Three Ways to Get Started with Hadoop

Get on the road to big data success with proven approaches

Executive summary

To extract value from an ever-growing onslaught of data, your organization needs next-generation data management, integration, storage and processing systems that allow you to collect, manage, store and analyze data quickly, efficiently and cost-effectively. That's the case with Dell™ | Cloudera® Apache™ Hadoop® solutions for big data.

These solutions provide end-to-end scalable infrastructure, leveraging open source technologies, to allow you to simultaneously store and process large datasets in a distributed environment for data mining and analysis, on both structured and unstructured data, and to do it all in an affordable manner.

While the opportunity is clear, the path forward can be a cloudy one for many organizations. The deployment of a large-scale Hadoop environment is a complex undertaking that comes with all the risks of a big technology project. Given the unknowns, it doesn't make sense to leap into a full-scale Hadoop deployment. Instead, you need to get started with Hadoop in a manner that gives your IT professionals hands-on experience with the software and a close understanding of what Hadoop can and cannot do for your organization. At that point, you're ready to move to a broader Hadoop solution.

If your organization is on this path, Dell has the experience, the knowledge and the expertise to help you to identify the best path for your Hadoop exploration. Dell offers three ways to initiate your journey: adoption and integration of the Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture, deployment of the Dell QuickStart for Cloudera Hadoop packaged solution and exploration of Hadoop software via a Dell Customer Solution Center.

This business-oriented white paper explains these three options for starting your Hadoop journey. This paper also outlines the benefits of Hadoop and highlights some of the many use cases for this new approach to managing, storing and processing big data.

The case for Hadoop

For organizations trying to extract value out of mountains of structured and unstructured data, the Hadoop data storage and processing system offers compelling benefits. It can store any kind of data from any source, inexpensively and at very large scale, and it can do very sophisticated analysis of that data easily and quickly.

Hadoop is scalable, fault-tolerant and distributed. The open source software was originally developed by the world's largest Internet companies to capture and analyze the massive amounts of data they generate. Unlike earlier platforms, Hadoop can store any kind of data in its native format and be used to perform a wide variety of analyses and transformations on that data.

Hadoop allows you to store terabytes, and even petabytes, of data inexpensively. Both robust and reliable, Hadoop handles hardware and system failures automatically, without losing data or interrupting data analyses. Better still, Hadoop runs on clusters of commodity servers. Each of those servers has local CPU and storage resources, and each has the flexibility to be configured with the proper balance of CPU, memory and drive capacity to meet your specific performance needs.

Business-driven results

Hadoop solves the hard-scaling problems that come with large amounts of complex data. As the amount of data in a cluster grows, new servers with local storage can be added incrementally and inexpensively. Thanks to the use of MapReduce technology that takes advantage of the processing power of the servers in the cluster, a 100-node Hadoop instance can answer questions on 100 terabytes of data just as quickly as a 10-node instance can answer questions on 10 terabytes.

The Hadoop edge

Hadoop delivers several key advantages:

- **Store anything.** Hadoop stores data in its native format, exactly as it arrives at the cluster. This allows you to avoid the downside of a common alternative, translating data on arrival so that it fits into a fixed data warehouse schema, which destroys information. Because Hadoop stores data without forcing that transformation, no information is lost. Downstream analyses run with no loss of fidelity.
- **Control costs.** Hadoop is open source software that runs on commodity hardware. That combination means that the cost per terabyte, for both storage and processing, is much lower than on older proprietary systems. As your storage and analytic requirements evolve, your Hadoop installation can, too.
- **Use with confidence.** The Hadoop community, including both developers of the platform and its users, is global, active and diverse. Companies across many industries participate, including social networking, media, financial services, telecommunications, retail, health care and others.
- **Scale with confidence.** You may not have petabytes of data that you need to analyze today. Nevertheless, you can deploy Hadoop with confidence because companies like Facebook, Yahoo! and others run very large Hadoop instances managing enormous amounts of data. Hadoop is ready to scale with your needs.

