Reference architecture

How to Build and Deploy an All-in-One Data Lake, Data Mart and Data Warehouse

Analytics at scale in the modern enterprise with VMware Tanzu Greenplum and Dell EMC Isilon storage with OneFS

Abstract

Businesses are generating data at an exponential rate. To compete effectively, IT leaders need to store, collect and analyze their enterprise data. Ideally, this "modern analytics stack" would feature open-source technology and pluggable hardware that can easily scale as the business continues to collect and analyze more data. A combination of VMware Tanzu Greenplum, VMware virtualization and Dell EMC infrastructure, can provide a flexible solution that is cost effective, easy to build and simple to manage.

February 2021

Revisions

Date	Description
November 3, 2020	First draft
December 4, 2020	Initial edit [JR]
February 12, 2021	Update

Acknowledgments

This paper was produced by the following authors:

- Chris Gully, Staff Solutions Architect — VMware Office of the CTO
- Ivan Novick, Manager, Product Management, VMware Tanzu Data Services — Greenplum
- Hamid Djam, Technical Staff, Product Technologist — Dell Technologies
- Scott Kahler, Product Line Manager, VMware Tanzu Data Services — Greenplum
- Suhail Gulzhar, Manager, Platform Architecture — Dell Technologies

Support:

All testing and validation were completed in the Dell Technologies Customer Solution Center with support from Guy Laporte, Solutions Consultant.

Executive summary

The effective management and analysis of data is a competitive advantage. To this end, IT leaders have evaluated many different concepts over the years: Apache® Hadoop®, data lakes, data marts and data warehouses leap to mind. All of these have their merits. But siloed, single-purpose solutions are counter-productive. More enterprises are turning to systems capable of handling a wide range of use cases.

VMware® Tanzu™ Greenplum® with Dell EMC PowerScale and Dell EMC Isilon storage is a popular choice. This combination provides a scale-out solution for data analytics that is performant and space-efficient. This configuration has proven to be flexible and addresses several industry-standard use cases in a single solution. Use cases solved with this architecture include landing zones for data ingestion, data lake storage, data science, data warehouses and data marts. IT leaders can support all these scenarios with a single architecture that is easy to deploy, scale and support.

This architecture can start small and scale out as data needs increase. Data is stored in multiple storage tiers. Storage can be expanded at multiple tiers using Dell EMC Isilon and VMware vSAN™. Virtual compute is scaled and managed by VMware vSphere®. Tanzu Greenplum, powered by massively parallel Postgres, handles SQL-based analytics of structured and unstructured data. Administrators can easily extend Tanzu Greenplum with additional analytical tools, such as Apache® Spark® and commercial data science engines, as requirements evolve.

In total, the solution can store and manage raw, unstructured data as well as cleansed, structured data; expose the data in multiple protocols (e.g., NFS, S3 and SQL); provide the data warehousing and data mart capabilities; and expose the raw data for access by other tools.

Table of Contents

Enterprise data architects face common data management challenges.

Enterprise data architects have a tough job of standardizing on common patterns, creating solutions for data storage and analysis across increasingly complex requirements. Crucial capabilities include the ability to make rapid changes to the analytics stack, deliver new capabilities to production quickly, and to maintain stability and scalability of the stack at a low cost.

How can data architectures address these challenges? We see three common paths: Do-it-yourself (DIY) solutions, public cloud deployments and integrated solutions. Let's examine some of the options available to data architects, and the potential pitfalls.

The DIY analytics stack: A growing maintenance burden

The modern data landscape has hundreds, if not thousands, of potential commercial and open-source technologies. Why not pick and choose from this landscape? In reality, this option comes with an explosion of cost and complexity over time.

If data architects are not careful, they can find themselves supporting a proliferation of technologies that are cost-prohibitive to maintain with brittle integration points. Each technology added to the enterprise data stack requires administrators to manage, upgrade and provide high availability (HA) for that portion of the stack. Some modules may even require in-house staff to develop custom management protocols. And this operational burden grows over time as the deployment scales, and as individual modules evolve and change.

What's more, each component may have different infrastructure demands. Some may require generous allotments of hardware resources, spanning many racks of data center space. Data center space on-premises becomes a scarce commodity.

The public cloud deployments

Hosting, scaling and processing analytics in the public cloud is an option for consideration. It should be noted that consumption of public cloud services has daily, monthly and annual recurring charges to support this large list of technologies.

If technologies are not portable — and many public cloud services feature proprietary APIs — then different processes and code may be required across environments. For example, Amazon Web Services® and Azure® have different disaster recovery processes. All of this toil adds cost and complexity. Costs can spiral ever higher and become burdensome to support an enterprise data architecture.

At the same time, the architect's job is to balance cost and complexity, with the fulfillment of the business users' functional requirements. Functional requirements include the APIs, interfaces and supported languages. Organizations also require performant processing of data and elastic scaling as business needs change. Some closed protocol data systems create data silos and prevent efficient sharing of data in an organization.

Integrated solutions
This option balances robust capabilities with a bare minimum of components. This path offers enterprise data architects a terrific value proposition: scalability, performance, security, and HA with fewer moving parts.

The power and capability of this reference architecture comes from the complementary nature of the parts that compose the full system. In the section titled, "Use cases where integrated stack provides key capabilities," we will cover several use cases that demonstrate the complementary benefits of the stack components.

- Dell EMC Isilon as a landing zone for incoming data to the platform
- Federated query processing of data in Isilon via the Greenplum engine
- Partitioning of Greenplum data for tiered storage
- Unloading of specific data sets from Greenplum to Isilon for future usage
- Spark processing of data living in Greenplum
- Spark processing of data living in Isilon
- Spark ingestion/ETL with data loading into Greenplum and Isilon
- Backup and restore of Greenplum data to/from Isilon

Why Tanzu Greenplum is the all-in-one solution for modern analytics at scale

VMware Tanzu Greenplum is a massively parallel database that excels at managing terabytes of data. More importantly for data architects, Greenplum provides core relational database features and a series of integrations, extensions, analytic modules and extensibility methods. Greenplum is a relational database and a platform for data storage and analytics. That's why it's the linchpin of the modern analytics stack.

Of all the big data products, Greenplum is unique in its support for diverse workloads. Let's consider them now.

- The cornerstone functionality of the Greenplum platform is ANSI SQL for aggregation, grouping, reporting and general traditional data warehouse workloads.
- Greenplum has a broad spectrum of in-database analytics methods that can be called natively from inside Greenplum SQL. Training of machine learning (ML) models is handled by Greenplum's Apache MADlib®, a bundled module. MADlib performs supervised ML, unsupervised ML and neural network–based deep learning (DL).
- Greenplum's geospatial modules support analysis of physical and logical spaces. Similarly, a graph model crunches business graph data.
- Windowing and time-series functions report on time-series data.
- Text-based unstructured data can be indexed, searched and analyzed with natural language processing (NLP) algorithms.

- JSON and XML data can be natively stored and parsed in Greenplum in data type-specific columns for these semi-structured formats.
- For any analytics outside of Greenplum's pre-defined analytics methods, programming languages such as Python®, R and Java® can be used to build custom analytical functions and modules.

The extensibility of Greenplum is inherited from open source PostgreSQL, the core technology that powers Greenplum. PostgreSQL is a leading Rational Database Management System (RDBMS) engine and provides for easy extensibility including the creation of user-defined types and functions. With this variety of analytical methods, all sorts and shapes of data stored in Greenplum can be analyzed.

It is important that data architectures avoid silos. In this undesirable state, data is consumed and accessed in isolation. Greenplum natively prevents this scenario. In fact, the system's architecture assumes that Greenplum will be deployed as part of a wider enterprise environment. The Platform Extension Framework (PXF) is a federated query framework that extends the SQL and analytical capabilities of Greenplum. Users tap into PXF to access data hosted via S3, HDFS or SQL protocols. From there, PXF provides Greenplum dynamic access to data living in multiple file formats, including CSV delimited, JSON, AVRO, ORC, and Parquet. PXF extends the reach of the Greenplum capability to external storage, including data stored on the Isilon platform.

The Greenplum Streaming Server connects Greenplum to Apache Kafka® and other high-speed message buses, to enable real-time streaming data ingestion. A diverse ecosystem of business intelligence and data science front-end tools are certified to work with Greenplum, adding to the utility of the platform.

Apache Spark is a popular compute engine for big data analytics and data science. The Greenplum platform comes with a native Apache Spark integration. Here, data stored in Greenplum can be rapidly ingested and processed as Spark jobs. Likewise, Spark computation results can be quickly pushed back into Greenplum with the integration.

Greenplum is one of a very few databases that run analytical and transactional workloads. The product can analyze data and perform traditional relational database indexing, multi-statement transactions, row-level locking and low latency transactions.

Operability and enterprise readiness come into focus as the analytics stack reaches scale and production usage. Enterprise readiness includes data security, HA, backup and restore, disaster recovery, upgrade, and monitoring. Greenplum has a remarkable track record of enterprise readiness, covering millions of hours of production workloads. Even when unexpected events occur, Greenplum has proven to stay online and available for critical business processing.

Portability of an analytics platform should also be considered. At its core, Greenplum is open-source software and an open platform, freely available for download and deployment on preferred infrastructure. The architecture described in this document is a recommended approach to Greenplum deployment; however, even this approach can be deployed on-premises, or in major cloud platforms such as Microsoft® Azure that support running vSphere ESX® server, vSAN and Isilon's OneFS.

Greenplum architecture overview

Here's a concise overview of the Greenplum parallel architecture, and how external data can be ingested and analyzed by Greenplum.

Greenplum massively parallel architecture

Greenplum is a large, clustered database where data is automatically divided and stored in many smaller PostgreSQL databases. There is a Master/Coordinator PostgreSQL instance that stores metadata. Further, this instance accepts user connections, parses incoming queries, creates optimized query plans and dispatches said query plans to other PostgreSQL segment instances.

