Cloud Search Service

Overview

Issue 01

Date 2025-04-17

HUAWEI CLOUD COMPUTING TECHNOLOGIES CO., LTD.

Trademarks and Permissions

HUAWEI and other Huawei trademarks are the property of Huawei Technologies Co., Ltd.
All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei Cloud and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

Huawei Cloud Computing Technologies Co., Ltd.

Address: Huawei Cloud Data Center Jiaoxinggong Road
Qianzhong Avenue
Gui'an New District
Gui Zhou 550029
People's Republic of China

Website: https://www.huaweicloud.com/intl/en-us/

Contents

1 What Is Cloud Search Service?

CSS

CSS is a fully managed, distributed search service based on open source Elasticsearch and OpenSearch. You can use it for structured and unstructured data search, and enable vector-based composite search, statistics generation, and reporting. CSS is a fully hosted cloud service of the ELK Stack and is compatible with open-source Elasticsearch, Logstash, Kibana, and Cerebro.

- Elasticsearch and OpenSearch

  Elasticsearch and OpenSearch are open-source distributed search engines that can be deployed in standalone or cluster mode. As the heart of the ELK Stack, Elasticsearch clusters support multi-condition search, statistical analysis, and create visualized reports of structured and unstructured text. For details about Elasticsearch, see the **Elasticsearch: The Definitive Guide**. For details about the OpenSearch search engine, see **OpenSearch Documentation**.

  CSS enables automatic deployment, allowing you to quickly provision Elasticsearch and OpenSearch clusters with zero O&M burdens. It offers built-in search engine optimization, helping to easily achieve optimal search performance. Additionally, its robust monitoring system tracks key metrics—such as cluster health and query performance—so you can manage your clusters effortlessly.

  Elasticsearch and OpenSearch clusters created in CSS can be accessed through Kibana, OpenSearch Dashboards, and Cerebro. For in-depth analysis and visualization of data, choose Kibana, as it provides rich visualization features and powerful analytical capabilities. For cluster management and monitoring, choose Cerebro, as it provides intuitive cluster status views and convenient management functions.

- Logstash

  Logstash is an open-source data processing pipeline that ingests data from a multitude of sources, transforms it, and then sends it to your desired destination.

  CSS Logstash is a fully managed data ingestion and processing service that is completely compatible with open-source Logstash. You can quickly create Logstash clusters in CSS. Data is scattered across many different systems in different formats. CSS Logstash helps you get insights by easily processing data from a variety of data sources and dumping it to CSS's Elasticsearch clusters or other systems.

Introduction Video

Functions

- Open-source compatibility

  Freely use native Elasticsearch and OpenSearch APIs and other software in the ecosystem, such as Logstash, Beats, and Kibana.

- Support for a variety of data sources

  A few simple configurations allow you to smoothly connect to multiple data sources, such as FTP, OBS, HBase, and Kafka. No extra coding is required.

- One-click operation

  One-click cluster application, capacity expansion, and restart from small-scale testing to large-scale rollout

- Flexible dictionary management

  You can custom your dictionaries. Modified settings take effect immediately without system restart.

- User-defined snapshot policies

  Trigger backup snapshots manually or configure an automated schedule.

2 Advantages

CSS has the following features and advantages.

Efficient and Ease of Use

You can get insights from terabyte-scale data in milliseconds. In addition, you can use the visualized platform for data display and analysis.

Flexible and Scalable

You can request resources as needed and perform capacity expansion online with zero service interruption.

Easy O&M

CSS is a fully-managed, out-of-the-box service. You can start using it with several clicks, instead of managing clusters.

Kernel Enhancement

- **Vector search**

  When you search for unstructured data, such as images, videos, and corpuses, the nearest neighbors or approximate nearest neighbors are searched based on feature vectors.
- **Decoupled storage and compute**

  CSS provides an API for freezing indexes. Hot data stored on SSD can be dumped to OBS to reduce data storage costs and decouple compute from storage.
- **Flow control**

  CSS can control traffic at the node level. You can configure the blacklist and whitelist, the maximum concurrent HTTPS connections, and the maximum HTTP connections for a node. Each function has an independent control switch.
- **Large query isolation**

  CSS allows you to separately manage large queries. You can isolate query requests that consume a large amount of memory or take a long period of time.

- **Index monitoring**

  CSS monitors various metrics of the running status and change trend of cluster indexes to measure service usage and handle potential risks in a timely manner, ensuring that clusters can run stably.

- **Enhanced monitoring**

  CSS supports enhanced cluster monitoring. It can monitor the P99 latency of cluster search requests and the HTTP status codes of clusters.

