Dell Technologies Validated Platform for Kamu Data Fabric

Abstract

This Dell Technologies validated white paper gives an overview of the Kamu Data Fabric platform deployed on Red Hat OpenShift Container Platform and Dell PowerFlex. Reference architectures are provided for operating the platform in a single environment as well as in a hybrid decentralized setting for cross-organizational data exchange and data processing at the edge, demonstrating unique capabilities of this Web3-native data solution.

Kamu Data Inc.: Sergii Mikhtoniuk, Özge Nilay Yalcin
Dell Technologies: Jaco van Dijk

DELL Technologies

The information in this publication is provided “as is.” Dell Inc. makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any software described in this publication requires an applicable software license.

This document may contain certain words that are not consistent with Dell's current language guidelines. Dell plans to update the document over subsequent future releases to revise these words accordingly.

This document may contain language from third party content that is not under Dell's control and is not consistent with Dell's current guidelines for Dell's own content. When such third-party content is updated by the relevant third parties, this document will be revised accordingly.

Copyright © April 2023 Dell Inc. or its subsidiaries. All Rights Reserved. Dell Technologies, Dell, EMC, Dell EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries. Other trademarks may be trademarks of their respective owners.

Table of contents

Executive Summary

Efficient use of data has become existential to many businesses. While the existing variety of enterprise solutions can satisfy basic needs of internal data management, a growing number of use cases like multi-cloud, edge deployments, and multi-party data exchange don't fit the traditional enterprise data model.

Kamu Data Fabric is a novel Web3-native data management solution that combines properties of modern enterprise data and blockchains. It's the first decentralized data platform that lets data scientists and analysts create logical data flows that can seamlessly span across edge devices, different clouds, business units, and even organizations, while delivering near real-time and verifiably trustworthy results. Using the principles of blockchain smart contracts, Kamu ensures that every data set, report, analysis, and ML model in your company will have an unbreakable link to the data it was produced from.

Kamu Data Fabric can help if your organization:

- Develops and maintains highly custom solutions to share internal data with partners and vendors in a secure, privacy-preserving way.
- Spends lots of resources integrating external data sources to enrich internal data.
- Operates in a multi-cloud, on-premise, and edge environment and wants to unify data access and processing without centralizing the data.
- Relies on people for recurrent data processing to produce analyses and reports.
- Deals with high-stakes data and strict audit requirements and needs strong accountability and verifiability guarantees for data.
- Drowning in dynamic data, multiple versions of data sets, reports, and models and wants to achieve better systematization, automation, reproducible data science, and unbreakable provenance throughout the entire life cycle of data.

Dell Technologies and Red Hat Solutions offer fully integrated and validated containerization solutions that transform IT infrastructure & reduce IT complexity. Combined with Red Hat OpenShift Container Platform and Dell PowerFlex software-defined infrastructure, the proposed data solution can easily scale according to the needs of your business. By separating data processing from data location and storage concerns, this validated design lets your analysts focus on extracting value from data while the IT department can independently manage data locality, replication, and storage tiering to optimize reliability, security, and costs.

Motivation

Today, efficient use of data is the top factor that separates market leaders from obsolescence. But transforming your organization to be data-centric remains an extremely challenging task. The choice of data management architecture alone can affect the performance of your company for decades to come.

While many internal analytical needs can be addressed by modern data lakes and federated platforms, there are several areas where all enterprise solutions struggle:

Autonomy vs. Ease of access: The best way to make data accessible is to put it in one big system. But centralizing data also means reducing the autonomy of your business units, adding complex layers of permissions, and sacrificing agility. Can we achieve the convenience of a single database while data remains decentralized?

Internal vs. External: If you need to share your sensitive customer data with vendors or enrich internal data with data from hundreds of external sources - you likely have developed highly custom and costly to support solutions for each individual integration. Can sharing and using external data be made as simple as accessing data internally?

Sharing vs. Ownership: When data is the most precious asset of your company, you may feel reluctant to upload it to proprietary and cloud-based solutions. Many companies chose to avoid intermediaries and develop their own data portals only to remain in full custody of data. Can data sharing be made non-custodial so that you never feel that your data is being held hostage by intermediaries?

