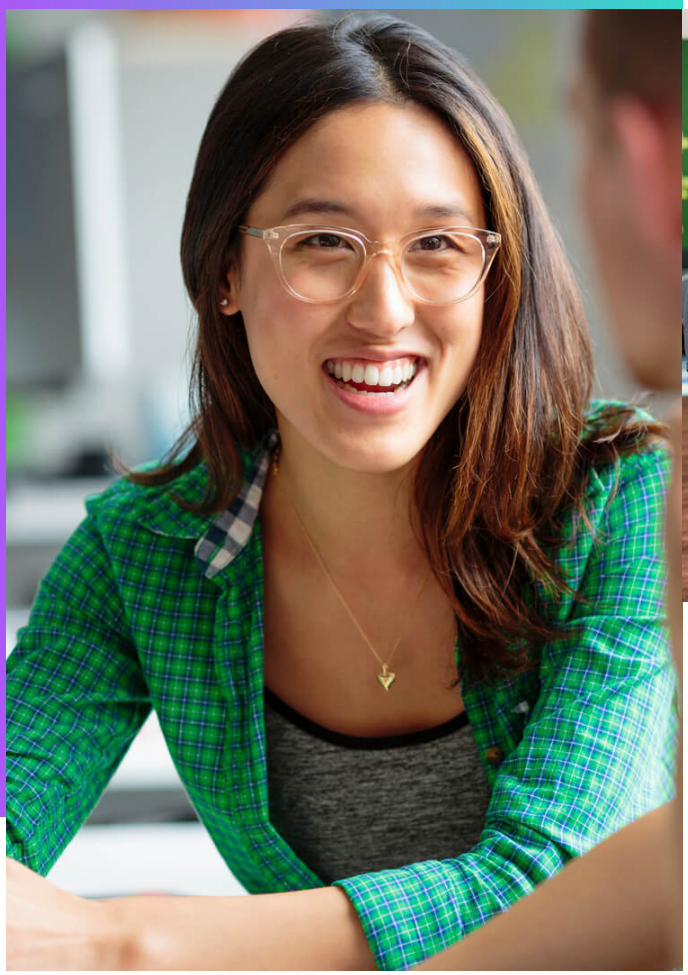

Unleash Apache Spark with HPE Ezmeral

Unified data analytics

HPE GreenLake

Exponential data growth has created an overwhelming flow of data across increasingly distributed sources. This data growth has outpaced many organizations' ability to effectively glean insights with data analytics to better understand their customers and deliver new services.

In the age of insight, Apache Spark is an open-source solution that has become the industry standard for developers and data science teams to create machine learning (ML) models that deliver the analytics required by data-driven organizations. The combination of Apache Spark and HPE Ezmeral software allows customers to quickly act on untapped value by simplifying data collection and analysis.

