SILICON VALLEY
DATA SCIENCE

Big Data Analytics with Silicon Valley Data Science and Red Hat

Stephen O'Sullivan
Vice President
Engineering
SVDS

Brent Compton
Sr. Director
Storage Solution Architectures
Red Hat

May 4, 2017

Key Takeaways

- Disaggregating compute from storage provides flexibility
- Many-to-one: multiple analytics clusters to one object store
- On-demand ephemeral compute enables speed to capability

A BIG DATA PLATFORM PARABLE

R&D / HOBBY

THIS HADOOP THING IS COOL

- A few servers “under a desk”
- Kicking tires with very small amounts of data
- Learning the new tool sets

Data Size: 100GB – 500GB **Cost:** $500-$3K, but no license cost

PROOF OF CONCEPT

MAYBE THIS WILL HELP OUR BUSINESS

- “Steel thread” use case to prove platform capabilities
- A few non-prod servers
- Only storing data just for the use case, not a full dataset
- Processing servers where data is stored

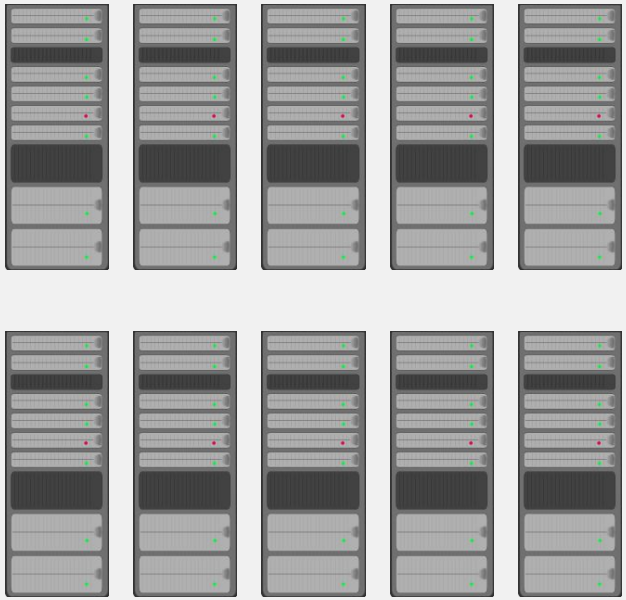
Data Size: 1TB – 5TB

Cost: $5K-$20K, still no license cost

INITIAL PRODUCTION

YES, WE CAN USE THIS

- Initial production Hadoop cluster
- Dedicated servers, network infrastructure
- Some operational support
- Full dataset volume and ingesting new data



Data Size: 50TB – 100TB

Cost: ~$500K, plus licenses at $4K-$8K per node … and ***growing***

VICTORY!

… OR JUST THE FIRST HURDLE

You're successful! But…

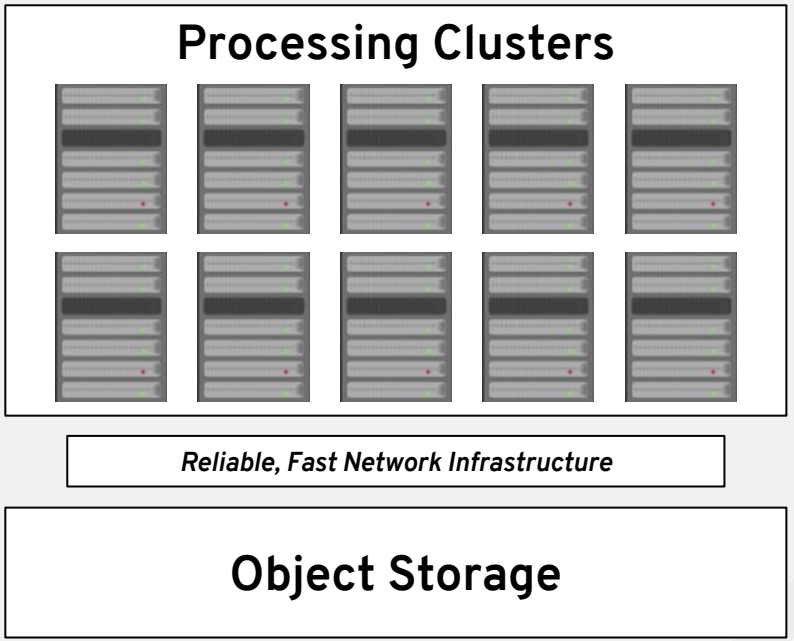
- More and more data
- Different workloads from different people
- Need to be able to explore, develop, and execute
- As you add nodes for the increased data, different consumers, and new workloads, server, licences, and support costs grow …

Do you need to scale all that compute ... or is it just storage?