Static vs. Dynamic: Before raw data can be used for decision making, it must be heavily processed and compiled into reports, analyses, and models. With data sources becoming more and more dynamic (e.g., IoT, market data, blockchains) by the time these artifacts are prepared, they are likely already outdated or completely obsolete. Can we produce autonomous data products that always stay up-to-date and provide real-time situational awareness without needing a human in the loop?

Reuse vs. Trust: If you often find yourself thinking that a certain data point in a report “looks wrong” or wondering what data it was based on - you know that tracing data to its origin and proving that it is correct can take weeks of engineering work. Just a few such issues left unchecked can make you lose confidence in your data entirely. Can we make questions about the quality and provenance of data resolvable in minutes and not require the help of engineers?

In this white paper we will look at Kamu Data Fabric - a novel Web3-native solution holistically designed to answer these questions and deliver unmatched agility and verifiability to your company data flows.

Solution	Non-custodial	ETL processing	Low code	Edge	Real-time	External sharing	Privacy preserving	Time travel	Auditability
Data Warehouses		✓	✓					*limited*	*limited*
Data Lakes	✓	✓						*limited*	*limited*
Stream Processing	*n/a*	✓		✓	✓				
Federated Learning	✓	✓		✓		✓	✓		*limited*
Data APIs	✓			✓	✓	✓	✓		
Hubs / Marketplaces			✓			✓			*limited*
Blockchains	✓			✓	✓	✓		✓	✓
Kamu Data Fabric	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1. Capability comparison of modern data management and dissemination solutions

Product

Kamu Data Fabric is a new-generation hybrid data lake and streaming ETL solution that brings blockchain-like properties to Big Data. Its novel Web3-native design simplifies data sharing within and between organizations, provides superior automation of data workflows, makes them more reliable and easier to maintain, and provides fine-grain provenance and auditability guarantees for all data - all this at IoT volumes and near real-time latencies.

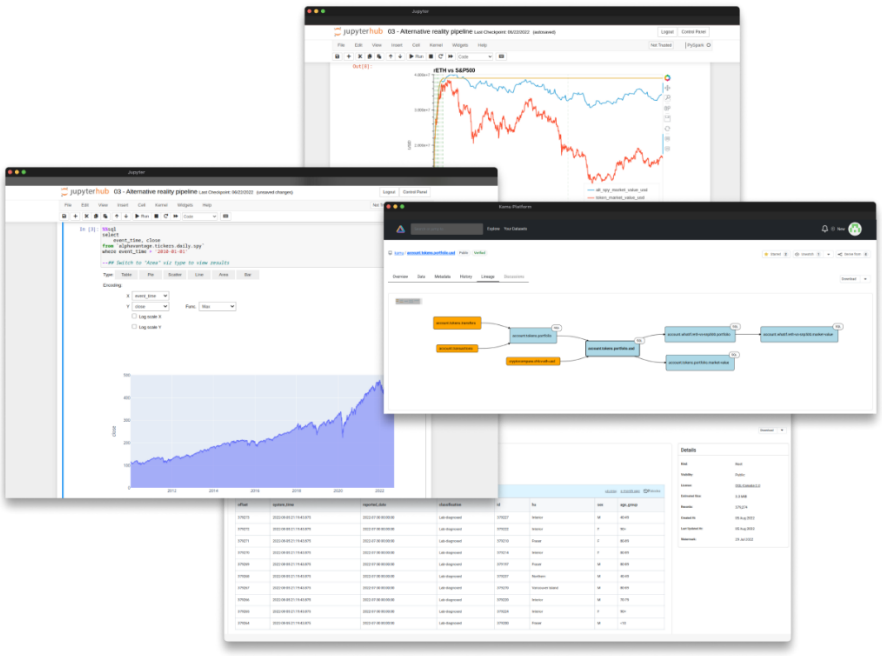

Figure 1. Data exploration and dynamic ETL pipelines in Kamu Platform

Red Hat OpenShift Container Platform provides developers and IT organizations with a hybrid cloud application platform for deploying both new and existing applications on secure, scalable resources with minimal configuration and management overhead. It provides enterprise-grade Kubernetes environments for building, deploying, and managing container-based applications, alongside virtualized workloads, across any on-premises data center where Red Hat Enterprise Linux is supported.

Dell PowerFlex Software-defined Infrastructure is designed to enable customers to modernize without constraints with features that allow extreme consolidation and flexibility. It enables customers to automate their infrastructure and processes to boost IT agility with intelligent software-driven automation that streamlines operations. It is designed to optimize IT outcomes with software-driven storage optimization that ensures extreme workload results.

