**Data Lake - Hadoop**

**Technical Design Specification**

**Market Requestor: *Michael Klein***

**Clarity Effort Id:** ***E0014535***

**CIT Intake Request:** [***Intake Requests: Data Lake POC***](https://itcomposites.lmig.com/Sites/CITIntake/Lists/Intake%20Requests/Item/displayifs.aspx?List=9c1936d9%2Dd89a%2D414c%2D9645%2D09ce19af63a2&ID=1410&ContentTypeId=0x01001C6E8100F55625479C734F30B39AFEB6)

**ETS Project Manager: Bonnie Rollins-Dumont**

**ETS Solutions Architect: Dan O’Sullivan**

**ETS Technologist: Claude Langlois**

**HS MRS Lead Engineer: John McFarland**

**HS MRS Engineering Manager: Elaine Landry**

…

**Table of Contents**

[Design Objective 3](#_Toc435175897)

[Design Requirements 3](#_Toc435175898)

[In Scope 3](#_Toc435175899)

[Hosting Services 4](#_Toc435175900)

[Out of Scope 4](#_Toc435175901)

[Architecture Design 4](#_Toc435175902)

[Node Configurations 4](#_Toc435175903)

[Network Requirements 6](#_Toc435175904)

[Implementation Steps 9](#_Toc435175905)

[Non-Production 9](#_Toc435175906)

[Production 10](#_Toc435175907)

[Security Requirements 10](#_Toc435175908)

[**Operational Support** 10](#_Toc435175909)

[**Operating System** 10](#_Toc435175910)

[Backup and Recovery 11](#_Toc435175914)

[Lifecycle Support 11](#_Toc435175915)

[Financial 11](#_Toc435175916)

[Appendix: 12](#_Toc435175917)

[Rack Configurations 12](#_Toc435175918)

[Hardware Monitoring 15](#_Toc435175919)

[References: 17](#_Toc435175920)

# Design Objective

A Data Lake is a centralized repository of large volumes of structured and unstructured data built on relatively inexpensive computer hardware. Hadoop, an open-source framework of tools for processing and analyzing big data, can be used to sift through the data in the repository. Hortonworks is a layered product that was selected to implement the Hadoop framework. The objective of this design is to provide a Non-Prod Hadoop cluster running Hortonworks Data Platform to implement a Hadoop based Data Lake for ETS Enterprise Platform & Services. The goal is to deliver a high performance, near real-time business analytics for Corporate data consumers such as Legal, Finance, and Enterprise Risk Management.

**This design is only for a limited, non-production deployment. Expansion into production or growth beyond 15 additional data nodes will invalidate this design. A production implementation will require another design review.**

# Design Requirements

## In Scope

* Hadoop cluster capable of MPP data analysis
* Hortonworks distribution for a Hadoop architecture
* Use published Hortonworks reference architectures to ensure a consistent performance benchmark comparison
  + x86 servers running RHEL6.6
  + DAS for Management Node, Name & Resource Node
  + DAS for Data Node configured as JDOB
  + 10GbE network connectivity
  + Dedicated Network TOR switches with isolated Data Nodes
  + Dual-homed Management Node
* Backups required for Management Nodes only
* Non-prod only, there is no production at this time
* Data Redundancy at cluster level is managed by the Hortonworks application stack
* Hadoop cluster is latency sensitive
  + Reduce number of racks
  + Keep racks close together
  + Use internal disk drives to reduce the data path
  + Use same VLAN for all Data Nodes
* Performance Requirements
  + Directly related and proportional to number of data nodes in cluster
  + Usable Data size 24TB (using default replication factor of 3 plus 1x for work space and usage <80%)
  + Cluster Processing Time: process a 10TB dataset in 4.5 hours
* Multi-tenancy of data and business users

## Hosting Services

* Standard, supported OS that is part of Roadmap
* Fully monitored systems that follow current processes and procedures
* OS and supporting management applications installed and configured per standard processes and procedures
* Delivered systems are managed by all Hosting Services teams: ID management, Security Compliance, Intrusion Detection, Asset Management, Operations and Lifecycle

## Out of Scope

* Hortonworks application installation, support and management including any application repositories
* Data Hub project/effort details and other initiatives such as Splunk, ELKstack, or EIS – Enhanced Visibility Program project
* Other cloud deployment options
* Production deployment design requirement

