There are a lot of problems with this old approach – not the least of which is its inflexibility with respect to change. Data is used in a diverse and ever-evolving multitude of ways across any company. The needs of the business are constantly evolving, and new data sources emerge and change daily. Any attempt to model everything before returning value to the business is doomed to lag and prevent innovation and agility.

Consider a fairly typical example – a company adds a new web ordering system, and a business analyst is tasked to look at how this data could be used to improve marketing programs. This new web order system tracks an extensive amount of data about the behavior of the user (e.g. clickstream through the site, offers and ads shown, etc). However the central warehouse wasn't modeled to support these large streams of event data, and only the customer's final order is transmitted to the warehouse. The remaining data, if kept at all, remains in a separate silo disconnected from the company's other data systems. As a result, the analyst has no way to explore this new data and combine it with existing data in support of marketing. The analyst could put in a request to model this data, but they find themselves at the back of a 12 month queue and required to make a business case for why this data should take up space in the expensive warehouse.

The problem here is that some parts of the business (i.e. both business requirements and data sources) are much more volatile than others. While much of the organization's data may be stable and cleanly centralized into a data model, at the edges of the organization are new use cases and untold amounts of volatile or emerging data. Value may emerge from this data only after exploration by a business analyst who can see the potential to analyze it in combination with other data toward some business end. The value in this data can rarely be found by modeling it in the abstract – it takes an analyst with an idea for how they want to use the data to pragmatically connect it with other data and forge value from it. In this way, the stable and the volatile can be pragmatically interwoven without the barrier of formal processes and long cycle times – and the model can organically emerge from the data.

However, today's reality looks a little different. It is a world of silos – some officially sanctioned data warehouses and marts, and an unknown number of 'shadow IT' systems to support everything that couldn't be done in the official systems for one reason or another.

## Promise vs Reality – The 'Enterprise Data Warehouse'

The technology and approach that came to be called 'Enterprise Data Warehousing' (or EDW) dates back to the 1980s and was introduced by Teradata. It is grounded in the idea of centralization – a single unifying data model implemented in a single central database containing the 'corporate view of data'. EDW is less about technology than data management practices – as described here by Inmon:

*"There is no point in bringing data ... into the data warehouse environment without integrating it. If the data arrives at the data warehouse in an unintegrated state, it cannot be used to support a corporate view of data. And a corporate view of data is one of the essences of the architected environment."*<sup>1</sup>

In other words, the EDW is the 'data mainframe', and every way that it is used must be carefully controlled and regimented. And when implemented by Teradata these solutions have been extremely expensive (i.e. as much as \$1million/TB), furthering the concern that everything be locked down and only the most essential business data be stored.

While there is clearly value in building well architected data warehouses, the promise that the EDW would be 'the database' for all data never came to pass. This should come as no surprise given the earlier discussion, since the principal of EDW is to solidify the data model and demand extensive process, planning and implementation hoops before introducing new data sources or supporting new business requirements.

Consequently, an EDW just can't move at the speed that the business requires. And since the business will find a way to get their work done, inevitably a multitude of shadow-IT systems will pop up to try to fill the cracks – leading right back to the silos and fragmentation that the company hoped to eliminate.

### Promise vs Reality – The 'Data Warehouse Appliances'

As a response to the inflexibility of the 'big-bang' EDW approach, Netezza introduced the idea of a 'Data Warehouse Appliance' (DWA) in the '90s. Rather than buying a mainframe, the customer can buy a 'minicomputer' for each department or application. The DWA is a pre-integrated hardware and software solution that provides a fixed amount of capacity and can be shipped to a customer in a ready-to-run state. The promise was that departments could now control their own destinies by writing a check, rather than waiting for IT to deliver an EDW.

But there is a problem with buying a data warehouse solution like it was a washing machine – you soon end up with lots of washing machines. From a business analyst perspective, you end up with a fragmented infrastructure in which each department or application's data is locked up in a box and there is no easy way to weave together data from across the company. From an IT operations perspective, you end up with a multitude of proprietary boxes that must each be managed and serviced, and that have hard capacity and performance bounds.

Again, the customer experiences silos and fragmentation, and DWA turns out to be as much of a problem as it is a solution.

