Multidomain MDM 10.4

Smart Search

Prashant Gupta

Informatica™

Housekeeping Tips

- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A 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 and Success Portal. 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.

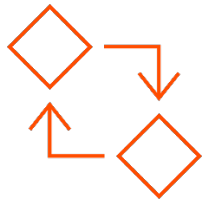
Success Portal

https://success.informatica.com

Learn. Adopt. Succeed.

Bootstrap product trial experience

Enriched Onboarding experience



FREE Product Learning Paths and weekly Expert sessions

Informatica Concierge with Chatbot integrations

Tailored training and content recommendations

Informatica

Safe Harbor

The information being provided today is for informational purposes only. The development, release, and timing of any Informatica product or functionality described today remain at the sole discretion of Informatica and should not be relied upon in making a purchasing decision.

Statements made today are based on currently available information, which is subject to change. Such statements should not be relied upon as a representation, warranty or commitment to deliver specific products or functionality in the future.

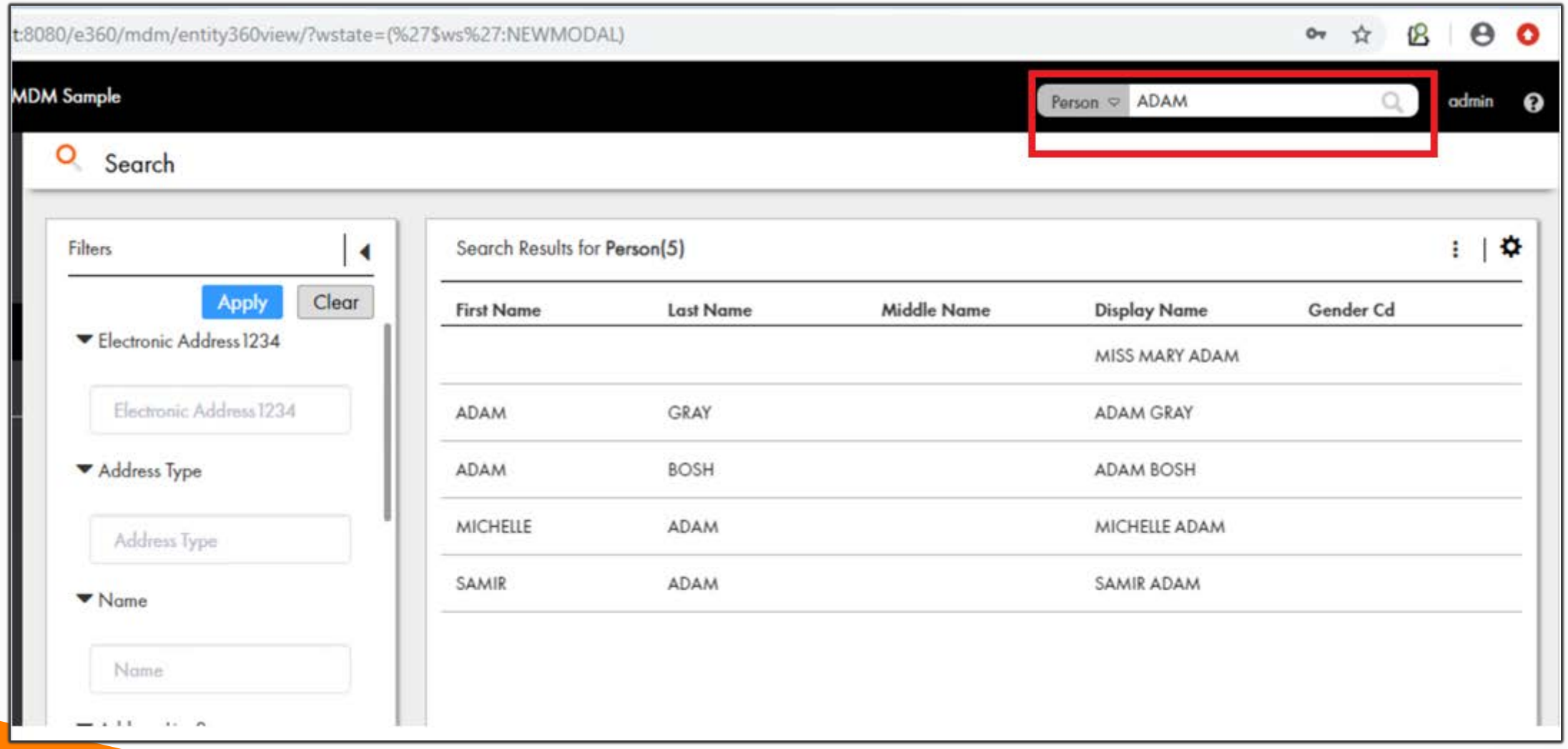
Agenda

- Overview
- SS Architecture
- Configuration of Smart Search
- Indexing Job Vs Load Job
- BE Search API’s
- Troubleshooting SS outside MDM via REST Calls
- Elastic Search Methodology
- MDM 10.4 Features
- Performance & Best Practices
- Known Limitations

Overview

- Smart Search is the search done in Entity 360 UI based on Business Entity.
- It was introduced starting MDM 10.x release.
- Business Entity service provide three different ways to search for entity records:
  - Match – SSA based search
  - Query – DB based search
  - Search – Google like text search
- Provides a Google-like text search engine for text data
- It searches within a specific searchable BE.
- SS request uses the indexes to search for the matching records.
- Filters can be applied to identify more relevant results.

Smart Search in E360 UI



Informatica

Migration Strategy for MDM 10.4

Upgrade

- From SOLR :

  Need to perform fresh installation & Configuration of Elastic Search. Indices will generate again.

- From Elastic Search :

  Upgrade the MDM Hub to 10.4, will not lose their ES setup from the previous release. No change for them.

Fresh Installation

- Download the ES package from web & extract
- Need to explicitly create the analysis and lang folder mentioned below.
- Reach out to Informatica GCS to get the below highlighted files
- Once all file in place, then we are good to Configure Smart Search

.

/config
--analysis(folder)
--stopwords.txt
--synonyms.txt
--lang (folder)
--mapping-FoldToASCII.txt
--stopwords_ja.txt

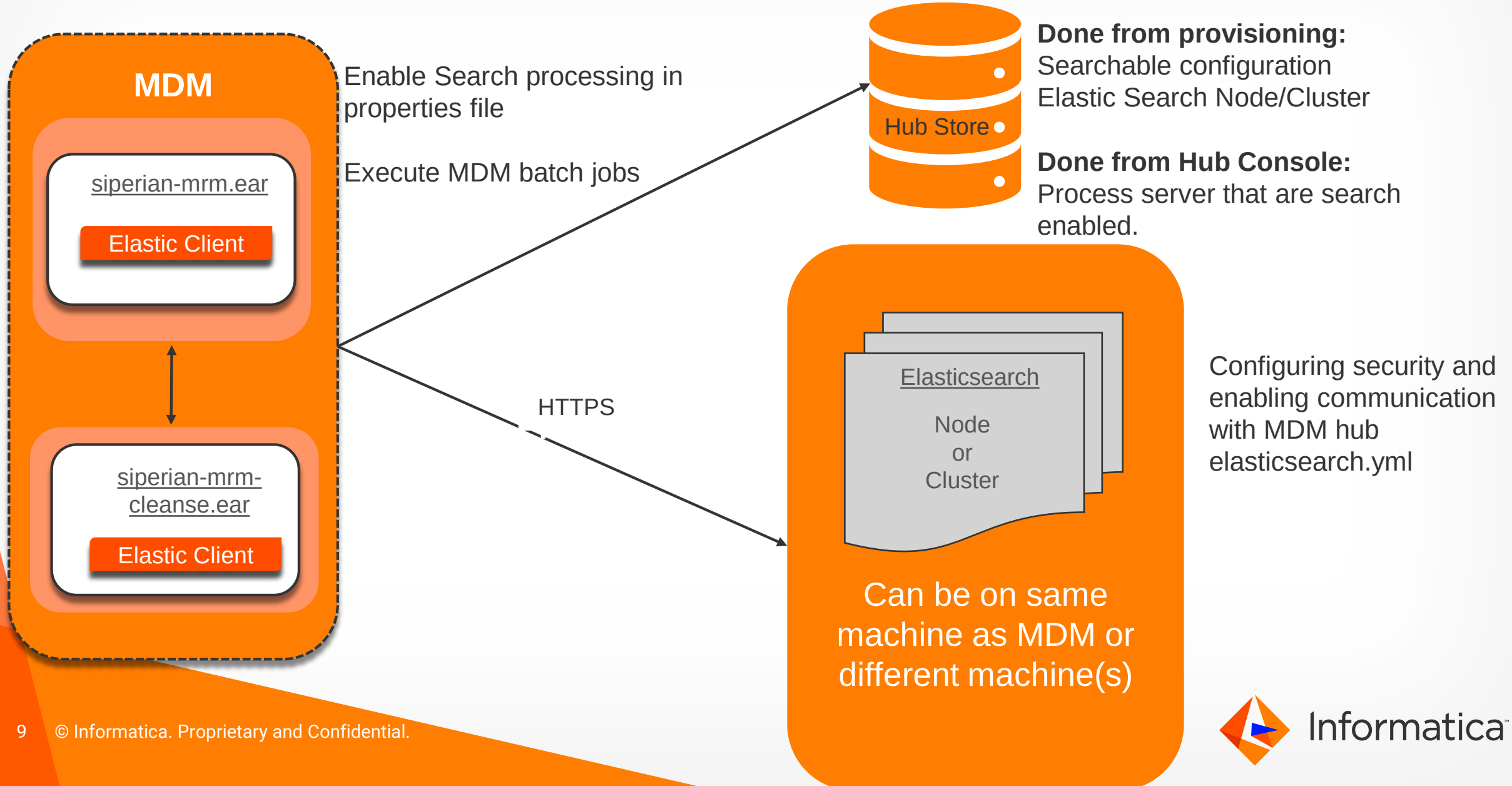
*When you perform a search, MDM can ignore common words such as "and", "an", and "is".

