

Oracle Maximum
Availability Architecture

Oracle Big Data Appliance

Maximum Availability Architecture

ORACLE WHITE PAPER | MARCH 2016



Table of Contents

Executive Overview	3
Big Data MAA Architecture	5
Inherent HA Benefits	7
Servers	7
Storage	7
Connectivity	8
InfiniBand Switches	8
Power Distribution Unit (PDU)	9
BDA Critical and Noncritical Nodes	9
BDA Software Components	9
NameNode	9
ResourceManager	10
BDA Service Locations for One or More CDH Clusters in a Single Rack	11
High Availability and Single Points of Failure	12
BDA Critical Service Locations	12
Big Data SQL Overview	13
Focus of HA Fault Testing	14
Application Load During Tests	14
MAA Test Scenarios	15
MAA Test Scenario Details	15
Failure of the Active NameNode	15
Failure with Service Migration	18

Failure of the Active NameNode and Standby NameNode	19
Failure of the First ResourceManager, Cloudera Manager and MySQL Database	22
Failure of the Second ResourceManager and Hive Metastore Server	25
InfiniBand Switch Failure	27
Cisco Management Switch Failure	33
Big Data SQL (BDS) Server Process Failure	33
Big Data SQL (BDS) Server Process Failure on All Nodes	36
Entire PDU Failure on the BDA Rack	40
BDA System Disk Failure	43
BDA Data Disk Failure	45
Exadata Big Data SQL HA Tests	46
Oracle RAC Database Node Failure	46
Oracle RAC Database Instance Failure	47
BDA Cluster Resource Failure on Exadata	48
Conclusion	50
Appendix A	51
MAA Test Scenarios Quick Reference	51



Executive Overview

Oracle Maximum Availability Architecture (MAA) is Oracle's best practices blueprint based on proven Oracle high availability technologies, along with expert recommendations and customer experiences. MAA best practices have been highly integrated into the design and operational capability of the Oracle Big Data Appliance, and together they provide the most comprehensive highly available solution for Big Data. Oracle MAA papers are published at the [MAA home page](#) of the Oracle Technology Network (OTN) website.

Oracle Big Data Appliance (BDA) Maximum Availability Architecture is a best-practices blueprint for achieving an optimal high-availability deployment using Oracle high-availability technologies and recommendations. The Oracle BDA MAA exercise for this paper was executed on Oracle Big Data Appliance and Oracle Exadata Database Machine to validate high availability and to measure downtime in various outage scenarios. The current release of this technical paper covers the first phase of the overall Oracle BDA MAA project. The project comprises the following two phases:

Phase 1: High Availability and Outage scenarios at a single site

Phase 2: Disaster Recovery Scenarios across multiple sites.

Oracle Big Data Appliance is an engineered system comprised of hardware and software components that are designed, tested, and optimized together according to Oracle MAA standards to provide the highest application availability and performance. It delivers:

- » A complete and optimized solution for big data
- » Single-vendor support for both hardware and software
- » An easy-to-deploy solution

Oracle Big Data Appliance provides a flexible, high performance, secure platform for running diverse workloads on Hadoop and NoSQL systems. The platform captures, organizes, and supports deep analytics on extremely large, complex data streams flowing into the enterprise from many data sources; and incorporates the ability to choose the best storage and processing location for the data depending on its structure, workload characteristics, and end-user requirements.



Oracle Big Data Appliance is tightly integrated with Oracle Database and Oracle Exadata Database Machine, and incorporates the same maximum availability architecture proven internally both by Oracle and mission critical customers worldwide. Oracle Exadata Database Machine provides outstanding performance in hosting data warehouses and transaction processing databases.

For maximum speed and efficiency, Oracle Big Data Appliance can be connected to Oracle Exadata Database Machine using InfiniBand technology. The InfiniBand connection between the engineered systems provides low latency and high throughput, which enables high-speed data transfer for batch and query workloads.

Oracle Big Data Appliance is the platform for acquiring and organizing big data, and enables exploration and analysis of data using the latest big data technologies. The addition of Oracle Database in front of the Big Data Appliance makes it easy to combine these insights with data in Oracle Database. Oracle Big Data SQL unifies data that spans multiple sources, and leverages Oracle's rich SQL dialect and security policies.

Oracle Exadata MAA best practices and unparalleled performance are leveraged to produce a tightly integrated, highly available, and high performing end-to-end system.

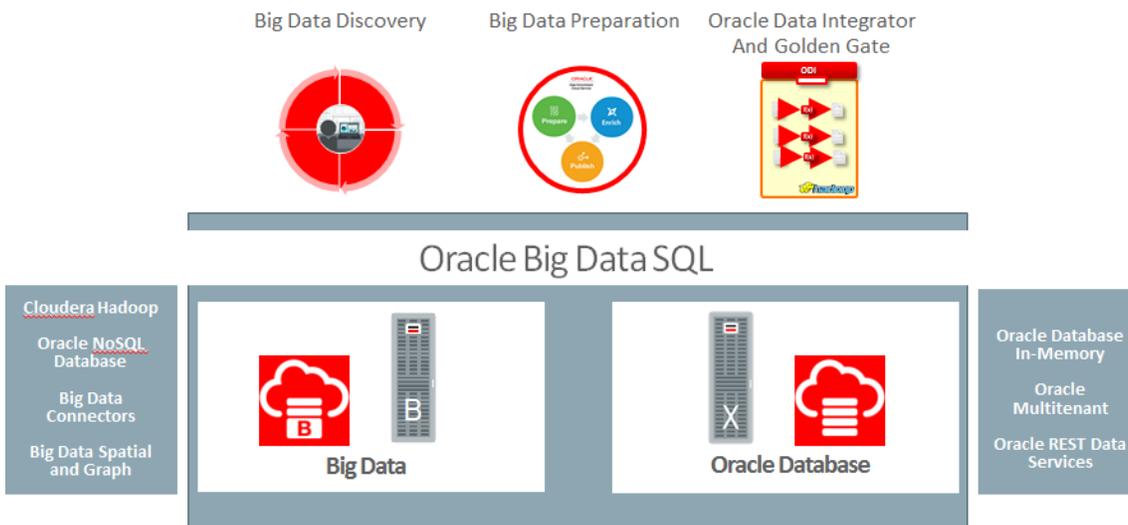


Fig 1. The Big Data Appliance and Exadata Database Machine are tightly integrated for performance both on-premise and in the cloud



Big Data MAA Architecture

The Oracle Big Data MAA architecture consists of the following technologies:

- » A primary Big Data Appliance used to house a data reservoir. The data reservoir acts as a repository for new and large sources of structured and unstructured data, and augments the data warehouse running on Exadata. The data reservoir may consist of one or more Big Data Appliances interconnected to address storage, performance, and growth needs. Up to 18 racks can be added without additional switches.

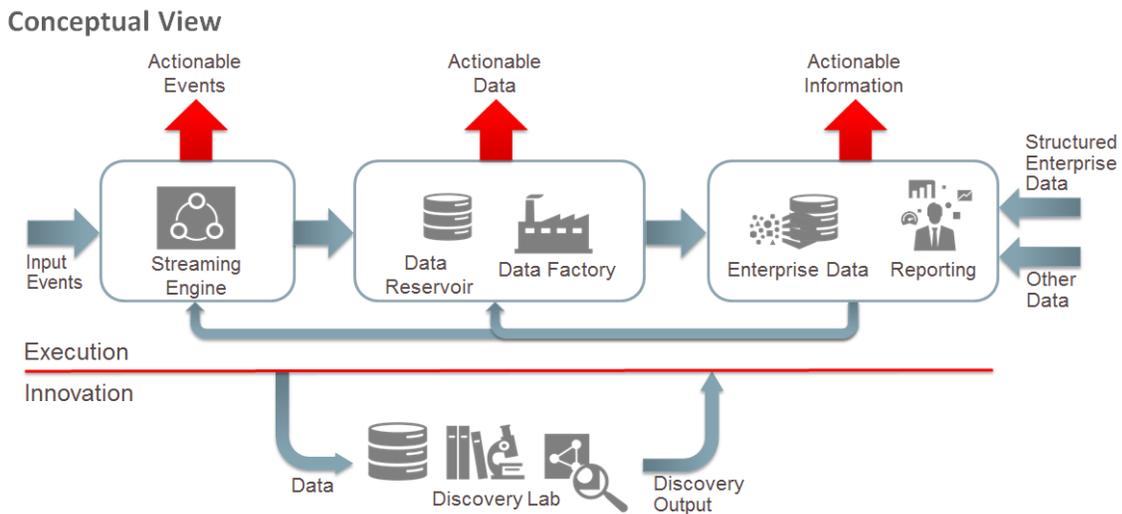


Fig 2. Information Management Reference Architecture

- » Tight integration with Oracle Exadata to create a Big Data Management System. Exadata should hold the main data warehouse system and store much of a company's core transactional data.
- » A standby Exadata system that is a replica of the primary running Oracle Data Guard is used to maintain synchronized databases that are exact physical replicas of the primary.
- » Replication of data to a second Big Data Appliance ensures high availability and added data consistency.

ORACLE®

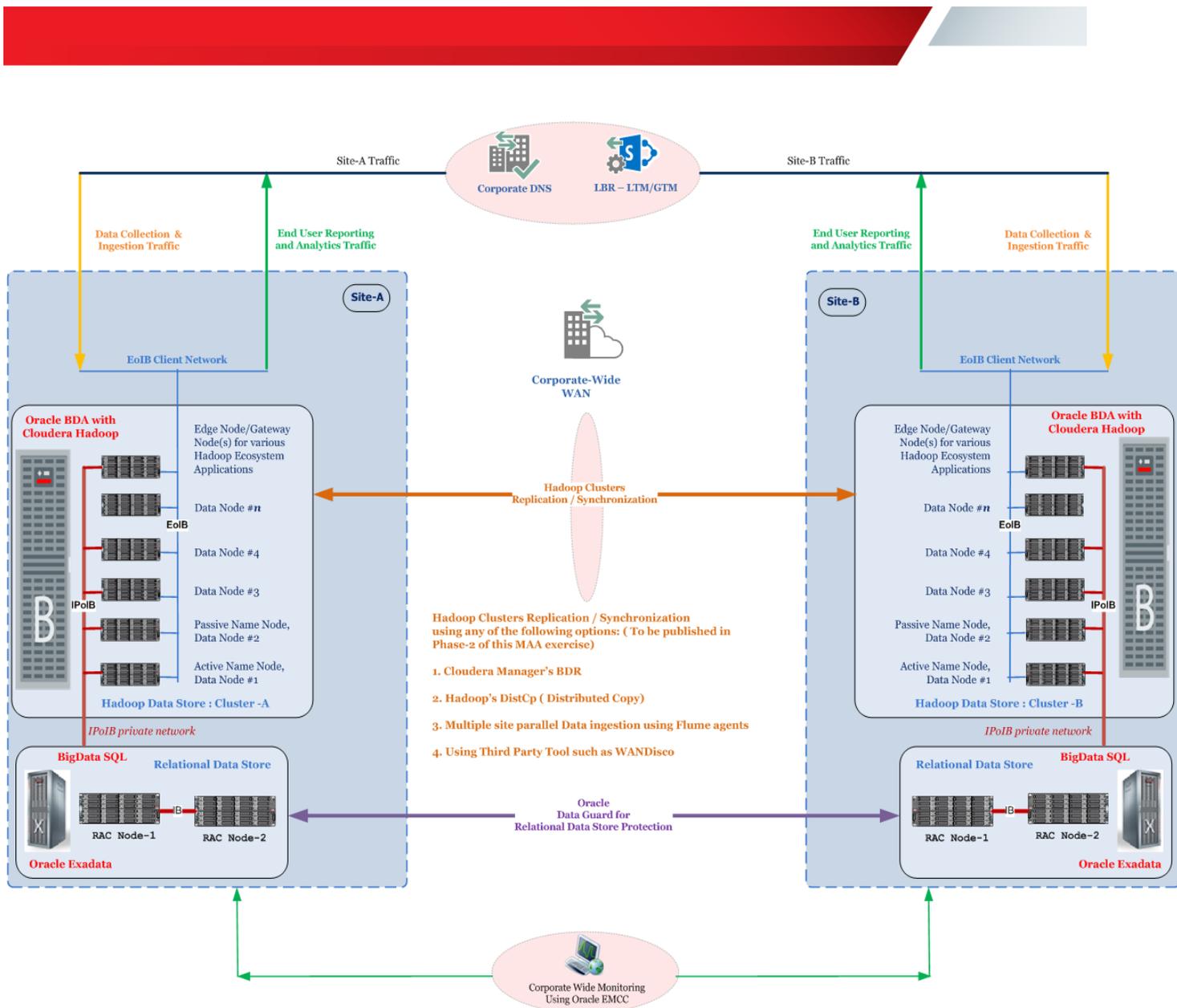


Fig 3. Big Data Appliance replication ensures high availability and added data consistency

- » Oracle Big Data SQL enables the full power of Oracle SQL to provide a single view of all data across Oracle databases, Hadoop, and NoSQL data sources.
- » The Big Data Management System can be used both in the cloud with Oracle's Big Data Cloud Service (BDCS) and on-premise. Oracle BDCS is integrated with Exadata Cloud Service enabling one fast query across all data sources.



Inherent HA Benefits

Oracle Big Data Appliance is designed to tolerate unplanned outages of all types. As part of this document, Oracle performed extensive testing to demonstrate the built-in HA capabilities of the Big Data Appliance and Exadata Database Machine. Specific failure scenarios are addressed in the [MAA Test Scenarios](#) section of this document.

Both the Oracle Big Data Appliance and Exadata are engineered and preconfigured to achieve end-to-end application availability after hardware faults on components such as fans, PDUs, batteries, switches, disks, flash, database server, motherboards, and DIMMs; extensive engineering and integration tests validate every aspect of the systems.

The full suite of Exadata HA capabilities and tests is covered in previous whitepapers and documentation. Please refer to <http://www.oracle.com/goto/maa> for additional information.

