Informatica MDM Match best practices

Anuvinda Kulkarni

Lead Support Engineer, MDM GCS Team

Informatica™

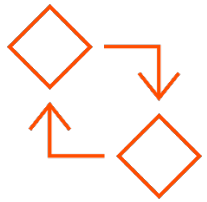
Housekeeping Tips

- 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.

Feature Rich Success Portal

Bootstrap trial and POC Customers

Enriched Customer Onboarding experience



Product Learning Paths and Weekly Expert Sessions

Informatica Concierge with Chatbot integrations

Tailored training and content recommendations

Informatica

More Information

Success Portal

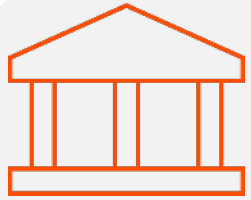
https://success.informatica.com

Communities & Support

https://network.informatica.com

Documentation

https://docs.informatica.com



University

https://www.informatica.com/in/services-and-training/informatica-university.html

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

- Introduction to MDM matching
- Walk through an example of match rules setup
- Match rules setup and tuning phases
  - Phase 1: Data discovery and analysis
  - Phase 2: Define Fuzzy Match Key, Key Width, Match Paths, Match Columns
  - Phase 3: Setup match rules: do’s and don'ts
  - Phase 4: A dry run of the match job using draft rules; review match results
  - Phase 5: Tune match rules with exact columns
  - Phase 6: Review final match results
- Tuning Process Server and Base Object properties
- Tuning cmxcleanse.properties
- Tuning the database
- Q&A

Introduction to MDM matching

- The match process helps consolidate records coming from multiple sources. There are 2 ways to do so:
  - Batch
  - API (SOAP -> searchMatch, BES REST -> action=match)
- Two types of matching:
  - Fuzzy
  - Exact

Key Level & Search Level

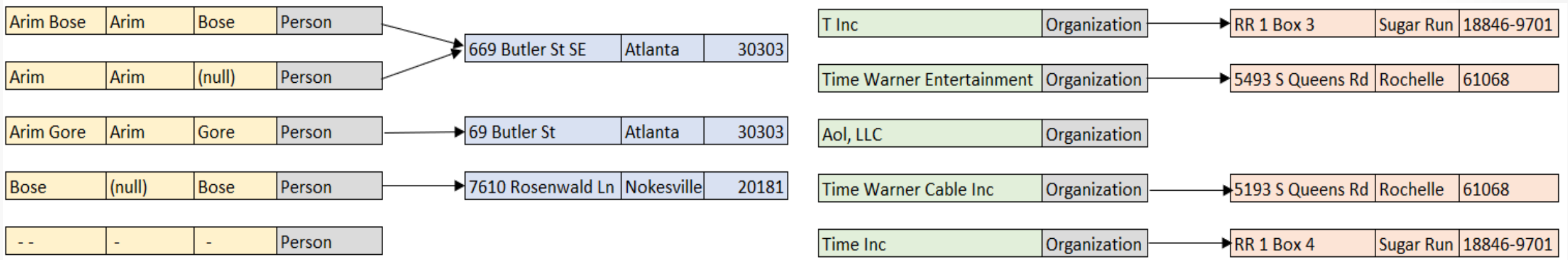
Optional match ruleset filter & Exact match columns

Match Purpose & Match Levels

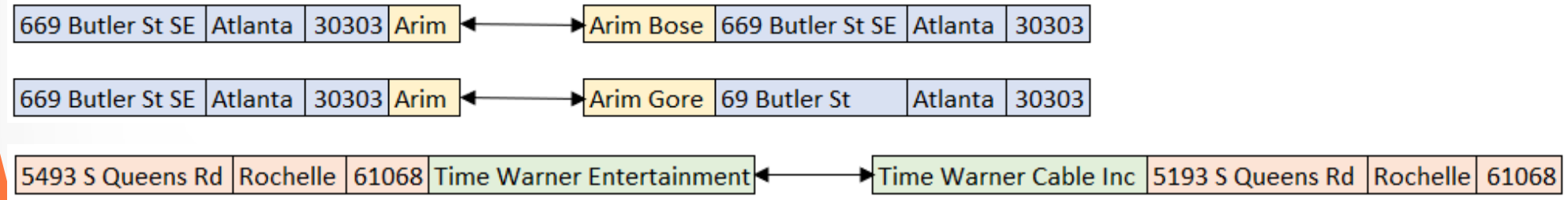
Informatica

Walk through an example of match rules setup

Consider the following data model:

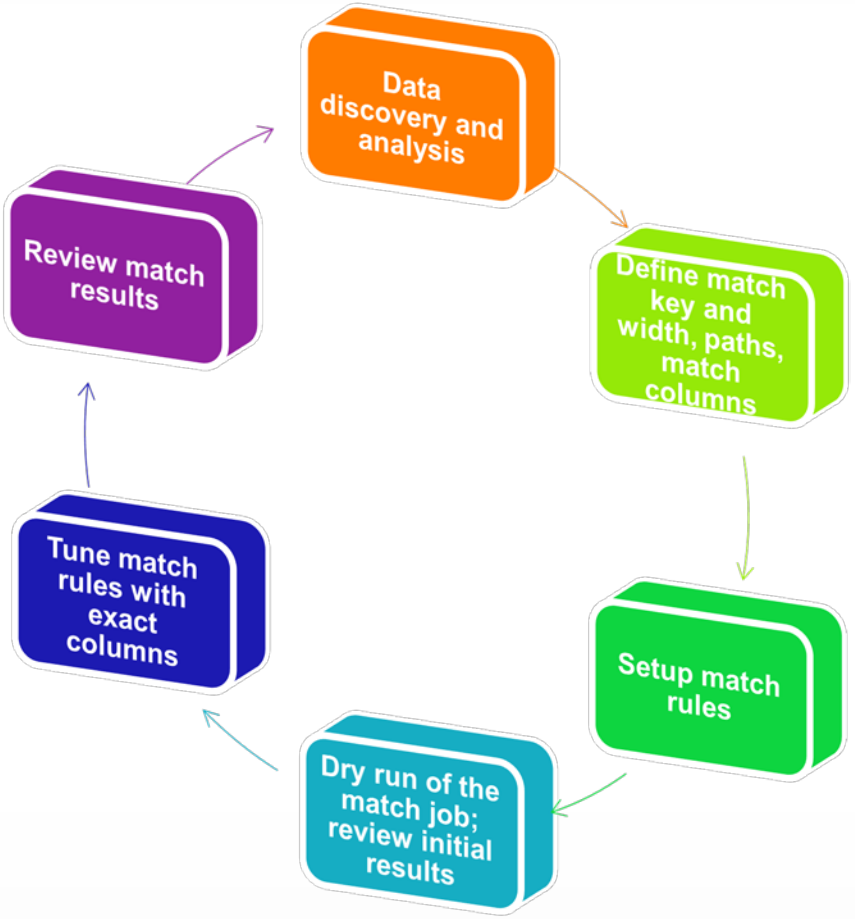


Expected match results:



Informatica

Match rules setup and tuning phases



Informatica

Phase 1: Data discovery and analysis

- **Auditing**
  - Get a reasonable-sized sample of data that best represents real or production-like data
  - Understand what needs to be considered for matching
  - Identify fields that will contribute to the match process, including Fuzzy Match Key
- **Quality and Profiling**
  - Ensure data completeness (e.g. Person records have both First Name and Last Name)
  - Ensure data accuracy (e.g. gender field has only gender values)
  - Use tools like Informatica Data Profiler, pattern analysis (SQL queries)
- **Standardization**
  - Standardize/format data as much as possible (e.g. Junior to JR, case & trim for exact fields)
  - Avoid non-ASCII characters
  - Use data quality tools such as MDM cleanse functions or Informatica Data Quality
  - Use an address cleansing tool to standardize and clean addresses

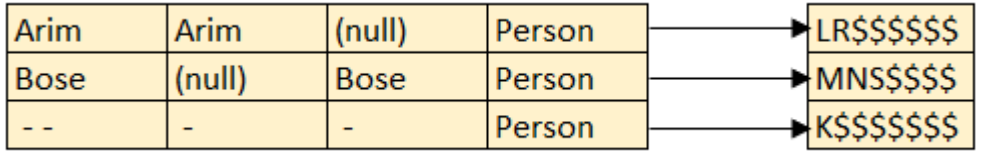
Phase 2. Define Fuzzy Match Key

- Any one of the following can be defined. Multiple fuzzy match keys are not supported

Fuzzy Match Key	Usage
Person Name	Data contains only Individuals
Organization Name	Data contains only Organizations, or if data contains both Individuals and Organizations
Address_Part1	Data has addresses that need to be consolidated

`

- First Name-only or Last Name-only fuzzy key (Person Name) can cause high number of candidates causing performance impact
- NULLs in fuzzy match key column produce null keys (K$$$$$$$ under SSA_KEY). They are potential candidates for each other
- Initials in First or Last Name can cause high number of candidates
- Irrelevant “noise words” in fuzzy match key column produce null keys



Informatica

Phase 2. Define Key Width

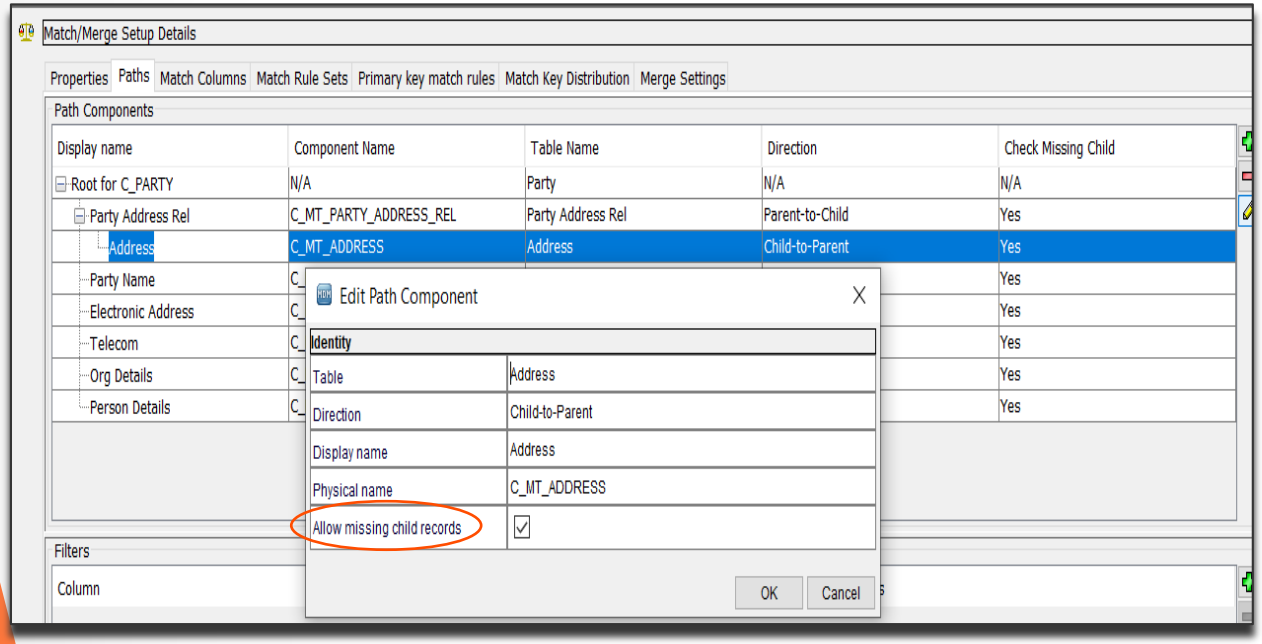
- How big is the dataset?
- How important is match quality VS performance?
- Wider key has higher chance of finding a search candidate, but it will lower the overall performance

