QCT | Cloudera Solution Reference Architecture

QCT Infrastructure with Cloudera Enterprise 5 for Apache Hadoop

CONTENTS

REVISIONS

Version	Date	Description	Authors
0.1	11/17/2015	First publish	QCT

QCT | Cloudera Solution Reference Architecture

Executive Summary

This white paper has been created to assist in the rapid design and deployment of Cloudera Enterprise software on QCT infrastructure for clusters of various sizes, which demonstrate the reference architecture of the QCT|Cloudera Solution for Apache Hadoop.

The Reference architecture is purposed to show the QCT|Cloudera Solution comprised of hardware and software that optimize Hadoop to service large enterprise deployments as well as smaller business implementations. It also showcases that this solution is an excellent choice for the compute and storage intensive Hadoop applications.

In this whitepaper, both QCT and Cloudera will be introduced with the highlights of their key product uniqueness in the industry. All the individual hardware and software components will be described as well as the benefits of (and use cases by) adopting QCT|Cloudera Solution.

1. Introduction

Big Data offers the potential for organizations to obtain valuable insight from large collections of structured, unstructured, and fast-moving data. The trends and technologies that make big data analytics possible have opened the door to conducting powerful analysts on massive data sets, petabytes or even exabytes in size.

Today, big data tends to be widely distributed and largely unstructured. Hadoop, the core technology of big data analytics, works by breaking data into pieces and assigning each "piece" to a specific cluster node for analysis. Apache Hadoop is an open-source solution framework that supports data-intensive distributed applications on large clusters of industrial standard hardware.

Hadoop is designed to scale out from a single server to thousands of machines, with a very high degree of fault tolerance. Rather than relying on high-end hardware, the resiliency of these clusters comes from the software's ability to detect and handle failures at the application layer. Hadoop has a massive parallel computation engine plus easily scalable storage architecture. It is used in the processing of streams of data that go well beyond even the largest enterprise data sets in size. Whether it's sensor, clickstream, social media, location-based or other data that is generated and collected in large batches, Hadoop is often on the scene in the service of processing and analyzing it.

2. Hadoop Challenge

While many enterprises find Hadoop to be cost-effective and useful, Hadoop's ecosystem and technology are relatively new and complex. As is the case with adopting any new technology in the enterprise, finding people who know the technology is difficult. Also, deploying and managing Hadoop requires a depth of knowledge in terms of operations and scaling. These include:

- Controlling and securing a large Hadoop cluster, including provisioning and management, is difficult.

- Data management regulations facing industries from financial services and insurance to healthcare and manufacturing require compliance with and preparation for internal and government audits of their practices, including the lineage of the data on which they are basing critical decisions.
- Choosing the appropriate Hadoop software distribution and extensions is key.
- Finding adequate hardware with the right configuration to run Hadoop for best efficiency is not trivial. This includes servers, storage, network devices, and supporting infrastructure.

3. What QCT | Cloudera Solution Can Offer

QCT (Quanta Cloud Technology) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to datacenter customers.

Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload.

Cloudera was found in 2008 to deliver the first enterprise-class implementation of Apache Hadoop.

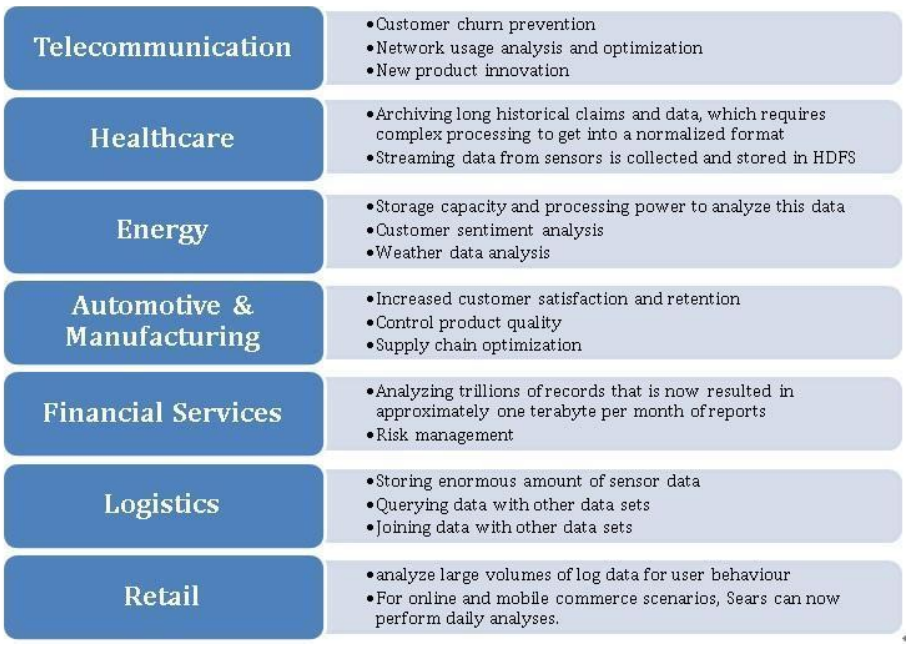
QCT has collaborated with Cloudera to develop a reference architecture that provides a solution to overcome the above challenges so that business can manage and analyze big data without spending excessive time on operations and maintenance. With the QCT|Cloudera Apache Hadoop Solution, an enterprise can adopt Apache Hadoop with minimal effort. The key benefits of adopting the QCT|Cloudera solution are:

