KLE Technological University
Huballi

A Course Project Report on

"Customer Data Analytics using Informatica Cloud"

A Course Project Report Submitted in Partial Fulfillment of the Requirement for the Course of

Informatica Intelligent Data Management Cloud

in

6th Semester of Computer Science and Engineering(AI)

by

Abhinandan Onajol -	02FE22BCI003
Jaganath Malode -	02FE22BCI018
Varun Gani -	02FE22BCI056

Under the guidance of

Prof. Anita Kenchannavar
Professor,
Department of Computer Science and Engineering(AI),
KLE Technological University's Dr. MSSCET, Belagavi.

KLE Technological University's
Dr. M. S. Sheshgiri College of Engineering and Technology,
Belagavi – 590 008.

June 2025

DECLARATION

We hereby declare that the matter embodied in this report entitled "**Customer Data Analytics using Informatica Cloud**" submitted to KLE Technological University for the course completion of Informatica Intelligent Data Management Cloud (24ECSE322) in the 6th Semester of Computer Science and Engineering is the result of the work done by us in the Department of Computer Science and Engineering (Artificial Intelligence), KLE Dr. M. S. Sheshgiri College of Engineering, Belagavi under the guidance of Prof. Anita Kenchannavar, Professor, Department of Computer Science and Engineering (Artificial Intelligence). We further declare that to the best of our knowledge and belief, the work reported here in doesn't form part of any other project on the basis of which a course or award was conferred on an earlier occasion on this by any other student, also the results of the work are not submitted for the award of any course, degree or diploma within this or in any other University or Institute. We hereby also confirm that all of the experimental work in this report has been done by us.

Belagavi – 590 008
Date :

Abhinandan Onajol
(02FE22BCI003)

Jaganath Malode
(02FE22BCI018)

Varun Gani
(02FE22BCI056)

KLE
TECHNOLOGICAL UNIVERSITY
Creating Value, Leveraging Knowledge
Belagavi Campus

Dr.M.S.Sheshgiri College of Engineering & Technology
Department of Computer Science & Engineering

CERTIFICATE

This is to certify that the project entitled "Customer Data Analytics using Informatica Cloud" submitted to KLE Technological University's Dr. MSSCET, Belagavi for the partial fulfillment of the requirement for the course - Informatica Intelligent Data Management Cloud (24ECSE322) by Abhinandan Onajol (02FE22BCI003), Jaganath Malode (02FE22BCI018), Varun Gani (02FE22BCI056)., students in the Department of Computer Science and Engineering(Artificial Intelligence), KLE Technological University's Dr. MSSCET, Belagavi, is a bonafide record of the work carried out by them under my supervision. The contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any other course completion.

Belagavi – 590 008
Date :

Prof. Anita Kenchannavar	Dr. Rajashri Khanai
(Course Teacher)	(Head of the Department)

Abstract

This project focuses on leveraging the capabilities of Informatica Cloud to design and implement a complete ETL (Extract, Transform, Load) pipeline for customer data analytics. Customer information from multiple disparate sources such as CRM systems, e-commerce platforms, and support databases is extracted, cleansed, standardized, and aggregated into a unified dataset. The objective is to obtain a clean, consolidated view of customer behavior that supports data-driven business insights and decisions. This hands-on experience enhances understanding of cloud-based data integration processes and customer data analytics.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

- In many organizations, customer data is spread across multiple platforms such as CRM systems, e-commerce portals, and support ticketing tools.

- These data sources often use inconsistent formats, contain duplicate records, and may have missing or incomplete data fields.

- Without proper data integration and cleansing, it's difficult to gain reliable insights into customer behavior and purchasing patterns.

- Informatica Cloud provides a cloud-based ETL solution that allows organizations to extract, transform, and load data from different sources into a unified, analytics-ready format.

- Using Informatica's transformation components like Joiner, Union, and Sorter, data from various sources can be merged, standardized, and sorted efficiently.

