Alibaba Cloud Elasticsearch

Product Introduction

Issue: 20190806

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. **You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.**
2. **No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company, or individual in any form or by any means without the prior written consent of Alibaba Cloud.**
3. **The content of this document may be changed due to product version upgrades , adjustments, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and the updated versions of this document will be occasionally released through Alibaba Cloud-authorized channels. You shall pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.**
4. **This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides the document in the context that Alibaba Cloud products and services are provided on an "as is", "with all faults " and "as available" basis. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity , applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not bear any liability for any errors or financial losses incurred by any organizations, companies, or individuals arising from their download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, bear responsibility for any indirect, consequential, exemplary, incidental, special, or punitive damages, including lost profits arising from the use**

or trust in this document, even if Alibaba Cloud has been notified of the possibility of such a loss.

5. By law, all the content of the Alibaba Cloud website, including but not limited to works, products, images, archives, information, materials, website architecture, website graphic layout, and webpage design, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of the Alibaba Cloud website, product programs, or content shall be used, modified , reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates . The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates).
6. Please contact Alibaba Cloud directly if you discover any errors in this document.

Generic conventions

Table -1: Style conventions

Style	Description	Example
	This warning information indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	 Danger: Resetting will result in the loss of user configuration data.
	This warning information indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	 Warning: Restarting will cause business interruption. About 10 minutes are required to restore business.
	This indicates warning information, supplementary instructions, and other content that the user must understand.	 Notice: Take the necessary precautions to save exported data containing sensitive information.
	This indicates supplemental instructions, best practices, tips, and other content that is good to know for the user.	 Note: You can use Ctrl + A to select all files.
>	Multi-level menu cascade.	Settings > Network > Set network type
Bold	It is used for buttons, menus , page names, and other UI elements.	Click OK.
`Courier font`	It is used for commands.	Run the `cd / d C :/ windows` command to enter the Windows system folder.
Italics	It is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	It indicates that it is a optional value, and only one item can be selected.	`ipconfig` *[-all|-t]*

Style	Description	Example
{} or {a|b}	It indicates that it is a required value, and only one item can be selected.	`swich {stand | slave}`

Contents

1 What is Elasticsearch

Elasticsearch is a Lucene-based data search and analysis tool that provides distributed services. Elasticsearch is an open-source product that complies with the Apache open standards. It is the mainstream search engine for enterprise data.

Alibaba Cloud Elasticsearch includes multiple versions, including Elasticsearch 5.5.3 with Commercial Feature, Elasticsearch 6.3.2 with Commercial Feature, and Elasticsearch 6.7.0 with Commercial Feature. It also contains the X-Pack plug-in. You can use Alibaba Cloud Elasticsearch to analyze and search data. Alibaba Cloud Elasticsearch provides enterprise-class access control, security monitoring and alarms, and automatic reporting based on the open-source Elasticsearch engine.

X-Pack is an Elastic Stack extension that bundles security, alarming, monitoring, reporting, and graph capabilities into one easy-to-install package. X-Pack is integrated into Kibana to provide services such as permission verification, role permission management, real-time monitoring, visualized reports, and machine learning.

Features

- Distributed real-time file storage. Every field is indexed and searchable.
- A distributed real-time analysis and search engine.
- X-Pack commercial edition for enterprise-class permission management and real-time system monitoring.
- Scaling to hundreds of servers for processing petabytes of structured and unstructured data.
- Supports the IK analyzer.
- 24/7 technical support from official Elasticsearch Technical Support.

Built-in plug-ins.

The built-in plug-ins provided by Alibaba Cloud Elasticsearch include but are not limited to the following:

- IK Analyzer: a lightweight Chinese analyzer kit based on java. It is a popular plug-in for language analysis in the open-source community,
- Pinyin Analyzer: Pinyin analyzer.
- Smart Chinese Analysis Plugin: the default Lucene Chinese analyzer.

- **ICU Analysis Plug-in: A Lucene ICU analyzer. ICU is a set of stable, tested, powerful , and easy to use libraries, providing Unicode and globalization support for software applications.**
- **Mapper Attachments Type Plug-in: allows Elasticsearch index file attachments in common formats by using the Apache text extraction library Tika.**

2 Concept

Cluster

A cluster consists of one or more nodes. Each cluster has a master node, which is automatically polled by the cluster. Master and subordinate nodes are in the scope of a cluster. Elasticsearch applies a decentralized model. It does not have a central node. Therefore, communicating with a node in a cluster is the same as communicating with the cluster.

Shards

Elasticsearch divides an index into multiple shards and distributes these shards across nodes, allowing you to search by using the index on any of the nodes. The number of shards for an index must be specified before the index has been created. After an index has been created, you can no longer change the number of shards for the index.

Replicas

You can create multiple index replicas to enhance the fault tolerance of the system. A replica can be restored to a shard when the shard has been damaged or lost. Using replicas also improves the search performance. Search requests can be load balanced by Elasticsearch among these replicas.

Recovery

Data recovery (or data redistribution) is the process of redistributing shards for a node to guarantee the integrity of data when the node joins or leaves a cluster, or when the node recovers from a failure.

Gateway

A gateway is used to store snapshots of indexes. By default, an Elasticsearch node stores all the indexes in memory. When the node memory is full, the node saves the indexes to local disks for persistent storage. Index snapshots stored on a gateway can be restored after a cluster restarts for fault recovery, which is faster than reading indexes from local disks. Elasticsearch supports multiple types of gateways, including local file system, distributed file system, Hadoop HDFS, and Alibaba Cloud Object Storage Service (OSS).

discovery.zen

discovery.zen is an automatic node discovery mechanism. Elasticsearch is a peer to peer (P2P) system that broadcasts to discover nodes. Nodes communicate with each other through multicast and P2P.

Transport

Transport is a method used for communication between nodes within a cluster, or between clusters and clients. By default, nodes communicate with each other over TCP. Elasticsearch also supports multiple transmission protocol plug-ins, including HTTP (JSON format), Thrift, Servlet, Memcached, and ZeroMQ.

3 Restful API

Introduction

Elasticsearch provides a RESTful Web API, which allows you to perform operations including addition, deletion, modification, search, and alias configuration.

For more information about the official Elasticsearch REST API, see Elasticsearch Restful API.

Elasticsearch Reference [5.5]

Single document APIs

- Index API
- Get API
- Delete API
- Update API

Multi-document APIs

- Multi Get API
- Bulk API
- Delete By Query API
- Update By Query API
- Reindex API

Use REST clients to communicate with clusters

You can use REST clients to access Elasticsearch clusters through HTTP or TCP. We recommend that you use the Elasticsearch official Java REST Client.

Use Java APIs to communicate with clusters

Elasticsearch provides a default client for Java users. For more information, see Java API.

Transport client

A transport client forwards requests to nodes in a cluster. However, it is not a part of a cluster.

A transport client uses the Elasticsearch Transport Protocol to communicate with clusters over port `9300`.

Nodes in a cluster also use port `9300` to communicate with each other. You must open port 9300 for your nodes before grouping them into a cluster.

Note:
Your Java clients and nodes must use the same Elasticsearch version for them to recognize each other.

RESTful API (HTTP)

All other languages can communicate with Elasticsearch over port `9200` using a RESTful API, with your favorite web client. You can use the curl command to communicate with Elasticsearch at the CLI.

Note:

Elasticsearch provides official clients for several languages, including `Groovy`, `Javascript`, `.NET`, `PHP`, `Perl`, `Python`, and `Ruby`.

For more clients and plug-ins provided by communities, see Document.

Structure of a curl request over HTTP

```
Curl - X  < VERB > '< PROTOCOL >: // < HOST  >:<  PORT >/< PATH >? <
QUERY_STRI  NG > '- d ' < BODY >'
```

- VERB: an HTTP method: `GET`, `POST`, `PUT`, `HEAD`, and `DELETE`.
- PROTOCOL: `http` or `https`. Use `https` only if HTTPS is enabled for Elasticsearch.
- HOST: the hostname of any node in your Elasticsearch cluster or `localhost` for a local node.
- PORT: the port that runs the Elasticsearch HTTP service. The default port number is `9200`.
- PATH: API endpoint, for example, `_count` returns the number of documents in a cluster. PATH may contain multiple components, such as `_cluster / stats` and `_nodes / stats / jvm`.
- QUERY_STRING: optional query parameters. For example, the `? pretty` parameter makes the JSON response much easier to read.