# Architecture Design

The following section outlines the architecture necessary to support the overall design objective. The focus of this design is to provide an **8-node non-prod cluster** in a single rack that is considered a POC ‘limited release’ deployment. A separate design would be required for a Production cluster which would be a multi-rack configuration. In order to meet the requirement to be able to compare the Hadoop cluster performance to published benchmarks, reference architectures must be used. This design is based on the [HP Verified Reference Architecture for Hortonworks HDP 2.2 on HP ProLiant DL380 Gen9 with RHEL](http://hortonworks.com/wp-content/uploads/2013/10/4AA5-9017ENW.pdf) The only modification is to use LM server standard of a DL380 is preferred over the suggested DL360 for Management Nodes. Hosting Services – Midrange will deliver the 8 physical servers built per the current standards for RHEL 6.x. These are fully managed systems that meet all Hosting Services operational, support, security and procurement requirements.

### Node Configurations

A Hadoop cluster is comprised of two types of systems, Management Nodes and the Data Nodes. Three Management Nodes are the minimum number required for a Hadoop cluster. Based on the amount required storage and computer, d five Data Nodes are requested for this effort. The three Management Nodes consist of a Name Node, Resource Manager and a cluster Management Node. These systems contain metadata and cluster configuration data. This unique information requires backups. The Name Node is a well-known single point of failure in a Hadoop cluster. If this node fails, the entire cluster ability to process data could be impacted. Therefore to help improve management nodes reliability, each system will have two RAID created, RAID 1 for the OS and RAID 5 for remaining disks.

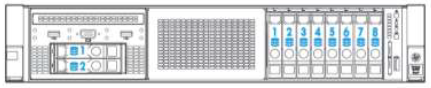
The HP reference architecture recommends using LFF drives instead of the normal SFF drives (same angular speed). LFF drives have better performance than the SFF drives due to faster tangential speed which results in higher disk I/O. The reference architecture uses internal disk drives (DAS) configured without any RAID (Just a Bunch of Disks – JBOD).

* HP DL 380 Gen9, dual 8-core Intel E5-2640v3, 128GB RAM, 8 x 900GB SAS disks for the Management Nodes
* HP DL 380 Gen9, dual 10-core Intel E5-2660v3, 256GB RAM, 12 x 2TB SATA disks for the Data Nodes

Management Node Configuration:

The Management Nodes system type is a DL380 which differs from the reference architecture. The reference architecture calls for DL360 which is not our standard HP model.

The overall usable amount of non-OS storage will be a single RAID 5 group yielding 5.8TB.

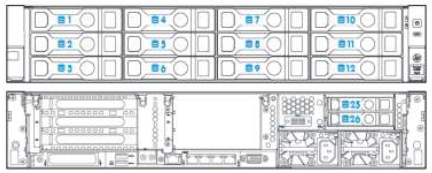


**Management Node = 128 GB RAM, 8x900GB SFF Drives, 2x300GB SFF Drives for OS**

| **Ln #** | **Qty** | **Mfg Part #** | **Description** |
| --- | --- | --- | --- |
| 1 | 1 | 719064-B21 | HP DL380 Gen9 8SFF CTO Server |
| 2 | 1 | 719049-L21 | HP DL380 Gen9 E5-2640v3 FIO Kit |
| 3 | 1 | 719049-B21 | HP DL380 Gen9 E5-2640v3 Kit |
| 4 | 8 | 726719-B21 | HP 16GB 2Rx4 Pc4-2133P-R Kit |
| 5 | 1 | 768857-B21 | HP DL380 Gen9 8SFF Cage Bay2/Bkpln Kit |
| 6 | 2 | 785067-B21 | HP 300GB 12G SAS 10K 2.5in SC ENT HDD |
| 7 | 8 | 652589-B21 | HP 900GB 6G SAS 10K 2.5in SC ENT HDD |
| 8 | 1 | 719073-B21 | HP DL380 Gen9 Secondary 3 Slot Riser Kit |
| 9 | 1 | 761874-B21 | HP Smart Array P840/4G FIO Controller |
| 10 | 1 | 783009-B21 | HP DL380 Gen9 8SFF SAS Cable Kit |
| 11 | 2 | 652503-B21 | HP Ethernet 10Gb 2P 530SFP Adptr |
| 12 | 1 | 733660-B21 | HP 2U SFF Easy Install Rail Kit |
| 13 | 2 | 720478-B21 | HP 500W FS Plat Ht Plg Pwr Supply Kit |
| 14 | 1 | 512487-B21 | HP iLO Adv Track incl 1yr TS/U SW |
| 15 | 1 | QL803B | HP Insight CMU 1yr 24x7 Flex Lic |

Data Node Configuration:

The Data Nodes require large amounts of local disk. HP recommends the use of the Large Form Factor (LFF) drives.



