April 14th 2020

MDM Fuzzy Match Deep Dive

Augustin Chan achan@informatica.com
Development Architect, MDM ACE Team

Informatica

Agenda

- Necessary Background
- Match Job Internals
- Match Pair Processing Details
- Match Batch Distribution
- Understanding the Cleanse Log
- Performance Tips
- Q&A

Note: All logs and screenshots are from MDM 10.3 GA

A Tale of Two Records

Results:

	PARTY_ROWID	DISPLAY_NAME	ADDRESS_LINE_1	ADDRESS_ROWID
1	966	AUTOMOTION CORPORATION	225 BRAE BLVD	1021
2	991	AUTOMOTION	1740 BROADWAY	1046
3	991	AUTOMOTION	1740 BROADWAY	1069
4	991	AUTOMOTION	1740 BROADWAY	1068
5	991	AUTOMOTION	1740 BROADWAY	1053

Necessary Background

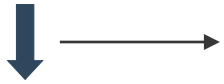
Fuzzy Keys Example

Example:

Automotion Corporation

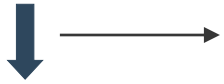
Character level cleaning

AUTOMOTION CORPORATION



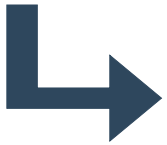
Edit List processing

AUTO MOTION



“Phonetics”

AT MATAN

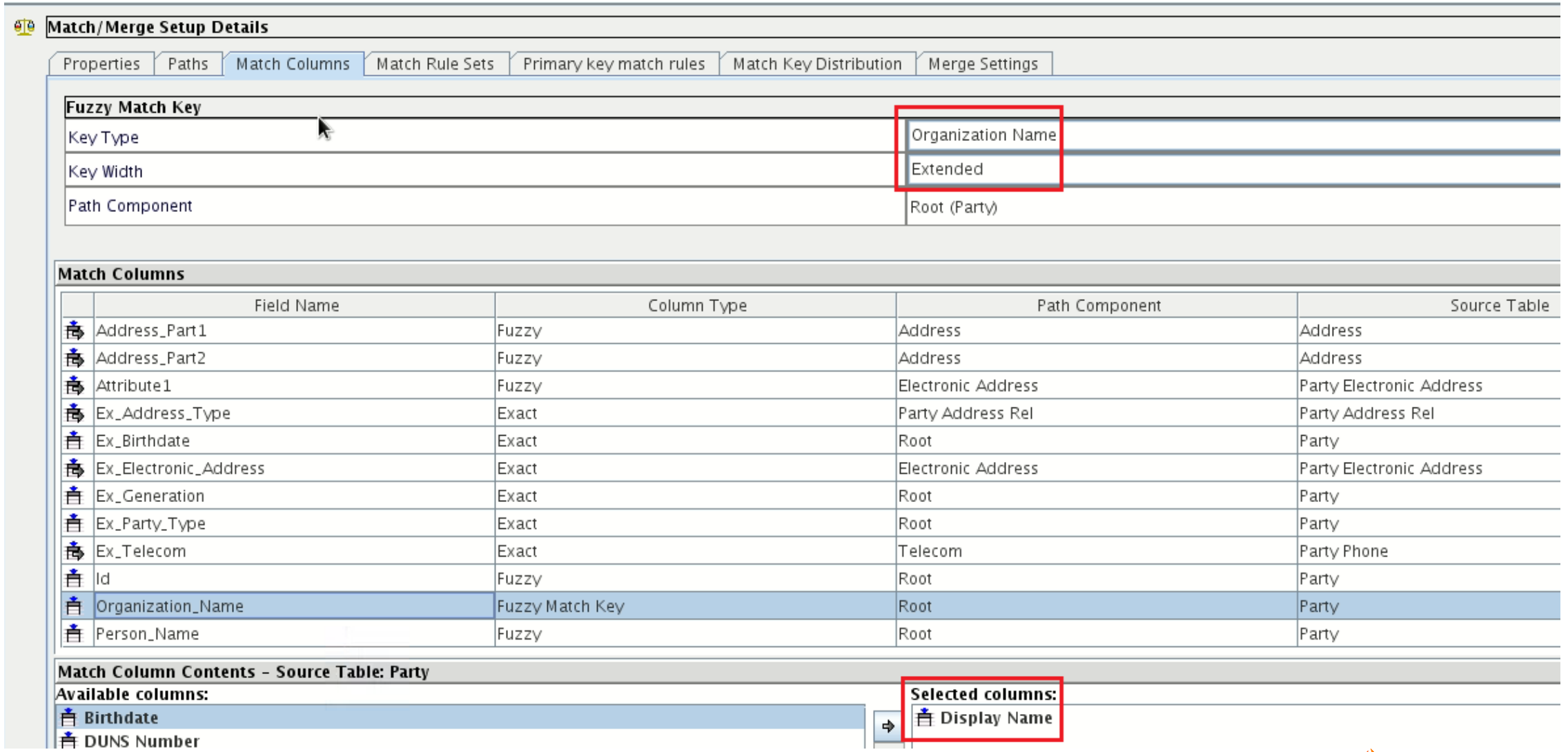


Key Building

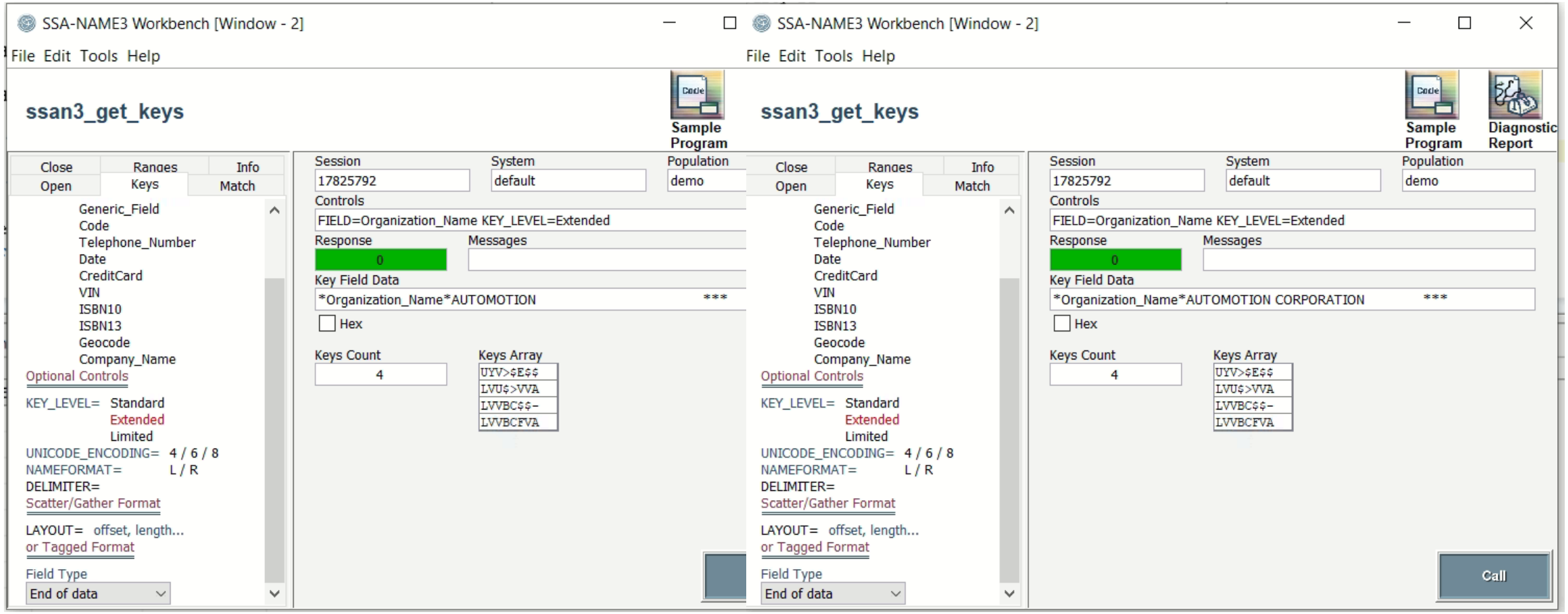
Key:	Based on:
UYV>$F$$	MOTION AUTO
LUU$>WVA	AUTO MOTION
LUVBC$$-	AUTO
LUVBCGVA	AUTOMOTION AUTO

Informatica

Hub Console – Key Level



Name3 Workbench Keys



Informatica

STRP Table Keys

```
select * from c_party_strp where rowid_object in ( '991', '966')
```