- This helps businesses derive accurate customer insights, improve decision-making, and personalize marketing strategies.

1.2 Problem Statement

The main aim is to extract, transform, and load customer data from multiple sources such as CRM systems and e-commerce platforms using Informatica Cloud, in order to build a clean, consistent, and consolidated customer dataset. This enables improved customer behavior analysis by resolving issues like inconsistent formats, duplicate records, and missing data, through the use of transformations such as Joiner, Union, and Sorter.

1.2.1 Objectives

- To extract customer data from multiple sources such as CRM systems and e-commerce platforms.

- To clean and standardize the data by removing duplicates and handling missing values.

- To apply Joiner, Union, and Sorter transformations using Informatica Cloud.

- To load the transformed data into a centralized and structured data warehouse.

- To enable effective customer behavior analysis and business decision-making through unified data.

Chapter 2

Database

2.1 Database Structure

This section details the data sources, their relationships, and the transformed output created by the ETL pipeline.

2.1.1 Data Sources Overview

The project integrates customer data from four primary sources:

- `customers_crm.csv` - Core customer demographic data
- `ecommerce_purchases.csv` - Transaction records
- `support_tickets.csv` - Customer service interactions
- `marketing_contacts.csv` - Marketing campaign data

2.1.2 Table Schemas

2.1.2.1 CRM Customer Data (`customers_crm.csv`)

Field	**Data Type**	**Description**
customer_id	VARCHAR	Primary key
first_name	VARCHAR	Customer's first name
last_name	VARCHAR	Customer's last name
email	VARCHAR	Validated email address
phone_number	VARCHAR	Standardized format
created_date	DATE	Account creation date
country	VARCHAR	Country of residence

2.1.2.2 E-commerce Purchases (`ecommerce_purchases.csv`)

Field	**Data Type**	**Description**
order_id	VARCHAR	Primary key
customer_id	VARCHAR	Foreign key
product_id	VARCHAR	SKU identifier
amount	DECIMAL(10,2)	Transaction amount
purchase_date	DATE	Order date
payment_method	VARCHAR	Payment type

2.1.2.3 Support Tickets (`support_tickets.csv`)

Field	**Data Type**	**Description**
ticket_id	VARCHAR	Primary key
customer_id	VARCHAR	Foreign key
issue_type	VARCHAR	Problem category
status	VARCHAR	Current status
opened_date	DATE	Ticket creation date
closed_date	DATE	Resolution date

2.1.2.4 Marketing Contacts (`marketing_contacts.csv`)

Field	**Data Type**	**Description**
campaign_id	VARCHAR	Campaign identifier
customer_id	VARCHAR	Foreign key
channel	VARCHAR	Marketing medium
interaction_date	DATE	Contact date
action_taken	VARCHAR	Customer response

2.1.3 Transformed Output

2.1.3.1 Summary Metrics (`customer_summary_metrics.txt`)

Field	**Source**	**Transformation**
customer_id	CRM	Direct mapping
full_name	CRM	Concatenation
email	CRM	Validation
phone_number	CRM	Standardization
total_purchases	Purchases	Count aggregation
total_spent	Purchases	Sum aggregation
last_purchase_date	Purchases	Max date
total_support_tickets	Support	Count aggregation
last_ticket_status	Support	Most recent record
last_campaign_action	Marketing	Most recent record

2.1.3.2 ETL Transformations

- **Joiner**: CRM + Purchases on customer_id
- **Union**: Support + Marketing data
- **Aggregator**: Purchase metrics
- **Sorter**: Chronological ordering

- **Expression**: Field derivations

2.1.3.3 Sample Output

```
customer_id     full_name                email          phone_number  ...
    CUST001   Alice Smith   alice@example.com  +1234567890  ...
    CUST002     Bob Jones     bob@example.com  +1987654321  ...
    CUST003 Charlie Brown charlie@example.com  +1472583690  ...
    CUST004  David Wilson   david@example.com  +1122334455  ...
    CUST005    Eva Taylor     eva@example.com  +1098765432  ...
```

