WHITE PAPER

Boosting Retail Revenue and Efficiency with Big Data Analytics

A Simplified, Automated Approach to Big Data Applications: StackIQ Enterprise Data Management and Monitoring

Abstract

Like many industries today, the world of retail has accumulated an array of computer applications over time from the warehouse to the point of sale and beyond to improve efficiency and competitiveness. In recent years, the number of data inputs has grown to include electronic social media, Web site clicks, mobile applications, and video systems that record and correlate individual and mass consumer behavior in the store. This increased volume of data available to businesses from both internal systems and external channels has led to a new category of application known as Big Data. For retailers, the benefits of using analytical applications that tap into the Big Data stream are considerable. These applications can provide information to significantly enhance sales, marketing, and customer relationships; introduce operational and merchandising efficiencies that reduce costs; and lead to strategies to better understand what customers want, even before they realize it.

Deploying Big Data applications and the Big Infrastructure to support them, however, entails a high level of complexity. These tasks have relied on a variety of legacy and newer tools to handle deployment, scalability, and management. Most of the tools require management of changes to configurations and other fixes through the writing of software scripts. Making changes to one appliance or several appliances entails a manual, time-intensive process that leaves the administrator unsure as to whether or not the changes have been implemented throughout the application cluster. Using homogenous, hardware/software appliances for Big Data applications from EMC, Teradata, Oracle, and others is another alternative, but it is very expensive and more limited.

Contents

Now, however, a paradigm shift in the design, deployment, and management of Big Data applications is underway, providing a faster, easier solution for retailers and other organizations interested in reaping the benefits of Big Data. For the first time in the IT industry, best-of-breed Hadoop implementation tools have been integrated with Hadoop and cluster management solutions in StackIQ Enterprise Data.

This paper summarizes the benefits of Big Data applications for the retail industry. It also presents the challenges of deploying and managing robust Hadoop clusters. Finally, the paper also includes information on the cost, efficiency, reliability, and agility features that make StackIQ Enterprise Data competitively unique and presents a reference architecture for Hadoop deployments using the product.

The Promise of Big Data Applications for the Retail Industry

Perhaps no other industry than retail reacts faster to changing and evolving trends that may impact customer demand for specific products, pricing and offers, customer service, and other facets of the customer experience of a retail brand. In the past, retailers analyzed transaction data and any other customer information that was available. Today, the data available to retailers from retail Web sites, in-store video and analytics systems, customer demographics, mobile location data from tablet computers and smart phones, social media, weather, and embedded sensors—to name a few sources—has increased dramatically. Today, the amount and variety of data available to retailers provides a wealth of new opportunities to increase revenue, control costs, and counter competitive threats.

Big Data applications, characterized by very large and complex data sets, have the potential to reveal new insights and patterns that can be used to better understand and capitalize on consumer behavior. A 2011 study by McKinsey & Company estimated that, “a retailer using Big Data to the full has the potential to increase its operating margin by more than 60 percent.”

According to a 2012 study by Gartner, by paying attention to this opportunity the retail industry now has one of the highest adoption rates of Big Data applications. Walmart, the world’s largest retailer, is using Big Data analytical applications to optimize its inventory based on the specific tastes and seasonal preferences of customers in each geographic area the company serves. Walmart is also using Big Data to develop an in-store, mobile navigation system that will alert customers to sales based on their preferences and location in the store. Another application uses data from a customer’s Facebook profile to suggest products of potential interest.

In Europe, a September 2012 study by MasterCard and the Economist Intelligence Unit found that 41% of the 330 retailers surveyed say they

will use data to deliver an improved customer experience in 2013. Of these retailers, 30% say they are increasing customer segmentation efforts and 39% say they are delivering personalized customer experiences across multiple communications channels using Big Data applications.

