



Taking control of big data

By Joey Jablonski and Aurelian Dumitru

Dive deeper

To discover how the Apache Hadoop solution from Dell and Cloudera meets the big data needs of enterprises, select the Content Type > White Paper check box for a list of in-depth white papers.

dell.com/hadoop

Organizations wresting insights from vast volumes and varied data types find that big data requires big power. Dell and Cloudera have teamed up to provide an end-to-end solution that leverages the ability of Apache Hadoop to process massive data sets.

Across a broad range of industries and agencies, organizations seek to extract value from mountains of structured and unstructured data. Retailers mine big data to discover trends, buying patterns, and the tastes of individual customers. Financial services companies analyze big data to detect fraudulent transactions. Web 2.0 companies use big data for marketing and matching potential buyers with sellers of goods and services, while national governments probe ever-growing data sets to connect the dots for national security and intelligence.

A common thread in these and countless other use cases is the need for extremely scalable environments that can store and analyze large and ever-growing amounts of data. These systems must be able to handle a wide range of data types and data structures and to process data at lightning-fast speeds to drive timely decisions. Today, analysts and administrators are collecting more data than ever before and want to search and analyze that data in new, faster ways.

That's the ideal. In reality, many organizations find themselves unable to derive valuable insights from their data. They are held back by rigid

data management systems that cannot accommodate the large data volumes and diverse data types that characterize big data. They cannot find answers to questions they are asking of their data fast enough to meet organizational needs. And all the while, costs are on the rise, in part due to vendor lock-in.

Many organizations are left asking, "How can we get better at this? How can we more effectively and efficiently store, process, analyze, and use all the data we are capturing?"

Addressing big data requirements

To effectively manage big data, one option is improving status quo—the systems that are already in place. In many cases, the status quo is a data management platform that brings together a relational database management system (RDBMS), business intelligence software, and tools that extract data from source systems, transform it for targeted uses, and load it into a data warehouse. The problem here, though, is that these legacy RDBMS platforms were not designed to handle today's onslaught of structured and unstructured data and widely varied data types.

Another option includes petabyte-scale enterprise data warehouses (EDWs) with massively parallel processing (MPP) capabilities. Available in several commercial varieties, these platforms offer features that are well suited for big data, such as

Use case	Description
Data storage	Collect and store unstructured and semi-structured data in a fault-resilient scalable data store that can be organized and sorted for indexing and analysis
Batch processing of unstructured data	Batch process (index, analyze, and so forth) large quantities of unstructured and semi-structured data
Data archive	Archive data from an EDW or database management system (DBMS) for 12 to 36 months to meet data-retention policies and compliance guidelines
Integration with data warehouse	Transfer data stored in Hadoop to a separate DBMS for advanced analytics and possibly transfer data from the DBMS back into Hadoop
Behavior simulation and prediction	Run mathematical models, such as Monte Carlo methods, to simulate and predict system response and behavior

Figure 1. Apache Hadoop is designed to handle use cases involving big data

shared-nothing architectures and the ability to run analytics directly in the database. Still, they may come with strings attached—in the form of vendor lock-in, costly software licenses, code that is controlled by the vendor, and a limited ecosystem to drive the technology forward.

For reasons like these, many organizations are looking to a third option—an open source analytics platform built from the ground up to address today's big data challenges. This is where the Apache™ Hadoop™ platform enters the picture. Hadoop enables organizations to load and consolidate data from various sources into an extremely scalable distributed file system for data storage, the Hadoop Distributed File System (HDFS). This data can then be processed using highly distributed compute jobs through the MapReduce framework.

Hadoop complements, rather than replaces, existing data management platforms (see Figure 1). Along with high scalability, it is designed to deliver analysis and processing tools and support for a much wider variety of data types than those supported by traditional databases.

Although a relatively young framework, Hadoop is used today across a broad spectrum of industries, including e-tailing and retailing, financial services, government, health and life sciences, telecommunications, and Web and digital media services (see the sidebar, "Putting Hadoop to work"). Industry observers have noted the rapid rise of Hadoop. In a post about Hadoop, James Kobielus, a senior analyst for data warehousing at Forrester Research Inc., declared, "The bottom line is that Hadoop is the future

Putting Hadoop to work

Organizations using Hadoop include some of the world's largest online retailers, social networking sites, and other service companies. Retailers and e-tailers use Hadoop to find patterns in customer purchase behavior and to suggest items customers might want to purchase based on past purchase history. Social media sites employ Hadoop to manage large amounts of data and recommend new connections based on a user's existing ones. Energy companies deploy Hadoop to store and process information gathered from sensors on the power grid. They then apply this information to plan capacity across the grid as conditions change over time.

Here are a few examples of how organizations capitalize on the ability of Hadoop to wrangle big data:

- **Search engine:** Analyzing search logs and performing data mining on its Web-page database
- **Auction site:** Optimizing searches and performing research
- **Professional networking site:** Creating models of potential candidates for job profiles, based on information in user profiles; discovering "people you may know" and "other fun facts"



Streamlining Hadoop integration

Cloudera's Distribution including Apache Hadoop (CDH) is a distribution of Hadoop that has been adopted in many commercial and noncommercial environments, including Fortune 500 companies, Web services companies, and government organizations. Ideal for enterprises seeking a stable, tested Hadoop distribution without proprietary vendor lock-in, CDH provides a bridge between the insights of organizations using Hadoop in production and the continuous stream of innovations from the Apache community.

