MASSIVEDATANEWS

Load and Go: Fast Data Loading with the Greenplum Data Computing Appliance (DCA)

Introduction: Why Fast and Flexible Data Loading Matters

Data loading is the beginning of the entire analytics process. Everything starts by getting data into the data warehouse. The minute the data loading process becomes difficult or slow, it becomes a barrier to a company's agility and ultimately its performance. The focus on fast data loading capabilities parallels a trend toward real-time Data Warehousing (DW), since DW is becoming the foundation for growing time-sensitive business practices such as operational business intelligence (BI), frequently refreshed management dashboards, just-in-time inventory, self-service information portals, and on-the-fly recommendations in e-commerce. When data loading isn't fast, business analysts scale down their data sets as a work-around to system rigidity. Ultimately, that situation leads to less-than-optimal business insight because analysis lacks all the data available to a business user. A DW solution that can expedite the data loading/ingestion step will—irrefutably—help increase business agility and performance.

The Old Way Is the Slow Way

Today the business imperative is to analyze data on-the-fly—not days or weeks after it has been stored in the database. The pressure is on IT to provide systems and service levels that support getting insight out of ever-larger data sets faster than traditional data warehouse systems. Traditional data warehousing is a relatively slow producer of information to business users who depend on analytic insight to do their job. Data warehouses are traditionally refreshed in a periodic manner, most often on a daily basis. Thus, there is some delay between a business transaction and its appearance in the DW. Fresh, relevant data is trapped in operational data stores, where it is unavailable for real-time analysis by the business user.

The traditional data warehouse approach is often too slow for rapid decision-making in organizations. Moreover, the traditional DW model falls short of including ALL the relevant information available to the business. Greenplum's fast data loading capabilities address the dataloading challenge by providing automatic and optimized data distribution and delivering source data to the data warehouse with lower latency.

How Greenplum Database Loads Data Fast

Because shared-nothing databases automatically distribute data and make query workloads parallel across all available hardware, they dramatically outperform general-purpose database systems on BI and analytical workloads.

Greenplum leverages Scatter/Gather Streaming™ technology to provide the highest-performing data loading capabilities in the industry. This technology eliminates the bottlenecks associated with other approaches to data loading. At a high level, Scatter/Gather Streaming:

- Manages the flow of data into all nodes of the database
- Does not require additional software or systems
- Takes advantage of the Greenplum Parallel Dataflow Engine

Greenplum utilizes a "parallel-everywhere" approach to loading, in which data flows from one or more source systems to every node of the database without any sequential choke points. This approach differs from traditional bulk loading technologies—used by most mainstream database and massively parallel processing (MPP) appliance vendors—which push data from a single source, often over a single channel or a small number of parallel channels, and result in fundamental bottlenecks and ever-increasing load times. Greenplum's approach also avoids the need for a "loader" tier of servers, as required by some other MPP database vendors, which can add significant complexity and cost while effectively bottlenecking the bandwidth and parallelism of communications into the database.