WHAT ARE WE SEEING NOW?

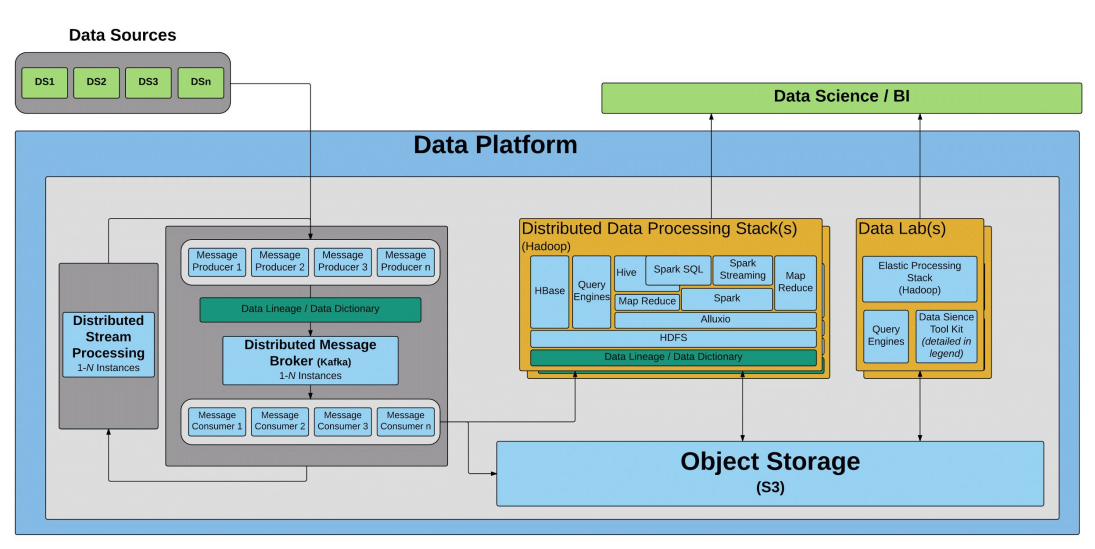
DISAGGREGATED COMPUTE AND STORAGE

- **Independently** scale compute and storage
- **Control** on licensing costs
- **Flexibility** for different datacenters / cloud regions or providers
- **Ephemeral** data labs for exploratory data science work
- **No dependency** on Hadoop for data access