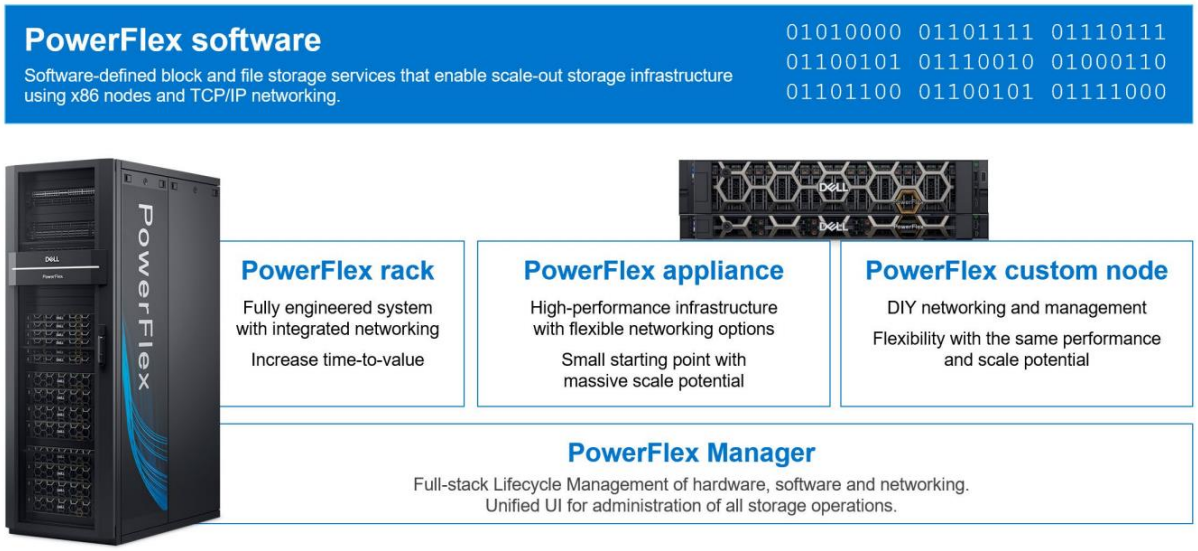

Figure 2. The Dell PowerFlex family

Solution

This section provides an overview of Kamu Data Fabric architecture in two different scenarios:

- First, we will look at how Kamu Data Fabric can be deployed in a single on-premises environment using Red Hat OpenShift and Dell PowerFlex and discuss the benefits it provides compared to conventional enterprise data solutions.
- Next, we will see how the same components can be deployed in hybrid and decentralized environments to process data at the edge, access data across different business units, and exchange data with partners, vendors, and even third parties at a global scale.

