Enterprise Strategy Group | **Getting to the bigger truth.™**

Technical Validation

Big Data Analytics Applications Using HPE Ezmeral Unified Analytics with NVIDIA RAPIDS Accelerator

Optimizing Performance and Management of Apache Spark Containerized Workloads

By Alex Arcilla, Senior Validation Analyst

December 2021

This ESG Technical Validation was commissioned by HPE and is distributed under license from ESG.

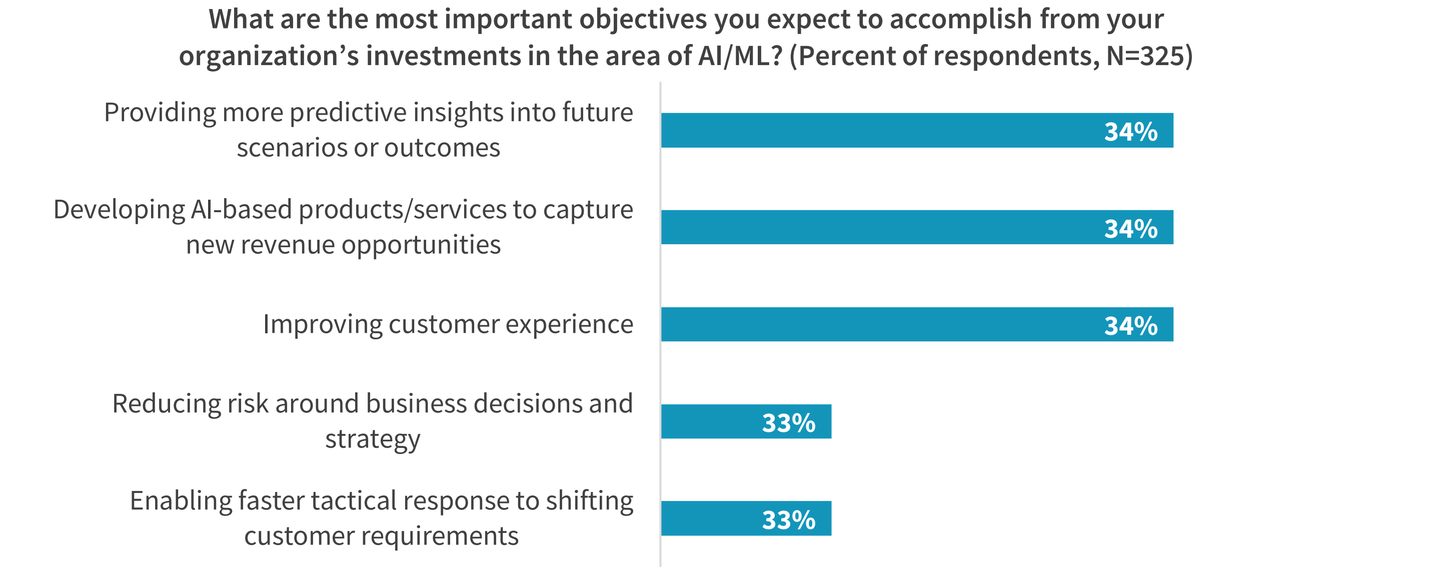
Introduction

This ESG Technical Validation documents our evaluation on HPE Unified Analytics. We focus on how the combination of HPE Ezmeral Runtime Enterprise and NVIDIA RAPIDS Accelerator can accelerate Apache Spark AI/ML workloads running on NVIDIA GPU-enabled containers. We also reviewed how organizations can use HPE Unified Analytics to simplify management and orchestration of multiple containerized applications simultaneously.

Background

According to ESG research, 34% of survey respondents cited providing more predictive insights into future scenarios or outcomes as among their most important objectives when investing in AI/ML initiatives, while another 34% cited developing AI-based products/services to capture new revenue opportunities (see Figure 1).[1] However, challenges remain in implementing an infrastructure that will achieve these objectives. The same ESG research revealed that resource sharing is the most-cited part of the infrastructure stack that respondents believe is one of the weakest links in delivering an effective AI environment.[2]

Figure 1. Most Important Objectives to Accomplish with AI/ML Investments



Source: Enterprise Strategy Group

Current AI/ML initiatives have been implemented differently across different business units, either on-premises or in the public cloud, using a wide variety of on-premises infrastructure, public cloud infrastructure and services, open-source tools, container engines, and libraries. This has resulted in multiple and disjointed AI/ML environments deployed across the organization. However, multiple data silos and compute resources remain separate, limiting how they can be fully utilized across the organization for maximum business advantage.

What if there was a solution that can maximize the usage of AI/ML compute and storage resources across the organization, whether those resources are located on-premises or in the public cloud, without sacrificing the performance required to gain insight and meet business needs?

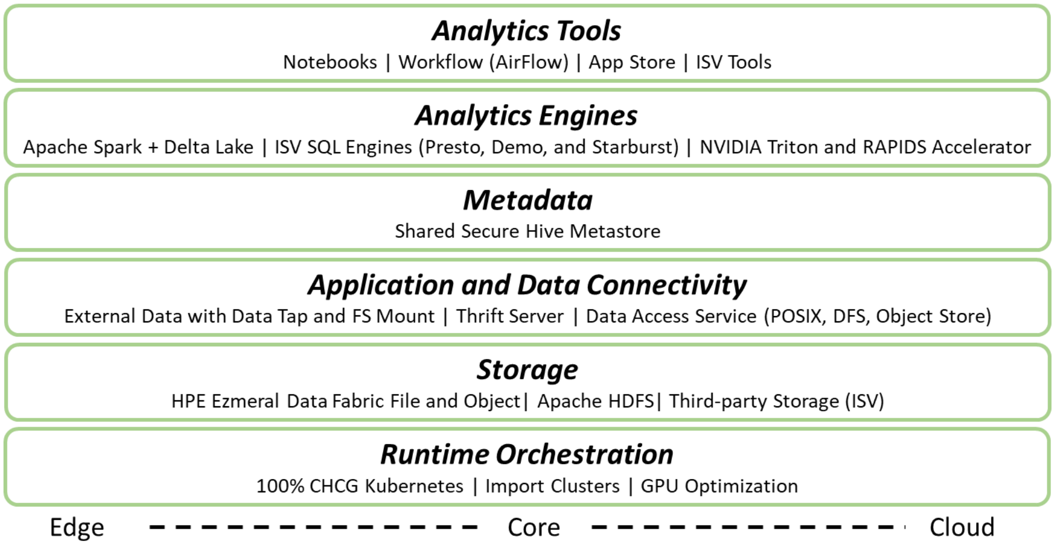
[1] Source: ESG Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021.
[2] Ibid.