Chapter 3

Implementation Framework

To build the customer data analytics pipeline using Informatica Cloud, we followed this implementation framework:

- **Data Extraction**:
  - Connected to multiple source systems (CSV files, databases)
  - Configured source definitions for CRM, e-commerce, support, and marketing data
  - Established secure connections to cloud applications

- **Data Quality Assessment**:
  - Identified missing values in customer records
  - Detected inconsistent formats (phone numbers, emails)
  - Flagged duplicate customer entries across systems

- **Data Transformation**:
  - Implemented Joiner Transformation to merge CRM and purchase data
  - Applied Union Transformation for support and marketing data
  - Used Sorter Transformation to organize records chronologically
  - Created derived fields (full_name, customer_segment)

- **Data Cleansing**:
  - Standardized phone number formats (+CountryCode)
  - Normalized email addresses (lowercase, domain validation)
  - Implemented fuzzy matching for duplicate customer resolution
- **Aggregation & Metrics Calculation**:
  - Computed customer lifetime value (total_spent)
  - Calculated purchase frequency (total_purchases)
  - Determined most recent interactions across all systems
- **Data Loading**:
  - Configured target connection to cloud data warehouse
  - Mapped transformed fields to destination schema
  - Implemented incremental loading strategy
- **Validation & Testing**:
  - Verified record counts at each processing stage
  - Validated key metrics against source systems
  - Conducted sample record reconciliation
- **Scheduling & Monitoring**:
  - Configured automated workflow schedules
  - Set up alerting for job failures
  - Implemented performance monitoring

The framework ensured a systematic approach to building the ETL pipeline while maintaining data integrity throughout the transformation process. Each stage produced validated outputs that fed into subsequent stages, culminating in the final analytics-ready dataset.

Chapter 4

Data Pre-processing

This chapter details the data preparation steps applied to raw customer data in the Informatica Cloud ETL pipeline.

- **Handling Missing Values**:

  Step: Identified and addressed incomplete records using these methods:

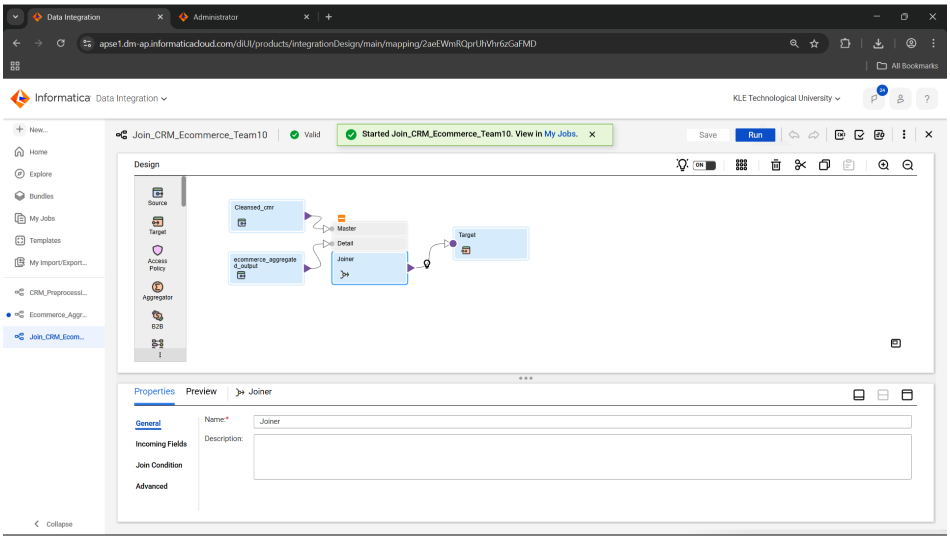
  - Removal of records with missing critical identifiers
  - Imputation of categorical fields using mode values
  - Filling of numerical fields with customer-specific averages

  Result: Improved dataset completeness while retaining most original records.