Applying analytical tools to the huge new volumes of data in retail requires a distinctly different infrastructure from traditional database architectures and query products. Big Data applications were once the exclusive domain of scientific research, national security, and oil and gas exploration. Today's Big Data applications can be run on hundreds or thousands of clustered computer servers instead of supercomputers. Some of the largest retailers in the world are already reaping the benefits from Big Data applications, but a large percentage of retail organizations with stores and/or Web sites have yet to deploy these applications. The reasons may include the perceived complexity of Big Infrastructure to support these applications and the lack of a clear understanding of use cases that may prove most beneficial.

How are Retailers Leveraging their Big Data Sets Today?

Williams Sonoma is mining Big Data repositories of customer purchase data, clicks, click-throughs, demographics, and Web browsing history and creating predictive models for each customer and product. Based on the models, targeted email offers are sent out. For example, if an item goes on sale an email goes to anyone who has previously purchased that item or browsed for it online. During Black Friday 2011, members of the MIT Media Lab captured and correlated location data from their smart phones to estimate the number of people in Macy's parking lots. They used the data to estimate the store's sales that day with great accuracy even before the retailer had crunched the sales numbers. Online fashion, shoe, and jewelry retailer BeachMint captures all the online activity of its customers and leverages analysis of the data for highly effective cross-selling and referrals .

Retail Use Cases: Big Data Analytics

Predicting Customer Purchasing Retailers can analyze hundreds, thousands, or millions of transactions and correlate purchase patterns with demographic profiles to look for trends that might lead to opportunities for increased revenue through special offers, different approaches to merchandising, and product bundling.

Customer Micro-segmentation Retailers can find opportunities by analyzing each customer's purchase history and online shopping data. Target Stores recently analyzed historical purchase data and identified 25 shopping items that can be used to assign a "pregnancy prediction" score to customers. The company was able to accurately predict that women purchasing some of those items were pregnant and it mailed coupons for special offers on baby products. The analysis also helped the company predict expected delivery time with a high degree of accuracy, as reported in the *New York Times*.

Cross selling and upselling Collecting and gathering data across multiple channels, including Web site click stream data, social media activities, account information, and other sources can help retailers suggest additional products to customers that match their needs and budgets. This type of application can also look at customer habits to assess risks and suggest alteration of habits to reduce risks.

Location-based Marketing Using mobile devices and geographic location data, retailers can target customer needs correlated with their location in stores. For example, while a customer is in a particular department, a mobile application tracks her location. A Big Data application correlates her location with her calendar and notices that her husband's birthday is coming up and offers a special promotion for a product in that department. Due to uncertainty in her body language, detected by a video analytics system, the Big Data application alerts a sales person, who offers assistance, increasing the potential for a sale.

Supply Chain and Logistics Optimization Retailers can optimize their supply chain for each product, channel, and location based on real-time events. Sensors in delivery trucks and radio frequency ID (RFID) tags on products can help retailers know the exact location of products in transit. This helps in planning for the optimal use of warehouse space and delivery methods.

Retail Fraud Through monitoring, modeling, and analyzing high volumes of data from transactions and extracting features and patterns, retailers can prevent credit card account fraud.

Pricing Optimization Smart shelves connected to Big Data applications will be able to automatically update prices based on dynamic market conditions.

Other Big Data analytics applications in retail include closed-loop analyses to gauge the effectiveness of marketing campaigns, online buying behavior, sales promotions, and inventory optimization efforts; sentiment analysis based on structured and unstructured data from multiple channels; and real-time price comparisons to support guaranteed lowest price campaigns aimed at countering competitor pricing. These applications can enhance marketing, merchandising, operations, customer service, and supply chain, and help generate new business models.

Deploying and Managing Big Data Analytical Applications

In addition to the server clusters provisioned in company data centers or leased from virtual cloud environments, the software to deploy and manage Big Data application environments is a crucial element. The complexity of deploying Big Infrastructure clusters has been somewhat lessened by a new generation of open-source software frameworks. Apache Hadoop is the leading solution and it has gained great popularity due to its maturity, its ease of scaling, its affordability as a non-proprietary data platform, its ability to handle both structured and unstructured data, and its many connector products.

