Think Global, Act Local!

Building future proof log archives with object storage

Prepared for RedHat
June 2016

TABLE OF CONTENTS

INTRODUCTION

Mobile devices, Big Data applications and IoT are drastically changing access patterns to large scale storage systems and, today, even the smallest of data centers can generate gigabytes of logs per day. In fact, many organizations are beginning to collect logs from remote offices and all kinds of devices under their control for the most varied reasons like auditing, security, monitoring, billing and so on. This trend will soon be characterized by an even greater flow of logs and sensors, due to the massive number of machines and devices connected to the internet, which continuously send home information.

Even the smallest of datacenters can generate gigabytes of logs per day, and many organizations are beginning to collect logs from remote offices and all kinds of devices for the most varied reasons.

In this scenario, object storage (which is historically considered suitable only for cheap, durable and cold data archives), is poised to become the perfect foundation for storing logs for both analytics and long term retention. Now, throughput with scalability should not be an issue - S3/Swift APIs are supported by an increasing number of software vendors and NFS/SMB gateways make it very easy to ingest data and remain compatible with legacy environments.

For object storage systems, the process of streaming data to and from the compute cluster is seamless, while storing original data sets or results is less costly and more reliable than for any other kind of storage; and even more so now with the flexibility provided by cloud tiering mechanisms implemented by most modern platforms.

Object storage is no longer relegated to the second or third tier in storage infrastructures. Some of its characteristics are perfectly suitable to be the backend of modern IT infrastructures which need scalability, performance and availability demanded by developers, end users and cloud applications.

Software Defined Solutions like Red Hat Ceph Storage are clearly going toward this direction:

- More performance: by eliminating intermediate layers like local file systems and including optimizations for All-Flash configurations;
- More security: improved thanks to better encryption options;
- More interfaces: thanks to the maturity of scale-out FS and S3 gateway.

Its open source and software-defined nature is another key to its success. In fact, it's no coincidence that Ceph is already prevalent in OpenStack clusters and growing in popularity in a large number of object (and block) use cases.

BUILDING SUSTAINABLE LOG ARCHIVES

The amount of logs that must be stored safely and accessed from anywhere is growing exponentially. Now, with IT organizations implementing multiple cloud strategies, finding a future-proof platform capable of serving different workloads simultaneosuly, to local and remote locations, is even more challenging. For example, logs could come from the most disparate sources and can be easily tagged and indexed during the ingestion process. Once data is safely stored as objects, it's available for metadata searching. Through specific API and commands, S3 based storage can also be leveraged as a primary store or as a secondary storage system. And if the object storage has tiering capabilities, it is possible to implement tiering (and retention) policies to move data across flash, disk and cloud or tape to improve the overall $/GB of the entire storage infrastructure.

Logs are collected for many different reasons and are processed accordingly. They are utilized in real time for monitoring or through batch jobs to build reports, trend analysis and so on. Sometimes they are archived as is.

Logs are collected for many different reasons and are processed accordingly. They are utilized in real time for monitoring or through batch jobs to build reports, trend analysis and so on. Sometimes they are archived as is, because of specific regulations or compliance standards. And in that case, it's highly likely that data must reside in the country of origin and cannot be uploaded to a public cloud storage (many European based firms have strict requirements for keeping their data 'in country' for example).

Finding the appropriate storage infrastructure to store log is very challenging, since its characteristics are very demanding and costs need to be kept low as well.

Finding the appropriate storage infrastructure for this task is very challenging, since its characteristics are very demanding and so is the cost, which needs to be relatively low: reliability, durability and scalability are on top of the list, but $/GB is crucial as well. The Total Cost of Acquisition (TCA) has to be very low and features which drive down the Total Cost of Ownership (TCO) are needed to make the infrastructure future proof.

In particular, looking at TCO, taking data protection and other basic capabilities for granted, the most significant features should be related to how the log lifecycle is managed:

- **Multi-protocol access:** Logs should be able to be imported by different methods depending on the source. Only the most modern applications and devices can use HTTP-based protocols and in many cases, for example, file access is still the only way to move logs from the original application/source.
- **Performance:** the number of logs collected simultaneously could be massive, and this can generate a very large amount of network and storage traffic. Many sequential streams can be joined by

uploads of single small and large objects depending on how logs are generated. In some cases, log analysis is performed in real time while data is saved on the storage system, or immediately after. In any case, the entire infrastructure has to be designed to avoid bottlenecks.

- **Caching or tiering capabilities:** It's highly likely that most recent logs are being accessed frequently for analytics and reports but will become inactive quite quickly. Mechanisms to grant faster access to most active data while managing a low $/GB for capacity growth are at the base of infrastructure sustainability.
- **Automation:** by automating retentions management it is possible to alleviate sysadmins from tedious tasks, improve efficiency and avoid mistakes. This is especially true when logs must be stored safely for a long time for compliance with laws and regulations.

OBJECT STORAGE FOR LOG ARCHIVES

A very flexible and cost effective storage system is necessary for an effective solution to the problem of log archiving.

Generally speaking, object storage can be a very effective platform for log archives, especially at scale. But Red Hat Ceph Storage has the right characteristics to be an ideal choice for this task. In fact, it is not only an object store - it can offer more when it comes to flexibility and efficiency. With the particular nature of logs, where a long term archive could be associated to a very active short term repository accessed by applications that need high throughput for analytics and reports, its architecture can make the difference.