· BODY: JSON-encoded request body (only if the request requires one).

Example

Count the number of documents in an Elasticsearch cluster:

```
 curl  - XGET  ' http :// localhost : 9200 / _count ? pretty ' - d  '
{
  " query ": {
    " match_all ": {}
  }
}'
```

The body of the response to the curl request:

```
{
    " count " :  0 ,
    " _shards " : {
        " total " :  5 ,
        " successful " :  5 ,
        " failed " :  0
     }
}
```

Use the curl `- i` command to display the HTTP header:

```
curl  - i  - XGET  ' localhost : 9200 /'
```

Full request:

```
 Curl - XGET  ' localhost :  9200 / _   count ?  pretty  '- d '
{
    " query ": {
        " match_all ": {}
    }
}'
```

Shorthand for the request:

```
 GET  / _count
{
    " query ": {
        " match_all ": {}
    }
}
```

4 Performance

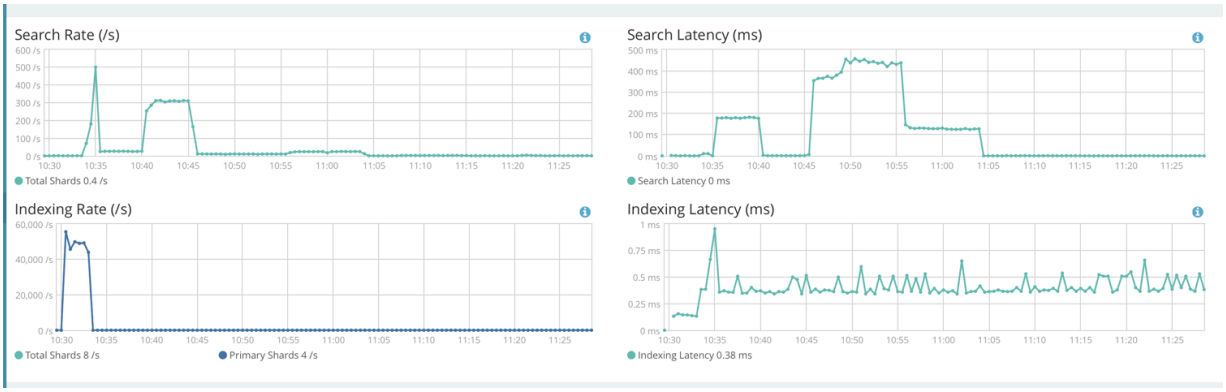
Performance benchmarking

All benchmarks are run by the Elasticsearch official macrobenchmarking framework Rally, on Alibaba Cloud Elasticsearch cluster version 5.5.3 in the Hangzhou region. The test reports are as follows:

> **Note:**
> Benchmarking uses the official geonames data, which is 3.3 GB in size and includes up to 11,520,617 documents.

Test on Cluster 1 (including three nodes, with four CPU cores and 16 GB memory per node)

Kibana metrics



Metric	Operation	Value	Unit
Indexing time		26.3543	min
Merge time		11.0297	min
Refresh time		3.05238	min
Flush time		0.04485	min
Merge throttle time		1.39282	min
Total Young Gen GC		92.902	s
Total Old Gen GC		0.4	s
Heap used for segments		18.7955	MB
Heap used for doc values		0.360752	MB

Metric	Operation	Value	Unit
Heap used for terms		17.2739	MB
Heap used for norms		0.0877075	MB
Heap used for points		0.241213	MB
Heap used for stored fields		0.831932	MB
Segment count		133	
Min Throughput	Index-append	51751.7	docs/s
Median Throughput	Index-append	52303	docs/s
Max Throughput	Index-append	54076.3	docs/s
50th percentile latency	Index-append	743.939	ms
90th percentile latency	index-append	1045.7	ms
99th percentile latency	index-append	1325.21	ms
100th percentile latency	index-append	1794.39	ms
50th percentile service time	index-append	743.939	ms
90th percentile service time	index-append	1045.7	ms
99th percentile service time	index-append	1325.21	ms
100th percentile service time	index-append	1794.39	ms
error rate	index-append	0	%
Min Throughput	Force-merge	0.95	ops/s
Median Throughput	force-merge	0.95	ops/s
Max Throughput	force-merge	0.95	ops/s
100th percentile latency	Force-merge	1052.54	ms

Metric	Operation	Value	Unit
100th percentile service time	force-merge	1052.54	ms
error rate	force-merge	0	%
Min Throughput	index-stats	100.04	ops/s
Median Throughput	index-stats	100.05	ops/s
Max Throughput	index-stats	100.09	ops/s
50th percentile latency	index-stats	4.85232	ms
90th percentile latency	index-stats	5.14185	ms
99th percentile latency	index-stats	77.3127	ms
99.9th percentile latency	index-stats	123.888	ms
100th percentile latency	index-stats	128.01	ms
50th percentile service time	index-stats	4.78006	ms
90th percentile service time	index-stats	4.9831	ms
99th percentile service time	index-stats	9.66475	ms
99.9th percentile service time	index-stats	48.4445	ms
100th percentile service time	index-stats	127.945	ms
error rate	index-stats	0	%
Min Throughput	node-stats	100.05	ops/s
Median Throughput	node-stats	100.1	ops/s
Max Throughput	node-stats	100.55	ops/s
50th percentile latency	node-stats	4.55259	ms
90th percentile latency	node-stats	4.78784	ms

Metric	Operation	Value	Unit
99th percentile latency	node-stats	18.8034	ms
99.9th percentile latency	node-stats	43.7684	ms
100th percentile latency	node-stats	48.1474	ms
50th percentile service time	node-stats	4.48138	ms
90th percentile service time	node-stats	4.69386	ms
99th percentile service time	node-stats	5.64618	ms
99.9th percentile service time	node-stats	27.8155	ms
100th percentile service time	node-stats	43.6905	ms
error rate	node-stats	0	%
Min Throughput	default	49.81	ops/s
Median Throughput	default	50	ops/s
Max Throughput	default	50	ops/s
50th percentile latency	default	19.7245	ms
90th percentile latency	default	94.1457	ms
99th percentile latency	default	133.091	ms
99.9th percentile latency	Default	137.285	ms
100th percentile latency	default	138.043	ms
50th percentile service time	default	19.1469	ms
90th percentile service time	default	19.9554	ms

Metric	Operation	Value	Unit
99th percentile service time	default	25.3462	ms
99.9th percentile service time	default	54.7931	ms
100th percentile service time	default	133.771	ms
error rate	default	0	%
Min Throughput	term	200.05	ops/s
Median Throughput	term	200.08	ops/s
Max Throughput	term	200.12	ops/s
50th percentile latency	term	3.07948	ms
90th percentile latency	term	3.37296	ms
99th percentile latency	term	22.3272	ms
99.9th percentile latency	term	26.9648	ms
100th percentile latency	term	28.1562	ms
50th percentile service time	term	3.00599	ms
90th percentile service time	term	3.15279	ms
99th percentile service time	term	4.22302	ms
99.9th percentile service time	term	26.9017	ms
100th percentile service time	term	28.0823	ms
error rate	term	0	%
Min Throughput	phrase	199.84	ops/s
Median Throughput	phrase	200.04	ops/s
Max Throughput	phrase	200.09	ops/s

Metric	Operation	Value	Unit
50th percentile latency	phrase	3.76927	ms
90th percentile latency	phrase	13.6055	ms
99th percentile latency	phrase	28.0245	ms
99.9th percentile latency	phrase	34.7198	ms
100th percentile latency	phrase	35.551	ms
50th percentile service time	phrase	3.67227	ms
90th percentile service time	phrase	4.08037	ms
99th percentile service time	Phrase	16.9256	ms
99.9th percentile service time	phrase	24.4886	ms
100th percentile service time	phrase	29.8604	ms
error rate	phrase	0	%
Min Throughput	country_agg_uncached	4.95	ops/s
Median Throughput	country_agg_uncached	4.99	ops/s
Max Throughput	country_agg_uncached	5	ops/s
50th percentile latency	country_agg_uncached	330.923	ms
90th percentile latency	country_agg_uncached	2780.17	ms
99th percentile latency	country_agg_uncached	2866	ms
99.9th percentile latency	country_agg_uncached	2880.39	ms