Single Environment Architecture

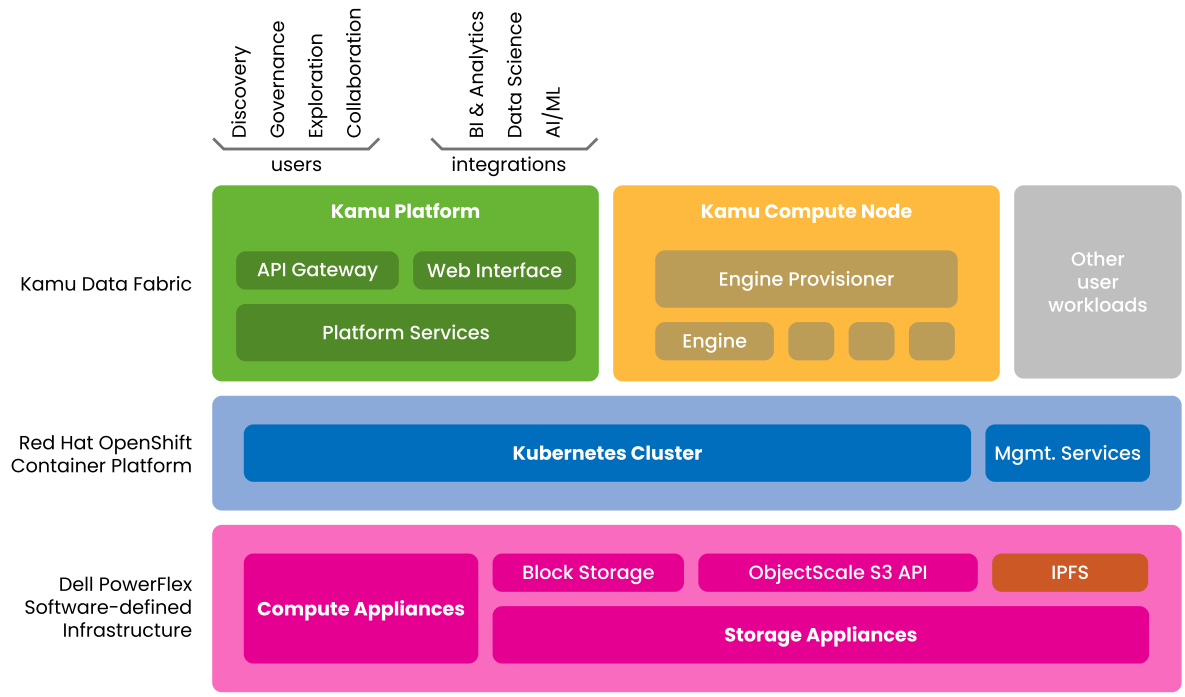

Figure 3. Logical architecture of Kamu Data Fabric with OpenShift and PowerFlex

Figure 3 shows the reference architecture that includes:

- A two-tier **Dell PowerFlex** appliance architecture is used that splits the storage and compute into two resources, allowing for independent scaling of both. Bare metal servers of the PowerFlex storage-only cluster provide both **block storage** and S3-compatible **object storage** using **Dell ObjectScale**. Object storage is used for the data lake, ensuring interoperability with modern data processing engines, while block storage can be dynamically provisioned for any other general-purpose workflows within the cluster.
- **Red Hat OpenShift Container Platform** running on top of PowerFlex compute appliances provides enterprise-grade Kubernetes environments for building, deploying, and managing container-based applications, alongside virtualized workloads.

- **Kamu Data Fabric** is deployed into the OpenShift Kubernetes cluster as containerized workloads. It consists of two major components:
    - **Kamu Platform** is an application through which users interact with data. It provides a web interface for users to browse the data catalogue (discovery), explore and visualize data (analytics, dashboards, alerts), create their own data sets (data ingestion, ETL pipelines), and collaborate with others to improve and make efficient use of existing data (governance, provenance, issue tracking, and more). It also provides a range of APIs (SQL, REST, Kafka) for connecting data to BI tools, AL/ML, data science projects, and various automation.
    - **Kamu Compute Node** is responsible for all data-intensive operations like executing API queries and running continuous stream processing operations for ETL pipelines. Based on the demand, it dynamically provisions the appropriate type and quantity of data processing engines (also as containerized workloads) that perform computations. Kamu integrates with several open-source data processing engines like Apache Spark, Apache Flink, and Apache DataFusion, so you can interchangeably use them to their best strengths at different stages of querying and processing.

This architecture scales easily according to organization needs. Storage and compute resources can be adjusted independently. Data processing can also scale linearly by provisioning more engines, which will be most optimally mapped to underlying hardware by the container platform. And while data is the heart and soul of an organization, the same underlying hardware and virtualization infrastructure can be used to run any other types of user services and workloads co-located with the data platform, allowing for the best resource utilization.

Comparisons

Kamu Data Fabric vs. Data Warehouses: Data warehouse model has two major drawbacks:

First, it conflates data storage with querying and processing - this makes your data inseparable from the database you are using, thus, extremely hard to switch to another solution. Kamu stores your data in open industry-standard formats so that your data is never locked-in to any specific vendor, and you can use multiple different processing solutions in parallel without duplicating the data.

Secondly, the warehouse model requires a high degree of harmonization - raw data is heavily processed into a form optimized for a very specific set of queries. This works great for some queries but makes it very hard to answer questions for which the data model was not designed for.

Kamu provides both - you can create high-quality harmonized and always up-to-date data sets using composable ETL steps, but also, at any point in time, access raw data to experiment and create new ways of using it.