- Providing visibility for administrators and analysts to secure, govern, and explore the large amounts of diverse data that land in Hadoop. Cloudera adds Enterprise ready features to provide end-to-end data lineage
- Deploying the appropriate Hadoop software distribution, fully tested and enterprise ready
- Providing employees access to data so that they can make more informed business decisions
- Finding a solution that is scalable and agile to support the growth of their business

4. Use Cases

A range of industry-specific use cases are demonstrated with Hadoop solutions. Enterprises in nearly every industry are discovering that they can use Hadoop to manage large quantities of data at a very low cost. Hadoop offers a way to store nearly unlimited amounts of data—structured or unstructured —and it can enable Enterprise to extract meaningful insights from that data using advanced analytics. Hadoop is already playing a key role in business intelligence and data warehousing projects in many industries.

Hadoop users access unstructured and semi-structured data from multiple sources including log files, social media feeds, sensors and internal data stores. Instead of storing this Big Data within one centralized database management system, Hadoop distributes the data across multiple machines arranged as a cluster. The cluster consists of commodity servers, which makes Hadoop relatively inexpensive to scale to petabyte levels when compared to traditional database technologies.



5. Cloudera Enterprise 5 Overview

Hadoop for the Enterprise

Cloudera Enterprise helps you become information-driven by leveraging the best of the open source community with the enterprise capabilities you need to succeed with Apache Hadoop in your organization. Designed specifically for mission-critical environments, Cloudera Enterprise includes CDH, the world's most complete, tested, and popular distribution of Apache Hadoop and related projects, as well as advanced system management and data management tools plus dedicated support and community advocacy from a world-class team of Hadoop developers and experts. Cloudera is your partner on the path to big data.

Cloudera Enterprise, with Apache Hadoop at the core, is:

Unified – one integrated system, bringing diverse users and application workloads to one pool of data on common infrastructure; no data movement required

Secure – perimeter security, authentication, granular authorization, and data protection

Governed – enterprise-grade data auditing, data lineage, and data discovery

Managed – native high-availability, fault-tolerance and self-healing storage, automated backup and disaster recovery, and advanced system and data management

Open – Apache-licensed open source to ensure your data and applications remain yours, and an open platform to connect with all of your existing investments in technology and skills

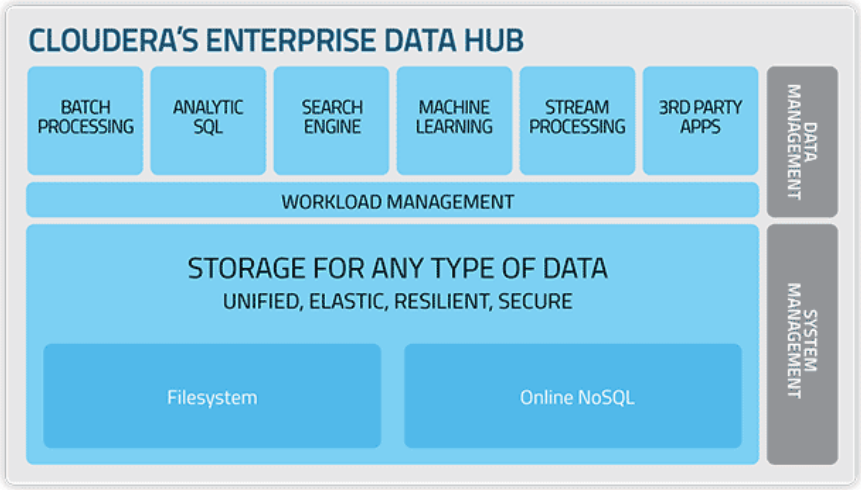
What's Inside?

CDH – At the core of Cloudera Enterprise is CDH, which combines Apache Hadoop with a number of other open source projects to create a single, massively scalable system where you can unite storage with an array of powerful processing and analytic frameworks.

Cluster Management – Cloudera Manager - Cloudera Enterprise includes Cloudera Manager to help you easily

deploy, manage, monitor, and diagnose issues with your cluster. Cloudera Manager is critical for operating clusters at scale. (*detail see Cloudera homepage)

Cloudera Support – Get the industry's best technical support for Hadoop. With Cloudera Support, you'll experience more uptime, faster issue resolution, better performance to support your mission critical applications, and faster delivery of the platform features you care about.



Cloudera Enterprise also offers support for several advanced components that extend and complement the value of Apache Hadoop:

NoSQL - HBase

Apache HBase is a distributed, scalable data store that runs on top of Apache Hadoop's file system, the Hadoop Distributed File System (HDFS). HBase is a key component of an enterprise data hub (EDH), as its design caters to applications that require fast, random access to significant data sets. HBase, which is modeled after Google's BigTable, can handle massive data tables containing billions of rows and millions of columns.

Analytic SQL - Impala

Impala is a fully integrated, state-of-the-art analytic database architected specifically to leverage the flexibility and scalability strengths of Hadoop - combining the familiar SQL support and multi-user performance of a traditional analytic database with the rock-solid foundation of open source Apache Hadoop and the production-grade security and management extensions of Cloudera Enterprise.

