Deploying and Managing Dell Big Data Clusters with StackIQ Cluster Manager

A Dell | StackIQ Technical White Paper

Dave Jaffe, PhD
Solution Architect
Dell Solution Centers

Greg Bruno, PhD
VP Engineering
StackIQ, Inc.

Tim McIntire
Co-Founder
StackIQ, Inc.
Executive Summary

The Dell | Intel Cloud Acceleration Program, a team within the Dell Solution Centers, provides customer engagements on the topics of Cloud and Big Data including proof-of-concept (POC) testing in the Dell Solution Center lab. In a recent POC for a large provider of financial data, Dell partnered with StackIQ to configure a 60-datanode cluster to compare two different Big Data technologies, Cassandra and HBase. Using StackIQ’s Cluster Manager software the customer was able to rapidly provision and re-provision the servers to run the two different applications with various configurations. The use of Cluster Manager enabled the customer to complete more and higher quality tests than originally expected.
Table of contents

1 Introduction ........................................................................................................................................ 4
2 Dell’s Big Data Cluster ....................................................................................................................... 5
3 StackIQ Cluster Manager .................................................................................................................. 8
4 Installing and Managing the Mixed-Use Cluster .............................................................................. 9
   “Step Zero” .................................................................................................................................. 9
Meeting the Customer Specifications ................................................................................................. 9
Automated Scaling ............................................................................................................................. 11
Dynamic Repurposing ......................................................................................................................... 11
Advanced Diagnostics and Repair ....................................................................................................... 12
Success ............................................................................................................................................. 12
To Learn More .................................................................................................................................. 12

Tables

Table 1. PowerEdge R720 Infrastructure Node Configuration ............................................................. 7
Table 2. PowerEdge R720XD Datanode Configuration ....................................................................... 7
Table 3. Cassandra/HBase Tests ........................................................................................................ 11

Figures

Figure 1. Dell 60-Datanode Big Data Cluster ..................................................................................... 5
Figure 2. Cluster networking diagram ............................................................................................... 6
Figure 3. StackIQ Cluster Manager Monitor Tab .............................................................................. 10
Figure 4. StackIQ Cluster Manager Attributes Tab .......................................................................... 10
1 Introduction

The Dell | Intel Cloud Acceleration Program (DICAP), a team within the Dell Solution Centers, has the mission of providing customer engagements on the topics of Cloud and Big Data. These engagements can be short briefings, longer architectural design sessions, or one to two-week proof of concept tests in which the customer may bring in their own data or workload to test the Dell solution.

For Big Data proof of concepts, the DICAP team maintains a laboratory on its Round Rock, Texas, campus with several hundred Dell™ PowerEdge™ R-series and C-series servers that function as management and data nodes for Big Data clusters. A typical Big Data proof of concept involves using Dell tools to stand up a Hadoop cluster with 20, 40 or even more data nodes.

Since the Dell customer set has many varied needs in the Big Data space it is often necessary to stand up a cluster with a specialized application, or even to build a cluster that may be repurposed quickly from one application to another. This was the case recently when a large financial data provider was interested in the performance scaling of two competing NoSQL databases: Apache Cassandra and Apache Hadoop’s HBase. The provider wanted to run tests on clusters of 10, 20 and 40 data nodes with one application and then quickly switch and run the second application on the same hardware.

For cases like this Dell has a strong partnership with StackIQ, whose Cluster Manager product provided the bare metal installation of the cluster nodes, followed by the installation and management of one application, and then the rapid changeover to the other application. Together the team of Dell, StackIQ, DataStax (provider of the Apache Cassandra distribution) and Cloudera (Apache Hadoop provider) was able to provide a large, flexible, test cluster to meet the customer’s needs.

