iStockphoto/Thinkstock

Accelerating big data analytics

By Mathew Eszenyi and Rahul Deshmukh

Vast data collections are a wellspring of information for organizations looking to bolster strategic planning efforts with big data analytics. Apache™ Hadoop® clusters powered by Dell and Intel technology help deliver fresh insights in near-real time.

By amassing and analyzing huge stores of unstructured data, organizations are discovering countless ways to grow their businesses and steer their destinies. Derived from a wide range of sources including social media, previously untapped stores of big data hold the promise to reveal unmatched insights, spark innovation and heighten competitive advantage while also accelerating responsiveness and efficiency goals.

However, gargantuan data stores with a preponderance of unstructured data, which may reach petabytes and beyond, become quite complex. Unstructured data is often text-heavy and unwieldy, and it does not fit well within traditional relational models. In addition, conventional tools often are proprietary and costly, requiring specialized expertise. Such limitations may put big data analytics out of reach for many organizations. And yet even for organizations that do have the capacity for big data analysis, the assembled information can fall significantly short of its full potential.

As a result, many organizations are deploying tools such as the open source Apache Hadoop framework to analyze large data volumes and transform the data into a manageable form that applications can handle with enhanced efficiency.

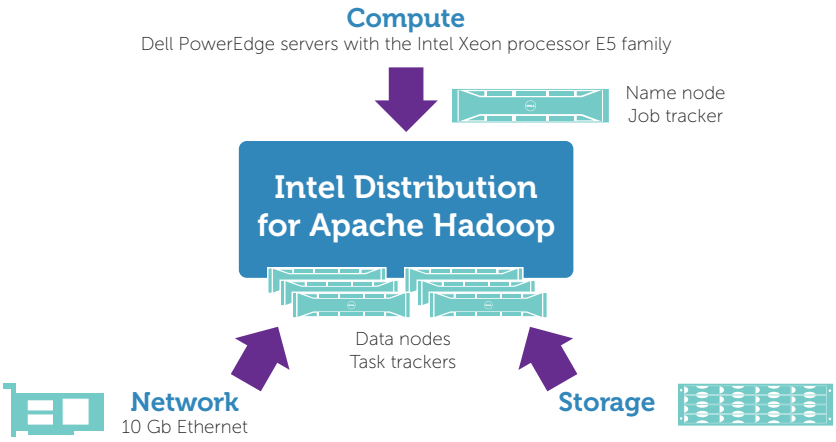
Hadoop is specifically designed to support very large volumes of unstructured data and is well suited to perform tasks typically associated with indexing, sorting, data mining, log analytics and image manipulation. It utilizes an open source framework designed to run on large clusters of cost-effective, industry-standard hardware.

Although the Apache Hadoop framework is powerful and flexible, large data sets often require hours to process. For example, a single sort of 1 TB of data on a Hadoop cluster based on legacy hardware can easily take up to four hours to complete. Moreover, import and export of large data sets into and out of a Hadoop framework may require a significant amount of time that could also undercut an organization's return on investment (ROI) in big data analytics deployments.

Because a substantial lag time in delivering meaningful results may diminish the immediacy and value that organizations expect to achieve through big data analytics, IT leaders should consider suitable technology alternatives to reach farther for those potential benefits. Just imagine how quickly scientists and researchers at the National Human Genome Research Institute (NHGRI) may assess, analyze and integrate results from test data run dozens of times instead of twice a day using conventional tools. The potential benefit — for example, advancing cures for diseases more rapidly than ever before — is huge.

Rising to technology challenges

To help organizations reduce their time requirements for big data analytics, Intel has designed and benchmarked a Hadoop cluster based on leading-edge Intel and Dell technology in an optimized configuration that performed a Hadoop sort of 1 TB of data in only seven minutes. This result represents a 97 percent reduction in search time compared to the four hours required by a configuration based on typical commodity-based components. The configuration also helped reduce data import and export times by up to six times compared to a typical configuration.[1]



Balanced server, storage and networking architecture to help maximize Hadoop cluster performance

