Technical White Paper

Dell EMC Technologies Powered by Dremio for Lightning-Fast Data Analytics

TPC-DS performance test results of Dremio with Dell EMC PowerEdge servers as compute cluster, PowerScale through HDFS protocol and ECS through S3 protocol as storage cluster.

Abstract

This paper describes the performance test results of running the Dremio cloud data lake query engine with Dell EMC™ PowerEdge R640 servers, PowerScale™, a scale-out NAS and ECS object storage for lightning-fast data processing.

October 2020

H18538

Revisions

Date	Description
October 2020	Initial release

Acknowledgements

Authors: Kirankumar Bhusanurmath, Analytics Solutions Architect, Dell EMC
Scott Gay, Senior Solutions Architect, Dremio
Serge Leontiev, Principal Technical Product Marketing Manager, Dremio

About Dremio Corporation

Dremio, the cloud data lake query engine, delivers ultra-fast query speed and a self-service semantic layer operating directly against data lake storage. Dremio eliminates the need to copy and move data to proprietary data warehouses or create cubes, aggregation tables, or BI extracts, providing flexibility and control for data architects, and self-service for data consumers. For more information, go to www.dremio.com.

Founded in 2015, Dremio is headquartered in Santa Clara, CA. Investors include Cisco Investments, Lightspeed Venture Partners, Norwest Venture Partners and Redpoint Ventures. Connect with Dremio on GitHub, LinkedIn, Twitter, and Facebook.

Table of contents

Executive summary

Dremio, the cloud data lake query engine, delivers lightning-fast query speed and a self-service semantic layer operating directly against data lake storage. It does this in a way that maintains the flexibility inherent in data lake storage; you don't have to move your data to a third-party solution or put it into a proprietary format. Dremio provides:

1. Lightning-fast queries, directly on data lake storage. Dremio technologies like data reflections, columnar cloud cache (C3), and predictive pipelining work alongside Apache Arrow to execute queries on data lake storage at interactive speed.
2. A self-service semantic layer. This abstraction layer enables line of business owners and IT to apply security and business meaning, while enabling analysts and data scientists to explore data and derive new virtual datasets.
3. Flexibility and openness. Dremio enables organizations to avoid vendor lock-in, query across clouds, and keep data in storage that they control.
4. Significant infrastructure cost savings. Dremio combines query acceleration and highly elastic compute resources to provide significant performance and cost benefits. For example, a 5x average increase in speed results in a 75% or more infrastructure cost reduction vs. Presto at the same level of performance.

With customer's most trusted data residing in Dell EMC PowerScale and ECS, unlock the full potential and maximum efficiency of Dell EMC PowerScale and ECS for business intelligence, analytics, and data science workloads with the Dremio data lake query engine. Together, Dell Technologies and Dremio offer a solution enabling customers to directly query data residing on Dell EMC PowerScale and ECS directly, eliminating the need for ETL, cubes, extracts, or any data movement away from Dell EMC PowerScale and ECS. The solution is easy to deploy, highly scalable, and provides significant performance and cost benefits to customers. A self-service semantic layer makes it easy for non-technical users to access and analyze data. With Dremio, users can easily migrate from legacy HDFS data warehouses and legacy HDFS database systems to Dell EMC PowerScale and ECS.

As you read through this paper, you will see that running Dremio with Dell EMC PowerScale F600 scale-out NAS and Dell EMC ECS EX3000 object stores provides excellent performance and significantly improved storage utilization with fewer hard drives and a smaller storage footprint.

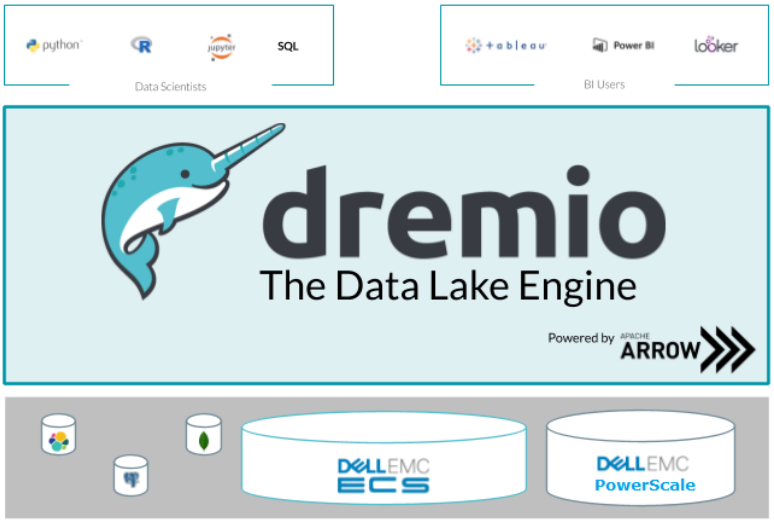


Figure 1 Dremio and Dell technologies solution architecture

1 Dremio

Dremio, the cloud data lake query engine, enables business analysts and data scientists to explore and analyze any data at any time, regardless of its location, size or structure, using their favorite tools such as Tableau, Power BI, Python, and R. Dremio leverages Apache Arrow, a patented data acceleration capability called Data Reflections, and highly-optimized native push-downs to provide interactive-speed query performance on any data volume, while enabling IT, data scientists, and business analysts to seamlessly shape the data according to the needs of the business.

1.1 Dremio Architecture

Dremio features a scale-out architecture. Dremio is designed to scale from one server to thousands of servers in a single cluster. Common deployment patterns include:

- Dedicated cloud infrastructure, such as Amazon EC2 instances, using AWS S3 for Dremio's data reflection store.
- In Docker containers, provisioned and managed via Kubernetes.
- On Hadoop, provisioned and managed as a YARN application, using HDFS for Dremio's data reflection store.

When using Dremio to analyze data in Hadoop, it is recommended to deploy Dremio directly on the Hadoop cluster. This enables Dremio to achieve data locality for the raw data as well as the data reflection store.

1.1.1 Instance node type

There are two distinct node-types in a Dremio instance, and the service's property determines whether the node is enabled with the master-coordinator or engine role. Each node type can be scaled independently. The two node types are:

• **Coordinators**. These nodes are responsible for coordinating query planning, managing metadata, serving Dremio's UI, and handling client connections. Client applications, such as BI or data science tools, connect to and communicate with coordinators. Coordinators are highly available, and can be scaled up to process more concurrent clients. Note that all Dremio instance nodes with Dremio coordinator services must have the master-coordinator role enabled. A node with only the coordinator role enabled is not supported.

• **Executors**. These nodes are responsible for query execution. Client applications do not connect to executors. Executors can be scaled up to process larger data volumes and more concurrent queries. Because Executors are stateless, deployments can treat these nodes as elastic resources and scale the system dynamically.

1.1.2 Acceleration

Dremio utilizes highly-optimized physical representations of source data called data reflections. A reflection store can live on HDFS, MapR-FS, cloud storage such as S3 or S3-compatible, or direct-attached storage (DAS). The reflection store size can exceed that of physical memory. This architecture enables Dremio to accelerate more data at lower cost, resulting in a much higher cache hit ratio compared to traditional memory-only architectures. Dremio data reflections are automatically utilized by the cost-based optimizer at query time.

When running Dremio on Hadoop, coordinators should be deployed on edge nodes so that external applications such as BI tools can connect to them. Furthermore, there is no need to manually deploy Dremio on the cluster because the coordinators can use YARN to provision the Executor nodes. To maximize

performance in a production deployment, it is recommended to have an Executor on every Hadoop node in the cluster. The following diagram outlines a typical on-Hadoop deployment:

