Informatica Data Quality Project Report

Project Overview

This project focuses on implementing data quality rules using Informatica Data Quality (IDQ) to cleanse and standardize customer data. The processed data is subsequently exported into Informatica PowerCenter for further integration. The project ensures data accuracy, completeness, and consistency by applying specific rules and validations.

Project Objectives

1. **Ensure Customer ID Consistency:**
   - Validate that Cust_id contains only digits and has a fixed length of 8.
   - Null values in Cust_id are rejected and moved to the rejection section.
2. **Name Splitting:**
   - Concatenate FNAME and LNAME attributes and split them into separate fields.
3. **Customer Tier Standardization:**
   - Replace invalid values in the cust_tier attribute with predefined valid values.
4. **State Code Standardization:**
   - Ensure that the state field contains exactly 2 characters (e.g., "NW").
5. **ZIP Code Validation:**
   - Ensure zip contains exactly 5 numeric digits.
6. **City, State, and Country Splitting:**
   - Separate concatenated city, state, and country fields into individual columns.
7. **Record Rejection Criteria:**
   - Reject records with null Cust_id.
   - Reject records with null city, state, and country fields.

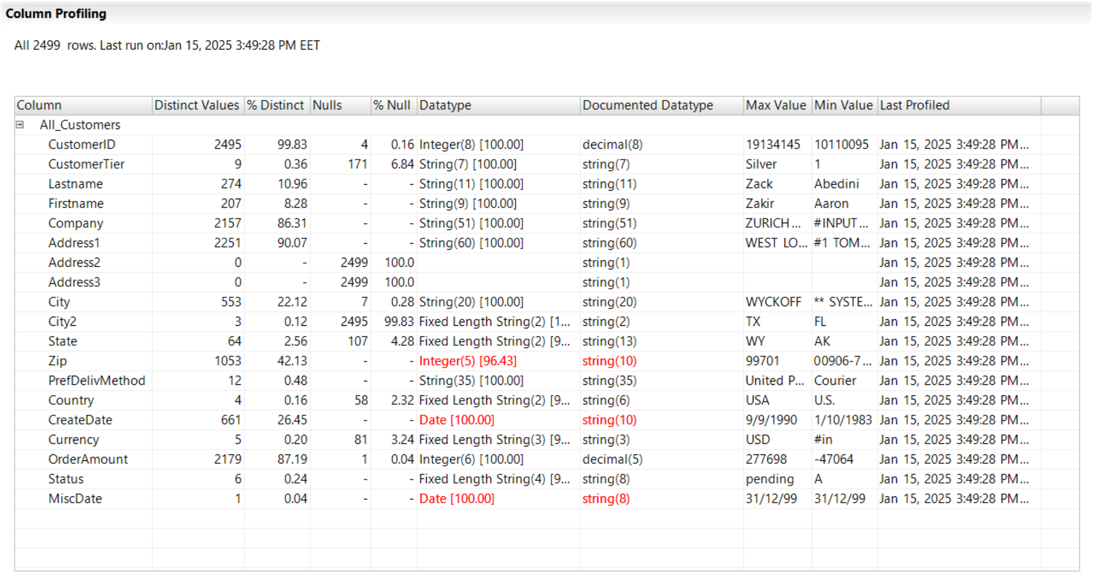
8. **Review Section Handling:**
    - If any attributes are null (except those leading to rejection), send records to the review section.
9. **Matching Process:**
    - Records not meeting rejection criteria are processed for matching.
    - Extract the first 3 characters from fname, lname, city, and state for matching.
    - Records with a match score of 80% or higher are auto-consolidated and loaded into the standard output.
    - Records with a match score between 60% and 80% are sent to the duplicate review section.

Data Profiling

Data profiling was conducted to assess the quality of the input dataset, identifying key issues such as:

- Null values in Address2, Address3.
- Inconsistent values in cust_tier.
- Length and format inconsistencies in zip and Cust_id.