HPE Ezmeral Unified Analytics with NVIDIA RAPIDS Accelerator

HPE Ezmeral Unified Analytics, available in January 2022, has been designed to deliver a unified data lake that connects distributed data stores, public cloud storage, files, and object stores, regardless of where they reside. The foundation of HPE Ezmeral Unified Analytics is HPE Ezmeral Runtime Enterprise, which delivers containerized orchestration of both cloud-native and non-cloud applications with a persistent data store. HPE Ezmeral Runtime Enterprise is designed to help organizations simplify the development, orchestration, and scaling of Kubernetes-based clusters deployed across hybrid environments.

The platform can help data scientists, data engineers, and data analysts to support multiple enterprise-grade containerized AI/ML applications via the following capabilities (see Figure 2):

Figure 2. HPE Unified Analytics


Analytics Tools
Notebooks | Workflow (AirFlow) | App Store | ISV Tools
Analytics Engines
Apache Spark + Delta Lake | ISV SQL Engines (Presto, Demo, and Starburst) | NVIDIA Triton and RAPIDS Accelerator
Metadata
Shared Secure Hive Metastore
Application and Data Connectivity
External Data with Data Tap and FS Mount | Thrift Server | Data Access Service (POSIX, DFS, Object Store)
Storage
HPE Ezmeral Data Fabric File and Object| Apache HDFS| Third-party Storage (ISV)
Runtime Orchestration
100% CHCG Kubernetes | Import Clusters | GPU Optimization
Edge
Core
Cloud

Source: Enterprise Strategy Group

- Multi-tenancy: With the platform's control plane, organizations can spin up and orchestrate multiple standards-based Kubernetes clusters using physical or virtual compute (CPU or GPU) and storage resources located on-premises or in the public cloud. Organizations can also automate orchestration of workflows, allowing many experiments to be scheduled and run "hands free." Thus, data science and development teams can share the same infrastructure stack with minimal impact on system performance.

- Unified data lake: The single global namespace provided by HPE Ezmeral Data Fabric provides a consolidated view into files that can reside in separate clusters across on-premises, cloud, and edge environments, thus enabling end-users and their applications to access remote data as if it is locally stored.

- Open platform/ecosystem: HPE Ezmeral Unified Analytics is built on an open source foundation that enables data analytics teams to deploy Kubernetes-based workloads into any AI/ML environment, comprising a wide range of open source tools, container engines, and libraries. NVIDIA Triton, RAPIDS Accelerator, and Apache Spark are included in the platform's ISV marketplace (see Figure 10 in the Appendix for details).

- Security: Organizations can integrate HPE Ezmeral Runtime Enterprise with enterprise LDAP and AD authentication systems to secure access to jobs, data, or clusters based on groups and/or roles. Access to the multiple applications supported by the platform can be restricted according to business needs. Risk and compliance tenant structure allows for granular security structures.

To simplify platform use, multiple versions of Kubernetes are supported, either to accommodate differing levels of expertise throughout the organization or to maintain current procedures and toolsets without disrupting business operations. End-users can use either API commands or a graphical web UI with this self-service platform, removing the complexity typically faced when using Kubernetes.

For ML initiatives, organizations have typically relied on GPU-based infrastructure to accelerate processing time, given the large data sets required to obtain a useful ML model. Organizations that employ Kubernetes clusters based on NVIDIA's A100 Tensor Core GPUs can leverage NVIDIA RAPIDS Accelerator for Apache Spark. While Apache Spark is designed to increase data processing performance via in-memory processing, NVIDIA RAPIDS Accelerator for Apache Spark can maximize GPU performance. RAPIDS is a suite of open-source software libraries and APIs for executing data science pipelines entirely on GPUs. The software has been designed to accelerate data preparation, feature extraction, training, and inference with model development and training.

ESG Technical Validation

ESG evaluated how HPE Ezmeral Runtime Enterprise and NVIDIA RAPID's Accelerator can accelerate the performance of Apache Spark workloads. We also observed the simplicity of Kubernetes cluster creation and the ability to access data regardless of location. Testing was performed via remote sessions.

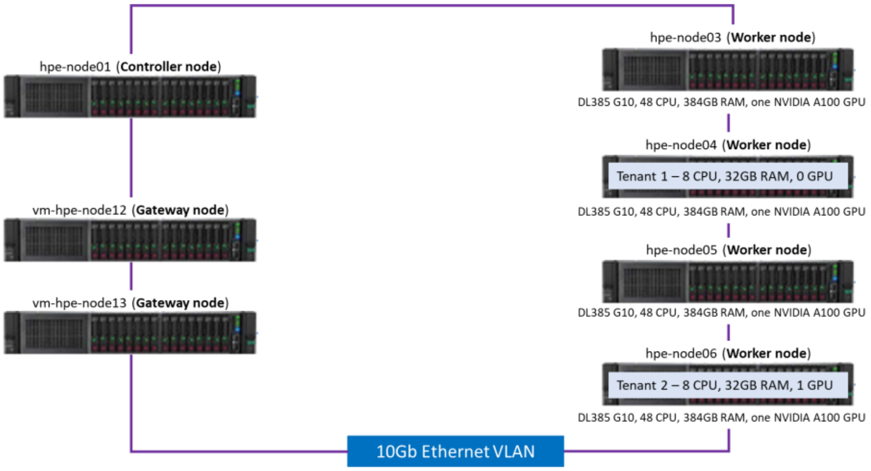
Performance

Apache Spark-based ML workloads require fast processing and response times to accommodate the very large data volumes required to create and train ML models. The combination of HPE Ezmeral Unified Analytics, NVIDIA RAPIDS Accelerator, and NVIDIA A100 GPU-based clusters can accelerate Apache Spark workloads, which can benefit applications such as fraud detection or financial transaction processing.