From there, the segment instances execute the query, and send the results from all the PostgreSQL segments to the Master/Coordinator. The Master/Coordinator then returns consolidated SQL responses to the user. Every PostgreSQL segment instance in a Greenplum cluster has a primary and a mirror instance that runs on different virtual machines (VMs) for HA.

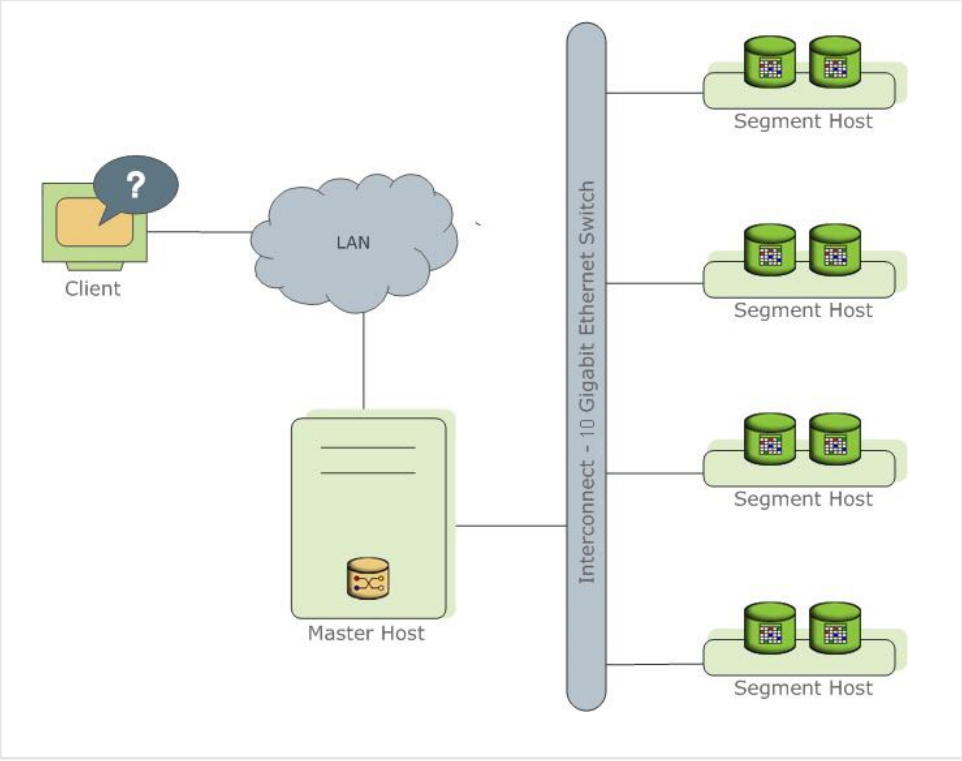


Figure 1: *The basic architecture of a typical Greenplum deployment.*

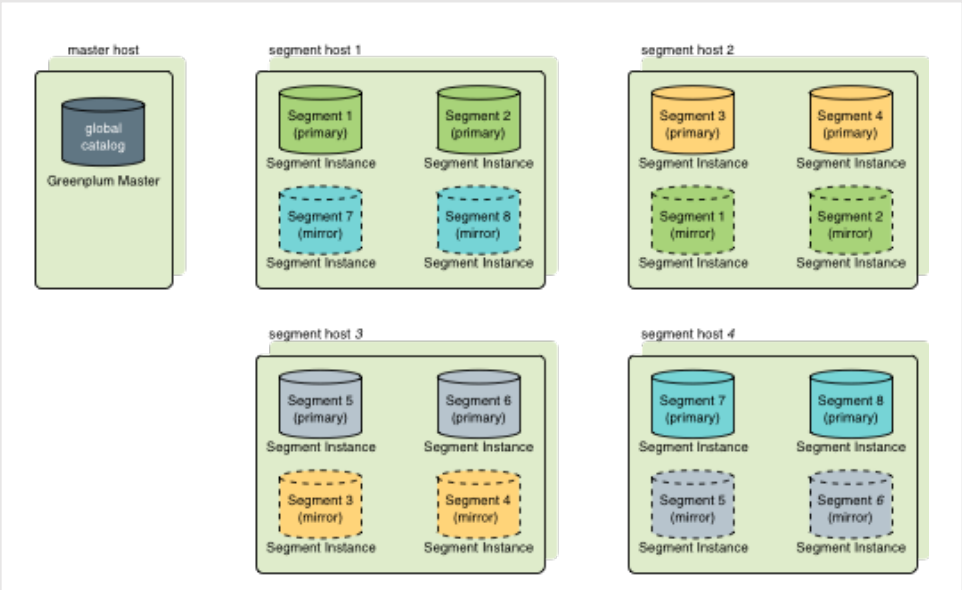


Figure 2: *The basic building blocks of a Greenplum cluster.*

External tables in Greenplum

In Greenplum, external tables have a schema definition created in SQL. As the name suggests, external tables reference database tables stored externally in an outside storage media. The outside storage media could be Isilon, HDFS, NFS, S3 or other options.

When doing a SELECT query from an external table, each PostgreSQL segment in the Greenplum cluster will create external scan operators to fetch data rows via a native protocol from the external system. After the scan operator has fetched the data, the rest of the query plan will be executed internally in Greenplum, just like any other SQL query. The primary difference is simply the mechanism of data scanning.

When doing an INSERT query statement in Greenplum to an external table, data is written to the pre-defined external storage media via a native protocol for the associated storage.

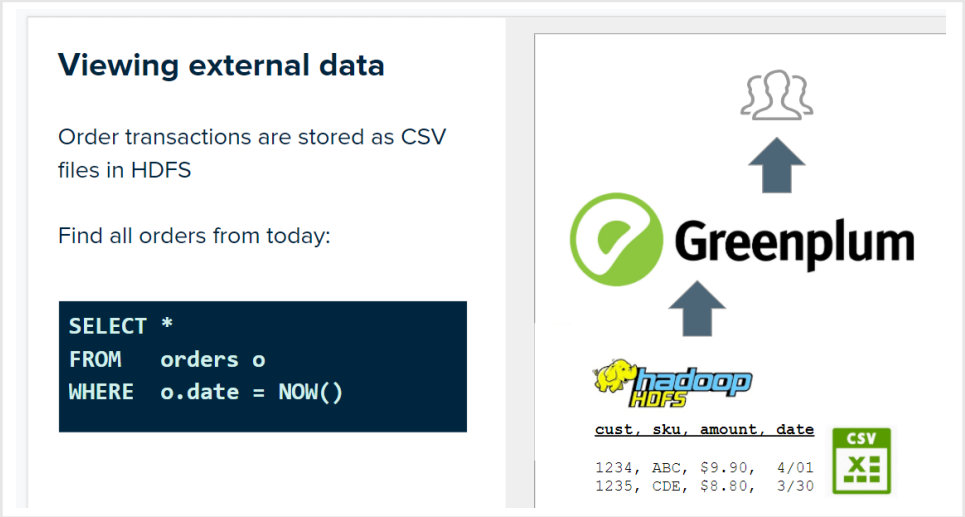


Figure 3: *Management of external tables in Greenplum.*

Gpfdist protocol

Greenplum uses the gpfdist protocol to read and write data with external tables. This protocol runs on top of the HTTPS protocol, and transfers data from external sources to and from each individual PostgreSQL segment in a Greenplum cluster. Many other modules within Greenplum use this protocol, including the Greenplum Spark Connector, the GemFire–Greenplum Connector, the Greenplum Streaming Server, and the aforementioned Platform Extension Framework (PXF). PXF has proven to be particularly useful technology, so let's review it in more detail.

Platform Extension Framework

With the explosion of data stores and cloud services, data now resides across many disparate systems and in a variety of formats.

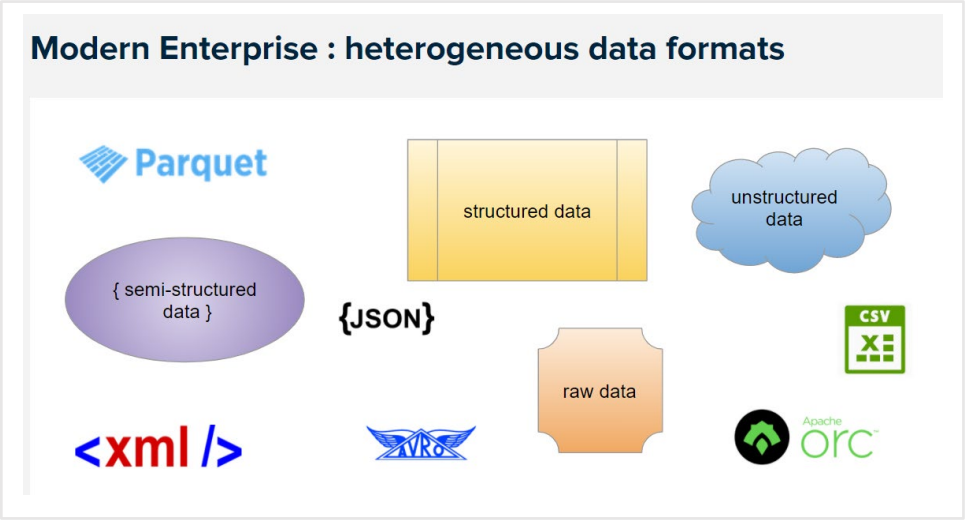


Figure 4: *Greenplum acknowledges that enterprise data exists in many places; PXF simplifies analysis of this data.*

When multiple data sets exist in external systems, it is often necessary to perform a lengthy ETL (extract, transform, load) operation to get data into the database. But what if we only needed a small subset of the data? What if we only want to query the data to answer a specific question or to create a specific visualization?

In this case, it's often more efficient to query data sets remotely and return only the results, rather than performing a full data load operation. The Greenplum Database Platform Extension Framework (PXF) helps accomplish this task. PXF is an open-source project that provides parallel, high-throughput data access and federated query processing across heterogeneous data sources. The framework uses built-in connectors that map a Greenplum external table definition to an external data source.

