Cloudera Data Warehouse Public Cloud

Resource Planning

Date published: 2024-01-01
Date modified: 2024-01-01

Legal Notice

Contents

Virtual Warehouse sizing requirements for public cloud environments

You need to understand how to estimate size requirements for Cloudera Data Warehouse (CDW) Public Cloud Virtual Warehouses to create a Virtual Warehouse that meets your needs.

Calculating public cloud requirements for on-premises data warehouse deployments

Selecting the correct size of public cloud environment before you migrate your workloads from CDH and HDP to Cloudera Data Warehouse (CDW) Public Cloud is critical for preserving performance characteristics. Consider the following workload characteristics when you plan for capacity on your public cloud environment:

- Query memory requirements
- CPU utilizations
- Disk bandwidth
- Working set size
- Concurrent query execution requirements

As part of this calculation, it is important to understand the core hardware difference between public cloud and on-premises hosts as explained in the following table:

Table 1: Hardware differences between public cloud and on-premises hosts

Hardware component	On-premises	AWS R5D.4xlarge instance	Azure E16 v3 instance
CPU cores	20 - 80	16	16
Memory	128 GB minimum 256 GB+ recommended	128 GB	128 GB
Network	10 Gbps minimum 40 Gbps recommended	Up to 10 Gbps	Up to 8,000 Mbps
Instance storage	12 x 2 TB drives (1,000 MB/s sequential)	2 x 300 GB NVMe SSD (1,100 MB/s sequential) **Note:** Optimally EBS volumes can be added to scratch and spool space for Impala.	400 GiB SSD
Persistent storage performance	At least 500 MB/s per disk With 20 disks, 10 GB/s per node	1,156 MB/s per EC2 instance	Maximum IOPS: 24,000 Maximum Read: 375 MB/s Maximum Write: 187 MB/s

An AWS R5D.4xlarge instance closely matches the CPU, memory, and bandwidth specifications that are recommended for CDH clusters. AWS R5D.4xlarge instance specifications are the default "instance type" for CDP. AWS EBS storage cannot be used as primary database storage because it is transient and lacks sufficient capacity. This core difference makes it necessary to use a different strategy for CDP than you used for CDH to achieve good scan performance.

Supported AWS instances

You can choose an instance type other than the default AWS R5D.4xlarge , and also a fallback instance type, using the UI or CDP CLI. Select the instance type using the UI when you activate the environment in CDP. To fetch a string of available instance types, use the following CLI command:

```
cdp dw describe-allowed-instance-types
```

Example output:

```
{
     "aws": {
     "default": [
     "r5d.4xlarge"
     ],
     "allowed": [
     "r5d.4xlarge",
     ...
```

For more information, see the DWX-1.5.1 November 22, 2022 release notes.

Public cloud sizing and scaling

Before you migrate data to CDW Public Cloud, plan for scaling and concurrency. In the cloud, scaling and concurrency can elastically respond to workload demands, which enables the system to operate at a lower cost than you might expect. If you configure your environment to accommodate peak workloads as a constant default configuration, you may waste resources and money when system demand falls below that level.

In CDW, the size of the Virtual Warehouse determines the number of executor instances for an individual cluster so it determines memory limits and performance capabilities of individual queries:

Table 2: Virtual Warehouse sizes

Size	Number of executors
X-Small	2
Small	10
Medium	20
Large	40

Warehouse size in combination with auto-scaling settings determine how many clusters are allocated to support concurrent query execution.

The Virtual Warehouse size must be at least large enough to support the memory used by the most data-intensive query. Usually, the Virtual Warehouse size does not need to be larger that the most data-intensive query. Better caching is provided if there is commonality between data sets accessed by queries. Increasing the Virtual Warehouse size can increase single-user and multi-user capacity because additional memory and resources allows larger datasets to be processed. Concurrent query execution is also supported by sharing resources. If too small a size is configured for the Virtual Warehouse, poor data caching and memory paging can result. If too large a size is configured, excessive public cloud costs are incurred due to idle executors.

