



**UNIVERSITI
MALAYA**

*Faculty of Computer Science
and Information Technology*

**WQD7005
DATA MINING**

ALTERNATIVE ASSESSMENT 1

**LECTURER NAME:
PROF. DR. TEH YING WAH**

**STUDENT NAME:
AMIRA HANEE BINTI SAIFULAZRI
17202918/1**

GitHub Link:

<https://github.com/amirahanee/WQD7005AltAsmnt1/tree/main>

Talend Data Integration

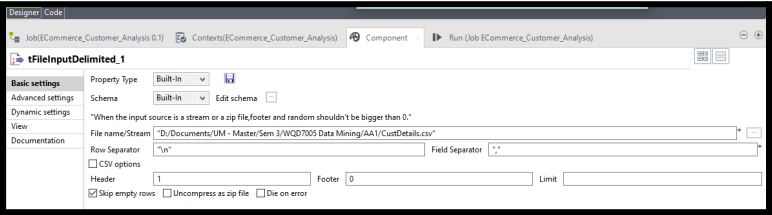
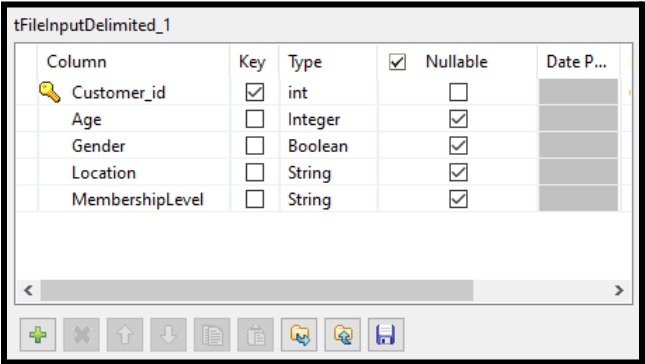
Talend Data Integration is a comprehensive open-source tool used for Extract, Transform, Load (ETL) processes. Once the data is imported into Talend, users can apply a variety of data integration tasks, including data cleansing, transformation, and loading into target systems. The importation of data into Talend Data Integration marks the initial phase of a workflow where the data will undergo various operations to meet specific business or analytical requirements.

I used multiple nodes to complete this integration, which are:

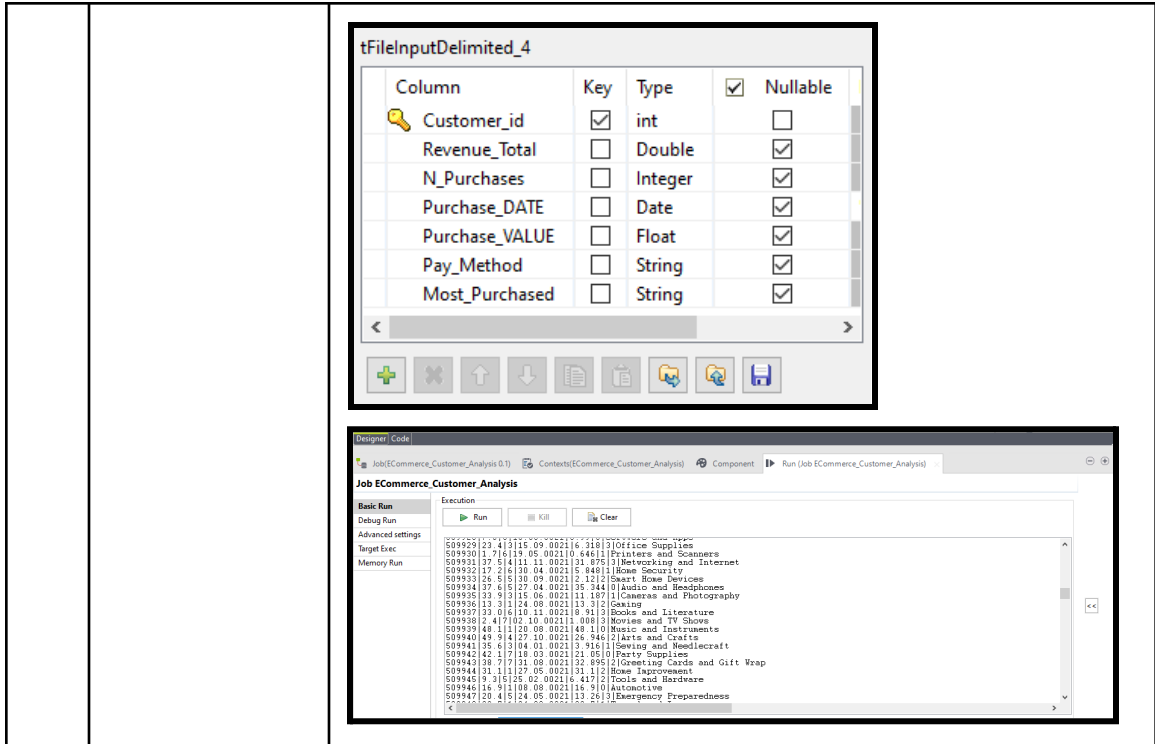
1. tFileInputDelimited

A component used for reading data from delimited text files. It is part of the input/output (I/O) family of components and is specifically designed to handle delimited files, where data is separated by a specified delimiter, such as a comma, tab, semicolon, etc.

It is being used three times to upload three different datasets.

