White Paper

High Performance Cloud

Accelerate Dremio Data Lakehouse Queries in the Cloud with AWS m6 Instances

TPC-DS benchmarks of Dremio show 3rd Gen Intel® Xeon® Scalable processor-based m6 instances deliver up to 29 percent faster queries than on m5 with 2nd Gen Intel® Xeon® Scalable processor[1]

Author

Martin Dimitrov
Systems Engineer, Intel

Mark Shainman
Principal Product Marketing Manager, Dremio

Nagesh Ramaiah
Staff Software Engineer, Dremio

Executive Summary

As enterprise data expands exponentially, resulting in siloed data stores in multiple forms and formats, accessing and managing the data has become increasingly complex and expensive. It is difficult for data scientists and analysts to find the data they need for analysis across highly distributed, unconnected data environments. At the same time, data engineers must create complex data pipelines to get the data to correct systems, in the correct format, where users can analyze it—delaying time-to-insight and critical decision making.

Dremio is a data lakehouse platform that enables self-service SQL queries at sub-second response times across all of a company's data, both in the data lake and in other repositories. Now, all users can easily run their own SQL queries to create unique views and reports for their specific departmental needs.

To evaluate performance of various cloud configurations, Dremio performed the TPC-DS benchmark on AWS EC2 instances running 2nd Gen Intel® Xeon® Scalable processors and 3rd Gen Intel® Xeon® Scalable processors. The results show that instances with the newer Intel Xeon processors deliver up to 1.29x faster queries than instances with the previous generation CPU. This means these newer instances can accelerate time-to-insight and decision-making for data-driven businesses that want to stay competitive.

This paper describes the Dremio platform, its architecture, and the benchmarks Dremio performed with significant results that can benefit businesses large and small.

Contents

Dremio Overview

From small business to enterprise and government, organizations continue to seek and gain critical insight from their data. The amount of data and number of data sources continue to grow exponentially, with companies continuing to struggle to not only manage all of the data, but to derive business value from it. With traditional data architectures that rely on data warehouses, companies must create complex pipelines to access, extract, and load data into data warehouses in order to analyze it. These pipelines are complex and time-consuming to create, and traditional data warehouses are expensive to license as well as maintain.

Dremio empowers enterprises with a seamless, open Unified Analytics Platform that unlocks the full potential of a company's data. Dremio delivers self-service analytics and data management with compelling price-performance and cost-effectiveness. With Dremio, organizations can power BI dashboards and interactive analytics directly on a data lake and across all of a company's data where it lives with a tightly integrated, highly performant SQL engine.

Additionally, Dremio's Apache-native Lakehouse Management capabilities simplify data discovery and automate data optimization and management, delivering high-performance analytics with Git-inspired data versioning.

Built on Open Standards, No Vendor Lock-In

Dremio is foundationally built-on and supports open source technologies (Figure 1), including Apache Parquet and Apache Iceberg, and Apache Arrow and Project Nessie (which Dremio helped create). Dremio uses no proprietary formats, and frictionlessly connects to dozens of data sources, including object stores like Amazon S3, other databases, and metastores.

Figure 1. Dremio is built on and supports open standards to help ensure performance with no vendor lock-in. (Source: Dremio).

Use Cases

Many companies, from small to large enterprises have deployed Dremio in their Business Intelligence and analytics infrastructure.

Amazon: Dremio's data lakehouse helped Amazon achieve 10x faster query performance, eliminate 60 hours of work per project, and reduce project completion times by 90 percent.[2]

Henkel: Dremio accelerated insights in the Henkel supply chain, improving manufacturing equipment productivity and lowering costs. By sharing data across silos, Henkle improved equipment efficiency by more than 10 percent across 250 manufacturing lines. Query response was reduced from three to four minutes to eight seconds, improving price performance by 30x.[3]

DB Cargo: After outgrowing their on-premises data warehouses, DB Cargo moved their data to AWS and deployed Dremio as part of a robust and reliable data lakehouse architecture. Dremio gave DB Cargo users—from data scientists and analysts to HR, sales, and management accounting—self-service access to all data, running queries quickly and easily. Dremio's open standards format allowed DB Cargo to integrate other open community projects, like Apache Spark and Apache Nifi, to modernize their new architecture without vendor lock-in. Dremio helped deliver near-real-time results and more efficient transportation planning to ensure higher shipment quality and faster decisions.[4]

Dremio's worldwide customers have achieved greater productivity and efficiencies using the Dremio platform. Find out more in their customer case studies.

