Hewlett Packard Enterprise

Reference Architecture

HPE Reference Architecture for data governance with SAP Data Hub on HPE Synergy and HPE Nimble Storage

Using Red Hat OpenShift Container Platform and Red Hat Virtualization

Contents

Executive summary

With the increased complexity of enterprise landscapes, which can now include Hadoop data lakes, enterprise data warehouses (EDWs), cloud storage, enterprise apps, etc., the need to manage and govern data across a landscape is well understood. SAP Data Hub provides organizations with an integration layer for data-driven processes across enterprise services to process and orchestrate data in the overall landscape for all user groups. It integrates and prepares data within a digital landscape to drive business decisions. Data and processes can be managed, shared, and distributed across the enterprise with seamless, unified, and enterprise-ready monitoring and landscape management capabilities.

For organizations to be able to trust their data's accuracy they must ensure end-to-end governance across all data sources. SAP Data Hub is designed to ensure data reliability, traceability, and compliance. The SAP Data Hub Metadata Explorer can be used to govern and manage an organization's metadata assets from various systems and distinct sources.

SAP Data Hub uses uniquely powerful data pipelines that are based on the serverless computing paradigm to address data integration, orchestration, and governance capabilities across a complex landscape. It is a containerized application that creates and scales out containers as more workflows, pipelines and processing power are needed.

Container technology provides the right application platform to help organizations become more responsive and iterate across multiple IT environments as well as develop, deploy, and manage applications faster. Red Hat OpenShift Container Platform on HPE Synergy provides an end-to-end fully integrated container solution that, once assembled, can be configured within hours. This eliminates the complexities associated with implementing a container platform across an enterprise data center and provides the automation of hardware and software configuration to quickly provision and deploy a containerized environment at scale. Red Hat OpenShift Container Platform provides organizations with a reliable platform for deploying and scaling container-based applications and HPE Synergy provides the flexible infrastructure you need to run that container platform to dynamically provision and scale applications, whether they run as VMs or containers, or are hosted on-premises, in the cloud, or in a hybrid environment.

This Reference Architecture provides an SAP Data Hub solution that delivers the end-to-end data governance required across complex diverse data landscapes. A Red Hat OpenShift deployment on HPE Synergy Composable Infrastructure and HPE Nimble Storage is utilized to deploy, scale, and manage the environment.

This Reference Architecture describes how to:

- Use SAP Data Hub to unify data management, centralize governance and pipelining capabilities of a complex data landscape.
- Use SAP Data Hub Metadata Explorer to help govern and manage metadata assets that are spread across diverse systems.
- Provide data protection, individual privacy, and security using the SAP Data Hub ecosystem.
- Efficiently lay out an OpenShift configuration using a mix of virtual machines and bare metal hosts.

This Reference Architecture demonstrates the following benefits of utilizing HPE Synergy and HPE Nimble for Red Hat OpenShift Container Platform:

- Synergy is the leading composable platform that supports any workload from virtualization, bare metal to containers. HPE Synergy uses fluid pools of resources, software-defined intelligence, and a unified application program interface (API) to enable composability of any configuration on demand, provides flexibility and eliminates over or under provisioning.
- The HPE Composable Infrastructure solution provides a layered view of security controls. The objective of choosing this layered security view is to ensure that consumers become aware of the depth of security risk that an infrastructure can have and also make them aware of the depth of defense that is built in to the HPE Composable Infrastructure design.
- Using Enterprise grade storage solution, for example HPE Nimble Storage for persistent container storage, enables speed, portability, and agility for traditional enterprise applications and data.
- A business-driven container application data protection architecture provided by HPE Nimble Storage with HPE InfoSight providing proactive management.

Target audience: This paper is intended to assist Chief Data Officers and Data VPs, Data Scientists, Data Engineers, AI/ML developers, and experienced users that are interested in simplifying and managing data and processes, shared and distributed across the enterprise with seamless, unified, and enterprise-ready monitoring and landscape management capabilities.

Document purpose: This Reference Architecture provides an overview of the deployment of SAP Data Hub on a Red Hat® OpenShift Container Platform environment running on HPE Synergy and HPE Nimble Storage. In addition to outlining the opportunity and key solution components, this paper focuses on using SAP Data Hub to ensure data governance and provides guidelines for configuring and deploying the combined solution.

This Reference Architecture describes solution testing performed in October 2019.

Introduction

This Reference Architecture describes a highly available and secure SAP Data Hub solution with a focus on data governance using a Red Hat OpenShift Container Platform deployment on HPE Synergy Composable Infrastructure. It includes key features of the SAP Data Hub application, focusing on end-to-end data governance and details on the design and configuration of the environment. It provides an architectural and solution overview of SAP Data Hub, the key benefits and design points, and how the SAP Data Hub Metadata Explorer can be used to govern and manage metadata assets. It also describes a comprehensive example of how Red Hat OpenShift Container Platform can be set up to take advantage of the HPE Synergy Composable Infrastructure and HPE Nimble Storage. It uses an HPE Synergy platform with HPE Synergy 480 Gen10 Compute Modules installed with a mix of Red Hat Virtualization and bare metal Red Hat Enterprise Linux®. HPE Nimble Storage provides persistent storage, which is required by SAP Data Hub, for containers and registry, virtual machine storage, and data management.

Due to the ephemeral nature of containers, protecting persistent data associated with the containers becomes a crucial task. In this Reference Architecture, the Red Hat OpenShift Container Platform pod's persistent volume is protected using HPE Nimble Storage array's data protection plan and is replicated to a remote HPE Nimble Storage array, thus providing end-to-end data protection and disaster recovery capability.

The HPE Synergy platform is designed to bridge traditional and cloud-native applications with the implementation of HPE Synergy Composable Infrastructure. HPE Synergy Composable Infrastructure combines the use of fluid resource pools made up of compute, storage, and fabric with software-defined intelligence. HPE Synergy platform provides the agility and scalability on the hardware layer to the overall Red Hat OpenShift Container Platform solution.

What is data governance?

Data governance is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods. The typical universal goals of a data governance program include enabling better decision making, protecting the needs of data stakeholders, building repeatable processes and ensuring transparency of processes.

The Data Governance Institute considers all the following principles as a requirement for a successful data governance project.

- **Integrity** - Data governance participants will practice integrity with their dealings with each other; they will be truthful and forthcoming when discussing drivers, constraints, options, and impacts for data-related decisions.
- **Transparency** - Data governance and Stewardship processes will exhibit transparency; it should be clear to all participants and auditors how and when data-related decisions and controls were introduced into the processes.
- **Auditability** - Data-related decisions, processes, and controls subject to data governance will be auditable; they will be accompanied by documentation to support compliance-based and operational auditing requirements.
- **Accountability** - Data governance will define accountabilities for cross-functional data-related decisions, processes, and controls.
- **Stewardship** - Data governance will define accountabilities for stewardship activities that are the responsibilities of individual contributors, as well as accountabilities for groups of Data Stewards.
- **Checks-and-Balances** - Data governance will define accountabilities in a manner that introduces checks-and-balances between business and technology teams as well as between those who create/collect information, those who manage it, those who use it, and those who introduce standards and compliance requirements.
- **Standardization** - Data governance will introduce and support standardization of enterprise data.

- **Change Management** - Data governance will support proactive and reactive Change Management activities for reference data values and the structure/use of master data and metadata.[1]

SAP Data Hub

It is getting harder and costlier for organizations to not only understand the data that they have, but to work across all the different systems that need to use it, and apply end-to-end governance, to capture the maximum value. SAP Data Hub delivers a simpler, more scalable approach to data landscape management. With enterprise-spanning data integration, processing, and governance, SAP Data Hub provides unprecedented visibility into and access across the complex network of data in the modern enterprise. By providing a broad, detailed, and easily understood view of the entire data landscape, from sources like Hadoop and Amazon S3 to SAP HANA and ERP, it helps organizations deeply understand data sources, uses, interconnections, quality, and impacts. This allows enterprises to see new opportunities from data, resolve emerging data issues, and ensure that data is flowing to where it needs to go.[2]

SAP Data Hub offers an integration layer for data-driven processes across enterprise services to process and orchestrate data in the overall landscape for all user groups – IT, business analysts, data scientists, data engineering, and data stewards. It integrates and prepares data within a digital landscape to drive business decisions. It offers an open big data-centric architecture with open source integration, cloud deployments, and third-party interfaces. It leverages massive distributed processing and serverless computing capabilities provided by the SAP Data Hub Pipeline Engine and SAP Vora, which are part of SAP Data Hub.

SAP Data Hub establishes a new category of software solution, and provides a comprehensive answer to an emerging and painful challenge for enterprise customers: integrating data and establishing data-driven processes across an increasingly diverse data landscape. The solution addresses data integration, data orchestration, and data governance capabilities across a complex landscape, and harnesses big data processing to create uniquely powerful data pipelines that are based on the serverless computing paradigm. Data and processes can be managed, shared, and distributed across the enterprise with seamless, unified, and enterprise-ready monitoring and landscape management capabilities.[3]

Figure 1 provides an architectural overview of the SAP Data Hub solution.

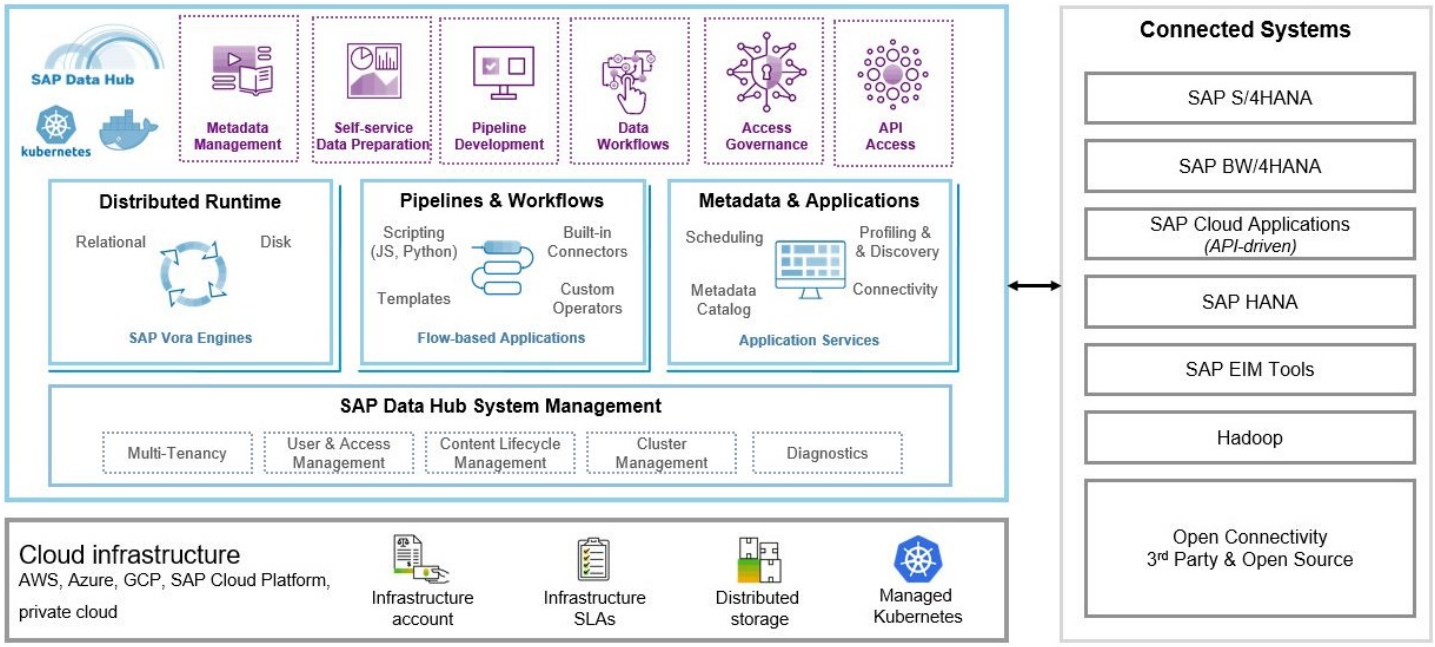

Figure 1. Architecture of SAP Data Hub

[1] www.datagovernance.com/adg_data_governance_goals/
[2] blogs.saphana.com/2017/10/04/what-is-sap-data-hub-and-answers-to-other-frequently-asked-questions/
[3] blogs.saphana.com/2017/10/16/sap-data-hub-solution-overview-youtube-channel-announcement/

Part of the power of the solution resides in its ability to leave the data where it is. The data does not have to be mass centralized with SAP Data Hub. This provides advantages in terms of ease of management and speed of data pipeline execution. Customers leverage their existing data stores and existing processing capabilities.

Key Features of SAP Data Hub

SAP Data Hub is built to deploy on-premise, on cloud, or in hybrid landscapes. Existing customers can continue to use the data integration tools; SAP Data Services, SAP Landscape Transformation, SAP HANA smart data integration for data integration, data virtualization, and data replication. In addition to enjoying the use of these integration tools, they will have the benefit of creating orchestration scenarios which involve these tools and the ability to monitor operations and processes end-to-end.

SAP Data Hub includes a modern UI for hub management to manage an end-to-end data landscape that connects enterprise data with big data. It enables the creation of copies of datasets for data scientists which can be used for training machine learning models. With fine grain, end-to-end security provided by policy management functionality, it will manage access control of all data, enterprise and big data, consistently.

To meet data governance requirements SAP Data Hub provides data discovery capabilities which includes profiling of data natively without leaving the source. Managing a repository of all forms of data and metadata, it offers data lineage and enables impact analysis. It offers rich data transformation, quality, and enrichment through the integration of e.g. SAP Enterprise Information Manager Tool capabilities, in SAP Data Hub.

One of the powerful features of SAP Data Hub is data pipelining, which connects data in different formats and makes them accessible to SAP Vora for analytics. In addition to allowing the use of existing data ingestion tools and orchestrating them within SAP Data Hub, it offers an easy-to-use advanced data pipeline to move data when needed from various sources, including Amazon S3 to Hadoop HDFS, HDFS to SAP Vora, and SAP Vora to HDFS.[4]

SAP Data Hub and data governance

End-to-end data governance is a requirement across complex landscapes. The need to manage and govern data across a landscape is well understood. Ensuring data lineage and impact analysis of changes, managing security and privacy requirements, etc. are all critical aspects of a trusted enterprise landscape. With the increased complexity of enterprise landscapes, which can now include Hadoop data lakes, EDWs, Cloud storage, enterprise apps, etc., the ability to appropriately provide effective governance is more difficult. Without end-to-end governance across all data sources, organizations cannot trust and rely on the data's accuracy, creating risk for anyone using analytics or operational applications that use the data.

