OpenAlex

ElasticSearch

Configurations:

✅ Elasticsearch security features have been automatically configured!
✅ Authentication is enabled and cluster connections are encrypted.

ℹ️ Password for the **elastic** user (reset with `bin/elasticsearch-reset-password -u elastic`):
nCHNY*2VK*DWndQCxNmZ

ℹ️ HTTP CA certificate SHA-256 fingerprint:
7399a71861f7b356927f3cfe7ced2a6d74a01436b488fa8044d3095adb028e05

ℹ️ Configure Kibana to use this cluster:
• Run Kibana and click the Configuration link in the terminal when Kibana starts.
• Copy the following enrollment token and paste it into Kibana in your browser (valid for the next 30 minutes):

eyJ2ZXIiOiI4LjYuMiIsImFkciI6WyIxNzIuMjIuMjI0LjE1MDo5MjAxIl0sImZnciI6IjczOTlhNzE4NjFmN2IzNTY5MjdmM2NmZTdjZWQyYTZkNzRhMDE0MzZiNDg4ZmE4MDQ0ZDMwOTVhZGIwMjhlMDUiLCJrZXkiOiJnbEdkdm9ZQlBfc3A4VmNqS0hFUzpNdjRWcTVsSFRmaWFJY3RsaTJPWUNRIn0=

ℹ️ Configure other nodes to join this cluster:
• On this node:
  - Create an enrollment token with `bin/elasticsearch-create-enrollment-token -s node`.
  - Uncomment the **transport.host** setting at the end of **config/elasticsearch.yml**.
  - Restart Elasticsearch.
• On other nodes:
  - Start Elasticsearch with `bin/elasticsearch --enrollment-token <token>`, using the enrollment token that you generated.

```
cluster.initial_master_nodes: ["master-node-1", "master-node-2", "master-node-3"]
discovery.seed_hosts: ["172.22.224.151", "172.22.224.152", "172.22.224.153"]
network.host: [_local_, _site_]
cluster.name: openalex_dhyeyhp2
```

```
# Enable security features
xpack.security.enabled: true

xpack.security.enrollment.enabled: true

# Enable encryption for HTTP API client connections, such as Kibana, Logstash, and Agents
xpack.security.http.ssl:
  enabled: true
  keystore.path: certs/http.p12

# Enable encryption and mutual authentication between cluster nodes
xpack.security.transport.ssl:
  enabled: true
  verification_mode: certificate
  keystore.path: certs/transport.p12
  truststore.path: certs/transport.p12
# Create a new cluster with the current node only
# Additional nodes can still join the cluster later
#cluster.initial_master_nodes: ["hawk1"]

# Allow HTTP API connections from anywhere
# Connections are encrypted and require user authentication
http.host: 0.0.0.0

# Allow other nodes to join the cluster from anywhere
# Connections are encrypted and mutually authenticated
#transport.host: 0.0.0.0
```

Local macbook elasticsearch configurations

✅ Elasticsearch security features have been automatically configured!
✅ Authentication is enabled and cluster connections are encrypted.

ℹ️ Password for the **elastic** user (reset with `bin/elasticsearch-reset-password -u elastic`):
E7AaHfVPLZ93b75ZSi=y

ℹ️ HTTP CA certificate SHA-256 fingerprint:
9cd149588d337aed62d90fcf9fca5570beb3b0b4e5730ba37c6ccb280fc52ef9

ℹ️ Configure Kibana to use this cluster:
• Run Kibana and click the configuration link in the terminal when Kibana starts.
• Copy the following enrollment token and paste it into Kibana in your browser (valid for the next 30 minutes):

eyJ2ZXIiOiI4LjYuMiIsImFkciI6WyIxOTIuMTY4LjAuMTk3OjkyMDAiXSwiZmdyIjoiOWNkMTQ5NTg4ZDMzN2FlZDYyZDkwZmNmOWZjYTU1NzBiZWIzYjBiNGU1NzMwYmEzN2M2Y2NiMjgwZmM1MmVmOSIsImtleSI6IlhBSVAwNFlCR1pBajNYSWhndXpuOmh2SnY3aTlpU0llWWY1Vmh1dDQ2OXcifQ==

ℹ️ Configure other nodes to join this cluster:
• On this node:
 - Create an enrollment token with `bin/elasticsearch-create-enrollment-token -s node`.
 - Uncomment the **transport.host** setting at the end of **config/elasticsearch.yml**.
 - Restart Elasticsearch.
• On other nodes:
 - Start Elasticsearch with `bin/elasticsearch --enrollment-token <token>`, using the enrollment token that you generated.

