Cloudera Semantic Search user guide for Cloudera on premises environments (Preview)

Date published: Dec 3, 2024
Date modified: Apr 16, 2026

Legal Notice

This document has been released as part of a technical preview for features described herein. Technical preview components are provided as a convenience to our customers for their evaluation and trial usage. These components are provided ‘as is’ without warranty or support. Further, Cloudera assumes no liability for the usage of technical preview components, which should be used by customers at their own risk.

Contents

Note:
This release of the Cloudera Semantic Search service for Cloudera on-premises deployments is currently available as a technical preview. Technical preview releases are considered developmental and are not intended for use in production environments.

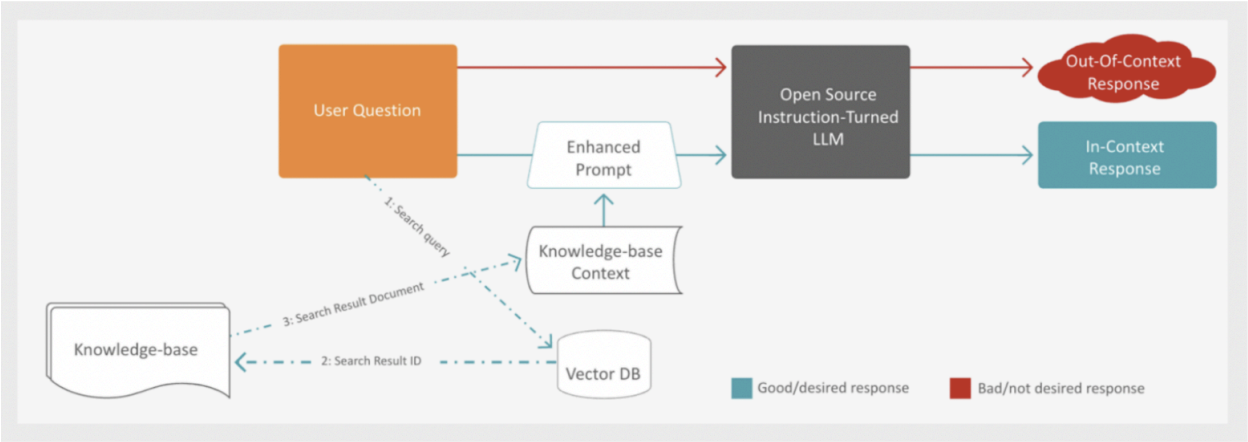
Cloudera Semantic Search service overview

Background information

Cloudera Semantic Search is the strategic evolution of our text search solution into a highly scalable, hybrid multimodal Vector Database and Semantic retrieval platform powered by OpenSearch. It acts as the secure "memory" and retrieval foundation for the entire Cloudera portfolio, transforming how enterprises power Retrieval-Augmented Generation (RAG) and autonomous AI agents, along with anomaly detection for time-series data. This functionality supports indexing, K-NN search, and Neural search. The underlying approach leverages an embedding model, which uses Cloudera AI to generate vector representations (embeddings) of documents or knowledge sources. These vectors are stored and indexed, enabling searching and retrieval to find documents semantically similar to an input query. Furthermore, Neural search extends K-NN search by supporting various sorts of searches, such as semantic, multimodal, and hybrid search, using pre-trained, custom, or external foundation models.

Workflow

The following diagram illustrates the Retrieval Augmented Generation (RAG) workflow:



The following typical Retrieval Augmented Generation workflow is valid for a search functionality:

1. Indexing of documents
    - An embedding model is selected to generate a vector representation of a source document, image, or video frame.
      Note:
      The term "document" refers to a data unit and does not imply only text. The document could be either structured or unstructured data.
    - The embedding model generates vector representations (embeddings) of the document collection.
    - The vectors are stored and indexed for fast search and retrieval.

2. Search and retrieval
    - A user enters an input query.
    - The same embedding model used for indexing the documents generates the embedding for the input query.
    - This embedding, which is a vector, is passed to the semantic search engine.
    - The semantic search engine searches across the stored vector representations of documents and identifies the top-k, which is the top 1 or more documents that are semantically closest to the input embedding.
    - The search engine returns the resultset, which can then be used to retrieve the source documents.

For more information, see Deploying an LLM ChatBot Augmented with Enterprise Data.