One of the key design points for SAP Data Hub is to ensure data reliability, traceability, and compliance in accordance with your business. SAP Data Hub helps manage your data across different systems by using the Metadata Explorer, which gathers information about the location, attributes, quality, and sensitivity of data. With this information, you can make informed decisions about which datasets to publish and determine who has access to use or view information about the datasets. The goal of the SAP Data Hub Metadata Explorer is to help govern and manage metadata assets that are spread across diverse systems and disparate sources. It can be used to:

- Preview data in the datasets
- Create indexes about the dataset contents to aid in searching for datasets
- Profile data to view information about the contents of different datasets
- Publish datasets to allow others to view and search the data
- Label the dataset with keywords, which also helps in searching for datasets
- Prepare the datasets by applying data quality enhancements to the data
- Conduct lineage analysis to learn where the dataset is used and how it is transformed
- Create validation rules to ensure that your data passes data quality standards
- Monitor the status of tasks

Effective data governance can be achieved through metadata management, data lineage and impact analysis, and data access and security. Metadata is managed through indexing, publishing, and profiling. Indexing extracts the metadata, so the data can be searched. The extracted

[4] https://blogs.saphana.com/2017/10/16/sap-data-hub-solution-overview-youtube-channel-announcement/

information is available in the search and browse pages. Publishing extracts the same metadata and the dataset is placed in the catalog where you can share the content with others and organize your datasets. The catalog is the target location for published datasets. Profiling produces extra metadata about the values in the dataset. For example, you can view the unique or distinct values, the minimum and maximum values, average length, and whether there are null, blank, or zero values. This information can help you determine which datasets may need cleansing, masking, or any number of options available in the SAP Data Hub Modeler.

The screenshots below demonstrate using the SAP Data Hub Metadata Explorer to profile, index, and publish a dataset.

From the Data Hub LaunchPad, select the Metadata Explorer application.

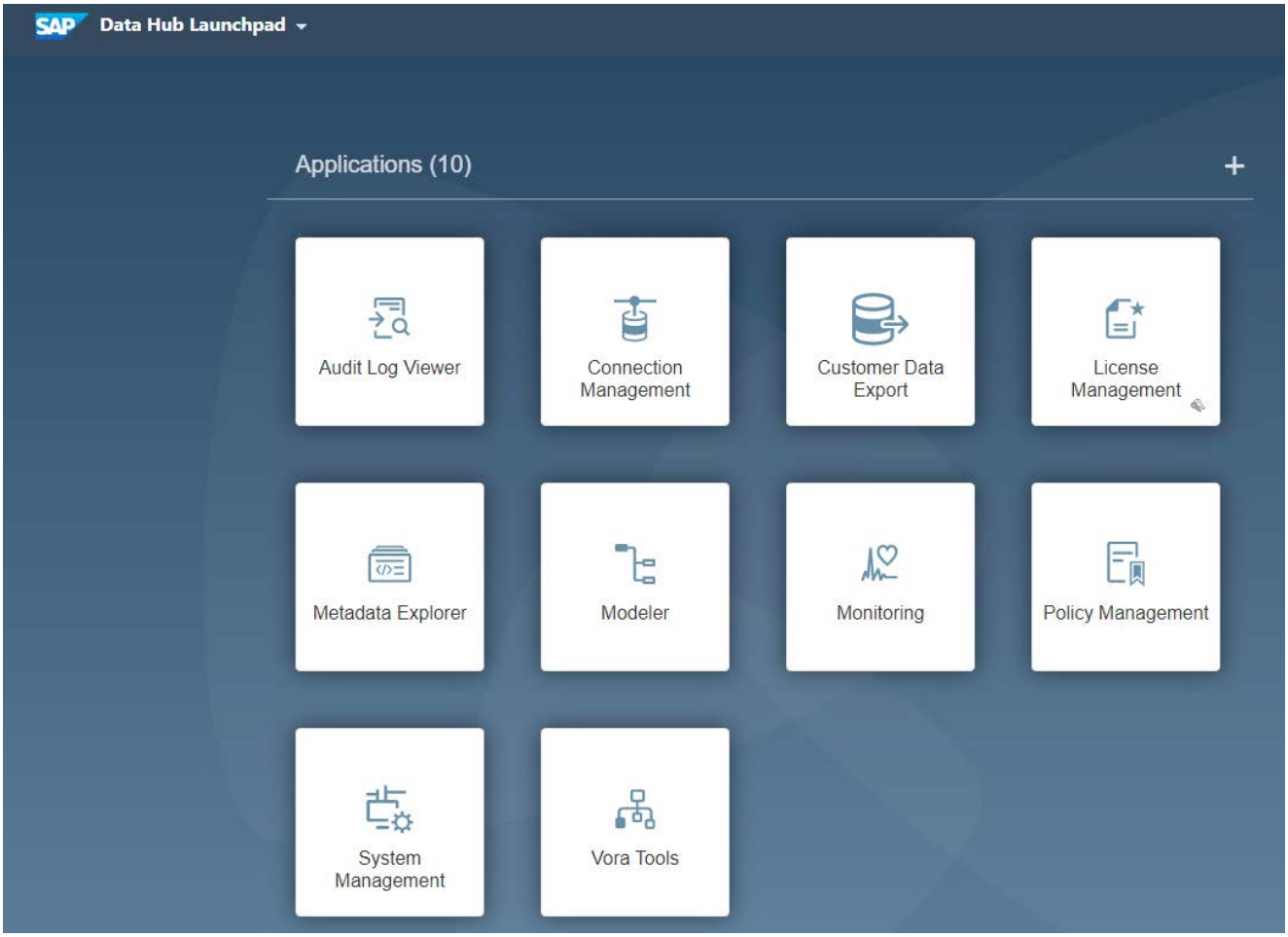

Figure 2. Data Hub Launchpad

Using Browse Connections from within the Metadata Explorer we can explore all of our dataset connections.

In Figure 3 below you can see we have a number of comma delimited files stored on a connection to an HDFS system.

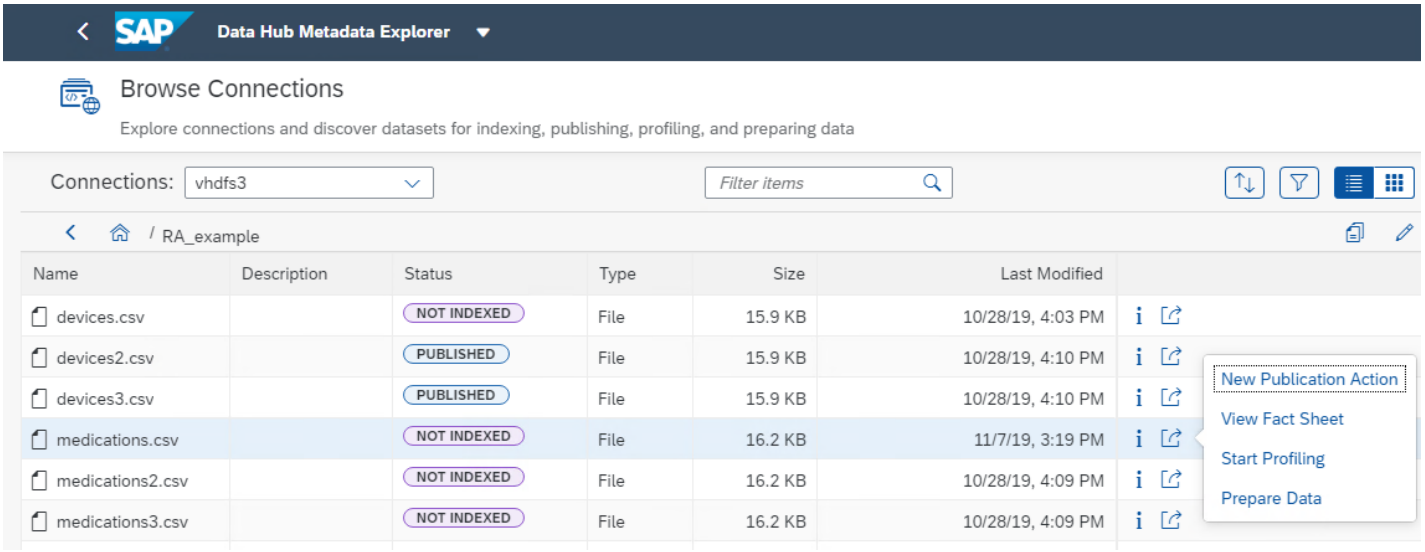

Figure 3. Files on an HDFS connection

Selecting the action icon on the medications.csv row shows us the possible actions to perform on that dataset. To view the metadata we select "View Fact Sheet".

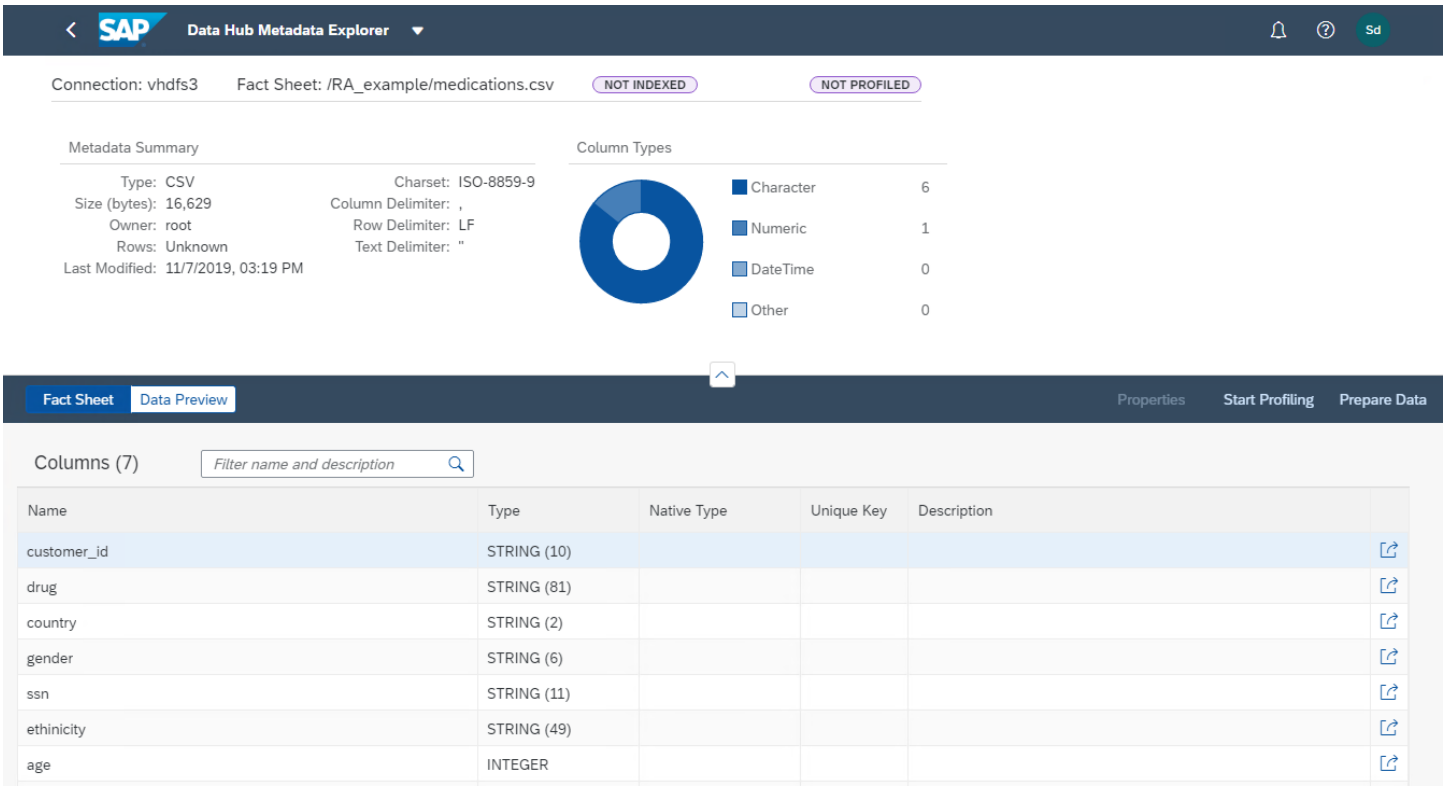

Figure 4. Fact Sheet for medications.csv

The fact sheet shows us that the data file has not been index or profiled, the metadata summary information, and column types. We can also preview the data, by selecting the "Data Preview" button.

	Aa customer_id	Aa drug	Aa country	Aa gender	Aa ssn	Aa ethinicity
1	85-9397130	Banana Boat Sport SPF 30	ZW	Female	220-01-3172	Chickasaw
2	46-6185014	fluticasone propionate	Null	Female	542-02-4345	Null
3	43-0564329	Childrens Pain Relief Bub...	Null	Male	351-04-1452	Seminole
4	09-3935124	Childrens Pain and Fever	CN	Male	647-25-0729	Null
5	03-3237071	HYSAN PAIN RELIEVER	MX	Male	535-66-4426	Delaware
6	87-5751656	OP Protective Gel 15	Null	Male	536-28-5114	Colombian

Figure 5. Fact Sheet data preview

Selecting the "Start Profiling" action will produced extra metadata about the values in the dataset, showing the unique or distinct values, the minimum and maximum values, average length, and whether there are null, blank, or zero values. This is the first step in determining which datasets may need cleansing, masking, or any number of options available in the SAP Data Hub Modeler. Once the profiling task is completed the fact sheet will display the additional metadata.