### The Way Forward

In all of the discussion above, the fundamental friction stems from the competing needs of business analysts and IT operations. Business analysts want the flexibility and power to combine and analyze any data, and get their work done without resource

<sup>1</sup> W. H. Inmon. Building the Data Warehouse. Wiley, 2005.

delays or long process hurdles. IT operations wants to centralize and consolidate to reduce costs, streamline support, and deliver better quality of service. Until now these have been largely irreconcilable – but no longer.

## Business and Technology Trends

There are a number of important trends taking hold today that are rewriting assumptions of the past. These are a catalyst for a shift in what is possible.

The first of these is the inflection in the price/performance of commodity hardware. When vendors such as Teradata or Netezza were coming to market, the only way they could achieve the necessary performance was through proprietary hardware solutions. However the predictable march of commodity technologies (many-core CPUs, low cost GigE/10GigE networking, low-cost SAS/SATA storage subsystems) has overwhelmed these proprietary approaches. It is now possible to buy servers with 1,000 cores of compute for less than US\$1million, which is orders of magnitude less expensive per-core than Teradata solutions. Instead of treating hardware as a scarce resource, it becomes cheap building blocks for the software to leverage.

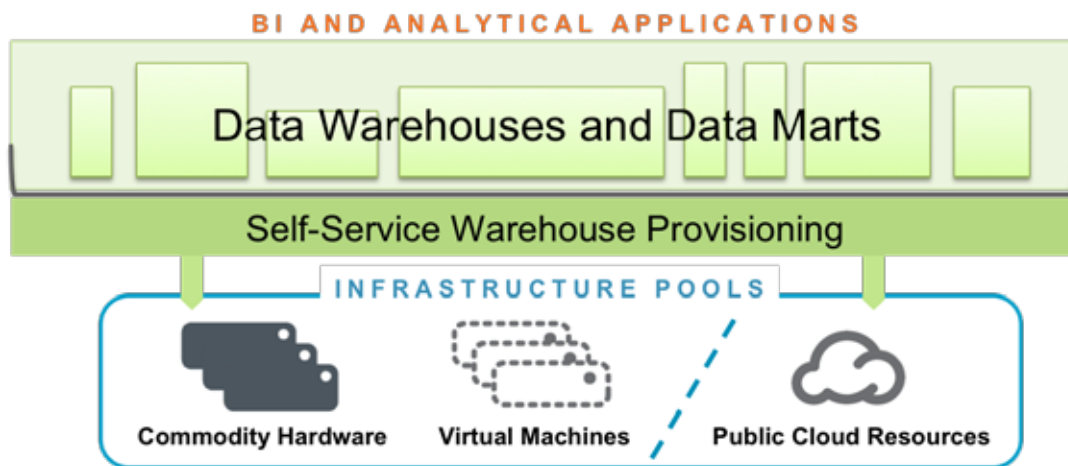
The second trend is the emergence and adoption of server virtualization and cloud computing. These technologies have demonstrated the benefits of a more flexible binding between application and hardware resources – i.e. reduced costs through consolidation and improved utilization, increased agility for rolling out new applications and services, and simplified management and operations. Today these technologies are better suited for less data-intensive applications, but they point to a direction that would be highly desirable in a data warehousing context.

The third trend is the complexity and unpredictable nature of the business landscape. While this trend is always a given, now more than ever it is clear that business intelligence and analytics have an essential role to play to help steer the ship and sense danger or opportunities sooner than competitors.

## The Missing Piece – Self Service

The EDW approach unsuccessfully pushed for grand centralization of the data and data model, while the DWA approach just accepted the creation of silos with walls between them. We can do better. There are numerous benefits to centralizing data on a common platform – as long as it doesn't make the business beholden to IT every time they want to get anything done. The answer is 'Self Service' – i.e. have IT provide a platform that allows business analysts to serve themselves without IT involvement.

In a Self-Service environment, IT gives the business the power and control to instantly provision and deploy data warehouses. A warehouse is provisioned from an 'infrastructure pool' that could consist of on-premise physical servers, virtual machines, or even public cloud resources. The Self-Service provisioning layer needs to provide a web interface to business analysts allowing them to create and manage warehouses. Meanwhile, IT operations can focus on assembling pools of 10s, 100s or 1000s of servers to hold the warehouses.