Use cases

Semantic search is a crucial technology in generative artificial intelligence (GenAI) use cases. Using RAG, enterprises can enhance the functionalities of large language models (LLMs), such as ChatGPT, with enterprise-specific knowledge. Support for an RAG architecture generally involves retrieving documents that are semantically similar to a given input. For example, the Cloudera Large Language Model (LLM) Applied Machine Learning Prototypes (AMPs), augmented with enterprise data, leverage semantic search and are built around an RAG (Retrieval Augmented Generation) architecture.

Semantic search use cases can provide more than use cases built around the RAG architecture and ChatBots, including recommendation engines, sentiment analysis, and various automated categorization tasks, such as customer feedback, product reviews, and news articles.

Having the semantic search engine and the underlying vector store capability as part of Cloudera helps maintain your data gravity within Cloudera. This means that you can leverage semantic search against the entire data estate, including all knowledge sources, such as websites, documentation, customer support systems, and internal data systems, such as data

lake houses, and Slack channels of your organization, without having to transfer this data or its vectorized representation outside the organization. This approach ensures that the security and authorization levels of the source data remain intact.

Cloudera Semantic Search includes vector search functionality, which is exposed as part of the Cloudera service, including the following functionalities:

- Indexing
  Indexes are the mechanism for storing and searching vector data.
- K-NN search
- Neural search.
  Neural search builds on K-NN search functionality by allowing users to use pre-trained models, upload custom models, or connect to foundation models hosted on external platforms. Various sorts of searches are supported, such as semantic, multimodal, and hybrid search.

System requirements

Refer to the following topics to learn more about the hardware, operating system, and database requirements for installing and deploying the Cloudera Semantic Search feature in your Cloudera on premises environments.

Hardware requirements

Learn the hardware requirements to install and deploy the Cloudera Semantic Search feature in your Cloudera on premises environments.

Note:

- The hardware configurations and sizes depend on several factors, including the number of documents, vector dimensions, replicas, shards, and the chosen engine type and method. Cloudera recommends the hardware configurations in the table Hardware configurations for general application deployments. However, these configurations must be validated for your specific size and scale requirements and must adhere to the provided sizing guidelines.
- Each role must be assigned to a separate machine. While the same role can be distributed across multiple environments based on size requirements, Cloudera recommends not running various Cloudera Semantic Search roles on the same production environment.

For more information, see the OpenSearch documentation.

Table 2.1 Hardware configurations

Role	Java Heap	CPU	Disk	Default HTTP Port	Purpose	Importance
cluster_manager	Min: 4 GB Max: 16 GB Recommended: 4 GB	Min 4 cores. Add more for large clusters or bulk loads.	Min: 20 GB	9200	It manages the cluster's overall operation and state by performing the following actions: creating and deleting indexes, tracking node joins and leaves, checking	Required

					node health (using ping requests), and allocating shards to nodes.	
Data	Min: 8 GB Max: 32 GB Recommended: 50% of system RAM	Min 4 cores. Add more for large clusters or bulk loads.	Min: 20 GB	9201	Worker nodes store and search data. They perform all data-related operations, such as indexing, searching, and aggregating, on local shards, and therefore require more disk space than any other node type.	Required
Coordinator	Min: 8 GB Max: 32 GB Recommended: 50%-70% of system RAM	Min 4 cores. Add more for large clusters or bulk loads.	Min 20 GB	9202	Delegates client requests to the shards on the data nodes, aggregates the results, and returns the final result to the client.	Optional
ML (Machine Learning)	Min: 8 GB Max: 32 GB Recommended: 50% of system RAM	Min 4 cores. Add more for large clusters or bulk loads.	Min 20 GB	9203	It delegates a specific node for custom work, such as machine learning (ML) tasks,	Optional

					to prevent resource consumption from data nodes and avoid affecting OpenSearch functionality.	
Ingest	Min: 8 GB Max: 32 GB Recommended: 50%-70% of system RAM	Min 4 cores. Add more for large clusters or bulk loads.	Min 20 GB	9204	It pre-processes data by running an ingest pipeline that transforms the data before storing it in an index.	Optional
Dashboard	Min: 4 GB Max: 16 GB	Min 4 cores. Add more for large clusters.	Min 20 GB	5601	Allows you to search, view, and interact with data, and easily create dynamic dashboards for monitoring and analyzing semantic search data.	Optional

Software requirements

Learn about the operating system, database, and third-party software requirements for installing and deploying the Cloudera Semantic Search feature in your Cloudera on premises environments.