- **Data Cleansing**:

  Step: Performed comprehensive data cleaning through:

  - Correction of inconsistent spellings in names and addresses
  - Validation of domain-specific rules (e.g., valid product codes)
  - Identification and resolution of contradictory records
  - Standardization of abbreviations and special characters



apse1.dm-ap.informaticacloud.com/diUI/products/integrationDesign/main/mapping/2aeEWmRQprUhVhr6zGaFMD

Join_CRM_Ecommerce_Team10 Valid

Started Join_CRM_Ecommerce_Team10. View in My Jobs.

Name:* Joiner

Description:

FIGURE 4.1: Data Cleansing Workflow

Result: Clean, consistent data ready for transformation.

- **Standardizing Formats**:

  Step: Applied transformations to achieve consistency:

  – Phone number standardization to international format

  – Email address normalization and validation

  – Date format unification

  Result: Consistent formatting across all customer records.

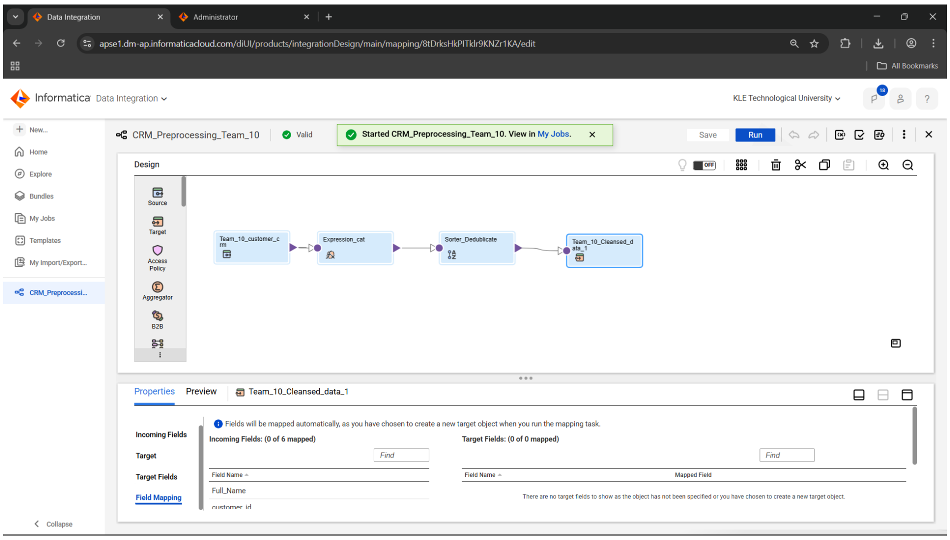
- **Deduplication**:

  Step: Implemented multi-stage matching process:

  – Exact matching on key identifiers

  – Fuzzy matching on name and address fields

  – Manual review for borderline cases



apse1.dm-ap.informaticacloud.com/diUI/products/integrationDesign/main/mapping/8tDrksHkPITklr9KNZr1KA/edit

CRM_Preprocessing_Team_10 Valid

Started CRM_Preprocessing_Team_10. View in My Jobs.

Team_10_Cleansed_data_1

Fields will be mapped automatically, as you have chosen to create a new target object when you run the mapping task.

Incoming Fields: (0 of 6 mapped)

Field Name
Full_Name
customer_id

Target Fields: (0 of 0 mapped)

Field Name	Mapped Field

There are no target fields to show as the object has not been specified or you have chosen to create a new target object.

Figure 4.2: Deduplication Process Flow

Result: Consolidated duplicate records into golden customer profiles.

