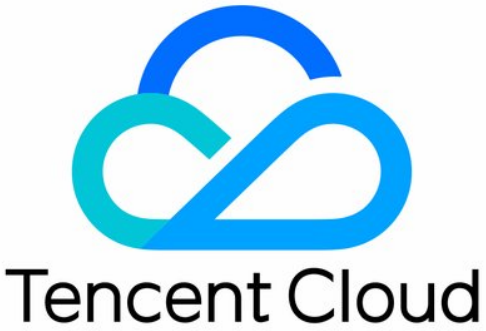
Elasticsearch Service Product Introduction



Copyright Notice

Trademark Notice

Service Notice

This document provides an overview of the as-is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre-sales consultation and technical after-sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

Product Introduction
Overview

Last updated: 2024-10-24 20:58:21

Tencent Cloud ES is a fully managed cloud service for huge data search and analysis, featuring a high-performance self-developed kernel and integrated X-Pack. ES supports easy cluster management with features like autonomous indexing, compute-storage separation, and cluster inspection. It also supports zero maintenance, automatic elasticity, and on-demand usage in a Serverless mode. With ES, you can efficiently build services for information retrieval, log analysis, and operations monitoring. Its unique vector retrieval can help you develop AI in-depth applications based on semantics and images.
While retaining Elasticsearch's compatibility and openness, ES incorporates Tencent Cloud's computing, storage, and security technologies. It has various cluster management functions and is secure, elastic, and highly available. Additionally, ES integrates the official Elastic Stack features (formerly X-Pack), which adds permission management, SQL, machine learning, and alert features to the open source foundation. These features simplify basic OPS tasks like cluster deployment and operation management, letting you focus on the business.
ES allows you to quickly build applications like massive data storage and search, and real-time log analysis, website search and navigation, search for enterprises, log monitoring, and click analysis.

Main Components

- Elasticsearch
  Elasticsearch is a distributed search engine capable of storing, full-text searching, and performing statistical analysis on massive data sets. It provides a RESTful API and various language clients, allowing flexible development according to business requirements.
- Kibana
  Kibana is a data visualization tool that enables easy querying and analysis of data stored in Elasticsearch clusters.
- Elastic Stack (formerly X-Pack)
  Elastic Stack is Elasticsearch's official plugin which has various advanced features. Its data permission management can be applied to the field level. It can be easily integrated to your existing business via SQL and JDBC connection. Its machine learning and alerts can analyze cluster data and fluctuations to predict data trends and send out huge fluctuation alarms.

Product Version

Last updated: 2024-10-24 20:58:40

We understand that every business has unique demands and challenges. Therefore, Tencent Cloud ES offers multiple product versions to meet different usage scenarios.
This article compares the features of each version, aiming to help you make the right choice based on your application scenario, business scale, performance requirements, and technology stack.

Differences between the Log Enhancement Edition and the General Edition

Item	Log Enhancement Edition	General Edition
Main Features	Higher write performance, lower cost	Suitable for various scenarios, faster community version updates
Applicable Scenario	• Log analysis: Web logs, App logs, database logs, etc • Metric analysis: Server metrics, network metrics, storage metrics, etc • Application performance tracking: Python, NodeJS, etc	Various scenarios such as logs, search, analysis, AI enhancement
Core Strength	• Cost: Lower cost based on COS compute-storage separation • Write: Unique index writing acceleration technology, performance improvement of 3 to 5 times	• Covers more of the community version, follows up faster • Also includes some proprietary Tencent Cloud ES technologies
Architecture Patterns	Separated Storage and Computing Architecture tailored for log scenarios, Hot Temperature Architecture	Classic ES Architecture suitable for various scenarios such as logs/search/analysis, Hot Temperature Architecture
Supported	7.14.2	8.13.3, 8.11.3, 7.17.7, 7.14.2, 7.10.1, 6.8.2, 5.6.4, etc

Editions		
Advanced features	Platinum Edition	Platinum Edition, Basic Edition, Open-source version
Billing	• Billing based on cluster node specifications, storage space, and number of nodes • Separated Storage and Computing Architecture will generate distributed shared storage costs (Lower costs for log scenarios)	Billing based on cluster node specifications, storage space, and number of nodes

Proprietary self-developed technology comparison

Log Enhancement Edition provides more advanced features at the kernel level compared to the General Edition, improving cluster write and query performance, and reducing massive data storage costs.

Classification	Features	Description	Log Enhancement Edition	General Edition
Stability	SLA	Service Level Agreement, committed to service quality, availability, and liability	Up to 99.9%	Up to 99.9%
	Full-link fuse and rate limiting	Combining elastic funnel and full-link memory fuse, comprehensively considering CPU, IO, memory contradiction, supporting high concurrency and large query in hybrid scenarios	√	√
Storage Cost	Storage-Compute Separation	Self-developed Hybrid Storage Architecture, Cache Layering, Intelligent Uninstallation, achieving integrated hot and cold search, reducing storage cost by 50% – 80%	√	-

	FST Off-Heap	Self-developed FST Off-heap Storage, reducing on-heap memory usage, significantly enhancing single-node disk management capability	√	√
	Deep Data Compression	Systematic encoding optimization for Row Storage, Columnar Storage, and Inverted Index Dictionary File, supporting fine-grained switch control, reducing storage costs by over 20%	√	√
Write Performance	Write Acceleration	Self-developed Share-free, Self-closed Loop Read-Write Separation Architecture, Isolated Resources while Enhancing Write Performance by 5 – 20 times	√	-
	Physical Replication	Master-Slave Replicas Physical Replication, Eliminating Redundant Computation of Replicas, Enhancing Write Performance by 50%	√	-
	Targeted Routing	Shard-directed Routing Optimization, Optimizing High Fan-out and Long-tail Issues, Enhancing Write Performance by 30%+	√	√
Query Performance	Query/IO Parallelization	Self-developed Multi-level Parallel Query Framework, Supporting All Query Scenarios, Enhancing Query Performance by 3 – 5 times	√	-
	Adaptive Bypass Merge Strategy	By merging and converging Segments, Reducing Random IO in Queries, Overall Query Performance Enhanced by more than 2 times	√	-