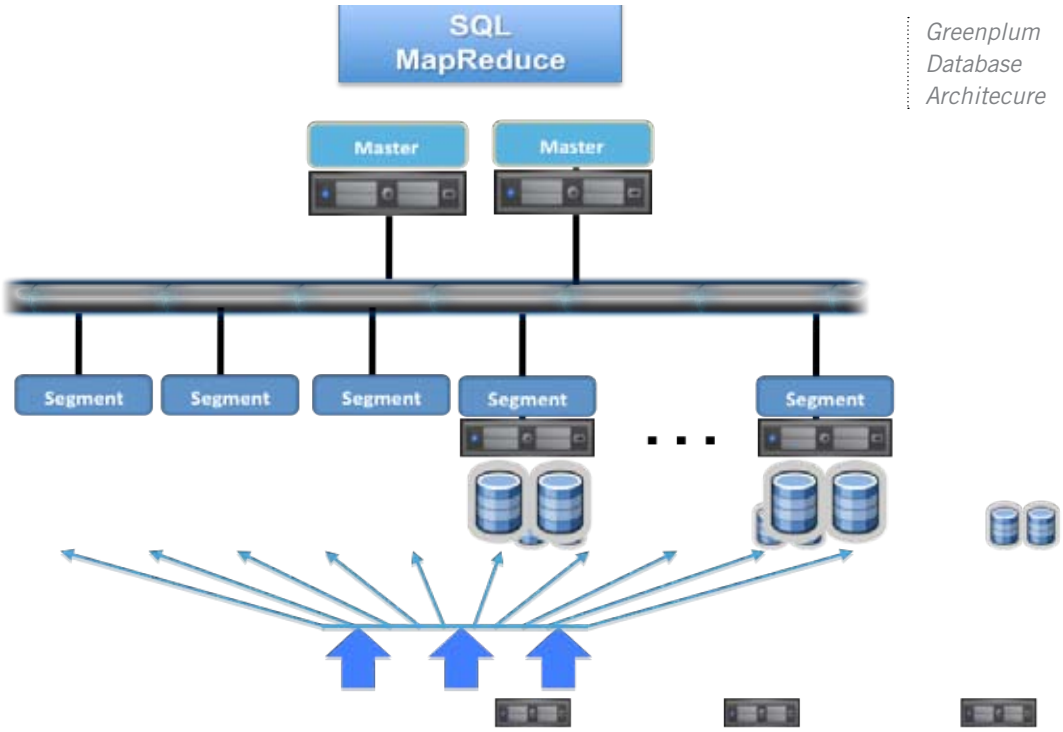
Industry Perspective by John Webster, Analyst, The Evaluator Group

The EMC Greenplum Data Computing Appliance

The Greenplum Data Computing Appliance (DCA) is built on a shared-nothing, massively parallel processing architecture. Each processing unit acts as a self-contained database management system that owns and manages a distinct subset of the database. The Greenplum DCA automatically distributes data and parallelizes query workloads across all available hardware within the server/network/storage stack, and it can scale to thousands of CPU cores.

This purpose-built product is a pre-integrated hardware and software framework. Its core query engine is the Greenplum Database, which offers both highly optimized SQL support and the option of programming with MapReduce directly against datasets.[1] A parallel query optimizer converts SQL or MapReduce into a physical execution plan. It also uses a cost-based optimization algorithm to evaluate potential plans and select the one that will lead to the most efficient query execution.

The Greenplum DCA offers system administrators the ability to tune the appliance to provide optimal performance in complex mixed-data environments. Administrators can set priorities to determine fair sharing of resources for all users. Queries can also be reprioritized on the fly to account for changing conditions.



Greenplum Database Architecure

[1] For more information see EGI Research Note entitled "EMC Builds a new Data Computing Division Around Greenplum.

Architecture

The Greenplum appliance is composed of Master Servers for user interface and control, Segment Servers for parallel query processing, and a gNet system communications backbone. Each segment server has a dedicated complement of direct-attached storage.

Segment Servers—These servers do the actual query processing in parallel. Storage for these servers is directly attached (no SAN or NAS is used) to minimize server-to-storage I/O latency. Each segment server is loaded directly using both internal and external data sources. Dedicated RAID storage can be mirrored to other storage that is attached to other Segment Servers.

Master Servers—These servers function as the control point between business analytics that application users provide and the Greenplum DCA business analytics tools. They also optimize process flow to optimize query efficiency. RAID storage is provided for Master Server data protection (RAID 0, 1, or 5). Master Server transaction logs are also replicated.

gNet—This high-speed bus pipelines data among Master and Segment Servers using Gigabit Ethernet and 10GigE switch technology. It supports the continuous pipelining of processing without blocking on all nodes of the system and can scale to tens of thousands of processors.

Parallel Dataflow Engine—One of the key features that enable the Greenplum database to scale linearly and achieve such high performance is its ability to utilize the full local disk I/O bandwidth of each system. The Parallel Dataflow Engine is an optimized parallel processing infrastructure that is designed to process data as it flows from disk or from other segments over the gNet interconnect. The engine is inherently parallel—it spans all segments of a Greenplum cluster and can scale effectively to thousands of commodity processing cores and no special tuning is required.

