ORACLE
Cloud Infrastructure

Cloudera Enterprise Data Hub Reference Architecture for Oracle Cloud Infrastructure Deployments

ORACLE WHITE PAPER | NOVEMBER 2018

ORACLE

Disclaimer

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

Revision History

The following revisions have been made to this white paper since its initial publication:

Date	Revision
November 29, 2018	Updated the minimum and suggested shapes for worker instances in the development deployment.
June 7, 2018	Initial publication of paper.

You can find the most recent versions of the Oracle Cloud Infrastructure white papers at https://cloud.oracle.com/iaas/technical-resources.

Table of Contents

Overview

Customers of both Cloudera and Oracle Cloud Infrastructure can now run Cloudera Enterprise Data Hub deployments in the cloud. Leveraging the power of Oracle Cloud Infrastructure bare metal instances, customers can drive flexible, easily scalable, and performant Enterprise Data Hub clusters in an automated fashion by using Terraform on Oracle Cloud Infrastructure.

This white paper details best practices for running Enterprise Data Hub on Oracle Cloud Infrastructure. Although individual use cases and requirements might vary and demand different approaches, the practices set forth in this paper represent the ideal configuration for both performance and security on Oracle Cloud Infrastructure. Topics covered in this paper include installation automation, automated configuration and tuning, and best practices for deployment and topology to support security and high availability.

The cloud reference architecture presented here represents best practices for sizing and deployment on Oracle Cloud Infrastructure. For more reference information about Cloudera, see the Appendix for links to the latest Cloudera documentation.

Oracle Cloud Infrastructure Terminology

This paper uses many terms specific to Oracle Cloud Infrastructure. For definitions of these terms, see “Oracle Cloud Infrastructure Terminology Reference” in the Appendix.

Infrastructure Guidance

All Enterprise Data Hub deployments on Oracle Cloud Infrastructure leverage either bare metal or virtual machine instances. The choice of which instances to use is yours, and this section provides some best practices to follow when making that choice. Terraform templates available on the Oracle Cloud Infrastructure Provider GitHub are preconfigured with the recommended instance types.

Note: Changing the instance types as part of the deployment could result in an unsupported cluster configuration, so consider this before making changes.

Compute Considerations

You have many options to consider when choosing the architecture for your Enterprise Data Hub deployment on Oracle Cloud Infrastructure. This section provides information about which instances are supported configurations for Cloudera.

Oracle Cloud Infrastructure Bare Metal Compute

Enterprise Data Hub on Oracle Cloud Infrastructure is validated by Cloudera for Bare Metal DenseIO worker instances only, using NVMe-based local storage for Apache Hadoop Distributed File System (HDFS). Two profiles are supported for bare metal instances running Enterprise Data Hub as workers, which differ based on compute, memory, and storage density.

- **BMDenseIO1.36 workers:** This instance provides 36 OCPUs (72 vCores), 512 GB of memory, and 28.8 TB in local NVMe storage. Additional block storage can be attached, up to 512 TB per host, but is not currently supported for HDFS.
- **BMDenseIO2.52 workers:** This instance provides 52 OCPUs (104 vCores), 768 GB of memory, and 51.2 TB in local NVMe storage. Additional block storage can be attached, up to 512 TB per host, but is not currently supported for HDFS.

More information about these compute profiles, including performance-related metrics, is located in the blog post High Performance X7 Compute Service Review and Analysis.

Oracle Cloud Infrastructure Virtual Machine Compute

Enterprise Data Hub can be deployed on virtual machines by using block storage for HDFS, but this is not currently supported by Cloudera. Oracle plans on getting vendor validation for this architecture in the near future. When you deploy using virtual machines, it is important to consider IOPS and bandwidth constraints when configuring your deployment.

Virtual machines are currently leveraged as non-worker elements of Enterprise Data Hub deployments that use the Terraform templates detailed later in this paper. Virtual machines are acceptable for bastion, utility, and master hosts, which do not require large compute, memory, or local disk capacity like worker nodes do for workload execution.

Note: It is possible to configure Enterprise Data Hub deployment to leverage BM.Standard and VM instances as workers. Although this configuration is not currently supported by Cloudera, we have found it to be extremely performant while also being cost effective. Before assigning an instance to a particular role, see the "Terraform Templates" section for minimum required shapes per role.

Storage Considerations

Oracle Cloud Infrastructure has several offerings to consider when choosing which storage to use for HDFS or for other purposes in your Enterprise Data Hub deployment.