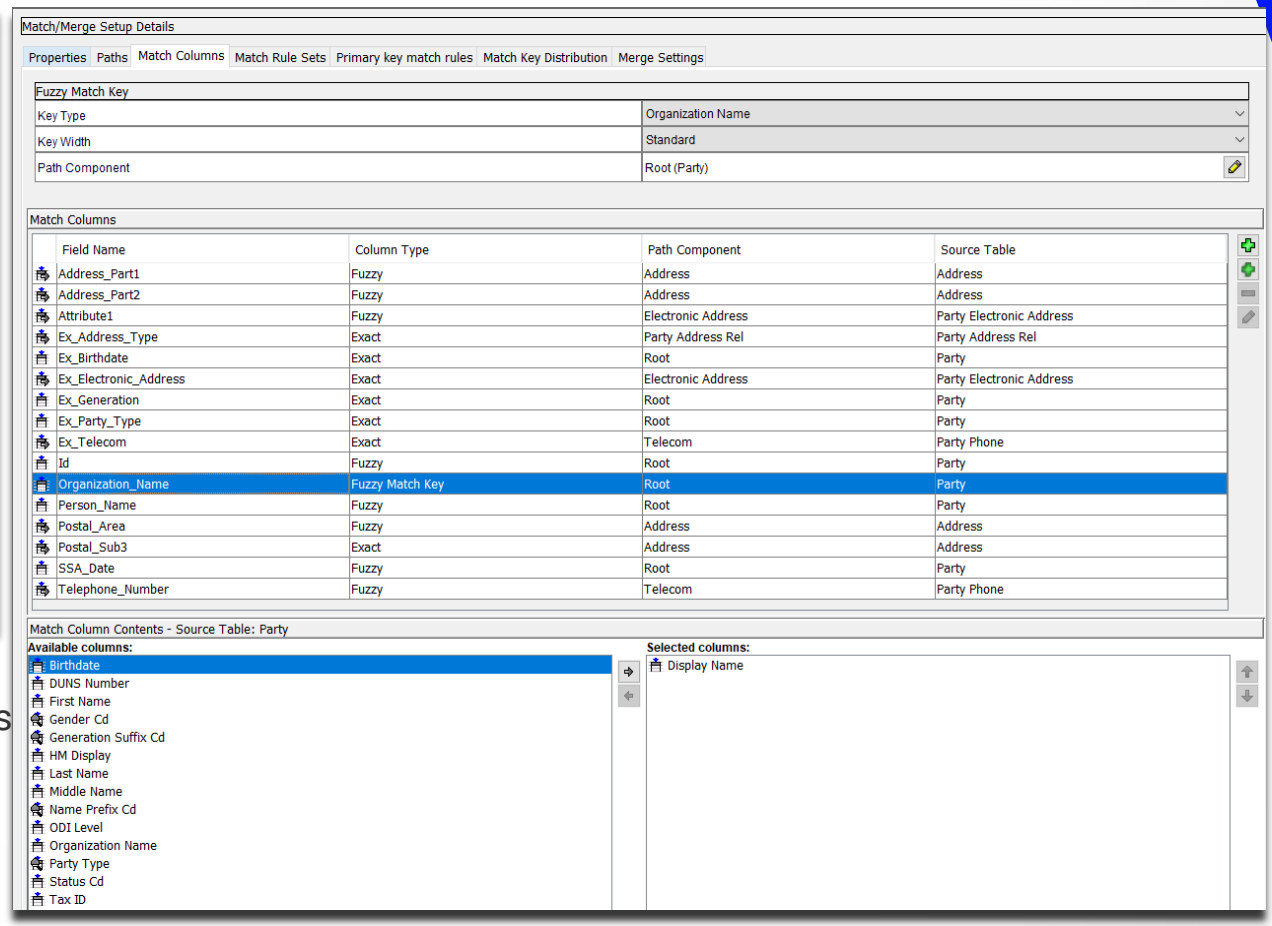
- **Limited** - Tradeoff between match quality and disk space. May cause fewer match candidates but faster searches.
  Use if disk space is limited or if data volume is extremely large
- **Preferred** - Single key per BO record. Might result in fewer match candidates
- **Standard** - Most appropriate; balances reliability and space usage
- **Extended** - Might result in more match candidates at the cost of longer processing time to generate keys. Works best:
  - Data set not extremely large
  - Data set not complete
  - Sufficient resources are available (disk space)
- For e.g., the SSA keys for names '**ASHLEY ROSENBERG**' and '**ASHLEY ROSEN BERG**' fall within the same SSA range for each of the width types, so they are possible match candidates for each other

Person	Limited	Standard	Extended	Preferred
ASHLEY ROSENBERG	KIT$AITA XGT>CJT$	XGT>CJT$ KIT$AITA	XGT>CJT$ KIT$AITA KIT-S$$- XGT>CJSU	XGT>CJT$
ASHLEY ROSEN BERG	KIT$AITA XB*Z>$/I LC*Z>$TS	LC*Z>$TS LCTV>K*Y XB*Z>$/I XB/IUK*Y KIT$AITA KIT$$VQQ KIT-SVQQ XGT>CJT-	LC*Z>$TS LCTV>K*Y XB*Z>$/I XB/IUK*Y KIT$AITA KIT$$VQQ KIT-SVQQ XGT>CJT- LC*Z/CTU	LC*Z>$TS

Informatica

Phase 2. Define Match Paths and Match Columns





- Enable ‘Allow missing child records’ helps matching on parent records that do not have child records in the child base object.