In the Developer :

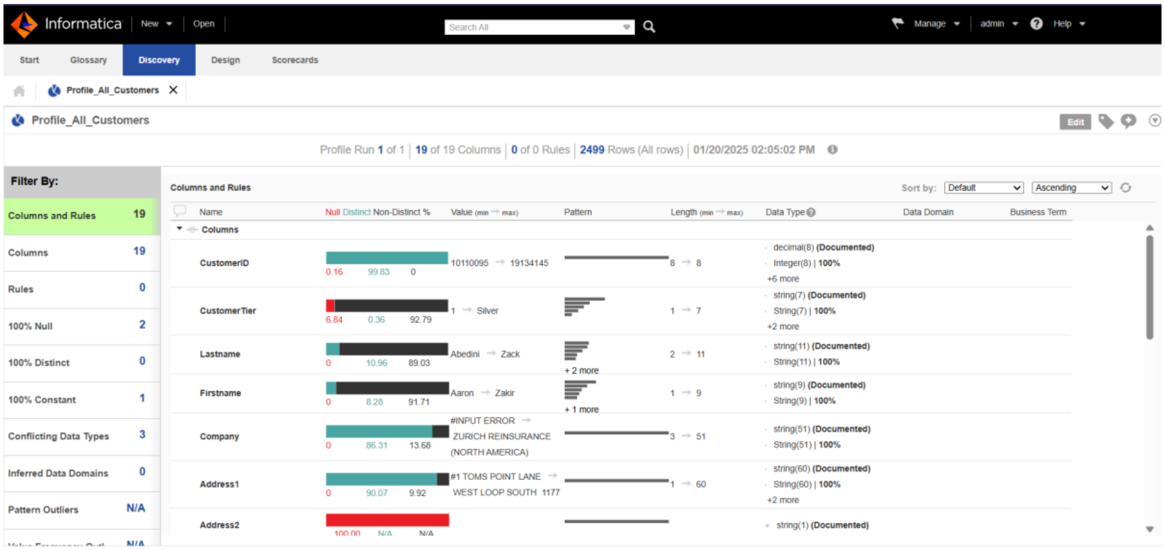


Column Profiling

All 2499 rows. Last run on:Jan 15, 2025 3:49:28 PM EET

Column	Distinct Values	% Distinct	Nulls	% Null	Datatype	Documented Datatype	Max Value	Min Value	Last Profiled
⊟ All_Customers									
CustomerID	2495	99.83	4	0.16	Integer(8) [100.00]	decimal(8)	19134145	10110095	Jan 15, 2025 3:49:28 PM...
CustomerTier	9	0.36	171	6.84	String(7) [100.00]	string(7)	Silver	1	Jan 15, 2025 3:49:28 PM...
Lastname	274	10.96	-	-	String(11) [100.00]	string(11)	Zack	Abedini	Jan 15, 2025 3:49:28 PM...
Firstname	207	8.28	-	-	String(9) [100.00]	string(9)	Zakir	Aaron	Jan 15, 2025 3:49:28 PM...
Company	2157	86.31	-	-	String(51) [100.00]	string(51)	ZURICH ...	#INPUT ...	Jan 15, 2025 3:49:28 PM...
Address1	2251	90.07	-	-	String(60) [100.00]	string(60)	WEST LO...	#1 TOM...	Jan 15, 2025 3:49:28 PM...
Address2	0	-	2499	100.0		string(1)			Jan 15, 2025 3:49:28 PM...
Address3	0	-	2499	100.0		string(1)			Jan 15, 2025 3:49:28 PM...
City	553	22.12	7	0.28	String(20) [100.00]	string(20)	WYCKOFF	** SYSTE...	Jan 15, 2025 3:49:28 PM...
City2	3	0.12	2495	99.83	Fixed Length String(2) [1...	string(2)	TX	FL	Jan 15, 2025 3:49:28 PM...
State	64	2.56	107	4.28	Fixed Length String(2) [9...	string(13)	WY	AK	Jan 15, 2025 3:49:28 PM...
Zip	1053	42.13	-	-	Integer(5) [96.43]	string(10)	99701	00906-7...	Jan 15, 2025 3:49:28 PM...
PrefDelivMethod	12	0.48	-	-	String(35) [100.00]	string(35)	United P...	Courier	Jan 15, 2025 3:49:28 PM...
Country	4	0.16	58	2.32	Fixed Length String(2) [9...	string(6)	USA	U.S.	Jan 15, 2025 3:49:28 PM...
CreateDate	661	26.45	-	-	Date [100.00]	string(10)	9/9/1990	1/10/1983	Jan 15, 2025 3:49:28 PM...
Currency	5	0.20	81	3.24	Fixed Length String(3) [9...	string(3)	USD	#in	Jan 15, 2025 3:49:28 PM...
OrderAmount	2179	87.19	1	0.04	Integer(6) [100.00]	decimal(5)	277698	-47064	Jan 15, 2025 3:49:28 PM...
Status	6	0.24	-	-	Fixed Length String(4) [9...	string(8)	pending	A	Jan 15, 2025 3:49:28 PM...
MiscDate	1	0.04	-	-	Date [100.00]	string(8)	31/12/99	31/12/99	Jan 15, 2025 3:49:28 PM...