Dremio Architecture

The Dremio platform comprises several modules (Figure 2), supporting access to a host of data formats and integrating with leading Business Intelligence dashboards and tools and analytics applications.

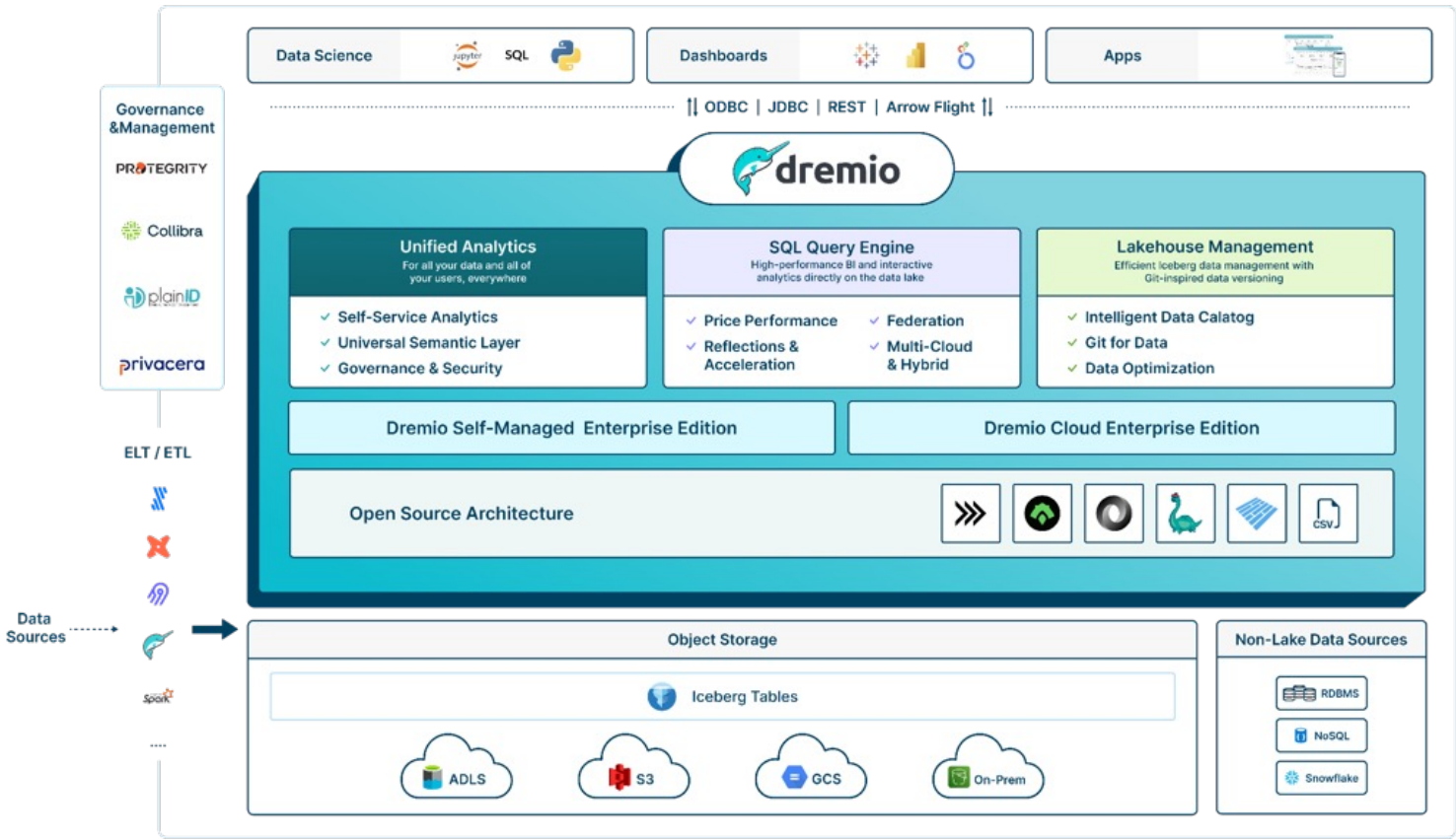

Figure 2. Dremio architecture is built on open standards with connectivity to non-lake data sources (Source: Dremio).