	Query Cache Optimization	Significantly improved query concurrency capability, QPS increased by over 50%	√	√
	Query Cropping	Node time-series index query cropping, time range cropping optimized to boundary point cropping, high-dimensional time-series retrieval performance increased by more than 10 times	√	√
	Sort Acceleration	Based on the bkd storage architecture, unnecessary data comparisons are skipped while traversing the inverted table, improving sorting performance by 5 to 10 times	√	√
	Authentication Acceleration	X-Pack authentication performance optimization, eliminating authentication hotspot overhead, improving query performance by over 30%	√	√
OPS Efficiency	Autonomous Index	Industry-exclusive, supports automatic shard tuning, intelligent rolling, query cropping, fault self-healing, achieving fully managed indexing	√	√

Enhanced Log Introduction

Last updated: 2024-10-24 21:11:19

Use Cases

Log analysis, such as Web logs, App logs, database logs, server metrics, network metrics, storage metrics, etc. These scenarios are generally characterized by more writes and fewer reads, and are cost-sensitive.

Version Features

- Keywords: Higher write performance, lower cost.
- By adopting storage-compute separation, write acceleration, and parallelization of queries/IO, we reduce storage costs and improve write performance. (Recommend the storage-compute separation architecture, but traditional hot-warm architecture can also be used based on actual business needs.)
- For operational efficiency, autonomous indexing can achieve fully automated index creation, rolling, archiving, deletion, and automatic fault repair, realizing intelligent tuning of shards and greatly reducing operational faults and resource investment.

Description: The main problem solved

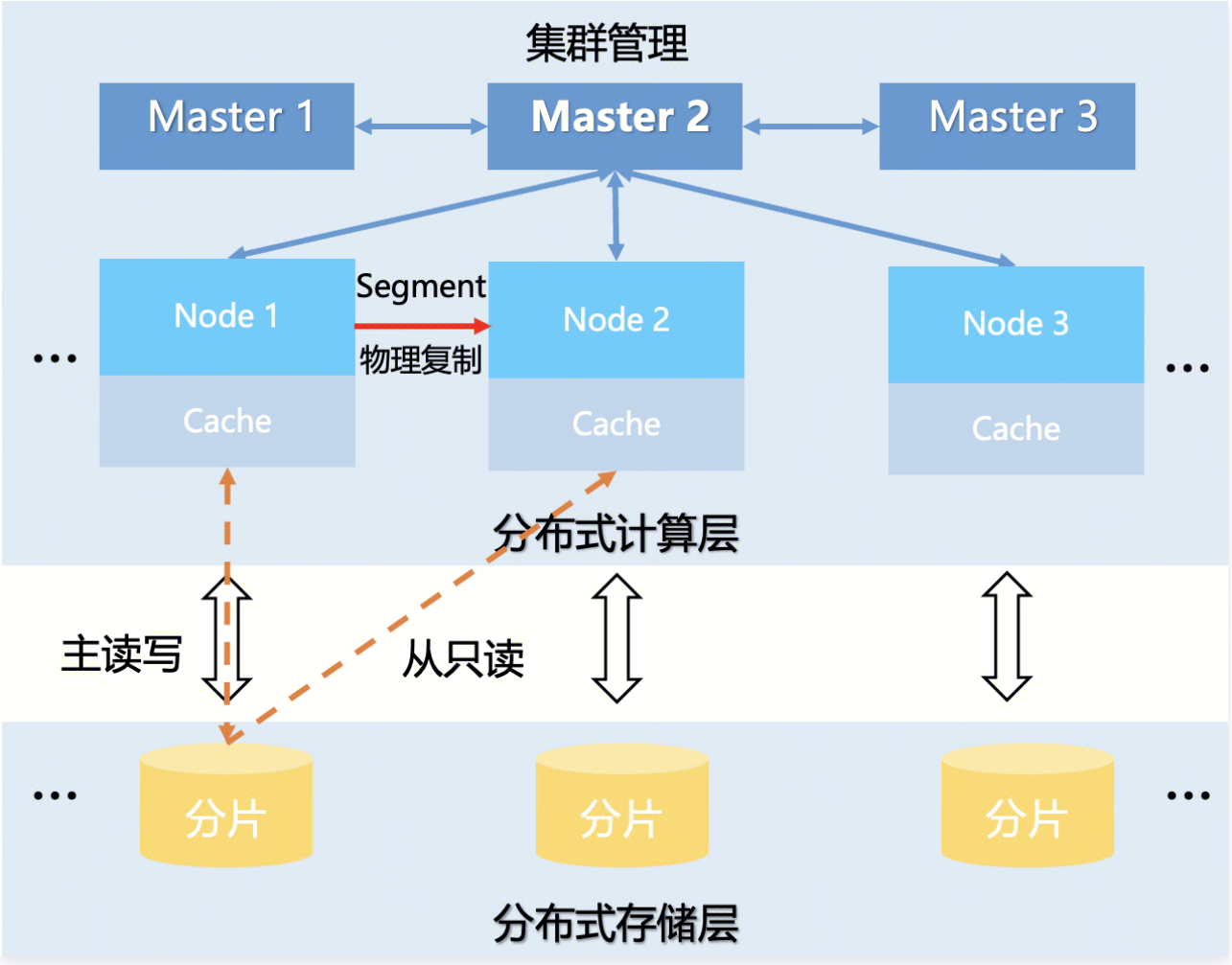
- In terms of storage, users usually set 2-3 replicas to ensure stability. If using cloud disks, the cloud disk layer also stores 3 replicas, resulting in significant storage redundancy.
- Regarding computation, since ES primary and secondary shards both parse documents, tokenize, and index, there are issues with repeated calculations and duplicate writings between primary and secondary replicas.
- Additionally, in the native architecture, storage and computation are coupled, with data and computation on the same node. Resources cannot independently scale up and down elastically. During cluster scaling, there will be extensive data relocation, which is not only time-consuming but also highly resource-intensive.