The Dell Big Data cluster hardware used in the test is described in the next section, followed by details of StackIQ’s Cluster Manager. Finally, the use of Cluster Manager to successfully implement the proof of concept is described in Section 4.
2 Dell’s Big Data Cluster

The cluster used for the Cassandra vs. HBase testing is shown below. Three 52-rack unit tall racks of Dell PowerEdge servers and Dell Force 10 network switches were utilized in the tests. The cluster was managed by three PowerEdge R720 servers with one running the StackIQ Cluster Manager software and two serving as name nodes for the HBase tests (see Table 1 for configuration details).

Figure 1. Dell 60-Datanode Big Data Cluster
Sixty PowerEdge R720XD servers with 24 500 GB disks each (see Table 2 for configuration) were employed as data nodes for HBase and Cassandra as well as load drivers for the customer’s test application.

All servers were connected via 1 Gb Ethernet links as well as 10 Gb Ethernet links using a set of Dell Force 10 S60 1 GbE switches and S4810 10 GbE switches. As shown below, the 1 GbE network provides server management as well as iDRAC (Dell Remote Access Controller) connectivity while the 10 GbE network serves as the data network. The two Force10 S60 switches within each rack are connected via stacking cables and uplinked to the 10 GbE infrastructure running on six stacked Force10 S4810 switches, two per rack. The S4810 switches are connected using two 40GbE ports per switch.

Figure 2. Cluster networking diagram
### Table 1. PowerEdge R720 Infrastructure Node Configuration

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>2 Rack Units (3.5&quot;)</td>
</tr>
<tr>
<td>Processor</td>
<td>2x Intel® Xeon® E5-2650 2 GHz 8-core procs</td>
</tr>
<tr>
<td>Memory</td>
<td>128 GB 1600 MHz</td>
</tr>
<tr>
<td>Disk</td>
<td>6x 600 GB 15K SAS Drives</td>
</tr>
<tr>
<td>Network</td>
<td>4x 1GbE LOMs, 2x PCIe 10GbE NICs</td>
</tr>
<tr>
<td>RAID Controller</td>
<td>PowerEdge RAID Controller H710 (PERC)</td>
</tr>
<tr>
<td>Management Card</td>
<td>Integrated Dell Remote Access Controller</td>
</tr>
<tr>
<td></td>
<td>Enterprise Edition (iDRAC)</td>
</tr>
</tbody>
</table>

### Table 2. PowerEdge R720XD Datanode Configuration

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>2 Rack Units (3.5&quot;)</td>
</tr>
<tr>
<td>Processor</td>
<td>2x Intel® Xeon® E5-2667 2.9 GHz 6-core procs</td>
</tr>
<tr>
<td>Memory</td>
<td>64 GB 1600 MHz</td>
</tr>
<tr>
<td>Disk</td>
<td>24x 500GB 7200 RPM Nearline SAS drives</td>
</tr>
<tr>
<td>Network</td>
<td>4x 1GbE LOMs, 2x 10GbE PCIe NICs</td>
</tr>
<tr>
<td>RAID Controller</td>
<td>PowerEdge RAID Controller H710 (PERC)</td>
</tr>
<tr>
<td>Management Card</td>
<td>Integrated Dell Remote Access Controller</td>
</tr>
<tr>
<td></td>
<td>Enterprise Edition (iDRAC)</td>
</tr>
</tbody>
</table>
3 StackIQ Cluster Manager

StackIQ takes a “software defined infrastructure” approach to provisioning and managing clusters. The product used herein, StackIQ Cluster Manager, was built around the Rocks project, which began in 2000 at the San Diego Supercomputer Center (a research division of the University of California) with the single purposed goal of “making clusters easy”. While Rocks initially focused on High Performance Computing, the underlying structure of the software allows it to easily define, deploy, and manage all types of Linux-based servers. The unique experience this project has had with clustered systems makes the Rocks Cluster Distribution even more interesting today for groups looking to deploy big data and cloud infrastructure. This paper discusses some of the enhancements StackIQ has made to the software for enterprise use, including a step-by-step description of a real-world scenario for installing, scaling, and managing Apache HBase and Apache Cassandra with StackIQ Cluster Manager. StackIQ Cluster Manager has been used in similar deployments between 2 nodes and 4,000+ nodes.