The primary difference between CDW Public Cloud and CDH on-premises deployments when choosing a warehouse size based on existing hardware:

- With CDW Public Cloud all resources on an executor are dedicated to query processing.
- With CDH on-premises deployments resources support other operations in addition to query processing. For example, these on-premises resources are shared with other services, such as HDFS or other locally hosted file systems. In particular Spark, HBase, or MapReduce. These other services might consume significant resources.

Consequently, you might be able to choose a much smaller Virtual Warehouse size in CDW Public Cloud because resources are isolated in their own pod in the CDW Public Cloud environment.

In the case of Impala, it is useful to look at the Cloudera Manager per-process metrics to isolate the impalad backend and the Impala front-end Java processes that hold the Catalog cache. In CDW Public Cloud, the Impala coordinator and executor roles are separated leaving the unused Catalog JVM memory free to support query execution. You should look at the memory utilization metrics for executor-only impalad nodes (those not also running the coordinator role) to estimate how much memory your current cluster of Impala executors requires.

Concurrency

Concurrency is the number of queries that can be run at the same time. Determine the size you need by considering the amount of resources your system needs to support peak concurrency.

By default, Impala Virtual Warehouses can run 3 large queries per executor group. Executors can handle more queries that are simpler and that do not utilize concurrency on the executor. When you enable legacy multithreading, the Virtual Warehouse can run 12 queries per executor group. For most read-only queries the default setting of 3 queries per executor group is sufficient. Occasional peaks are handled transparently by the auto-scaling feature. When auto-scaling is triggered an additional executor group is added thereby doubling query concurrency capacity. Scaling the Virtual Warehouse by adding more clusters enables additional concurrent queries to run, but does not improve single-user capacity or performance. Concurrently executed queries are routed to the different clusters and execute independently. The number of clusters can be changed to match concurrent usage by changing the auto-scaling parameters. For more details about auto-scaling settings, see the links at the bottom of this page.

Important: The number of very large queries that can be run might also be impacted by memory limits on the executors.

For Hive on LLAP Virtual Warehouses, each size setting indicates the number of concurrent queries that can be run. For example, an X-Small Hive on LLAP Virtual Warehouse can run 2 concurrent queries. A Small Virtual Warehouse can run 10 concurrent queries. To run 20 concurrent queries in a Hive on LLAP Virtual Warehouse choose Medium size.

Caching "Hot Datasets"

Frequently accessed data is sometimes referred to as a "hot dataset." CDH supports caching mechanisms on the compute nodes to cache the working set that is read from remote file systems, such as remote HDFS data nodes, S3, ABFS, or ADLS. This offsets the input/output performance difference.

In CDW Public Cloud, frequently accessed data is cached in a storage layer on SSD so that it can be quickly retrieved for subsequent queries. This boosts performance. Each executor can have up to 200 GB of cache. For example, a Medium-sized Virtual Warehouse can keep 200 * 20 = 4 TB of data in its cache. For columnar formats, such as ORC, data in the cache is decompressed, but not decoded. If the expected size of the hot dataset is 6 TB, which requires about 30 executors, you can over-provision by choosing a Large-sized warehouse to ensure full cache coverage. A case can also be made to under-provision by choosing a Medium-sized warehouse to reduce costs at the expense of having a lower cache hit rate. To offset this, keep in mind that columnar formats allow optimizations such as column projection and predicate push-down, which can significantly reduce cache requirements. In these cases, under-provisioning the cache might have no negative performance effect.

Scanned dataset size

Scanning large datasets on Amazon S3 or Azure ADLS can be slow. A single R5D.4xlarge EC2 instance can only scan data at 1,156 MB/s maximum throughput according to standard S3 benchmarks. If a query must read 100 GB from S3, S3 scanning takes a minimum of 88 seconds on just one node. Depending on the number of files in the S3 directory, it might take more than 3 minutes. In this case, if a query needs to scan 100 GB of data, if you use 10 nodes, you can get the scan time down to approximately 20 seconds.

However, keep in mind that with columnar storage, minimum/maximum statistics in files, and other conditions, often the amount of data read is significantly less than the total size of the files for certain queries.

