Sizing Guidelines for Using Cloudera Navigator XConnect with Metadata Manager

Abstract

This document describes the performance of Metadata Manager when you use Cloudera Navigator XConnect for data management. Use the data provided in this document as a performance baseline when you configure Metadata Manager with Cloudera Navigator XConnect in your organization.

Supported Versions

- Informatica Metadata Manager 10.1.1

Table of Contents

Overview

This document defines the requirements and measurements of resources for the Cloudera Navigator XConnect environment to run the Cloudera Navigator load with reasonable performance for 120 million entities and 240 million relationships assuming a moderate level of data and relationship complexity.

Metadata Manager Configuration and Projection

Describes the server and database details of the Informatica domain where Metadata Manager was installed and tested. The load projection is for a Cloudera load of 120 Million entities based on Informatica's testing of 5.5 million load.

If you want to increase the volume of the load, you must be able to predict future data sizes based on the needs of your organization. Forecasting the future data size helps you predict in advance the changes you will need to make and the point of time at which the growing data size will exceed the capacity of your environment. The predictive sizing guidelines will help you avoid load failures. Informatica recommends that you forecast every month the future data volume for a number of future points in time.

The expected data volume and capacity based on your organization's needs should always have a three month buffer. You can calculate the buffer based on the present data loading capacity. The buffer should exceed the immediate expected data volume by some margin on top of the future data size that you forecasted for three months in the future. If the time to acquire new hardware takes long, the margin on top of the future forecast should be increased.

Informatica Domain

This section describes the test results of the hardware used to install Informatica. Metadata Manager runs as a service in an Informatica domain. The hardware testing is based on the loading of 120 million entities from Cloudera using

simple data. If your organization uses complex data with a large number of relationships, you will have to increase the hardware configuration to something more powerful than the configuration used for testing.

The following table describes the hardware details of the server in which the Informatica domain was configured and the Metadata Manager service was enabled:

Operating System Details	Value for 120 Million Entities	Increment Value for Additional 10 Million Entities
RAM	64 GB	Not Applicable
CPU	64 cores of minimum 2.5 GHz	2 cores of minimum 2.5 GHz
Disk Type	SATA or RAID 5	-
Disk I/0	1.5 GB /sec	-
Disk space (minimum requirement)	1 TB	100 GB

Database

You can use an Oracle database for the Metadata Manager warehouse.

The following table describes the database server component values:

Parameter Name	Value for 120 Million Entities	Increment for Additional 10 Million Entities
Total RAM	64 GB	5 GB
CPU Cores	72 CPU cores	-
CPU Speed	2.30 GHz	-
SGA/PGA	40 GB/10 GB	10 GB/4 GB
Automatic Memory Management	Enabled	-
audit_trail	DB	-
Compatbible	11.2.0.4.0	-
control_management_pack_access	DIAGNOSTIC + TUNING	-
db_block_size	8192	-
open_cursors	3000	-
plsql_warnings	DISABLE:ALL	-
Processes	500	-
Sesions	792	-

The following table describes the tablespace configuration that was used:

Tablespace Configuration	Value for 120 Million Entities	Increment for Additional 10 Million Entities
Data Files Spindle Distribution	Must be on different disk spindles	Must be on different disk spindles
Data File Location	Data files must be on local disks	Data files must be on local disks
Data File I/O Speed	Disk I/O Rate >= 1.5 GB/sec	Disk I/O Rate >= 1.5 GB/sec
Users Tablespace	300 GB	30 GB
Undo Tablespace	32 GB	32 GB
Temp	32 GB	32 GB

Parameter Projection

This section projects the parameters for a Cloudera load of 120 Million entities based on Informatica's testing of 5.5 million load.

The version of Cloudera used was version 5.8. The number of entities extracted was 5,521,873. The number of relationships extracted was 3,682,13.