Key Features

Storage-Compute Separation

The core idea of Storage-Compute Separation is to eliminate replica redundant computing based on physical replication and ensure complete consistency between primary and replica shards. Meanwhile, it adopts a Delta + Base Architecture, with Local SSD bearing high

concurrency writes and handling Merge computational overhead. Larger data files are merged in real-time and offloaded to the massive COS, achieving high availability through COS and significantly reducing storage costs compared to cloud disks. The cache module also caches high-frequency access data, reducing the access frequency to COS. To address performance differences between COS and local disks, IO parallelization technology combined with multi-level caching is used to achieve integrated hot and cold hybrid search capability.
Search for the required CAM policy as needed, and click to complete policy association.

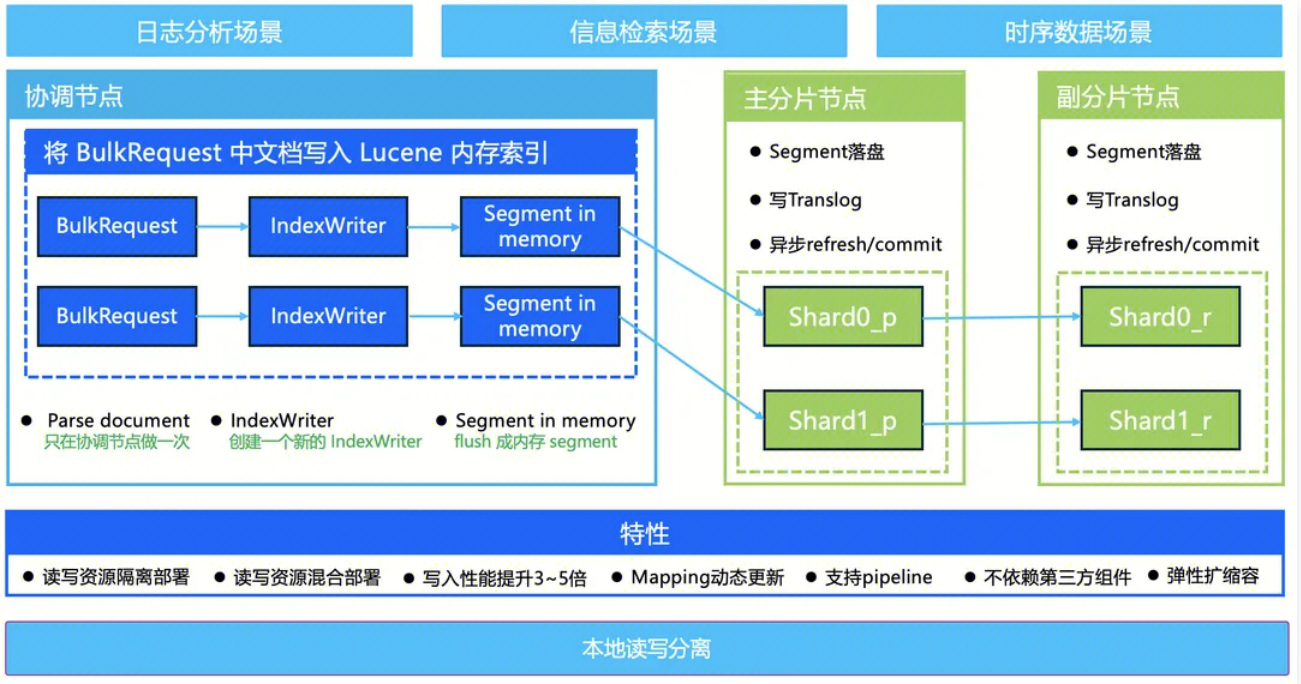


For specific feature details, please refer to Storage-Compute Separation Features.

Write Acceleration

The core idea of Write Acceleration is to eliminate replica redundant computing, network overhead, and lock contention based on memory segments. When ES writes data, it ultimately writes to memory through Lucene and refreshes into segments after a while. We can pre-build segments using Lucene's API in the Coordinator and store them in memory, then forward the memory segments to specific index shards. After receiving the memory segments, the index shards periodically append them to Lucene. Through memory segment generation and copying, memory merge, self-Definition merge policy, directed routing, and other highlighted ideas and technologies, the write throughput can be increased by 3-5 times.

Search for the required CAM policy as needed, and click to complete policy association.



The write acceleration optimization here can be understood as local read-write separation or single-cluster read-write separation architecture. The Coordinator constructs the segment in memory and physically copies it to the primary and replica shards. The cluster can be deployed with read and write partition nodes or mixed deployment.

Note:

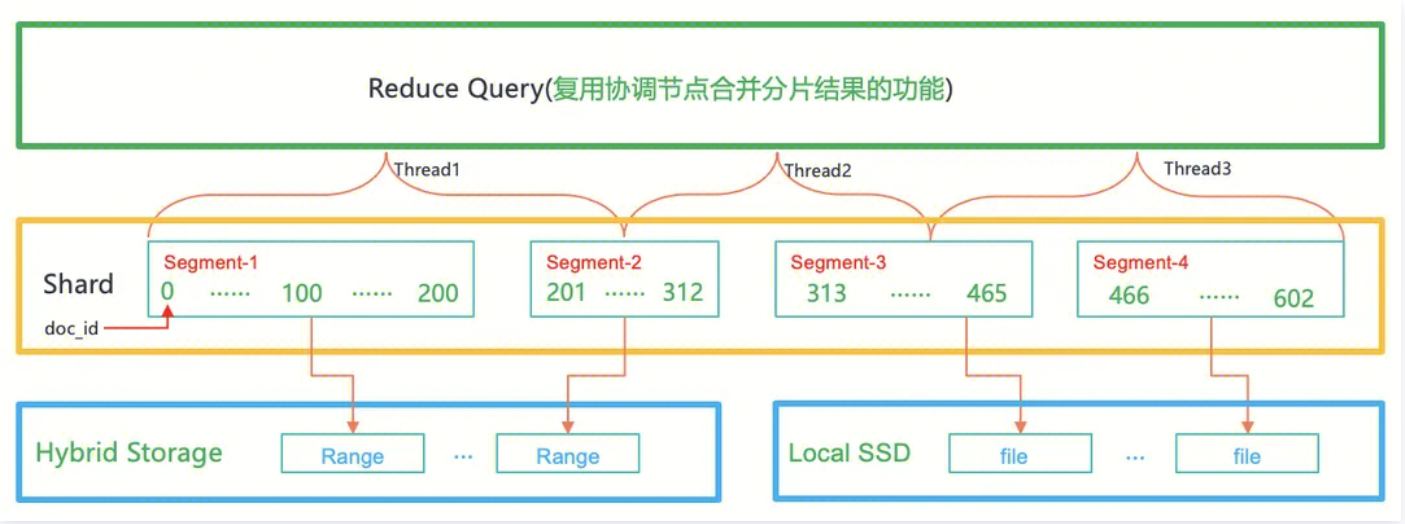
- Read-Write Partition Deployment: Designate some nodes in the cluster as dedicated Coordinators, which provide write computing capabilities. This setup is suitable for scenarios that prioritize improving write performance while also focusing on resource isolation.
- Read-Write Hybrid Deployment: Each node in the cluster provides write computing capabilities. Compared to partition deployment, this approach offers the advantage of more nodes providing write computing capabilities, resulting in a greater improvement in write performance. It is suitable for scenarios where the aim is solely to improve write performance without concern for resource isolation.

