Aug 31, 2023

MDM SaaS Match and Merge Best Practices

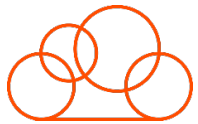
- Ankit Mahajan, Principal SME, GCS
- Madhu Immadi, Lead SME, GCS

Informatica

Housekeeping Tips

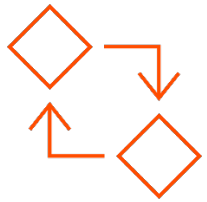
- Today's Webinar is scheduled for 1 hour
- The session will include a webcast and then your questions will be answered live at the end of the presentation
- All dial-in participants will be muted to enable the speakers to present without interruption
- Questions can be submitted to "All Panelists" via the Q&A option and we will respond at the end of the presentation
- The webinar is being recorded and will be available on our Success Portal - where you can also download the slide deck for the presentation. The link to the recording 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

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.

Informatica

Aug 31, 2023

MDM SaaS Match and Merge Best Practices

- Ankit Mahajan, Principal SME, GCS
- Madhu Immadi, Lead SME, GCS

Informatica

Agenda

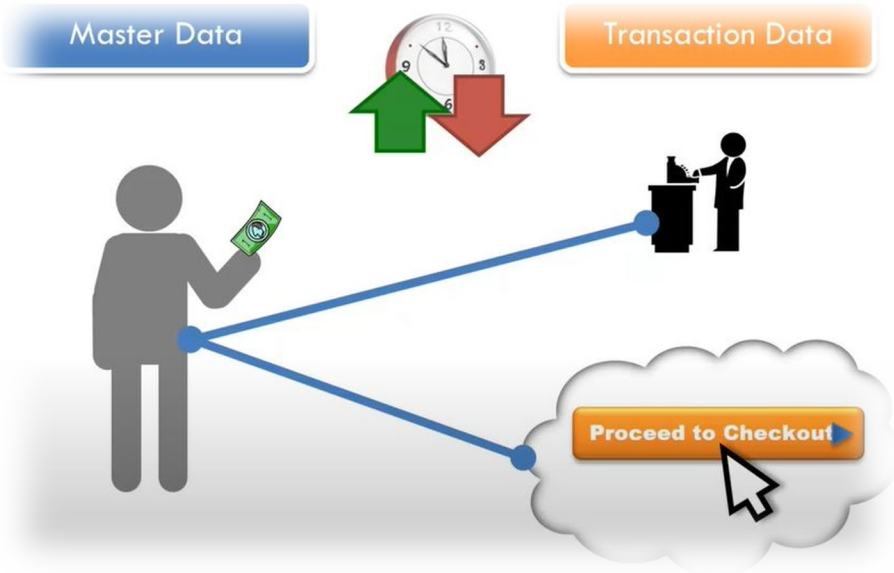
Best practices for:

1. Match Process Overview
2. Data Profiling
3. Match Model Key Points
4. Match Rule Key Points
5. Match Fields Key Points
6. Match Tuning Key Points

Informatica

What Data Should be Mastered?

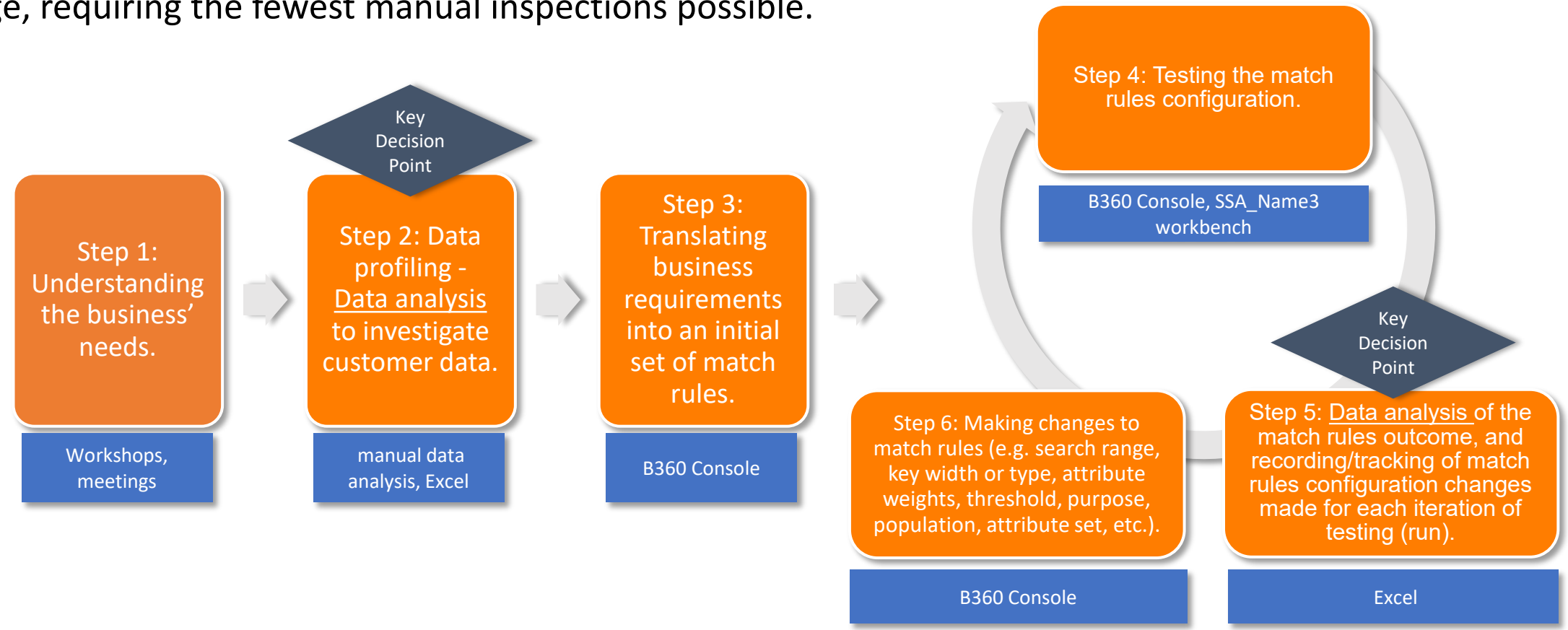
- **Fundamental Entities Upon Which a Business is Based**
  - SPACE = Suppliers/Vendors, Products/Services, Assets, Customers, Employees/Contractors, (+ Contracts + Accounts …)
  - **Company Structure** – Divisions, Regions, Territories, Departments, etc.
  - **Related Data (Child Tables)** – Addresses, Communication Methods, Certifications, etc.
  - Represents people, places and things related to organization
  - Customer details, product details, employee details



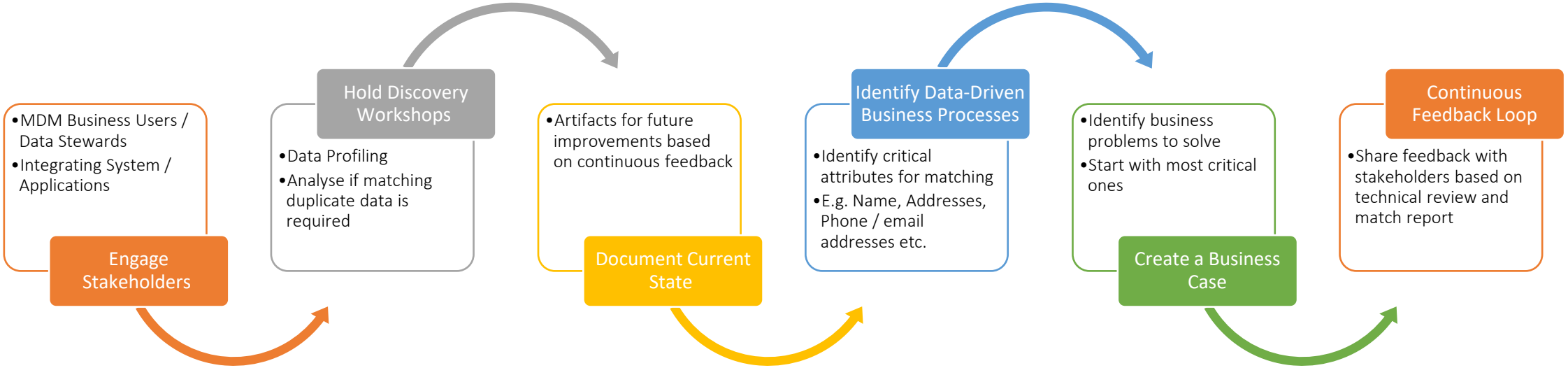
- Master Data is Relatively Static – Non-Transactional.
  - Filter out transactional data through CDI jobs
  - Generates excessive matches
  - Impacts performance negatively