Servers

All servers in a BDA rack are homogeneous, and there are no specialized nodes. This enables every node to potentially fulfill every role, and makes service migrations and repurposing nodes to other tasks much easier.

The servers feature dual hot-swappable power supplies and fans, as well as an embedded Integrated Lights Out Manager (ILOM.)

Storage

The first two disks on every BDA server contain the Linux operating system. These disks contain a mirrored copy of the operating system, a swap partition, a mirrored boot partition, and an HDFS data partition. This allows the operating system to sustain the loss of one system disk and keep the system available while the faulty disk is replaced. The state of the mirrored devices can be verified with the following commands

```
# cat /proc/mdstat
Personalities : [raid1]
md2 : active raid1 sda2[2] sdn2[0]
      488150016 blocks super 1.1 [2/2] [UU]
      bitmap: 3/4 pages [12KB], 65536KB chunk
md0 : active raid1 sda1[3] sdn1[2]
      194496 blocks super 1.0 [2/2] [UU]

unused devices: <none>
```

The Oracle logo, consisting of the word "ORACLE" in white, uppercase, sans-serif font, centered within a red rectangular background.

All drives are part of either a CDH (Cloudera Distribution Including Apache Hadoop) cluster (HDFS) or an Oracle NoSQL Database cluster. HDFS (Hadoop Distributed File System) is a highly scalable file system that stores large files across multiple servers, and can scale rapidly as additional storage is needed. HDFS data is replicated in three locations, reducing the chance of any data loss.

Connectivity

Each BDA server is fitted with an InfiniBand dual-ported Quad Data Rate (QDR) Host Channel Adapter (HCA) card. Each port on the HCA is connected to a redundant InfiniBand switch in the rack. If a port or an entire InfiniBand switch fails, a short brownout may be experienced in the order of a second or less, as traffic is redirected over the active interface.

InfiniBand Switches

All BDA nodes are redundantly connected to the InfiniBand fabric, which acts as a high performance and highly available backplane for BDA operations.

The InfiniBand fabric is constructed of two fully redundant InfiniBand Gateway switches. If one switch fails the other working switch ensures the overall system continues to function. The InfiniBand Gateways act as both “leaf” switches, and InfiniBand to 10Gb Ethernet gateways. There is also a 36-port “spine” switch that enables expansion of the BDA system.

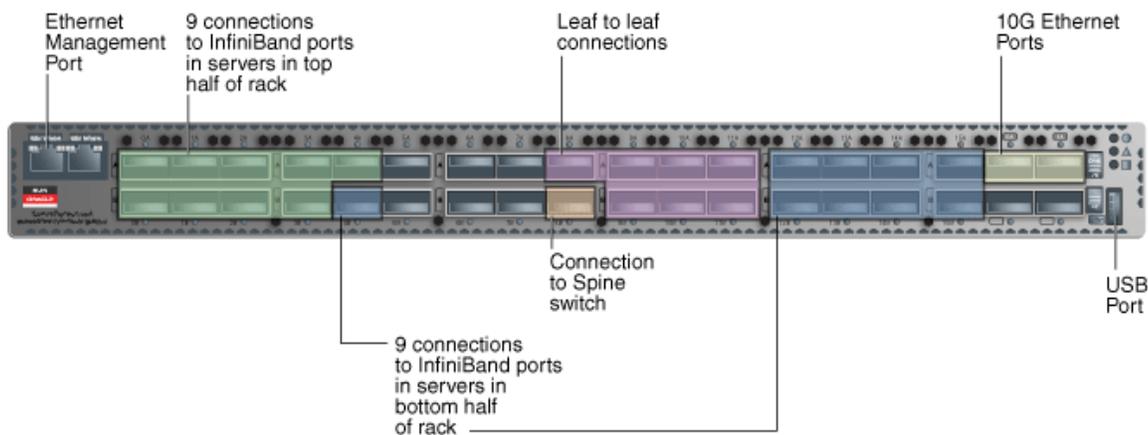


Fig 4. Port locations on the InfiniBand Gateway switch

Power Distribution Unit (PDU)

The Big Data Appliance has redundant power distribution units (PDUs) for high availability. The PDUs provide redundant power to the following rack components

- » BDA Nodes
- » InfiniBand Switches
- » Cisco Network Switch

All of the power supplies for the above components are hot swappable. Each PDU can be monitored to provide insight into how much power, energy and current the connected equipment uses, as well as the voltage level powering the equipment.

BDA Critical and Noncritical Nodes

Critical nodes are required for the cluster to operate normally and provide all services to users. In contrast, the cluster continues to operate with no loss of service when a noncritical node fails.

On single-rack clusters, the critical services are installed initially on the first four nodes of the cluster. The remaining nodes (node05 up to node18) only run noncritical services. If a hardware failure occurs on one of the critical nodes, then the services can be moved to another noncritical server. For example, if node02 fails, its critical services can be moved to node05.

See Oracle Big Data Documentation for information about where [services run in a multi-rack CDH cluster](#)¹.

BDA Software Components

NameNode

The NameNode is the most critical process because it keeps track of the location of all data. Without a healthy NameNode, the entire cluster fails. Apache Hadoop v0.20.2 and earlier are vulnerable to failure because they have a single name node.

CDH5 reduces this vulnerability by maintaining redundant NameNodes. The data is replicated during normal operation as follows:

- » CDH maintains redundant NameNodes on the first two nodes of a cluster. One of the NameNodes is in active mode, and the other NameNode is in hot standby mode. If the active NameNode fails, the standby NameNode automatically assumes the role of the active NameNode.

¹ http://docs.oracle.com/cd/E69290_01/doc.44/e65665/admin.htm#BIGUG270

- » The NameNode data is written to a mirrored partition so that the loss of a single disk can be tolerated. This mirroring is done at the factory as part of the operating system installation.
- » The active NameNode records all changes to the file system metadata in at least two JournalNode processes, which the standby NameNode reads. There are three JournalNodes which run on the first three nodes of each cluster.
- » The changes recorded in the journals are periodically consolidated into a single fsimage file in a process called checkpointing.

The diagram below shows the relationships among the processes that support automatic failover of the NameNode.

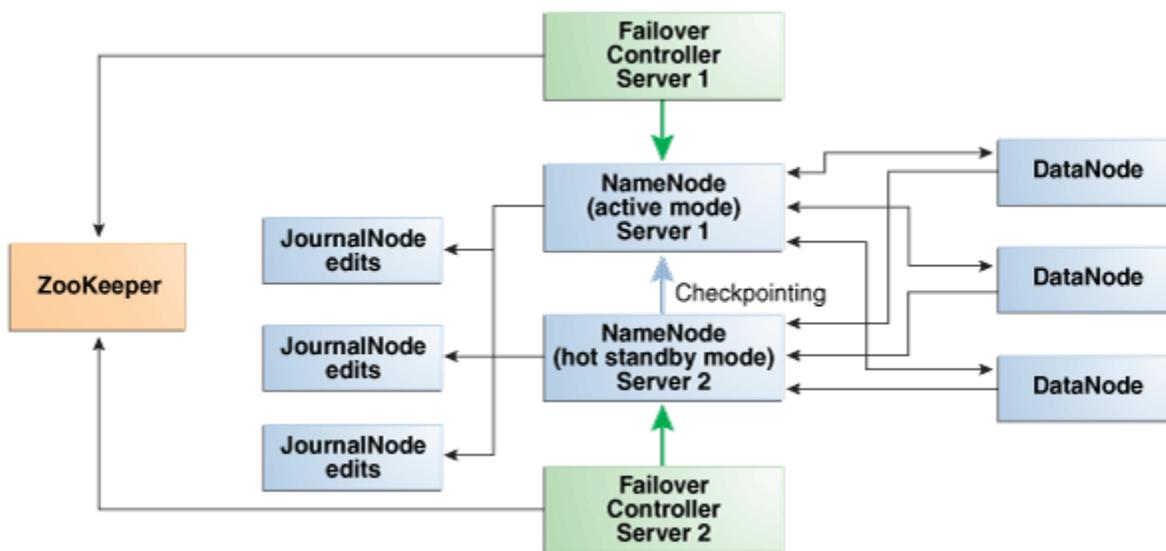


Fig 5. Automatic Failover of the NameNode on Oracle Big Data Appliance

ResourceManager

The ResourceManager allocates resources for application tasks and application masters across the cluster. Like the NameNode, the ResourceManager is a critical point of failure for the cluster. If all ResourceManagers fail, then all jobs stop running. High Availability is supported by Oracle Big Data Appliance 3.0 and above in CDH5 to reduce this vulnerability.

CDH maintains redundant ResourceManager services on node03 and node04. One of the services is in active mode and the other service is in hot standby mode. If the active service fails then the role of the active ResourceManager automatically fails over to the standby service. No failover controllers are required.

The diagram below shows the relationships among the processes that support automatic failover of the ResourceManager.



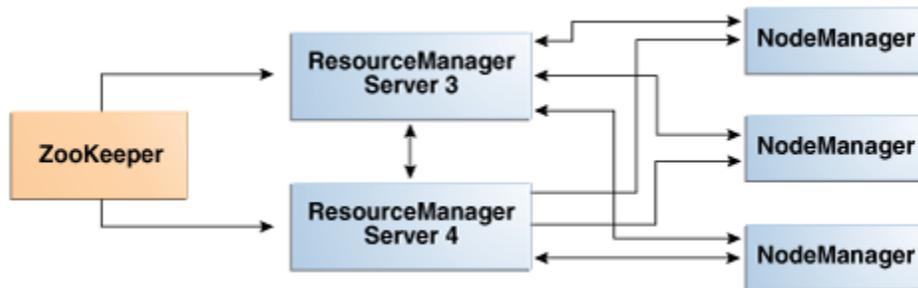


Fig 6. Automatic Failover of the ResourceManager on Oracle Big Data Appliance

BDA Service Locations for One or More CDH Clusters in a Single Rack

Node01	Node02	Node03	Node04	Node05 to nn
Balancer		Cloudera Manager Server		
Cloudera Manager Agent	Cloudera Manager Agent	Cloudera Manager Agent	Cloudera Manager Agent	Cloudera Manager Agent
DataNode	DataNode	DataNode	DataNode	DataNode
Failover Controller	Failover Controller	JobHistory	Hive, Hue, Oozie, Solr	
JournalNode	JournalNode	JournalNode		
	MySQL Backup	MySQL Primary		
NameNode	NameNode			
NodeManager	NodeManager	NodeManager	NodeManager	NodeManager
			Oracle Data Integrator	
Puppet	Puppet	Puppet	Puppet	Puppet
Puppet Master		ResourceManager	ResourceManager	
ZooKeeper	ZooKeeper	ZooKeeper		



High Availability and Single Points of Failure

Some services have built-in high availability and automatic failover, and other services have a single point of failure. The following list summarizes the critical services and their vulnerabilities:

- » NameNodes: High availability with automatic failover
- » ResourceManagers: High availability with automatic failover
- » MySQL Database: Primary and backup databases are configured with replication of the primary database to the backup database. There is no automatic failover. If the primary database fails, the functionality of the cluster is diminished, but no data is lost.
- » Cloudera Manager: The Cloudera Manager server runs on one node. If it fails then Cloudera Manager functionality is unavailable.
- » Oozie server, Hive server, Hue server, and Oracle Data Integrator agent: These services have no redundancy. If the node fails, then the services are unavailable.

BDA Critical Service Locations

Node Name	Critical Functions
First NameNode	Balancer, Failover Controller, JournalNode, NameNode, Puppet Master, ZooKeeper
Second NameNode	Failover Controller, JournalNode, MySQL Backup Database, NameNode, Puppet, ZooKeeper
First ResourceManager Node	Cloudera Manager Server, JobHistory, JournalNode, MySQL Primary Database, ResourceManager, ZooKeeper
Second ResourceManager Node	Hive, Hue, Oozie, Solr, NodeManager, Oracle Data Integrator Agent, ResourceManager

The information above is valid for a BDA single rack configuration. In a [multi-rack scenario](#)² service locations may be different. The *BDA Software User's Guide* contains detailed information on service locations for different rack scenarios.

² http://docs.oracle.com/cd/E69290_01/doc.44/e65665/admin.htm#BIGUG270

Big Data SQL Overview

An increasing volume of enterprise data resides in Hadoop and NoSQL systems. Big Data SQL allows customers to leverage the industry standard SQL language to seamlessly access data stored in their Hadoop and NoSQL systems, and combine this with data stored in Oracle Database.

Oracle Big Data SQL allows queries for huge amounts of data to be issued using Oracle's rich SQL dialect against multiple data sources, including Hive, HDFS, Oracle NoSQL, and HBase. Using Big Data SQL, organizations can integrate data from relational databases, Hadoop, and NoSQL sources in a single query, and extend their Oracle Database security policies to these external sources.

Big Data SQL leverages the tight integration between Oracle Big Data Appliance and Oracle Exadata Database Machine, and InfiniBand connectivity provides low latency, high bandwidth transfer between the systems. Big Data SQL utilizes the significant performance advantages of Exadata, providing Exadata-inspired smart scan technology such as filter-predicate offloads, and storage indexes to SQL queries spanning diverse underlying technologies.

Oracle Big Data SQL provides external tables with next generation performance gains. An external table is an Oracle Database object that identifies and describes the location of data outside of a database. You can query an external table using the same SQL SELECT syntax that is used for any other database tables.