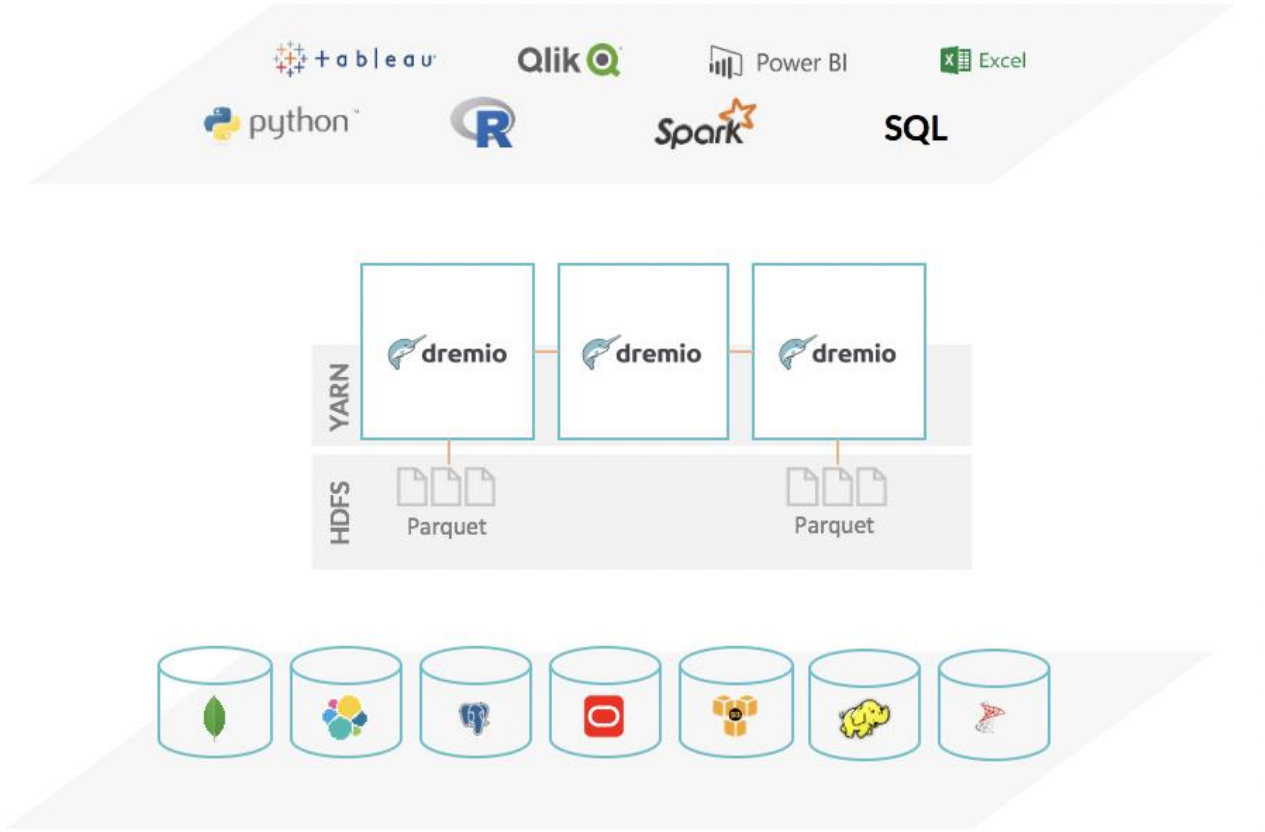


Figure 2 Dremio Architecture

2 Dell EMC PowerEdge R640 server

The PowerEdge R640 is a general-purpose platform expandable up to 7.68TB of memory, up to twelve 2.5-inch drives, and flexible I/O options. The R640 can handle demanding workloads such as virtualization, dense private cloud, High Performance Computing (HPC) and software-defined storage.

The Dell EMC PowerEdge R640 is the ideal dual-socket, 1U platform for dense scale-out data center computing. The R640 combines density, performance, and scalability to optimize application performance and data center density.

The PowerEdge R640 features:

1. 2nd Generation Intel® Xeon® Scalable Processor product family (with up to 28 cores and two threads per core)
2. Up to six DDR4 memory channels with two DIMMs per channel per CPU and 24 DIMMs (supports DDR4 RDIMM/LRDIMM/NVDIMM-N/DCPMM)
3. PCI Express® (PCIe) 3.0 enabled expansion slots (with up to 48 lanes per CPU)
4. Networking technologies, such as Ethernet, InfiniBand, OCP, OPA

Figure 3 Dell EMC PowerEdge R640 server

3 Dell EMC PowerScale

PowerScale is the next evolution of OneFS –the operating system powering the industry's leading scale-out NAS platform. The PowerScale family includes Dell EMC PowerScale platforms and Dell EMC Isilon platforms configured with the PowerScale OneFS operating system. OneFS provides the intelligence behind the highly scalable, high–performance modular storage solution that can grow with your business. A OneFS-powered cluster is composed of a flexible choice of storage platforms including all-flash, hybrid and archive nodes. These solutions provide the efficiency, flexibility, scalability, security, and protection for you to store massive amounts of unstructured data within a cluster. The new PowerScale all-flash platforms co-exist seamlessly in the same cluster with existing Isilon nodes to drive traditional and modern applications.

New PowerScale all-flash storage platforms - powered by the OneFS operating system - provide a powerful yet simple scale-out storage architecture to speed up access to massive amounts of unstructured data while dramatically reducing cost and complexity. They deliver extreme performance and efficiency for the most demanding unstructured data applications and workloads. Powered by the new OneFS 9.0 operating system, the all-flash platforms are available in four product lines:

3.1 PowerScale F200

Provides the performance of flash storage in a cost-effective form factor to address the needs of a wide variety of workloads. Each node allows you to scale raw storage capacity from 3.84 TB to 15.36 TB per node and up to 3.8 PB of raw capacity per cluster. The F200 includes in-line compression and deduplication. The minimum number of PowerScale nodes per cluster is three while the maximum cluster size is 252 nodes. The F200 is best suited for remote offices, small M&E workloads, small hospitals, retail outlets, IoT, factory floor and other similar deployment scenarios.

Figure 4 PowerScaleF200

3.2 PowerScale F600

Results captured in this white paper are from PowerScale F600. With new NVMe drives, the F600 provides larger capacity with massive performance in a cost-effective compact form factor to power the most demanding workloads. Each node can scale raw storage capacity from 15.36 TB to 61.4 TB per node and up to 15.48 PB of raw storage per cluster. The F600 includes inline software data compression and deduplication. The minimum number of nodes per cluster is three while the maximum cluster size is 252 nodes. The F600 is best suited for M&E studios, hospitals and financials that need performance and capacity for demanding workloads.

Figure 5 PowerScale F600

3.3 Isilon F800 and Isilon F810

Isilon F800 provides massive performance and capacity. It delivers up to 250,000 IOPS and up to 15 GB/s aggregate throughput in a single chassis configuration and up to 15.75M IOPS and up to 945 GB/s of aggregate throughput in a 252-node cluster. Each chassis houses 60 SSDs with a capacity choice of 1.6 TB, 3.2 TB, 3.84 TB, 7.68 TB or 15.36 TB per drive. This allows you to scale raw storage capacity from 1 to 96 TB to 924 TB in a single 4U chassis and up to 58 PB raw storage in a single cluster.

Isilon F810 provides massive performance and capacity along with inline data compression and deduplication capabilities to deliver extreme efficiency. The F810 delivers up to 250,000 IOPS and up to 15 GB/sec aggregate throughput in a single chassis configuration and up to 15.75M IOPS and up to 945 GB/s of aggregate throughput in a 252-node cluster. Each F810 chassis houses 60 SSDs with a capacity choice of 3.84 TB, 7.68 TB or 15.36 TB per drive. This allows you to scale raw storage capacity from 230 TB to 924 TB in a 4U chassis and up to 58 PB of raw storage in a single cluster.

Figure 6 Isilon F800 and F810

4 Dell EMC ECS

Dell EMC ECS is a software-defined, cloud-scale, object storage platform. With ECS, any organization can deliver scalable public cloud services with the reliability and control of a private-cloud infrastructure. ECS provides comprehensive protocol support for unstructured—object and file—workloads on a single modern storage platform. Using ECS, organizations can easily manage globally distributed storage infrastructure under a single global namespace with anywhere access to content. ECS features a flexible software-defined architecture that is layered to promote limitless scalability. Each layer is completely abstracted and independently scalable with high availability and no single points of failure. ECS also comes in a fully integrated turnkey appliance that bundles software and Dell PowerEdge servers into an easily deployed object system. ECS is currently in its third generation of hardware appliances, the EX-Series, building on the legacy of Dell EMC's Centera and Atmos object storage platforms which predate ECS. The ECS EX-Series is comprised of three unique hardware products: the EX300, EX500 and EX3000.

4.1 ECS EX300

As a starter edition, the EX300 lowers object storage adoption entry barriers with 60TB starting cluster options. With the capacity to grow to exabyte scale, this is the ideal sandbox for inhouse, cloud-native, mobile and web application storage. It's also the optimal system to modernize existing Centera or Atmos deployments.

Figure 7 ECS EX300

4.2 ECS EX500

The perfect blend of economy and density, the EX500 injects even greater flexibility into the ECS appliance portfolio. With rack capacity ranging from 480TB to 6.1PB, the EX500 is a versatile option for midsized enterprises looking to support either modern application or deep archive use cases.

Figure 8 ECS EX500

4.3 ECS EX3000

Results captured in this white paper are from EX 3000. EX 3000 is a high density, hot disk-swappable, object storage system, the EX3000 packs up to 11.5PB per rack and can grow into exabyte-scale with ease. It's an ideal platform for long-term retention, storage consolidation and multi-purpose object storage requirements that span S3, HDFS and archive workloads.

