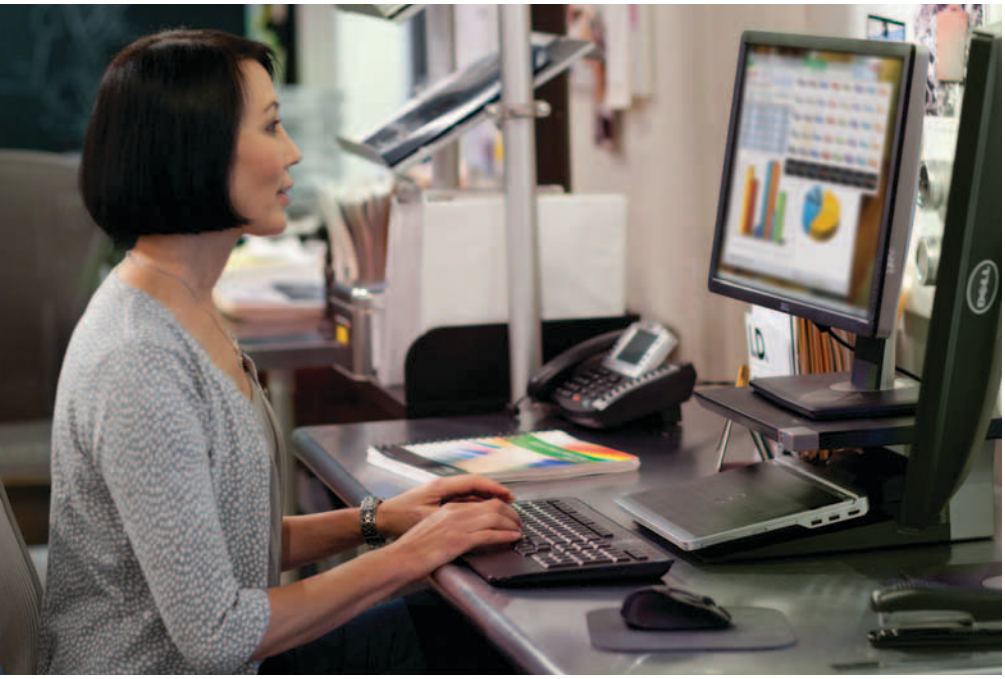
Meeting big data analytics challenges through a scalable framework

By Julie Lockner

By building a highly scalable open source infrastructure, enterprises can cost-effectively support big data analytics. One exceptional approach combines the Apache™ Hadoop™ framework and its MapReduce distribution running on Dell™ PowerEdge™ servers.



Organizations are taking big data to the next level. When asked in a recent Enterprise Strategy Group (ESG) research survey, most respondents stated that large data sets are the top reason why their organization is considering a MapReduce implementation in 2012, with experimentation a close second (see Figure 1).

The Apache Hadoop platform, a data processing and analytics framework for large data sets, combines a MapReduce programming approach with the Hadoop Distributed File System (HDFS). Using a cluster of server-class compute nodes, such as 12th-generation Dell PowerEdge servers, developers can build programs designed to process or analyze terabytes of any type of data—structured or unstructured—at once, potentially saving time and expense.

As a result, Hadoop helps organizations reduce the costs of completing large data processing and analytics tasks compared to traditional architectures that require powerful computers with database software. And because Hadoop is designed to process and analyze an entire data set at once, it enables organizations to glean value from the whole data set rather than from data samples, which may otherwise introduce a margin of error depending on the sampling method.

Because Hadoop MapReduce tasks scale linearly with the size of the cluster, application developers are taking a close look at it when faced with large data sets and the need to process the entire data set, rather than a sample. ESG research

indicates that the Hadoop architecture is a clear choice when the data volumes, the variety of data sources and types, and the speed in which an organization expects results from big data analytics exceed the current data analytics platform's capabilities. The reality is, besides those who recognize the need for a Hadoop-based platform, more than one-third of respondents to the ESG survey are currently looking at this emerging technology in an experimental sandbox environment.

