

BEST PRACTICES FOR AN OPEN SOURCE SOLUTION TO BIG DATA CHALLENGES

EXECUTIVE SUMMARY

For many of today's IT specialists, "big data" isn't just another meaningless buzz phrase. It's something closer to a crisis, and it can't be wished away. The reason is simple: Big data just keeps getting bigger. For a large number of organizations – especially for those in data-intensive industries such as retail – finding cost-effective ways to gain real value from an overabundance of information will be a key factor in determining which companies remain relevant in the next few critical years.

Fortunately, practical solutions are available. Leading enterprises are turning to open, software-defined storage as a sensible way to deploy web-scale IT architectures, giving them more flexibility in how they manage a wide variety of data, even on a massive scale. As often happens with new technologies, the definitions for this innovative approach to storage may vary. Most experts, however, seem to agree that a software-defined storage environment is characterized by hardware agnosticism, distributed architectures, converged storage, and native support for standard data protocols.

But where to start? In this paper, we offer a step-by-step guide for adopting software-defined storage as part of a practical, proactive strategy for managing enterprise data. The process will always be different from company to company and application to application. Flexibility is inherent in this technology. And now more than ever, a flexible IT architecture might be the key to facing the once unthinkable challenges of big data – not to mention the challenges we haven't even thought of yet.

CHALLENGE: MASSIVE AMOUNTS OF DATA

Unprecedented data growth is the biggest challenge for many IT departments across every industry – and the difficulties just keep coming. Companies of every size are scrambling to capture and analyze a massive amount of information from a growing number of sources, doing what they can to accommodate increases in enterprise data volume that would have seemed unfathomable less than a decade ago. Given these circumstances, it should come as little surprise that analysts at IT research firm Gartner estimate that 40% of all organizations will double the size of their on-premise storage infrastructure by 2016.¹

Yet even as costs continue to explode, many IT budgets keep getting smaller, leaving internal teams unable to deliver the innovative services necessary to compete in a crowded global marketplace. As data growth becomes even more difficult to manage in the coming years, some companies may keep relying on legacy solutions that deliver severely diminished returns. These companies will not be poised for long-term success. To fully engage in our data-driven world, any organization with serious claims to industry leadership in 2020 and beyond will need to make a significant investment in a modernized storage infrastructure. Decisive action may come at a considerable cost in the short term. But make no mistake: The cost of doing nothing will certainly be much higher.



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

redhat.com

¹ "Making the hybrid cloud storage work," Gartner, 2014.

RED HAT AND CEPH: BETTER TOGETHER

In May 2014, Red Hat acquired Inktank, the provider of Ceph and Inktank Ceph Enterprise. Together, these industry-leading technologies enable customers to store and manage a whole new spectrum of data – from “hot” mission-critical data to “cold” archival data.

RED HAT APPROACH: OPEN SOURCE

When we talk about open source at Red Hat, we’re talking about a proven way of collaborating to create technology – the freedom to see the code, learn from it, ask questions, and offer improvements. We prefer to work and participate in communities of rigorous meritocracy, where the process is open, everyone has the same information, and everyone is free to make improvements. In these communities, the best ideas will always win.

For us, open source isn’t really a movement with a manifesto. It’s just the best method available. And we believe it provides the utmost agility for enterprises today.

TECHNOLOGY SOLUTIONS: MORE AGILE DATACENTER

When technology runs up against its own limitations, maybe it’s time to build a better technology. Instead of relying on older, less flexible solutions based on an outdated set of assumptions, today’s enterprises have many options for deploying web-scale IT architectures to bridge the public and private cloud. From a storage perspective, that means separating storage hardware from the software that manages it – an approach that’s come to be known as “software-defined storage.” Adoption may not be widespread at the moment, but don’t expect it to remain that way. Gartner estimates that open source storage will account for more than 20% of enterprise market share by 2018.²

One of the ironies of software-defined storage is that no one can seem to agree on a complete definition. That’s perfectly normal for any emerging technology; and vendor-biased definitions are already starting to give way to a common set of defining characteristics:

Hardware agnosticism. A software-defined storage solution should be able to run on any standard server platform not tied to a specific hardware platform. This explains, at least in part, the emergence of x86 storage servers with direct-attached disks – a clear outgrowth of the cloud and web-scale IT movements.