Figure 9 ECS EX3000

5 Solution Overview

5.1 Overview of PowerScale for Big Data Analytics

The Dell EMC PowerScale scale-out platform combines modular hardware with unified software to provide the storage foundation for data analysis. PowerScale scale-out NAS is a fully-distributed system that consists of nodes of modular hardware arranged in a cluster. The distributed PowerScale OneFS operating system combines the memory, I/O, CPUs, and disks of the nodes into a cohesive storage unit to present a global namespace as a single file system.

The nodes work together as peers in a shared-nothing hardware architecture with no single point of failure. Each node adds capacity, performance, and resiliency to the cluster, and each node acts as a NameNode and DataNode. The NameNode daemon is a distributed process that runs on all the nodes in the cluster. A compute client can connect to any node in the cluster to access NameNode services.

As nodes are added, the file system expands dynamically and redistributes data, eliminating the work of partitioning disks and creating volumes. The result is a highly efficient and resilient storage architecture that brings all the advantages of an enterprise scale-out NAS system to storing data for analysis.

An Isilon cluster optimizes data protection. OneFS more efficiently and reliably protects data than HDFS. The HDFS file system, by default, replicates a block of data three times. In contrast, OneFS stripes the data across the cluster and protects the data with forward error correction codes, which consumes less space than replication with better protection.

An Isilon cluster also includes enterprise features to back up data and to provide high availability. For example, in managing your DataNode data, a best practice with a traditional Hadoop system is to back up data to another system—an operation that must be performed with brute force by using a tool like DistCP. OneFS includes support for NDMP backups, cluster synchronization, geo-replication, snapshots, file system journal, virtual hot spare, antivirus, Integrity Scan, dynamic sector repair, and accelerated drive rebuilds.

The enterprise features of OneFS ease data management. OneFS includes storage pools, deduplication, automated tiering, quotas, high-performing SSDs, capacity-optimized HDDs, and cluster monitoring and forecasting with InsightIQ.

SmartPools, for example, provides tiered storage so that you can store current data in a high-performance storage pool while storing older data in a lower, more cost-effective pool in case you need to analyze it again later.

For security, OneFS can authenticate HDFS connections with Kerberos. SmartLock can protect sensitive data from malicious, accidental, or premature alteration or deletion to help comply with SEC 17a-4 regulations.

5.1.1 Dell EMC Isilon and Dremio solution architecture

Below is the solution architecture used in this performance testing.

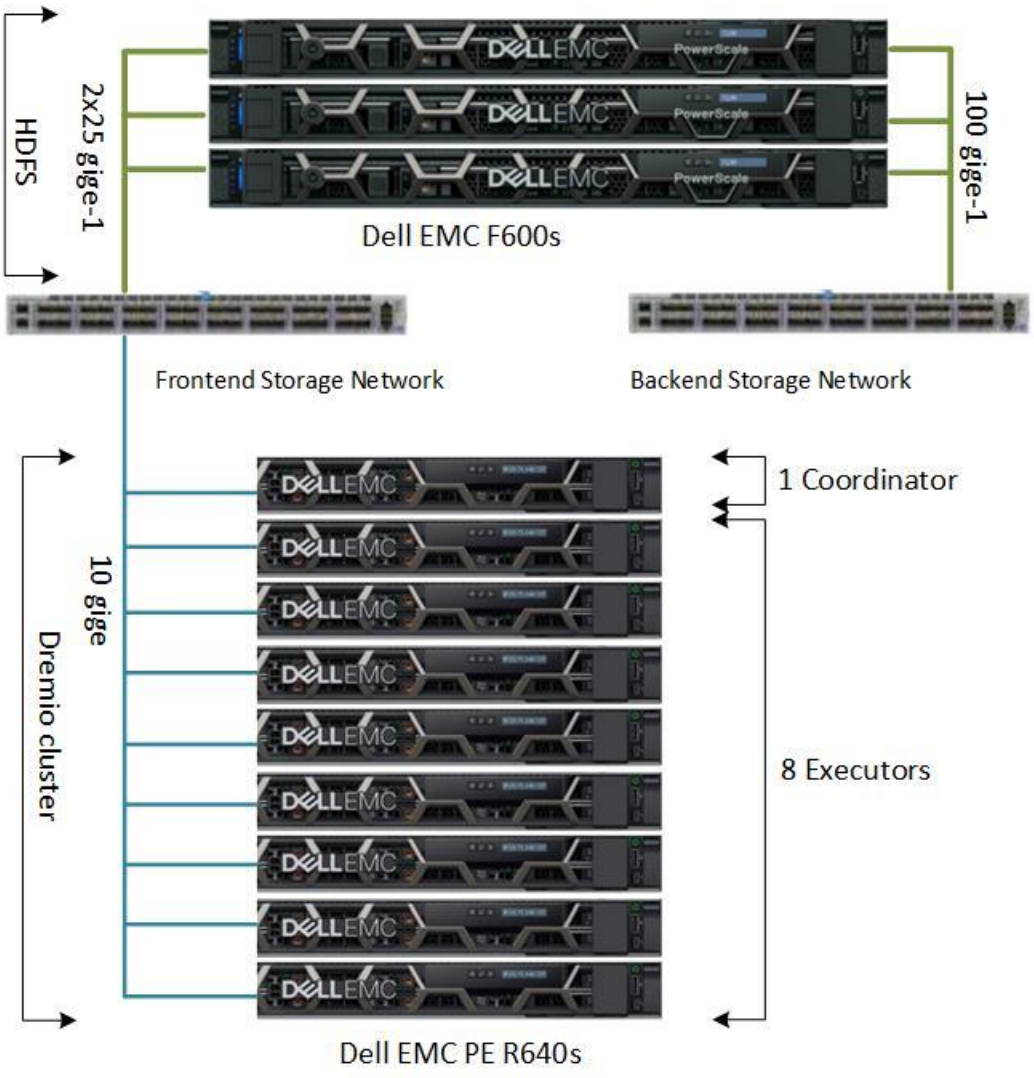


Figure 10 Dell EMC Isilon F600 and Dremio cluster

5.2 Overview of ECS for Big Data Analytics

As data volumes continue to explode, finding that “needle in the haystack” becomes increasingly more difficult. Business users continue to demand new ways to drive value from their data – which in turn drives increased demands on IT organizations to meet critical SLA’s.

With Dell EMC ECS, large-petabyte and exabyte size data lakes often hold critical data for many lines of business-like market data for capital market quant research, fraud data for payment fraud risk management, and analytic workloads. Users are becoming more sophisticated in the types of questions they want/need to ask from this data. Building indexes, aggregations, and views only scales so far so the ability to do “what-if” analysis becomes next to impossible. So, what is the answer?

Dell Technologies has partnered with Dremio to enable our customers to drive value from structured, semi-structured and unstructured data. This myriad of data can now be leveraged on-premise without a forklift to another environment.

5.2.1 Dell EMC ECS and Dremio solution architecture

Below is the solution architecture used in this performance testing.