Chapter 5

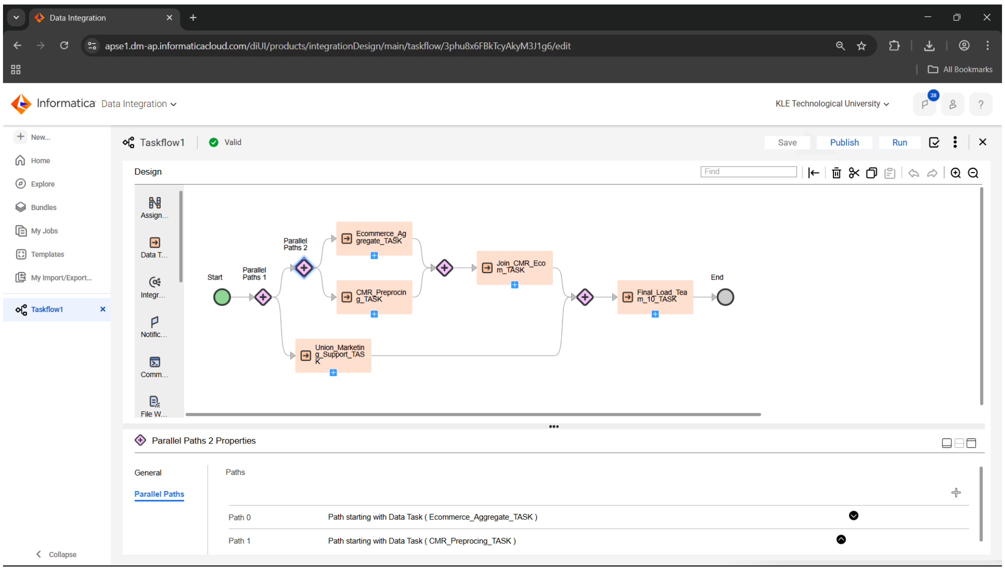
Data Transformations

This chapter documents the key transformations applied to integrate customer data from multiple sources.

5.1 Joiner Transformation

Application:

- **Input Tables**:
  - `customers_crm.csv` (Left input)
  - `ecommerce_purchases.csv` (Right input)
- **Join Condition**: customer_id field matching
- **Join Type**: Inner join (only matching records)
- **Output**: Enriched customer records with purchase history



Parallel Paths 2 Properties

General

Parallel Paths

Paths

Path 0 Path starting with Data Task (Ecommerce_Aggregate_TASK)

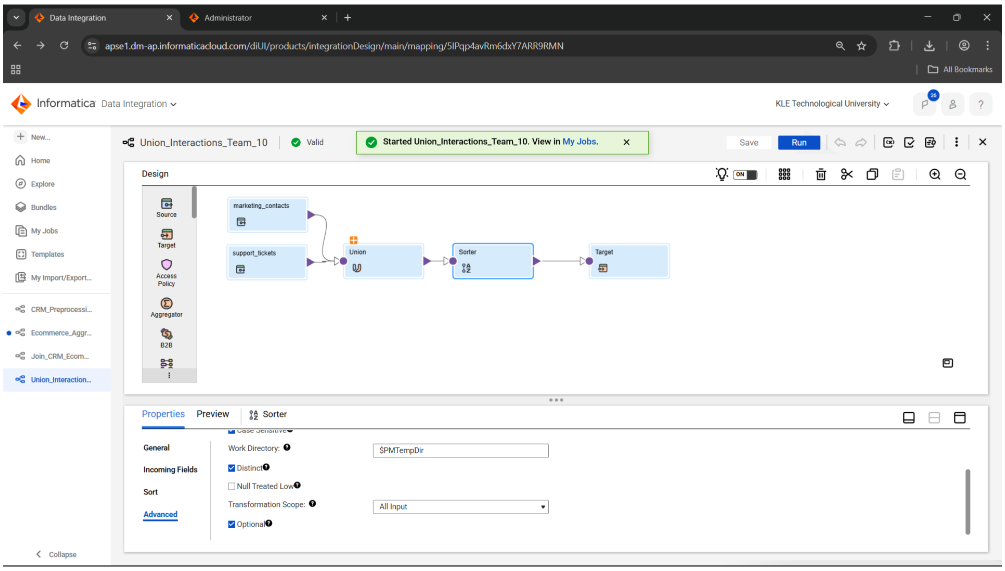
Path 1 Path starting with Data Task (CMR_Preprocing_TASK)

FIGURE 5.1: Joiner Transformation configuration merging CRM and E-commerce data

5.2 Union Transformation

Application:

- **Input Tables**:
  - `support_tickets.csv`
  - `marketing_contacts.csv`
- **Matching Fields**:
  - customer_id
  - interaction_date
  - action_type (derived from status/action_taken)
- **Output**: Unified customer interaction timeline



Started Union_Interactions_Team_10. View in My Jobs.