*when you search for "William", the search result can include the synonyms "Will" and "Willy".

*To use the default Elasticsearch analyzers for languages such as Chinese, Japanese, and Korean

Informatica

Elastic Search - High Level Architecture :



Install & Configure Smart Search

- Configure the elasticsearch.yml file ([MDMHubServer]\thirdparty\elasticsearch-6.2.3\config)
- Configure SS in cmxserver and cmxcleanse properties.
- Enable Search Processing on Process Servers in MDM Hub Console.
- Register the ES Node/Cluster in Provisioning Tools
- Restart the ES server followed by MDM server.
- Execute Indexing Job

For more information, detailed steps are present in KB 564677

SS Properties

Cmxserver & Cmxcleanse:

- **cmx.ss.enabled :** Indicates whether to enable search. In a new installation, the default is true. When upgrading, if this property is set, the value remains set to the pre-upgrade value.
- **cmx.ss.engine :** Required if you want to use the Elasticsearch engine for search. Manually add the property and set to es.
- **ssl.keyStore, ssl.keyStore.password, ssl.trustStore, ssl.trustStore.password :** Required if you use the HTTPS port of the application server to configure the Hub Server. Manually add the property. Absolute path and file name of the KeyStore file.
- **ex.max.conn.per.host :** This property is the maximum number of connections for a given host in the pool. Default is 100. MDM maintains a pool of http connections specific to Elastic Search communication.
- **ex.max.threads :** Sets the maximum number of threads that you want the Apache asynchronous non-blocking receiver to use for each node in the Elasticsearch cluster. Default is 1.

Only Cmxserver :

- **es.index.refresh.interval :** Sets the interval, in seconds, for Elasticsearch to commit the changes to the data after an Initially Index Smart Search Data batch job is run. The data is available for search after this time interval. Default is 30 seconds.
- **cmx.server.batch.smartsearch.initial.block_size :** Indicates the block size during indexing job.
- **cmx.ss.dirtyIndex.disable=true :** This property allows the Batch Viewer tool to start and prevents the warning icon that indicates an outdated index from being displayed in the Batch Viewer tool.

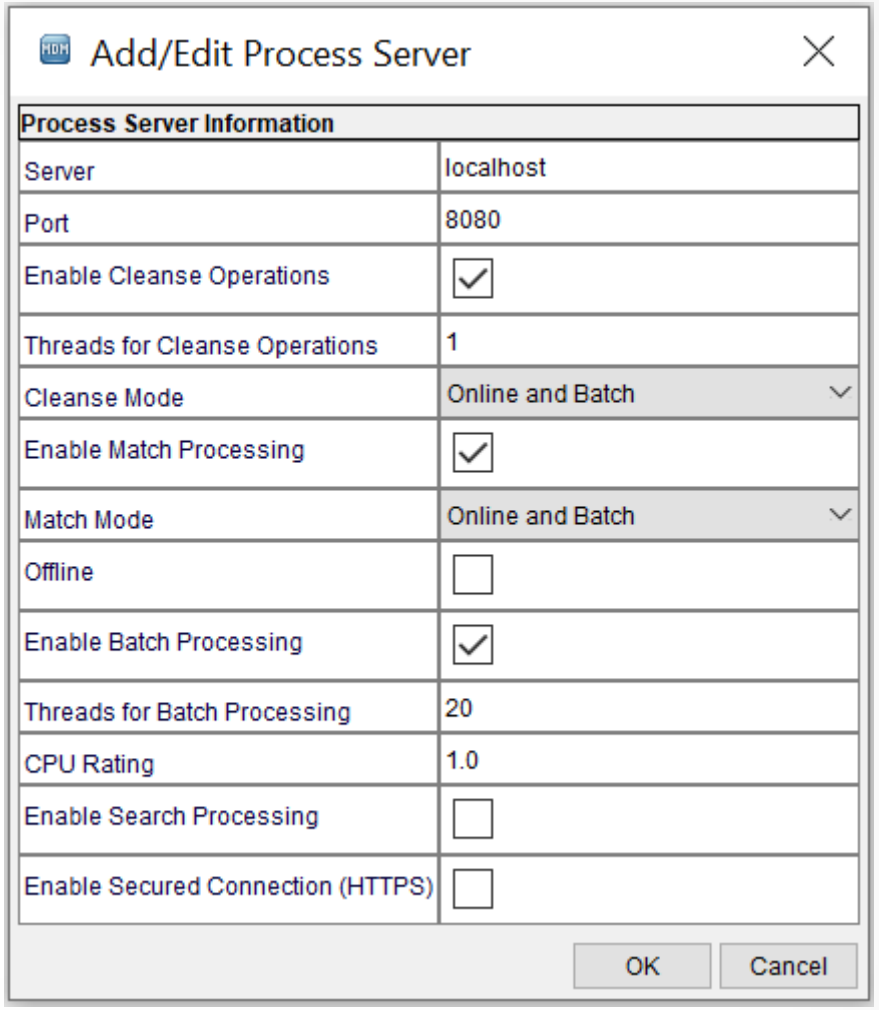
JVM Parameter :

- **-Dmdm.es.index.client.ttl :** If there are any connection issues between the AppServer and Elasticsearch processes, this value will set the timeout as to how long the AppServer process should wait before getting a response. This property is in seconds and should be set in the AppServer startup properties to 86400 which translates to 24 hours.

Informatica

Configuring Process Servers

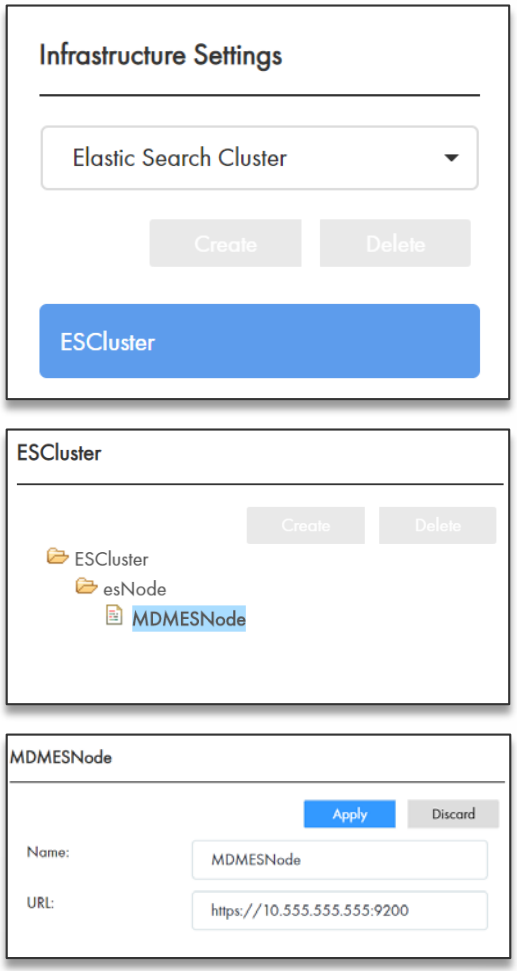
- Process Servers act as clients for Search Requests
- No indexing is done on Process Servers
- You will need to register at least one Process Server as search enabled
- You can register more than one process server



Informatica

Configuring Node(s) in Provisioning

- After installation, start the elastic search process.
- Once the process is available register each ES Node under infrastructure settings.
- Ensure that ES Node name and the HTTPS URL is reachable.
- If you have more than one ES node, register each one individually.
- Save and publish changes. Validate ORS



Informatica

Configuring Searchable BE in Provisioning

Searchable – Enables search for that field.

