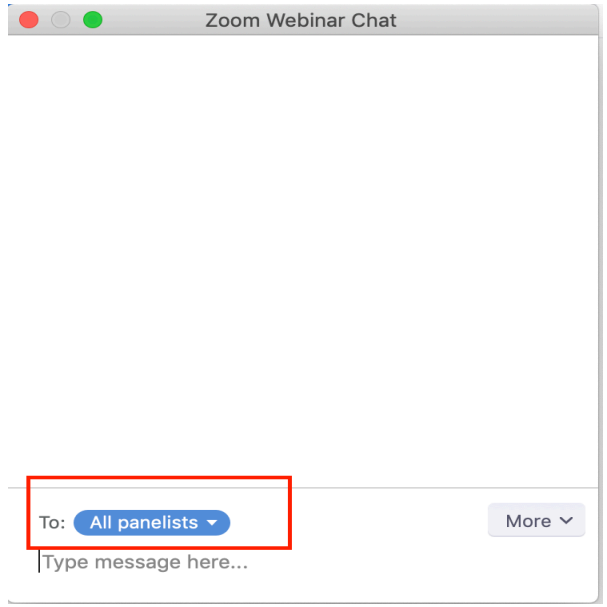
MDM Best Practices
&
Performance Parameters

Informatica, Global Customer Support

Informatica™

Housekeeping Tips



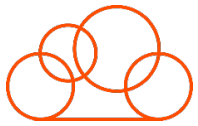
- Todays Webinar is scheduled to last 1 hour including Q&A
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to “All Panelists" via the Chat option and we will respond at the end of the presentation
- The webinar is being recorded and will be available to view on our INFASupport YouTube channel. The link will be emailed as well.
- Please take time to complete the post-webinar survey and provide your feedback and suggestions for upcoming topics.

Informatica

Success Portal

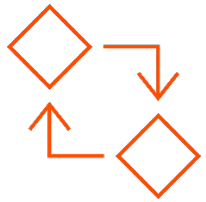
https://success.informatica.com

Learn. Adopt. Succeed.



Bootstrap product trial experience

Enriched Onboarding experience



Product Learning Paths and weekly Expert sessions

Informatica Concierge with Chatbot integrations

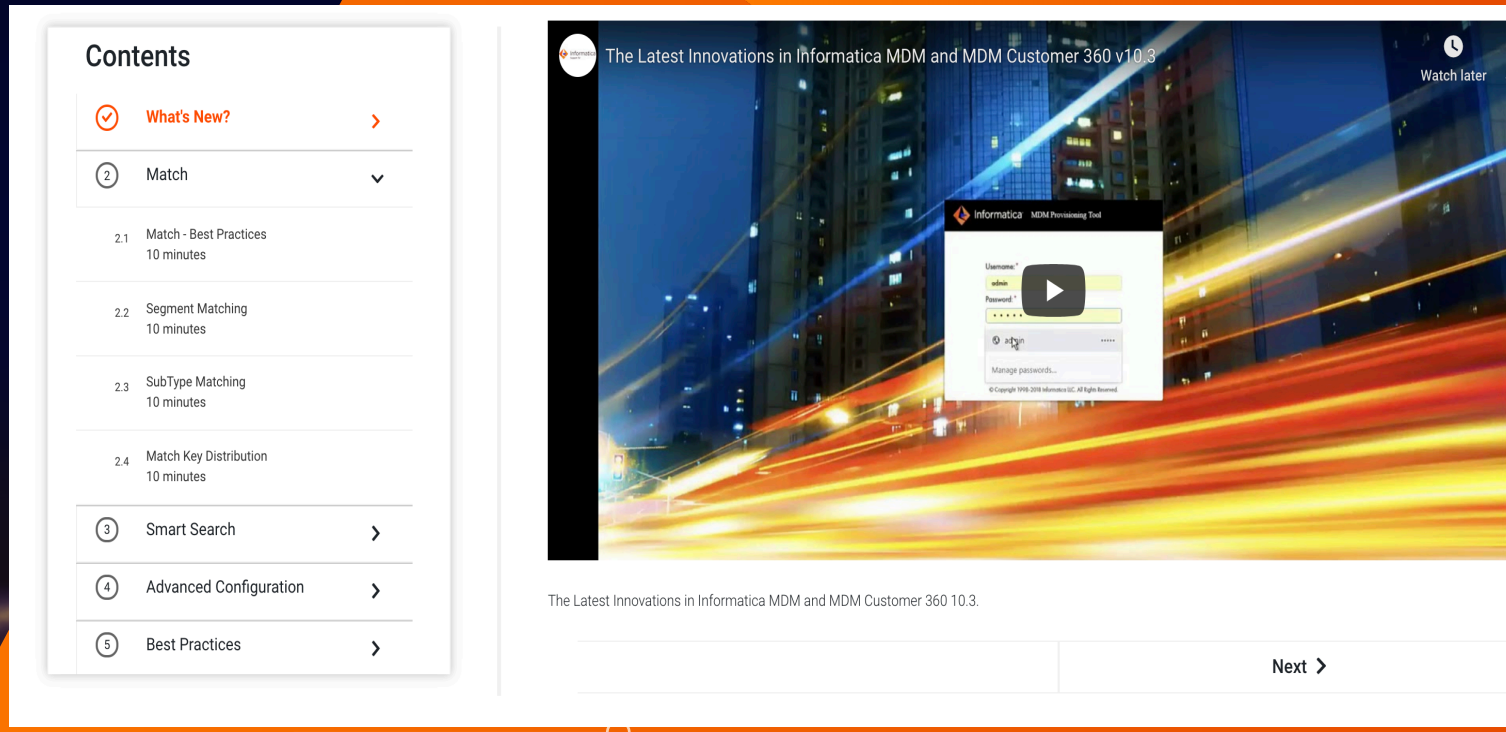
Tailored training and content recommendations