**Data Node = 256GB RAM, 12 x 2 LFF SATA and 2 x 300GB SFF SAS Hot Pluggable Rear Drive Bays**

| **Ln #** | **Qty** | **Mfg Part #** | **Description** |
| --- | --- | --- | --- |
| 1 | 1 | 719061-B21 | HP DL380 Gen9 12LFF CTO Server |
| 2 | 1 | 762764-L21 | HP DL380 Gen9 E5-2660v3 FIO Kit |
| 3 | 1 | 762764-B21 | HP DL380 Gen9 E5-2660v3 Kit |
| 4 | 8 | 728629-B21 | HP 32GB 2Rx4 PC4-2133P-R Kit |
| 5 | 1 | 724864-B21 | HP DL380 Gen9 2SFF Bay Kit |
| 6 | 12 | 628061-B21 | HP 2TB 6G SATA 7.2k 3.5in |
| 7 | 2 | 785067-B21 | HP 300GB 12G SAS 10K 2.5in SC ENT HDD |
| 8 | 1 | 719073-B21 | HP DL380 Gen9 Secondary 3 Slot Riser Kit |
| 9 | 1 | 700751-B21 | HP FlexFabric 10Gb 2P 534FLR-SFP+ Adptr |
| 10 | 1 | 761874-B21 | HP Smart Array P840/4G FIO Controller |
| 11 | 1 | 783007-B21 | HP DL380 Gen9 P840/440 SAS Cable Kit |
| 12 | 1 | 785991-B21 | HP DL380 Gen9 12LFF SAS Cable Kit |
| 13 | 2 | 652503-B21 | HP Ethernet 10Gb 2P 530SFP Adptr |
| 14 | 1 | 733662-B21 | HP 2U LFF Easy Install Rail Kit |
| 15 | 2 | 720479-B21 | HP 800W FS Plat Ht Plg Pwr Supply Kit |
| 16 | 1 | 512487-B21 | HP iLO Adv Track incl 1yr TS/U SW |
| 17 | 1 | QL803B | HP Insight CMU 1yr 24x7 Flex Lic |

Due to future potential cluster growth and additional inter-rack latency it is recommended to use a 3 cabinet pod design. A benefit to the pod design is the additional available cooling which means more nodes can be put into one rack.

# Network Requirements

The single rack configuration required for this effort consists of a pair of 10GbE TOR switches for Hadoop VLAN isolation and a single 1GbE switch for iLO/management connectivity. The data nodes will be placed on a firewall protected VLAN instead of a private, non-routable VLAN. The standard set of LM pre-filters will be applied so the systems can be managed properly.

**Initial Non-Prod Hadoop- Datalake network Design:**



# Implementation Steps

The following section outlines the steps necessary to implement an 8-node Hadoop cluster in Non-Prod. The reference architecture requires the use of a non-standard HP DL380 configuration; a large disk form factor to accommodate the additional internal disk drives. The hardware, OS and management software installation will follow as many of the current Hosting Services – Midrange practices and procedures as possible. The data nodes or worker nodes require the use of non-Raid protected data drives. The configuration and operational support practices and procedures need to be modified to meet this requirement.

## Prerequisite Tasks

* Task-1 Network design including TOR switches and VLANs information completed.
* Task-2 All network components and configuration - including TOR switches and VLAN definitions is completed. New VLANs must be available on LIN.
* Task-3 Security Operations and Perimeter Control have reviewed the network and compute design and have signed off.
* Task-4 Standard LM Firewall Pre-filters have been applied to the new secure VLANs.