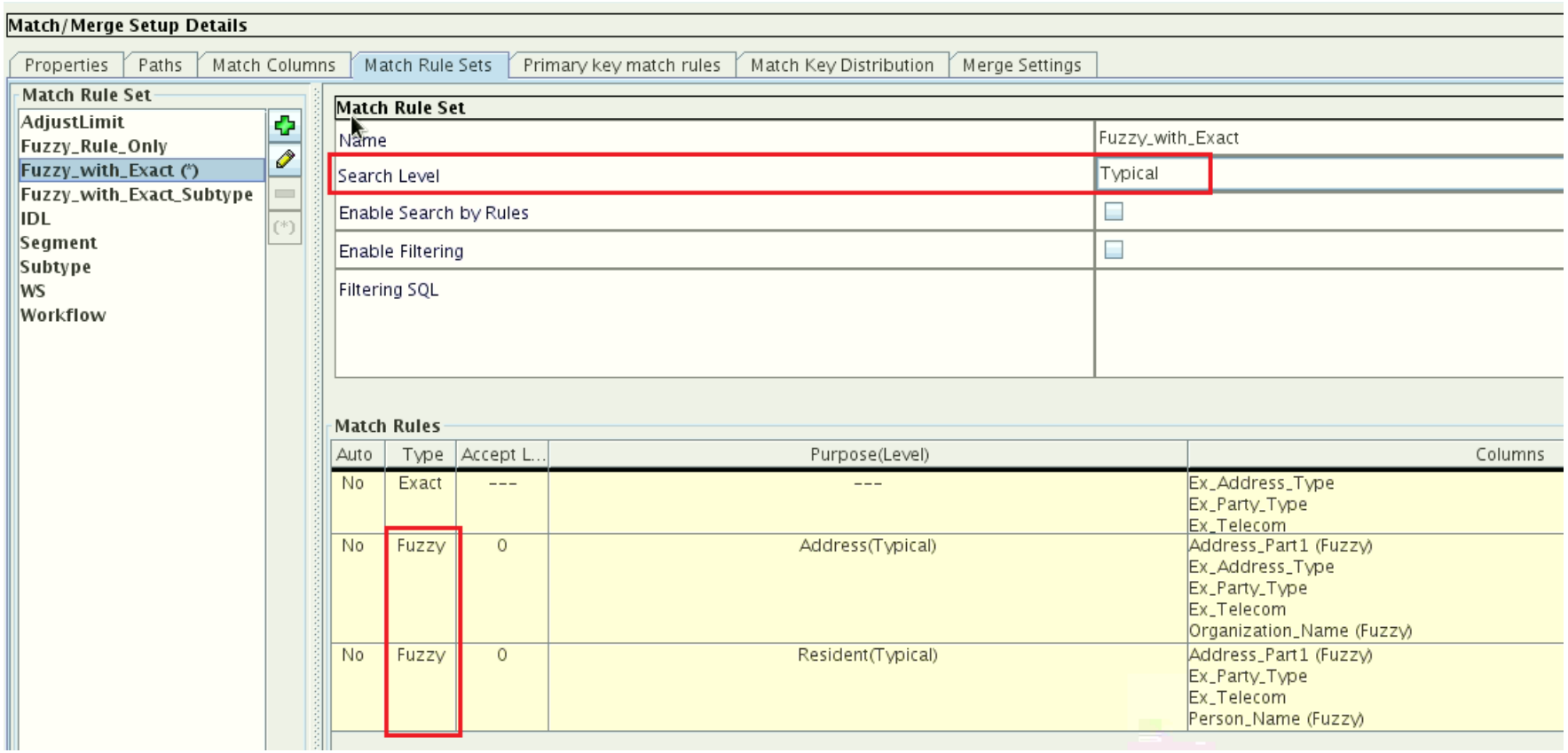
Results | Script Output | Explain | Autotrace | DBMS Output | OWA Output

Results:

	SSA_KEY	ROWID_OBJECT	DATA_ROW	DATA_COUNT	SSA_DATA	
1	LVU$>VVA	966	1	1	a05	AUTOMOTION CORPORATION
2	LVVBC$$-	966	1	1	a05	AUTOMOTION CORPORATION
3	LVVBCFVA	966	1	1	a05	AUTOMOTION CORPORATION
4	UYV>$E$$	966	1	1	a05	AUTOMOTION CORPORATION
5	LVU$>VVA	991	1	1	a05	AUTOMOTION b000c00
6	LVVBC$$-	991	1	1	a05	AUTOMOTION b000c00
7	LVVBCFVA	991	1	1	a05	AUTOMOTION b000c00
8	UYV>$E$$	991	1	1	a05	AUTOMOTION b000c00

Informatica

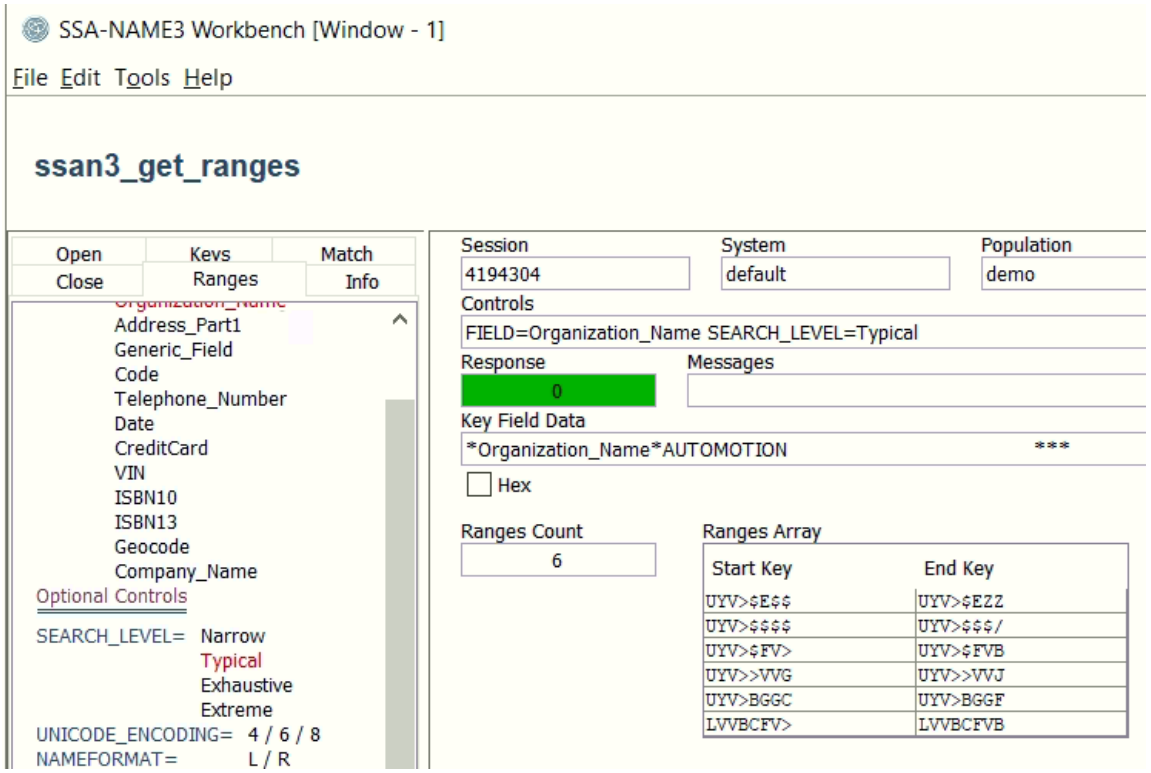
Hub Console – Search Level

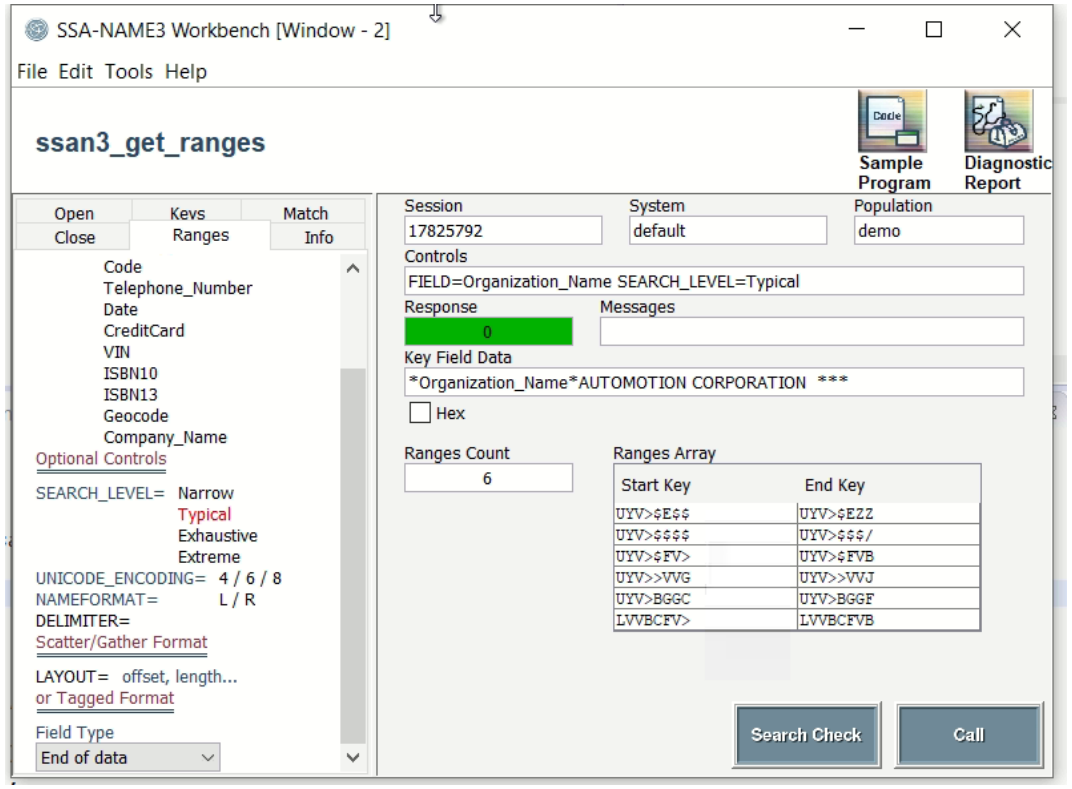


Search Ranges

- A range is a pair of 8 character strings
- Can be thought of as the fuzziness around a key
  - Give me all keys between `'UYV>$E$$' and 'UYV>$EZZ'`
- Ranges are not persisted in any table!
- Some ranges can be seen from ThreadMonitor, or Match Summary in cleanse log
- MDM generates ranges at runtime with an ssa call

Name3 Workbench – Search Ranges





Informatica

RangerWorker Summary – Top 10 Range Comparisons

```
[Ranger0] [INFO ] com.siperian.mrm.match.RangerWorker:
Top 10 Range Comparisons counts
Ranger0 Comparison Max Range 0 = 10 Q:2 DB:5 between 'UYV>$E$$' and 'UYV>$EZZ'
Ranger0 Comparison Max Range 1 = 8 Q:2 DB:4 between 'LVVBCFV>' and 'LVVBCFVB'
Ranger0 Comparison Max Range 2 = 0 Q:2 DB:0 between 'UYV>BGGC' and 'UYV>BGGF'
Ranger0 Comparison Max Range 3 = 0 Q:2 DB:0 between 'UYV>>VVG' and 'UYV>>VVJ'
Ranger0 Comparison Max Range 4 = 0 Q:2 DB:0 between 'UYV>$FV>' and 'UYV>$FVB'
Ranger0 Comparison Max Range 5 = 0 Q:2 DB:0 between 'UYV>$$$$' and 'UYV>$$$/'
Ranger0 Comparison Max Range 6 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Comparison Max Range 7 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Comparison Max Range 8 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Comparison Max Range 9 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Total Ranges Processed  = 6
Ranger0 Total Candidates        = 14
Ranger0 Total Matches           = 1
Matcher Summary :total_calls: 14 SSA Matches: 14
```

‘Candidates’ really means candidate comparisons done by this thread (Ranger0).