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. A cluster-aware management framework leverages this database to define server configuration, deploy software, manage cluster services, and monitor the environment. Specific features of the Cluster Manager include:

- Provisioning and managing the operating system from bare metal
- Configuring host-based network settings throughout the cluster
- Leveraging hardware resource information (such as CPU, memory, and disk layout) to set cluster application parameters
- Setting up disk controllers and using this information to programmatically partition disks for specific cluster services
- Installing and configuring a cluster monitoring system
- Providing a unified interface (CLI and GUI) for controlling and monitoring all of the above

In addition to managing the underlying cluster infrastructure, StackIQ Cluster Manager’s capabilities also handle the day-to-day operation of cluster services. Other competing products fail to integrate the management of the cluster stack with the application stack. By contrast, StackIQ leverages its knowledge and control of the underlying cluster infrastructure to manage cluster services such as HDFS, MapReduce, HBase, and Cassandra. Some benefits of this integration include:

- Faster time to value through automation
- Consistent, dependable application deployment and management
- Simplified operation that can be quickly learned without cluster experience
- Reduced downtime due to configuration errors
- Reduced total cost of ownership
- Increased flexibility from the ability to quickly repurpose resources
4 Installing and Managing the Mixed-Use Cluster

“Step Zero”
Most installation instructions for cluster applications start with the assumption that a running cluster is already in place, thereby skipping over the complex, time-consuming process of building and managing cluster infrastructure. Since these instructions generally start with “Step 1”, StackIQ refers to the actual first step as “Step Zero”.

For the Apache Cassandra and Apache HBase tests the StackIQ Cluster Manager node was installed from bare-metal (i.e. there was no pre-requisite software and no Operating System was previously installed) by burning the StackIQ Cluster Core Roll ISO to DVD and booting from it (the StackIQ Cluster Core Roll can be obtained from the “Rolls” section after registering at http://www.stackiq.com/download). The Core Roll led the user through a few simple forms (e.g., set the IP address of the Cluster Manager, set the gateway, DNS server, etc.) and then asked for a base OS DVD (for this installation, Oracle Linux 6.3 was supplied; similar distributions such as Red Hat Enterprise Linux & CentOS are supported as well). The installer copied all the bits from both DVDs and automatically created a new Oracle Linux distribution by blending the packages from all “Rolls” together. The user can also optionally select additional “Rolls” for the cluster; in this case, the Cloudera Roll and the Cassandra Roll were used, which were purpose built for the end-user.

After the Cluster Manager was installed and booted (in about 30 minutes), it was put into “discovery” mode using the GUI and all backend nodes were set to PXE-first and booted. The Cluster Manager discovered and installed each backend node in parallel (in about 15 minutes) -- no manual steps were required.

As shown in Figure 3, the default view from StackIQ Cluster Manager provides an interface to cluster-wide monitoring and management. The left-hand pane is used to change the view from Global, to Appliance, Rack, or Host, each of which provide context sensitive tabs for monitoring and management.

Meeting the Customer Specifications
Of the 60 data nodes used in the tests 20 were dedicated to running the customer’s workload generator program. The remaining 40 were configured with either Cassandra or the Cloudera Distribution of Hadoop (CDH) with HBase, using the StackIQ Cluster Manager (StackIQ CM) to meet the customer’s specifications. This means this configuration could be changed or automatically replicated on any cluster managed by a StackIQ CM. For example, if the current StackIQ CM node were to fail, one could build a new StackIQ CM from bare metal and it would contain all customer-specific settings on first boot. An initial 10-node Cassandra instance and an initial 10-node HBase instance were installed concurrently on separate racks by the StackIQ CM.