Polymorphic Data Storage—This feature provides built-in flexibility that allows customers choose the right strategy for their environment. Data can be stored in rows or column with optional compression, so users can specify the storage option that is best for their workload. These settings are NOT systemwide, so multiple strategies can be addressed in the same platform.

Scatter/Gather Streaming

Scatter/Gather Streaming is a significant step forward in business analytics because it does more than just parallelize the simultaneous ingest of multiple data streams from multiple sources. This capability alone represents a big advancement over the single-source, extract-transform-load (ETL) process common to the more traditional data warehouses. The Scatter/Gather Streaming approach also allows the sources for that data to be located off the premises. These sources can reside on other websites or within other organizations connected to the user—a supply chain partner, for example.

With Scatter/Gather Streaming, data flows in parallel from multiple source systems to every node of the database without intervention on the part of other operations that add latency. For example, this capability eliminates the need for an additional tier of servers that function only as data warehouse “loaders.” Rather, Database Nodes can be added to scale up performance on the data ingest side. This technology also supports both large batch and continuous near-real-time loading. Data is transformed and processed on the fly, utilizing all available Database Nodes. The final “gathering” of data—in other words, when data is written to disk—is done by all nodes simultaneously. Data is automatically partitioned across nodes and optionally compressed.

As mentioned previously, though, the ability to leverage external systems is especially noteworthy because it allows data that resides within—and is potentially owned by—other systems and other organizations to be used as sources. These sources can exist in the form of flat files, web pages, and other databases. The Scatter/Gather load process selects from these sources and extracts data in parallel streams, loading the database without using a dedicated loading node or the intervention of the Greenplum DCA Master Server to load the Segment Servers.

Why Fast Data Loading Matters: Real World Applications

Consider a large retailer with outlets across the United States. One of the biggest problems facing retailers is unaccountable product loss, otherwise known as "shrinkage." Product loss can occur at multiple points along the supply chain, from the producer of an item to the checkout counter. Technologies such as optical scanning devices and RFID can be used to monitor the product flow along the supply chain. But two challenges must be met to effectively control loss of product as it moves along the supply chain:

1. The volume of data that the digital scanning and RFID devices produce can be huge for a large national retailer with hundreds of outlets using these technologies at checkpoints along the supply chain.
2. Time to information is critical when trying to recover a lost shipment, for example. The longer it goes unaccounted for, the harder it will be to find. To harness the value of this data, relevant data points have to be extracted quickly and often in real time from large volumes of data as they are produced.

To extract relevant data, the integrated hardware/software stack must be able to load and parse massive amounts of data from multiple sources, in a way that maximizes the value of transient data.

Other applications of fast data loading include:

- An IP telecommunications services provider who wants to ensure quality of service by monitoring networks in real time, understand customer behavior patterns, and offer targeted customer loyalty programs.
- A financial services provider who needs to quickly detect and prevent fraudulent activity as it is occurring.
- A services provider for smartphone users who wants to combine flight-delay records data with current, real-time conditions "on the ground" to send flight-delay warnings to subscribers.

Conclusion

Traditional data warehousing is now a relatively slow producer of information to business analytics users. It draws from limited data resources and depends on time-consuming, reiterative ETL processes. Data loading no longer has to be a slow, batch-oriented process anymore, however. In today's world, real-time insight has a high value, and that value decays the longer it takes to get data to the business. What business analytics users are looking for is speed to information using multiple data sources concurrently.

Traditional data warehousing is now a relatively slow producer of information to business analytics users. It draws from limited data resources and depends on time-consuming, reiterative ETL processes. But, data loading no longer has to be a slow, batch-oriented process. Innovations in database technologies can now deliver extreme performance improvement in the data loading step. The Greenplum DCA offers the ability to parse large data sets from multiple sources and produce information in real or near-real time.