In the Analyst :



Profile_All_Customers

Profile Run 1 of 1 | 19 of 19 Columns | 0 of 0 Rules | 2499 Rows (All rows) | 01/20/2025 02:05:02 PM

Filter By:	
Columns and Rules	19
Columns	19
Rules	0
100% Null	2
100% Distinct	0
100% Constant	1
Conflicting Data Types	3
Inferred Data Domains	0
Pattern Outliers	N/A
Value Frequency Outl…	N/A

Columns and Rules

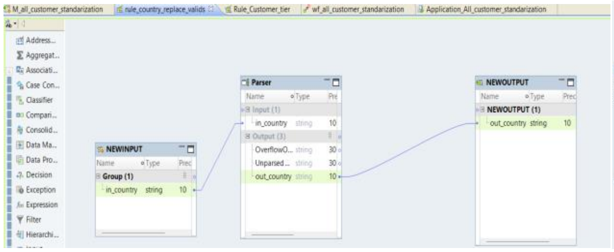
Sort by: Default Ascending

Name	Null Distinct Non-Distinct %	Value (min → max)	Pattern	Length (min → max)	Data Type	Data Domain	Business Term
Columns							
CustomerID	0.16 99.83 0	10110095 → 19134145		8 → 8	decimal(8) (Documented) Integer(8) | 100% +6 more		
CustomerTier	6.84 0.36 92.79	1 → Silver		1 → 7	string(7) (Documented) String(7) | 100% +2 more		
Lastname	0 10.96 89.03	Abedini → Zack	+ 2 more	2 → 11	string(11) (Documented) String(11) | 100%		
Firstname	0 8.28 91.71	Aaron → Zakir	+ 1 more	1 → 9	string(9) (Documented) String(9) | 100%		
Company	0 86.31 13.68	#INPUT ERROR → ZURICH REINSURANCE (NORTH AMERICA)		3 → 51	string(51) (Documented) String(51) | 100%		
Address1	0 90.07 9.92	#1 TOMS POINT LANE → WEST LOOP SOUTH 1177		1 → 60	string(60) (Documented) String(60) | 100% +2 more		
Address2	100.00 N/A N/A				string(1) (Documented)		

Data Quality Rules Implementation

Key transformations and rules implemented include:

- **Parsing and Validation:**
  - Cust_id parsed and checked for numeric values.
  - zip length validated to ensure 5 digits.
- **Rule-Based Transformations:**
  - Standardizing cust_tier values.
  - Separating concatenated fields.



NEWINPUT

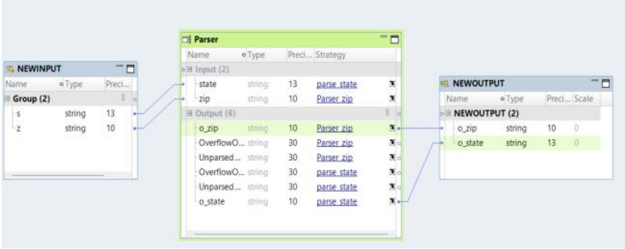
Name	Type	Prec
Group (1)		
in_country	string	10

Parser

Name	Type	Pre
Input (1)		
in_country	string	10
Output (3)		
OverflowO...	string	30
Unparsed ...	string	30
out_country	string	10

NEWOUTPUT

Name	Type	Prec
NEWOUTPUT (1)		
out_country	string	10


NEWINPUT

Name	Type	Preci...
Group (2)		
s	string	13
z	string	10

Parser

Name	Type	Preci...	Strategy
Input (2)			
state	string	13	parse state
zip	string	10	Parser zip
Output (6)			
o_zip	string	10	Parser zip
OverflowO...	string	30	Parser zip
Unparsed ...	string	30	Parser zip
OverflowO...	string	30	parse state
Unparsed ...	string	30	parse state
o_state	string	10	parse state