No	Dataset Name	Evidence
1	CustDetails	<div></div> <div></div>

		<div><div><div><div>Designer Code</div><div>Job forCustDetails 0.1</div><div>Contexts forCustDetails</div><div>Component</div><div>Run (Job forCustDetails)</div></div><div><div>Job forCustDetails</div><div>Execution</div><div>Basic Run</div><div>Debug Run</div><div>Advanced settings</div><div>Target Exec</div><div>Memory Run</div></div><div><div>511221 27 true Wisconsin Nova</div><div>511223 41 false Maine Nova</div><div>511224 44 false Rhode Island Nova</div><div>511225 52 true Missouri Nova</div><div>511226 28 true Iowa Nova</div><div>511227 24 true New Jersey Nova</div><div>511228 62 true Kentucky Nova</div><div>511229 60 false California Nova</div><div>511230 25 true Massachusetts Nova</div><div>511231 59 false Maine Nova</div><div>511232 59 true New York Nova</div><div>511233 48 true Mississippi Nova</div><div>511234 55 true Alaska Nova</div><div>511235 49 true Delaware Nova</div><div>511236 21 true Connecticut Nova</div><div>511237 57 true South Carolina Nova</div><div>511238 54 true Minnesota Nova</div><div>511239 59 true Utah Nova</div><div>511240 72 ..</div></div></div></div>																																										
2	CustLog	<div><div><div><div>Designer Code</div><div>Job ECommerce_Customer_Analysis 0.1</div><div>Contexts ECommerce_Customer_Analysis</div><div>Component</div><div>Run (Job ECommerce_Customer_Analysis)</div></div><div><div>tFileInputDelimited_3</div><div>Basic settings</div><div>Advanced settings</div><div>Dynamic settings</div><div>View</div><div>Documentation</div></div><div><div>Property Type</div><div>Built-In</div><div>Schema</div><div>Built-In</div><div>Edit schema</div><div>"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."</div><div>File name/Stream</div><div>"D:/Documents/UM - Master/Sem 3/WQD7005 Data Mining/AA1/CustLog.csv"</div><div>Row Separator</div><div>"\n"</div><div>Field Separator</div><div>","</div><div>Header</div><div>1</div><div>Footer</div><div>0</div><div>Limit</div><div></div><div><input checked="" type="checkbox"/> Skip empty rows</div><div><input type="checkbox"/> Uncompress as zip file</div><div><input type="checkbox"/> Die on error</div></div></div></div> <div><div><div>tFileInputDelimited_3</div><table><tr><th>Column</th><th>Key</th><th>Type</th><th><input checked="" type="checkbox"/></th><th>Nullable</th><th>Date</th></tr><tr><td> Customer_id</td><td><input checked="" type="checkbox"/></td><td>int</td><td><input type="checkbox"/></td><td><input type="checkbox"/></td><td></td></tr><tr><td>Time_Spent</td><td><input type="checkbox"/></td><td>Integer</td><td><input type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td></td></tr><tr><td>Browser</td><td><input type="checkbox"/></td><td>Boolean</td><td><input checked="" type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td></td></tr><tr><td>Newsletter</td><td><input type="checkbox"/></td><td>Boolean</td><td><input type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td></td></tr><tr><td>Voucher</td><td><input type="checkbox"/></td><td>Boolean</td><td><input type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td></td></tr><tr><td>Churn</td><td><input type="checkbox"/></td><td>Boolean</td><td><input type="checkbox"/></td><td><input checked="" type="checkbox"/></td><td></td></tr></table></div></div> <div><div><div>Designer Code</div><div>Job ECommerce_Customer_Analysis 0.1</div><div>Contexts ECommerce_Customer_Analysis</div><div>Component</div><div>Run (Job ECommerce_Customer_Analysis)</div></div><div><div>Job ECommerce_Customer_Analysis</div><div>Execution</div><div>Basic Run</div><div>Debug Run</div><div>Advanced settings</div><div>Target Exec</div><div>Memory Run</div></div><div><div>505126 1888 false false true false</div><div>505127 680 false false false true</div><div>505128 455 false true false false</div><div>505129 312 false false false true</div><div>505130 448 false true false false</div><div>505131 957 false true false false</div><div>505132 397 false false false false</div><div>505133 978 false false false false</div><div>505134 167 false true false false</div><div>505135 240 false false false false</div><div>505136 964 true true false false</div><div>505137 628 false false false false</div><div>505138 701 false false false false</div></div></div>	Column	Key	Type	<input checked="" type="checkbox"/>	Nullable	Date	Customer_id	<input checked="" type="checkbox"/>	int	<input type="checkbox"/>	<input type="checkbox"/>		Time_Spent	<input type="checkbox"/>	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>		Browser	<input type="checkbox"/>	Boolean	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Newsletter	<input type="checkbox"/>	Boolean	<input type="checkbox"/>	<input checked="" type="checkbox"/>		Voucher	<input type="checkbox"/>	Boolean	<input type="checkbox"/>	<input checked="" type="checkbox"/>		Churn	<input type="checkbox"/>	Boolean	<input type="checkbox"/>	<input checked="" type="checkbox"/>	
Column	Key	Type	<input checked="" type="checkbox"/>	Nullable	Date																																							
Customer_id	<input checked="" type="checkbox"/>	int	<input type="checkbox"/>	<input type="checkbox"/>																																								
Time_Spent	<input type="checkbox"/>	Integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>																																								
Browser	<input type="checkbox"/>	Boolean	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>																																								
Newsletter	<input type="checkbox"/>	Boolean	<input type="checkbox"/>	<input checked="" type="checkbox"/>																																								
Voucher	<input type="checkbox"/>	Boolean	<input type="checkbox"/>	<input checked="" type="checkbox"/>																																								
Churn	<input type="checkbox"/>	Boolean	<input type="checkbox"/>	<input checked="" type="checkbox"/>																																								
3	CustSales	<div><div><div><div>Designer Code</div><div>Job ECommerce_Customer_Analysis 0.1</div><div>Contexts ECommerce_Customer_Analysis</div><div>Component</div><div>Run (Job ECommerce_Customer_Analysis)</div></div><div><div>tFileInputDelimited_4</div><div>Basic settings</div><div>Advanced settings</div><div>Dynamic settings</div><div>View</div><div>Documentation</div></div><div><div>Property Type</div><div>Built-In</div><div>Schema</div><div>Built-In</div><div>Edit schema</div><div>"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."</div><div>File name/Stream</div><div>"D:/Documents/UM - Master/Sem 3/WQD7005 Data Mining/AA1/CustSales.csv"</div><div>Row Separator</div><div>"\n"</div><div>Field Separator</div><div>","</div><div>Header</div><div>0</div><div>Footer</div><div>0</div><div>Limit</div><div></div><div><input checked="" type="checkbox"/> Skip empty rows</div><div><input type="checkbox"/> Uncompress as zip file</div><div><input type="checkbox"/> Die on error</div></div></div></div>																																										



