

# Data Integration Using Talend Open Studio

Import CustomerData.xlsx and Online\_Sales\_preprocessed.xlsx, and justify the data pattern:

The image displays four screenshots of the Talend Open Studio interface, showing the steps to add metadata and schema files to a repository.

**Top Left: Edit an existing Excel File - Step 2 of 4**  
This window shows the 'File Settings' section where the file path is set to 'C:/Users/JSY/Desktop/WQD7005/final/Online\_Sales\_preprocessed.xlsx'. The 'Read excel2007 file format(xlsx)' checkbox is checked, and the 'Generation mode' is set to 'Memory-consuming(User mode)'. The 'File Viewer and Sheets setting' section shows 'All sheets/DSelect sheet' selected, with 'sheet1' chosen from the list.

**Top Right: New Excel File - Step 4 of 4**  
This window shows the 'Schema' section where the schema name is 'Online\_Sales'. The 'Description of the Schema' table lists columns and their properties:

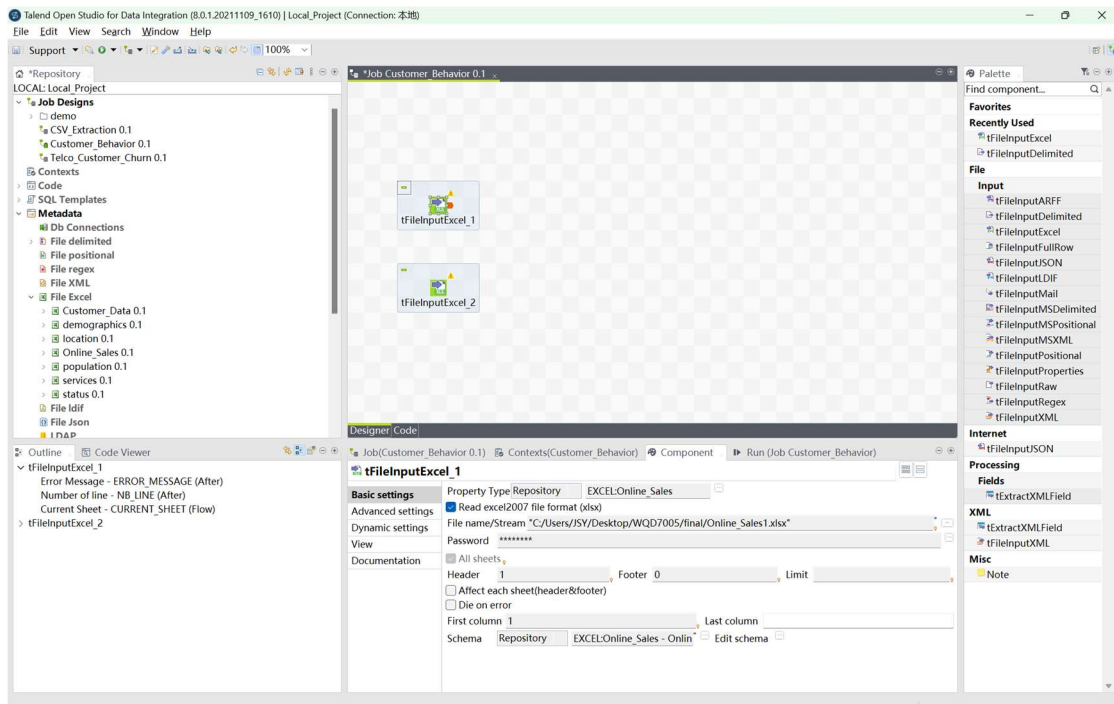
Column	K...	Type	N	Date Pattern...	Length	Precisi...	Defa...	Comm...
CustomerID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
Transaction_ID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
Transaction_Date	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	*yyy-MM-	10	0		
Product_SKU	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		14	0		
Product_Description	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		54	0		

**Bottom Left: Edit an existing Excel File - Step 2 of 4**  
This window shows the 'File Settings' section where the file path is set to 'C:/Users/JSY/Desktop/WQD7005/final/CustomersData.xlsx'. The 'Read excel2007 file format(xlsx)' checkbox is checked, and the 'Generation mode' is set to 'Memory-consuming(User mode)'. The 'File Viewer and Sheets setting' section shows 'All sheets/DSelect sheet' selected, with 'Customers' chosen from the list.