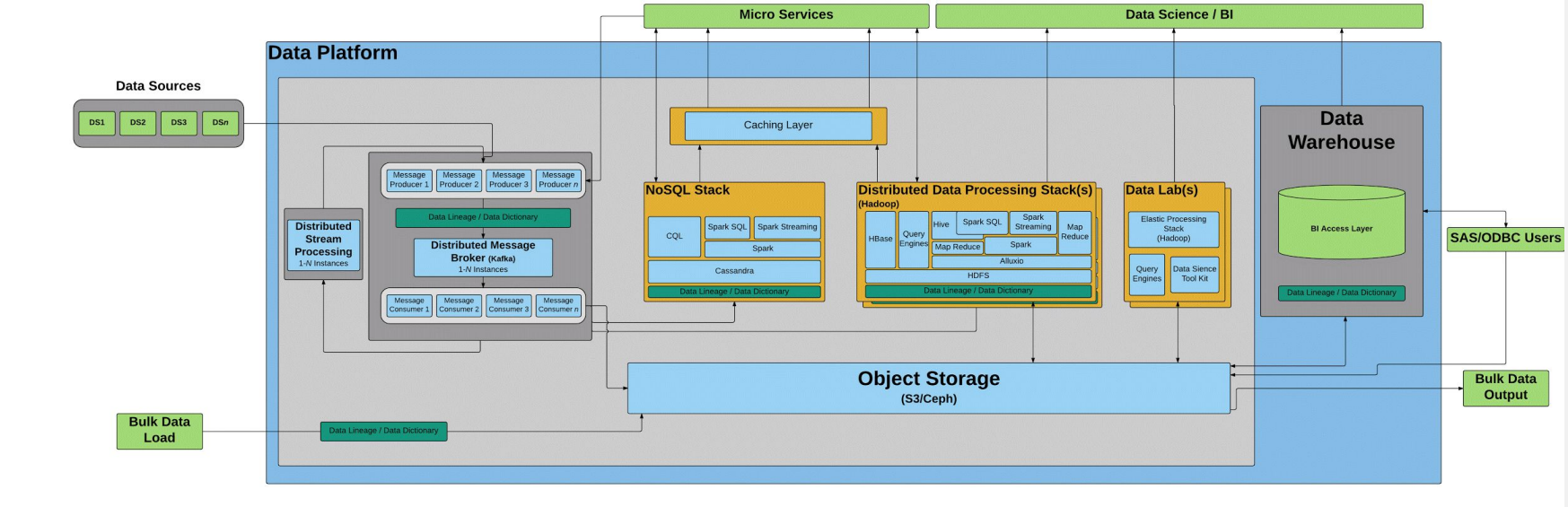


WHAT ARE WE SEEING NOW

- Multiple analytics clusters on unified object storage
- Supports production, data science, and data warehouse activities
- Read-only copies of data with lineage
- Ephemeral on-demand environments



WHAT ARE WE SEEING NOW



Unified object storage gives enterprises flexibility, control, and scalability across workloads

WHAT'S POSSIBLE—*Large European grocery company*

BEFORE

- Board-level imperative to:
    - Understand customers better
    - Run stores more effectively

CHALLENGE

- Not all data captured, or being deleted quickly
- Data silos: no central view, lots of different technology, little communication

SOLUTION

- Unified platform with full metadata and data lineage
- Data consumable by BI and Data Science users using Hadoop, Spark, SAS, Data Warehouse

RESULTS

- No waiting for months to get access
- Company-wide data catalog of all data assets
- Effective cost control means ability to store data for longer periods of time

WHAT'S POSSIBLE—*Large Global Retailer*

BEFORE

- Group had different data silos in Oracle
- Already over-subscribed for on-premise Hadoop cluster with a limited data set
- Held hostage by third-party data sources

CHALLENGE

- Certain customer data sources were being captured but were difficult to use
- Different teams needed access to data in one place

SOLUTION

- Moved to cloud with all data persisted in object store
- Analytics Hadoop clusters and toolkits
- Configurable framework for Load + Transform into different end user reporting clusters