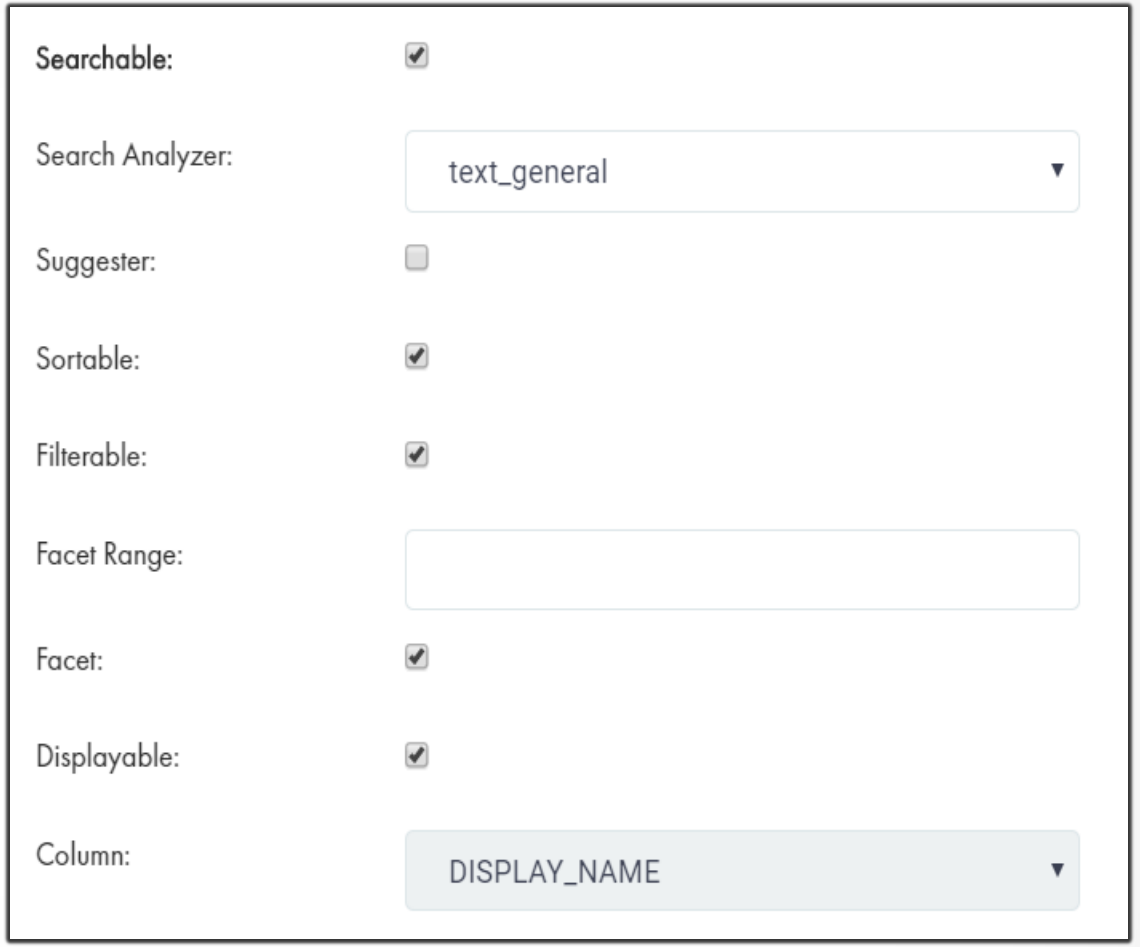
Suggester – Type-ahead search for Searchable field.

Do not enable suggester for sensitive data.

Filterable – Enables Filtering on a Field.

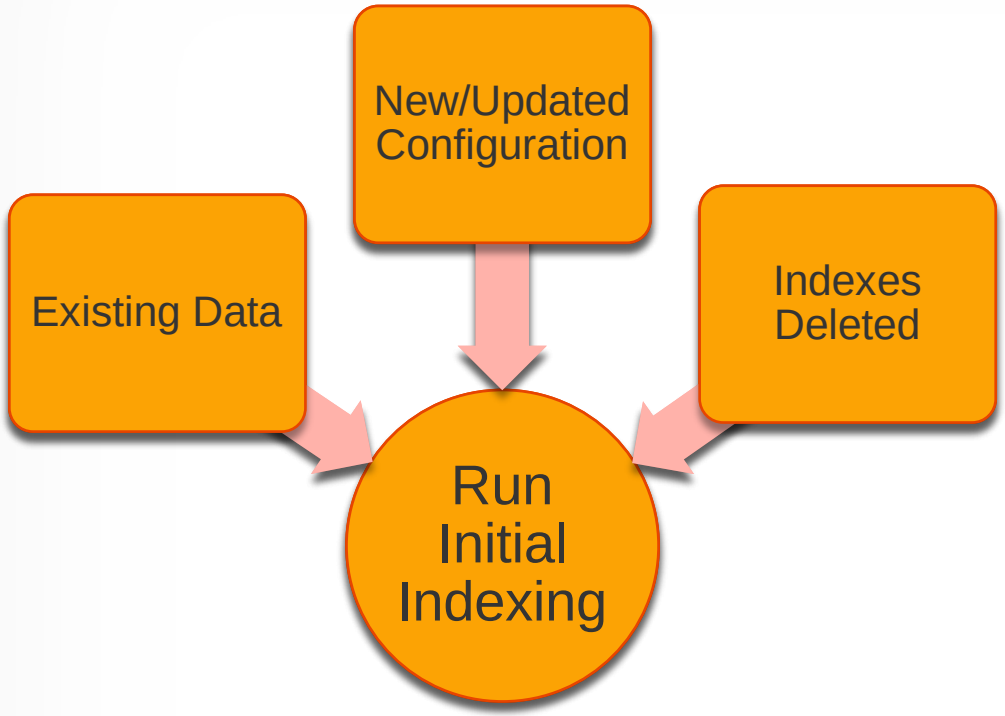
Search Analyzer – Specifies the custom search analyzer for the field

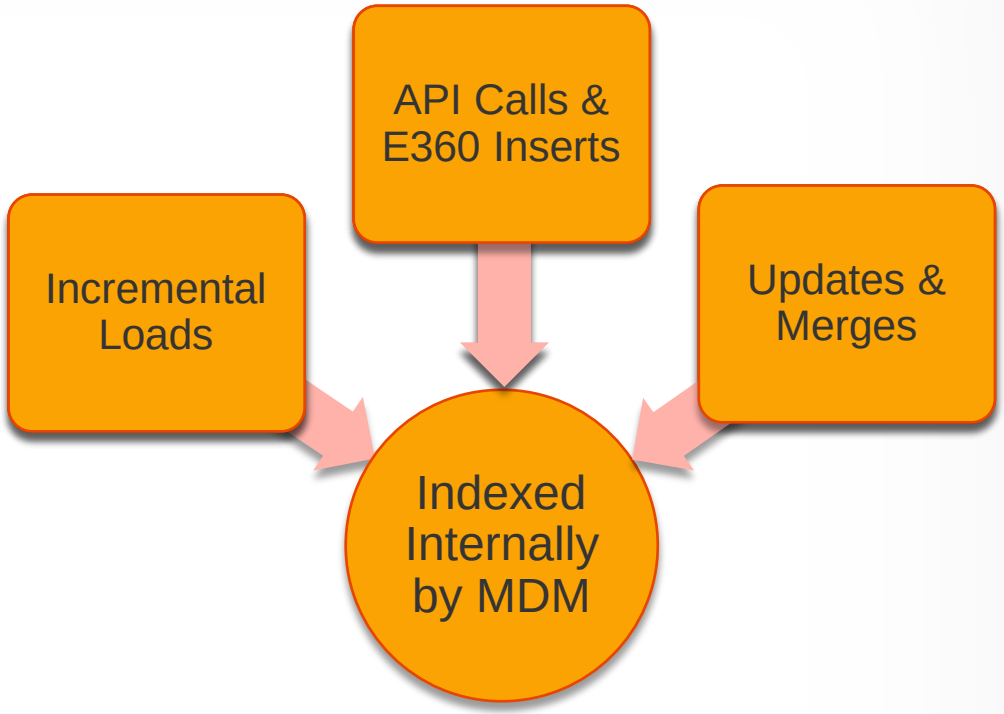
Facets – Grouping of similar data.



Informatica

Indexing Job or Load Job?

