Blazingly Fast Elasticsearch

不完全指南

Bin Wu (吴斌) - Elastic Community Member

JVM Heap

- 设置 Xmx、Xms 为同样定值且不超过可用内存50%
  - 堆外缓存
    - 网络
    - 文件
  - 26GB - 30GB
- Compressed Oops
  - https://docs.oracle.com/javase/8/docs/technotes/guides/vm/performance-enhancements-7.html
  - XX:+UseCompressedOops 32GB
  - Zero-Based Compressed Ordinary Object Pointers
    - For Java heap sizes up around 26 gigabytes, any of Solaris, Linux, and Windows operating systems will typically be able to allocate the Java heap at virtual address zero.
- https://gist.github.com/bindiego/3a0e73aa2e7ec17188f1c9c4cc8b7198

JVM - Compressed Oops

https://gist.github.com/bindiego/3a0e73aa2e7ec17188f1c9c4cc8b7198

Check JVM compressed oops option

even_better.md — Raw

try to stay below the threshold for zero-based compressed oops; the exact cutoff varies but 26 GB is safe on most systems, but can be as large as 30 GB on some systems. You can verify that you are under the limit by starting Elasticsearch with the JVM options -XX:+UnlockDiagnosticVMOptions -XX:+PrintCompressedOopsMode and looking for a line like the following

```
heap address: 0x000000011be00000, size: 27648 MB, zero based Compressed Oops
```

showing that zero-based compressed oops are enabled instead of

```
heap address: 0x0000000118400000, size: 28672 MB, Compressed Oops with base: 0x00000001183ff000
```

jvm_compressed_oops.sh — Raw

```
1  #!/bin/bash -ex
2
3  java -Xmx32766m -XX:+PrintFlagsFinal 2> /dev/null | grep UseCompressedOops
4
5  java -Xmx32767m -XX:+PrintFlagsFinal 2> /dev/null | grep UseCompressedOops
```


elastic

JVM Heap - Zero-Based Compressed Oops

- http://hg.openjdk.java.net/jdk/jdk11/file/f729ca27cf9a/src/hotspot/cpu/x86/macroAssembler_x86.cpp

```
6217 // Algorithm must match oop.inline.hpp encode_heap_oop.
6218 void MacroAssembler::encode_heap_oop(Register r) {
6219 #ifdef ASSERT
6220   verify_heapbase("MacroAssembler::encode_heap_oop: heap base corrupted?");
6221 #endif
6222   verify_oop(r, "broken oop in encode_heap_oop");
6223   if (Universe::narrow_oop_base() == NULL) {
6224     if (Universe::narrow_oop_shift() != 0) {
6225       assert (LogMinObjAlignmentInBytes == Universe::narrow_oop_shift(), "decode alg wrong");
6226       shrq(r, LogMinObjAlignmentInBytes);
6227     }
6228     return;
6229   }
6230   testq(r, r);
6231   cmovq(Assembler::equal, r, r12_heapbase);
6232   subq(r, r12_heapbase);
6233   shrq(r, LogMinObjAlignmentInBytes);
6234 }
6235
```

JVM Heap - Less is More

Machine type

n1-highmem-8 (8 vCPUs, 52 GB memory)