## Non-Production

* Step-1 Rack and connect both management and data nodes per standard procedures referring to the provided racking diagrams along with standard iLO connection. Reserve rack space in datacenter to handle 10 more data nodes.
* Step-2a Data node: Determine and document how to configure the internal disks so the 300GB disks are hardware mirrored for the OS and the remaining disks do not have any RAID
* Step-2b Management node: Determine and document how to configure the internal disks so the two 300GB drives are mirrored (RAID 1) for the OS and the 900GB are configured to be a RAID 5 set.
* Step-3 standard RHEL 6.6 installation per PDP build package on the mirrored 300GB disk only
* Step-4a Data Nodes: Configure all data drives for JBOD (each drive is a RAID 0), mounted without LVM as /grid/1 … /grid/12 Recommended mount options

**mkfs -t ext4 -m 1 -O dir\_index,extent,sparse\_super /dev/sdXN**

* Step-4b Management Nodes: Configure the RAID 5 set of disks into on LVM file system
* Step-5 Configure OS per Hortonworks, see section 1.2 in the following document <http://docs.hortonworks.com/HDPDocuments/Ambari-2.1.2.0/bk_Installing_HDP_AMB/bk_Installing_HDP_AMB-20150930.pdf>
* Step-6 Install and configure HP Cluster Management Utility
* Step-7 Determine and document on how to detect and send disk drive errors to standard monitoring and alerting applications, this may include HP OneView integration

[Adding a DL Server to HP OneView](https://itcomposites.lmig.com/Sites/Midrange/Integration/InternalDocuments/Project%20Documents%20and%20Deliverables/Data%20Lake%20-%20Hadoop/Adding%20a%20DL%20Server%20to%20HP%20OneView.docx)

* Step-8 Develop and document procedure on how to replace failed JDOB disks, including removing disks to test alerting. Please refer to the Hardware Monitoring in the Appendix for more details.
* Step-9 Develop and document a procedure on how perform QA in systems with DAS

Use at running OS to verify /opt/compaq/hpacucli/bld/hpacucli ctrl all show status

* Step-10 Perform OS level tuning for internal disk followed by running disk IO benchmarks using concurrent dd commands or vendor recommended tools.
* Step-11 Turn over systems to Requester/ETS for additional loading and configuring of Hortonworks
* Step-12 Document the list of missing system based RPMs required by Hortonworks software stack
* Step-13 If needed, modify the RHEL installation to include the custom list of Hortonworks pre-requisite RPMs
* Step-14 Coordinate with ETS any specific monitoring parameters and requirements associated with the Hortonworks applications, for example Ambari
* Step-15 Determine if HP Call-home if beneficial and how to implement

## Production

Not in scope, but would follow same process with the exception/addition of additional rack/network configurations for a multi-rack (prod) design.

# Security Requirements

Data set insolation is important to provide multi-tenancy, otherwise more data nodes will be required to isolate the use-cases at a physical layer. Security Operations needs to be consulted to determine if the standard enterprise tools can be used. Security Operations also needs to be consulted to determine their requirements for managing the data nodes if these are on a private, non-accessible network.

**Operational Support**

**Operating System**

* Red Hat(RHEL) 6.6 systems, with a specific list of RPMs as pre-requisites for Hortonworks HDP V2.3 application stack
* OS drives use RAID 1, data disks are JBOD (no RAID) and each has its own mounted file system
* No redundancy at the node level, redundancy is created at cluster level

**Server Platform Support & Node-level Redundancy – A** Hadoop cluster ensures that a certain number of block copies are consistently available at the cluster level. This number is configurable in the block replication factor setting, which is typically set to three. If a Hadoop worker node goes down, Hadoop will replicate the blocks that had been on that server onto other servers in the cluster to maintain the consistency of the number of block copies. Using a JBOD (Just a Bunch of Disks) configuration and node-level redundancy, eliminates the need for a traditional RAID level on the data drives. This introduces a single point of failure that creates operational instability. Processes and procedures need to be modified to mitigate this risk.

* The pair of OS drives will be mirrored in a traditional RAID 1 configuration for the OS.
* The expectation is that these DL380 Gen9 servers can be monitored via HP OneView for component level failures for standard SNMP alert generation. This is relevant when considering the number of HDDs used.
* Failed drives would be replaced in standard platform support process with the possible addition of re-adding the new drive in the ACU (Array Configuration Utility) on the server. The specific steps are need to be determined.