By getting this right, each party can now focus on doing what it does best. Business analysts can spin up warehouses in minutes as projects dictate, and bring together the data they need in their own project space. IT operations can manage the infrastructure pools and Self-Service provisioning platform as one infrastructure, without the need to concern themselves with the contents and usage of any particular warehouse.

The introduction of Self-Service is a breath of fresh air for everyone involved, and opens the door to new kinds of practices, use cases and collaboration that wouldn't have been possible otherwise.

## Greenplum's Enterprise Data Cloud™ Initiative

Greenplum's EDC Initiative comprises 3 core elements: Platform, Methodology and Ecosystem. Each is as important as the others in propelling the initiative forward.

### EDC Platform

The first element focuses on the product and technology aspects of the EDC Initiative – Greenplum's EDC Platform. This layer provides the technology to go beyond a single warehouse and support the provisioning and operation of 10s, 100s, or 1000s of warehouses on one or more common pools of hardware. At the highest level it is built around three key enabling features:

1. **Extreme Scale and Elastic Expansion** – The ability to handle massive scale and dynamic growth or shrinking of the volume of data. i.e. For any one warehouse within the EDC, be able to support any size from TBs up to multiple PBs of data (given enough hardware resources) and rapid system expansion as more capacity or performance is needed.
2. **Self-Service Provisioning** – Allow IT to provide pools of resources (compute + storage) and provide business analysts with a Self-Service interface (i.e. web console) to create new marts and warehouses in minutes with a few clicks. Also allow business analysts to access and combine data from any other marts in the EDC into their newly created mart.

- 3. Massively Parallel Analytic Processing** – Just storing data with the EDC is only the first step towards querying and analyzing that data. Each mart or warehouse in the EDC is a fully-featured Greenplum Database that provides massively parallel SQL, MapReduce and analytical processing. Whether the warehouse spans 2, 10 or 100s servers, Greenplum Database will leverage all available CPU, I/O and network performance to maximize the performance of each query or analytical program.

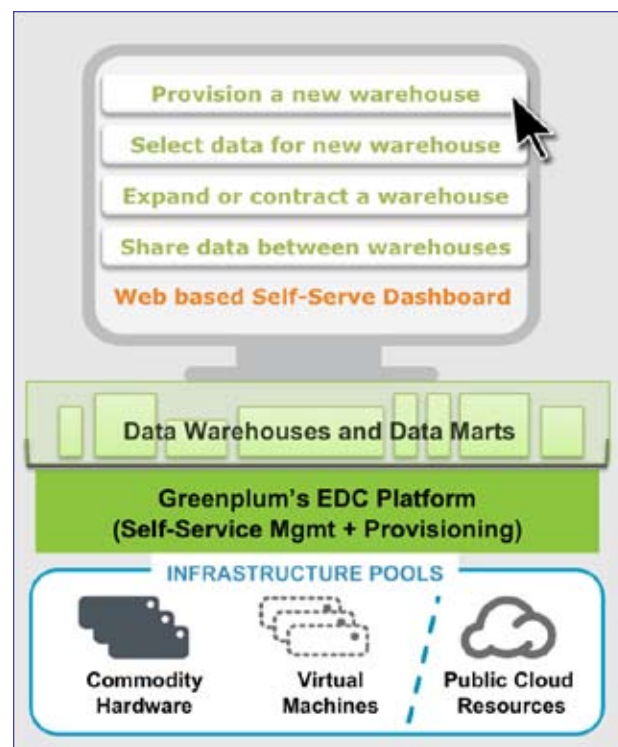
Greenplum's EDC is only possible because the Greenplum Database is a software technology. The capabilities go beyond what is possible from a proprietary hardware based Data Warehouse Appliance. Greenplum's EDC allows Self-Service provisioning from consolidated resource pools of 100s or 1000s of servers - leveraging the latest in commodity hardware, and the ability to run in the enterprise on physical or virtual hardware or deploy out to the public cloud.

### Self-Service Provisioning

Greenplum's EDC is designed to empower both the business analyst and the IT organization, and allow each to focus on getting its job done without unnecessary friction with the other. By cleanly abstracting the interface between the organizations, IT can standardize and centralize infrastructure and provide a Self-Service 'menu of service' to the business, while the business can select from those services and have them delivered in an automated fashion without requiring hands-on IT involvement.