Overview: Match Rules Implementation

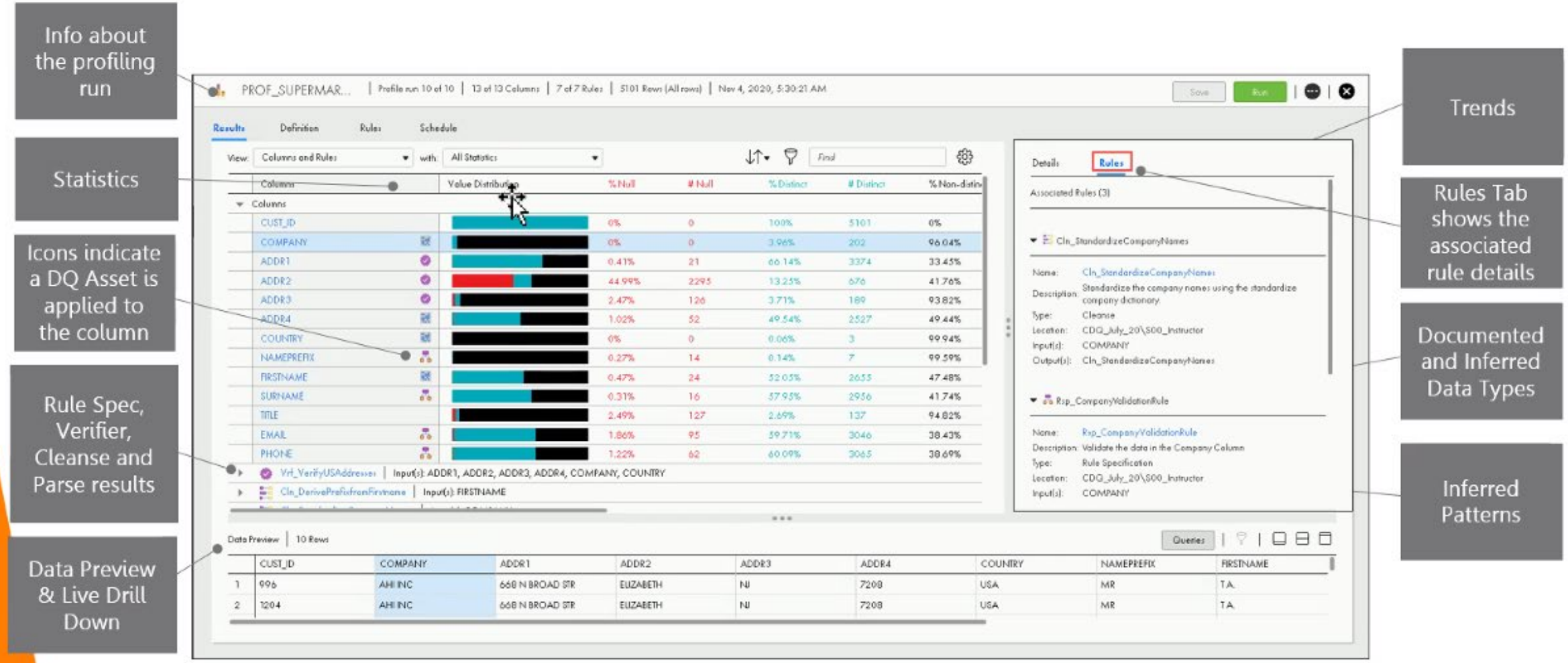
Goal: to find the match rules configuration that uses the least number of rules, of which most are auto match-merge, requiring the fewest manual inspections possible.



Understand Business Needs



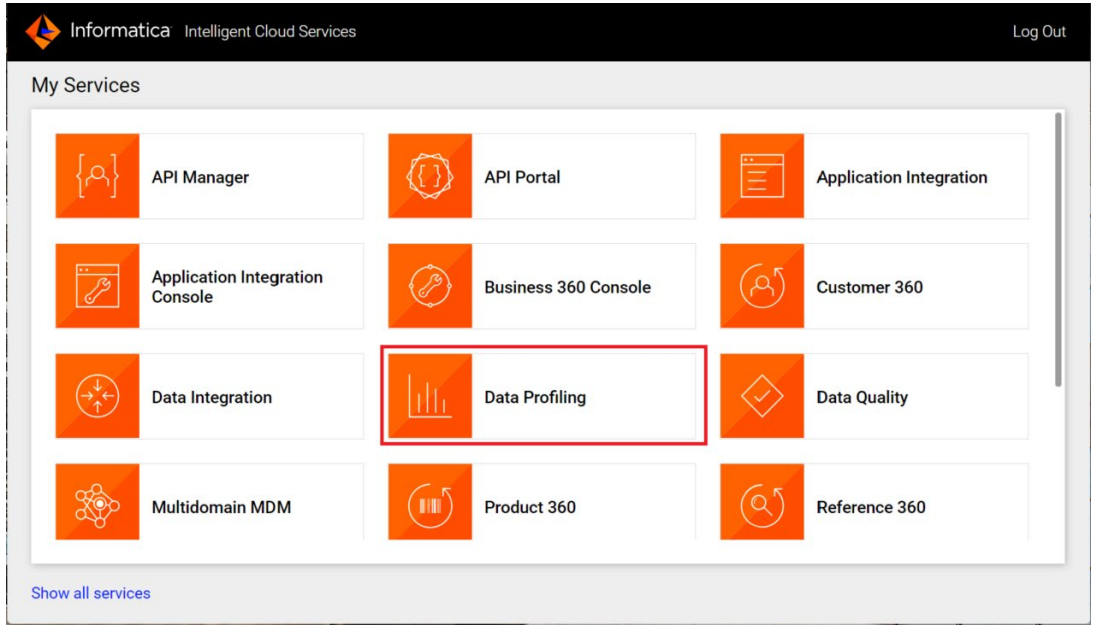
Data Profiling (contd.)



Informatica

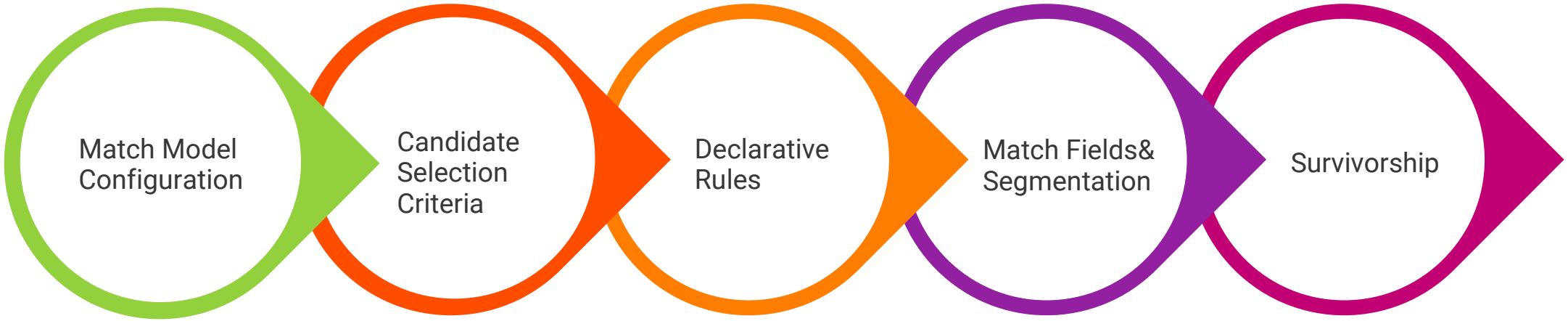
Data Profiling Key Points

- Use data profiling at project start to discover if data is suitable for MDM
- Identify duplicates, and missing values
- Identify inconsistences in data formats
- Identify stale data & contradicting data
- Identify Distinct count and percent
- Helps in making the fuzzy match key(s) and creation of match rules.