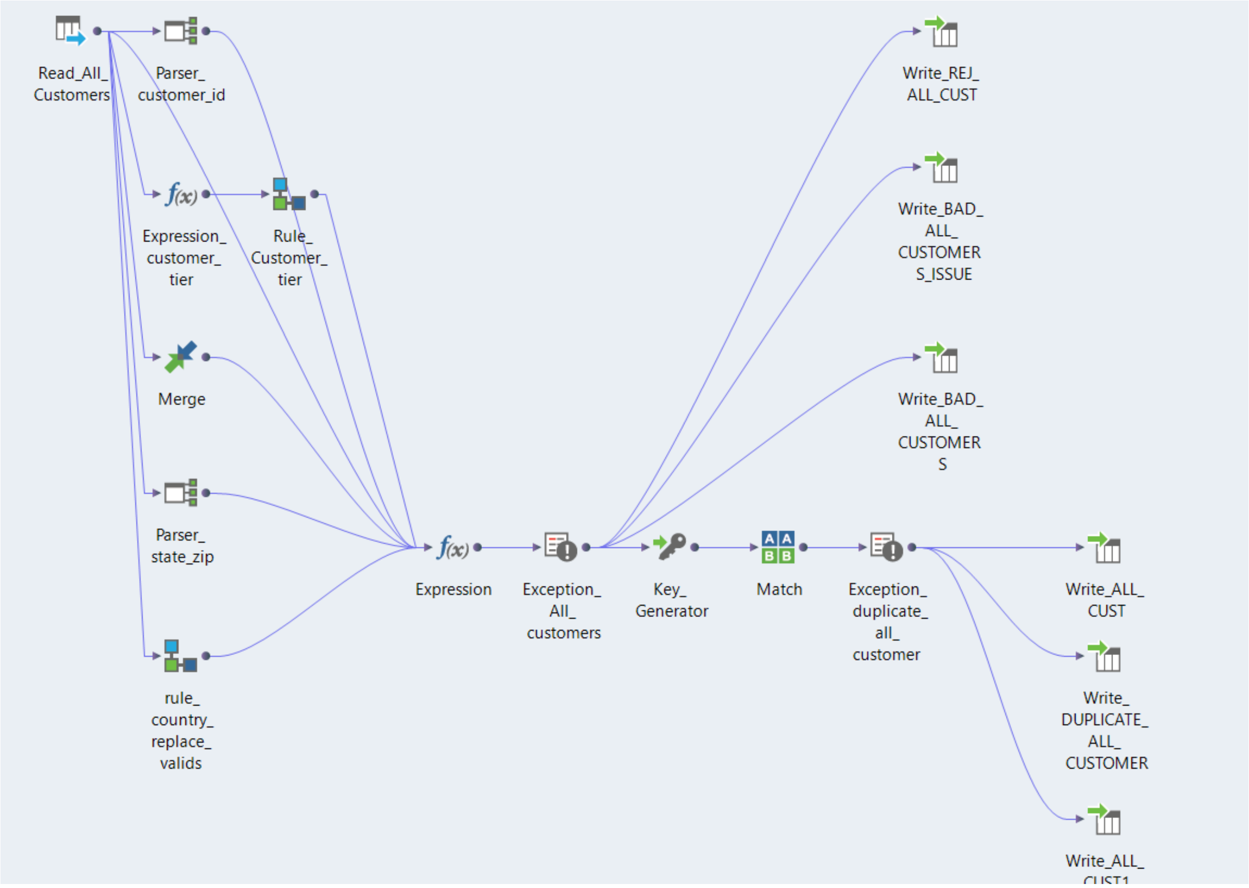
NEWOUTPUT

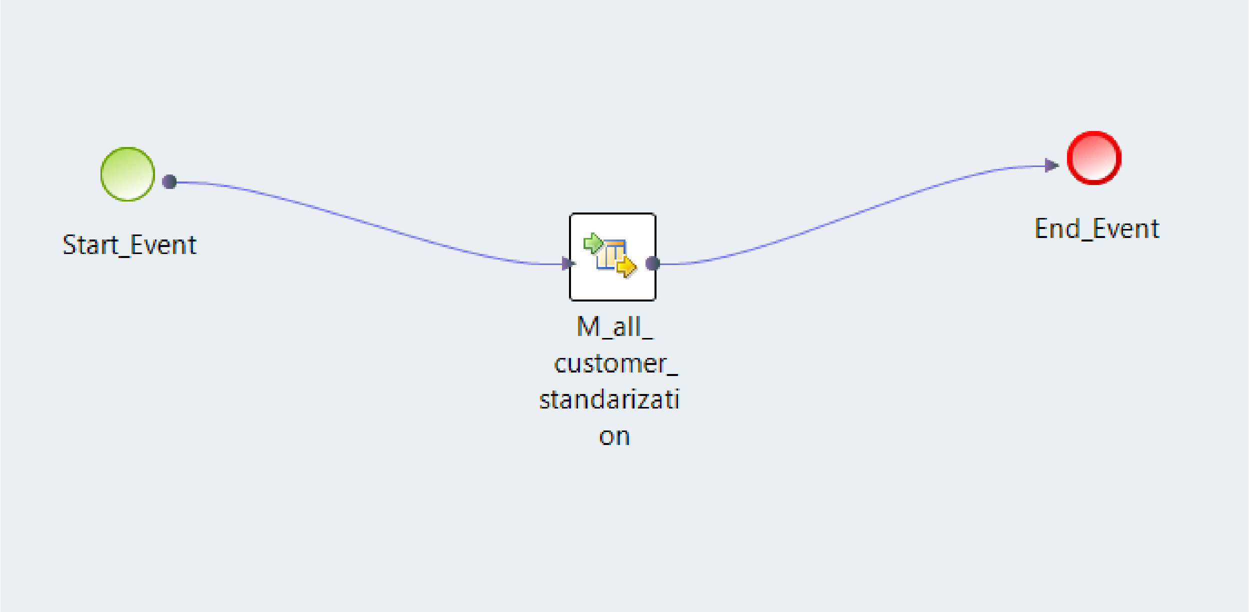
Name	Type	Preci...	Scale
NEWOUTPUT (2)			
o_zip	string	10	0
o_state	string	13	0

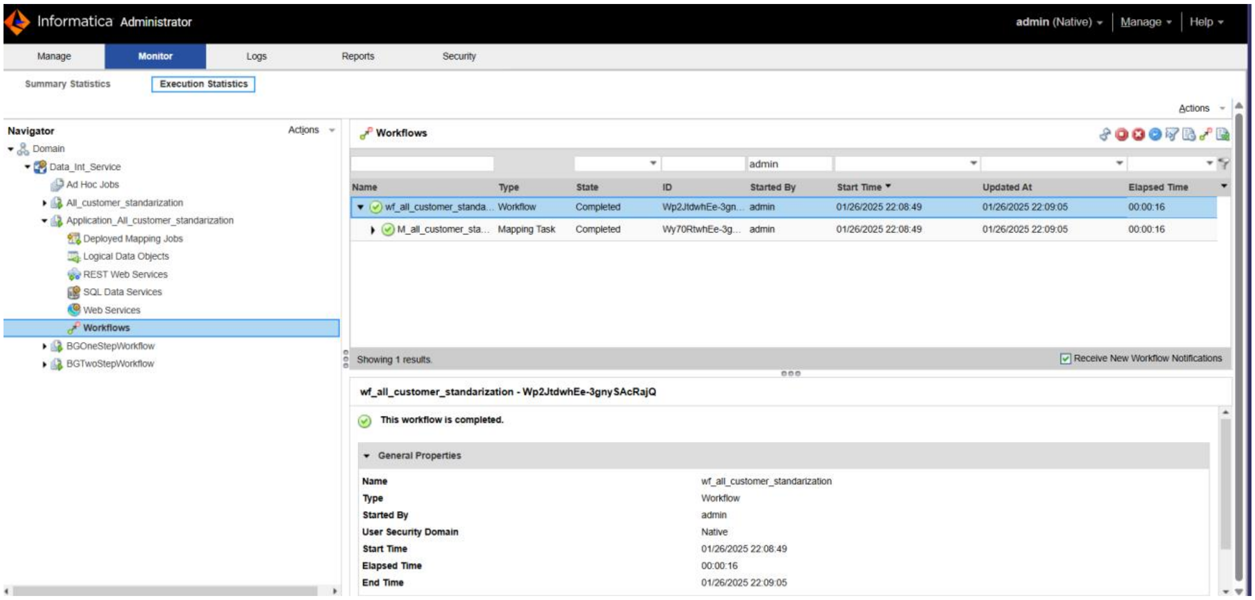
Data Standardization Workflow

The workflow designed in Informatica Data Quality follows the steps below:

1. Reading customer data from the source.
2. Applying various data quality transformations.
3. Sending rejected records to the rejection table.
4. Consolidating valid records and exporting them to PowerCenter.


