Hewlett Packard Enterprise

Solution brief

Check if the document is available in the language of your choice.

THE FOUNDATION FOR A MODERN DATA LAKE

HPE Apollo 4000 systems

HPE Apollo 4200 Gen10+ LFF

HPE Apollo 4200 Gen10+ SFF

HPE Apollo 4510 Gen10

INTRODUCTION

The availability of open-source frameworks and distributed file systems have enabled enterprises to integrate structured and unstructured data from a wide array of sources. This has better equipped enterprises to create and manage data lakes on-premises or in the cloud, while enabling analytics on that data, providing new insights, and accelerating business decisions.

The emergence of IoT and the instrumentation of everything is contributing further to the exponential growth of data lakes, exacerbating the challenges enterprises face as they search for better ways to manage all the data and to utilize it in achieving critical business objectives.

"Increasingly, data and analytics has become a primary driver of business strategy and the potential for data-driven business strategies and information products is greater than ever."

– Gartner[1]

ARE DATA LAKES STILL RELEVANT?

Despite the presence of data lakes, enterprises continue to struggle with managing the exponential growth in the volume of data, often resulting in the proliferation of data silos that have little or no governance. The legacy of application-centric architectures inhibits the move to more data-centric architectures. Untrusted data due to poor or missing metadata management is also a challenge, along with a lack of adequate governance and oversight of data, and the inability to fully persist and use data coming in off event streams due to inadequate data pipeline architectures.

The importance and need for the data lake remain, but the objectives and requirements have expanded. What was originally designed for only extreme scale and batch analytics, now needs to also support different levels of performance, real-time data streaming and analytics, machine learning, horizontal data movement and accessibility, and the ability to securely store and catalog data, which is critical to understanding context on the data origin, when it was last updated.

WHAT IS A MODERN DATA LAKE?

Early data lakes for Hadoop were constructed using cookie cutter, symmetric node profiles with the primary objective of making it quick and easy to store structured and unstructured data at any scale, and to manage batch analytics on top of this data. That worked initially, but the introduction of new workloads along with the need for

[1] "Over 100 Data and Analytics Predictions through 2025," Gartner, March 2021

real-time streaming and analytics demanded more flexibility and elasticity in the building blocks used to create a data lake architecture.

The modern data lake must be built on infrastructure that provides the ability to optimize for both performance and cost-optimized capacity, must be extremely scalable, modular, reliable, and highly available. It should also support a variety of analytic engines, workloads, software-defined storage (SDS), as well as interfaces (for example, S3 API, Hadoop Distributed File System [HDFS], NFS, and SMB). With these diverse requirements in mind, many enterprises have also begun adopting object storage, either cloud-based or on-premises as an alternative or supplementary to HDFS.

A new data management paradigm

AI and analytics are at the core data driven organizations and digital transformation, so the modern data lake and the data management paradigm must be focused on optimizing data accessibility for analytics and AI applications taking into account the following considerations:

- Data lake architectures must be flexible to support heterogeneous hardware environments, supporting hardware accelerators (for example, GPU, persistent memory [PMEM]), as well as fast storage and SDS.
- Data mobility is no longer defined as only a primary to secondary storage, but must be able to persist data across the data pipeline, from the data lake to the optimal datastore (e.g., when you want to save the results from model testing).

HPE APOLLO 4000 FAMILY PROVIDES A DENSITY AND PERFORMANCE-OPTIMIZED BUILDING BLOCK FOR THE MODERN DATA LAKE

Constructing a modern data lake demands having flexible infrastructure building blocks to support a variety of workloads, a centralized data repository, larger archives, backup repositories, all through cost-effective storage capacity and density in a smaller form factor. The HPE Apollo 4000 family of intelligent data storage servers are architected to accommodate both ends of the data-centric workload spectrum, from deeper data lakes and archives to performance-demanding machine learning (ML), data analytics, hyperconverged infrastructure (HCI), and cache-intensive workloads.

The HPE Apollo 4200 (Gen10 and Gen10 Plus) is designed for throughput intensive workloads that require the ability to cache data, and to support flexible storage tiers with a superior balanced system architecture with faster I/O, all in an incredibly dense, 2U chassis.

The HPE Apollo 4510 Gen10 is a density-optimized, bulk capacity platform needed for deep data lakes in a 4U chassis. Unlike competitive products, the HPE Apollo 4510 Gen10 supports SAS and NVMe drives, provides easy drive access via front drawers, and integrates HPE ProLiant security and management features, offering lower cost/GB density.

WHAT'S NEW WITH THE HPE APOLLO 4200 GEN10 PLUS SERVER

Introduced in June 2021, the HPE Apollo 4200 Gen10 Plus complements the Gen10 model with advanced capabilities for even more demanding data lakes and analytics workloads, still in a data-center-friendly 2U rackmount server footprint.

- More data capacity for deep data lakes:
  - 28 LFF drive bays for cost-economic data lakes using ultra-large NL HDDs.
  - 60 SFF drive bays for All Flash data lakes using Very Read Optimized (aka QLC) SSDs.
- Higher throughput balanced architecture plus more low latency cache capacity for performance-demanding data lakes:
  - PCIe Gen4 with up to 200 Gb NICs and up to 4 storage controllers.
  - Up to 8 Persistent Memory DIMMs and up to 12 SFF NVMe drive bays.
- Superior compute for accelerated analytics within the data lake itself:
  - Capable of supporting select GPUs and FPGAs.
  - Up to two 3rd generation Intel® Xeon® Scalable processors.

CONCLUSION

The HPE Apollo 4000 family of servers provide storage optimized, purpose built servers designed for the modern data lake and data pipeline, offering flexibility, elasticity, and performance tiers to accommodate a wide array of storage-intensive workloads.

LEARN MORE AT

buy.hpe.com/us/en/p/1013422400

a50004344ENW, June 2021

Hewlett Packard Enterprise