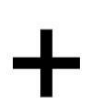
Lenovo and Red Hat Ceph Storage for HPC

HPC use cases for Object and file clusters

Kristoffer Nærland
Data Services Specialist

Sander Snel
Senior Solutions Architect



What we'd like to talk about today

"Providing digital solutions in a changing world"

1. Who are we and what is Red Hat Data Services?
2. What's connecting applications and supercomputers in 2022?

1. What is Red Hat Data Services ?

Red Hat Data Services in a nutshell

Data efficiency

- Erasure coding
- Compression
- Performance

Data resilience

- Snapshots
- Clones
- Backup
- Recovery
- Business continuity
- Disaster recovery

Data security

- At rest encryption
- In flight encryption
- Key management

Data governance

- WORM
- Auditing
- Compliance
- SEC & FINRA
- GDPR

Data discovery

- Cataloging
- Tagging
- Search

intel + Lenovo Red Hat Data Services

Multiple cephs tier 1 use cases

Red Hat Ceph Storage

Ceph storage cluster
Leading on-prem for S3 at scale

- **Object storage**
- Block storage
- File storage
- Leading the on-premise object market at 10-Petabyte+ scale
- Setting the standard for object compatibility outside of AWS

Red Hat OpenStack Platform

Red Hat Ceph Storage

Ceph for OpenStack
 # 1 in OpenStack storage

- Cinder block storage
- Nova ephemeral storage
- Glance image storage
- Swift object store
- Manila file storage
- Advanced integration
- Unified management
- Hyperconverged and Edge capabilities

Red Hat OpenShift Data Foundation

Ceph for OpenShift
Self-managing storage

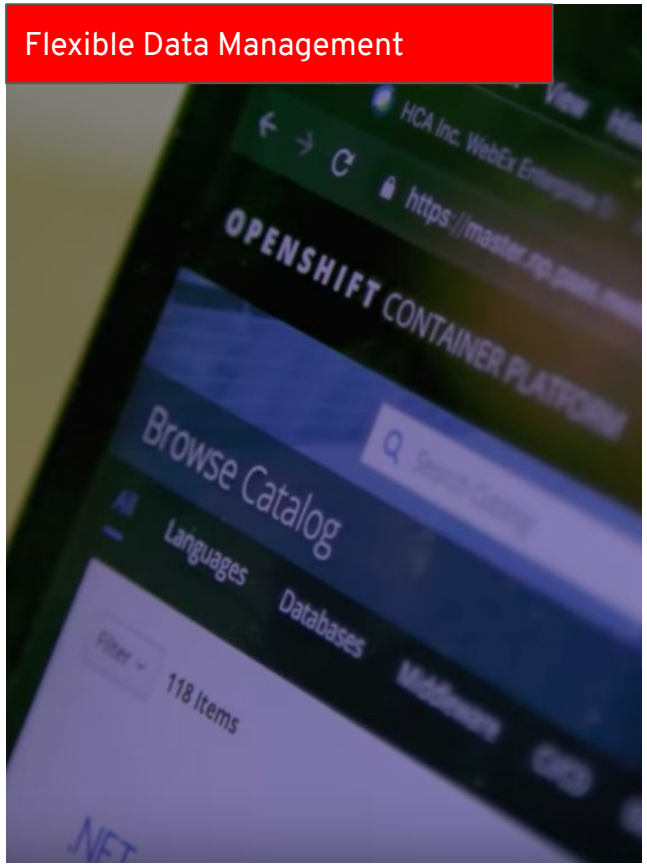
- Powered by Red Hat Ceph Storage
- Automated by Rook and completed with Multicloud object gateway
- Advanced integration and ease of use
- Adds support for stateful workloads to OpenShift

What does a Data Engineer Want?

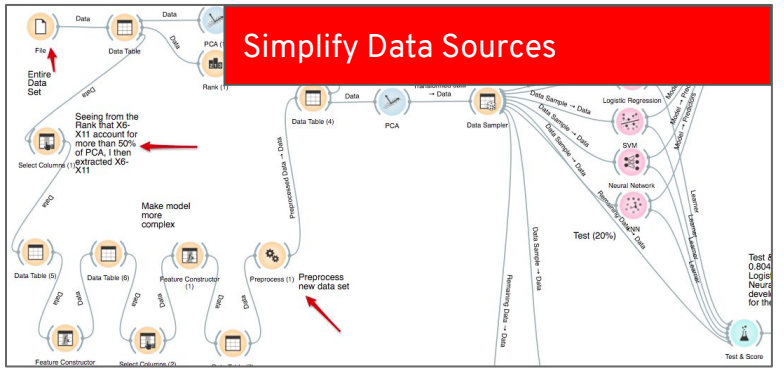
As a Data Engineer, I want to provide my users with a “seamless data” ingestion experience, where I can enable and manage data flows to without disturbing the data environment.