The following table describes the parameter projection:

Volume	5.5 Million Load	120 Million Load
Metadata Manager heap space	8 GB	40 GB
Temporary file space (JSON)	9 GB	250 GB
Temporary file space (IME)	18 GB	450 GB
JSON extraction	50 minutes	24 hours
Writing to staging table	50 minutes	24 hours
IME File Creation	20 minutes	8 hours
ETL	48 minutes	24 hours
Post load tasks (includes graph creation, linking and indexing)	30 minutes	20 hours
Total load time	3 hours and 27 minutes	Approximately 4 days

Cloudera Navigator

Cloudera Navigator is a data management tool for the Hadoop platform. If your organization uses complex data with a large number of relationships, you will have to increase the hardware configuration to something more powerful than the configuration used for testing.

The following table describes the hardware details of the server in which Metadata Manager service was enabled and 120 million entities of simple data was loaded to Metadata Manager using Clouder Navigator:

Parameter	Value for 120 Million Entities	Increment Value for Additional 10 Million Entities
RAM available on the Data Nodes	32 GB	Not Applicable
Cloudera Version	Refer to the Informatica Product Availability Matrix (PAM)	Not Applicable
Cloudera Navigator Heap Size	16 GB	Not Applicable
Network Speed (Max Usage)	1.5 Mbps	-

Parameters to Monitor

Monitor the following parameters to optimally utilize the resources:

- Cloudera REST API calls – The network latency must be less than 100 milliseconds. Contact Informatica for the utility to measure various metrics while the job runs.
- ETL – The database write speed must be 12,000 records/sec or faster. You can monitor the write speed using AWR reports.
- Disk space

The following table describes the disk space for JSON files, temp files, and IME files that the Metadata Manager Service requires:

Parameter Name	Value for 120 Million Entities	Increment for Additional 10 Million Entities
Metadata Manager Heap Space	40 GB	4 GB
Temporary File Space (JSON)	250 GB	20 GB
Temporary File Space (IME)	450 GB	40 GB
User Table Space in the Database	400 GB	30 GB

Use the following command to measure some of the parameters like Disk I/O:

```
hdparm -Td /dev/sda1
```

Recommendations

- The recommended I/O must be at least 1.5 GB/sec in the database server.
- The recommended I/O must be at least 1 GB/sec in the Informatica domain server. This determines the write speed for JSON and IME files.

Performance Test Results

An Oracle database was used in the Informatica performance testing lab environment to load 11 million entities and 22 million relationships of fairly simple data. The version of Oracle tested was 11.2.0.4.0 standalone.

The following table describes the tablespace configuration that was used:

Tablespace Configuration	Value for 11 Million Entities
Data Files Spindle Distribution	Must be on different disk spindles
Data File Location	Data files must be on local disks
Data File I/O Speed	Disk I/O Rate >= 1.5 GB/sec
Users Tablespace	300 GB
Undo Tablespace	32 GB
Temp	32 GB

Load Comparison

Describes the performance of a Cloudera load job which was run in an in-house installation of Metadata Manager verison 9.6.1 with performance enhanced code.

The following table shows the comparison in hardware between the lab environment used in Informatica and the User Acceptance Test (UAT) environment for a large financial organization:

Parameters	Informatica Instance	Large Financial Organization Instance
CPU	24 Core	4 Core
RAM	62 GB	48 GB
Server	Physical Machine	Virtual Machine

The following table shows the comparison in test results from the lab environment used in Informatica and the UAT environment for a large financial organization:

Parameters	Informatica Instance	Large Financial Organization Instance
Cloudera Metadata Manager load duration for 3 million entities and 6 million relationships (end to end completion)	~ 4 hours	> 12 hours
Cloudera MM load duration for 11 million entities and 22 million relationships	~ 15 hours	> 24 hours

Author

Pratap J
Lead Technical Writer

Acknowledgements

The author would like to acknowledge Srinivasa Raghavan, Rashmi Mani, Karthick Kandaswamy, and Abhishek Asthana for their contributions to this article.