**Monitoring Hadoop Functionality & Alerts**

* Hadoop has its own monitor via Apache Ambari. Ambari can monitor at the infrastructure level and at the application level. There is the possibility Ambari monitoring will at the hardware level will be required. This will require additional training as this is a new management tool.
* The standard Midrange monitoring alerting tools will be used. There is the possibility integration with Ambari needs to be determined. This will require additional training and implementation.

**Capacity Management –** The standard, single system capacity and performance tools do not operate on a Hadoop cluster level. This level of capacity management must be done with Hortonworks tools. The specifics around Capacity Management (when to add nodes) needs to be determined. The assumption is there will be an agreed upon process for determining capacity expansions will be needed. The current design does not require HS-Midrange to learn or use the Hortonworks management tools, however, this is subject to change.

# Backup and Recovery

The management nodes require normal backups. The data nodes do not store any unique data plus the cluster keeps replica copies to protect against hardware failures. It will be incumbent on application teams to insure that input data is properly backed up in the event data needs to be reconstructed.

# Lifecycle Support

Standard hardware and OS software lifecycle support will be used. However, because Hortonworks applications are Open Source based there is the potential for an increased frequency of OS upgrades or individual OS library upgrades that falls outside of the Midrange Roadmaps.

# Financial

The table lists the hardware based on HP’s published reference architecture for a Hortonworks Hadoop Cluster. We are still in negotiations with HP to procure a consulting contract to assist the installation and support of this new type of clustering and data processing model.

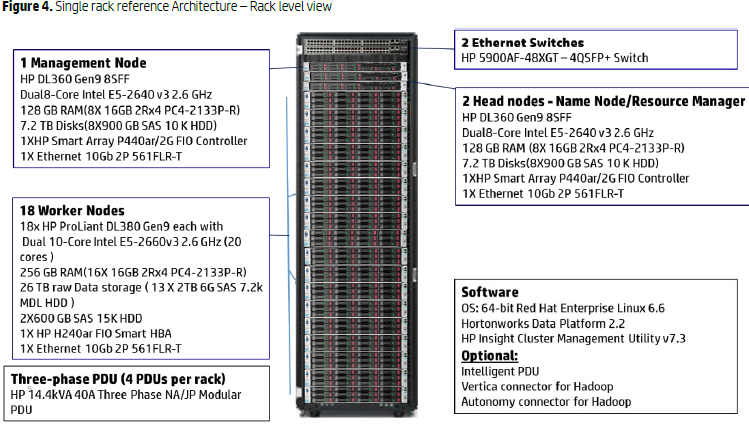
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PDC** | Quantity | Config | Notes | USD Cost |
|  | 5 | DL380 20CPU 2.6GHz/256GB Memory  12 x2TB SATA disks | Worker/data nodes  Connected to management nodes only | $   75,470 |
|  | 3 | DL380 16CPU 2.6GHz/128GB Memory  8 x900GB SAS disks | Management nodes  Connected to LIN  Connected to data nodes | $   37,263 |
|  |  |  | **Grand Total** | **$  112,733** |

The usable amount of storage on the data nodes is 6TB per system at an approximate cost of $2500 per TB.

# Appendix:

# Rack Configurations

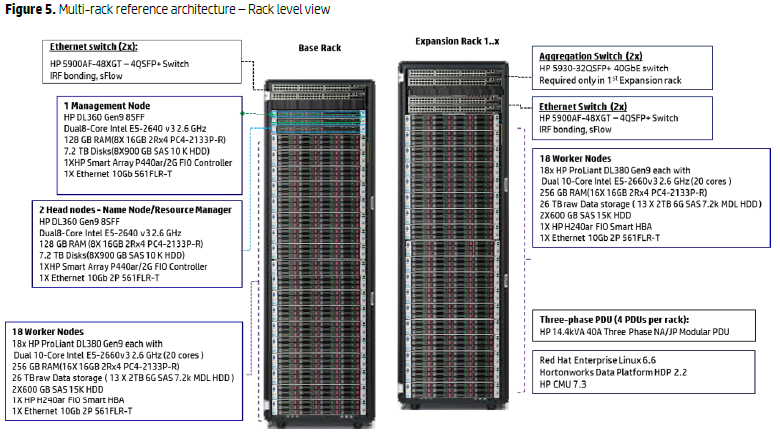
Single rack reference architecture for a non-prod Limited Release/POC deployment would initially consist of 8 nodes (3 management nodes and 5 data nodes) w/ the option and capacity to scale up to 10 data nodes. These images are for component functionality reference. Specific LM server component & cabling diagrams will be incorporated when a Network Resource has been engaged.