Bare Metal NVMe Storage

Oracle Cloud Infrastructure's bare metal NVMe storage provides a fast option for use as HDFS, and is currently supported by Cloudera for Enterprise Data Hub on Oracle Cloud Infrastructure. This model uses bare metal instances that have local NVMe-based storage as the underlying capacity for HDFS. This model is the highest performant storage option for running Cloudera Enterprise Data Hub on Oracle Cloud Infrastructure, and is recommended for production deployments.

When you deploy Enterprise Data Hub on bare metal, consider the HDFS replication factor for environments that require data redundancy. We recommend a replication factor of 3 when you use bare metal NVMe for HDFS.

Block Storage

The Oracle Cloud Infrastructure Block Volume service provides a cost effective means for securely and reliably storing data while maintaining performance. Block storage volumes are completely flexible in configuration, from 50 GB to 16 TB per volume, in 1 GB increments. Each instance can have a maximum of 32 volumes attached.

Oracle has a guaranteed SLA on block storage, ensuring 3K IOPS and 24 MB/s per 50 GB of block storage, up to a maximum of 25K IOPS and 320 MB/s per volume. This means that a block storage volume peaks at 700 GB for IOPS and throughput. This bandwidth aggregates at the host level and is something that you should consider if you choose to use block storage as HDFS. If the aggregate volume bandwidth is not high enough, HDFS stability during load will be a concern. Although this does not usually affect smaller deployments, it can become problematic for instance types with large CPU and memory capacity, or for large clusters.

Block storage does provide a unique advantage when used for HDFS. Because redundancy is built into the platform, the requirement for running an HDFS replication factor of 3 for physical redundancy is not necessary. HDFS can be run at a replication factor of 1 with block storage, allowing for performance gains, while still being redundant because of the underlying replication of block storage volumes on Oracle Cloud Infrastructure.

Network Considerations

Oracle provides a guaranteed networking SLA for instance and block storage bandwidth. For detailed bandwidth information for each instance, see the Compute service documentation.

Networking on Oracle Cloud Infrastructure uses virtual cloud networks (VCNs) as the basis for all connectivity. For basic information about VCNs, read the FAQ.

VCNs support the concept of *security lists* to manage security and network access. Security lists are used with host-level firewalls to limit or permit access to services run on instances in Oracle Cloud Infrastructure.

VCNs are local to each region and can span multiple availability domains. Multiple subnets can exist inside a single VCN and availability domain. Subnets must have a unique CIDR inside each VCN.

Instances have virtual network interface cards (VNICs), which are attached to specific subnets inside a availability domain. Instances and VNICs can only be a part of the same availability domain.

- BMDenseIO1.36 instances support 10Gbps, with a maximum of 16 VNICs per instance.
- BMDenseIO2.52 instances support dual 25Gbps, with a maximum of 24 VNICs per instance (12 per physical NIC).

On-Premises Connectivity

Oracle Cloud Infrastructure supports private connectivity across your on-premises and cloud networks, allowing you to extend your IT infrastructure with connectivity services that offer predictable and consistent performance, isolation, and availability.

This connectivity gives you the ability to leverage a hybrid deployment model, allowing for versatile uses of cloud infrastructure as part of your big data ecosystem.

For more information about this connectivity, see the Oracle Cloud Infrastructure Fast Connect FAQ.

Enterprise Data Hub on Oracle Cloud Infrastructure: Deployment Recommendations

This section provides detailed best practices for cluster and network architecture, and deployment topology for Enterprise Data Hub on Oracle Cloud Infrastructure.

Cluster Architecture

Enterprise Data Hub cluster architecture on Oracle Cloud Infrastructure follows the supported reference architecture from Cloudera. A basic cluster consists of a utility host, master hosts, worker hosts, and one or more bastion hosts.

- The **utility host** is the primary host in the cluster used for core administrative services. It hosts the Cloudera Manager, Hue server, and Job History server UI. It is also leveraged during initial cluster setup, and runs a ZooKeeper daemon for cluster service coordination.
- **Master hosts** run core cluster service daemons for NameNode, Failover Controller, Resource Manager, HBase, and ZooKeeper. These daemons drive workloads on the worker hosts.
- **Worker hosts** run HDFS and Apache Hadoop YARN, and are the target for all jobs inside the cluster. These hosts facilitate compute and memory resources for all job execution, and HDFS for file storage and replication.
- The **bastion host** acts as an edge node for user interaction and job submission for the cluster. It's also where third-party software should be installed for use with the Enterprise Data Hub cluster.

Bastion and utility hosts should have public IP addresses so that they can be accessed outside the VCN, and access should be restricted though security lists. Master and worker hosts should be deployed on a private network and not be directly accessible from the internet.