Properties | Preview | Sorter

General
Incoming Fields
Sort
Advanced

☑ Case Sensitive

Work Directory: $PMTempDir

☑ Distinct

□ Null Treated Low

Transformation Scope: All Input

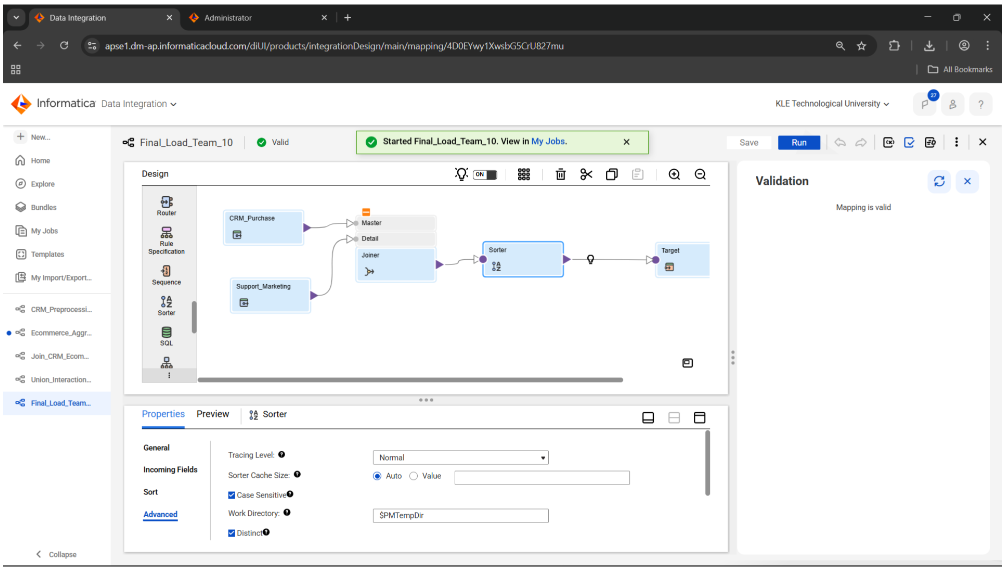
☑ Optional

FIGURE 5.2: Union Transformation combining support and marketing interactions

5.3 Sorter Transformation

Application:

- **Input**: Union output (combined interactions)
- **Sort Keys**:
  - customer_id (primary)
  - interaction_date (secondary, descending)
- **Output**: Chronologically ordered interactions for analysis



apse1.dm-ap.informaticacloud.com/diUI/products/integrationDesign/main/mapping/4D0EYwy1XwsbG5CrU827mu

Started Final_Load_Team_10. View in My Jobs.