Distributed architecture. This is essential because one of the primary purposes of software-defined storage is to break through the limitations of more traditional scale-up network-attached storage (NAS) and storage area network (SAN) architectures.

Support for standard data protocols. The emerging industry consensus is that any comprehensive solution should include support for block, file, and object data services.

Convergence of compute and storage. As more companies move to unified computing, software-defined storage solutions should be able to run application workloads on storage nodes.

Management control plane. As with any technology that routes or forwards information across a complex network, the success of a software-defined storage solution will depend on the sophistication of its control plane to help streamline and simplify data access.

COMPETITIVE ADVANTAGE: EIGHT STEPS TO GETTING STARTED

The journey to software-defined storage is not on a predefined path. It’s flexible by nature and often iterative in practice. As you read the following steps, try to think of the process as circular rather than linear: Start by focusing on a single set of applications, then go through every step. Once you’ve finished, you can move to the next set of applications to begin again:

1. **Find your key pressure points.** What are you currently spending on storage? What are your capacity needs and constraints? Make sure to consider cost, along with requirements around flexibility, availability, and agility.
2. **Identify the workloads based on these pressure points.** Are some workloads based on unstructured data? If so, software-defined storage might be a very good choice. If, however, you have relatively small workloads based on structured data, software-defined storage might not be the right way to go.

² “The five-year storage scenario—Why storage in 2019 won’t look like what’s on the floor today,” Gartner, 2014.

**APPLIED EXAMPLES:
COMPANIES SUCCEEDING
WITH SOFTWARE-DEFINED
STORAGE**

Metro de Madrid. The legacy storage system at the eighth-longest metro system in the world, was divided into two distinct control centers. Unfortunately, the centers didn't share the same network address, so Metro staff often relied on time-consuming workarounds to meet their storage requirements. With help from Red Hat® Storage, Metro de Madrid was able to put its existing resources to optimal use, ultimately meeting its technology challenges despite considerable budgetary constraints.

Intuit. The software giant found itself attached to large proprietary systems dictating the compatibility of all storage components. This made it difficult to achieve performance levels adequate for peak periods, and it limited Intuit's ability to protect against localized or sitewide failures with a replication-ready architecture. What the company needed was a solution that would scale easily to keep pace with accelerating growth. Intuit chose Red Hat Storage and got exactly what it needed, at a cost far lower than proprietary storage systems.

3. **Determine how many applications are in play.** If you're primarily using just one monolithic application, that's usually not a good use case for software-defined storage. But if there's a broader mix of applications, software-defined storage might make a great deal of sense for you.
4. **Migrate noncritical workloads or new applications first.** To gain experience and confidence in your new storage platform, start with a few of your less-critical applications, just in case you experience some downtime. And don't worry if you encounter difficulties at the beginning – the first migrations will always be the most problematic. Once you've refined and standardized your migration strategy and processes, you'll find it much easier to migrate even your most critical applications.
5. **Decide whether your workloads will be virtualized or in the cloud.** Are you planning to run your applications on physical servers, in a virtualized environment, or in the public cloud? If you'll be using more than one of these deployment models, you'll want to choose a flexible storage platform that supports these models. That way, you won't encounter the unnecessary complication of managing separate storage islands based on incompatible technologies.
6. **Determine what analytics you'll require.** If the rise of big data is a particularly pressing concern for your company, you might consider a software-defined storage solution that relies on leading data analytics technologies such as the Apache Hadoop implementation of MapReduce. With tools such as this at your disposal, you'll be able to directly store and share data from legacy applications without moving information between silos. Not only will this enable you to extract more useful information from giant ponds of data in less time, but you'll also have the potential to save significantly on IT costs by extracting the maximum value from your existing infrastructure.
7. **Determine the right level of data protection and replication for your needs.** Which disaster recovery scenarios need to be covered? What are the likely outcomes for each scenario, including estimated cost? Is it possible to achieve the highest levels of protection for your most important data without over investing in excessive security for information that is neither sensitive nor critical?
8. **Decide how long you need to maintain your data.** Are there any regulatory requirements that might require you to maintain it for longer than you otherwise would? If not, look for ways to delete data that is no longer required – after all, why store it if you don't need it?