2. tMap1

The "tMap" component is a versatile transformation component that allows you to define mappings between input and output columns. It is commonly used for data transformations, lookups, and calculations within a Talend Job. The tMap component can have multiple input flows, and can apply various operations such as filtering, aggregating, and joining data in a visually intuitive way.

In this case study, those three datasets are being connected with Customer_id as it is the primary keys for all three datasets. Refer to image below:

CustDetails	
Column	
Customer_id	
Age	
Gender	
Location	
MembershipLevel	

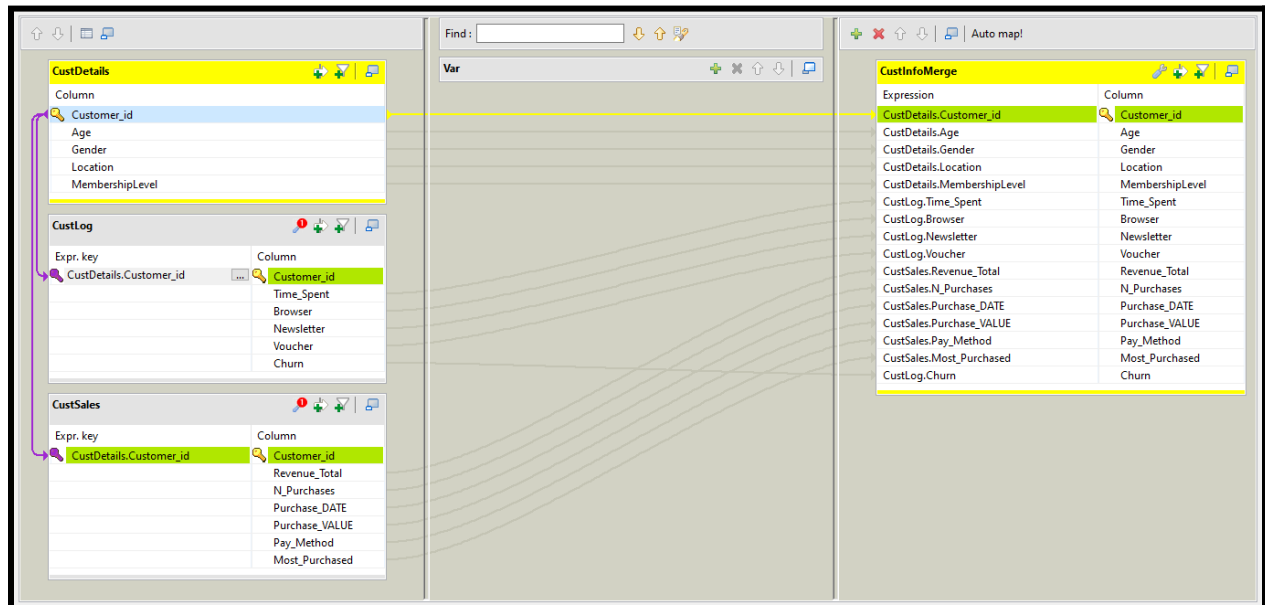
CustLog	
Expr. key	Column
CustDetails.Customer_id	Customer_id
	Time_Spent
	Browser
	Newsletter
	Voucher
	Churn

CustSales	
Expr. key	Column
CustDetails.Customer_id	Customer_id
	Revenue_Total
	N_Purchases
	Purchase_DATE
	Purchase_VALUE
	Pay_Method
	Most_Purchased

As result, one table is created with those variables, like image below:

CustInfoMerge	
Expression	Column
CustDetails.Customer_id	Customer_id
CustDetails.Age	Age
CustDetails.Gender	Gender
CustDetails.Location	Location
CustDetails.MembershipLevel	MembershipLevel
CustLog.Time_Spent	Time_Spent
CustLog.Browser	Browser
CustLog.Newsletter	Newsletter
CustLog.Voucher	Voucher
CustLog.Churn	Churn
CustSales.Revenue_Total	Revenue_Total
CustSales.N_Purchases	N_Purchases
CustSales.Purchase_DATE	Purchase_DATE
CustSales.Purchase_VALUE	Purchase_VALUE
CustSales.Pay_Method	Pay_Method
CustSales.Most_Purchased	Most_Purchased

This is how the connection looks like:



3. tFileOutputDelimited

In Talend Data Integration, tFileOutputDelimited is a component used for writing data to delimited text files. This component is part of the extensive set of tools provided by Talend for Extract, Transform, and Load (ETL) processes. It is primarily used for writing data to delimited text files, such as CSV (Comma-Separated Values) or other custom-delimited formats.