Search – Cloudera Search

Cloudera Search brings full-text, interactive search and scalable, flexible indexing to CDH and your enterprise data hub. Powered by Apache Hadoop and Apache Solr, the enterprise standard for open source search, Cloudera Search brings scale and reliability for a new generation of integrated, multi-workload search. Through its unique integrations with CDH, Cloudera Search gains the same fault tolerance, scale, visibility, security, and flexibility provided to other enterprise data hub workloads

In-Memory Machine learning and Stream Processing - Apache Spark

Apache Spark is an open source, general data processing framework in the

Apache Hadoop ecosystem that make it easy to develop fast, end-to-end Big Data applications combining batch, streaming, and interactive analytics on all your data. Apache Spark is a key component inside CDH, Cloudera's open source platform, with full enterprise support and capabilities available via Cloudera Enterprise.

Data Management – Cloudera Navigator

Cloudera Navigator is the only native end-to-end governance solution for Apache Hadoop-based systems. Through a single user interface, it provides visibility for administrators, data managers, data scientists, and analysts to secure, govern, and explore the large amounts of diverse data that land in Hadoop. Cloudera Navigator is part of Cloudera Enterprise's comprehensive data security and governance offering and is a key part to meeting compliance and regulatory requirements.

6. Overview of QCT

QCT (Quanta Cloud Technology) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to datacenter customers.

Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload.

QCT offers a full spectrum of datacenter products and services from engineering, integration and optimization to global supply chain support, all under one roof.

The parent of QCT is Quanta Computer Inc., a Fortune Global 500 technology engineering and manufacturing company.

QCT covers full lineup of the product portfolios, including:

Servers

1U/2U/4U general-purpose servers, multi-node high-density servers, and 4-way high-end business critical servers for applications such as HPC, virtualization, big data analytics, distributed computing, and various cloud services.

Storage

High-density converged storage systems and JBOD/JBOF systems optimized for various storage architectures and a range of I/O performance requirements.

Rack Systems

Rackgo X is a rack solution that integrates QCT servers, storage and top-of-rack switches inspired by Open Compute Project (OCP) standards. Designed for low CAPEX & OPEX with simplicity, scalability, serviceability, manageability, and high density, Rackgo X features the highest level of efficiency required by hyperscale datacenters.

Rackgo M is a rack solution based on the Open CloudServer version 2 (OCS v2) standard contributed to the Open Compute Project (OCP) by Microsoft. It is an innovative solution offering performance, density, availability, flexibility, scalability, and ease of use/deployment. Highly optimized with a unified, flexible design, Rackgo M offers superior performance with minimized TCO for cloud-scale datacenters that provide a diverse range of cloud services and/or deploy services in global datacenters with different infrastructure requirements.

Networking

Emerging trends such as cloud computing, big data and parallel calculation redefine the network infrastructure from three layers (core, aggregation and access) to two layers (spine and leaf). The QuantaMesh Ethernet switch product lineup of QCT is

designed for two-layer architecture in modern datacenters. It features low latency, low power consumption, and high port density and offers various speed options from 1G, 10G/40G to 25G/100G.

It also offers a wide range of software support, including virtualization, L3 fabric, OpenFlow-based SDN and ONIE/ONL. The integrated management interface provides unified control and provision plane to the rack scale solution.

Solution

Hyperscale engineering excellence is in the QCT "DNA." With a deep understanding of different cloud workloads, systems integration and optimization experiences, QCT offers the most efficient, innovative, scalable and easy to deploy hardware to meet demand of different scales of datacenters.

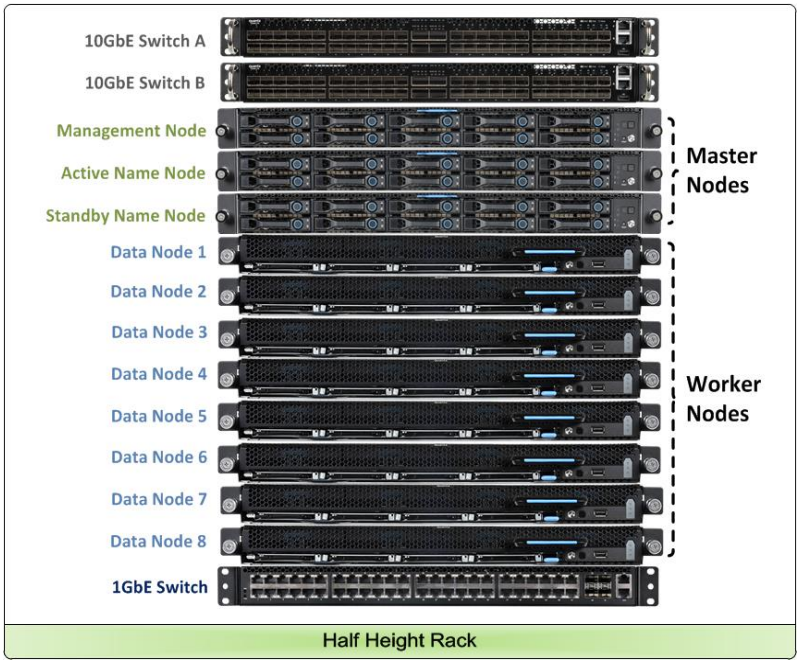
QCT product lines use standard and open hardware, so customers are never locked into proprietary hardware, firmware and management software.

Our cloud infrastructure and big data solutions will range and scale to thousands of servers, all with software pre-installed, pre-racked and pre-cabled, well verified, integrated, and optimized on our most efficient hardware platforms, easy for datacenters to deploy.