Kamu Data Fabric vs. Data Lakes: Kamu builds on top of the data lake model and is a superset of capabilities:

- Its ledger-like data format records every change that happens to data and provides infinite resolution "time travel" capabilities, so you can go back to any point in time and see exactly how your data looked.
- Using new-generation stream processing engines like Apache Spark and Apache Flink, Kamu provides levels of automation unseen in most enterprise solutions. With Streaming SQL queries your data scientists and analysts can build infinitely composable ETL pipelines without the help of engineers, and have results produced much faster, with superior consistency and accuracy, and often with 1000x less computational costs compared to traditional batch processing.
- All data produced by ETL process contains an internal record of all operations that created the final result. These blockchain-like properties provide best-in-class provenance - ability to tell where every single bit of information came from - and make all data 100% tamper-proof, verifiable, and auditable.

If you already operate a data lake - Kamu Data Fabric can be integrated alongside, gradually and without disruption. It can ingest data from other data lake formats, while data in Kamu's open-source ledger format remains compatible with other modern analytical processing engines.

Decentralized Architecture

Most enterprise data solutions were designed only with internal company data in mind. But as the world continues to move towards a data-centric economy, companies often need to share internal data with other parties, process data at the edge, or organize their data flows in a way that doesn't fit the standard centralized mode. This results in architectures stitched from dozens of poorly fitting pieces, were moving data between solutions, managing permissions, and safeguarding an ever-growing system against attacks occupies 90% of time of data engineering and IT teams.

Kamu Data Fabric is the first data platform designed for cross-organizational data exchange. It allows data to be safely exchanged and processed by parties that don't necessarily trust one-another, while data remains fully verifiable. Figure 4 shows how Kamu can be deployed in such a decentralized scenario.

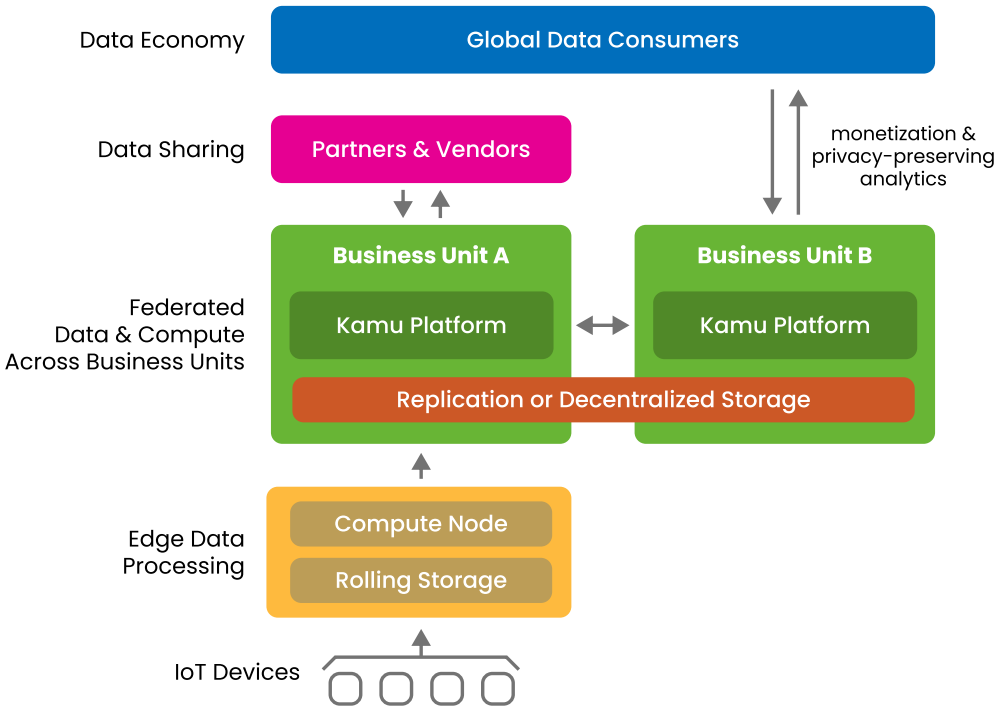

Figure 4. Hybrid architecture with edge deployment and cross-organizational data exchange

Edge data processing: High volumes and frequencies of IoT data pose significant challenges for centralized data architectures as transferring and storing such raw data in the main data lake can become costly. The streaming data processing model of Kamu can operate in a hybrid mode where autonomous compute nodes are deployed at the edge, close to the IoT devices, and can pre-process and aggregate data before transferring abbreviated results to the main data lake. Depending on the task, the processing configuration can be modified dynamically to adjust sampling rates and filters.