SSA Matches = SSA calls

Informatica

Range Queries and DB Counts

```
select * from c_party_strp where ssa_key between 'UYV>$E$$' and 'UYV>$EZZ'   5
select * from c_party_strp where ssa_key between 'LVVBCFV>' and 'LVVBCFVB'   4

select * from c_party_strp where ssa_key between 'UYV>BGGC' and 'UYV>BGGF'   0
select * from c_party_strp where ssa_key between 'UYV>>VVG' and 'UYV>>VVJ'   0
select * from c_party_strp where ssa_key between 'UYV>$FV>' and 'UYV>$FVB'   0
select * from c_party_strp where ssa_key between 'UYV>$$$$' and 'UYV>$$$/'   0
```

Results | Script Output | Explain | Autotrace | DBMS Output | OWA Output

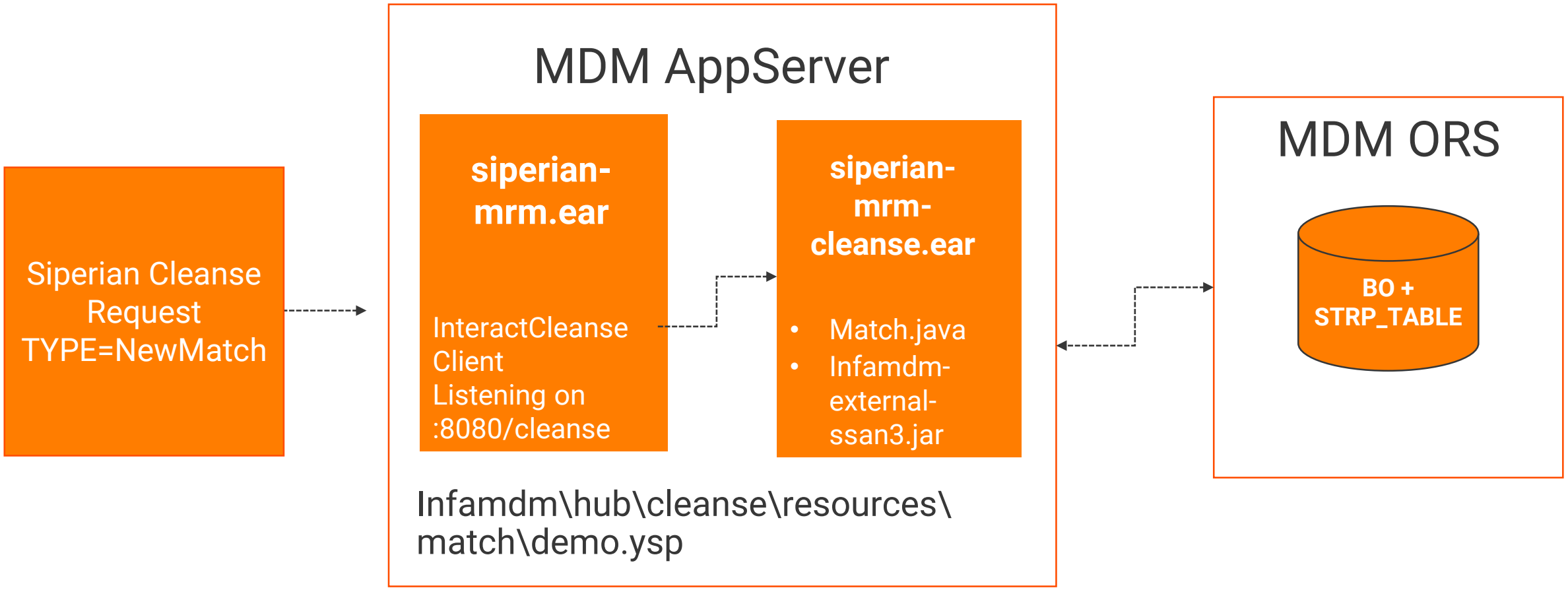
Results:

	SSA_KEY	ROWID_OBJECT	DATA_ROW	DATA_COUNT	SSA_DATA
1	UYV>$E$$	1042	1	1	a050AUTOMOTION CORPORATION
2	UYV>$E$$	1044	1	1	a050AUTOMOTION CORPORATION
3	UYV>$E$$	966	1	1	a050AUTOMOTION CORPORATION
4	UYV>$E$$	991	1	1	a050AUTOMOTION b000c0(
5	UYV>$EBM	971	1	1	a050AUTOMOTION CAR RENTAL CORP

Informatica

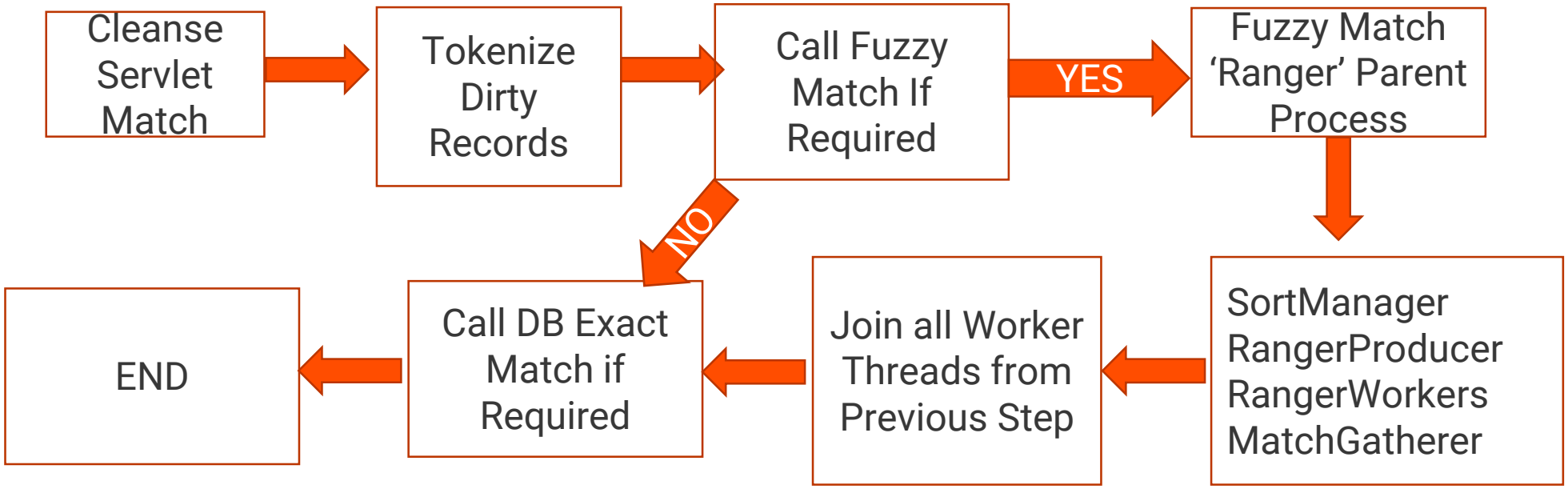
Match Job Internals

MDM Fuzzy Match Architecture



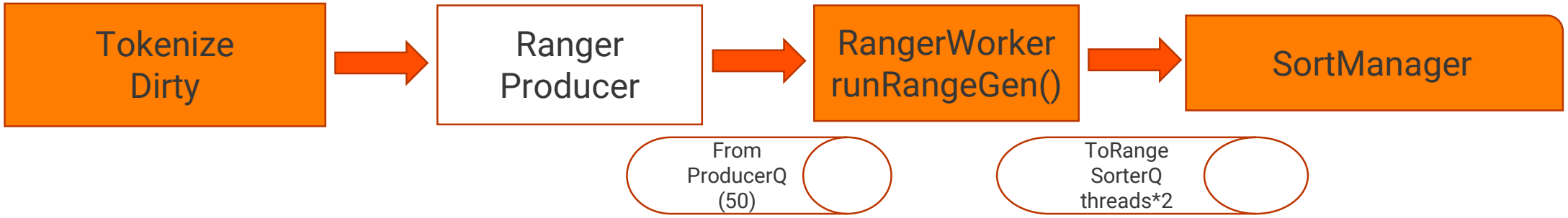
If cmx.server.match.distributed_match=1 then the job will be distributed across registered Process Servers.

Match Job Execution Overview



- Tokenize – If COMPLETE_STRIP_RATIO or STRIP_CTAS_DELETE_UPPER_LIMIT or STRIP_CTAS_DELETE_RATIO are exceeded, the entire STRP is rebuilt, with an exclusive lock on BO which prevents puts and merges!
  STRIP_CTAS merges existing and newly tokenized records into new STRP.

MDM Match Process – Data Prep, Range Gen



- TokenizeWorker – Generates fuzzy keys based on dirty records in Base Object. (**ssan3_get_keys_encoded**)
- RangerProducer – Reads data from _STRP and packages them into RangerNodeTransports containing 100 rangerNodes each (1 STRP row per rangerNode)
- RangerWorker – runRangeGen() reads these RangerNodeTransports fromProducerQ and calls **ssan3_get_ranges_encoded** to assign all search ranges within work range to their rangerNodes. These processed rangerNodeTransports are placed onto the ToRangeSorterQ for sorting.