Object storage can be a very effective platform for log archives, especially at scale. Red Hat Ceph Storage has the right characteristics to be an ideal choice as an object store for log archives.

The scalability of this type of platform is another key point. In fact, end users usually start collecting logs from few sources at the beginning with more and more consolidation happening over time. In this particular case, a scale-out architecture, like Red Hat Ceph Storage, is critical when aiming for a low TCO. The ability to start very small, in the order of a few tens of Terabytes, and grow to the Petabyte scale without changing the way the storage infrastructure is operated, added to the flexibility brought by the freedom of choice in its configuration, is a great advantage when compared to traditional proprietary solutions.

Furthermore, Red Hat Ceph Storage is a software-defined solution which runs on Linux and, for larger environments, makes it possible to build a specialized hyperconverged-like infrastructure for log analytics and archive. Third party log analytics solutions as well as NoSQL databases can easily be installed alongside Red Hat Ceph Storage, perhaps in the same cluster, simplifying the infrastructure while optimizing the process.

BOTTOM LINE

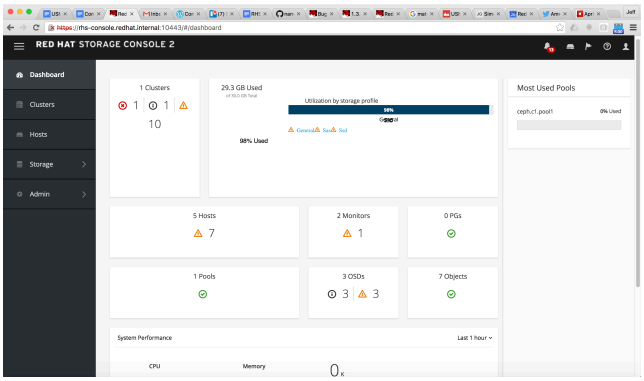
Thanks to object storage technology it is possible to build a data lake capable of concurrently streaming data, being the back-end for Big Data analytics and storing any other form of unstructured and semi-structured data, including logs.

By adopting Red Hat Ceph Storage, enterprises can effectively face the challenges posed by data growth and diversity with a future proof and cost effective infrastructure. It can be the ideal solution for building the core of sustainable data-driven infrastructures for today's and future needs.

By adopting Red Hat Ceph Storage, enterprises can effectively face the challenges posed by data growth and diversity with a sustainable cost effective infrastructure.

The multi-protocol front-end, the object storage architecture back-end, it's modern distributed design, hardware independence and open, community based development model, all contribute to making Red Hat Ceph Storage a unique solution in the market, covering a growing number of use cases.

In a modern, two tier storage strategy which sees latency-sensitive workloads served by All-Flash Arrays and capacity-driven applications covered by high throughput scale-out storage, Red Hat Ceph Storage can play a perfect role in the latter category and, thanks to the latest improvements with Red Hat Ceph Storage 2.0, it can most definitely aspire to something more in the future.



Red Hat Ceph Storage is ready to support multiple applications and workloads at the same time. In fact, all log analytics/archiving can fit in the Big Data category and they can all contribute to building a single, large and accessible enterprise data lake.

Red Hat Ceph Storage enables a smart data infrastructure that can start very small, in the order of tens of Terabytes, and grow to any size (into the multi Petabytes range) while being easy to use and manage, thanks to tools like the Red Hat Storage Console (see side graphic). New in Red Hat Ceph Storage 2.0, the Red Hat Storage Console simplifies both the deployment and operational management of Ceph, making scale-out storage accessible to a wider range of users while reducing deployment time.

Think Global, Act Local!

JUKU

WHY JUKU

Jukus are Japanese specialized cram schools and our philosophy is the same. Not to replace the traditional information channels, but to help decision makers in their IT environments, to inform and to discuss the technological side that we know better: IT infrastructure virtualization, cloud computing and storage.

Unlike the past, today those who live in the IT environment need to be aware of their surroundings: things are changing rapidly and there is a need to be constantly updated, to learn to adapt quickly and to support important decisions - but how? Through our support, our ideas, the result of our daily global interaction on the web and social networking with vendors, analysts, bloggers, journalists and consultants. But our work doesn't stop there - the comparison and the search is global, but the sharing and application of our ideas must be local and that is where our daily experience, with companies rooted in local areas, becomes essential in providing an honest and productive vision. That's why we have chosen: "think global, act local" as a payoff for Juku.

AUTHOR

Enrico Signoretti is an analyst, trusted advisor and passionate blogger (not necessarily in that order). He has been immersed in IT environments for over 20 years. His career began with Assembler in the second half of the 80′s before moving on to UNIX platforms until now when he joined the "Cloudland". During these years his job has changed from highly technical roles to management and customer relationship management. In 2012 he founded Juku consulting SRL, a new consultancy and advisory firm deeply focused on supporting end users, vendors and third parties in the development of their IT infrastructure strategies. He keeps a vigil eye on how the market evolves and is constantly on the lookout for new ideas and innovative solutions. You can find Enrico's social profiles here: http://about.me/esignoretti