PXF's architecture enables users to efficiently query large data sets from multiple external sources, without requiring those data sets to be loaded into Greenplum. Using this framework, analysts can access Greenplum through the standard SQL interface. From there, users can instruct the system to perform complex processing on data sets that live externally in a wide range of storage devices and file formats, such as CSV, JSON, Parquet, ORC and Avro.

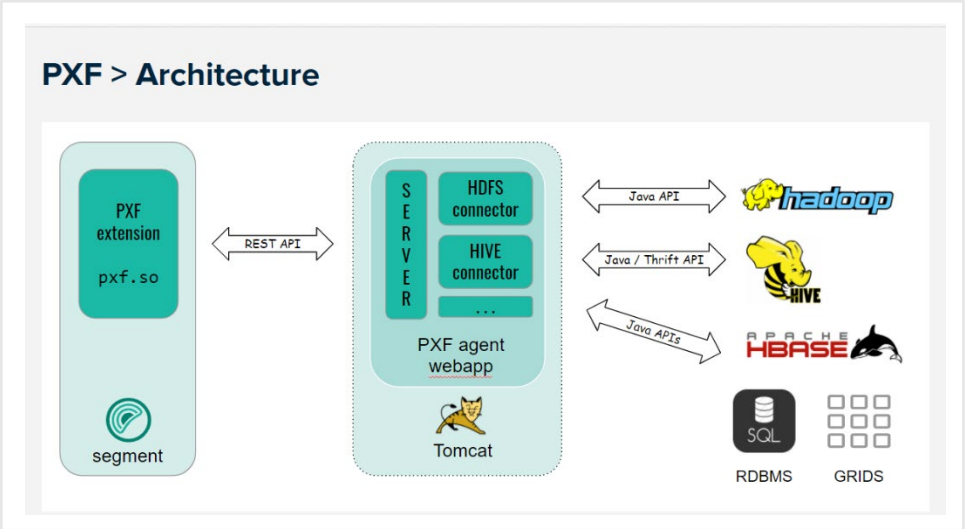


Figure 5: *The Platform Extension Framework allows external data sources to be quickly processed natively within Greenplum.*

Greenplum Streaming Server

Historically, many data warehouse and big data environments relied on scheduled batch ETL to ingest data into a data warehouse, data mart or data lake. But today, streaming data ingestion has become state of the art. Now, it's relatively simple to load data to your analytics stack in real time. Important attributes of a real-time streaming data ingestion platform include:

- Guaranteed once-only data delivery
- High-speed ingestion that can scale with increased data loads
- Error handling for malformed data
- Fault tolerance for the infrastructure and resumption ingest operation if a failure occurs
- Multiple structured and semi-structured input message formats
- Post-load triggers for data processing after ingestion.

The Greenplum Streaming Server provides best-in-class capabilities in this area by reading from Apache Kafka (or other real-time streaming system), and ingesting all data records atomically. In parallel, Greenplum records a database transaction reflecting that the messages have been successfully loaded. This approach helps you avoid data load duplication, and allows for a stateless streaming ingestion infrastructure that's easy to manage and scale.

The Greenplum Streaming Server leverages the gpfdist protocol for parallel, high-speed data loading to all distributed PostgreSQL worker segments in Greenplum.

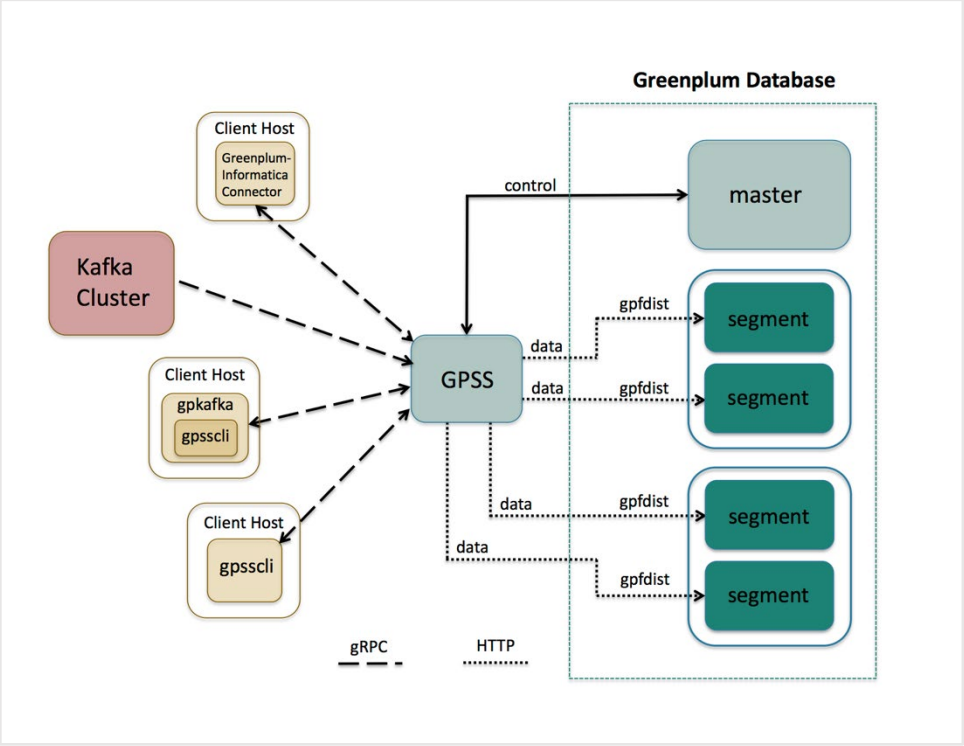


Figure 6: *Parallel data ingest into Greenplum.*

Elastic infrastructure for Greenplum

In years past, users deployed Greenplum atop bare-metal servers to maximize performance. Today, virtualization now offers superior performance for Greenplum deployments, in addition to the expected benefits of cost savings and availability.

Virtualization and the abstraction of resources allows for consolidated manageability of resources, and optimizes personnel time to achieve the same work. The end result: Lower total cost of ownership when compared to bare metal and a modern user administration experience that unlocks the agility required to meet dynamic business needs.

We will now introduce the all-in-one elastic infrastructure, featuring VMware vSphere, vSAN and Dell EMC Isilon. Together, these products create a modern solution to abstract and manage not only the infrastructure elements, but also provide a robust and cloud-ready approach for enterprise data needs.

Building an architecture that enables you to meet the dynamic needs of a modern enterprise is no simple task. While there is no silver bullet, there are ways to adopt and assimilate infrastructure to reduce friction, and establish and maintain centers of operational excellence. The architecture outlined in this paper is one example that streamlines scalability, provides better integration into data systems, and evolves your systems to be more cloud-ready.

In the architecture diagram below, we designed a four-node vSAN cluster connected to a pair of switches in an HA configuration to meet enterprise standards. Each server has four network interfaces, and these are split off into the VMware services network and the Greenplum network.

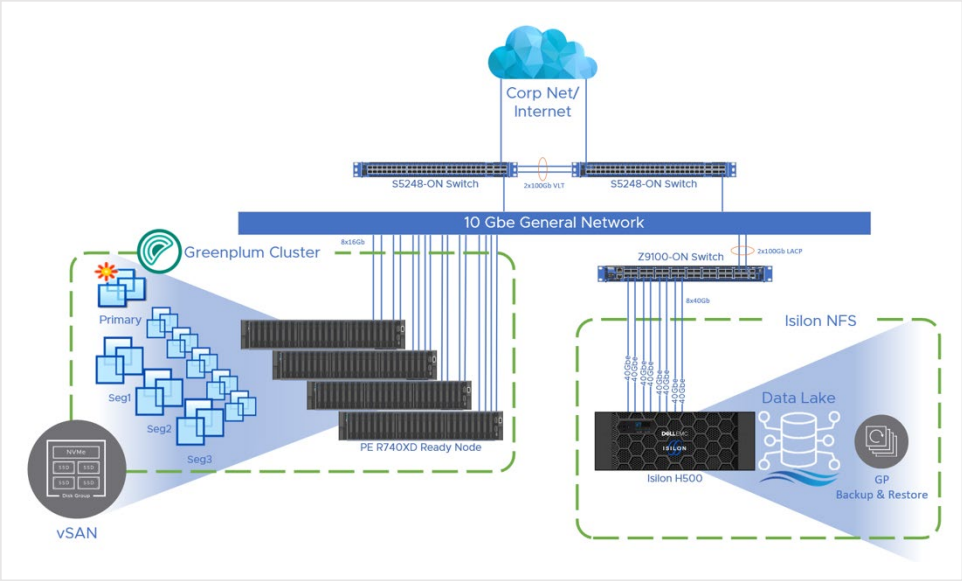


Figure 7: *A four-node vSAN cluster connected to a pair of switches in a high-availability configuration.*

The Greenplum VM cluster was implemented with one Greenplum coordinator VM (a standby coordinator should be added in a production environment) and eight Greenplum segment host VMs. The majority of resources of the ESXi™ hosts were divided between two segment host VMs running on each system because they will process the tasks. The VMs were created with parameters such that no oversubscription was done in the system during testing. Greenplum is a parallel processing platform, optimally making use of all available resources at once to generate results. Any contention for those resources affects the speed that results can be generated.

Enough resource headroom needed to be on each ESXi so that if any host in the cluster failed, the VMs on the failed host could be started up on other hosts within the system. With this in mind, each segment host VM was assigned 32 CPUs, 128GB RAM and the two separate networks. Greenplum Database can pass large amounts of traffic between the segment hosts in a system. This internode traffic was directed across the Greenplum-specific network while outside access to the hosts and vSAN and VMware services were allocated to a different network. Three disk devices were given to each host from the vSAN cluster, one for the OS drive, and two large mount points were exposed for most of the available space in the vSAN. With the following resource on each VM, we determined that running four primary segments and four mirror segments per segment host was the best match for the workload.

The following section outlines the architecture in more detail to help provide a clear understanding of how and why the infrastructure was set up this way.