Reservation

Automatically choose

CPU platform

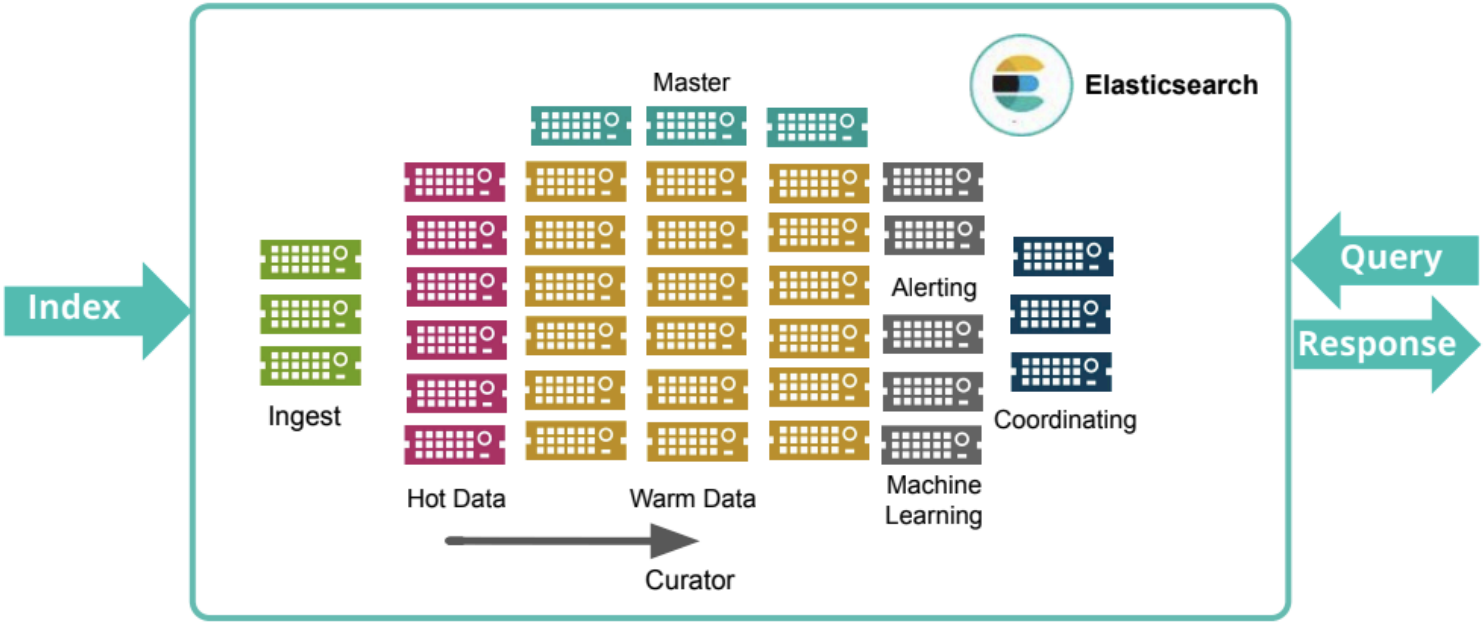
Intel Broadwell

```
➜  JavaMemory java -version
java version "11.0.4" 2019-07-16 LTS
Java(TM) SE Runtime Environment 18.9 (build 11.0.4+10-LTS)
Java HotSpot(TM) 64-Bit Server VM 18.9 (build 11.0.4+10-LTS, mixed mode)
```

```
binwu@tmp1:~/JavaMemory$ ./run.sh
+ m=31
+ rm -rf Memory.class 'Memory$Entity.class'
+ /home/binwu/jdk/bin/javac Memory.java
+ /home/binwu/jdk/bin/java -Xms31g -Xmx31g -Xmn50m Memory
Total Memory (in GB): 31
Free Memory (in GB): 30
Max Memory (in GB): 31
Elements created and added to LinkedList: 538158393
binwu@tmp1:~/JavaMemory$
```

```
binwu@tmp1:~/JavaMemory$ ./run.sh
+ m=32
+ rm -rf Memory.class 'Memory$Entity.class'
+ /home/binwu/jdk/bin/javac Memory.java
+ /home/binwu/jdk/bin/java -Xms32g -Xmx32g -Xmn50m Memory
Total Memory (in GB): 32
Free Memory (in GB): 31
Max Memory (in GB): 32
Elements created and added to LinkedList: 351350824
binwu@tmp1:~/JavaMemory$
```