Table 2.2 Software configurations

	Supported versions	Note
Cloudera Runtime	Cloudera on premises from 7.3.2	-
Cloudera Manager	7.13.2	
Operating System	RHEL 8 RHEL 9	Supported on the operating system with the same minor versions for the Cloudera installation, with which the Cloudera Semantic Search feature is installed.
JDK	Java 17	Support for the Cloudera install, with which the Cloudera Semantic Search feature is installed.

Important settings

- For production workloads executing on a Linux operating system, it is requisite that the Linux kernel setting `vm.max_map_count` is configured with a minimum value of 262144.

  To check the current value, run the following command:

```
None
cat /proc/sys/vm/max_map_count
```

  To increase the value, add the following line to `/etc/sysctl.conf`:

```
None
vm.max_map_count=262144
```

  Then run sudo `sysctl -p` to reload.

- You must enable the memory locking for the semantic search user. On Linux systems, the memory can be locked to prevent the operating system from paging out heap or off-heap memory. This is a recommended setting for semantic search.
  To ensure that the semantic search user can lock an unlimited amount of memory, add the following lines to `/etc/security/limits.conf` on all hosts.
  ```
  opensearch soft memlock unlimited
  opensearch hard memlock unlimited
  ```

  When the Cloudera Semantic Search service is installed, you can verify these settings using the following API call:

```
None
GET _nodes?filter_path=nodes.*.process.mlockall
```

You must get the following response:

```
None
"nodes": {
"IjE4z1fFTnKjhc01W4Upqw": {
"process": {
"mlockall": true
}
},
```

Getting started

Install Cloudera on premises

Before installing the Cloudera Semantic Search parcel, you must install Cloudera on premises and add the mandatory services.

The Cloudera service is only supported on Cloudera on premises. For a full cluster installation, Cloudera recommends following the Cloudera Base on premises Installation Guide.

Preparing for installing Cloudera Semantic Search

Before installing the Cloudera Semantic Search service in your Cloudera on premises environment, you must perform the following steps:

1. Install the Cloudera Manager.
   The following is an example URL of the Cloudera Manager instance.

```
None
https://<CM_host.domain_name>:7183/cmf/
```

2. Install Cloudera cluster services.
3. Install the Cloudera Semantic Search CSD.
4. Install the Cloudera Semantic Search parcel.
5. Verify your installation.

Install and deploy

Installing the Cloudera Semantic Search CSD file

Cloudera Semantic Search is an add-on service, so you must install its Custom Services Descriptor (CSD) file to manage it on the Cloudera service.

About this task

A CSD file contains all the configurations needed to describe and manage a new service. A CSD is provided in the form of a JAR file. For more information about CSD files, see the Add-on Services documentation, *Custom Service Descriptor Files* section.

Before you begin

The Cloudera Manager must be installed.

Steps

1. Download the jar file for the CSD that provides the Cloudera Semantic Search on the Cloudera service.

   For example,
   You can download the JAR artifact from the following location:

```
None
https://archive.cloudera.com/p/solr/0.1.0/CSD
```

2. Copy the jar file to the location where Cloudera Manager is configured to look for CSD files. By default, that is the `/opt/cloudera/csd` folder on the host where Cloudera Manager is running.

3. Restart Cloudera Manager as described in the *Configuring the Location of Customer Service* section in the Add-on Services documentation by using the following command:

```
None
sudo systemctl restart cloudera-scm-server
```

After you finish

Install Cloudera Semantic Search parcel.

Installing the Cloudera Semantic Search parcel

After you have installed the Cloudera Semantic Search CSD and the required Cloudera cluster and services, you must install the Cloudera Semantic Search parcel.

There are two options to install the Cloudera Semantic Search parcel:

- Using a Local Parcel Repository
- Using Remote Parcel Repository URLs

Install Option 1: Installing Cloudera Semantic Search parcel using the Local Parcel Repository

You can install the Cloudera Semantic Search parcel using your Local Parcel Repository.

Before you begin

- The Cloudera Manager must be installed.
- The Cloudera Semantic Search CSD file must be installed.

Steps

4. Download both the parcel and the `sha` file from the following location.

```
None
https://archive.cloudera.com/p/solr/0.1.0/parcel
```

1. Copy the parcel and the sha file associated with the parcel file to the Local Parcel Repository. By default, the Local Parcel Repository is located in the `/opt/cloudera/parcel-repo` directory on the host where the Cloudera Manager is running.