7. Solution Overview

Hadoop momentum has been building rapidly over the past several years. QCT has designed the solution with the best of practice providing the topologies for the deployment of management of Hadoop. The QCT|Cloudera Solution helps enterprise with the intention to adopt Hadoop in production. As with any emerging technology, implementing and managing Hadoop clusters and performing advanced analytics on large volumes of data requires significant expertise. The QCT|Cloudera Solution offers commercial, enterprise-ready Hadoop tools and applications that help automate deployment in a production environment. The faster you can deploy, the faster you can derive value from the data.

Building Blocks Reference Architecture



8. HA Considerations

Hadoop NameNode HA

The configurations in this white paper utilize quorum-based journaling high-availability features in Cloudera CDH5.

Resource Manager HA

Resource Manager could be made by a YARN cluster and is responsible for tracking the resources in a cluster. And the High Availability feature adds redundancy in the form of an Active/Standby Resource Manager pair to remove this single point of failure.

OS availability and reliability

For the reliability of the server, the OS disk is configured in a RAID 1(two root disks) configuration thus preventing failure of the system from OS hard disk failures.

Network reliability

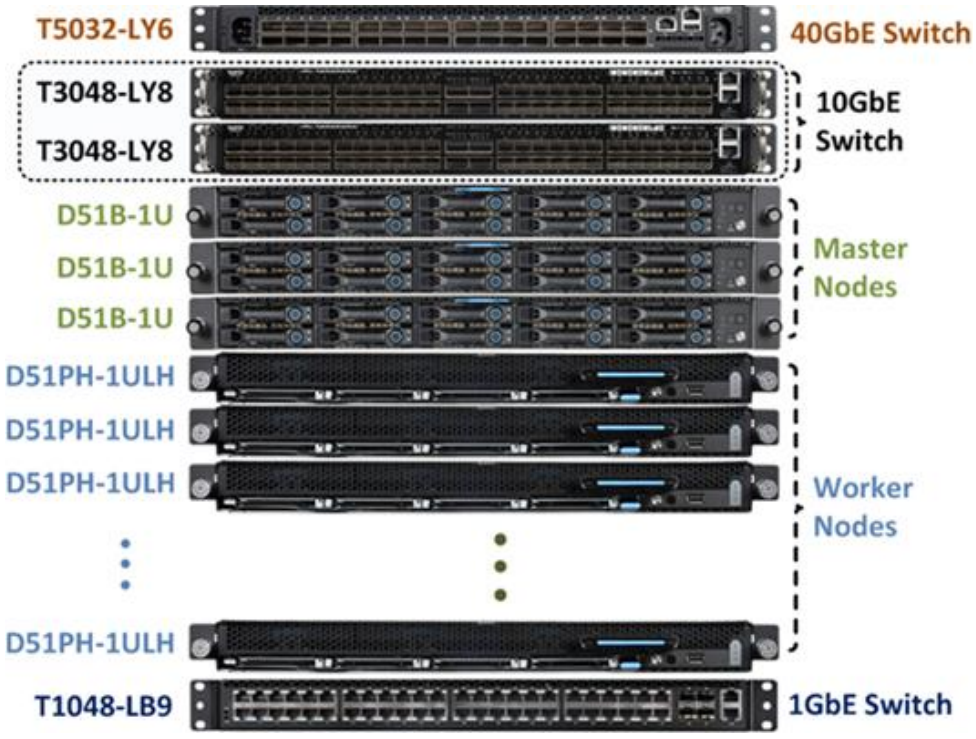
The reference architecture configuration uses QCT T3048-LY8 switches for redundancy, with redundant power supplies.

Power supply HA

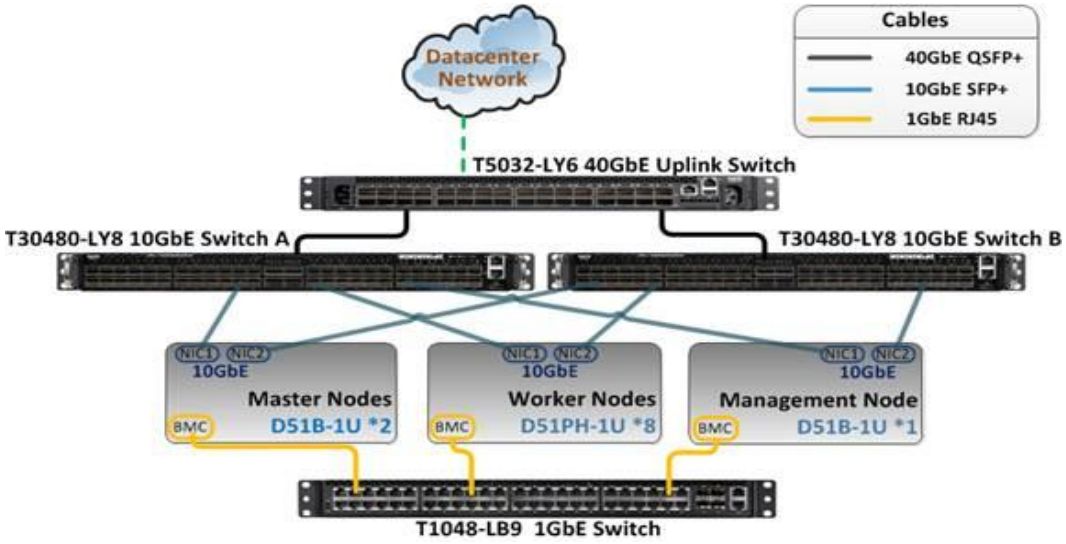
Each server has a backup power supply, and each rack have at least two Power Distribution Units (PDUs) for adequate power redundancy.