Informatica

Product Learning Path

"Product Learning Path" *is an unique structured product enablement to help bootstrap you with your product journey*

- Informatica's ***Free*** Training Platform
- Catalog of microlearning tutorials
- Provides foundational to advanced Informatica product knowledge
- Supplements your existing skills
- Breadth of enablement covering developers, administrators and architects and many more personas



Informatica

Disclaimer

- *This service is intended for educational and informational purposes only. Informatica does not offer any warranty or representation, expressed or implied, with respect to the accuracy, completeness or usefulness of the information provided through this service. The materials and other information provided through this service are not designed or intended to replace the services offered by Informatica Professional Services, Informatica partners and any third-party service providers.*

- *This service offering is not a commitment to deliver any environment, material, code, or functionality, and should not be relied upon in making major implementation decisions. If such scenario arises, you may need to work with Informatica Professional Services. This service is also not meant for analyzing individual support cases or debugging specific product behavior.*

- *Informatica intends this service to meet the intended purpose but doesn't guarantee accuracy. Therefore, Informatica disclaims any warranty of any kind, whether express or implied as to any matter whatsoever relating to this service, including without limitation merchantability or fitness for any particular purpose. Thus, by using this service, the user releases Informatica from and against any and all such liability.*

Agenda

- MDM – Best Practice
  - ORS Level
- Performance Parameters
  - Oracle settings ,SQL server in progress
  - JAVA / App Server settings
- MDM Jobs Properties
  - Stage , tokenization, Match
  - Load, AutoMerge, Unmerge, Recalculate BVT, Revalidation
- MDM Batch processing threads
- Smart Search
- MDM - Housekeeping
- MDM Job repository tables

MDM – Best Practice

- Register only in-use ORS
- Unwanted Base Objects can lead to slow ORS validation
- LDAP can be added to Security providers - mostly used on top
- SAM is powerful tool to control User Access – Avoid inheriting roles and creating too many roles.
- Data Steward options are more in IDD/e360 as compared to MDM HubConsole
- Batch execution can be automated using utility present in Resource Kit

Best Practice – ORS Level

- **HUB & Process Server Logging**
  - Set to INFO/Error mode in production in the file appender and the following categories
    - com.delos
    - com.siperian
    - com.informatica
    - Siperain.performance (off)
  - It can be enabled at DEBUG mode in development and testing environments
- **Auditing**
  - Auditing introduces additional overhead and should be disabled
  - Can be enabled only for debugging purposes
- **Archive Logging**
  - It can be disabled during IDL for large volume of data