VALUE OF PUTTING DATA SERVICES FIRST: MANAGED WORKLOADS ON ANY SCALE

Software-defined storage, even if narrowly defined, can take a number of different forms depending on a company's environment, business requirements, budget, and more. But a truly scalable approach to software-defined storage requires decoupling data services from the underlying data structure, enabling more services to be used across a wider range of workloads. These services may include the following:

- **File services.** Traditional file systems suffer from limited scalability due to their hierarchical structures, making tasks such as data protection and capacity optimization more difficult.
- **Object services.** Object storage systems have become a common alternative to file-based and block-based systems, often providing enough metadata while simplifying the indexing of unstructured information. They provide service based on the representational state transfer (REST)

Casio. The multinational electronics manufacturer with a long tradition of innovation was committed to making web-scale storage a reality. Yet despite the company's attempts to virtualize its storage environment, it couldn't escape from vendor lock-in. At last, the organization addressed its storage-related challenges with Red Hat Storage, enabling the internal disks of multiple commodity servers to be integrated and used as one large storage pool.

Cisco. The company's networking solutions are designed to intelligently connect people, processes, data, and things. It required a high-performance storage platform that could support key use cases around block, object, and file storage capabilities in one unified platform.

architecture as a functional element in abstracting storage away from the application. However, object-based systems are often incompatible with file-based systems, requiring companies to undergo costly application rewrites.

- **Shared file-and-object services.** These services enable organizations to take advantage of the file-based applications currently in use. At the same time, they make that data available to object-based applications, often through a REST-based method. This enables maximum flexibility in standards.
- **Block services.** Block storage, often used by SANs, manages data as blocks within sectors and tracks. The OpenStack® project Cinder is one of today's more commonly used block storage services.

By having a full set of data services that can freely interact, organizations are able to desegregate information, achieving the agility they need to manage more complex workloads on whatever scale may be required.

CONCLUSION

In the coming years, as more IT specialists explore the potential of software-defined storage, our understanding of this technology will inevitably change. That's how it ought to be. Even at Red Hat, we're still learning about it ourselves, doing everything we can to consider its implications based on our initial deployments at leading-edge organizations. But those efforts will only take us so far. Technology is powered by innovation, and truly innovative ideas usually exceed whatever expectations were once attached to them.

We're comfortable with the active exploration of new ideas. Our customers seem comfortable with that, too. That's because we do what we do as a community, and we know from experience that our vision for open, software-defined storage doesn't just belong to us. We happen to be long-time believers in the power of ownership, not just ownership of a metal box in a locked cabinet, but ownership of a more interesting kind – the kind that comes from active contribution to the creation of something entirely new.

We welcome your help in creating the future of open, software-defined storage. To learn more, visit www.redhat.com/storage.

ABOUT RED HAT

Red Hat is the world's leading provider of open source solutions, using a community-powered approach to provide reliable and high-performing cloud, virtualization, storage, Linux, and middleware technologies. Red Hat also offers award-winning support, training, and consulting services. Red Hat is an S&P company with more than 80 offices spanning the globe, empowering its customers' businesses.

NORTH AMERICA
1 888 REDHAT1

**EUROPE, MIDDLE EAST,
AND AFRICA**
00800 7334 2835
europa@redhat.com

ASIA PACIFIC
+65 6490 4200
apac@redhat.com

LATIN AMERICA
+54 11 4329 7300
info-latam@redhat.com



facebook.com/redhatinc
[@redhatnews](https://twitter.com/redhatnews)
linkedin.com/company/red-hat

redhat.com
INC0204410_V2_0115

Copyright © 2015 Red Hat, Inc. Red Hat, Red Hat Enterprise Linux, the Shadowman logo, and JBoss are trademarks of Red Hat, Inc., registered in the U.S. and other countries. Linux® is the registered trademark of Linus Torvalds in the U.S. and other countries.

The OpenStack® Word Mark and OpenStack Logo are either registered trademarks / services marks or trademarks / service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundations' permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation or the OpenStack community.