Hadoop makes it possible to conduct the types of analysis that would be impossible or impractical using any other database or data warehouse. Along the way, Hadoop helps you reduce costs and extract more value from your data.

Diverse use cases

Hadoop is different from older database and data warehousing systems, and those differences can be confusing to IT professionals. What data belongs in a Hadoop cluster? What kinds of questions can the system answer?

Cloudera, the leading provider of Hadoop-based software and services, offers these examples of common use cases for Hadoop, and the questions they can answer.

1. **Risk modeling:** How can your company better understand your customers and markets?
2. **Customer churn analysis:** Why does your company really lose customers?
3. **Recommendation engine:** How can your company predict customer preferences?
4. **Ad targeting:** How can your company increase the efficiency of your ad campaigns?
5. **Point of sale transaction analysis:** How can you target retail promotions that are sure to make customers buy?
6. **Analyzing network data to predict failure:** How can your IT organization use machine-generated data to identify potential trouble?
7. **Threat analysis:** How can your company detect threats and fraudulent activity?
8. **Trade surveillance:** How can you spot a rogue trader in the financial services industry?
9. **Search quality:** What's in your search?
10. **Data sandbox:** What can you do with new data?

One of the keys to getting started with Hadoop is to identify the specific use cases that are right for your organization, and then identify the workloads that will best leverage a Hadoop environment. A related key to getting started is to identify the metrics you will use to gauge the success of your Hadoop deployment.

Three ways to get started with Hadoop

To help your organization get on the path to the benefits of a Hadoop environment, Dell offers three ways to start your journey: Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture, Dell QuickStart for Cloudera Hadoop and Dell Customer Solution Centers. Let's walk through these options.

Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture

Processing and storing ever-increasing data volumes with traditional enterprise data warehouses (EDWs) and related data-integration technologies are overloading systems and taxing IT budgets. These are key reasons many organizations are offloading extract-

transform-load (ETL) processes from the EDW to Hadoop.

The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture enables this shift. It helps your organization lower data transformation costs and build operational efficiencies while laying a robust, cost-effective, secure and scalable foundation for managing data while maturing into advanced data analytics.

Jointly designed by Dell, Cloudera, Intel and Syncsort, this tested and validated Reference Architecture outlines the end-to-end components for a complete ETL offload solution. The solution includes all the hardware, software, resources and services you need to turn Hadoop into a robust ETL environment. With this industry-first end-to-end approach, you can be in production with Hadoop for ETL offload in a shorter time than would be typically possible with a homegrown solution.

The core components of the solution include the reference architecture, software from Cloudera and Syncsort, and hardware and services from Dell. Specific solution components include:

- Cloudera Distribution for Apache Hadoop (CDH), the leading Hadoop distribution
- Syncsort DMX-h ETL software, including SILQ, a web-based utility that helps you shift ELT processing into Hadoop with an easy-to-use "no coding" approach
- Dell™ PowerEdge™ R series servers with Intel® Xeon® processors
- Dell™ networking
- Optional Dell consulting and integration services

Dell QuickStart for Cloudera Hadoop

Dell QuickStart for Cloudera Hadoop offers an easy entry point for your organization to begin managing and analyzing data. It's an all-in-one system designed to reduce the complexity of deploying, configuring and managing Hadoop systems. QuickStart includes the hardware, software and services needed to deliver a Hadoop cluster that enables your organization to quickly engage in Hadoop testing, development and proof of concept (PoC) work.

Through the combination of Dell™ PowerEdge™ servers, Cloudera Enterprise Basic Edition and Dell Professional Services, your organization can quickly deploy Hadoop and enable your development and application teams to test business processes, data analysis methodologies and operational needs against a fully functioning Hadoop cluster.

Dell QuickStart for Cloudera Hadoop builds on Dell's deep expertise and working relationship with Cloudera and Intel. The solution represents a unique collaboration in the big data ecosystem to collectively deliver an easy and affordable way to get started with Hadoop.