Unified Analytics for All Your Users

Dremio Unified Analytics enables easy-to-use self-service analytics for all users, from data scientists to analysts.

With a Universal Semantic Layer, Dremio facilitates a business-friendly language for defining federated, consistent, and secure data views, ensuring accessibility and comprehension for all users. The platform promotes collaboration through the curation, analysis, and sharing of data views, fostering consistent and collaborative data projects.

In addition, Dremio provides centralized data governance, striking a balance between data access and control. It offers fine-grained access control at every level, from raw data sources to shared and governed data views. Dremio also helps audit and monitor data access to help ensure data security and regulatory compliance. Extensive data lineage capabilities make it easy to view the full data lineage for all datasets and queries, enhancing data use transparency.

Frictionless connector integrations enhance versatility, featuring native Business Intelligence connectors. It also integrates compatibility with various data sources, spanning object storage, metastores, and both cloud-based and on-premises databases. This unified approach accelerates business insight, innovation, and results without the need for data movement.

SQL Query Engine for Sub-second Performance Directly on Your Data Lake

The Dremio SQL Query Engine stands out as a high-performance solution tailored for seamless analytics directly on the data lake, offering sub-second performance for BI workloads across diverse data sources. The solution delivers lightning-fast query performance at scale, leveraging a multi-engine architecture for sophisticated workload management, intelligent autoscaling, and cost-based optimization. Dremio's intelligent SQL query acceleration technology, Reflections, significantly speeds up query performance, enabling near-instantaneous query performance.

Furthermore, the platform facilitates flexible, fast, and lightweight data transformation with SQL capabilities, including filtering, sorting, aggregating, joining, and casting. It thus reduces the reliance on costly ETL processes and complex data pipelines. Dremio supports cloud, multi-cloud, on-premises, and hybrid environments, enabling users to analyze data wherever it resides and harness its analytical potential within seconds.

The Dremio SQL Query Engine includes:

Columnar Cloud Cache (C3) – C3 allows Dremio to achieve NVMe-level I/O performance on S3, ADLS, and GCS by selectively caching the data required to satisfy workloads. C3 also eliminates over 90 percent of S3, ADLS, and GCS I/O costs, saving up to 15 percent of the costs for each query.[5]

Reflections – Reflections intelligently precompute aggregations and other operations on your data, eliminating the need to do these on the fly. Reflections are transparent to end users: Dremio's built-in optimizer picks the best Reflections to accelerate a query.

Cost-Based Optimizer – Dremio understands the deep statistics about your data, including location, cardinality, and distribution. A built-in Cost-Based Optimizer accurately predicts how much data will flow through a query's operators and will choose the best plan to accelerate the query.

Apache Arrow Gandiva and Apache Arrow Flight – These are two open source standards that further accelerate in-memory computing using Apache Arrow's in-memory computing columnar table format. Gandiva maximizes CPU utilization and leverages optimizations, such as vectorized processing and SIMD execution. Flight enables parallelism in data transfers, accelerating query results up to 100x compared to JDBC and ODBC interfaces.[6]

For more details on the Sonar query engine, visit www.dremio.com/platform/sonar/query-engine.

Lakehouse Management Automates Data Optimization and Operations

Dremio Lakehouse Management is a catalog and data management and optimization service with innovative Git-for-data functionality. Lakehouse Management provides Git-like branching and versioning, allowing analysts to isolate and manage their data. Analysts and scientists can engineer and experiment on data without creating physical copies, helping to prevent contamination of production data sources while exploring approaches to further insight.

Dremio also automates data management tasks like compaction and garbage collection to optimize query performance and minimize storage costs. Users can govern data by securing and tracking access to data with Role-Based Access Control (RBAC) privileges and a built-in commit log. Built on open source Project Nessie, Dremio Lakehouse Management can use a variety of engines to work with data, including the Dremio SQL Engine, Spark, Flink, and Trino.

For more details on Dremio Lakehouse Management and use case examples, visit https://docs.dremio.com/cloud/arctic and www.dremio.com/platform/arctic.

Deploying Dremio

Dremio can be deployed as a self-managed platform in the cloud or on-premises, or as a fully managed data lakehouse with Dremio Cloud.

- Self-managed cloud or on-premises: Dremio can be deployed as a self-managed service in the cloud on bare-metal or other instances, such as AWS m5 and m6, on Azure, and Google Cloud.
- Fully managed cloud: Dremio Cloud is currently available on AWS and Microsoft Azure.