ESG Testing

The following test bed was used (shown in Figure 3). Six HPE ProLiant servers (one master node, one controller node, and four worker nodes) were used. Each worker node was equipped with one AMD EPIC 7352 24-core CPU and one NVIDIA A100 40G GPU. The master and controller node were equipped with two AMD EPYC 7702 64-core CPUs. All nodes were interconnected by a 10Gb Ethernet VLAN configured on a Cisco Nexus Model 9000 switch.

Figure 3. Test Bed for Apache Spark Workload Performance Testing

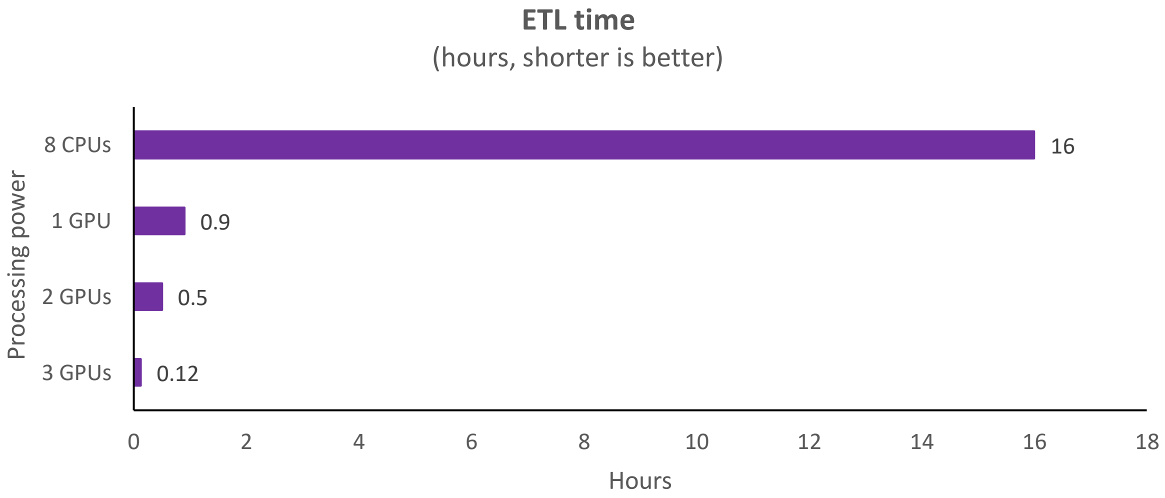

hpe-node01 (Controller node)
vm-hpe-node12 (Gateway node)
vm-hpe-node13 (Gateway node)
hpe-node03 (Worker node)
DL385 G10, 48 CPU, 384GB RAM, one NVIDIA A100 GPU
hpe-node04 (Worker node)
Tenant 1 – 8 CPU, 32GB RAM, 0 GPU
DL385 G10, 48 CPU, 384GB RAM, one NVIDIA A100 GPU
hpe-node05 (Worker node)
DL385 G10, 48 CPU, 384GB RAM, one NVIDIA A100 GPU
hpe-node06 (Worker node)
Tenant 2 – 8 CPU, 32GB RAM, 1 GPU
DL385 G10, 48 CPU, 384GB RAM, one NVIDIA A100 GPU
10Gb Ethernet VLAN

Source: Enterprise Strategy Group

Clusters were created with HPE Ezmeral Runtime Enterprise and Kubernetes v1.20. Container size for CPU-based workloads was limited to eight CPUs and 32GB RAM and one container with 48 CPUs, 384 GB of RAM, and one NVIDIA A100 GPU. Two separate namespaces were configured for the CPU and GPU-based clusters, respectively. All jobs were automatically submitted via Apache Spark Submit. Both extract-transform-load (ETL) and query response times were measured using a 5TB data set.

ESG first reviewed the test results that measured ETL processing time, covering the creation of data sets and tables, as well as conversion of file format to Apache Parquet. ETL times were recorded when processing data sets on one container (equivalent to eight CPUs) were measured against another container with GPUs scaling from one, two, and three NVIDIA A100 GPUs respectively. The NVIDIA RAPIDS Accelerator ran when using GPU-based containers. Results are shown in Figure 4.

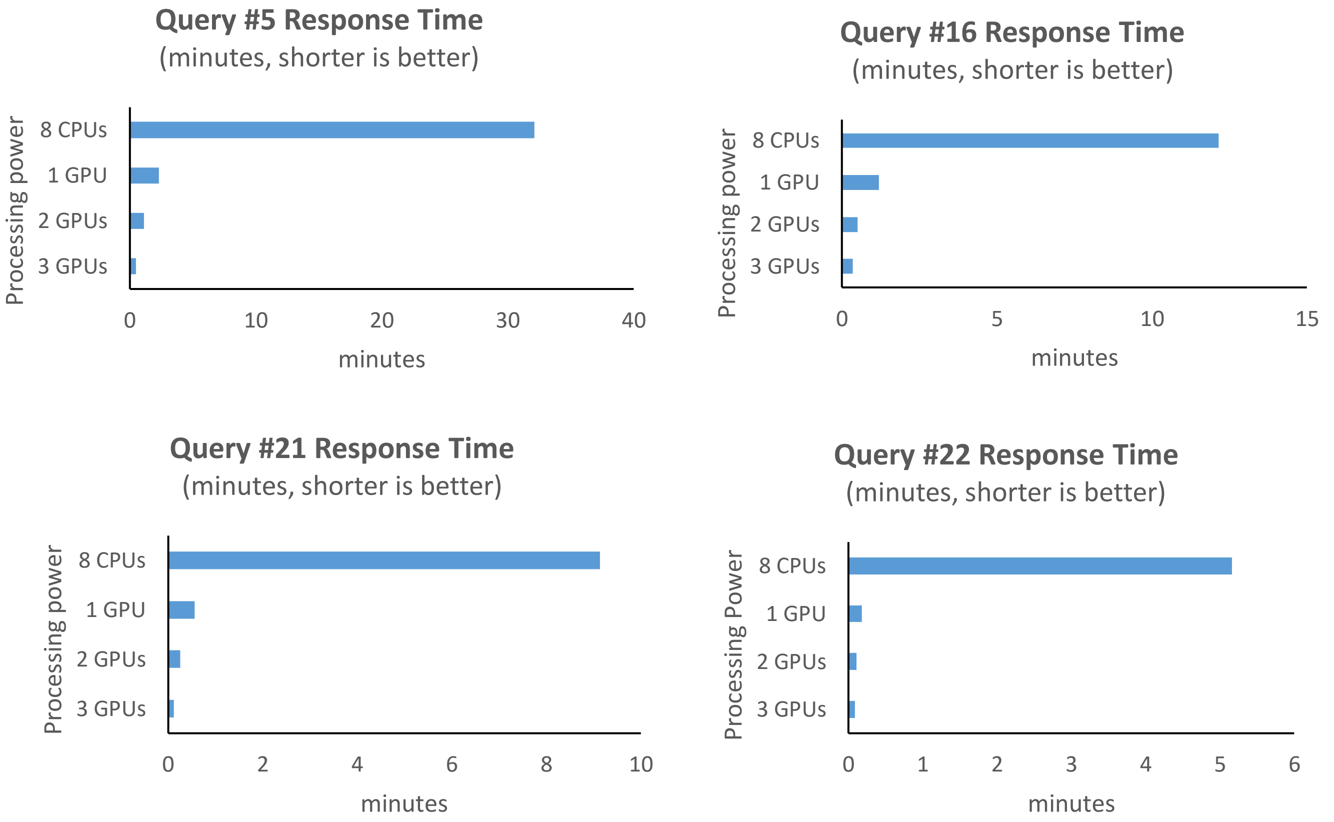
Figure 4. ETL Times with Apache Spark and NVIDIA RAPIDS Accelerator on HPE Ezmeral Runtime