For specific features, refer to the Write Acceleration Features documentation.

Query/IO Parallelization

The ES query model splits the query request into shard-level subrequests, which are forwarded to each shard for execution. Finally, the Coordinator merges the results from each shard. Each shard contains multiple segments. By default, when ES executes shard-level queries, it processes each segment sequentially, using a single thread. Since the segments

within a shard are independent, it is possible to further split the subrequests and leverage multiple threads to process them in parallel. After the data nodes merge the results from multiple threads, they return the results to the Coordinator. The merging process at the data nodes is analogous to how the Coordinator merges results from each shard. This method aims to utilize idle CPU resources by breaking down a single shard-level request into 3-5 subrequests to be processed in parallel within the segments or docs of that shard. Each thread handles only a portion of the docs or segments. After the data nodes merge the results from each thread, they return them to the Coordinator, which then merges the results from each shard and sends them back to the client. This approach can lead to a significant performance improvement.
Search for the required CAM policy as needed, and click to complete policy association.



Test

Overall Effect

Tencent Cloud ES Brand-new Technology Stack: It adopts advanced technologies like storage-compute separation, write acceleration, and Query/IO parallelization. It is widely used in log scenarios to achieve integrated search of hot and cold data and AS capabilities, generally reducing costs by 30% – 80%.

- Storage-Compute Separation: Self-developed hybrid storage architecture achieves integrated search of hot and cold data, saving costs by 50% – 80%.
- Write Acceleration: Independent, self-closed loop, improves write throughput by over 3 times while isolating read and write resources.
- Query/IO Parallelization: Self-developed multi-level parallel query framework supports all query scenarios, improving query performance by 3 – 5 times.

Performance Stress Testing

Storage-Compute Separation

Stress Testing Environment:

- Cluster: 3 standard SA2 instances with 16 cores and 64GB, 1500GB SSD CBS x 1.
- Data: http_logs.
- Tools: esrally.

Stress Test Results:

As can be seen from the query performance loss, taking local disk as the baseline, Elastic's self-developed Searchable Snapshot has substantial performance loss. The log enhancement edition's self-developed storage-compute separation has acceptable query performance loss compared to local disks. In most scenarios, after adding parallel stress testing, it is 2 – 3 times faster than local disks. (Query latency unit: ms)

Query Type	Local SSD Disk	Searchable Snapshot	Storage-Compute Separation	Storage-Compute Separation + Parallelization
match_all (full match)	3	59	3	4
term (exact single value match)	6	71	16	7
terms(Multi-value Exact Match)	5	45	4	5
range(Range queries)	12	28	23	9
aggs(Aggregated Query)	1544	2575	2020	580
desc_sort_timestamp(Sort by Time Field in Descending Order)	65	187	81	33
asc_sort_timestamp(Sort by Time Field in Ascending Order)	71	256	54	8
desc_sort_with_after_timestamp (Add search_after in descending sort)	1140	1863	1968	440
asc_sort_with_after_timestamp (Add search_after in ascending sort)	936	1692	1389	395

Write Acceleration

Stress Testing Environment:

- Cluster: 3 standard SA2 instances with 16 cores and 64GB, 1500GB SSD CBS x 1.
- Data: Random generation.
- Tools: Code compilation to consume Kafka.

Stress Test Results:

Write Type	Replica	Batch Size	Write Performance	Note:
Default ES	1	5000	31w/s	Default ES Write Performance
Write Acceleration	1	5000	119w/s	3.8 times of ES

Query/IO Parallelization

Stress Testing Environment:

- Cluster: 3 standard SA2 instances with 16 cores and 64GB, 1500GB SSD CBS x 1.
- Data: geonames.
- Tools: esrally.
- Concurrency: IO Parallelization concurrency set to 3.

Stress Test Results:

- With IO Parallelization concurrency set to 3, performance generally increased by about 3 times. Stress test comparison showed that P50, P90, and P99 time consumption generally decreased by 5 – 10 times, with less and more stable query jitter.
- If CPU cores are more, subrequests can be split further, resulting in better performance. If the concurrency is set to 5, theoretically, performance would improve by around 5 times.

Category	Phrase Query (ms)	Aggregate Query (ms)	Script Query (ms)	Custom Scoring (ms)	View Documentation (ms)
Disable Parallelization	43	43	374	408	143
Enable Parallelization	5	4	132	125	44

Product Features

Last updated: 2024-10-24 21:10:55

Real-time logs of other cloud products such as CVM, TencentDB, and TKE, along with stored and incremental business data, can be aggregated and transferred to the ES Cluster for distributed data storage and query analysis.

Data Collection and Synchronization

- Users can use the Beats feature in Elasticsearch to transmit data to Elasticsearch for storage. Or, they can first transmit data to Logstash for custom conversion and parse, and then, they can transmit the processed data to Elasticsearch.
- ES provides the easy-to-use RESTful API for users to develop their own clients. They can call the data storage API to store data in ES clusters.
- ES is built within a Virtual Private Cloud (VPC), allowing users to easily use various data synchronization plugins to sync existing Cloud Services data to the ES cluster.

