Eya Hadj Hassen

ETL Process using Talend:

I. Introduction

1. Purpose of the Talend Jobs

The purpose of these Talend jobs is to **extract** various raw data tables from the **staging area** stored in a MySQL database, apply **data cleaning and transformation**, and then **load** the transformed data into a **data warehouse** that follows a **star schema** structure, also managed in MySQL.

2. ETL Overview

- **Extraction:** Retrieve raw data from the MySQL staging area.
- **Transformation:** Apply necessary data cleaning, filtering, and restructuring to align with the star schema.
- **Loading:** Store the cleaned and transformed data into the MySQL data warehouse.

II. Talend Job Descriptions:

1. Customer Data Cleaning and Transformation Job

a. Objective

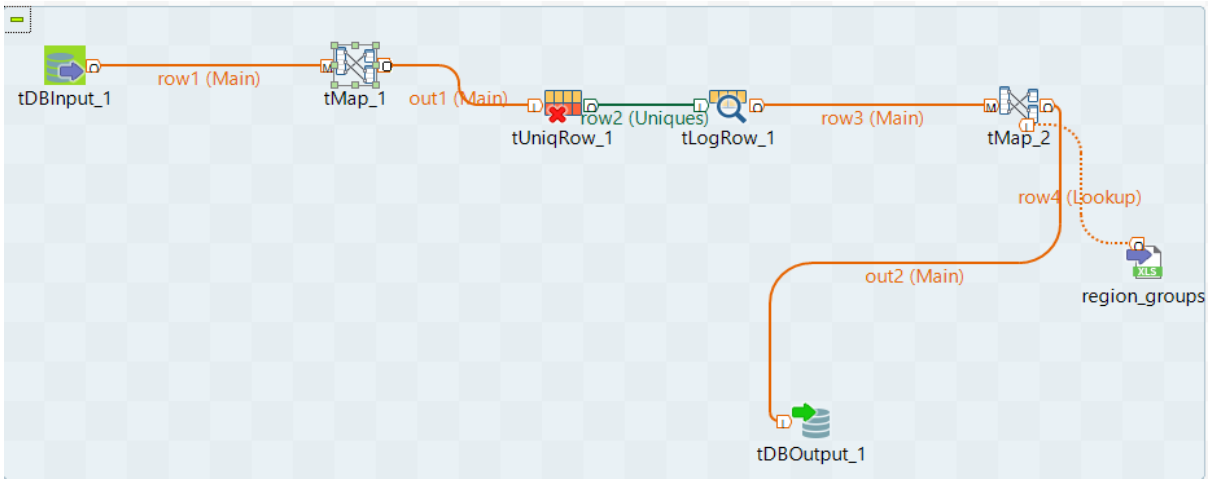
The purpose of this Talend job is to clean and transform customer data extracted from the staging area in the MySQL database before loading it into the data warehouse. The job ensures data consistency, removes duplicates, and enriches customer records with region group classification.