For self-managed cloud deployments, Dremio recommends AWS m6 instances with 3rd Gen Intel Xeon Scalable processors. The benchmarks on the following page illustrate these performance gains and performance/dollar.

As a result of the benchmarking, Dremio is migrating its Dremio Cloud offering on AWS to m6 instances to deliver improved performance.

Hadoop Modernization with Dremio

Hadoop was once a cutting-edge framework for big data, but it's no longer the most efficient infrastructure. With the availability of Apache Spark and other frameworks, Hadoop's limitations are apparent. It is becoming increasingly necessary to modernize any Hadoop infrastructure. Hadoop modernization is a crucial step for organizations to unlock the full potential of big data, improve performance, and stay competitive in today's data-driven landscape.

There are several reasons to modernize a Hadoop framework, including eliminating outdated technologies, high costs, reduced efficiency, and security vulnerabilities. Its lack of standardization across distributions can result in many challenges, including bottlenecks with HDFS.

Migrating from Hadoop to a modern lakehouse can help improve performance, reduce latency, and optimize resource utilization, while enhancing security, data governance, and compliance. By leveraging today's modern frameworks, storage, and cloud-based solutions, organizations improve agility, scalability, and efficiency, while embracing a self-service analytics culture. Such a transformation can reduce the total cost of ownership (TCO) and breed innovation and growth across an organization.

Dremio is dedicated to helping organizations simplify their Hadoop modernization and migration process with a phased approach. For further information on the phased approach to the data lakehouse, visit our Hadoop Migration page.

Learn more about simplifying your journey to Dremio's data lakehouse with our Hadoop migration and modernization playbook.

Benchmarks

Dremio benchmarked its software on AWS EC2 cloud instances to evaluate any performance gains offered by EC2 instances based on the 3rd Gen Intel Xeon Scalable processor. The industry-standard TPC-DS benchmark was used to evaluate performance on both m5dn.8xlarge (2nd Gen Intel Xeon Scalable processors) and m6id.8xlarge (3rd Gen Intel Xeon Scalable processors) instances. For details about the TPC-DS benchmark, see the TPC-DS page of the Transaction Processing Council website.

Benchmark Configuration

To normalize the tests as much as possible, the benchmarks were run using out-of-box commercial Dremio software on nearly identical instance configurations. The differences in the instances centered around processor choice and resources (memory and disk size) built into each instance (Table 1). Refer to the AWS website for further details of instance configurations.

Instance Size	vCPU	Memory (GB)	Instance Storage (GB)	Network Bandwidth (Gbps)	EBS Bandwidth (Mbps)
m5dn.8xlarge	32	128	2 x 600 NVMe SSD	25	6,800
m6id.8xlarge	32	128	1 x 1,900 NVMe SSD	12.5	10

Table 1. AWS instance configurations.[7]

For the benchmarks, Dremio configured 8-node and 16-node clusters for each of the m5dn.8xlarge and m6id.8xlarge configurations (Table 2).

Cluster Size	vCPUs	Memory (GB)	Storage (GB)
8 nodes	256	1,024	19,661
16 nodes	512	2,048	30,400

Table 2. Cluster configurations for TPC-DS benchmark tests.

Six types of benchmarks were completed, measuring query times between the two instances:

- 1 TB non-partitioned data on an 8-node cluster
- 1 TB non-partitioned data on a 16-node cluster
- 1 TB partitioned data on an 8-node cluster
- 10 TB non-partitioned data on a 16-node cluster
- 10 TB partitioned data on a 16-node cluster
- 1 TB non-partitioned data on an 8-node cluster running a concurrency benchmark

For the concurrency benchmark, five concurrent users executed 100 TPC-DS queries for a total of 500 queries.

Results: 1.11 to 1.29x Faster Queries on AWS m6id.8xlarge Instances

Across the six tests, the m6id.8xlarge instances queries ran 1.11x to 1.29x faster, as shown in Figure 3. The chart presents a normalized baseline representative of each of the m5dn.8xlarge instance results and relative speedup for each of the m6id.8xlarge instance results. The query times and speedup data for the benchmark are presented in Table 3.

