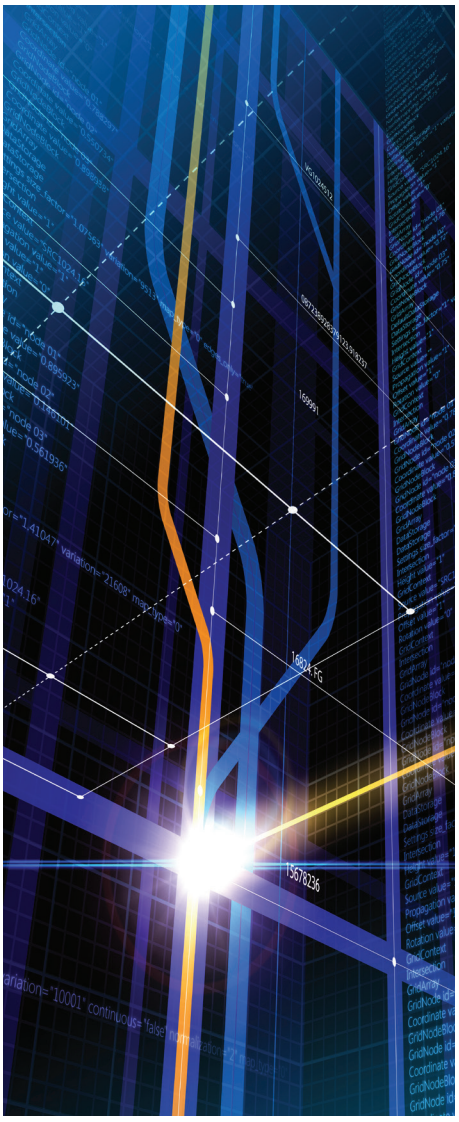
Solution Brief

The Hadoop Active Archive

It's time to take the first step: Initiate your big data strategy



Big data means big challenges for IT

As the volume and velocity of data grows, many organizations find they have to offload months or even years of historical transactions to secondary repositories or, worse still, to offline tape. What's more, they often find they need to discard much of the detail in the data, keeping only the aggregates. This results in a short window of time for the use of the full body of data and, on an ongoing basis, limited amounts of data to use in making crucial business decisions.

This loss of data can hurt the business. Most analysts today would argue that having access to data that represents exactly what happened three years back, five years back or even further back in time can help drive better decisions for the future. Ultimately, the ability to keep limitless volumes of data can lead to better intelligence and a competitive advantage.

This new reality for businesses points to the need for solutions that allow organizations to store ever-larger volumes of data in a cost-effective manner while keeping all of that data available for business uses. This isn't a need that can be met with yesterday's approaches—notably to flow all data into an expensive enterprise data warehouse. The costs of continual EDW expansions are already too high, and will soon become unsustainable for many companies.

To complicate matters, in an era of diverse data types—including structured, unstructured and semi-structured data—today's EDW costs are driven even higher by the need to continually cleanse and parse diverse types of data to prepare it for archiving and use in the EDW.

A path to the future: the Hadoop opportunity

For organizations facing these challenges, the Apache Hadoop data storage and distributed processing platform offers an affordable alternative to traditional approaches to the storage and archiving of large amounts of data. Research indicates that enterprises can store data in Hadoop for about a tenth of the cost of using a traditional EDW.