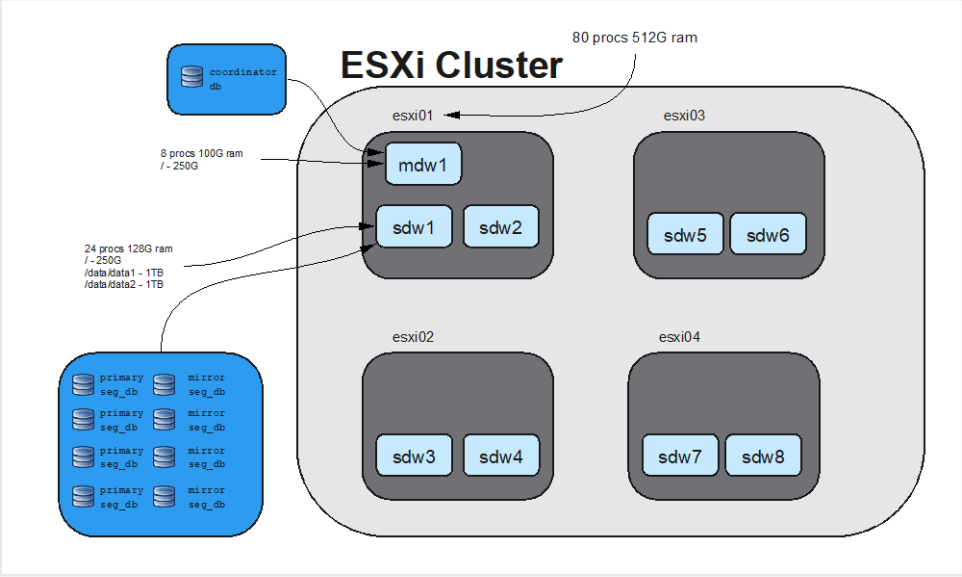


Figure 8: *ESXi cluster.*

Compute

Dell EMC PowerEdge R740xd vSAN Ready Node

The reference architecture of Tanzu Greenplum with VMware vSAN was tested using Dell EMC PowerEdge R740xd vSAN Ready Nodes in an NVMe configuration. The PowerEdge R740xd platform is a 2U dual socket server equipped with the latest Intel® Xeon® Scalable processors. It has the flexibility to support configurations such as 24x 2.5-inch NVMe drives, as well as support for up to 24x 2.5-inch traditional SSD flash drives. Dell EMC vSAN Ready Nodes are pre-configured, tested and certified to run VMware vSAN. Each Ready Node includes just the right amount of CPU, memory, network I/O controllers, HDDs and SSDs. Dell EMC also offers premiere vSAN Ready Node configurations; each model boasts an Identity Module that self-identifies the server as a vSAN Ready Node upon boot-up to streamline deployment, updates and more.

The PowerEdge R740xd Server is the ideal building block for a hyperconverged data warehouse solution to power workloads in the data center. The graphic below outlines the server specification used in this reference architecture.

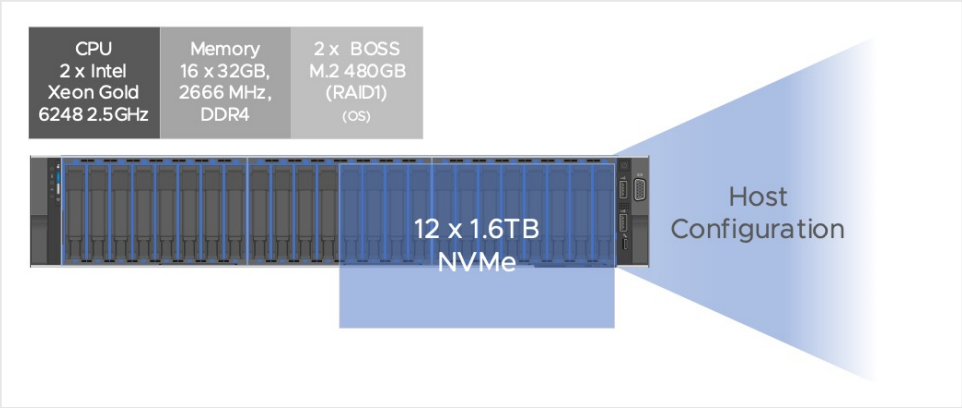


Figure 9: *Dell EMC PowerEdge R740xd vSAN Ready Node.*

Compute node networking

To achieve optimal network performance for workloads on Greenplum, we recommend separating the control plane (VMware services) from the Greenplum data plane. In this example, we use a 10Gbe control plane that adheres to vSAN best practices to ensure the proper throughput and resiliency required for production environments. For the Greenplum data plane (GP Net), we allocate an additional dedicated 10Gbe dual port card to improve the reliability of the data path and reduce congestion from other application services.

We geared the configuration for small to medium sized use cases (4 to 24 vSAN Ready Nodes). If a customer needs to scale from medium to large (24+ vSAN Ready Nodes), we suggest higher bandwidth networking (25/40/100Gbe) to ensure the proper throughput. The figure below outlines the vSAN Ready Node configuration and how the interfaces have been allocated.

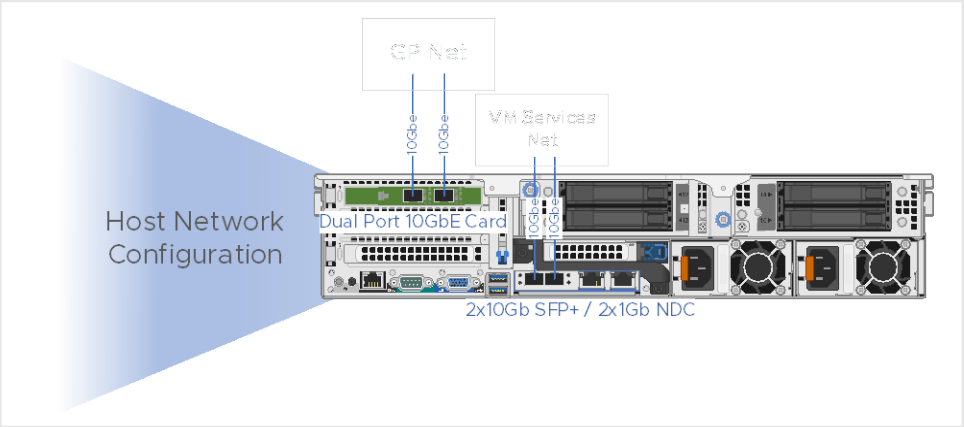


Figure 10: *The vSAN Ready Node configuration and interface allocation for this testing.*

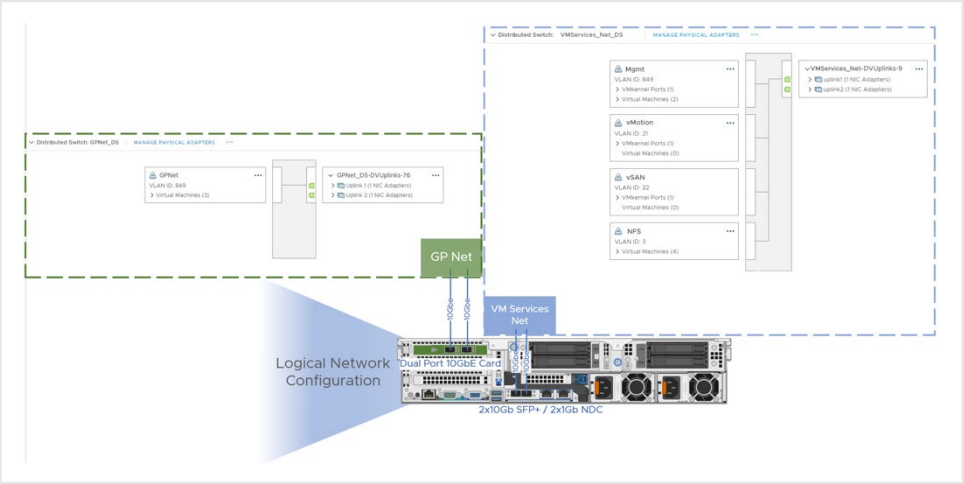
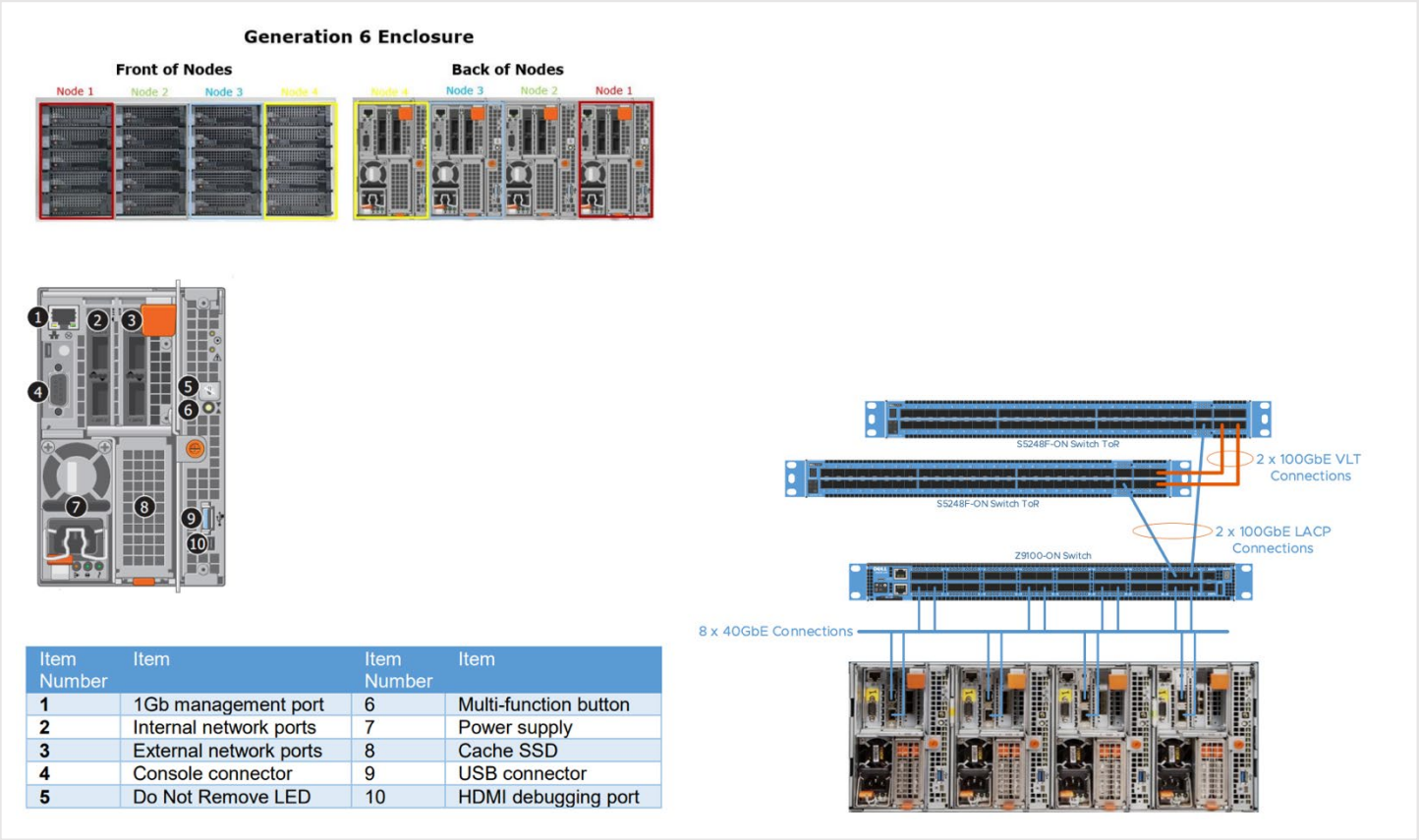