Source: Enterprise Strategy Group

ESG then reviewed the results of query processing time. We selected four jobs—Queries #5, #16, #21, and #22—to illustrate performance. Each job ran on four Kubernetes clusters with container sizes equivalent to eight CPUs and one, two, and three GPUs, respectively. Again, the NVIDIA RAPIDS Accelerator ran when using GPU-based clusters. We compared query response times, shown in Figure 5 and Table 1.

Figure 5. Query Response Times with Apache Spark and NVIDIA RAPIDS Accelerator on HPE Ezmeral Runtime



Source: Enterprise Strategy Group

Table 1. Query Processing Times Using Apache Spark and NVIDIA RAPIDS Accelerator

	Query #5 Query response time (min)	Query #16 Query Response Time (min)	Query #21 Query Response Time (min)	Query #22 Query Response Time (min)
8 CPUs	32.13	2.31	1.1	0.5
1 GPU	12.15	1.2	0.5	0.35
2 GPUs	9.13	0.56	0.25	0.12
3 GPUs	5.16	0.18	0.11	0.09

Source: Enterprise Strategy Group

What the Numbers Mean

- ESG confirmed that HPE Ezmeral Runtime Enterprise with NVIDIA RAPIDS Accelerator, and NVIDIA A100 GPU-enabled Kubernetes clusters can improve performance of Apache Spark ML workloads by up to 29x compared to running similar workloads on an eight CPU-enabled cluster. (Results will vary in real-world production environments.)

- ESG also found that when leveraging the NVIDIA RAPIDS Accelerator, query processing times on a single GPU-based cluster decreased between 90% and 97%. Response times decreased as more GPUs were added to the cluster.

- Optimizing Apache Spark workload performance translates into minimizing both ETL and query response times. This is critical to helping organizations quickly process large data sets to deliver business insights quickly.

Why This Matters

While organizations embracing containerized applications find it important to simplify their deployment and management, application performance must be maintained. This is especially critical for those wanting to use Kubernetes clusters to support big data and AI/ML applications and workloads, which require high-performance and large-scale data processing.

ESG confirmed that processing of Apache Spark workloads was accelerated with the addition of HPE Ezmeral Runtime Enterprise, NVIDIA RAPIDS Accelerator and NVIDIA A100 GPUs. Customers can experience even greater acceleration when NVIDIA A100 GPU capacity is increased. We saw how the combination of HPE Ezmeral, NVIDIA, and Apache Spark allows enterprise customers to get insights faster for more data-driven business decisions.

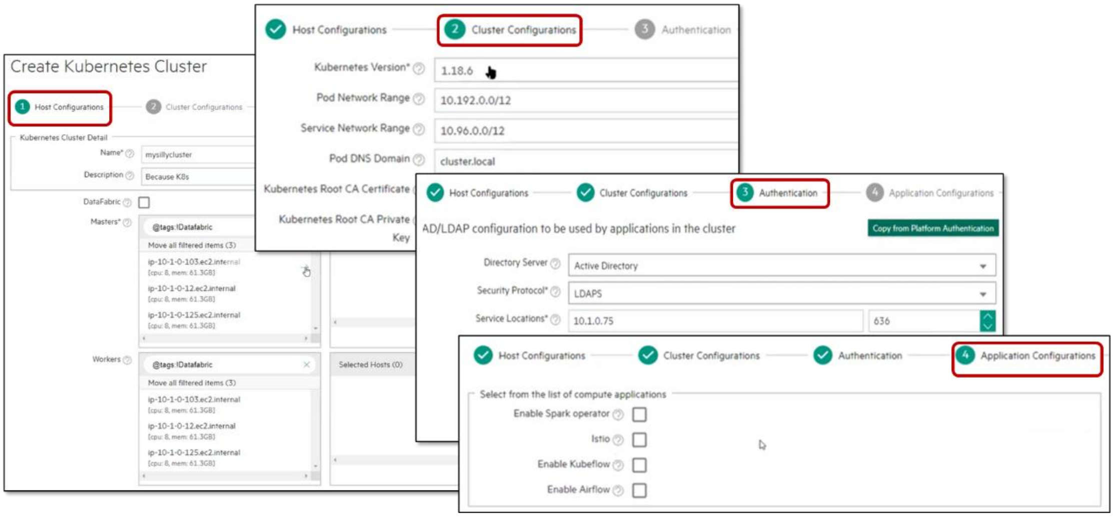
Simple to Use

With HPE Ezmeral Runtime Enterprise, organizations can create and manage multiple Kubernetes clusters and containerized applications within a single web UI, without the need to purchase or configure the underlying infrastructure and supporting toolsets.