Based on the image below, it presents how the details in Component of one tFileOutputDelimited looks like where we can select the path and name of the CSV file that we want to export and save to local.

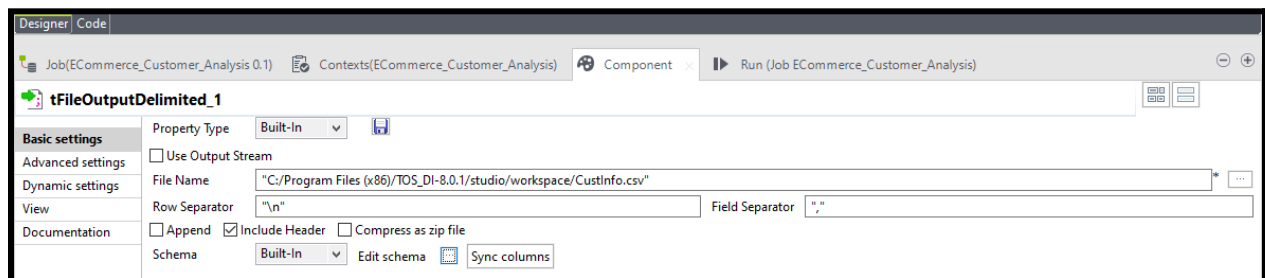
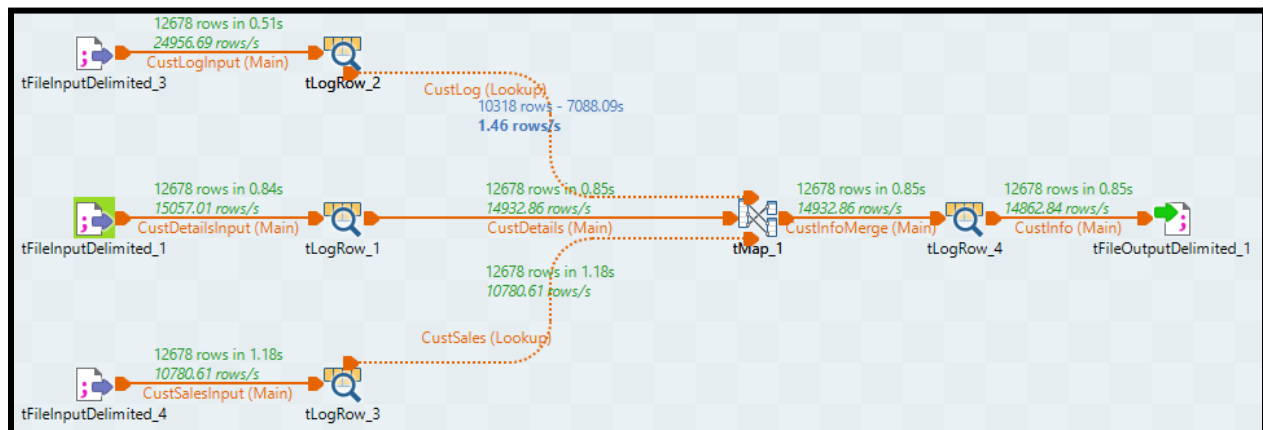


Image below shows overall view in Talend Data Integration:



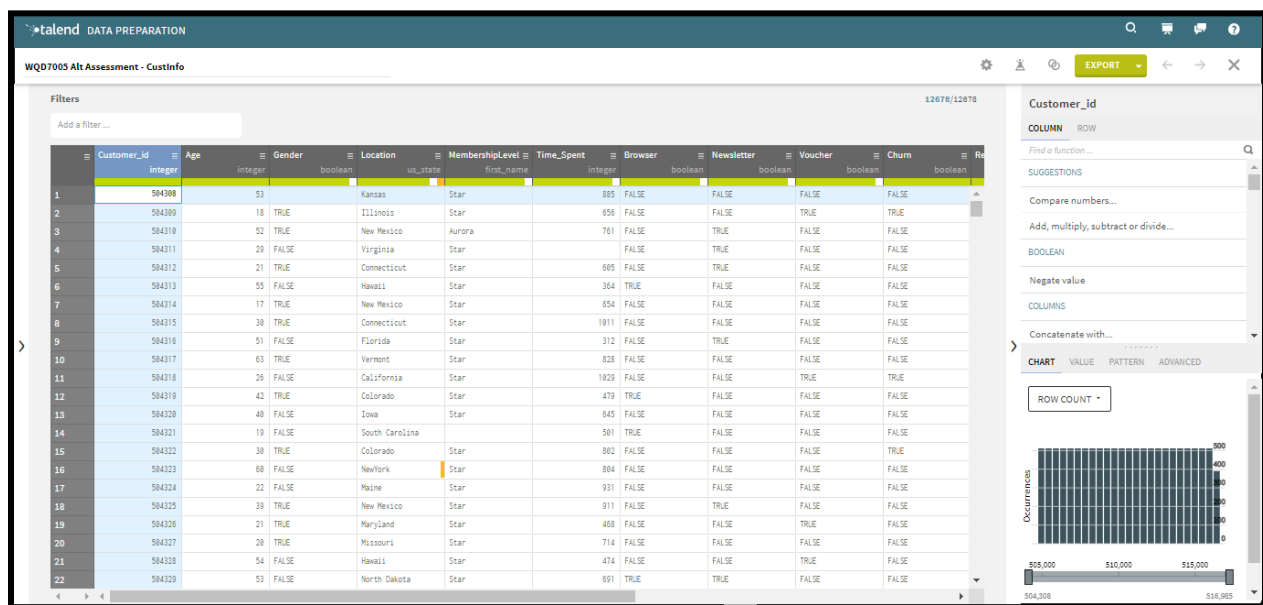
In this particular step, the "tFileOutputDelimited" component is applied by using CustDetails, CustLog and CustSales dataset, then go through "tMap" component to merge those three dataset and come out with one dataset as "customerinfo". This step involves configuring the component to write or transform data from the preceding stages and mapping it to the appropriate structure for the CSV file format. I define the file name, directory, and delimiter, ensuring that the output adheres to the desired specifications. The versatility of "tFileOutputDelimited" enables me to handle column mapping, include headers or footers, and manage various advanced settings to tailor the output file according to specific requirements. This step plays a crucial role in the ETL workflow, as it finalizes the process by persisting the transformed or processed data into the specified CSV file, ready for further analysis, reporting, or sharing. Image below whos the view of CustInfo dataset.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Customer	Age	Gender	Location	Members	Time_Spe	Browser	Newsletters	Voucher	Churn	Revenue	N_Purchase	Purchase	Purchase	Pay_Meth	Most_Purchased					
2	504308	53		Kansas	Star	885	FALSE	FALSE	FALSE	FALSE	45.3	2	22.06.0021	24.915		1	Fresh Fruits				
3	504309	18	TRUE	Illinois	Star	656	FALSE	FALSE	TRUE	TRUE	36.2	3	10.12.0021	2.896		2	Fresh Vegetables				
4	504310	52	TRUE	New Mexi	Aurora	761	FALSE	TRUE	FALSE	FALSE	10.6	1	14.03.0021	10.6		0	Dairy and Eggs				
5	504311	29	FALSE	Virginia	Star		FALSE	TRUE	FALSE	FALSE	54.1	5	25.10.0021	43.28		1	Meat and Poultry				
6	504312	21	TRUE	Connectic	Star	605	FALSE	TRUE	FALSE	FALSE	56.9	1	14.09.0021	56.9			Seafood				
7	504313	55	FALSE	Hawaii	Star	364	TRUE	FALSE	FALSE	FALSE		6	14.05.0021	12.467		1	Bread and Bakery				
8	504314	17	TRUE	New Mexi	Star	654	FALSE	FALSE	FALSE	FALSE	30.7	6	09.01.0021	2.456		0	Pantry Staples				
9	504315	30	TRUE	Connectic	Star	1011	FALSE	FALSE	FALSE	FALSE	8.1	7	28.03.0021	6.561		3	Beverages				
10	504316	51	FALSE	Florida	Star	312	FALSE	TRUE	FALSE	FALSE	18	4	04.08.0021	11.88		0	Snacks and Chips				
11	504317	63	TRUE	Vermont	Star	828	FALSE	FALSE	FALSE	FALSE	19.2	4	06.10.0021	11.904		3	Frozen Foods				
12	504318	26	FALSE	California	Star	1029	FALSE	FALSE	TRUE	TRUE	36.5	5	31.12.0021	31.39		2	Breakfast Foods				
13	504319	42	TRUE	Colorado	Star	479	TRUE	FALSE	FALSE	FALSE	14	4	22.11.0021	4.34		3	Canned Goods				
14	504320	40	FALSE	Iowa	Star	645	FALSE	FALSE	FALSE	FALSE	14.7	2	02.08.0021	2.94		3	Pasta and Grains				
15	504321	19	FALSE	South Carolina		501	TRUE	FALSE	FALSE	FALSE	37.4	4	07.05.0021	15.334		3	Cooking Oils and Condiments				
16	504322	30	TRUE	Colorado	Star	802	FALSE	FALSE	FALSE	TRUE	15.4	1	02.05.0021	15.4		3	Baking Ingredients				
17	504323	60	FALSE	New York	Star	804	FALSE	FALSE	FALSE	FALSE	28.7	7	04.06.0021	0.861		3	Spices and Seasonings				
18	504324	22	FALSE	Maine	Star	931	FALSE	FALSE	FALSE	FALSE	39.7	3	22.02.0021	30.569		2	International Foods				
19	504325	39	TRUE	New Mexi	Star	911	FALSE	TRUE	FALSE	FALSE	5.1	3	13.07.0021	1.53		3	Organic Products				
20	504326	21	TRUE	Maryland	Star	468	FALSE	FALSE	TRUE	FALSE	43.9	6	13.09.0021	19.755		1	Gluten-Free Products				
21	504327	20	TRUE	Missouri	Star	714	FALSE	FALSE	FALSE	FALSE	36.4	6	16.01.0021	14.56		2	Vegan and Plant-Based				
22	504328	54	FALSE	Hawaii	Star	474	FALSE	FALSE	TRUE	FALSE	23.2	1	03.07.0021	23.2		0	Ready-to-Eat Meals				
23	504329	53	FALSE	North Dak	Star	691	TRUE	TRUE	FALSE	FALSE	26.3	7	22.12.0021	15.517		3	Baby and Toddler				

Talend Data Preprocessing

The initial step in the data processing journey using Talend Data Preparation involves importing the raw data into the platform. This step is crucial as it sets the foundation for subsequent transformations and analyses. Users typically leverage Talend's intuitive interface to connect to various data sources, including databases, spreadsheets, or flat files. Upon establishing the data connection, the imported dataset is visually presented within the Talend Data Preparation environment, providing users with an overview of the raw data's structure and content. This step lays the groundwork for the subsequent stages of data cleaning, enrichment, and transformation, empowering users to seamlessly transition from raw data to a refined and prepared dataset for more advanced analytics and decision-making processes.