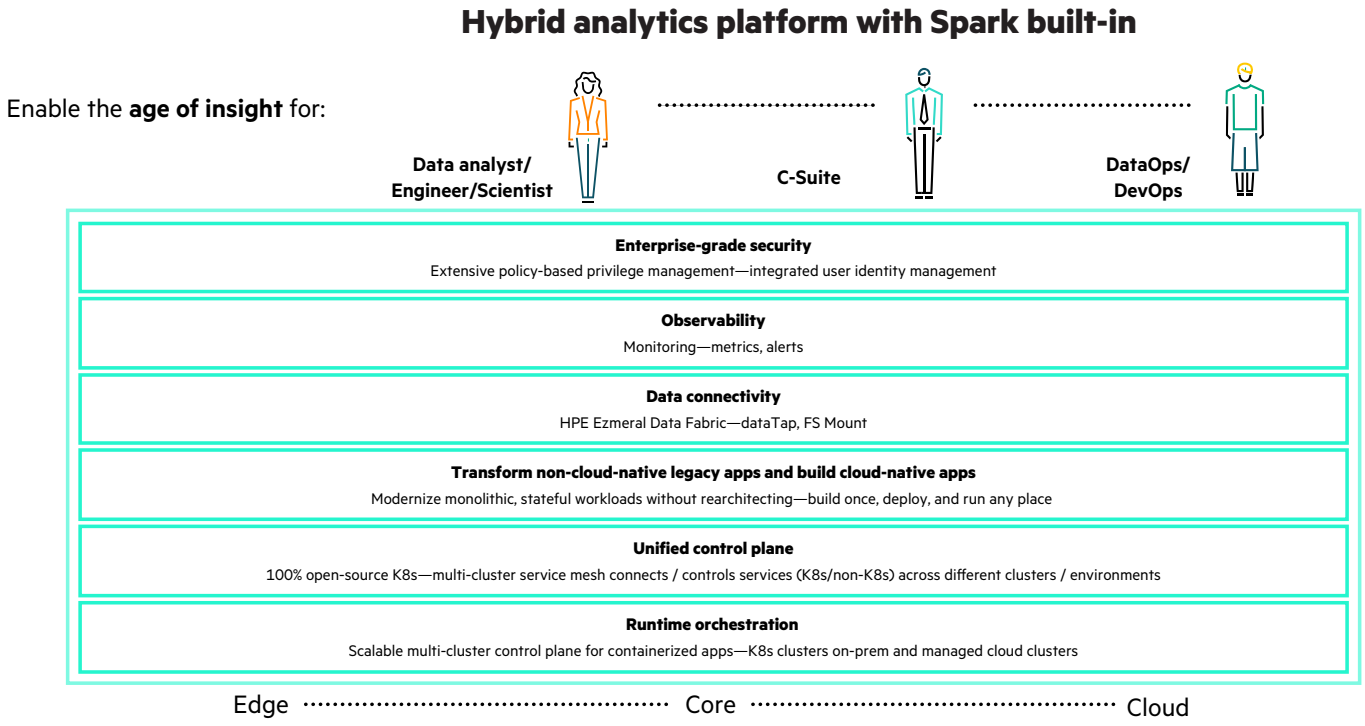

Figure 1. Deploy modern applications using open-source Kubernetes from edge to cloud

HPE Ezmeral and Apache Spark benefits:

- Elastic platform that scales Apache Spark workloads
- Zero trust security—Secure Production Identity Framework for Everyone (SPIFFE), SPIFFE Runtime Environment (SPIRE)
- Any cloud (Azure, HPE GreenLake, Google™, or AWS)
- HPE Ezmeral Data Fabric with single global namespace
- HPE Ezmeral marketplace-validated solutions for Apache Spark
- Cloud Native Computing Foundation (CNCF)-certified open-source Kubernetes
- Multi-tenant control plane
- GPU acceleration
- Full isolation and control
- Prebuilt data analytic tools

Often the strategy and use case for Apache Spark is defined by the implementation. Here is an outline of the key personas and use cases for Apache Spark and the framework that provides the focus of this document.

Common use cases

- Parallel processing data sets from the small to the very large, often across a cluster of compute resources
- Ad hoc and or interactive query sets to visualize data sets
- Model construction and training for ML and artificial intelligence (AI)
- Data pipeline development and implementations
- Data analysis and data transformation

Common personas

- **Data engineers:** Background in software engineering, methodologies for building scalable data pipelines, along with familiarity with NoSQL, relational database management system (RDMS), data sources, and more
- **Data scientist:** Familiar with SQL, NumPy, Pandas, R, and Python, along with the ability to develop classification, regression, and algorithms to build models enabling the building, testing of hypotheses

As data analytics in general becomes more mainstream, we start to see more traditional roles leveraging the power of Apache Spark.

- **Data architects:** Like data engineers but more focused on how data is used and linked across a business and/or business units; deep understanding of specific database technologies and the means to access that information
- **Database manager:** Not a common role for Apache Spark, with the freedom and flexibility to choose whether to deploy across on-premises or multicloud and keep the data coordinated on the cloud, on-premises, and at the edge; this capability brings Apache Spark and data management together
- **Data/Application developers:** Analytics has driven the use of massive scaled up compute and storage clusters to work on complex problems

HPE Ezmeral Unified Analytics

HPE Ezmeral Unified Analytics is the industry's first unified, modern, hybrid analytics, and data lakehouse platform.[1] Enterprises looking to embrace a digital-first strategy and unlock data's value have been limited to hyperscale environments that often require code refactoring, surprise costs, and forced data migration. HPE Ezmeral Unified Analytics uses open-source software to help ensure as-needed data portability and enhancement of on-premises, hybrid, and cloud deployments.