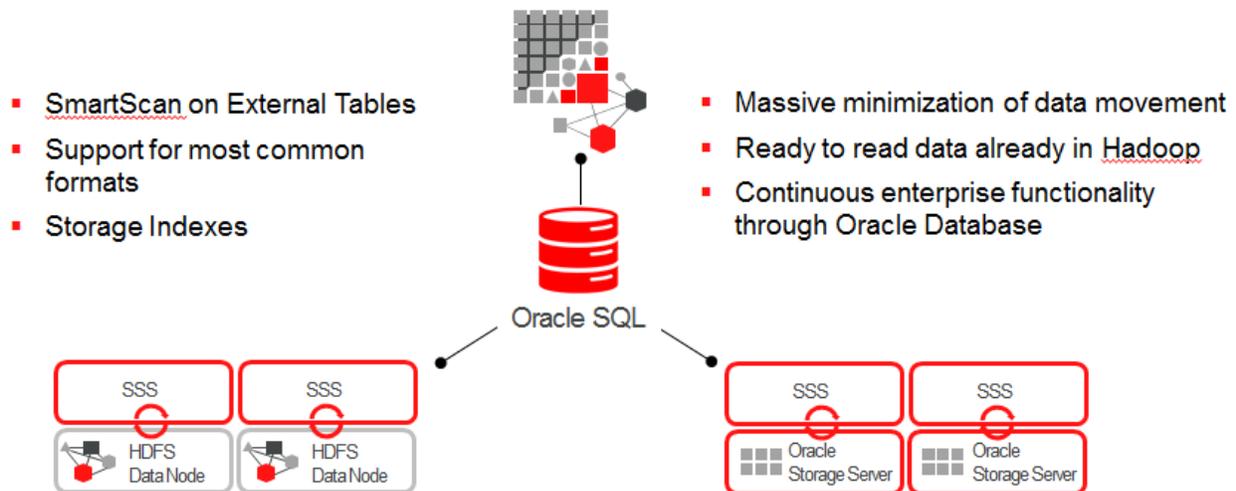


Fig 7. Big Data SQL enables Exadata Smart Scan performance against diverse technologies

Big Data SQL provides a unified query methodology using Smart Scan technology. Agents local to the data on the Hadoop system provide fast scans and data filtering, while joins are handled by the querying database.

ORACLE®



Focus of HA Fault Testing

The focus of testing for this paper was to deliberately cause the failure of certain key components within the BDA and Exadata racks, and to demonstrate the inherent high availability and MAA architecture built into the overall Big Data Management System.

Where elements in either rack were considered highly available, the operational impacts of the component failure were tested and documented; interruptions of service and downtime were also noted where appropriate.

Application Load During Tests

Oracle “Copy to BDA” (a component of Oracle Big Data SQL,) was used to copy tables from an Oracle database into Hadoop. As part of this process, Data Pump files were created for each table, and then copied to HDFS on the BDA using Hadoop commands. Hive external tables were created over the Data Pump files to provide access to the data. To understand end-to-end impacts, external tables were created for Hive data in an Oracle Database located on an Exadata platform, and a combination of load tests were executed against local tables in Exadata and remote tables located on the BDA.

The Swingbench load generation tool and a SQL*Plus workload were run on the Exadata platform. On the BDA platform, Hadoop tools included in the Cloudera distribution such as TeraGen, TeraSort, TestDFSIO and NNBench were utilized. The BigBench tool was also utilized; BigBench is adapted from TPC-DS, and is a proposed industry standard performance benchmark for big data.

MAA Test Scenarios

High availability and application impacts were assessed by injecting the following hardware and software failures -

- (1) Failure of the Active NameNode
- (2) Failure with Service Migration
- (3) Failure of the Active NameNode and Standby NameNode
- (4) Failure of the First ResourceManager, Cloudera Manager and MySQL Database
- (5) Failure of the Standby ResourceManager and Hive Metastore Server
- (6) InfiniBand Switch Failure
- (7) Cisco Management Switch Failure
- (8) Big Data SQL (BDS) Server process Failure
- (9) Big Data SQL (BDS) Server Process Failure on All Nodes
- (10) Entire PDU Failure on the BDA rack
- (11) BDA System Disk Failure
- (12) BDA Data Disk Failure
- (13) Exadata Database Node Failure
- (14) Exadata RAC Database Instance Failure
- (15) Exadata Failure of the BDA Clusterware resource

MAA Test Scenario Details

Failure of the Active NameNode

High availability is provided by two redundant NameNodes in a Big Data Appliance deployment. One NameNode is active and the other is in a standby state. For this test, a crash of the active NameNode was simulated by removing both redundant power supply cables from the back of the active NameNode server.

The event and its impacts were monitored from Cloudera Manager, Oracle Enterprise Manager, BDA MapReduce jobs, and Big Data SQL operations using Swingbench and SQL*Plus.

The expected behavior in this situation is for the second NameNode to automatically take over and become the active NameNode. Once the second NameNode is active, it will continue without a backup. However, with only one NameNode the cluster is then vulnerable to failure, and no longer has the redundancy needed for automatic failover.

During a failure of the first NameNode, the following services are also unavailable on that node; however, they do not affect the availability of the cluster. Refer to the [BDA Service Locations](#) section for more information.

The Oracle logo, consisting of the word "ORACLE" in white, uppercase letters on a red rectangular background.

Additional services on Node1:

- » Balancer
- » Cloudera Manager Agent
- » DataNode
- » Failover Controller
- » NodeManager
- » Puppet
- » Puppet Master
- » ZooKeeper

As the puppet master runs on this node, and the Mammoth utility uses Puppet, it is not possible to install or reinstall software if this node is down; for example, if a disk drive must be replaced elsewhere in the rack.

The active NameNode was deliberately crashed by removing both redundant power supply cables from the server; all query processing on the cluster stopped for a short time.



Status	Name	IP	Roles
●	scaj42bda01.us.oracle.com	192.168.23.73	10 Role(s)
●	scaj42bda02.us.oracle.com	192.168.23.74	10 Role(s)
●	scaj42bda03.us.oracle.com	192.168.23.75	15 Role(s)
●	scaj42bda04.us.oracle.com	192.168.23.76	10 Role(s)

Fig 8. The Cloudera Manager Hosts screen shows that the NameNode is down (node 1)

Instance	Log File
<input checked="" type="radio"/> BDS Server	Log File
<input type="radio"/> Balancer	Log File
<input checked="" type="radio"/> DataNode	Log File ↗
<input checked="" type="radio"/> Failover Controller	Log File ↗
<input checked="" type="radio"/> JournalNode	Log File ↗
<input checked="" type="radio"/> NameNode	Log File ↗
<input type="radio"/> Gateway	Log File
<input type="radio"/> Gateway	Log File
<input checked="" type="radio"/> NodeManager	Log File ↗

Fig 9. Cloudera Manager confirms all roles are down

After a short time (under a minute,) the second NameNode became the active NameNode as shown below.

maacuster-a > hdfs > **NameNode, scaj42bda02 (Active)** Status Configuration

Summary

Host: ● scaj42bda02 [Log File ↗](#) [Stacks Logs ▾](#)

Quick Links: [NameNode Web UI ↗](#)

Event Search: [Alerts ↗](#), [Critical ↗](#), [All ↗](#)

Health Tests [Expand All](#) [Create Trigger](#)

▶ ● 16 good.

⊘ Test disabled because role is not configured to dump heap when out of memory. Test of whether this role's heap dump directory has enough free space. [Details](#)

Health History

- ▶ ● 11:01:13 AM 2 Became Good [Show](#)
- ▶ ● 11:01:08 AM RPC Latency Good [Show](#)
- ▶ ● Feb 17 10:01 AM RPC Latency Disabled [Show](#)
- ▶ ● Feb 17 9:56:24 AM 3 Became Good [Show](#)
2 Became Disabled
- ▶ ● Feb 17 9:56:09 AM 2 Became Good [Show](#)

Fig 10. The standby NameNode (node 2) assumed the active NameNode role



Once the active NameNode (the former standby) was running, queries continued to run, and results were returned. The duration of the brownout is related to the length of time it takes the second NameNode to become active; however there was no outage in this situation.

For testing purposes, the crashed node was then brought back online and assumed the role of the standby NameNode.

Failure with Service Migration

In a real-life outage situation, if the original active node can be repaired in a timely manner, it can be powered on and will become the new standby NameNode. However, if the node failure is more critical and will take time to repair, services can be migrated to a non-critical node using `bdacli` commands. The high level steps to perform a migration are:

- (1) If the failed node is where Mammoth is installed, the Mammoth bundle must be downloaded to another noncritical node in the same cluster.

Note: This isn't necessary if other critical nodes are lost.

- a. Download the Mammoth patch to the non-critical node where services are to be located.
- b. Unzip both patches to the non-critical node.
- c. Install Mammoth to the non-critical node. Below is an example of the steps.

```
# cd <patchdir>/BDAMammoth-ol6-4.3.0
# ./BDAMammoth-ol6-4.3.0.run
<output truncated>
LIST OF STEPS:
    Step 1 = PreinstallChecks
    Step 2 = SetupPuppet
    Step 3 = PatchFactoryImage
    Step 4 = CopyLicenseFiles
    Step 5 = CopySoftwareSource
    Step 6 = CreateUsers
    Step 7 = SetupMountPoints
    Step 8 = SetupMySQL
    Step 9 = InstallHadoop
    Step 10 = StartHadoopServices
    Step 11 = InstallBDASoftware
    Step 12 = SetupKerberos
```

This will take some time and run through all the listed steps.





(2) Perform the node migration from the new/chosen non-critical node.

```
# bdacli admin_cluster migrate node1
```

(3) Once node1 has been repaired or replaced it can be re-provisioned and introduced back into the system. Run the `bdacli reprovision` command from the node where services were migrated to -

```
# bdacli admin_cluster reprovision node1
```

Specific instructions are documented in the [Managing a Hardware Failure](#)³ section of the *BDA Software Users Guide*.

Failure of the Active NameNode and Standby NameNode

Since the NameNode is a critical role, the objective of this test was to observe the operational impact on the BDA if both the active (first) NameNode and the standby (second) NameNode both crashed within a short space of time and remained down. Both servers were deliberately crashed by removing their redundant power supply cables. The events were monitored from Cloudera Manager and the client applications to observe the effects on the workload.

The expected behavior in this case is for the cluster to fail as neither of the NameNodes are available. Additional services running on the active and standby NameNodes are also affected. However, this is of lesser concern since the cluster cannot continue with both NameNodes unavailable.

Once power was removed from the active NameNode, queries were interrupted for a short time (typically under one minute) as the standby NameNode assumed the active role; queries then resumed processing once the role transition was complete. Power was then removed from the second (now active) NameNode. Once this was done, both NameNodes were down and all query processing came to a halt.

³ http://docs.oracle.com/cd/E69290_01/doc.44/e65665/admin.htm#BIGUG76674

Status	Name	IP	Roles
	scaj42bda01.us.oracle.com	192.168.23.73	> 10 Role(s)
	scaj42bda02.us.oracle.com	192.168.23.74	> 10 Role(s)
	scaj42bda03.us.oracle.com	192.168.23.75	> 15 Role(s)
	scaj42bda04.us.oracle.com	192.168.23.76	> 10 Role(s)

Fig 11. Cludera Manager shows the active NameNode is down and the second NameNode in transition

Status	Name	IP	Roles
	scaj42bda01.us.oracle.com	192.168.23.73	> 10 Role(s)
	scaj42bda02.us.oracle.com	192.168.23.74	> 10 Role(s)
	scaj42bda03.us.oracle.com	192.168.23.75	> 15 Role(s)
	scaj42bda04.us.oracle.com	192.168.23.76	> 10 Role(s)

Fig 12. After another minute both nodes changed to red



Roles	
Instance	Log File
<input checked="" type="radio"/>  BDS Server	Log File
<input type="radio"/>  Balancer	Log File
<input checked="" type="radio"/>  DataNode	Log File ↗
<input checked="" type="radio"/>  Failover Controller	Log File ↗
<input checked="" type="radio"/>  JournalNode	Log File ↗
<input checked="" type="radio"/>  NameNode	Log File ↗
<input type="radio"/>  Gateway	Log File
<input type="radio"/>  Gateway	Log File
<input checked="" type="radio"/>  NodeManager	Log File ↗
<input checked="" type="radio"/>  Server	Log File ↗

Fig 13. Cloudera Manager view with all roles showing red

Big Data SQL queries against Hive tables on the Big Data Appliance may return errors similar to this example:

```
SQL> ERROR at line 5:
ORA-29913: error in executing ODCIEXTTABLEOPEN callout
ORA-29400: data cartridge error
KUP-11504: error from external driver:
oracle.hadoop.sql.xcat.common.XCatException : 321 : Error getting hive metadata.
Cause : java.lang.RuntimeException:
    MetaException(message:org.apache.hadoop.hive.serde2.SerDeException
java.net.NoRouteToHostException: No Route to Host from exaadm06.us.oracle.com to
bdanode01.us.oracle.com:8020
    failed on socket timeout exception: java.net.NoRouteToHostException: No
route to host; For more details see: http://wiki.apache.org/hadoop/NoRouteToHost)
```

One of the NameNodes was restarted after ten minutes and it became the active NameNode; query processing on the BDA was then able to continue. Any failed Big Data SQL queries were resubmitted and then also finished successfully. In a real-life outage of this nature it is critical to recover one of the NameNodes in a timely manner.

Note: The NameNode is not a single point of Failure (SPOF) in a BDA rack as there is an active and standby NameNode deployed in a highly available configuration.



Failure of the First ResourceManager, Cloudera Manager and MySQL Database

The ResourceManager is a critical service within CDH. It is responsible for allocating resources for application tasks across the cluster, and if it fails then all jobs stop running. High availability for the ResourceManager is provided by a hot standby that automatically assumes the active role if the original ResourceManager fails.

During a failure of the first ResourceManager the following services are disrupted :

» MySQL Database:

Cloudera Manager, Oracle Data Integrator, Hive, and Oozie use MySQL Database. The MySQL database houses the Hive Metastore database that stores metadata for Hive tables. The Hive Metastore Service runs on a different node, but uses the MySQL database for its metadata.

The primary MySQL database is configured to automatically replicate to a backup database located on the second NameNode server (in the case of a single rack.) This provides a backup to the primary database instance. There is no automatic failover however, and the standby is not activated by default.

Cloudera Manager:

This tool provides central management for the entire CDH cluster. Without it, it is still possible to monitor activities using the native Hadoop utilities described in [Using Hadoop Monitoring Utilities](#)⁴ section of the *Oracle Big Data Appliance Software User's Guide*.