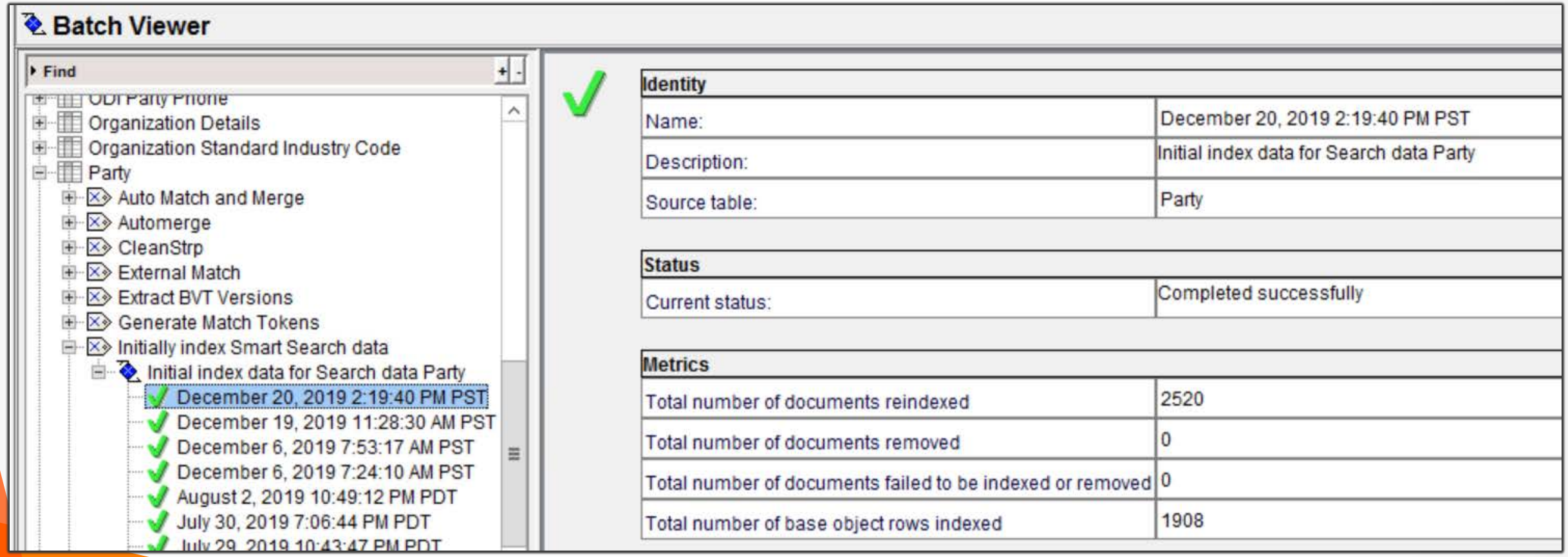




Informatica

Execute Smart Search Indexing Job

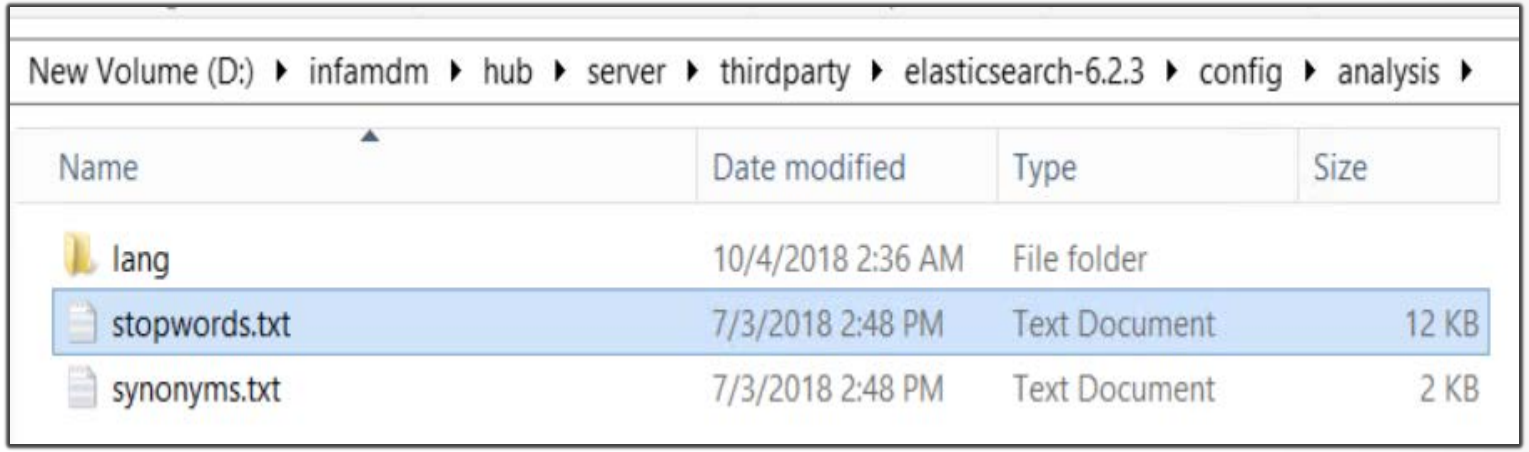
- Indexing job is being executed from Batch Viewer in MDM Hub Console.
- Need to configure at least one process server marked as ‘Enable Search Processing’



Informatica

Stopwords

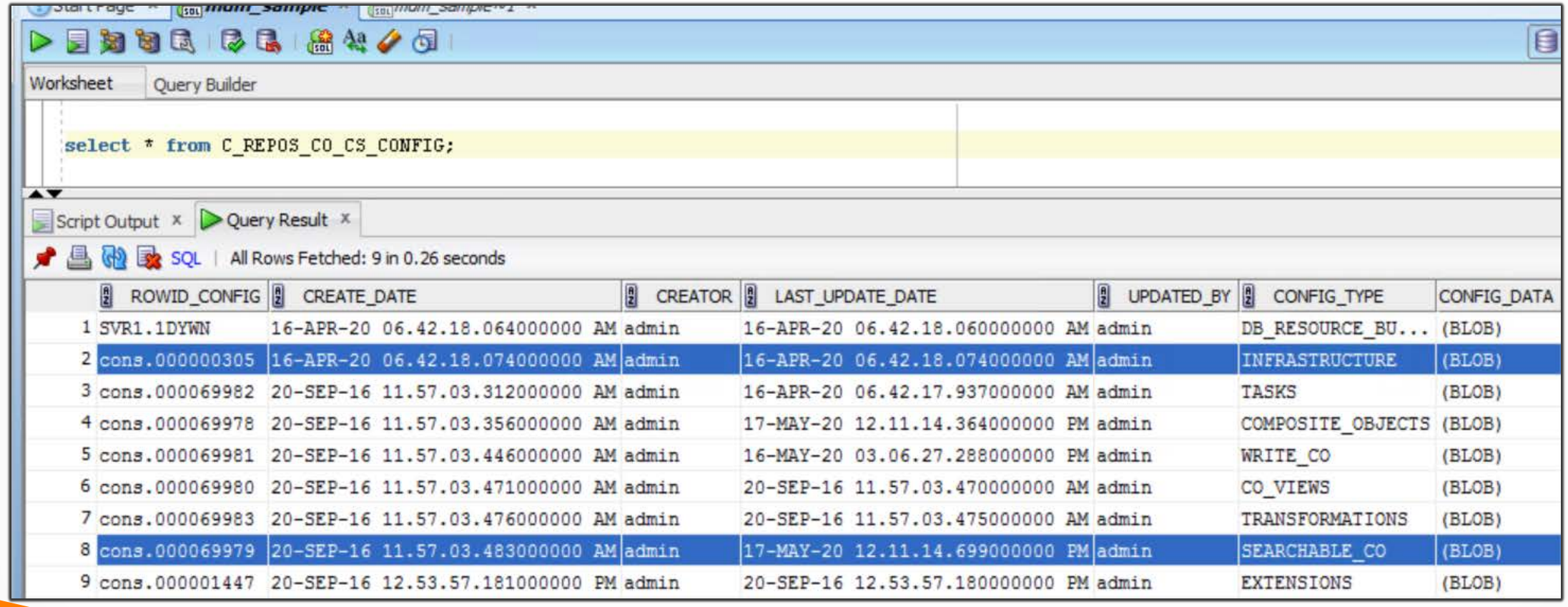
- Common list of words that search engines filter out after processing.
- It improves indexing performance, i.e. doesn't index the common words present in stopwords.txt.
- Can use the default list of common words to ignore in a search or even can customize the list.



- Configure Stop words :
  a. Use a text editor to open the stopwords.txt file in the following location: <elasticsearch installation directory>/config/analysis
  b. Edit and save the stopwords.txt file.
  c. If data is indexed before you edit the stopwords.txt file, manually delete the indexes, restart Elasticsearch, restart MDM and then, reindex the data.