Data engineers consume **data services.**

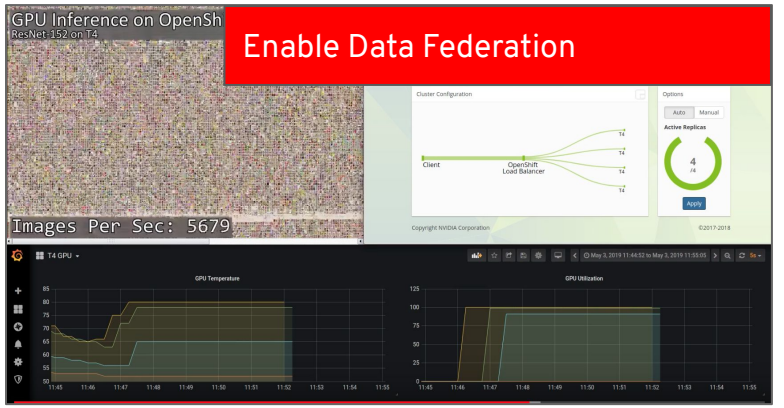
Flexible Data Management



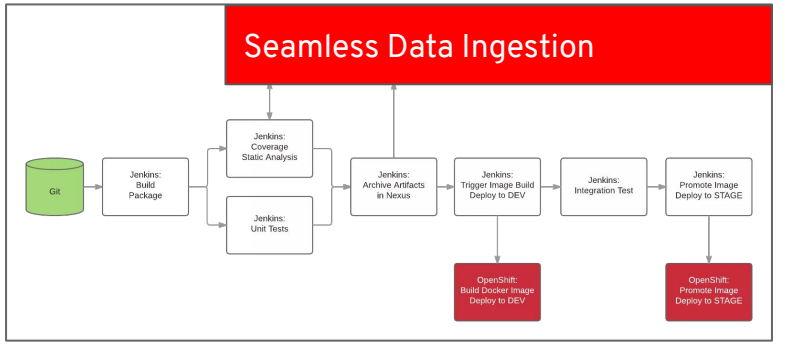
Simplify Data Sources



Enable Data Federation



Seamless Data Ingestion



intel + Lenovo Red Hat Data Services

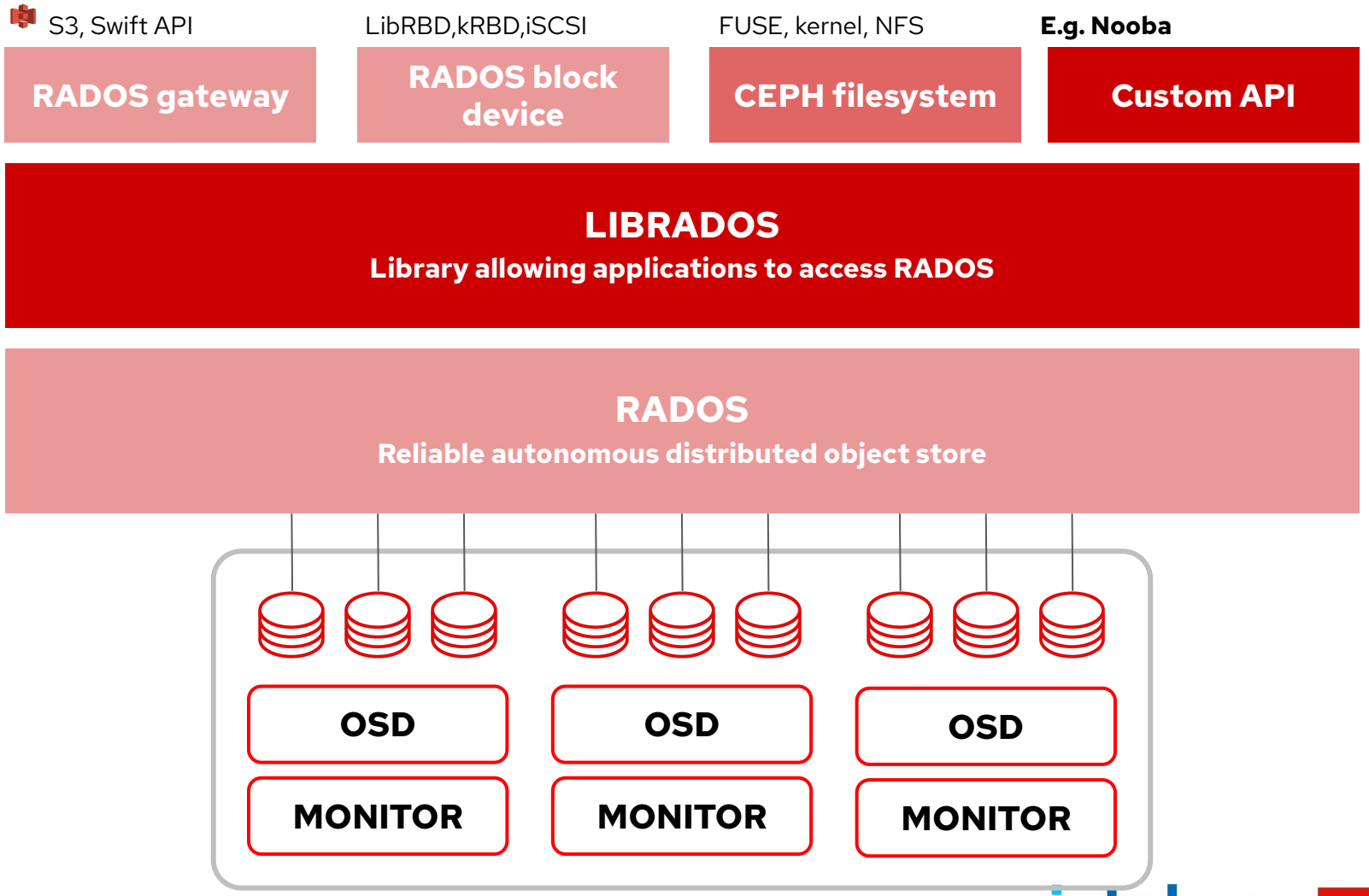
Data is the most significant asset in today's businesses—give it data services

Red Hat
Data Services

- Data services focuses on infrastructure **and application** needs so they can run and interact with ease and efficiency
- Data services provides a foundational layer for applications to function and interact with data in a simplified, consistent and scalable manner

Why is Red Hat Ceph Storage is **the foundational** component to drive data services?

Ceph architecture



intel + Lenovo

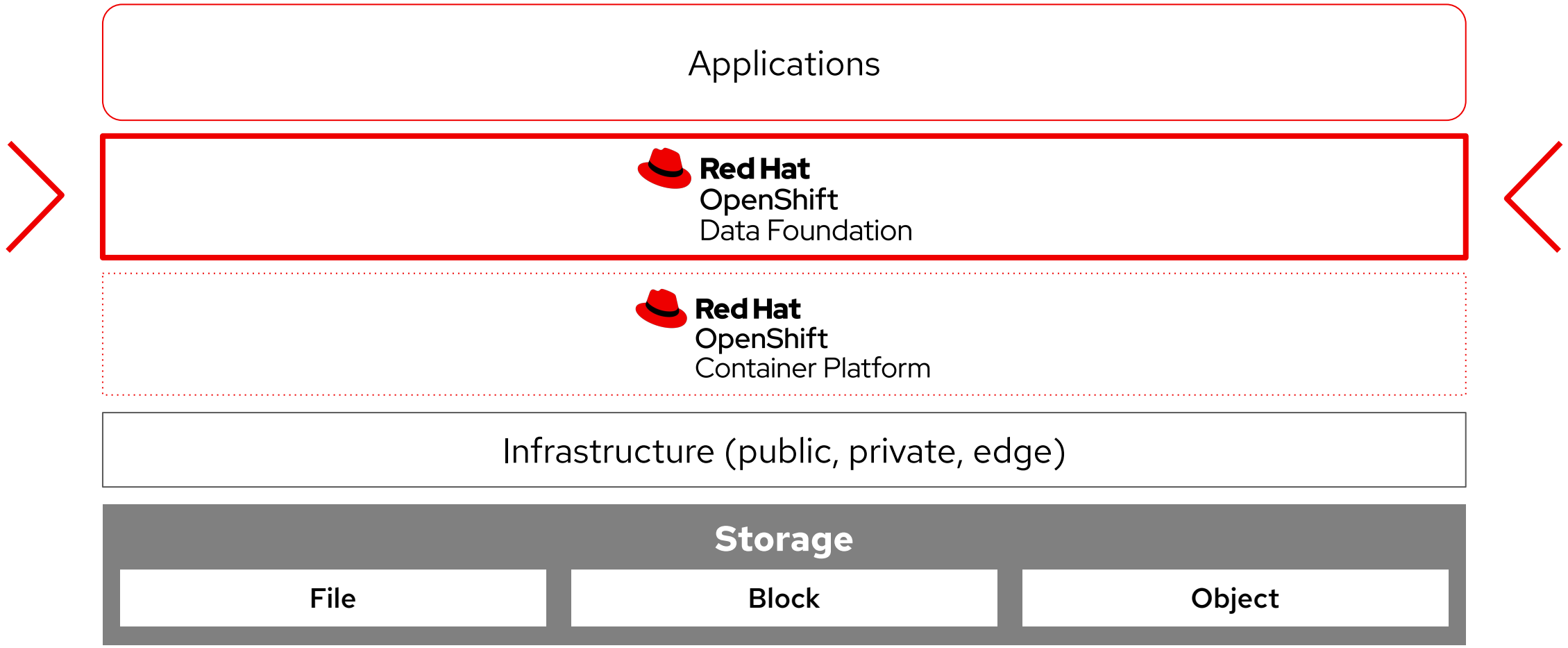
Red Hat Data Services

What’s connecting applications and supercomputers in 2022?

File block or object?

Onprem solutions prefer file storage for blazing fast performance. Increasingly, it becomes cheaper for data scientists to rely on external data services to find new analytics capabilities

Persistent storage, the Kubernetes way



intel + Lenovo Red Hat Data Services

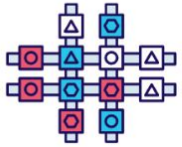
What role does Rook play?