High Reliability

You can choose to trigger snapshots manually or on a periodic basis for backup and restore snapshots to the current or other clusters. Snapshots of a cluster can be restored to another cluster to implement cluster data migration.

- Automatic backup using snapshots

  CSS provides the backup function. You can enable the automatic backup function on the CSS management console and set the backup period based on the actual requirements.

  Automatic backup is to back up the index data of a cluster. Index backup is implemented by creating cluster snapshots. For backup of the first time, you are advised to back up all index data.

  CSS allows you to store the snapshot data of Elasticsearch instances to OBS, thereby achieving cross-region backup with the cross-region replication function of OBS.

- Restoring data using snapshots

  If data loss occurs or you want to retrieve data of a certain period, click **Restore** in the **Operation** column in the **Snapshots** area to restore the backup index data to the specified cluster by using existing snapshots.

High Security

CSS uses network isolation in addition to various host and data security measures.

- Network isolation

  The network is divided into two planes, service plane and management plane. The two planes are deployed and isolated physically to ensure the security of the service and management networks.

  - Service plane: refers to the network plane of the cluster. It provides service channels for users and delivers data definition, index, and search capabilities.
  - Management plane: This is mainly the management console, where you manage CSS.
  - VPC security groups or isolated networks ensure the security of hosts.

- Access control

  - Using the network access control list (ACL), you can permit or deny the network traffic entering and exiting the subnets.
  - Internal security infrastructure (including the network firewall, intrusion detection system, and protection system) can monitor all network traffic that enters or exits the VPC through the IPsec VPN.

  - User authentication and index-level authentication are supported. CSS also supports interconnection with third-party user management systems.
- Data security
  - In CSS, a multi-replica mechanism is used to ensure data security.
  - Communication between the client and server can be encrypted using SSL.
- Operation audit

  Cloud Trace Service (CTS) can be used to perform auditing on key logs and operations.

High Availability

To prevent data loss and minimize the cluster downtime in case of service interruption, CSS supports cross-AZ cluster deployment. When creating a cluster, you can select two or three AZs in the same region. The system will automatically allocate nodes to these AZs. If an AZ is faulty, the remaining AZs can still run properly, significantly enhancing cluster availability and improving service stability.

3 Product Components

CSS supports Kibana and Cerebro.

Kibana

Kibana is an open-source data analytics and visualization platform that works with Elasticsearch. You can use Kibana to search for and view data stored in Elasticsearch indexes and display data in charts and maps. For details about Kibana, visit **https://www.elastic.co/guide/en/kibana/current/index.html**.

By default, the Elasticsearch cluster of CSS provides the access channel to Kibana. You can quickly access Kibana without installing it. CSS is compatible with Kibana visualizations and Elasticsearch statistical and analysis capabilities.

- Over 10 data presentation modes
- Nearly 20 data statistics methods
- Classification in various dimensions, such as time and tag

Cerebro

Cerebro is an open-source Elasticsearch web admin tool built using Scala, Play Framework, AngularJS, and Bootstrap. Cerebro allows you to manage clusters on a visualized page, such as executing REST requests, modifying Elasticsearch configurations, monitoring real-time disks, cluster loads, and memory usage.

By default, the Elasticsearch cluster of CSS provides the access channel to Cerebro. You can quickly access Cerebro without installing it. CSS is fully compatible with the open-source Cerebro and adapts to the latest 0.8.4 version.

- Elasticsearch visualized and real-time load monitoring
- Elasticsearch visualized data management

4 Scenarios

CSS can be used to build search boxes for websites and apps to improve user experience. You can also build a log analysis platform with it, facilitating data-driven O&M and business operations. CSS vector search can help you quickly build smart applications, such as AI-based image search, recommendation, and semantic search.

Site Search

CSS can be used to search for website content by keyword as well as search for and recommend commodities on e-commerce sites.

- Real-time search: When site content is updated, you can find the updated content in your search within minutes, or even just seconds.
- Categorized statistics: You can apply search filters to sort products by category.
- Custom highlight style: You can define how the search results are highlighted.

All-Scenario Log Analysis

Analyze the logs of Elastic Load Balance (ELB), servers, containers, and applications. In CSS, the Kafka message buffer queue is used to balance loads in peak and off-peak hours. Logstash is used for data extract, transform and load (ETL). Elasticsearch retrieves and analyzes data. The analysis results are visualized by Kibana and presented to you.