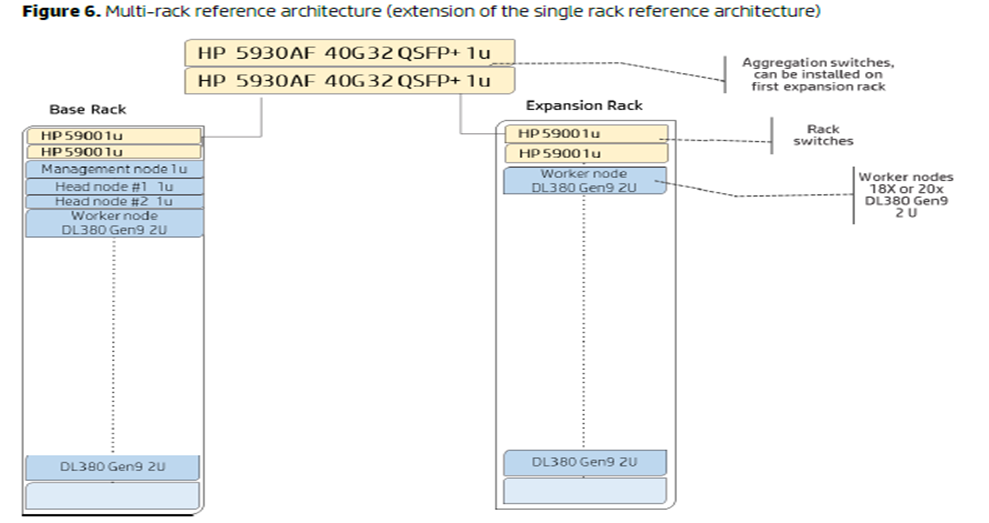


**Top of Rack Switches**

Configuring a single Top of Rack (ToR) switch per rack introduces a single point of failure for each rack. In a multi-rack system such a failure will result in a very long replication recovery time as Hadoop rebalances storage, and in a single-rack system such a failure could bring down the whole cluster. Consequently, configuring two ToR switches per rack is recommended for all production configurations as it provides an additional measure of redundancy. This can be further improved by configuring link aggregation between the switches.

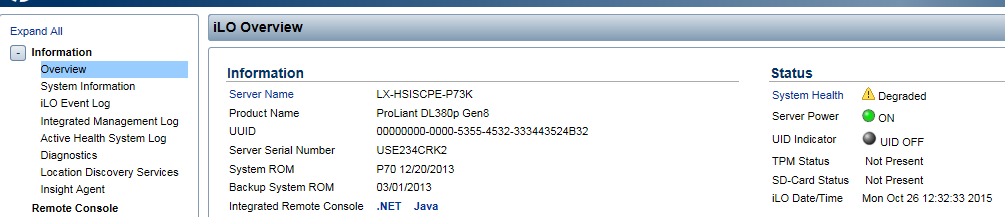
Multi-rack reference architecture for a prod deployment would employ additional data nodes and racks. The initial set of three management nodes (Name Node, Resource Manager & Management Node) are presumably capable of accommodating up to 1280 data nodes, thus additional racks would require an additional pair of aggregation switches as well as the pair of TOR and iLO switches. See *a typical Hadoop Infrastructure Architecture* diagram below.