The growing popularity of Hadoop has also shined the spotlight on its shortcomings—specifically the complexity of deploying and managing Hadoop infrastructure. Early adopters of Hadoop have found that they lack the tools, processes, and procedures to deploy and manage it efficiently. IT organizations are facing challenges in coordinating the rich clustered infrastructure necessary for enterprise-grade Hadoop deployments.

The current generation of Hadoop products was designed for IT environments in which different groups of skilled personnel are required to deploy them. One group installs and configures the cluster hardware, networking components, software, and middleware that form the foundation of a Hadoop cluster. Another group of IT professionals is responsible for deploying the Hadoop software as part of the cluster infrastructure. Until now, cluster management products have been mainly focused on the upper layers of the cluster (e.g., Hadoop products, including the Hadoop Distributed File System [HDFS], MapReduce, Pig, Hive, HBase, and Zookeeper). The installation and maintenance of the underlying server cluster is handled by other solutions. Thus, the overall Hadoop infrastructure is deployed and managed by a collection of disparate products, policies,

and procedures, which can lead to unpredictable and unreliable clusters.

StackIQ combines the leading Apache Hadoop software stack with the leading cluster management solution. In doing so it has engineered a revolutionary new solution that makes Hadoop deployments of all sizes much faster, less costly, more reliable, and more flexible. StackIQ Enterprise Data optimizes and automates the deployment and management of underlying cluster infrastructures of any size for any Hadoop distribution or Big Data application.

With StackIQ Enterprise Data, physical or virtual Hadoop clusters can be quickly provisioned, deployed, monitored, and managed. System administrators can manage the entire system using a single pane of glass. New nodes are also configured automatically from bare metal—with a single command—without the need for complex administrator assistance. If a node needs an update, it will be completely re-provisioned by the system to ensure that it boots into a known good state. Since StackIQ Enterprise Data places every bit on every node, administrators have complete control and consistency across the entire infrastructure. Now administrators have the integrated, holistic Hadoop tools and the control they need to more easily and swiftly meet their enterprise Big Data application requirements.

StackIQ Enterprise Data

StackIQ Enterprise Data is a complete, integrated Hadoop solution for enterprise customers. For the first time, enterprises get everything they need to deploy and manage Hadoop clusters throughout the entire operational lifecycle in one product. StackIQ Enterprise Data includes:

Apache Hadoop is an open-source, massively scalable, highly stable and extensible platform based on the most popular and essential Hadoop projects for storing, processing, and analyzing large volumes of structured and unstructured data. StackIQ Enterprise Data can support all major Apache Hadoop Distributions from vendors including Hortonworks, Cloudera, and MapR.

StackIQ offers a *Hortonworks Edition* of StackIQ Enterprise Data that comes bundled with *Hortonworks Data Platform* making it easier than ever to integrate Apache Hadoop into existing data architectures. The platform includes HDFS, MapReduce, Pig, Hive, HBase, and Zookeeper, along with open source technologies that make the Hadoop platform more manageable, open, and extensible. These include HCatalog, a metadata management service for simplifying data sharing between Hadoop and other enterprise information systems, and a complete set of open APIs such as WebHDFS to make it easier for ISVs to integrate and extend Hadoop.

StackIQ Hadoop Manager manages the day-to-day operation of the Hadoop software running in the clusters, including configuring, launching, and monitoring HDFS, MapReduce, ZooKeeper, Hbase, and Hive. A unified single pane of glass—with a command line interface (CLI) or graphical user interface (GUI)—is used to control and monitor all of these elements, as well as manage the infrastructure components in the cluster.

Easy to use, the StackIQ Hadoop Manager allows for the deployment of Hadoop clusters of all shapes and sizes (including heterogeneous

hardware support, parallel disk formatting, and multi-distribution support). Typically, the installation and management of a Hadoop cluster has required a long, manual process. The end user or deployment team has had to install and configure each component of the software stack by hand, causing long setup times. Ongoing management of these systems is time-intensive and problematic, resulting in security and reliability implications. StackIQ Enterprise Data completely automates the process.