Features of Rook

Simple and reliable automated resource management

Hyper-scale or hyper-converge your storage clusters

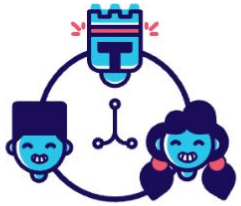
Efficiently distribute and replicate data to minimize loss



Provision, file, block, and object with multiple storage providers

Manage open-source storage technologies

Easily enable elastic storage in your datacenter

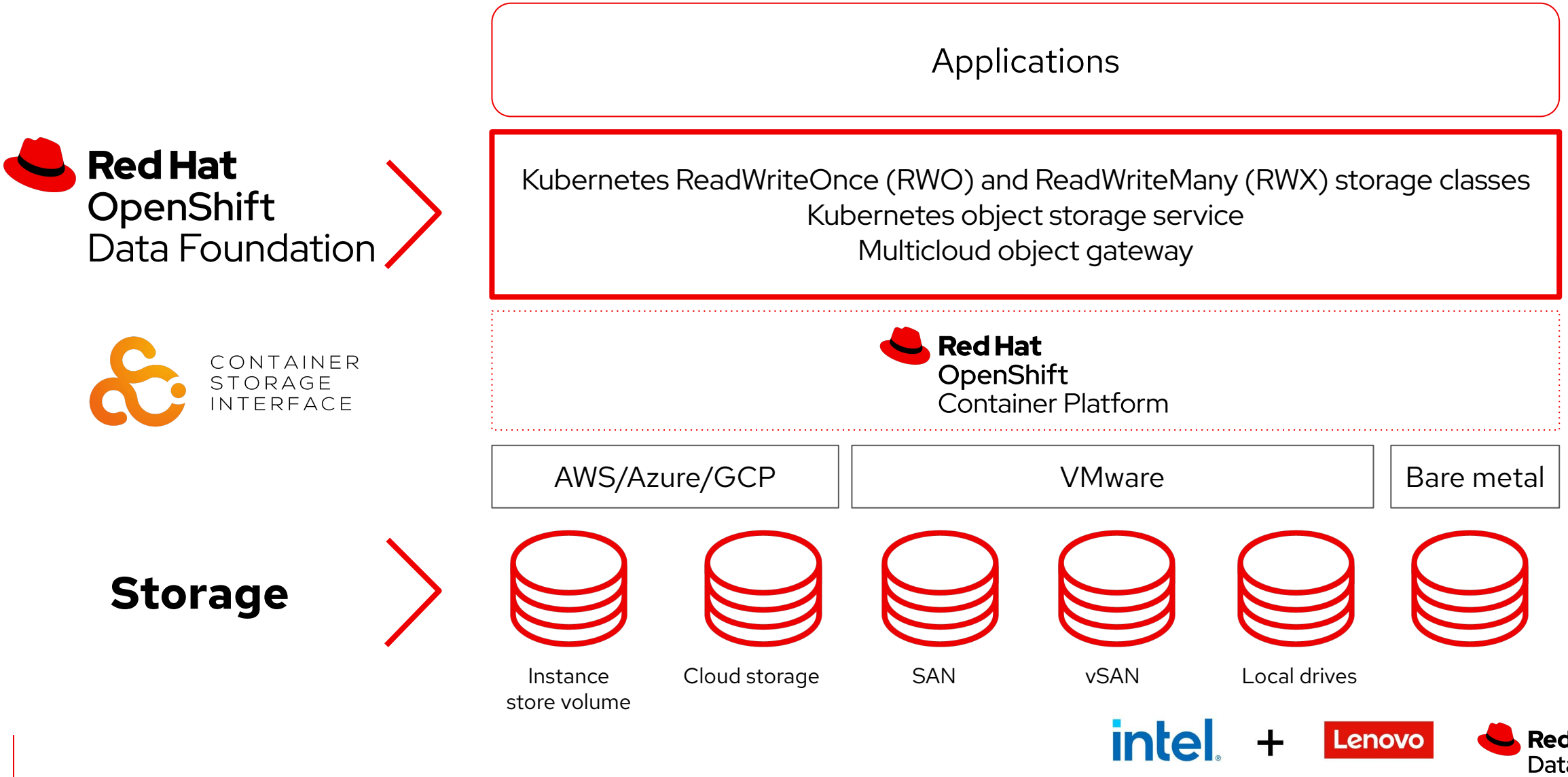


Open source software released under the Apache 2.0 license

Optimize workloads on commodity hardware

intel + Lenovo Red Hat Data Services

The Red Hat OpenShift Data Foundation stack



Red Hat OpenShift Data Foundation workloads

Workload specialized data foundation

For data at rest

Databases, warehouses and lakes

- Microsoft SQL Server
- crunchy data
- mongoDB
- Red Hat Ceph Storage
- MySQL
- IBM Db2 Warehouse

For data in motion

Streaming and messaging

- Red Hat Ceph Storage
- ceph
- Red Hat AMQ Streams
- kafka
- Red Hat OpenShift Serverless
- Knative

For data in action

Data analytics, intelligence, AI/ML

- Apache Spark
- jupyter notebook
- TensorFlow
- PyTorch
- Starburst ANALYTICS ANYWHERE
- Jenkins
- TEKTON
- Microsoft SQL Server Big Data Clusters
- SAP Data Intelligence

Cloud-native infrastructure data foundation

For any stateful app

Infrastructure services

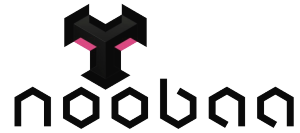
- Red Hat Quay
- Prometheus
- elasticsearch
- fluentd
- kibana

intel + Lenovo

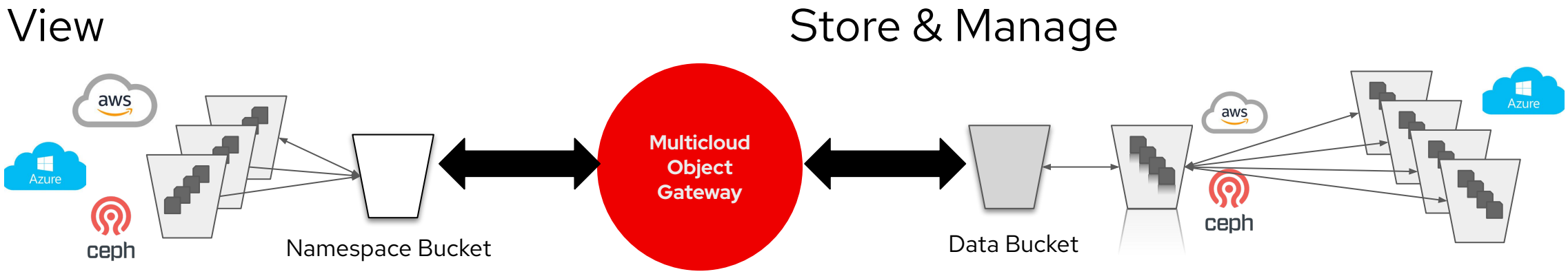
Red Hat Data Services

What is Multicloud Object Gateway?

- Data management layer for Object Storage
- Part of OpenShift Data Foundation
- Based on the community project NooBaa



Data bucket VS Namespace bucket.