The advantages include:

- **No vendor lock-in:** Apache Spark workloads offer the freedom to choose deployment environments, tools, and partners needed to innovate faster.
- **Performance acceleration:** Using industry-standard analytics and Big Data benchmarks, HPE completed testing the use of NVIDIA® RAPIDS Accelerator for Apache Spark. The results of this test indicate that HPE Ezmeral accelerates Spark analytic workloads by 29x.[2]

Figure 2 CPU/GPU testing for eight CPUs and one NVIDIA GPU.

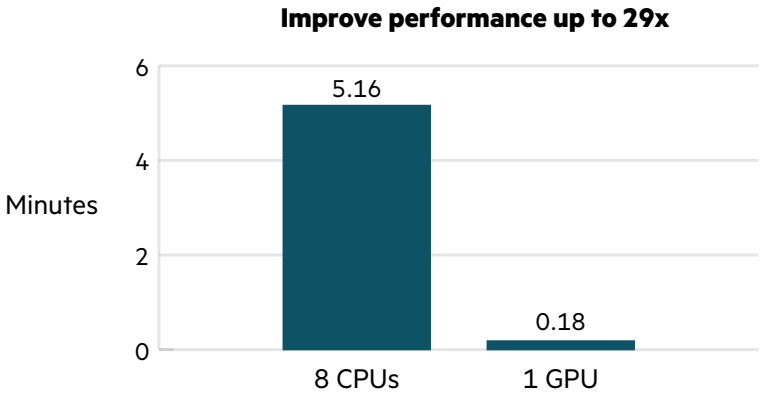

Figure 2. ETL performance with GPU

- **Next-generation architecture:** Multi-tenant Kubernetes environment supports a compute-storage separation cloud model, providing the elasticity and performance required for advanced analytics, enabling users to create unified real-time and batch analytics with Delta Lake integration and data lakehouses.
- **Enhanced collaboration:** Integrated workflows from analytics to ML/AI span hybrid clouds and edge locations, including native open-source integrations with Airflow, MLflow, and Kubeflow technologies to help data science, data engineering, and data analytics teams collaborate and deploy models faster.
- **Enriched for data analytics:** Enterprises can create a unified data repository for use by data scientists, developers, and analysts, including usage and sharing controls, creating the foundation for a silo-free digital transformation that scales with the business as it grows, and reaches new data sources.

[1] Based on HPE internal analysis, September 2021

[2] HPE standardized the testing models based on industry-standard analytic and Big Data benchmarks. This benchmark was based on leveraging Kubernetes Pods and storage managed by HPE Ezmeral, NVIDIA A100 40 GB GPUs, and HPE ProLiant DL385 servers. In July 2021, all tests were conducted in a laboratory setting to simulate real-world conditions. User experiences may differ from those in our test scenarios.

Data analytics ecosystems often address data locality or the idea of data gravity. HPE Ezmeral unifies the Inmon and Kimball[3] data architectures paradigm. The HPE Ezmeral Runtime ecosystem enables customers to choose the best model based on their different perspectives and different solutions.

Apache Spark on HPE Ezmeral Runtime gives the Inmon and Kimball followers equal opportunity to develop their Big Data, analytics, and workflow models based on the deterministic model that works best.

Apache Spark

Its strength lies in the ability to complete work in-memory—keeping the data as close to the CPUs as possible. While HPE Ezmeral provides access to the data regardless of its location, Apache leverages the power of high-performance computing (HPC) technology and supports efficient use of resources by distributing work to additional nodes.

