Automating data pipelines

Real-time data ingest processing and event generation with Red Hat OpenShift

Highlights:

Deploy a fully integrated cloud-native data ingest and automated streams processing solution based on Red Hat OpenShift and Red Hat AMQ.

Insert custom metadata with ingest events from Red Hat OpenShift Container Storage, enhancing data richness and making it more actionable.

Automatically scale applications up or down based on data ingest rates with Red Hat OpenShift Serverless.

Event-driven data processing

Artificial intelligence and machine learning (AI/ML) have triggered a revolution with their ability to increase data value and make diverse information more useful and actionable. Serverless models and distributed event streaming technologies like Apache Kafka make AI/ML technologies even more powerful and scalable. Given the massive amounts of data flowing into organizations, software-defined storage technology has emerged as a critical component for automating and scaling data processing, pipeline applications, and infrastructure.

Red Hat® OpenShift® Container Storage combined with Red Hat OpenShift and Red Hat AMQ (part of Red Hat Integration) delivers a powerful foundation and architecture for automating data pipelines. Along with storage-based event generation in Red Hat Ceph® Storage, these technologies can aggregate, ingest, prepare, and manage data from its inception—automating data pipelines and scaling infrastructure based on demand.

Scalable event-driven data pipeline architectures

Notification-driven or event-driven architectures are increasingly important data processing tools. By connecting data ingest services to Red Hat OpenShift Serverless functions, data pipelines can automatically scale to meet requirements, growing or shrinking application infrastructure to process incoming data based on your specific organizational needs. Ingest notifications can move fresh data into data pipelines automatically, with the ability to customize metadata insertion upon ingest. With these innovations, organizations can respond rapidly to changing information or variations in customer behavior, providing the organization with timely insights based on the latest data for a range of use cases.

- Manufacturers can detect anomalies for quality assurance, manage retail product replenishment, and understand logistics.
- Healthcare facilities can automatically process images, using AI inference for detection and alerting—simultaneously anonymizing data for researchers to improve processes or accelerate cures.
- Financial services institutions can accelerate payments or utilize fraud detection to better serve their customers.
- Public sector use cases range from equipment maintenance to automatically detecting geospatial changes in satellite imagery.

Containers and Kubernetes orchestration are vital for deploying AI/ML in hybrid clouds.[1] Figure 1 shows how a combination of Red Hat technologies can be used to build an ingest data pipeline for a financial services application. The solution includes integrated object bucket notifications in the Red Hat Ceph Storage RADOS gateway (RGW), data streaming services provided by Red Hat AMQ streams, and Serverless capabilities in Red Hat OpenShift.

www.facebook.com/redhatinc/
@RedHat
linkedin.com/company/red-hat

[1] For more background, read the e-book "Top considerations for building a production-ready AI/ML environment."

Object bucket notifications. *When certain events are triggered on an object bucket in Red Hat Ceph Storage, notifications can be sent to HTTP, Advanced Message Queueing Protocol, and Kafka endpoints.*

With this fully-integrated architecture, organizations can create data pipelines that automatically process incoming data in real time. User-defined functions can perform operations such as data anonymization, tagging, metadata enrichment, and other modifications that can be defined and placed as desired. Object bucket notifications can be used throughout to automate processing for different stages of the pipeline.