The native Hadoop utilities that will help if Cloudera Manager is down are:

- » The YARN resource manager interface for monitoring MapReduce jobs
- » The DFS Health utility for monitoring the Hadoop file system
- » Cloudera Hue to perform a number of useful interactions with Hadoop, such as querying Hive data stores, working with Hive tables, working with HDFS files, and monitoring MapReduce jobs.

To simulate a ResourceManager failure, power was removed from both redundant power supplies on the first ResourceManager node while an application workload was running. Cloudera Manager was unavailable while the node was down, and SQL sessions using Hive tables were unable to continue; Hive operations were interrupted due to the Hive metastore in the MySQL database being unavailable.

There is a standby MySQL database instance, but it is not activated by default. Primary and backup databases are configured with replication of the primary database to the backup database; however, there is no automatic failover.

Note: HDFS access to the data is unaffected by the MySQL database Hive metastore issue, and data access is still possible using HDFS. MapReduce jobs resume successfully after the standby ResourceManager becomes active.

⁴ http://docs.oracle.com/cd/E69290_01/doc.44/e65665/admin.htm#BIGUG385

SQL*Plus sessions accessing Hive tables threw an error similar to this example:

```

SELECT order_mode,
*
ERROR at line 1:
ORA-29913: error in executing ODCIEXTTABLEOPEN callout
ORA-29400: data cartridge error
KUP-11504: error from external driver:
org.apache.thrift.transport.TTransportException:
java.net.SocketTimeoutException: Read timed out
    
```

For monitoring tools when Cloudera Manager is down, the YARN resource manager, DFS, and Hue tools are still accessible, as shown in the following figures.

The YARN resource manager is accessible from the URL:

```

http://bda1node04.example.com:8088
    
```

Where bda1node03 is the server where resource manager runs.

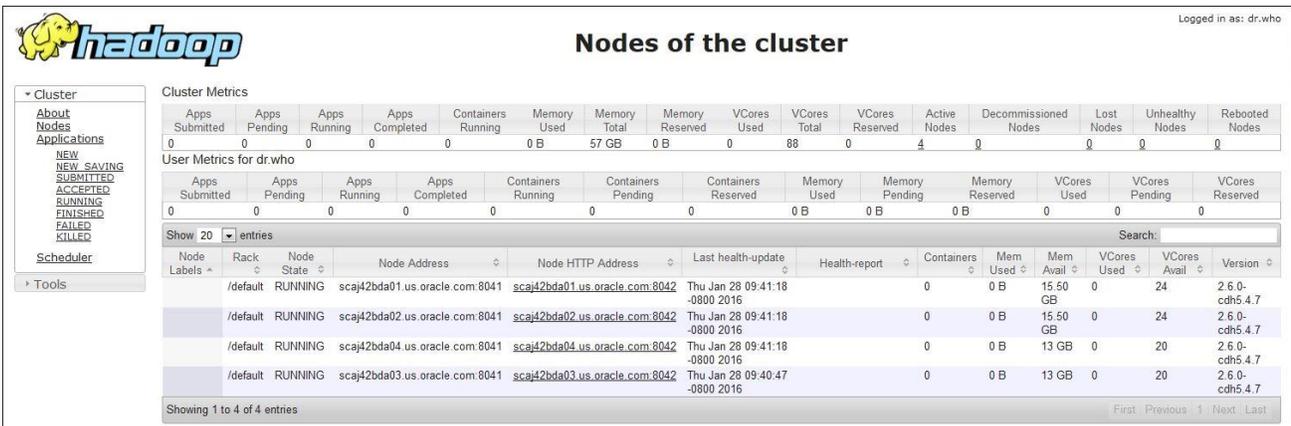


Fig 14. The YARN Resource Manager Interface.



Hue is also still accessible. Hue is access from the following server and port number where Hue runs:

<http://bda1node04.example.com:8888>

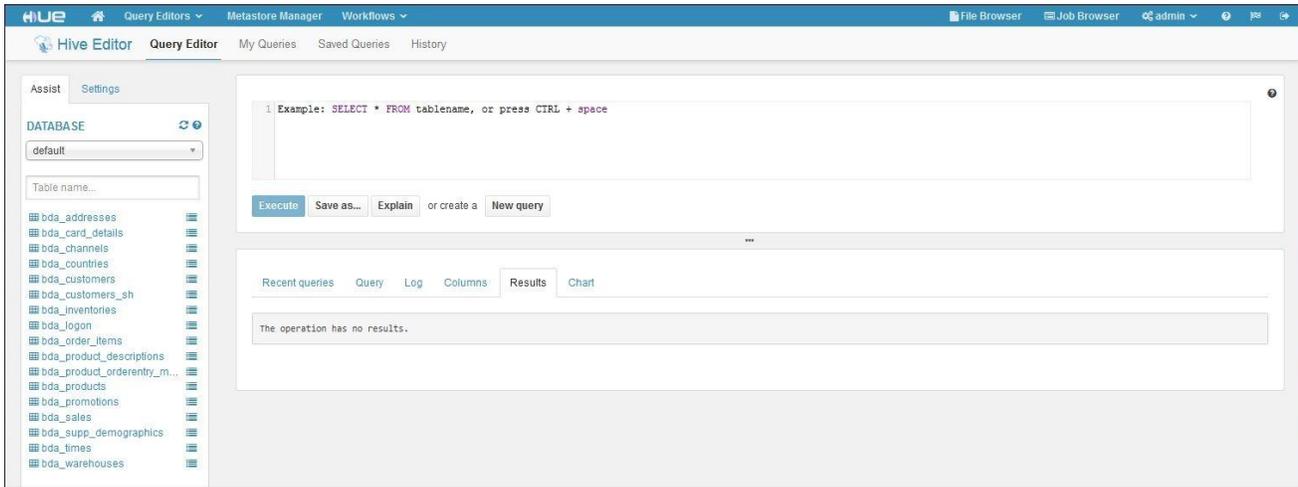


Fig 15. The Hue Interface

DFS is used for monitoring the health of the Hadoop file system, and is available from the server where DFS runs, similar to the following:

<http://bda1node01.example.com:50070>

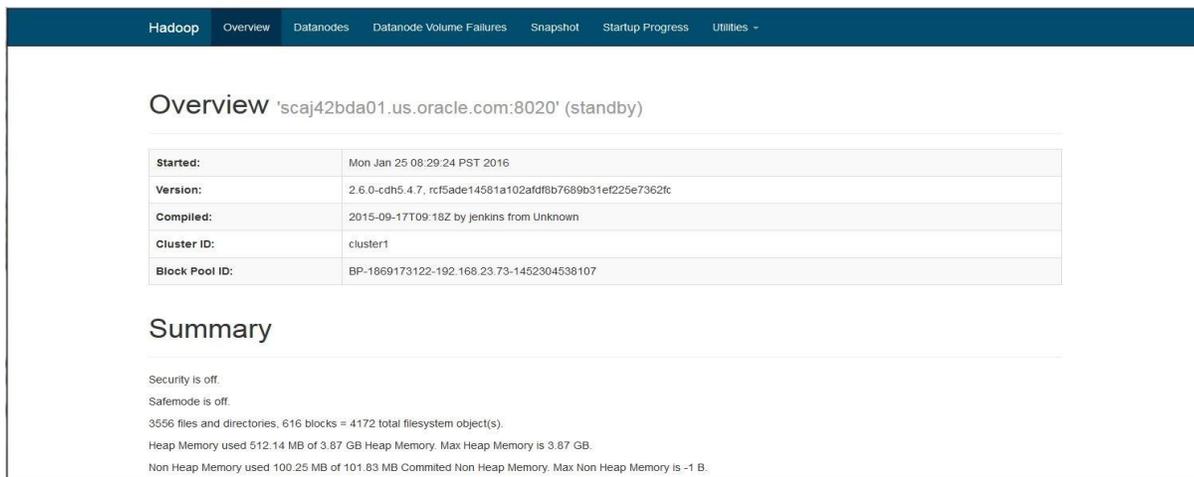


Fig 16. DFS is available for monitoring the health of the Hadoop file system

Note: Once node 3 with the MySQL primary database was restarted, SQL queries continued against Hive tables.



Failure of the Second ResourceManager and Hive Metastore Server

The objective of this test was to simulate a crash of the BDA node running the Hive Metastore service, HiveServer2 and the second ResourceManager, and to determine the operational and performance impacts of the crash. The standby ResourceManager node (node 4 in a single rack deployment,) was deliberately crashed by removing both redundant power cables from the server. This server also hosts the services listed below, which are disrupted when the server crashes.

- » Oracle Data Integrator Agent.
This service supports Oracle Data Integrator, which is one of the Oracle Big Data Connectors. You cannot use Oracle Data Integrator when the ResourceManager node is down.
- » Hive:
Hive provides a SQL-like interface to data that is stored in HDFS. Oracle Big Data SQL and most of the Oracle Big Data Connectors can access Hive tables, which are not available if this node fails. The HiveServer2 interface communicates with the Hive Metastore Server on this node.
- » Hue:
This administrative tool is not available when the ResourceManager node is down.
- » Oozie:
This workflow and coordination service runs on the ResourceManager node, and is unavailable when the node is down.

Once the server crashed, Cloudera Manager detected the failure as shown in the figure below, and reported the health of HiveServer2 and the Hive Metastore Server.



The screenshot shows the 'Health Tests' section in Cloudera Manager. It features a 'Collapse All' link and a 'Create Trigger' button. A dropdown menu is open, showing '3 concerning' items. Each item is preceded by a yellow circle icon. The items are:

- Healthy HiveServer2: 0. Concerning HiveServer2: 1. Total HiveServer2: 1. Percent healthy: 0.00%. Percent healthy or concerning: 100.00%. Warning threshold: 99.00%. [Details](#)
- Healthy Hive Metastore Server: 0. Concerning Hive Metastore Server: 1. Total Hive Metastore Server: 1. Percent healthy: 0.00%. Percent healthy or concerning: 100.00%. Warning threshold: 99.00%. [Details](#)
- Healthy WebHCat Server: 0. Concerning WebHCat Server: 1. Total WebHCat Server: 1. Percent healthy: 0.00%. Percent healthy or concerning: 100.00%. Warning threshold: 99.00%. [Details](#)

Fig 17. The *Clusters -> Hive* screen in Cloudera Manager

The health tests change from yellow to red as shown in figure 18 below.

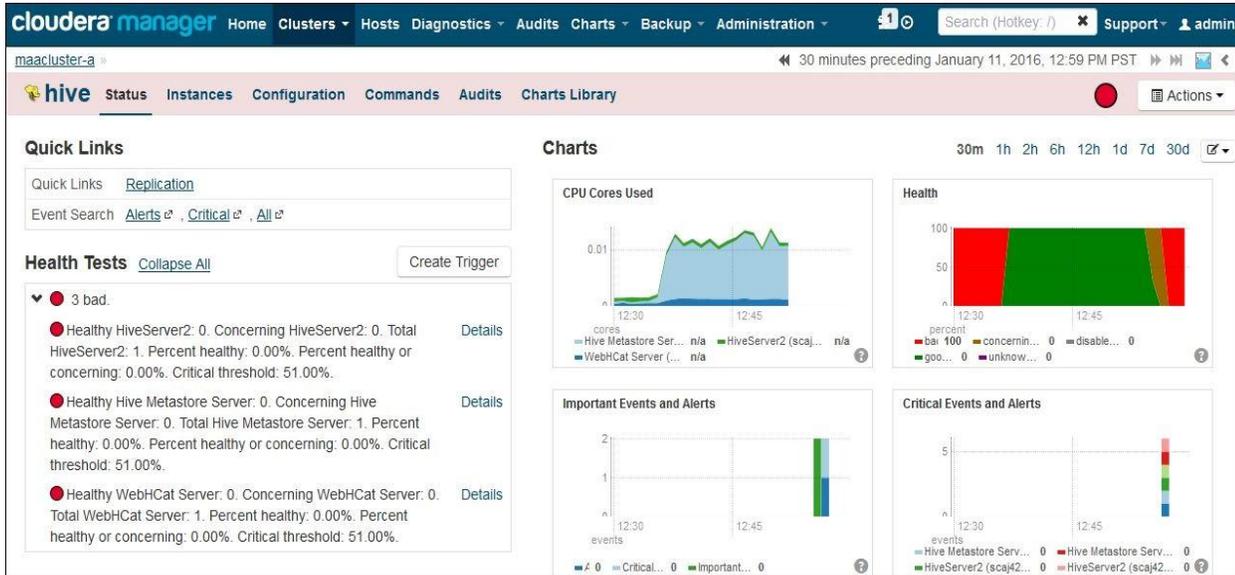


Fig 18. Hive service status in Cloudera Manager after Node 4 crashed

The “hosts” screen verifies that node 4 is down as shown below.

Status	Name	IP
●	scaj42bda01.us.oracle.com	192.168.23.73
●	scaj42bda02.us.oracle.com	192.168.23.74
●	scaj42bda03.us.oracle.com	192.168.23.75
●	scaj42bda04.us.oracle.com	192.168.23.76

Fig 19. Node 4 is verified as down within Cloudera Manager

There was an outage for accessing data in Hive tables using SQL while the Hive node was down. SQL queries in flight when the node crashed were still able to complete successfully and return data, however new SQL queries issued while the node was down threw an error similar to the one shown below.



```
SELECT order_mode,
*
ERROR at line 1:
ORA-29913: error in executing ODCIEXTTABLEOPEN callout
ORA-29400: data cartridge error
KUP-11504: error from external driver: MetaException(message:Could not connect
to meta store using any of the URIs provided. Most recent failure:
org.apache.thrift.transport.TTransportException:
java.net.NoRouteToHostException: No route to host
at org.apache.thrift.transport.TSocket.open(TSocket.java:187)
at
org.apache.hadoop.hive.metastore.HiveMetaStoreClient.open(HiveMetaStoreClient.jav
a:414)
at
org.apache.hadoop.hive.metastore.HiveMetaStoreClient.<init>(HiveMetaStoreClient.j
ava:234)
at oracle.hadoop.sql.xcat.hive.XCatHive.open(XCatHive.java:158)
```

Once the node was restarted, queries were able to continue.