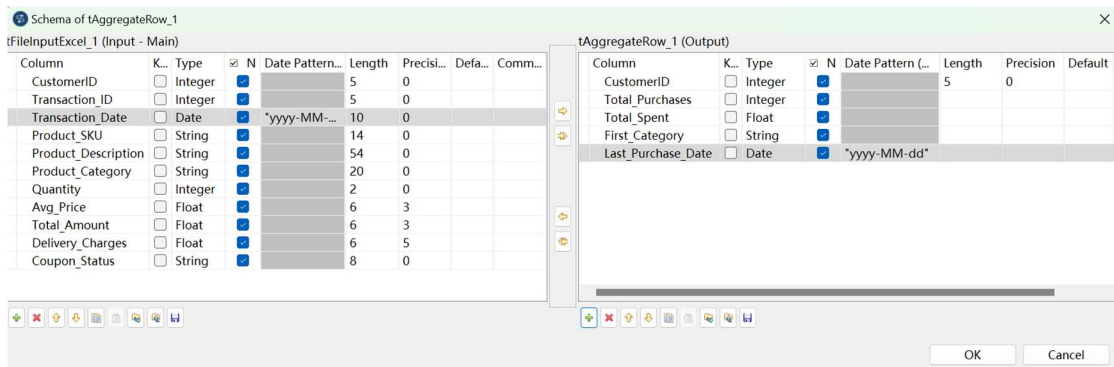
**Bottom Right: New Excel File - Step 4 of 4**  
This window shows the 'Schema' section where the schema name is 'Customer\_Data'. The 'Description of the Schema' table lists columns and their properties:

Column	K...	Type	N	Date Pattern ..	Length	Precisi...	Defa...	Comm...
CustomerID	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		5	0		
Gender	<input type="checkbox"/>	Charac...	<input checked="" type="checkbox"/>		1	0		
Location	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		13	0		
Tenure_Months	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		

Create a new job named 'Customer\_Behavior' and load Online\_Sales and Customer\_Data files in the tFileInputExcel components:



Add columns of Total\_Purchases, Total\_Spent, First\_Category, Last\_Purchased\_Date, and get these data from count of TransactionID, sum of Total\_Amount, first of Product\_Category, and max of Transaction\_Date, separately using Operations in tAggregateRow component:



The screenshot displays the Talend Studio interface for a job named "Job Customer\_Behavior 0.1". The top pane shows the job design with two input components, "tFileInputExcel\_1" and "tFileInputExcel\_2", connected to an aggregation component "tAggregateRow\_1". The "tFileInputExcel\_1" component is linked to "tAggregateRow\_1" via a "row1 (Main)" link. The bottom pane shows the configuration for "tAggregateRow\_1".

**Basic settings**

- Schema: Built-In
- Edit schema
- Sync columns

**Group by**

Output column	Input column position
CustomerID	CustomerID

**Operations**

Output column	Function	Input column position	Ignore null values
Total_Purchases	count	Transaction_ID	<input type="checkbox"/>
Total_Spent	sum	Total_Amount	<input type="checkbox"/>
First_Category	first	Product_Category	<input type="checkbox"/>
Last_Purchase_Date	max	Transaction_Date	<input type="checkbox"/>

Join the output of Aggregation and the CustomerData file using tMap component, and run the whole working process, then we get one combined dataset named Customer\_Behavior.xlsx, having 8 columns of CustomerID, Gender, Location, Total\_Purchases, Total\_Spent, First\_Category, Last\_Purchased\_Date, and Tenure\_Months, with 1468 customers' rows.