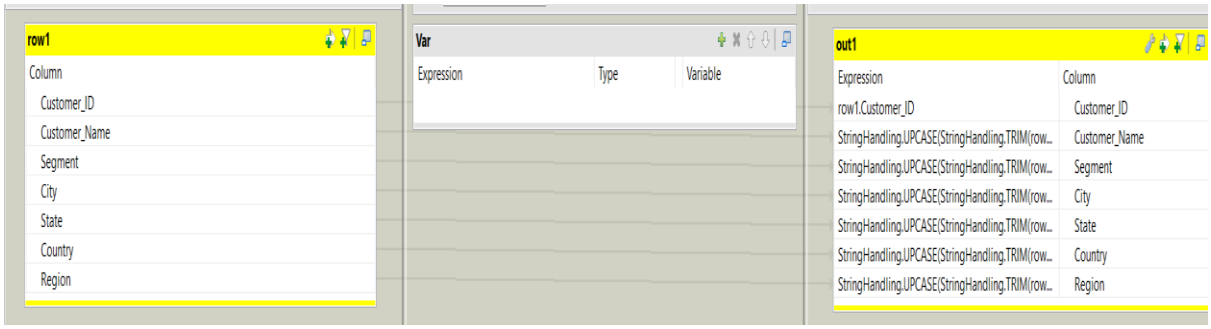
b. Process Overview

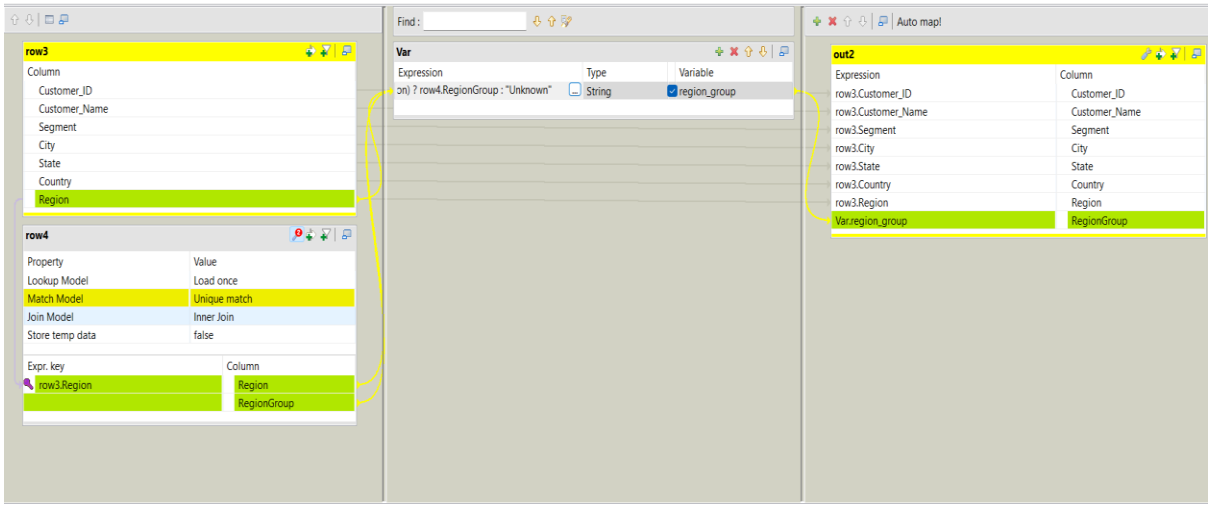
The job consists of multiple components that handle the data extraction, transformation, and loading (ETL) process:

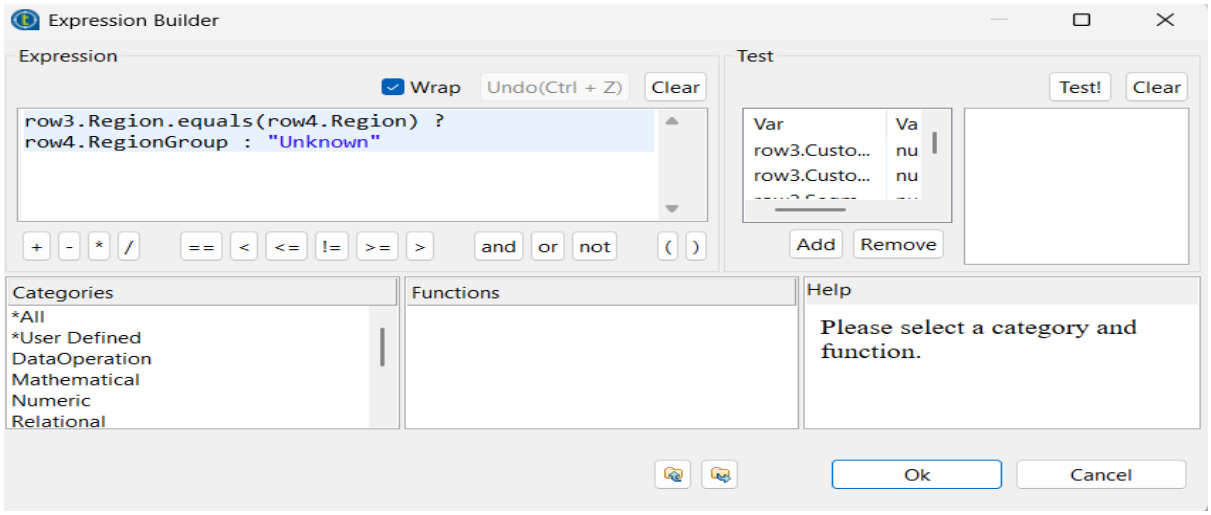
- **Extracting Data (tDBInput)**
    - The **tDBInput** component retrieves raw customer data from the MySQL staging database. The extracted data includes:
        - Customer ID
        - Customer Name
        - Segment
        - City
        - State
        - Country
        - Region