From the business analyst's perspective this is a huge step forward than today. Rather than lobbying for months for database resources and access to data, a business analyst can fill out a web form to automatically create a new warehouse in minutes. The analyst can grow the size and performance of the warehouse elastically as needed. And the analyst can keep it indefinitely or use it just for a short period before allowing the resources to flow back to the available pool.

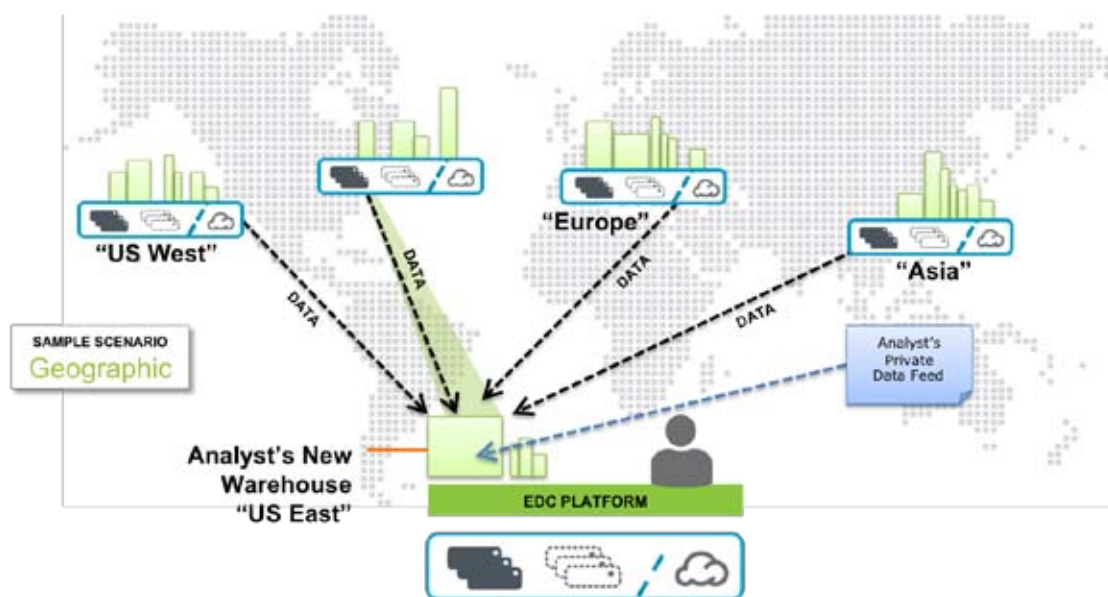
The business analyst can treat this as their dedicated warehouse, with the confidence that they are fully isolated from other warehouses and can query it as intensively as they like without impacting any other warehouses. They can load their own data, and can also bring together data from any other EDC warehouses into their warehouse with a few clicks. From their perspective, they are in control.



From IT's perspective the platform presents huge benefits over the status quo. Instead of supporting and managing hundreds of separate data warehouse marts and appliances, IT can treat the EDC as a single infrastructure composed of one or more consolidated server resource pools (each utilizing physical hardware, virtual machines, or public cloud infrastructure). Each pool could just as easily be a single machine or 1000 servers, running Greenplum's EDC Platform services, and serving 10s or 100s of marts and warehouses to the business.

Instead of being concerned about free space and servicing of each individual appliance, IT can plan capacity and track utilization at the granularity of a pool. Additional commodity servers can be added periodically as demand dictates, and become available to be provisioned into new or expanding warehouses. IT can treat the EDC pools as a utility – focusing on utilization, power efficiency and cost reduction, and can deliver higher quality of service to the business.