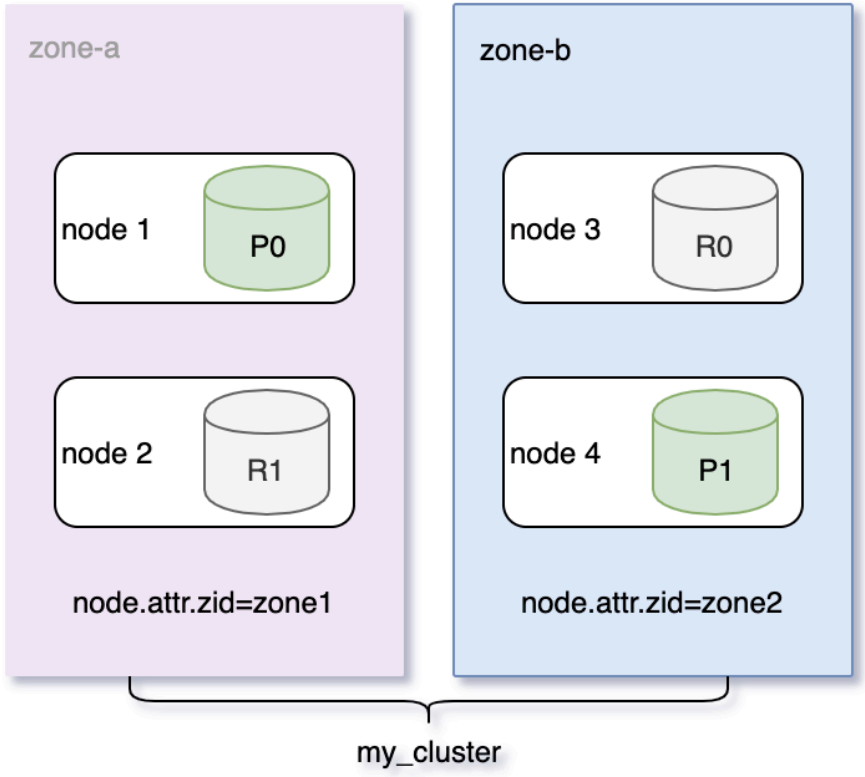
节点架构



elastic

集群节点和数据的 zonal/rack（机架）感知

```
PUT _cluster/settings
{
  “Persistent”: {
    “Cluster”: {
      “Routing”: {
        “Allocation.awareness.attributes: “zid”,
        “Allocation.awareness.force.zid.values: “zone1,zone2”,
      }
    }
  }
}
```



elastic

Indexing

- sudo swapoff -a
  - /etc/fstab 永久关闭
- mmapfs
  - sudo sysctl -w vm.max_map_count=262144
  - /etc/sysctl.conf 永久关闭
- File descriptors
  - ulimit -n 65535
  - /etc/security/limits.conf -> nofile
- Bulk requests in size of 100, 200, 400 etc.
  - Multi-threads till 429 (EsRejectedExecutionException in Java)
- Use auto IDs to skip uniqueness check
- Index.refresh_interval -> 30s or -1
- Index.number_of_replicas -> 0 or 1
- *Indices.memory.index_buffer_size deal with cautious*

查询

- Off-heap memory
  - 检索/搜索
  - 计算/聚合
- query_string / multi_match 字段越多越慢
  - Copy_to
- “Feature Engineering”
  - 预处理，空间换时间
    - E.g. price: $10 => pricer_range: “10-99”
  - Numbers with no meanings is actually a keyword
    - E.g. IDs, ISBN etc. 当作keyword处理
- Rounded dates
  - now-1h -> now-1h/m, now -> now/m
- Force-merge 时序数据 (按时间分的索引且不更新了)

查询 - 数据预热

- Global originals 查询时加载，提升indexing，打开后加速查询。
- 针对keyword 和 text
- 要aggs的field
- Cautious: force-merged index & frozen index

```
PUT my_index/_mapping
{
  "properties": {
    "blahblah": {
      "type": "keyword",
      "eager_global_ordinals": true
    }
  }
}
```

- 文件缓存预热

```
PUT /my_index
{
  "settings": {
    "index.store.preload": ["*"]
  }
}

# preload norms, doc values, terms dictionaries,
  postings lists and points
PUT /my_index
{
  "settings": {
    "index.store.preload": ["nvd", "dvd", "tim",
      "doc", "dim"]
  }
}
```