ESG Testing

ESG began by observing the creation of a Kubernetes cluster using a wizard-driven process (see Figure 6). For this cluster, we used Amazon Web Services (AWS) Elastics Compute Cloud (EC2) instances as hosts. We navigated to the HPE Ezmeral Runtime Enterprise web UI; selected the virtual hosts for the cluster; assigned AWS EC2 instances as master and worker nodes; selected the Kubernetes version to run, IP network address ranges, LDAP and AD settings; and added other services, such as Apache Spark Operator.

Figure 6. Creating a Kubernetes Cluster with HPE Ezmeral Container Platform Web UI



Create Kubernetes Cluster

Kubernetes Cluster Detail

Name*: mysillycluster

Description: Because K8s

DataFabric: □

Masters*: @tags:!Datafabric

Move all filtered items (3)

ip-10-1-0-103.ec2.internal [cpu: 8, mem: 61.3GB]
ip-10-1-0-12.ec2.internal [cpu: 8, mem: 61.3GB]
ip-10-1-0-125.ec2.internal [cpu: 8, mem: 61.3GB]

Workers: @tags:!Datafabric

Move all filtered items (3)

ip-10-1-0-103.ec2.internal [cpu: 8, mem: 61.3GB]
ip-10-1-0-12.ec2.internal [cpu: 8, mem: 61.3GB]
ip-10-1-0-125.ec2.internal [cpu: 8, mem: 61.3GB]

Selected Hosts (0)

Kubernetes Version*: 1.18.6

Pod Network Range: 10.192.0.0/12

Service Network Range: 10.96.0.0/12

Pod DNS Domain: cluster.local

Kubernetes Root CA Certificate

Kubernetes Root CA Private Key

AD/LDAP configuration to be used by applications in the cluster

Directory Server: Active Directory

Security Protocol*: LDAPS

Service Locations*: 10.1.0.75 636

Select from the list of compute applications

Enable Spark operator □

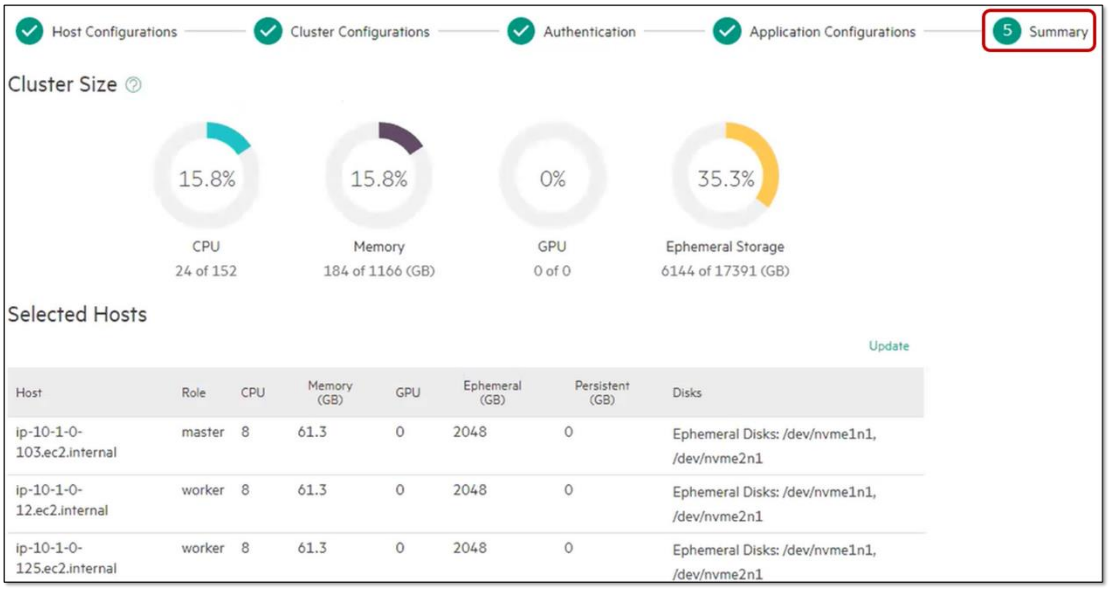
Istio □

Enable Kubeflow □

Enable Airflow □

Once we completed this process, we viewed the details of the Kubernetes Cluster in a Summary screen (see Figure 7). The summary detailed the amount of compute and storage capacity available to any application running on this cluster and noted the specific physical and virtual hosts consumed by this cluster.

Figure 7. Summary Details of Kubernetes Cluster



Cluster Size

CPU 15.8% 24 of 152

Memory 15.8% 184 of 1166 (GB)

GPU 0% 0 of 0

Ephemeral Storage 35.3% 6144 of 17391 (GB)

Selected Hosts

Host	Role	CPU	Memory (GB)	GPU	Ephemeral (GB)	Persistent (GB)	Disks
ip-10-1-0-103.ec2.internal	master	8	61.3	0	2048	0	Ephemeral Disks: /dev/nvme1n1, /dev/nvme2n1
ip-10-1-0-12.ec2.internal	worker	8	61.3	0	2048	0	Ephemeral Disks: /dev/nvme1n1, /dev/nvme2n1
ip-10-1-0-125.ec2.internal	worker	8	61.3	0	2048	0	Ephemeral Disks: /dev/nvme1n1, /dev/nvme2n1

ESG observed that this process completed in minutes, resulting in significant savings in both time and effort. Creating a Kubernetes cluster on premises involves many steps. Hardware needs to be purchased and configured with the proper compute and storage capacity to support gateway, controller, master, and worker nodes. Networking and security requirements need to be addressed. All open source software, especially Kubernetes and supporting tools, need to be downloaded from code repositories. Furthermore, additional scripts need to be written to customize the software to satisfy specific business needs. The net impact of all these steps is that time, effort, and cost are wasted, and time to value is long.

Deploying containerized applications on Kubernetes clusters may be simpler when working with public cloud providers. However, considering the cost of virtual compute and storage resources, services required for containerization and cluster management, this may not be a cost-effective alternative for many organizations. Also, time and effort can be excessive as organizations need to coordinate tasks between multiple interfaces, regardless of the public cloud provider.

Why This Matters

ESG validated that HPE Ezmeral Runtime Enterprise reduces the amount of time to operationalize Kubernetes clusters with wizards and automated prechecks to deliver business outcomes faster.