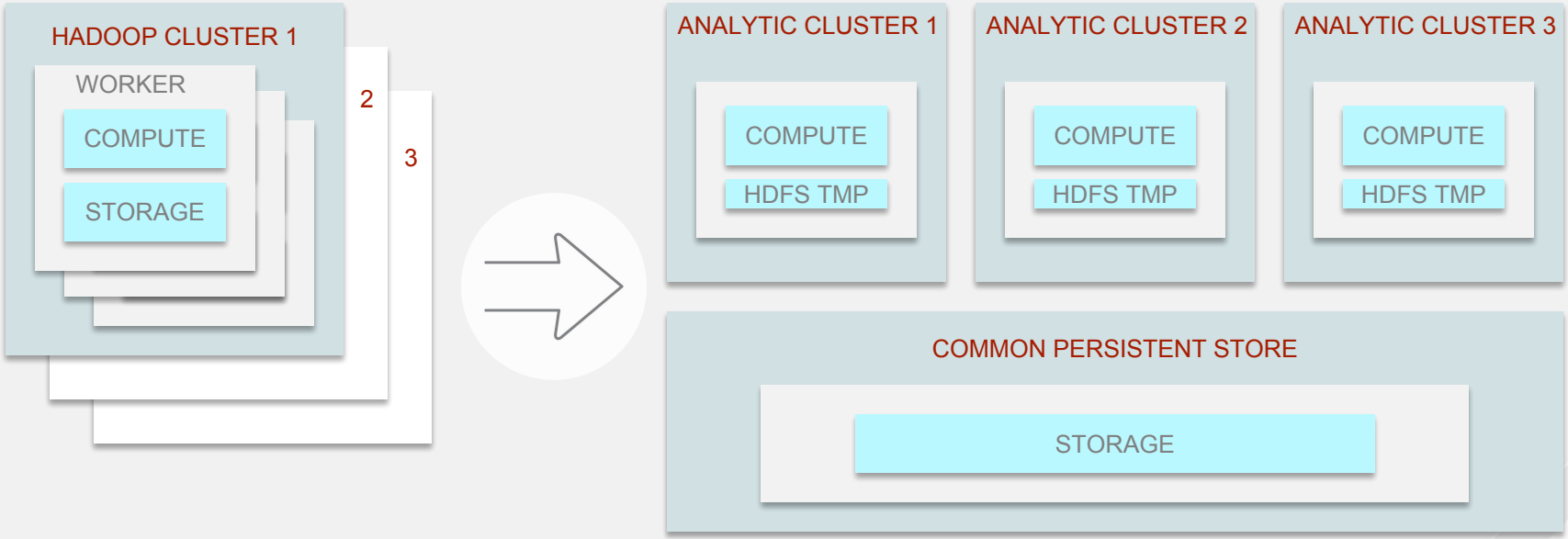
RESULTS

- Frictionless environment/tools to move data science insights to end user reporting
- Can spin up additional compute environments faster, with data from the object store

INSTANTIATING EMERING PATTERNS WITH RED HAT TECHNOLOGY

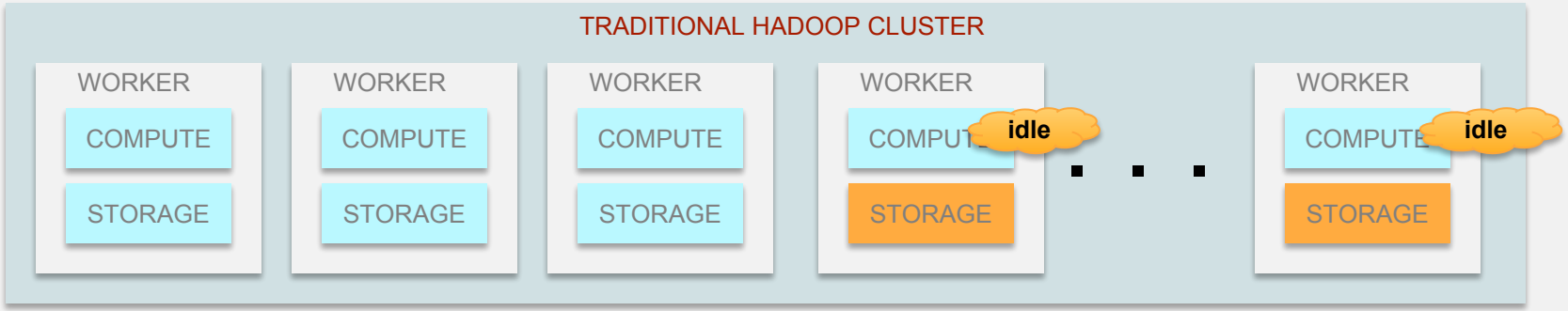
Emerging Patterns

Multiple analytic clusters, provisioned on-demand, sourcing from a common object store