查询 - 排序健

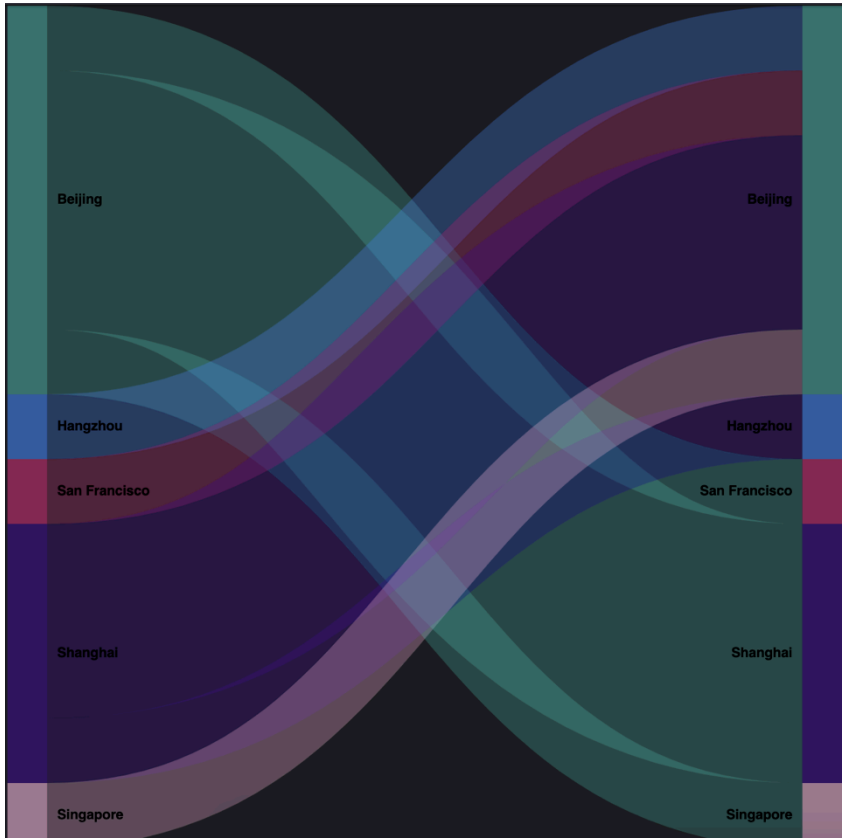
```
PUT events
{
    "settings" : {
        "index" : {
            "sort.field" : "timestamp",
            "sort.order" : "desc"
        }
    },
    "mappings": {
        "properties": {
            "timestamp": {
                "type": "date"
            }
        }
    }
}

GET /events/_search
{
    "size": 10,
    "sort": [
        { "timestamp": "desc" }
    ]
}
```

```
PUT events
{
    "settings" : {
        "index" : {
            "sort.field" : ["username", "date"],
            "sort.order" : ["asc", "desc"]
        }
    },
    "mappings": {
        "properties": {
            "username": {
                "type": "keyword",
                "doc_values": true
            },
            "date": {
                "type": "date"
            }
        }
    }
}
```

查询 - 副本。。Er。。可能有帮助。。可能还。。

- 副本的核心功能
  - 融灾
  - 增加吞吐
- 有时会更慢?
  - 是的 e.g. 同样的或者类似filter的请求到了不同shard
- 副本个数?
  - max(max_failures, ceil(num_nodes / num_primaries) - 1)

查询 - Profiler




```
GET activity/_search
{
  "profile": "true",
  "size": 0,
  "aggs": {
    "table": {
      "composite": {
        "size": 100,
        "sources": [
          {
            "stk1": {
              "terms": {"field": "travelled_from
                .keyword"}
            }
          },
          {
            "stk2": {
              "terms": {"field": "travelled_to
                .keyword"}
            }
          }
        ]
      }
    }
  }
}
```