Learn More

Take The Test: Real World Data Loading Performance

When examining performance in any system, most administrators look to benchmarks for guidance on how the system will eventually perform. In theory, benchmarks provide an example of real-world performance, but the reality is that these numbers are often suspect or just plain wrong. It's no secret that most benchmarking results cannot be re-created in production environments. Vendors cook the books by spending a huge amount of time having the engineering experts fine-tune the system to create the best numbers. While these benchmarking tactics are defensible (barely), most organizations are unlikely to achieve the published results unless they have an amazing team or a huge consulting budget.

Greenplum and the EMC Data Computing products Division are now producing real world benchmarks. No more obscure tests against formula-one tuning. Instead we would like to present the beginning of what we are calling real-world benchmarks. These benchmarks are designed to reflect the true customer experience and conform to the following guiding principles:

1. Test on the system as it leaves the factory, not the laboratory.
2. Create data types and schemas that match real-world use cases.
3. Consider options beyond raw bulk loading.

In partnership with their customers, Greenplum created a number of database schemas and sample data sets that are designed to mimic customer environments. By loading a sample set based on a standard, customers can instantly understand how many records the system can ingest and then use this information to better plan updates and reporting schedules. Greenplum has defined the first two examples as follows:

1. Internet and Media—Web clicks and behavior tracking
2. Retail—Line-item entries

For retail we have defined the schema as follows:

```
CREATE TABLE lineitems (

Order_ID                    INTEGER,
Order_Item_ID               INTEGER,
Ordering_Session_ID         INTEGER,
Product_ID                  INTEGER,
Customer_ID                 INTEGER,
Store_ID                    INTEGER,
Item_Quantity               INTEGER,
Order_Datetime              TIMESTAMP,
Ship_Datetime               TIMESTAMP,
Return_Datetime             TIMESTAMP,
Refund_Datetime             TIMESTAMP,
Coupon_ID                   INTEGER,
Payment_Method_ID           INTEGER,
Product_Group_ID            INTEGER,
Tax_Amount                  DECIMAL(10,3),
Discount_Amount             DECIMAL(10,3),
Coupon_Amount               DECIMAL(10,3),
Product_Group_Name          VARCHAR(200),
Product_Name                VARCHAR(200),
```

```
Ship_Address_Line1            VARCHAR(200),
Ship_Address_Line2            VARCHAR(200),
Ship_Address_Line3            VARCHAR(200),
Ship_Address_City             VARCHAR(200),
Ship_Address_State            VARCHAR(200),
Ship_Address_Postal_Code      VARCHAR(20),
Ship_Address_Country          VARCHAR(200),
Ship_Phone_Number             VARCHAR(20),
Billing_Address_Line1         VARCHAR(200),
Billing_Address_Line2         VARCHAR(200),
Billing_Address_Line3         VARCHAR(200),
Billing_Address_City          VARCHAR(200),
Billing_Address_State         VARCHAR(40),
Billing_Address_Postal_Code   VARCHAR(20),
Billing_Address_Country       VARCHAR(200),
Billing_Phone_Number          VARCHAR(20),
Customer_Name                 VARCHAR(200),
Customer_Email_Address        VARCHAR(200),
Website_URL                   VARCHAR(500)
```

Using this schema, we tested the Data Computing Appliance loading speeds and achieved the following results.

Rows per Second	4.77 Million Rows per Second
TB per Hour	10.4TB per Hour

These results are from a factory default system. The number of segments was not changed, and we did not load multiple tables at the same time to cook the results. Given a similar schema, Greenplum DCA customers can expect 4.77 million rows per second on a typical retail schema. That number equates to a billion rows every 3.5 minutes. Imagine the just-in-time inventory decisions that could be driven by leveraging a micro-batching loading approach.

More detail on the schema samples, data generators, and initial benchmarking results are available on the EMC Greenplum website at the following URL: www.greenplum.com

As we continue with this project, we will look to identify additional schemas based on customer feedback. Currently we are looking at adding:

1. Telecommunications—Call Data Records (CDRs)
2. Financial Services—Transaction records

Going forward, Greenplum will identify additional schemas based on customer feedback.

EMC Greenplum has commisioned Massive Data News to conduct an on-going real world benchmark and performance testing program. They will continue to expand the real world data types and schemas to include additional use cases. For more information visit www.massivedatanews.com.