The following table shows the services that run on each type of host:

Service	Utility Host	Master Hosts (2)	Worker Hosts	Bastion Hosts
HDFS	• Journal Node • HTTP File server	• NameNode • Journal Node • Failover Controller	Data Host	
YARN	Job History server	Resource Manager	Host Manager	
Hive	• MetaStore • WebHCat • Hive Server 2			
Hue	Hue server			
Spark	History server			

Service	Utility Host	Master Hosts (2)	Worker Hosts	Bastion Hosts
Impala	Catalog server		Impala Daemon	
Cloudera Search			Solr	
HBase	Thrift server	HBase Master	Region server	
ZooKeeper	ZooKeeper Service	ZooKeeper Service		
Flume				Flume Agent
Gateway Role				• HDFS • YARN • Hive • Sqoop • Hue
Management Role	• Cloudera Manager and Service • Cloudera Manager Agent • Oozie	Cloudera Manager Agent	Cloudera Manager Agent	Cloudera Manager Agent

Network Architecture

The recommended network architecture for Enterprise Data Hub deployment on Oracle Cloud Infrastructure consists of a VCN containing three subnets, which are duplicated across all availability domains in a target region. This architecture enables you to deploy an Enterprise Data Hub cluster in any availability domain in the region and have the same topology and security lists associated with each network.

Bastion Network

The bastion network is used as an edge network, has direct access to the internet, and is where the bastion hosts are deployed. Instances in this network have both a public and a private IP address. This network acts as an entry point for accessing cluster resources while not exposing those services directly to the internet.

Public Network

The public network is secondary to the bastion network and also has direct access to the internet, along with public and private IP addresses for each instance associated with it. This network is where the utility node is deployed, and it provides additional services like Cloudera Service Manager, Job History, Hue, and other UIs that require external access to interact with.

Private Network

The private network should have only private IP addresses for all instances associated with it. This network is more secure because the instances on it can't be accessed directly from the internet. This network is where master and worker instances are deployed, which provides additional security for services and data on these instances.

Network Access

Access to all of these networks is controlled by Security Lists. Security lists are whitelists that allow network connectivity between the internet and subnets, and subnet interaction inside a VCN. For more information about security lists, see the Networking service documentation.

There is no deny rule for network traffic on Oracle Cloud Infrastructure in a VCN because the default behavior is to deny. The only way for traffic to route is to create a security list rule that allows the traffic, whether it is allowing the entire network segment internal access between subnets in the VCN, or allowing a specific host IP/network access to the Cloudera Service Manager UI on the utility node.

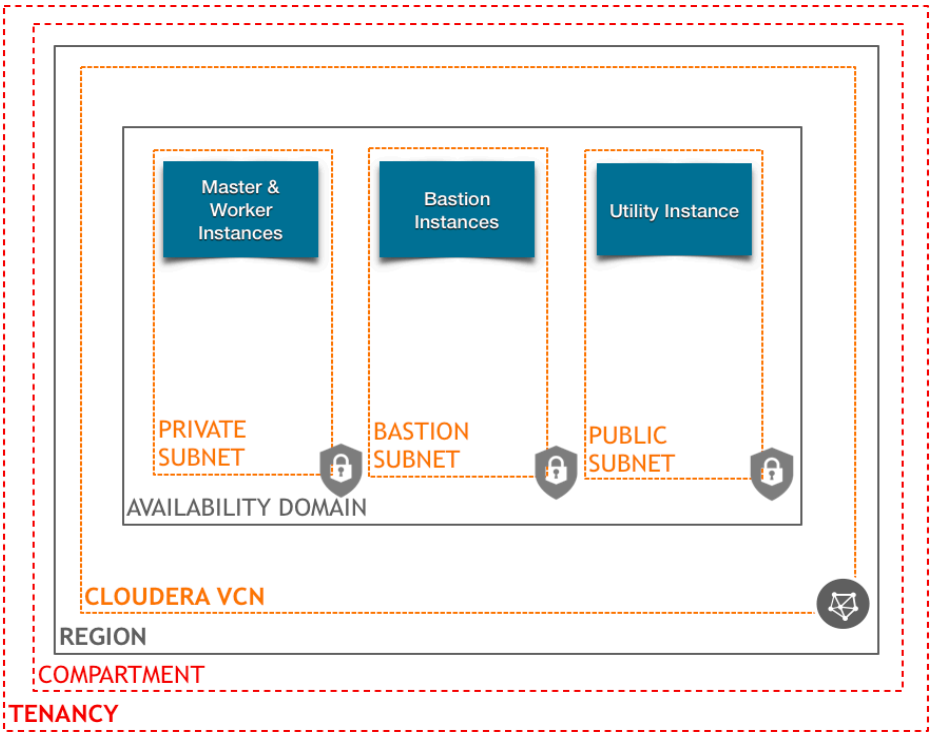
Automation deployment for Enterprise Data Hub on Oracle Cloud Infrastructure using Terraform creates the Cloudera VCN and associated subnets automatically. The network CIDR used for the VCN is an entire Class B 10-net, and each subnet is programmatically set as a unique Class C network member.