StackIQ Cluster Manager manages all of the software that sits between bare metal and a cluster application, such as Hadoop. A dynamic database contains all of the configuration parameters for an entire cluster. This database is used to drive machine configuration, software deployment (using a unique Avalanche peer-to-peer installer), management, and monitoring.

The Cluster Manager includes the following features:

- Provisions and manages the operating system from bare metal, capturing networking information (such as MAC addresses).
- Configures host-based network settings throughout the cluster.
- Captures hardware resource information (such as CPU and memory information) and uses this information to set cluster application parameters.
- Captures disk information and uses this information to programmatically partition disks across the cluster.
- Installs and configures a cluster monitoring system.
- Provides a unified interface (CLI and GUI) to control and monitor everything.

The StackIQ Cluster Manager for Hadoop is based on StackIQ's open source Linux cluster provisioning and management solution, Rocks®, originally developed in 2000 by researchers at the San Diego Supercomputer Center at the University of California, San Diego. Rocks was initially designed to enable end users to easily, quickly,

and cost-effectively build, manage, and scale application clusters for High Performance Computing (HPC). Thousands of environments around the world now use Rocks.

In StackIQ Enterprise Data, the Cluster Manager's capabilities have been expanded to not only handle the underlying infrastructure, but to also handle the day-to-day operation of the Hadoop software running in the cluster. Other competing products fail to integrate the management of the hardware cluster with the Hadoop software stack. By contrast, StackIQ Enterprise Data operates from a continually updated, dynamic database populated with site-specific information on both the underlying cluster infrastructure and the running Hadoop services. The product includes everything from the operating system on up, and it packages CentOS Linux or Red Hat Enterprise Linux, cluster management middleware, libraries, compilers, and monitoring tools.

Key Benefits of StackIQ Enterprise Data

- Complete, integrated, Hadoop solution for the enterprise
- Faster time to deployment
- Automated, consistent, dependable deployment and management
- Simplified operation that can be quickly learned without systems administration experience
- Reduced downtime due to configuration errors
- Reduced total cost of ownership for Hadoop clusters
- Works with all major Hadoop Distributions

Enterprise Hadoop Use Cases

Hadoop enables organizations to move large volumes of complex and relational data into a single repository where raw data is always available. With its low-cost, commodity servers and storage repositories, Hadoop enables this data to be affordably stored and retrieved for a wide variety of analytic applications that can help organizations increase revenues by extracting value such as strategic insights, solutions to challenges, and ideas for new products and services. By breaking up Big Data into multiple parts, Hadoop allows for the simultaneous processing and analysis of each part on servers throughout the cluster, greatly increasing the efficiency and speed of queries. The use cases for Hadoop are many and varied, impacting disciplines as varied as public health, stock and commodities trading, sales and marketing, product development, and scientific research. For the business enterprise, Hadoop use cases include:

Data Processing Hadoop allows IT departments to extract, transform, and load (ETL) data from source systems and to transfer data stored in Hadoop to and from a database management system for the performance of advanced analytics; it is also used for the batch processing of large quantities of unstructured and semi-structured data.

Network Management Hadoop can be used to capture, analyze, and display data collected from servers, storage devices, and other IT hardware to allow administrators to monitor network activity and diagnose bottlenecks and other issues.

Retail Fraud Through monitoring, modeling, and analyzing high volumes of data from transactions and extracting features and patterns, retailers can prevent credit card account fraud.

Recommendation Engine Web 2.0 companies can use Hadoop to match and recommend users to one another or to products and services based on analysis of user profile and behavioral data.

Opinion Mining Used in conjunction with Hadoop, advanced text analytics tools analyze the unstructured text of social media and social networking posts, including Tweets and Facebook posts, to determine the user sentiment related to particular companies, brands, or products; the focus of this analysis can range from the macro-level down to the individual user.