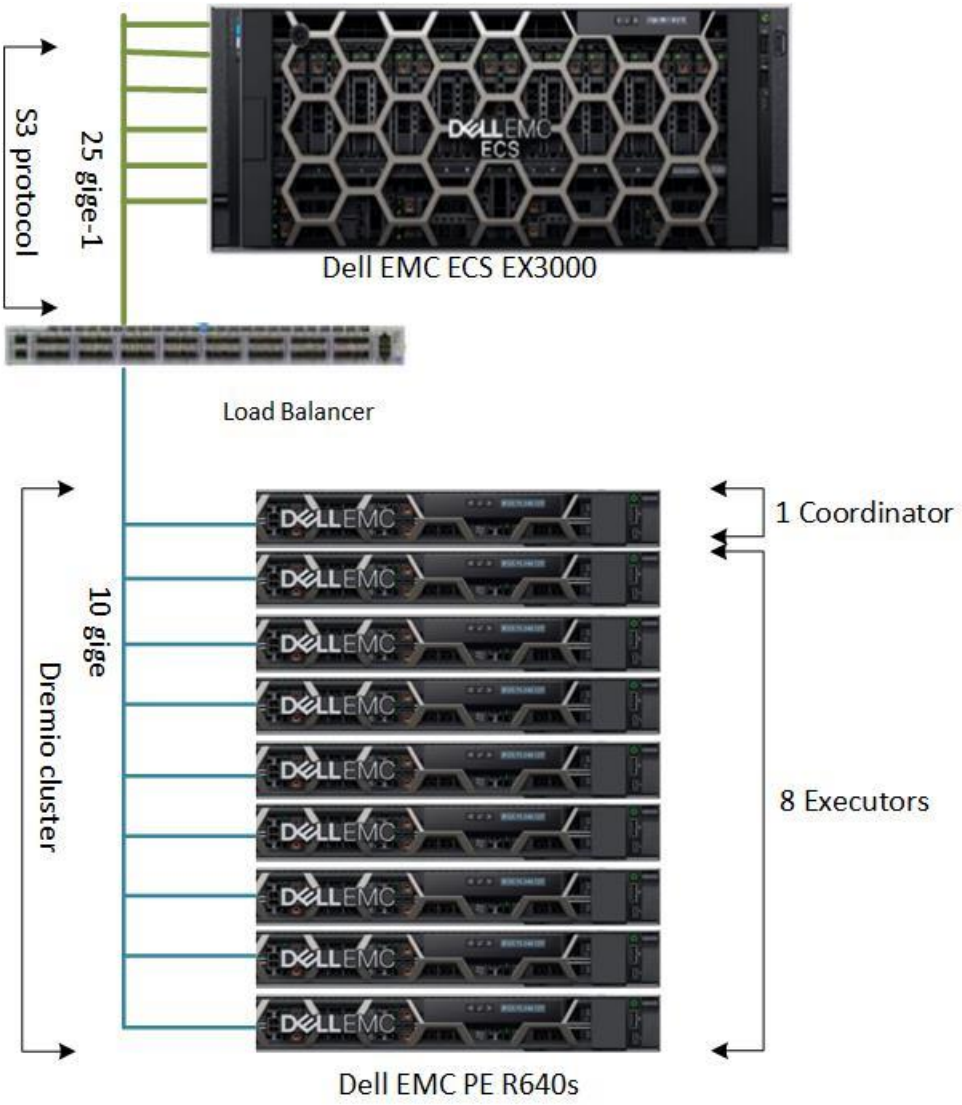


Figure 11 Dell EMC ECS and Dremio cluster

6 Dell EMC and Dremio: tested configuration

6.1 Dremio cluster

Dremio on-prem is installed and configured as a standalone cluster with only (1) coordinator and (8) executor nodes.

The following diagram illustrates a simple nine-node deployment architecture.

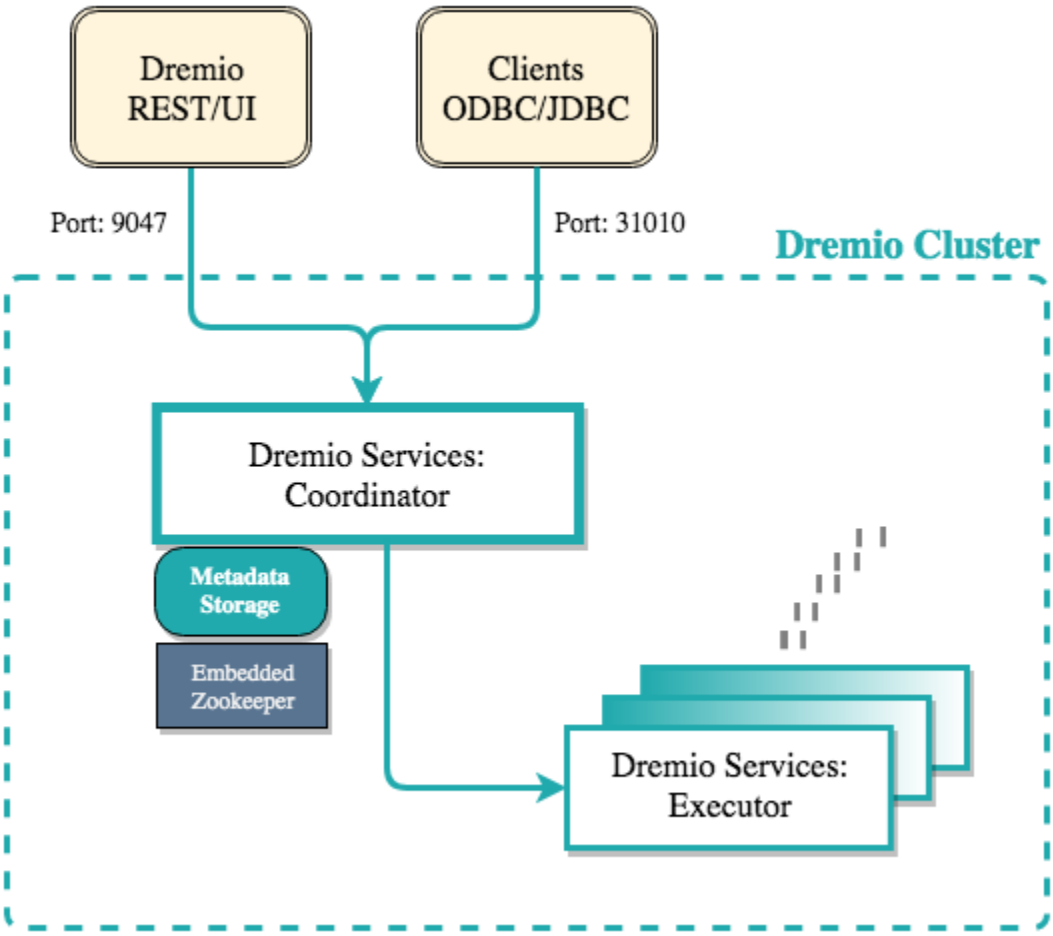


Figure 12 Standalone Dremio cluster

Coordinator node: A single node with the Dremio service configured with the master-coordinator role.
Executor nodes: Eight (8) nodes with the Dremio service configured with the executor role.
Metadata storage: Local on the coordinator node. Default: no configuration required.
Zookeeper: Embedded - it is on the coordinator node. Default: no configuration required.

6.2 Compute nodes

All the compute nodes are identical Dell EMC PowerEdge R640 servers with 18 cores, 256G RAM, and 10G NIC running CentOS Linux release 7.6.1810 (Core). Total 2x 1.6 TB NVMe drives.

6.3 Dell EMC PowerScale F600

PowerScale	3 x F600 (OneFS 9.0)
Per Node	32 Core CPU
	128 GB RAM
	15 x 2 TB SSD
	2x25 GbE Ethernet front-end

6.4 Dell EMC ECS EX3000

ECS	6 x EX3000D (ECS Object Version 3.4.0)
Per Node	32 Core CPU
	64 GB RAM
	30 x 12 TB HDD
	1 x 25GbE Ethernet front-end

6.5 TPC-DS test suite

The industry-standard TPC-DS test suite was used to benchmark Dremio performance. The TPC-DS benchmark data was modeled on the decision support functions of a retail product supplier. TPC-DS consists of seven fact tables and 17 dimensions.

TPC-DS consists of 99 queries which are divided into four broad classes:

1. Reporting queries
2. Ad-hoc queries
3. Iterative OLAP queries
4. Data Mining queries

The data set size of 10TB in Apache Parquet file format is used to test the scale-out factor. First, a 10TB text file format data set is generated by using TPC-DS dsdgen utility, and then all the tables are converted to Apache Parquet file format by the Dremio cloud data lake query engine using CTAS (CREATE TABLE AS SELECT) command. Data is placed on PowerScale and ECS storage and connected as a data source to the Dremio instance. The Apache JMeter test suite is used to execute benchmarking tests directly against the Dremio instance through JDBC protocol.

Note: Dremio data reflections are NOT enabled, and all data is pulled directly from PowerScale and ECS during the benchmarking.

There are 99 queries in the TPC-DS benchmark that include a variety of BI/reporting and ad-hoc queries representing typical analytical workloads. The subset of 58 queries used in this benchmark includes queries that can be successfully executed on Dremio without any modifications to query syntax. While a 100% query success rate is achievable with logical query rewrites, the benchmarking effort focused on the 58 unmodified queries because of the primary objective to provide easily-reproducible results.

7 Dell EMC technologies and Dremio performance

In this section, we will review the results captured for TPD-DS benchmarking on Dremio for the data stored on Dell EMC PowerScale F600 and Dell EMC ECS EX3000.

Below are the steps followed to benchmark and capture the results.

1. 10TB data set is generated using TPC-DS dsdgen utility.
2. Dremio is connected to this data source as an NFS mount and verified.
3. Dremio is connected to PowerScale through HDFS protocol and ECS through S3A protocol.
4. Using Dremio CTAS (Create Table as Select) to convert 10TB of generated data from flat files to Parquet file format. Tables are created on the PowerScale and ECS locations with Dremio via HDFS protocol and S3A connections, respectively.
5. A total of 58 TPC-DS queries were run successfully; only those are used in this benchmarking and same are run sequentially three times to warm the C3 feature of Dremio.
6. Apache JMeter application is used to sequentially execute the queries and capture the performance results.
7. The test is conducted with (4) Dremio executors and scaled to (8) Dremio executors.

Note: The captured results of TPC-DS benchmarking are for the default configuration Dremio cluster without any reflections enabled. All of the 10TB data set resides on the Dell EMC PowerScale F600 with Cloud Cache enabled on the Dremio compute nodes.