Note: Access to the data via HDFS remains unaffected. MapReduce jobs continued to run successfully until completion.

InfiniBand Switch Failure

Each BDA rack contains two InfiniBand Gateway switches. The switches are fully redundant and do not present a single point of failure. The connections are evenly distributed by default between the BDA nodes and the switches, with half of the connections for each node having their primary interface connected to one leaf switch, and the other half having their primary interface connected to the other leaf switch.

Each switch also has fully redundant power supplies, and the status of each power supply can be verified from the switch command line as shown here, or using the switch ILOM.

```
# checkpower
PSU 0 present OK
PSU 1 present OK
All PSUs OK
```

The Oracle logo, consisting of the word "ORACLE" in white, uppercase, sans-serif font, centered on a red rectangular background.

The switches and network configuration are customized, validated, and tested internally by Oracle. If one of the switches were to fail, the bonded interface on each BDA node fails over to the redundant port with minimum service level impact.

The bonded InfiniBand interface on each BDA host is bondib0. The active slave is ib0 and passive slave is ib1 by default. These change once the primary InfiniBand switch fails during a switch failure test or during a real life scenario. A short brownout may be seen as traffic is redirected over the new active interface.

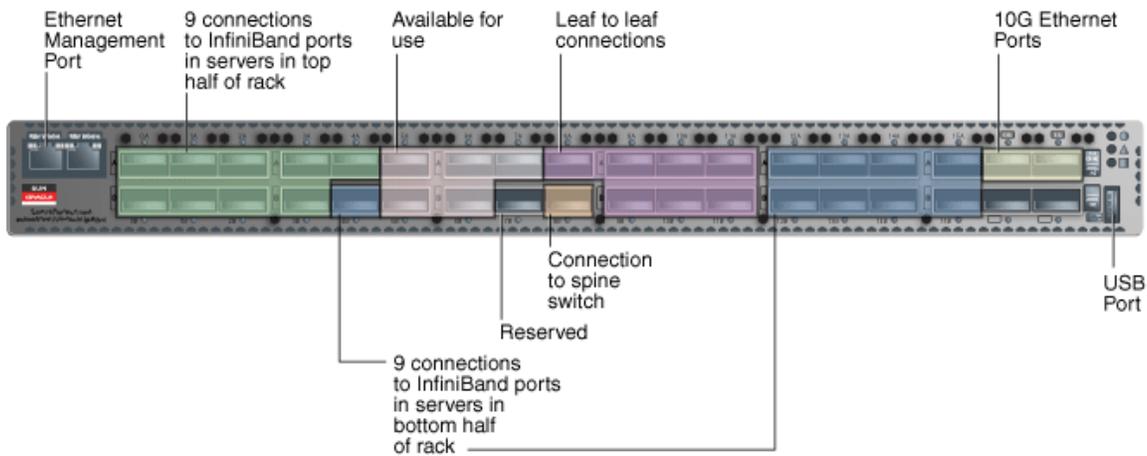


Fig 20. InfiniBand Switch Connections

The current active slave InfiniBand interface on a BDA node can be checked using the following command.

```
# cat /sys/class/net/bondib0/bonding/active_slave  
ib0
```

(also by checking /proc/net/bonding/bondib0)

By comparing output from the `ip addr` and `ibstat` commands on a BDA node, the GUID and port number can be determined, as below. This information can be used to determine the switch that BDA node port 1 is physically connected to.



```
# ip addr show ib0
7: ib0: <BROADCAST,MULTICAST,SLAVE,UP,LOWER_UP> mtu 65520 qdisc pfifo_fast master
bondib0 state UP qlen 1024

link/infiniband 80:00:00:4a:fe:80:00:00:00:00:00:00:00:00:10:e0:00:01:32:f3:b9 brd
00:ff:ff:ff:ff:12:40:1b:ff:ff:00:00:00:00:00:00:00:ff:ff:ff:ff

# ibstat
CA 'mlx4_0'
  CA type: MT4099
  Number of ports: 2
  Firmware version: 2.11.1280
  Hardware version: 0
  Node GUID: 0x0010e0000132f3b8
  System image GUID: 0x0010e0000132f3bb
  Port 1:
    State: Active
    Physical state: LinkUp
    Rate: 40
    Base lid: 12
    LMC: 0
    SM lid: 594
    Capability mask: 0x02514868
    Port GUID: 0x0010e0000132f3b9
    Link layer: IB
```

With an application load running against the Big Data Appliance, an InfiniBand switch failure was simulated by pulling both redundant power supply cables from one of the switches, causing it to fail immediately.

When an InfiniBand link goes down, the active slave will switch to "ib1", and port "ib0" connected to the failed switch will be marked as disabled. The time taken to fail over to the redundant port is determined by the `downdelay` parameter in the `BONDING_OPTS` line for the `bondib0` interface. In this case it is 5000ms (2 to 3 seconds.)

The InfiniBand Switch is shown as down within Oracle Enterprise Manager, as seen in the following Enterprise Manager screens.

The Oracle logo, consisting of the word "ORACLE" in white, uppercase letters on a red rectangular background.

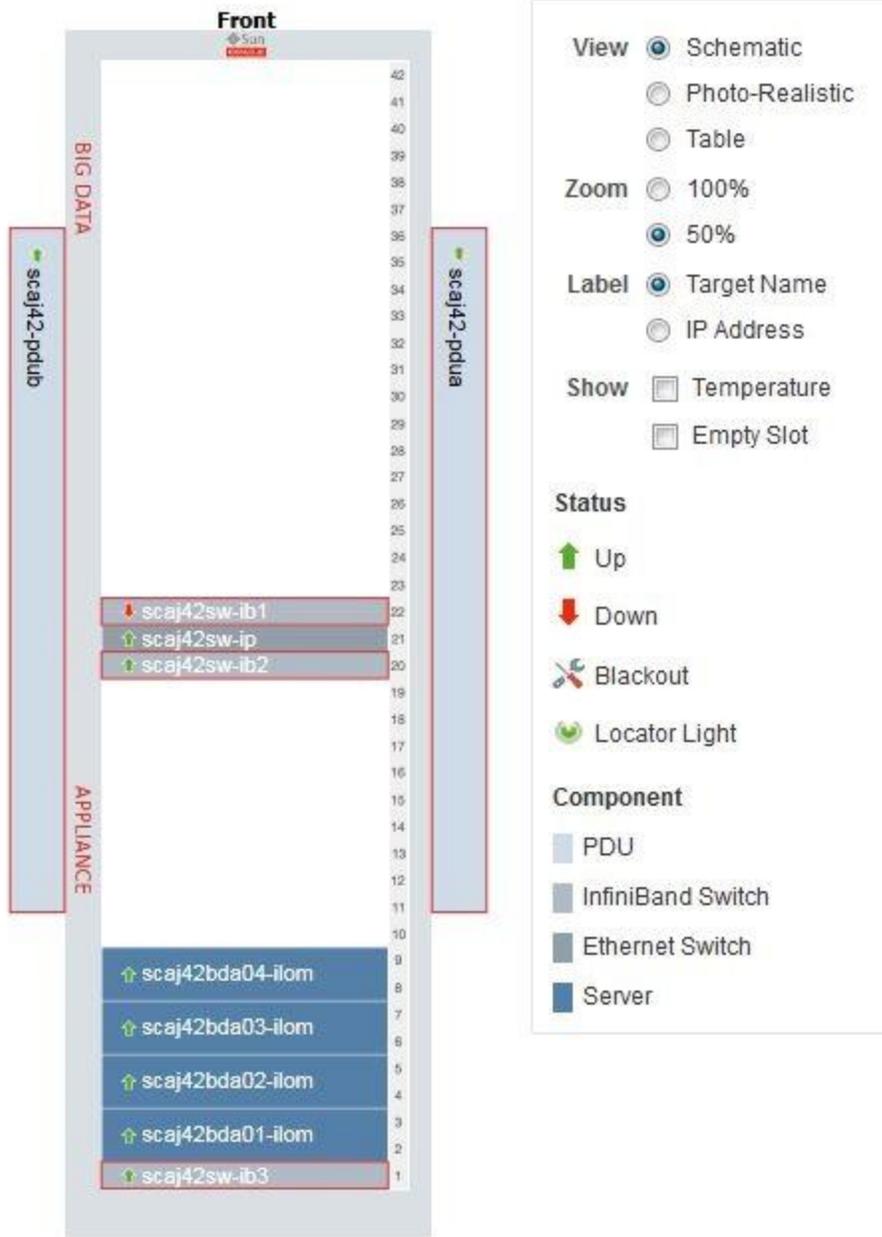


Fig 21. Enterprise Manager reports the failed switch as down



Additional information about the failed switch failure can be determined from within Enterprise Manager.

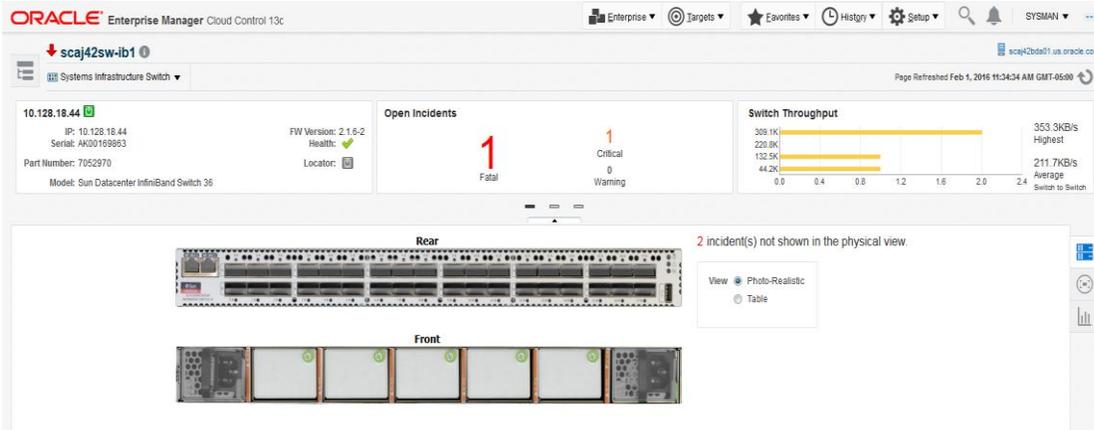


Fig 22. Enterprise Manager shows an open incident for the failed switch.

The failover can be seen on the BDA node. The active slave changed to *ib1* following the switch failure. The failed connection *ib0* is marked as down.

```
# cat /sys/class/net/bondib0/bonding/active_slave
ib1

# ibstat | egrep -i 'Port|State'

Number of ports: 2
Port 1:
    State: Down
    Physical state: Disabled
    Port GUID: 0x0010e0000132f3b9
Port 2:
    State: Active
    Physical state: LinkUp
    Port GUID: 0x0010e0000132f3ba
```

The port to switch mappings and open incidents can also be seen from within Enterprise Manager on the IBFabric screen, as shown below.

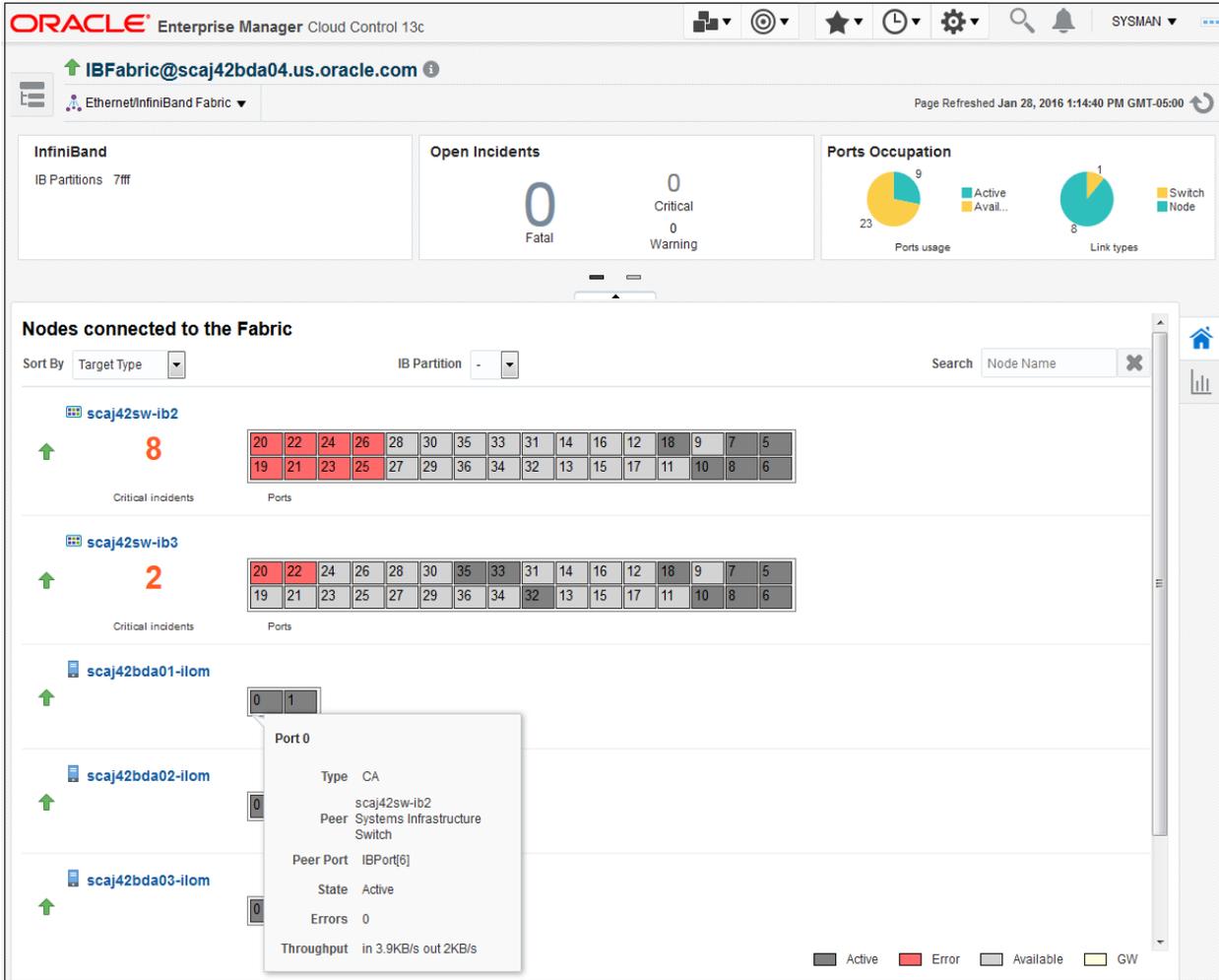


Fig 23. IB Fabric and switch to port mappings from within EM

Once the failed switch was brought back online by reconnecting the power supplies, the active port on the BDA nodes switches back to *ib0* again. From an application perspective, there was a short brownout as the passive slave became the active slave on the bonded InfiniBand interface.