Metric	Operation	Value	Unit
100th percentile latency	country_agg_uncached	2882.11	ms
50th percentile service time	country_agg_uncached	197.883	ms
90th percentile service time	country_agg_uncached	213.402	ms
99th percentile service time	country_agg_uncached	256.649	ms
99.9th percentile service time	country_agg_uncached	290.496	ms
100th percentile service time	country_agg_uncached	296.875	ms
error rate	country_agg_uncached	0	%
Min Throughput	country_agg_cached	99.92	ops/s
Median Throughput	country_agg_cached	100.06	ops/s
Max Throughput	country_agg_cached	100.11	ops/s
50th percentile latency	country_agg_cached	3.30479	ms
90th percentile latency	country_agg_cached	3.52514	ms
99th percentile latency	country_agg_cached	52.8258	ms
99.9th percentile latency	country_agg_cached	112.895	ms
100th percentile latency	country_agg_cached	119.435	ms
50th percentile service time	country_agg_cached	3.23149	ms
90th percentile service time	country_agg_cached	3.41319	ms

Metric	Operation	Value	Unit
99th percentile service time	country_agg_cached	7.60955	ms
99.9th percentile service time	country_agg_cached	26.2229	ms
100th percentile service time	country_agg_cached	119.365	ms
error rate	country_agg_cached	0	%
Min Throughput	scroll	61.59	ops/s
Median Throughput	scroll	61.67	ops/s
Max Throughput	scroll	61.94	ops/s
50th percentile latency	scroll	164549	ms
90th percentile latency	scroll	237443	ms
99th percentile latency	scroll	253860	ms
100th percentile latency	scroll	255710	ms
50th percentile service time	scroll	399.964	ms
90th percentile service time	scroll	424.303	ms
99th percentile service time	scroll	523.877	ms
100th percentile service time	scroll	639.45	ms
error rate	scroll	0	%
Min Throughput	expression	2	ops/s
Median Throughput	expression	2	ops/s
Max Throughput	expression	2	ops/s
50th percentile latency	expression	409.927	ms

Metric	Operation	Value	Unit
90th percentile latency	expression	434.544	ms
99th percentile latency	expression	532.412	ms
100th percentile latency	expression	537.618	ms
50th percentile service time	expression	409.812	ms
90th percentile service time	expression	428.156	ms
99th percentile service time	expression	532.33	ms
100th percentile service time	expression	537.495	ms
error rate	expression	0	%
Min Throughput	painless_static	2	ops/s
Median Throughput	painless_static	2	ops/s
Max Throughput	painless_static	2	ops/s
50th percentile latency	painless_static	497.626	ms
90th percentile latency	painless_static	643.32	ms
99th percentile latency	painless_static	700.559	ms
100th percentile latency	painless_static	704.679	ms
50th percentile service time	painless_static	490.705	ms
90th percentile service time	painless_static	500.663	ms
99th percentile service time	painless_static	642.124	ms
100th percentile service time	painless_static	683.621	ms

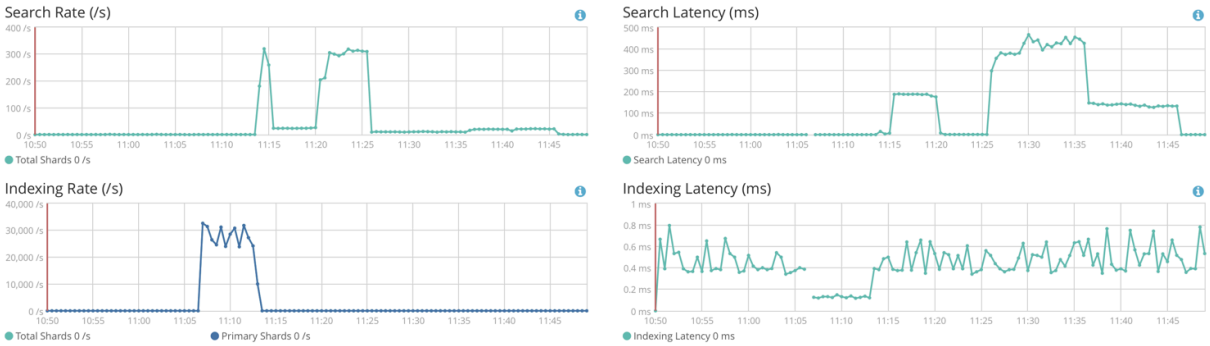
Metric	Operation	Value	Unit
error rate	painless_static	0	%
Min Throughput	painless_dynamic	2	ops/s
Median Throughput	painless_dynamic	2	ops/s
Max Throughput	painless_dynamic	2	ops/s
50th percentile latency	painless_dynamic	473.087	ms
90th percentile latency	painless_dynamic	554.729	ms
99th percentile latency	painless_dynamic	668.363	ms
100th percentile latency	painless_dynamic	706.557	ms
50th percentile service time	painless_dynamic	469.145	ms
90th percentile service time	painless_dynamic	501.774	ms
99th percentile service time	painless_dynamic	606.61	ms
100th percentile service time	painless_dynamic	624.751	ms
error rate	painless_dynamic	0	%
Min Throughput	large_filtered_terms	1.64	ops/s
Median Throughput	large_filtered_terms	1.64	ops/s
Max Throughput	large_filtered_terms	1.65	ops/s
50th percentile latency	large_filtered_terms	33013.5	ms
90th percentile latency	large_filtered_terms	40869	ms
99th percentile latency	large_filtered_terms	42644	ms

Metric	Operation	Value	Unit
100th percentile latency	large_filtered_terms	42936.2	ms
50th percentile service time	large_filtered_terms	598.001	ms
90th percentile service time	large_filtered_terms	626.81	ms
99th percentile service time	large_filtered_terms	771.815	ms
100th percentile service time	large_filtered_terms	796.884	ms
error rate	large_filtered_terms	0	%
Min Throughput	large_prohibited_terms	1.69	ops/s
Median Throughput	large_prohibited_terms	1.69	ops/s
Max Throughput	large_prohibited_terms	1.7	ops/s
50th percentile latency	large_prohibited_terms	27732.3	ms
90th percentile latency	large_prohibited_terms	34305.5	ms
99th percentile latency	large_prohibited_terms	35840.4	ms
100th percentile latency	large_prohibited_terms	35993.5	ms
50th percentile service time	large_prohibited_terms	586.382	ms
90th percentile service time	large_prohibited_terms	618.185	ms
99th percentile service time	large_prohibited_terms	661.378	ms
100th percentile service time	large_prohibited_terms	823.782	ms

Metric	Operation	Value	Unit
error rate	large_prohibited_terms	0	%

Test on Cluster 2 (including three nodes, with two CPU cores and 8 GB memory per node)

Kibana metrics



Metric	Operation	Value	Unit
Indexing time		23.9479	min
Merge time		14.3001	min
Refresh time		5.26405	min
Flush time		0.0308333	min
Merge throttle time		1.27945	min
Total Young Gen GC		183.74	s
Total Old Gen GC		1.125	s
Heap used for segments		18.8167	MB
Heap used for doc values		0.452751	MB
Heap used for terms		17.2004	MB
Heap used for norms		0.0852051	MB
Heap used for points		0.241465	MB
Heap used for stored fields		0.836876	MB

Metric	Operation	Value	Unit
Segment count		140	
Min Throughput	index-append	28115.4	docs/s
Median Throughput	index-append	28645.5	docs/s
Max Throughput	index-append	30037.8	docs/s
50th percentile latency	index-append	1447.76	ms
90th percentile latency	index-append	1847.05	ms
99th percentile latency	index-append	2264.68	ms
99.9th percentile latency	index-append	2515.95	ms
100th percentile latency	index-append	2608.68	ms
50th percentile service time	index-append	1447.76	ms
90th percentile service time	index-append	1847.05	ms
99th percentile service time	index-append	2264.68	ms
99.9th percentile service time	index-append	2515.95	ms
100th percentile service time	index-append	2608.68	ms
error rate	index-append	0	%
Min Throughput	force-merge	2.1	ops/s
Median Throughput	force-merge	2.1	ops/s
Max Throughput	force-merge	2.1	ops/s
100th percentile latency	force-merge	475.984	ms
100th percentile service time	force-merge	475.984	ms
error rate	force-merge	0	%