The Economics of Data

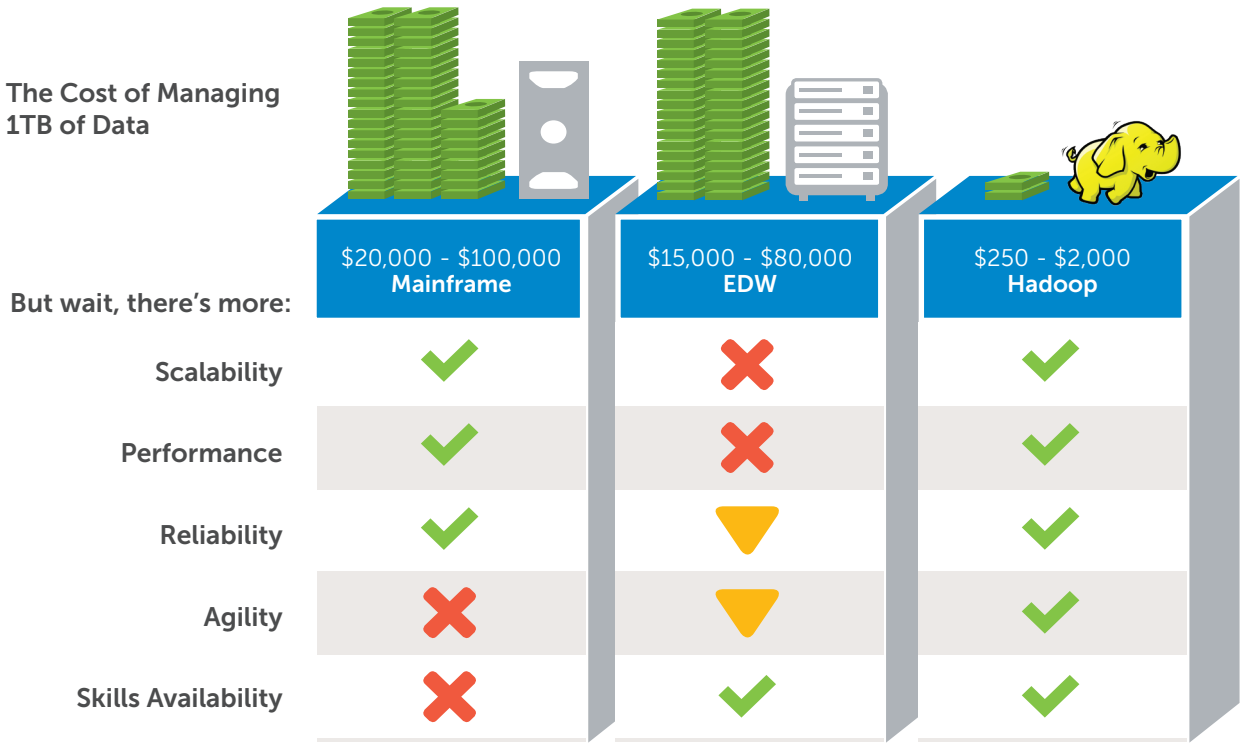


Figure 1. The cost advantages of the Hadoop platform.[1]

While the open source Hadoop platform is attractive from a cost-benefit standpoint, it exists in unchartered waters for many organizations. If they haven't worked with Hadoop before, IT managers are understandably leery of launching large-scale Hadoop projects. They can't be sure of what they are getting into and what sort of skills gaps they could face.

So how do you move forward into this new realm for big data storage and processing, given all of the uncertainties? At Dell, we recommend that organizations think in terms of strategic steps, and not giant leaps. You can take the first step into the Hadoop world by starting with a narrowly targeted use case that any organization could benefit from. That's active archiving.

What is an active archive?

An active archive is a system that enables your organization to capture, retain, search and query data within an online archiving environment. Unlike deep archiving solutions and offline archives, which tuck data away mainly for safekeeping, an active archive keeps data readily accessible to business users.

The Hadoop platform is ideally suited to serve as a platform for active archiving. It gives you the ability to store any type of data, in any format, from any source, inexpensively and at very large scale, without doing a lot of data cleansing and parsing before archiving. Hadoop allows you to avoid this upfront work, thanks to its ability to store structured, unstructured and semi-structured data in native formats.

While providing cost-effective data archiving, a Hadoop environment can enable broad organizational access to varied data sets for ad-hoc analysis. Unlike conventional archives, the data stored in a Hadoop environment can be queried without pulling it into an EDW. And it makes it possible to retain all of your data for long periods of time in a cost-effective manner—which wouldn't be possible with a costly EDW as your archive.