Smart Search Configuration Table

- Searchable columns configurations are stored as SEARCHABLE_CO xml in C_REPOS_CO_CS_CONFIG table
- ES node URL configurations are stored as INFRASTRUCTURE xml in C_REPOS_CO_CS_CONFIG table



	ROWID_CONFIG	CREATE_DATE	CREATOR	LAST_UPDATE_DATE	UPDATED_BY	CONFIG_TYPE	CONFIG_DATA
1	SVR1.1DYWN	16-APR-20 06.42.18.064000000 AM	admin	16-APR-20 06.42.18.060000000 AM	admin	DB_RESOURCE_BU...	(BLOB)
2	cons.000000305	16-APR-20 06.42.18.074000000 AM	admin	16-APR-20 06.42.18.074000000 AM	admin	INFRASTRUCTURE	(BLOB)
3	cons.000069982	20-SEP-16 11.57.03.312000000 AM	admin	16-APR-20 06.42.17.937000000 AM	admin	TASKS	(BLOB)
4	cons.000069978	20-SEP-16 11.57.03.356000000 AM	admin	17-MAY-20 12.11.14.364000000 PM	admin	COMPOSITE_OBJECTS	(BLOB)
5	cons.000069981	20-SEP-16 11.57.03.446000000 AM	admin	16-MAY-20 03.06.27.288000000 PM	admin	WRITE_CO	(BLOB)
6	cons.000069980	20-SEP-16 11.57.03.471000000 AM	admin	20-SEP-16 11.57.03.470000000 AM	admin	CO_VIEWS	(BLOB)
7	cons.000069983	20-SEP-16 11.57.03.476000000 AM	admin	20-SEP-16 11.57.03.475000000 AM	admin	TRANSFORMATIONS	(BLOB)
8	cons.000069979	20-SEP-16 11.57.03.483000000 AM	admin	17-MAY-20 12.11.14.699000000 PM	admin	SEARCHABLE_CO	(BLOB)
9	cons.000001447	20-SEP-16 12.53.57.181000000 PM	admin	20-SEP-16 12.53.57.180000000 PM	admin	EXTENSIONS	(BLOB)

Informatica

BE Search API’s

- Search call to a Smart Search Enabled Entity can be made using a Generic Query or a Fielded Query.

  -> In Simple Search Query, the search term will be looked up or searched for in all the Business Entity fields, while In fielded query, the search is only carried out on that particular business entity field

- Generic Search or a Simple Search can be made using 'q’:

  -> http://host:port/cmx/cs/<orsId>/Person?q=John

- Fielded Search can be made using the 'fq’:

  -> http://host:port/cmx/cs/<orsId>/Person?fq=displayName=‘JOHN’

- Exact Queries Search on BE field Parameters.

  -> http://host:port/cmx/cs/<orsId>/Person?action=query&filter=firstName=JIM

- Filtered Query Search : Adds a filter to the search results

  -> http://host:port/cmx/cs/<orsId>/Person?q=John&filters=Addresses.Address.cityName=Washington

- Facets : provide an option of group-based filtering

  -> http://host:port/cmx/cs/<orsId>/Person?fq=firstName=John&facets=displayName

Informatica

Contd..

- To get the data from child entities, use the depth option. Note the search is always done on all searchable fields (including child). The depth option is to include the data from the child in the response.

  -> http://host:port/cmx/cs/<orsId>/Person?q=John&depth=5

- To sort the data on a certain field

  -> http://host:port/cmx/cs/<orsId>/Person?q=John&order=lastName

- To fetch the higher/lower number of records Or restricts the number of records (parent).

  -> http://host:port/cmx/cs/orsId/Person?q=John&recordsToReturn=3

- Range Filter

  -> http://host:port/cmx/cs/<orsId>/Person?fq=displayName=STEVE&filters=age=[35,45]

- To suppress the links in results

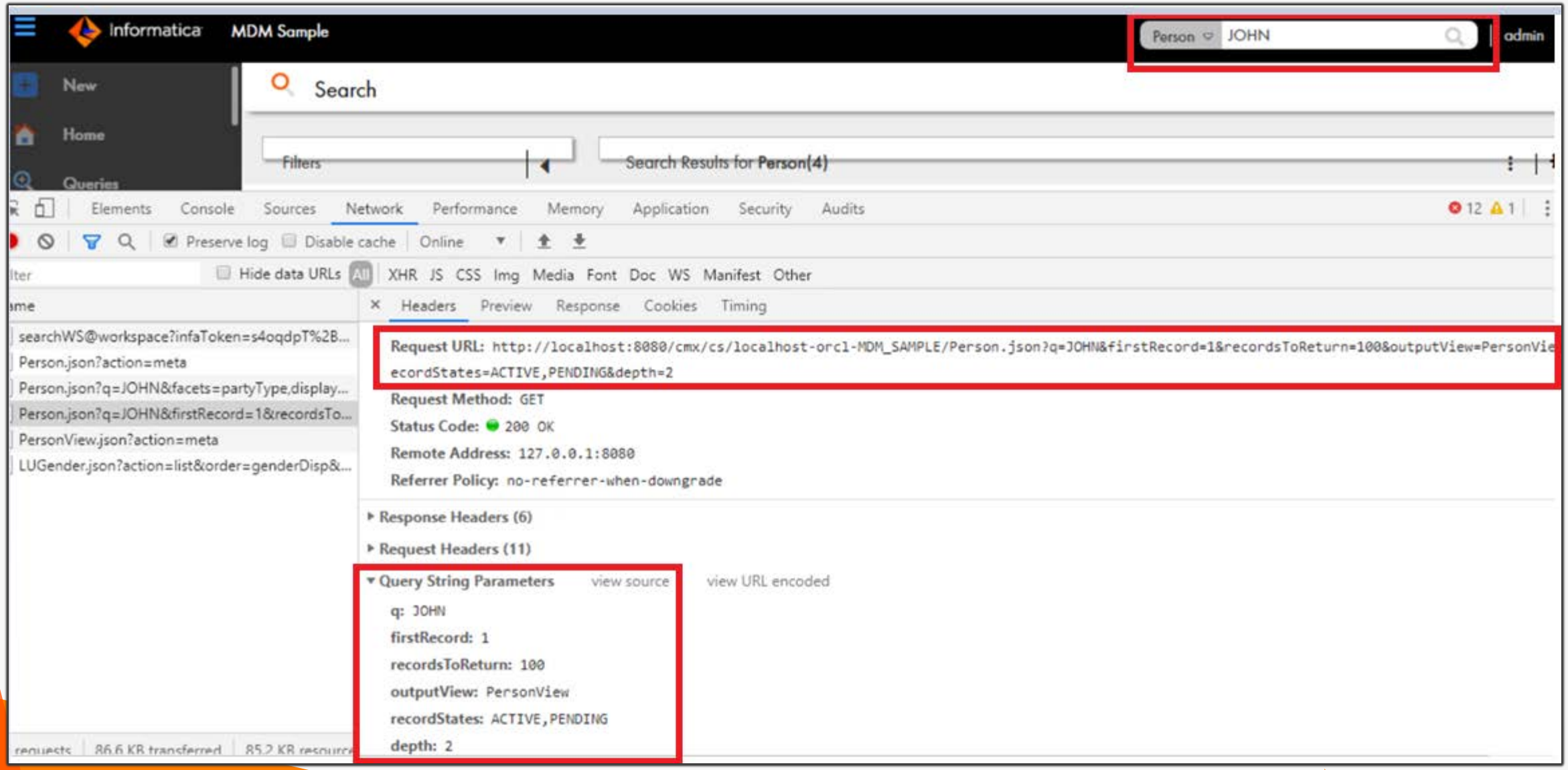
  -> http://host:port/cmx/cs/<orsId>/Person?q=John&depth=5&suppressLinks=true

- defaultPageSize : Changes the number of records for each returned page.