Metric	Operation	Value	Unit
Min Throughput	index-stats	97.75	ops/s
Median Throughput	index-stats	100.05	ops/s
Max Throughput	index-stats	100.07	ops/s
50th percentile latency	index-stats	5.09015	ms
90th percentile latency	index-stats	10.7365	ms
99th percentile latency	index-stats	234.761	ms
99.9th percentile latency	index-stats	277.393	ms
100th percentile latency	index-stats	281.866	ms
50th percentile service time	index-stats	5.01096	ms
90th percentile service time	index-stats	5.30021	ms
99th percentile service time	index-stats	12.0005	ms
99.9th percentile service time	index-stats	141.631	ms
100th percentile service time	index-stats	150.153	ms
error rate	index-stats	0	%
Min Throughput	node-stats	100.01	ops/s
Median Throughput	node-stats	100.08	ops/s
Max Throughput	node-stats	100.49	ops/s
50th percentile latency	node-stats	4.90659	ms
90th percentile latency	node-stats	5.29285	ms
99th percentile latency	node-stats	29.3245	ms

Metric	Operation	Value	Unit
99.9th percentile latency	node-stats	43.3885	ms
100th percentile latency	node-stats	44.6019	ms
50th percentile service time	node-stats	4.83552	ms
90th percentile service time	node-stats	5.12694	ms
99th percentile service time	node-stats	9.08739	ms
99.9th percentile service time	node-stats	39.744	ms
100th percentile service time	node-stats	44.5383	ms
error rate	node-stats	0	%
Min Throughput	default	47.83	ops/s
Median Throughput	default	48.28	ops/s
Max Throughput	default	48.73	ops/s
50th percentile latency	default	617.465	ms
90th percentile latency	default	1033.98	ms
99th percentile latency	default	1083.23	ms
99.9th percentile latency	default	1095.4	ms
100th percentile latency	default	1097.14	ms
50th percentile service time	default	18.646	ms
90th percentile service time	default	24.9381	ms
99th percentile service time	default	35.7667	ms

Metric	Operation	Value	Unit
99.9th percentile service time	default	57.3679	ms
100th percentile service time	default	151.505	ms
error rate	default	0	%
Min Throughput	term	199.43	ops/s
Median Throughput	term	200.07	ops/s
Max Throughput	term	200.13	ops/s
50th percentile latency	term	2.9728	ms
90th percentile latency	term	7.10648	ms
99th percentile latency	term	22.4487	ms
99.9th percentile latency	term	29.0737	ms
100th percentile latency	term	29.6253	ms
50th percentile service time	term	2.87833	ms
90th percentile service time	term	3.08983	ms
99th percentile service time	term	19.9777	ms
99.9th percentile service time	term	29.0082	ms
100th percentile service time	term	29.5597	ms
error rate	term	0	%
Min Throughput	phrase	199.71	ops/s
Median Throughput	phrase	200.04	ops/s
Max Throughput	phrase	200.07	ops/s
50th percentile latency	phrase	3.61484	ms

Metric	Operation	Value	Unit
90th percentile latency	phrase	16.5523	ms
99th percentile latency	phrase	31.394	ms
99.9th percentile latency	phrase	33.902	ms
100th percentile latency	phrase	34.5784	ms
50th percentile service time	phrase	3.47402	ms
90th percentile service time	phrase	3.90958	ms
99th percentile service time	phrase	19.3773	ms
99.9th percentile service time	phrase	22.7947	ms
100th percentile service time	phrase	27.8164	ms
error rate	phrase	0	%
Min Throughput	country_agg_uncached	4.63	ops/s
Median Throughput	country_agg_uncached	4.65	ops/s
Max Throughput	country_agg_uncached	4.67	ops/s
50th percentile latency	country_agg_uncached	14864.3	ms
90th percentile latency	country_agg_uncached	21046	ms
99th percentile latency	country_agg_uncached	22902	ms
99.9th percentile latency	country_agg_uncached	22997.6	ms
100th percentile latency	country_agg_uncached	23018.7	ms

Metric	Operation	Value	Unit
50th percentile service time	country_agg_uncached	204.174	ms
90th percentile service time	country_agg_uncached	242.492	ms
99th percentile service time	country_agg_uncached	345.382	ms
99.9th percentile service time	country_agg_uncached	378.302	ms
100th percentile service time	country_agg_uncached	422.53	ms
error rate	country_agg_uncached	0	%
Min Throughput	country_agg_cached	98.37	ops/s
Median Throughput	country_agg_cached	100.06	ops/s
Max Throughput	country_agg_cached	100.13	ops/s
50th percentile latency	country_agg_cached	3.2638	ms
90th percentile latency	country_agg_cached	4.69259	ms
99th percentile latency	country_agg_cached	189.143	ms
99.9th percentile latency	country_agg_cached	249.851	ms
100th percentile latency	country_agg_cached	256.028	ms
50th percentile service time	country_agg_cached	3.18679	ms
90th percentile service time	country_agg_cached	3.42086	ms
99th percentile service time	country_agg_cached	20.4171	ms

Metric	Operation	Value	Unit
99.9th percentile service time	country_agg_cached	117.273	ms
100th percentile service time	country_agg_cached	255.951	ms
error rate	country_agg_cached	0	%
Min Throughput	scroll	59.16	ops/s
Median Throughput	scroll	60.44	ops/s
Max Throughput	scroll	61.02	ops/s
50th percentile latency	scroll	168347	ms
90th percentile latency	scroll	240658	ms
99th percentile latency	scroll	257048	ms
100th percentile latency	scroll	258853	ms
50th percentile service time	scroll	402.962	ms
90th percentile service time	scroll	431.267	ms
99th percentile service time	scroll	455.632	ms
100th percentile service time	scroll	601.214	ms
error rate	scroll	0	%
Min Throughput	expression	2	ops/s
Median Throughput	expression	2	ops/s
Max Throughput	expression	2	ops/s
50th percentile latency	expression	409.417	ms
90th percentile latency	expression	434.858	ms

Metric	Operation	Value	Unit
99th percentile latency	expression	501.498	ms
100th percentile latency	expression	517.438	ms
50th percentile service time	expression	409.165	ms
90th percentile service time	expression	434.749	ms
99th percentile service time	expression	498.681	ms
100th percentile service time	expression	517.332	ms
error rate	expression	0	%
Min Throughput	painless_static	1.96	ops/s
Median Throughput	painless_static	1.97	ops/s
Max Throughput	painless_static	1.97	ops/s
50th percentile latency	painless_static	3163.94	ms
90th percentile latency	painless_static	3679.27	ms
99th percentile latency	painless_static	3994.52	ms
100th percentile latency	painless_static	4006.5	ms
50th percentile service time	painless_static	503.588	ms
90th percentile service time	painless_static	528.807	ms
99th percentile service time	painless_static	600.103	ms
100th percentile service time	painless_static	623.666	ms
error rate	painless_static	0	%
Min Throughput	painless_dynamic	2	ops/s

Metric	Operation	Value	Unit
Median Throughput	painless_dynamic	2	ops/s
Max Throughput	painless_dynamic	2	ops/s
50th percentile latency	painless_dynamic	611.305	ms
90th percentile latency	painless_dynamic	786.806	ms
99th percentile latency	painless_dynamic	973.432	ms
100th percentile latency	painless_dynamic	982.484	ms
50th percentile service time	painless_dynamic	494.097	ms
90th percentile service time	painless_dynamic	518.082	ms
99th percentile service time	painless_dynamic	606.748	ms
100th percentile service time	painless_dynamic	638.903	ms
error rate	painless_dynamic	0	%
Min Throughput	large_filtered_terms	1.39	ops/s
Median Throughput	large_filtered_terms	1.4	ops/s
Max Throughput	large_filtered_terms	1.4	ops/s
50th percentile latency	large_filtered_terms	65601.1	ms
90th percentile latency	large_filtered_terms	82494.7	ms
99th percentile latency	large_filtered_terms	86452.2	ms
100th percentile latency	large_filtered_terms	86857.3	ms
50th percentile service time	large_filtered_terms	707.17	ms