Best Practice – ORS Level

- **Batch API Inter-operability**
  - Enable ONLY if both real time and batches are used
  - If your application uses neither real time API updates nor IDD, do not enable API Batch Interoperability
  - Enabling this will have a slight impact in performance
  - During Initial Data Load, disable this property for faster loading of data

Best Practice – ORS Level- cache

- **Production Mode**
  - Enable this property in Production, acquiring schema write lock can lead to slow performance .
  - Can be configured from the HUB
  - MDM internally uses infinispan cache

Best Practice – ORS Level - cache

- **User Profile Cache**
  - cmx.server.provider.userprofile.cacheable = true (default is true)
  - When this flag is set to true, the user profile once authenticated is cached.
  - User is authenticate for every API call

- **User Profile Life Span**
  - Time to retain the cached user (in milliseconds) before refreshing, default is 1 min
  - In production Set cmx.server.provider.userprofile.**lifespan** =3600000
  - cmx.server.provider.userprofile.**expiration**=3600000

Best Practice -Infinispan setup

- Hub/server/resources/infinispanceConfig.xml, jgroups-tcp, jgroups-udp
- Ensure each environment has a unique multicast port, avoids overhead
- On a multi-node clustered environment /or load balanced, infinispan clustering must be established.
  - make sure cmx_system.c_repos_database.database_host is not set to localhost.
    - Directly update this value to a hostname or ip that is uniquely resolvable (won't affect anything but the cache name), update cmx_system.c_repos_database set database_host=<appserver1's IP> **where rowid_database='CMX.MASTER'**
  - Make sure the ports used for multicast are open
  - KB 568046. for infinispan set up in clustered environment
- In production for better performance, update the lifespan values as recommended in KB 509572

Best Practice –Snippet of infinispanConfig.xml

```
<transport stack="udp" cluster="infinispan-cluster"

                    ...

      <replicated-cache name="MDM_SESSION_MANAGEMENT" statistics="true" mode="SYNC"
remote-timeout="20000">

                        <expiration interval="5000" lifespan="360000" />

                    </replicated-cache>
```


Informatica

Best Practice – ORS Level

- **Dynamic Match Analysis Threshold (DMAT)**
  - DMAT will help improve performance when large ranges are causing slowness
  - It is recommended to analyze the data to assess why a given search range contains a large count
  - Setting the level too low may cause under matching
  - KB 90740 - in-depth explanation of DMAT

Best Practice – ORS Level

- **Try to keep matching on the Process server**
  - Matching in process server is much faster as it can take advantage of parallel threads and multiple process servers
  - Fuzzy rules - matching is done in the process server
  - Exact rules -When all match columns used are Exact ,Match occurs on the Database
  - Filtered rules – When all columns are exact but fuzzy key column is used for search, matching is done on the process server

Best Practice – ORS Level

- **Custom/Backup tables created with name starting with C_REPOS%**
  - Do not name the custom/backup tables with names starting with C_REPOS%
  - Degrades Hub Console performance
- **High Volume of Data in repository tables with historical data**
  - Regular maintenance on the below METADATA tables
    - C_REPOS_AUDIT
    - C_REPOS_MQ_DATA_CHANGE
    - C_REPOS_JOB_CONTROL - cmx.server.jobControl.noOfDays=n to control the display of historical batch job data
    - C_REPOS_JOB_METRIC
    - C_REPOS_MET_VALID_RESULT
    - C_REPOS_MET_VALID_MSG
    - C_REPOS_TASK_ASSIGNMENT_HIST
  - **NOTE: Don’t truncate C_REPOS_JOB_* tables if RAW retention is enabled; Contact support**

Best Practice -Tasks

- Cmxserver.properties related to merge task creation
  - sip.task.assignment.start.delay=0 – Number of minutes before Automated task assignments starts after hub initialization. Set it to a non-zero value
  - sip.task.assignment.interval=0 - Number of Minutes between automated task assignments. 0 means that assignment is disabled
  - task.creation.maximum=50 – Will stop creating tasks after 50 tasks are created for a BO
  - task.creation.batch.size=1000 – how many tasks can be created per execution when task.creation.maximum has not been reached.