Figure 11: *VMware logical networking.*

Isilon architecture

The diagrams below outline the Dell EMC Isilon H500 hardware architecture and the network topology for the front-end (external) services that are exposed to the Greenplum vSAN cluster. The two top-of-rack switches (ToRs) on the right-side diagram are also the switches that the four Dell EMC vSAN Ready Nodes are connected to (front-end access) and are outlined in the Figure 9 architecture diagram shown in the section titled, "Compute." The back-end (internal) network ports were implemented using the standard Isilon best practices but are not shown here.



Item Number	Item	Item Number	Item
1	1Gb management port	6	Multi-function button
2	Internal network ports	7	Power supply
3	External network ports	8	Cache SSD
4	Console connector	9	USB connector
5	Do Not Remove LED	10	HDMI debugging port

Figure 12: *System components and architecture details.*

VMware vSAN

Overview

VMware vSAN is a widely deployed option for hyperconverged infrastructure (HCI), hence its inclusion here. In fact, vSAN has proven to be an excellent fit for all types of workloads. The success of vSAN can be attributed to many factors, including performance, flexibility, ease of use, robustness and pace of innovation.

Traditional infrastructure deployment is often mired by disjointed operations and maintenance workstreams. Practitioners must be proficient with various disaggregated tools and experts in specialized skills. The hyperconverged approach of vSphere and vSAN solves this inefficiency with familiar tools to deploy, operate and manage private-cloud infrastructure. VMware vSAN provides best-in-class enterprise storage and provides the elasticity that modern enterprises demand.

VMware vSAN enables customers to prime their business for growth through seamless evolution, leading flexibility and hybrid cloud capabilities. vSAN helps customers seamlessly evolve, as it is integrated into vSphere and requires no new tools. vSAN's industry-leading ecosystem empowers customers to run HCI on certified solutions with their preferred vendor, and hybrid cloud capabilities provide customers consistent operations from edge to core to cloud, with intrinsic security throughout.

VMware HCI, powered by vSAN, is the cornerstone for modern data centers whether they are on the premises or in the cloud. vSAN runs on standard x86 servers and provides several Ready Node configurations from Dell Technologies that provide customers peace of mind and confidence so they can focus on their business outcomes.

Scaling made easy

One of our customers' biggest hurdles is scaling a solution to meet business needs. vSAN tackles this problem head on with "Cluster Quickstart," a guided cluster creation wizard.

With Cluster Quickstart, vSAN deployment and setup are easy. This step-by-step configuration wizard guides users through the creation of a production-ready vSAN cluster. The module then handles the initial deployment and the process of expanding the cluster as needs change.

To enable vSAN, simply click the "Configure" option in the Configure tab of the vSphere cluster. This will start the process.

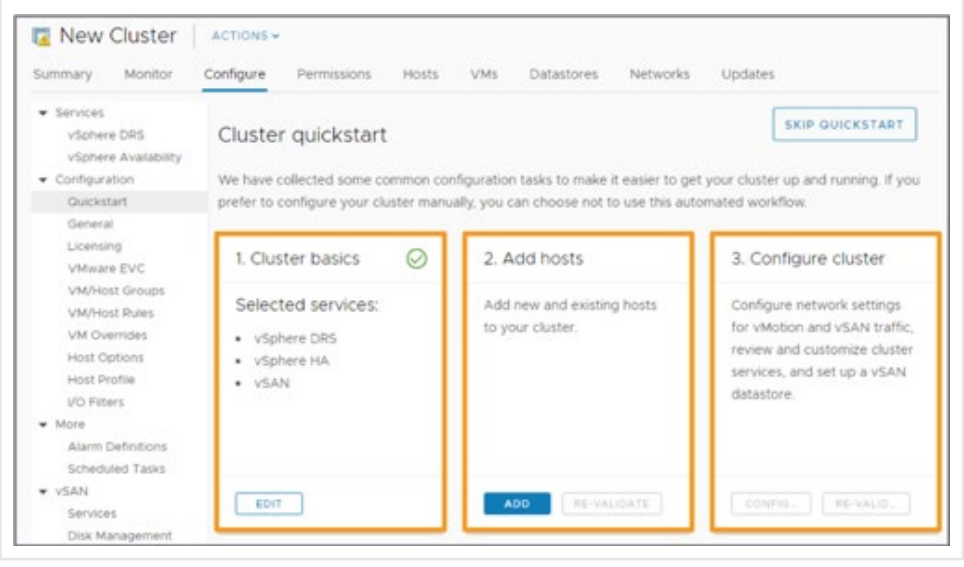


Figure 13: *Easily create and scale clusters with Cluster Quickstart.*

The Cluster Quickstart wizard workflow includes three areas to ease the deployment process:

1. **Cluster basics** — Select services like vSphere DRS, vSphere HA and vSAN.
2. **Add hosts** — Add multiple hosts simultaneously.
3. **Configure cluster** — Configure cluster, network and vSAN data store settings.

The ability to scale your vSAN ecosystem is easier than ever. This simple approach provides customers the flexibility and automation to ensure proper configuration.

It also helps address challenges that a non-virtualized Greenplum architecture might struggle with. For example, one such challenge is the ability to easily add more capacity. Expanding your Greenplum environment onto new bare-metal nodes typically requires weeks of planning, and can require a maintenance window to maintain data integrity. With vSAN, this scaling effort is as simple as adding more HCI nodes to your cluster, and spinning up new Greenplum segments.

This simplified scale-up allows Greenplum to provide a more elastic approach to growing your data warehouse infrastructure and enables you to deliver a highly resilient and secure solution that can grow with your business needs.

Advantages of vSAN storage policies + Greenplum HA capabilities

Storage Policy Based Management (SPBM) from VMware enables precise control of storage services. Like other storage solutions, vSAN provides services such as availability levels, capacity consumption and data placement techniques to optimize performance. A storage policy contains one or more rules that define service levels.

Storage policies are primarily created and managed through vCenter Server®. Policies can be assigned to VMs and individual objects such as a virtual disk. Storage policies are easily changed or reassigned if application requirements change. These modifications are performed with no downtime and without the need to migrate VMs from one data store to another. SPBM makes it possible to assign and modify service levels with precision on a per-VM basis.

In our testing, we used a storage policy that was set to FTT=1 (failures to tolerate) RAID 5 erasure coding (minimum four hosts). This allows for the use of erasure coding to enable the same level of data protection as mirroring (RAID 1), while using less storage capacity (minimum of four hosts required). If performance is your focus, then a storage policy of FTT=1 RAID 1 mirroring (minimum three hosts) will provide the best solution and is the default setting for vSAN. The diagram below demonstrates how the data is striped across your vSAN cluster nodes in an FTT=1 RAID 5 configuration.

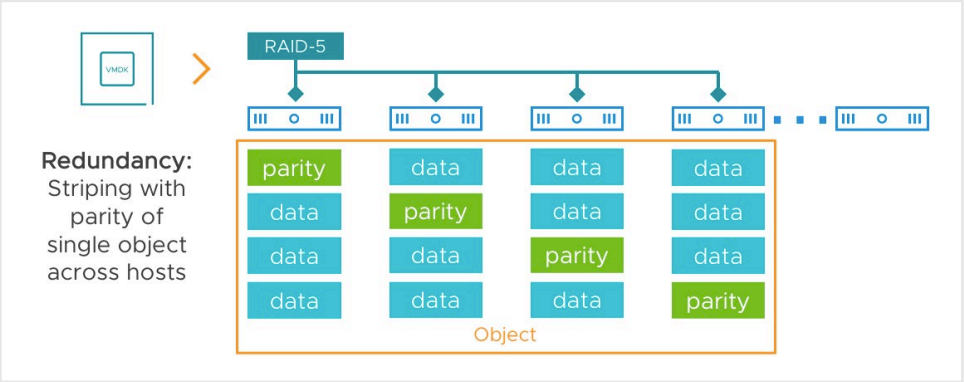


Figure 14: *A look at how the data was striped across vSAN cluster nodes in the Greenplum reference architecture.*

So how does this work with Greenplum HA features? Balancing performance, space optimization and data resiliency may vary. Often, the driving factor is capacity, not performance. Coupling the right mix of vSAN storage policies with Greenplum's replication strategy provides a robust physical and logical layer to achieve the best data integrity and resilience. Greenplum Database supports highly available, fault-tolerant database services when you enable and properly configure Greenplum high availability features. To guarantee a required level of service, each component must have a standby host ready to take its place if it should fail. With the Greenplum Database "shared-nothing" MPP architecture, the primary host and segment hosts each have their own dedicated memory and disk storage, and each master or segment instance has its own independent data directory.