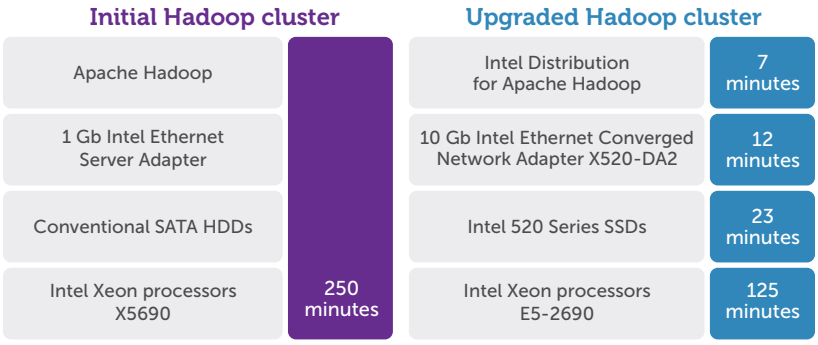
Incremental benchmarking of the configuration also revealed that extracting optimal performance out of a Hadoop cluster is not dependent on a single component. Instead, a balanced design matching server, storage and networking capabilities helps avoid bottlenecks and maximize performance throughout the infrastructure (*see figure*). As a result, a Hadoop cluster can be configured to capitalize on advanced Intel and Dell hardware and software that works optimally together, offering a near-real-time analytics capability that helps organizations greatly enhance ROI for big data analytics implementations.

Benchmarking cluster configuration performance

Intel engineers conducted benchmark testing in 2012 at Intel Labs using two Apache Hadoop cluster configurations: an initial Hadoop cluster based on a baseline configuration and an upgraded Hadoop cluster based on leading-edge server, processor, storage, networking and software technologies (*see figure on following page*). The Hadoop TeraSort benchmark test was used for both configurations. This test performs a sort operation on 1 TB of workload data.

Intel also tested the initial Hadoop configuration at intermediate stages of upgrade to assess the

[1] Based on benchmark testing performed in 2012 by Intel engineers at Intel Labs using baseline and upgraded Hadoop cluster configurations documented in "Big data technologies for near-real-time results," Intel white paper, Intel, December 2012, qrs.ly/f73gwpw.



Workloads may vary based on data set and configuration.

Workload speed enhancements with each intermediate component upgrade (right) to the initial Hadoop cluster configuration (left)

Adopting a balanced approach

Many organizations are gaining competitive traction by deriving actionable business insights from big data analytics. View this video for a high-level demonstration of how Hadoop works and why its scalability and reliability are particularly well suited for adding compute, storage and networking throughput to accommodate growing volumes of data.

qrs.ly/6b3gwpy

performance improvements enabled by individual component upgrades. After the initial benchmark testing, component types were upgraded in the following sequence: processors, storage, network adapters and software. The configuration was tested after each component upgrade to help quantify the incremental improvement achieved when each was integrated into the configuration. In addition to benchmarking the cluster performance for a sort operation, Intel also tested the time required to import and export data to and from both the initial and upgraded Hadoop clusters.

The initial Hadoop cluster configuration was composed of ten 1U servers with two 3.46 GHz Intel® Xeon® processors X5690 and 48 GB RAM running Cloudera's Distribution Including Apache Hadoop (CDH) with Oracle® Java® Development Kit (JDK) software. It also included five 700 GB, 7,200 rpm Serial ATA (SATA) hard disk drives (HDDs) for local storage and a single 1 Gb Intel Ethernet Server Adapter — 1000BASE-T — for server connectivity.

The upgraded Hadoop cluster configuration consisted of ten Dell™ PowerEdge™ R720 2U servers with two 2.9 GHz Intel Xeon processors E5-2690 and 128 GB RAM configured as worker nodes. Each worker node included five Intel 520 Series solid-state drives (SSDs), an Intel Ethernet Converged Network Adapter X520-DA2 and a 10 Gb small form-factor pluggable plus (SFP+) server adapter. One PowerEdge R710 1U server with two 2.9 GHz Intel Xeon processors X5690 and 48 GB RAM was configured as a head node. The head node also had one 10,000 rpm Serial Attached SCSI (SAS) drive; an Intel Ethernet Converged Network Adapter X520-DA2, 10 Gb; and an Intel Ethernet Server Adapter, 1 Gb. The configuration ran the Intel Distribution for Apache Hadoop 2, Apache Hadoop 1, Red Hat® Enterprise Linux® 6.3 and Oracle Java.

Minimized time for sorting

The initial configuration required 250 minutes, or approximately four hours, to complete the TeraSort 1 TB workload. To measure performance enhancement that could be gained by upgrading processing power, servers in the initial configuration were replaced with the PowerEdge R720 servers equipped with the Intel Xeon processor E5 family. These processors provide eight cores per socket and Intel® Data Direct I/O (Intel® DDIO) technology that helps boost the ability of Hadoop clusters to process data-intensive workloads. This processor upgrade helped reduce the time required to process the 1 TB workload from 250 minutes to 125 minutes — a 50 percent reduction in processing time.

Boosting processor performance can lead to bottlenecks in storage deployments and network I/O. As a result, Intel upgraded the local worker node storage from traditional HDDs to SSDs. Because this upgrade helps avoid the mechanical spinning disks and read/write heads in HDDs, SSDs are designed to significantly improve data access time and reduce latency. Upgrading to all-SSD storage is also designed to reduce the time required to perform the 1 TB workload processing even further, from 125 minutes to about 23 minutes, which represents an additional 80 percent reduction in data processing time.