Server Configurations

- 3 master eligible nodes
    - 1 voting only master eligible role, data
    - 2 master eligible nodes with data role
- 3 dedicated Data nodes + 2 master eligible
    - 1 node with master eligible, voting only data node
    - 2 dedicated Data nodes with node role as data

- network.host: 192.168.1.10 # change ip address for each node
- `discovery.seed_hosts:`
- `  - 192.168.1.10:9300`
- `  - 192.168.1.11`
- `  - seeds.mydomain.com`
- `  - [0:0:0:0:0:ffff:c0a8:10c]:9301`
- `cluster.initial_master_nodes:`
- `  - master-node-a`
- `  - master-node-b`
- `  - master-node-c`
-
- After the cluster forms successfully for the first time, remove the `cluster.initial_master_nodes` setting from each node's configuration. Do not use this setting when restarting a cluster or adding a new node to an existing cluster.
- Default index settings:
    - 5 primary shards
    - 1 replica shard
    - 1000 fields

- 1s refresh interval

Progress

Data Info

Entity	No. of Records (from manifest)	Size of Data (compressed) **du -sh ./**
Authors ✅	102 Million	**40** GB
Works ✅	~250 Million	**268** GB
Institutions ✅	108618	56 MB
Sources✅	226726	63 MB
Publishers ✅	3675	548 KB
concepts ✅	65,000	85 MB

compressed Works = 268GB
Uncompressed works ~ 1TB (primary shards)
If replication factor = 2

(Per elastic search cluster) Total space for indexing entire works = **~ 2TB**, with just 1 replica shard/s.

Total needed **4TB.**

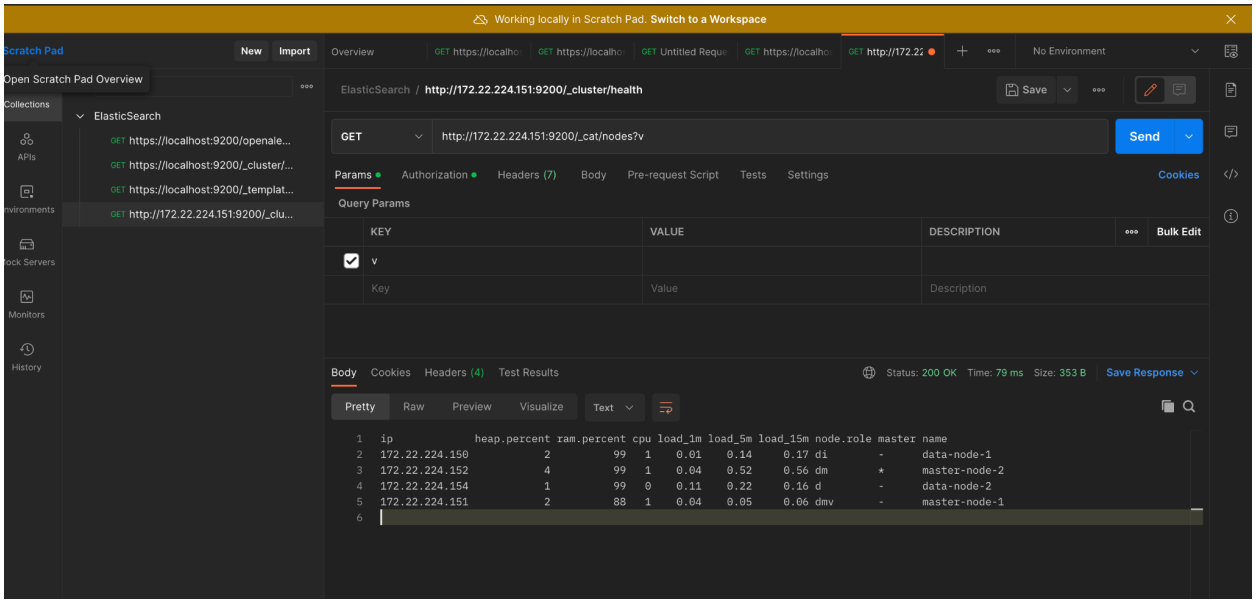
—--- Available Disk Space for combining all the machines—
4.4TB *2 = Extra 4.4 TB on all machines = 4.5.