7.1 Dell EMC PowerScale and Dremio performance

The bar chart below shows the total execution time of 58 queries run sequentially for a different number of Dremio executor nodes; the shorter the bar, the better. As expected, the execution time linearly reduces with an increase in the number of Dremio executor nodes. At this scale, Dremio enjoys a 25% performance boost with increased number of nodes, linearly scaling and delivering faster execution time. For a data set size of 10TB in open source parquet file format, this demonstrates excellent execution time for faster time-to-insight and interactive query performance at this scale and the query performance is improvement at larger node counts.

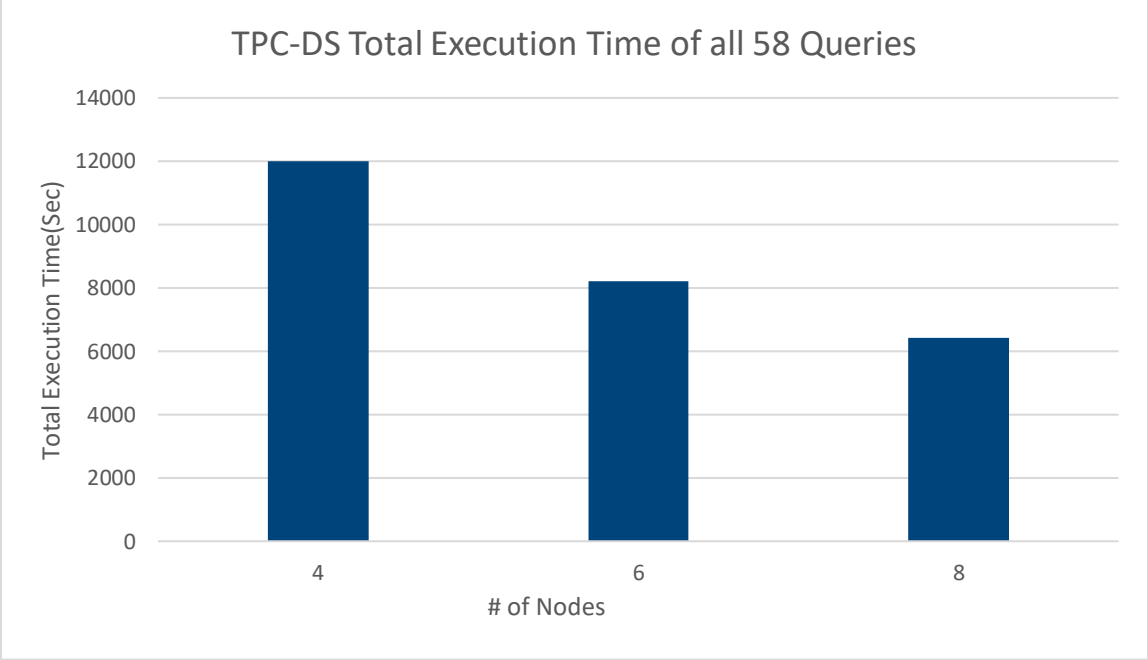


Figure 13 Dremio cluster TPC-DS total execution time

The below bar chart shows the individual query execution time for 58 TPC-DS queries for the data residing on the PowerScale F600. The shorter bars show faster execution time. Like the chart above, this demonstrates great performance improvement with larger node counts.

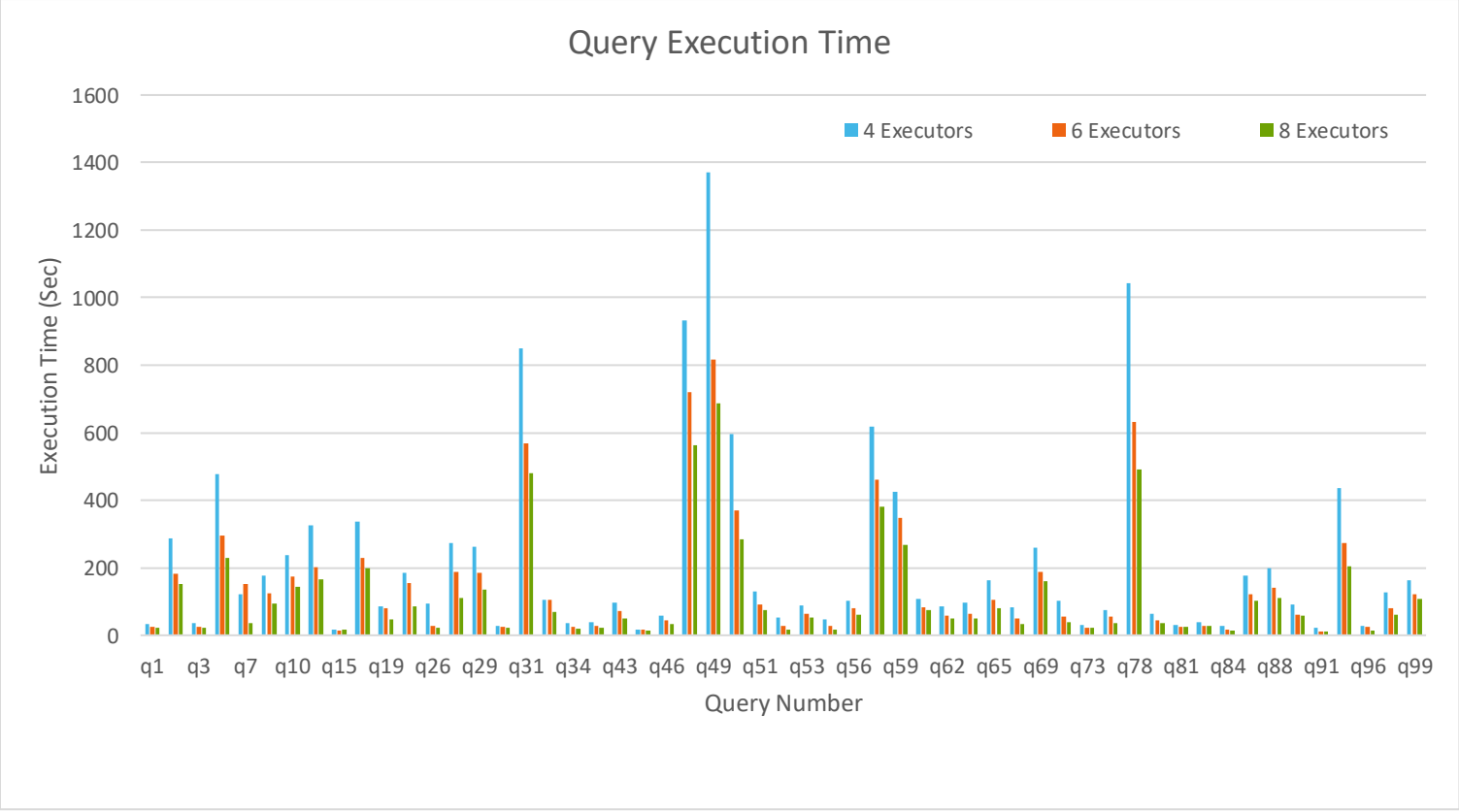


Figure 14 Dremio cluster TPC-DS individual query execution time

7.2 Dell EMC ECS and Dremio performance

The bar chart below shows the total execution time of 58 queries run sequentially for a different number of Dremio executor nodes; the shorter the bar, the better. As expected, the execution time linearly reduces with the increase in the number of Dremio executor nodes. At this scale, Dremio gains a 50% performance boost with increased number of nodes from four to eight, linearly scaling and delivering faster execution time. For a data set size of 10TB in open source parquet file format, this demonstrates excellent performance for faster time-to-insight and interactive query performance at this scale and the query performance improvement at larger node counts.