Metric	Operation	Value	Unit
90th percentile service time	large_filtered_terms	747.949	ms
99th percentile service time	large_filtered_terms	847.069	ms
100th percentile service time	large_filtered_terms	927.917	ms
error rate	large_filtered_terms	0	%
Min Throughput	large_prohibited_terms	1.46	ops/s
Median Throughput	large_prohibited_terms	1.46	ops/s
Max Throughput	large_prohibited_terms	1.46	ops/s
50th percentile latency	large_prohibited_terms	55916.3	ms
90th percentile latency	large_prohibited_terms	70529.7	ms
99th percentile latency	large_prohibited_terms	73769.1	ms
100th percentile latency	large_prohibited_terms	74143.9	ms
50th percentile service time	large_prohibited_terms	679.394	ms
90th percentile service time	large_prohibited_terms	717.476	ms
99th percentile service time	large_prohibited_terms	782.085	ms
100th percentile service time	large_prohibited_terms	822.723	ms
error rate	large_prohibited_terms	0	%

Benchmarking result

Cluster 1 (baseline) against Cluster 2 (contender)

Metric	Operation	Baseline	Contender	Diff	Unit
Indexing time		26.3543	23.9479	-2.40645	min
Merge time		11.0297	14.3001	3.27042	min
Refresh time		3.05238	5.26405	2.21167	min
Flush time		0.04485	0.0308333	-0.01402	min
Merge throttle time		1.39282	1.27945	-0.11337	min
Total Young Gen GC		92.902	183.74	90.838	s
Total Old Gen GC		0.4	1.125	0.725	s
Heap used for segments		18.7955	18.8167	0.02126	MB
Heap used for doc values		0.360752	0.452751	0.092	MB
Heap used for terms		17.2739	17.2004	-0.07343	MB
Heap used for norms		0.0877075	0.0852051	-0.0025	MB
Heap used for points		0.241213	0.241465	0.00025	MB
Heap used for stored fields		0.831932	0.836876	0.00494	MB
Segment count		133	140	7	
Min Throughput	index-append	51751.7	28115.4	-23636.2	docs/s
Median Throughput	index-append	52303	28645.6	-23657.5	docs/s
Max Throughput	index-append	54076.3	30037.8	-24038.5	docs/s

Metric	Operation	Baseline	Contender	Diff	Unit
50th percentile latency	index-append	743.939	1447.76	703.818	ms
90th percentile latency	index-append	1045.7	1847.05	801.342	ms
99th percentile latency	index-append	1325.21	2264.68	939.47	ms
100th percentile latency	index-append	1794.39	2608.68	814.293	ms
50th percentile service time	index-append	743.939	1447.76	703.818	ms
90th percentile service time	index-append	1045.7	1847.05	801.342	ms
99th percentile service time	index-append	1325.21	2264.68	939.47	ms
100th percentile service time	index-append	1794.39	2608.68	814.293	ms
error rate	index-append	0	0	0	%
Min Throughput	force-merge	0.950072	2.10087	1.1508	ops/s
Median Throughput	force-merge	0.950072	2.10087	1.1508	ops/s
Max Throughput	force-merge	0.950072	2.10087	1.1508	ops/s
100th percentile latency	force-merge	1052.54	475.984	-576.556	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	force-merge	1052.54	475.984	-576.556	ms
error rate	force-merge	0	0	0	%
Min Throughput	index-stats	100.037	97.7524	-2.28456	ops/s
Median Throughput	index-stats	100.049	100.048	-0.00112	ops/s
Max Throughput	index-stats	100.085	100.068	-0.01745	ops/s
50th percentile latency	index-stats	4.85232	5.09015	0.23784	ms
90th percentile latency	index-stats	5.14185	10.7365	5.59466	ms
99th percentile latency	index-stats	77.3127	234.761	157.448	ms
99.9th percentile latency	index-stats	123.888	277.393	153.505	ms
100th percentile latency	index-stats	128.01	281.866	153.856	ms
50th percentile service time	index-stats	4.78006	5.01096	0.23091	ms
90th percentile service time	index-stats	4.9831	5.30021	0.31711	ms
99th percentile service time	index-stats	9.66475	12.0005	2.33576	ms
99.9th percentile service time	index-stats	48.4445	141.631	93.186	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	index-stats	127.945	150.153	22.2078	ms
error rate	index-stats	0	0	0	%
Min Throughput	node-stats	100.054	100.007	-0.04689	ops/s
Median Throughput	node-stats	100.098	100.085	-0.01341	ops/s
Max Throughput	node-stats	100.551	100.494	-0.0566	ops/s
50th percentile latency	node-stats	4.55259	4.90659	0.354	ms
90th percentile latency	node-stats	4.78784	5.29285	0.50501	ms
99th percentile latency	node-stats	18.8034	29.3245	10.5211	ms
99.9th percentile latency	node-stats	43.7684	43.3885	-0.3799	ms
100th percentile latency	node-stats	48.1474	44.6019	-3.54548	ms
50th percentile service time	node-stats	4.48138	4.83552	0.35414	ms
90th percentile service time	node-stats	4.69386	5.12694	0.43308	ms
99th percentile service time	node-stats	5.64618	9.08739	3.44121	ms
99.9th percentile service time	node-stats	27.8155	39.744	11.9285	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	node-stats	43.6905	44.5383	0.84783	ms
error rate	node-stats	0	0	0	%
Min Throughput	default	49.8129	47.8334	-1.97948	ops/s
Median Throughput	default	50.0009	48.281	-1.71982	ops/s
Max Throughput	default	50.0045	48.7269	-1.2776	ops/s
50th percentile latency	default	19.7245	617.465	597.74	ms
90th percentile latency	default	94.1457	1033.98	939.834	ms
99th percentile latency	default	133.091	1083.23	950.137	ms
99.9th percentile latency	default	137.285	1095.4	958.114	ms
100th percentile latency	default	138.043	1097.14	959.1	ms
50th percentile service time	default	19.1469	18.646	-0.50082	ms
90th percentile service time	default	19.9554	24.9381	4.98271	ms
99th percentile service time	default	25.3462	35.7667	10.4206	ms
99.9th percentile service time	default	54.7931	57.3679	2.57481	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	default	133.771	151.505	17.7337	ms
error rate	default	0	0	0	%
Min Throughput	term	200.055	199.431	-0.62401	ops/s
Median Throughput	term	200.076	200.072	-0.00349	ops/s
Max Throughput	term	200.119	200.13	0.01076	ops/s
50th percentile latency	term	3.07948	2.9728	-0.10668	ms
90th percentile latency	term	3.37296	7.10648	3.73353	ms
99th percentile latency	term	22.3272	22.4487	0.12153	ms
99.9th percentile latency	term	26.9648	29.0737	2.10889	ms
100th percentile latency	term	28.1562	29.6253	1.46915	ms
50th percentile service time	term	3.00599	2.87833	-0.12766	ms
90th percentile service time	term	3.15279	3.08983	-0.06296	ms
99th percentile service time	term	4.22302	19.9777	15.7546	ms
99.9th percentile service time	term	26.9017	29.0082	2.10648	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	term	28.0823	29.5597	1.4774	ms
error rate	term	0	0	0	%
Min Throughput	phrase	199.842	199.711	-0.13145	ops/s
Median Throughput	phrase	200.04	200.038	-0.00174	ops/s
Max Throughput	phrase	200.087	200.074	-0.0125	ops/s
50th percentile latency	phrase	3.76927	3.61484	-0.15442	ms
90th percentile latency	phrase	13.6055	16.5523	2.94681	ms
99th percentile latency	phrase	28.0245	31.394	3.36944	ms
99.9th percentile latency	phrase	34.7198	33.902	-0.81778	ms
100th percentile latency	phrase	35.551	34.5784	-0.97253	ms
50th percentile service time	phrase	3.67227	3.47402	-0.19825	ms
90th percentile service time	phrase	4.08037	3.90958	-0.17079	ms
99th percentile service time	phrase	16.9256	19.3773	2.45168	ms
99.9th percentile service time	phrase	24.4886	22.7947	-1.69386	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	phrase	29.8604	27.8164	-2.04399	ms
error rate	phrase	0	0	0	%
Min Throughput	country_agg_uncached	4.95005	4.6328	-0.31724	ops/s
Median Throughput	country_agg_uncached	4.99422	4.65258	-0.34163	ops/s
Max Throughput	country_agg_uncached	5.00022	4.67361	-0.32661	ops/s
50th percentile latency	country_agg_uncached	330.923	14864.3	14533.3	ms
90th percentile latency	country_agg_uncached	2780.17	21046	18265.8	ms
99th percentile latency	country_agg_uncached	2866	22902	20036	ms
99.9th percentile latency	country_agg_uncached	2880.39	22997.6	20117.2	ms
100th percentile latency	country_agg_uncached	2882.11	23018.7	20136.6	ms
50th percentile service time	country_agg_uncached	197.883	204.174	6.29064	ms
90th percentile service time	country_agg_uncached	213.402	242.492	29.0907	ms
99th percentile service time	country_agg_uncached	256.649	345.382	88.7335	ms
99.9th percentile service time	country_agg_uncached	290.496	378.302	87.8056	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	country_agg_uncached	296.875	422.53	125.655	ms
error rate	country_agg_uncached	0	0	0	%
Min Throughput	country_agg_cached	99.9249	98.3659	-1.55896	ops/s
Median Throughput	country_agg_cached	100.064	100.056	-0.00795	ops/s
Max Throughput	country_agg_cached	100.112	100.135	0.02245	ops/s
50th percentile latency	country_agg_cached	3.30479	3.2638	-0.04099	ms
90th percentile latency	country_agg_cached	3.52514	4.69259	1.16745	ms
99th percentile latency	country_agg_cached	52.8258	189.143	136.317	ms
99.9th percentile latency	country_agg_cached	112.895	249.851	136.956	ms
100th percentile latency	country_agg_cached	119.435	256.028	136.593	ms
50th percentile service time	country_agg_cached	3.23149	3.18679	-0.0447	ms
90th percentile service time	country_agg_cached	3.41319	3.42086	0.00767	ms
99th percentile service time	country_agg_cached	7.60955	20.4171	12.8075	ms