Informatica

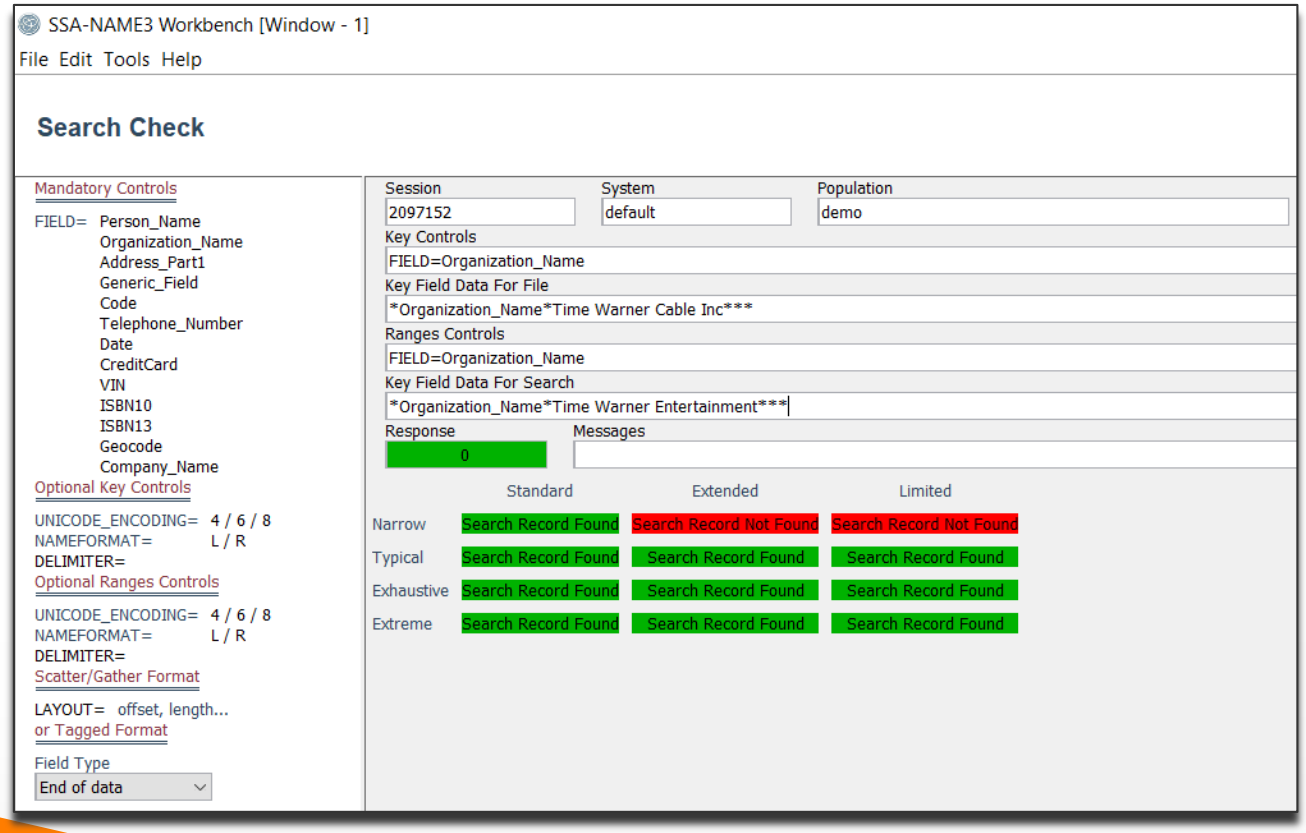
Phase 3. Setup Match rules: Match Level

Typical	**Appropriate for most matches**
Conservative	**Tighter than Typical, causing undermatching**
Loose	**More matches than Typical, causing overmatching. Good to use this in a match rule for manual merges**

- SSA Workbench tool, available as part of MDM Resource Kit, helps decide on the appropriate match level
- Demo to look at how records “Arim Bose” matches with “Arim Gore” along with their addresses, using different match levels and their scores
- U(Undecided)/R(Rejected) are considered as rejected matches in MDM

Phase 3. Setup Match rules: Search Level

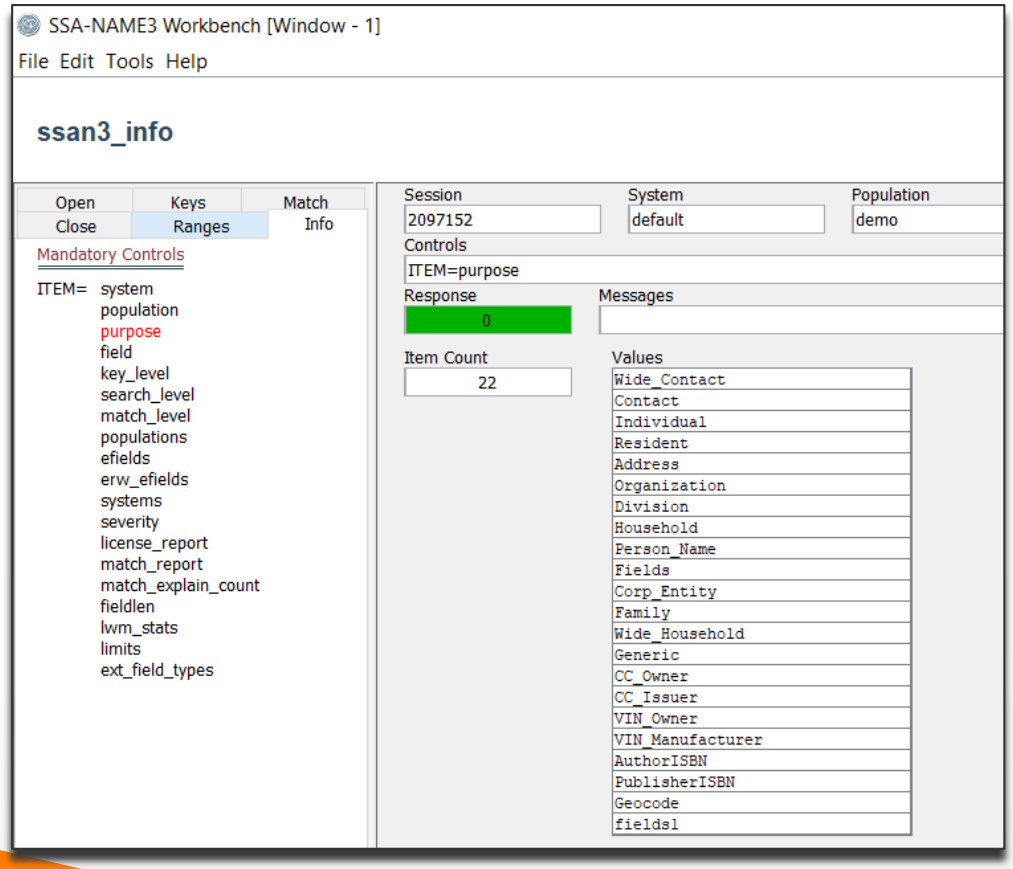
Narrow	Typical	Exhaustive	Extreme
• Most stringent, faster, undermatching • Correct and complete datasets and highly matchy datasets	• Apt for most match rulesets	• More match candidates than Typical, more time, overmatching • Smaller, less complete, less reliable datasets	• More match candidates than Exhaustive, much more time, overmatching • Datasets that are even less reliable and less complete