### Unified Data Access



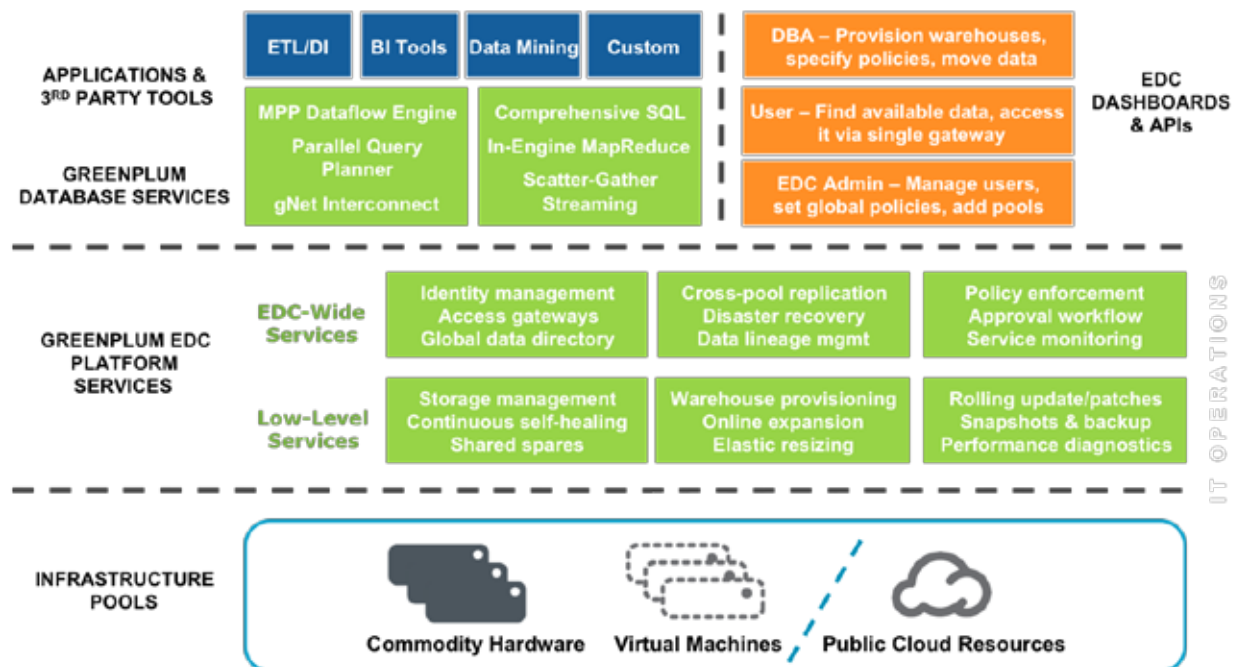
The value of EDC to the business goes beyond Self-Service provisioning of independent marts and warehouses. Unlike a data warehouse appliance, the walls between marts and warehouses within an EDC are not impermeable. Analysts can select data they are interested in anywhere in the EDC, and access it directly or bring slices of it together in their warehouse. This data movement is very fast because it leverages Greenplum's Scatter/Gather Streaming technology to move data between marts (whether local or geographically remote) using all available parallelism for maximum efficiency.

So, for example, a business analyst could take three months of clickstream data from one mart, combine that with customer data from a central warehouse, and add to the mix an interesting new data feed that they load directly. In this way they could

combine and analyze data that on their terms without being constrained by the silos and fragmentation that are endemic today. They could also make their data available for use by other analysts in the EDC.

Security and access control play an important role in all of this. Not all data should be available to all analysts, and some users will naturally have access to data that others don't. The EDC Platform provides appropriate controls to allow data to be shared appropriately in a way that reflects enterprise policies.

### EDC Technical Stack



Fully realizing the promise of Self Service data warehousing for an enterprise requires a new set of platform services that go beyond traditional database or virtualization infrastructure. This is the EDC Platform Services layer.