With Hadoop as your data lake, you get the best of all worlds: You can store all types of data at a fraction of the cost of EDW storage while retaining the ability to query and analyze that data at any point in the future. And, better still, with the addition of analytics tools, Hadoop allows you to perform sophisticated analysis of data quickly and easily. This capability opens the door to predictive analytics and other advanced use cases.

And looking ahead, your Hadoop-based active archiving initiative can serve as a first step in your big data journey. It can help your team develop Hadoop experience and build a skills base for

1 Syncsort Inc. "5 Steps to Offload Your Data Warehouse with Hadoop." 2014

follow-up projects. Once your active archive is in place, you can build on your Hadoop successes with a wide range of high-value use cases, such as offloading ETL workloads from your EDW and using data analytics to gain a 360-degree view of your customers.

Moving forward with Dell, Cloudera and Intel

Together with partners Cloudera and Intel, Dell offers everything you need to initiate your big data journey with the deployment of a Hadoop-based active archive.

Reference architectures

Dell helps you accelerate your Hadoop deployment with tested and validated reference architectures that have been leveraged in enterprises around the world. These reference architectures give your organization a blueprint that covers all of the components for a complete Hadoop environment, including software, hardware, networking and services.

Dell reference architectures for Hadoop bring together leading-edge Dell™ PowerEdge™ R730xd servers with the Intel® Xeon® processor E5 v3 family, Dell Networking gear and the Cloudera Enterprise offering, which includes the Cloudera distribution of Apache Hadoop open source software.

Proven performance

The performance of Hadoop solutions based on the PowerEdge R730xd server with Intel Xeon E5 v3 processors has been proven in benchmark testing. In 2014, the Dell Server Performance Analysis Lab ran a generational performance comparison between multiple processor configurations on the PowerEdge R720xd server and the PowerEdge R730xd server. With the new generation of Intel Xeon E5 v3 processors, the PowerEdge R730xd server delivered significant performance increases in the Hadoop benchmarks for K-Means Clustering, Teragen, Terasort and Test-DFSIO.[2]

Powerful software options

In addition to a robust Hadoop environment, Dell offers sophisticated tools for data analytics, integration and management. These include:

- Dell Statistica Big Data Analytics for integrated information modeling and visualization in a big data search and analytics platform
- Dell SharePlex Connector for Hadoop for loading and continuously replicating changes from an Oracle database to a Hadoop cluster
- Dell Boomi for synchronizing data between mission-critical applications—on-premises and in the cloud

Expert services and hands-on trials

To help you gain the greatest value from your Hadoop investments and solve any skills gaps, Dell offers the expertise of consulting engineers and a full portfolio of professional services for Hadoop deployments.

In addition, you can leverage the resources of a Dell Customer Solution Center to explore your Hadoop opportunities. Located in key sites around the globe, these technical centers give you the opportunity to experience Dell solutions and technology in a dedicated, hands-on environment equipped with state-of-the-art labs and teams of solution experts.

Key takeaways

The creation of a Hadoop active archive is an ideal first step down the road to a big data strategy and powerful analytics applications. Dell can get you there today.

When you work with Dell to build your active archive, you can have the confidence that comes with an organization that has worked with Hadoop for years and maintains close working relationships with Intel and Cloudera, the leading provider of Hadoop-based software and services. You can also have the confidence that comes with a technology partner that offers proven reference architectures for Hadoop and all of the components for an end-to-end solution.

Ultimately, Dell, together with Cloudera and Intel, delivers everything you need to launch your big data journey with a Hadoop-based active archive tailored your specific requirements.

2 For details on benchmark results, see the Dell white paper "Hadoop Reference Configurations — PowerEdge R730/R730xd."

Explore leading-edge technologies and services for big data and Hadoop environments: Dell.com/Hadoop | Dell.com/BigData

For questions: Hadoop@Dell.com

© 2016 Dell Inc. All rights reserved. Dell, the DELL logo, the DELL badge and PowerEdge are trademarks of Dell Inc. Other trademarks and trade names may be used in this document to refer to either the entities claiming the marks and names or their products. Dell disclaims proprietary interest in the marks and names of others. Intel and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

April 2016 | Version 1.0