The Greenplum Database primary instance is the client's single point of access to the system. The primary instance stores the global system catalog, the set of system tables that store metadata about the database instance, but no user data. If an unmirrored primary instance fails or becomes inaccessible, the Greenplum instance is effectively offline, since the entry point to the system has been lost. For this reason, a standby primary host must be ready to take over if the primary host fails as shown in the figure below. Next, we will discuss some best practices for implementing HA, and how the mechanism works with segment hosts and maintains coordination with the primary host.

Primary host HA view:

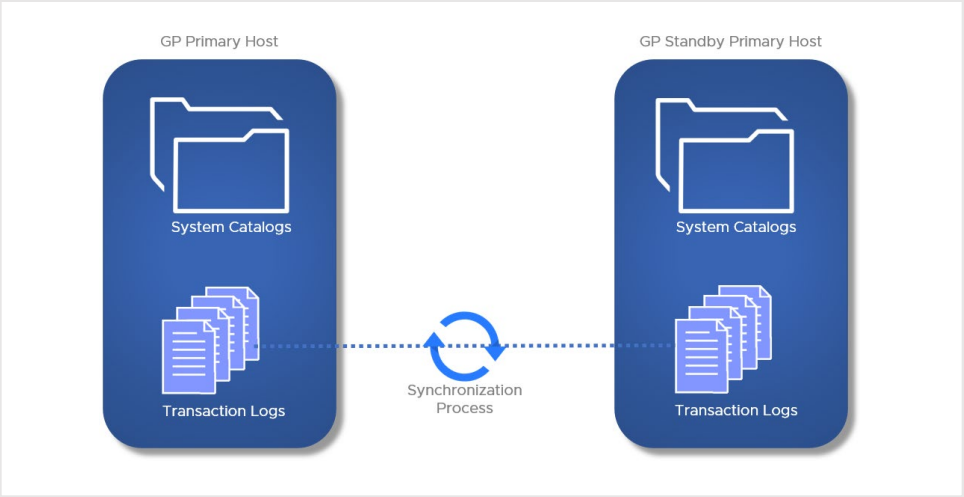


Figure 15: *Primary host high-availability view.*

Primary host best practices

- Set up a standby primary instance (mirror) to take over if the primary host fails.
- The standby can be on the same host or on a different host, but it is best practice to place it on a different host from the primary master to protect against host failure.
- Plan how to switch clients to the new master instance when a failure occurs, for example, by updating the master address in DNS.
- Set up monitoring to send notifications in a system monitoring application or by email when the primary fails.

Greenplum Database segment instances each store and manage a portion of the database data, with coordination from the primary instance. If any unmirrored segment fails, the database may have to be shut down and recovered, and transactions occurring after the most recent backup could be lost. Mirroring segments is, therefore, an essential element of an HA solution.

A segment mirror is a hot standby for a primary segment. Greenplum Database detects when a segment is unavailable and automatically activates the mirror. During normal operation, when the primary segment instance is active, data is replicated from the primary to the mirror in two ways:

- The transaction commit log is replicated from the primary to the mirror before the transaction is committed.
- Second, segment mirroring uses physical file replication to update heap tables. Greenplum Server stores table data on disk as fixed-size blocks packed with tuples.

When the acting primary is unable to access its mirror, replication stops and state of the primary changes to “Change Tracking.” The primary saves changes that have not been replicated to the mirror in a system table to be replicated to the mirror when it is back online.

The master automatically detects segment failures and activates the mirror. Transactions in progress at the time of failure are restarted using the new primary. Depending on how mirrors are deployed on the hosts, the database system may be unbalanced until the original primary segment is recovered. The diagram below demonstrates the segment host primary and mirrored layout with four segment hosts and is followed by some best practices to keep in mind.

Segment host view:

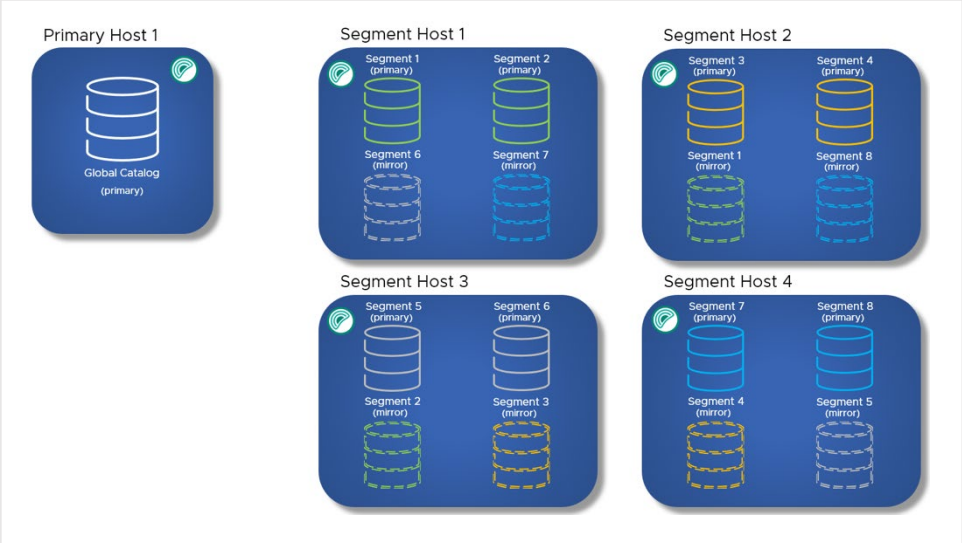


Figure 16: *Segment host view.*

Segment hosts best practices

- Set up mirrors for all segments.
- Locate primary segments and their mirrors on different hosts to protect against host failure.
- Mirrors can be on a separate set of hosts or co-located on hosts with primary segments.
- Set up monitoring to send notifications in a system monitoring application or by email when a primary segment fails.
- Recover failed segments promptly, using the **gprecoverseg** utility, to restore redundancy and return the system to optimal balance.

So, as you can see Greenplum HA has many advantages and when coupled with VMware vSAN storage policies, the two provide a potent one-two punch for protecting customer data and ensuring availability if and when there are unforeseen issues. Below are the key highlights of Greenplum HA:

- Two copies of each segment data
- Automatic mirroring and failover when a failure occurs
- Proven in production for over a decade of runtime

Dell EMC Isilon data lake technology

When thinking about analytics and artificial intelligence (AI), the first ingredient that should be top of mind is data, particularly unstructured data. (Without data, there is no analytics.)

The status quo for most IT organizations is highly siloed, a phenomenon we discussed earlier. There are four fundamental issues with this siloed approach:

- **Higher cost:** Redundant replication of data across silos increases hardware budgets and the day-to-day administrative burden.
- **Data inconsistency:** Data is replicated and transformed based on various application needs, resulting in inconsistency between different system reports.
- **Security concerns:** Securing and tracking access across multiple silos is difficult. Enterprises have a broad attack surface for bad actors, further complicated by various different requirements, access rights, user counts and physical security.
- **Maintenance concerns:** Basic administration of data infrastructure is tedious, often involving maintaining numerous copies of the same data.

Modern analytics applications require terabytes, even petabytes, of data. Replicating — and securing — this volume of data is hard to imagine, let alone achieve. The pragmatic solution is a sound data strategy to service all analytics and AI requirements. A useful solution needs to provide four basic pillars:

- Scalability
- Performance
- Flexibility
- Enterprise readiness

PowerScale storage

The Dell Technologies modern analytics stack based on Isilon/PowerScale is the basis of a sound data strategy.

PowerScale helps you unlock the structure within your data, and to address the challenges with unstructured data management. PowerScale is the next evolution of the OneFS — the operating system powering the industry's leading scale-out NAS platform.

Key characteristics of PowerScale include:

- The PowerScale family includes Isilon nodes, PowerScale nodes, and PowerScale OneFS running across all of them.
- The software-defined architecture of OneFS gives you simplicity at scale, intelligent insights and the ability to have any data anywhere it needs to be. Whether it is hosting file shares or home directories or delivering high-performance data access for applications like analytics, video rendering and life sciences, PowerScale can seamlessly scale performance, capacity and efficiency to handle any unstructured data workload.
- PowerScale brings a new version of OneFS to our Isilon nodes as well as two new all-flash PowerScale nodes that deliver application requirements like S3 protocol and performance needs like NVMe, from the edge to the cloud.

- The new PowerScale all-flash platforms coexist seamlessly in the same cluster with your existing Isilon nodes to drive your traditional and modern applications.

A large portion of your data is unstructured data, and that data set is growing exponentially — not just in the data center but at the enterprise edge and in the cloud. PowerScale OneFS powered scale-out storage solutions are designed for organizations that want to manage their data, not their storage. Our storage systems are powerful yet simple to install, manage and scale to virtually any size. The storage includes a choice of PowerScale all-flash nodes along with Isilon all-flash, hybrid or archive nodes to meet the most demanding business needs. And, unlike traditional enterprise storage, these solutions stay simple no matter how much storage capacity is added, how much performance is required, or how business needs change in the future.

Data strategy considerations

Developing a data-first mindset

In a world where unstructured data is growing rapidly and taking over the data center, organizations are looking for ways to get more out of their data. Whether it is driving innovation, getting to market faster or creating differentiation, they want the data to start creating value. Instead of thinking of destinations for your data, you think about what the data is going to be used for, who will be using it, and how the data will help you solve for business needs. When you have a data-first mindset, the goal is to get any data to where it needs to be for business needs. Whether it is all-flash edge offerings or the cloud to take advantage of the tools and access, data must be located where it needs to be for the business.