Cisco Management Switch Failure

The Cisco switch features fully redundant power supplies; if one power supply fails the switch is able to continue running on the remaining power supply. The Cisco switch does not affect client application availability, and is not critical for retrieving data; however, the management interfaces are unavailable while the switch is down. The Cisco switch is also necessary for monitoring components in the rack, so alerts such as ILOM event delivery will be unavailable if the switch is down.

The management switch is not considered a single point of failure as it does not affect client operations.

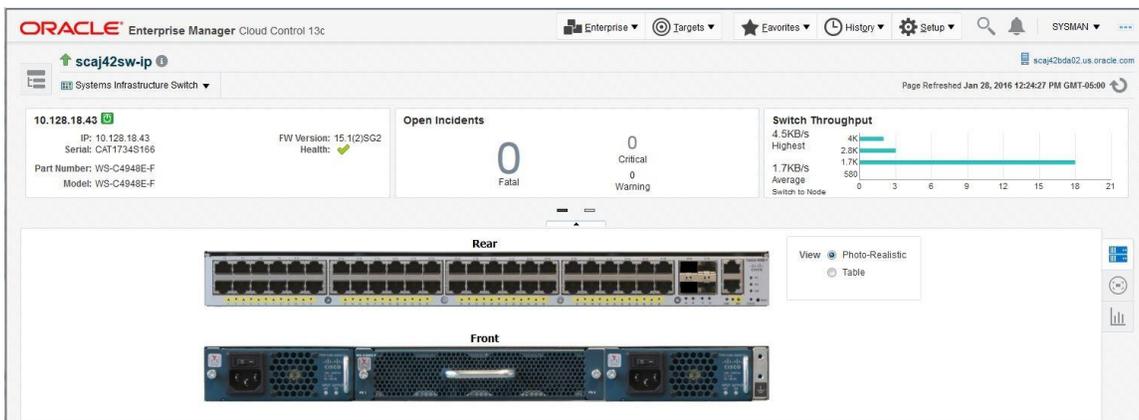


Fig 24. The Cisco switch is monitored by Enterprise Manager

As a test, power was removed from both power supplies on the Cisco switch to simulate a switch failure. The application workload continued to run, and no client impacts were seen.

Big Data SQL (BDS) Server Process Failure

The BDS Server process runs on each Big Data Appliance node, and enables high performance SQL access from Exadata to data stored on BDA nodes. The process supports smart scan capabilities such as storage indexes and predicate pushdown on the BDA nodes.

If the BDS Server process dies unexpectedly, it will be restarted automatically by the restart process that runs on the same node. Query processing will not be interrupted, and the process is restarted automatically. All results will be returned successfully.

For this test, a script was used to identify the BDS Server process and then kill it on a BDA node. An application workload was started and remained running for the duration of the test.

The running state of the BDS process can be verified from `bdscli` before any processes are killed.

ORACLE®

```
# bdscli -e list bdsq1 detail | grep bds
      bdsVersion:           OSS_PT.EXADOOP3_LINUX.X64_150912.1
      bdsq1srvStatus:      running
      bdsq1msStatus:       running
      bdsq1rsStatus:       running
```

The process ID of the BDS Server process is first identified

```
# ps -ef | grep -i 'bdsq1srv 100' | grep -v grep | awk '{print $2}'
20595
```

The short kill script contains the following line.

```
# cat bdsq1srv_crash.sh
ps -ef | grep -i 'bdsq1srv 100' | grep -v grep | awk '{print $2}' | xargs kill -11
```

The script was executed on a particular cell as follows.

```
# dcli -l root -c cellname -x bdsq1srv_crash.sh
```

Following the scripts execution, the BDS Server process was restarted within a second or two by the restart (RS) process. The process ID changed to reflect the newly restarted process as shown below.

```
# ps -ef | grep -i 'bdsq1srv 100' | grep -v grep | awk '{print $2}'
19638
```



A quick check from `bdscli` confirms the process is running again.

```
# bdscli -e list bdsqldb detail | grep bds
      bdsVersion:                OSS_PT.EXADOOP3_LINUX.X64_150912.1
      bdsqldbStatus:             running
      bdsqldbmsStatus:          running
      bdsqldbmsStatus:          running
```

Big Data SQL processes can also be started and stopped gracefully on different hosts if desired using Cloudera Manager.

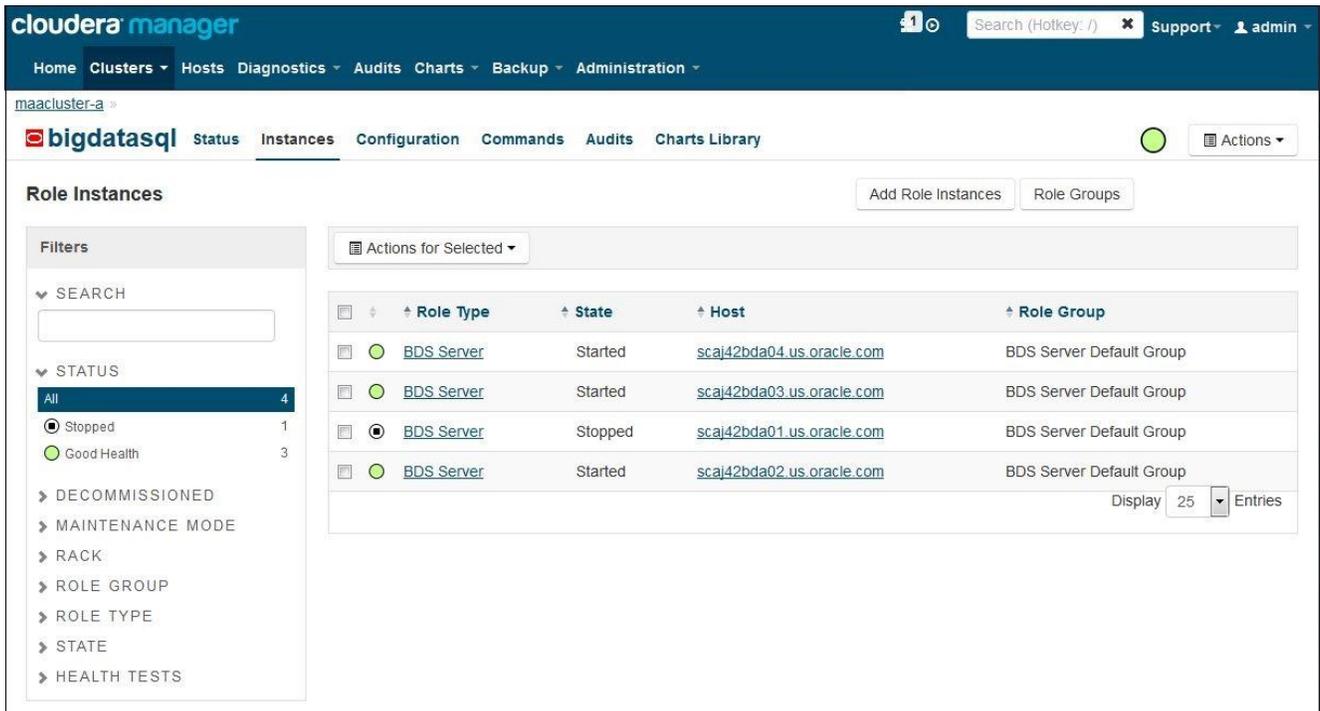
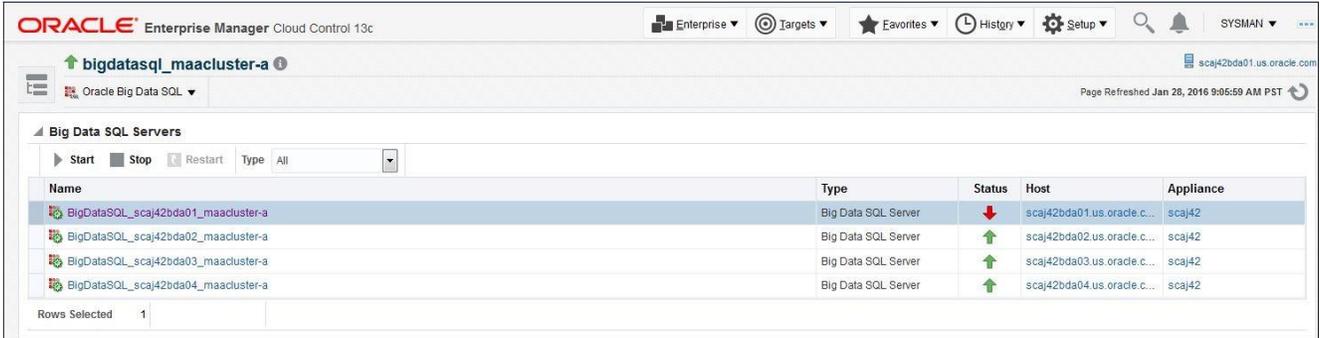


Fig 25. Cloudera Manager shows Big Data SQL Server stopped on node 1



Similar information can be viewed from Enterprise Manager.



Name	Type	Status	Host	Appliance
BigDataSQL_scaj42bda01_maacluster-a	Big Data SQL Server	↓	scaj42bda01.us.oracle.c...	scaj42
BigDataSQL_scaj42bda02_maacluster-a	Big Data SQL Server	↑	scaj42bda02.us.oracle.c...	scaj42
BigDataSQL_scaj42bda03_maacluster-a	Big Data SQL Server	↑	scaj42bda03.us.oracle.c...	scaj42
BigDataSQL_scaj42bda04_maacluster-a	Big Data SQL Server	↑	scaj42bda04.us.oracle.c...	scaj42

Fig 26. Enterprise Manager shows the status of the stopped process on node 1

Big Data SQL Server processes run on each node, and are restarted automatically if they fail.

Big Data SQL (BDS) Server Process Failure on All Nodes

In order to observe the impact and benefits of Big Data SQL, the BDS Server process was stopped on all Big Data nodes. Since the process is monitored and restarted on all nodes automatically, Cloudera Manager was used to stop the BDS service to ensure that it remained down for the duration of the test.

The impact of stopping the BDS Server process is that cell offload capabilities such as smart scan and storage indexes are not available when querying data and performance may be impacted. In this test, Big Data SQL queries initiated from Exadata ran successfully, although at a slower rate due to offloading features not being available.



Command Details: Stop Last Refreshed: Jan 28, 2016 8:51:05 AM PST ✕

Command	Context	Status	Started at	Ended at
✓ Stop	bigdatasql	Finished	Jan 28, 2016 8:50:43 AM PST	Jan 28, 2016 8:51:05 AM PST

Command completed with 4/4 successful subcommands

Child Commands

All Failed Only Active Only

Command (Child commands)	Context	Status	Started at	Ended at
✓ Stop this BDS Server	BDS Server_scaj42bda03	Finished	Jan 28, 2016 8:50:44 AM PST	Jan 28, 2016 8:51:05 AM PST
✓ Stop this BDS Server	BDS Server_scaj42bda02	Finished	Jan 28, 2016 8:50:44 AM PST	Jan 28, 2016 8:51:04 AM PST
✓ Stop this BDS Server	BDS Server_scaj42bda01	Finished	Jan 28, 2016 8:50:44 AM PST	Jan 28, 2016 8:51:04 AM PST
✓ Stop this BDS Server	BDS Server_scaj42bda04	Finished	Jan 28, 2016 8:50:44 AM PST	Jan 28, 2016 8:51:04 AM PST

[All Recent Commands](#) [Close](#)

Fig 27. BDS Server process shutdown from Cloudera Manager

Enterprise Manager confirms that all BDS Server processes are down.

ORACLE Enterprise Manager Cloud Control 13c Enterprise Targets Favorites History Setup SYSMAN

bigdatasql_macluster-a scaj42bda01.us.oracle.com

Oracle Big Data SQL

Big Data SQL Servers

Start Stop Restart Type All

Name	Type	Status	Host	Appliance
BigDataSQL_scaj42bda01_macluster-a	Big Data SQL Server	↓	scaj42bda01.us.oracle.c...	scaj42
BigDataSQL_scaj42bda02_macluster-a	Big Data SQL Server	↓	scaj42bda02.us.oracle.c...	scaj42
BigDataSQL_scaj42bda03_macluster-a	Big Data SQL Server	↓	scaj42bda03.us.oracle.c...	scaj42
BigDataSQL_scaj42bda04_macluster-a	Big Data SQL Server	↓	scaj42bda04.us.oracle.c...	scaj42

Rows Selected 1

Fig 28. Enterprise Manager also shows the Big Data SQL Server process status



Offload statistics were viewed before running SQL queries.

```
SQL> SELECT sn.name,ms.value FROM V$MYSTAT ms, V$STATNAME sn WHERE
ms.STATISTIC#=sn.STATISTIC# AND sn.name LIKE '%XT%';
```

NAME	VALUE
cell XT granules requested for predicate offload	0
cell XT granule bytes requested for predicate offload	0
cell interconnect bytes returned by XT smart scan	0
cell XT granule predicate offload retries	0
cell XT granule IO bytes saved by storage index	0

SQL queries were started, and statistics were viewed again to verify that no offloading took place.