Match Configuration Steps

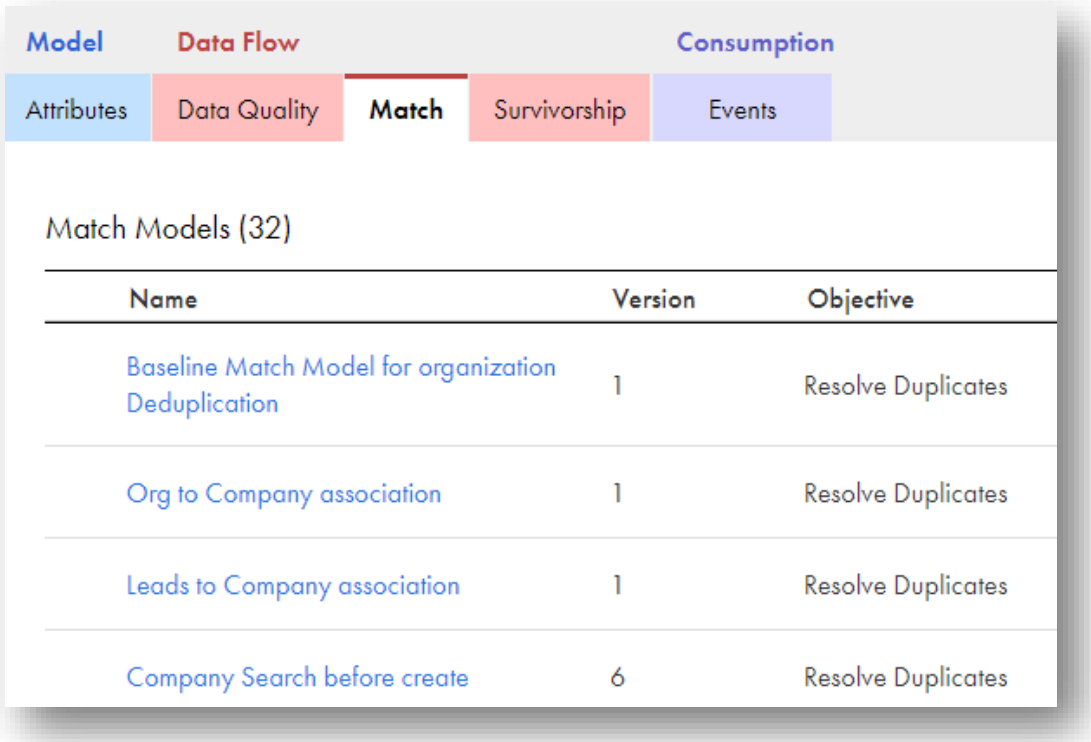
Master data identification



Informatica

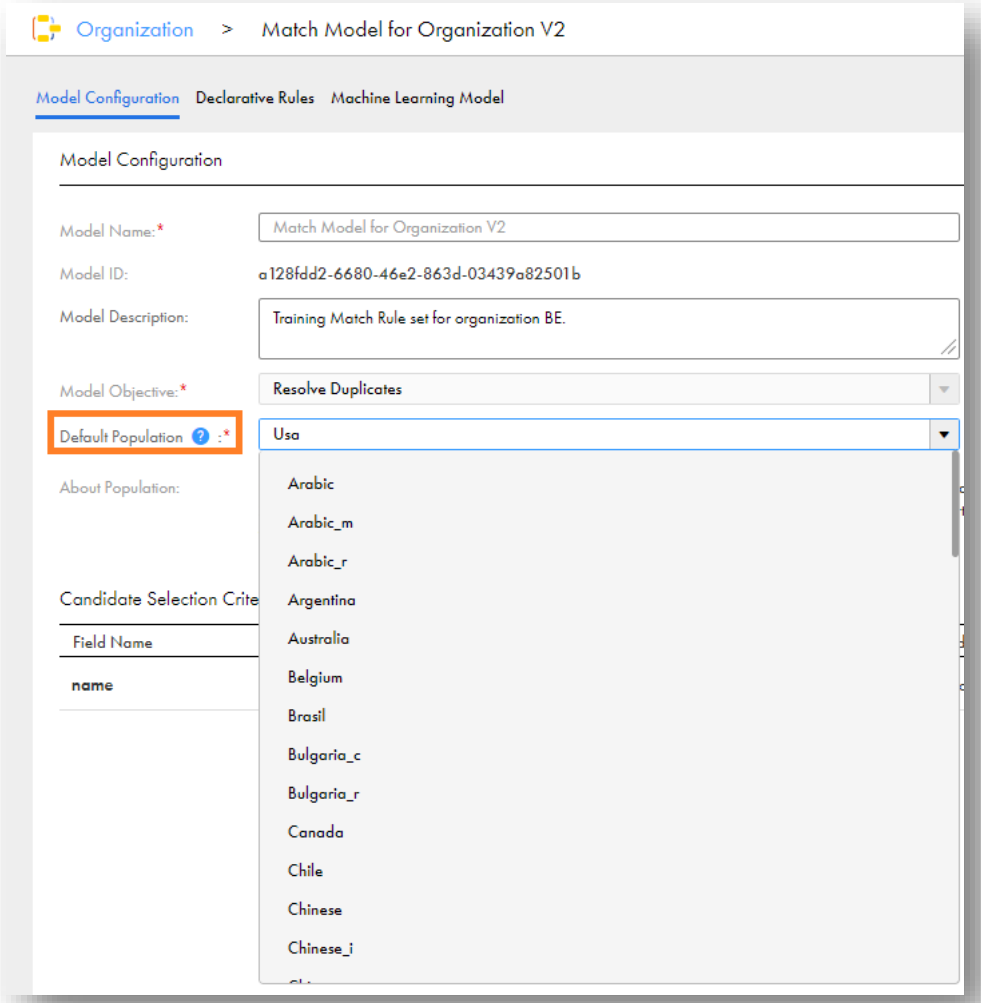
Match Model Key Points

- Configuration that determines how your data is matched
- Use a Meaningful model name
- You cannot change the model's name after the model is published
- A match model can be copied to new one.
- Predefined models can’t be deleted
- Before deleting the match model, remove it from its dependencies (ex. Job, Search Match API)
- Maximum 25 published Match models can be created.



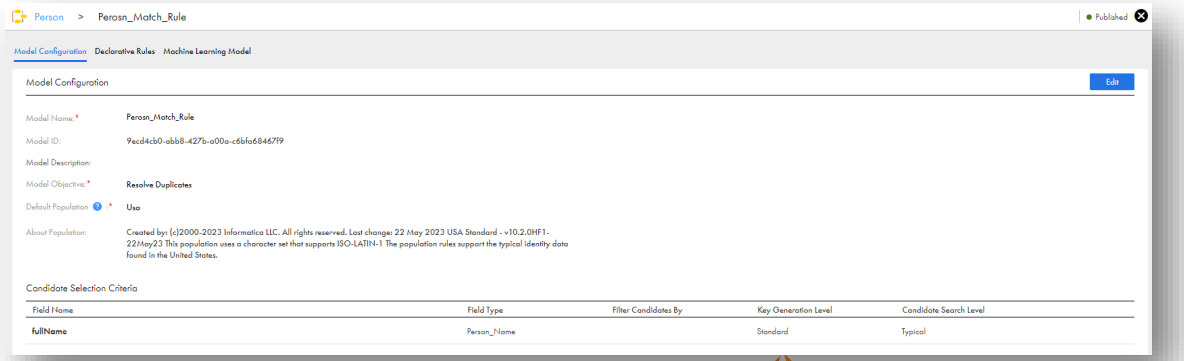
Population Key Points

- Population is a definition of certain characteristics of data
- Choose the right population file based on the requirement, default is USA
- Populations contain the logic to generate match keys
- The system matches similar records that belong to the same population
- A population set encapsulates intelligence about name, address, and other identification information
- If you have mixed data from different languages consider using multiple populations (USA, Japan, Chinese etc.). Population Name field can be mapped during ingress/IFI.



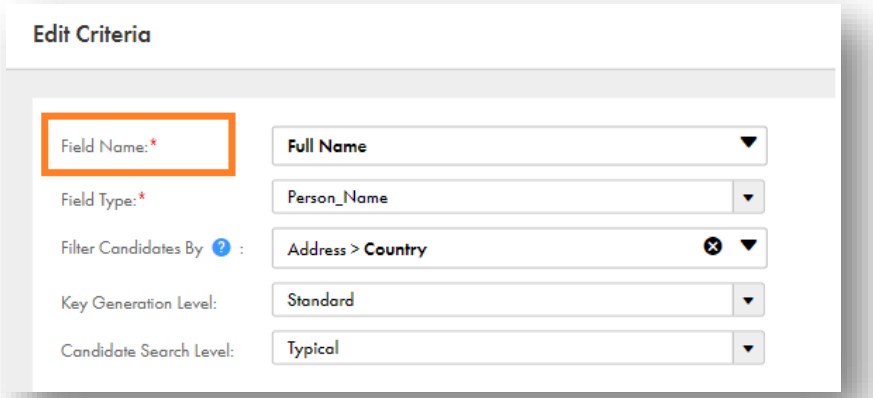
Candidate Selection Criteria

- Match candidates are record pairs that are possible matches
- Avoid choosing multiple Candidate Selection Criteria unless it is essential, and you have such a requirement for your use case.
- Having multiple criteria improves quality but affects the performance of the candidate selection process
- Multiple candidate selection criteria: Does a union of all candidates from all the criterion.
- Configure candidate selection criteria that are necessary to enhance the quality of the match candidates
- Remove extraneous/bad data to improve candidate selection and avoid large number of candidates returned.
- Ensure that you publish when you modify Population file and Candidate selection criteria
- Ensure that you regenerate the Match keys for the records when you modify Population file and Candidate selection criteria



Candidate Selection Criteria Key Points

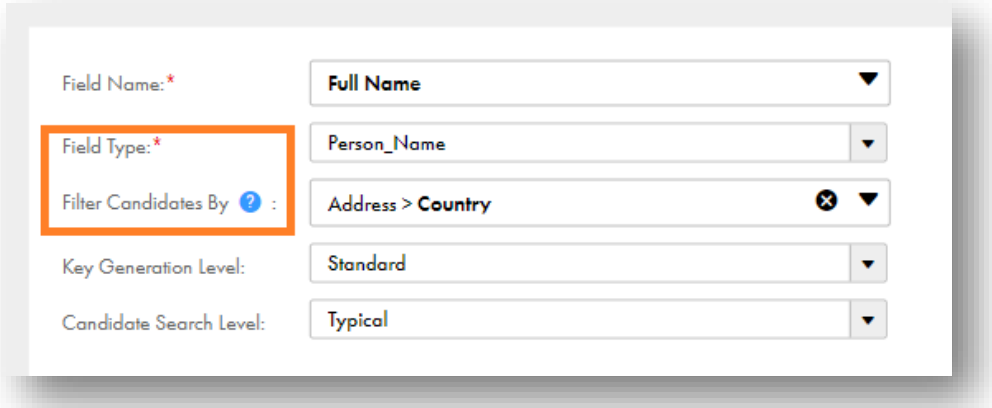
- Business entity field use for generating the match keys and match candidates
- You can use numeric/alpha numeric columns depending on the cardinality level – for example SSN, passport numbers are valid. Phone numbers are commonly used. Use filter criteria, where applicable, to limit the number of candidates returned.
- As an example, use full Name instead of firstname or lastname
- Based on your discussion with the business users and data audit, as a rule of thumb, you would use the following as match keys:
  - If data contains organization names use organization name as field
  - If data contains individual names only use person name as the field
  - If the data contains addresses only use address part1 as the field



Candidate Selection Criteria Key Points (cont.)