elastic

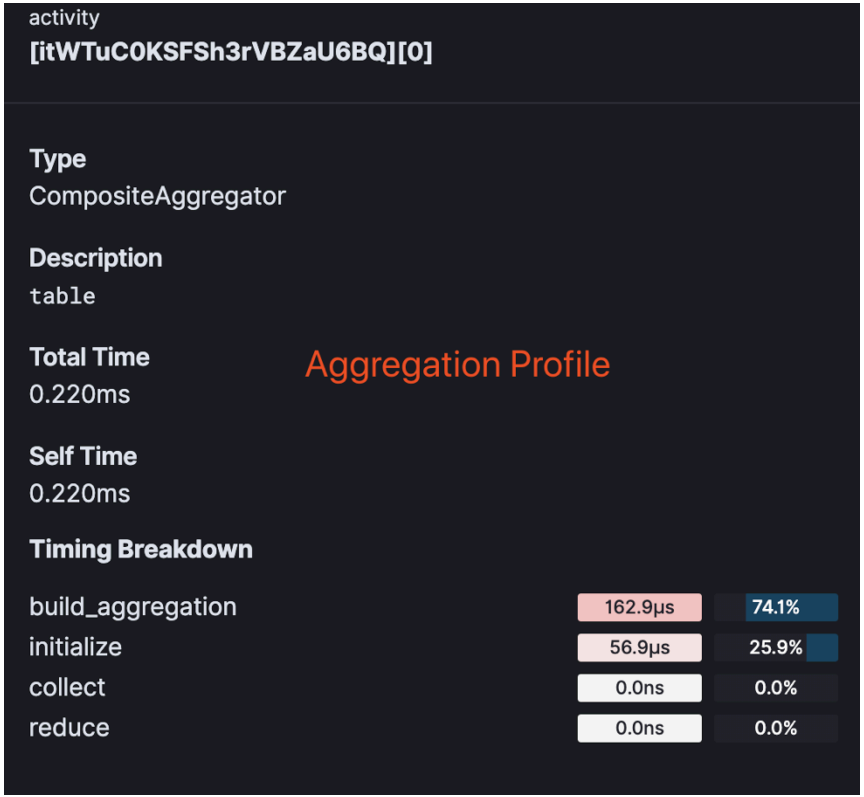
查询 - Profiler

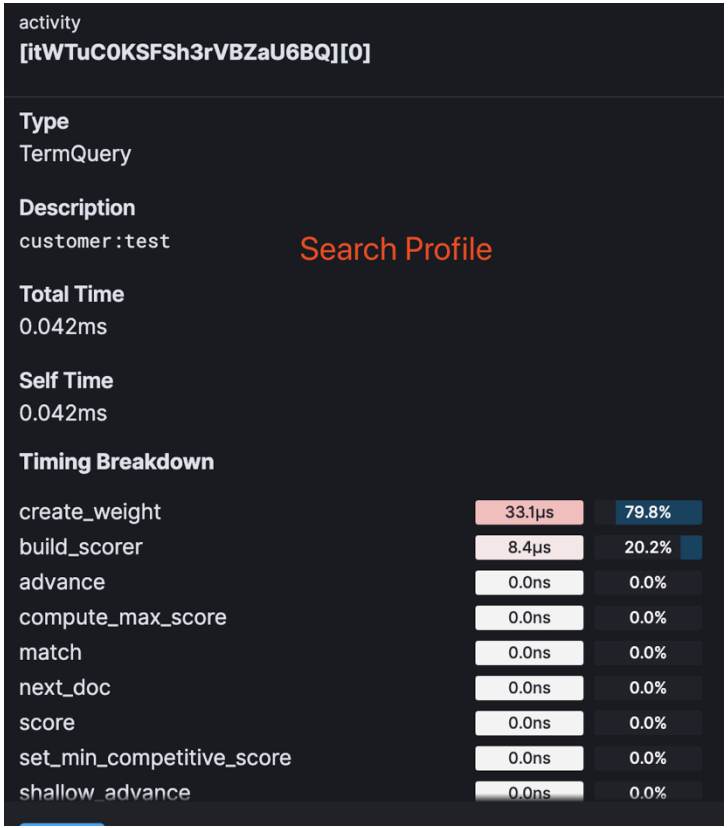
重要返回指标

elastic

查询 - Profiler

https://www.elastic.co/guide/en/elasticsearch/reference/current/search-profile.html





elastic

谢谢

elastic
中文社区

专业、垂直、纯粹的 Elastic 开源技术交流社区
https://elasticsearch.cn/