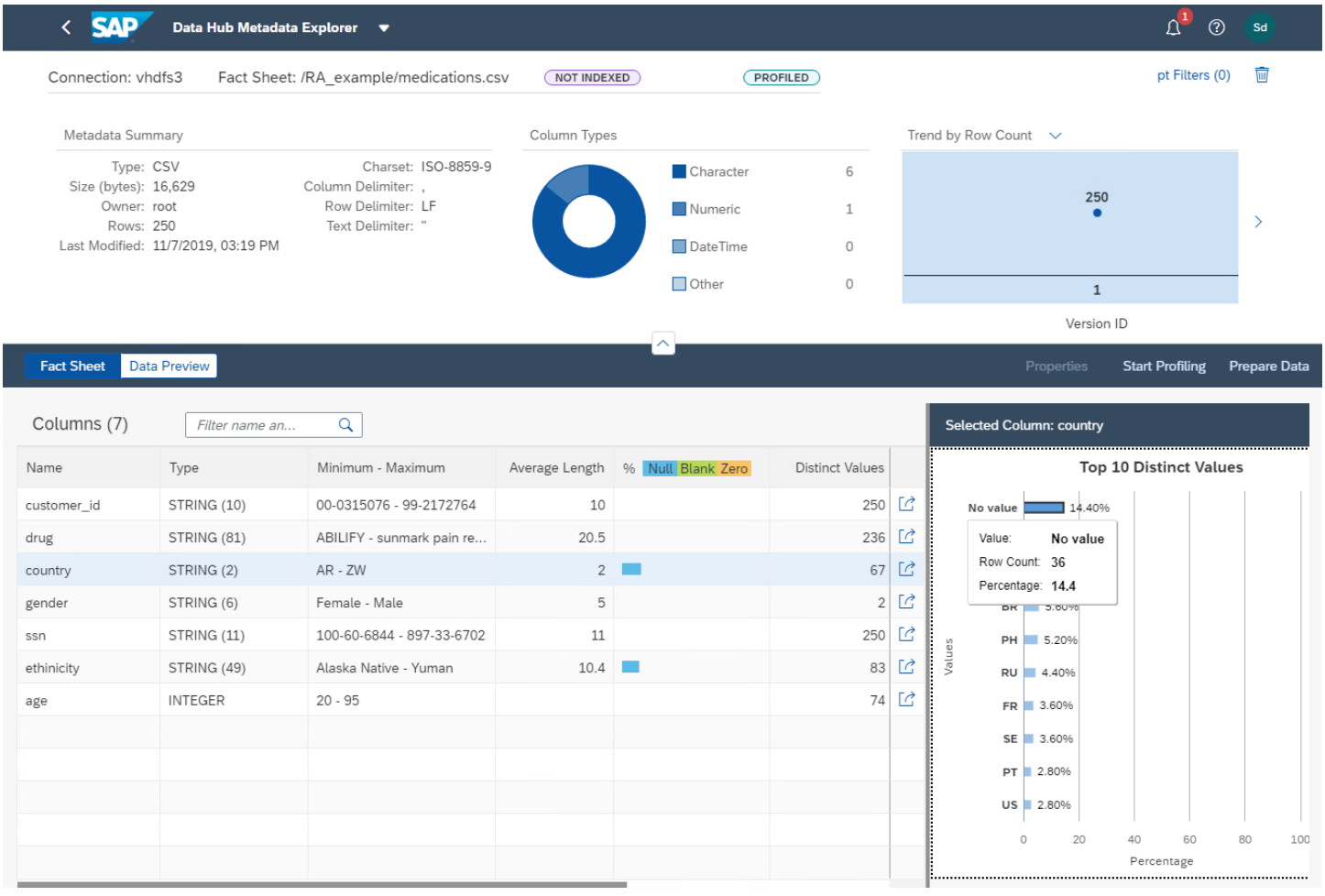

Figure 6. Fact Sheet after dataset is profiled

After profiling, in addition to the metadata the fact sheet will display trend by row count. In this example the profiling version id is 1 and the row count is 250. Columns can be selected to show an analytic graph of the distinct values. This example highlights the fact that both the country and the ethnicity columns have null values. The "% Null Blank Zero" column displays a visual indicator as to the percentage of each. Selecting the country column shows that 14% of the country values are "No value" with a row count of 36.

Indexing extracts the metadata, so the data can be searched. The extracted information is available in the search and browse pages. A catalog is the target location for published datasets. A published dataset contains the extracted metadata and is placed in the catalog where you can share the content with others and organize your datasets. To index and publish the dataset we select "New Publication Action" as shown in Figure 3 above.

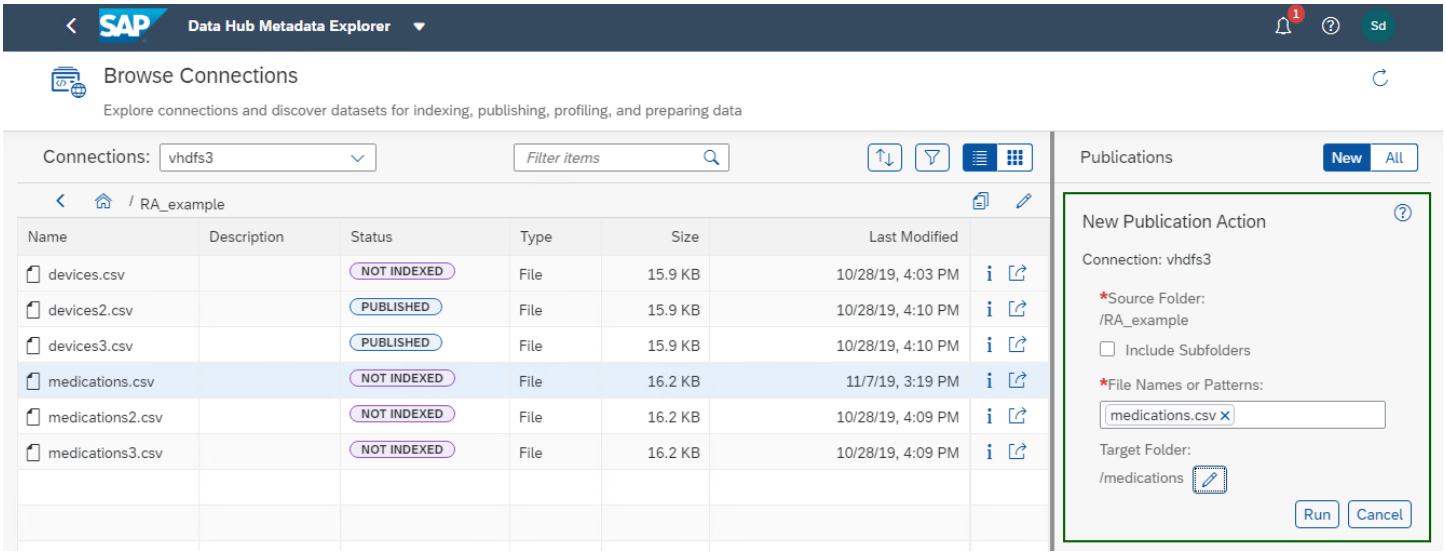

Figure 7. New Publication Action

From the New Publication Action frame:

1. Select the catalog target folder in which to place the published dataset. This example uses the medications target folder.
2. Then press the "Run" button to start the publishing task.

Once the publishing task is completed the fact sheet will show the dataset as published and it will be listed in the catalog folder. Figure 8 below shows the fact sheet for medications.csv as published.

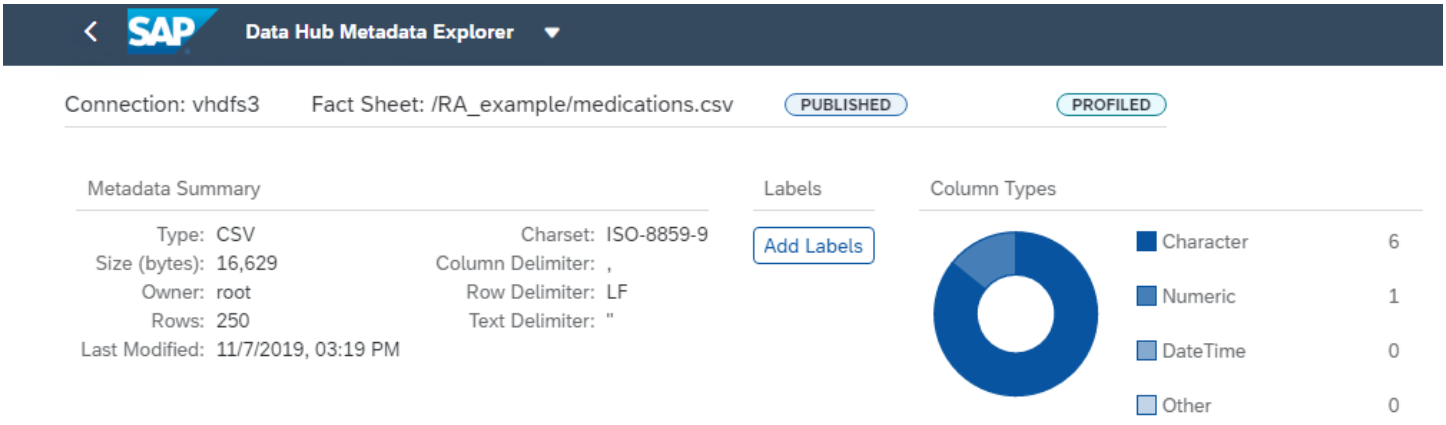

Figure 8. Fact Sheet after New Publication Action

Figure 9 below shows the medications.csv file listed within the medications catalog folder.

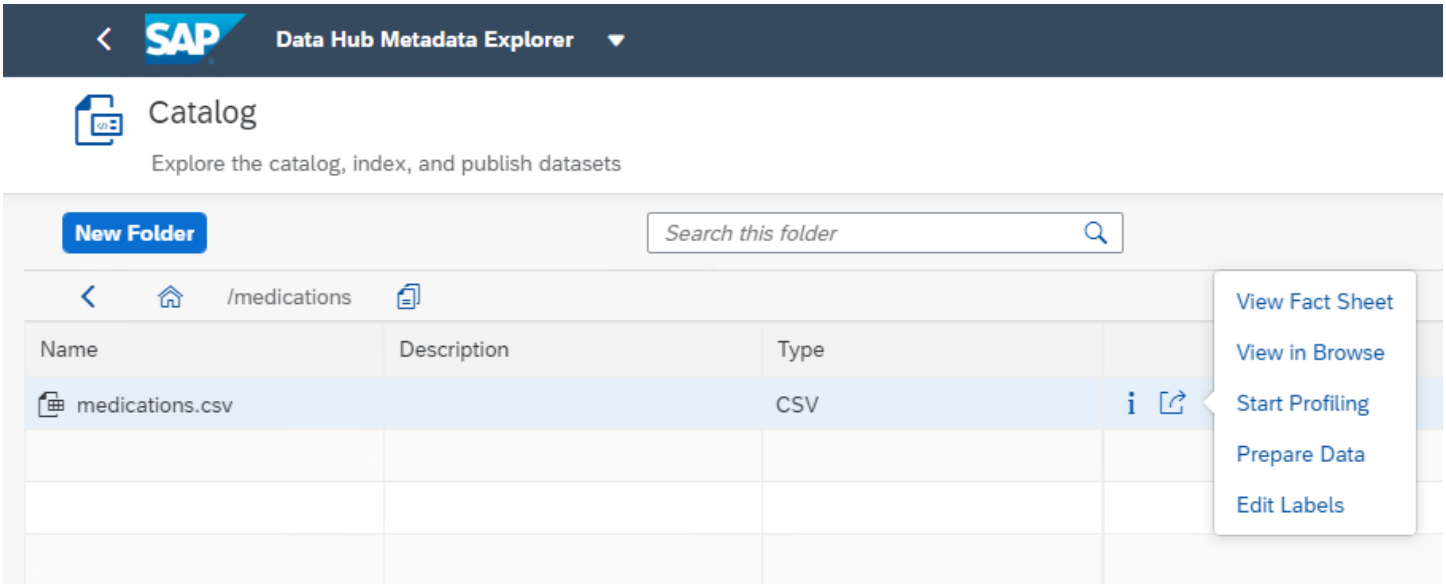

Figure 9. Catalog with indexed and published datasets

3. Once the dataset is published labels can be used to help you search for datasets or identify the contents of your datasets. Select the "Edit Labels" action to add a label to the published medications.csv dataset.

   We can then use the label to search for the dataset.

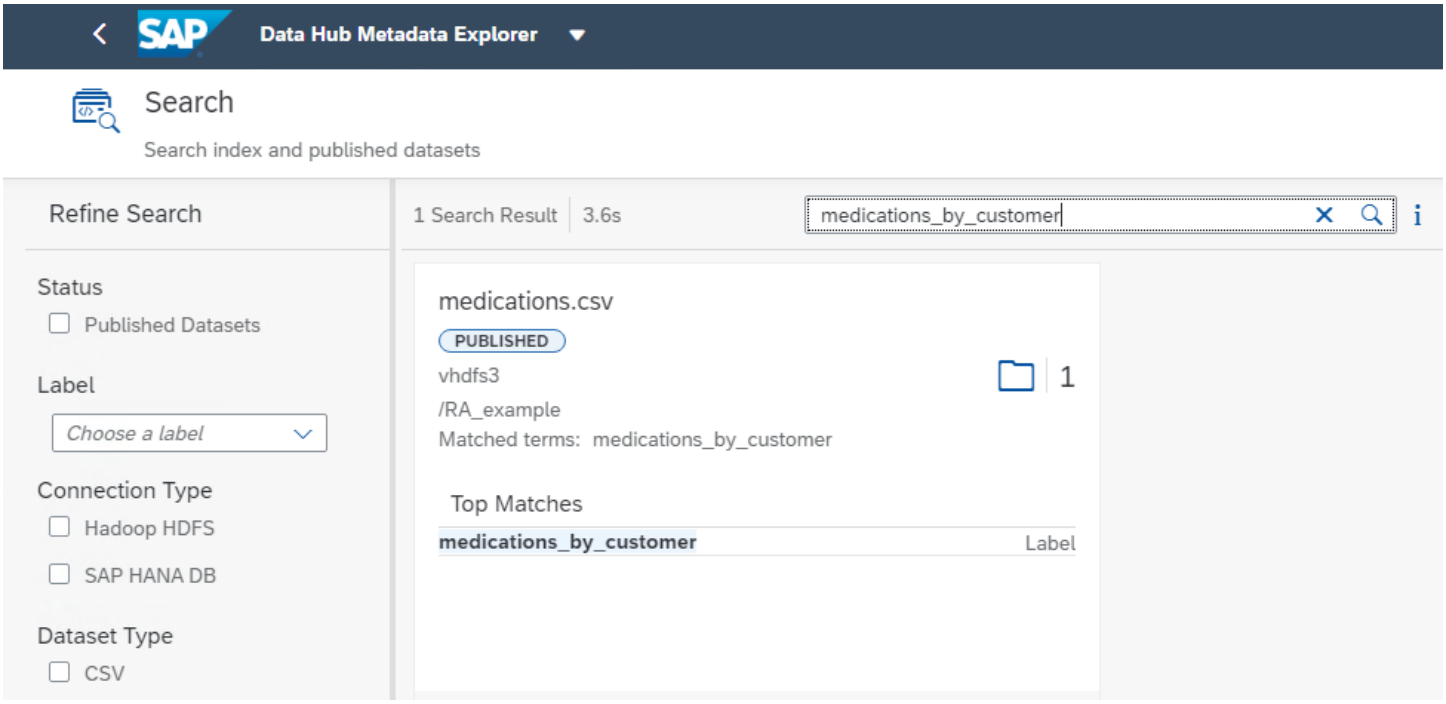

Figure 10. Searching for a published dataset using a label

When you are looking for more information about the origin of a dataset or want to learn where a dataset has been used, you can run lineage analysis when indexing or publishing a dataset. After lineage is processed, you can see a graphical representation that shows the input and output of transformations performed on the dataset. Use data lineage to trace back from a dataset to the source and to view and navigate through various dependencies between objects. For example, if you have some data that has been transformed or enhanced, you may want to find where the data originated to learn how the dataset may have been modified.[5]

SAP Data Hub provides policy management to set resource access to a user for defined policies. A policy is a defined control structure where access rights are granted to users via the specific use of "policies". The policies can use multiple attributes such as user attributes and resource attributes. The SAP Data Hub administrator creates policies, which defines a resource and who has access to it. Each policy includes a policy id, a description, and an exposed field, which specifies whether the policy can be assigned to the users or not.