- **Data Cleaning and Standardization (tMap)**
    - The extracted data flows through a **tMap** component where cleaning operations are applied:
        - **String Trimming and Uppercasing**: The customer name, segment, city, state, country, and region fields are standardized using StringHandling.TRIM() and StringHandling.UPCASE() functions to remove leading/trailing spaces and ensure consistent casing.
- **Removing Duplicates (tUniqRow)**
    - The **tUniqRow** component filters out duplicate records based on unique Customer IDs, ensuring data integrity.
- **Logging Unique Records (tLogRow)**
    - The unique records are logged for debugging and verification before further processing.
- **Region Group Lookup (tMap with Lookup)**
    - The transformed data is enriched by joining with a **region group reference dataset** (Excel file).
    - The lookup operation:
        - Matches the Region column from the customer data with the RegionGroup column from the reference dataset.
        - Assigns a **RegionGroup** value to each customer record.
        - If no match is found, a default value of "Unknown" is assigned.
- **Loading Data into Data Warehouse (tDBOutput)**
    - The final cleaned and enriched customer dataset is loaded into the MySQL data warehouse.




row1
Column
Customer_ID
Customer_Name
Segment
City
State
Country
Region
Var
Expression
Type
Variable
out1
Expression
Column
row1.Customer_ID
Customer_ID
StringHandling.UPCASE(StringHandling.TRIM(row...
Customer_Name
Segment
City
State
Country
Region


row3
Column
Customer_ID
Customer_Name
Segment
City
State
Country
Region
row4
Property
Value
Lookup Model
Load once
Match Model
Unique match
Join Model
Inner Join
Store temp data
false
Expr. key
Column
row3.Region
Region
RegionGroup
Find :
Var
Expression
Type
Variable
on) ? row4.RegionGroup : "Unknown"
String
region_group
Auto map!
out2
Expression
Column
row3.Customer_ID
Customer_ID
row3.Customer_Name
Customer_Name
row3.Segment
Segment
row3.City
City
row3.State
State
row3.Country
Country
row3.Region
Region
Var.region_group
RegionGroup


Expression Builder
Expression
Wrap
Undo(Ctrl + Z)
Clear
row3.Region.equals(row4.Region) ?
row4.RegionGroup : "Unknown"
+ - * / == < <= != >= > and or not ( )
Test
Test!
Clear
Var
Va
row3.Custo...
nu
row3.Custo...
nu
Add
Remove
Categories
*All
*User Defined
DataOperation
Mathematical
Numeric
Relational
Functions
Help
Please select a category and function.
Ok
Cancel