Data Storage

- ES provides different types of nodes and high-performance SSDs, ensuring the data read/write performance.
- Supports elastically scaling out to hundreds of nodes for data storage at the petabyte level. Satisfies the users' needs of different business scenarios.
- Supports detecting and replacing faulty nodes. Ensures cluster high availability.
- It features full-text search, vector retrieval, and hybrid search capabilities.

Visualization for Data Query and Analysis

- Elasticsearch has search features like full-text search, structured search, data filtering, and metric statistics. It can be applied to various scenarios such as information search and data analysis.
- Elasticsearch provides the easy-to-use RESTful API and clients in various programming languages for users to build their own search services.
- With Kibana, users can easily search and statistically analyze cluster data in a browser.

Service Performance Overview

Last updated: 2024-10-24 21:11:36

This section mainly describes the results of stress testing on different specifications of Tencent Cloud ES (Elasticsearch) instances (v7.10.1) using the benchmark rally script officially provided by Elasticsearch.
This section provides the stress test results of ES clusters with the following specifications:

- 4-Core 16 GB 3-Node Cluster Performance Test
- 8-Core 32 GB 3-Node Cluster Performance Test

It also provides the comparison of stress test results of 4-core 16 GB and 8-core 32 GB ES clusters. For more information, please see Stress Test Result Comparison Between 4-Core 16 GB 3-Node Cluster and 8-Core 32 GB 3-Node Cluster.

Elastic Stack (X-Pack)

Last updated: 2024-10-30 17:41:41

Overview

X-Pack features are Elasticsearch's official commercial features, including security, SQL, machine learning, and monitoring. It facilitates the application development and OPS management of Elasticsearch services. ES offers editions that come with such features, which you can select when purchasing and creating a cluster. The features in different editions are detailed below.

Purchase Guide

As shown in the figure above, there are options for the X-Pack features on the ES purchase page. ES offers three editions that have different X-Pack features as follows:

Comparison Item	Basic or Platinum Edition	Platinum Edition	Open Source
X-Pack included	✓	✓	×
X-Pack completeness	Partial	All	No

Purchase recommendation In order to be able to use more advanced features in ES, we recommend that you choose the **Platinum Edition** when you create a cluster. The specific features and differences of each edition are detailed below. For pricing information, please see ES Cluster Pricing.

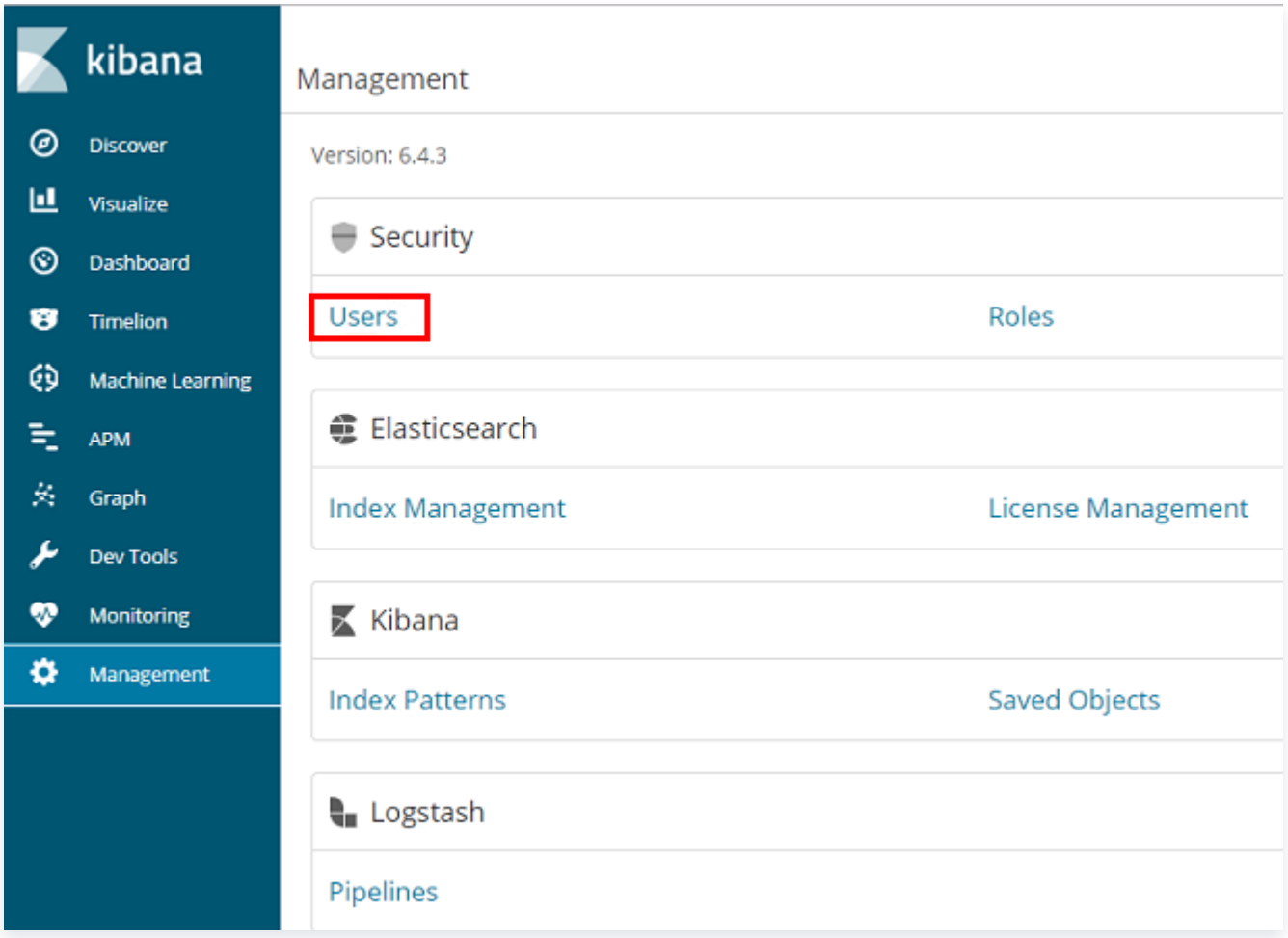
Advanced features introduction

This document describes some of the commonly used X-Pack features. For more information, please see the official Elastic Stack subscriptions and API documentation.