Anonymization Operator

To support privacy requirements SAP Data Hub provides an anonymization operator for individual's privacy protection. You can use the anonymization operator to mask out all or a portion of the data that contains sensitive information. You can further protect the privacy of each individual identity by grouping similar records into a category and discover insights from your data without risking re-identification of individuals.

Figure 11 is a sample graph that shows how the anonymization operator is used within a pipeline to hide the individual identities within a group of records.

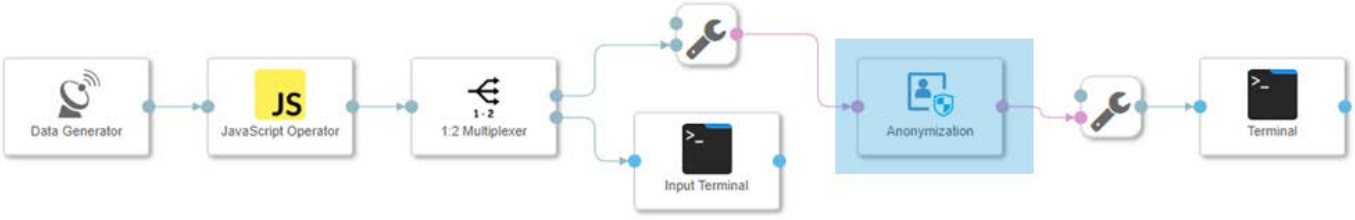

Figure 11. Anonymization operator used within a pipeline

The anonymization operator will use various techniques such as masking, generalization, and global suppression to reduce the granularity of the data representation. As an example, in the table below the patient name is completely hidden, zip code is partially masked, age information is being generalized, and the patient data is categorized in 3 groups.

Table 1. Anonymization example tables

Row Id	Zip	Age	Issue
1	546*	[40,60]	Pink eye
2	538*	[20,40]	Ingrown toenail
3	546*	[40,60]	Severed finger
4	555*	[60,80]	Heart Disease
5	546*	[40,60]	Concussion
6	538*	[20,40]	Twisted ankle
7	538*	[20,40]	Broken toe

Masked input data

Anonymized Group	Count	Zip	Age
1	3	546*	[40,60]
2	3	538*	[20,40]
3	1	555*	[60,80]

Anonymized groups

Row Id	Zip	Age	Issue
1	546*	[40,60]	Pink eye
2	538*	[20,40]	Ingrown toenail
3	546*	[40,60]	Severed finger
5	546*	[40,60]	Concussion
6	538*	[20,40]	Twisted ankle
7	538*	[20,40]	Broken toe

Masked output data

In this example the anonymization operator was configured to suppress records with a count less than 2, to further protect the privacy of an individual. The resulting data set does not include the anonymized group containing less than 2 records. [6]

SAP Data Hub software supports data privacy by providing security features and specific functions relevant to data protection, such as functions for the simplified blocking and deletion of personal data. SAP Data Hub does not own or store any sensitive data but the features for working with data sources, flowgraphs, and so on. Therefore, it is the user that knows, for example, which tables in the database contain sensitive personal

[5] help.sap.com/viewer/bd9aff2447b24109bccbd7618cfd542e/2.6.latest/en-US
[6] https://blogs.saphana.com/2019/01/18/whats-new-in-sap-data-hub-2-4/

data, or how business level objects, such as sales orders, are mapped to technical objects in the SAP Data Hub ecosystem. Many data protection and privacy functions are available to enable a company to process personal data in a clear and compliant manner in addition to a variety of security-related features to implement general security requirements.

For additional information on SAP Data Hub security features refer to the Administration Guide for SAP Data Hub help.sap.com/viewer/1246f58c6a74412580877af1f484f41a/2.6.latest/en-US

For additional information on SAP Data Hub refer to the official SAP documentation site help.sap.com/viewer/product/SAP_DATA_HUB/2.6.latest/en-US

Solution overview

The Red Hat OpenShift Container Platform on HPE Synergy and HPE Nimble Storage provides the infrastructure to support the SAP Data Hub solution. Figure 12 provides an overview of the solution components.

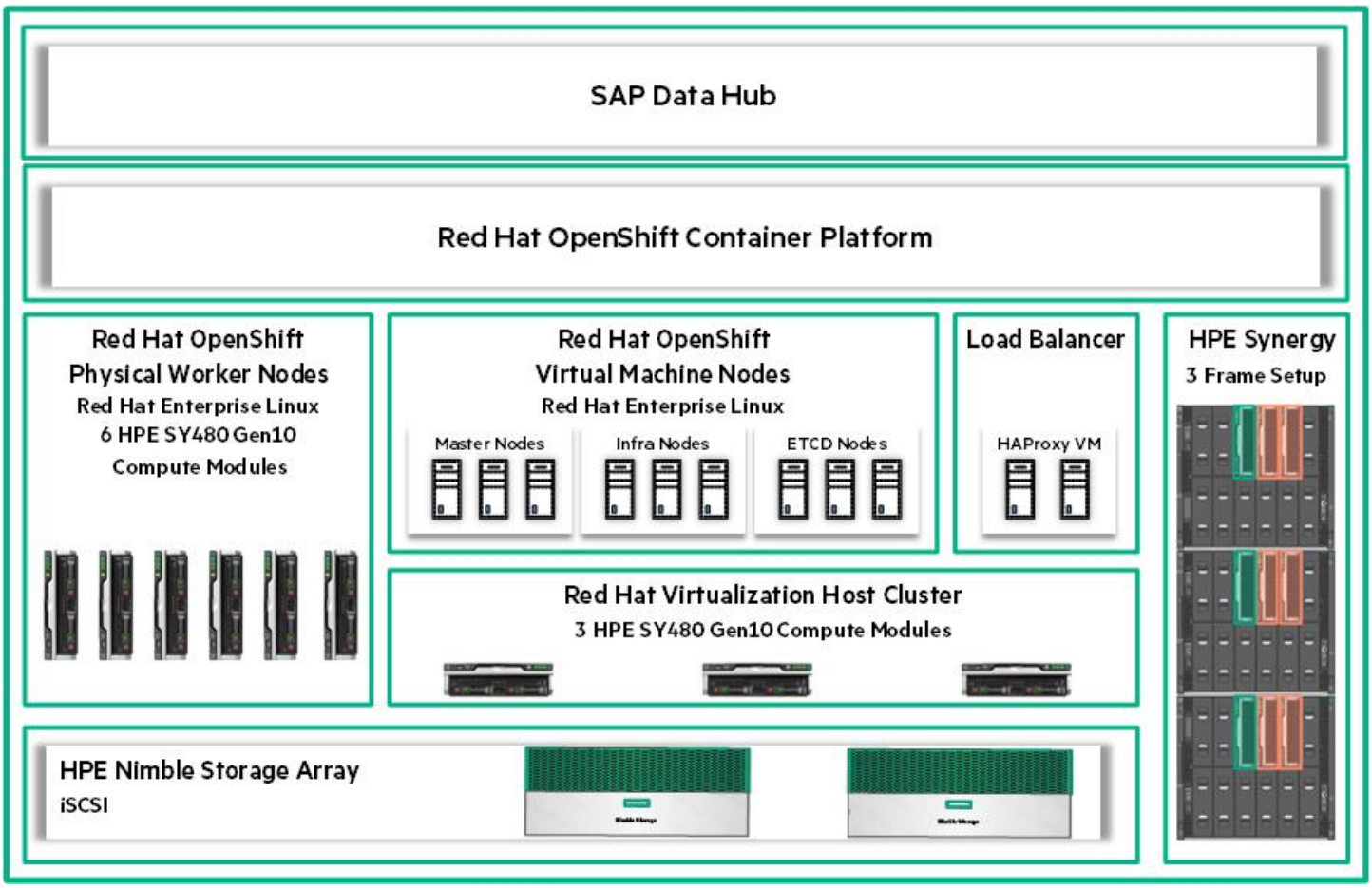

Figure 12. Solution overview

The HPE Synergy with HPE Nimble Storage solution deploys Red Hat OpenShift Container Platform 3.11 as a combination of virtual and physical resources. The OpenShift master, etcd, and infrastructure nodes are deployed as virtual machines running on three HPE Synergy 480 Gen10 Compute Modules running Red Hat Virtualization Host 4.2 (RHVH) and managed by Red Hat Virtualization Manager (RHV-M). Red Hat Enterprise Virtualization (RHVH) is the hypervisor used. Load balancing can be deployed as a virtual machine or as physical appliances. Red Hat OpenShift worker nodes are deployed on bare metal on six HPE Synergy 480 Gen10 Compute Modules running Red Hat Enterprise Linux 7.6. The operating system for the Red Hat OpenShift worker nodes is booted from HPE Image Streamer. HPE Nimble Storage provides support for both ephemeral and persistent container volumes using the Nimble Linux toolkit. The number of servers used in this solution overview is only an example. The number required depends on the organization's workload requirements.

SAP Data Hub Foundation is installed on the Red Hat OpenShift Container Platform. A Kubernetes cluster infrastructure is a prerequisite to install SAP Data Hub Foundation. The HPE Synergy is certified for Red Hat Enterprise Linux versions and provides an ecosystem to run Red Hat OpenShift. A container registry is required to mirror the SAP Data Hub container images and to deploy images created by the Pipeline Engine. For detailed instructions on installing SAP Data Hub on the Red Hat OpenShift Container Platform refer to the latest SAP Data Hub Installation Guide and the Red Hat Customer Portal.

HPE Synergy Provides the flexibility, scalability, easy management and security

Synergy Composable Infrastructure offers fluid resource pools to be composed, released to the pool and recomposed at any compute and storage ratio to scale and adapt to the needs of the workloads or apps.

Synergy software defined intelligence allows template driven, frictionless configuration to reduce management complexity and manual tasks. For example, adding more drives is simply adjusting server profile and zoning more drives from the available pool of resources.

Synergy adopted the HPE pioneered silicon root of trust technology and it's the corner stone of our continued efforts to harden your infrastructure with the most robust and innovative security solutions. Synergy Composer2 is a management appliance that manages the Secure Boot processes.

- 'Secure Start': validates the iLO5 FW & the UEFI BIOS
- 'Secure Boot': validates the OS Bootloader and OS kernel (and kernel modules & drivers) of the Composer2 appliance
- Synergy-specific validation of the OS Bootloader

With these security enhancements, you can rest easy knowing your firmware is secure down to the silicon and that Synergy Composer2 can never boot compromised code.

Synergy Layered Security

Data governance requires the overall management of the availability, usability, integrity and security of data used in an enterprise. As threats move from network security to the hardware and firmware layers, HPE Synergy Gen10 security features help protect your hardware, firmware, and network components from unauthorized access and unapproved use. HPE offers an array of embedded and optional software and firmware for HPE Gen10 that enables you to institute the best mix of remote access and control for your network and data center. HPE Synergy Gen10 servers contain the following security aware components.

HPE iLO 5 -The HPE iLO subsystem, a standard component of HPE Synergy servers, simplifies server setup, health monitoring, power and thermal optimization, and remote server administration. With an intelligent microprocessor, secure memory, and dedicated network interface, iLO offers varying degrees of encryption and security. Ranging from a standard open level up to the Federal Information Processing Standard and the Commercial National Security Algorithm security, iLO offers administrators a reliable way to integrate HPE Synergy servers into existing security environments.

UEFI System Utilities -The UEFI System Utilities is embedded in the system ROM. Unified Extensible Firmware Interface (UEFI) defines the interface between the operating system and platform firmware during the boot, or start-up process. UEFI supports advanced pre-boot user interfaces and extended security control. Features such as Secure Boot enable platform vendors to implement an OS-agnostic approach to securing systems in the pre-boot environment. The ROM-Based Setup Utility (RBSU) functionality is available from the UEFI System Utilities along with additional configuration options.

The HPE Synergy Gen10 servers with iLO 5 and its silicon root of trust undergo a server boot process that authenticates from the hardware itself and undergoes a series of trusted handshakes before fully initializing the UEFI and the OS. The silicon root of trust enables the detection of previously undetectable compromised firmware or malware. The advanced capabilities of iLO 5 enable daily automatic scanning of firmware and

automatic recovery to authentic good states. Combining the Gen10 security features with selected server options allows you to design a resilient and hardened industry-standard server infrastructure.

Figure 13 contains the key security features provided within the HPE Synergy infrastructure.
For more information on HPE server infrastructure security refer to hpe.com/security.

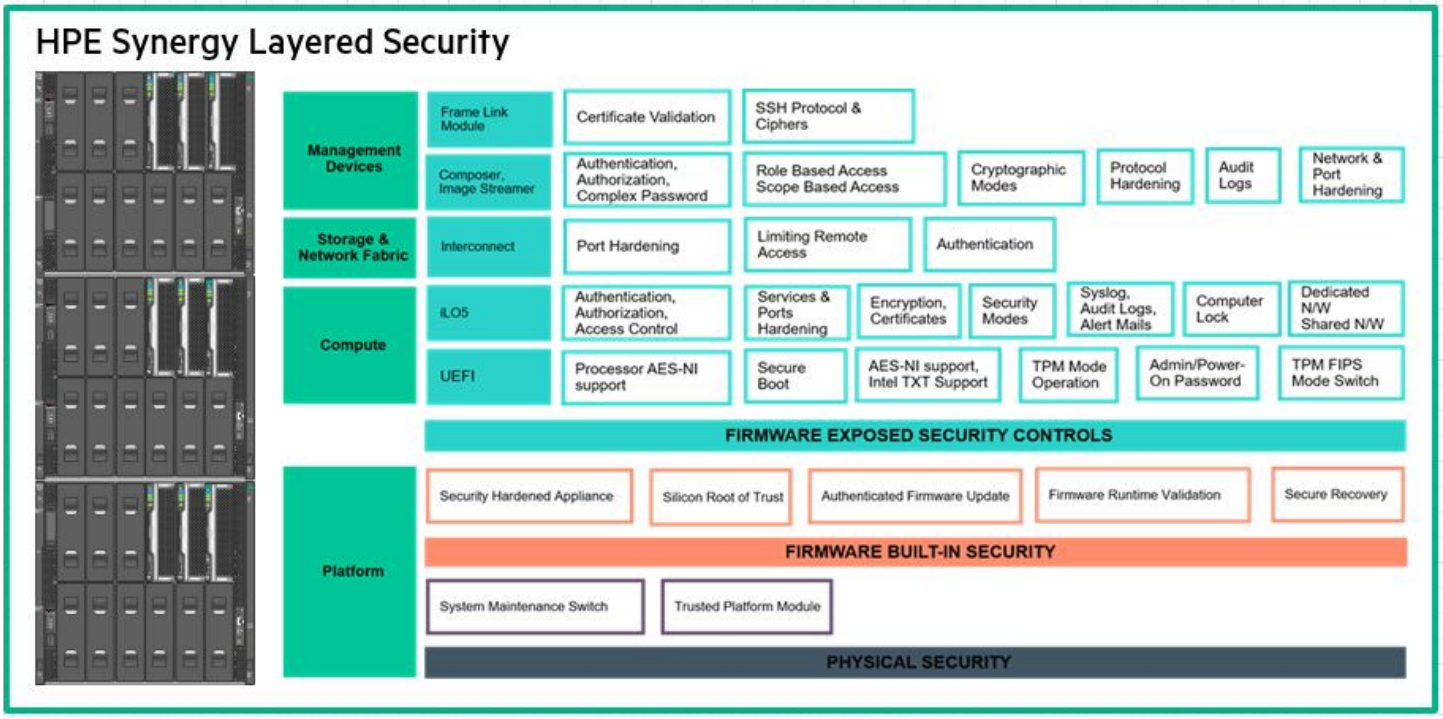

Figure 13. Overview of the key security features within the HPE Synergy infrastructure

Data protection

A container application data protection architecture should be defined by business requirements. These requirements include factors such as the speed of recovery, the maximum permissible data loss, and data retention needs. The data protection plan must also take into consideration various regulatory requirements for data retention and restoration. Finally, different data recovery scenarios must be considered, ranging from the typical and foreseeable recovery resulting from user or application errors to disaster recovery scenarios that include the complete loss of a site. For this solution, HPE recommends the use of a remote HPE Nimble Storage array as a backup target for Red Hat OpenShift components including persistent volumes.