   Restart the Cloudera Manager Server by running the following command on the RHEL 8/9 compatible Cloudera Manager Server host:

```
None
sudo systemctl restart cloudera-scm-server
```

   and ensure that the cluster is accessible again.

2. In the Cloudera Manager, navigate to **Parcels**.

3. Click the **Check for New Parcels** button.
   A Semantic Search parcel appears in the **Parcel** list.
4. In the Semantic Search parcel row, click the **Distribute** button.
5. After the distribution is completed, click the **Activate** button.

   Result:
   The Cloudera Semantic Search parcel is installed.

After you finish

1. Address the stale configurations by restarting the Cloudera Management Service. For more information, see Restarting the Cloudera Management Service.
2. Add the Cloudera Semantic Search to the Cloudera service.

Install Option 2: Installing Cloudera Semantic Search using the Remote Parcel Repository

You can install the Cloudera Semantic Search parcel using the archive.cloudera.com Remote Parcel Repository.

Before you begin

- The Cloudera Manager must be installed.
- The Cloudera Semantic Search CSD file must be installed.

Steps

1. In the Cloudera Manager, navigate to **Parcels**.
2. Click the **Parcel Repositories & Network Settings** button.
3. In the Remote Parcel Repository URLs section, click + to include another remote parcel repository URL.
4. Add the following URL:

   For RHEL9:

```
None
https://archive.cloudera.com/p/solr/0.1.0/parcel/redhat9
```

   For RHEl8:

```
None
https://archive.cloudera.com/p/solr/0.1.0/parcel/redhat8
```

5. Click the Save & Verify Configuration button.
6. In the Semantic Search parcel row, click the Download button.
7. After the download is completed, click the Distribute button.
8. Once the distribution is completed, click the Activate button.

After you finish

1. Address the stale configurations by restarting the Cloudera Management Service. For more information, see Restarting the Cloudera Management Service.
2. Add the Cloudera Semantic Search to the Cloudera service.

Adding Cloudera Semantic Search parcel as a Cloudera service

After you have installed the Cloudera Semantic Search CSD and the Cloudera Semantic Search parcel, you can add the Cloudera Semantic Search on the Cloudera service to your cluster.

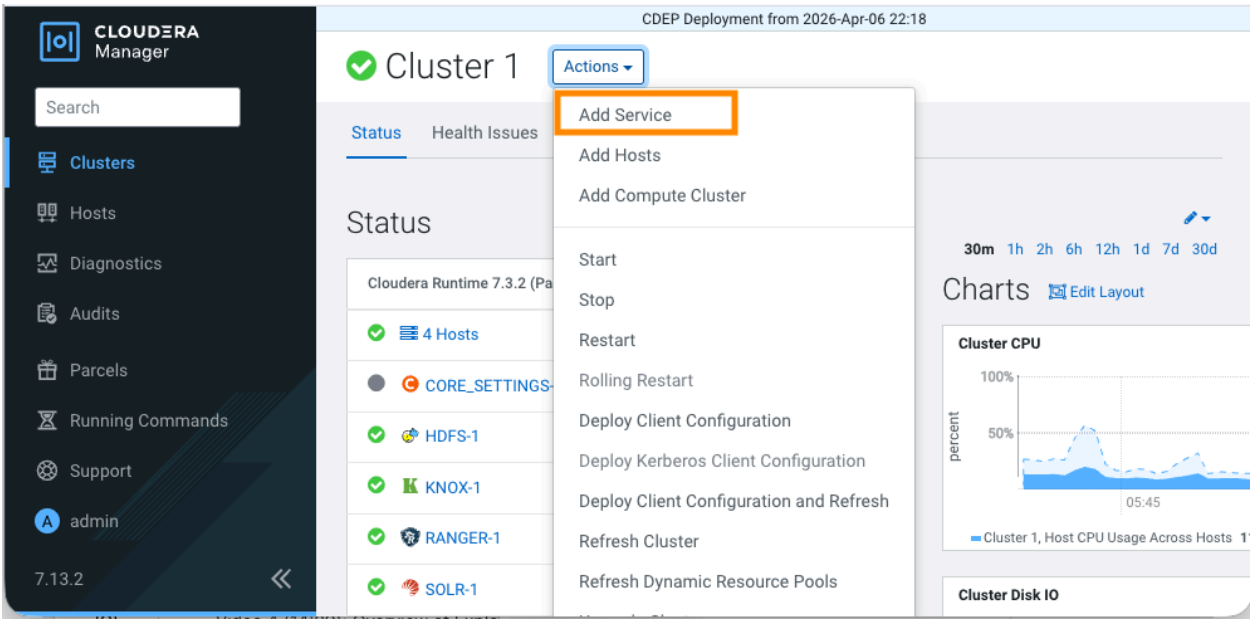
When installing the Cloudera Semantic Search service, you can add an unsecured Cloudera Semantic Search on a Cloudera service to your Cloudera Base on premise deployment.