Field Type:

- Indicates the type of data contained in the "Field Name" field
- Use a field that serves to capture the potential candidates
- Ensure that the selected field meets the selection criteria best practices



Filter Candidates:

- Adding a filter might reduce the number of match candidates
- Fewer match candidates improve the performance and accuracy
- Find filter column that correctly removes the incorrect matches without losing good matches
- This can effectively be used to ensure that 10K limit is not reached.

Key Generation Level Key Points

- Define the thoroughness with which match keys are generated to identify the records for matching.
- Decide key generation level based on:
  - Size and quality of the data
  - Reliability of the matched records
  - Processing time
- Standard is Default and recommended
- Use Extended for high-risk or critical search applications
- Larger candidate sets at search time
- Using Limited reduces the search reliability

Key Generation

	Key Quantity	Data Volume / Quality / Time
Extended	Large	Not large / Poor / Lots of time
Limited	Small	Very Large / Fairly Good / Not much time
Standard	Medium	Medium / Medium / Medium

Key Width (Length)

Example: John Alexander Smith

Limited	Standard	Extended	Based on
YBKS/$$$	YBKS/$$$	YBKS/$$$	Smith+John+Alexander
K?KVKTU-	YB$U/$$-	YB$U/$$-	Smith+Alexander+John
S$NQCAWA	K?KVKTU-	K?KVKTU-	Alexander+John+Smith
	K?WAIKKU	K?WAIKKU	Alexander+Smith+John
	S$NQCAWA	S$NQCAWA	John+Alexander+Smith
	S$NQJ*NQ	S$NQJ*NQ	John+Smith+Alexander
		S?N/AZNQ	JohnAlexander+Smith
		KHLMGA$-	AlexanderSmith + John

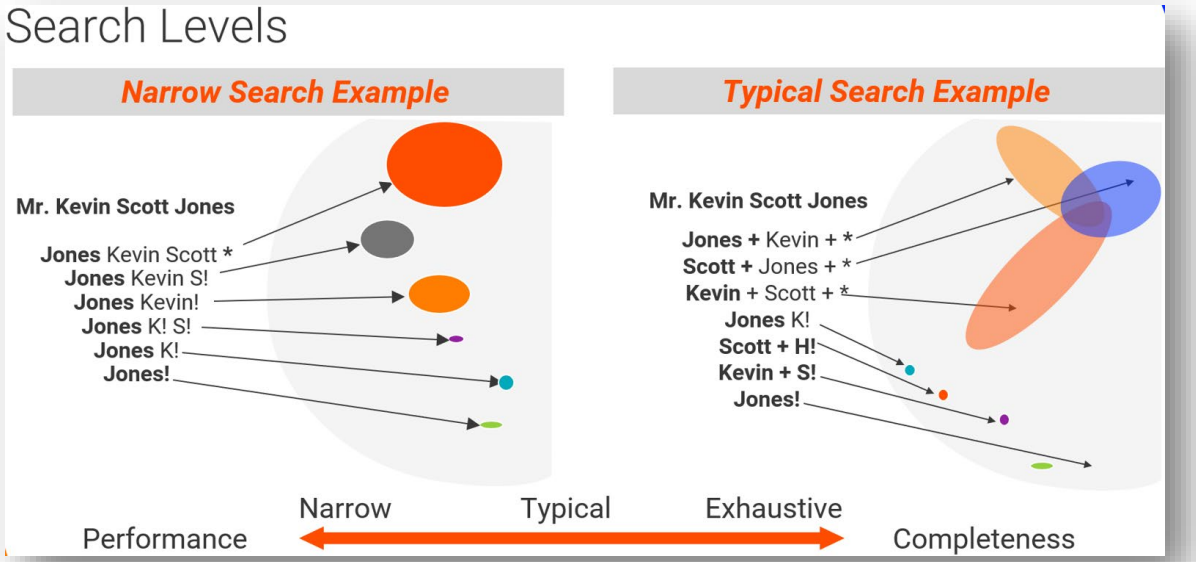
Performance ⟷ Completeness

Candidate Search level Key Points

- Defines how stringently and thoroughly to search for match candidates
- Declarative rules applies on the match candidates and not on the entire data
- Decide on a candidate search level based on the following considerations:
    - Size and quality of the data
    - Criticality of the matches
    - Time constraints

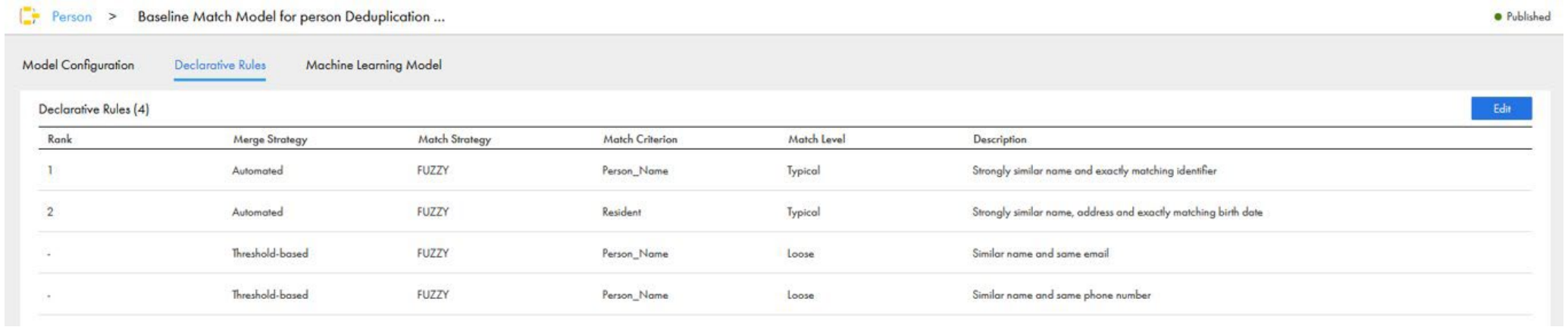
Candidate Search Level

	Candidates	Data Volume / Criticality / Time
Extreme	The most	Small / Very High / Takes the most time
Exhaustive	Many	Small / High / Lots of time
Typical	Moderate	Medium to Large / Moderate / Moderate
Narrow	Few	Very Large / Low / Very fast



Declarative Rules Key Points

- Set of conditions and BE fields to identify the duplicate records
- Configure a match model with the least possible number of match rules for optimal match results
- A meaningful description helps identify a declarative rule and its purpose.
- Separate out rules per business logic
- Start with simple tighter rules



Rank	Merge Strategy	Match Strategy	Match Criterion	Match Level	Description
1	Automated	FUZZY	Person_Name	Typical	Strongly similar name and exactly matching identifier
2	Automated	FUZZY	Resident	Typical	Strongly similar name, address and exactly matching birth date
-	Threshold-based	FUZZY	Person_Name	Loose	Similar name and same email
-	Threshold-based	FUZZY	Person_Name	Loose	Similar name and same phone number

Match Strategy Key Points

- You can define a declarative match rule for exact or fuzzy matching.
- Configure the exact match strategy if the quality of the data is good
- Configure the Fuzzy strategy, for a probabilistic match based on data patterns
  - ‘Tech Corp’
  - Technology Corporation
- Rank rules in the order you want to run them for matching
- Unranked rules get applied if the matching pair doesn’t meet ranked rule conditions

Model Configuration | Declarative Rules | Machine Learning Model

Declarative Rules (8)

Rank	Merge Strategy	Match Strategy	Match Criterion	Match Level	Description
1	Automated	EXACT	-	-	Rule 1: Exact DUNS, Exclude from Match
2	Automated	FUZZY	Division	Typical	Rule 2: Fuzzy on Address_Part_1, Org_Name and Exact on Partial_Postal_Code, Country Description, Exclude from Match, ...
3	Automated	FUZZY	Division	Typical	Rule 3: Fuzzy on Address_Part_1, Org_Name and Exact on City, State Description, Country Description, Exclude from Match, SIP_POP_International
4	Automated	FUZZY	Organization	Conservative	Rule 4: Fuzzy on Org_Name and Exact on City, Partial_Postal_Code, Country Description, Exclude from Match, SIP_POP_Japan
5	Automated	FUZZY	Corp_Entity	Conservative	Rule 5: Exact City, Country, Partial Zip code, Exclude from match and Fuzzy Organization Name with Threshold(100%)
6	Manual	FUZZY	Division	Typical	Rule 6: Fuzzy on Address_Part_1, Org_Name and Exact on Partial_Postal_Code, Country Description, SIP_POP_International
7	Manual	FUZZY	Division	Typical	Rule 7: Fuzzy on Address_Part_1, Org_Name and Exact on City, State Description, Country Description, SIP_POP_International
-	Threshold-based	FUZZY	Organization	Typical	