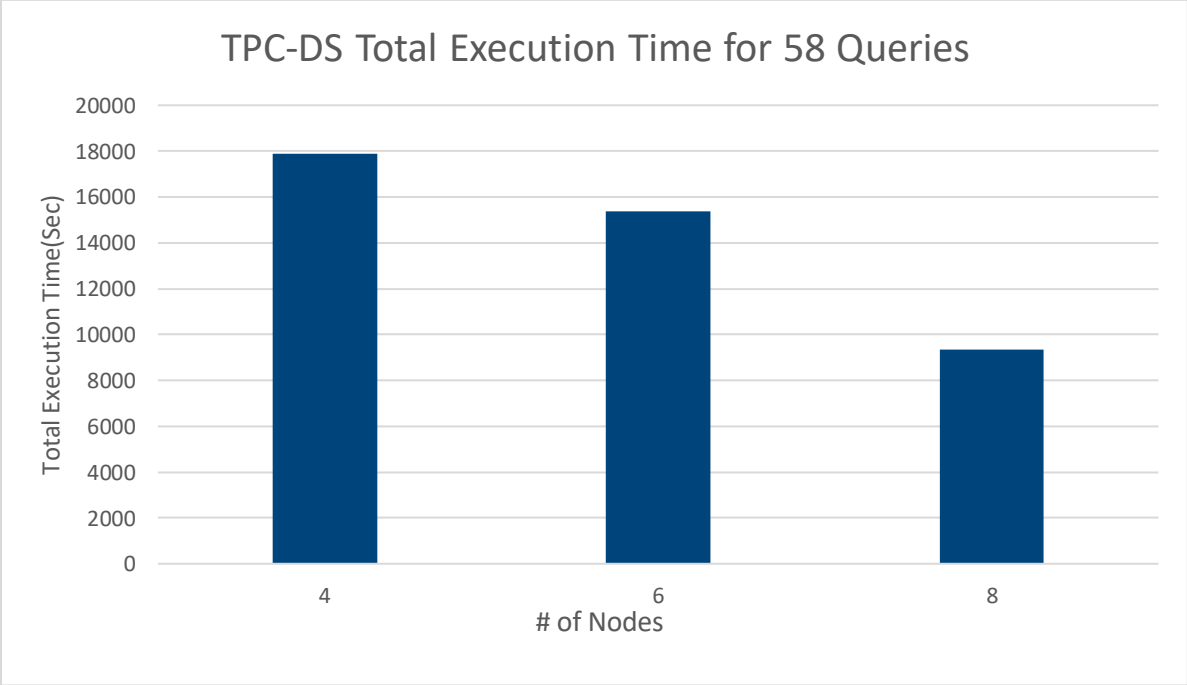


Figure 15 Dremio cluster TPC-DS total execution time

The below bar chart shows the individual query execution time for 58 TPC-DS queries for data residing on ECS EX3000; the shorter the bar, the better. Like the chart above, this demonstrates great performance improvement with larger node counts.

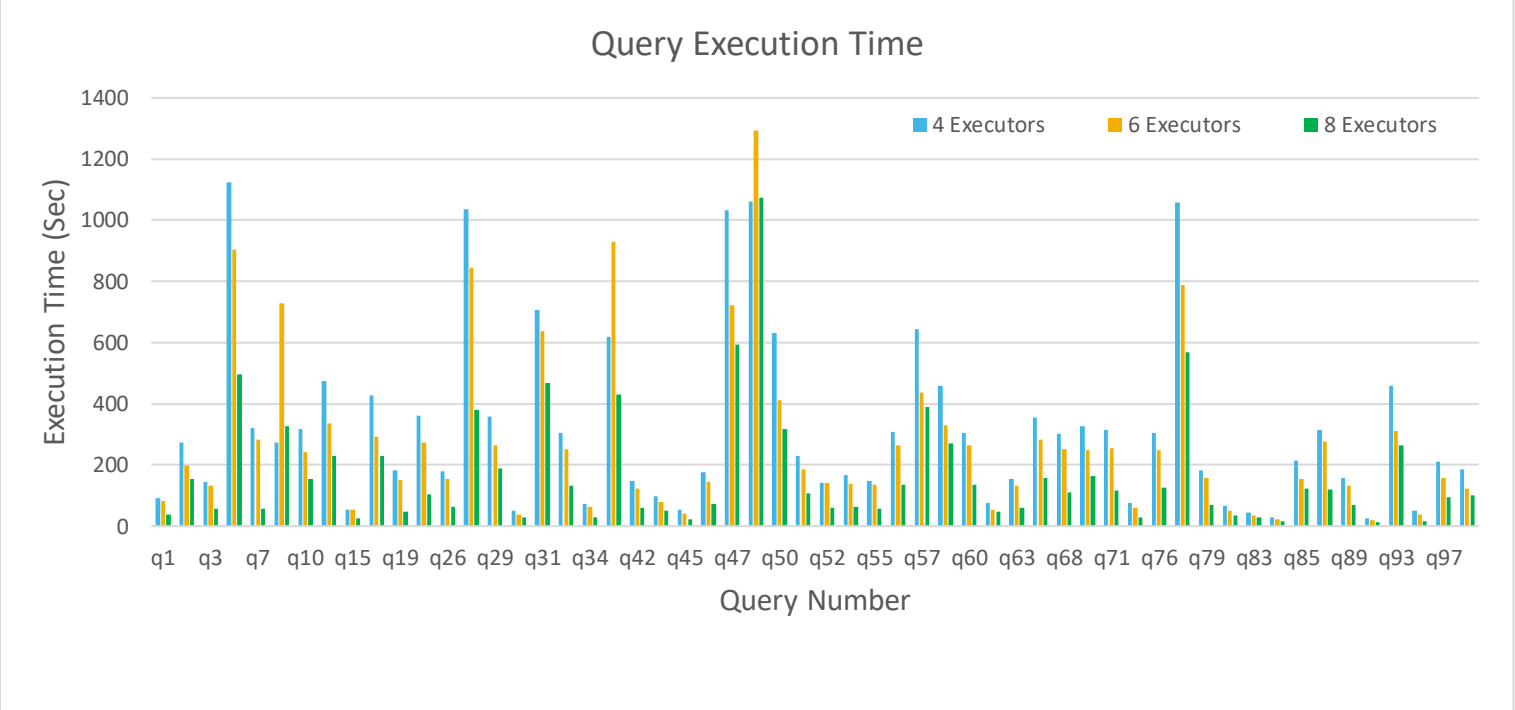


Figure 16 Dremio cluster individual TPD-DC query execution time

7.3 TPC-DS results summary

Based on the above, the results of the TPC-DS benchmark Dremio with Dell EMC PowerScale and ECS solutions delivers:

1. Lightning-fast queries directly on top of Dell EMC PowerScale and ECS
2. Up to 1,700x faster ad-hoc queries vs traditional data lake query engines with data reflections
3. Up to 3,000x faster BI queries vs traditional data lake query engines with data reflections
4. No data movement or transformation
5. Built-in Dell EMC connector
6. Easy-to-manage Dell EMC infrastructure

7.3.1 Performance and efficiency benchmark

To help clarify the performance and efficiency of Dremio with Dell EMC technologies solution, let's review the results of the publicly-available TPC-DS benchmark from Dremio's "Dremio vs Presto distros – Performance and Efficiency Benchmark" report. The report compares Dremio with various Presto distributions running on the AWS public cloud. The benchmark measures and highlights query execution cost and performance, evaluates execution time for BI/reporting and ad-hoc queries, and offers additional analysis of performance improvements with Dremio's data reflections.

We will compare the number of queries that Dremio was able to execute in 10 minutes at a storage scale factor of 10TB, and compare those numbers for Dremio on-prem with Dell EMC Technologies versus Dremio on AWS.