Performance Parameters

- Database
  - Parameters that can be fine-tuned in database layer
  - Applicable for both CMX_SYSTEM and ORS schema.
- Java
  - Parameters that can be fine-tuned in Java Layer.
  - Applicable for both Hub and Process servers, including clustered environment.

Performance Parameters - DB settings

- **Init.ORA Parameters**
  - We do hash join between large tables which require large SGA memory allocation (Ref. P&T guide)
  - Recommended parameters are in KB 323799
- **Use Dedicated environment for Database**
  - At least for production
- **Database IO speed**
  - TEST_IO utility is devised specifically to test database disk I/O response based on select/insert/delete/update with a known data set.
  - It is based on the benchmark disk IO performance.
  - The Total elapsed time for Test_IO, is dependent on the Load on the system and other activity, including network contention.
  - Share the output of the tool with Informatica Support

Performance Parameters – SQL server

- **Locks -**5000 before lock promotion
- PARAMETERIZATION **FORCED**
- READ_COMMITTED_SNAPSHOT **ON**
- AUTO_UPDATE_STATISTICS_ASYNC **ON**

Performance Parameters – JAVA / App Server settings

- **JVM Max heap**
  - Hub Server: min 4GB- (if >6G G1GC garbage collection policy)
  - Process Server: min 4G - (if >6G G1GC garbage collection policy)
- **Maximum Connection (datasource)**
  - For ORS the recommendation is 2.5 connections per IDD/SIF users/connections
    - So for 50 concurrent users, it should be set to 125
    - For 100 concurrent users, connection pool of 250 and so on
  - For CMX_SYSTEM the recommendation is around 1.5 connections per IDD/SIF/REST users
    - So for 50 concurrent users, it should be set to 75

Performance Parameters – App Server settings Contd..

- **JBOSS Pool Threads**
  - Async thread pool - Default value is 10 but recommendation is to increase it to 20 or 30

<async thread-pool-name="default>

<thread-pools>

<thread-pool name="default">

<max-threads count="10"/>

<keepalive-time time="100" unit="milliseconds"/>

</thread-pool>

- EJB Connection Pool - Default value is 20 but recommendation is to increase at least 50 or more, if lot of EJB connections are being utilized in the environment.

<strict-max-pool name="slsb-strict-max-pool" **max-pool-size="20**" instance-acquisition-timeout="5" instance-acquisition-timeout-unit="MINUTES"/>

Informatica

MDM Jobs Properties

- Batch groups can be grouped into two categories
  - Stage, tokenization & match jobs – uses legacy multithreading & distribution
  - Load,merge,recalculateBVT, batchUnmerge, Revalidate, Recalculate, batchDelete

MDM Jobs Properties – Stage Jobs

- **Console**
  - Threads for Cleanse Operations (hubConsole->Process Server)
    - should be less than number of cpu's, start with n-1 and check the cpu load on both application server and database and tune appropriately.
    - Used by stage, tokenization and match jobs

- **cmxcleanse.properties**
  - cmx.server.cleanse.min_size_for_distribution
    - Specifies the minimum size for distribution.. Default is 1000.

MDM Jobs Properties – Match Jobs

- **Console**
  - For parallelism uses “Threads for Cleanse Operation”, same as stage & tokenization.
- **cmxcleanse.properties**
  - cmx.server.match.distributed_match (Default is 1)
    - Set to 1 to enable a Process Server for distributed cleanse and match. Set to 0 to disable..
  - cmx.server.cleanse.min_size_for_distribution
    - Specifies the minimum size for distribution. Default is 1000.
  - cmx.server.match.file_load
    - Set to true to use an intermediate file to load data. Set to false for direct data load. Default is true for Oracle and IBM DB2 environments. Default is false for Microsoft SQL Server environments
  - cmx.server.ma Default is 1 tch.loader_batch_size
    - Maximum number of insert statements to send to the database during direct load of the match process. Default is 1000, when file_load is set to false.
  - ‘Maximum matches for manual consolidation’ (default 1000)– increase it as needed to avoid match job failure
  - ‘Number of rows per match job batch cycle’ (default 10)– increase as needed for better performance, start with 1million

MDM Jobs Properties – Match Jobs cont..