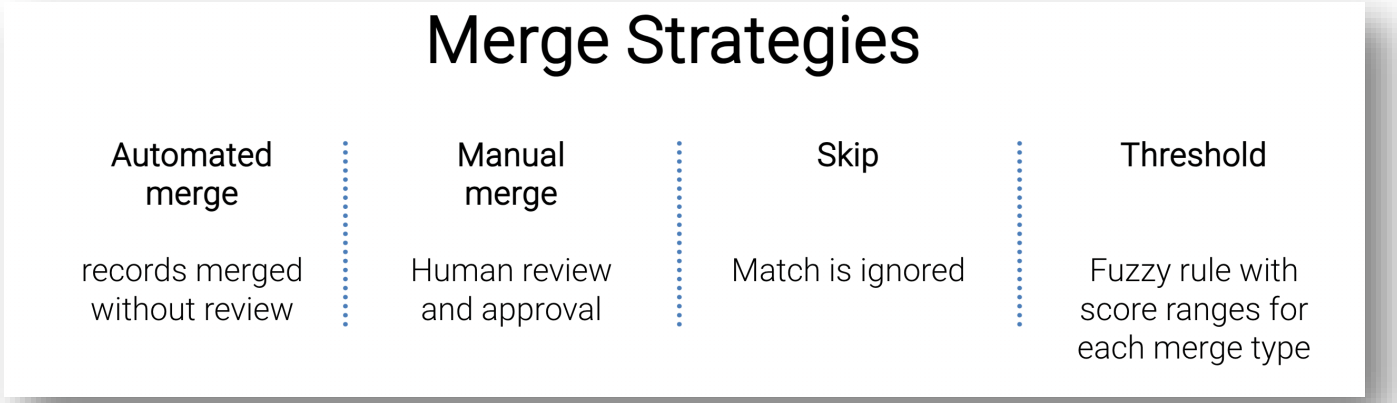
Match Criteria Strategy Key Points

- Applicable for Fuzzy based rules, and every Fuzzy rule must have a match criteria
- Two rules with properties but different match criteria result in different set matching records.
- For example, if address is important to determine match of a two person records, use the Resident match criteria, and use Division for Organization and address.
- Refer complete details of Match criteria via Doc link

Match Criteria	Description
Address	Identifies an address match. The address might be postal, residential, delivery, descriptive, formal, or informal.
AuthorISBN	Matches an author for the specified book with an International Standard Book Number (ISBN).
CC_Issuer	Matches the organization that issues credit cards to the customers. Use this match criteria to match credit cards or the credit card-issuing organizations.
CC_Owner	Matches the credit card owner and the address.
Contact	Matches a contact within an organization at a specific location.
Corp_Entity	Matches an organization based on its legal corporate name, including the legal endings such as INC and LTD. Use the Corp_Entity match criteria to honor differences between names, such as ABC TRADING INC and ABC TRADING LTD.
Division	Matches an organization based on the address. This match criteria matches organization X at address Y or Z if multiple addresses are provided.
Family	Matches individuals with the same or similar family names, who share the same address or the same telephone number.
Fields	For general nonspecific use. Use this criteria with any field type.
Geocode	Matches records based on geocode, which is the geographic coordinates and elevation.
Household	Matches individuals with the same or similar family names, who share the same address.
Individual	Matches individuals by name and with either the same ID or date of birth.
Organization	Matches organizations primarily by name. Use this match criteria for online searches where a human is available to review the search results and make a choice.
Person_Name	Matches a person by name. Use this match criteria for online searches where a human is available to review the search results and make a choice.
PublisherISBN	Matches the publishing organization for a specified book with an International Standard Book Number (ISBN). The ISBN can be a 10-digit number or a 13-digit number.
Resident	Matches a person based on an address.
VIN_Manufacturer	Matches the manufacturer of a vehicle for a specified vehicle number.
VIN_Owner	Identifies and matches the owner of a vehicle.
Wide_Contact	Identifies a contact within an organization without regard to the location.
Wide_Household	Identifies individuals who share the same address, the same family name, or the same phone number.

Merge Strategy Key Points

- A merge strategy indicates the action to be performed on the records that have gone through the match process.
- Use Automated for rules that are based on unique identifier fields, such as social security number
- Use manual strategy for data steward review. Avoid rules that generate too many matches. You do not want to flood the Data Stewards with too many matches



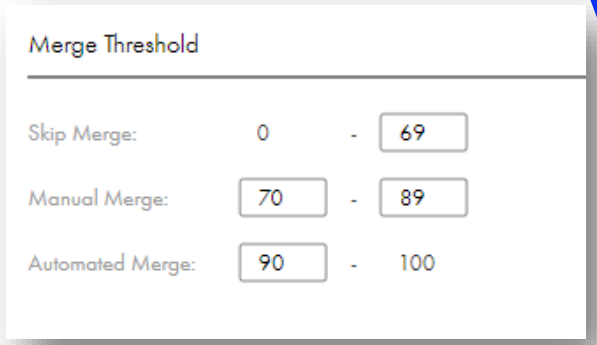
Declarative Rules Key Points (cont.)

- Always have some exact match filters on every match rule especially fuzzy match rules
- Excessive use of fuzzy matching algorithms without tight filters can impact system performance.
- Remove redundant rules that don't capture any matches
- Move rules up or down in match set to see if order improves results

Declarative Rules (8)

Rank	Merge Strategy	Match Strategy	Match Criterion	Match Level	Description
1	Automated	EXACT	-	-	Rule 1: Exact DUNS, Exclude from Match
2	Automated	FUZZY	Division	Typical	Rule 2: Fuzzy on Address_Part_1, Org_Name and Exact on Partial_Postal_Code, Country Description, Exclude from Match, ...
3	Automated	FUZZY	Division	Typical	Rule 3: Fuzzy on Address_Part_1, Org_Name and Exact on City, State Description, Country Description, Exclude from Match, SIP_POP_International
4	Automated	FUZZY	Organization	Conservative	Rule 4: Fuzzy on Org_Name and Exact on City, Partial_Postal_Code, Country Description, Exclude from Match, SIP_POP_Japan
5	Automated	FUZZY	Corp_Entity	Conservative	Rule 5: Exact City, Country, Partial Zip code, Exclude from match and Fuzzy Organization Name with Threshold(100%)
6	Manual	FUZZY	Division	Typical	Rule 6: Fuzzy on Address_Part_1, Org_Name and Exact on Partial_Postal_Code, Country Description, SIP_POP_International
7	Manual	FUZZY	Division	Typical	Rule 7: Fuzzy on Address_Part_1, Org_Name and Exact on City, State Description, Country Description, SIP_POP_International
-	Threshold-based	FUZZY	Organization	Typical	

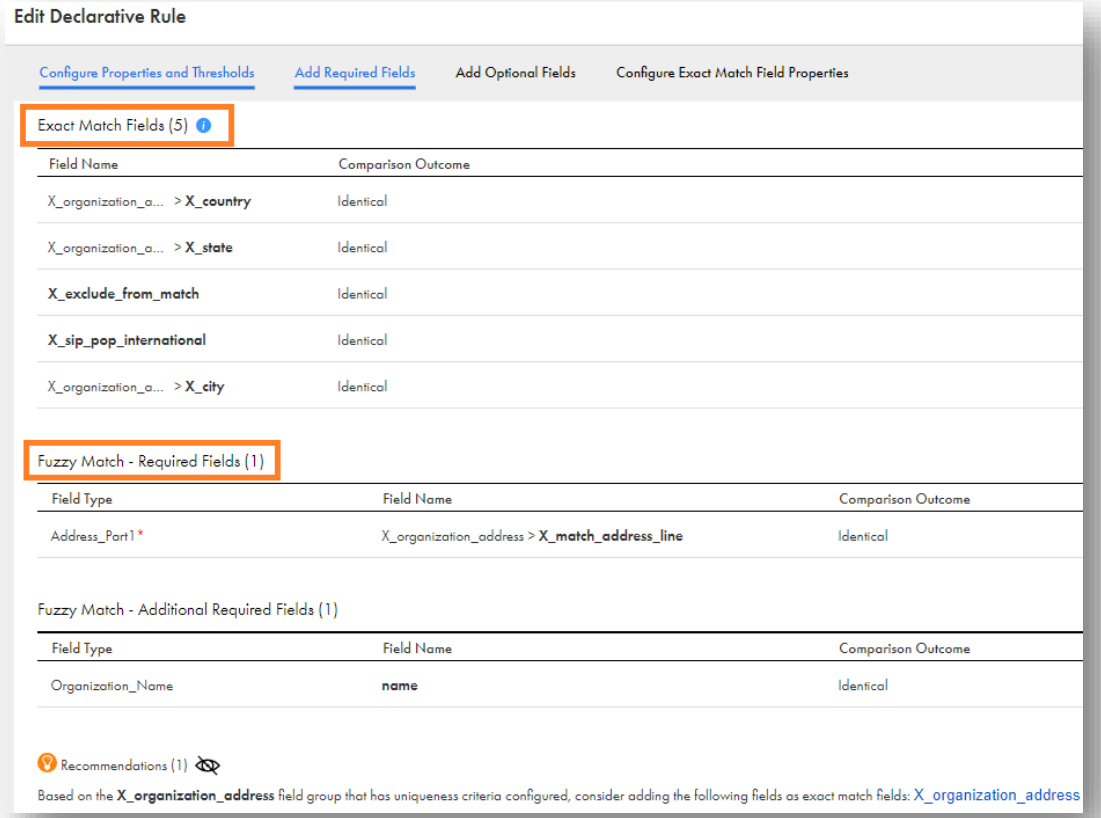
Threshold Based Rules Key Points



- Use thresholds to control the level of similarity for a match
- A lower threshold will result in higher recall but lower precision, while a higher threshold will increase precision but reduce recall.
- Configure tight thresholds to avoid false positives for name and address match without unique identifiers
- Review documentation for predefined rules while using Threshold-based Merge Strategy
- When there is a tie between manual merge and automated merge, Match outcome always go for REVIEW so that Data stewards can take a necessary decision in terms of conflicts. Refer Article for example
- All THB rules get executed and using high number impacts the performance.