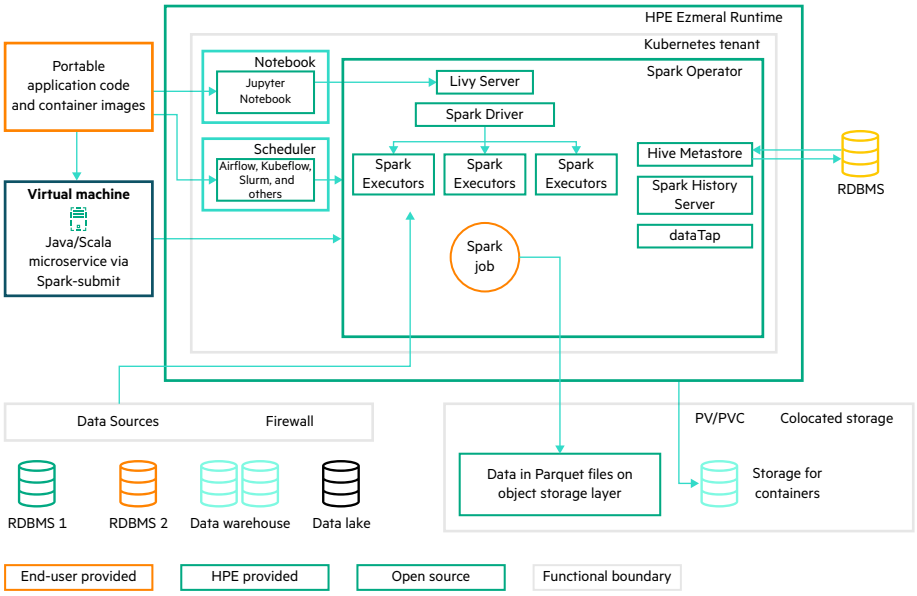

Figure 3. HPE Ezmeral Runtime and Spark architecture

With Apache Spark powered by HPE Ezmeral Runtime, you are enabled to build data solutions that can scale to meet any demand. What's more, shared storage solutions can be created with HPE Ezmeral Data Fabric, MinIO, NFS, and other third-party providers and made available to tenants within HPE Ezmeral Runtime.

Data transformation

Spark truly shines when it comes to data transformation. Before data scientists can build models, they need data engineers to perform many complex extract, transform, and load (ETL)/extract, load, and transform (ELT) processes on the data. Building from the same framework as Inmon and Kimball, we often need to work with raw data that is available in many formats and distill it into a usable format. This process in the past has been one of single-threaded workstreams, taking hours to process and stage the data for the next phase.

[3] Inmon vs. Kimball zentut.com/data-warehouse/kimball-and-inmon-data-warehouse-architectures/

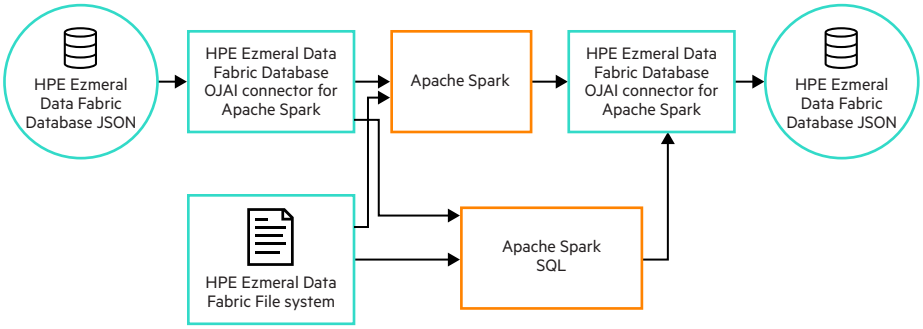

Figure 4. Apache Spark ETL

Spark, with its ability to work with native SQL database structures, traditional RDBMS or NoSQL platforms, unstructured data, and such makes it ideal for accelerating this arduous data transformation phase. Spark can work with multiple sources of data in parallel greatly increasing overall data refresh and or cycle times.

By incorporating streaming services such as Kafka and workflow tools such as Delta Lake, Airflow, and Kubeflow much of the ETL processes can be automated. Providing in some instances, near-real-time or real-time updates.

Machine learning/ML Ops

The HPE Ezmeral ML Ops solution supports every stage of the ML lifecycle—data preparation, model building, model training, model deployment, collaboration, and monitoring. HPE Ezmeral ML Ops is an end-to-end data science solution with the flexibility to run on-premises, in multiple public clouds, or in a hybrid model and respond to dynamic business requirements in a variety of use cases.