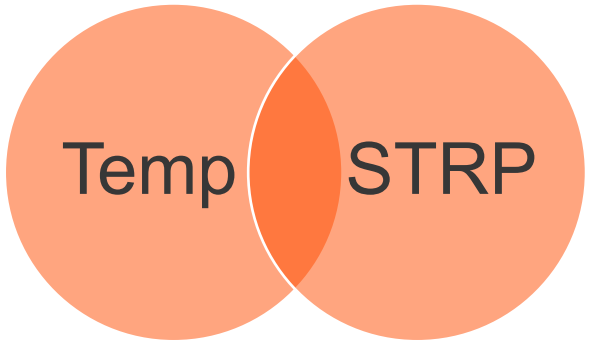
- **Console**
  - ‘Maximum matches for manual consolidation’ (default 1000) – increase it as needed to avoid match job failure
  - ‘Number of rows per match job batch cycle’ (default 10) – increase as needed for better performance, start with 1million
  - Max elapsed match minutes (Default 20 mintues) – Match job will stop after 20 minutes
    - hubConsole , BaseObject -> advanced properties

MDM Jobs Properties – Tokenize Jobs

- **Console**
  - For parallelism uses “Threads for Cleanse Operation”, same as stage job.
- **cmxcleanse.properties**
  - cmx.server.tokenize.file_load =true
    - Set to true to use an intermediate file to load data. Set to false for direct data load. Default is true for Oracle and IBM DB2 environments. Default is false for Microsoft SQL Server environments
    - For multibyte characters on oracle set c_repos_db_release.column_length_in_bytes_ind=0. This will enable sqlldr to use char instead of bytes to load the data
  - cmx.server.tokenize.loader_batch_size
    - Maximum number of insert statements to send in a batch, to the database during direct load of the tokenization process. Default is 1000

Best Practice – Tokenize jobs continued. Blocking

- **Avoid real time call during tokenization**
  - STRIP_CTAS_DELETE_RATIO (Default 10%)
  - STRIP_CTAS_DELETE_UPPER_LIMIT (Default value 40,000);
  - Tokens are deleted/inserted into STRP table if both of the above thresholds are not met
  - With BatchAPI on, as long as the records to be tokenized are less than above mentioned properties, Dirty records will be deleted from STRP table and inserted using parallel threads.
  - Else STRP table is renamed, temp table renamed to STRP and unchanged records inserted into new STRP table from old STRP in a single transaction, may cause blocking for API calls
  - Best to change STRIP_CTAS_DELETE_UPPER_LIMIT , (rule of thumb 10% of the total BO records)
  - cmx.server.strip.DML.useDeleteInsertLock=true (KB 573020)



MDM Jobs Properties – Tokenize Jobs Contd..

- **cmxcleanse.properties**
  - cmx.server.stripDML.noOfThreadsForDelete
    - Number of threads that the MDM Hub uses to delete records from the match key tables. Default is 30.
  - cmx.server.stripDML.noOfThreadsForInsert
    - Number of threads that the MDM Hub uses to insert records into the match key tables. Default is 50.
  - cmx.server.stripDML.noOfThreadsForUpdate
    - Number of threads that the MDM Hub uses to update records in the match key tables. Default is 30.
  - cmx.server.stripDML.blockSize
    - Number of records that the MDM Hub processes in each block. Default is 100.
  - cmx.server.stripDML.useUpdate (applicable only for DB2)
    - Update the match tokens as invalid. Do not delete the tokens in real time.