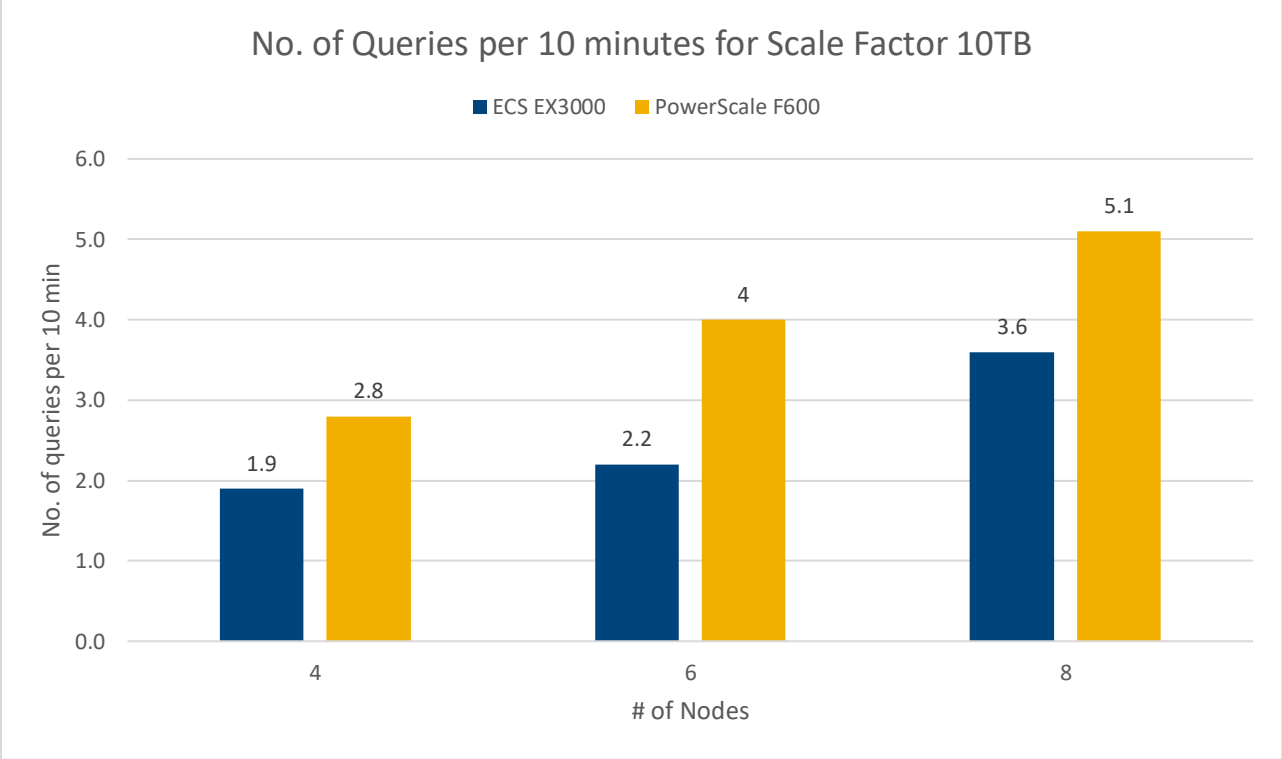


Figure 17 Number of TPC-DS queries per 10 minutes

From the bar chart above, we will compare the eight-node Dremio results using Dell ECS against the eight-node Dremio results on AWS, and forecast the linear scalability for 12, 16, and 20 nodes as shown in the below chart from Dremio vs Presto distros blog.

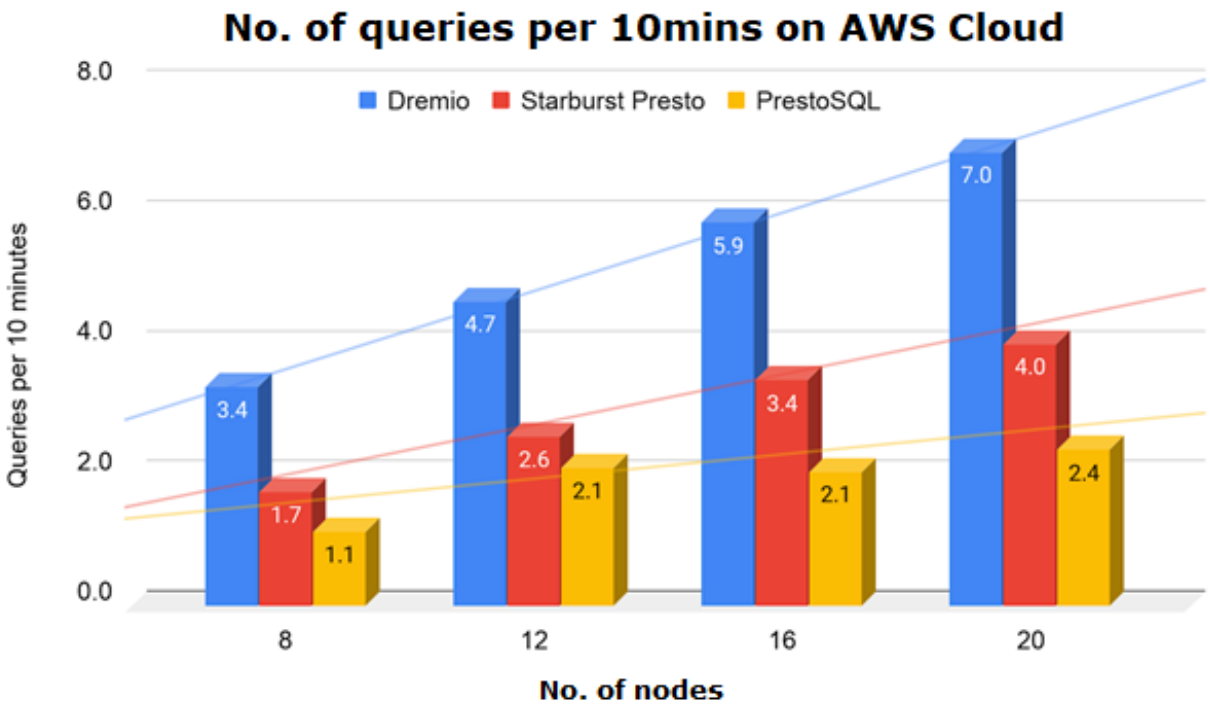


Figure 18 Number of queries per 10mins on AWS

Based on both bar charts, it is clear that at eight nodes, Dremio with Dell EMC ECS EX3000 via the S3 connector delivers slightly better performance (3.6 versus 3.4 queries per 10 minutes) compared to Dremio running on EC2 with data stored in AWS S3.

Also, Dremio with PowerScale F600 via an HDFS connector with the same node count shows excellent performance results compared to Dremio running on EC2 (5.1 versus 3.4 queries per 10 minutes). This represents 30% more queries per minutes for the same number of nodes. This clearly shows how much faster the Dremio and PowerScale F600 solution is compared to AWS.

8 Conclusion

Dremio is designed to enable sub-second SQL analytics on various storage systems in anticipation of vast increases in data volumes. The ability of PowerScale OneFS and ECS to scale to multi-petabytes while delivering high performance I/O makes PowerScale and ECS storage ideal for Dremio to offer greatly improved workload portability when offloading workloads from legacy data lakes is considered.

This paper clearly shows that Dell EMC PowerScale F600 All-Flash NAS storage and Dell EMC ECS Object storage solutions meet data lake storage and performance requirements and performs very well under I/O load with Dremio used for both small and large data sets. The TPC-DS test results for all 58 TPC-DS queries are included in this paper and show excellent results in terms of execution time for 10TB scale factor.

With Dell EMC PowerScale and ECS, enterprise organizations and Dremio administrators can effortlessly scale from tens of terabytes to tens of petabytes within a single file system, single volume, and with a single point of administration. Dell EMC PowerScale and ECS delivers high-performance and high throughput without adding management complexity.