For more information on HPE Nimble Storage data protection for Red Hat OpenShift Container Platform, refer to the backup and recovery document at github.com/HewlettPackard/hpe-solutions-openshift/tree/master/synergy/scalable/backup_and_recovery.

Architecture design

Figure 14 highlights the solution layout of storage resources at a high level. This includes a reflection of the relationship between hosts, OS/hypervisor, boot volumes, and SAN storage.

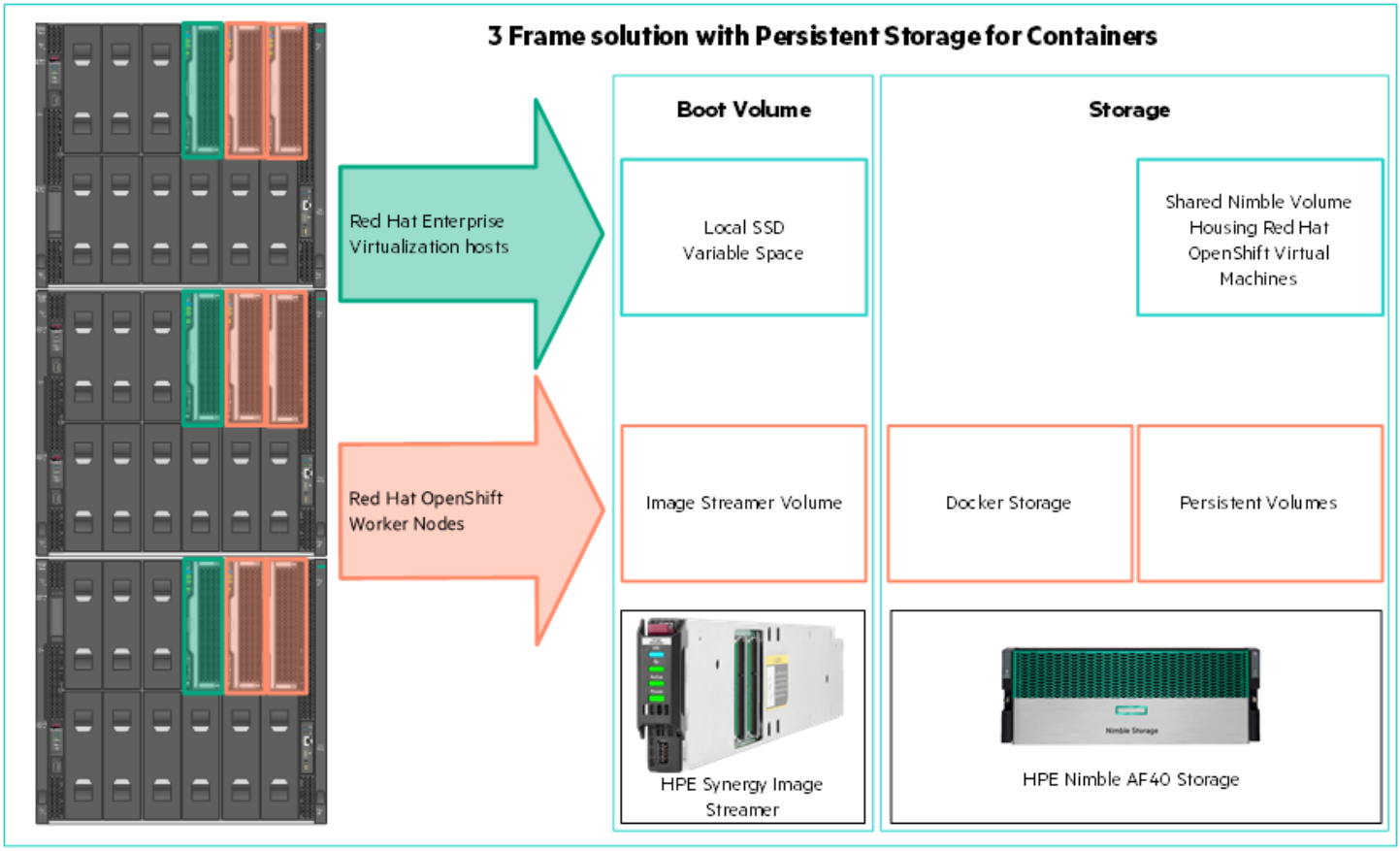

Figure 14. High level solution layout of storage resources

The solution environment assumes the presence of certain products and services to ensure proper functionality. Table 2 lists the products and services utilized in the creation of this solution and provides a high-level explanation of their function.

Table 2. Products and services used in the creation of this solution.

Service	Description/Notes
Ansible Engine	Provides Ansible engine to run playbooks.
DNS	Provides name resolution on management and data center networks, optionally on iSCSI networks.
DHCP	Provides IP address leases on PXE, management and usually for data center networks. Optionally used to provide addresses on iSCSI networks.
TFTP/PXE	Required to provide network boot capabilities to virtualized hosts.
NTP	Required to ensure consistent time across the solution stack
Active Directory/LDAP	May be used for authentication functions on various networks.

Figure 15 shows the configuration of the racks used to build this solution. For simplicity, the master and remote HPE Nimble Storage arrays are shown in the same rack. In a production environment, storage would be separated into separate physical racks with physical location being determined by the level of protection desired.

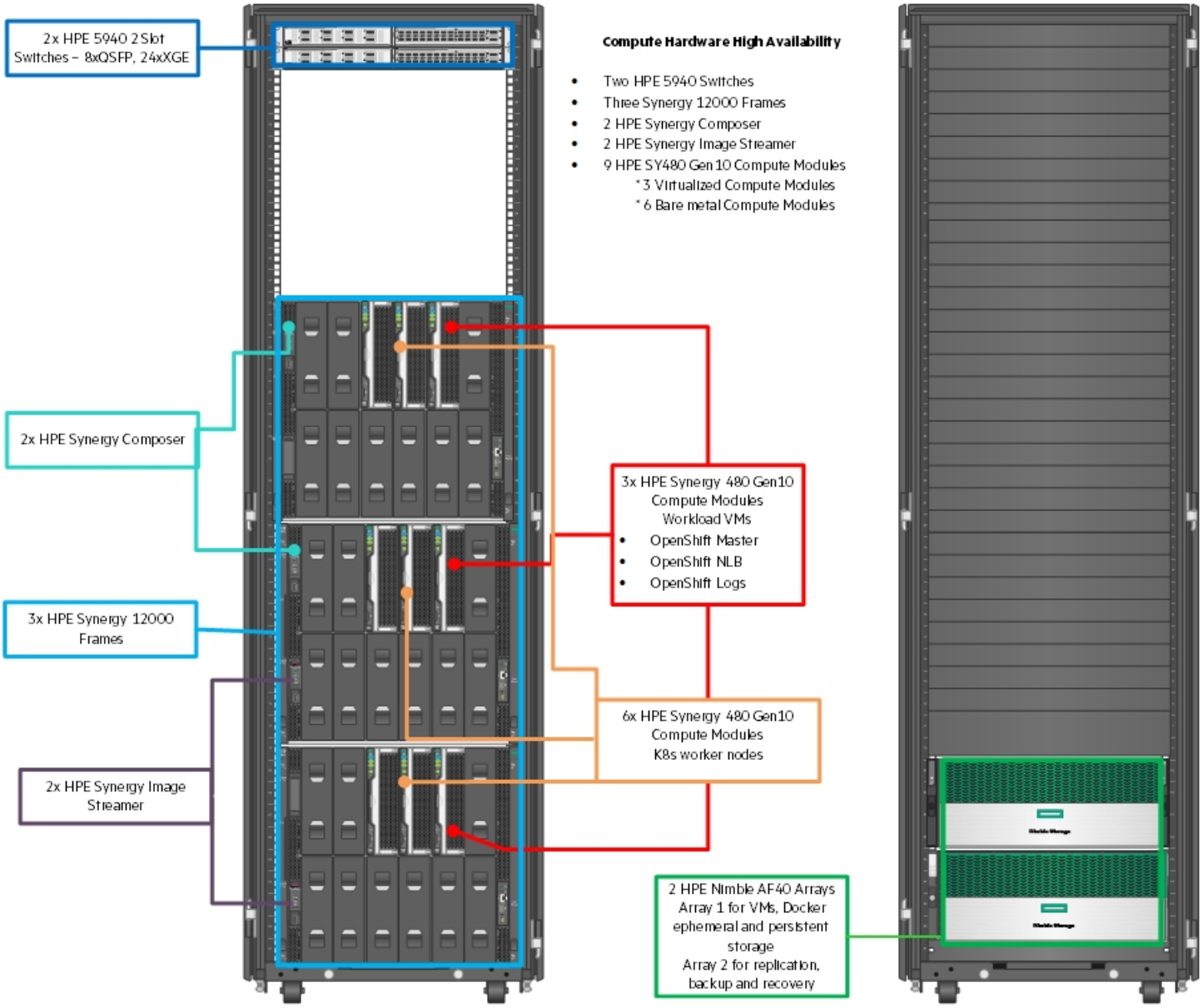

Figure 15. Physical layout of the compute within the solution

For detailed instructions on installing Red Hat OpenShift on HPE Synergy and HPE Nimble refer to the deployment guide document at github.com/hewlettpackard/hpe-solutions-openshift/tree/master/synergy/scalable/nimble.

Solution components

Hardware

The following hardware components were utilized in this Reference Architecture as listed in Table 3.

Table 3. Components utilized in the creation of this solution

Component	QTY	Description
HPE Synergy 12000 Frame	3	HPE Synergy 12000 Frames house the infrastructure used for the solution
• HPE Synergy Composer2	2	HPE Synergy Composers for core configuration and lifecycle management for the Synergy Components
• HPE Synergy Image Streamer	2	HPE Synergy Image Streamers provide Red Hat Enterprise Linux to the worker nodes in the solution
• HPE Virtual Connect SE 40Gb F8 Module	2	HPE Virtual Connect SE 40Gb F8 Modules provide network connectivity into and out of the frames
• HPE Synergy 20Gb Interconnect Link Module	4	HPE Synergy 20Gb Interconnect Link Modules provide network connectivity between frames
HPE Synergy 480 Gen10 Compute Module	9	Three virtualized management hosts and six bare metal or virtualized hosts for worker nodes
HPE FlexFabric 2-Slot Switch	2	Each switch contains one each of the HPE 5940 modules listed below
• HPE 5930 24p SFP+ and 2p QSFP+ Module	2	One module per HPE FlexFabric 2-Slot Switch
• HPE 5930 8p QSFP+ Module	2	One module per HPE FlexFabric 2-Slot Switch
HPE Nimble Storage AF40	2	One array for virtual machines, Docker storage and persistent volumes, one array for replication, backup and recovery of configuration files and persistent data

HPE Synergy

HPE Synergy, the first platform built from the ground up for composable infrastructure, empowers IT to create and deliver new value instantly and continuously. This single infrastructure reduces operational complexity for traditional workloads and increases operational velocity for the new breed of applications and services. Through a single interface, HPE Synergy composes compute, storage, and fabric pools into any configuration for any application. It also enables a broad range of applications from bare metal to virtual machines to containers, and operational models like hybrid cloud and DevOps. HPE Synergy enables IT to rapidly react to new business demands.

HPE Synergy Frames contain a management appliance called the HPE Synergy Composer2 which hosts HPE OneView. HPE Synergy Composer2 manages the composable infrastructure and delivers:

- Fluid pools of resources, where a single infrastructure of compute, storage and fabric boots up ready for workloads and demonstrates self-assimilating capacity.
- Software-defined intelligence, with a single interface that precisely composes logical infrastructures at near-instant speeds; and demonstrates template-driven, frictionless operations.
- Unified API access, which enables simple line-of-code programming of every infrastructure element; easily automates IT operational processes; and effortlessly automates applications through infrastructure deployment.

HPE Synergy Composer2 provides the enterprise-level management to compose and deploy system resources to meet your application needs. This management appliance uses software-defined intelligence to aggregate compute, storage, and fabric resources in a manner that scales to your application needs, instead of being restricted to the fixed ratios of traditional resource offerings. HPE Synergy template-based provisioning enables fast time to service with a single point for defining compute module state, pooled storage, network connectivity, and boot image.

HPE OneView is a comprehensive unifying platform designed from the ground up for composable infrastructure management. A unifying platform increases the productivity of every member of the internal IT team across servers, storage, and networking. By streamlining processes, incorporating best practices, and creating a new holistic way to work, HPE OneView provides organizations with a more efficient way to work. It is designed for open integration with existing tools and processes to extend these efficiencies.

HPE OneView is instrumental for the deployment and management of HPE servers and enclosure networking. It collapses infrastructure management tools into a single resource-oriented architecture that provides direct access to all logical and physical resources of the solution. Logical resources include server profiles and server profile templates, enclosures and enclosure groups, and logical interconnects and logical interconnect groups. Physical resources include compute modules, interconnects and storage modules.

The HPE OneView offers a uniform way for administrators to interact with resources by providing a RESTful API foundation. The RESTful APIs enable administrators to utilize a growing ecosystem of integrations to further expand the advantages of the integrated resource model that removes the need for the administrator to enter and maintain the same configuration data more than once and keep all versions up to date. It encapsulates and abstracts many underlying tools behind the integrated resource model, so the administrator can operate with new levels of simplicity, speed, and agility to provision, monitor, and maintain the solution.