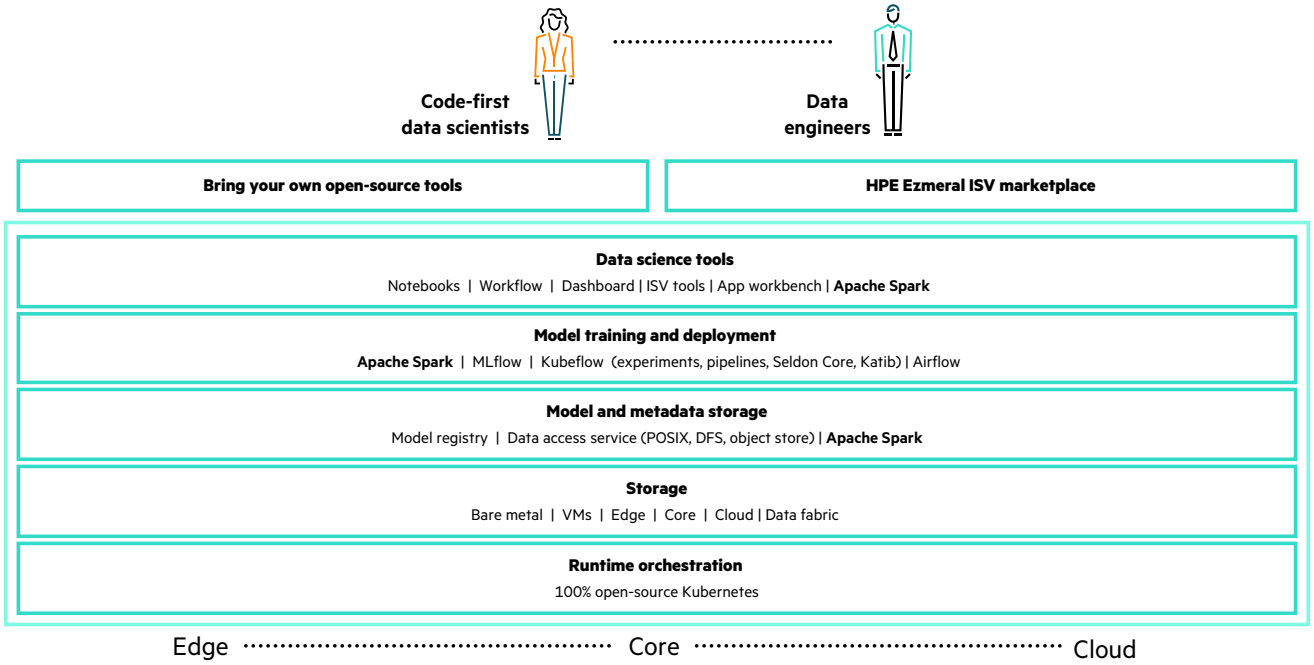

Figure 5. ML/ML Ops architecture within HPE Ezmeral

Figure 5 shows the various phases where Spark is utilized to facilitate ML and ML operations (ML Ops) projects. HPE Ezmeral provides curated solutions for ML Ops.

Performance
To provide some relevant information in terms of NVIDIA GPU performance, a benchmark was defined using TPC-DS. The in-lab testing conducted in July 2021, indicates it was possible to demonstrate the power of NVIDIA GPU accelerated Spark with HPE Ezmeral. For example, when looking at the complex analytics queries, we were able to showcase a 10–29x performance improvement when comparing an 8 CPU cluster to a single accelerated GPU cluster. The use of TPC-DS framework had been to standardize the test results and highlight the simplicity of building a solution with HPE Ezmeral and adding in the GPU capability.

- Spark 3.1.1 with NVIDIA RAPIDS
- NVIDIA Tesla and A100 40 GB
- HPE ProLiant DL385 Gen10
- Red Hat® Enterprise Linux® 7
- TPC-DS
- RAPIDS Accelerator for Apache Spark
- Storage provided by HPE Ezmeral tenant storage
- Query performance: AMD CPU versus GPU performance that runs specific queries based on the TPC-DS benchmark