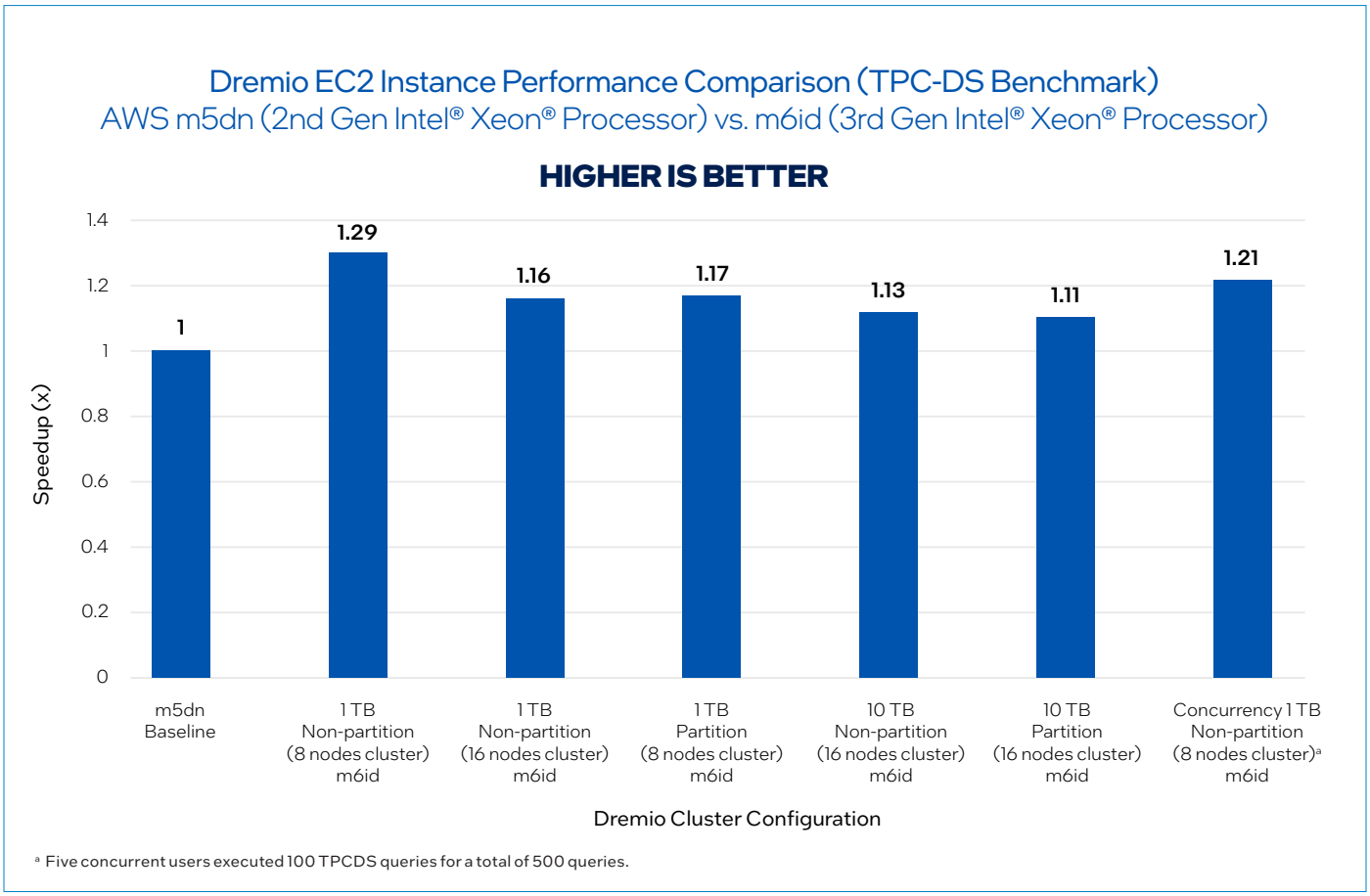

Figure 3. Benchmark results show from 11 to 29 percent faster queries with m6id.8xlarge instances.