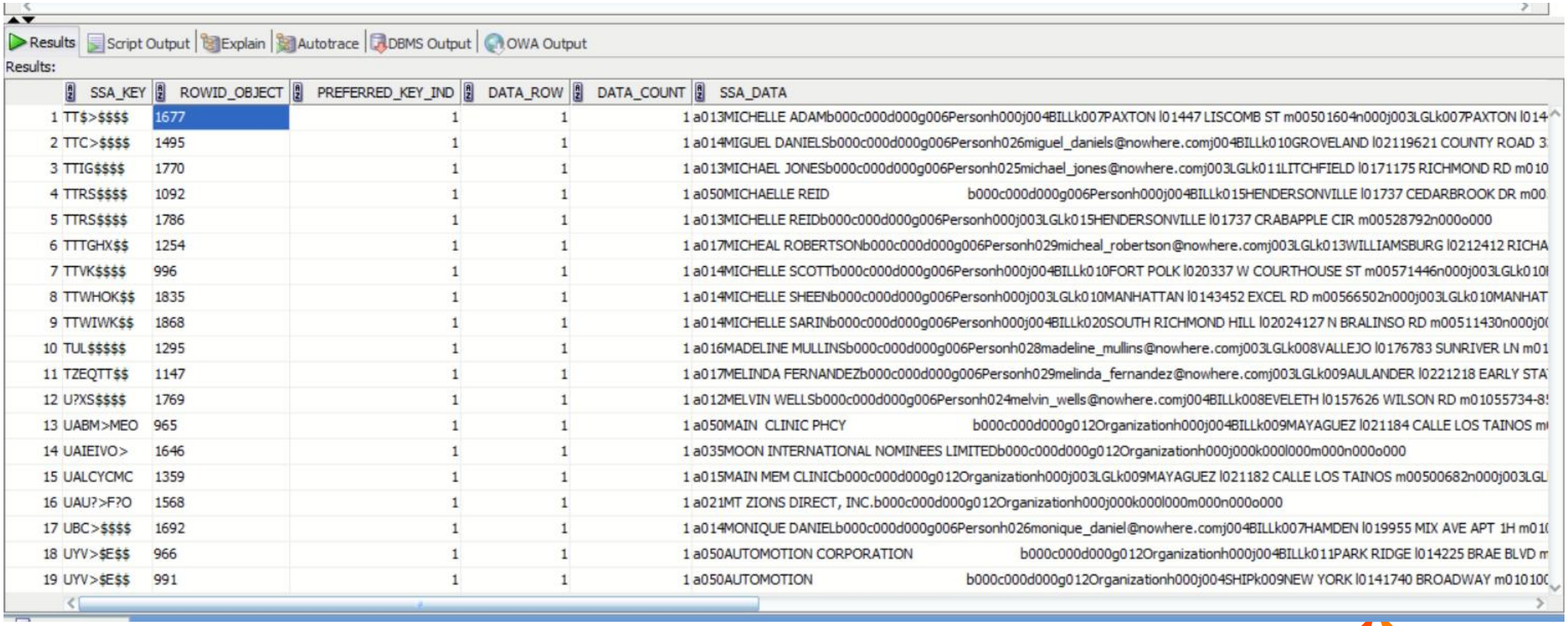
For Distributed matching, only Ranges that fall within the work range for that Process Server are processed by the downstream SortManager and placed on the work queue.

Boxes in ORANGE are multi-threaded

Informatica

RangerProducer – Prepare for Range Generation

- Reads _STRP rows where PREFERRED_KEY_IND = 1 and packages them into 100 rangerNode chunks inside RangerNodeTransports (1 row per rangerNode)



Results | Script Output | Explain | Autotrace | DBMS Output | OWA Output

Results:

	SSA_KEY	ROWID_OBJECT	PREFERRED_KEY_IND	DATA_ROW	DATA_COUNT	SSA_DATA
1	TT$>$$$$	1677	1	1	1	a013MICHELLE ADAMb000c000d000g006Personh000j004BILLk007PAXTON l01447 LISCOMB ST m00501604n000j003LGLk007PAXTON l014
2	TTC>$$$$	1495	1	1	1	a014MIGUEL DANIELSb000c000d000g006Personh026miguel_daniels@nowhere.comj004BILLk010GROVELAND l02119621 COUNTY ROAD 3
3	TTIG$$$$	1770	1	1	1	a013MICHAEL JONESb000c000d000g006Personh025michael_jones@nowhere.comj003LGLk011LITCHFIELD l0171175 RICHMOND RD m010
4	TTRS$$$$	1092	1	1	1	a050MICHAELLE REID b000c000d000g006Personh000j004BILLk015HENDERSONVILLE l01737 CEDARBROOK DR m00
5	TTRS$$$$	1786	1	1	1	a013MICHELLE REIDb000c000d000g006Personh000j003LGLk015HENDERSONVILLE l01737 CRABAPPLE CIR m00528792n000o000
6	TTTGHX$$	1254	1	1	1	a017MICHEAL ROBERTSONb000c000d000g006Personh029micheal_robertson@nowhere.comj003LGLk013WILLIAMSBURG l0212412 RICHA
7	TTVK$$$$	996	1	1	1	a014MICHELLE SCOTTb000c000d000g006Personh000j004BILLk010FORT POLK l020337 W COURTHOUSE ST m00571446n000j003LGLk010l
8	TTWHOK$$	1835	1	1	1	a014MICHELLE SHEENb000c000d000g006Personh000j003LGLk010MANHATTAN l0143452 EXCEL RD m00566502n000j003LGLk010MANHAT
9	TTWIWK$$	1868	1	1	1	a014MICHELLE SARINb000c000d000g006Personh000j004BILLk020SOUTH RICHMOND HILL l02024127 N BRALINSO RD m00511430n000j0(
10	TUL$$$$$	1295	1	1	1	a016MADELINE MULLINSb000c000d000g006Personh028madeline_mullins@nowhere.comj003LGLk008VALLEJO l0176783 SUNRIVER LN m01
11	TZEQTT$$	1147	1	1	1	a017MELINDA FERNANDEZb000c000d000g006Personh029melinda_fernandez@nowhere.comj003LGLk009AULANDER l0221218 EARLY STA
12	U?XS$$$$	1769	1	1	1	a012MELVIN WELLSb000c000d000g006Personh024melvin_wells@nowhere.comj004BILLk008EVELETH l0157626 WILSON RD m01055734-8
13	UABM>MEO	965	1	1	1	a050MAIN CLINIC PHCY b000c000d000g012Organizationh000j004BILLk009MAYAGUEZ l021184 CALLE LOS TAINOS m
14	UAIEIVO>	1646	1	1	1	a035MOON INTERNATIONAL NOMINEES LIMITEDb000c000d000g012Organizationh000j000k000l000m000n000o000
15	UALCYCMC	1359	1	1	1	a015MAIN MEM CLINICb000c000d000g012Organizationh000j003LGLk009MAYAGUEZ l021182 CALLE LOS TAINOS m00500682n000j003LGL
16	UAU?>F?O	1568	1	1	1	a021MT ZIONS DIRECT, INC.b000c000d000g012Organizationh000j000k000l000m000n000o000
17	UBC>$$$$	1692	1	1	1	a014MONIQUE DANIELb000c000d000g006Personh026monique_daniel@nowhere.comj004BILLk007HAMDEN l019955 MIX AVE APT 1H m01(
18	UYV>$E$$	966	1	1	1	a050AUTOMOTION CORPORATION b000c000d000g012Organizationh000j004BILLk011PARK RIDGE l014225 BRAE BLVD m
19	UYV>$E$$	991	1	1	1	a050AUTOMOTION b000c000d000g012Organizationh000j004SHIPk009NEW YORK l0141740 BROADWAY m010100

Informatica

RangerWorker Range Generation and SortManager

RangerWorker Range Generation

```
Automotion Corporation
UYV>$E$$        UYV>$EZZ
UYV>$$$$        UYV>$$$/
UYV>$FV>        UYV>$FVB
UYV>>VVG        UYV>>VVJ
UYV>BGGC        UYV>BGGF
LVVBCFV>        LVVBCFVB

Automotion
UYV>$E$$        UYV>$EZZ
UYV>$$$$        UYV>$$$/
UYV>$FV>        UYV>$FVB
UYV>>VVG        UYV>>VVJ
UYV>BGGC        UYV>BGGF
LVVBCFV>        LVVBCFVB
```

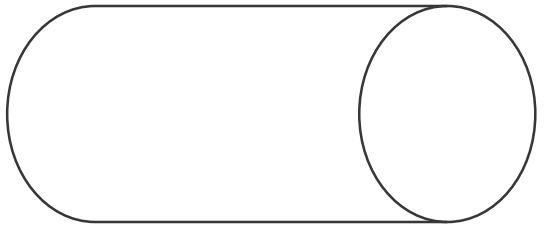
2 Records in match batch generate 12 total Ranges

SortManager packaging RangerNodes

```
UYV>$E$$    UYV>$EZZ
    Automotion Corporation
    Automotion
UYV>$$$$    UYV>$$$/
    Automotion Corporation
    Automotion
UYV>$FV>    UYV>$FVB
    Automotion Corporation
    Automotion
UYV>>VVG    UYV>>VVJ
    Automotion Corporation
    Automotion
UYV>BGGC    UYV>BGGF
    Automotion Corporation
    Automotion
LVVBCFV>    LVVBCFVB
    Automotion Corporation
    Automotion
```

Sorted into 6 RangerNodes

RangerNodes placed onto work queue to be processed by RangerWorkers



Each RangerNode contains max 3000 search records (max_records_per_rangernode cleanse prop)

Informatica™

Ranges vs RangerNodes

RANGE GENERATION

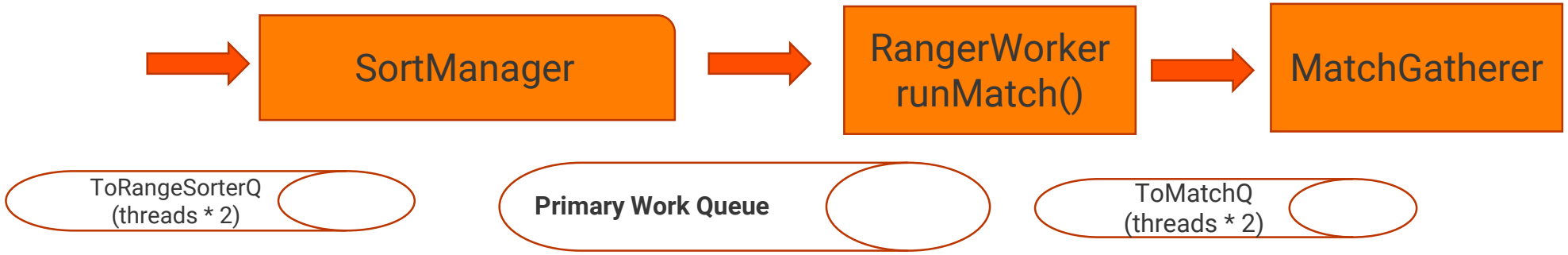
```
[2020-03-28 00:49:58,170] [Ranger0] [DEBUG] com.siperian.mrm.match.RangerWorker:
Total Records read    = 2
[2020-03-28 00:49:58,170] [Ranger0] [DEBUG] com.siperian.mrm.match.RangerWorker:
Total Ranges created  = 12 (Range+Record combinations)
[2020-03-28 00:49:58,170] [Ranger0] [DEBUG] com.siperian.mrm.match.RangerWorker:
Time Range Generation = 90 ms
```