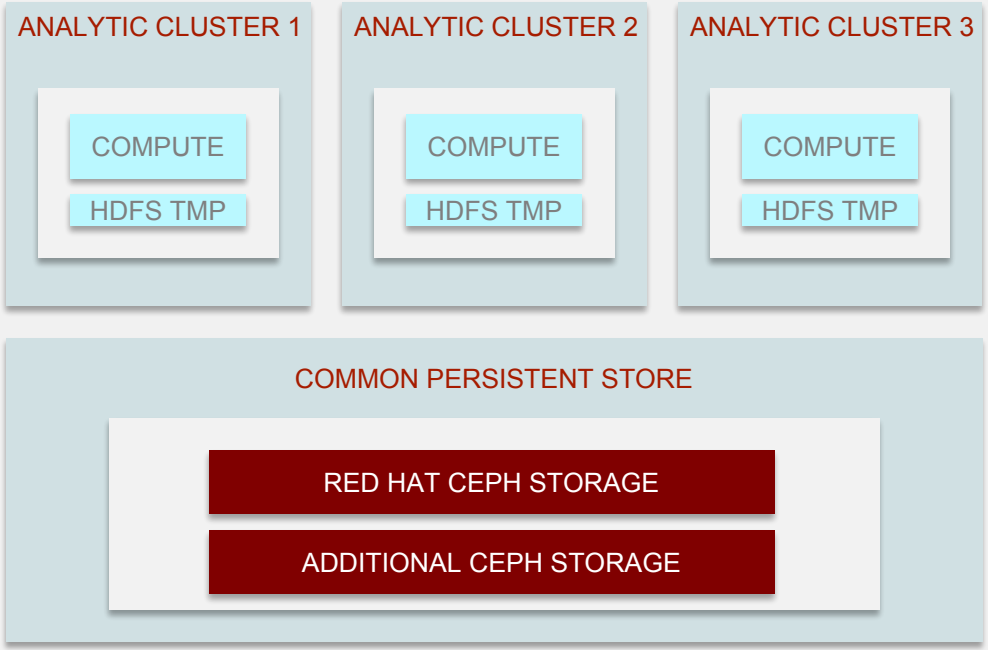
Addressing Cost Inefficiency at Scale

Adding storage capacity frequently means adding idle compute capacity too



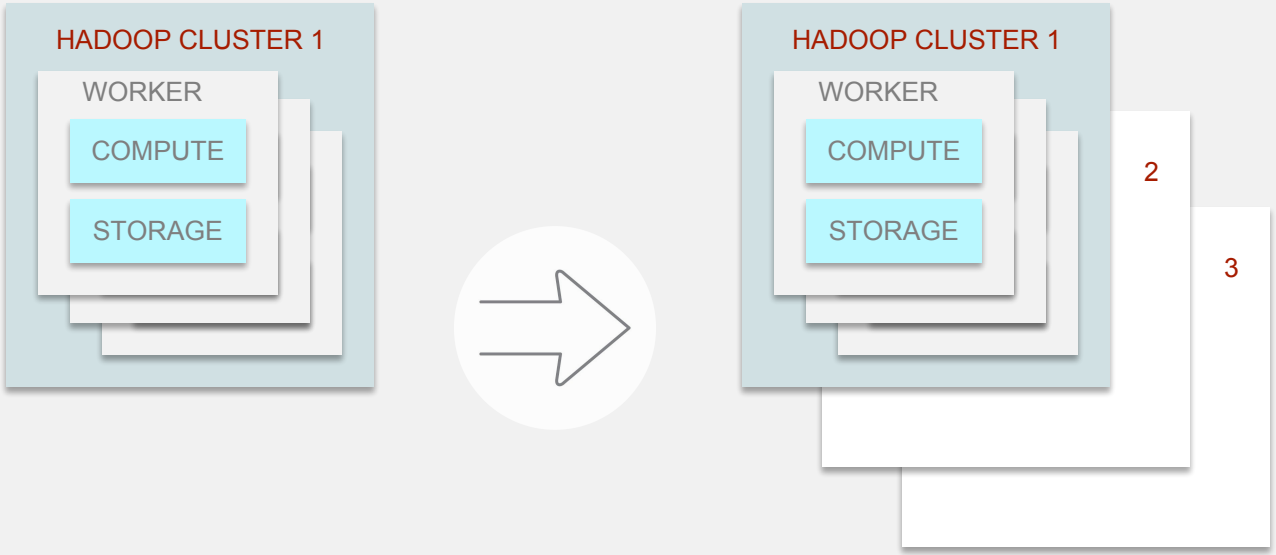
Addressing Cost Efficiency at Scale