Related Information

Tuning Hive Virtual Warehouses on public clouds

Tuning Impala Virtual Warehouses on public clouds

Virtual Warehouse IP address and cloud resource requirements for public cloud environments

Learn about how many IP addresses and cloud resources are required to run Virtual Warehouses efficiently in public cloud environments for Cloudera Data Warehouse (CDW) Public Cloud.

IP address and cloud resource requirements for Virtual Warehouses running on AWS environments

Learn about the estimated number of IP addresses and cloud resources required to run Virtual Warehouses on AWS environments for Cloudera Data Warehouse (CDW) Public Cloud.

Important: These requirements are estimated. Your particular workloads and configurations can affect the number of IP addresses and cloud resources required to run CDW Virtual Warehouses efficiently. In addition, these requirements are for AWS environments that use the default AWS VPC Container Networking Interface (CNI) plugin. To reduce the number of required IP addresses, you can enable the overlay networking feature for AWS environments in CDW. For further details, see the link to "Overlay networks for AWS environments in CDW" at the bottom of this page.

Virtual Warehouse requirements:

Each compute node in a Virtual Warehouse that runs on AWS environments requires 8 IP addresses. Each executor needs one compute node, so the size of your Virtual Warehouse contributes to the number of IP addresses required. To calculate the number of IP addresses required for custom sizes, multiply the number of executors by 8 and add for the shared services as specified in the following sections.

Shared services requirements for Database Catalogs:

Virtual Warehouses also require shared services for the Database Catalog.

Additional shared services requirements:

The usage for other shared services is different for Hive Virtual Warehouses and Impala Virtual Warehouses. Here are the different requirements for Hive versus Impala Virtual Warehouses:

- Hive Virtual Warehouses add 1 compute node for each executor and 1 shared services node for the HiveServer for each Virtual Warehouse.
- Impala Virtual Warehouses add 1 compute node for each executor, 1 or 2 compute nodes for the coordinator, depending on the HA configuration, and 1 shared services node per Virtual Warehouse for Impala catalogd.

The following tables summarize the approximate number of IP addresses and cloud resources you should plan for Virtual Warehouses on AWS environments.

Table 3: Hive Virtual Warehouses running on AWS environments

Size (# executors)	# Compute nodes for executors	# Shared services nodes Database Catalog	# Shared services nodes for HiveServer	Total IP addresses required
XSMALL (2)	2	3	1	Executor nodes: 2 nodes X 8 = 16 Shared services nodes: 3 + 1 = 4 nodes X 25 = 100 TOTAL = ~116 IP addresses
SMALL (10)	10	3	1	Executor nodes: 10 nodes X 8 = 80 Shared services nodes: 3 + 1 = 4 nodes X 25 = 100 TOTAL = 180 IP addresses
MED (20)	20	3	1	Executor nodes: 20 nodes X 8 = 160 Shared services nodes: 3 + 1 = 4 nodes X 25 = 100 TOTAL = 260 IP addresses
LARGE (40)	40	3	1	Executor nodes: 40 nodes X 8 = 320 Shared services nodes: 3 + 1 = 4 nodes X 25 = 100 TOTAL = 420 IP addresses

Table 4: Impala Virtual Warehouses running on AWS environments

Size (# executors)	# Compute nodes for executors	# Compute nodes for coordinator	# Shared services nodes for Impala catalogd	# Shared services nodes for Database Catalog	Total IP addresses required
XSMALL (2)	2	1-2	1	3	Executor/ coordinator nodes: 3-4 nodes X 8 = 24-32 Shared services nodes: 1 + 3 = 4 nodes X 25 = 100 TOTAL = 124 to 132 IP addresses
SMALL (10)	10	1-2	1	3	Executor/ coordinator nodes: 11-12 nodes X 8 = 88-96 Shared services nodes: 1 + 3 = 4 nodes X 25 = 100 TOTAL = 188 to 196 IP addresses