With the added flexibility of the Dell Professional Services, you can choose a combination of training, installation and application development that is right for your organization. By adjusting the services combination to the unique needs of your organization, you can ensure that your team gets the most out of QuickStart, and quickly determines how best to grow to production usage with Hadoop.

Dell Customer Solution Centers

Another way to get started down the path to Hadoop is to leverage the resources of a Dell Customer Solution Center. 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. To date, approximately 750 organizations have used Dell Customer Solution Centers to investigate next-generation, scale-out computing technologies, including Hadoop.

Organizations that leverage the resources of Dell Customer Solution Centers to investigate Hadoop solutions often work in tandem with Dell or third-party database and business consultants who help them explore their data challenges and identify use cases that are candidates for a Hadoop environment. When you take this approach, you gain an upfront view of your ideal Hadoop use cases.

Key benefits

The Dell | Cloudera | Syncsort Data Warehouse Optimization – ETL Offload Reference Architecture helps you:

- Reduce Hadoop deployment to weeks, develop Hadoop ETL jobs within hours, and become fully productive within days after deployment
- Achieve significant improvements in business agility
- Avoid unsustainable EDW upgrade costs just to keep the lights on
- Optimize your EDW by freeing up valuable storage and processing capacity for faster queries and other workloads more suitable for the EDW
- Start building your enterprise data hub (EDH)

Key benefits

Dell QuickStart for Cloudera Hadoop:

- Simplifies procurement and deployment with a bundled Hadoop solution
- Enables your organization to respond quickly to business needs
- Allows your IT team to focus on gaining strategic value from the platform vs. deploying the platform
- Includes services for help where needed
- Offers an aggressively priced entry point

Key benefits

Dell Customer Solution Centers give you the power to:

- **Collaborate**—Dell solution experts employ industry use cases to help you explore your Dell Hadoop solution in state-of-the-art labs.
- **Validate**—Together, we'll test your Dell Hadoop solution against current business objectives and future scalability needs.
- **Innovate**—With the experience you gain in a Dell Customer Solution Center, you can deploy your Hadoop solution with confidence.

At that point, you can access the resources of a Dell Customer Solution Center, at no cost, to learn about Dell solutions for Hadoop and other big data needs and to execute a proof of concept with clear metrics for success. The IT professionals in the Dell Solution Center can help you set up the required infrastructure and execute your PoC.

Your PoC experience in the Dell Solution Center enables you to validate your chosen Hadoop solution and reduce the risk associated with the next steps in your Hadoop journey—a test implementation on your own premises with your own data followed by deployment of a production environment.

With any of these paths forward, Dell, Cloudera and Intel deliver enterprise-grade Hadoop. Here are a few of the benefits:

- Dell, in partnership with Intel and Cloudera, delivers a broad range of solutions for the enterprise optimized for key workloads and powered by Intel compute, networking, and storage technologies
- In 2014 Intel and Cloudera jointly announced a cooperation that saw Intel making investments in Cloudera and also an intent to work with Cloudera to integrate Intel contributions into core Apache Hadoop projects and Cloudera's CDH
- You can now adopt Hadoop more rapidly and with confidence as key capabilities for security, performance, and management are jointly addressed by these leaders in Big Data Hadoop

Start your Hadoop journey with Dell

The open source Hadoop software platform gives your organization the ability to store and analyze data more affordably than ever before. With its power and flexibility, Hadoop offers an ideal complement to your existing data warehousing infrastructure.

When you partner with Dell for your Hadoop exploration and deployment, you have the confidence that comes with an organization that has worked with Hadoop since 2008 and maintains a close working relationship with Cloudera, the leading provider of Hadoop-based software and services.

Dell has what it takes to help you gain hands-on experience with Hadoop through a proof of concept and then take your solution into a full production environment—guided by proven reference architectures, enabled by package solutions and supported by the Dell Professional Services organization.

Authors: Armando Acosta is a senior product line consultant at Dell, specializing in Dell big data and Hadoop solutions. Brandon Draeger is director of marketing and strategy for Big Data solutions within Intel's Datacenter Group.

To learn more, visit Dell.com/Hadoop or Dell.com/BigData.