Match fields Key Points

- Add Exact match field to match identical data
- Exact fields Used for filtering that reduces the number of rows that go to through fuzzy matching
- Use fuzzy match fields for identifying data that are similar
- Pay attentions to Recommendations from CLAIRE Engine
  - Suitable for exact match -> 'Name ', 'phone', 'SSN'
  - Unsuitable for exact match -> 'DOB', 'email', 'AddressLine1', 'City', 'ZiP'
  - Hot spot-> DOB, ZiP, City or State

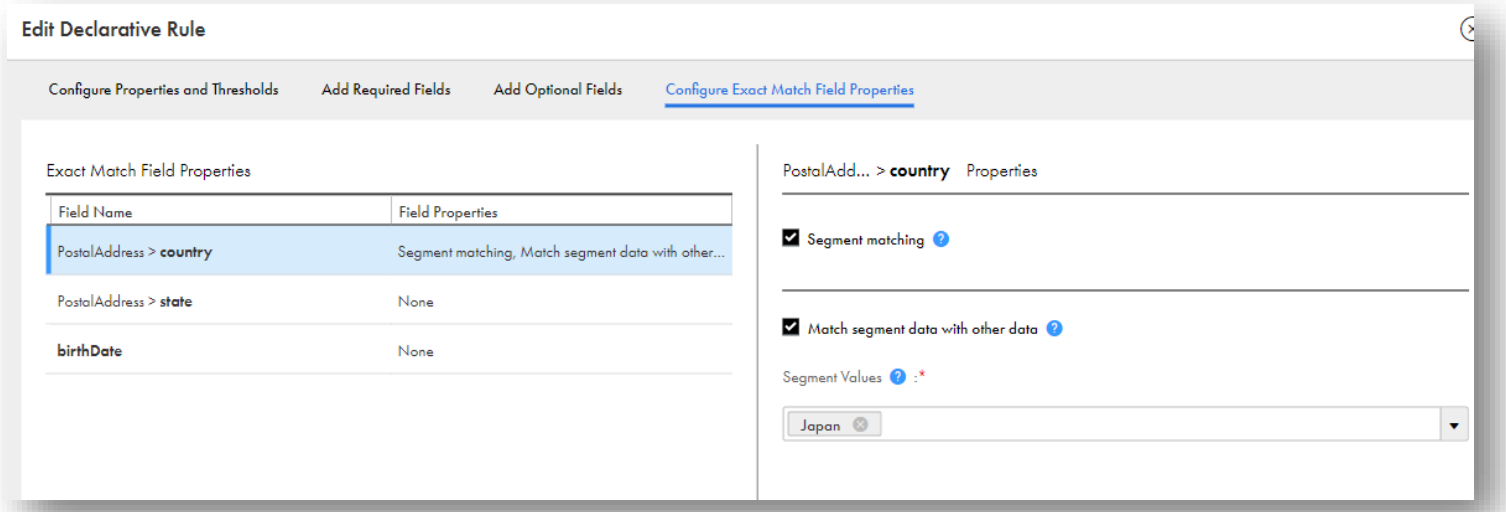


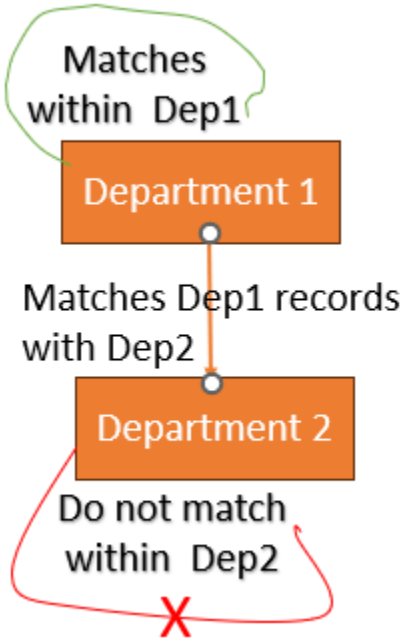
Match fields Key Points (Contd.)

RECID	BUSINESS_NAME	ADDRESS	POSTAL_AREA	PHONE	TAXID	CONTACT_NAME
3705154	Villa St Vincent Inc	516 WALSH ST OFC	567162750	555555342	9999992227	LYNN GELLER
3705164	Villa St Vincent Inc	516 WALSH ST OFC	567162750	555555342	9999992227	DIANE AMIOT
3705174	SISTERS ST BENDCT CRKSTN	516 WALSH ST OFC	567162750		9999992227	CORNELIA GUST
3705184	VILLA ST VINCENT SUMMIT	27148 233RD AVE SW	567169125	555555342	9999992227	JUDY HULST
3705194	SUMMIT	516 WALSH ST OFC	567162750	555555342	9999992227	REBECCA LANGLOIS
3705204	SUMMIT	516 WALSH ST OFC	567162750		9999992227	LORNA TOFSLEY

Segment Matching Key Points

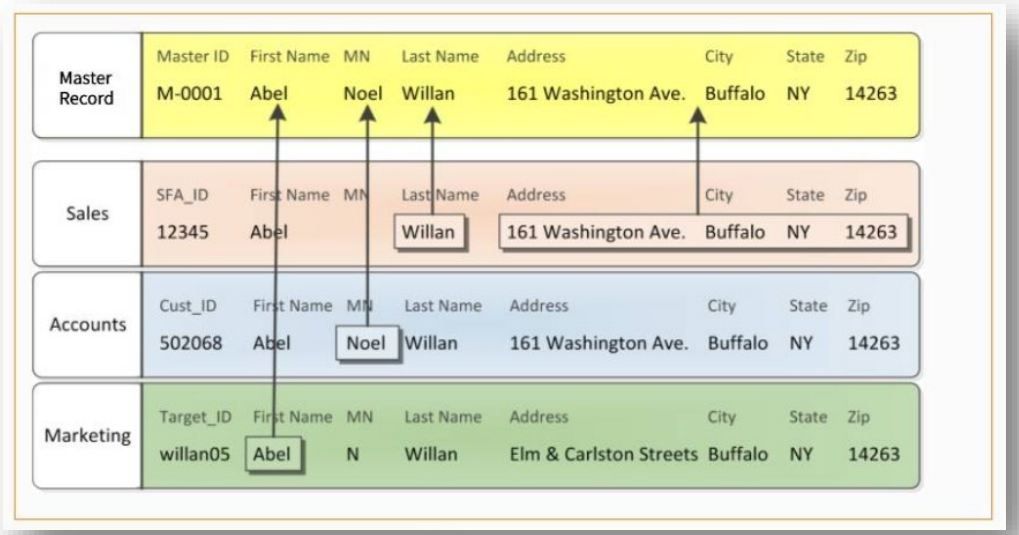
- Use Segment Matching to limit match rules to specific subsets of data (e.g., country code, dpt code).
- Segment matching can only be enabled on exact match fields
- For example, If you specify the segment value as Japan, the rule uses the records that have Japan as a country value for matching.
- Select Match segment data with other data to match the segment data with the rest of the data
- Achieve Distinct matching using workaround mentioned in Article

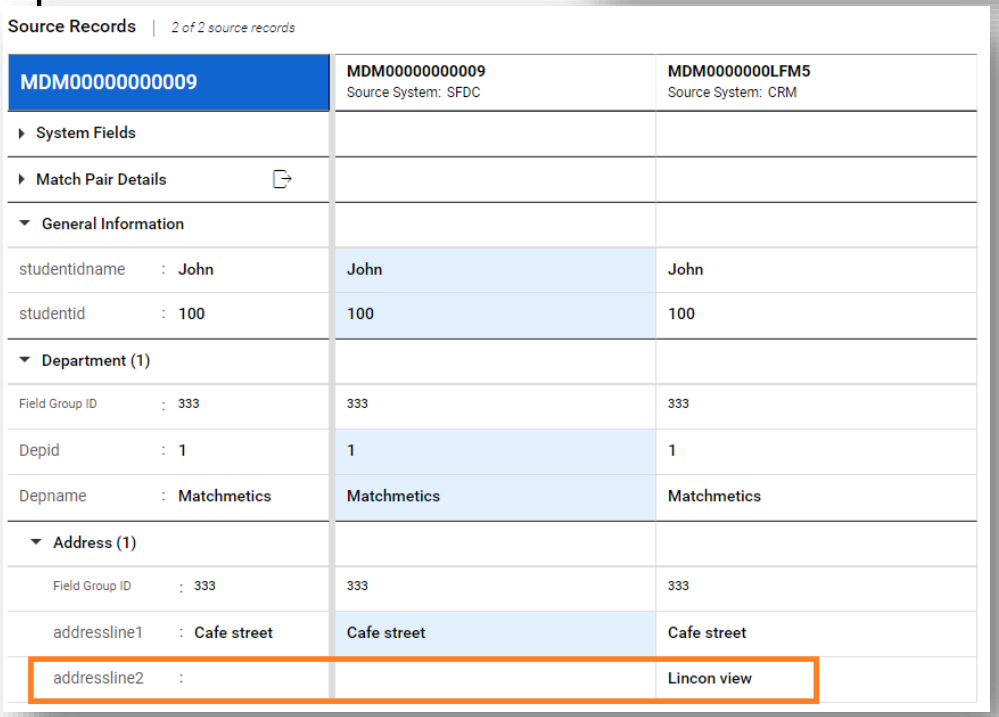




Survivorship Rule Key Points

- Implement survivorship rules to determine how conflicting data from matched records should be resolved
- Must be configured before records Ingress
- By default, Informatica customer 360 source as Rank 1 and Default system as Rank 2
- The ability to survive a field group as a block is available for preview from July