Informatica

Phase 3. Setup Match rules: Match Purpose

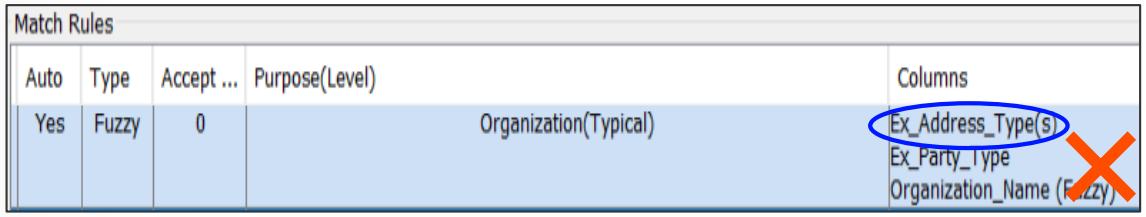
- For data with both Organizations and Individuals, use appropriate match purpose based on the party/customer type
- If there is no customer type indicator, you can use Organization. Or use Division as match purpose for mixed data types
- If you are trying to identify matches for people where address is important to determine if two records are for the same person, you can use Resident match purpose
- Different match purposes available:

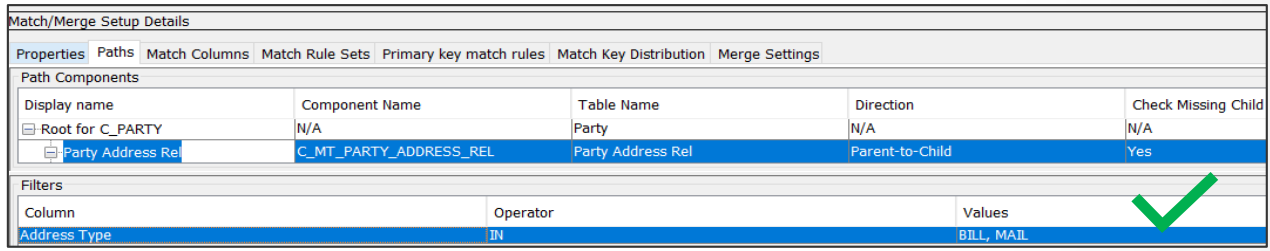


Informatica

Phase 3. Setup Match rules:– do’s and don'ts

- Start with rules that will provide the tightest matches
- Fuzzy match rules are evaluated first, followed by exact match rules
- For each fuzzy match rule, exact columns are evaluated first. Use exact match columns when you can. Saves fuzzy calls made to SSA
- Exact match rules are processed almost exclusively on the database. If database performance is not sufficient, convert them to **Filtered** match rules. Comes with trade-off between match quality and performance
- Run SQL queries on exact match columns to find rough estimate of potential candidates returned
- Loose filters will pass more potential candidates to SSA, creating more work and decreasing performance.
  Examples of tight filters – Id, Date Of Birth, Postal Code. Loose filters – City, State
- Avoid subtype match; makes multiple SSA calls for each type. Use a match path filter instead