- High cost-effectiveness: CSS separates cold and hot storage, and decouples computing and storage resources, achieving high performance and reducing costs by over 30%.
- Ease of use: Perform queries in a GUI editor. Easily create reports using drag-and-drop components.
- Powerful processing capability: CSS can import hundreds of terabytes of data per day, and can process petabytes of data.

Database Query Acceleration

CSS can be used to accelerate database queries. E-commerce and logistics companies have to respond to a huge number of concurrent order queries within a short period of time. Relational databases, although having good transaction atomicity, are weak in transaction processing, and can rely on CSS to enhance OLTP and OLAP capabilities.

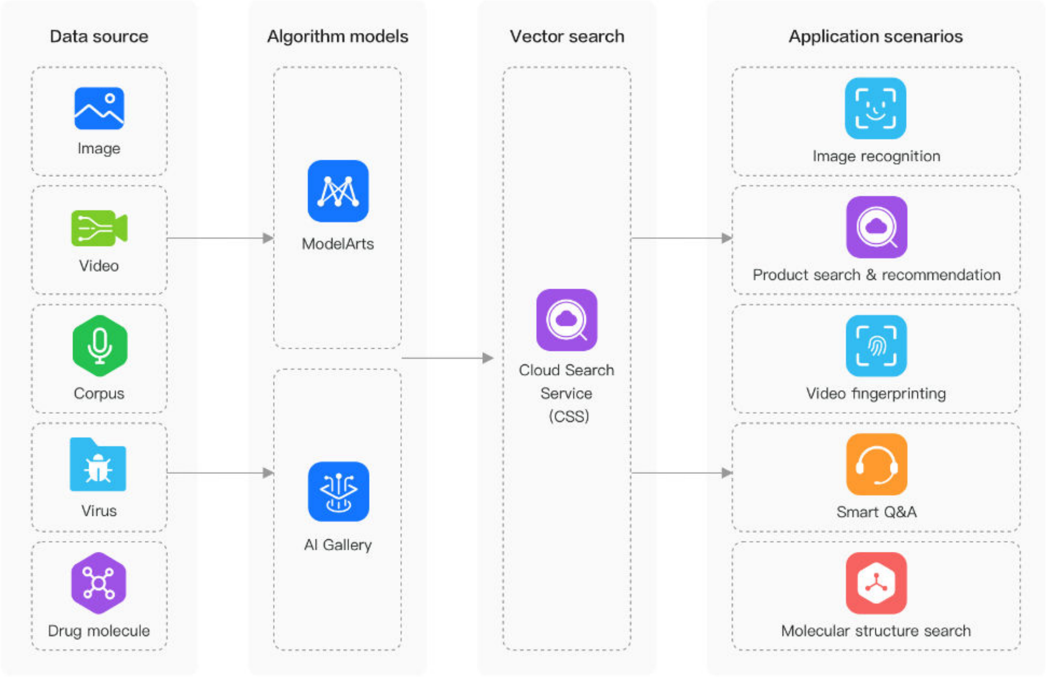
- High performance: Retrieve data from hundreds of millions of records within milliseconds. Text, time, numeric, and spatial data types are supported.
- High scalability: CSS can be scaled to have over 200 data nodes and over 1000 columns.
- Zero service interruption: The rolling restart and dual-copy mechanisms can avoid service interruption in case of specifications change or configuration update.

Vector Search

When you search for unstructured data, such as images, videos, and corpuses, the nearest neighbors or approximate nearest neighbors are searched based on feature vectors. This has the following advantages:

- Efficiency and reliability: The Huawei Cloud vector search engine provides ultimate search performance and distributed disaster recovery capabilities.
- Abundant indexes: Multiple indexing algorithms and similarity measurement methods are available and can meet diverse needs.
- Easy learning: CSS is fully compatible with the open-source Elasticsearch ecosystem.

Figure 4-1 Vector search



5 Constraints

This topic describes limits on the node quantity of a CSS cluster. For details about the limits and limitations of different features provided by CSS, see relevant topics in the CSS User Guide.

Maximum and Minimum Numbers of Nodes in a Cluster

The following tables provide the maximum and minimum numbers of nodes each CSS cluster can have.

Table 5-1 Maximum and minimum numbers of nodes in an Elasticsearch or OpenSearch cluster

Number of Nodes	Limit
Maximum number of nodes in a cluster	
Minimum number of nodes in a cluster	1

Table 5-2 Maximum and minimum numbers of nodes in a Logstash cluster

Number of Nodes	Limit
Maximum number of nodes in a cluster	100
Minimum number of nodes in a cluster	1

Quotas

CSS uses the following resource quotas:

- Number of instances

- CPUs
- Memory capacity in GB
- Number of disks
- Disk size (GB)

For details about how to check and modify quotas, see **Quotas**.