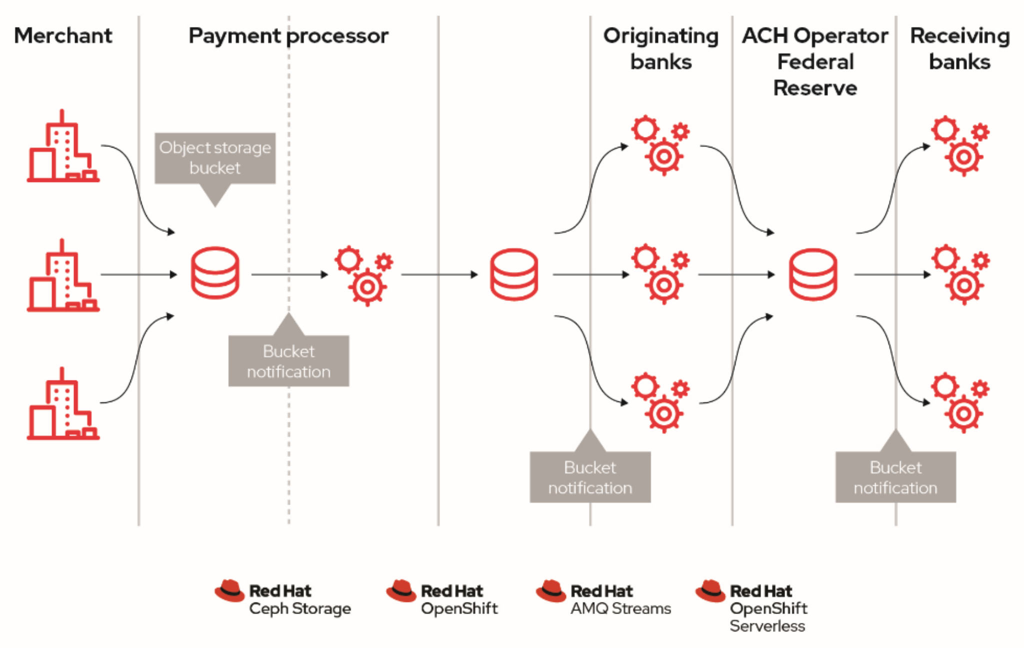

Figure 1. Object bucket notifications in Red Hat Ceph Storage drive streams processing events

Red Hat Ceph Storage (a component of OpenShift Container Storage) helps create scalable, event-driven architectures by connecting data ingest to Red Hat OpenShift Serverless functions.

This cloud-native architectural approach delivers tangible benefits, including:

- **Faster real-time processing.** Once the ingest data pipeline is configured, data can be processed in real time as it is ingested. This significantly speeds the process of making datasets available for machine learning, and ensures that data scientists have the latest data to train their models.
- **Scalability and flexibility.** Each component of the architecture can be customized and independently scaled depending on administrative and user needs. Within Red Hat OpenShift Serverless, kNative functions can be customized to suit the specific dataset and organizational processes, requirements, and goals.
- **Extensibility.** An extensible solution architecture lets organizations add more functionality and automation in their data life-cycle processes, without having to re-architect. The combination of OpenShift Container Storage, Red Hat OpenShift Serverless, and Red Hat AMQ streams can extend to other areas of data life-cycle management such as data cataloging and audit logging.
- **Efficiency.** Ingest data pipeline automation lets data engineers codify and automatically perform many operations related to data preparation for machine learning. Once configured, pipelines can scale to manage large and varied amounts of incoming data, leaving engineers to focus on other high-value activities.
- **A foundation for tracking the data lifecycle.** The ability to process data at ingest in real time lets you embed the right tags about the data source as well as any source- and time-specific information. This valuable information can be used later to enhance data management and machine learning processes such as data classification, feature engineering, and cataloging.

OpenShift Container Storage provides highly-available cloud-native persistent storage for Red Hat AMQ.

- **Persistent and resilient storage for Red Hat AMQ.** OpenShift Container Storage provides the data persistence you need with Red Hat AMQ components, including Apache Zookeeper and Apache Kafka, in the form of highly-available block storage persistent volume claims (PVCs).

Apache Kafka on Red Hat OpenShift with Red Hat AMQ

Red Hat AMQ is a massively scalable, distributed, and high-performance data streaming platform based on the Apache Kafka project, and designed to run on Kubernetes (Red Hat OpenShift Container Platform). Used in a wide variety of use cases, Apache Kafka is ideal for applications that include web activity tracking, eventing, metrics, logging, streaming data to a data lake, and others. Kafka can handle real time streams of data, either to collect large amounts of data or to perform real-time analysis, or both.

As illustrated in Figure 2, Red Hat AMQ provides three Kubernetes operators:

- A **cluster operator**, responsible for deploying and managing Apache Kafka clusters within a Red Hat OpenShift Container Platform cluster.
- A **topic operator**, responsible for managing Kafka topics within a Kafka cluster running with a Red Hat OpenShift Container Platform cluster.
- A **user operator**, responsible for managing Kafka users within a Kafka cluster running within a Red Hat OpenShift Container Platform cluster.