SSH access to hosts with public IP addresses is enabled by default, and a few specific ports are enabled with global access through security lists for ease of access. These configurations are customizable in post-deployment, and we recommend that you review the rules and adjust them to meet your network security requirements.

Network Topology

The recommended network topology for an Enterprise Data Hub deployment consists of a single VCN in the region that you choose. This VCN should contain nine subnets, three per availability domain for the bastion, public, and private networks. This model allows for granular control of hosts deployed in each subnet by using security lists.

The following diagram shows a single availability domain in this model, with host associations at the subnet level.



Connectivity and Security

Connectivity between hosts inside the VCN is controlled by a combination of security lists and local firewalls. This means that any connection between hosts is required to exist both in a security list and the local firewall on the hosts where the connection is needed. Security list rules are global in the sense that they allow a particular port or port range across all hosts inside the subnets associated with the security list. There is no host-level control at the security-list level; host-level control is applied only at the local firewall level. This makes it important to manage security lists in a manner that is most restrictive to allowed traffic into subnets that are publicly addressable.

This is why we recommend keeping host-level firewalls in place across all deployed hosts. Many Hadoop vendors suggest disabling the local firewall for connectivity, but that security model is appropriate only for non-cloud deployments. Connectivity at the host level can be whitelisted for internal networks in a broad manner, and fine-grained control for external access can also be applied. This is done with iptables (EL6) or firewalld (EL7). Primers for how to leverage connectivity using these firewalls can be found readily online.

Find out more about security best practices with automation by reading the readme.

Automated Cluster Deployment with Terraform and the Oracle Cloud Infrastructure Provider

Cloudera Enterprise Data Hub on Oracle Cloud Infrastructure using Terraform and the Oracle Cloud Infrastructure Provider allows for flexible deployments in several preset configurations. These configurations are available on GitHub. From provisioning to a fully ready cluster typically takes a half hour and requires minimal user interaction after setting up a few configuration values in the Terraform template.

Detailed steps for deploying Enterprise Data Hub on Oracle Cloud Infrastructure are located in the readme file available in Oracle's public GitHub repository. Deployment templates there leverage Terraform by Hashicorp. Detailed setup instructions for Terraform are located on the Terraform website, and complementary information is located on the Oracle Cloud Infrastructure Provider GitHub page.