SORTING+PACKAGING

```
[2020-03-28 00:49:58,175] [RangeSorter] [DEBUG]
com.siperian.mrm.match.SortManager: Sort of 12 Records Done:8ms
[2020-03-28 00:49:58,175] [RangeSorter] [DEBUG]
com.siperian.mrm.match.SortManager: ship the sorted ranges in memory  back to
the workers to match with
[2020-03-28 00:49:58,266] [RangeSorter] [DEBUG]
com.siperian.mrm.match.SortManager: Time to distribute, from memory,  12 ranges
with 6 candidates = 91 ms (RangerNodes)
```


Informatica

MDM Match Process – Sorting, Matching

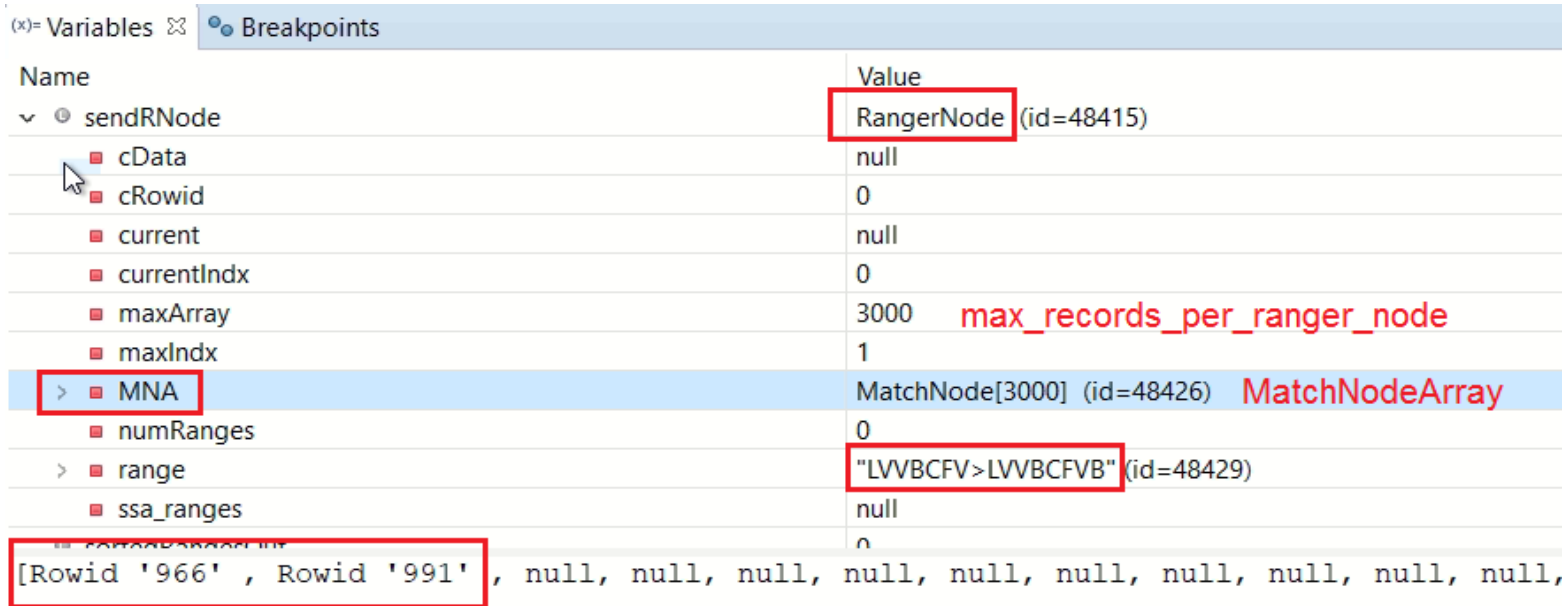


- SortManager – Extracts the rangeGen() rangerNodes from transports, sorts them, and creates new transports each with a single rangerNode that contains up to 3000 match records in their matchNodeArray (cleanse property **max_records_per_ranger_node**).
- RangerWorker – Performs Fuzzy match processing (**ssan3_match_encoded)** and exact string comparisons for exact fields. Threads = number of cleanse threads.
- MatchGatherer – gathers and removes duplicate match rows produced by the RangerWorkers, persists them to temp file for loading into mtch table

Match Processing Details

RangerNodes, Rulesets, and Match Pairs

RangerNode Contents



Each element of the MatchNodeArray has the STRP data for its rowid:

Rowid 966: `a050AUTOMOTION CORPORATION b000c000d000g012Organizationh000j004BILLk011PARK RIDGE l014225 BRAE BLVD m01007656-1870n000o0135551212`

Rowid 991: `a050AUTOMOTION b000c000d000g012Organizationh000j004SHIPk009NEW YORK l0141740 BROADWAY m01010019-4315n000j004SHIPk009NEW YORK l0141740 BROADWAY m01010019-4315n000j004BILLk00l0141740 BROADWAY m000n000j004BILLk009NEW YORK l0141740 BROADWAY m01010019-4315n000o0135551212 o0137771111 o0135557890`

Informatica

RangerNode Processing Details

RangerNode

Search Range: e.g., 'UYV>$E$$' - 'UYV>$EZZ'

Match Node Array: Up to 3000 Search Records from match batch which generate that range

And other stuff…..

RangerNode Search Records
991
966

File Rowid	File SSA_KEY
1042	UYV>$E$$
1044	UYV>$E$$
966	UYV>$E$$
991	UYV>$E$$
971	UYV>$EBM

Outer Loop Range Query for File Records:

```
SELECT ROWID_OBJECT, DATA_COUNT, SSA_DATA,
DATA_ROW FROM " +
stripTableName +
" WHERE SSA_KEY BETWEEN ? AND ? "+
" AND INVALID_IND = 0"+
" GROUP BY ROWID_OBJECT, DATA_COUNT, SSA_DATA,
DATA_ROW " +
" ORDER BY ROWID_OBJECT,DATA_ROW");
```

Comparison Matrix (8 Actual Comparisons)	
1042	991
	966
1044	991
	966
966	991
	~~966~~
991	~~991~~
	966
971	991
	966

Informatica

Fuzzy Matching on Distinct ‘Ruleset Nodes’

```
[2020-04-06 15:23:42,459] [Ranger0] [DEBUG]
com.siperian.mrm.match.RangerWorker: Matcher Rules:
Ruleset 'Fuzzy_with_Exact' has 3 rule(s), Search Call:false
Exact Rule :1 AutoMerge Ind :false Asymetrical Ind:false
     Node Num :1 Exact Match, Match Column:'Ex_Party_Type'
Match Column Id:6 Anti Match Ind:false
     Node Num :0 Exact Match, Match
Column:'Ex_Address_Type' Match Column Id:9 Anti Match
Ind:false
     Node Num :2 Exact Match, Match Column:'Ex_Telecom'
Match Column Id:14 Anti Match Ind:false
Fuzzy Rule :2 AutoMerge Ind :false Asymetrical Ind:false
     Node Num :1 Exact Match, Match Column:'Ex_Party_Type'
Match Column Id:6 Anti Match Ind:false
     Node Num :0 Exact Match, Match
Column:'Ex_Address_Type' Match Column Id:9 Anti Match
Ind:false
     Node Num :2 Exact Match, Match Column:'Ex_Telecom'
Match Column Id:14 Anti Match Ind:false
     Node Num :3 SSA Matching on
          'Address_Part1 Address_Part1' Column Id:11
          'Organization_Name Organization_Name' Column Id:0
          Match Level:Typical Geocode Radius:0 Match
Purpose:Address
```

Match Rules

Auto	Type	Accept L...	Purpose(Level)	Columns
No	Exact	---	---	Ex_Address_Type Ex_Party_Type Ex_Telecom
No	Fuzzy	0	Address(Typical)	Address_Part1 (Fuzzy) Ex_Address_Type Ex_Party_Type Ex_Telecom Organization_Name (Fuzzy)
No	Fuzzy	0	Resident(Typical)	Address_Part1 (Fuzzy) Ex_Party_Type Ex_Telecom Person_Name (Fuzzy)

```
Fuzzy Rule :3 AutoMerge Ind :false Asymetrical Ind:false
     Node Num :1 Exact Match, Match Column:'Ex_Party_Type'
Match Column Id:6 Anti Match Ind:false
     Node Num :2 Exact Match, Match Column:'Ex_Telecom'
Match Column Id:14 Anti Match Ind:false
     Node Num :4 SSA Matching on
          'Address_Part1 Address_Part1' Column Id:11
          'Person_Name Person_Name' Column Id:5
          Match Level:Typical Geocode Radius:0 Match
Purpose:Resident
Total Nodes :12 Actual Nodes:5
```


Informatica

Ruleset Optimization for Fuzzy Match Pair Evaluation

- A ruleset (rs) node is either an exact match field (Java String.regionMatches) or an SSA Purpose and its fuzzy fields (ssa match call).
- MDM determines the distinct set of rs nodes across all rules in ruleset. Fuzzy fields are associated with their Purpose.
- Fuzzy rules are evaluated in order for a given search+file record match pair.
- For each fuzzy rule - exact rs nodes are evaluated first
- If any rs node evaluates as false - current rule is non-match. Other fuzzy rules with this rs node are removed from further evaluation.
- If any rs node evaluates as true – store true result in case this rs node is present in a subsequent fuzzy rule.
- Any Exact rules are processed in a later phase.