Metric	Operation	Baseline	Contender	Diff	Unit
99.9th percentile service time	country_agg_cached	26.2229	117.273	91.0502	ms
100th percentile service time	country_agg_cached	119.365	255.951	136.586	ms
error rate	country_agg_cached	0	0	0	%
Min Throughput	scroll	61.5897	59.1628	-2.42689	ops/s
Median Throughput	scroll	61.6735	60.4406	-1.23292	ops/s
Max Throughput	scroll	61.9387	61.019	-0.91967	ops/s
50th percentile latency	scroll	164549	168347	3798.13	ms
90th percentile latency	scroll	237443	240658	3214.79	ms
99th percentile latency	scroll	253860	257048	3187.91	ms
100th percentile latency	scroll	255710	258853	3143.03	ms
50th percentile service time	scroll	399.964	402.962	2.99858	ms
90th percentile service time	scroll	424.303	431.267	6.96397	ms
99th percentile service time	scroll	523.877	455.632	-68.2449	ms

Metric	Operation	Baseline	Contender	Diff	Unit
100th percentile service time	scroll	639.45	601.214	-38.236	ms
error rate	scroll	0	0	0	%
Min Throughput	expression	1.9994	1.9998	0.0004	ops/s
Median Throughput	expression	2.00113	2.00113	0	ops/s
Max Throughput	expression	2.00186	2.00189	2e-05	ops/s
50th percentile latency	expression	409.927	409.417	-0.5091	ms
90th percentile latency	expression	434.544	434.858	0.31406	ms
99th percentile latency	expression	532.412	501.498	-30.914	ms
100th percentile latency	expression	537.618	517.438	-20.1798	ms
50th percentile service time	expression	409.812	409.165	-0.64674	ms
90th percentile service time	expression	428.156	434.749	6.59297	ms
99th percentile service time	expression	532.33	498.681	-33.6493	ms
100th percentile service time	expression	537.495	517.332	-20.1637	ms
error rate	expression	0	0	0	%
Min Throughput	painless_static	1.99752	1.96306	-0.03446	ops/s

Metric	Operation	Baseline	Contender	Diff	Unit
Median Throughput	painless_static	1.99998	1.96607	-0.03391	ops/s
Max Throughput	painless_static	2.00041	1.96914	-0.03127	ops/s
50th percentile latency	painless_static	497.626	3163.94	2666.31	ms
90th percentile latency	painless_static	643.32	3679.27	3035.95	ms
99th percentile latency	painless_static	700.559	3994.52	3293.97	ms
100th percentile latency	painless_static	704.679	4006.5	3301.82	ms
50th percentile service time	painless_static	490.705	503.588	12.8834	ms
90th percentile service time	painless_static	500.663	528.807	28.1439	ms
99th percentile service time	painless_static	642.124	600.103	-42.021	ms
100th percentile service time	painless_static	683.621	623.666	-59.9546	ms
error rate	painless_static	0	0	0	%
Min Throughput	painless_dynamic	1.99721	1.99513	-0.00209	ops/s
Median Throughput	painless_dynamic	2.00032	1.99838	-0.00194	ops/s
Max Throughput	painless_dynamic	2.00089	2.00053	-0.00036	ops/s

Metric	Operation	Baseline	Contender	Diff	Unit
50th percentile latency	painless_dynamic	473.087	611.305	138.218	ms
90th percentile latency	painless_dynamic	554.729	786.806	232.077	ms
99th percentile latency	painless_dynamic	668.363	973.432	305.069	ms
100th percentile latency	painless_dynamic	706.557	982.484	275.926	ms
50th percentile service time	painless_dynamic	469.145	494.097	24.9528	ms
90th percentile service time	painless_dynamic	501.774	518.082	16.3086	ms
99th percentile service time	painless_dynamic	606.61	606.748	0.13817	ms
100th percentile service time	painless_dynamic	624.751	638.903	14.1524	ms
error rate	painless_dynamic	0	0	0	%
Min Throughput	large_filtered_terms	1.64076	1.38866	-0.2521	ops/s
Median Throughput	large_filtered_terms	1.6443	1.39554	-0.24876	ops/s
Max Throughput	large_filtered_terms	1.65048	1.39764	-0.25283	ops/s
50th percentile latency	large_filtered_terms	33013.5	65601.1	32587.5	ms

Metric	Operation	Baseline	Contender	Diff	Unit
90th percentile latency	large_filtered_terms	40869	82494.7	41625.7	ms
99th percentile latency	large_filtered_terms	42644	86452.2	43808.2	ms
100th percentile latency	large_filtered_terms	42936.2	86857.3	43921.1	ms
50th percentile service time	large_filtered_terms	598.001	707.17	109.169	ms
90th percentile service time	large_filtered_terms	626.81	747.949	121.139	ms
99th percentile service time	large_filtered_terms	771.815	847.069	75.2534	ms
100th percentile service time	large_filtered_terms	796.884	927.917	131.032	ms
error rate	large_filtered_terms	0	0	0	%
Min Throughput	large_prohibited_terms	1.6893	1.45607	-0.23323	ops/s
Median Throughput	large_prohibited_terms	1.69452	1.46074	-0.23379	ops/s
Max Throughput	large_prohibited_terms	1.69856	1.46248	-0.23608	ops/s
50th percentile latency	large_prohibited_terms	27732.3	55916.3	28184	ms
90th percentile latency	large_prohibited_terms	34305.5	70529.7	36224.2	ms

Metric	Operation	Baseline	Contender	Diff	Unit
99th percentile latency	large_prohibited_terms	35840.4	73769.1	37928.7	ms
100th percentile latency	large_prohibited_terms	35993.5	74143.9	38150.4	ms
50th percentile service time	large_prohibited_terms	586.382	679.394	93.0121	ms
90th percentile service time	large_prohibited_terms	618.185	717.476	99.2908	ms
99th percentile service time	large_prohibited_terms	661.378	782.085	120.707	ms
100th percentile service time	large_prohibited_terms	823.782	822.723	-1.05804	ms
error rate	large_prohibited_terms	0	0	0	%

5 Overdue payments

Pay-As-You-Go

1. A notification will be sent to you after 8 days, 12 days, and 14 days of overdue payments.
2. Your service will be stopped after 15 days of overdue payments.
3. Your Elasticsearch instance will be released 15 days after your service is stopped . The data on the released instance will be permanently deleted and cannot be restored.
4. A notification will be sent to you 9 days before your Elasticsearch instance is released.