= 1.2TB disk space on each server

dhyeyhp2@hawk1:/scratch$ df -h ./
Filesystem Size Used **Avail** Use% Mounted on
/dev/mapper/vghawk1-root 1.8T 901G **831G** 53% /scratch

Cluster Name: **openalex_dhyeyhp2**

hostname	IP Address	Node-type	Node Name
Hawk1 (least space)	172.22.224.150	Data	data-node-1
Hawk2	**172.22.224.151**	Master-egiible, data, vote	master-node-1
Hawk3	172.22.224.152	Master-egiible, data	master-node-2
Hawk4 (most space)	172.22.224.153	Master-egiible, data	master-node-3
Hawk5	172.22.224.154	Data	data-node-2



ip heap.percent ram.percent cpu load_1m load_5m load_15m node.role master name
172.22.224.150 2 99 1 0.01 0.14 0.17 di - data-node-1
172.22.224.152 4 99 1 0.04 0.52 0.56 dm * master-node-2
172.22.224.154 1 99 0 0.11 0.22 0.16 d - data-node-2

172.22.224.151 2 88 1 0.04 0.05 0.06 dmv - master-node-1

Copy data from one server to other for openalex indexing:

```
scp -r ./publishers dhyeyhp2@hawk2.csl.illinois.edu:/scratch/dhyeyhp2/openalex_dataset/
```

Running of python ingestion script

- Current Ingestion speed of python script is **roughly 1 million records in 4 mins in local machine.**

Ingesting publishers data in the `openalex_publishers` index

```
(venv) dhyeyhp2@hawk2:/scratch/dhyeyhp2/scripts$ python3 data_ingest.py
{'name': 'master-node-1', 'cluster_name': 'openalex_dhyeyhp2', 'cluster_uuid': 'lX7M4p2DQum6dOlufKlClA', 'version': {'number': '8.6.2', 'build_flavor': 'defau
lt', 'build_type': 'tar', 'build_hash': '2d58d0f136141f03239816a4e360a8d17b6d8f29', 'build_date': '2023-02-13T09:35:20.314882762Z', 'build_snapshot': False, '
lucene_version': '9.4.2', 'minimum_wire_compatibility_version': '7.17.0', 'minimum_index_compatibility_version': '7.0.0'}, 'tagline': 'You Know, for Search'}
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-17
0 processed
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-15
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-07
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-16
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-09
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-11
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-13
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-06
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-21
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-20
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-05
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-02
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-14
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-12
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-19
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-03
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-18
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-04
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-08
Indexing /scratch/dhyeyhp2/openalex_dataset/publishers/updated_date=2023-02-10
Completed 3675 records in 8.52 seconds
{'_shards': {'total': 2, 'successful': 2, 'failed': 0}}
(venv) dhyeyhp2@hawk2:/scratch/dhyeyhp2/scripts$
```

Ingested Publishers records in an **openalex_publishers** index
3675 Records: 8.52 seconds

Sources - Ingestion speed

```
215000 processed
220000 processed
Indexing /scratch/dhyeyhp2/openalex_dataset/sources/updated_date=2023-02-10
225000 processed
Completed 226726 records in 89.11 seconds
{'_shards': {'total': 2, 'successful': 2, 'failed': 0}}
(venv) dhyeyhp2@hawk2:/scratch/dhyeyhp2/scripts$
```

Institutions - Ingestion speed

```
Indexing /scratch/dhyeyhp2/openalex_dataset/institutions/updated_date=2023-01-16
Indexing /scratch/dhyeyhp2/openalex_dataset/institutions/updated_date=2022-07-29
Indexing /scratch/dhyeyhp2/openalex_dataset/institutions/updated_date=2023-02-13
Completed 108618 records in 103.76 seconds
{'_shards': {'total': 2, 'successful': 2, 'failed': 0}}
(venv) dhyeyhp2@hawk1:/scratch/dhyeyhp2/scripts$
```

Authors Ingestion speed:

```
54335000 processed
54340000 processed
54345000 processed
54350000 processed
Completed 54351569 records in 24939.41 seconds
{'_shards': {'total': 2, 'successful': 2, 'failed': 0}}
```

Concepts Ingestion Speed:

```
Indexing /scratch/dhyeyhp2/openalex_dataset/concepts/updated_date=2023-01-05
Indexing /scratch/dhyeyhp2/openalex_dataset/concepts/updated_date=2023-01-27
Indexing /scratch/dhyeyhp2/openalex_dataset/concepts/updated_date=2022-11-01
Indexing /scratch/dhyeyhp2/openalex_dataset/concepts/updated_date=2023-02-20
Indexing /scratch/dhyeyhp2/openalex_dataset/concepts/updated_date=2023-01-13
Completed 65073 records in 144.77 seconds
{'_shards': {'total': 2, 'successful': 2, 'failed': 0}}
(venv) dhyeyhp2@hawk3:/scratch/dhyeyhp2/scripts$
```

```
(venv) dhyeyhp2@hawk2:/scratch/dhyeyhp2/scripts$ curl http://172.22.224.151:9200/_cat/shards/openalex_publishers
openalex_publishers 0 p STARTED 3675   2mb 172.22.224.154 data-node-2
openalex_publishers 0 r STARTED 3675 2.1mb 172.22.224.150 data-node-1
```