Size (# executors)	# Compute nodes for executors	# Compute nodes for coordinator	# Shared services nodes for Impala catalogd	# Shared services nodes for Database Catalog	Total IP addresses required
MED (20)	20	1-2	1	3	Executor/ coordinator nodes: 21-22 nodes X 8 = 168-176 Shared services nodes: 1 + 3 = 4 nodes X 25 = 100 TOTAL = 168 to 176 IP addresses
LARGE (40)	40	1-2	1	3	Executor/ coordinator nodes: 41-42 nodes X 8 = 328-336 Shared services nodes: 1 + 3 = 4 nodes X 25 = 100 TOTAL = 428 to 436 IP addresses

Related Information

Overlay networks for AWS environments in CDW

IP address and cloud resource requirements for Virtual Warehouses running on Azure environments

Learn about the estimated number of IP addresses and cloud resources required to run Virtual Warehouses on Azure environments for Cloudera Data Warehouse (CDW) Public Cloud.

Important: These requirements are estimated. Your particular workloads and configurations can affect the number of IP addresses and cloud resources required to run CDW Virtual Warehouses efficiently. In addition, these requirements are for Azure environments that use the default Azure Container Networking Interface (CNI) plugin. To reduce the number of required IP addresses, you can enable the kubenet networking feature for Azure environments in CDW.

Virtual Warehouse requirements:

Each compute node in a Virtual Warehouse that runs on Azure environments requires 16 IP addresses (for pods and for the node itself). Each executor needs one compute node, so the size of your Virtual Warehouse contributes to the number of IP addresses required. To calculate the number of IP addresses required for custom sizes, multiply the number of executors by 16 and add for the shared services as specified in the following sections.

Shared services requirements for Database Catalogs:

Virtual Warehouses also require shared services for Database Catalogs.

Additional shared services requirements:

The usage for other shared services is different for Hive Virtual Warehouses and Impala Virtual Warehouses. Here are the different requirements for Hive versus Impala Virtual Warehouses:

- Hive Virtual Warehouses add 1 compute node for each executor and 1 shared services node for the HiveServer for each Virtual Warehouse.
- Impala Virtual Warehouses add 1 compute node for each executor, 1 or 2 compute nodes for the coordinator, depending on the HA configuration, and 1 shared services node per Virtual Warehouse for Impala catalogd.

The following tables summarize the approximate number of IP addresses and cloud resources you should plan for Virtual Warehouses on Azure environments.

Table 5: Hive Virtual Warehouses running on Azure environments

Size (# executors)	# Compute nodes for executors	# Shared services nodes for Database Catalog	# Shared services nodes for HiveServer	Total IP addresses required
XSMALL (2)	2	3	1	Executor nodes: 2 nodes X 16 = 32 Shared services nodes: 3 + 1 = 4 nodes X 31 = 124 TOTAL = 156 IP addresses
SMALL (10)	10	3	1	Executor nodes: 10 nodes X 16 = 160 Shared services nodes: 3 + 1 = 4 nodes X 31 = 124 TOTAL = 284 IP addresses
MED (20)	20	3	1	Executor nodes: 20 nodes X 16 = 320 Shared services nodes: 3 + 1 = 4 nodes X 31 = 124 TOTAL = 444 IP addresses
LARGE (40)	40	3	1	Executor nodes: 40 nodes X 16 = 640 Shared services nodes: 3 + 1 = 4 nodes X 31 = 124 TOTAL = 764 IP addresses

Table 6: Impala Virtual Warehouses running on Azure environments

Size (# executors)	# Compute nodes for executors	# Compute nodes for coordinator	# Shared services nodes for Impala catalogd	# Shared services nodes for Database Catalog	Total IP addresses required
XSMALL (2)	2	1-2	1	3	Executor/ coordinator nodes: 3-4 nodes X 16 = 48-64 Shared services nodes: 1 + 3 = 4 nodes X 31 = 124 TOTAL = 172 to 188 IP addresses
SMALL (10)	10	1-2	1	3	Executor/ coordinator nodes: 11-12 nodes X 16 = 176-192 Shared services nodes: 1 + 3 = 4 nodes X 31 = 124 TOTAL = 300 to 316 IP addresses