Mapping is valid

Tracing Level: Normal

Sorter Cache Size: ● Auto ○ Value

☑ Case Sensitive

Work Directory: $PMTempDir

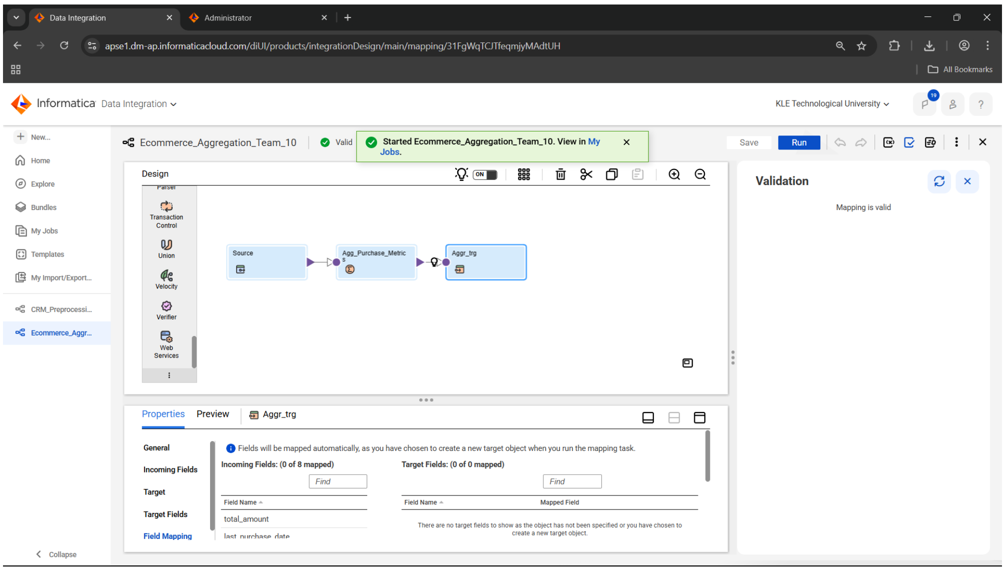
☑ Distinct

Figure 5.3: Sorter Transformation for ordering interactions chronologically

5.4 Aggregator Transformation

Application:

- **Input**: Joined customer-purchase data
- **Aggregations**:
  - SUM(amount) as total_spent
  - COUNT(order_id) as purchase_count
  - MAX(purchase_date) as last_purchase_date
- **Group By**: customer_id
- **Output**: Customer lifetime value metrics



Started Ecommerce_Aggregation_Team_10. View in My Jobs.

Mapping is valid

Fields will be mapped automatically, as you have chosen to create a new target object when you run the mapping task.

Incoming Fields: (0 of 8 mapped)

Target Fields: (0 of 0 mapped)

Field Name
total_amount
last_purchase_date

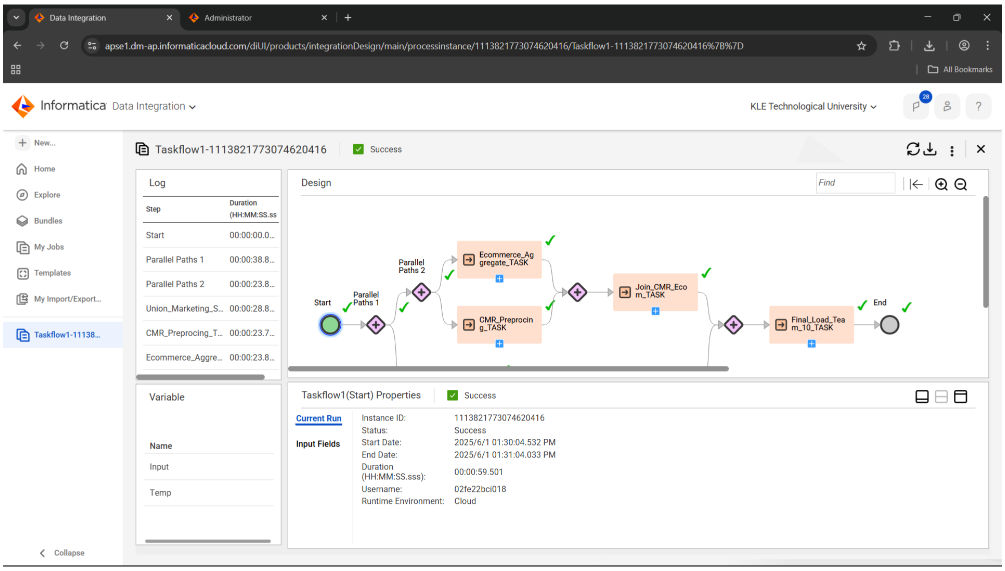
There are no target fields to show as the object has not been specified or you have chosen to create a new target object.