Within the context of the solution, HPE OneView for Synergy is utilized to:

- Configure the profiles of the HPE Synergy Compute Modules.
- Apply and maintain compliance for firmware across the HPE Synergy infrastructure.
- Configure networking from the HPE Synergy Compute Modules to internal and outbound destinations.

HPE Synergy Image Streamer

HPE Synergy Image Streamer implements rapid image/application changes to multiple compute modules in an automated manner. HPE Synergy Image Streamer works with HPE Synergy Composer2 to rapidly deploy and update multiple physical compute modules. Operating environment images for bare-metal use might boot directly into a running OS, or VM hosts might perform quick image changeovers. "Infrastructure-as-code" capability enables fast delivery of applications and services, including the ability to perform rapid workload switching (using Linux, VMware® ESX, or Microsoft® Windows®). Enhanced profiles provide true stateless images, which integrate the server hardware configuration with operating environment images. Enhanced profiles are stored in redundant image repositories and are automatically integrated for simplicity of use. The unified API enables integration, automation, and customization of operations and applications with HPE Synergy Image Streamer.

HPE Synergy Image Streamer was used in this solution to provide Red Hat Enterprise Linux images to the OpenShift Container Platform worker nodes.

HPE Synergy 480 Gen10

The HPE Synergy 480 Gen10 Compute Module delivers an efficient and flexible two-socket server to support the most demanding workloads. Powered by Intel® Xeon® Scalable Family of processors, up to 3TB DDR4, and large storage capacity within a composable architecture. HPE Synergy 480 Gen10 Compute Module:

- Is the most secure server with exclusive HPE Silicon Root of Trust. Protect your applications and assets against downtime associated with hacks and viruses.
- Offers customer choice for greater performance and flexibility with Intel Xeon Scalable Family of processors on the Synergy 480 Gen10 architecture
- Offers Intelligent System Tuning with processor smoothing and workload matching to improve processor throughput/overall performance up to 8% over previous generation.
- Features a maximum memory footprint of 3TB for large in-memory database and analytic applications.
- Features a hybrid HPE Smart Array for both RAID and HBA zoning in a single controller.

The HPE Synergy 480 Gen10 provides the needed compute to power this solution running both Red Hat Virtualization for the core management pieces of Red Hat OpenShift and Red Hat Enterprise Linux to host the worker nodes.

The bill of materials found in Appendix A of this document outlines the configuration of the HPE Synergy Compute Modules used in this solution.

HPE Nimble Storage and HPE InfoSight

HPE Nimble Storage All Flash Arrays combine a flash-efficient architecture with HPE InfoSight predictive analytics to achieve fast, reliable access to data and 99.9999% guaranteed availability.[7] Radically simple to deploy and use, the arrays are cloud-ready, providing data mobility to the cloud through HPE Cloud Volumes. Your storage investment made today will support you well into the future, thanks to the technology and business-model innovations. HPE Nimble Storage All Flash Arrays include all-inclusive licensing, and easy upgrades, while also being future-proofed for new technologies, such as NVMe and SCM.

[7] HPE Get Six Nines Guarantee: https://h20195.www2.hpe.com/v2/GetPDF.aspx/4AA5-2846ENN.pdf

HPE InfoSight is an industry-leading predictive analytics platform that brings software-defined intelligence to the data center with the ability to predict and prevent infrastructure problems before they happen. Features include:

By learning from problems across the entire HPE storage install base, HPE InfoSight can predict and prevent the same problems from occurring on other arrays.

- Utilizing cloud-based predictive analytics, HPE InfoSight can predict and resolve 86% of problems before they are "discovered" by the end user, of which 54% are not directly related to the storage array.
- HPE InfoSight analyzes storage and server performance patterns, and can recommend optimizations automatically, taking away a lot of the guesswork or trial and error of manual optimizations.
- Using trends over time, HPE InfoSight can accurately predict future capacity, performance, and bandwidth needs.

HPE Nimble Storage abstracts a rich feature set to Red Hat OpenShift Container Platform to enable a breadth of modern use cases. Features are exposed as parameters to a Kubernetes StorageClass that users may leverage through Persistent Volume Claims.

Features include:

- Advanced lifecycle controls to enable ephemeral clones and volume annotation through custom metadata
- Application optimized performance policies and quality of service (QoS) controls
- Volume placement directives, per cluster or StorageClass, to enforce performance caps, storage capacity or media (Flash or Hybrid)
- Protection templates to fulfill data protection SLA/SLO defined by the business, including replication to HPE Cloud Volumes
- Provision data at rest (DAR) encrypted volumes to meet compliance and regulations
- Satisfy a wide range of stateful application by provisioning volumes up to 127 TB – either thick or thin-provisioned
- Toggle data reduction features such as variable block size, deduplication, and adaptive compression per workload
- Enable users to clone and restore Persistent Volume Claims using native tools and APIs
- Import existing Nimble volumes into a container from traditional virtualized environments when modernizing applications

The bill of materials found in Appendix A of this document outlines the configuration of the HPE Nimble Storage used in this solution.

Red Hat Software

Red Hat Enterprise Linux

Red Hat Enterprise Linux Server powers the applications that run your business with the control, confidence, and freedom that comes from a consistent foundation across hybrid cloud deployments. As the premier platform provider for enterprise workloads, Red Hat works side by side with engineers from major hardware vendors and cloud providers to make sure that the operating system takes full advantage of the latest innovations. This leadership with partners, as well as Red Hat's influence and contributions to upstream communities, provides a stable, secure, and performance driven foundation for the applications that run the business of today and tomorrow. Red Hat Enterprise Linux is at the core of this solution, each of the Red Hat OpenShift Container Platform control plane nodes running as virtual machines are running Red Hat Enterprise Linux. This solution is designed to use dedicated physical servers for the Kubernetes compute nodes. Each Kubernetes compute node is a dedicated SY480 Compute Module running Red Hat Enterprise Linux 7.6 for maximum application and pod scalability.

Red Hat Virtualization

Red Hat Virtualization is an open, software-defined, efficient platform for virtualized Linux and Microsoft Windows workloads, built on Red Hat Enterprise Linux and Kernel-based Virtual Machine (KVM) technologies. Automated management, scaling, and security features let you build a streamlined, reliable environment for virtualized applications and cloud-native workloads. Virtual-to-virtual (V2V) conversion tooling lets you easily migrate workloads from other hypervisors. A RESTful application programming interface (API) lets you integrate Red Hat Virtualization into your existing infrastructure and easily add innovative new technologies as business requirements change.[8]

This solution is designed to optimize the use of physical resources by using a combination of virtual machines and physical servers. Red Hat Virtualization provides a robust highly available virtualization platform for the Red Hat OpenShift Container Platform virtual machines. The Red

[8] redhat.com/cms/managed-files/vi-virtualization-datasheet-f12281-201805-en.pdf

Hat OpenShift Container Platform control plane, including the master, etcd, infrastructure, and load balancer roles, is deployed on virtual machines that are distributed across a three node RHV cluster.

Red Hat OpenShift Container Platform

Red Hat OpenShift Container Platform unites developers and IT operations on a single platform to build, deploy, and manage applications consistently across hybrid cloud and multi-cloud infrastructures. Red Hat OpenShift helps businesses achieve greater value by delivering modern and traditional applications with shorter development cycles and lower operating costs. Red Hat OpenShift is built on open source innovation and industry standards, including Kubernetes and Red Hat Enterprise Linux, the world's leading enterprise Linux distribution.[9]

Figure 16 highlights how the individual Red Hat OpenShift Container Platform pieces are laid out within the solution.

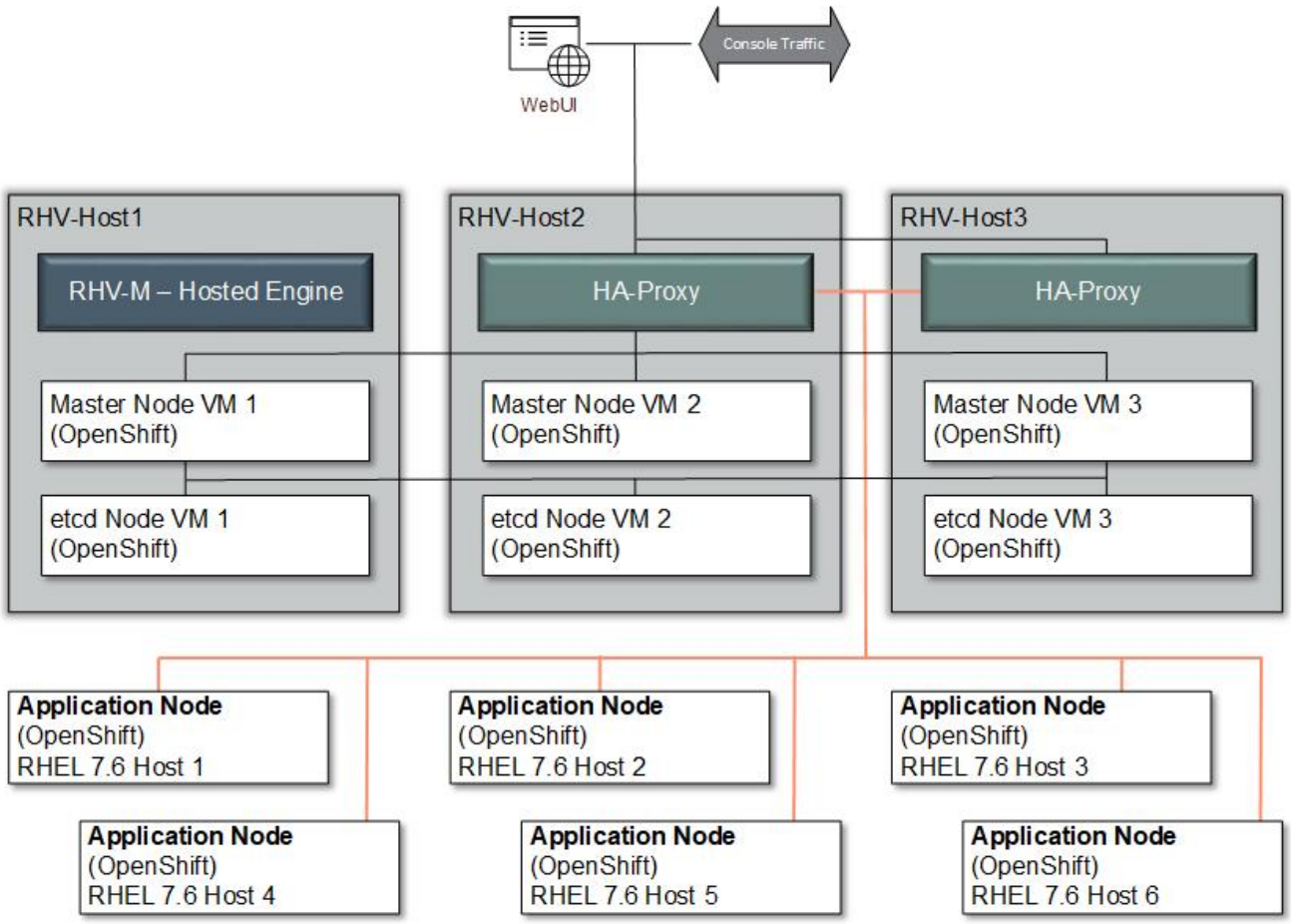

Figure 16. Red Hat OpenShift Container Platform solution component layout

Best practices and configuration guidance for the solution

Red Hat OpenShift virtual machines

Red Hat OpenShift Container Platform requires a number of redundant functions. These functions may be hosted on either physical compute modules or on virtual machines, both of which run Red Hat Enterprise Linux 7.6. For this solution, Hewlett Packard Enterprise chose to implement these functions as virtual machines. This approach reduces the amount of infrastructure required while introducing enhanced options for management and high availability. Three HPE Synergy 480 Gen10 Compute Modules host the virtual machines shown in Figure 17. This

[9] redhat.com/cms/managed-files/cl-openshift-container-platform-datasheet-f8610kc-201708-en_0.pdf

figure also shows that the worker nodes run on bare metal. While the solution was tested with six worker nodes, it is scalable to include many more.

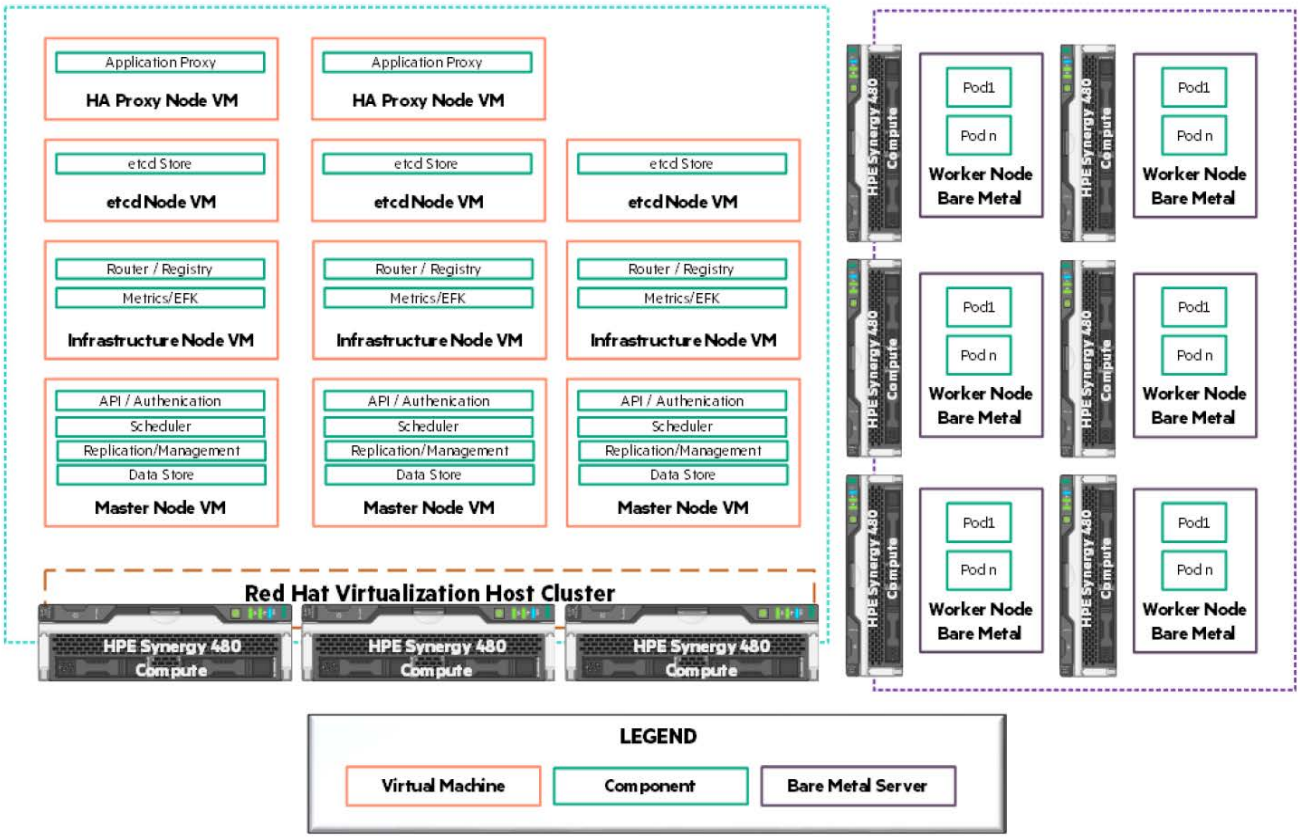

Figure 17. Red Hat OpenShift implementation with virtual machines for core management functions

Software versions

This section describes the software versions utilized in the solution. Table 4 lists the versions of required software used in the creation of this solution at the time of testing. As newer versions are released the installer should verify compatibility with other software component versions.