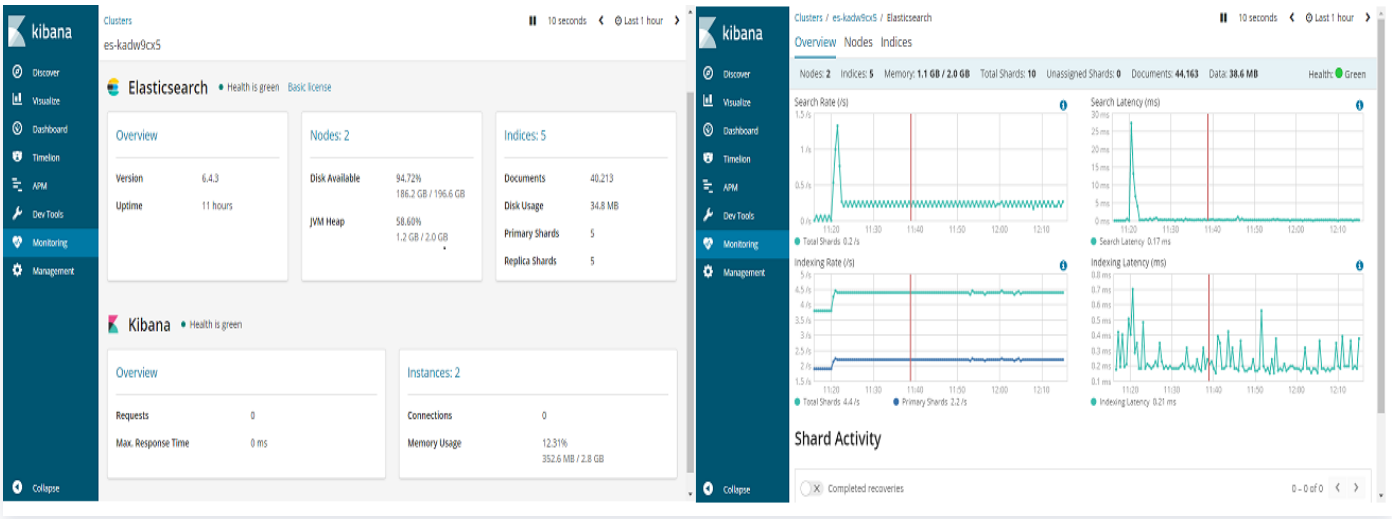
> **Note**
> - Some features vary by editions (Basic, Platinum, and Open Source).

- Some features are unavailable in earlier ES versions. If you encounter this issue, please consult through after-sales support.

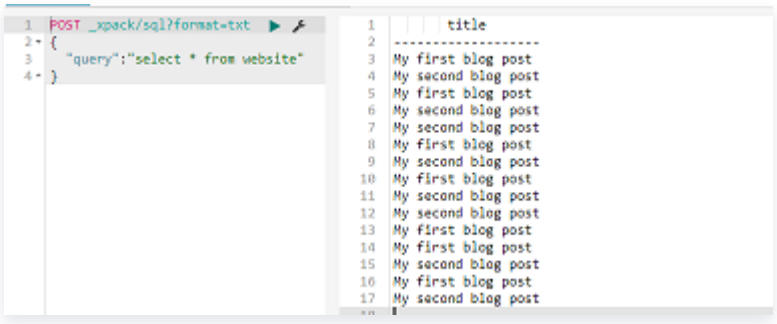
- **Security**This feature supports refined read/write permission control at the index and field levels and effectively protects data security by enabling data security protection and business access isolation, granting access to the right people, and preventing malicious attacks and data leakage.



- **Machine learning**In the application scenario of custom data alerting, it is sometimes difficult to set rules and thresholds to define the changes. In this case, the trend in data changes and reasonable fluctuation range can be predicted by the unattended machine learning feature, and when the data deviates from the normal trend, alarms will be triggered and notifications sent.
- **Monitoring** clusters, nodes, and indexes from multiple dimensions, providing comprehensive monitoring, real-time understanding of cluster operations, and supporting application development and operations.



- **SQL** offers full-text search, data statistical analysis features on Elasticsearch data through traditional SQL tools, supporting CLI, REST, and other access methods, **with the Platinum edition also supporting JDBC connections**. This allows for seamless integration with existing business systems, lowering the learning curve for new technologies.



> **Note**
> Regarding SQL support, the open-source version integrates other SQL plugins. For detailed information and usage, see elasticsearch-sql.

Detailed comparison among editions

This section mainly compares and highlights some key features of different Elasticsearch versions. As Elasticsearch is in a stage of rapid development, and the support for various features by different versions is constantly adjusted, we do not guarantee that the following content can stay in sync with the changes in the community.
To understand the latest and accurate feature comparison, refer to the official Elasticsearch introduction of Elastic Stack subscriptions.

> **Note**

In the table below●,◑,— are used to indicate the completeness of the corresponding feature, ⬚: includes all features; ◑: includes some features; —: does not include.