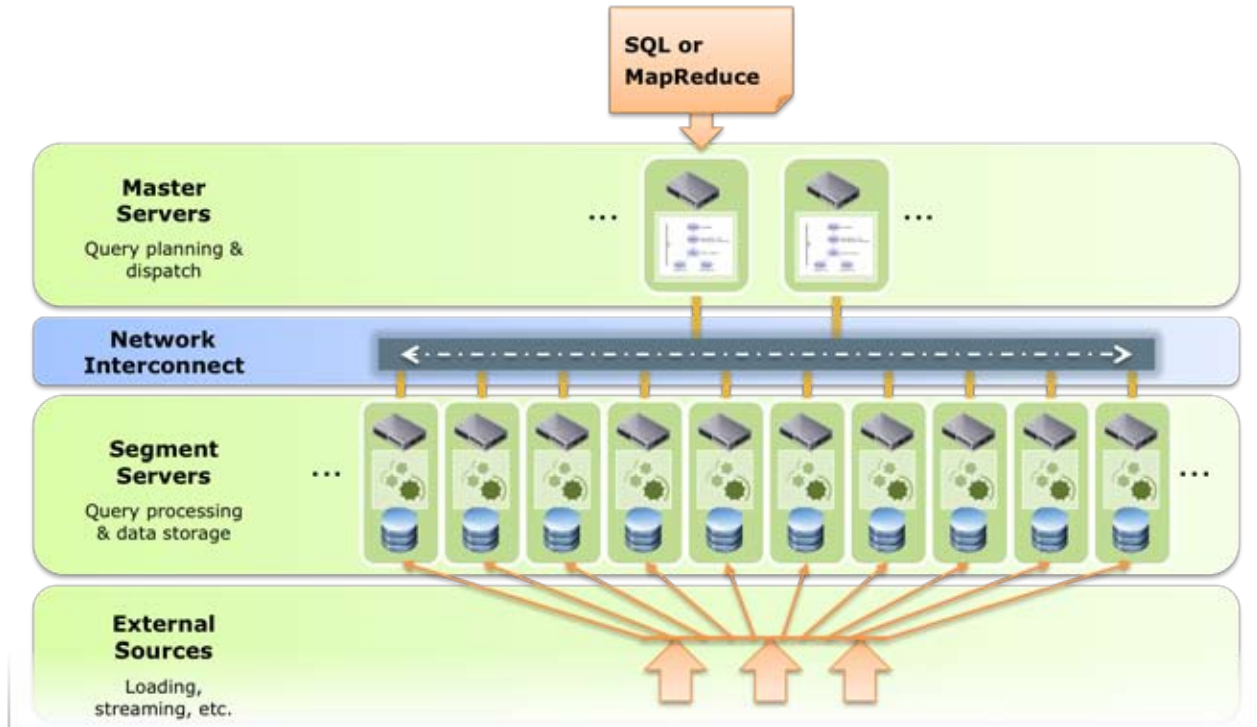
The EDC Platform Services layer is actually two layers of services:

'Low-level services' span each infrastructure pool (i.e. pool of servers) and provide aggregated storage management, fault-tolerant self-healing, low-level provisioning and expansion mechanisms, and operational capabilities like updating and patching.

'EDC-wide services' provide those services that span multiple infrastructure pools across the enterprise. These higher-level services include access and identity management, a global data directory, cross-pool replication including disaster recovery support, data lineage tracking for data replicated between marts, and policy and approval mechanisms.

Above this layer are the EDC dashboards and APIs that DBAs/business analysts, EDC admins, and end users can use to interact with the platform.

## At the Core – Greenplum’s MPP Database



At the heart of every mart or warehouse on the EDC is the Greenplum Database. Greenplum Database utilizes a shared nothing, massively parallel processing architecture that has been designed for business intelligence and analytical processing. The database is designed for massive scalability and multi-level fault tolerance using commodity CPUs, storage and networking. It natively supports SQL queries (SQL-92, SQL-99 and SQL 2003 OLAP extensions) and MapReduce processing. It provides automatic parallelization of data and queries – all data is automatically partitioned across all nodes of the system, and queries are planned and executed using all nodes working together in a highly coordinated fashion.

Greenplum Database supports elastic resizing via its online system expansion capability. New servers can be added to a database in minutes, and data is automatically rebalanced across available storage while the system is online and serving queries. Each new server increases the database’s storage capacity, query performance, and loading performance.

For more details about the Greenplum Database, see the Greenplum Database Whitepaper (available at [www.greenplum.com](http://www.greenplum.com)).

## EDC Methodology

As important as the platform are the new data warehousing methodologies that are being demonstrated by leading Greenplum customers. EDC makes possible new approaches to data management that fundamentally challenge the orthodoxy of the past 20+ years.

The traditional practice of data warehousing, as described earlier, focuses on establishing a single 'corporate view' of data via up-front modeling of all data before it is allowed into the data warehouse. This is analogous to requiring every business user in the company to share a single page of an Excel spreadsheet for all their work, and force endless meetings to discuss the organization of the page. It is unrealistic and is a drag on both business and IT.