Read_All_ Customers
Parser_ customer_id
Expression_ customer_ tier
Rule_ Customer_ tier
Merge
Parser_ state_zip
rule_ country_ replace_ valids
Expression
Exception_ All_ customers
Key_ Generator
Match
Exception_ duplicate_ all_ customer
Write_REJ_ ALL_CUST
Write_BAD_ ALL_ CUSTOMER S_ISSUE
Write_BAD_ ALL_ CUSTOMER S
Write_ALL_ CUST
Write_ DUPLICATE_ ALL_ CUSTOMER
Write_ALL_ CUST1





Name	Type	State	ID	Started By	Start Time	Updated At	Elapsed Time
wf_all_customer_standa…	Workflow	Completed	Wp2JtdwhEe-3gn…	admin	01/26/2025 22:08:49	01/26/2025 22:09:05	00:00:16
M_all_customer_sta…	Mapping Task	Completed	Wy70RtwhEe-3g…	admin	01/26/2025 22:08:49	01/26/2025 22:09:05	00:00:16

Showing 1 results.

☑ Receive New Workflow Notifications

wf_all_customer_standarization - Wp2JtdwhEe-3gnySAcRajQ

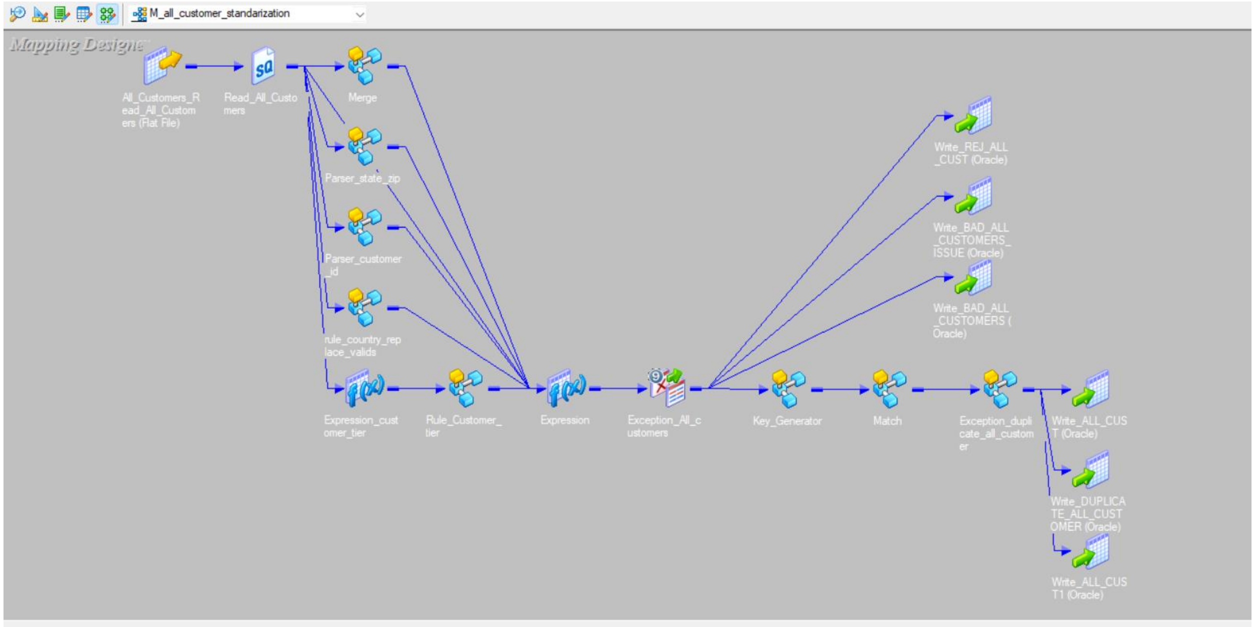
This workflow is completed.