Before you begin

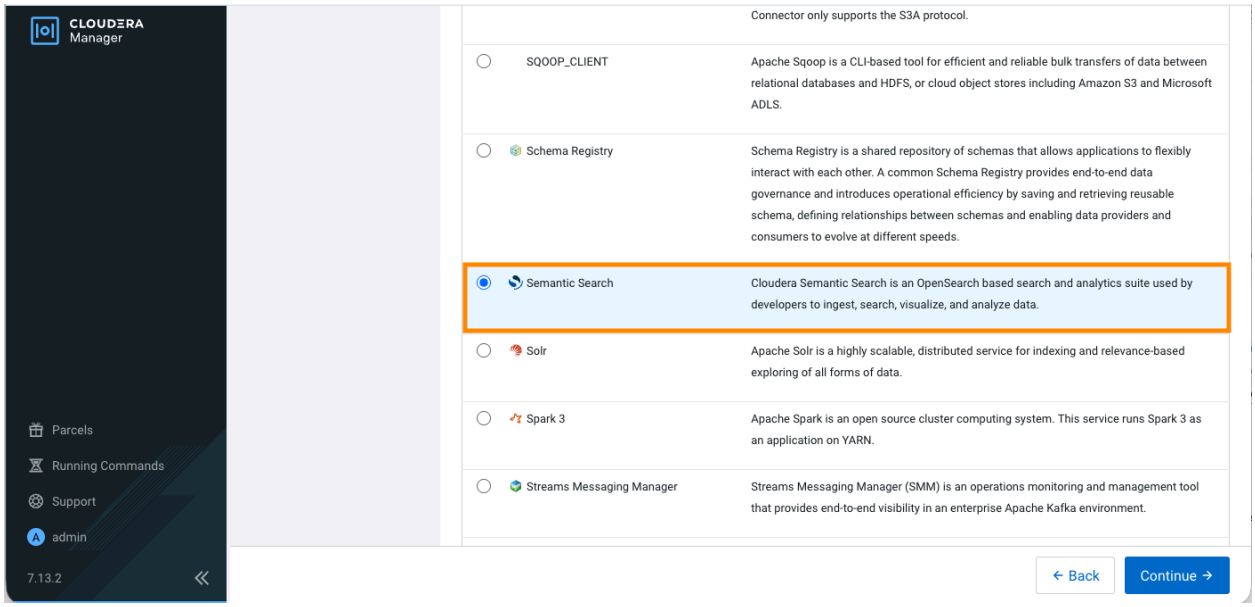
- The Cloudera Semantic Search CSD file must be installed.
- The Cloudera Semantic Search parcel must be installed.

Steps

1. Navigate to the Cloudera Manager Admin Console home page.
2. Click the action menu of the cluster to which you want to add the Cloudera Semantic Search as a Cloudera service, for example, Cluster 1, and select the **Add Service** option from the list of actions.