9. Hardware / Network / Cluster Architecture

Topology for Hardware



Network Topology



Switches	Description
 T5032-LY6 40GbE uplink Switch	Uplink or TOR switches
 T30480-LY8 10GbE Switch *2	Dual switches with HA for cluster data network (1) Each switch attached one 10GbE cable from NIC0 of each node, it's master connection and load balanced. (2) Another cable from node NIC1 is slave and load balanced.
 T1048-LB9 1GbE Switch	Switch per rack, dedicated for BMC(Mgmt.) network
Cables	**Description**
40GbE QSFP+	Provide connectivity and between the router/gateway/firewall and cluster data network infrastructure.
10GbE SFP+	1. Provides connectivity from each node to HA cluster data switches. Dual connections with active load balancing are used from each node. 2. Two NICs on each nodes are bonded by mode 6(teaming) which support failover and load balanced.
1GbE RJ45	Between BMC network ports and the out-of-band management ports of the switches

D51B-1U - Hardware Architecture

QuantaGrid D51B-1U is a general-purpose rackmount server designed for optimal performance and power efficiency. It is based on the dual Intel® Xeon® processor E5-2600 v3 product family and features up to 1.5 TB memory capacity in a 1U chassis.

Greener and More Powerful:

- Powered by the Intel® Xeon® processor E5-2600 v3 product family and DDR4 memory technology, the QuantaGrid D51B-1U allows owners to upgrade computing performance without overextending power consumption. This allows owners to save unnecessary costs associated with datacenter cooling needs and to achieve higher data center infrastructure efficiency value

Flexible and Scalable I/O Options:

- QuantaGrid D51B-1U provides flexible I/O scalability for today's diverse data center application requirements. It features OCP LAN mezzanine card solutions in addition to dual GbE or 10GbE (LAN on motherboards (LoM)). With various controller vendors and different speed and technology options, customers can choose from 1GbE to 56GbE bandwidth, copper or fiber-optic cabling, basic Ethernet function or FCoE and iSCSI SAN connectivity. The on-board SAS controller offers multiple QCT SAS mezzanine card options, with different RAID levels, and data transfer bandwidth, so customers can tailor the SAS controller for specific application needs.

D51PH-1ULH - Hardware Architecture

Tailored for hyper-scale datacenters and software defined storage solution, QCT's innovative D51PH-1ULH features hybrid tiered storage architecture in an ultra-dense hot-swappable 1U platform. D51PH-1ULH is a rack mount server based on Intel® Xeon® processor E5-2600 v3 product family and with up to 1TB memory capacity. It is equipped with 12x hot-swappable 3.5" disk drives and 4x hot-swappable 2.5" SATA SSD (7mm).

Ultra Dense 1U High computing storage server

- D51PH-1ULH features 12 hot-swappable 3.5" disk drives and 4 SSDs in a 1U chassis, delivering

excellent computing performance using a dual Intel® Xeon® E5-2600 v3 processor architecture. D51PH-1ULH provides both extreme storage density and computing power to hyper scale cloud data centers which require not only the large storage capacity but also enterprise class computing capability. It helps IT managers to optimize overall rack and data center space. With its innovative hot-swappable drive design, D51PH-1ULH allows service technicians to service drives with minimum efforts and little to no downtime.

Sophisticated Hybrid Architecture

- Caching is a requirement not only in today's software defined storage solution but also virtualization applications to boost performance and IOPs. D51PH-1ULH is equipped with 4x SSDs in addition to 12x 3.5" high capacity disk drives, such hybrid architecture is ideal for tier storage planning which solid-state drives are required to accelerate IOPs and throughput, and not sacrificing large data storage capacity.

Hot-Swappable and Easy Service

- D51PH-1ULH is designed with delicate architecture to support 12 hot-swappable 3.5" hard drives within a 1U chassis, which provides optimized rack density and also deliver non-disruptive on-site service. With this design, there is no external cable management arm required which significantly reduce the system deployment and rack assembly time, and also simplify the rack cable routing in the rear.

Network Architecture

The QuantaMesh T3048-LY8 is a high performance and low latency layer 2/3/4 Ethernet switch with 48 SFP+ ports and up to 6 QSFP+ port in a 1U form factor. Each 40G QSFP+ port can be independently configured as 40GbE or 4 x 10GbE ports. Built for Infrastructure-as-a-service (IaaS) datacenter deployment, high performance computing clusters, and financial applications, the very high port density and high performance as well as ultra-low latency characteristic makes QuantaMesh T3048-LY8 ideal for demanding workloads and provides the best TCO.

Automation

- Accompanied by the application of cloud computing, big data and parallel calculation, datacenter network devices continue to grow fast and make network automation a critical factor. Supporting auto installation and integration with orchestration tools like Chef and Puppet, QuantaMesh T3048-LY8 helps for easy deployment of mass datacenter IaaS build-up.