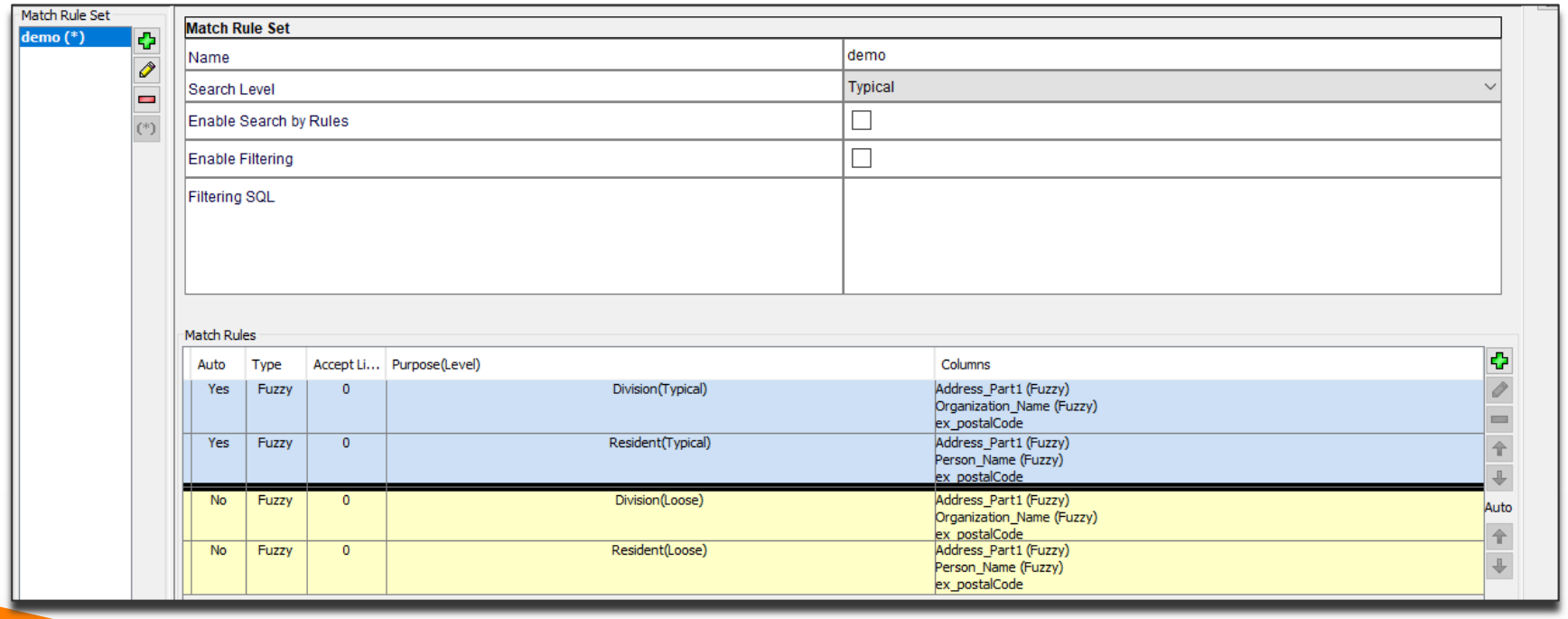


- Use filter on root path filter to exclude records from match, instead of filtering on match rule level. Saves those records from being tokenized and thus will not participate in match

Informatica

Phase 4: A dry run of the match job using draft rules

- Avoid having tighter match rules during this phase. Below example has ex_postalCode as exact
- This will give you a feel of how fuzzy name and address matches look like. Gives you an idea on the quality of matches. Helps assess any underlying data issues



Informatica

Phase 4: Review match results from the dry run

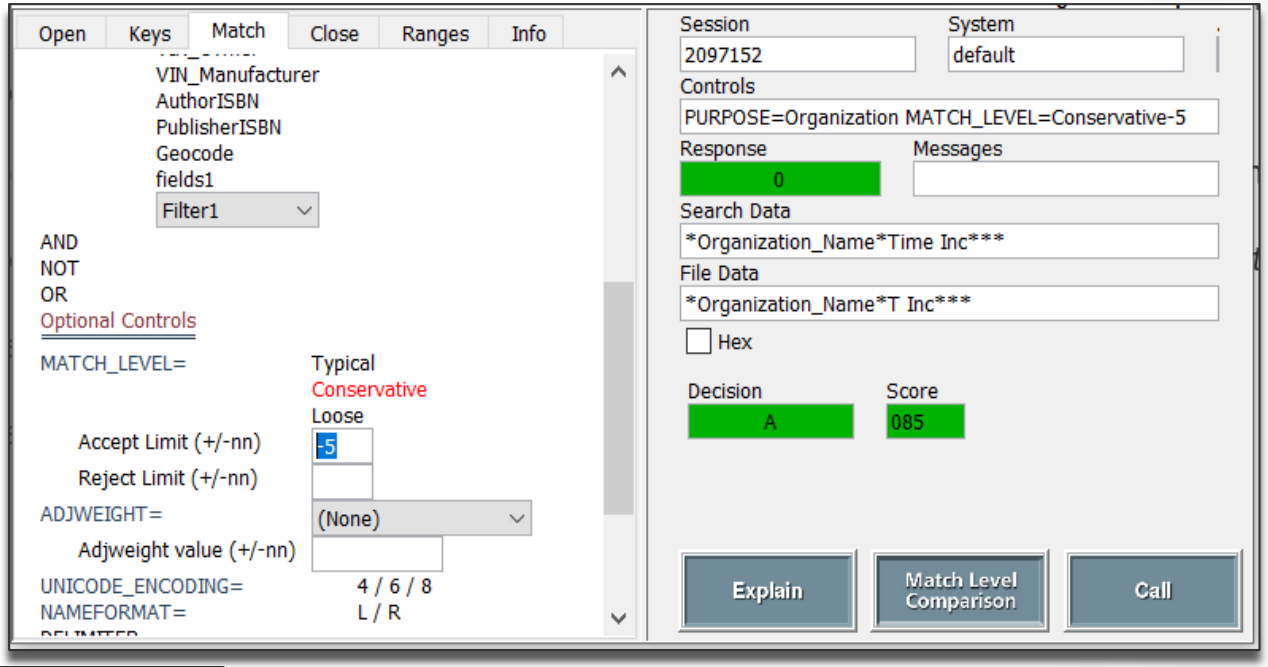
- Match results:

Time Warner Entertainment - 5493 S Queens Rd Rochelle 61068	Time Warner Cable Inc - 5193 S Queens Rd Rochelle 61068	Match Rule 3	Manual
Arim - 669 Butler St SE Atlanta 30303	Arim Gore - 69 Butler St Atlanta 30303	Match Rule 1	Auto
Arim - 669 Butler St SE Atlanta 30303	Arim Bose - 669 Butler St SE Atlanta 30303	Match Rule 1	Auto

- Run a query against MTCH table group by match rule; helps revise rules that have gained less matches
- Make a copy of MTCH table for each iteration
- Review undermatches VS overmatches

E.g. *"Time Inc" did not match with "Time Warner Cable Inc" as their addresses are different*

- Use SSA Workbench to know why certain records matched and did not match
- SSA workbench tool also helps to make adjustments on accept limits
- To change accept limits in MDM:

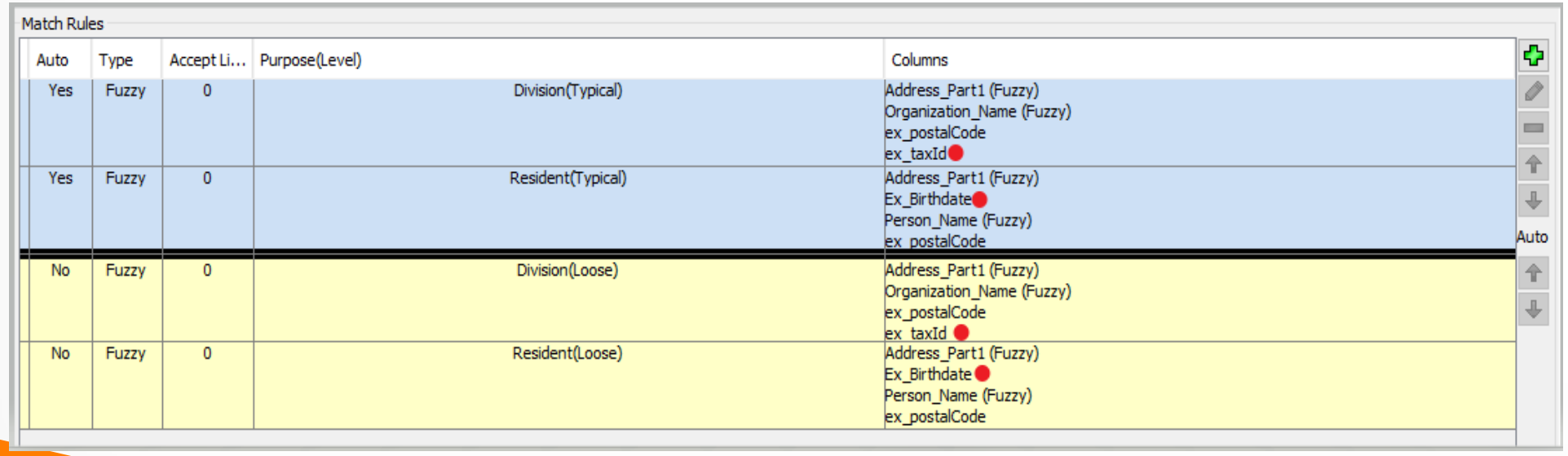


Auto	Type	Accept ...	Purpose(Level)	Columns
Yes	Fuzzy	-5	Organization(Conservative)	Organization_Name (Fuzzy) ex_postalCode ex_taxId

Informatica

Phase 5: Tune match rules with exact columns

- Introduce unique identifies (as exact match column) to further qualify matches and to further tighten the rules
- If there's no unique identifier, then use exact column such as DateOfBirth
- Prevent performance issues by including at least one exact match column in each match rule
- Use several identical match rules with varying exact match columns



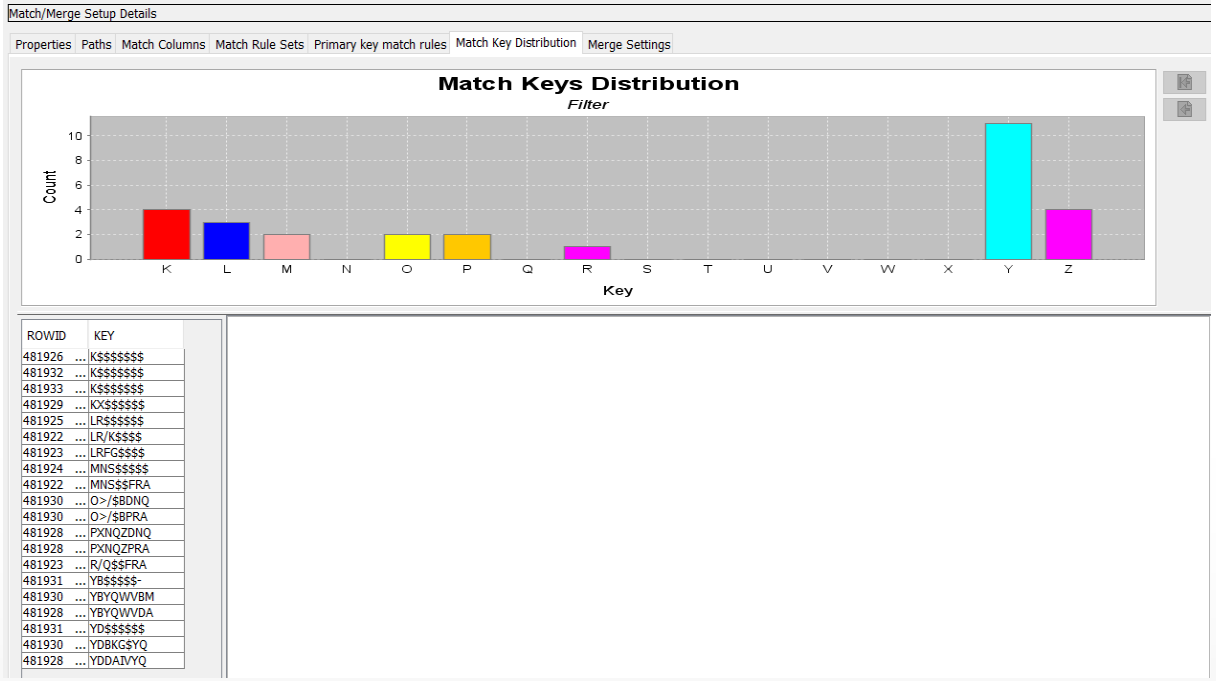
Informatica

Phase 6: Review match results

- What to review?
  - **STRP table**
    - If there's a large set of data (outliers; for e.g. records more than 50K) residing between a set of SSA keys