Following the storage upgrade, network connectivity was upgraded from the 1 Gb Intel Ethernet Server Adapter to the 10 Gb Intel Ethernet Converged Network Adapter X520-DA2. The 10 Gb Intel Ethernet Converged Network Adapters are designed to substantially increase I/O throughput between nodes and help reduce data wait time. After upgrading to 10 Gigabit Ethernet (10GbE) connectivity, the 1 TB workload was completed in only 12 minutes, representing another 50 percent reduction in processing time.

The testing also measured performance enhancement that could be attained by

running Intel Distribution for Apache Hadoop as a replacement for the open source Hadoop distribution. The Intel Distribution for Apache Hadoop includes a range of enhancements designed to boost performance, enhance efficiency and streamline management of Hadoop clusters.[2] This upgrade to the testing configuration helped reduce the time required to complete the 1 TB workload processing to seven minutes, representing an additional 40 percent reduction in processing time.

By deploying these processor, storage, networking and software upgrades to the testing configuration, the Intel team observed benchmark results that demonstrated a significant reduction in the time required to process a 1 TB workload to sort data. The benchmark testing resulted in a reduction from four hours to seven minutes, representing an overall performance improvement of nearly 97 percent.

Optimized data import and export

Data import and export are critical operations when analyzing massive stores of big data. Data has to be imported into a Hadoop cluster at least once, and potentially many more times if the cluster is deployed as a service. Furthermore, results of Hadoop processing — such as sorting — need to be extracted from the cluster to be of any value for analysis. Import and export operations can place intense demands on both networking and storage I/O.

In addition to measuring the time required for baseline and upgraded Hadoop cluster configurations to perform sort operations, the benchmark testing also measured the time required to import data into and export results out of both configurations. Utilizing the same datastores to perform the TeraSort benchmark testing, Intel compared the time required to import data into the initial baseline configuration against the time required to import data into the upgraded configuration.

Upgrading the initial configuration from Gigabit Ethernet (GbE) networking to 10GbE networking helped improve import speed by four times. Additionally, upgrading local storage from traditional HDDs to SSDs helped increase import speed relative to the initial configuration by six times. Similarly, the time required to export the TeraSort benchmark results from both the initial and upgraded configurations was measured. The same connectivity and storage drive updates resulted in a sixfold reduction in time required to export data.

Balanced configuration

As the benchmark testing showed, performance in a Hadoop cluster can be significantly optimized with a balanced approach to configuration using innovative processors, storage, networking and software. Organizations considering a Hadoop cluster deployment should explore balanced configurations to suit their big data analytics needs within prevailing budget constraints.

For example, although all-SSD configurations are designed for highly optimized performance, an alternative deployment of tiered storage using both SSDs and traditional HDDs may offer a suitable way to help contain costs. Organizations should also keep a balanced approach in mind for components. The benefits of a high-end processor, for example, can be challenging to achieve if storage and/or networking bottlenecks impede data throughput. In addition, all three data processing stages — sort, import and export — should be taken into consideration when configuring a Hadoop cluster for enhanced data collection and analytics performance.

Boosting responsiveness and efficiency

Big data analytics open up tremendous possibilities for extracting actionable business insights quickly and cost-effectively. Yet these opportunities can be diminished when organizations take a substantial time-efficiency hit while processing vast data collections that must be sorted, imported and exported during analysis.

Adopting a balanced approach to component selection and configuration enables organizations to leverage an Intel Distribution for Apache Hadoop cluster with advanced servers, storage, networking and software in a way that is designed to dramatically improve their ROI for big data analytics implementations. Recent benchmark tests of the Intel Distribution for Apache Hadoop on a cluster configured with advanced Intel and Dell hardware working optimally together in a balanced architecture demonstrated substantial performance enhancements when sorting, importing and exporting data for analytic workloads. PS

Authors

Mathew Eszenyi is a product marketing manager, specifically for Intel Ethernet 10GbE and 40GbE products, and a big data technology strategist at Intel.

Rahul Deshmukh is a senior technical marketing manager at Dell specializing in server peripheral devices.

Learn more

Intel and Dell Ethernet:
intelethernet-dell.com

Intel big data intelligence:
intel.com/bigdata

Dell governance and retention:
Dell.com/bigdata

[2] For more information on the Intel Distribution for Apache Hadoop, see "Optimizing performance for big data analysis," by Armando Acosta and Maggie Smith, in Dell Power Solutions, 2013 Issue 3, qrs.ly/vx3gwq0.