Talend Open Studio for Data Integration - tMap - tMap\_1

Find:

Var:

Auto map!

row3

Column
CustomerID
Total_Purchases
Total_Spent
First_Category
Last_Purchase_Date

row4

Expr. key	Column
row3.CustomerID	CustomerID
	Gender
	Location
	Tenure_Months

out1

Expression	Column
row3.CustomerID	CustomerID
row4.Gender	Gender
row4.Location	Location
row3.Total_Purchases	Total_Purchases
row3.Total_Spent	Total_Spent
row3.First_Category	First_Category
row3.Last_Purchase_Date	Last_Purchase_Date
row4.Tenure_Months	Tenure_Months

Schema editor

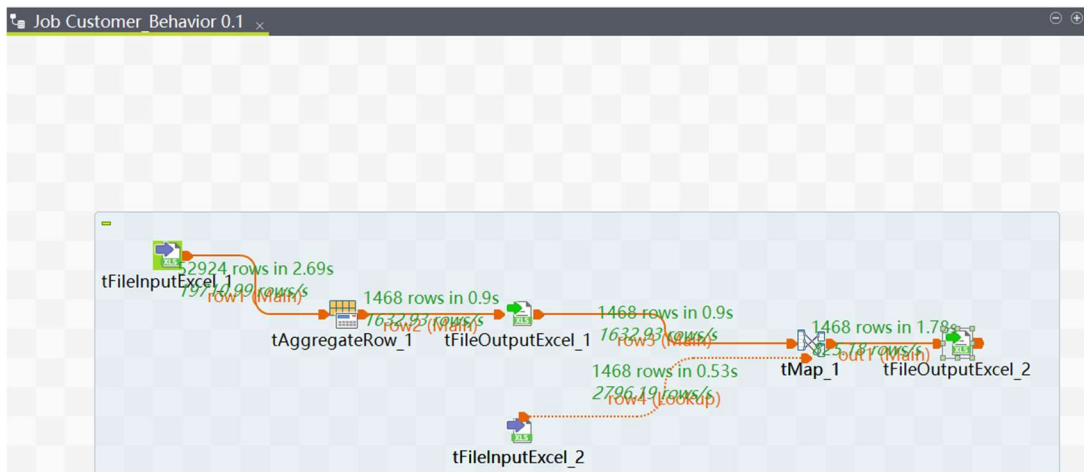
row4

Column	K..	Type	N	Date Pattern (Ctrl..)	Length	Precision	Default	Comment
CustomerID		Integer			5	0		
Gender		Character			1	0		
Location		String			13	0		
Tenure_Months		Integer			2	0		

out1

Column	K..	Type	N	Date Pattern (Ctrl..)	Length	Precision	Default	Comment
CustomerID		Integer			5	0		
Gender		Character			1	0		
Location		String			13	0		
Total_Purchases		Integer						
Total_Spent		Float						
First_Category		String						
Last_Purchase_Date		Date		yyyy-MM-dd				
Tenure_Months		Integer			2	0		

Apply Ok Cancel



Designer Code

Job(Customer\_Behavior 0.1) Contexts(Customer\_Behavior) Component Run (Job Customer\_Behavior)

tFileOutputExcel\_2

Basic settings

Advanced settings

Dynamic settings

View

Documentation

Use Output Stream

File Name "D:/Talend Studio/Customer\_Behavior.xlsx"

Sheet name "Sheet1"

Include header

Append existing file

Is absolute Y pos.

Font Default

Define all columns auto size

Define column auto size

Column	Auto size
CustomerID	<input type="checkbox"/>
Gender	<input type="checkbox"/>
Location	<input type="checkbox"/>
Total_Purchases	<input type="checkbox"/>
Total_Spent	<input type="checkbox"/>
First_Category	<input type="checkbox"/>
Last_Purchase_Date	<input type="checkbox"/>
Tenure_Months	<input type="checkbox"/>

Protect file

Schema Built-In Edit schema Sync columns