MDM Batch processing threads

- **Console**
  - Threads for Batch Processing, set via hubConsole->ProcessServer
    - This property set the maximum number of threads that can be used by batch jobs per process server (other than stage, tokenization, match)
    - Set this to a value > number_of_parallel_jobs * max_threads_each_job_uses/number of batch enabled process server. Example if 3 load jobs are running and cmx.server.batch_threads_per_job is set to 10, with 2 process servers that can handle batch jobs then set the value to 3*10/2 =16 on each process server.
    - If there are multiple process servers then on each you can set this value.
- **cmxserver.properties**
  - cmx.server.batch.threads_per_job (used by load, recalculate, revalidate)
  - cmx.server.batch.unmerge.threads_per_job
  - cmx.server.batch.acceptunmatchedrecordsasunique.threads_per_job (default is 20)
  - cmx.server.batch.recalculate.block_size

    cmx.server.batch.batchunmerge.block_size

MDM Jobs Properties – Load Jobs

- **Console**
  - Threads for Batch Processing
    - hubConsole->ProcessServer (c_repos_cleanse_match_server.capacity)

- **cmxserver.properties**
  - cmx.server.batch.threads_per_job (Default is 10)
    - Number of threads that the MDM Hub uses to process the load, recalculate BVT, and revalidate batch jobs.
  - cmx.server.batch.load.block_size (Default is 250)
    - Number of records to process in each block for the load job.

- **cmxcleanse.properties**
  - None

MDM Jobs Properties – AutoMerge Jobs

- **Console**
  - Threads for Batch Processing
    - hubConsole->ProcessServer (c_repos_cleanse_match_server.capacity)
- **cmxserver.properties**
  - cmx.server.automerge.threads_per_job
    - Number of threads that the MDM Hub uses to process the automerge batch jobs. Default is 1.
  - cmx.server.automerge.block_size
    - Number of records to process in each block for the automerge job. The default is 250.
  - If “accept unmatached records as unique” is set , add the following to the cmxcleanse.properties. cmx.server.batch.acceptunmatchedrecordsasunique.threads_per_job=n, by default it uses 20 threads and might run into capacity issues

MDM Jobs - Unmerge / Recalculate BVT / Revalidation

- **cmxserver.properties**
  - cmx.server.batch.threads_per_job (Default is 10)
    - Number of threads that the MDM Hub uses to process the load, recalculate BVT, and revalidate batch jobs. Can be set for individual batch jobs by adding the following to cmxserver.properties
  - cmx.server.batch.recalculate.block_size. (Default is 250)
    - Number of records to process in each block for the recalculate BVT and revalidate jobs..
  - cmx.server.batch.batchunmerge.block_size
    - Custom flag to only dictate the block size for batch unmerge jobs
  - cmx.server.batch.unmerge.threads_per_job

MDM Smart Search

- Avoid enabling search on entities that have the same BO as the parent
- Keep indexed columns to a minimum required.
- Search is automatically enabled on lookup columns on the BE level, if search is enabled on the reference entity level
- Recommendations to reduce performance impact
  - In one ORS, total searchable columns should be under 100
  - One BE should not have more than 15 searchable configurations
  - Avoid configuring searchable columns at depth more than 4th/5th level of children
  - Index parent BO first, for e.g. lookup BO's, ideal to create batch group that will run the initial index jobs for all search enabled BO's
- Best to have a dedicated server for elasticSearch for high volume data, Heavy in Read/Write, SSD will help.
- Batch Indexing jobs also use cmx.server.batch.threads_per_job property for batch processing.
- Avoid indexing uniquely identifiable columns. e.g. PersistentID. Use queries for uniquely identifiable columns, index on the database has to be created.

MDM – Housekeeping

- High Resource Usage – monitor
  - CPU, monitor during each batch job to come up with optimal numbers for parallelism settings.
  - Memory Consumption
  - Disk space consumption (Hub and Process Server both)
    - cmx.server.datalayer.cleanse.working_files=DELETE (cmxcleanse.properties to remove temp files created during stage/match/tokenization)
- Job Execution Failure
- Job Execution SLA Violation
  - Completed job time exceeds threshold or recent average
- Data Governance Surveillance
  - Completed Job Matched Record Count Deviates from Recent Average
  - Too many rejects during stage job
  - Unusual high cluster of xref records