We observed the ease and simplicity of creating a cluster, removing the multiple tasks and complexity of setting up the infrastructure on premises or using multiple public cloud-based services concurrently.

Access Data Regardless of Data Location

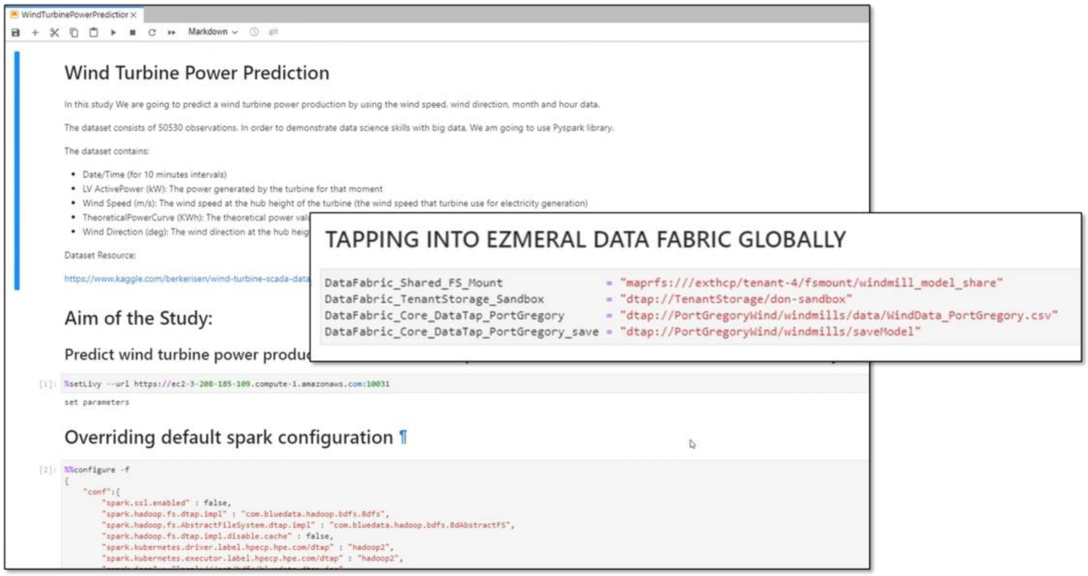
Application performance partly depends upon how quickly data can be accessed. However, replicating and transmitting data to local storage may not always be an option. For example, organizations want to analyze data from IoT deployments in real time. That cannot happen when data is being gathered at the network edge while the analytics application is located within a data center. Applications placed into containers by HPE Ezmeral Runtime Enterprise can ingest any data type or source, regardless of the data's physical location.

ESG Testing

ESG reviewed the deployment of a Kubernetes cluster using a Jupyter notebook and HPE Ezmeral Runtime Enterprise. As shown in Figure 8, this application was designed to predict power output from wind turbines in Australia. The application used Apache Spark for data processing and accessed data via a secure connection to HPE Ezmeral Data Fabric, enabling end-users to ingest data from remote IoT or cloud deployments.

We observed that a few lines of declarative script that directed path names to data sources were all that was required for this application to access the IoT data. The data fabric was also deployed as a Kubernetes cluster, allowing it to transparently scale as data is ingested.

Figure 8. Jupyter Notebook Application Accessing IoT Data



Wind Turbine Power Prediction

In this study We are going to predict a wind turbine power production by using the wind speed, wind direction, month and hour data.

The dataset consists of 50530 observations. In order to demonstrate data science skills with big data, We am going to use Pyspark library.

The dataset contains:

- Date/Time (for 10 minutes intervals)
- LV ActivePower (kW): The power generated by the turbine for that moment
- Wind Speed (m/s): The wind speed at the hub height of the turbine (the wind speed that turbine use for electricity generation)
- TheoreticalPowerCurve (KWh): The theoretical power val
- Wind Direction (deg): The wind direction at the hub heig

Dataset Resource:

https://www.kaggle.com/berkerisen/wind-turbine-scada-data

Aim of the Study:

Predict wind turbine power produc

```
[1]: %setLivy --url https://ec2-3-208-185-109.compute-1.amazonaws.com:10031
set parameters
```

Overriding default spark configuration ¶

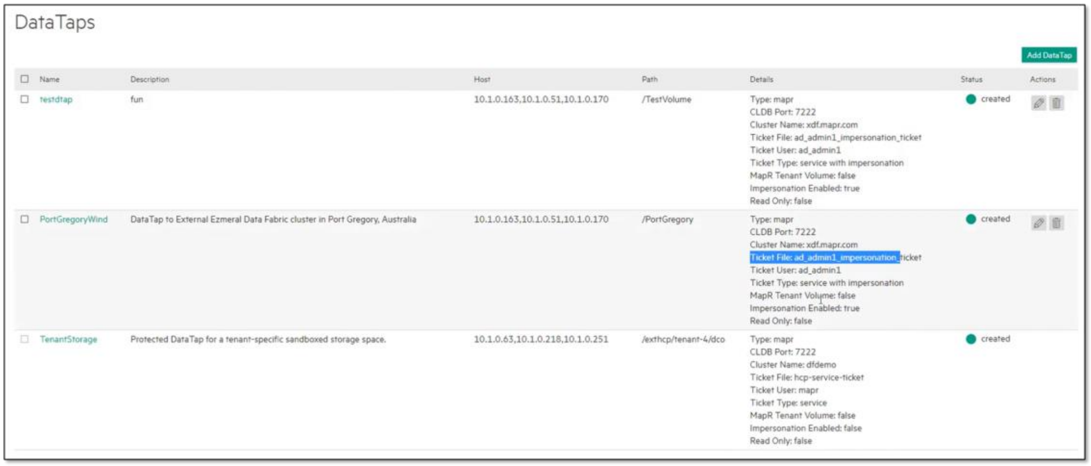
```
[2]: %%configure -f
     {
         "conf":{
             "spark.ssl.enabled" : false,
             "spark.hadoop.fs.dtap.impl" : "com.bluedata.hadoop.bdfs.Bdfs",
             "spark.hadoop.fs.AbstractFileSystem.dtap.impl" : "com.bluedata.hadoop.bdfs.BdAbstractFS",
             "spark.hadoop.fs.dtap.impl.disable.cache" : false,
             "spark.kubernetes.driver.label.hpecp.hpe.com/dtap" : "hadoop2",
             "spark.kubernetes.executor.label.hpecp.hpe.com/dtap" : "hadoop2",
```

TAPPING INTO EZMERAL DATA FABRIC GLOBALLY

```
DataFabric_Shared_FS_Mount              = "maprfs:///exthcp/tenant-4/fsmount/windmill_model_share"
DataFabric_TenantStorage_Sandbox        = "dtap://TenantStorage/don-sandbox"
DataFabric_Core_DataTap_PortGregory     = "dtap://PortGregoryWind/windmills/data/WindData_PortGregory.csv"
DataFabric_Core_DataTap_PortGregory_save = "dtap://PortGregoryWind/windmills/saveModel"
```

Creating secure connections to IoT data was also simple, as shown in Figure 9. Using another wizard-driven process, we observed the simplicity of creating secure connections to any data source by inputting key information such as the IP addresses of the hosts collecting the data.