A class of 'new practitioners' – thought leaders who are challenging these principles in the way they implement warehouses – has started to emerge. Greenplum is fortunate to count many of these 'new practitioners' amongst its customers. The key principles they espouse include:

- Get data into the EDC and out to the business teams ASAP
  - Focus on providing access to the raw data as quickly as possible, rather than waiting for up-front modeling. The value of the data will only become valuable through use, so allow that to proceed organically and use modeling as a way to pragmatically organize these findings.
- Model Less, Iterate More
  - Look for pragmatic value to the business from day one, and allow the model to emerge iteratively based on the usage that emerges.
- Optimize for operations instead of performance
  - Don't complicate the data model by prematurely optimizing for performance. The database is so fast that most of these optimizations are irrelevant, and it is better to organize the data in a way that is easily accessed and consumed by business analysts.
- Embrace Self-Service data infrastructure
  - Reduce friction between business and IT to achieve faster results at lower cost
- Implement analyst sandboxes
  - Empower analyst to explore data on their own terms without slowing them down with processes
- Close alignment of analysts and DW team
  - Ensure that business analysts and the data warehousing DBAs work closely and collaborate. Each play an important role in this agile iterative cycle.

## EDC Ecosystem

The move towards EDC creates a profound shift in the way that companies think about their data infrastructure. It also creates a new frontier of opportunities for disruptive new technologies, capabilities and standards – in areas such as servers and storage, high-speed and wide-area networking, BI/analytics (visualization, in-memory processing, data discovery and metadata, collaboration) and data processing (Map-Reduce, real-time event processing, etc).

While Greenplum is excited to be leading the charge towards EDC, there will be countless companies that play a role in fully realizing the vision of EDC and the new opportunities it presents.

## Use Cases

### Use Case: Data Mart Consolidation



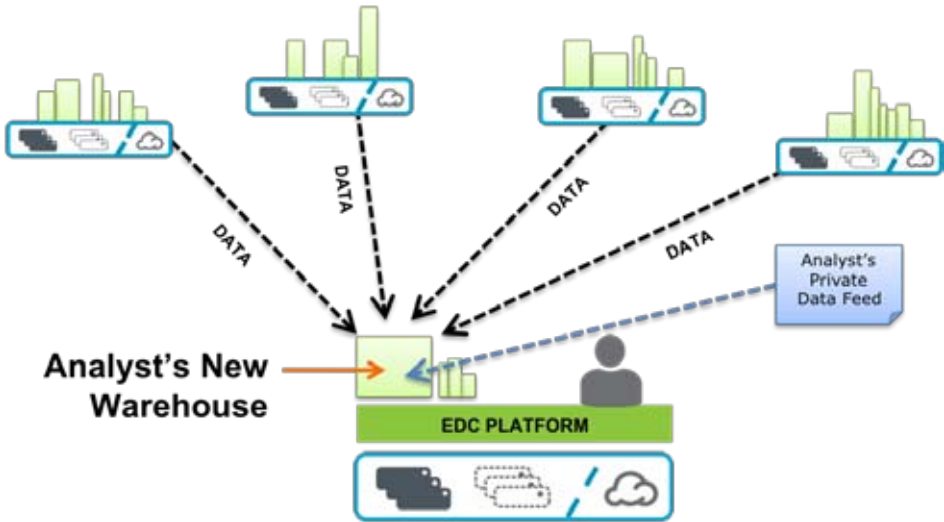
### Goals:

- Reduce maintenance and support costs from proliferation of data mart platforms
- Reduce risks and exposure due to data in shadow IT systems
- Break down silo walls - provide a unified way to find and access all data

### Approach:

- Embrace data – encourage 'physical consolidation' in advance of data model unification
- Provide Self-Service model to bring shadow IT into the light
- Allow unified data access and pragmatic 'logical' data model unification incrementally

### Use Case: Project Sandboxes



**Goals:**

- Remove IT barriers to analyst productivity and value creation
- Dramatically reduce IT resource constraints and delays – i.e. realize ideas sooner
- Combine centralized 'EDW' data with freshly discovered feeds and other useful sources

**Approach:**

- Self-Service creation of project warehouses in minutes – and elastically expand as needed
- Load new data feeds without requiring formal modeling
- Bring together any data within the EDC – even if globally distributed – and analyze

### Conclusion

Greenplum's Enterprise Data Cloud™ Initiative is built around a bold new vision for bringing the power of Self-Service to data warehousing and analytics.

It comprises 3 core elements: Platform, Methodology and Ecosystem. It empowers both business and IT to do more and work together without friction.

Greenplum's EDC™ Initiative is being embraced by major customers and partners today.