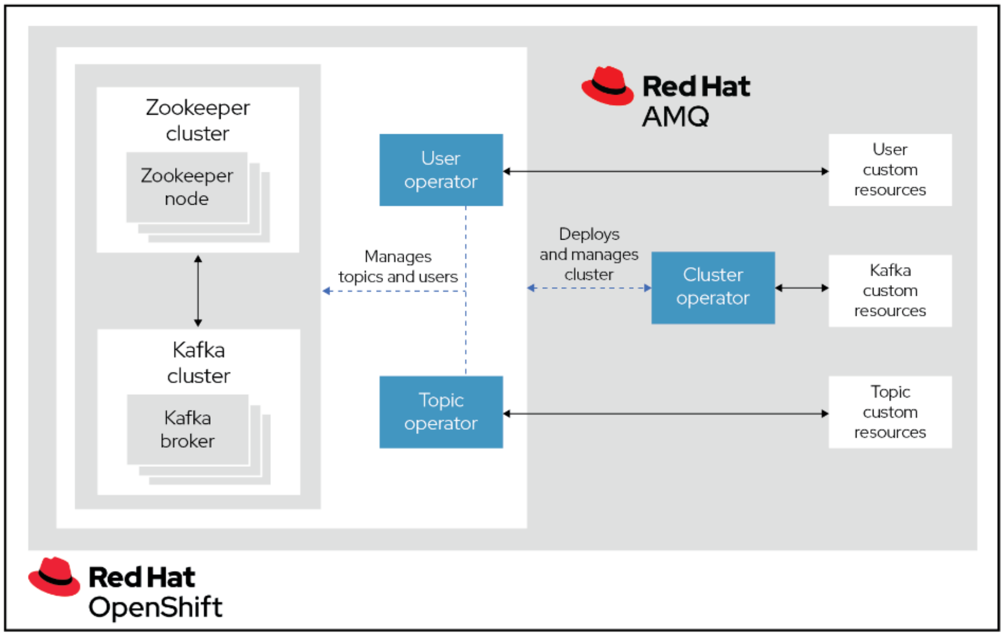

Figure 2. Red Hat AMQ implements three Kubernetes operators

Resilient persistent storage with OpenShift Container Storage

OpenShift Container Storage is a unified petabyte-scale storage solution providing block, object, and file interfaces to OpenShift applications. The storage platform has been tightly integrated and designed to work with Red Hat OpenShift Container Platform. OpenShift Container Storage PVCs support both Kafka and Zookeeper clusters (Figure 3), with object bucket notifications generating events that trigger Apache Kafka topics.

About Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions, using a community-powered approach to deliver reliable and high performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers integrate new and existing IT applications, develop cloud-native applications, standardize on our industry leading operating system, and automate, secure, and manage complex environments. Award winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500. As a strategic partner to cloud providers, system integrators, application vendors, customers, and open source communities, Red Hat can help organizations prepare for the digital future.

In addition, unlike typical public cloud block storage offerings, OpenShift Container Storage can provide the highest levels of resiliency across multiple public cloud availability zones. Apache Kafka has built-in data resiliency capabilities, but OpenShift Container Storage persistent volume claims play a vital role when deploying on a cloud-native environment like Kubernetes. Kafka messages can also be moved to scalable, distributed Amazon Simple Storage Service (S3) compatible object storage for data analytics and long-term retention purposes.

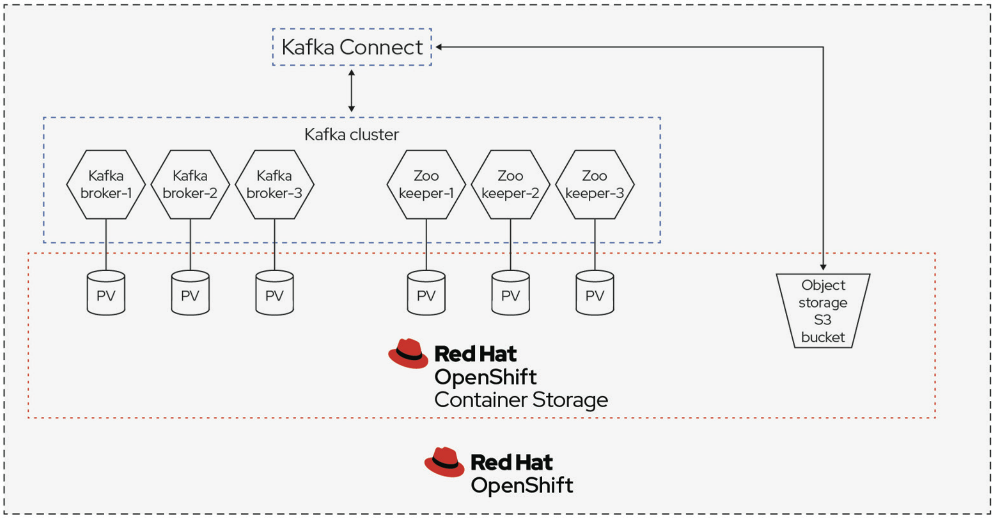

Figure 3. Storage use case for Red Hat AMQ streams

Conclusion

Red Hat offers a fully integrated cloud-native solution for deploying automated stream processing. Red Hat Ceph Storage and OpenShift Container Storage let organizations ingest and process massive amounts of data in real time. Object bucket event notifications let data engineers and data scientists automate streams processing and drive their insights with the freshest data, inserting custom metadata to increase data value for their AI/ML applications. With availability across cloud provider availability zones, the solution is robust and ready for the most demanding enterprise environments.

North America
1 888 REDHAT1
www.redhat.com

Europe, Middle East, and Africa
00800 7334 2835
europe@redhat.com

Asia Pacific
+65 6490 4200
apac@redhat.com

Latin America
+54 11 4329 7300
info-latam@redhat.com

facebook.com/redhatinc
@Redhat
linkedin.com/company/red-hat