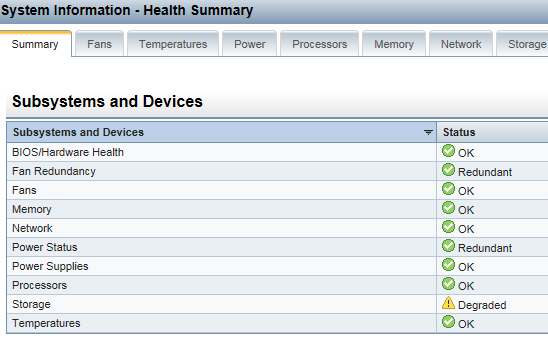


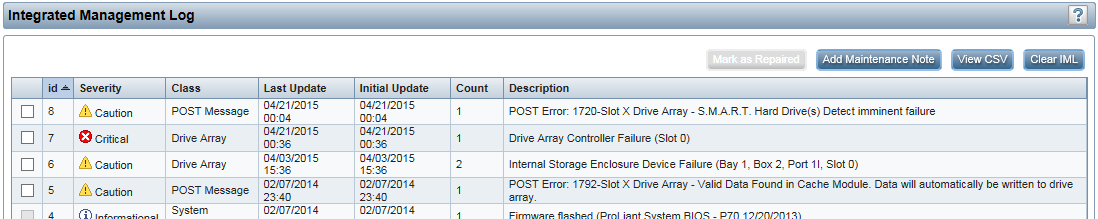


# Hardware Monitoring

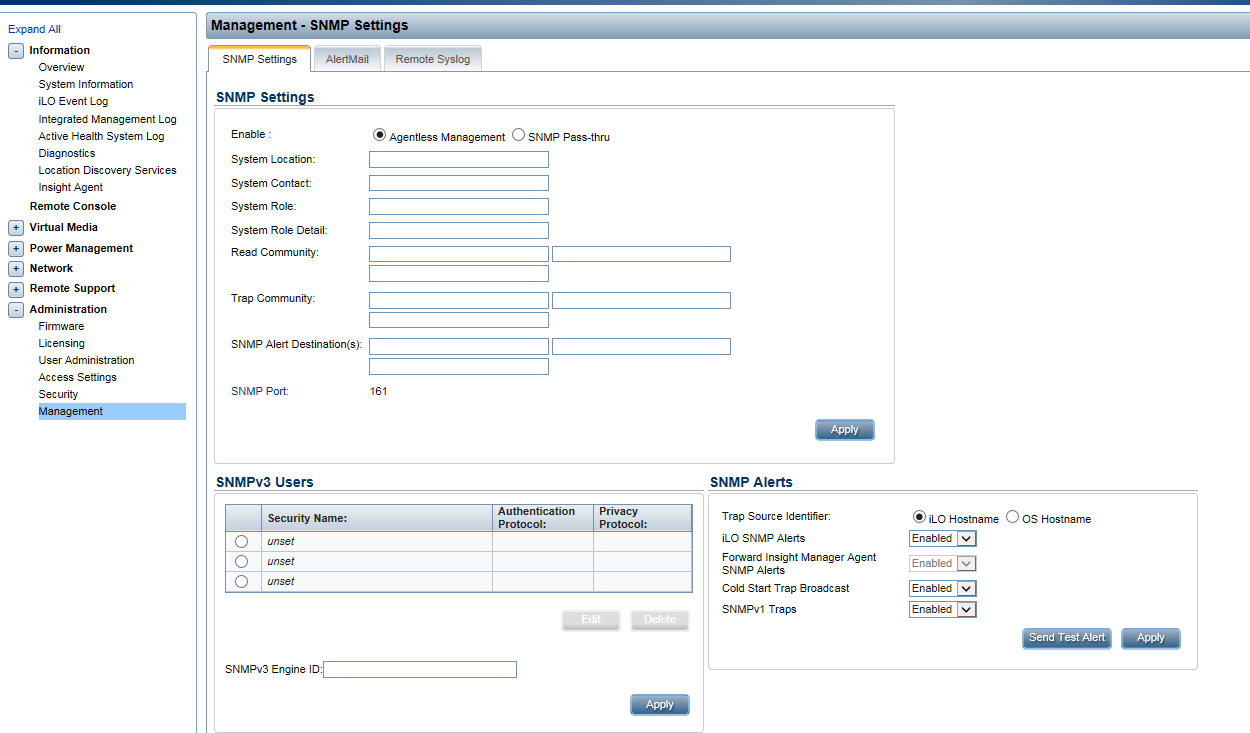
The following are screenshots showing the HP hardware logs that are captured