2. Product Data Cleaning and Transformation Job

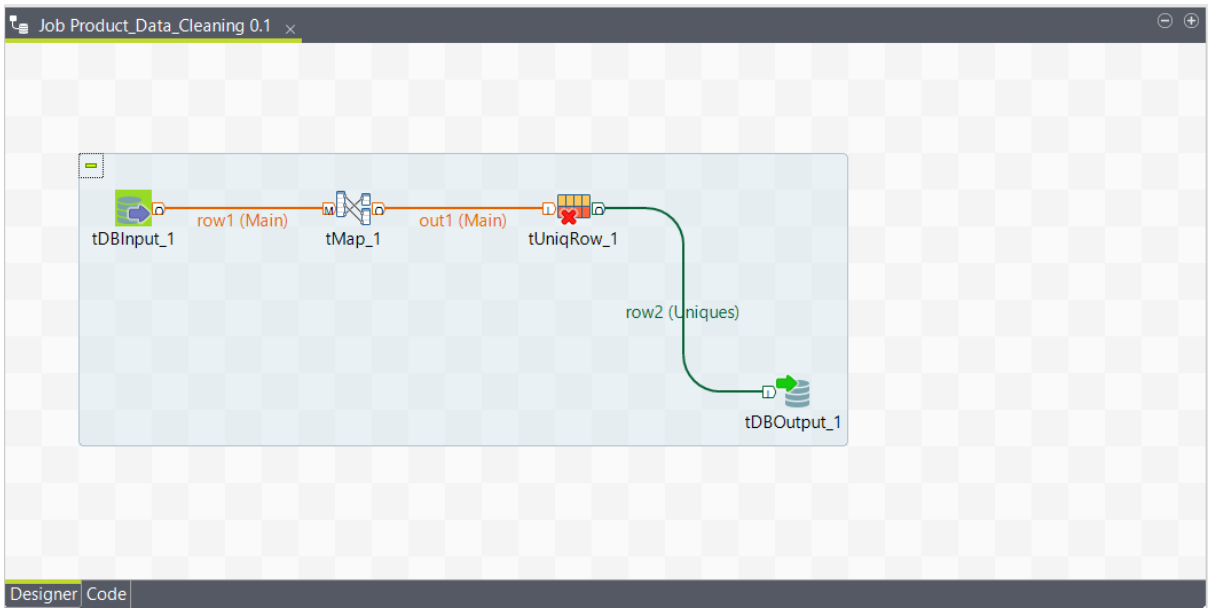
a. Objective

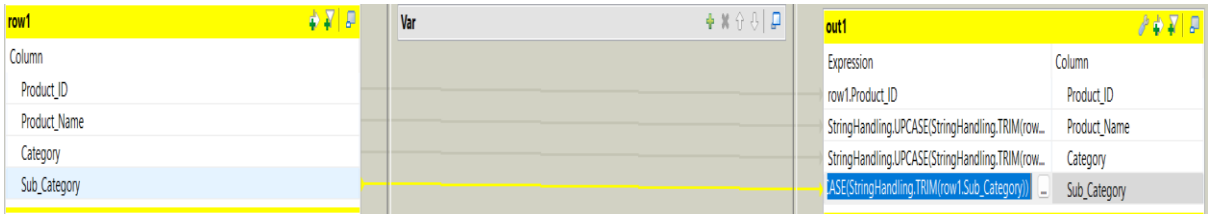
The purpose of this Talend job is to clean and transform product data extracted from the staging area in the MySQL database before loading it into the data warehouse. The job ensures data consistency, removes duplicates, and standardizes data.

b. Process Overview

The job consists of multiple components that handle the data extraction, transformation, and loading (ETL) process:

- **Extracting Data (tDBInput)**
    - The **tDBInput** component retrieves raw product data from the MySQL staging database. The extracted data includes:
        - Product ID
        - Product Name
        - Category
        - Sub_Category
- **Data Cleaning and Standardization (tMap)**
    - The extracted data flows through a **tMap** component where cleaning operations are applied:
        - **String Trimming and Uppercasing**: The product name, Category, and Sub_Category fields are standardized using StringHandling.TRIM() and StringHandling.UPCASE() functions to remove leading/trailing spaces and ensure consistent casing.
- **Removing Duplicates (tUniqRow)**
    - The **tUniqRow** component filters out duplicate records based on unique Product IDs, ensuring data integrity.
- **Loading Data into Data Warehouse (tDBOutput)**
    - The final cleaned and enriched Product dataset is loaded into the MySQL data warehouse.