Read-centric

- Data is pulled (read) from multiple underlying sources on demand, from a single endpoint

Single Endpoint - No Siloes

- Provides a single endpoint 'view' across all underlying storage sources to all applications

Data aggregation

- Underlying source data doesn't change, but can be replicated

Balanced Read/Write

- Data is mirrored, replicated, or spread to multiple underlying storage destinations (read/write).

Single Endpoint - No Siloes

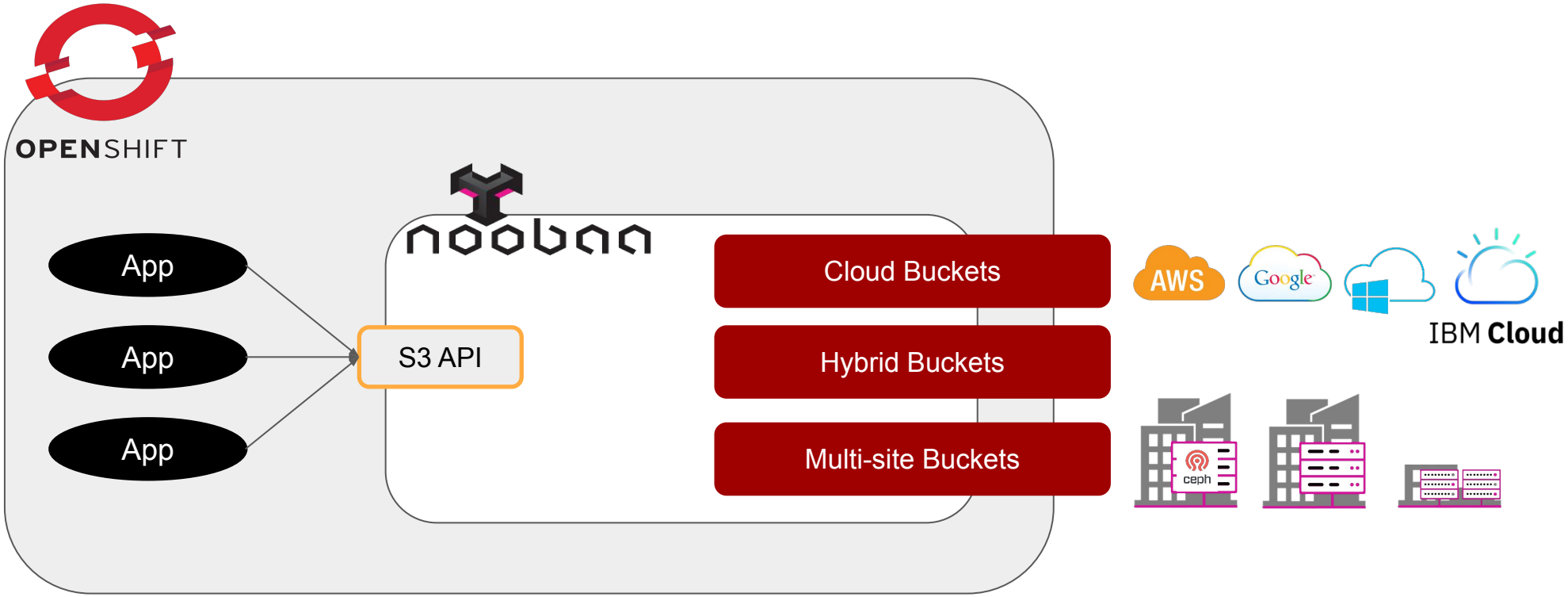
- Provides a single endpoint across all underlying storage destinations to all applications

Data Tiering

- Combining multiple layers of mirroring and spreading

intel + Lenovo Red Hat

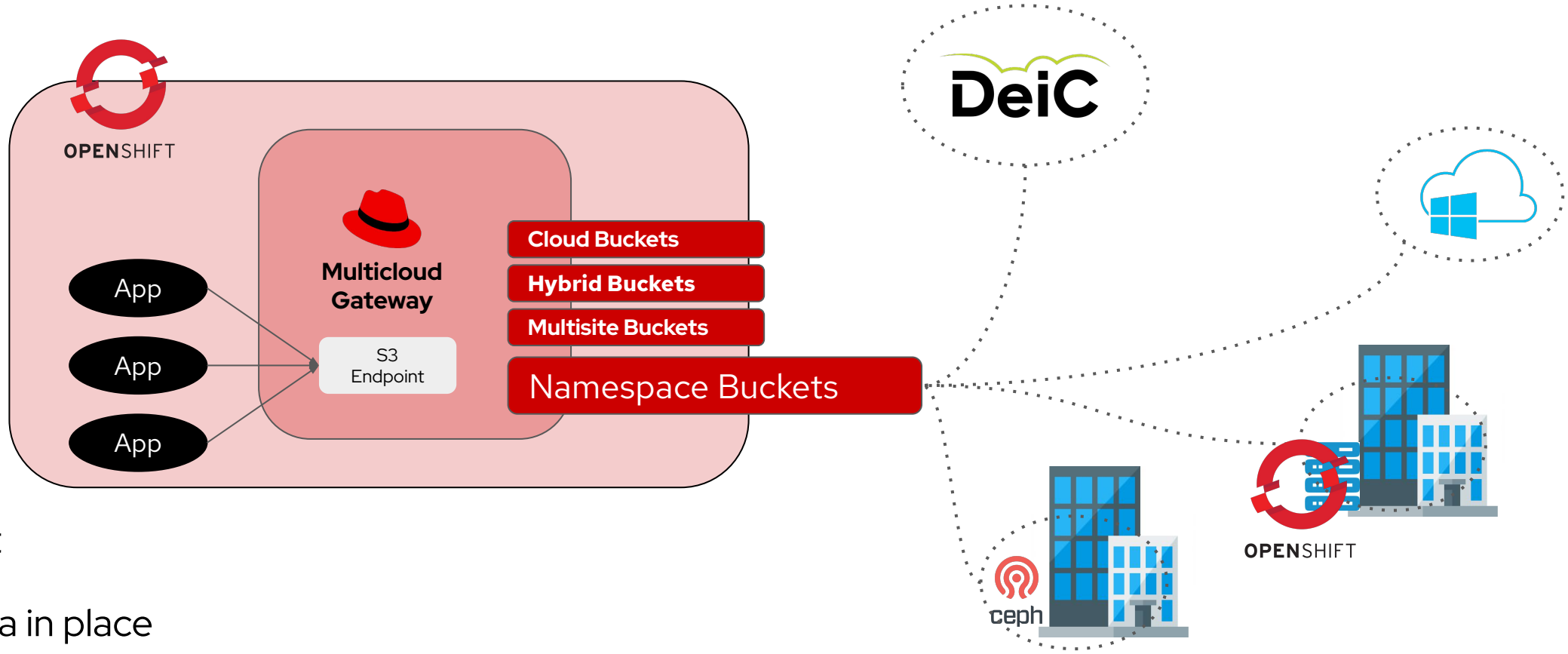
Data Buckets



- Inline deduplication, compression and encryption
- Data stored on local PVs, S3 compatible storage or Cloud
- Data can be mirrored and served from single endpoint