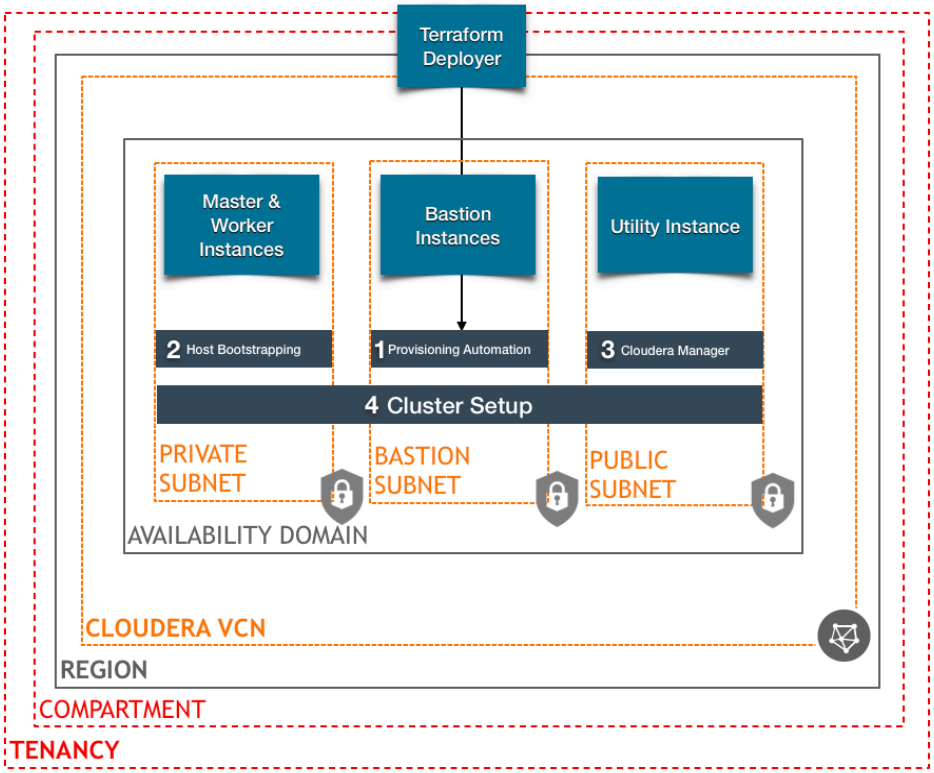
Installation Model Overview

At a high level, the deployment process leverages the Terraform Deployer to invoke Oracle Cloud Infrastructure API calls, which provision infrastructure inside the customer tenancy. A compartment is targeted for the deployment, where a VCN is set up with three subnets, which are duplicated across each availability domain to allow deployment to any availability domain in the region. A bastion subnet is set up for the bastion hosts, a public subnet is set up for a utility host, and a private subnet is set up for master and worker hosts. Hosts are then provisioned in these subnets in the target availability domain.

When all the infrastructure provisioning is complete, the following steps occur:

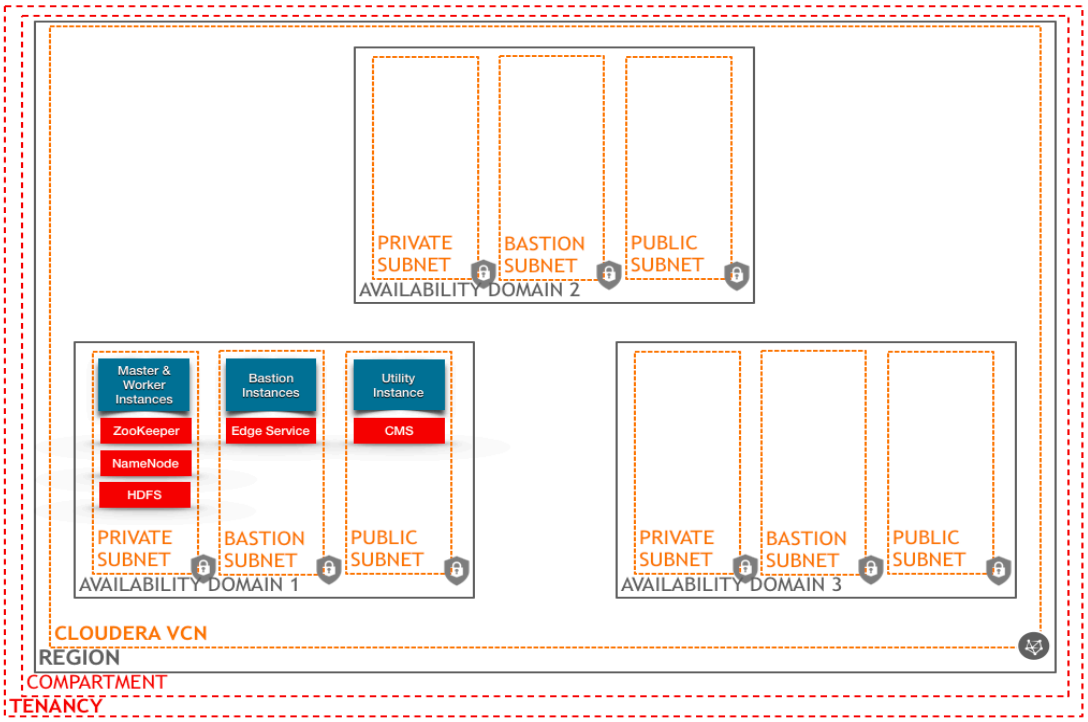
1. An automated setup script is triggered to run on the bastion host.
2. All the hosts in the deployment are bootstrapped.
3. The Cloudera Manager is installed and set up.
4. The Cloudera Manager sets up the cluster through a Python script, which invokes the Cloudera Manager API to configure and deploy Enterprise Data Hub.

This process is illustrated in the following figure.



Single Availability Domain Deployment Model

In the supported automated installation architecture, hosts are deployed and configured for Enterprise Data Hub in a single availability domain, as shown in the following figure.



Note: It is also possible to install Enterprise Data Hub to span availability domains in a region, which allows for additional fault tolerance and high availability. Although this model is not currently supported in the automated Terraform deployment, Oracle plans to release availability-domain spanning templates to support this automated deployment model in the future. For more information, see "Availability Domain Spanning Deployment Model" in the Appendix.

Terraform Templates

This section describes the templates that are available via Terraform and the Oracle Cloud Infrastructure Provider to automatically stand up an Enterprise Data Hub cluster on Oracle Cloud Infrastructure.

Sandbox

The sandbox deployment consists of a single instance running the Cloudera Docker Container. This deployment is a great starting point for customers who want to explore the power and functionality of Enterprise Data Hub on Oracle Cloud Infrastructure while maintaining a cost-

effective bottom line. This deployment is not a good fit for multiple users, development efforts, or large datasets.