MDM Job repository tables

C_REPOS_JOB_CONTROL

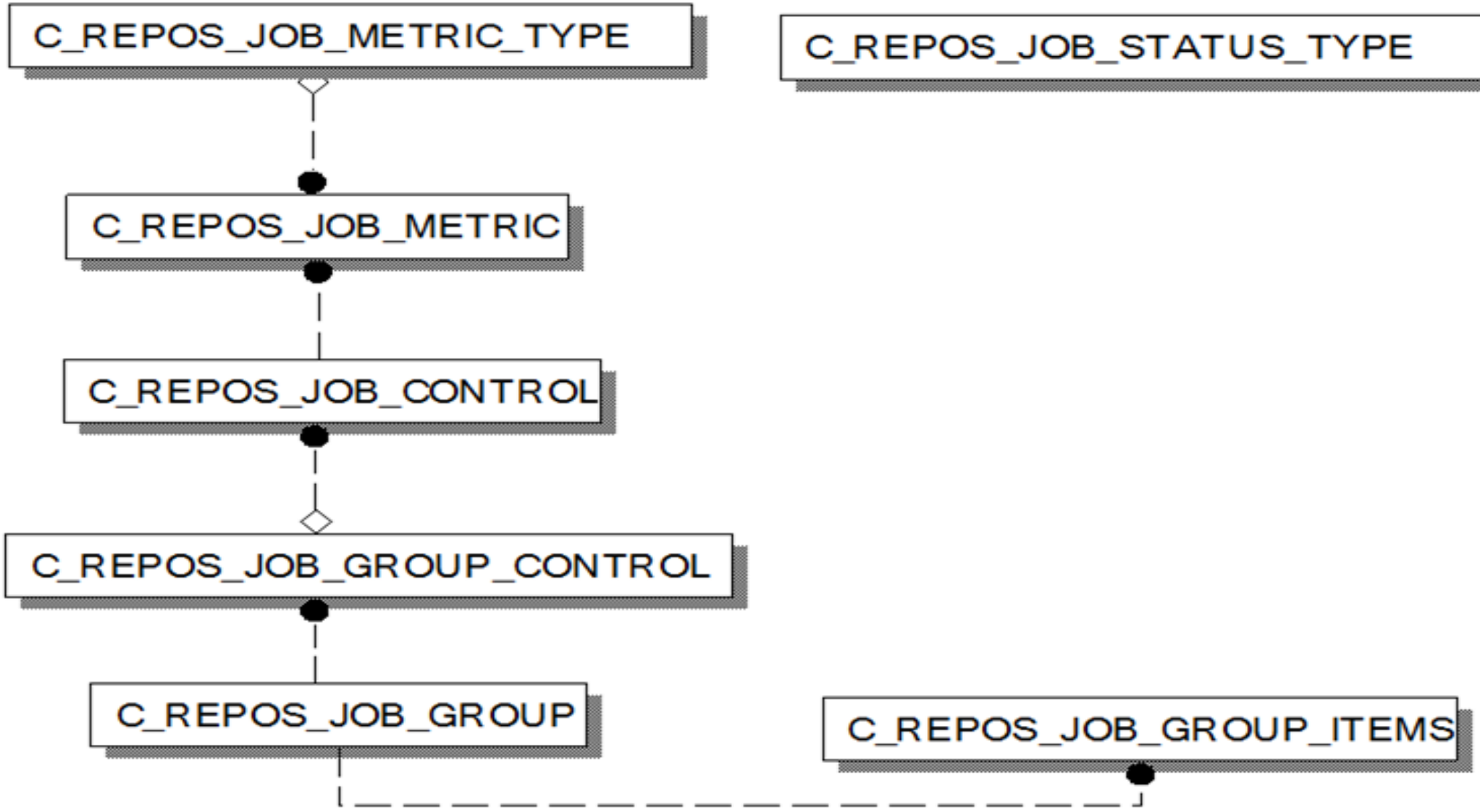
Name	Comment
ROWID_JOB	Primary key
ROWID_TABLE	ROWID of the table that owns this job item. To get the table details join to C_REPOS_TABLE on ROWID_TABLE
TABLE_DISPLAY_NAME	Table name that owns the job
SYSTEM_NAME	For Load/Stage jobs, this is the name of the system for the staging table. For all other processes, the system name is "Admin".
START_RUN_DATE	The date/time at which this batch process started
END_RUN_DATE	The date/time at which this batch process ended. Note: if RUN_STATUS is 3 or 4 (i.e. Incomplete), then END_RUN_DATE will be null
RUN_STATUS	Indicates the status of batch job. When a job starts, its RUN_STATUS is 2 (i.e. Processing). When it ends, RUN_STATUS will be updated to reflect success, warning or failure., See C_REPOS_JOB_STATUS_TYPE for details of the valid job status values
RETURN_CODE	The return code of the job. If the job completed successfully, then RETURN_CODE is 0
STATUS_MESSAGE	Message providing an overall explanation of the status of the batch. If the job returned an error or warning run status, then this column will contain the error or warning text
OBJECT_FUNCTION_TYPE_CODE	Defines the type of process that was executed (e.g. Load, Stage, Synchronize, etc). Join to C_REPOS_OBJ_FUNCTION_TYPE For further details
ROWID_JOB_GROUP_CONTROL	FK to C_REPOS_JOB_GROUP_CONTROL. ROWID_JOB_GROUP_CONTROL