As shown in Figure 4, the “Attributes” tab provides the ability to add, remove, or change values in the StackIQ database that are used for application and service configuration. This tab is present for Global, Appliance, and Host views so the user can choose the appropriate granularity of each attribute.
Figure 3. StackIQ Cluster Manager Monitor Tab

Figure 4. StackIQ Cluster Manager Attributes Tab
Automated Scaling
Both instances went through a "step-up" procedure where 10 nodes were incrementally added via the StackIQ CM until the node count reached 20. Both Cassandra and HBase were "stepped-up". When the 20-node tests completed, the 20 HBase nodes were converted into Cassandra nodes via reprovisioning using the StackIQ CM. The Dell PowerEdge RAID Controllers (PERC) were automatically reconfigured for Cassandra (a single RAID 0 boot disk and 12 drives configured as a RAID 10). The new 20-nodes were "perfectly shuffled" into the existing 20-node Cassandra instance to bring the instance up to 40 nodes with the StackIQ CM. A perfect shuffle is the fastest way in which to expand an existing Cassandra instance.

Dynamic Repurposing
When the customer completed their 40-node Cassandra testing, the 40 Cassandra nodes were "flipped" into HBase nodes. Using the StackIQ CM, a bare-metal reinstalltion of all 40-nodes was kicked off in order to automatically and correctly optimize the underlying software stack and PERCs for HBase. It took 1 hour and 38 minutes to completely reconfigure the system and bring a 40-node instance of HBase online. This required configuring and starting a 40-node HDFS instance (4 commands), a 40-node MapReduce instance (3 commands), a 3-node ZooKeeper instance (3 commands) and a 40-node HBase instance (3 commands). During the switchover, StackIQ Cluster Manager coordinated the parallel reconstruction of almost a quarter petabyte of disk from RAID10 to RAID Bunch of Disks (RBOD) formatted with XFS.

After the customer completed their 40-node HBase testing the 40 HBase nodes were flipped back into Cassandra nodes with a third disk configuration, again using the StackIQ CM to perform a bare-metal reinstalltion of all 40-nodes in order to automatically and correctly configure the PERCs to the customer’s new specifications. During the switchover, StackIQ Cluster Manager again coordinated the parallel reconstruction of almost a quarter petabyte of disk in under two-hours (note: had the cluster contained 400-nodes, rather than 40-nodes the time would have stayed approximately the same due to the parallel nature of StackIQ’s installation and management system).

See Table 3 for a complete listing of all tests performed.

Table 3. Cassandra/HBase Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Workload Driver Nodes</th>
<th>Cassandra Nodes</th>
<th>HBase Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>40</td>
<td>0</td>
</tr>
</tbody>
</table>
Advanced Diagnostics and Repair
After the 10-node HBase instance was running, the customer’s engineers decided to try a different caching policy on the virtual disks storing the test data. The StackIQ CM was used to change each of the virtual disks from write-through mode to write-back mode, in parallel. Additionally, the new settings were stored in the StackIQ CM local database so when a node is installed, its PERC will be automatically configured with these new settings (for existing nodes and for any new nodes assigned to run HBase).

During the 40-node test, one node suffered a hardware failure, as expected in any large-scale environment, which caused the Cassandra service on that node to fail. That node was removed from the Cassandra configuration and a spare node was installed as a Cassandra node and then added to the Cassandra instance. Again, the procedure was fully automated with the StackIQ Cluster Manager.

Success
At the end of the proof-of-concept tests, the customer commented that the rapid redeployment capabilities of the StackIQ Cluster Manager enabled them to: 1) experiment with more configurations than they thought possible in the available time frame; and 2) completely focus on testing Cassandra and HBase rather than spend time thinking about how to deploy and configure the services. In other words, the joint Dell | StackIQ solution successfully enables customers to drive more value from their investment and focus on results; a win for all parties.

To Learn More
For more information on installing and using the StackIQ Cluster Manager, please visit http://www.stackiq.com/support or http://www.youtube.com/stackiq. The StackIQ Cluster Manager demonstration video is a good place to start: http://www.youtube.com/watch?v=qVPZcA-yHQY

To learn more about Dell and Hadoop: www.dell.com/hadoop

To learn more about the Dell Solution Centers: www.dell.com/solutioncenters