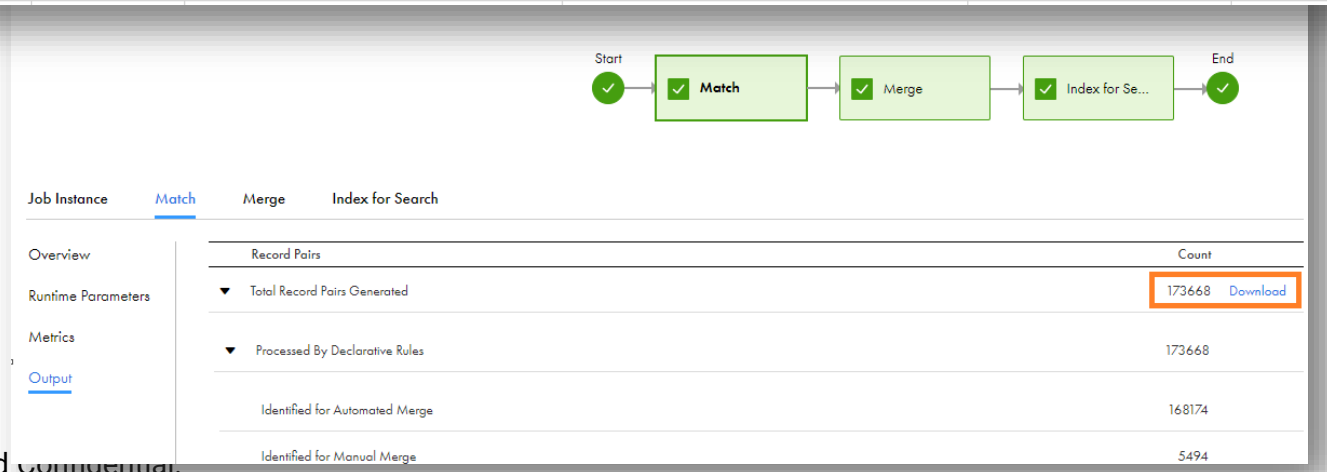




Match Report Key Points

- After an initial run, review and assess the match output
- Match report with 150K pairs information can be downloaded and reach out to GCS for complete report.

Record I	Rec	Matched Record ID	Mat	Match Score	Match Outcome	Record PostalAddress.addressLine1	Matched Record PostalAddress.addressLine1	Record PostalAddress.city	Matched Record PostalAddress.city
444E6B916	bvd	A43696841A3E89237CF03	sap	87	REVIEW	BEBELSBERGSTR. 27	Bebelsbergstr.	HOLZGERLINGEN	Holzgerlingen
444EF71951	bvd	29A0C8136C6E21AE2E478	sap	99	AUTO_MERGE	GARBSENER LANDSTR. 10	Garbsener Landstr.	HANNOVER	Hannover
445097DBE	bvd	7FFE778DC36CAC6B5240	sap	98	AUTO_MERGE	BISCHOFSWEG 2 B	Bischofsweg	WERMSDORF	Wermsdorf
4450A1A7E	bvd	FCF3BEF293FF67E733C7B	sap	98	AUTO_MERGE	DANZIGER STR. 8	Danziger Str.	BOUS	Bous
445143F87	bvd	6C73079F86F5EEF63ACFB	sap	85	REVIEW	BETTENHAEUSER STR. 3	BettenhÃƒÂƒÃ‚Â¤user Str.	LICH	Lich
445163BC9	bvd	2DCC48EF1DF820A7E09C	sap	98	AUTO_MERGE	BAHNHOFSTR. 8 A	Bahnhofstr.	KIRCHLINTELN	Kirchlinteln
44557C33A	bvd	E5A6FD678278DCB3405D	sap	96	AUTO_MERGE	WERNER-VON-SIEMENS-STR. 5	Werner-von-Siemens-Str.	KOELLEDA	KÃƒÂƒÃ‚Â¶lleda
4457F69FD	bvd	DA9183915DB01A5090E9	sap	96	AUTO_MERGE	LUDWIG-RICHTER-STR. 6	Ludwig-Richter-Str.	WUPPERTAL	Wuppertal
44595D080	bvd	84438F6545F23D89EF7EC	sap	90	AUTO_MERGE	RHEINSTR. 9	Rheinstr.	TELTOW	Teltow
445A4B539	bvd	46CE026D6755CA92B01C	sap	94	AUTO_MERGE	CARL-ZEISS-STR. 18	Carl-Zeiss-Str.	METZINGEN	Metzingen
445AA0DE9	bvd	698D6808140A62A7CD88	sap	96	AUTO_MERGE	UNTERFELDRING 3 A	Unterfeldring	VIERKIRCHEN	Vierkirchen
445B037A6	bvd	AC10EC56A6B760939002	sap	96	AUTO_MERGE	ZUM ELBERSKAMP 24	Zum Elberskamp	FINNENTROP	Finnentrop
445BC0157	bvd	B88F7B27D725920AE40C	sap	99	AUTO_MERGE	MAYBACHSTR. 30	Maybachstr.	SCHLIERBACH	Schlierbach
445D21AF3	bvd	50F874FE023322098511F	sap	99	AUTO_MERGE	MONT-CENIS-STR. 357	Mont-Cenis-Str.	HERNE	Herne
445D4899B	bvd	CAAB5E778CAA7792E926	sap	95	AUTO_MERGE	LAUBACHER WEG 31	Laubacher Weg	GRUENBERG	GrÃƒÂƒÃ‚Â¼nberg



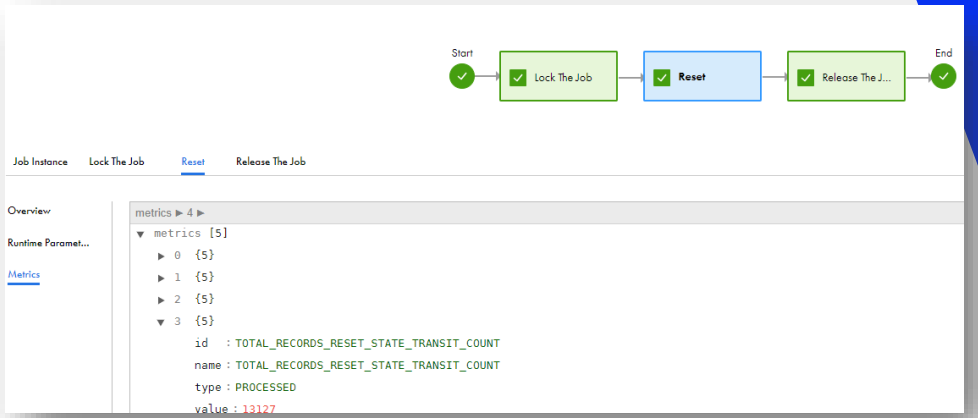
Informatica

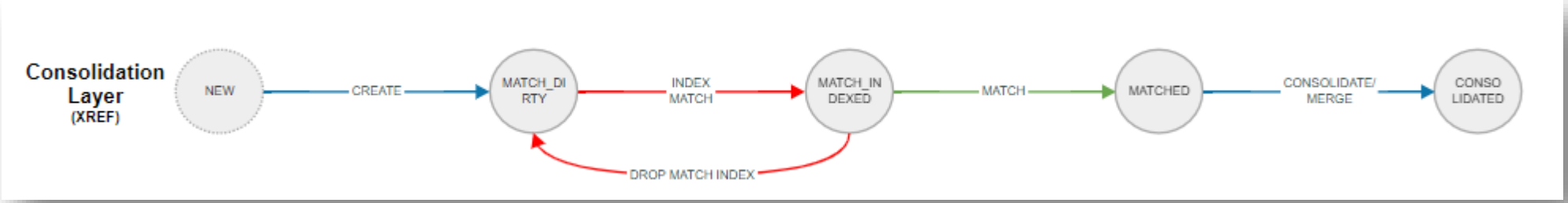
Match Operation Key Points

- In MDM SAAS Matching happens at XREF level and not on Master level
- If the record pair meets the conditions set by the first match rule, then it doesn't process further.
- Null to Null and Null to Non-Null matching available in August release, on demand.
- By default, the match process identify up to first 10,000 candidates for each record to improve the performance and accuracy
- Currently 1 master can only have 1000 Source XREF records