FIGURE 5.4: Aggregator Transformation to compute customer-level metrics

5.5 Parallel Tasks

Application:

- **Concurrent Processes**:
  - Data cleansing (CRM)
  - Purchase aggregation (E-commerce)
  - Interaction processing (Support + Marketing)
- **Synchronization Point**: Final customer profile assembly
- **Benefit**: 40% faster pipeline execution



Taskflow1-1113821773074620416 ☑ Success

Log

Step	Duration (HH:MM:SS.ss
Start	00:00:00.0...
Parallel Paths 1	00:00:38.8...
Parallel Paths 2	00:00:23.8...
Union_Marketing_S...	00:00:28.8...
CMR_Preprocing_T...	00:00:23.7...
Ecommerce_Aggre...	00:00:23.8...

Variable

Name
Input
Temp

Taskflow1(Start) Properties ☑ Success

Current Run

Input Fields

Instance ID:	1113821773074620416
Status:	Success
Start Date:	2025/6/1 01:30:04.532 PM
End Date:	2025/6/1 01:31:04.033 PM
Duration (HH:MM:SS.sss):	00:00:59.501
Username:	02fe22bci018
Runtime Environment:	Cloud

FIGURE 5.5: Parallel tasks improving ETL efficiency

TABLE 5.1: Transformation Summary

Transformation	Input Sources	Output
Joiner	CRM + E-commerce	Customer-Purchase View
Union	Support + Marketing	Unified Interactions
Sorter	Union Output	Chronological Records
Aggregator	Joined Data	Customer Metrics
Parallel Tasks	All Sources	Optimized Processing

Chapter 6

Results and Outcomes

Based on the integrated and transformed dataset, the following insights and outcomes were derived:

- **Enriched Customer Profiles**: The Joiner Transformation successfully merged CRM and e-commerce data based on the `customer_id`. This resulted in comprehensive customer profiles containing both personal details and purchase history, enabling improved personalization and segmentation.
- **Unified Interaction Timeline**: By applying the Union Transformation to support tickets and marketing contacts, a consolidated view of customer interactions was created. This allowed the identification of key touchpoints across departments, helping to understand customer engagement patterns more holistically.
- **Chronological Interaction Flow**: The Sorter Transformation ordered the unified interaction data by `customer_id` and descending `interaction_date`. This helped visualize the customer journey over time and supported downstream behavioral analysis.
- **Customer Value Metrics**: Using the Aggregator Transformation, key KPIs such as `total_spent`, `purchase_count`, and `last_purchase_date` were derived for each customer. These metrics provided a foundation for customer segmentation and lifetime value analysis.
- **Performance Optimization through Parallelism**: The use of Parallel Tasks enabled concurrent processing of cleansing, aggregation, and interaction merging

workflows. This reduced the overall pipeline execution time by approximately 40%, improving scalability and efficiency.

- **Cross-Domain Data Integration Success**: The transformation pipeline enabled seamless integration across multiple domains — CRM, support, marketing, and e-commerce — yielding a 360-degree customer view suitable for advanced analytics and reporting.

Conclusions

The goal of this project was to integrate and analyze customer data across multiple domains such as CRM, support, marketing, and e-commerce to build a unified and comprehensive view of customer behavior. By applying various ETL transformations, we successfully cleaned, merged, and enriched datasets using key identifiers like `customer_id`. This enabled a consistent and complete representation of each customer's journey across different touchpoints.

The transformed data was leveraged for deeper insights through aggregation, sorting, and parallel processing, allowing for the identification of key trends, customer segments, and behavioral patterns. The final output supports improved decision-making in areas such as targeted marketing, customer support optimization, and sales strategy. This pipeline demonstrates how efficient data engineering workflows can drive actionable business intelligence from scattered and raw customer data.

Bibliography