	Instance Type	Execution Time	Speedup (x)
1 TB Non-partition (8 Nodes cluster)	m5dn	0:14:36	
	m6id	0:11:19	1.29
1 TB Non-partition (16 Nodes cluster)	m5dn	0:09:57	
	m6id	0:08:23	1.16
1 TB Partition (8 nodes cluster)	m5dn	0:10:18	
	m6id	0:08:33	1.17
10 TB Non-partition (16 Nodes cluster)	m5dn	1:04:46	
	m6id	0:56:02	1.13
10 TB Partition (16 Nodes cluster)	m5dn	0:43:51	
	m6id	0:38:55	1.11
Concurrency 1 TB Non-partition (8 Nodes cluster) 5 concurrent users executed 100 TPCDS queries ; total queries executed by concurrent workload is 500	m5dn	0:54:21	
	m6id	0:43:11	1.21

Table 3. TPC-DS Benchmark query measurements and speedup for AWS m5dn.8xlarge and m6id.8xlarge instances.

Analysis

Running on m6id.8xlarge instances shows that the 3rd Gen Intel Xeon Scalable processor-based instances deliver faster query response on large data lakes, whether partitioned or not. Partitioned database performance is faster than non-partitioned as expected. However, it's interesting to note that the 16-node clusters do not perform as well as the 8-node clusters. This is something to consider for further study.

The 3rd Gen Intel Xeon Scalable processor contains a new microarchitecture and is manufactured on a new process technology (10nm) compared to the 2nd Gen Intel Xeon Scalable processor. The new CPU cores have deeper and wider pipelines, larger caches, and built-in capabilities for AI, crypto, database compression, and security. In addition, I/O is faster due to PCIe Gen 4 support. A new memory controller enables larger memory capacities with 8 memory channels of DDR4 3200. Dremio query performance benefits from these improvements, which deliver considerable business benefits, including faster time-to-insight.

Run Dremio on 3rd Gen Intel Xeon Scalable Processor-Based Instances

Intel® architecture is a trusted foundation that provides what businesses need to build, scale, and transform. Intel Xeon processors are backed by a proven history of four successful platform generations. Each generation of Intel Xeon processors offers a wide range of features for computing on-premises or in the cloud. From more cores and threads to faster clock speeds, better power efficiency, and next-generation memory and I/O, Intel Xeon processors provide the right solution for any use case.

In addition, advanced features, such as built-in accelerators for AI, analytics, image processing, and data streaming can help improve workload processing and server efficiency for many workloads, including Business Intelligence applications such as Dremio. These features help drive down total cost of ownership, while enabling innovation and the delivery of new services.

When it comes to a company's data, hardware-enabled security capabilities in the CPU can help protect data in any deployment by keeping it encrypted and isolated from other workloads. Intel Xeon processors integrated powerful security acceleration and enablement technologies that deliver performance while keeping data and code secure against today's threats.

3rd Gen Intel Xeon Scalable Processor Capabilities Overview

The 3rd Gen Intel Xeon Scalable processors are based on a new balanced, efficient architecture that increases core performance, memory, and I/O bandwidth compared to previous generations. These CPUs integrate built-in acceleration and advanced security capabilities, designed through decades of innovation for the most in-demand workload requirements. They are optimized for cloud, enterprise, HPC, network, security, and IoT workloads, with 8 to 40 powerful cores and a wide range of frequency, feature, and power levels.

Compared to the previous generation, these CPUs deliver:

- 1.46x average performance improvement.[8]
- Up to 1.60x higher memory bandwidth.[9]
- Up to 2.66x higher memory capacity.[10]
- Up to 1.33x more PCI Express lanes per processor.[11]

3rd Gen Intel Xeon Scalable processors are at the core of strong, capable platforms—on-premises and in the cloud—for the data-fueled enterprise. Key features and capabilities include the following:

- Infused with Intel® Crypto Acceleration, enhancing data protection and privacy by increasing the performance of encryption-intensive workloads, while reducing the performance impact of pervasive encryption.
- Built-in AI acceleration, end-to-end data science tools, and an ecosystem of smart solutions.
- Engineered for the demands of cloud workloads and to fuel a wide range of XaaS environments.
- Fueled by Intel® Software Guard Extensions (Intel® SGX), which protects data and application code while in use from the edge to the data center and multi-tenant public cloud.
- Built-in workload acceleration features include Intel® Deep Learning Boost (Intel® DL Boost), Intel® Advanced Vector Extensions 512 (Intel® AVX-512), and Intel® Speed Select technology (Intel® SST).

These capabilities and features help power Dremio performance on AWS m6 instances. When designing your data and Business Intelligence infrastructure, Dremio recommends you select these instances with 3rd Gen Intel Xeon Scalable processors.