Leveraging the OneFS operating system

With OneFS-powered clusters consisting of PowerScale or Isilon nodes, you can eliminate storage silos, consolidate all your unstructured data, store petabytes of file data, and analyze them in a data-first world. With up to 252 nodes in a cluster, you can scale both capacity and performance in a few minutes to meet your specific business needs with no additional IT burden. With the performance of all-flash nodes that are configured with NVMe, you can drive demanding workloads like AI, ML and DL. The OneFS operating system powers scale-out storage solutions.

OneFS provides the intelligence behind the highly scalable, high performance modular storage solution that can grow with your business. With support for all-flash and NVMe, OneFS can help you accelerate processes and workflows while scaling easily to handle massive growth and providing the highest levels of data protection. This is all provided in a storage solution designed for unmatched ease of use.

Orchestrated by OneFS, all components in a cluster work to create a unified pool of highly efficient storage — with a storage utilization rate of up to 80%.

Deduplicating data
With SmartDedupe data deduplication, you can further reduce your data storage requirements by up to 35%. The F810, F200 and F600 all-flash platforms and the H5600 hybrid platform deliver improved data reduction with features like inline compression and deduplication to dramatically increase the effective storage capacity and density of your storage solution. The unmatched efficiency of the storage systems means that less physical storage and space are required to house the same amount of data — reducing both CapEx and OpEx.

Adding storage nodes
With the OneFS AutoBalance function, you can quickly and easily add nodes without downtime, manual data migration or application logic reconfiguration, saving precious IT resources. Because the storage is so easy to manage, it requires fewer IT resources for storage administration than traditional storage systems, which further reduces overall operating costs.

Eliminating storage silos
You can streamline your storage infrastructure by consolidating large-scale unstructured data assets thus eliminating silos of storage. OneFS-powered solutions include integrated support for a wide range of industry-standard protocols, including internet protocols IPv4 and IPv6, NFS, SMB, S3, HTTP, FTP and HDFS. As a result, you can simplify workflows, accelerate business analytics projects, support cloud initiatives and get more value from your enterprise applications and data. With new support for the high performance, multi-protocol S3, data can simultaneously read and write through any protocol, and there is no longer a need to migrate and copy data from a secondary source to run modern cloud-enabled applications.

Protecting data
Massive stores of data present unique management challenges including disaster recovery, quota management and off-site replication. OneFS data protection and management software provides you with powerful tools that help you protect your data assets, control costs and optimize the storage resources and system performance of your big data environment.

Use cases where integrated stack provides key capabilities

The power and capability of this reference architecture comes from the complementary nature of the parts that compose the full system. In this section, we cover several use cases that demonstrate the complementary benefits of the stack components.

- Isilon as a landing zone for incoming data to the platform
- Federated query processing of data in Isilon via Greenplum engine
- Partitioning of Greenplum data for tiered storage
- Unloading of specific data dets from Greenplum to Isilon for future usage
- Spark processing of data living in Greenplum
- Spark processing of data living in Isilon
- Spark ingestion/ETL with data loading into Greenplum and Isilon
- Backup and restore of Greenplum data to/from Isilon

Isilon as a landing zone for incoming data to the platform

When data first enters an analytical environment (such as a data lake, data mart or data warehouse), it first must be stored temporarily on a large, scalable file system. In fact, this raw data is often stored in its raw format for several weeks in case the need to re-process the ingested data arises. This use case is called a “data landing zone,” where the data first lands before processing.

In this reference architecture, the Isilon platform serves as a data landing zone. After landing on Isilon, further processing can be done for ingestion formatting and cleansing into a Greenplum Database. Or administrators can opt to perform processing in-memory in a tool like Spark. From there, data can then be stored back to a new location in Isilon, or in Greenplum for downstream processing.

Federated query processing of data in Isilon via Greenplum engine

Data stored in the Isilon storage layer can be exposed via HDFS, NFS or S3 protocol to the Greenplum query and analytical engine. This way, the data does not need to be ingested into the Greenplum and vSAN storage. Storing data in Isilon could be cheaper than storing it in the Greenplum layer, and also would make the data accessible in raw file formats to both Greenplum and other systems such as Spark for analytics and reporting. The PXF federated query engine scans the data on Isilon dynamically as warranted by user queries.

Partitioning of Greenplum data for tiered storage

When Isilon is paired with Greenplum, a tiered storage approach is often desirable. Here, “hot” data is stored internally in storage that’s managed and optimized by Greenplum. “Warm” data is stored in open file formats on Isilon for query access or third-party application access. This approach lowers the long-term cost of storage, but keeps the data accessible for queries.

Let’s consider an example. A financial institution stores billions of historical financial transactions. The organization could opt to store a rolling 180 days of data live in the Greenplum engine. The next 15 years of data would be stored on Isilon. Queries accessing the last 180 days of financial transactions would be served from the local Greenplum optimized storage. Queries accessing older dates using the SQL WHERE clause would be automatically served from the Isilon storage.

Another example could be historical sales data. Greenplum could divide data into row-based storage and compressed columnar storage based on date. Older data could be compressed more, while the oldest data is still stored in Isilon storage. Again, all the data can be seamlessly queried by the user without knowledge of the data storage model.

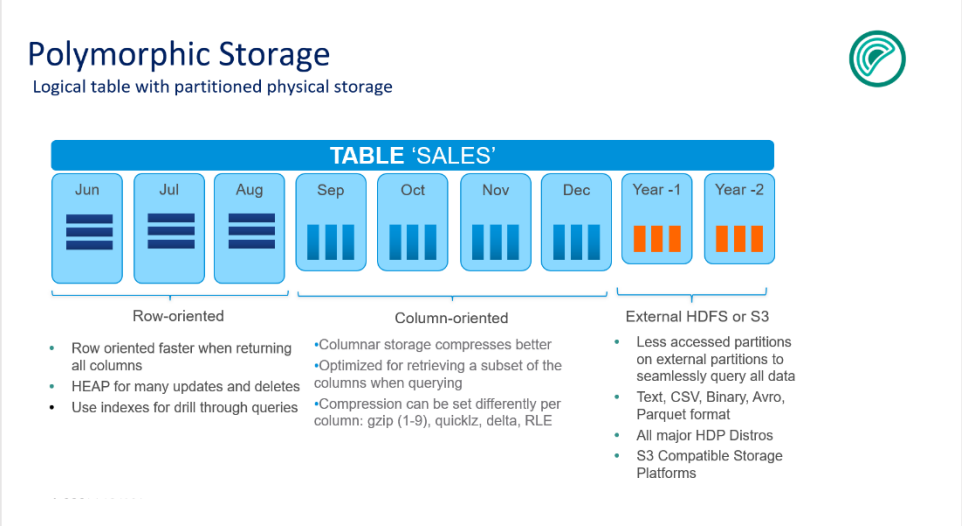


Figure 17: *A closer look at optimized storage performance tiers in the reference architecture.*

Unloading of specific data sets from Greenplum to Isilon for future usage

The Greenplum Database Platform Extension Framework (PXF) has the ability write data to external storage as part of a SQL INSERT command. A typical workflow would be to generate a result data in Greenplum and then do an SQL INSERT command to push the result set to external storage for usage by other systems or future reloading into Greenplum. Greenplum has the ability to export data in standard open file formats to ensure third-party tools like Apache Spark could also process the resulting data sets living on the Isilon system.

Apache Spark plus Greenplum Database

Greenplum has many built-in analytical libraries, beyond ANSI SQL and Apache MADlib. Users often find it useful and convenient to do data science work in Apache Spark. In fact, the usage of Greenplum and Apache Spark works in a complementary fashion. Greenplum is fundamentally a relational database management system, and Apache Spark is an in-memory compute framework. Therefore, customers often combine these compute frameworks to solve their analytical business needs.

Greenplum is often used for high-speed computation of well-known, published models on structured data living in the database. Apache Spark would be used for a wide variety of ML models on data that may not be in a fully structured format. New models and variations could be more easily coded in Apache Spark and combined with downstream processing. Together, both tools give a full solution for any ML computation.

Apache Spark processing of data living in Greenplum

Greenplum has a rich set of libraries, including MADlib, that provide a full suite of ML and DL algorithms that can be leveraged to process data in the Greenplum engine. In addition to Greenplum, users often find it useful and convenient to do some of their data science work in Apache Spark.

In fact, the usage of Greenplum and Apache Spark are not mutually exclusive but work in a complementary fashion. Although both systems have generic open ended compute functionality and specific ML libraries, the use pattern and compute models of Greenplum and Apache Spark are different. Greenplum is fundamentally a relational database management system and Apache Spark is an in-memory compute framework. Therefore, customers often combine these models to solve their analytical business needs.

Spark processing of data living in Isilon

Spark is a fundamental tool for the data science practitioner and was leveraged to analyze and process both the hot data living in Greenplum, as well as the data lake and warm data living in Isilon. All data sets from both Greenplum and Isilon can be seamlessly accessed via Apache Spark from stateless VMs on the connected network. In practice, it’s quite simple to differentiate from the analytics to Greenplum and Isilon; only the connection string to the data source needs to be modified.

Backup and restore of Greenplum data to/from Isilon storage

The Greenplum Backup Manager can directly back up data from the Greenplum Database to the Isilon storage platform. NFS or S3 mount points can be used for the backup job to store backups directly to Isilon. As noted earlier, a Greenplum cluster consists of individual PostgreSQL databases. The Greenplum Backup Manager will initiate a consist snapshot of all data. Administrators can also opt to snapshot a subset of data directly from each PostgreSQL segment to the Isilon storage. The set of files in a backup set will be tracked by the Greenplum Backup Manager. When restores are initiated, the data will be retrieved from the storage system into each PostgreSQL segment in parallel. The parallel access of each PostgreSQL segment running on multiple VMs provides for scale-out bandwidth for backup and restore operations.