Subscription

1. A notification will be sent to you 7 days, 3 days, and 1 day before your subscription expires.
2. Your service will be stopped 15 days after your subscription has expired.
3. Your Elasticsearch instance will be released 15 days after your service is stopped . The data on the released instance will be permanently deleted and cannot be restored.
4. A notification will be sent to you 7 days, 3 days, and 1 day before your Elasticsearch instance is released.

6 High reliability

This topic introduces the high reliability of Alibaba Cloud Elasticsearch based on auto-creation, restoration, and storage of snapshots and load balancing.

Auto snapshot

Alibaba Cloud Elasticsearch instances support the auto snapshot feature. You can enable auto snapshot on the Snapshots page in the Alibaba Cloud Elasticsearch console and then set the snapshot creation cycle. Elasticsearch then creates a snapshot daily at the scheduled time. This feature allows you to back up your data for disaster recovery. For more information, see Snapshots.

```
Restore   snapshots
```

On the Snapshots page of the Alibaba Cloud Elasticsearch console, you can use a specified snapshot to restore data. For more information, see Auto snapshot guide.

Note:

- Alibaba Cloud Elasticsearch only stores snapshots that are created within the last three days.
- A snapshot created by the auto snapshot feature can only be restored to the Alibaba Cloud Elasticsearch instance where the snapshot is created.

Store snapshots on OSS

Alibaba Cloud Elasticsearch allows you to store the snapshots of your Elasticsearch instance on Alibaba Cloud Object Storage Service (OSS). To store snapshots on OSS, you must first purchase the OSS service in the same region as your Elasticsearch instance. You can call the snapshot creation operation to create a snapshot of the specified index data. For more information, see Snapshots and data restoration.

```
Restore
```

Alibaba Cloud Elasticsearch allows you to call the restore operation to restore index data from a specified snapshot. This feature enables support for disaster recovery. For more information, see Snapshots and data restoration.

Note:

- We recommend that you use OSS standard buckets to store snapshots. OSS Archive buckets are not supported.
- A snapshot stored on OSS can be restored to an Alibaba Cloud Elasticsearch instance in the same region as OSS.
- You can call the corresponding operation to create a snapshot or restore the index data in a specified snapshot.
- By default, each Alibaba Cloud Elasticsearch data node can process 40 MB of data per second. You can reference the Snapshot And Restore page on the official Elasticsearch site and set the `max_restor e_bytes_pe r_sec` parameter to tune the data processing capability of the data nodes.

Load balancing

Alibaba Cloud Elasticsearch instances support load balancing. You can specify the public or internal network address of your Elasticsearch instance on your client to access the Elasticsearch instance. Your requests are evenly distributed to all data nodes of the Elasticsearch instance based on load balancing.

 Notice:

The load balancing among these data nodes depends on the number and size of index shards. We recommend that you reference Calculate shard size and then determine the number and size of the index shards properly when you create indexes.

7 High security

Access an Alibaba Cloud Elasticsearch instance through its internal IP address

You can access an Alibaba Cloud Elasticsearch instance through its internal IP address from a VPC network. If you want to enhance the security of your access, create an Alibaba Cloud Elastic Compute Service (ECS) instance in the same region and VPC network where the Alibaba Cloud Elasticsearch instance is created. You can then deploy applications on this ECS instance and use the ECS instance to access the internal IP address of the Alibaba Cloud Elasticsearch instance.

Note:

A VPC network is isolated from the public network and provides a more secure access environment.

Access control

Configure a whitelist

You can configure a whitelist or blacklist to limit the access to the internal IP address of an Alibaba Cloud Elasticsearch instance. Only whitelisted or non-blacklisted IP addresses can access the Alibaba Cloud Elasticsearch instance. For more information, see Elasticsearch cluster configuration.

You can configure a whitelist to limit the access to the public IP address of an Alibaba Cloud Elasticsearch instance. Only whitelisted IP addresses can access the Alibaba Cloud Elasticsearch instance. For more information, see Security configuration.

RAM-based access control

Alibaba Cloud Elasticsearch allows you to create RAM users to manage access permissions. Resources of different RAM users are isolated from each other. RAM users can only manage and view Alibaba Cloud Elasticsearch instances created under their own accounts. For more information, see Access authentication rules.

X-Pack role-based access control

Alibaba Cloud Elasticsearch allows you to use X-Pack. X-Pack is an Elastic Stack extension that bundles security, alerting, monitoring, reporting, and graph capabilities into one easy-to-install package. X-Pack can be installed in Kibana and provides a wide range of features, such as authentication, permission control, real-

time monitoring, visualized reports, and machine learning. X-Pack role-based access control allows you to authenticate access requests to indexes. For more information, see Security APIs.

System security

- You can access an Alibaba Cloud Elasticsearch instance through a VPC network. A VPC network provides a more secure access environment.
- You cannot log on to any servers of the nodes that are contained in an Elasticsearch instance.
- No IP addresses are allowed to access the public IP address of an Elasticsearch instance by default. To allow access requests to the public IP address, you must configure a public IP whitelist. For more information, see Public network whitelist.
- You can create a whitelist to limit the access to the public and internal IP addresses of an Alibaba Cloud Elasticsearch instance.
- An Alibaba Cloud Elasticsearch instance opens ports 9200 and 9300 only, and enables you to access its public and internal IP addresses.

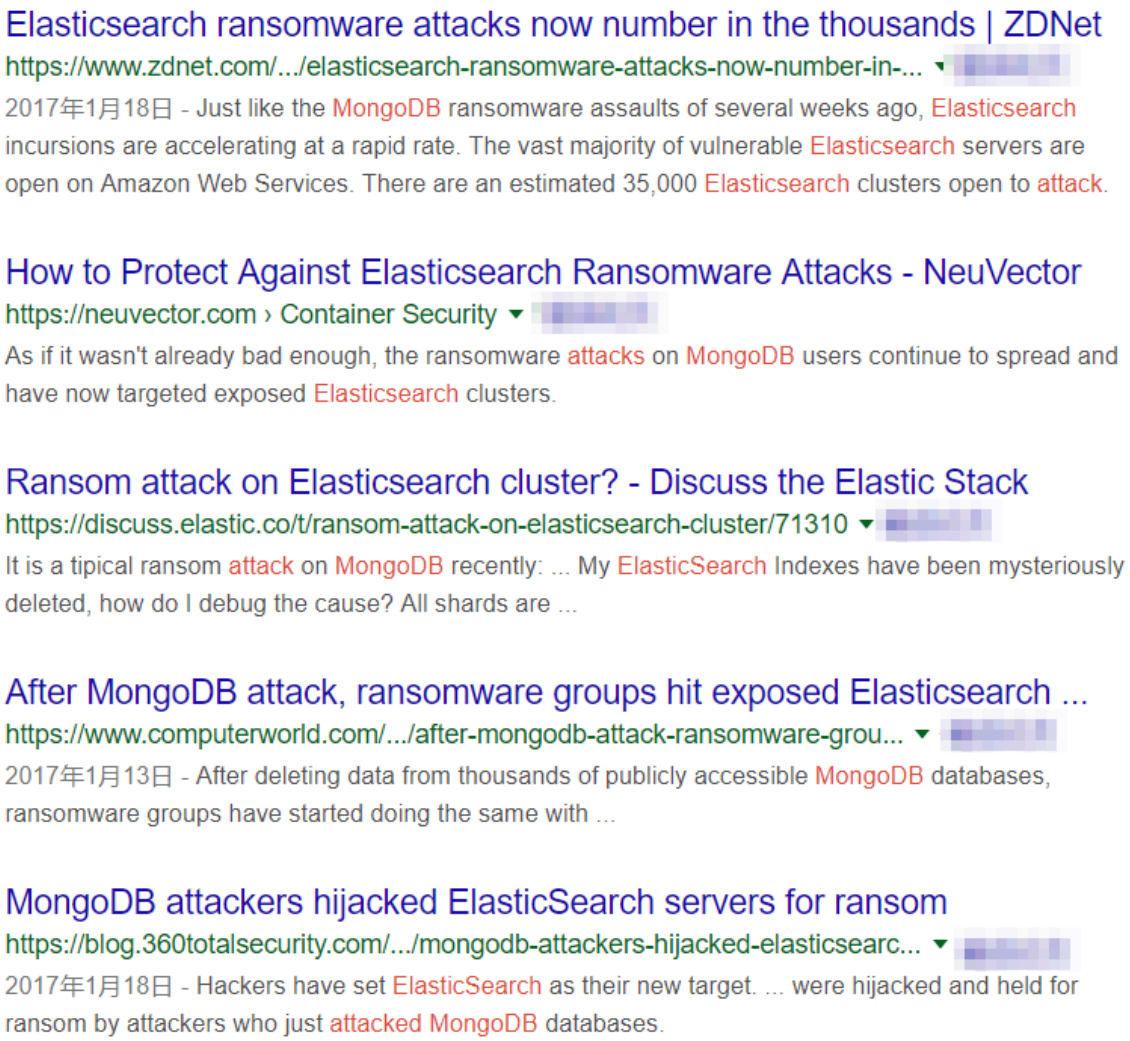
 Note:
Port `9300` is closed for Alibaba Cloud Elasticsearch 6.3.2 that has X-Pack installed.