Summary and Conclusion

Dremio revolutionizes the landscape of data analytics by eliminating the complexities and costs associated with data integration and ETL processes, offering a seamless enterprise-scale analytics solution without the need for data movement. The Dremio Unified Analytics Platform empowers users with direct access to data lakes through a unified semantic layer, fostering self-service SQL queries for swift insights crucial to business decision-making.

Dremio's architecture optimizes queries through features like Reflections query acceleration and a built-in cost optimizer. Dremio Lakehouse Management, featuring its own data catalog built on Apache Iceberg, helps automate lakehouse operations, including versioning and branching capabilities. Versioning and branching allow data scientists and analysts to experiment effortlessly with production and new data sources without creating duplicate copies or compromising the integrity of production data lakes.

With its combination of ultra-fast, self-service SQL queries and an efficient management platform, Dremio becomes a transformative force, reshaping how companies leverage their data for greater insights.

Dremio can be deployed on-premises and in the cloud, and Dremio offers its own cloud platform on AWS and Azure.

Through Dremio's benchmarking using the TPC-DS industry-standard benchmark, Dremio achieves up to over 1.29x faster queries on AWS m6id.8xlarge instances with the 3rd Gen Intel Xeon Scalable processors. Based on these benchmarks, Dremio is migrating its Dremio Cloud Platform to 3rd Gen Intel Xeon Scalable processor-based instances. The company recommends that for self-managed cloud or on-premises deployments, customers also choose infrastructure with this newer Intel CPU.

To find out more about Dremio and the Dremio Cloud Platform, visit Dremio.com.

To get started with Dremio Cloud, visit www.dremio.com/get-started/.

[1] See Table 1 and Table 2 in Benchmarks section for configurations. Testing done by Intel for Dremio.

[2] See https://www.dremio.com/customers/amazon/

[3] See https://www.dremio.com/customers/henkel/

[4] See https://www.dremio.com/customers/db-cargo/

[5] https://docs.dremio.com/current/get-started/cluster-deployments/customizing-configuration/dremio-conf/cloud-cache-config/ and https://www.dremio.com/press-releases/announcing-the-data-lake-engine-dremio-4-0/

[6] See https://www.dremio.com/blog/announcing-gandiva-initiative-for-apache-arrow/

[7] https://aws.amazon.com/ec2/instance-types/m5/ and https://aws.amazon.com/ec2/instance-types/m6i/

[8] Please visit www.intel.com/3gen-xeon-config and use the corresponding performance number [#] to access full system configuration and performance detail.

- Process up to 1.64x more database transactions per minutes vs. prior gen [81]
- Up to 1.72x higher virtualization performance vs. prior gen [84]
- 1.62x average performance improvement across network and communications workloads vs. prior gen [91]
- 2x Massive MIMO throughput in a similar power envelope for a best-in-class 3x100MHz 64T64R configuration [91]
- 1.76x enhanced DPDK L3 forwarding vs. prior gen [91]
- Up to 1.63x increased throughput enabling you to serve the same number of users at a higher resolution or a greater number of subscribers at the same resolution vs. prior gen [91]
- Increase 5G UPF performance by 1.42x vs. prior gen [91]
- Up to 1.48x faster encryption performance with Intel Crypto Acceleration vs. prior gen [97]
- 1.58x higher performance on cloud-based microservices vs. prior gen [98]
- 1.53x higher HPC performance vs. prior gen [108]
- 1.56x improvement in AI inference for image classification with enhanced Intel Deep Learning Boost vs. prior gen [119]
- Up to 1.74x more AI inference performance with enhanced Intel Deep Learning Boost vs. prior gen [120]

[9] 3rd Gen Intel Xeon Platinum 8380 CPU: 8 channels, 3200 MT/s (2 DPC) vs. 2nd Gen Intel Xeon Platinum 8280 CPU: 6 channels, 2666 MT/S (2 DPC).

[10] 3rd Gen Intel Xeon Platinum 8380 CPU: 8 channels, 2 DPC (256GB DDR4) vs. 2nd Gen Intel Xeon Platinum 8280 CPU: 6 channels, 2 DPC (128GB DDR4).

[11] 3rd Gen Intel Xeon Platinum 8380 CPU: 64 lanes of PCI Express 4 per processor vs. 2nd Gen Intel Xeon Platinum 8280 CPU: 48 lanes of PCI Express 3 per processor.