Job Product_Data_Cleaning 0.1
tDBInput_1
row1 (Main)
tMap_1
out1 (Main)
tUniqRow_1
row2 (Uniques)
tDBOutput_1
Designer
Code


row1
Column
Product_ID
Product_Name
Category
Sub_Category
Var
out1
Expression
Column
row1.Product_ID
Product_ID
StringHandling.UPCASE(StringHandling.TRIM(row...
Product_Name
StringHandling.UPCASE(StringHandling.TRIM(row...
Category
ASE(StringHandling.TRIM(row1.Sub_Category))
Sub_Category

3. Shipping Data Cleaning and Transformation Job

a. Objective

The purpose of this Talend job is to clean and transform shipping data extracted from the staging area in the MySQL database before loading it into the data warehouse. The job ensures data consistency, removes duplicates, and categorizes shipping costs for better analysis.

b. Process Overview

The job consists of multiple components that handle the data extraction, transformation, and loading (ETL) process:

- **Extracting Data (tDBInput)**

The **tDBInput** component retrieves raw shipping data from the MySQL staging database. The extracted data includes:

  - Order ID
  - Ship Date
  - Ship Mode
  - Delivery Days
  - Shipping Cost

- **Removing Duplicates (tUniqRow)**

The **tUniqRow** component filters out duplicate records based on unique **Order ID**, ensuring data integrity.

- **Data Cleaning and Transformation (tMap)**

The extracted data flows through a **tMap** component where cleaning operations and transformations are applied:

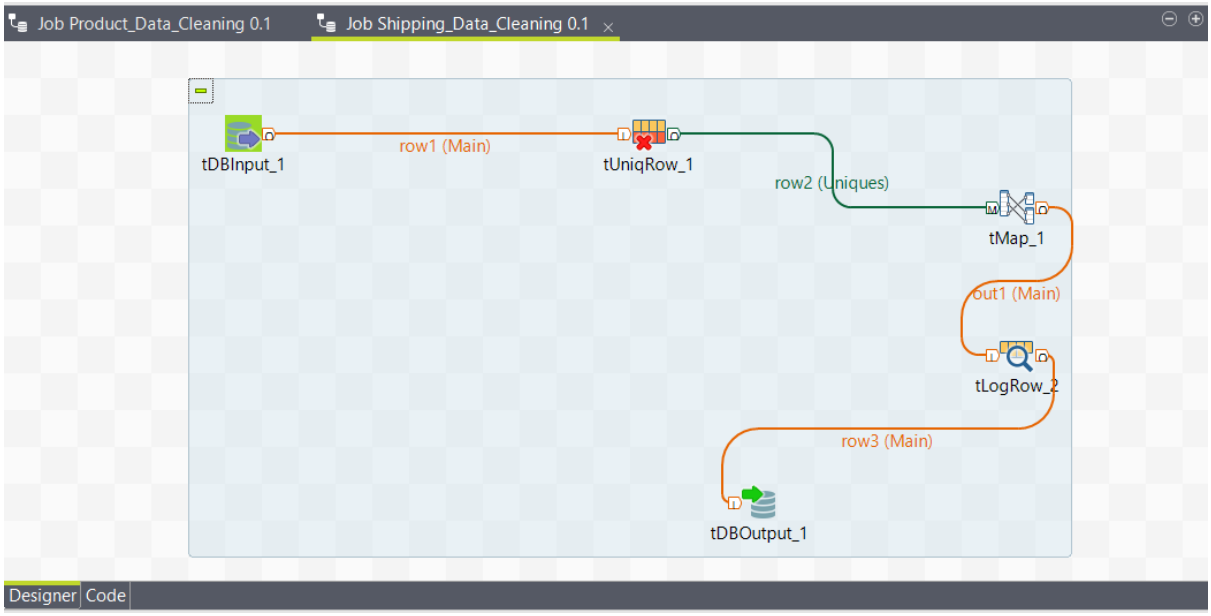
- **Shipping Cost Categorization**
  - The **Shipping Cost** is categorized into three levels:
    - **Low**: Shipping cost ≤ 311
    - **Medium**: Shipping cost between 311 and 622
    - **High**: Shipping cost > 622
    - If the shipping cost is null, it is assigned a default value of **"Unknown"**.