3. Select the Semantic Search option and click the Continue button.

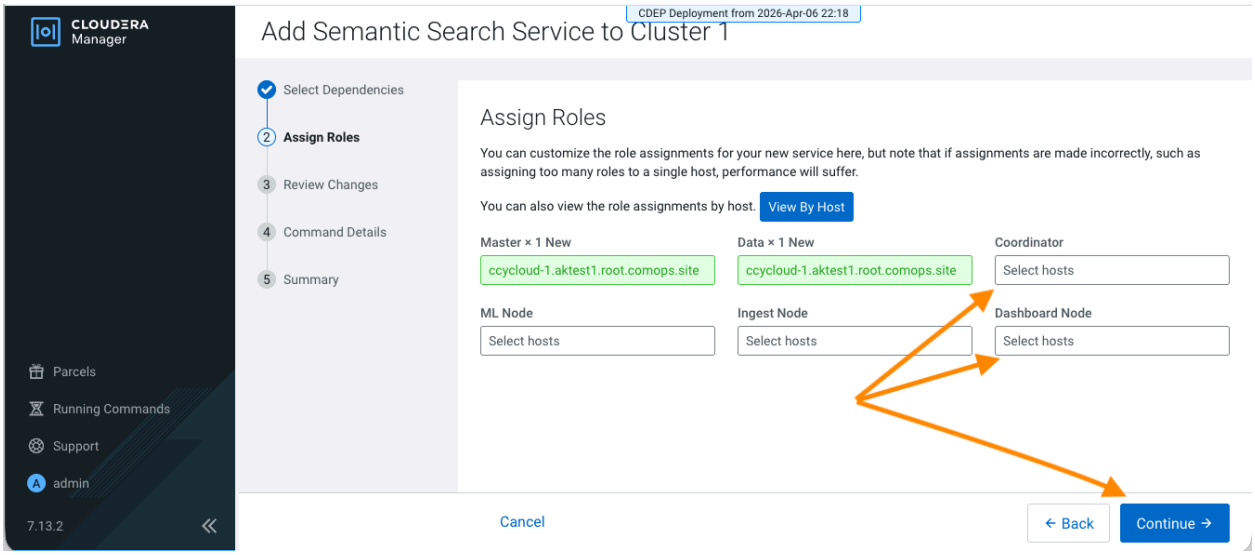


4. Select the dependent services and click the Continue button.
5. Assign the Cloudera Semantic Search roles to the hosts in your cluster.

   Assign the following roles to the appropriate hosts:

   - Cluster_manager (Master): Manages the overall operation of a cluster and keeps track of the cluster state. This includes creating and deleting indexes, keeping track of the nodes that join and leave the cluster, checking the health of each node in the cluster (by running ping requests), and allocating shards to nodes.
   - Data
   - Coordinating: Delegates client requests to the shards on the data nodes, collects and aggregates the results into one final result, and sends this result back to the client. This role is optional.
   - Ingest: Pre-processes data before storing it in the cluster. Run an ingest pipeline that transforms your data before adding it to an index. This role is optional.
   - Machine Learning: By default, the Cloudera AI tasks and models only run on Cloudera AI nodes. When configured without the data node role, the Cloudera AI nodes do not store any shards and instead calculate resource requirements at runtime. This role is optional.
   - Dashboard: To access and view the Semantic Search Dashboard, you need to configure the appropriate security role within the Cloudera Semantic Search cluster. The role must grant read permissions for the necessary indices and cluster-level actions required by the dashboard. This role is optional.

   Cloudera does not recommend assigning multiple Semantic Search roles to a single host.



6. Click the Continue button.
7. Wait while the Cloudera Manager does the initial service setup.
8. Click the Continue button.
9. Click the Finish button.
10. Address the stale configurations by restarting the Ranger service.

After you finish

Configure the administrator certificates for the Cloudera Semantic Search cluster. For more information, see *Using Administrator certificates for Cloudera Semantic Search cluster management* document.

Verifying Cloudera Semantic Search installation

Use the Cloudera Semantic Search shell to verify your Cloudera Semantic Search installation.

Before you begin

- The Cloudera Semantic Search CSD file must be installed.
- Cloudera must be installed.
- The Cloudera Semantic Search parcel must be installed.
- The Cloudera Semantic Search must be added on the Cloudera service to your deployment.
- Ensure that the Semantic Search service is up and running.

Steps

1. Open a terminal on your Linux machine.
2. Use the following “curl” command to check whether the Cloudera Semantic Search is installed correctly for the respective nodes:

```
None
curl https://<Master_Host_Name>:<Master_Port>/_cat/nodes\?v --cacert
cm-auto-global_cacerts.pem -u admin:admin
```

For example,

```
None
curl https://<Master_Host_Name>:<Master_Port>/_cat/nodes?v --cacert
cm-auto-global_cacerts.pem -u admin:admin
```

Expected output:

The output contains basic host information and statistics such as IP address, CPU load, and so on.

3. Use the Semantic Search dashboard to verify the installation. For more information, see the FAQs section in the *OpenSearch Dashboards* document.

FAQ and Troubleshooting

A collection of frequently asked questions and debugging steps while installing and deploying the Cloudera Semantic Search.