NAME	VALUE
cell XT granules requested for predicate offload	69
cell XT granule bytes requested for predicate offload	15005982720
cell interconnect bytes returned by XT smart scan	0
cell XT granule predicate offload retries	0
cell XT granule IO bytes saved by storage index	0



BDS was then restarted on all nodes.

Command Details: Start Last Refreshed: Jan 28, 2016 8:55:13 AM PST

Command	Context	Status	Started at	Ended at
✓ Start	bigdatasql	Finished	Jan 28, 2016 8:54:51 AM PST	Jan 28, 2016 8:55:13 AM PST

Service started successfully.

Child Commands

All Failed Only Active Only

Command (Child commands)	Context	Status	Started at	Ended at
✓ Start this BDS Server	BDS Server_scaj42bda01	Finished	Jan 28, 2016 8:54:51 AM PST	Jan 28, 2016 8:55:13 AM PST
✓ Start this BDS Server	BDS Server_scaj42bda02	Finished	Jan 28, 2016 8:54:51 AM PST	Jan 28, 2016 8:55:13 AM PST
✓ Start this BDS Server	BDS Server_scaj42bda03	Finished	Jan 28, 2016 8:54:51 AM PST	Jan 28, 2016 8:55:13 AM PST

Supervisor returned RUNNING
Program: csd/csd.sh ["start"]
Recent Log Entries Links to full logs: [Stderr](#) [Stdout](#)

[All Recent Commands](#) [Close](#)

Fig 29. BDS processes restarted on all nodes using Cloudera Manager

ORACLE Enterprise Manager Cloud Control 13c

Enterprise Targets Favorites History Setup

bigdatasql_macluster-a scaj42bda01.us.oracle.com

Oracle Big Data SQL

Page Refreshed Jan 28, 2016 9:02:58 AM PST

Big Data SQL Servers

Start Stop Restart Type All

Name	Type	Status	Host	Appliance
BigDataSQL_scaj42bda01_macluster-a	Big Data SQL Server	↑	scaj42bda01.us.oracle.c...	scaj42
BigDataSQL_scaj42bda02_macluster-a	Big Data SQL Server	↑	scaj42bda02.us.oracle.c...	scaj42
BigDataSQL_scaj42bda03_macluster-a	Big Data SQL Server	↑	scaj42bda03.us.oracle.c...	scaj42
BigDataSQL_scaj42bda04_macluster-a	Big Data SQL Server	↑	scaj42bda04.us.oracle.c...	scaj42

Rows Selected 1

Fig 30. Enterprise Manager Confirms all BDS server Processes started successfully





When additional SQL queries are run, and the statistics checked again, it can be seen that offloading is working as the BDS server processes are up.

NAME	VALUE
cell XT granules requested for predicate offload	500
cell XT granule bytes requested for predicate offload	129947230208
cell interconnect bytes returned by XT smart scan	371752960
cell XT granule predicate offload retries	0
cell XT granule IO bytes saved by storage index	11005853696

The Big data SQL Server processes enable Exadata like smart scan performance on BDA nodes, and can be controlled from Cloudera Manager. Failed Big Data SQL Server processes are restarted automatically.

Entire PDU Failure on the BDA Rack

Each BDA rack features redundant Power Distribution Units (PDU's.) When a PDU fails, or is powered down, the second PDU maintains power to all components in the rack.

The objective of this test was to demonstrate that all components within the rack keep running after the failure of an entire PDU. With an application load running against the BDA, and entire PDU was turned off from within the rack to verify there were no application impacts.

PDU-A was switched off as shown in Figure 31. From within Enterprise Manager, the PDU fault was detected and raised as an alert.



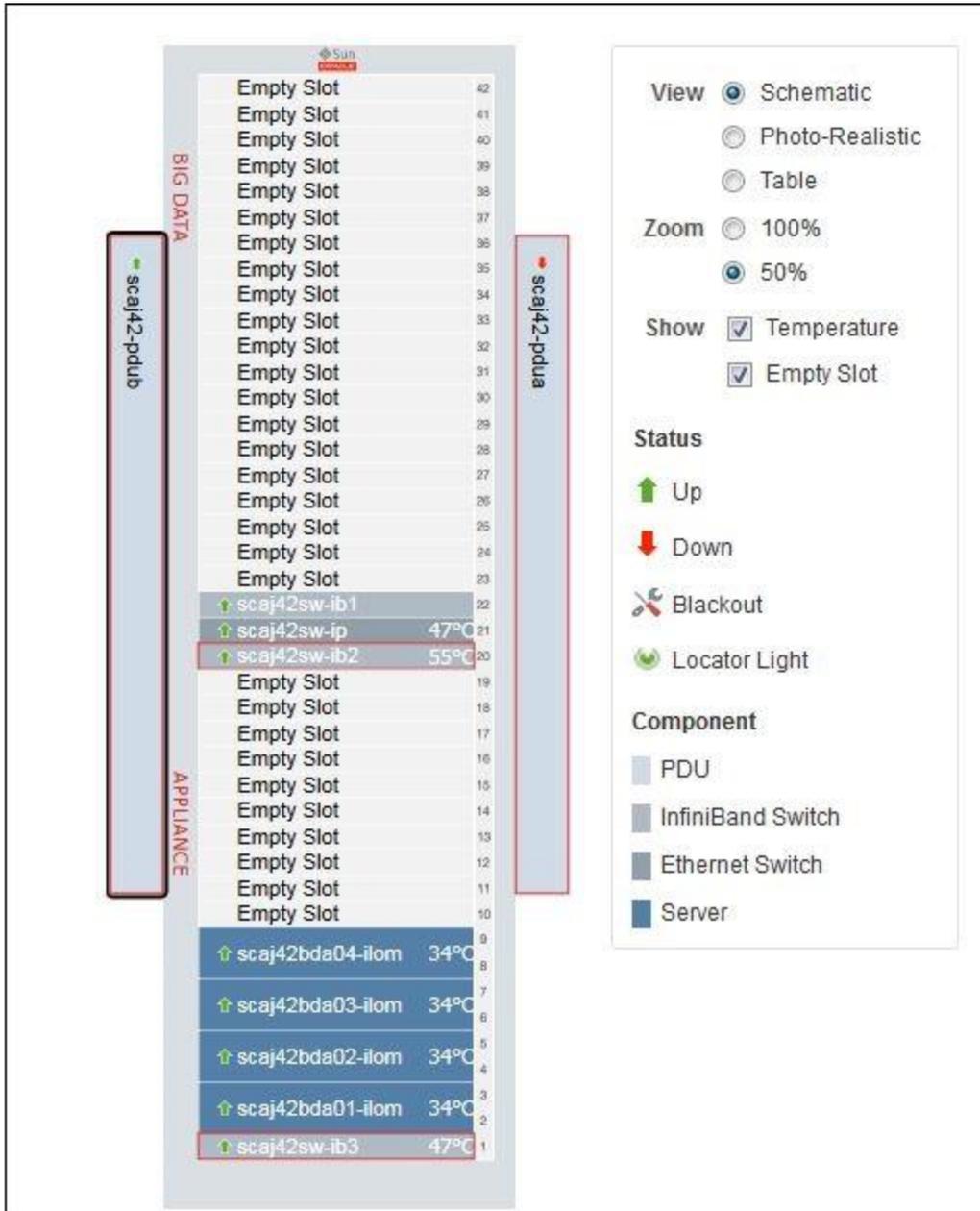


Fig 31. Enterprise Manager shows the PDU Failure



Cloudera Manager verified that all BDA nodes are still running.

Status	Name	IP	Roles	Last Heartbeat	Load Average	Disk Usage	Physical Memory
●	scaj42bda01.us.oracle.com	192.168.23.73	10 Role(s)	1.57s ago	1.00 2.52 1.51	238.4 GiB / 42.9 TiB	8.3 GiB / 62.7 GiB
●	scaj42bda02.us.oracle.com	192.168.23.74	10 Role(s)	9.45s ago	1.25 2.38 1.58	203 GiB / 42.9 TiB	9.6 GiB / 62.7 GiB
●	scaj42bda03.us.oracle.com	192.168.23.75	15 Role(s)	11.36s ago	0.83 0.81 0.80	140 GiB / 42.4 TiB	13.8 GiB / 62.7 GiB
●	scaj42bda04.us.oracle.com	192.168.23.76	10 Role(s)	14.17s ago	2.22 3.92 2.36	223.5 GiB / 42.9 TiB	10.4 GiB / 62.7 GiB

Fig 32. Cloudera Manager confirms all BDA nodes are running

The power failure was detected by the BDA nodes; however, there was no impact to the system as each component is equipped with redundant power supplies, The ILOM Logs from the BDA nodes (bdanode1 is shown below) demonstrate the power failure was detected by the ILOM.

19877	Sensor	Log	minor	Tue Jan 12 15:30:08 2016	Power Supply : /SYS/PS0/STATE : Power Supply AC lost : Asserted
19878	Fault	Fault	critical	Tue Jan 12 15:30:18 2016	Fault detected at time = Tue Jan 12 15:30:18 2016. The suspect component: /SYS/PS0 has fault.chassis.power.ext-fail with probability=100. Refer to http://www.sun.com/msg/SPX86-8003-73 for details.



Once PDU-A was powered back on, the ILOM Log from bdanode1 confirmed that power was restored.

19879	Sensor	Log	minor	Tue Jan 12 15:41:50 2016	Power Supply : /SYS/PS0/STATE : Power Supply AC lost : Deasserted
19880	Fault	Repair	minor	Tue Jan 12 15:41:50 2016	Fault fault.chassis.power.ext-fail on component /SYS/PS0 cleared
19881	Fault	Repair	minor	Tue Jan 12 15:41:50 2016	Component /SYS/PS0 repaired
19882	Fault	UUID Repaired	minor	Tue Jan 12 15:41:50 2016	Fault with UUID 95be9791-e28a-e687-e509-9304d6cf3135 repaired

BDA System Disk Failure

There are two system disks containing the operating system on each BDA node. The system disks are mirrored using Linux MD RAID-1 devices. Each disk contains a copy of the operating system, a swap partition, a mirrored boot partition, and an HDFS partition. Each node can stand the loss of one system disk without experiencing any downtime; however, the failed disk should be replaced as soon as possible to restore redundancy to the system.

If a system disk is inadvertently pulled out and then returned to its original position, the disk will not be recognized by the system and must be partitioned and reintroduced. See the *Big Data Appliance Owner's Guide* for more information on [replacing a BDA server disk](#)⁵.

To verify high availability, a BDA server system disk (disk 0) was pulled out and then pushed back into place after ten seconds. The BDA server remained online and the application workload was unaffected by the system disk failure. The state of the mirrored devices were reviewed and the appropriate action taken to restore redundancy to the system.

The state of the disk was obtained from the controller after pushing it back into place.

```
# /opt/MegaRAID/megacli/MegaCli64 pdlist a0
Firmware state: Unconfigured(bad)
Foreign State: Foreign
```

⁵ http://docs.oracle.com/cd/E69290_01/doc.44/e65664/disks.htm#BIGOG76981



The steps in the Big Data Appliance Owner's Guide were followed to change the disk status to "Unconfigured(good), Spun Up" status, and then to "Online, Spun Up."

The status of the RAID partitions can be seen with the `mdadm` command, specifying the partition number similar to the following example.

```
# mdadm -Q --detail /dev/md2
/dev/md2:
    Version : 1.1
  Creation Time : Wed Jan  6 10:39:43 2016
    Raid Level : raid1
    Array Size : 488150016 (465.54 GiB 499.87 GB)
  Used Dev Size : 488150016 (465.54 GiB 499.87 GB)
    Raid Devices : 2
  Total Devices : 2
 Persistence : Superblock is persistent
 Intent Bitmap : Internal
  Update Time : Tue Feb  2 08:11:23 2016
    State : active
 Active Devices : 2
 Working Devices : 2
 Failed Devices : 0
 Spare Devices : 0
    Name : bdanode01-adm.us.oracle.com:2
   UUID : b8d19e74:4cc75e7a:1lead157:789e1074
  Events : 29383
  Number  Major   Minor   RaidDevice State
     0     8     210         0   active sync  /dev/sdn2
     2     8         2         1   active sync  /dev/sda2
```

The high level steps to rebuild an operating system disk are:

- » Partition the Operating System Disk
- » Repair the RAID Arrays
- » Format the HDFS Partition of an Operating System Disk
- » Restore the Swap Partition
- » Restore the GRUB Master Boot Records and HBA Boot Order





See detailed information to [configure an operating system disk](#)⁶ in the *Big Data Appliance Owner's Guide* for more information.

BDA Data Disk Failure

Hadoop maintains three copies of the data across BDA nodes. The loss of a single data disk does not cause data loss as there are still two copies of the data available to read, and Hadoop automatically restores the correct number of data blocks to maintain redundancy. However, the failed drive should be replaced as soon as possible.

To simulate a hard disk failure, a data disk was pulled from a BDA server. There was no impact to the application workload caused by the disk removal.

In order to replace the failed disk, the new drive must be partitioned and formatted after it has been inserted. The high level steps for replacing a failed drive are

- » Dismount any HDFS partitions
- » Verify the Firmware State (using MegaCLI)
- » Partition the New Drive
- » Format the Partition

Refer to the *Big Data Appliance Owner's Guide* for specific details to [configure a data disk](#)⁷

The steps for a data disk replacement are similar to those in the previous section, as the newly inserted disk is treated in a similar manner by the operating system. The full list of steps can be found in the *Big Data Appliance Owner's Guide* at the link above.

⁶ http://docs.oracle.com/cd/E69290_01/doc.44/e65664/disks.htm#BIGOG76713

⁷ http://docs.oracle.com/cd/E69290_01/doc.44/e65664/disks.htm#BIGOG76719

Exadata Big Data SQL HA Tests

Oracle RAC Database Node Failure

Exadata is a part of the overall Big Data MAA Architecture, and Big Data SQL enables high performance smart scan technology to retrieve results from the BDA. Additional testing has been performed to assess the impact on Big Data SQL operations of an Oracle RAC database node crash.