At the same time, Hadoop is rapidly growing in popularity within the technology community because the platform is so versatile. When new data sets are introduced, Hadoop is designed to store and support analytics without further modification. With the Hadoop ecosystem of integrated development environments, organizations can replace custom scripts and replace or complement extract, transform, load (ETL) processes. Some are even looking at Hadoop to replace their current search mechanisms.

Many organizations are building Hadoop compute clusters based on a do-it-yourself mentality. Organizations investing in teams to learn Hadoop are often skilled at standing up distributed compute grids—a skill that is necessary to implement a Hadoop compute cluster. These same organizations tend to have an open-source culture as well. What will be interesting is to see how the do-it-yourself trend shifts now that additional options are available in the market, such as preconfigured, prebuilt Hadoop clusters and appliances. These offerings help speed up deployment as well as streamline the support model, making them attractive to enterprises looking for a simple, efficient way to handle large data volumes.

A major shortage of experienced developers introduces another challenge; finding someone who can program a MapReduce application to run on Hadoop may be tricky. As more applications are designed to make it easy to build programs that run on Hadoop, the need for hard-core Hadoop developers may be reduced. But in the meantime, organizations should start training developers now or find a consulting firm that specializes in developing Hadoop-based applications.

What is driving your organization's decision to consider a MapReduce framework to support its data analytics activities?

(Percent of respondents, N = 130, multiple responses accepted)*

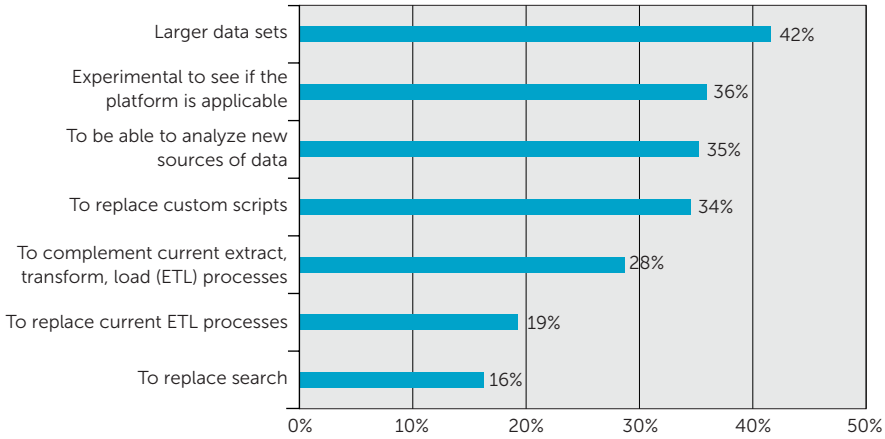


*Figure based on research that appears in "ESG research: The impact of big data on data analytics," September 2011, qrs.ly/hc1p6ze.

Figure 1. Large data sets are driving many organizations to implement a MapReduce framework

Integrated, high-performance options for big data analytics

In a recent interview, an administrator of a major Hadoop deployment identified the benefits of using Dell PowerEdge servers for the Hadoop cluster. The organization appreciated the ability to leverage the Dell do-it-yourself architecture—for example, color-coded components that indicate which items are hot swappable (orange) and which items should be handled by a certified administrator (purple). The administrator chose Dell servers and Hadoop because the organization's use case was to store terabytes of scientific data on a single file system that scaled—and that same platform could support their scientists' analytics needs with varying workload requirements. The net result: store-once, cost-effective storage for flexible data-analytics use cases. The administrator also liked the fact that a PowerEdge server could simply be pulled from its rack to easily plug in additional disk drives to increase capacity.

Given the release of 12th-generation Dell PowerEdge servers together with the company's recent announcement to sell Cloudera's Distribution including Apache Hadoop version 3 (CDH3), organizations have expanded options to service their Hadoop needs. As many organizations consider Hadoop to address their big data requirements, ESG anticipates a surge in demand for servers that not only are equipped to handle the processing load but also simplify support for do-it-yourself IT shops. PS

Author

Julie Lockner is senior analyst and vice president of Data Management Solutions at Enterprise Strategy Group.

Learn more

Enterprise Strategy Group:
enterprisestrategygroup.com

Dell | Cloudera Hadoop solution:
dell.com/hadoop