- **Minimum shape:** VM.Standard1.8
- **Suggested shape:** VM.Standard2.8

Development

The development deployment consists of five instances: one bastion host, one utility host, and three workers. This environment provides a much higher HDFS storage capacity than the sandbox environment, with a good size pool of compute and memory resources for use with a variety of big data workloads. This environment is not a good fit for customers who want highly available services because the reduced infrastructure footprint does not support high availability.

Minimum Shapes

- **Worker:** BM.Standard1.36 with three 700-GB block storage devices per worker
- **Bastion:** VM.Standard1.4
- **Utility:** VM.Standard1.8

Suggested Shapes

- **Worker:** BM.Standard2.52 with three 1-TB block storage devices per worker
- **Bastion:** VM.Standard2.4
- **Utility:** VM.Standard2.8

Production Starter

The largest preset configuration for Enterprise Data Hub on Oracle Cloud Infrastructure, this deployment contains 10 instances: one bastion host, one utility host, two master hosts, and six workers. This environment provides the most density and best performance for Enterprise Data Hub on Oracle Cloud Infrastructure. This environment provides high availability and is an appropriate entry point for scaling up a production big data practice.

Minimum Shapes

- **Worker:** BM.DenseIO1.36
- **Bastion:** VM.Standard1.4
- **Utility and master:** VM.Standard1.8

Suggested Shapes

- **Worker:** BM.DenseIO2.52
- **Bastion:** VM.Standard2.4
- **Utility and master:** VM.Standard2.8

Custom (N-Node)

Oracle Cloud Infrastructure also supports N-Node Enterprise Data Hub implementations for customers whose needs might exceed the performance or capacity limitations of the largest preset cluster configuration. Contact Oracle Cloud Infrastructure for more information. We are happy to work with you to determine the optimal cluster deployment for your needs, and we have an automated solution to support dynamic cluster sizes scaling into the thousands of nodes.

Minimum Shapes

- **Worker:** BM.DenseIO1.36
- **Bastion:** VM.Standard1.4
- **Utility and master:** VM.Standard1.8

Suggested Shapes

- **Worker:** BM.DenseIO2.52
- **Bastion:** VM.Standard2.4
- **Utility and master:** VM.Standard2.16

Enterprise Data Hub Configuration Recommendations

The following recommendations are considered best practice when deploying Cloudera Enterprise Data Hub on Oracle Cloud Infrastructure.

HDFS

We recommend that HDFS be configured with a replication factor of 3 for bare metal Enterprise Data Hub clusters. Because bare metal hosts use local NVMe storage for HDFS, redundancy should be built in to the HDFS topology to ensure high availability and failure tolerance.

ZooKeeper

ZooKeeper is set up by default on the utility host and master hosts. An odd number of ZooKeeper instances should always be maintained to prevent split brain for service election.

NameNode

For high availability, multiple NameNodes should be provisioned as part of the Enterprise Data Hub deployment. This typically consists of a primary and secondary NameNode in active-standby configuration.

Summary

Automated deployment with Terraform on Oracle Cloud Infrastructure provides a flexible, highly scalable framework for Cloudera Enterprise Data Hub. Combined with bare metal performance on Oracle Cloud Infrastructure's fast network, this solution is excellent for customers who want to explore Cloudera on the Oracle Cloud Infrastructure platform, leverage cloud for a low-cost alternative to on-premises deployments, or offload entire Hadoop ecosystems to the cloud.

Cloudera Enterprise Data Hub on Oracle Cloud Infrastructure delivers a cost-effective, performant means to enable customer big data workloads in Cloud.

Appendix

Benefits of Running Cloudera on Oracle Cloud Infrastructure

Cloudera on Oracle Cloud Infrastructure is a joint solution between Cloudera and Oracle Cloud Infrastructure that combines the power of Oracle Cloud Infrastructure with the performance of Cloudera Enterprise Data Hub. This joint solution allows for large, scalable data management using Enterprise Data Hub, deployed to leverage the flexibility and performance of Oracle Cloud Infrastructure. This solution provides a powerful, cost-efficient, easy-to-manage platform for running diverse big data workloads in the cloud.

Cloudera is fast for business, providing support for the broadest range of use cases and verticals, including all of the Apache Hadoop Distribution core components. Cloudera offers the fastest SQL on Hadoop in the industry, with multiple frameworks to support diverse workloads.

Cloudera provides enterprise security, including audit and governance capabilities, enterprise encryption and key management, uniform access policy enforcement (role-based access control), and the capability to isolate specific environments to provide compartmentalization.