  -> http://host:port/cmx/cs/< orsId >/Person/1415?depth=2&defaultPageSize=20

Informatica

E360 Search Call – BE View



Informatica

Troubleshoot SS outside MDM – via REST calls

- To configure ES rest call in web browser, refer KB 566891

- List all the indices :

  https://host:9200/_cat/indices?v=

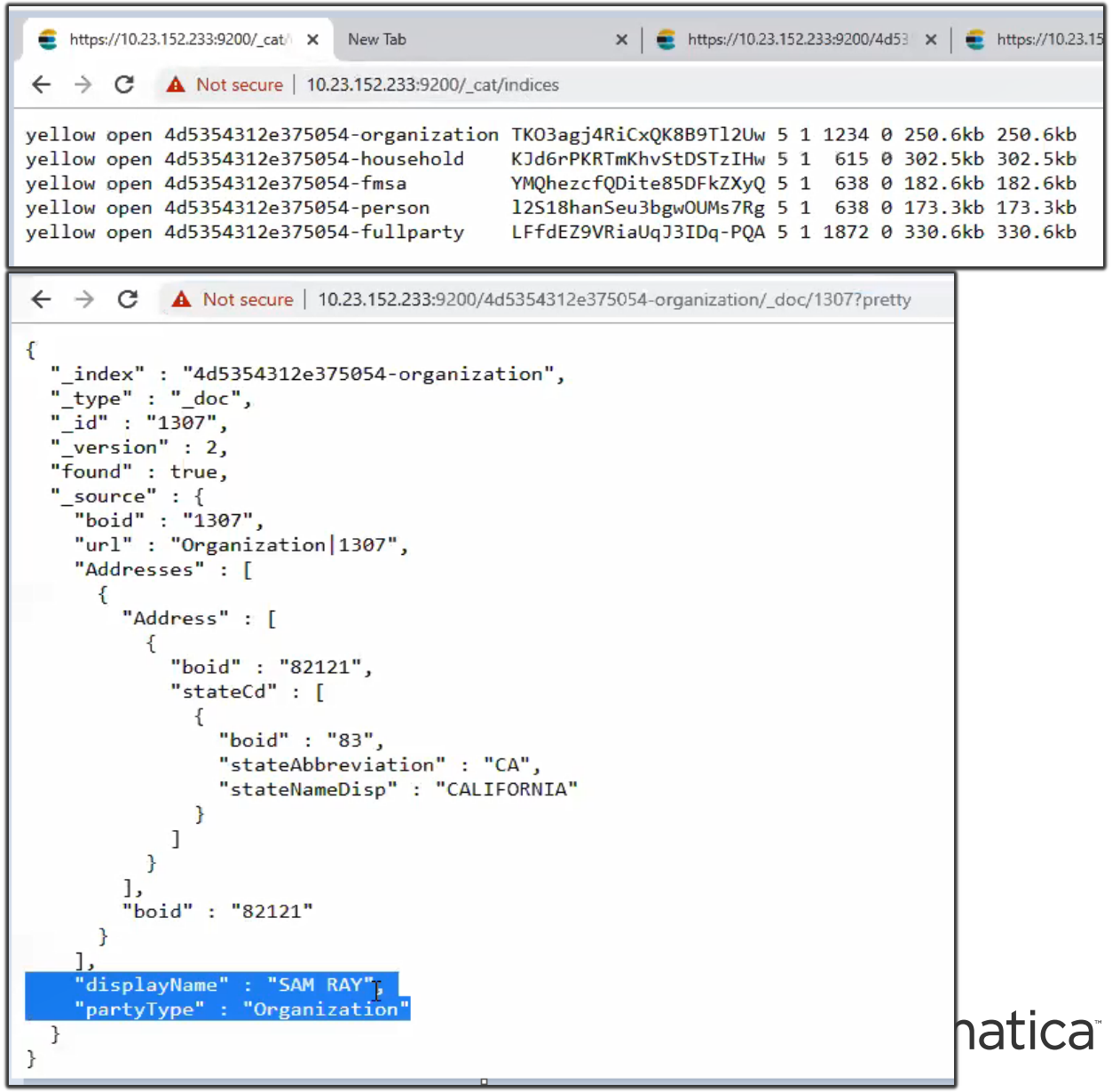
This will list down all the BE indices along with their health, status & docs count.

- To check for all the columns associated into Specific Indices:

  https://<hostname>:9200/4d5354312e375054-organization/_mapping/_doc/?pretty=

- Check the BE data record

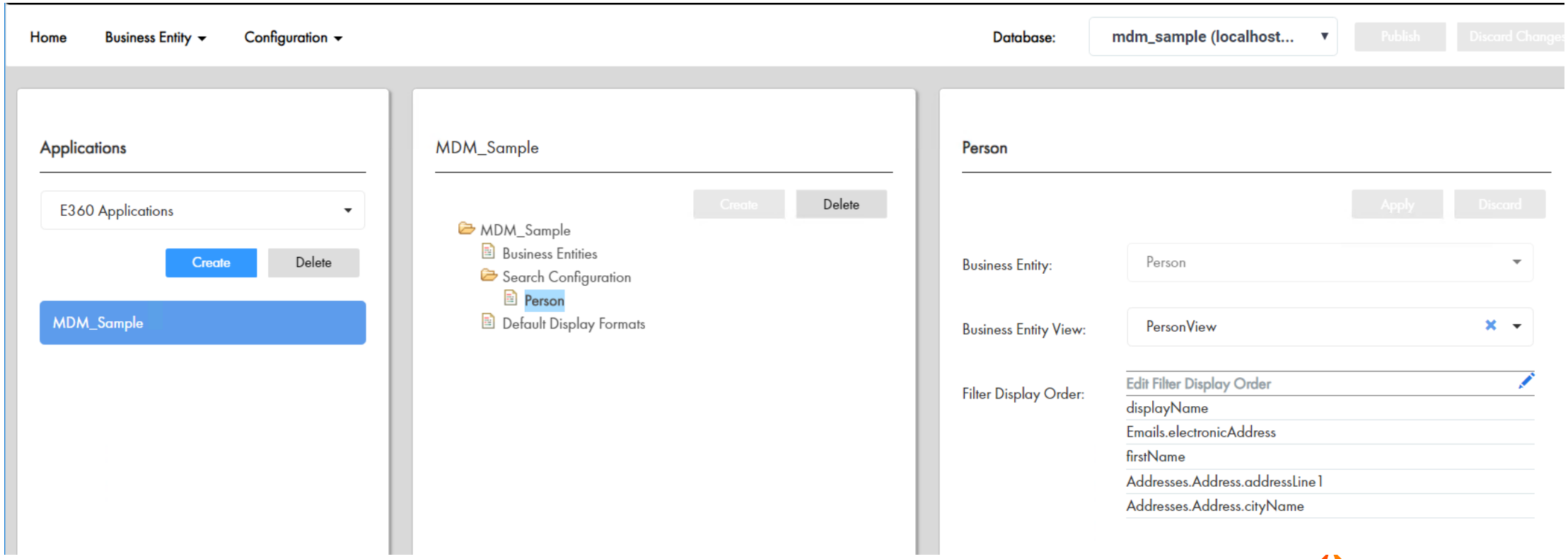
  https://<hostname>:9200/4d5354312e375054-organization/_doc/<BO RowID>?pretty=



MDM 10.4 Features

❑ Re Ordering of Filters column in Smart Search

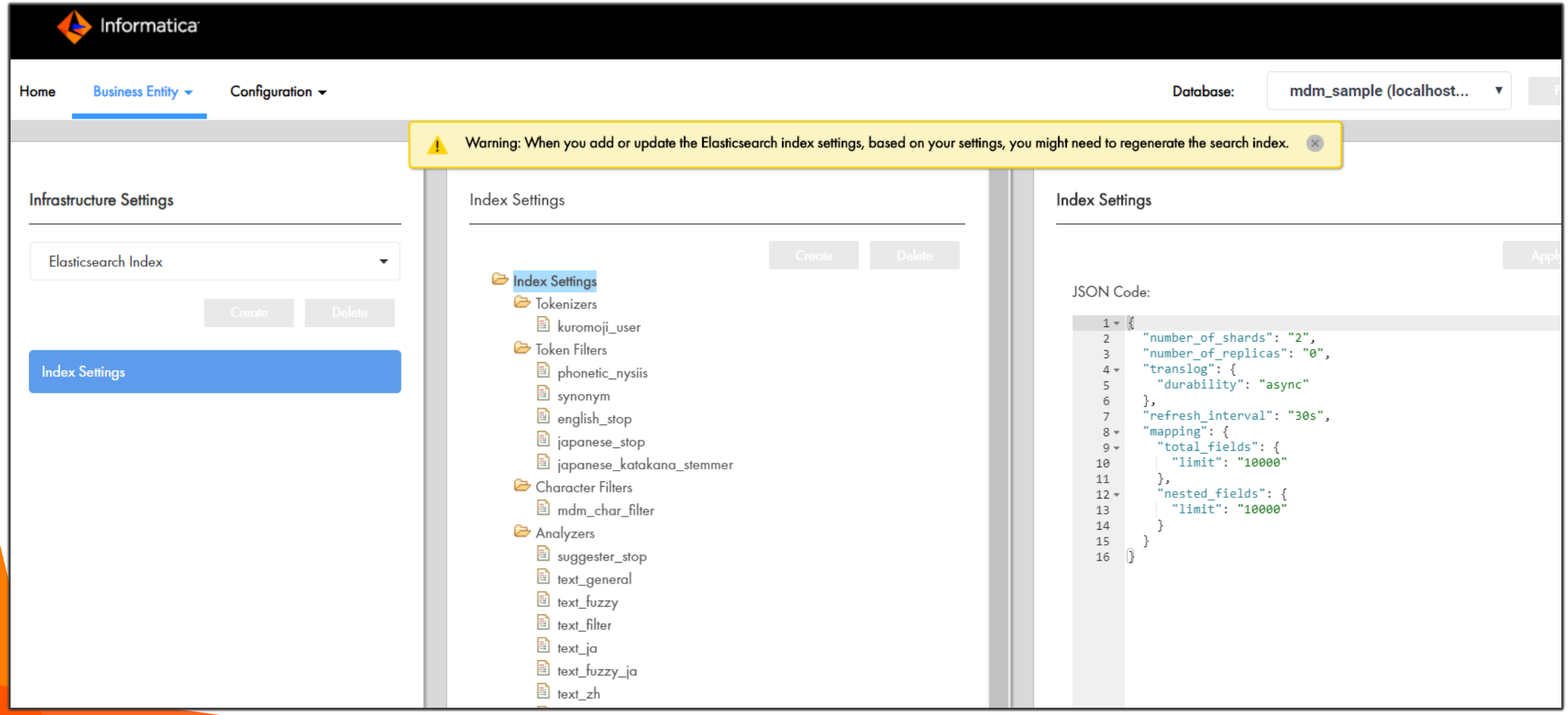
- Can use the Provisioning tool to configure search filters and the order in which these appear in the Data Director Search page
- Configuration gets stored in C_REPOS_COMPONENT_INSTANCE table.