Fully documented and ready to go, CDH consists of open source Apache Hadoop and a comprehensive set of related open source software components, which can be activated on an as-needed basis. Dell, working with Cloudera, thoroughly tested and certified Apache Hadoop to work with a wide range of operating systems, hardware, databases, data warehouses, business intelligence systems, and extract, transform, and load (ETL) systems. This broad compatibility helps organizations to integrate Hadoop with existing tools and resources.

of the cloud EDW, and its footprint in companies' core EDW architectures is likely to keep growing throughout this decade.¹

Supporting big data analytics through an open source framework

Organizations looking to leverage Hadoop for processing massive application data sets can turn to the Dell | Cloudera Solution, a comprehensive Hadoop-based stack. This solution combines a validated reference architecture based on Dell™ PowerEdge™ C Series servers, Dell-developed Crowbar software framework, Dell networking components, and Cloudera's Distribution including Apache Hadoop (CDH) with Cloudera® Enterprise management tools, training, technology support, and professional services. (For more information about

CDH, see the sidebar, "Streamlining Hadoop integration.")

The Dell | Cloudera Solution is tailored to the needs of a diverse range of organizations, including financial services; energy, utility, and telecommunication companies; research institutions; retail businesses; and Internet media groups. It is designed to reduce the complexity and cost of deploying, configuring, and managing Hadoop environments. The solution is based on a reference architecture that helps ensure consistency in rapid deployments with minimal differences in the network configuration.

Key building blocks of the Dell | Cloudera Solution include the following:

- **Cloudera Enterprise:** This subscription service delivers an integrated, highly

optimized solution for massive distributed processing against a single file system namespace. It includes CDH as well as expert support and tools that help organizations streamline the monitoring and management of the Hadoop ecosystem.

- **Dell PowerEdge C2100 rack server:** This server was designed for data analytics clusters, with an appropriate mix of memory, disk, and I/O capacity needed for MapReduce, Web analytics, database, and cloud computing environments. (For more information, see the sidebar, "Handling high data volumes.")
- **Dell PowerConnect™ 6248 Ethernet switch:** This 48-port Gigabit Ethernet Layer 3 switch delivers top-of-rack connectivity to Hadoop-related nodes. It provides significant rack density and is built on open standards, which allows organizations to use other brands and configurations of switches for the Hadoop environment.
- **Crowbar software:** A Dell-developed, open source tool, the Crowbar software framework helps manage a Hadoop deployment from the initial server boot to the configuration of the primary Hadoop components. Using Crowbar, administrators can complete a bare-metal deployment of multi-node Hadoop environments in a matter of hours,

Handling high data volumes

The Dell | Cloudera Solution is designed to take advantage of the Intel® processor-based Dell PowerEdge C2100 rack server. The PowerEdge C2100 is part of the hyperscale-inspired Dell PowerEdge C Series line of servers, which is designed to maximize compute density and energy efficiency.

The PowerEdge C2100 features two four- and six-core Intel Xeon® processor 5500 series or Intel Xeon processor 5600 series and up to 18 Double Data Rate 3 (DDR3) memory slots to maximize memory density. And with support for up to 26 TB of storage, the PowerEdge C2100 provides both the memory and disk capacity required to handle big data analytics.

¹"Hadoop: What Is It Good For? Absolutely ... Something," by James Kobielus, June 6, 2011, blog post, blogs.forrester.com/james_kobielus/11-06-06-hadoop_what_is_it_good_for_absolutely_something.

as opposed to days as typically required for manual installation. After completing the initial deployment, administrators can use Crowbar continuous integration and management features, including BIOS configuration, network discovery, status monitoring, performance data gathering, and alerting.

- **Services and support:** Dell and Cloudera work cooperatively to provide a full suite of services, including consulting, deployment assistance, and solution support for the Dell | Cloudera Solution that can be tailored to specific organizational needs.


Accelerating the deployment of big data analytics

The Dell | Cloudera Solution offers a high-performance Hadoop environment that helps organizations address their big data challenges. Cloudera Enterprise and Crowbar enable organizations to

manage the complete operational life cycle of Hadoop systems and streamline the deployment and management of Apache Hadoop services.

Dell PowerEdge C Series servers offer maximum density, performance, and serviceability for large cluster and cloud environments. They are designed to help lower total cost of ownership.

The Dell | Cloudera Solution has been tested and validated and is supported by Dell. Through Dell, organizations have access to comprehensive and collaborative service and support for the entire solution—from installation, configuration, and deployment to optimization, tuning, and integration with existing IT infrastructures.

Ultimately, the end-to-end Dell | Cloudera Solution facilitates the quick deployment of a predictable, production-level Hadoop environment—enabling organizations to accelerate time to value of big data to best competitive advantage. 

Authors

Joey Jablonski and **Aurelian "A.D." Dumitru** are part of the Dell Cloud and Big Data Solutions team. They lead the architecture and strategy efforts behind the Dell | Cloudera Solution.

Learn more

Dell | Cloudera Solution:
dell.com/hadoop