Pre-trained Model Registration Intermittently Fails Due to Certificate Path Error

Problem description

The registration of a pre-trained model may intermittently fail. When attempting to register a model, the task status is reported as `FAILED`.
The registration task returns the following error:

```
None
"error" : "unable to find valid certification path to requested
target"
```

The corresponding error message in the Machine Learning log is:

```
None
sun.security.validator.ValidatorException: PKIX path building
failed:
sun.security.provider.certpath.SunCertPathBuilderException:
unable to find valid certification path to requested target.
```

Possible Cause

The issue is caused by OpenSearch attempting to download resources, such as the pre-trained models list from https://artifacts.opensearch.org/ or packages from https://download.pytorch.org/. The certificates for these sites, which are issued by Amazon, are not accepted because of the Java Development Kit (JDK)'s strict list of trusted root and intermediate certificates.

Workaround

Add the Amazon root certificate to the Cloudera Manager (CM)'s trusted global cacerts, as the certificates for both OpenSearch and PyTorch download sites are issued by Amazon.

Alternatively, add the specific certificates of https://download.pytorch.org/ and https://artifacts.opensearch.org/.

Steps

The following are the steps to update the Cloudera Manager truststore.

1. Export the root certificate ("Amazon Root CA1") of either https://download.pytorch.org/ or https://artifacts.opensearch.org/ through a browser. The file is typically exported as `Amazon Root CA 1.pem`.
2. Get the current truststore from CM through the API:

```
None
curl -u admin:<password> -X 'GET'
'https://<cm-host>:7183/api/v58/certs/truststore?type=PEM'
-H 'accept: application/octet-stream' -o
cm-auto-global_cacerts.pem
```

3. Concatenate the new certificate with the existing truststore:

```
None
cat "Amazon Root CA 1.pem" cm-auto-global_cacerts.pem >
combined_cert.pem
```

4. Update the truststore with the combined certificate:

```
None
curl -u admin:<password> -H "Content-Type: application/json"
-X POST -d '{ "newCertLocation": "/tmp/combined_cert.pem",
"newCertContent": "" }'
```

```
"https://<cm-host>:7183/api/v57/cm/commands/updateGlobalTruststore"
```

5. Restart Semantic Search.
6. Retry the model registration.

Semantic Search Cluster Health is Yellow with Single Data Node Configuration

Problem description

When the Semantic Search cluster is configured with only one data node, the service health status is yellow. A yellow status indicates that all primary shards are allocated to nodes, but some replicas are not. Although the cluster retains its full functionality, the yellow status might suggest an issue to the user.

Possible cause

The default replica count for the security-auditlog index is 1. When only one data node is configured in the cluster, there is no available node for the replica to be assigned, resulting in unassigned replicas and the yellow health status.

Workaround

The yellow health status is expected when configuring only one data node, and the Semantic Search service still retains its full functionality in this state. To achieve a green health status, it is recommended to configure 2 or more data nodes.

How to create an unsecure Cloudera Semantic Search cluster

You should only create an unsecure Cloudera Semantic Search cluster in well-protected, isolated, and trusted environments. You must use the predefined `admin` as the metrics user.

After you install the Cloudera on premises cluster as mentioned in the Cloudera Base on premises Installation Guide, **do not enable** the Auto-TLS and Kerberos.

Appendix

1. To secure the Cloudera Semantic Search cluster using Kerberos authentication, see *Securing the Cloudera Semantic Search cluster* document.
2. To check how to access the OpenSearch dashboard, see the “Accessing the OpenSearch dashboard” section in the *Securing the Cloudera Semantic Search cluster* and *OpenSearch Dashboards* documents.
3. To check how to enable the Ranger plugin for Cloudera Semantic Search cluster, see *Enabling the Ranger plugin for the Cloudera Semantic Search service* document.
4. Client-side examples that demonstrate how to connect to Cloudera Semantic Search and authenticate with Kerberos, see *Client-side Examples* document.
5. To check the usage of the `securityadmin.sh` script while working with the security plugin, see *Applying changes to the security configuration files* document.
6. To check how to build charts and add them to the respective dashboards, see *Adding Cloudera Semantic Search chart to the Cloudera Manager* document.
7. To know how to integrate Knox with Cloudera Semantic Search, see *Integrating Knox with Cloudera Semantic Search* document.
8. To know how to create and use administrator’s certificates for cluster management, see *Using Administrator certificates for Cloudera Semantic Search cluster management* document.