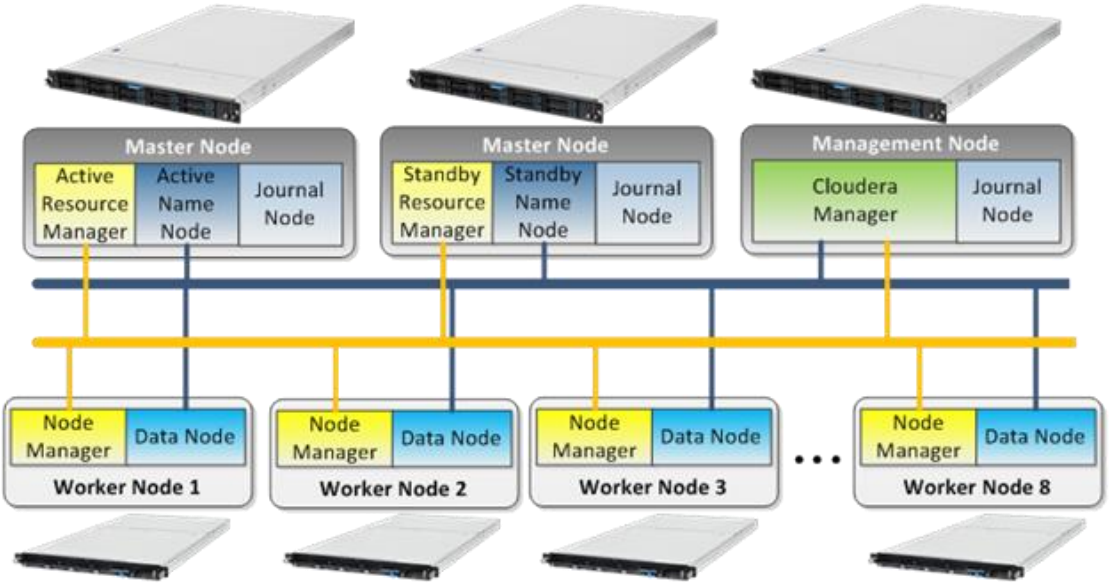
High Availability

- 1+1 hot-swappable power supplies , 2+1 hot-swappable fans

- Out-of-band management supported. Multi-chassis LAG for preventing the risks of single point failure .Up to 32 paths ECMP routing for load balancing and redundancy

- x86 CPU Board Support
- MLAG, OSPF, BGP, & ECMP
- VXLAN/NVGRE
- OpenFlow Support
- OpEN API

Cluster Node Architecture



**ABOVE LINKING LINES ARE NOT FOR REAL NETWORK CABLES BUT SHOWS A ROLE FOR BASIC CLUSTER

The cluster environment consists of multiple software services running on multiple physical server nodes. The implementation divides the server nodes into several roles, and each node has a configuration optimized for its role in the cluster. A 10GbE network connects the cluster nodes together, and separates the core data network from management functions.

Logical Topology

Node		Master Node	Master Node	Master Node	DataNode 1..n
Service/ Roles	ZooKeeper	ZooKeeper	ZooKeeper	ZooKeeper	
	HDFS	NN,QJN	NN,QJN	QJN	Data Node
	YARN	RM	RM	History Server	Node Manager
	Hive			MetaStore, WebHCat, HiveServer2	
	Management (misc)			Oozie, CM, Management Services	Cloudera Agent
	Navigator			Navigator, KMS	
	HUE			HUE	
	SOLR				Search
	Spark				
	Impala			statestore, catalog	Impalad
	HBASE	HMaster	HMaster	HMaster	Region Servers

10. Software Architecture

The QCT | Cloudera Solution is based on Cloudera Enterprise, which includes Cloudera's distribution for Hadoop (CDH) and Cloudera Manager.

Cloudera Manager

Cloudera Manager is designed to make administration of CDH simple and straightforward, at any scale. With Cloudera Manager, you can easily deploy and centrally operate the complete Hadoop stack. The application automates the installation process, reducing deployment time from weeks to minutes; gives you a cluster-wide, real-time view of nodes and services running; provides a single, central console to enact configuration changes across your cluster; and incorporates a full range of reporting and diagnostic tools to help you optimize performance and utilization.

Cloudera RTQ (Impala)

Cloudera Impala is an open source Massively Parallel Processing (MPP)query engine that runs natively in Apache Hadoop. The Apache-licensed Impala project brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache HBase without requiring data movement or transformation. Impala is integrated from the ground up as part of the Hadoop ecosystem and leverages the same flexible file and data formats, metadata, security and resource management frameworks used by MapReduce, Apache Hive, Apache Pig and other components of the Hadoop stack.

Cloudera Search

Cloudera Search delivers full-text, interactive search to CDH, Cloudera's 100% open source distribution including Apache Hadoop. Powered by Apache Solr, Cloudera Search enriches the Hadoop platform and enables a new generation of search –Big Data search – through scalable indexing of data within HDFS and Apache HBase. Cloudera Search gains the same fault tolerance, scale, visibility, and flexibility provided

to other Hadoop workloads, due to its integration with CDH.

Cloudera Enterprise BDR

BDR is an add-on subscription to Cloudera Enterprise that provides end-to-end business continuity. When you add BDR to your Cloudera Enterprise subscription, you'll get the management capabilities and support you need to get maximum value from the powerful disaster recovery features available in CDH. BDR makes it easy to configure and manage backup disaster recovery policies for data stored in CDH.