Raw data can also be stored at the edge with configurable retention and accessed through the platform interface seamlessly, as if already a part of the main data lake. Kamu's verifiability properties combined with digital signing of data by IoT devices enable architectures where summary data from a vast IoT device fleet can be made verifiable and tamper-proof.

Federation across business units: Different business units within one organization may want to have autonomy in managing data, but with most data architectures this also results in siloing and difficulty of analyzing data across sites. Standardizing the company on a single data warehouse makes data easier to access, but also reduces the agility and freedom of experimentation. The federation feature of Kamu Data Fabric provides users an experience of a single large data lake.

Different units can run their independent instances of the platform and choose which data sets are stable and ready to be shared with a larger organization. The ETL pipelines can work across all sites and combine data regardless of its location – Kamu minimizes data transfers between different locations.

Kamu's ledger-like data format was designed for extreme simplicity of replication, which you can set up across business units to protect yourself from catastrophic data loss. It's also the first data platform to natively support content-addressable file systems like IPFS. It allows your IT team to utilize these novel technologies to further optimize data replication, locality, and storage tiering across large hybrid multi-cloud and on-premises environments.

Data sharing with partners and vendors: Same mechanisms that Kamu uses to transfer data between business units let you share data far across organizational boundaries. Your partners and vendors can use your data as an extension of their own data lake (and vice versa). At the same time every party can hold each other accountable for the data they provide and collaborate on ETL processing even without fully trusting one another.

Kamu gives you control over which data sets you want to share and the granularity at which other parties can query the data. The privacy-preserving analytics features of Kamu allow you to share sensitive customer data with partners while limiting which attributes and at what minimal level of aggregation they are allowed to query data. With Kamu, data always has a single source of truth, you no longer need to copy data across multiple solutions, which can go out of sync, and your partners always get access to the most up-to-date data and can connect it directly to their own ETL pipelines.

Participating in the global data economy: With the advance of IoT, AI/ML, and InsurTech revolution - companies across all industries can benefit from monetizing their internal data. While previously sharing data was so complex, to the point of requiring dedicated engineering teams, with Kamu's privacy-preserving analytics you can share data in a simple yet very secure and responsible way, and easily participate in a global data economy as both publisher and consumer of data.

Comparisons

Kamu Data Fabric vs. Federated Learning: In Federated Learning consumers of data send the ML model training code to publishers, so that training process could happen on the data owner's side without ever exposing sensitive data externally. This process requires specialized infrastructure and a team of ML experts to review submitted models for possible attempts by consumers to extract personal data, making such solutions expensive to integrate and operate. For analytical use cases, where full power of ML is not needed, Kamu provides a simpler privacy-preserving data exchange functionality that, once configured, does not require any maintenance. For ML training, Kamu provides a foundation for novel automated privacy techniques such as differential privacy.

Kamu Data Fabric vs. Data Hubs and Marketplaces: While specialized solutions for cross-company data sharing exist, they are often custodial and require you to copy your data to be stored by intermediary. Many organizations are rightfully reluctant to entrust their data to intermediaries, and in many cases may be legally prevented from doing so. Audit records provided by data hub solutions are often limited to data modification and downloads logs. With Kamu, data can be shared directly from your infrastructure without duplication or reliance on intermediaries (non-custodial), while also providing superior accountability and provenance trail.

Summary

Bringing Web3 properties to data can improve auditability, autonomy, systematization, and governance in data-centric companies, and can open entirely new avenues for exchanging data with other parties. Kamu Data Fabric provides a realistic and gradual transition path where Web3 properties can be adopted by companies with minimal disruption to existing infrastructure and workflows.

The Dell Technologies validated solution based on Dell PowerFlex and Red Hat OpenShift Container Platform presented in this paper provides a modern foundation not only for your data needs, but for your entire IT infrastructure. It delivers optimal performance and scaling characteristics that can adapt to your business needs and streamline and simplify the deployment and ongoing operations.

References

Kamu documentation:

- Kamu Data Fabric - Website
- Kamu Data Fabric - Developer Documentation
- Open Data Fabric - Protocol Specification

Red Hat documentation:

- Red Hat OpenShift Container Platform Documentation

Dell Technologies documentation:

- Dell Validated Platform for Red Hat OpenShift
- Dell ObjectScale: Overview and Architecture