Financial Risk Modeling Financial firms, banks, and other companies use Hadoop and data warehouses for the analysis of large volumes of transactional data in order to determine risk and exposure of financial assets, prepare for potential "what-if"

scenarios based on simulated market behavior, and score potential clients for risk.

Marketing Campaign Analysis Marketing departments across industries have long used technology to monitor and determine the effectiveness of marketing campaigns; Big Data allows marketing teams to incorporate higher volumes of increasingly granular data, like click-stream data and call detail records, to increase the accuracy of analysis.

Customer Influencer Analysis Social networking data can be mined to determine which customers have the most influence over others within social networks; this helps enterprises determine which customers are most important and influential.

Analyzing Customer Experience Hadoop can be used to integrate data from previously siloed customer interaction channels (e.g., online chat, blogs, call centers) to gain a complete view of the customer experience; this enables enterprises to understand the impact that one customer interaction channel has on another in order to optimize the entire customer lifecycle experience.

Research and Development Enterprises like pharmaceutical manufacturers use Hadoop to comb through enormous volumes of text-based research and other historical data to assist in the development of new products.

StackIQ Enterprise Data Reference Architecture

Table 1 shows the StackIQ Enterprise Data reference architecture hardware using Dell PowerEdge servers.

Using 3 TB drives in 18 data nodes in a single rack, this configuration represents 648 TB of raw storage. Using the standard HDFS replication factor of 3 yields 216 TB of usable storage.

Table 1. StackIQ Enterprise Data Reference Architecture (Hardware)

Reference Hardware Configuration on Dell™ PowerEdge Servers				
Machine Function	Management Node	Name Node	Secondary Name Node	Data Node
Platform	PowerEdge R410	PowerEdge R720xd		
CPU	2 x E5620 (4 core)	2 x E5-2640 (6-core		
RAM	16 GB	96 GB		
Network	1 x Dell PowerConnect 5524 Switch, 24-ports 1 Gb Ethernet (per rack)			
	1 x Dell PowerConnect 8024F 10Gb Ethernet switch (For rack interconnection in multi-rack configurations)			
Disk	2 x 1 TB SATA 3.5″	12 x 3TB SATA 3.5″		
Storage Controller	PERC H710			
RAID	RAID 1	NONE		
Minimum per Pod	1	1	1	3*

* Based on HDFS's standard replication factor of 3

Table 2 shows the software components of the StackIQ Enterprise Data (Hortonworks Edition) reference architecture.

Table 2. StackIQ Enterprise Data Reference Architecture (Software)

Reference Architecture (Software)	
StackIQ Enterprise Data 2.0 ISO Contents	
Hadoop Roll	Hortonworks Data Platform 1.0.3 (Included in SED Hortonworks Edition)
Base Roll	Base Cluster Management and Command Line Interface (CLI)
Kernel Roll	Installation Support for Latest x86 chipsets
Core Roll	StackIQ Diagnostics and GUI
OS Roll	CentOS 6.3
Ganglia Roll	Cluster Monitoring
Web Server Roll	Apache Web Server and WordPress

The management node is installed with StackIQ Enterprise Data management software, which automatically installs and configures Hortonworks Data Platform software on the Name Node, Secondary Name Node, and all Data Nodes.

Rolls are pre-packaged software modules that integrate software components for site-specific requirements. They may be selected and automatically configured in StackIQ Enterprise Data and are available from StackIQ at http://www.stackiq.com/download/.

StackIQ Enterprise Data is designed to manage a number of Big Data software products, including Apache Hadoop, and MapR. The Hortonworks Edition of StackIQ Enterprise Data includes the Hortonworks Data Platform software (HDP).