Adding more storage should require the cost of … adding more storage



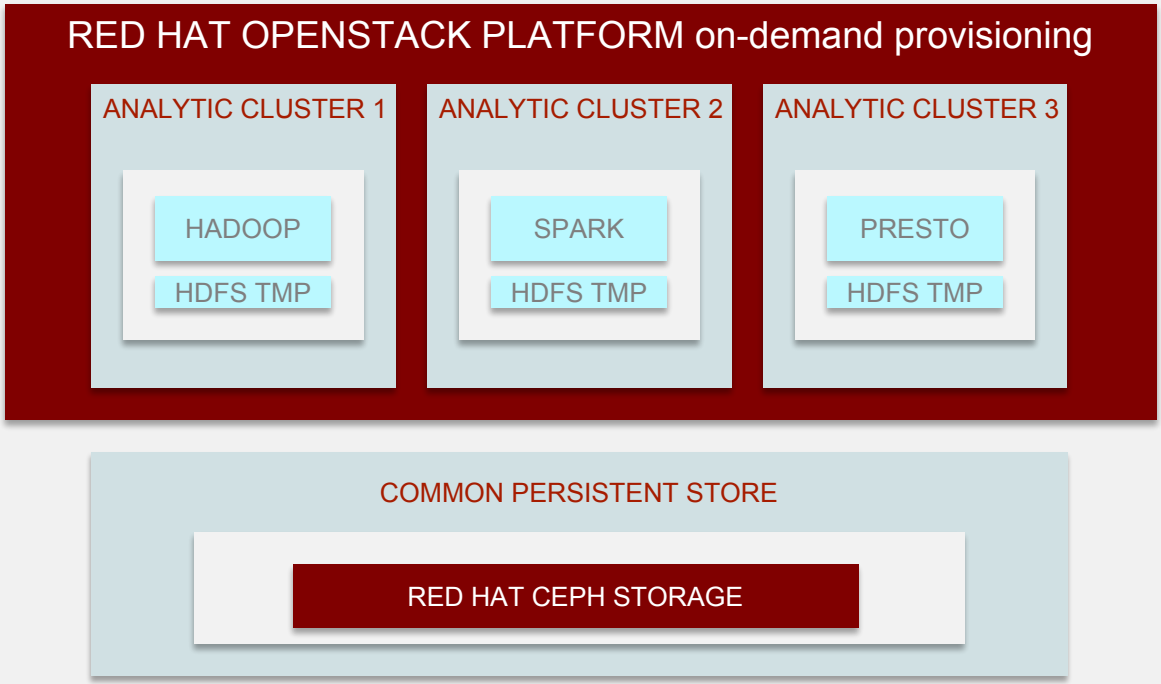
Addressing Agility Inefficiency at Scale

New analytics projects frequently require spinning up separate Hadoop/Spark clusters



