FLOWx.AI

PLATFORM SETUP GUIDES / FLOWX.AI Engine setup guide / Process instance indexing / process-instance-indexing-config-guidelines

PLATFORM SETUP GUIDES / FLOWX.AI Engine setup guide /
Process instance indexing / process-instance-indexing-config-
guidelines

Contents

PLATFORM SETUP GUIDES / FLOWX.AI Engine setup guide / Process instance indexing / Configuration guidelines

The configuration of Elasticsearch for process instances indexing depends on various factors related to the application load, the number of process instances, parallel requests, and indexed keys per process. Although the best approach to sizing and configuring Elasticsearch is through testing and monitoring under load, here are some guidelines to help you get started:

Indexing strategy

When deleting data in Elasticsearch, it's recommended to delete entire indices instead of individual documents. Creating multiple smaller indices provides the flexibility to delete entire indices of old data that are no longer needed.

- Advantages of Multiple Small Indices:
    - Fast indexing process.

PLATFORM SETUP GUIDES / FLOWX.AI Engine setup guide /
Process instance indexing / process-instance-indexing-config-
guidelines

- Flexibility in cleaning up old data.
- Potential Drawbacks:
  - Hitting the maximum number of shards per node, resulting in exceptions when creating new indices.
  - Increased search response time and memory footprint.

Alternatively, you can create fewer indices that span longer periods of time, such as one index per year. This approach offers small search response times but may result in longer indexing times and difficulty in cleaning up and recovering data in case of failure.

» What is indexing?

Shard and replica configuration

The solution includes an index template that gets created with the settings from the process-engine app (name, shards, replicas) when running the app for the first time. This template controls the settings and mapping of all newly created indices.

» What is sharding?

» Index template

Once an index is created, you cannot update its number of shards and replicas. However, you can update the settings from the index template at runtime in

PLATFORM SETUP GUIDES / FLOWX.AI Engine setup guide /
Process instance indexing / process-instance-indexing-config-
guidelines

Elasticsearch, and new indices will be created with the updated settings. Note that the mapping should not be altered as it is required by the application.

Recommendations for resource management

To manage functional indexing operations and resources efficiently, consider the following recommendations:

- Sizing indexes upon creation
- Balancing
- Delete unneeded indices
- Reindex large indices
- Force merge indices
- Shrink indices
- Combine indices

Sizing indexes upon creation

Recommendations:

- Start with monthly indexes that have 2 shards and 1 replica. This setup is typically sufficient for handling up to 200k process instances per day; ensures a parallel indexing in two main shards and has also 1 replica per each main shard (4 shards in total). This would create 48 shards per year in the elastic search nodes; A lot less than the default 1000 shards, so you will have enough space for other indexes as well.
  - If you observe that the indexing gets really, really slow, then you should look at the physical resources / shard size and start adapting the config.

PLATFORM SETUP GUIDES / FLOWX.AI Engine setup guide /
Process instance indexing / process-instance-indexing-config-
guidelines

- If you observe that indexing one monthly index gets massive and affects the performance, then think about switching to weekly indices.

- If you have huge spikes of parallel indexing load (even though that depends on the Kafka connect cluster configuration), then think about adding more main shards.

- Consider having at least one replica for high availability. However, keep in mind that the number of replicas is applied to each shard, so creating many replicas may lead to increased resource usage.

- Monitor the number of shards created and estimate when you might reach the maximum shards per node, taking into account the number of nodes in your cluster.

Balancing

When configuring index settings, consider the number of nodes in your cluster. The total number of shards (calculated by the formula: primary_shards_number * (replicas_number +1)) for an index should be directly proportional to the number of nodes. This helps Elasticsearch distribute the load evenly across nodes and avoid overloading a single node. Avoid adding shards and replicas unnecessarily.

Delete unneeded indices

Deleting unnecessary indices reduces memory footprint, the number of used shards, and search time.

Reindex large indices

If you have large indices, consider reindexing them. Process instance indexing involves multiple updates on an initially indexed process instance, resulting in multiple versions of the same document in the index. Reindexing creates a new

PLATFORM SETUP GUIDES / FLOWX.AI Engine setup guide /
Process instance indexing / process-instance-indexing-config-
guidelines

index with only the latest version, reducing storage size, memory footprint, and search response time.

Force merge indices

If there are indices with no write operations performed anymore, perform force merge to reduce the number of segments in the index. This operation reduces memory footprint and response time. Only perform force merge during off-peak hours when the index is no longer used for writing.

Shrink indices

If you have indices with many shards, consider shrinking them using the shrink operation. This reindexes the data into an index with fewer shards. Perform this operation during off-peak hours.

Combine indices

If there are indices with no write operations performed anymore (e.g., process_instance indices older than 6 months), combine these indices into a larger one and delete the smaller ones. Use the reindexing operation during off-peak hours. Ensure that write operations are no longer needed from the FLOWX platform for these indices.

Was this page helpful?