Here is the view after we upload the CSV file from Talend Data Integration to Talend Data Preprocessing:



To make the dataset become more valuable and precise, there are few steps need to be done for some field, which are:

1. Location

The screenshot displays the Talend Data Preparation interface. The main window shows a dataset with columns: Customer_id, Age, Gender, Location, MembershipLevel, Time_Spent, Browser, Newsletter, Voucher, Churn, and Revenue. A filter is applied to the 'Location' column, showing rows with invalid values. The right sidebar shows the 'Location' column details, including a 'FIND AND GROUP SIMILAR TEXT' dialog box. The dialog box is titled 'FIND AND GROUP SIMILAR TEXT' and contains a table with columns 'These values have been found' and 'This value will be kept'. The table lists several locations: Arizona, Arkansas, North Carolina, North Dakota, Washington, and Wisconsin. The 'Replace value' dropdown is set to 'Arizona'. The 'SUMMIT' button is visible at the bottom of the dialog box.

Customer_id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter	Voucher	Churn	Revenue
16	584223	68	FALSE	NewYork	Star	884	FALSE	FALSE	FALSE	FALSE
206	584513	18	FALSE	Arizona	Star	506	FALSE	FALSE	FALSE	FALSE

Filters: Location: rows with invalid values

Location

COLUMN ROW

Find a function...

SUGGESTIONS

- Delete these filtered rows
- Keep these filtered rows
- Delete the rows with invalid cell
- Fill invalid cells with value...
- Clear the cells with invalid values

Apply changes to: ☐ All rows ☒ Filtered rows

CHART VALUE PATTERN ADVANCED

Count: 12678 Avg length: 8.45

Distinct: 53

Duplicate: 12625 Min length: 0

Valid: 12675

Empty: 1 Max length: 14

Invalid: 2

FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

	These values have been found	This value will be kept
<input type="checkbox"/>	<input checked="" type="checkbox"/> Arizona	Replace value: Arizona
<input checked="" type="checkbox"/>	<input type="checkbox"/> Arkansas	
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Arizona	North Carolina
<input type="checkbox"/>	<input checked="" type="checkbox"/> North Carolina	
<input type="checkbox"/>	<input checked="" type="checkbox"/> North Dakota	Washington
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Washington	
<input type="checkbox"/>	<input checked="" type="checkbox"/> Wisconsin	

SUMMIT

FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

These values have been found	This value will be kept
<input checked="" type="checkbox"/> North Carolina	North Carolina
<input checked="" type="checkbox"/> North Dakota	
<input checked="" type="checkbox"/> Washington	Washington
<input checked="" type="checkbox"/> Wisconsin	
<input checked="" type="checkbox"/> New York	Replace value: New York
<input checked="" type="checkbox"/> New York	

SUBMIT

CustInfo PREPARATION

1 Find and group similar text on column Location

Location: rows with invalid values

	Customer_Id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter
	Integer	Integer	boolean	string	string	Integer	boolean	boolean
1	504300	53		Kansas	Star	885	FALSE	FALSE
2	504300	18	TRUE	Illinois	Star	656	FALSE	FALSE
3	504310	52	TRUE	New Mexico	Aurora	761	FALSE	TRUE
4	504311	20	FALSE	Virginia	Star		FALSE	TRUE
5	504312	21	TRUE	Connecticut	Star	605	FALSE	TRUE
6	504313	55	FALSE	Hawaii	Star	364	TRUE	FALSE
7	504314	17	TRUE	New Mexico	Star	654	FALSE	FALSE
8	504315	30	TRUE	Connecticut	Star	1011	FALSE	FALSE
9	504316	51	FALSE	Florida	Star	312	FALSE	TRUE
10	504317	63	TRUE	Vermont	Star	828	FALSE	FALSE
11	504318	26	FALSE	California	Star	1029	FALSE	FALSE
12	504319	42	TRUE	Colorado	Star	479	TRUE	FALSE
13	504320	40	FALSE	Iowa	Star	645	FALSE	FALSE
14	504321	19	FALSE	South Carolina		501	TRUE	FALSE
15	504322	30	TRUE	Colorado	Star	802	FALSE	FALSE
16	504323	60	FALSE	New York	Star	804	FALSE	FALSE
17	504324	22	FALSE	Maine	Star	931	FALSE	FALSE
18	504325	30	TRUE	New Mexico	Star	911	FALSE	TRUE
19	504326	21	TRUE	Maryland	Star	468	FALSE	FALSE
20	504327	20	TRUE	Missouri	Star	714	FALSE	FALSE
21	504328	54	FALSE	Hawaii	Star	474	FALSE	FALSE
22	504329	53	FALSE	North Dakota	Star	691	TRUE	TRUE

To rectify the problem of invalid location data, a systematic solution is applied by grouping and modifying the dataset within Talend. Initial identification of records with invalid location entries is followed by grouping the data based on relevant criteria, facilitating a focused approach. Talend's transformation components are then employed to cleanse and rectify the location data within each group. This involves operations such as standardizing formats, filling missing values, or applying geocoding techniques. Rigorous validation and testing are integral to ensuring the effectiveness of the cleansing process, and an iterative approach may be adopted for large or complex datasets. By documenting the changes made throughout this process, the dataset undergoes a structured transformation, resolving the invalid location issues and establishing a reliable foundation for subsequent data analyses.

2. Date

The image displays two screenshots of the Talend Data Preparation interface, illustrating the process of standardizing date data.