Internal Match Record (aka SearchNode/FileNode)

```
Rowid '991'
ORGANIZATION_NAME
    0 0, AUTOMOTION
EX_GENERATION
    0 0, SIP null
DATE
    0 0, SIP null
ID
    0 0, SIP null
EX_BIRTHDATE
    0 0, SIP null
PERSON_NAME
    0 0, AUTOMOTION
EX_PARTY_TYPE
    0 0, Organization
ATTRIBUTE1
    1 0,0, SIP null
EX_ELECTRONIC_ADDRESS
    1 0,0, SIP null
EX_ADDRESS_TYPE
    2 0,0,0, SHIP
    2 0,0,1, SHIP
    2 0,0,2, BILL
    2 0,0,3, BILL
ADDRESS_PART2
    3 0,0,0,0, NEW YORK
    3 0,0,1,1, NEW YORK
    3 0,0,2,2, SIP null
    3 0,0,3,3, NEW YORK
ADDRESS_PART1
    3 0,0,0,0, 1740 BROADWAY
    3 0,0,1,1, 1740 BROADWAY
    3 0,0,2,2, 1740 BROADWAY
    3 0,0,3,3, 1740 BROADWAY
POSTAL_AREA
    3 0,0,0,0, 10019-4315
    3 0,0,1,1, 10019-4315
    3 0,0,2,2, SIP null
    3 0,0,3,3, 10019-4315
POSTAL_SUB3
    3 0,0,0,0, SIP null
    3 0,0,1,1, SIP null
    3 0,0,2,2, SIP null
    3 0,0,3,3, SIP null
EX_TELECOM
    4 0,0,3,3,0, 5551212
    4 0,0,3,3,1, 7771111
    4 0,0,3,3,2, 5557890
TELEPHONE_NUMBER
    4 0,0,3,3,0, 5551212
    4 0,0,3,3,1, 7771111
    4 0,0,3,3,2, 5557890
```

```
Rowid '966'
ORGANIZATION_NAME
    0 0, AUTOMOTION CORPORATION
EX_GENERATION
    0 0, SIP null
DATE
    0 0, SIP null
ID
    0 0, SIP null
EX_BIRTHDATE
    0 0, SIP null
PERSON_NAME
    0 0, AUTOMOTION CORPORATION
EX_PARTY_TYPE
    0 0, Organization
ATTRIBUTE1
    1 0,0, SIP null
EX_ELECTRONIC_ADDRESS
    1 0,0, SIP null
EX_ADDRESS_TYPE
    2 0,0,0, BILL
ADDRESS_PART2
    3 0,0,0,0, PARK RIDGE
ADDRESS_PART1
    3 0,0,0,0, 225 BRAE BLVD
POSTAL_AREA
    3 0,0,0,0, 07656-1870
POSTAL_SUB3
    3 0,0,0,0, SIP null
EX_TELECOM
    4 0,0,0,0,0, 5551212
TELEPHONE_NUMBER
    4 0,0,0,0,0, 5551212
```


Informatica

Fuzzy Rules - Exact Multi-Field Handling in MDM

	Record 1	Record 2
EX_TELECOM	5557890	2225555
EX_TELECOM	5551212	5551212
EX_TELECOM	7771111	

We first check if file record string is null, then we check if string lengths are the same before we try to find string match

All combinations will be matched until an exact string match is found:

- 5557890 to 2225555
- 5551212 to 2225555
- 7771111 to 2225555
- 5557890 to 5551212
- 5551212 to 5551212 – exact match found. stop

We will do all 6 comparisons only if we cannot stop early with 100%.

SSA Layout for Fuzzy Match

Rowid 991

```
a050AUTOMOTION
b000c000d000g012Organizationh000j004SHIPk009NEW YORK l0141740 BROADWAY
m01010019-4315n000j004SHIPk009NEW YORK l01441740 BROADWAY m01010019-
4315n000j004BILLk0001014l740 BROADWAY m000n000j004BILLk009NEW YORK l01441740
BROADWAY m01010019-4315n000o0135551212      o0137771111      o0135557890

=Address_Part1,222,28,Address_Part2,196,18,Telephone_Number,598,26,Address_Part1
,336,28,Address_Part2,310,18,Telephone_Number,632,26,Address_Part1,432,28,Teleph
one_Number,666,26,Address_Part1,526,28,Address_Part2,500,18
```

NOTE: Layout string lengths are doubled because cmx.server.match.server_encoding=1

Match Call Pseudo-Code:

```
ssa.match(searchNode, searchLayout, fileNode, fileLayout)
```


Informatica

Multi-Field Handling for Match in SSA

	Record 1	Record 2
Address #1	1740 Broadway	1 Fern St
Address #2	6555 Quince Rd Ste 400	1740 Broadway
Address #3	12012 N Mo Pac Expy	109 E Main St

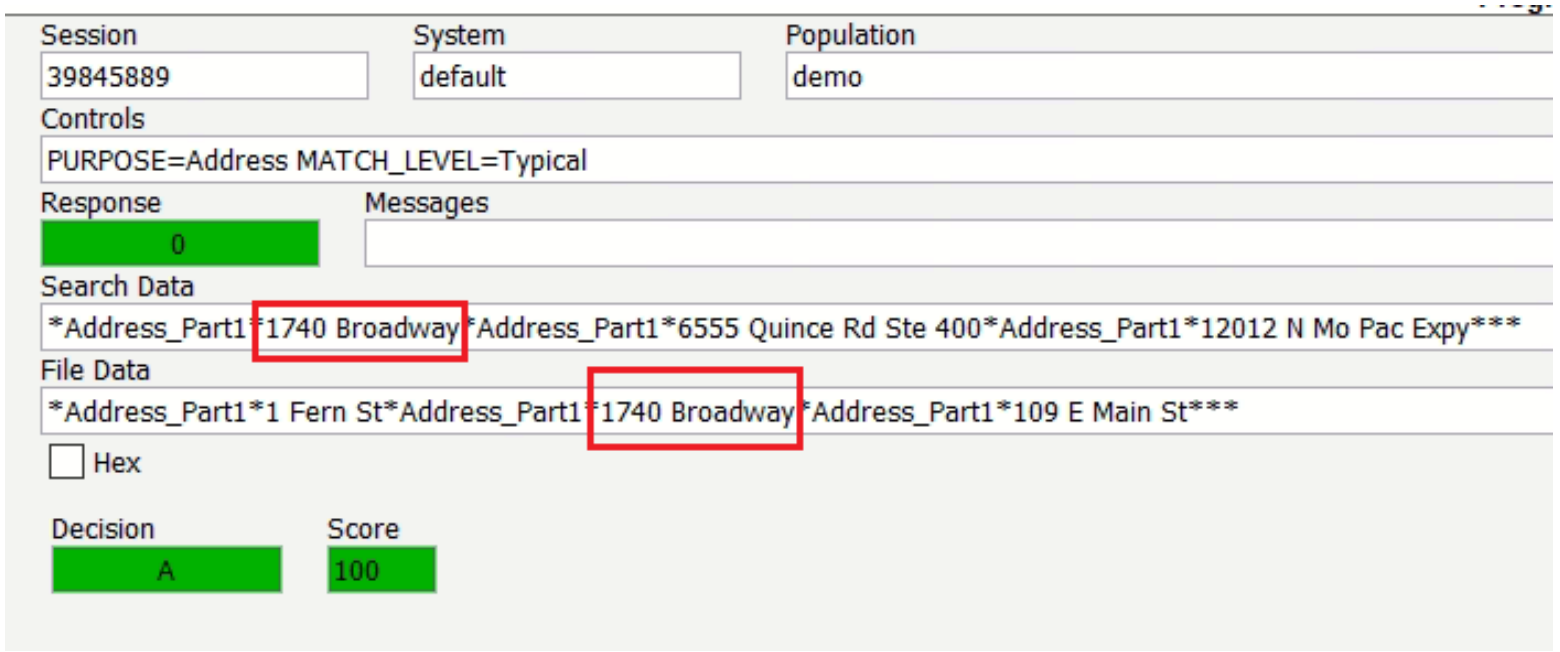
All combinations will be matched until a 100 pct match is found:

- 1740 Broadway to 1 Fern St
- 6555 Quince Rd Ste 400 To 1 Fern St
- 12012 N Mo Pac Expy To 1 Fern St
- 1740 Broadway to 1740 Broadway – 100 pct match, we stop

We will do all 9 comparisons only if we cannot stop early with 100%.
Only 1 combination needs to match for the records to match.

SSA Multi-Field Matching in Workbench

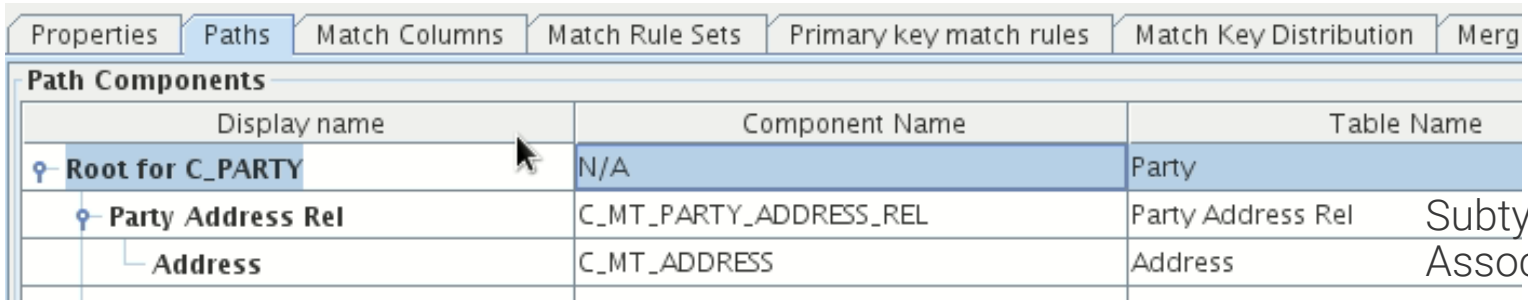
	Record 1	Record 2
Address #1	1740 Broadway	1 Fern St
Address #2	6555 Quince Rd Ste 400	1740 Broadway
Address #3	12012 N Mo Pac Expy	109 E Main St