      SELECT DISTINCT ROWID_OBJECT, DATA_COUNT,SSA_DATA, DATA_ROW FROM C_PARTY_STRP WHERE SSA_KEY BETWEEN 'YBJ>$$$$' AND 'YBLVZZZZ' AND INVALID_IND = 0 ORDER BY ROWID_OBJECT, DATA_ROW

    - Match key distribution



Informatica

Phase 6: Review final match results (continued)

- **Review cleanse server log**

Ranger5 Matching TCan:167038393 Tgr:167038393 TSSA:98428393 TM:0 TR:1 Cur RI:1800219 Cur Range:YBJ>$$$$ to YBLVZZZZ CompsPerRange:167043000

Ranger7 Matching TCan:173858958 Tgr:173858958 TSSA:104410228 TM:14 TR:1 Cur RI:1802487 Cur Range:YBJ>$$$$ to YBLVZZZZ CompsPerRange:173862000

*[RangerManger] [INFO] com.siperian.mrm.util.threads.ThreadMonitor: RangerProducer Candidates Read:**2020666***
[RangerManger] [INFO] com.siperian.mrm.util.threads.ThreadMonitor: MatchGatherer received 6544
*[RangerManger] [INFO] com.siperian.mrm.util.threads.ThreadMonitor: RangeSorter Sorting: Recs in:**2,020,666** with **13,621,427** ranges.*
*SortManager: Ranges in: **13,621,427** Sorted Ranges out: **81,000** file Count: 137 Sort Count: 1263*
*[RangerManger] [INFO] com.siperian.mrm.util.threads.ThreadMonitor: run minutes:**778** Max minutes:2880*

What does this mean?

- It is processing **2,020,666** records and that those records produced **13,621,427** search ranges that need to be evaluated to complete the matching
- It has currently only processed **81,000** of the ranges yet. It has taken **778** minutes to do that
- Range: YBJ>$$$$ to YBLVZZZZ keeps appearing on the log and is a potential hotspot. The comparison count (CompsPerRange) is over 167 million and counting
- This job will take a long time to complete. Maybe there's a high frequency word (e.g. 'Medical') in the data within this range. Clean up this data

Phase 6: Review final match results (continued)

- **Review if matches are slow**
  - Slow DB read

    *Ranger0 Matching TCan:**156020763** Tgr:156020763 TSSA:2188740 TM:2165385 TR:186577 Cur RI:7511404 Cur Range:YKMGBBQ$ to YKMGBBQ/ CompsPerRange:160*
    *Ranger0 Matching TCan:**160711852** Tgr:160711852 TSSA:2268773 TM:2244506 TR:193761 Cur RI:9600897 Cur Range:YKVAVA to YKVA$VA/ CompsPerRange:2196*
    *(**156,020,763** – **160,711,852**) = **4,691,089** ← total number of candidates read from DB from one minute to another – a low count could indicate a potential database or network issue – expect millions*

  - High number of candidates going to SSA; poor exact match columns are used

    TCan:70056898 Tgr:70056898 TSSA:1,023,821 TM:1012237 TR:88571 Cur RI:7502399 Cur Range:S>M$$$$$ to S>M/ZZZZ CompsPerRange:236474

Note: *How to track progress of a match job?* KB - https://kb.informatica.com/howto/6/Pages/19/503645.aspx -- Helps *determine the approximate time taken by the job to run and complete eventually*

Tuning Process Server and Base Object properties

Property	Usage
Threads for Cleanse Operations	To achieve parallelism
Number of rows per match job batch cycle	Start with 10% of volume of records to be matched and adjust upwards
Maximum matches for manual consolidation	Increase it as needed to avoid match job failure
Max Elapsed Match Minutes	Default is 20. Increase only if match rules and data is complex
Dynamic Match Analysis Threshold (DMAT)	Helps improve performance when large ranges are causing it *[2015-03-13 20:16:05,306] [RangerManger] [INFO] com.siperian.mrm.util.threads.ThreadMonitor: Dist:Ranger4 Matching TCan:56892103 Tgr:50659217 TSSA:12285983 TM:7230 TR:22165 Cur RI:100207398 Cur Range:OG$$$$$$ to OGZZZZZZ CompsPerRange:97999* …….. *[2015-03-13 22:09:52,611] [RangerManger] [INFO] com.siperian.mrm.util.threads.ThreadMonitor: Dist:Ranger4 Matching TCan:82741526 Tgr:73413451 TSSA:21394992 TM:7250 TR:22165 Cur RI:99784575 Cur Range:OG$$$$$$ to OGZZZZZZ CompsPerRange:25947538 ⟶ 25 Million comparisons* Analyze the data to assess why a given search range contains a large count; maybe matchy data Setting the DMAT level too low may cause under matching Note: Any DMAT changes on Production should be reviewed with Informatica GCS

Informatica

Tuning cmxcleanse.properties

Property	Usage
cmx.server.match.distributed_match	Set to 1 to enable. Default is 0 (disabled)
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.match.loader_batch_size	Default is 1000, when file load is set Maximum number of insert statements to send to the database during direct load of the match process

Tuning the database

- Exact rules are converted to SQL queries based on exact match columns in the match rule and their match paths. Look for CREATE/INSERT for T$MLE and T$MT tables
  - If you find the exact match query running slow, query related to T$MLE or T$MT
- Ensure all tables in the exact match query are analyzed
- Create index on one or more exact match columns

References

- MDM Fuzzy Match Deep Dive - https://www.youtube.com/watch?v=_T6x24bMnP8&feature=youtu.be
- How to configure SSA Workbench on MDM Resource Kit - https://youtu.be/Jp2gcFgE_5Q
- How to use SSA Name3 workbench - https://youtu.be/lLwTHA0SnY4
- How to track progress of a Match job in MDM - https://kb.informatica.com/howto/6/Pages/19/503645.aspx

Thank You

Informatica™