intel + Lenovo Red Hat

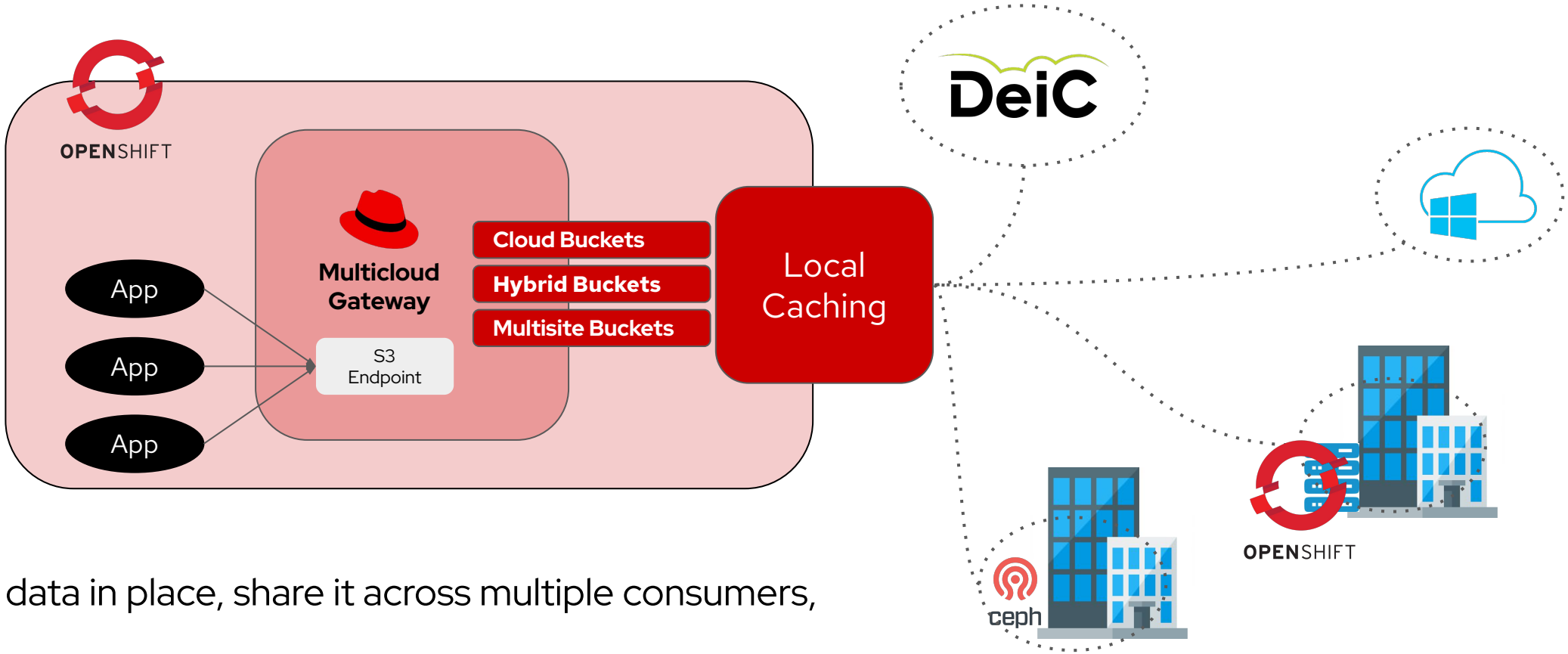
Namespace buckets



- Plain text
- Keep data in place
- Flexible virtualization across multiple varied storage backends with a local caching

intel + Lenovo Red Hat

Local Caching

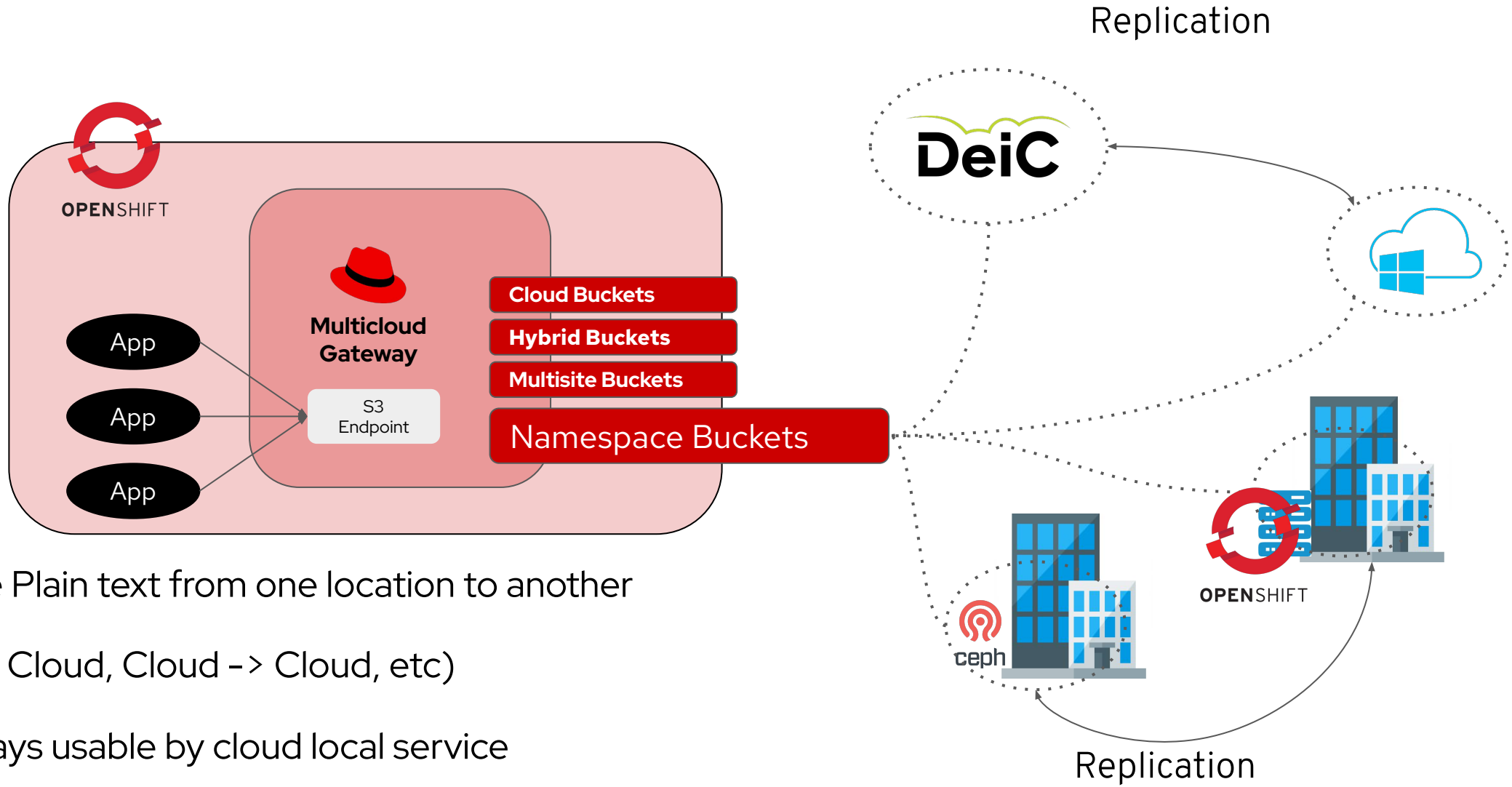


- Plain text
- Keep raw data in place, share it across multiple consumers,

  based on their credentials. Multiple accounts can be used to expose data variations.
- Locality caching layer reduces network bandwidth and egress cost

intel + Lenovo Red Hat

Namespace Bucket Replication



- Replicate Plain text from one location to another

  (MCG -> Cloud, Cloud -> Cloud, etc)