Size (# executors)	# Compute nodes for executors	# Compute nodes for coordinator	# Shared services nodes for Impala catalogd	# Shared services nodes for Database Catalog	Total IP addresses required
MED (20)	20	1-2	1	3	Executor/ coordinator nodes: 21-22 nodes X 16 = 336-352 Shared services nodes: 1 + 3 = 4 nodes X 31 = 124 TOTAL = 460 to 476 IP addresses
LARGE (40)	40	1-2	1	3	Executor/ coordinator nodes: 41-42 nodes X 16 = 656-672 Shared services nodes: 1 + 3 = 4 nodes X 31 = 124 TOTAL = 780 to 796 IP addresses

Related Information

Overlay networks for Azure environments in CDW

Managing costs in the public cloud environments for Cloudera Data Warehouse

Cost optimization in cloud environments is top priority for enterprises. Eighty percent of cloud costs are determined by the number of compute instances you use. Compute instances can be virtual machines or containers. Cloudera Data Warehouse (CDW) Public Cloud and your cloud provider give you ways to monitor and control the cloud resources you use.

Setting resource limits with Cloudera Data Warehouse service

Cloudera Data Warehouse service provides the following ways to manage your cloud costs:

- Choose Virtual Warehouse size: Virtual Warehouse size specifies the number of executor nodes used by the Virtual Warehouse, which translates to compute instances. Before you create a Virtual Warehouse, determine the number of concurrent queries or users your Virtual Warehouse must serve during peak periods. This information helps you determine what size of Virtual Warehouse you need. Choose the size based on the number of nodes you typically use for clusters in an on-premises deployment.
- Set auto-scaling thresholds: When you create a Virtual Warehouse, you can define auto-scaling, which sets limits on how many cloud resources can be consumed to meet workload demands. In addition, you can also set the maximum time a Virtual Warehouse idles before shutting down. Both settings ensure that you only use cloud resources that you need when you need them, helping you to manage your costs in the cloud.

Setting resource limits with your cloud provider

Another way to manage your cloud costs is by setting effective resource limits with your cloud provider. Cloud providers offer ways to set the overall limit on numbers of virtual machines or containers that can be used for your account. You can also save on cloud expenses by shutting down resources when they are not in use. CDP has built-in functionality to shut down cloud resources when not in use. If necessary, you can also terminate resources to save on costs related to disk snapshots, reserved IP addresses, and so on. In addition, cloud providers might offer tools to help you save. For example, AWS offers AWS Trusted Advisor and CloudWatch. Microsoft Azure offers Azure Cost Management and Azure Advisor.

Always active, shared services

Several shared service nodes are always active in a Cloudera Data Warehouse environment. A number of infrastructure pods are required at all times, and sufficient nodes must be deployed for these pods. The following table lists actions and the impact of those actions on shared service nodes.

Action	Impact on Shared Services Nodes	Comments
Activating a Cloudera Data Warehouse environment	Yes	Starts with minimum 3 nodes
Creating Database Catalog	Maybe	It depends on the pod placement logic
Creating Impala/Hive Virtual Warehouse	Yes	This also adds more shared service pods to support compute nodes (weavenet, kube-proxy etc.)
Deleting Impala/Hive Virtual Warehouse	Maybe	It depends on the pod placement logic. “Cooldown” period also plays a role.
Compute autoscaling	Yes	Autoscaling compute nodes will add hue, hs2 pods.

The following scaling factors also affect shared services nodes.

- The number of Virtual Warehouses you run increases the required shared service resources.
- The T-shirt you set for Virtual Warehouses and autoscaling on executors has impact as shown in the following example:

  Hue instance (part of shared service) count = (query-executors/10) + 1

  => from 2 to 10 query executors then 1 Hue server Pod instance,

  => from 10 to 19 query executors then 2 Hue server Pod instances

  => from 20 to 29 query executors then 3 Hue server Pod instances

Related Information

AWS Trusted Advisor
AWS CloudWatch
Azure Cost Management
Azure Advisor
Adding a new Virtual Warehouse
Tuning auto-scaling for Virtual Warehouses