Table 4. Major software versions used in solution creation.

Component	Version
Red Hat Enterprise Linux Server	7.6
Red Hat Virtualization, Red Hat Virtualization Manager	4.2
Red Hat OpenShift Container Platform	3.11
HPE Nimble Storage Linux Toolkit	2.4.1.13
Nimble Kube Storage Controller	2.4.1
SAP Data Hub	2.6.99
SAP Data Hub Launchpad	2.6.17
Kubernetes	1.11.0
Docker	1.13.1

Capacity and sizing

SAP Data Hub system sizing

Sizing for an SAP Data Hub environment varies depending on the requirements of the specific organization and type of deployment. This section discusses sizing considerations for SAP Data Hub running on Red Hat OpenShift Container Platform. SAP Data Hub capabilities can be partitioned into three functional areas, data governance, distributed data management, and data pipelines and workflows. As mentioned earlier, SAP Data Hub provides data governance by helping you to manage and gather information about the location, attributes, quality, and sensitivity of data across different systems by using the Metadata Explorer. SAP Data Hub also contains a distributed database system for big data processing, the SAP Vora database, providing distributed data management. The SAP Data Hub Modeler is used to build data pipelines and data workflows for data ingestion and transformation.

While each of these functional areas has different sizing requirements, sizing SAP Data Hub is primarily based on data volume and workload characteristics. Data volume includes data and metadata stored in SAP Data Hub as well as data processed but not stored in SAP Data Hub. Workload characteristics can include rate of data ingestion, types of queries on data stored in SAP Data Hub, complexity of data pipelines and workflows, and the amount of profiling, indexing, and preparation jobs. The number of concurrent users can also effect the amount of memory and processing power required.

When sizing an SAP Data Hub solution, in addition to the functional area components one must consider system components that are independent of the functional areas. These include components such as the log viewer, connection manager, launchpad, etc. The total amount of memory, CPU, disk, # of pods, and # of IPs required is a summation of the amounts needed for system components, data governance, distributed data management, and data pipeline and workflows. Table 5 lists the formulas to calculate the solution hardware requirements.

Table 5. Formulas for calculating SAP Data Hub hardware requirements

Resource	Formula
RAM	= RAM for system components + RAM for data governance + RAM for distributed data management + RAM for data pipeline and workflows
CPU	= CPU for system components + CPU for data governance + CPU for distributed data management + CPU for data pipeline and workflows
Disk	= Disk for system components + Disk for data governance + Disk for distributed data management + Disk for data pipeline and workflows
# of Pods	= Pods for system components + Pods for data governance + Pods for distributed data management + Pods for data pipeline and workflows
# of IPs	= IPs for system components + IPs for data governance + IPs for distributed data management + IPs for data pipeline and workflows
Ephemeral Disk	Number of Kubernetes nodes * 100 GB
Checkpoint Store	3.3 * Streaming Data Footprint / Compression Factor
Container Registry Free Space	60 GB

Sizing the system components is based on the number of tenants, number of users across all tenants, and the number of Kubernetes nodes in the cluster that will be used for SAP Data Hub. The maximum number of tenants per SAP Data Hub system is 25. Table 6 lists the formulas to calculate the hardware requirements for the system components.

Table 6. Formulas for calculating system component hardware requirements

Resource	Formula
RAM	= Number of tenants (not including the system tenant) * 3,000 MB + Number of users * 2,900 MB + Number of Kubernetes nodes * 400 MB + 38,100 MB
CPU	= Number of tenants (not including the system tenant) * 0.9 + Number of users * 0.5 + Number of Kubernetes nodes * 0.6 + 8,7
Disk	= Number of Kubernetes nodes * 10 GB + 200 GB
# of Pods	= Number of tenants (not including the system tenant) * 5 + Number of users * 5 + Number of Kubernetes nodes * 3 + 20
# of IPs	= Number of tenants (not including the system tenant) * 10 + Number of users * 10 + Number of Kubernetes nodes * 3 + 44
Ephemeral Disk	= Number of Kubernetes nodes * 100 GB

Sizing the data governance functional area is based on the number concurrently running jobs and how many of those jobs are for profiling, indexing / publishing, and preparation. Table 7 lists the formulas to calculate the hardware requirements for the data governance functional area.

Table 7. Formulas for calculating data governance hardware requirements

Resource	Formula
RAM	= Number of profiling jobs * 3,500 MB + (Number of indexing and publishing jobs + Number of preparation jobs) * 1,000 MB
CPU	CPU = Number of profiling jobs * 2.0 + (Number of indexing and publishing jobs + Number of preparation jobs) * 1.4
Disk	= 0
# of Pods	= Number of pods = 2 * Number of profiling jobs + Number of indexing and publishing jobs + Number of preparation jobs
# of IPs	= Number of IPs = 2 * Number of profiling jobs + Number of indexing and publishing jobs + Number of preparation jobs

The SAP Vora database is the distributed database system for SAP Data Hub. It supports two table types, data source and streaming. Data source tables are modified by adding and loading data sources from shared storage locations. Streaming tables are modified by INSERT, UPSERT, UPDATE and DELETE statements. They do not support loading of data sources. The content of streaming tables is stored in the distributed log and checkpoint store. Sizing the distributed data management functional area is based on the amount of data stored in the relational disk engine, how much of this data is in streaming tables, the ingestion rate into the streaming tables, and the compression factor for the data. Table 8 lists the formulas to calculate the hardware requirements for the distributed data management functional area.

Table 8. Formulas for calculating distributed data management hardware requirements

Resource	Formula
RAM	= RAM (in MB) for Relational Disk Engine + RAM (in MB) for Distributed Log + RAM (in MB) for Miscellaneous Components
CPU	CPU for Relational Disk Engine + CPU for Distributed Log + CPU for Miscellaneous Components
Disk	= Disk (in GB) for Relational Disk Engine + Disk (in GB) for Distributed Log + Disk (in GB) for Miscellaneous Components
# of Pods	= Pods for Relational Disk Engine + Pods for Distributed Log + Pods for Miscellaneous Components
# of IPs	= IPs for Relational Disk Engine + IPs for Distributed Log + IPs for Miscellaneous Components
Checkpoint Store	= 3.3 * Streaming Data Footprint / Compression Factor

Sizing the data pipelines and workflows functional area is based on the number of pipelines and workflows and their complexity. Data pipelines process data, receiving and sending data. Data workflows orchestrate tasks using workflow operators which send and receive status messages. Table 9 lists the formulas to calculate the hardware requirements for the data pipelines and workflows functional area.

Table 9. Formulas for calculating data pipelines and workflows hardware requirements

Resource	Formula
RAM	= Number of low complexity pipelines * 500 MB + Number of medium complexity pipelines * 1,000 MB + Number of high complexity pipelines * 1,500 MB + 2,000 MB
CPU	= Number of low complexity pipelines * 0.5 + Number of medium complexity pipelines * 1.0 + Number of high complexity pipelines * 1.5 + 2
Disk	= 0
# of Pods	= Number of low complexity pipelines + Number of medium complexity pipelines + Number of high complexity pipelines + 1
# of IPs	= Number of low complexity pipelines + Number of medium complexity pipelines + Number of high complexity pipelines + 2

Minimum Sizing Requirements

A minimum SAP Data Hub deployment on Red Hat OpenShift Container Platform requires a minimum of 4 Kubernetes worker nodes for a production environment and 3 for a test or development environment. The SAP Data Hub installation requires an image registry where images are first mirrored from an SAP registry and then delivered to the OCP cluster nodes. The integrated OCP registry is not appropriate for this purpose and an external image registry needs to be setup. The server used for the external registry could be a server used for one of the required OCP deployment services listed in table 1 or an additional server.

The recommended minimum SAP Data Hub sizing is listed in table 10.

Table 10. Minimum SAP Data Hub deployment sizing

	Test / Development Environment	Production Environment
Kubernetes worker nodes	Minimum 3	Minimum 4
RAM	96 GB (32 GB per worker node)	256 GB (64 GB per worker node)
CPU	24 (8 CPUs per worker node)	64 (16 CPUs per worker node)
Disk	500 GB	2,100 GB
External storage for Checkpoint Store	55 GB	5,500 GB
Ephemeral disk	100 GB per worker node	100 GB per worker node
Container registry free space	Minimum 60 GB	Minimum 60 GB

Potential Configurations for SAP Data Hub

In lieu of calculating the hardware requirements for each of the functional areas and the system components, the potential configurations listed in table 11 could be a good starting point. Table 10 lists small, medium, and large potential configurations for SAP Data Hub. These are general use configurations that can give you a rough estimate of what you may need for your specific business requirements.

Table 11. Potential configurations for SAP Data Hub

Configuration	Data Volume (TB)	Number of concurrent pipelines	Number of concurrent users	RAM (GB)	CPUs	Nodes	Disk (TB)	Checkpoint Store (TB)
Small	5	3	30	256	64	4	2.1	5.5
Medium	50	150	50	1,152	288	18	17.6	55
Large	100	300	100	2,240	560	35	35	110

In addition to the hardware requirements in table 10, each worker node needs at least 100 GB ephemeral disk and a container registry with at least 60 GB of free space.

These configurations assume the following:

- The data volume is the uncompressed data volume stored in the Relational Disk Engine in data source tables and/or streaming tables
- That each worker node has 64GB main memory and 16 CPUs
- Disk refers to the size of the persistent volumes
- Checkpoint Store refers to the size of the external storage and is only needed for streaming tables
- Compression factor of the disk engine is assumed to be 3 and the maximum ingestion rate for streaming tables is assumed to be 50 MB/s

Mapping the potential configurations listed in table 10 to HPE Synergy 480 Gen 10 Compute Module resources can be done easily. Let's assume that each compute module contains the processor, memory, and disk resources listed in table 12.

Table 12. Example HPE Synergy 480 Gen10 Compute Module components

Quantity	Description	
2	Intel Xeon Gold 6242 (2.8GHz/16-core/150W)	32 total CPU cores
4	HPE 32GB 2Rx4 PC4-2933Y-R	128 GB total memory
1	HPE Smart Array P416ie-m SR Gen10 12G SAS Mezzanine controller	
2	HPE 480GB SATA 6G Mixed Use SFF SSD	

Using this example configuration each compute module contains 32 processor cores and 128 GB of memory. Table 12 lists the calculated node count required based on the example compute module configuration in table 13. The small configuration requires only 2 nodes to meet the processor core and memory requirements but OCP requires a minimum of 3 or 4 depending on the type of environment deployed.

Table 13. Example configuration required resources and node count

Configuration	Required RAM (GB)	Required CPU Cores	Required Storage (TB)	Nodes	Notes
Small	256	64	12.6	3 or 4	Minimum of 3 for test/development or 4 for production environment
Medium	1,152	288	122.6	9	
Large	2,240	560	245	18	

The storage architecture use in this solution provides support for both ephemeral and persistent container volumes on the HPE Nimble Storage All Flash Arrays. These arrays can be configured with different capacity drives and additional expansion shelves if desired. Each of the potential configuration storage requirements can be met using the HPE Nimble Storage arrays. They can be configured to meet your capacity, expansion, and data protection needs.

Table 14. HPE Nimble Storage All Flash Arrays

HPE Nimble Storage Model	Raw capacity (TB)	Max Expansion shelves
AF20	11-46	1
AF40	11-184	1
AF60	11-553	2
AF80	23-1,106	2

For additional detail on configuring HPE Nimble Storage All Flash Arrays refer to hpe.com/us/en/storage/nimble.html.

For more detailed information on sizing SAP Data Hub refer to the Sizing Guide for SAP Data Hub.

SAP Data Hub Licensing

SAP Data Hub licensing is based on Data Hub Units. The number of Data Hub Units is equal to the number of connected systems and the number of distributed runtime nodes. Connected systems include big data storage (Hadoop, Amazon S3, SAP HANA, etc.) Distributed runtime nodes can be physical or virtual and each 256GB of RAM is considered a Data Hub Unit.

Figure 18 below explains the current SAP Data Hub licensing model. Contact SAP for further SAP Data Hub licensing details, https://www.sap.com/products/data-hub.html#pricing-packaging.

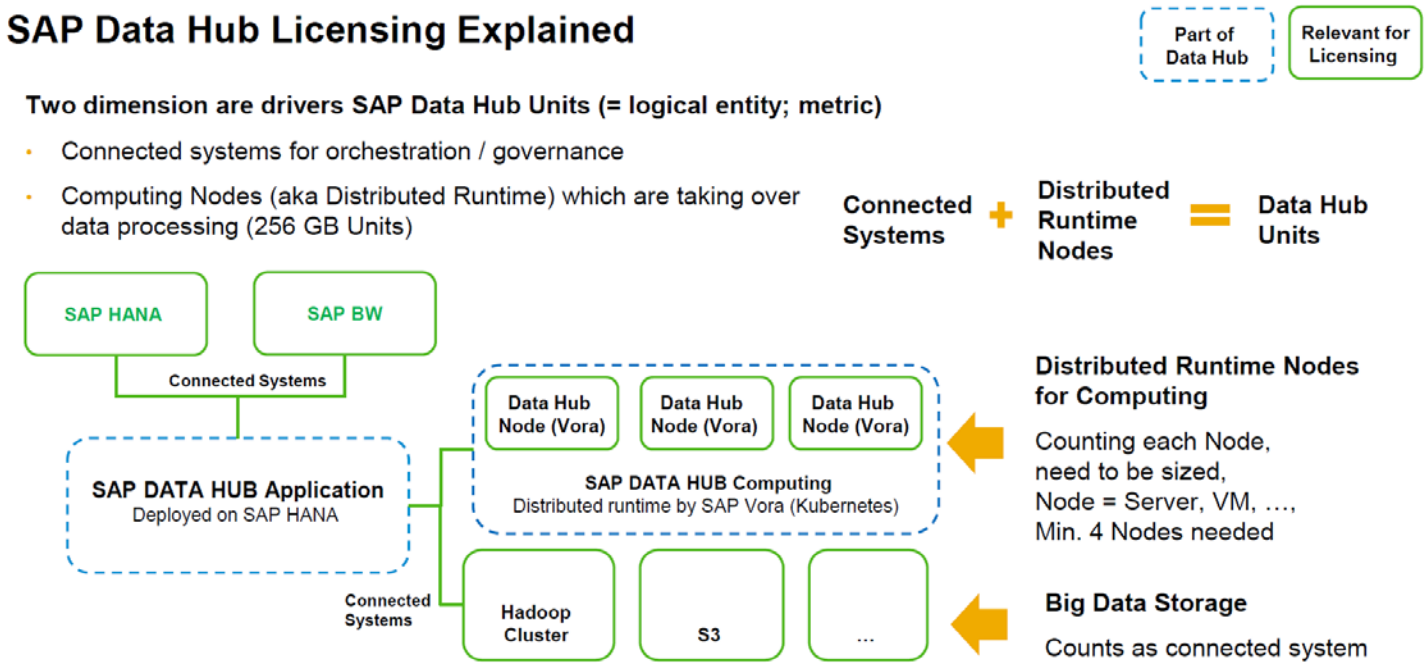

Figure 18. SAP Data Hub licensing model

The sizing considerations for Red Hat OpenShift Container Platform should also be considered.