Check on which nodes are the shard in of index

command: **curl http://172.22.224.151:9200/_cat/shards/openalex_publishers**

Issues and Solutions

Problems	Solutions
Same username `elastic` will it work for both me and Sid?	Change the port of each instance

To change the number of replicas	{
	"index": {
	"number_of_replicas":0,
	"auto_expand_replicas": false
	}
	}

The huge Works dataset is not getting indexed, fully. The servers face a resource crunch. The garbage collector crashes again and again and it takes forever to index the data.	
Multiple Indexes are creating an issue though, the query search time is deprecating	

Notes:

1. Might use Asynchronous I/O for serving search service using an API https://elasticsearch-py.readthedocs.io/en/v8.6.2/async.html#getting-started-with-async

2. Elasticsearch prepares the data for search/query operations while ingesting the data. It is also called "Schema on Write." It is like creating the schema dynamically/explicitly while defining and inserting the record in a SQL Table.

https://aravind.dev/everything-index-elastic/

3. So, how many recommended fields per doc in a single index? 1000 fields are the max limit. [2]
4. Of course, by modifying the **index.mapping.total_fields.limit** setting field limit can be changed.
5. Index templates contain settings like how many shards and replicas the index should initialize with, what mapping settings, and aliases to use. One can also assign priority to the index template. 100 is the default priority.
6. High availability (HA) clusters require at least three master-eligible nodes, at least two of which are not voting-only nodes. Such a cluster will be able to elect a master node even if one of the nodes fails.
7. Voting-only master-eligible nodes may also fill other roles in your cluster. For instance, a node may be both a data node and a voting-only master-eligible node. A *dedicated* voting-only master-eligible node is a voting-only master-eligible node that fills no other roles in the cluster. To create a dedicated voting-only master-eligible node, set:
8. {
   "id": "XwI104YBGZAj3XIhC-x2",
   "name": "python_example",
   "api_key": "xVE1zNGjS1acurq6WjOj_Q",
   "encoded":
"WHdJMTA0WUJHWkFqM1hJaEMteDI6eFZFMXpOR2pTMWFjdXJxNldqT2pfUQ=="

9. It is recommended to keep the size of each index below 50GB to ensure optimal performance, as larger indexes can lead to longer query response times and increased resource usage. It is also important to consider the number of shards per index, as having too many or too few shards can also affect performance.
10. https://www.elastic.co/guide/en/elasticsearch/reference/current/tune-for-search-speed.html
11. **filesystem cache is probably Elasticsearch's number 1 performance factor.** https://www.elastic.co/guide/en/elasticsearch/reference/current/tune-for-search-speed.html
12. the trade-off between throughput and availability
13. https://www.elastic.co/guide/en/elasticsearch/reference/8.6/search-profile.html **To measure the performance of search queries**

14. Had to increase the fields limit number when indexing concepts data, since it exceeded the elastic search by default 1000 fields limit.

Commands/API:

Configurations

1. Copy open alex data from one server to another:
   a. scp -r ./publishers dhyeyhp2@hawk2.csl.illinois.edu:/scratch/dhyeyhp2/openalex_dataset/

2. Checking cluster health using curl
   a. curl http://172.22.224.151:9200/_cluster/health?pretty
3. Ressentialpython script in the background, with logs being generated
   a. nohup venv/bin/python -u data_ingest.py > authors_output.log &
   b. tail -f authors_output.log

4. Checking the node configurations
   a. curl http://172.22.224.151:9200/_cat/nodes?v

5. Checking Disk usage of each node in the cluster
   a. curl http://172.22.224.151:9200/_cat/allocation?v

6. Check which nodes are the shard in index
   a. **curl http://172.22.224.151:9200/_cat/shards/openalex_publishers**

7. How to ssh into a remote machine without a password:
   a. ssh-keygen
   b. ssh-copy-id remoteuser@hostname

8. Changing the fields limit size in concepts index
   a. curl -s -XPUT http://172.22.224.151:9200/openalex_concepts/_settings -H 'Content-Type: application/json' -d '{"index.mapping.total_fields.limit": 2000}'
9.

Keyword Search

Works:

a. **Find all the Works that have any author from either RUSSIA (RU) or the US**. Also,
   Just selective output fields in the response.
   i. http://172.22.224.152:9200/openalex_works/_search?q=authorships.institutions.country_code:(RU OR FR)&_source=title,publication_year,authorships.institutions.country_code,primary_location.landing_page_url