Module	Feature	Open Source	Basic or Platinum Edition	Platinum Edition
Elasticsearch	Scalability and resiliency	⬚	⬚	●
	Query and analytics	◑	◑	●
	Data Enrichment	●	●	●
	Management and Tools	◑	◑	●
	Security	—	⬚	⬚
	Machine Learning	—	—	●
Kibana	Explore and visualize	◑	◑	⬚
	Stack management and tooling	◑	◑	●
	Stack monitoring	—	◑	●
	Share and collaborate	◑	⬚	●
	Security	—	—	●
	Machine Learning	—	—	●
Beats	Data collection	◑	◑	●
	Data shipping	◑	◑	⬚

	Module	◑	◑	●
	Monitoring and management	—	◑	●
Logstash	Data collection	⬚	●	⬚
	Data Enrichment	●	⬚	●
	Data shipping	●	●	●
	Module	◑	●	⬚
	Monitoring and management	—	◑	●
ELASTIC APM	APM server	●	●	●
	APM agents	●	●	●
	APM dashboards in Kibana	●	●	●
	APM UI	—	●	●
	Distributed tracing	—	●	●
	Machine learning integration	—	⬚	●
ELASTIC Logs	Log shipper (Filebeat)	⬚	⬚	⬚
	Dashboards for common data sources	⬚	⬚	●
	Logs UI	⬚	●	●
Elastic Infrastructure	Metricbeat Indicator Collector	●	⬚	⬚

	Dashboards for common data sources	●	⬚	●
	Infrastructure UI	—	●	⬚
ELASTIC Operation Status Monitoring	Uptime monitor (Heartbeat)	●	●	⬚
	Uptime dashboards in Kibana	●	●	●
	Uptime UI	—	⬚	●

Detailed description of some Elasticsearch features:

> **Note**
> In the table below, ✓ means the feature is available, – means not available.

Elasticsearch Feature Modules	Details	Open Source	Basic or Platinum Edition	Platinum Edition
Management and Tools	RESTful APIs	✓	✓	✓
	Language clients	✓	✓	✓
	Snapshot/restore	✓	✓	✓
	_source only snapshot	–	✓	✓
	SQL interpreter CLI	–	✓	✓
	Data aggregation	–	✓	✓
	Index lifecycle management	–	✓	✓
	Frozen indices	–	✓	✓
	Upgrade Assistant APIs	–	✓	✓
	JDBC client	–	–	✓

	ODBC client	–	–	✓
Security	Encrypted communications	–	✓	✓
	Role-based access control	–	✓	✓
	File and native authentication	–	✓	✓
	Audit logging	–	–	–
	Attribute-based access control	–	–	✓
	Field- and document-level security	–	–	✓
	LDAP Authentication	–	–	✓
Machine Learning	Anomaly detection on time series	–	–	✓
	Input/Entity Analysis	–	–	✓
	Log message categorization	–	–	✓
	Root cause indication	–	–	✓
	Alerting on anomalies	–	–	✓
	Forecasting on time series	–	–	✓

Advantages

Last updated: 2024-10-24 21:10:33

Tencent Cloud ES offers cloud-hosted services, enabling users to easily create and manage Elasticsearch clusters, while ensuring high availability in production environments. Below are the core advantages of the product:

Ease of Deployment and Management

An ES cluster can be created with simple operations in a few minutes without the complex processes of software and hardware deployment and debugging. Additionally, ES comes with convenient cluster operation management tools, Kibana pages, and comprehensive cluster monitoring and alarm systems to facilitate customers' daily cluster operation and management needs.

Auto Scaling

ES offers various types of node specifications and storage media for you to choose from so that they can meet your business needs. As your business grows, clusters can be dynamically adjusted through scale-out and scale-in to meet your business needs and control costs effectively.

Elasticsearch X-Pack

ES integrates Elasticsearch X-Pack, which has advanced features such as security, SQL, and machine learning to improve the efficiency of security management, usage, and OPS of Elasticsearch clusters.

High Availability

ES can be deployed in multiple availability zones, guaranteeing non-stop service in the event of force majeure such as network or power failure in a single availability zone. A COS data backup policy can periodically back up data to ensure rapid recovery in case of data loss due to unexpected conditions. Dispersed Placement Groups place cluster nodes on different underlying hardware to reduce the risk of simultaneous failures of nodes with the same underlying hardware. In addition, ES features kernel optimization strategies that help comprehensively ensure data security and service stability.

Security Reinforcement

ES can be deployed in a logically isolated VPC, giving you full control over your environment configuration and the ability to customize network access control lists and security groups. It features a blocklisting/allowlisting mechanism for Kibana and IP access requests, and the

security capability of X-Pack enables access control at the field level, helping ensure the security of your resources in the cloud.

Openness and Service Integration

ES supports the complete system of ELK products and is compatible with standard open-source RESTful APIs and ecosystem components. It can be integrated with Tencent Cloud products such as COS, FL, CMQ, and TencentDB to implement data transfer and backup to meet your needs in different business scenarios.

Use Cases

Last updated: 2024-10-24 21:10:15

Log Analysis

During the operation of business systems, servers, databases, and containers generate a large amount of logs and monitoring data. These are dispersed, diverse, and massive in scale, making retrieval and analysis difficult. ES, through rich data collection tools and distributed storage, facilitates unified log management and real-time monitoring of indicators. The one-stop full observation advantage helps users quickly pinpoint issues and improve operational efficiency.

Information Retrieval

Elasticsearch is ideal for website search, mobile app search, and other scenarios, especially when dealing with large data volumes, high concurrency, and high requirements for search flexibility and relevance. It can return search results in milliseconds from PB-scale structured and unstructured data using flexible keywords, query conditions, and fuzzy matching.

Vector Search

Vector Retrieval is a retrieval technique based on the Vector Space Model. It converts data such as text, images, and videos into numerical vectors to conduct similarity searches in the vector space, overcoming the limitations of traditional text searches which can only be based on keywords and not on semantic searches. ES provides a one-stop solution from Vector Generation to Vector Indexing, Storage, and Retrieval, helping users efficiently build applications such as Semantic Search, Image Search, and Product Recommendation.

Retrieval-Augmented Generation (RAG)

Large Language Models (LLMs) face numerous challenges in enterprise applications, including the lack of enterprise private domain knowledge, hallucinations, and knowledge updates. RAG combines retrieval and generation technology, utilizing inputs from the enterprise knowledge base to improve the accuracy of LLM responses. ES provides one-stop services around RAG, including data slicing, vector retrieval, text and vector hybrid search, rerank, and large model integration, surpassing the single-point solution of traditional Tencent Cloud VectorDB and helping enterprises easily build AI assistants, knowledge Q&A, and other scenarios.