Informatica

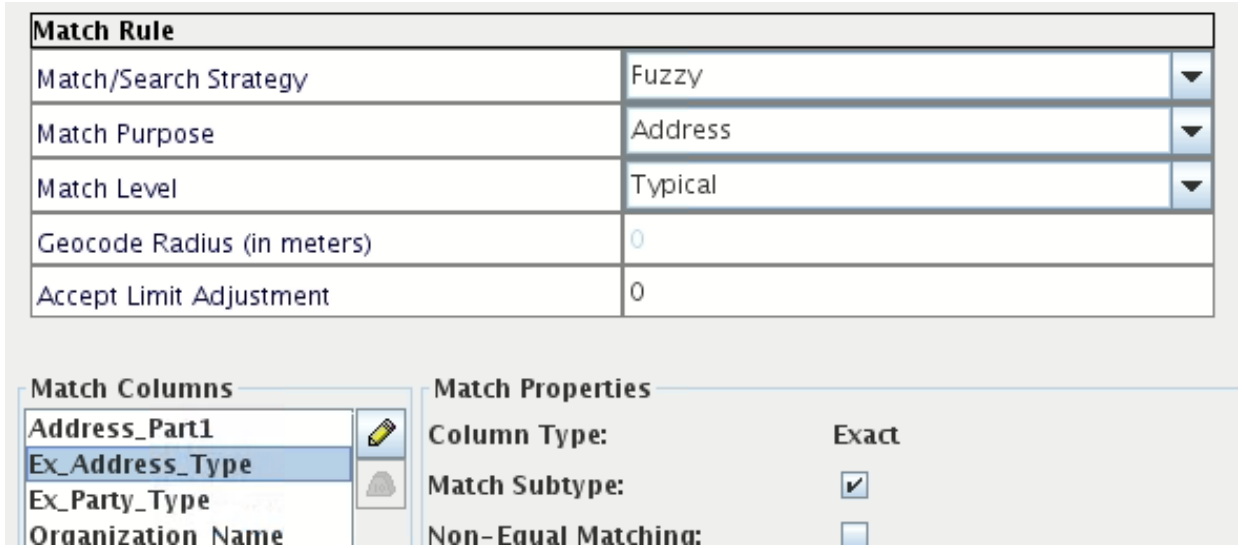
Subtype Match = Matching with Subsets of Child Data

```
Rowid '991'
ORGANIZATION_NAME
    0 0, AUTOMOTION
EX_GENERATION
    0 0, SIP null
DATE
    0 0, SIP null
ID
    0 0, SIP null
EX_BIRTHDATE
    0 0, SIP null
PERSON_NAME
    0 0, AUTOMOTION
EX_PARTY_TYPE
    0 0, Organization
ATTRIBUTE1
    1 0,0, SIP null
EX_ELECTRONIC_ADDRESS
    1 0,0, SIP null
EX_ADDRESS_TYPE
    2 0,0,0, SHIP
    2 0,0,1, SHIP
    2 0,0,2, BILL
    2 0,0,3, BILL
ADDRESS_PART2
    3 0,0,0,0, NEW YORK
    3 0,0,1,1, NEW YORK
    3 0,0,2,2, SIP null
    3 0,0,3,3, NEW YORK
ADDRESS_PART1
    3 0,0,0,0, 1740 BROADWAY
    3 0,0,1,1, 1740 BROADWAY
    3 0,0,2,2, 1740 BROADWAY
    3 0,0,3,3, 1740 BROADWAY
```



Subtype column

Associated fields



Each color represents a logical group tied to a subtyped value. Each group will have its own layout for the ssa call.

Matching is done for each common subtype in the match pair.

Matching stops when a match is found.

Only Fuzzy rules can have subtypes.

Informatica

Match Batch Distribution

Process Server Work Range (scaled to 1000)

```
[2020-03-31 12:55:45,205] [default task-12] [DEBUG]
com.siperian.mrm.util.distributed.DistManager:

This server is:http://torapp2:8380/cleanse/
Server:Port is torapp1:8380 Match true Cleanse true Match Mode 3 online
flag true
Included MatchServer-- Server:Port is torapp1:8380 Node Count 24 node
Capability Multiplier 1.0
Server:Port is torapp2:8380 Match true Cleanse true Match Mode 3 online
flag true
Included MatchServer-- Server:Port is torapp2:8380 Node Count 24 node
Capability Multiplier 1.0
Number of servers :2 Number of Nodes :48.0
Work Range from 1000 for Server 0 is 0 to 499
Work Range from 1000 for Server 1 is 500 to 999
```


Informatica

MDM Match Job Distribution – Based on Search Range

- All Search Ranges have their hashCode calculated during Range Generation, scaled to 1000. If the value falls within that nodes UoW range, then the search range is passed to the SortManager.
- Every process server generates all ranges and determines its own ranges concurrently

```
"UYV>$FV>UYV>$FVB".hashCode() % 1000          Server 1
    (int) 648
"UYV>>VVGUYV>>VVJ".hashCode() % 1000          Server 0
    (int) 341
```

Understanding the Cleanse Log

Producer – Reading Key Data from the STRP

```
[RangerProducer] [DEBUG] com.siperian.mrm.match.RangerProducer: Starting RangerProducer
[RangerProducer] [DEBUG] com.siperian.mrm.match.MatchProperties: No Prefetch setting in
properties file.  Setting to Default :1000
[RangerProducer] [DEBUG] com.siperian.mrm.match.RangerProducer: Before SQL Max Memory =
6442450944 Total Memory = 2456813568 Free Memory = 1342555808
[RangerProducer] [INFO ] com.siperian.mrm.match.RangerProducer: Start reading Data from
STRP Table, sql is:SELECT /*+ PARALLEL ORDERED USE_HASH (B, A) */ S.ROWID_OBJECT,
S.DATA_COUNT, S.SSA_DATA FROM T$MAQ_PARTY B INNER JOIN C_PARTY_STRP S ON (S.ROWID_OBJECT
= B.ROWID_OBJECT AND S.PREFERRED_KEY_IND = 1 AND S.INVALID_IND = 0) ORDER BY S.SSA_KEY,
S.ROWID_OBJECT, S.DATA_ROW
[RangerProducer] [DEBUG] com.siperian.mrm.match.RangerProducer: After SQL Max Memory =
6442450944 Total Memory = 2456813568 Free Memory = 1342501032
[RangerProducer] [DEBUG] com.siperian.mrm.match.MatchProperties: No
ranger_producer_to_ranger_worker_buffer setting in properties file.  Setting to Default
:100
[RangerProducer] [INFO ] com.siperian.mrm.match.RangerProducer: Finished fetching data
from database. Fetched 3 BOs
[RangerProducer] [DEBUG] com.siperian.mrm.match.RangerProducer: Total Search Candidates
:2
[RangerProducer] [DEBUG] com.siperian.mrm.match.RangerProducer: Processed 2
[RangerProducer] [DEBUG] com.siperian.mrm.match.RangerProducer: RangerProducer completed
in 0.099 ( 0.099 sec)
```

RangerWorker – Generating Search Ranges

```
com.siperian.mrm.match.RangerWorker: Search Level Set for Execution
:Narrow
com.siperian.mrm.match.RangerWorker: File loading is used
com.siperian.mrm.match.RangerWorker: Starting Ranger0
com.siperian.mrm.match.RangerWorker: Total Records read    = 2
com.siperian.mrm.match.RangerWorker: Total Ranges created  = 12
com.siperian.mrm.match.RangerWorker: Time Range Generation = 87 ms
```

NOTE: Above Search Level log message is incorrect and shows the search level for realtime SearchMatch. Look further up in the log for the Search Level for the ruleset to see the actual search level used for Range Generation, e.g:

```
[2020-04-01 01:38:03,698] [HTTP-276] [DEBUG]
com.siperian.mrm.match.SSAMeta: MatchRuleSet Fuzzy_Rule_Only Search
Level is:Typical
```


Informatica

SortManager – Packaging RangerNodes

```
[2020-03-28 00:49:58,086] [RangeSorter] [DEBUG]
com.siperian.mrm.match.RangeSorter:
RangeSorter: Starting RangeSorter
RangeSorter: Start gathering output from worker threads
RangeSorter: got end of loading, number closed:1
RangeSorter: Finished sort. Time to gather 2 records, 12 ranges =
1585327798167 ms. Start handing back to workers for matching
SortManager: Sort Starting
SortManager: Sort of 12 Records Done:8ms
SortManager: ship the sorted ranges in memory  back to the workers to
match with
SortManager: Time to distribute, from memory,  12 ranges with 6
candidates = 91 ms
RangeSorter: RangeSorter completed in 0.182 ( 0.182 sec )
```

Candidates = rangerNodes!!! Log statement is misleading.

Informatica

Opening the SSA Session

```
[2020-03-28 00:49:58,203] [Ranger0] [INFO ]
com.siperian.mrm.match.SsaBase:
*** SSA Session opened: s_mdt>     Mar 29 2017 16:59:20 10.0.0.100
s_mdt   MDT     1.8.2.11MSVS2008 2014-02-19 18:11:41
Population File = /home/infa/infamdm/hub/cleanse/resources/match/demo
SECTION: E1                                SSA-NAME3 00302n3sgxx  E1
YY0031           0000EXPDAT  2014-02-19 18:11:43.397000
```

Red – SSA Library Version (loaded from cleanse/lib) – 10.0.0.100

Blue – Population Version – 2014-02-19

SSA Client Jar Version (from siperian-mrm.ear):

09/25/2018 04:27 PM 200,658 ssan3-10.1.0.jar

ThreadMonitor – Totals and Current Snapshots

```
[2019-02-10 22:49:58,342] [RangerManger] [INFO ]
com.siperian.mrm.util.threads.ThreadMonitor: Dist:Ranger15 Matching TCan:1891941329
Tgr:1891941329 TSSA:5482230 TM:660029 TR:57720 Cur RI:137572855 Cur Range:S?DGAA$$ to
S?DGAAZZ CompsPerRange:12408025
```

- TCan: total number of database candidates retrieved across all rangerNodes processed by this thread
- Tgr: Total number of comparisons (exact and ssa) where the search record has a lower rowid than the file record performed across all rangerNodes processed by this thread. Only accurate if "Match Only Previous Rowid Objects" is enabled
- TSSA: Total number of ssa comparisons performed across all rangerNodes processed by this thread
- TM: Total matches found across all rangerNodes processed by this thread
- TR: Total rangerNodes processed by this thread up to this point
- Cur RI: The rowid of the db file record currently being processed
- Cur Range: The search range of the current rangerNode being processed by this thread at this point in time
- CompsPerRange: The number of search records * db comparisons done so far for the current rangerNode

RangerWorker Summary – Top 10 Range Counts

```
[Ranger0] [INFO ] com.siperian.mrm.match.RangerWorker:
Top 10 Range counts:
Ranger0 Max Range 0          = 5 between 'UYV>$E$$' and 'UYV>$EZZ'
Ranger0 Max Range 1          = 4 between 'LVVBCFV>' and 'LVVBCFVB'
Ranger0 Max Range 2          = 0 between 'UYV>BGGC' and 'UYV>BGGF'
Ranger0 Max Range 3          = 0 between 'UYV>>VVG' and 'UYV>>VVJ'
Ranger0 Max Range 4          = 0 between 'UYV>$FV>' and 'UYV>$FVB'
Ranger0 Max Range 5          = 0 between 'UYV>$$$$' and 'UYV>$$$/'
Ranger0 Max Range 6          = 0 between 'null' and 'null'
Ranger0 Max Range 7          = 0 between 'null' and 'null'
Ranger0 Max Range 8          = 0 between 'null' and 'null'
Ranger0 Max Range 9          = 0 between 'null' and 'null'
```