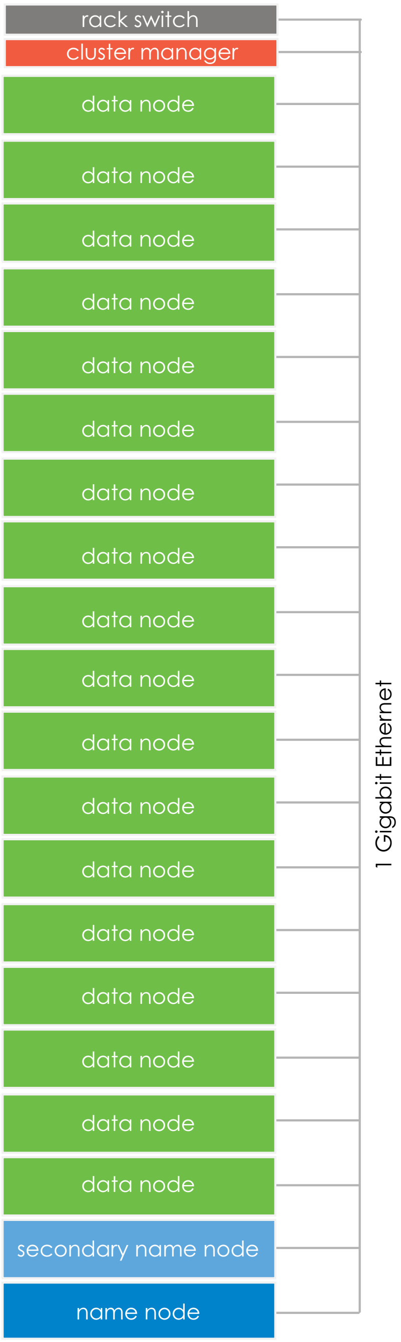

Figure 2. Single Rack Configuration

Single Rack Configuration

In the single rack configuration, there is one Cluster Manager Node, one Name Node, and one Secondary Name Node. This configuration may include between one and 18 Data Nodes, depending upon how much storage is needed for the cluster. The top-of-rack switch connects all of the nodes using Gigabit Ethernet. A sample single-rack configuration of StackIQ Enterprise Data is shown in Figure 2.

Multi-Rack Configuration

More racks may be added to build a multi-rack configuration. Each rack may contain between one and 20 Data Nodes, depending upon how much storage is needed for the cluster. A multiport 10 GE switch should be added to the second rack, with all of the top-of-rack switches connected to it via one of the 10 GE ports. For simplicity, a step and repeat layout is shown in the multi-rack sample configuration in Figure 3.