A Configuration details

A.1 Dremio cluster configurations

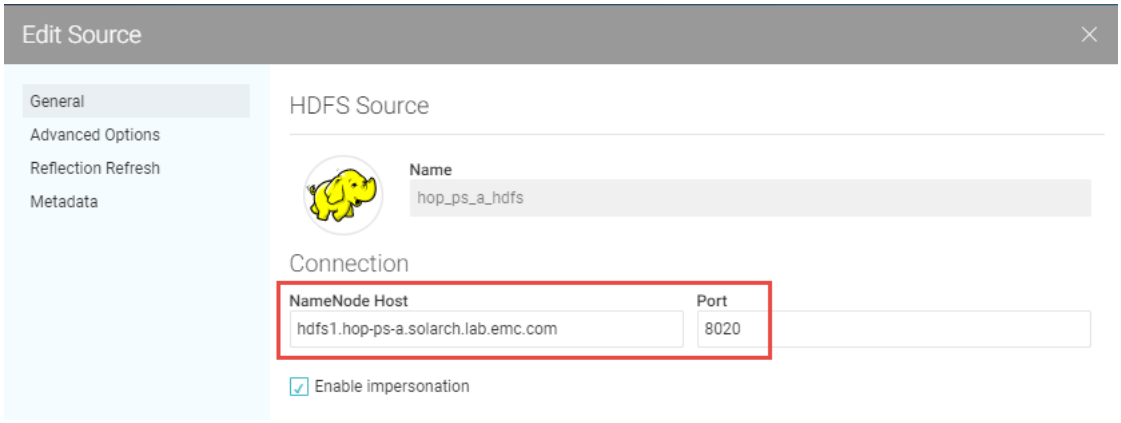
A.1.1 Dremio coordinator conf file

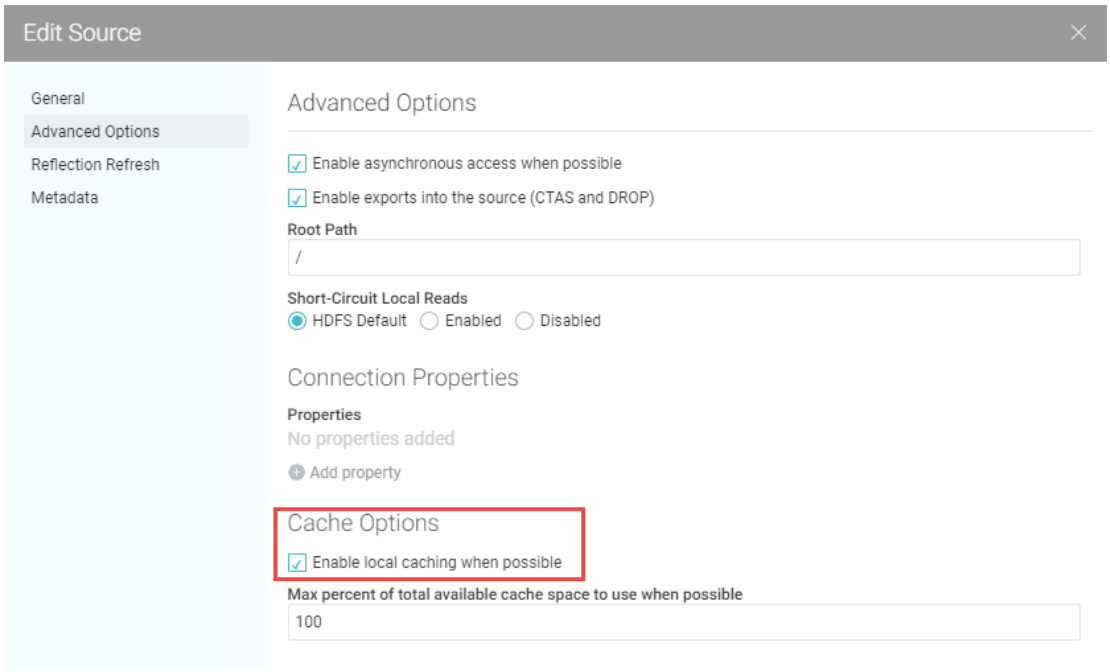
```
paths: {
  # the local path for dremio to store data.
  local: "/data1/customDremio"

  # the distributed path Dremio data including job results, downloads, uploads, etc
  #dist: "pdfs://"${paths.local}"/pdfs"
}

services: {
  coordinator.enabled: true,
  coordinator.master.enabled: true,
  executor.enabled: false
  executor.cache.path.db : "/data2/customDremio/cachemanagerdisk/db",
  executor.cache.path.fs : [ "/data2/customDremio/cachemanagerdisk/dir1"]
}

zookeeper: "hop-r640-01.solarch.lab.emc.com:2181"
[root@hop-r640-01 conf]#
```

A.1.2 Dremio executor conf file

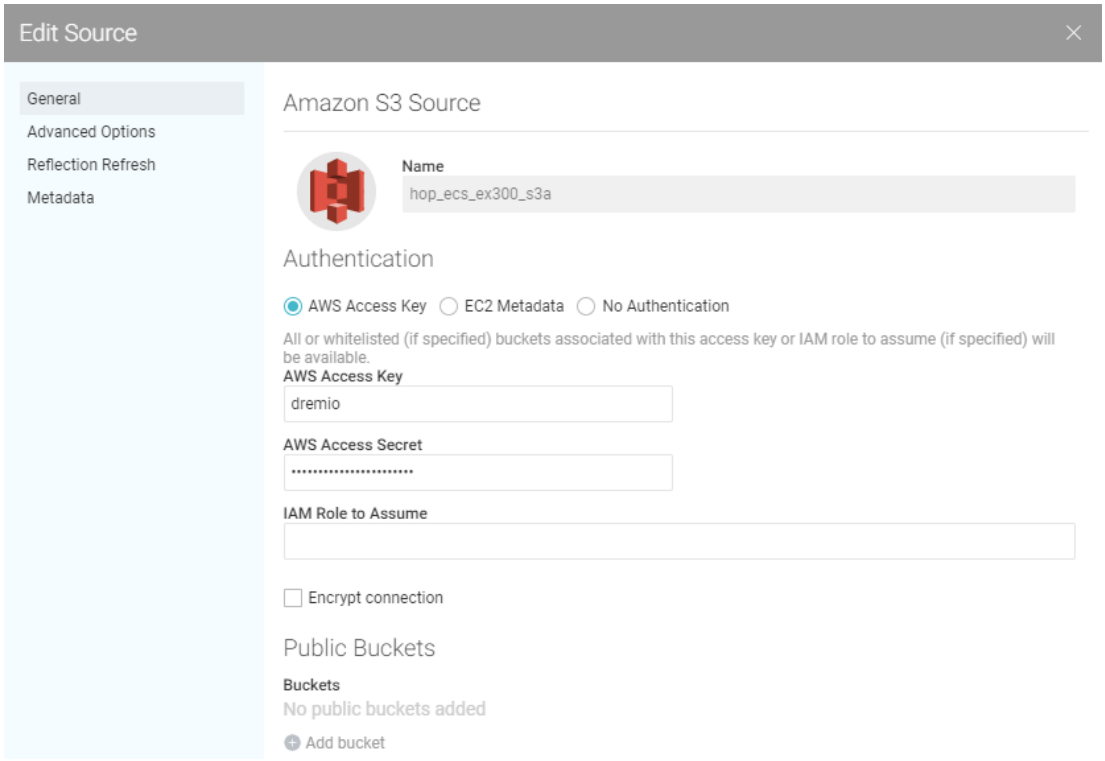
```
paths: {
  # the local path for dremio to store data.
  local: "/data1/customDremio"

  # the distributed path Dremio data including job results, downloads, uploads, etc
  #dist: "pdfs://"${paths.local}"/pdfs"
}

services: {
  coordinator.enabled: false,
  coordinator.master.enabled: false,
  executor.enabled: true
  executor.cache.path.db : "/data2/customDremio/cachemanagerdisk/db",
  executor.cache.path.fs : [ "/data2/customDremio/cachemanagerdisk/dir1"]

}

zookeeper: "hop-r640-01.solarch.lab.emc.com:2181"
[root@hop-r640-02 conf]#
```

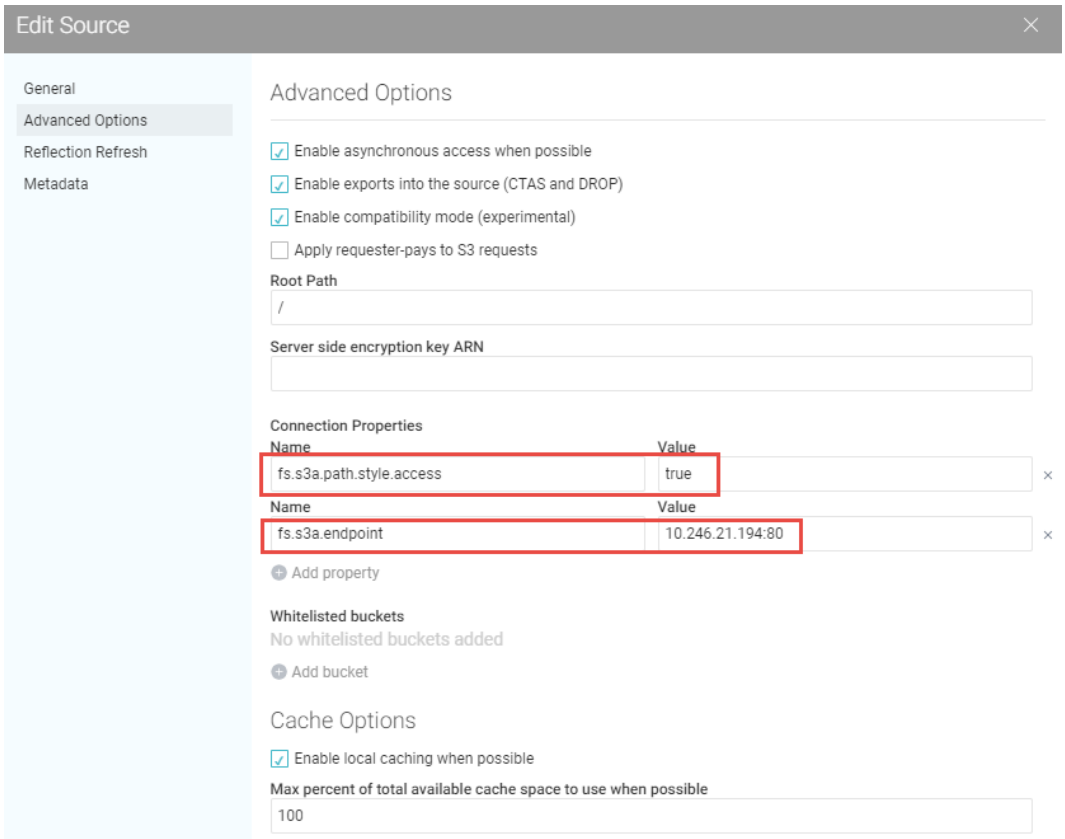
A.2 Dremio Dell EMC PowerScale HDFS connection





A.3 Dremio Dell EMC ECS S3 connection





B Technical support and resources

Dell.com/support is focused on meeting customer needs with proven services and support.

Storage and data protection technical white papers and videos provide expertise that helps to ensure customer success with Dell EMC storage and data protection products.

B.1 Related resources

- Dremio docs
- Dremio tutorial
- Dremio university
- Apache Jmeter
- TPC-DS Benchmarking
- Make Dell EMC PowerScale and ECS with Dremio the Centerpiece of your Data Analytics Infrastructure
- Dell EMC PowerScale
- Dell EMC ECS