Top Screenshot: The interface shows a data table with columns: Browser, Newsletter, Voucher, Churn, Revenue_Total, N_Purchases, Purchase_DATE, Purchase_VALUE, Pay_Method, and Most_Purchased. The 'Purchase_DATE' column contains dates in various formats (e.g., 22/06/0021, 10/12/0021, 14/03/0021). The right sidebar shows the 'Purchase_DATE' column selected, with a 'ROW COUNT' chart.

Bottom Screenshot: The interface shows the 'Find and group similar text' and 'Change date format' steps applied to the 'Purchase_DATE' column. The 'Current format' is 'I don't know, best guess' and the 'New format' is 'French standard'. The 'Purchase_DATE' column now displays dates in the 'French standard' format (e.g., 22/06/21, 10/12/21, 14/03/21). The right sidebar shows the 'Purchase_DATE' column selected, with a 'ROW COUNT' chart.

Following the resolution of invalid location data, the subsequent step involves standardizing the format of the date column within the dataset using Talend. This crucial data preparation step ensures consistency and coherence in date representations. Utilizing these components, users can define the desired date format, convert data types if necessary, and handle any anomalies in the date column. This process guarantees uniformity, making it easier for downstream analyses, reporting, and integration with other systems.

3. Most Purchased

talend DATA PREPARATION

CustInfo PREPARATION

1 Find and group similar text on column Location

2 Change date format on column Purchase_DATE

Filters

Add a filter...

id	Churn	Revenue_Total	N_Purchases	Purchase_DATE	Purchase_VALUE	Pay_Method	Most_Purchased
id	boolean	decimal	integer	date	decimal	integer	text
1	FALSE	45.3	2	22/06/21	24.915	1	Fresh Fruits
2	TRUE	36.2	3	10/12/21	2.896	2	Fresh Vegetables
3	FALSE	18.6	1	14/03/21	18.6	0	Dairy and Eggs
4	FALSE	54.1	5	25/10/21	43.28	1	Meat and Poultry
5	FALSE	56.9	1	14/09/21	56.9	0	Seafood
6	FALSE		6	14/05/21	12.487	1	Bread and Bakery
7	FALSE	30.7	6	09/01/21	2.456	0	Pantry Staples
8	FALSE	8.1	7	28/03/21	6.561	3	Beverages
9	FALSE	18	4	04/08/21	11.88	0	Snacks and Chips
10	FALSE	19.2	4	06/10/21	11.984	3	Frozen Foods
11	TRUE	36.5	5	31/12/21	31.39	2	Breakfast Foods
12	FALSE	14	4	22/11/21	4.34	0	Canned Goods
13	FALSE	14.7	2	02/08/21	2.94	3	Pasta and Grains
14	FALSE	37.4	4	07/05/21	15.334	3	Cooking Oils and Com
15	TRUE	15.4	1	02/05/21	15.4	3	Baking Ingredients
16	FALSE	28.7	7	04/06/21	9.861	3	Spices and Seasoning
17	FALSE	39.7	3	22/02/21	38.569	2	International Foods
18	FALSE	5.1	3	13/07/21	1.53	3	Organic Products
19	FALSE	43.9	6	13/09/21	19.795	1	Gluten-Free Products
20	FALSE	36.4	6	16/01/21	14.56	2	Vegan and Plant-Based
21	FALSE	23.2	1	03/07/21	23.2	0	Ready-to-Eat Meals
22	FALSE	26.3	7	22/12/21	15.517	3	Baby and Toddler

Most_Purchased

COLUMN ROW

grou

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

ROW COUNT

0 20 40 60 80 100 120

Fresh Fruits

Pasta and Grains

Frozen Foods

Meat and Poultry

Bread and Bakery

Snacks and Chips

Breakfast Foods

Seafood

talend DATA PREPARATION

CustInfo PREPARATION

1 Find and group similar text on column Location

2 Change date format on column Purchase_DATE

FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

	These values have been found	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Baby and Toddler Toys <input checked="" type="checkbox"/> Baby and Toddler	Replace value: Baby and Toddler
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Party Games and Supplies <input checked="" type="checkbox"/> Party Supplies	Replace value: Party Games and Supplies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Audio and Headphones <input checked="" type="checkbox"/> Automotive	Audio and Headphones
<input type="checkbox"/>	<input checked="" type="checkbox"/> Personal Care	

SUBMIT

Most_Purchased

COLUMN ROW

grou

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

Count: 12678

Distinct: 111

Duplicate: 12567

Valid: 12678

Empty: 0

Invalid: 0

Avg length: 18.43

Min length: 6

Max length: 32

FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

	These values have been found	This value will be kept
<input type="checkbox"/>	<input checked="" type="checkbox"/> Snacks and Chips	Snacks and Chips
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Cleaning Tools and Supplies <input checked="" type="checkbox"/> Cleaning Supplies	Replace value: Cleaning Supplies
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Electronics Accessories <input checked="" type="checkbox"/> Electronics and Appliances	Replace value: Electronics Accessories
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Baby Gear and Nursery <input checked="" type="checkbox"/> Baby Care Products	Replace value: Baby Care Products

SUBMIT

Most_Purchased

COLUMN ROW

group

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

Count: 12678

Distinct: 111

Duplicate: 12567

Valid: 12678

Empty: 0

Invalid: 0

Avg length: 18.43

Min length: 6

Max length: 32

FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

	These values have been found	This value will be kept
<input type="checkbox"/>	<input checked="" type="checkbox"/> Baby Care Products	
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> IndividualSports <input checked="" type="checkbox"/> Individual Sports	Replace value: Individual Sports
<input type="checkbox"/>	<input checked="" type="checkbox"/> Health and Wellness <input checked="" type="checkbox"/> Health Monitors	Health Monitors
<input type="checkbox"/>	<input checked="" type="checkbox"/> Fresh Vegetables <input checked="" type="checkbox"/> Fresh Fruits	Fresh Fruits