MDM Job repository tables

C_REPOS_JOB_METRIC

Name	Comment
ROWID_JOB	FK to C_REPOS_JOB_CONTROL. ROWID_JOB
METRIC_TYPE_CODE	FK to C_REPOS_JOB_METRIC_TYPE. METRIC_TYPE_CODE
METRIC_VALUE	The value associated with a specific job metric. For example, this could be the number of records inserted, or the number of records updated - i.e. the number of records affected by a particular process or in a particular way within the batch job

MDM Job repository tables



MDM Job tables – Query / Output

```
SELECT
c_repos_job_control.ROWID_JOB,
c_repos_job_control.OBJECT_DESC,
c_repos_job_control.RUN_STATUS,
c_repos_job_control.STATUS_MESSAGE,
c_repos_job_control.TABLE_DISPLAY_NAME,
to_char(c_repos_job_control.START_RUN_DATE, 'MON-DD-YYYY HH24.MI.SS') start_date,
to_char(c_repos_job_control.END_RUN_DATE, 'MON-DD-YYYY HH24.MI.SS') end_date,
c_repos_job_control.OBJECT_FUNCTION_TYPE_CODE,
c_repos_job_metric_type.METRIC_TYPE_DESC, c_repos_job_metric.METRIC_VALUE
FROM
(c_repos_job_control
INNER JOIN c_repos_job_metric
ON c_repos_job_control.ROWID_JOB = c_repos_job_metric.ROWID_JOB)
INNER JOIN c_repos_job_metric_type
ON c_repos_job_metric.METRIC_TYPE_CODE =
c_repos_job_metric_type.METRIC_TYPE_CODE
WHERE (
START_RUN_DATE >= TRUNC(SYSDATE) - 150
)
ORDER BY START_RUN_DATE DESC, ROWID_JOB;
;
```

MDM – Miscellaneous

- Hub Console performance
  - jnlp.max-heap-size=512 – increase it to prevent outofmemory errors during Met promotion
  - cmx.appserver.console.mode=HTTPS or HTTP for hubConsole to use http instead of EJB
  - cmx.appserver.jnlp.protocol=HTTPS

MDM – References

- Infinispan setup on multinode environment
  https://kb.informatica.com/faq/7/Pages/22/568046.aspx?myk=568046
- Recommended infinispan configurations
  https://kb.informatica.com/faq/7/Pages/18/509572.aspx?myk=509572
- What is DMAT and how does it work
  https://kb.informatica.com/solution/16/Pages/90740.aspx?myk=90740
- MDM performance Foundation Guide
  https://kb.informatica.com/whitepapers/4/Pages/1/323799.aspx?myk=323799
- unique constraint violation during tokenization
  https://kb.informatica.com/solution/23/Pages/70/573020.aspx?myk=573020
- MDM configuration guide https://docs.informatica.com/master-data-management/multidomain-mdm/10-3-hotfix-1/configuration-guide/introduction.html

Thank you

Informatica™