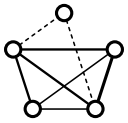
RED HAT DATA ANALYTICS INFRASTRUCTURE SOLUTION

A cost-effective shared data repository for financial analytics

BRIEF



BENEFITS

- Rapidly deploy and decommission analytics clusters on demand with Red Hat hybrid multicloud infrastructure.
- Share analytics datasets through object storage, avoiding unnecessary duplication of large datasets and eliminating data access delays.
- Use space-efficient erasure coding for data protection, saving up to 50% of per-cluster storage costs over 3x replication.[1]

INTRODUCTION

Financial services organizations face uncertainty and risk in a dynamic marketplace governed by stringent and complex regulations. Data analytics now plays an essential role, and it is increasingly a part of virtually every decision in the financial sector. Unprecedented access to information, including historical data and real-time inputs, is driving important areas like risk analysis and anti-money laundering. Effective data strategies are required to accurately predict pricing and risk profiles on an intraday or multiday basis.

As analytics adoption grows, sharing large datasets and infrastructure between analysts and teams can be a challenge. Different teams have distinct priorities, and they require different tools and software versions. Creating separate copies of large datasets is often not feasible, with data hydration and destaging consuming valuable time—not to mention the storage costs for petabytes of redundant data. These challenges are complicated by data retention policies that are now a vital component of regulatory compliance.

The Red Hat® data analytics infrastructure solution offers a novel approach based on accepted cloud models. Integrating key components of the Red Hat stack yields the ability to rapidly spin up and spin down analytics clusters while giving them access to the same shared data repositories.

RED HAT DATA ANALYTICS INFRASTRUCTURE SOLUTION

With the growing prevalence of cloud-based analytics, data scientists and analysts have grown accustomed to dynamically deploying clusters on public cloud services. By design, these clusters have access to shared datasets, avoiding time-consuming data hydration periods after initializing a new cluster or destaging cycles upon cluster termination. Many analysts now expect or desire these capabilities on-premise.

Technology is evolving as well. Most organizations initially deployed analytics on bare-metal systems, hoping to get the most performance while taking advantage of data locality. Many now employ virtual machines, provisioned by OpenStack®, allowing for some of the flexibility of cloud-based analytics infrastructure. Containers represent the latest evolution for on-premise analytics deployments, increasing both performance and flexibility.

The Red Hat data analytics infrastructure solution builds on all of these trends and implements a software-defined shared datastore. The solution supports Amazon Simple Storage Service (Amazon S3) compatibility. With the S3A filesystem client connector, Apache Spark and Apache Hadoop jobs and queries can run directly against data held within a shared S3-compatible datastore.

1 *Testing by Red Hat and QCT, 2017-2018. https://www.redhat.com/en/blog/why-spark-ceph-part-3-3*

The solution integrates key components of the Red Hat stack:

- **Red Hat OpenShift® Container Platform** is an optional element of this solution for those who are interested in containerizing Spark clusters. It is a reliable, enterprise-grade platform that combines the industry-leading container orchestration engine with advanced application build and delivery automation features.
- **Red Hat OpenStack Platform** is an optional element of this solution and is a cloud computing platform that virtualizes resources from industry-standard hardware, organizes those resources into clouds, and manages them. In the context of this solution, it provides on-demand provisioning of virtualized analytics clusters.
- **Red Hat Ceph® Storage** is an open and massively scalable S3-compatible software-defined storage solution for modern workloads like cloud infrastructure and data analytics. It provides the shared object repository for the solution.

THE BENEFITS OF A SHARED DATA REPOSITORY

The Red Hat data analytics infrastructure solution is a natural choice for financial services organizations that want to provide an S3-compatible shared data repository experience to their analysts on-premise. Supporting Spark or Hadoop analytics provides a number of benefits over traditional Hadoop Distributed File System (HDFS), including:

- **Lower capital expenditures (CapEx) by reducing duplication.** Petabytes of redundant storage capacity can be reduced or eliminated while allowing access to the same datasets by multiple clusters.
- **Lower CapEx by improving data durability efficiency.** Using erasure coding for data protection potentially reduces the CapEx of purchased storage capacity by 50% over typical 3x HDFS replication.[2]
- **Right-size CapEx for infrastructure.** Shared object storage promotes right-sizing of compute needs and avoids overprovisioning of either compute or storage resources.
- **Lower operating expenses and risk.** With shared storage, clusters can retain access to the same data without costly and time-consuming scripting and scheduling of dataset copies between HDFS instances.
- **Accelerated insights and better compliance.** Analyzing data in place within a shared Ceph data repository can reduce time to insight and help maintain compliance within established deadlines.
- **Support for different tool and version needs of diverse data teams.** With a shared datastore, cluster users can choose the toolsets and versions appropriate to their jobs without disrupting users from other teams requiring different tools and versions.

CONCLUSION

The Red Hat data analytics infrastructure solution helps financial services organizations embrace data analytics as part of market risk and compliance efforts. Allowing teams to effectively share historical and new real-time data improves organizational competitiveness while reducing acquisition and operational costs. By adopting software-defined storage models from the public cloud, organizations can anticipate market trends while supporting essential data retention policies that help maintain regulatory compliance.

2 *Testing by Red Hat and QCT, 2017-2018. https://www.redhat.com/en/blog/why-spark-ceph-part-3-3*

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.

NORTH AMERICA
1 888 REDHAT1

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

redhat.com
f15496_0419