Reset Job Key Points

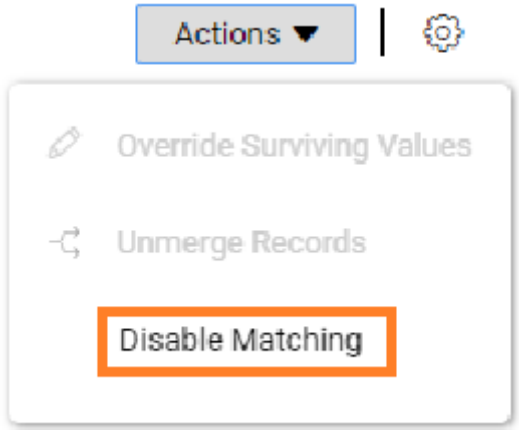
- The Reset step resets the already matched records.
- Reset job is not recommended for Production environments
- The record state changes from MATCHED to MATCHED_INDEXED
- Reset job doesn't remove the CONSOLIDATED and MATCH_INDEXED records.
- Already merged/consolidated records are not unmerged. It affects only the cross-reference state

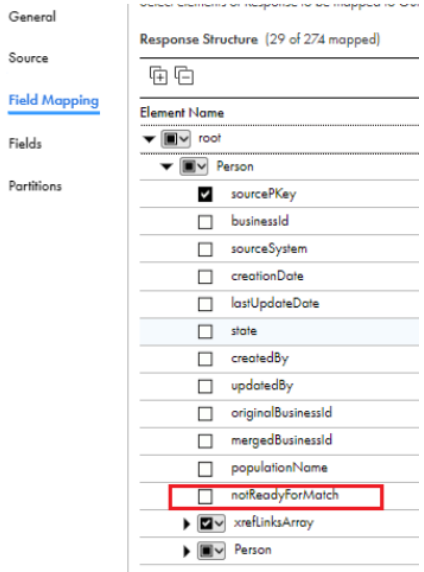




Job Definitions Key Points

- During match tuning, select Match Only when executing match. This will allow you to reset the data for re-match in case you make match configuration changes.
- If required, ensure to rerun key generation after making match changes
- Use NotReadyForMatch to specify whether a source record can participate in the match process.
- Use NotReadyForMatch to avoid a large cluster record from participating in match Via API, UI and CDI Ingress.
- Use search match instead of Match Job for real time matching.



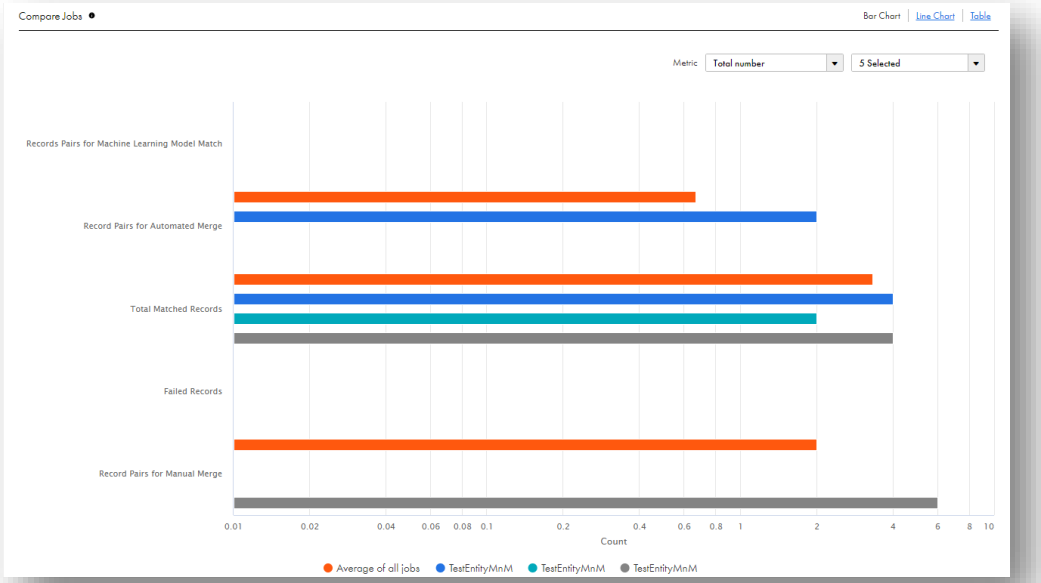


Operational Insights Key Points

- Use Operational Insights to view data processing analytics and monitoring statistics for MDM SaaS.
- You can analyze the match job metrics and decide whether to modify the match model configuration.
- You can analyze the merge job metrics to determine the number of record pairs and record pair groups that were created by the job.
- You can compare two to five jobs and view their key metrics.

Jobs (95) | 5 Selected | Find | Compare

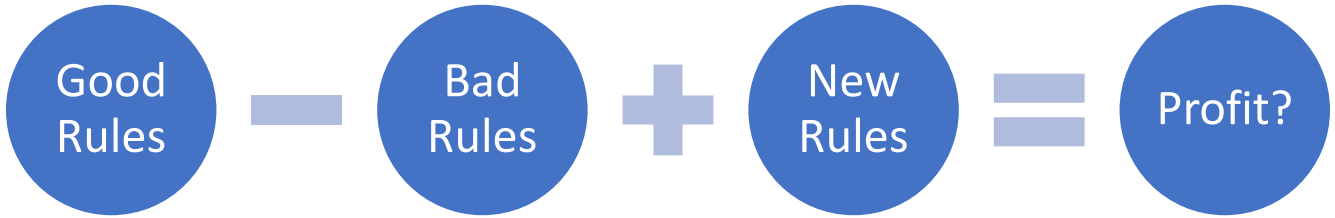
Job Name	Duration (hrs)	Start time	Records read	Records failed	Records updated	Name3 keys generated	Search records read
<match for person...	5.8	Jan 01, 2023, 5:40:00 PM	1000000	30000	30000	30000	30000
<job name 2>	5.1	Jan 01, 2023, 4:40:00 PM	1000000	30000	30000	30000	30000
<job name 3>	3.9	Jan 01, 2023, 5:40:00 PM	1000000	30000	30000	30000	30000
<job name 4>	3.7	Jan 01, 2023, 7:40:00 PM	1000000	30000	30000	30000	30000
<job name 5>	3.5	Jan 01, 2023, 8:40:00 PM	1000000	30000	30000	30000	30000
<job name 6>	5.8	Jan 01, 2023, 5:40:00 PM	1000000	30000	30000	30000	30000
<job name 7>	5.1	Jan 01, 2023, 4:40:00 PM	1000000	30000	30000	30000	30000
<job name 8>	3.9	Jan 01, 2023, 5:40:00 PM	1000000	30000	30000	30000	30000



Informatica

Tuning Exercise Key Points

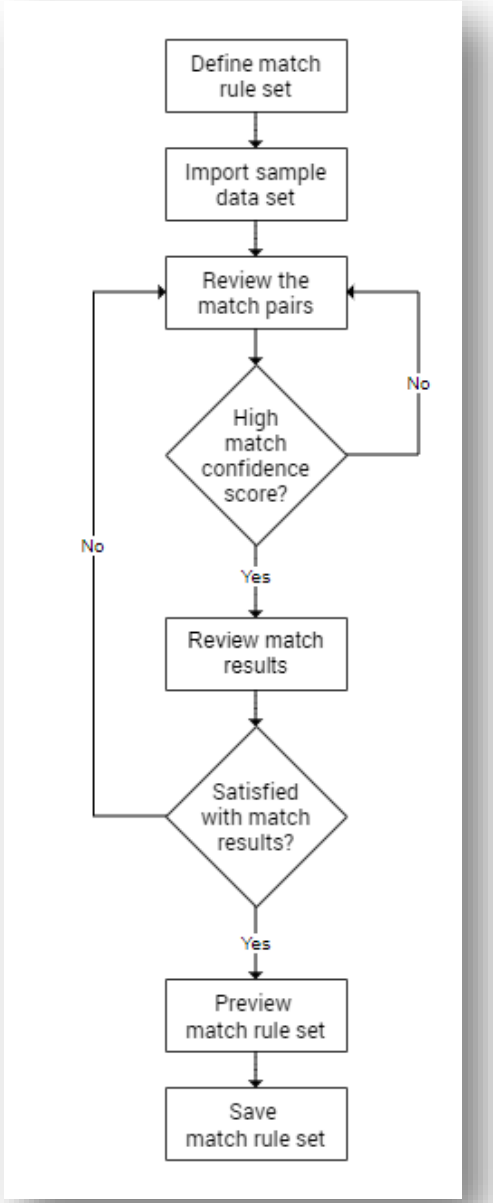
- **What rules do we start with?**



Starting Ruleset

- Rules that work already
- Rules that are troublesome today
- Rules utilizing new attributes

Don't overthink it! The more iterations you execute, the less important your starting point becomes.



Match Best Practice Checklist

- Know your Data
- Define Candidate Selection Criteria
- Identify Hotspots in Data
- Define Fuzzy Column
- Define Exact Column
- AUTO, Manual, Threshold rules
- Leverage Segment Matching

Informatica

Thank You

Informatica™