Cloudera Navigator

Navigator is an add-on subscription to Cloudera Enterprise that provides the first fully integrated data management tool for Cloudera Enterprise. It's designed to provide all of the capabilities required for administrators, data managers and analysts to secure, govern, and explore the large amounts of diverse data that land in CDH. Cloudera Navigator is the only native end-to-end governance solution for Apache Hadoop-based systems. Through a single user interface, it provides visibility for administrators, data managers, data scientists, and analysts to secure, govern, and explore the large amounts of diverse data that land in Hadoop. The first release of Cloudera Navigator (v1.0) was developed specifically to address data security concerns most typically associated with highly regulated industries, such as financial services, healthcare and government. It includes a full suite of auditing capabilities across all CDH components that store data.

11. System Configuration Guidance

As QCT|Cloudera Solution distributes data through a cluster of machines and uses replication to ensure redundancy and data integrity in production. All the machines are configured to optimize the performance to meet the requirement of, performance, storage capacity, as well as all the networking capability. Below is the suggested Hardware and software configuration suggested by QCT.

The Best Practice

a. Hardware Configuration

Server Hardware	Item	Description	Qty	Total Qty
Master & Mgmt. Nodes	CPU	Intel Xeon E5-2660 v3	2	6
	Memory	16GB DDR4 2400MHz RDIMM	24	72
	SSD	SSD SATA 800GB for OS(RAID 1)	2	6
	Hard drive	HDD 2.5" 900G	8	24
	RAID	LSI QS-3008-8i-IR for Non-OS disk	1	3
	NIC	Dual Ports 10GbE (SFP+)	1	3
Worker Nodes	CPU	Intel Xeon E5-2660 v3	2	16
	Memory	Samsung 16GB DDR4	16	128
	SSD	SSD SATA 800GB for OS(RAID 1)	2	16
	Hard drive	3.5" SAS 7.2K 3TB	12	96
	RAID	LSI QS-3008-8i-IR for Non-OS disk	1	8
	NIC	At least one 10GbE(SFP+)	1	8
10GbE	LY8	T3048-LY8 48/10Gb,4/40Gb	1	2

Switch				
1GbE Switch	LB9	T1048-LB9 48 100/100/1000BASE-T and 4 1/10Gbe SFP+ ports	1	1

b. Software Components

Mgmt. Node	CentOS 6.6	Operating System
	Oracle JDK 1.7.0_65	Java Development Kit
	Cloudera Manager 5	End-to-end Application for Managing CDH Clusters
	NameNode HA	Journal Node for HDFS
	HiveServer 2	Data warehouse Infrastructure
	Hue Server	Web Applications to Interact with CDH Cluster
	ZooKeeper	Coordination Service
	Oozie Server	Oozie Workflow Scheduler
	Sqoop 2 Server	Connectivity tool for Moving Data

Master Node 1	CentOS 6.6	Operating System
	Oracle JDK 1.7.0_65	Java Development Kit
	NameNode	Active NameNode / Journal Node for HDFS
	ResourceManager	ResourceManager for YARN
	ZooKeeper	Coordination Service

Master Node 2	CentOS 6.6	Operating System
	Oracle JDK 1.7.0_65	Java Development Kit
	NameNode HA	Standby NameNode / Journal Node for HDFS
	JobHistoryServer	JobHistory for YARN
	ZooKeeper	Coordination Service
	Impala Server	Analytic Database for Hadoop

Worker Nodes	CentOS 6.6	Operating System
	Oracle JDK 1.7.0_65	Java Development Kit
	DataNode	DataNode for HDFS
	NodeManager	NodeManager for YARN

12. Summary

Traditional data management and business analytics tools and technologies are straining under the added weight of Big Data, but new approaches are emerging to help enterprises gain actionable insights from Big Data via tools like Hadoop. Because a Hadoop cluster is optimized for data that is highly distributed, loosely structured, and increasingly large in volume, it is ideal for processing Big Data. Organizations that embrace Big Data as the foundation of their business analytics practices stand to gain significant competitive advantage. Business intelligence and application users who are not Hadoop specialists may, however, find their lack of knowledge a significant barrier.

The QCT|Cloudera solution allow enterprise to derive new business insights from Big Data by providing a platform to store, manage and process data with a fully integrated and ready-to-deploy solution. This solution also enables Enterprise Hadoop to deliver a next-generation enterprise data platform that makes extracting value from big data analytics faster and more reliable.

In conclusion, the QCT|Cloudera solution provides optimized performance for both hardware and software that help enterprises deploying Hadoop and big data succeed.

Appendix – Glossary of Terms

Abbreviation	Definition
Apache Hadoop	A free, open source software framework that supports data-intensive distributed applications. The core components of Apache Hadoop are the HDFS and the MapReduce processing framework. The term is also used for an ecosystem of projects related to Hadoop that fall under the umbrella of infrastructure for distributed computing and large-scale data processing.
Big data	Data sets whose input/output velocity, variety of data structure, and volume is beyond the beyond the capabilities of systems which were designed assuming smaller data sets to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are an expanding target, currently ranging from terabytes to many petabytes of data in a single data set.
BigTable	A compressed, high performance, column-oriented database built on Google File System (GFS). The BigTable design was the inspiration for Apache HBase and Apache Accumulo, but the implementation, unlike other Google projects such as Protocol Buffers, is proprietary.
CDH	Cloudera Apache Hadoop distribution containing core Hadoop and the following related projects: Apache Avro, DataFu, Apache Flume, Fuse-DFS, Apache HBase, Apache Hive, Hue, Apache Mahout, Apache MRv1, Apache Oozie, Apache Pig,Apache Spark, Apache Sqoop, Apache Whirr, and Apache ZooKeeper
CAPEX	Capital Expenditure
Flume	A service designed to collect data and put it into a Hadoop environment
Hadoop	An open source platform, developed at Yahoo! that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. The platform particularly suited to large volumes of unstructured data such as Facebook comments and Twitter tweets, email and instant messages, and security and application logs
HBase	HBase is an open-source, distributed, versioned, non-relational database modeled after Google's Bigtable: A Distributed Storage System for Structured Data
HDFS	Hadoop's Distributed File system breaks large application workloads into smaller data blocks that are replicated and distributed across a cluster of commodity hardware for faster processing
Hive	A data warehousing solution built on top of Hadoop
Hue	A browser-based desktop interface for interacting with Hadoop.