Deploying Cloudera Enterprise Data Hub on Oracle Cloud Infrastructure provides the following advantages:

- The lowest compute pricing from a pay-as-you-go (PAYG) perspective
- Additional discounts
- The lowest network egress costs in the industry
- The only provider with bare metal server performance, on demand
- Enhanced security inherent in bare metal architecture
- Reduced complexity of lift-and-shift with bare metal
- The only cloud provider that offers performance SLAs for:
    - Network throughput from an instance
    - Block storage IOPS and throughput
    - Local NVMe storage IOPS
- Industry leading storage capabilities, including:
    - Local NVMe storage: No other cloud provider recommends allowing data to be stored locally. Oracle ensures data locality to the bare metal instance, and provides service for the local storage.
    - Block storage: Guaranteed IOPS, at half the price of similar storage in the cloud provider market, ensures that Oracle is has the best price per performance storage offering available. This is backed by independent storage review testimonials.
    - Oracle bare metal instances leverage a 25-gigabit network backbone, allowing our biggest bare metal server to achieve unparalleled throughput on the cloud.
- Cloudera clusters that are spun up in the cloud sit next to Exadata or Oracle Database environments over private networks, allowing easy data sharing for analytics purposes
- Gartner regards Oracle as one of the top three vendors in the Data Management Storage Analytics space, making Cloudera on Oracle Cloud Infrastructure a great choice for running analytics workloads.

Oracle Cloud Infrastructure Terminology Reference

This section provides definitions for some of the terminology specific to Oracle Cloud Infrastructure.

Region and Availability Domain

Oracle Cloud Infrastructure is hosted in regions and availability domains. A region is a localized geographic area, and an availability domain is one or more data centers located within a region. A region is composed of several availability domains. Most Oracle Cloud Infrastructure resources are either region specific, such as a virtual cloud network, or availability domain specific, such as a compute instance or block storage volume.

Virtual Cloud Network

A virtual cloud network (VCN) is a customizable and private network in Oracle Cloud Infrastructure. Just like a traditional data center network, the VCN provides you with complete control over your network environment. This control includes assigning your own private IP address space, creating subnets and route tables, and configuring stateful firewalls. A single tenant can have multiple VCNs, thereby providing grouping and isolation of related resources.

Security List

A security list provides a virtual firewall for an instance, with ingress and egress rules that specify the types of traffic allowed in and out. Each security list is enforced at the instance level. However, you configure your security lists *at the subnet level*, which means that all instances in a given subnet are subject to the same set of rules. The security lists apply to a given instance whether it's talking with another instance in the VCN or a host outside the VCN.

Compute Instances

Oracle Cloud Infrastructure Compute lets you provision and manage compute hosts, known as instances. You can launch instances as needed to meet your compute and application requirements. After you launch an instance, you can access it securely from your computer, restart it, attach and detach volumes, and terminate it when you're done with it. Any changes made to the instance's local drives are lost when you terminate it. Any saved changes to volumes attached to the instance are retained.

Oracle Cloud Infrastructure offers both bare metal and virtual machine instances:

- **Bare metal:** A bare metal compute instance gives you dedicated physical server access for highest performance and strong isolation.
- **Virtual machine:** A virtual machine (VM) is an independent computing environment that runs on top of physical bare metal hardware. The virtualization makes it possible to run multiple VMs that are isolated from each other. VMs are ideal for running applications that do not require the performance and resources (CPU, memory, network bandwidth, storage) of an entire physical machine.

An Oracle Cloud Infrastructure VM compute instance runs on the same hardware as a bare metal instance, leveraging the same cloud-optimized hardware, firmware, software stack, and networking infrastructure.

Service Limits

When you sign up for Oracle Cloud Infrastructure, a set of service limits are configured for your tenancy. A service limit is the quota or allowance set on a resource. For example, your tenancy is allowed a maximum number of compute instances per availability domain. These limits are generally established with your Oracle sales representative when you purchase Oracle Cloud Infrastructure. If you did not establish limits with your Oracle sales representative, or if you signed up through the Oracle Store, default or trial limits are set for your tenancy. You can request to have a service limit raised.

Default service limits are available in the Oracle Cloud Infrastructure service documentation.

Identity and Access Management

Oracle Cloud Infrastructure Identity and Access Management (IAM) lets you control who has access to your cloud resources. You can control what type of access a group of users has and to which specific resources. You can write policies to control access to all of the services within Oracle Cloud Infrastructure, including Audit, Block Volume, Container Engine for Kubernetes, Compute, Database, DNS, Email Delivery, File Storage, IAM, Load Balancing, Object Storage, Networking, and Registry.

For more information, see the IAM documentation.