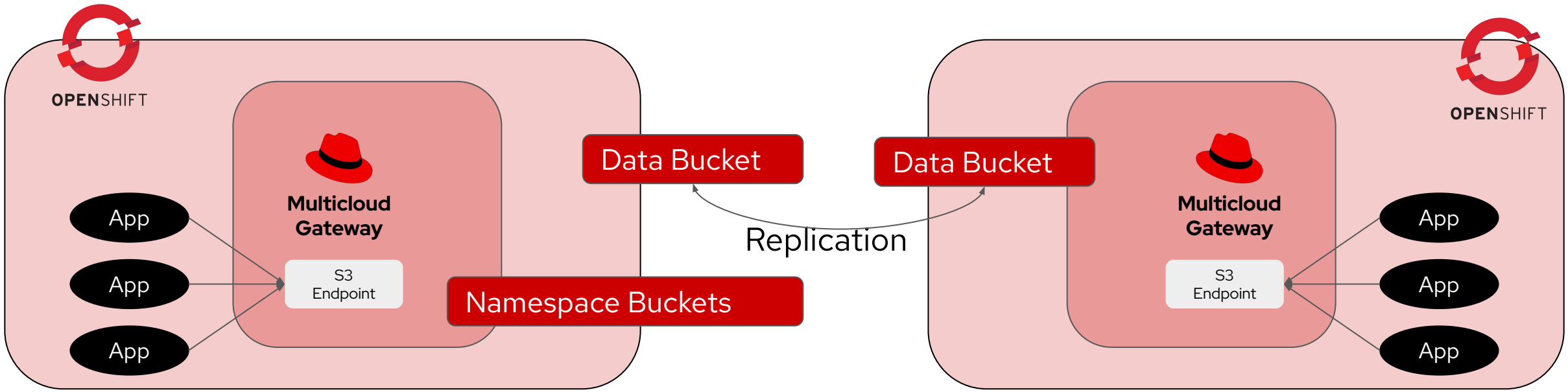
- Data always usable by cloud local service

intel + Lenovo Red Hat

Namespace Bucket Replication

- Configurable - Unidirectional or Bidirectional Asynchronous replication
- Support for AWS S3, S3 compatible, Azure
- Metrics for total replicated objects, bytes and more. Nothing in the dashboard yet.

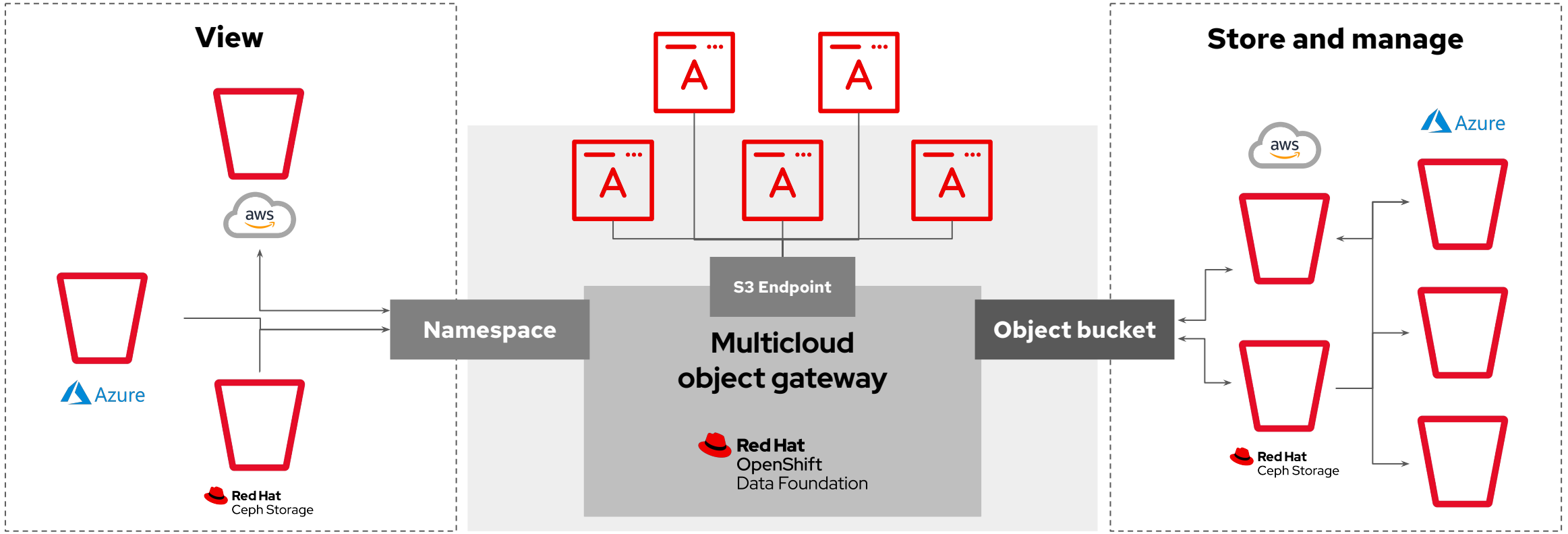
Use case - Data collaboration and resiliency



- Each site produces and consumes its own data
- The entire data set is available in both sites
- Data is replicated to ensure availability in case of disconnection