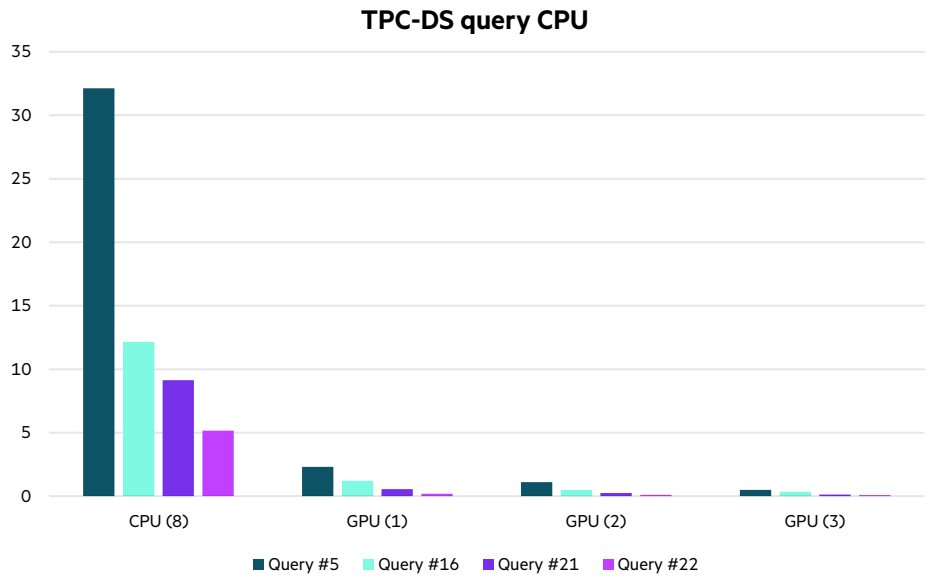

Figure 6. TPC-DS query performance

As working nodes come online Apache Spark can dynamically adjust (dynamic scaling) to include additional GPUs resource to improve performance.

The elasticity of HPE Ezmeral Unified Analytics provides the tools necessary to build dynamic scalable solutions on-premises, at the edge, and in the cloud.

Summary
Hewlett Packard Enterprise continues to disrupt the data analytics market and the latest editions continue that trend with HPE Ezmeral Runtime and HPE Ezmeral Data Fabric Object Store.

Incorporating technology partners such as NVIDIA and their accelerator technology further enhances the HPE Ezmeral solution base and provides opportunity for customers to get the most from their Apache Spark solutions.

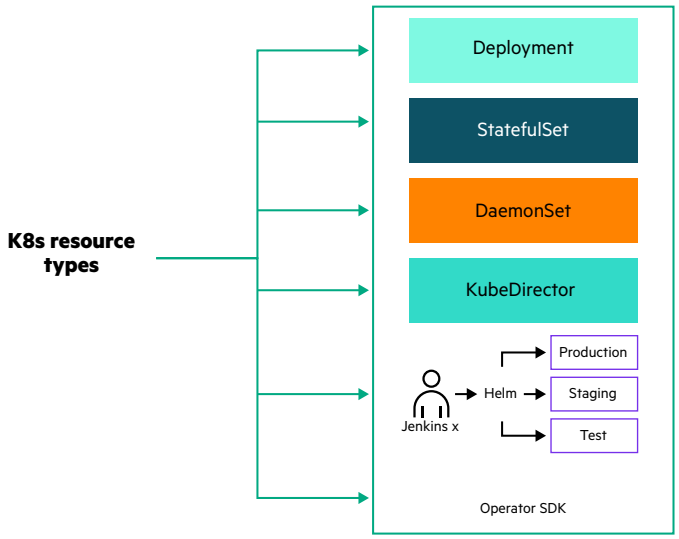

Figure 7. HPE Ezmeral application framework

Several key features include:

- Enhanced performance for analytics
  - GPU acceleration—NVIDIA RAPIDS Accelerator for Apache Spark, Multi-Instance GPU support
  - Scalable object stores—supports file, streams, database, object within a common persistent store
  - Designed for high-performance across edge-to-cloud analytics workloads
- Globally synchronized edge-to-cloud data
  - Clusters and data are orchestrated together to support dispersed edge operations
  - Single global namespace provides simplified access to edge-to-cloud topologies from any application or interface
- Near-continuous scaling—enterprises can grow as needed by adding nodes and configuring policies
- Performance and cost balance—adapting to small or large objects, auto-tiering policies automatically move data from high-performance storage to low-cost storage

There are many platforms to build your data analytics solutions, Apache Spark and HPE Ezmeral provide a single platform that provides the elasticity to grow with your business and provide the same user experience you and your teams are more comfortable using.

Learn more at

hpe.com/us/en/software

Visit **HPE GreenLake**

Find more information about the HPE Ezmeral portfolio

HPE Ezmeral Runtime Enterprise

Request a demo

HPE Ezmeral software

HPE Ezmeral Runtime Enterprise online documentation

HPE Ezmeral Data Fabric online documentation

HPE Ezmeral Runtime Enterprise software architecture overview—white paper

HPE Developer—HPE Ezmeral Runtime Enterprise website

HPE Developer—HPE Ezmeral Data Fabric website

Learning portal

Make the right purchase decision. Contact our presales specialists.

Chat now (sales)

Call now

Get updates

Hewlett Packard Enterprise

a00118439ENW, Rev. 1