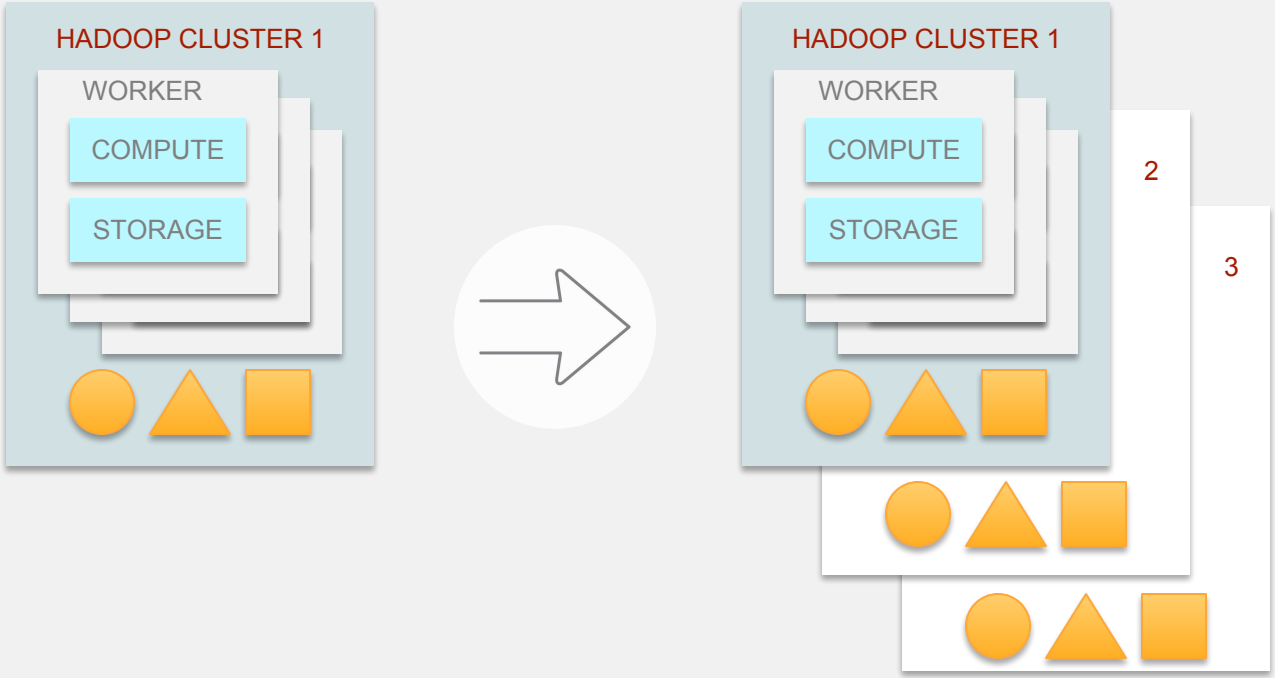
Addressing Agility Efficiency at Scale

Data teams need on-demand provisioning of analytic clusters right-sized for the job



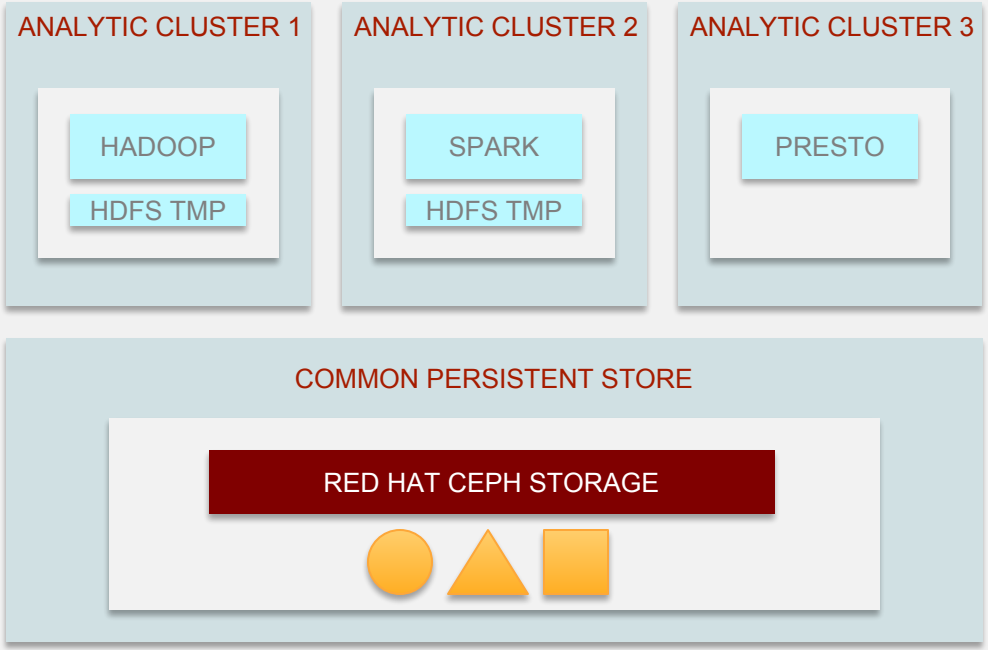
Addressing Cost Inefficiency at Scale

Multiple Hadoop/Spark clusters frequently means buying storage for full datasets on each cluster