Informatica

❑ Externalize the ElasticSearch Analyzer Chain

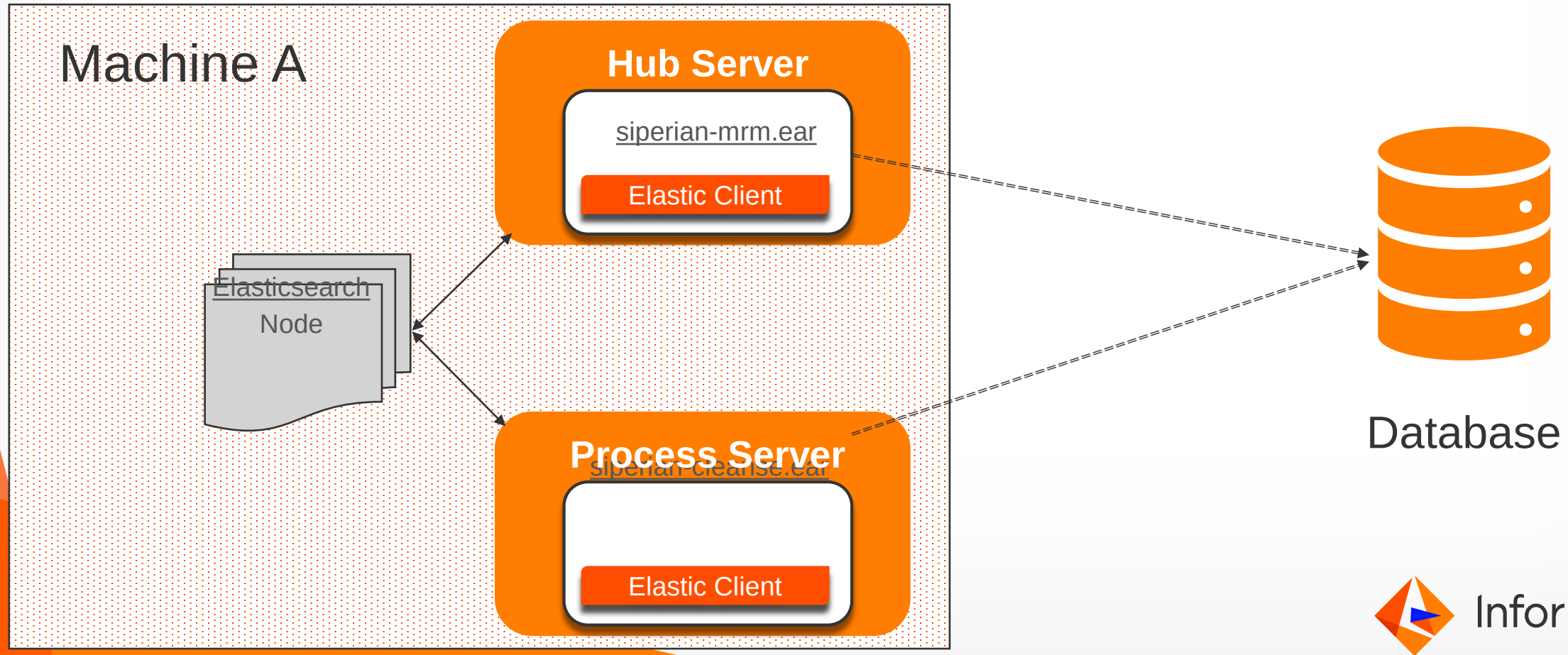
- Can create custom Elasticsearch index settings from the Provisioning tool. You can configure custom analyzers, tokenizers, token filters, and character filters for the Elasticsearch index.



ES – Different Architecture Set Up

Option 1 – Co-located with Hub/Process Servers

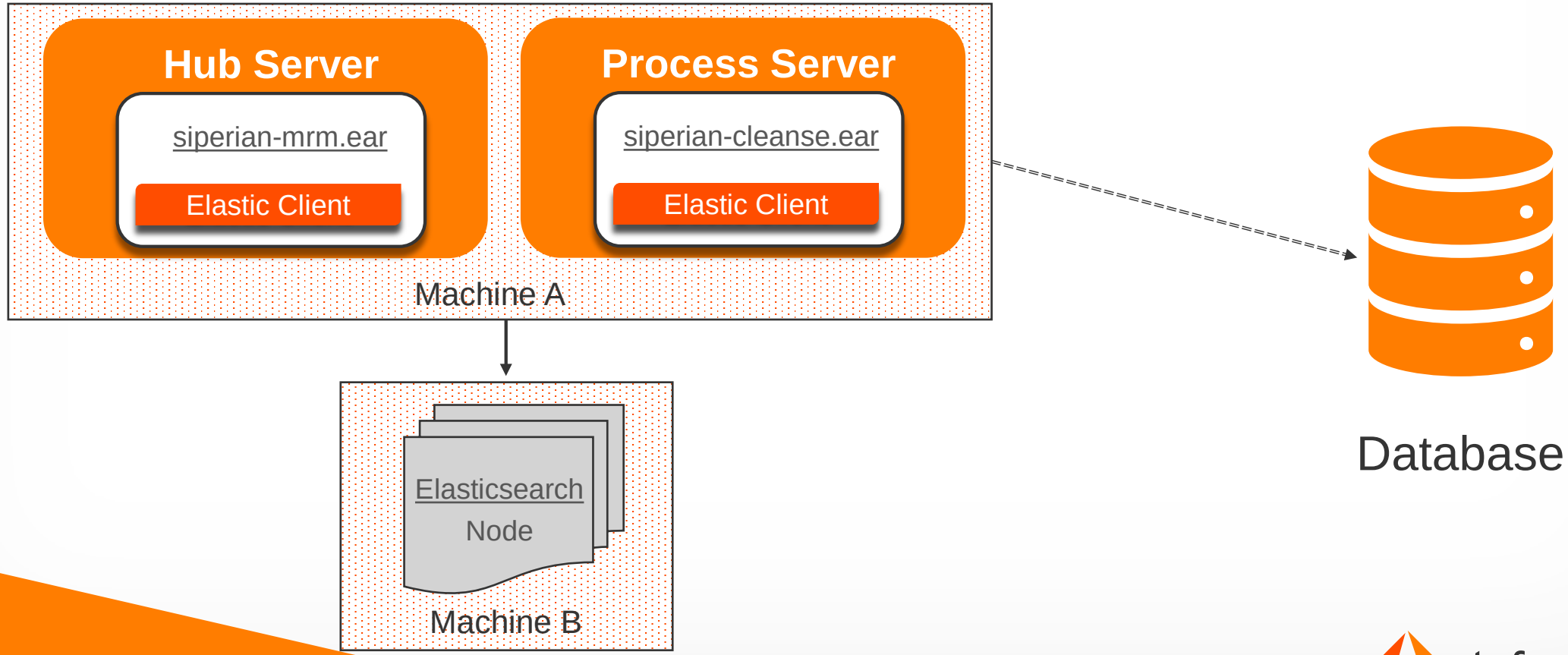
- Smaller Implementations – Fewer BEs and lower volumes.
- Good for smaller incremental loads and when E360 is used for searches.