JBOD/JBOF	Just a bunch of disk/Just a bunch of flash
MapReduce	A software framework for easily writing applications that process vast amounts of data (multi-terabyte data-sets) in parallel on large clusters of commodity hardware in a reliable, fault-tolerant manner. Hadoop acts as a platform for executing MapReduce.
NIC	Network interface card
OPEX	Operating Expense
Oozie	A workflow engine for Hadoop.
Pig	A platform for analyzing large data that leverages parallel computation
QJN - Quorum JournalNode	Maintains a directory to log the modifications to the namespace metadata when using the Quorum-based Storage mechanism for providing High Availability. In the event of a failover, the NameNode standby will ensure that it has applied all of the edits from the JournalNodes before promoting itself to the active state.
Spark	Spark is an open source framework for machine learning, interactive analytics, and streaming analytics capabilities for large datasets. Spark is appropriate for low-latency computations and iterative algorithms that employ its in-memory computing capabilities. At Altiscale, Spark is fully integrated into the larger Hadoop ecosystem, so customers benefit not only from Spark, but also from Hive, Pig, and MapReduce, and tools like MLlib, SparkSQL, Adatao and H2O.
Sqoop	A tool designed to transfer data between Hadoop and relational databases
TCO	Total Cost of Ownership
Thrift	An interface definition language, runtime library, and a code generation engine to build services that can be invoked from many languages. Thrift can be used for serialization and RPC, but within Hadoop is mainly used for RPC.
ToR	Top-of-rack switch/router
Whirr	A set of libraries for running cloud services. It's ideal for running temporary Hadoop clusters to carry out a proof of concept, or to run a few one-time jobs
YARN	Allows Hadoop administrators to track and coordinate distributed applications.
Zookeeper	Allows Hadoop administrators to track and coordinate distributed applications.

References

This paper references the following documents and resources

Cloudera Website:

http://www.cloudera.com

About QCT

QCT (Quanta Cloud Technology) is a global datacenter solution provider extending the power of hyperscale datacenter design in standard and open SKUs to all datacenter customers.

Product lines include servers, storage, network switches, integrated rack systems and cloud solutions, all delivering hyperscale efficiency, scalability, reliability, manageability, serviceability and optimized performance for each workload.

QCT offers a full spectrum of datacenter products and services from engineering, integration and optimization to global supply chain support, all under one roof.

The parent of QCT is Quanta Computer Inc., a Fortune Global 500 technology engineering and manufacturing company.

http://www.QCT.io

United States

QCT LLC., Silicon Valley office
1010 Rincon Circle, San Jose, CA 95131
TOLL-FREE: 1-855-QCT-MUST
TEL: +1-510-270-6111
FAX: +1-510-270-6161
Support: +1-510-270-6216

QCT LLC., Seattle office
13810 SE Eastgate Way, Suite 190, Building 1,
Bellevue, WA 98005
TEL: +1-425-633-1620
FAX: +1-425-633-1621

China

云达科技, 北京办公室（Quanta Cloud Technology）
北京市朝阳区东三环中路 1 号，环球金融中心东楼 1508 室
Room 1508, East Tower 15F, World Financial Center
No.1, East 3rd Ring Zhong Rd., Chaoyang District, Beijing, China
TEL: +86-10-5920-7600
FAX: +86-10-5981-7958

云达科技, 杭州办公室（Quanta Cloud Technology）
浙江省杭州市西湖区古墩路浙商财富中心 4 号楼 303 室
Room 303，Building No.4，ZheShang Wealth Center
No. 83 GuDun Road, Xihu District, Hangzhou, Zhejiang , China
TEL: +86-571-2819-8660

Japan

Quanta Cloud Technology Japan 株式会社
日本国東京都港区芝大門二丁目五番八号
牧田ビル 3 階
Makita Building 3F, 2-5-8, Shibadaimon ,
Minato-ku, Tokyo 105-0012, Japan
TEL: +81-3-5777-0818
FAX: +81-3-5777-0819

Taiwan

雲達科技（Quanta Cloud Technology）
桃園市龜山區文化二路 211 號 1 樓
1F, No. 211 Wenhua 2nd Rd., Guishan Dist.,
Taoyuan City 33377, Taiwan
TEL: +886-3-286-0707
FAX: +886-3-327-0001

Other regions

Quanta Cloud Technology
No. 211 Wenhua 2nd Rd., Guishan Dist.,
Taoyuan City 33377, Taiwan
TEL: +886-3-327-2345
FAX: +886-3-397-4770