intel + Lenovo Red Hat

Multicloud object gateway



- Read-centric
- Single endpoint view—no siloes
- Data does not move

- Balanced read/write
- Single endpoint—no siloes
- Data tiering

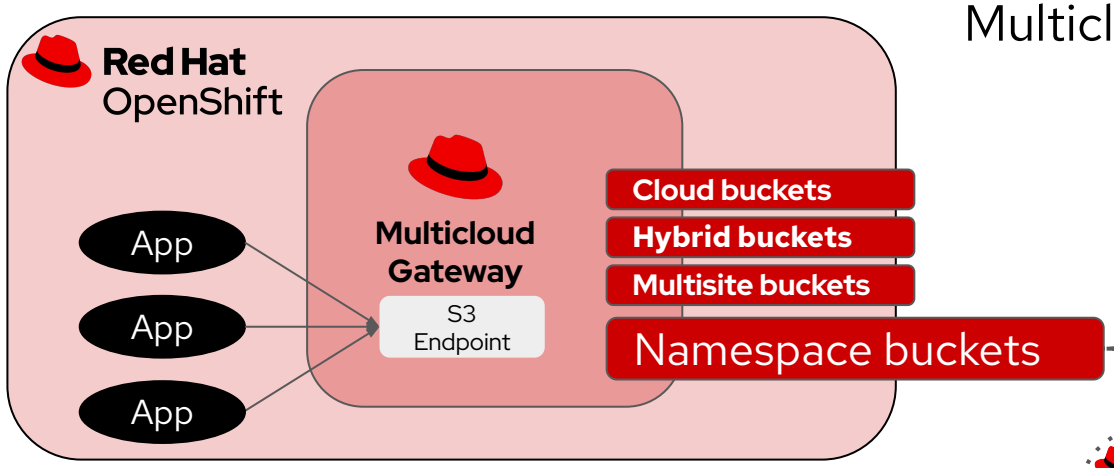
intel + Lenovo Red Hat Data Services

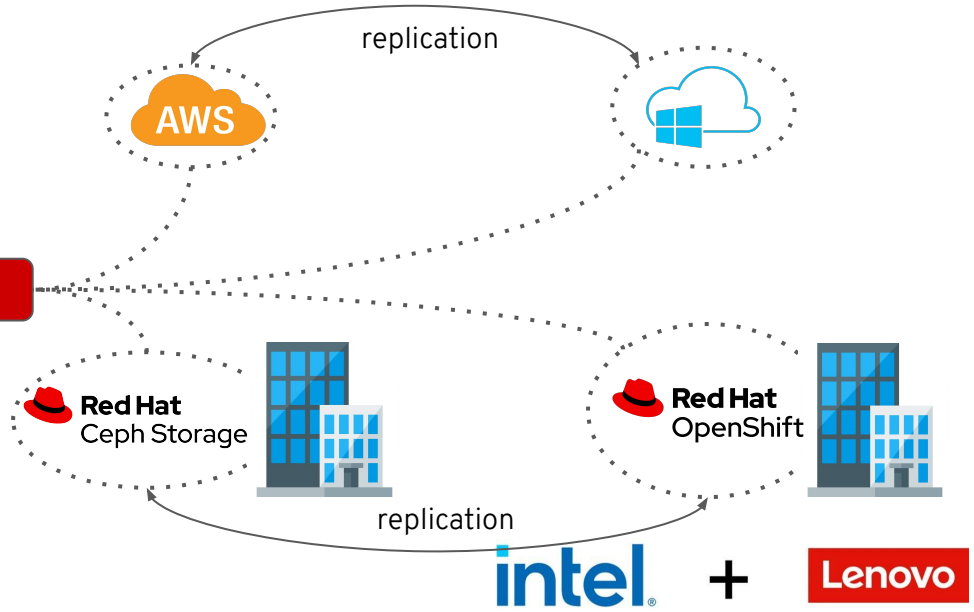
FUNCTIONALITY

Multicloud Object Gateway
Namespace bucket replication

Provides improved resiliency and more collaboration options by replicating data to other locations.

This could be S3 or S3 compatible, including other Multicloud Object Gateway instances.



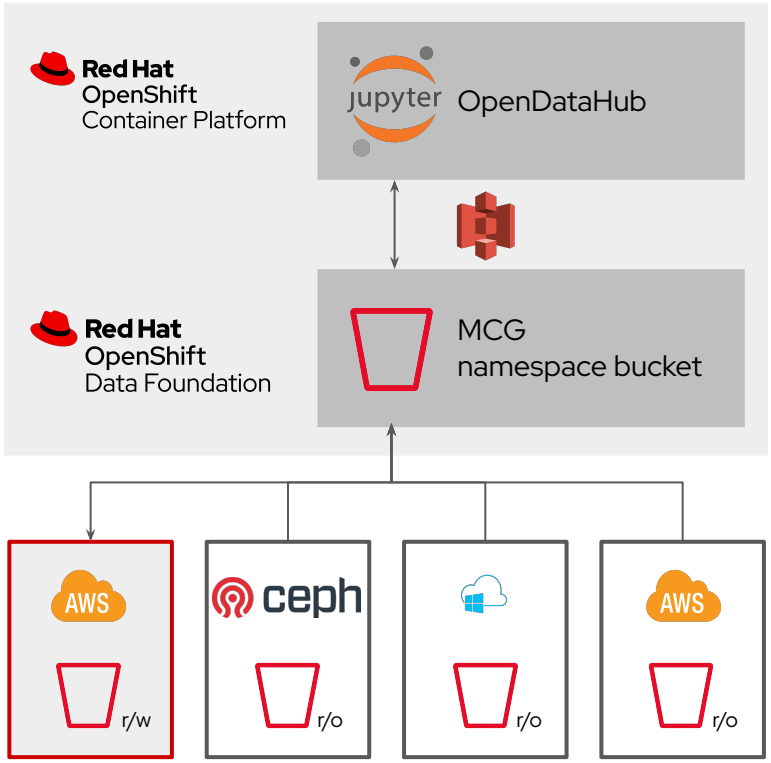


intel + Lenovo

Red Hat
Data Services

What's new?
Red Hat OpenShift Data Foundation 4.7

FUNCTIONALITY



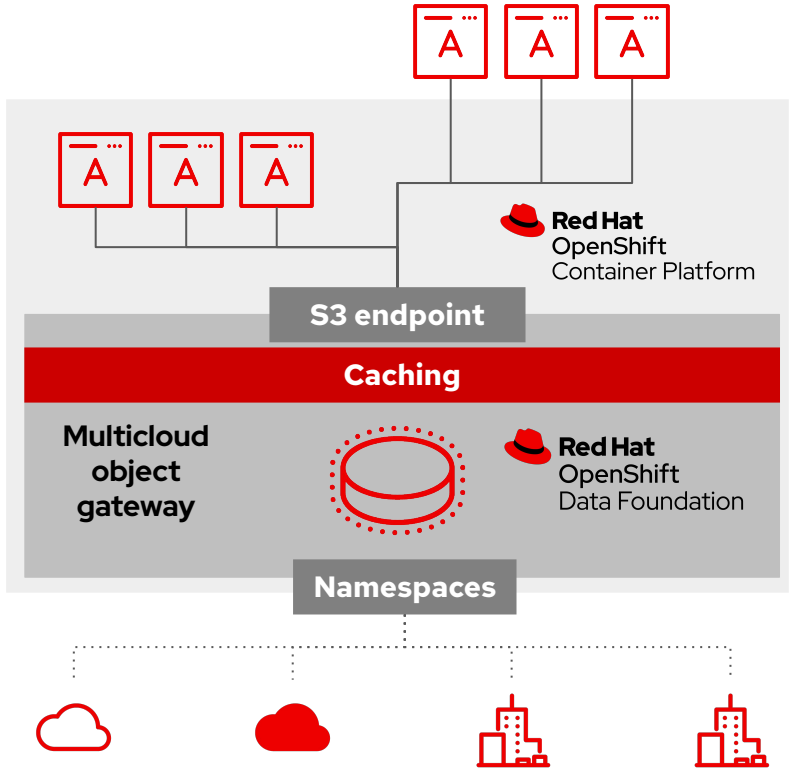
Multicloud object gateway (MCG)
Namespaces support

Jupyter Notebook example:

- Jupyter Notebook reads and writes to the same (namespaced) bucket
- Namespaced bucket has several underlying resources
- Writes are funnelled to a single bucket
- Namespace resources (object stores) are still available outside of the MCG namespaced bucket

intel + Lenovo Red Hat Data Services

FUNCTIONALITY



Multicloud object gateway (MCG)

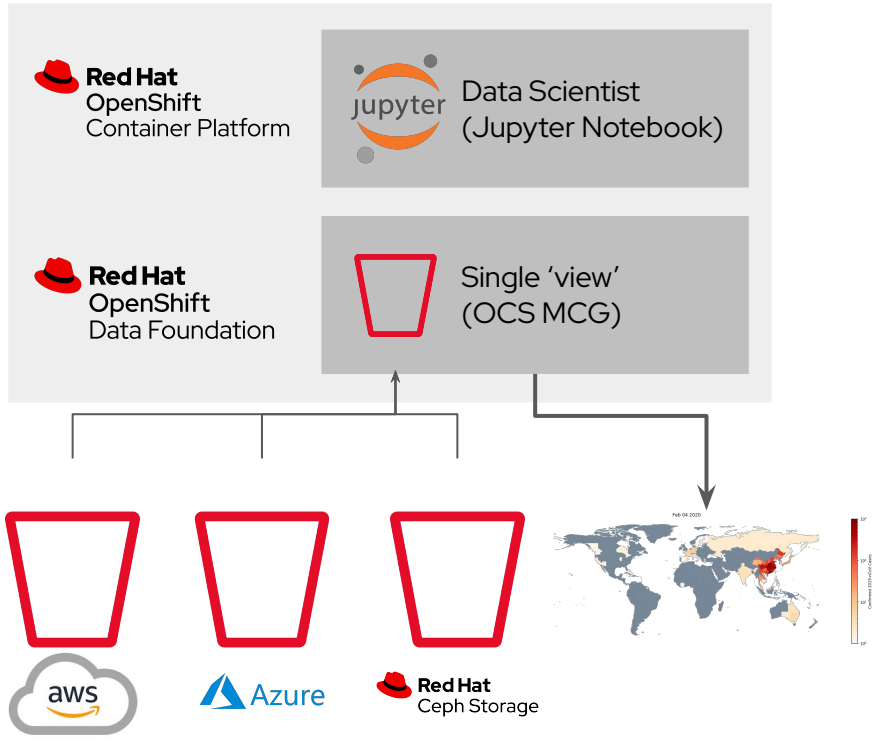
Caching support

A caching object solution for customers where data gravity is required. This is particularly useful for those using artificial intelligence/machine learning (AI/ML) platforms.