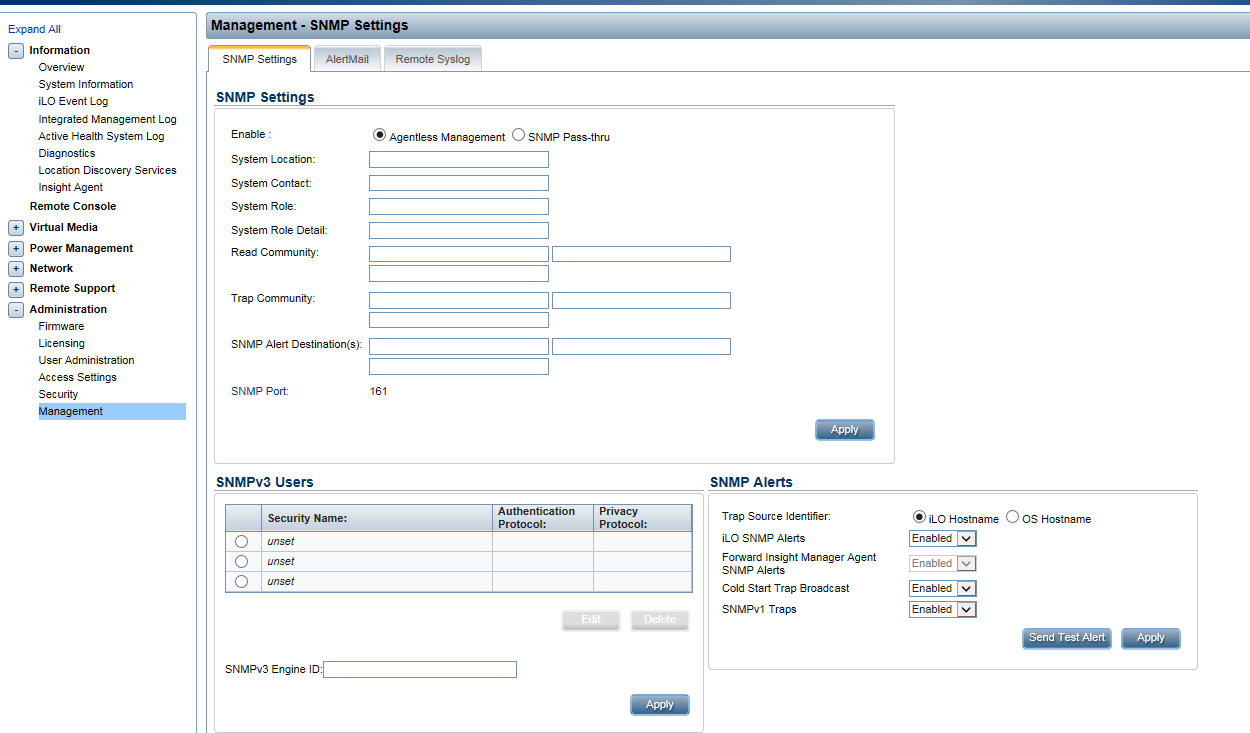






The Integrated Management Logs need to be collected sent to the standard monitoring and alerting tools.





# References:

[HP Verified Reference Architecture for Hortonworks](https://itcomposites.lmig.com/Sites/Midrange/Integration/InternalDocuments/Project%20Documents%20and%20Deliverables/Data%20Lake%20-%20Hadoop/HPHadoop.pdf)

[HDP Cloud Deployment](https://itcomposites.lmig.com/Sites/Midrange/Integration/InternalDocuments/Project%20Documents%20and%20Deliverables/Data%20Lake%20-%20Hadoop/HDP%20AWS%20Deployment%20Guide%20June2015.docx)  Reference Architecture and How-To for HDP on Amazon AWS

[Proper Care and Feeding of Drives in a Hadoop Cluster](http://hortonworks.com/blog/proper-care-and-feeding-of-drives-in-a-hadoop-cluster-a-conversation-with-stackiqs-dr-bruno/)

[The HP/Hortonworks Relationship](http://hortonworks.com/partner/hp/) Includes additional resource whitepapers

[Virtualized Hadoop Performance with VMWare](http://www.vmware.com/files/pdf/techpaper/Virtualized-Hadoop-Performance-with-VMware-vSphere6.pdf)

<http://www.slideshare.net/Hadoop_Summit/radia-srinivas-june261120amroom210c>

[Hortonworks system configuration](http://docs.hortonworks.com/HDPDocuments/Ambari-2.1.2.0/bk_Installing_HDP_AMB/bk_Installing_HDP_AMB-20150930.pdf)

[Configuring monitoring and alerting](Information%20about%20monitoring%20and%20alerting%20can%20be%20found%20here:%20http:/docs.hortonworks.com/HDPDocuments/Ambari-2.1.2.0/bk_Ambari_Users_Guide/content/_configuring_notifications.html)