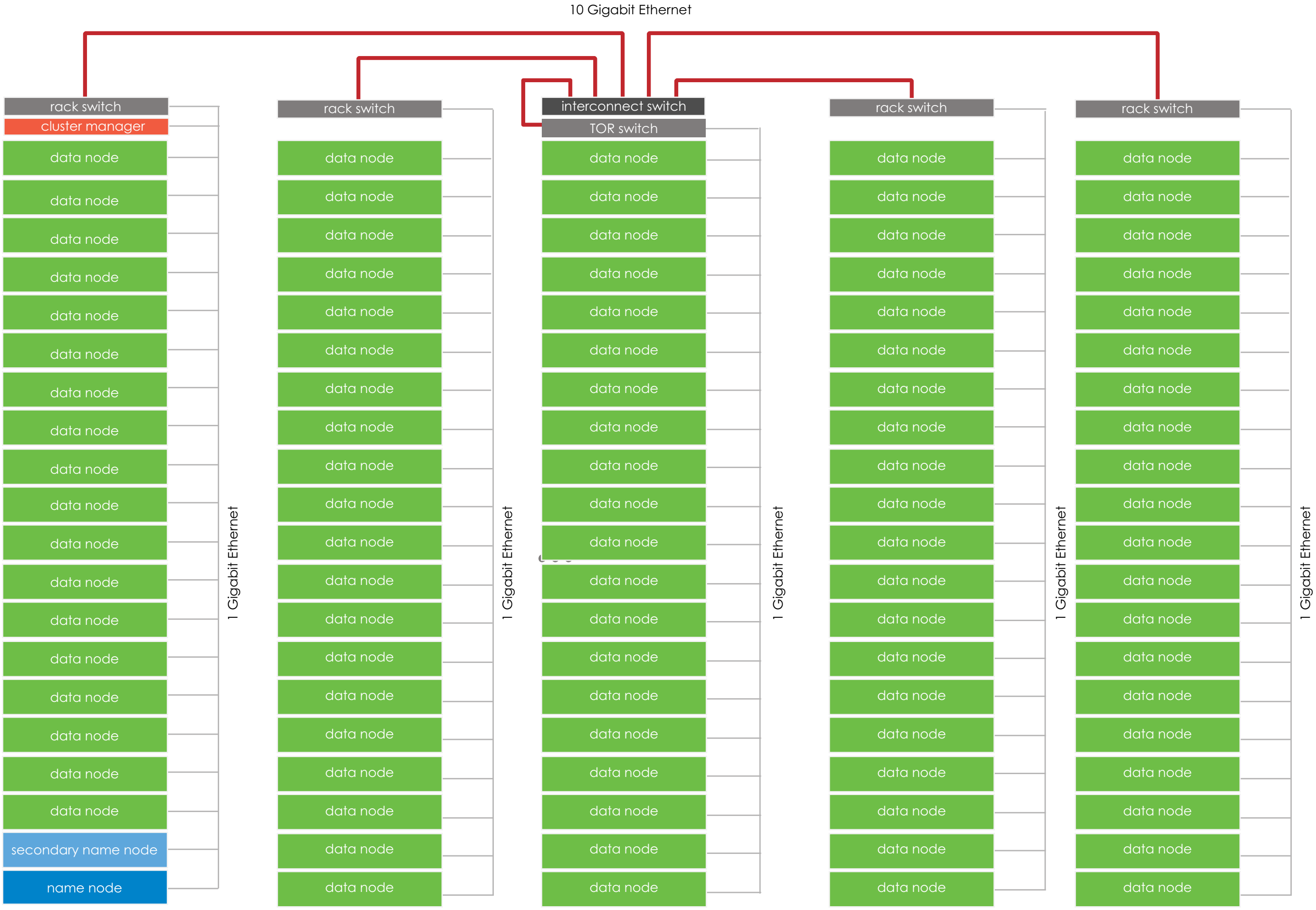

Figure 3. Multi-Rack Configuration

Summary

As the leading software framework for massive, data-intensive, distributed applications, Apache Hadoop has gained tremendous popularity, but the complexity of deploying and managing Hadoop server clusters has become apparent. Early adopters of Hadoop, moving from proofs-of-concept in labs to full-scale deployment, are finding that they lack the tools, processes, and procedures they need to deploy and manage these systems efficiently. For reliable, predictable, simplified, automated Hadoop enterprise deployments, StackIQ has created StackIQ Enterprise Data. This powerful, holistic, simplified tool for Hadoop deployment and management combines the leading Apache Hadoop software stack with the leading cluster management solution. StackIQ Enterprise Data makes it easy to deploy and manage consistent Hadoop installations of all sizes, and its automation, powerful features, and ease of use lower the total cost of ownership of Big Data systems.

For More Information

StackIQ White Paper: "Optimizing Data Centers for Big Infrastructure Applications"
bit.ly/N4haaL

Intel® Cloud Buyers Guide to Cloud Design and Deployment on Intel® Platforms
bit.ly/L3xXWI

StackIQ Product Information
www.StackIQ.com/products

About StackIQ

StackIQ is a leading provider of Big Infrastructure management software for clusters and clouds. Based on open-source Rocks® cluster software, StackIQ's Rocks+ product simplifies the deployment and management of highly scalable systems. StackIQ is based in La Jolla, California, adjacent to the University of California, San Diego, where the open-source Rocks Group was founded. Rocks+ includes software developed by the Rocks Cluster Group at the San Diego Supercomputer Center at the University of California, San Diego, and its contributors. Rocks® is a registered trademark of the Regents of the University of California.

StackIQ

4225 Executive Square
Suite 1000
La Jolla, CA 92037
858.380.2020
info@stackIQ.com