Figure 9. Details of Embedded Connections to IoT Data



DataTaps

☐	Name	Description	Host	Path	Details	Status	Actions
☐	testdtap	fun	10.1.0.163,10.1.0.51,10.1.0.170	/TestVolume	Type: mapr CLDB Port: 7222 Cluster Name: xdf.mapr.com Ticket File: ad_admin1_impersonation_ticket Ticket User: ad_admin1 Ticket Type: service with impersonation MapR Tenant Volume: false Impersonation Enabled: true Read Only: false	● created	
☐	PortGregoryWind	DataTap to External Ezmeral Data Fabric cluster in Port Gregory, Australia	10.1.0.163,10.1.0.51,10.1.0.170	/PortGregory	Type: mapr CLDB Port: 7222 Cluster Name: xdf.mapr.com Ticket File: ad_admin1_impersonation_ticket Ticket User: ad_admin1 Ticket Type: service with impersonation MapR Tenant Volume: false Impersonation Enabled: true Read Only: false	● created	
☐	TenantStorage	Protected DataTap for a tenant-specific sandboxed storage space.	10.1.0.63,10.1.0.218,10.1.0.251	/exthcp/tenant-4/dco	Type: mapr CLDB Port: 7222 Cluster Name: dfdemo Ticket File: hcp-service-ticket Ticket User: mapr Ticket Type: service MapR Tenant Volume: false Impersonation Enabled: false Read Only: false	● created	

We ran this Jupyter application without replicating the data to a local data store or impacting application performance.

Why This Matters

Businesses need to run analytics in-place where the data is located. This prevents hurdles that slow down insights due to the need to move or manipulate the data from an edge or cloud environment.

ESG verified that HPE Ezmeral Runtime Enterprise uses APIs and POSIX to allow developers and data scientists to use script language and path names to access data wherever it is located. This saves the time and effort of copying data to secondary systems before analysis can begin.

The Bigger Truth

As more organizations employ AI/ML applications to gain business advantage, they are faced with a wide variety of infrastructure and tools for implementing an AI/ML environment. However, the options available for building these environments are not standardized. Multiple business units end up implementing their own AI/ML infrastructure stacks, leading to disjointed compute and data resources that are not fully leveraged. Orchestrating and managing these diverse resources and the underlying infrastructure leads to operational inefficiency and decreased time to insight and value.

HPE Ezmeral Unified Analytics is an open, hybrid platform that unifies analytics across disjointed data stores to provide data teams with a single data repository, simplifying data analytics. HPE Ezmeral Runtime Enterprise, built into HPE Ezmeral Unified Analytics, eliminates steep learning curves, as well as the time and resources required to create a Kubernetes infrastructure across hybrid deployments.

ESG validated that HPE Ezmeral Runtime Enterprise, the foundation of HPE Unified Analytics, can:

- Accelerate performance of Apache-Spark workloads when using the combination of HPE Ezmeral Runtime Enterprise, NVIDIA A100 GPU-enabled Kubernetes clusters and NVIDIA RAPIDS Accelerator.

- Simplify the use of Kubernetes via wizard-enabled processes to remove the complexity of creating and managing Kubernetes clusters.

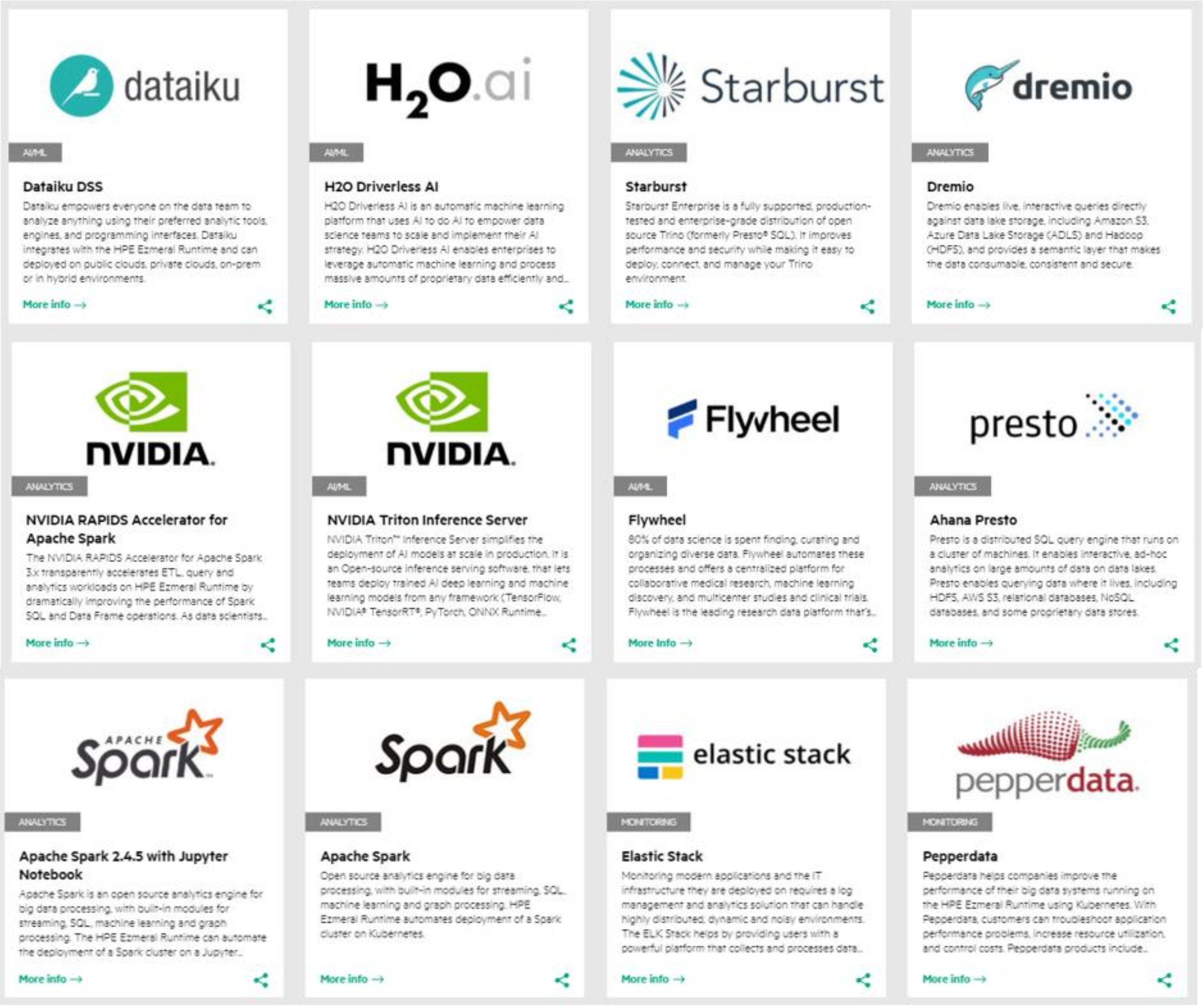
- Enable data access regardless of data location, thus not sacrificing overall application performance.