Data analysis

In the context of data-driven operations, industries such as e-commerce, mobile applications, and advertising media need to leverage data analysis and data mining to assist business decisions. The massive scale of business data brings significant challenges to statistical analysis. ES has structured query capabilities, supporting complex filtering and aggregation statistical features, helping clients efficiently perform personalized statistical analysis on massive data, identify problems and opportunities, and assist in business decisions, thus realizing the true value of data.

Database Query Acceleration

Relational databases tend to focus on transactional queries and often face challenges such as insufficient query performance and poor scalability in scenarios with massive data scale. ES offers elastic scalability and high concurrency as well as low latency query capabilities for massive data, using data synchronization tools to maintain database synchronization. It supports SQL capabilities to meet clients' database query acceleration needs, addressing the shortcomings of traditional databases.

Capabilities and Restrictions

Last updated: 2024-10-24 21:09:57

Tencent Cloud ES is a cloud-based PaaS service developed on the open-source software Elasticsearch. Through Tencent Cloud ES, you can quickly set up Elasticsearch cluster services and develop applications such as log analysis and data search. Below are the product capabilities and usage restrictions.

Product Components

Tencent Cloud ES consists of core components: the Elasticsearch Cluster and the data visualization analysis tool Kibana. For data collection and transmission to the ES cluster, you can deploy data collection tools like Beats or Logstash according to your business needs, or develop applications to write the data into the ES cluster.

Configuration Options

SSD Cloud Disk, Single Node, Disk Size Limitation are related to instance specifications. For ES versions 6.8 and above, the purchasable disk size range is as follows:

Type	Model Specification	Capacity Limit (GB)
Standard instance family (S1,SA5,SA2)	2-core 4 GB	20 - 2000
	2-core 8 GB	20 - 2000
	4-core 8 GB	20 - 4000
	4-core 16 GB	20 - 6000
	8 cores and 16 GB or above	20 - 30000
Memory-optimized Instance Family (M1)	All Specifications	20 - 30000

Note:
As the ES cluster is often constructed in a distributed multi-node form, a primary node is required for unified cluster management. To prevent split-brain issues caused by potential node failures, it is recommended to choose at least three nodes to build the

> cluster. **For configuration options,** refer to Evaluation of Cluster Specification and Capacity Configuration.

Network Access

VPC Private Network Access

To ensure data security, Tencent Cloud ES is built within the user's VPC and can only be accessed through the VPC for data writing and querying. If you need to access the ES cluster via the public network for development and debugging, you can do so through VPN Connections to connect the local IDC with the cloud VPC, or via External Network Proxy. Please also take necessary measures to protect data security.

Kibana Page

Public network access to the Kibana Page is supported. To ensure data security, the Kibana Page requires setting a password and configuring an access IP allowlist.

VPC Network Selection

Once Tencent Cloud ES is created, switching VPC networks is not supported. Please plan your business deployment in advance when creating a cluster.

Relevant Concepts

Last updated: 2024-10-24 21:09:39

Elasticsearch clusters are usually composed of multiple nodes to form a distributed cluster. The nodes communicate and cooperate with each other to collectively provide search and indexing services (nodes are capable of redirecting client requests to the appropriate node). Different nodes take on different roles; some may handle a single role, while others might handle multiple roles. In Elasticsearch, there are various node roles like data nodes, primary nodes, machine learning nodes, and coordinators.

Data Node

It is mainly responsible for operations related to the storing, processing, and manipulating of data and index shards, such as I/O-, memory-, and CPU-intensive operations like CRUD, search, and aggregation. During the use of cluster, you should closely monitor the resource utilization of the data nodes and ensure cluster stability by adding more nodes to scale the cluster up when the service is overloaded.

primary node (Master Node)

Responsible for lightweight, cluster-wide operations such as creating or deleting indices, tracking which nodes are part of the cluster, and deciding which shards to assign to which nodes. It is important to have a stable master node for the cluster health.

Master-eligible Node

Refers to a node that is eligible to be selected as a master node. Any node that meets the requirements for a master node (by default, all nodes) can be selected as a master node through the master selection process.
By default, all nodes are data nodes and also candidate master nodes, which is very convenient for small clusters. Because the requests for index processing and data searching are I/O-, memory-, and CPU-intensive for data nodes, they may cause pressure on the node resources. As the cluster grows, in order to ensure that the master nodes are stable and free from pressure and to ensure cluster stability, the master nodes should be separated from the data nodes.

Dedicated Master Node

Refers to a node set to serve only as a master node in an Elasticsearch cluster.

Recommendations for Dedicated Master Node Configuration

Setting dedicated master nodes mainly ensures cluster stability as it scales. It is recommended to have at least 3 dedicated master nodes.

- If the number of dedicated master nodes is 1, there is only one master-eligible node. discovery.zen.minimum_master_nodes can only be set to 1, and there is no backup in case of network failure.
- If the number of dedicated master nodes is 2, there are 2 master-eligible nodes. If minimum_master_nodes is set to 1, although there is a backup node, there may be a risk of split-brain (i.e., each master-eligible node sets itself as the master node) when the master node is re-selected in case of network failure. If minimum_master_nodes is set to 2, as the number of master-eligible nodes falls short, no master node can be selected in case of failure.
- If the number of dedicated master nodes is 3, there are 3 master-eligible nodes. If discovery.zen.minimum_master_nodes is set to 2, even if one master-eligible node is lost in case of network failure, there is still one master node that can be re-selected.

Machine Learning Nodes

Machine Learning Nodes are used to create machine learning tasks, automatically perform data analysis, identify abnormal data, load vector models to enhance vector generation and vector retrieval capabilities, and isolate from business operations to improve cluster stability.

Coordinator (Coordinating Node)

Coordinator is mainly responsible for coordinating client requests, such as search requests or batch index requests, distributing received requests to the appropriate data nodes. Each data node executes the request locally and gathers the results to the Coordinator. Every ES node can serve as a Coordinator. In CPU-intensive scenarios like high concurrency read/write and multi-aggregation queries, having an independent Coordinator helps reduce the load on primary nodes and data nodes.

For more detailed explanation, see ES Node.