The Exadata Database Machine and Big Data Appliance are connected using high bandwidth, low latency InfiniBand technology. A two node Oracle RAC system on Exadata was used for the test, and an application workload was started against the Oracle RAC database on the Exadata side, with Hive tables located on the Big Data Appliance that were accessible using external tables within the Oracle database.

A database node failure simulates the impact of hardware failures, reboots, motherboard, and component failures that can cause the system to go down unexpectedly. Existing MAA tests performed around high availability on Exadata were also leveraged for this test. The following MAA whitepaper and Oracle documentation discuss Exadata high availability.

- » [Deploying Oracle Maximum Availability Architecture with Exadata Database Machine](#)⁸
- » [Database High Availability Best Practices](#)⁹

Oracle Exadata testing ensures the hardware, firmware, and software delivered with the engineered system is tuned for optimum performance and availability. Oracle RAC technology is used on Exadata to provide high availability, and ensures that applications tolerate a complete node failure with minimal impact.

Typically, an Oracle RAC database node failure results in a brownout period for the application, as the remaining nodes wait for 30 to 60 seconds before declaring that the crashed node is dead. The waiting period is configurable using the CSS miscount parameter. However, on Exadata using Grid Infrastructure 12.1.0.2 BP7 and later, the InfiniBand network is leveraged to reduce node failure detection to just 2 seconds or less.

⁸ <http://www.oracle.com/technetwork/database/features/availability/exadata-maa-131903.pdf>

⁹ <http://docs.oracle.com/database/121/HABPT/toc.htm>

Database Node Power Failure with a Read Mostly Workload and CSS Misscount=60

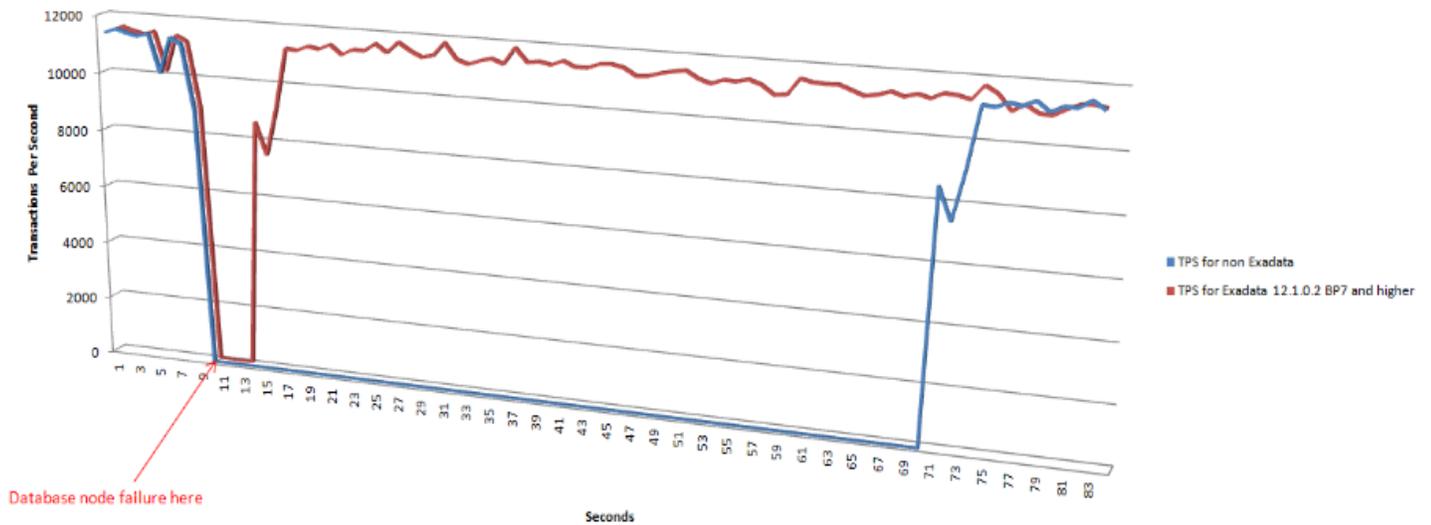


Fig 33. The reduced application brownout time of a node crash on Exadata compared to non-Exadata systems

During tests running Big Data SQL queries from Exadata against data on the Big Data Appliance; application throughput decreased and response time increased following an Exadata database node crash for a short duration of 2 seconds, and then returned to normal previous levels. There was no database outage associated with the node crash.

Oracle RAC Database Instance Failure

The end-to-end impact on Big Data SQL operations of an Oracle RAC Database instance failure has undergone considerable testing by Oracle. An application load with connections balanced across all Oracle RAC instances was the basis for the test.

The database PMON process was killed on an Oracle RAC database node to simulate an instance crash. The process was killed using a script that contained the logic shown here.



```
ps -ef | grep -i ora_pmon | grep -v grep | awk '{print $2}' | xargs kill -11
```

In an instance failure such as this, sub-second failover of connections to the surviving instance is possible. Recovery consists of a two-step process.

- (1) Oracle RAC Reconfiguration.
- (2) Instance Recovery

The surviving instance performs instance recovery for the crashed instance. The low latency InfiniBand network and the use of Write Back Flash Cache on the Exadata cells significantly reduce instance recovery time.

A short application brownout of a few seconds is expected for the connections on the failed instance. During this test, application response time increased for 2 seconds. The brownout consists of the cluster reconfiguration and instance recovery times. No database downtime was expected or experienced during this test.

BDA Cluster Resource Failure on Exadata

A Big Data SQL agent managed by Oracle Clusterware runs on each of the Exadata database nodes. The agent is registered with Clusterware during Big Data SQL installation, and the state of the agent can be verified using the following Clusterware commands.

```
# <GI_HOME>/bin/crsctl stat res -t
-----
Cluster Resources
-----
bds_maadb_maclustera
  1          ONLINE  ONLINE          exadbadm05          STABLE
  2          ONLINE  ONLINE          exadbadm06          STABLE

# <GI_HOME>/bin/mtactl check bds_maadb_maclustera
Process "extprocbds_maadb_maclustera -mt" running!
```





For high availability, there is an agent on each Exadata database node. The agent is restarted automatically by Oracle Clusterware if it should fail for any reason.

For this test, the impact on queries from Exadata Big Data SQL was assessed if the BDS agent under Clusterware control fails. An application workload was started on the Exadata and Big Data Appliance, and then a script to kill the agent was run. The script performed the following action.

```
ps -ef | grep -i extprocbds | grep -v grep | awk '{print $2}' | xargs kill -11
```

The immediate impact of killing the `extproc` process is that new SQL queries started to fail as shown below. These failures occurred in SQL executed after the BDS agent was killed, whereas queries in flight continued to process data successfully.

```
SELECT SUM(amount_sold),
*
ERROR at line 1:
ORA-29913: error in executing ODCIEXTTABLEOPEN
callout
ORA-29400: data cartridge error
ORA-28575: unable to open RPC connection to external
procedure agent

Elapsed: 00:01:02.90""
```

It is expected that Clusterware will restart the Big Data SQL agent in about 10 to 20 seconds, and queries in flight will continue. This was validated in testing, and queries in flight completed successfully. SQL executing on other Exadata nodes where the Big Data SQL agent was still running were unaffected.



Conclusion

Oracle Big Data MAA and the Big Data Management System provide the most comprehensive, integrated, and highly available solution available for Big Data. Proven high availability capabilities are delivered pre-configured with every Oracle Big Data Appliance and Oracle Exadata system, and can be deployed both on-premise and in the cloud with Oracle's Big Data Cloud Service.

When the Oracle Big Data Appliance and Exadata platforms are integrated to form a Big Data Management System, Oracle Big Data SQL technology delivers high performance access to data, and enables the full power of Oracle SQL to provide a unified view across Oracle Database, Hadoop, and NoSQL sources.



ORACLE®

Appendix A

MAA Test Scenarios Quick Reference

Failure Test Scenario	Simulation Process	Application Impact and Duration
Active NameNode	Removed the Redundant Power Supply Cables from the Server	<p>Short duration brownout of typically under one minute. The duration depends on the amount of time it takes for the standby NameNode to assume the Active NameNode role.</p> <p>Additional services on this node are unavailable; however they do not affect the availability of the cluster.</p> <p>All queries continued and returned data once the standby NameNode assumed the active role. No data errors were observed.</p>
Active NameNode with service Migration	Removed the Redundant Power Supply Cables from the Server	<p>There is a short brownout as described above while the standby NameNode becomes Active.</p> <p>If the failed server cannot be returned to duty promptly, services can be migrated to a non-critical node using the</p> <pre>bdacli admin_cluster migrate</pre> <p>syntax.</p>
Active NameNode and Standby NameNode	Removed the Redundant Power Supply Cables from the Server	<p>After power is pulled from the first NameNode, the previous standby NameNode takes over and processing continues.</p> <p>Once power is also pulled from the remaining NameNode all processing stops in the cluster.</p> <p>After power is restored to a NameNode it resumes as the active NameNode and client processing continues.</p>



Failure Test Scenario	Simulation Process	Application Impact and Duration
First ResourceManager, Cloudera Manager and MySQL Database	Removed the Redundant Power Supply Cables from the Server	<p>The second ResourceManager takes over if the first ResourceManager fails.</p> <p>Big Data SQL queries that access Hive tables on the BDA stopped running due to Hive Metastore metadata being unavailable, as it is stored in the MySQL database that runs on this node.</p> <p>Cloudera Manager was also down due to this node failure.</p>
Second ResourceManager and Hive Metastore Server	Removed the Redundant Power Supply Cables from the Server	<p>If the Second ResourceManager fails, client access continues, and the first ResourceManager continues without a backup.</p> <p>Big Data SQL queries accessing Hive tables stopped running due to the Hive Metastore Server and HiveServer2 interface being unavailable. Existing queries in flight did complete successfully, new Hive queries were unable to run.</p> <p>Access to data through HDFS is unaffected.</p> <p>Hue, Oozie, and ODI are also unavailable.</p>
InfiniBand Switch	Removed the Redundant Power Supply Cables from the Switch	The switches are fully redundant and do not present a single point of failure. Each BDA server has connections to both switches. A brownout of 2 to 3 seconds may be seen as traffic is redirected over the active interface.
Cisco Management Switch	Removed the Redundant Power Supply Cables from the Switch	The management switch is not considered a single point of failure as it does not affect client operations.
Big Data SQL (BDS) Server Process	Killed the Server Process	There is a BDS Server process on each node. It is restarted automatically in case of failure.
Big Data SQL (BDS) Server Process Failure on All Nodes	Shutdown the BDA Server Process on all Nodes with Cloudera Manager	Cell offload capabilities such as smart scan and storage indexes are not available; however, there is no outage and query processing continues.

Failure Test Scenario	Simulation Process	Application Impact and Duration
Entire PDU	Turned off Power to one of the PDUs	Each BDA rack contains redundant PDUs. Every BDA rack component has a connection to each PDU, ensuring there is no impact if a PDU fails.
BDA System Disk Failure	Pulled a System Disk (disk 0 or 1), wait 10 seconds and then pushed the same disk back into the slot	The BDA system disks are fully redundant, using RAID mirroring. There is no impact to the system if a system disk is removed, although redundancy is reduced. If a disk is pushed back into the same slot, the RAID device and disk partitions must be recreated.
BDA Data Disk Failure	Pulled a Data Disk (disks 2 to 11) and then replaced it with a brand new disk	HDFS maintains a replication factor of three across BDA nodes. The loss of a single data disk will not cause any data loss, as there are still two copies of the data available to read. Client operations will continue without interruption. Hadoop automatically restores the correct number of data blocks to maintain redundancy.
Exadata Database Node Failure	Removed the Redundant Power Supply Cables from the Server	An Oracle RAC Database node failure will result in a short brownout period for the application according to cluster detection, reconfiguration and instance recovery. The remaining nodes wait for 30 to 60 seconds before declaring that the crashed node is dead, according to the CSS miscount parameter on non-Exadata environments. On Exadata with Grid Infrastructure 12.1.0.2 BP7 and higher, the InfiniBand network is leveraged to reduce the detection time to 2 seconds or less. Application throughput reduced and response times increased for 2 seconds.



Failure Test Scenario	Simulation Process	Application Impact and Duration
Exadata Oracle RAC Database Instance Failure	Killed the Database PMON Process with a script to simulate an instance failure	<p>Sub-second failover of connections to the surviving instance is possible. A short application brownout of a few seconds is expected for the connections on the failed instance. The brownout will consist of the cluster reconfiguration and instance recovery times. No database downtime was expected, or experienced during this test.</p> <p>The low latency InfiniBand network and the use of Write Back Flash Cache on the Exadata cells significantly reduce instance recovery time.</p>
Exadata Failure of the BDA Clusterware resource	Killed the Big Data SQL agent running under Oracle Clusterware control on a database node.	<p>There is an agent on each Exadata database node.</p> <p>The agent is restarted automatically by Oracle Clusterware if it should fail for any reason. Clusterware will restart the agent in 10 to 20 seconds.</p> <p>Query processing will continue once the agent has fully restarted.</p> <p>SQL queries on other Exadata database nodes are unaffected.</p>





Oracle Corporation, World Headquarters

500 Oracle Parkway
Redwood Shores, CA 94065, USA

Worldwide Inquiries

Phone: +1.650.506.7000
Fax: +1.650.506.7200

CONNECT WITH US

-  blogs.oracle.com/oracle
-  facebook.com/oracle
-  twitter.com/oracle
-  oracle.com

Integrated Cloud Applications & Platform Services

Copyright © 2016, Oracle and/or its affiliates. All rights reserved. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle and Java are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD, Opteron, the AMD logo, and the AMD Opteron logo are trademarks or registered trademarks of Advanced Micro Devices. UNIX is a registered trademark of The Open Group. 0615

Oracle Big Data Appliance: Oracle Maximum Availability Architecture

March 2016

Authors: Richard Scales, Lingaraj Nayak

Contributors: Jean-Pierre Dijcks, Martin Gubar, Ravi Ramkissoon