Red Hat OpenShift Container Platform role sizing

- Master – The minimum size for a physical or virtual machine running the master node is 4 vCPU and 16 GB RAM with a 40 GB disk space for /var, 1 GB disk space for /usr/local/bin, and 1 GB disk space for the system's temporary directory. Master nodes should be configured with an additional 1 CPU core and 1.5 GB RAM for each additional 1,000 pods. Three master node virtual machines were deployed for this solution.
- Infrastructure nodes – Application and infrastructure nodes require a minimum of 1 vCPU and 8 GB RAM. The node requires at least 15 GB of disk space for /var/, 1 GB of disk space for /usr/local/bin, and 1 GB of disk space for the system's temporary directory. A total of three infrastructure nodes were deployed for this solution.
- etcd – etcd nodes should be configured with a minimum of 4 vCPU, 15 GB RAM, and 20 GB for etcd data. A total of three etcd nodes were deployed for this solution.
- HAProxy – A total of two HAProxy load balancer VMs were deployed for this solution. Default values for CPU and memory were utilized. In a production environment, the expectation is that a dedicated software or hardware load balancing solution will be implemented.

Red Hat OpenShift Container Platform cluster sizing

The number of application nodes in an OpenShift Cluster depends on the number of pods that an organization is planning to deploy. Red Hat OpenShift Container Platform can support the following:

- Maximum of 2000 nodes per cluster
- Maximum of 150,000 pods per cluster
- Maximum of 250 pods per node
- Maximum of 10 pods per CPU core

To determine the number of nodes required in a cluster, estimate the number of pods the organization is planning to deploy and divide by the maximum number of pods per node. For example, if the organization expects to deploy 5000 pods, then the organization should expect to deploy 20 application nodes with 250 pods per node (5000 / 250 =20). In this environment with a default configuration of six physical application nodes, the Red Hat OpenShift Cluster should be expected to support 1500 pods (250 pods x 6 nodes = 1500 pods).

For more information about Red Hat OpenShift Container Platform sizing, refer to the Red Hat OpenShift Container Platform Scaling and Performance Guide which can be found at: https://docs.openshift.com/container-platform/3.11/scaling_performance/index.html.

Summary

The complexity of enterprise landscapes, which can now include Hadoop data lakes, EDWs, Cloud storage, etc., makes the ability to appropriately provide effective data governance difficult. Organizations need to be able to trust and rely on their data's accuracy, or they create risk for anyone using analytics or operational applications that use the data. SAP Data Hub can ensure data reliability, traceability, and compliance in accordance with your business. It delivers a simpler, more scalable approach to data landscape management. With enterprise-spanning data integration, processing, and governance, SAP Data Hub provides unprecedented visibility into and access across the complex network of data in the modern enterprise. It is a data sharing, pipelining, and orchestration solution that helps your company to accelerate and expand the flow of data across your diverse data landscape. Metadata assets that are spread across diverse systems and disparate sources are governed and managed through the SAP Data Hub Metadata Explorer. Providing policy management and security features for data protection.

Red Hat OpenShift Container Platform on HPE Synergy provides an end-to-end fully integrated container solution that, once assembled, can be configured within hours. This eliminates the complexities associated with implementing a container platform across an enterprise data center and provides the automation of hardware and software configuration to quickly provision and deploy a containerized environment at scale. Red Hat OpenShift Container Platform provides organizations with a reliable platform for deploying and scaling container-based applications. HPE Synergy provides the flexible infrastructure you need to run the container platform to dynamically provision and scale applications.

The HPE Synergy Composable Infrastructure solution provides a layered view of security controls. The objective of choosing this layered security view is to ensure that customers become aware of the depth of security risk that an infrastructure can have and also make them aware of the depth of defense that is built into the HPE design.

Using an enterprise grade storage solution such as HPE Nimble Storage for persistent storage with containers enables speed, portability, and agility for traditional enterprise applications and data. The HPE Nimble storage array's data replication for protection and disaster recovery helps in faster and affordable recovery.

Appendix A: Bill of materials

The following bill of materials contains the core components utilized in the creation of this solution. Services, support, and software are not included in the BOM and should be customized based on customer needs.

Note

Part numbers are at time of publication/testing and subject to change. The bill of materials does not include complete support options or other rack and power requirements. If you have questions regarding ordering, please consult with your HPE Reseller or HPE Sales Representative for more details. hpe.com/us/en/services/consulting.html

Table A1. Bill of materials

Qty	Part number	Description
		Rack and Network Infrastructure
2	P9K10A	HPE 42U 600mmx1200mm G2 Kitted Advanced Shock Rack with Side Panels and Baying
2	P9K10A 001	HPE Factory Express Base Racking Service
2	H6J85A	HPE Rack Hardware Kit
2	BW932A	HPE 600mm Rack Stabilizer Kit
2	BW932A B01	HPE 600mm Rack include with Complete System Stabilizer Kit
6	AF533A	HPE Intelligent Modular 3Ph 14.4kVA/CS8365C 40A/208V Outlets (6) C19/Horizontal NA/JP PDU
2	H6J85A	HPE Rack Hardware Kit
		HPE Synergy Composable Infrastructure
3	797740-B21	HPE Synergy 12000 Configure-to-order Frame with 1x Frame Link Module 10x Fans
4	779218-B21	HPE Synergy 20Gb Interconnect Link Module
2	794502-B23	HPE Virtual Connect SE 40Gb F8 Module for Synergy
2	804937-B21	HPE Synergy Image Streamer
3	798096-B21	HPE 6x 2650W Performance Hot Plug Titanium Plus FIO Power Supply Kit
2	804353-B21	HPE Synergy Composer
3	804938-B21	HPE Synergy Frame Rack Rail Kit
3	804942-B21	HPE Synergy Frame Link Module
1	813567-001	HPE Synergy Frame 4x Lift Handles
1	859493-B21	Synergy Multi Frame Master1 FIO
1	859494-B22	Synergy Multi Frame Master2 FIO
8	804101-B21	HPE Synergy Interconnect Link 3m Active Optical Cable
2	720199-B21	HPE BladeSystem c-Class 40G QSFP+ to QSFP+ 3m Direct Attach Copper Cable
2	861412-B21	HPE Synergy Frame Link Module CAT6A 1.2m Cable
1	861413-B21	HPE Synergy Frame Link Module CAT6A 3m Cable
		Virtualization hosts
3	871940-B21	HPE Synergy 480 Gen10 Configure-to-order Compute Module
3	873381-L21	HPE Synergy 480/660 Gen10 Intel Xeon-Gold 6130 (2.1GHz/16-core/125W) FIO Processor Kit
3	873381-B21	HPE Synergy 480/660 Gen10 Intel Xeon-Gold 6130 (2.1GHz/16-core/125W) Processor Kit

Qty	Part number	Description
54	815097-B21	HPE 8GB (1x8GB) Single Rank x8 DDR4-2666 CAS-19-19-19 Registered Smart Memory Kit
18	815098-B21	HPE 16GB (1x16GB) Single Rank x4 DDR4-2666 CAS-19-19-19 Registered Smart Memory Kit
6	875478-B21	HPE 1.92TB SATA 6G Mixed Use SFF (2.5in) SC 3yr Wty Digitally Signed Firmware SSD
3	P01367-B1	HPE 96W Smart Storage Battery (up to 20 Devices) with 260mm Cable Kit
3	804424-B21	HPE Smart Array P204i-c SR Gen10 (4 Internal Lanes/1GB Cache) 12G SAS Modular Controller
3	777430-B21	HPE Synergy 3820C 10/20Gb Converged Network Adapter
		Worker Nodes
6	871943-B21	HPE Synergy 480 Gen10 6130 2P 64GB-R P204i-c SAS Performance Compute Module
6	873381-L21	HPE Synergy 480/660 Gen10 Intel Xeon-Gold 6130 (2.1GHz/16-core/125W) FIO Processor Kit
6	873381-B21	HPE Synergy 480/660 Gen10 Intel Xeon-Gold 6130 (2.1GHz/16-core/125W) Processor Kit
108	815097-B21	HPE 8GB (1x8GB) Single Rank x8 DDR4-2666 CAS-19-19-19 Registered Smart Memory Kit
36	815098-B21	HPE 16GB (1x16GB) Single Rank x4 DDR4-2666 CAS-19-19-19 Registered Smart Memory Kit
6	P01367-B1	HPE 96W Smart Storage Battery (up to 20 Devices) with 260mm Cable Kit
6	804424-B21	HPE Smart Array P204i-c SR Gen10 (4 Internal Lanes/1GB Cache) 12G SAS Modular Controller
6	777430-B21	HPE Synergy 3820C 10/20Gb Converged Network Adapter
		Primary HPE Nimble Storage
1	Q8H41A	HPE Nimble Storage AF40 All Flash Dual Controller 10GBASE-T 2-port Configure-to-order Base Array
1	Q8B88B	HPE Nimble Storage 2x10GbE 2-port FIO Adapter Kit
1	Q8G27B	HPE Nimble Storage NOS Default FIO Software
1	Q8H47A	HPE Nimble Storage AF40 All Flash Array R2 11.52TB (24x480GB) FIO Flash Bundle
2	R0P84A	HPE Nimble Storage NEMA IEC 60320 C14 to C19 250V 15 Amp 1.8m FIO Power Cord
1	Q8F56A	HPE Nimble Storage 10GbE 2-port Spare Adapter
2	P9Q66A	HPE G2 IEC C20 Input/(8) C13 Expansion Outlets/PDU Extension Bar Kit
		Replication target HPE Nimble Storage
1	Q8H41A	HPE Nimble Storage AF40 All Flash Dual Controller 10GBASE-T 2-port Configure-to-order Base Array
1	Q8B88B	HPE Nimble Storage 2x10GbE 2-port FIO Adapter Kit
1	Q8G27B	HPE Nimble Storage NOS Default FIO Software
1	Q8H47A	HPE Nimble Storage AF40 All Flash Array R2 11.52TB (24x480GB) FIO Flash Bundle
2	R0P84A	HPE Nimble Storage NEMA IEC 60320 C14 to C19 250V 15 Amp 1.8m FIO Power Cord
1	Q8F56A	HPE Nimble Storage 10GbE 2-port Spare Adapter
2	P9Q66A	HPE G2 IEC C20 Input/(8) C13 Expansion Outlets/PDU Extension Bar Kit
		HPE 5940 FlexFabric Switching
2	JH397A	HPE FF 5940 2-Slot Switch
2	JH180A	HPE 5930 24p SFP+ and 2p QSFP+ Module
2	JH183A	HPE 5930 8-port QSFP+ Module
4	JG553A	HPE X712 Back (Power Side) to Front (Port Side) Airflow High Volume Fan Tray
4	JC680A	HPE 58x0AF 650W AC Power Supply
4	JC680A B2B	INCLUDED: Jumper Cable - NA/JP/TW

Qty	Part number	Description
2	JG326A	HPE X240 40G QSFP+ QSFP+ 1m DAC Cable
4	JG327A	HPE X240 40G QSFP+ QSFP+ 3m DAC Cable
		Red Hat OpenShift Container Platform
6	R1Z92AAE	Red Hat OpenShift Container Platform for HPE Synergy 1-32 Cores 1yr Subscription 24x7
		Red Hat Enterprise Linux Server
6	J8J36AAE	Red Hat Enterprise Linux Server 2 Sockets 1 Guest 1 Year Subscription 24x7 Support
		Red Hat Enterprise Linux for Virtual Datacenters
11	G3J22AAE	Red Hat Enterprise Linux for Virtual Datacenters 2 Sockets 1 Year Subscription 24x7 Support
		Red Hat Enterprise Virtualization
3	J1U48AAE	Red Hat Enterprise Virtualization 2 Sockets 1 Year Subscription 24x7 Support

Resources and additional links

Red Hat, www.redhat.com

Red Hat OpenShift Container Platform 3.11 Documentation, docs.openshift.com/container-platform/3.11/welcome/index.html

Red Hat OpenShift Container Platform Documentation on SAP Data Hub, access.redhat.com/articles/3630111

SAP Data Hub, help.sap.com/viewer/product/SAP_DATA_HUB/2.6.latest/en-US

SAP Data Hub Installation Guide, help.sap.com/viewer/e66c399612e84a83a8abe97c0eeb443a/2.6.latest/en-US/9f866d8ef9a94c30947f12e73eaf0dd9.html

Data Governance User Guide for SAP Data Hub, help.sap.com/viewer/bd9aff2447b24109bccbd7618cfd542e/2.6.latest/en-US/f1db03f08d994523b5d822a6c5257b48.html

SAP Data Hub Administration Guide, help.sap.com/viewer/1246f58c6a74412580877af1f484f41a/2.6.latest/en-US

SAP Data Hub Sizing Guide, Sizing Guide for SAP Data Hub

HPE Synergy, https://www.hpe.com/info/synergy

HPE Nimble Storage, hpe.com/us/en/storage/nimble.html

HPE Solutions for OpenShift GitHub, github.com/hewlettpackard/hpe-solutions-openshift

HPE FlexFabric 5940 switching, www.hpe.com/us/en/product-catalog/networking/networking-switches/pip.hpe-flexfabric-5940-switch-series.1009148840.html

HPE Reference Architectures, hpe.com/info/ra

HPE Servers, hpe.com/servers

HPE Storage, hpe.com/storage

HPE Networking, hpe.com/networking

HPE Technology Consulting Services, hpe.com/us/en/services/consulting.html

a50000637enw, December 2019