Cloudera Runtime 7.2.18

Planning for Infra Solr

Date published: 2021-06-29
Date modified:

https://docs.cloudera.com/

Legal Notice

Contents

Calculating Infra Solr resource needs

Based on the number of audit records per day, you can make sizing assumptions for your Infra Solr servers.

About this task

Infra-Solr is not intended for custom workloads. As Infra-Solr typically hosts sensitive, high-traffic collections such as ranger_audits, its utilization for custom collections could result in:

- Performance degradation
- Resource contention
- Disruption of audit logging

Cloudera strongly advises the deployment of a distinct Solr instance for any workloads beyond those automatically deployed on Infra-Solr. Employing a separate instance ensures:

- A clear delineation between audit-related and workload-related functionalities
- Enhanced resource management
- Safeguarding of critical services, including Ranger audit logging

It is recommended to refrain from adding custom Solr instances on the same nodes as Infra-Solr instances to avert resource conflicts.

Procedure

1. To determine how many audit records are indexed per day, execute the following command using a HTTP client, such as curl:

   ```
   curl -g "http://[***INFRA_SOLR_HOSTNAME***]:[***PORT***]/solr/ranger_aud
   its/select?q=(evtTime:[NOW-7DAYS+TO+*])&wt=json&indent=true&rows=0"
   ```

   Replace *[***INFRA_SOLR_HOSTNAME***]* and *[***PORT***]* with values valid in your environment. The default port is 8886.

   Tip:

   You can also replace the ‘7DAYS’ in the curl request with a broader time range, if necessary, using the following key words:

   - 1MONTHS
   - 7DAYS

   You receive a message similar to the following:

   ```
   {
     "responseHeader":{
       "status":0,
       "QTime":1,
       "params":{
         "q":"evtTime:[NOW-7DAYS TO *]",
         "indent":"true",
         "rows":"0",
         "wt":"json"}},
     "response":{"numFound":306,"start":0,"docs":[]
     }}
   ```

2. Take the numFound element of the response and divide it by 7 to get the average number of audit records being indexed per day.

   Make sure you divide by the appropriate number if you change the event time query. The average number of records per day will be used to identify which recommendations below apply to your environment.

What to do next

These recommendations assume that you are using a 12 GB heap per Solr server instance. In each situation we have recommendations for co-locating Solr with other master services, and for using dedicated Solr servers. Testing has shown that Solr performance requires different server counts depending on whether Solr is co-located or on dedicated servers. Based on our testing with Ranger, Solr shard sizes should be around 25 GB for best overall performance. However, Solr shard sizes can go up to 50 GB without a significant performance impact.

Note: If you want to reduce the number of audits, see Ranger Audit Filters.

In each recommendation, the time to live, or TTL, which controls how long a document should be kept in the index until it is removed is taken into consideration. The default TTL is 90 days, but some customers choose to be more aggressive, and remove documents from the index after 30 days. Due to this, recommendations for both common TTL settings are specified.

- **Less Than 50 Million Audit Records Per Day**

  Based on the Solr REST API call if your average number of documents per day is less than 50 million records per day, the following recommendations apply.

  This configuration is our best recommendation for just getting started with Ranger and Infra Solr so the only recommendation is using the default TTL of 90 days.

  Default Time To Live (TTL) 90 days:

  - Estimated total index size: ~150 GB to 450 GB
  - Total number of shards: 6
  - Total number of replicas, including 2 replicas (1 leader, 1 follower) for each shard: 12
  - Total number of co-located Solr nodes: ~3 nodes, up to 2 shards per node (does not include replicas)
  - Total number of dedicated Solr nodes: ~1 node, up to 12 shards per node (does not include replicas)

50 - 100 Million Audit Records Per Day

50 to 100 million records ~ 5 - 10 GB data per day.Default Time To Live (TTL) 90 days:

- Estimated total index size: ~ 450 - 900 GB for 90 days
- Total number of shards: 18-36
- Total number of replicas, including 1 replica for each shard: 36-72
- Total number of co-located Solr nodes: ~9-18 nodes, up to 2 shards per node (does not include replicas)
- Total number of dedicated Solr nodes: ~3-6 nodes, up to 12 shards per node (does not include replicas)

Custom Time To Live (TTL) 30 days:

- Estimated total index size: 150 - 300 GB for 30 days
- Total number of shards: 6-12
- Total number of replicas, including 1 replica for each shard: 12-24
- Total number of co-located Solr nodes: ~3-6 nodes, up to 2 shards per node (does not include replicas)
- Total number of dedicated Solr nodes: ~1-2 nodes, up to 12 shards per node (does not include replicas)

100 - 200 Million Audit Records Per Day

100 to 200 million records ~ 10 - 20 GB data per day. Default Time To Live (TTL) 90 days:

- Estimated total index size: ~ 900 - 1800 GB for 90 days
- Total number of shards: 36-72
- Total number of replicas, including 1 replica for each shard: 72-144
- Total number of co-located Solr nodes: ~18-36 nodes, up to 2 shards per node (does not include replicas)
- Total number of dedicated Solr nodes: ~3-6 nodes, up to 12 shards per node (does not include replicas)

Custom Time To Live (TTL) 30 days:

- Estimated total index size: 300 - 600 GB for 30 days
- Total number of shards: 12-24
- Total number of replicas, including 1 replica for each shard: 24-48
- Total number of co-located Solr nodes: ~6-12 nodes, up to 2 shards per node (does not include replicas)
- Total number of dedicated Solr nodes: ~1-3 nodes, up to 12 shards per node(does not include replicas)

- If you choose to use at least 1 replica for high availability, then increase the number of nodes accordingly. If high availability is a requirement, then consider using no less than 3 Solr nodes in any configuration.
- As illustrated in these examples, a lower TTL requires less resources. If your compliance objectives call for longer data retention, you can use the SolrDataManager to archive data into long term storage (HDFS, or S3), which also provides Hive tables allowing you to easily query that data. With this strategy, hot data can be stored in Solr for rapid access through the Ranger UI, and cold data can be archived to HDFS, or S3 with access provided through Ranger.