6 Related Services

Figure 6-1 shows the relationships between CSS and other services.

Figure 6-1 Relationships between CSS and other services

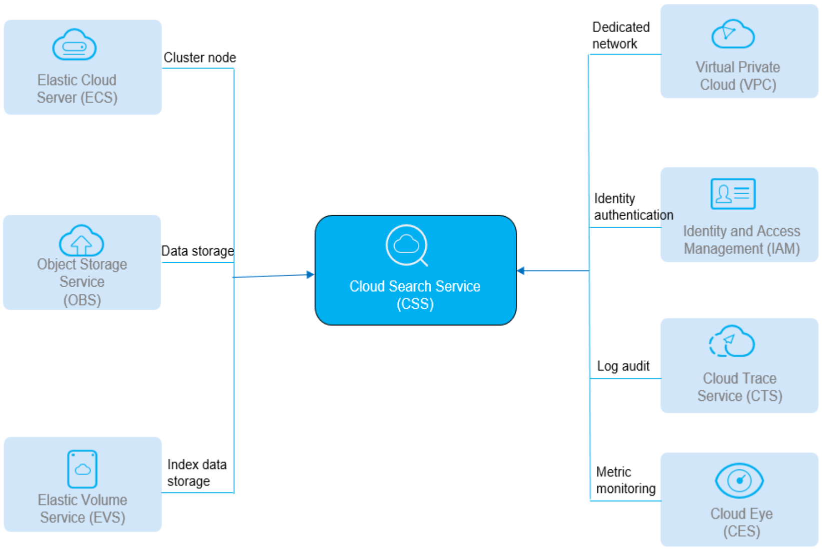

Elastic Cloud Server (ECS)
Cluster node
Object Storage Service (OBS)
Data storage
Elastic Volume Service (EVS)
Index data storage
Cloud Search Service (CSS)
Dedicated network
Virtual Private Cloud (VPC)
Identity authentication
Identity and Access Management (IAM)
Log audit
Cloud Trace Service (CTS)
Metric monitoring
Cloud Eye (CES)

Table 6-1 Relationships between CSS and other services

Service	Description
Virtual Private Cloud (VPC)	CSS clusters are created in the subnets of a VPC. VPCs provide a secure, isolated, and logical network environment for your clusters.
Elastic Cloud Server (ECS)	In a CSS cluster, each node is an ECS. When you create a cluster, ECSs are automatically created.
Elastic Volume Service (EVS)	CSS uses EVS to store index data. When you create a cluster, EVSs are automatically created for cluster data storage.
Object Storage Service (OBS)	Snapshots of CSS clusters are stored in OBS buckets.

Service	Description
Identity and Access Management (IAM)	IAM authenticates access to CSS.
Cloud Eye	CSS uses Cloud Eye to monitor cluster metrics in real time. Supported CSS metrics include disk usage and cluster health status. You can learn about the disk usage of the cluster based on the disk usage metric. You can learn about the health status of a cluster based on the cluster health status metric.
Cloud Trace Service (CTS)	With CTS, you can record operations associated with CSS for query, auditing, and backtracking operations.

7 Basic Concepts

Cluster

CSS provides functions on a per cluster basis. A cluster represents an independent search service that consists of multiple nodes.

Index

An index stores Elasticsearch data. It is a logical space in which one or more shards are grouped.

Shard

An index can potentially store a large amount of data that can exceed the hardware limits of a single node. To solve this problem, Elasticsearch provides the ability to subdivide your index into multiple pieces called shards. When you create an index, you can simply define the number of shards that you want. Each shard is in itself a fully-functional and independent "index" that can be hosted on any node in the cluster.

You need to specify the number of shards before creating an index and cannot change the number after the index is successfully created.

Replica

A replica is a copy of the actual storage index in a shard. It can be understood as a backup of the shard. Replicas help prevent single point of failures (SPOFs). You can increase or decrease the number of replicas based on your service requirements.

Document

An entity for Elasticsearch storage. Equivalent to the row in the RDB, the document is the basic unit that can be indexed.

Document Type

Similar to a table in the RDB, type is used to distinguish between different data.

In versions earlier than Elasticsearch 7.*x*, each index can contain multiple document types. Elasticsearch defines a type for each document.

Elasticsearch 7.*x* and later versions only support documents of the .doc type.

Mapping

A mapping is used to restrict the type of a field and can be automatically created based on data. It is similar to the schema in the database.

Field

The field is the minimum unit of a document. It is similar to the column in the database.