General Properties

Name	wf_all_customer_standarization
Type	Workflow
Started By	admin
User Security Domain	Native
Start Time	01/26/2025 22:08:49
Elapsed Time	00:00:16
End Time	01/26/2025 22:09:05

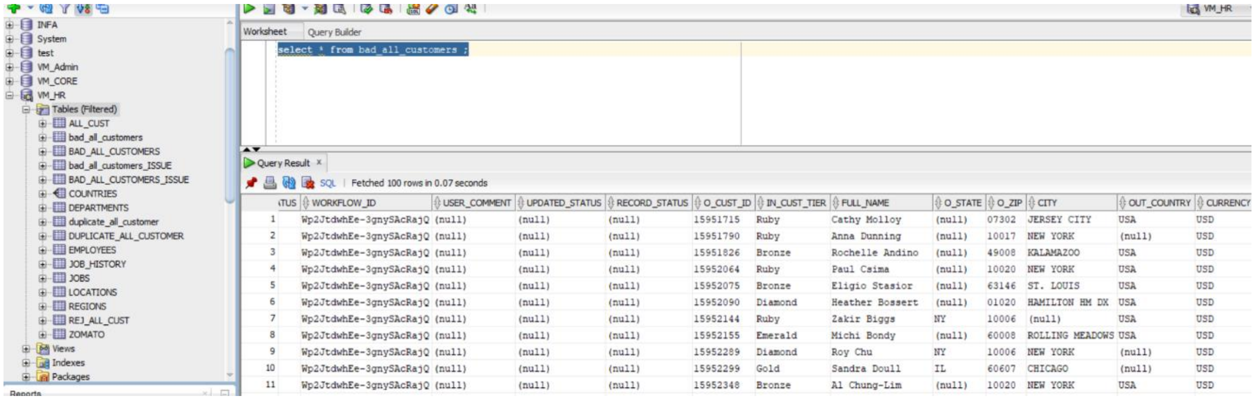
Exporting to PowerCenter

After standardization, clean data is exported to Informatica PowerCenter for integration into downstream systems.


M_all_customer_standarization
Mapping Designer
All_Customers_Read_All_Customers (Flat File)
Read_All_Customers
Merge
Parser_state_zip
Parser_customer_id
rule_country_replace_valids
Expression_customer_tier
Rule_Customer_tier
Expression
Exception_All_customers
Key_Generator
Match
Exception_duplicate_all_customer
Write_REJ_ALL_CUST (Oracle)
Write_BAD_ALL_CUSTOMERS_ISSUE (Oracle)
Write_BAD_ALL_CUSTOMERS (Oracle)
Write_ALL_CUST (Oracle)
Write_DUPLICATE_ALL_CUSTOMER (Oracle)
Write_ALL_CUST1 (Oracle)

Results and Observations

- Successfully processed 2499 records.
- Identified and rejected 5% of records due to null values.
- 10% of records were flagged for review due to partial completeness.
- The matching process identified duplicate records with a 70% match score.



```
select * from bad_all_customers ;
```

Fetched 100 rows in 0.07 seconds

	TUS	WORKFLOW_ID	USER_COMMENT	UPDATED_STATUS	RECORD_STATUS	O_CUST_ID	IN_CUST_TIER	FULL_NAME	O_STATE	O_ZIP	CITY	OUT_COUNTRY	CURRENCY
1		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15951715	Ruby	Cathy Molloy	(null)	07302	JERSEY CITY	USA	USD
2		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15951790	Ruby	Anna Dunning	(null)	10017	NEW YORK	(null)	USD
3		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15951826	Bronze	Rochelle Andino	(null)	49008	KALAMAZOO	USA	USD
4		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952064	Ruby	Paul Csima	(null)	10020	NEW YORK	USA	USD
5		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952075	Bronze	Eligio Stasior	(null)	63146	ST. LOUIS	USA	USD
6		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952090	Diamond	Heather Bossert	(null)	01020	HAMILTON HM DX	USA	USD
7		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952144	Ruby	Zakir Biggs	NY	10006	(null)	USA	USD
8		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952155	Emerald	Michi Bondy	(null)	60008	ROLLING MEADOWS	USA	USD
9		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952289	Diamond	Roy Chu	NY	10006	NEW YORK	(null)	USD
10		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952299	Gold	Sandra Doull	IL	60607	CHICAGO	(null)	USD
11		Wp2JtdwhEe-3gnySAcRajQ	(null)	(null)	(null)	15952348	Bronze	Al Chung-Lim	(null)	10020	NEW YORK	USA	USD

Conclusion

The Informatica Data Quality project successfully improved the accuracy and consistency of customer data. The project followed a structured approach to validate, standardize, and cleanse data, ensuring high-quality data for business processes.