Informatica

Option 2 – Dedicated ES Node

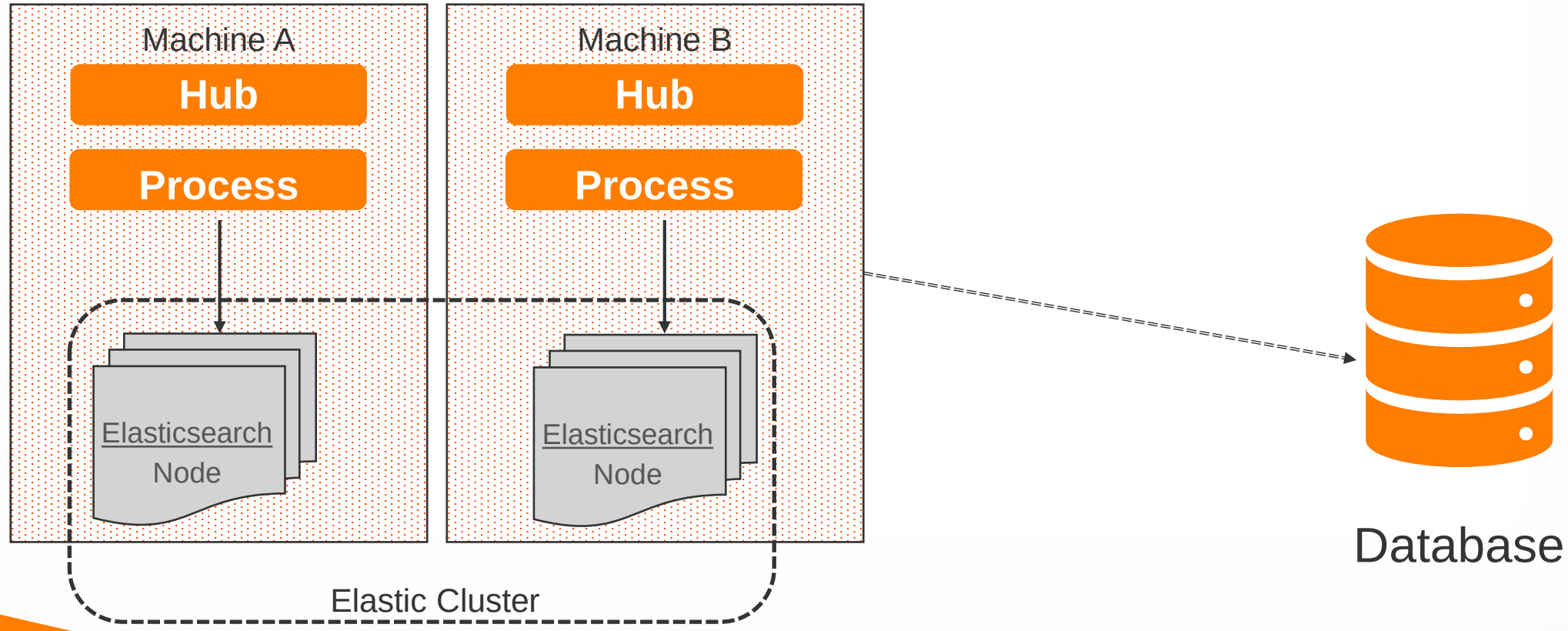
- Fewer BEs with high volume OR More BEs with lower volume.
- Daily load with some API activity for Create/Update/Search.



Informatica

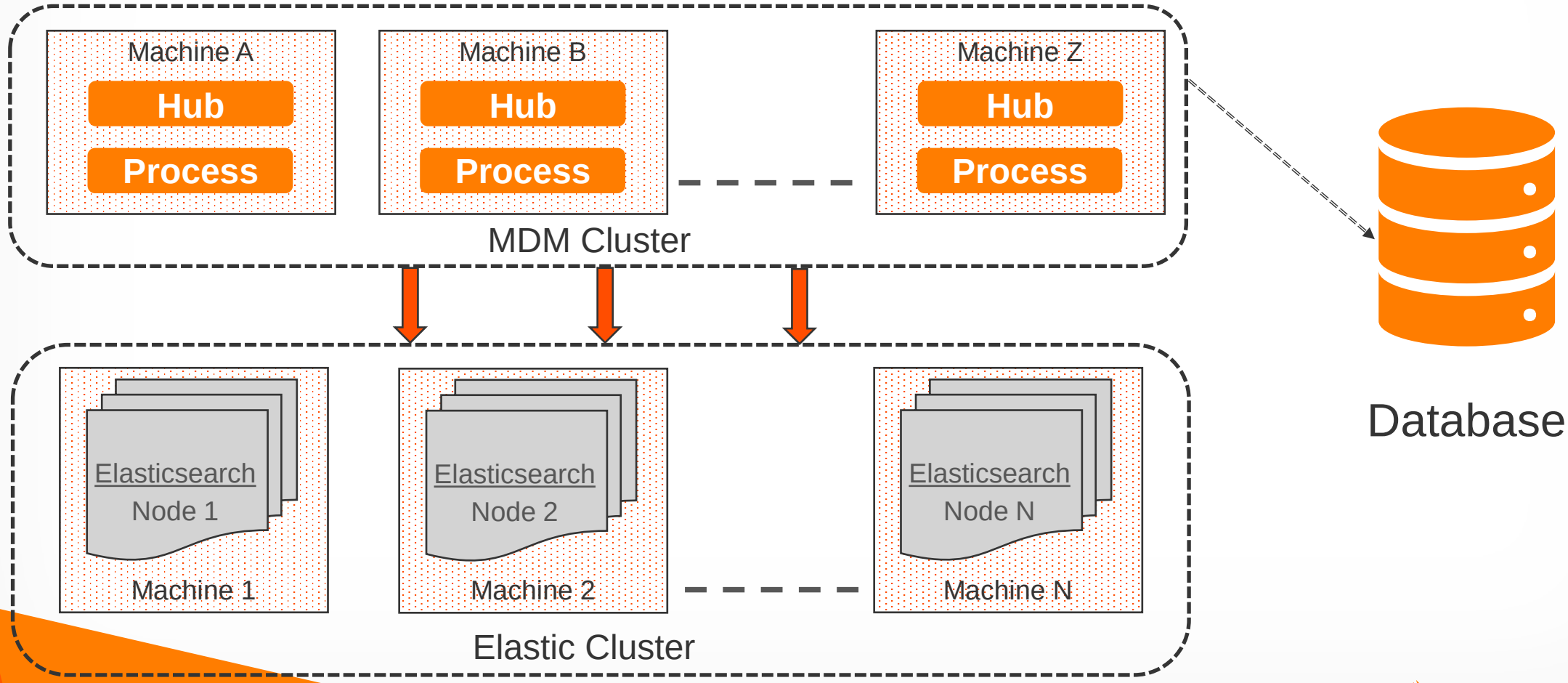
Option 3 – Co-located Cluster in a multi-node MDM

- Multiple BEs with medium volume data
- Good amount of daily loads, E360 search users, API activity for creates/updates/searches.



Option 4 – Clustered ES in multi-node MDM

- High Volume Systems with Frequent Loads, Heavy API usage and E360 Searches.
- N should always be an Odd number.



Informatica

Performance & Best Practices

- Run indexing on the Look Up tables first.
- Need to execute indexing job on all respective BO which are participating in the Searchable BE.
- CleanTable API does not delete the indexes, we need to manually delete indices & regenerate them.
  - Delete ES REST API
  - Simulate a change to searchable column through provisioning.
- Avoid running Indexing jobs in parallel.
- Avoid running 2 load jobs in parallel if they have searchable fields.
- Do not use too many search fields i.e. limit searchable fields on BE (parent/child).
  - There are limitations on the number of columns in a document in ES (Limit of nested fields (default = 50) has been reached)
  - There are limitations on the amount of data that is returned by ES REST calls. The more the number of indexed columns, the more the chances of hitting this limit (104857600 bytes of data as a response to a REST call).

Contd..

- Remove the searchable checkbox on non used/extra BE's. (e.g. Customer360 Accelerator)
  - This slows down the performance of jobs that need to delete the indexes.
- Limit the use of facets.
- Facet field should be chosen on a field with low entropy, that is, fields with a low set of unique values.
- Retrieve records till a usable depth.
- Increase auto commit property. (es.index.refresh.interval)
- There should be a monitoring tool which keep an eye on the disk space of ES server.
  - ES will mark all indexes read only if there is less than 15% free disk space
  - {"index.blocks.read_only_allow_delete": false}'
- To increase the mmapfs count which is used by Elasticsearch to store its indices. (KB# 589503)

Known Limitations

- Case sensitive mismatch in UI & DB values when configured Facets on Searchable columns
- Search results doesn’t get displayed when ‘#’ is used in Search.
- No way to perform search on multiple BE’s in a single call.
- Provision to enable role based restriction to access specific BEs sharing same root BO for Search, Queries and BES
- *For Auto merge of child records, it performs GET (instead of MultiGet) for each record – which leads to job failure
- *Auto merge job doesn't obsolete redundant ES indices, resulting in duplicate search results

* Fixed Limitation

Thank You

Informatica™