Leveraging the Isilon storage system for operational and backup data provides consolidation around the storage infrastructure. Administrators have fewer systems to support, and dedicated infrastructure isn't required for backups. Also, capacity can be accumulated together into a large pool for backups and operational data for easier expansion and storage space management.

All-in-one analytics solution in action

To quantify the benefits of this architecture, we configured the Greenplum and Isilon cluster in the Dell Technologies Customer Solution Center and performed a series of tests. The setup stored 20% of data locally on Greenplum and the remaining 80% on Isilon. This ratio reflects a best practice in the real world; data growth is never a challenge, as the compute and storage are independent. Customers enjoy the flexibility to expand compute and storage based on the demands of the system. Another important aspect of the test was to measure the performance of run queries on the external Isilon data storage. To keep the testing challenging, the queries leveraged advanced analytics and ML functions. In this section, we'll review the lab setup and the results achieved.

Data model

The first step to build the environment was to get good quality data. We used the Telecom DPI data, which captures subscriber data usage and usage patterns. The total data size was about 4TB with 14 billion rows from DPI web data and 400 million rows from DPI streaming. 80% of the fact and atomic data was stored on Isilon as external partitions. Dimension tables and recent partitions were stored locally on Greenplum within the vSAN. In all, 3.2TB of data was stored on Isilon and 800GB of data was stored on Greenplum locally. Data stored on Isilon was accessed via gpfdist external table.

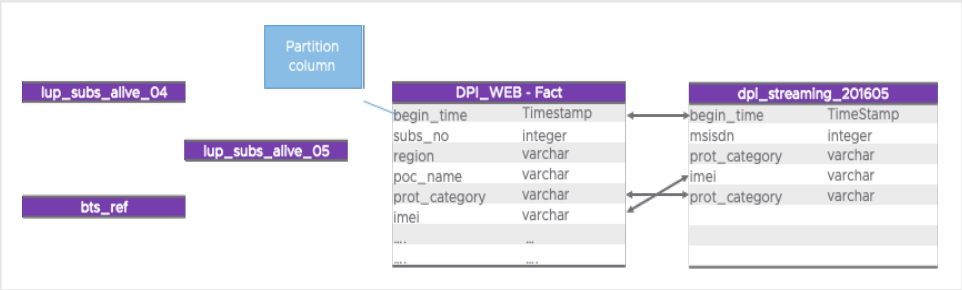


Figure 18: *Data model.*

Tiered storage

As mentioned, the architecture supports multi-temperature data storage via Greenplum partitions. These can either be vertical or polymorphic partitions. With vertical partitions, large fact tables are divided into time ranges for efficient data access and retention policies. Polymorphic partitions involve different ranges in partitioned tables that use different storage parameters and mediums.

For this environment, out of 12 partitions, 10 vertical partitions were stored on Isilon using gpfdist external tables. There were no changes in the data model or the queries.

```
CREATE writable external TABLE public.dpi_web_ext_write (like dpi_web
)
LOCATION (
    'gpfdist://sdw1:8086/dpi_web_Jul2016_ext1.txt',
    'gpfdist://sdw2:8086/dpi_web_Jul2016_ext2.txt',
    'gpfdist://sdw3:8086/dpi_web_Jul2016_ext3.txt',
    'gpfdist://sdw4:8086/dpi_web_Jul2016_ext4.txt',
    'gpfdist://sdw5:8086/dpi_web_Jul2016_ext5.txt',
    'gpfdist://sdw6:8086/dpi_web_Jul2016_ext6.txt',
    'gpfdist://sdw7:8086/dpi_web_Jul2016_ext7.txt',
    'gpfdist://sdw8:8086/dpi_web_Jul2016_ext8.txt'
)
FORMAT 'TEXT' ( DELIMITER '|' NULL '')
DISTRIBUTED by (subs_no)
;
CREATE EXTERNAL TABLE
Time: 96.026 ms
insert into dpi_web_ext_write select * from dpi_web_1_prt_5;
INSERT 0 1404590968
Time: 4387563.339 ms
```

Figure 19: *Example.*

Our test users — data scientists — accessed all the data via Greenplum. All the complexity of joining data on Greenplum and Isilon was seamlessly managed by Greenplum. Over 350GB of user data was loaded from Isilon into Greenplum in 11 minutes. Greenplum external writable tables were used to move the data from Greenplum internal table to Isilon. About 350GB of data was loaded in 1 hour and 20 minutes from Greenplum to Isilon.

```
insert into dpi_web_col select * from ext_r_dpi_web;

NOTICE:  found 355 data formatting errors (355 or more
input rows), rejected related input data
INSERT 0 1404590968
Time: 664312.150 ms
```

Figure 20: *Example.*

SQL query workload

Performance was measured for data loads and data queries. DPI flat files were ingested into Greenplum using the GPLoad utility. The duration of each load and query was captured on both Greenplum local storage and storage on Isilon. The results are noted below.

Type of the query	Complexity	Time duration - Isilon	Time duration – Local storage
Data loading from flat files to Greenplum via GPload	1.4 billion rows	11.04 mins	1.2 mins
Calculate Data Upload + Download by Subscriber number, month and Data Category	Group by operation on 14 billion rows	21.56 mins	6 to 8 mins
Calculate duration in seconds of streaming data by joining base tables dpi_web and dpi_streaming.	Join on 14 billions rows with 400 million rows	3.6 mins	1 min

Figure 21: *Results.*

Data loading and querying on local storage was faster than expected. However, the same activity on Isilon (which was an external storage to Greenplum) also had very promising results. The queries included aggregations and table joins.

To make the testing more challenging, the data scientist team developed an analytical use case of “Subscriber Behavior Modeling” using Greenplum MADlib libraries. As mentioned, MADlib is the powerful open-source library of scalable in-database algorithms for ML. It provides data-parallel implementations of ML, mathematical, statistical and graph methods on Greenplum.

MADlib uses the full compute power of Greenplum’s massively parallel Postgres architecture to process very large data sets. (Other options are often limited by the amount of data that can be loaded into memory on a single node.) MADlib algorithms are invoked from a familiar SQL interface, so they are easy to use.

Below are the published numbers after executing the ML queries:

Type of the query	Complexity	Time duration
Calculate Data Upload + Download by Subscriber number, month and Data Category (IM,Social_Networking, Web_Browsing, Streaming, Miscellaneous)	Group by operation on 14 billion rows	21.56 Minutes
Calculate duration in seconds of streaming data by joining base tables dpi_web and dpi_streaming.	Join on 14 billions rows with 400 million rows	3.6 Minutes
Standardize the metrics for subscriber clustering	Multi-level aggrates and statisitcal functions	28 Seconds
Cluster subscriber for month of March based on data usage	Machine learning - Clustering operation using K-means.	12 second
Consolidate Subscriber cluster id, data usage, location details for month of march	Join on 14 billions rows with 171 million rows and 600K.	4 Minutes
Consolidate Subscriber cluster id, data usage, location details for month of April	Join on 14 billions rows with 171 million rows and 600K.	4 Minutes
Subscriber Behaviour (Data Usage) Modelling Queries Total Time took.	Subscriber behaviour modeling	68.16 Minutes
Ad-hoc Cube Processing Query generates 8 Group By Combination operations	Cube processing on 14 billions with 8 group by operations and statistical functions.	190 Minutes
Ad-hoc Cube Processing query generating 4 Group By Combination operations and computing Statistical functions	Cube processing on 14 billions with 4 group by operations and statistical functions.	75 minutes

Figure 22: *Published numbers after executing ML queries.*

Accessing the lake directly

Once the data is available on Isilon, it can be directly accessible by Greenplum using the external tables operation. Data in different formats (structured, semi-structured, unstructured) can be stored on the lake. Other than Greenplum accessing the data, this data can also be accessed and organized directly using the Isilon interface. This gives added flexibility to manage the data at the storage level.

Data protection backups

Greenplum Database supports parallel and non-parallel methods for backing up and restoring databases. Parallel operations scale regardless of the number of segments in your system. The Greenplum Database parallel dump utility "gpcrondump" backs up the Greenplum primary instance and each active segment instance at the same time.

Greenplum backups can be done in three ways:

1. Locally on Greenplum cluster
2. On external servers/backup solutions using NFS mounts
3. On object storage using S3 protocol

In our tests, we chose option B. Backups were pointed to Isilon. A full backup achieved a throughput of 5.6TB/hour in 38 minutes.

Backup	Result
Started at	02:21 PM
Backup Size	3.5 TB
Completed at	02:59 PM
Total Backup Time	38 Minutes

Figure 23: *Backups*

Conclusion

Together the small set of components in the architecture described in this report, consisting of Greenplum Database, VMware vSAN, VMware vSphere, Dell EMC Isilon, Apache Spark and optionally Apache Kafka provide almost an entire data ecosystem. This architecture consists of all components that are scalable from small to nearly unlimited data sizes and come with enterprise management capability to simplify the daily required operational duties and lower the total cost of ownership. This all-in-one solution, when compared to sprawling sets of software components across hundreds of instances of servers or VMs, can be described as simple and effective at reasonable cost.

Technical resources

VMware documentation

- VMware vSAN: https://www.vmware.com/products/vsan.html
- VMware vRealize Suite: https://www.vmware.com/products/vrealize-suite.html
- Greenplum on VxRail Reference Implementation: https://core.vmware.com/resource/running-greenplum-vmware-cloud-foundation-dell-emc-vxrail
- VMware Tanzu Greenplum documentation: http://gpdb.docs.pivotal.io/
- Greenplum open source project homepage: https://greenplum.org/
- Greenplum 101 tutorials: https://greenplum.org/greenplum-101/

Dell Technologies documentation

- Test this solution in Dell Technologies Customer Solution Centers
- Learn more at https://tanzu.vmware.com/greenplum-reference-architecture
- Find out why Dell Technologies uses Greenplum
- Dell EMC Solutions for VMware
- Dell Technologies InfoHub, Data Analytics Solutions
- Dell EMC Storage for Analytics
- delltechnologies.com/referencearchitectures

DELL Technologies

vmware®