Informatica™

RangerWorker Summary – Top 10 Range Comparisons

```
[Ranger0] [INFO ] com.siperian.mrm.match.RangerWorker:
Top 10 Range Comparisons counts
Ranger0 Comparison Max Range 0 = 10 Q:2 DB:5 between 'UYV>$E$$' and 'UYV>$EZZ'
Ranger0 Comparison Max Range 1 = 8 Q:2 DB:4 between 'LVVBCFV>' and 'LVVBCFVB'
Ranger0 Comparison Max Range 2 = 0 Q:2 DB:0 between 'UYV>BGGC' and 'UYV>BGGF'
Ranger0 Comparison Max Range 3 = 0 Q:2 DB:0 between 'UYV>>VVG' and 'UYV>>VVJ'
Ranger0 Comparison Max Range 4 = 0 Q:2 DB:0 between 'UYV>$FV>' and 'UYV>$FVB'
Ranger0 Comparison Max Range 5 = 0 Q:2 DB:0 between 'UYV>$$$$' and 'UYV>$$$/'
Ranger0 Comparison Max Range 6 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Comparison Max Range 7 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Comparison Max Range 8 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Comparison Max Range 9 = 0 Q:0 DB:0 between 'null' and 'null'
Ranger0 Total Ranges Processed  = 6
Ranger0 Total Candidates        = 14
Ranger0 Total Matches           = 1
Matcher Summary :total_calls: 14 SSA Matches: 14
```

‘Candidates’ really means candidate comparisons done by this thread (Ranger0).

These comparison counts can be used to choose a relevant Dynamic Match Analysis Threshold. A DMAT of 8 would still process rangerNode `'LVVBCFV>' and 'LVVBCFVB'` but skip `'UYV>$E$$' and 'UYV>$EZZ'`

Informatica

Fuzzy Match Summary

```
[2020-03-31 15:23:39,256] [HTTP-219] [INFO ] com.siperian.mrm.match.Ranger:
Total Records to Match       :2
Total DB Records Read        :9
Total Possible Matches       :14
Total Greater                :14
Total SSA                    :14
Total Matches                :1
Total Auto Matches           :0
Total Manual Matches         :1
Total Ranges Created         :12
Total Ranges Processed       :6
Total Rejects From Exact     :0
Total Rejects From Ssa       :13
Total Match Calls            :14
Total time used 1 secs
```

Ranges Created = total search ranges generated across all records in match batch

Ranges Processed = rangerNodes processed across all rangerWorkers

Informatica

Fuzzy Match Summary Legend

Total Records to Match: Number of records in match batch

Total DB Records Read: Candidates read from STRP table across all RangerWorker threads

Total Possible Matches: Number of candidates evaluated across all RangerWorker threads

Total Greater: Candidates whose rowid is greater than the search record's rowid (only correct if using Match Only Previous Rowid Objects, otherwise same as Possible Matches)

Total SSA: Number of SSA Purpose evaluations across all RangerWorker threads

Total Matches: # of match rows collected by MatchGatherer, net of any dupes found by MatchGatherer

Total Auto Matches: of the Total Matches found, how many are from fuzzy automerge rules

Total Manual Matches: of the total matches found, how many are from fuzzy manual merge rules

Total Ranges Created: total number of search ranges generated by rangeGen() across all RangerWorkers

Total Ranges Processed: total number of rangerNodes processed across all rangerWorkers

Total Rejects From Exact: exact comparisons that failed evaluation, including child data (does not count: exact only rules, null matching, or segment matching)

Total Rejects From Ssa: fuzzy comparisons that failed ssa purpose evaluation

Total Match Calls: Total search rec to file rec comparisons across all RangerWorkers

Informatica™

Begin Exact Match Phase

```
[2020-04-06 15:23:42,918] [HTTP-209] [INFO ] com.siperian.mrm.match.cmxma.Match:
Proceeding with exact match rules.
Exact Rule :1 AutoMerge Ind :false Asymetrical Ind:false
        Node Num :1 Exact Match, Match Column:'Ex_Party_Type' Match Column Id:6 Anti Match
Ind:false
        Node Num :0 Exact Match, Match Column:'Ex_Address_Type' Match Column Id:9 Anti
Match Ind:false
        Node Num :2 Exact Match, Match Column:'Ex_Telecom' Match Column Id:14 Anti Match
Ind:false
This rule will be processed with a normal select join clause
[2020-04-06 15:23:42,952] [HTTP-209] [DEBUG] com.siperian.mrm.match.SSAMeta:
Node is         Node Num :1 Exact Match, Match Column:'Ex_Party_Type' Match Column Id:6
Anti Match Ind:false
Match Column is Ex_Party_Type Depth 1 Table:C_PARTY_MTIP
Node is         Node Num :0 Exact Match, Match Column:'Ex_Address_Type' Match Column Id:9
Anti Match Ind:false
Match Column is Ex_Address_Type Depth 2 Table:C_MT_PARTY_ADDRESS_REL
Node is         Node Num :2 Exact Match, Match Column:'Ex_Telecom' Match Column Id:14 Anti
Match Ind:false
Match Column is Ex_Telecom Depth 2 Table:C_MT_TELECOM
```


Informatica

Improving Performance

Performance Tips

- Use exact fields as much as possible
- Avoid subtype match – try filtered match path as workaround
- If few straggling RangerWorkers finish much later
  - Consider decreasing max_records_per_ranger_node to smooth out uneven rangerNode processing times and improve concurrency
  - Configure Dynamic Match Analysis Threshold if match quality can be sacrificed for performance
- Analyze RangerWorker Summary Top Counts
  - Check unusually large range counts against STRP table
    - If SSA_DATA shows keys from noise, add noise words with Population Override Manager
    - If SSA_DATA shows numerous keys from valid data, adjust frequency table with Population Override Manager to mark this data as 'common'

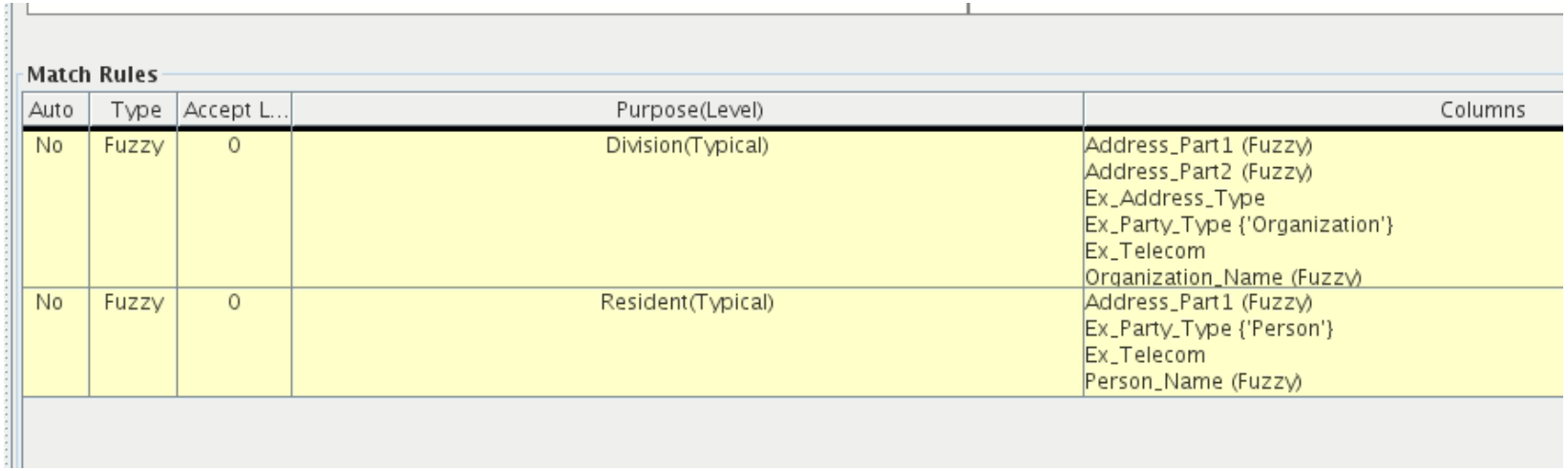
CAUTION: Any population changes should be well tested before promotion. Reach out to IPS/GCS/ACE for help as needed.

Q&A

Thank You

achan@informatica.com

Matching on Distinct ‘Ruleset Nodes’



Match Rules

Auto	Type	Accept L...	Purpose(Level)	Columns
No	Fuzzy	0	Division(Typical)	Address_Part1 (Fuzzy) Address_Part2 (Fuzzy) Ex_Address_Type Ex_Party_Type {'Organization'} Ex_Telecom Organization_Name (Fuzzy)
No	Fuzzy	0	Resident(Typical)	Address_Part1 (Fuzzy) Ex_Party_Type {'Person'} Ex_Telecom Person_Name (Fuzzy)

```
Ruleset 'Fuzzy_Rule_Only' has 2 rule(s), Search Call:false
Fuzzy Rule :1 AutoMerge Ind :false Asymetrical Ind:false
      Node Num :1 Segment Match, Match Column:'Ex_Party_Type'
Match Column Id:6 Segment Value/s:'Organization'
      Node Num :0 Exact Match, Match Column:'Ex_Address_Type'
Match Column Id:9 Anti Match Ind:false
      Node Num :2 Exact Match, Match Column:'Ex_Telecom' Match
Column Id:14 Anti Match Ind:false
      Node Num :3 SSA Matching on
            'Address_Part1 Address_Part1' Column Id:11
            'Address_Part2 Address_Part2' Column Id:10
            'Organization_Name Organization_Name' Column Id:0
            Match Level:Typical Geocode Radius:0 Match
Purpose:Division
```

```
Fuzzy Rule :2 AutoMerge Ind :false Asymetrical Ind:false
      Node Num :4 Segment Match, Match Column:'Ex_Party_Type'
Match Column Id:6 Segment Value/s:'Person'
      Node Num :2 Exact Match, Match Column:'Ex_Telecom' Match
Column Id:14 Anti Match Ind:false
      Node Num :5 SSA Matching on
            'Address_Part1 Address_Part1' Column Id:11
            'Person_Name Person_Name' Column Id:5
            Match Level:Typical Geocode Radius:0 Match
Purpose:Resident
Total Nodes :10 Actual Nodes:6
```


Informatica