Potential issues that ESG believes organizations should examine when considering the use of HPE Ezmeral Unified Analytics include application size and the number and type of applications to be containerized. Small applications are not the best candidates for containerization; better candidates skew toward web-scale applications. Also, the platform may be better suited for organizations planning to modernize multiple applications (both re-factoring existing ones and net-new).

If your organization seeks to simplify how containers are deployed and managed while maximizing performance of workloads running on NVIDIA A100 GPU workloads, ESG recommends taking a closer look at how HPE Ezmeral Unified Analytics, NVIDIA RAPIDS Accelerator and NVIDIA A100 GPUs.

Appendix

Figure 10: ISV Marketplace



AI/ML

Dataiku DSS

Dataiku empowers everyone on the data team to analyze anything using their preferred analytic tools, engines, and programming interfaces. Dataiku integrates with the HPE Ezmeral Runtime and can deployed on public clouds, private clouds, on-prem or in hybrid environments.

More info →

AI/ML

H2O Driverless AI

H2O Driverless AI is an automatic machine learning platform that uses AI to do AI to empower data science teams to scale and implement their AI strategy. H2O Driverless AI enables enterprises to leverage automatic machine learning and process massive amounts of proprietary data efficiently and...

More info →

ANALYTICS

Starburst

Starburst Enterprise is a fully supported, production-tested and enterprise-grade distribution of open source Trino (formerly Presto® SQL). It improves performance and security while making it easy to deploy, connect, and manage your Trino environment.

More info →

ANALYTICS

Dremio

Dremio enables live, interactive queries directly against data lake storage, including Amazon S3, Azure Data Lake Storage (ADLS) and Hadoop (HDFS), and provides a semantic layer that makes the data consumable, consistent and secure.

More info →

ANALYTICS

NVIDIA RAPIDS Accelerator for Apache Spark

The NVIDIA RAPIDS Accelerator for Apache Spark 3.x transparently accelerates ETL, query and analytics workloads on HPE Ezmeral Runtime by dramatically improving the performance of Spark SQL and Data Frame operations. As data scientists...

More info →

AI/ML

NVIDIA Triton Inference Server

NVIDIA Triton™ Inference Server simplifies the deployment of AI models at scale in production. It is an Open-source inference serving software, that lets teams deploy trained AI deep learning and machine learning models from any framework (TensorFlow, NVIDIA® TensorRT®, PyTorch, ONNX Runtime...

More info →

AI/ML

Flywheel

80% of data science is spent finding, curating and organizing diverse data. Flywheel automates these processes and offers a centralized platform for collaborative medical research, machine learning discovery, and multicenter studies and clinical trials. Flywheel is the leading research data platform that's...

More Info →

ANALYTICS

Ahana Presto

Presto is a distributed SQL query engine that runs on a cluster of machines. It enables interactive, ad-hoc analytics on large amounts of data on data lakes. Presto enables querying data where it lives, including HDFS, AWS S3, relational databases, NoSQL databases, and some proprietary data stores.

More info →

ANALYTICS

Apache Spark 2.4.5 with Jupyter Notebook

Apache Spark is an open source analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing. The HPE Ezmeral Runtime can automate the deployment of a Spark cluster on a Jupyter...

More info →

ANALYTICS

Apache Spark

Open source analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing. HPE Ezmeral Runtime automates deployment of a Spark cluster on Kubernetes.

More info →

MONITORING

Elastic Stack

Monitoring modern applications and the IT infrastructure they are deployed on requires a log management and analytics solution that can handle highly distributed, dynamic and noisy environments. The ELK Stack helps by providing users with a powerful platform that collects and processes data...

More info →

MONITORING

Pepperdata

Pepperdata helps companies improve the performance of their big data systems running on the HPE Ezmeral Runtime using Kubernetes. With Pepperdata, customers can troubleshoot application performance problems, increase resource utilization, and control costs. Pepperdata products include...

More info →

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of The Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at 508.482.0188.

The goal of ESG Validation reports is to educate IT professionals about information technology solutions for companies of all types and sizes. ESG Validation reports are not meant to replace the evaluation process that should be conducted before making purchasing decisions, but rather to provide insight into these emerging technologies. Our objectives are to explore some of the more valuable features and functions of IT solutions, show how they can be used to solve real customer problems, and identify any areas needing improvement. The ESG Validation Team's expert third-party perspective is based on our own hands-on testing as well as on interviews with customers who use these products in production environments.

Enterprise Strategy Group is an IT analyst, research, validation, and strategy firm that provides market intelligence and actionable insight to the global IT community.