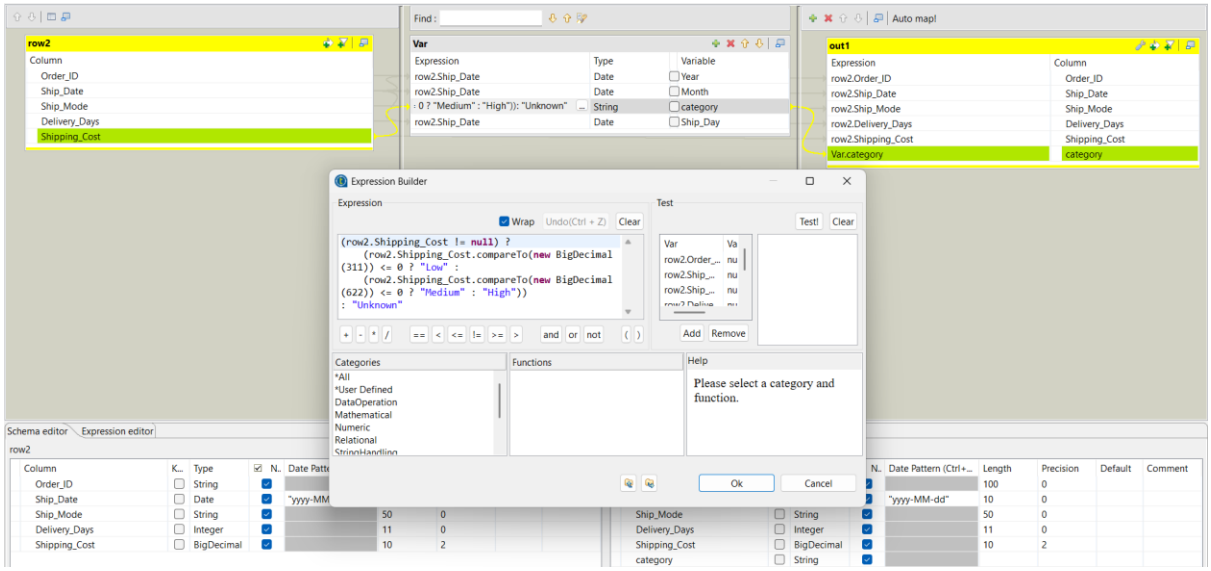
- **Logging Processed Data (tLogRow)**

The transformed records are logged for debugging and verification before further processing.

- **Loading Data into Data Warehouse (tDBOutput)**

The final cleaned and transformed shipping dataset is loaded into the MySQL data warehouse.





4. Time Data Cleaning and Transformation Job

a. Objective

The purpose of this Talend job is to clean and standardize time-related data extracted from the MySQL staging area before loading it into the data warehouse. The job ensures data consistency, removes duplicates, and prepares time-related records for further analysis.

b. Process Overview

This ETL process consists of multiple components for extracting, cleaning, and loading data:

- **Extracting Data (tDBInput)**

The **tDBInput** component retrieves raw time-related data from the MySQL staging database. The extracted data includes:

  - Time ID
  - Order Date
  - Ship Date
  - Delivery Date

- **Removing Duplicates (tUniqRow)**

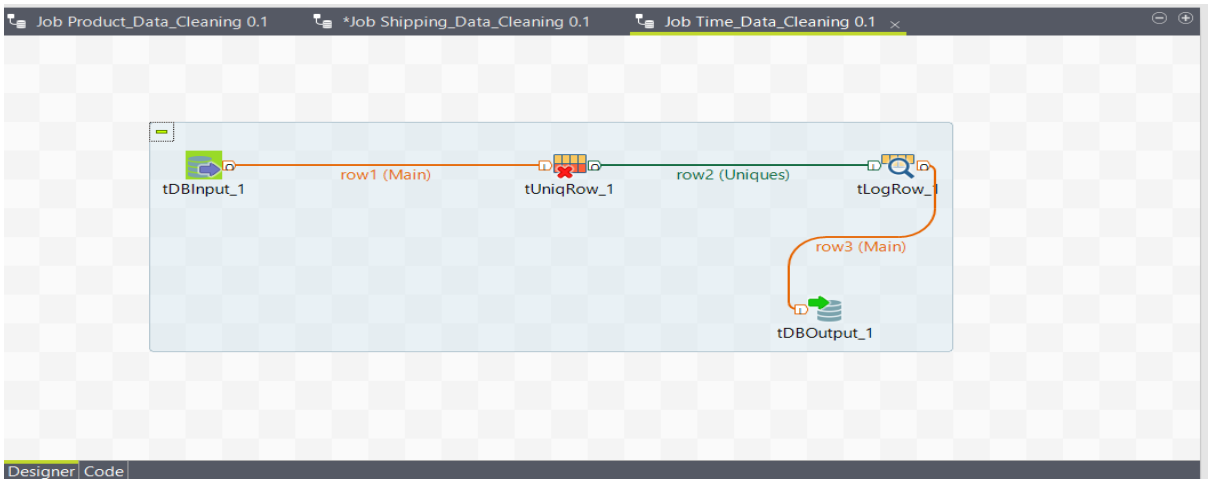
The **tUniqRow** component ensures that only unique **Time ID** records are retained, removing any duplicate entries to maintain data integrity.

- **Logging Processed Data (tLogRow)**

The unique records are logged for debugging and verification before they are loaded into the data warehouse.

- **Loading Data into Data Warehouse (tDBOutput)**

The cleaned and deduplicated time-related dataset is stored in the MySQL data warehouse for further analysis and reporting.



5. Sales Data Cleaning and Transformation Job

a. Objective

The purpose of this Talend job is to clean and standardize sales transaction data extracted from the MySQL staging area before loading it into the data warehouse. The job ensures data consistency, removes duplicates, and prepares sales records for accurate reporting and analysis.

b. Process Overview

This ETL process consists of multiple components for extracting, cleaning, and loading data:

- **Extracting Data (tDBInput)**

The **tDBInput** component retrieves raw sales data from the MySQL staging database. The extracted data includes:

  - Order ID
  - Product ID
  - Customer ID
  - Time ID
  - Sales
  - Profit
  - Quantity
  - Discount
- **Removing Duplicates (tUniqRow)**

The **tUniqRow** component ensures that only unique sales transaction records are retained by filtering out duplicates based on the following key attributes:

  - Order ID
  - Product ID
  - Customer ID
  - Time ID

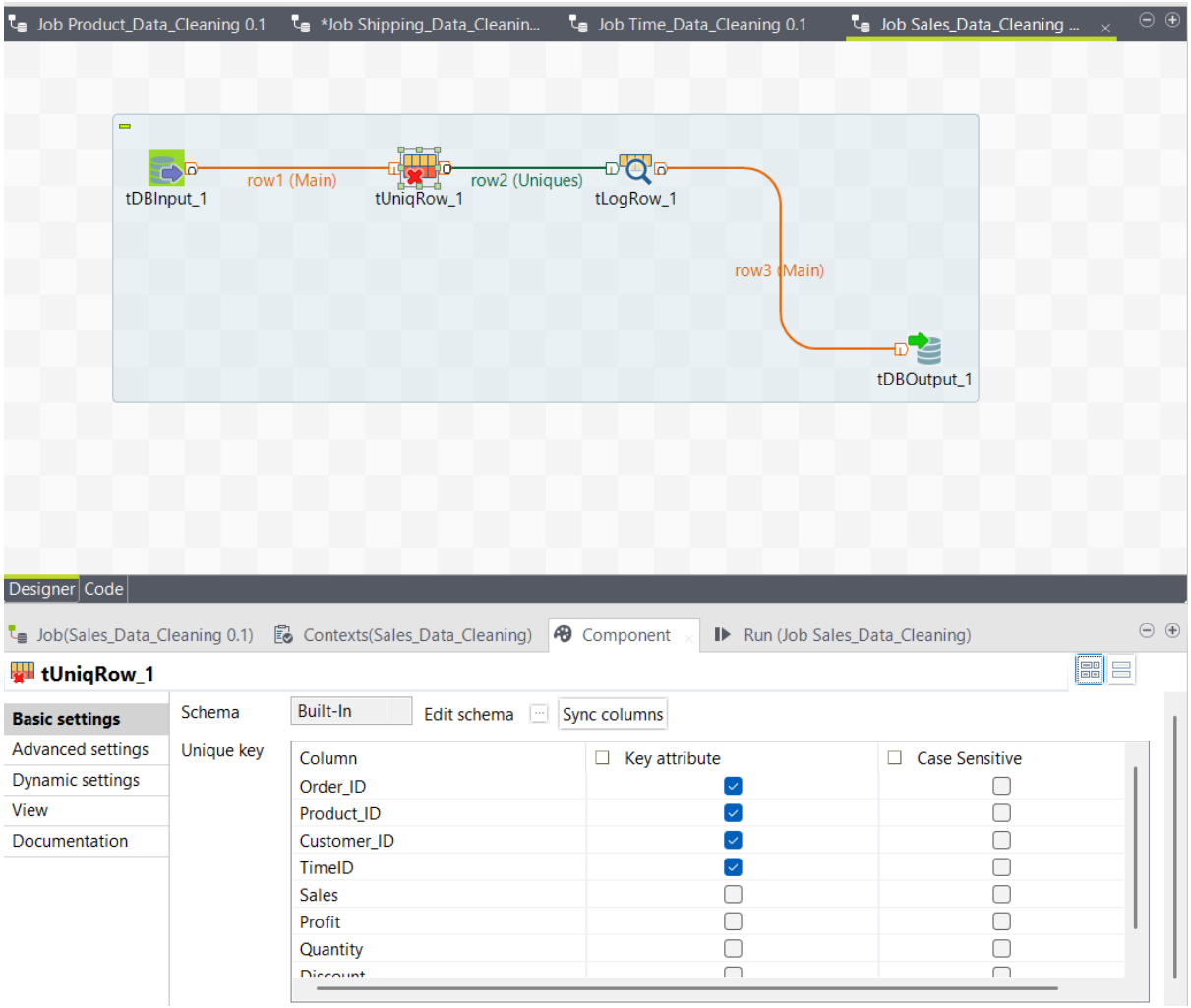
This step guarantees data integrity and avoids duplicate sales reporting.

- **Logging Processed Data (tLogRow)**

The unique records are logged for debugging and verification before they are loaded into the data warehouse.

- **Loading Data into Data Warehouse (tDBOutput)**

The cleaned and deduplicated sales dataset is stored in the MySQL data warehouse for further analysis and business intelligence reporting.


Job Product_Data_Cleaning 0.1
*Job Shipping_Data_Cleanin...
Job Time_Data_Cleaning 0.1
Job Sales_Data_Cleaning ...
tDBInput_1
row1 (Main)
tUniqRow_1
row2 (Uniques)
tLogRow_1
row3 (Main)
tDBOutput_1
Designer
Code
Job(Sales_Data_Cleaning 0.1)
Contexts(Sales_Data_Cleaning)
Component
Run (Job Sales_Data_Cleaning)
tUniqRow_1
Basic settings
Advanced settings
Dynamic settings
View
Documentation
Schema
Built-In
Edit schema
Sync columns
Unique key
Column
Key attribute
Case Sensitive
Order_ID
Product_ID
Customer_ID
TimeID
Sales
Profit
Quantity