Availability Domain Spanning Deployment Model

Oracle Cloud Infrastructure has many regions where customers can deploy Enterprise Data Hub clusters. Each cluster is localized to that specific region but can span availability domains inside each region. This is the recommended deployment model for customers who want high availability and redundancy for their Enterprise Data Hub deployment. Hosts in each availability domain should be treated as a traditional “rack” when HDFS is deployed. This means that if you span all three availability domains in a region, you will have three “racks” in your HDFS topology. Replicas can then be distributed using this model, so that in the event of an outage at the availability-domain level, your cluster will maintain redundancy and availability.

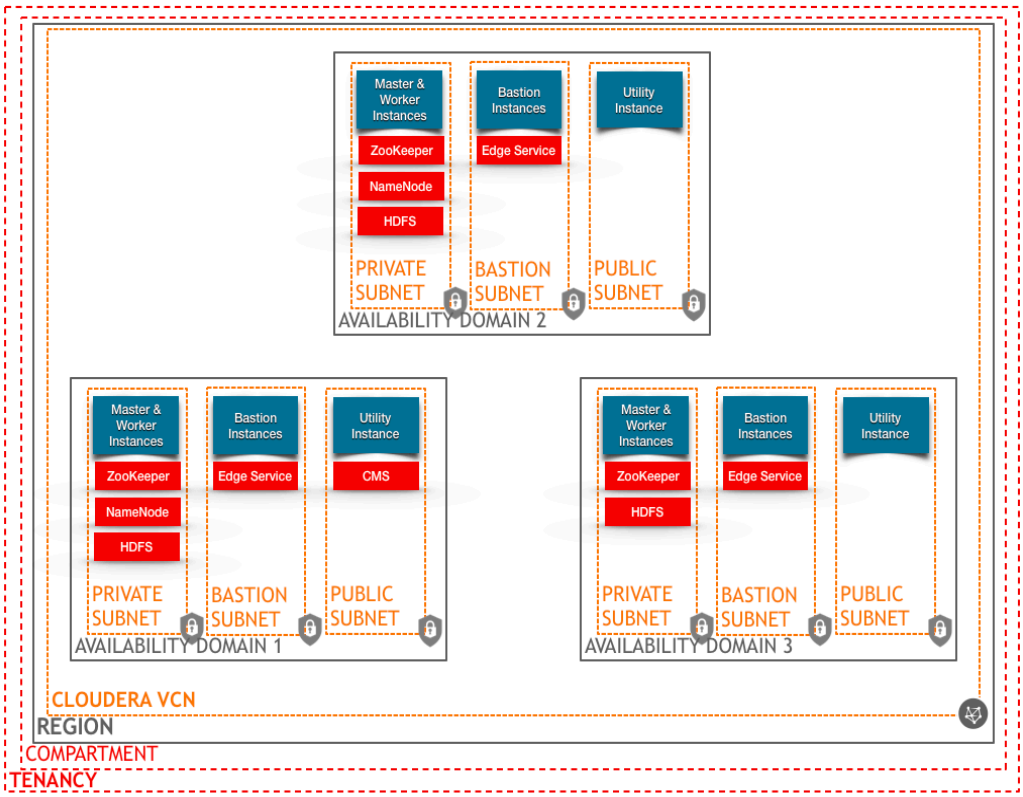
This functionality will be further enhanced with the forthcoming release of fault domains inside each availability domain, which allow for even more “rack topology” to be configured, providing enhanced high availability.

Although this model is not currently supported in the automated Terraform deployment, Oracle plans to release availability-domain spanning templates to support this automated deployment model in the future.

Availability-domain spanning is similar to the single-availability-domain model described in this paper, with a few differences.

- A bastion host is deployed in each availability domain to provide high availability for edge services.
- ZooKeeper hosts are spread out, one per availability domain, and NameNode hosts are deployed so that they are not in the same availability domain.
- Workers are distributed across availability domains, which allows them to be treated as distinct “racks” for setting up rack-awareness topology when using triple replication to distribute HDFS data copies for redundancy.

The following figure shows the architecture for this model.



References

For more information about Oracle Cloud Infrastructure, Cloudera, and topics discussed in this white paper, see the following resources:

- Cloudera website
- Cloudera documentation
- Cloudera Enterprise Reference Architecture for Bare Metal Deployments
- Oracle Cloud Infrastructure Documentation
- Oracle Cloud Infrastructure Provider GitHub
- Terraform automation templates for Cloudera on Oracle Cloud Infrastructure
- Bare metal and virtual machine shape reference charts

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

Cloudera Enterprise Data Hub Reference Architecture for Oracle Cloud Infrastructure Deployments
November 2018
Author: Zachary Smith