intel + Lenovo Red Hat Data Services

Multicloud object gateway (MCG)

COVID-19 mapping spread example



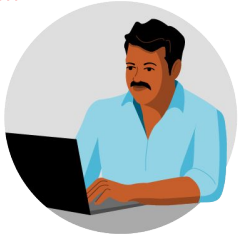
- Simple image generator for tracking outbreak distribution and spread
- Data source from John Hopkins University
- **Inputs:** CSV data (daily counts) from multiple public cloud data sources
- **Outputs:** map images, each image is an outbreak intensity heat map

intel + Lenovo Red Hat Data Services

What's new?

Red Hat OpenShift Data Foundation 4.9

BUSINESS VALUE



Data engineer

Can now replicate on-premise plain data to the cloud to take advantage of in example, cloud native AI\ML services

System administrator

Can now allow collaboration on the same data between two different locations and replicate objects data to another location in order to access the data, whenever a site location experiences outage

intel + Lenovo Red Hat Data Services

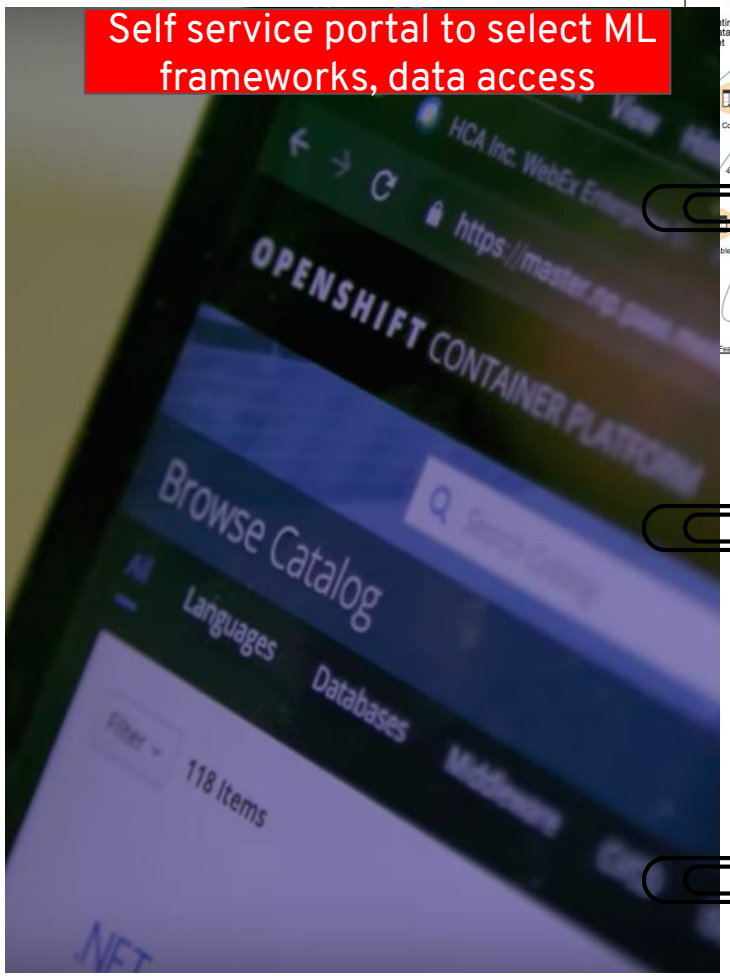
What do Data Scientists Want?

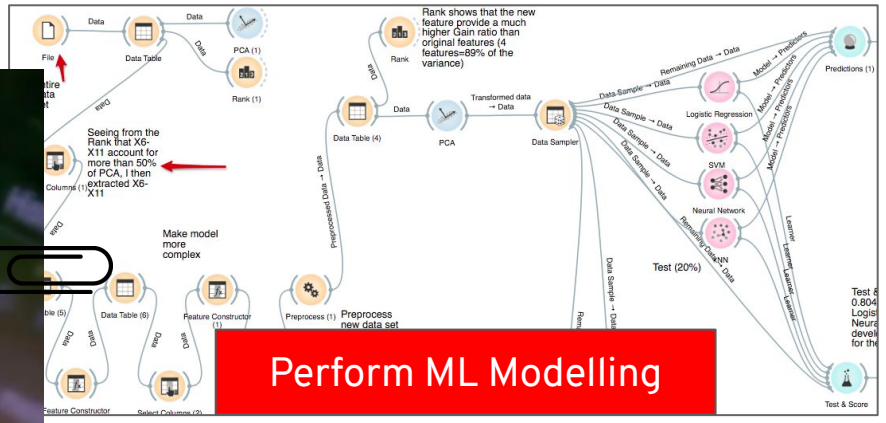
"Self-service cloud like" experience for my machine learning projects

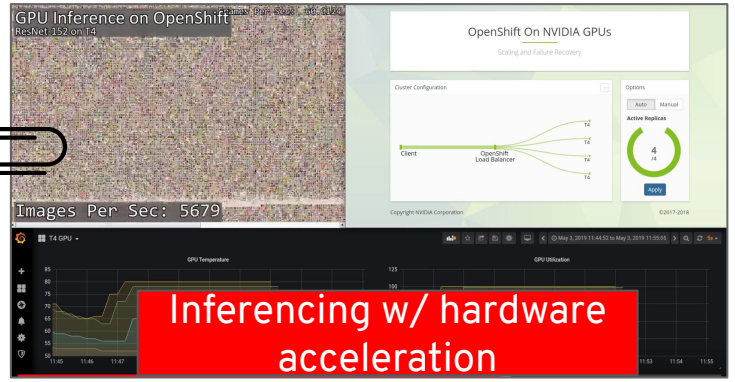
Access to a rich set of modelling frameworks, data, and computational resources

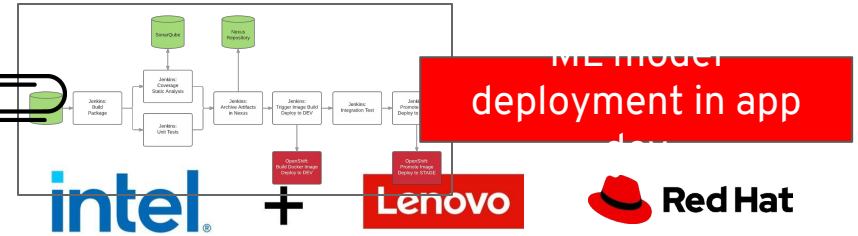
Collaborate with colleagues

Deliver my work into production with speed, agility and repeatability to drive business value









Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted advisor to the Fortune 500.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

intel + Red Hat