8 Security features

This topic compares Alibaba Cloud Elasticsearch instances with user-built Elasticsearch instances to describe the advantages of Alibaba Cloud Elasticsearch in security protection.

Background

Open-source software has typically been the first choice of attackers, such as the MongoDB ransomware attacks event. Elasticsearch has also become the target of the attackers. They may attack user-created Elasticsearch services that do not have professional security protection, and then delete important data or intrude into the business system.

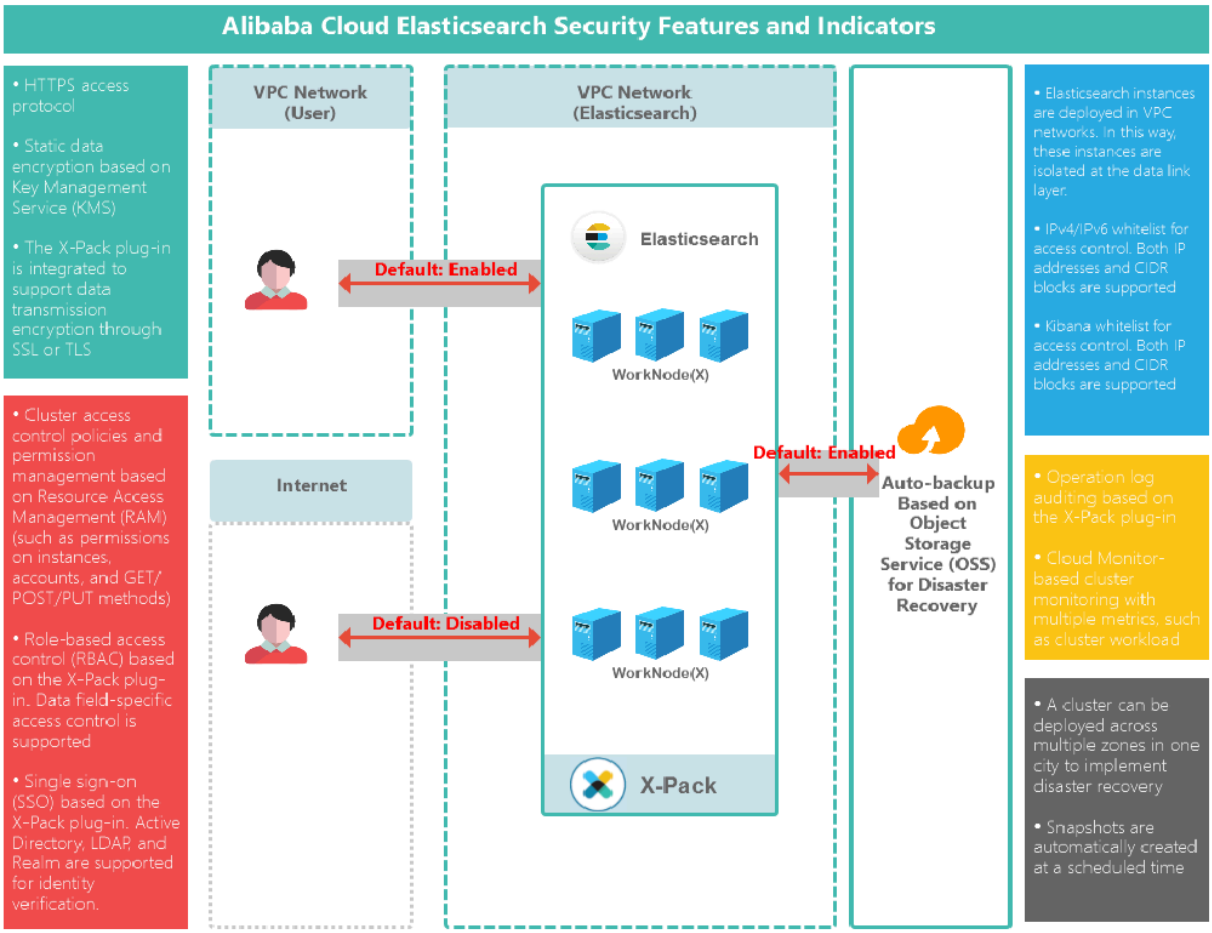


Alibaba Cloud Security Center has released a note about security risk warning on Elasticsearch and provided multiple security hardening strategies and solutions.

Compared with the security protection for user-built Elasticsearch instances, the solutions provided by Alibaba Cloud Elasticsearch for data and service security are more reliable and professional.

Security feature descriptions

Alibaba Cloud has released the fully-hosted Elasticsearch service in November 2017. Alibaba Cloud Elasticsearch provides security protection features for you to safeguard your services.



The following table compares Alibaba Cloud Elasticsearch with user-built Elasticsearch services in security protection:

Security indicator	Security protection for user-built Elasticsearch service	Integrated security features for Alibaba Cloud Elasticsearch
Access control	· Purchase cloud security products, such as security groups or firewalls, to control and quarantine source IP addresses. · Disable port 9200 unless it is necessary. · Bind source IP addresses. · Change the default port.	· Alibaba Cloud Elasticsearch instances are deployed in VPC networks. In this way, they can be isolated at the data link layer. · IPv4 and IPv6 whitelists for access control. Both IP addresses and CIDR blocks are supported. · Kibana whitelist for access control. Both IP addresses and CIDR blocks are supported.
Authentication and authorization	Install third-party security plug-ins, such as Searchguard and Shield.	· Cluster access control policies based on Resource Access Management (RAM), such as the ReadOnlyAccess policy for granting the read-only permission and the FullAccess policy for granting the administrator permission. · Permission control based on RAM, such as permissions on instances, accounts, and GET, POST, and PUT methods. · Role-based access control (RBAC) based on the X-Pack plug-in. Data field-specific access control is supported. · Single sign-on (SSO) based on the X-Pack plug-in. Active Directory, LDAP, and Elasticsearch native Realm are supported for identity verification.
Data encryption	· Use storage media that support static data encryption. · Disable HTTP in YML configuration.	· HTTPS is supported. · Static data encryption based on Key Management Service (KMS). · The X-Pack plug-in is integrated to support data transmission encryption through SSL or TLS.

Security indicator	Security protection for user-built Elasticsearch service	Integrated security features for Alibaba Cloud Elasticsearch
Monitoring and auditing	Use third-party tools to audit logs and monitor services.	· Operation log auditing based on the X-Pack plug-in. · CloudMonitor-based cluster monitoring with multiple metrics, such as cluster workload.
Disaster recovery	· Purchase file systems to back up data periodically. · Use multiple clusters to implement disaster recovery.	· A cluster can be deployed across multiple zones in one city to implement disaster recovery. · Snapshots are automatically created at a scheduled time.