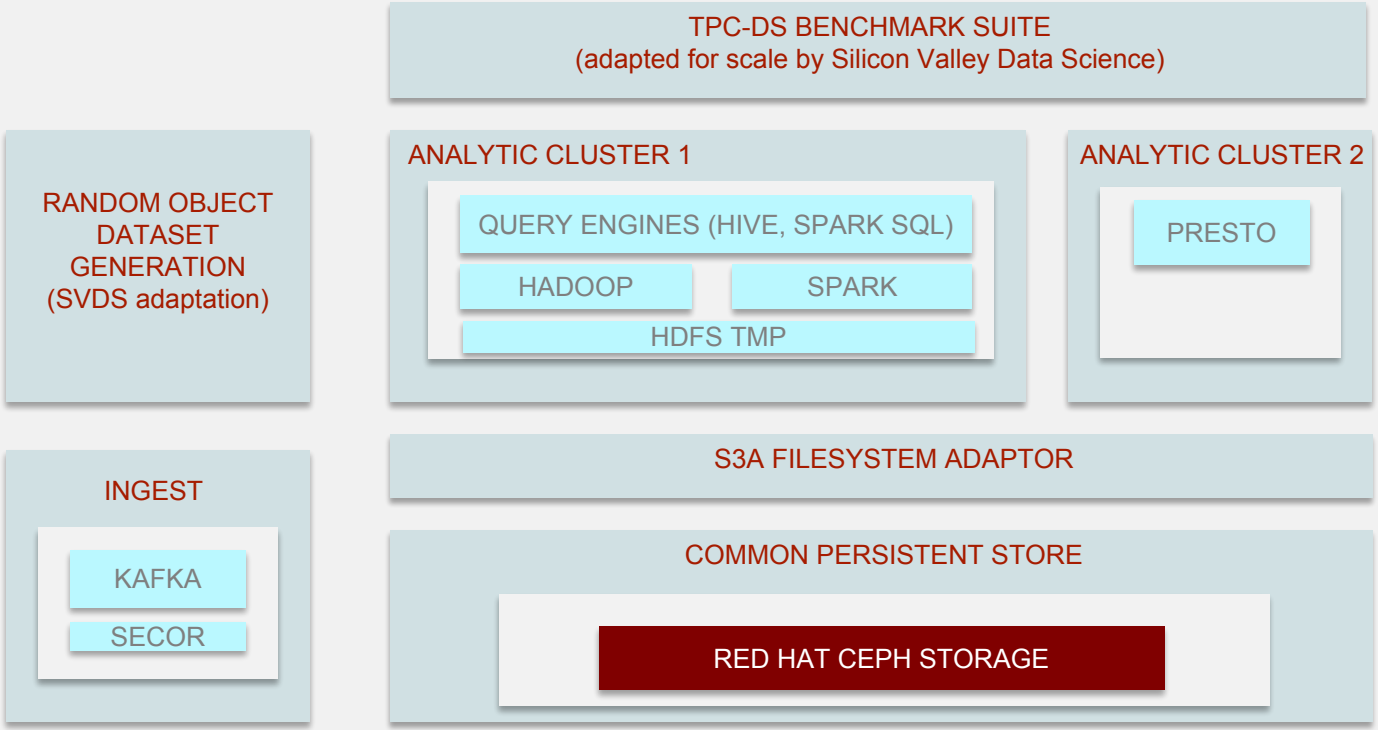
Addressing Cost Efficiency at Scale

Adding more analytic clusters doesn't require storing duplicate copies of datasets



Lab Validation and Benchmarking Underway

BIG DATA
LAB PARTNER



RED HAT SUMMIT

THANK YOU

plus.google.com/+RedHat

facebook.com/redhatinc

linkedin.com/company/red-hat

twitter.com/RedHatNews

youtube.com/user/RedHatVideos

LEARN. NETWORK.
EXPERIENCE
OPEN SOURCE.