SUBMIT

Most_Purchased

COLUMN ROW

group

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

Count: 12678

Distinct: 111

Duplicate: 12567

Valid: 12678

Empty: 0

Invalid: 0

Avg length: 18.43

Min length: 6

Max length: 32

The subsequent step involves grouping and modifying the data within the "most_purchased" column to ensure uniformity in format throughout the dataset using Talend. This process addresses any inconsistencies, variations, or irregularities in the "most_purchased" column, fostering standardized representations for ease of analysis and interpretation. Whether it involves categorization, renaming, or standardizing values, the objective is to bring cohesion to the "most_purchased" column. This harmonization facilitates a more seamless integration of the dataset for downstream processes, such as reporting or machine learning, where consistency in data formats is paramount. Through careful grouping and data modification, Talend ensures that the "most_purchased" column adheres to a standardized format, enhancing the overall quality and coherence of the dataset.

4. Gender

The screenshot displays the Talend Data Preparation interface for a job named 'CustInfo PREPARATION'. The interface is divided into several sections:

- Left Panel (Job Design):** Contains three steps:
 - Find and group similar text on column Location
 - Change date format on column Purchase_DATE
 - Find and group similar text on column Most_Purchased
- Filters:** A section for adding filters to the data stream.
- Table View:** A grid showing 22 rows of data. The columns are: Customer_id, Age, Gender, Location, MembershipLevel, Time_Spent, Browser, and Newsletter. The 'Gender' column is highlighted in blue, indicating it is the active column for the current transformation.
- Right Panel (Transformation Configuration):** Titled 'Gender', it shows the configuration for the transformation. The 'COLUMN' tab is selected, displaying a list of columns. The 'VALUE' tab is also visible, showing a row count of 12678. Below the row count, there are statistics for the 'Gender' column: Count: 12678, Distinct: 3, Duplicate: 12675, Valid: 12677, Empty: 1, Invalid: 0.

The 'Gender' transformation is configured to replace the values in the 'Gender' column. The 'Use with:' dropdown is set to 'Value'. The 'Value:' field contains 'Men'. A 'SUBMIT' button is located at the bottom of the configuration panel.

To uniformly represent gender in the dataset, Talend's tMap component is utilized. A new output column, "Gender," is created using a ternary expression, assigning "Men" for true and "Women" for false in the specified boolean column. Executing the Talend job applies this transformation across all rows, ensuring consistent gender values by replacing "true" with "Men" and "false" with "Women" in the designated column.

5. Fill empty cells

talend DATA PREPARATION

Custinfo PREPARATION

Most_Purchased

4 Replace the cells that match on column Gender

5 Replace the cells that match on column Gender

6 Fill empty cells with text on column Gender

7 Fill empty cells with text on column Location

8 Fill empty cells with text on column MembershipLevel

9 Fill empty cells with text on column Browser

10 Fill empty cells with text on column Newsletter

11 Fill empty cells with text on column Pay_Method

12 Fill empty cells with text on column Voucher

Filters

Add a filter ...

	Customer_id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter
	integer	integer	gender	us_state	first_name	integer	boolean	boolean
1	504308	53	Men	Kansas	Star	885	FALSE	FALSE
2	504309	18	Men	Illinois	Star	656	FALSE	FALSE
3	504310	52	Men	New Mexico	Aurora	761	FALSE	TRUE
4	504311	29	Woman	Virginia	Star		FALSE	TRUE
5	504312	21	Men	Connecticut	Star	685	FALSE	TRUE
6	504313	55	Woman	Hawaii	Star	364	TRUE	FALSE
7	504314	17	Men	New Mexico	Star	654	FALSE	FALSE
8	504315	38	Men	Connecticut	Star	1011	FALSE	FALSE
9	504316	51	Woman	Florida	Star	312	FALSE	TRUE
10	504317	63	Men	Vermont	Star	828	FALSE	FALSE
11	504318	26	Woman	California	Star	1029	FALSE	FALSE
12	504319	42	Men	Colorado	Star	479	TRUE	FALSE
13	504320	40	Woman	Iowa	Star	645	FALSE	FALSE
14	504321	19	Woman	South Carolina	Star	501	TRUE	FALSE
15	504322	38	Men	Colorado	Star	802	FALSE	FALSE
16	504323	60	Woman	New York	Star	804	FALSE	FALSE
17	504324	22	Woman	Maine	Star	931	FALSE	FALSE
18	504325	39	Men	New Mexico	Star	911	FALSE	TRUE
19	504326	21	Men	Maryland	Star	468	FALSE	FALSE
20	504327	20	Men	Missouri	Star	714	FALSE	FALSE
21	504328	54	Woman	Hawaii	Star	474	FALSE	FALSE
22	504329	53	Woman	North Dakota	Star	691	TRUE	TRUE

Pay_Method

COLUMN ROW

Find a function ...

Negate value

COLUMNS

Concatenate with...

Delete column

Swap columns...

CONVERSIONS

Convert distance...

CHART VALUE PATTERN ADVANCED

ROW COUNT

Occurrences

talend DATA PREPARATION

Custinfo PREPARATION

Filters

Add a filter ...

	Customer_id	Age	Gender	Location	MembershipLevel	Time_Spent	Browser	Newsletter	Voucher	Churn	Revenue_Total	IL_Purchases	Purchase_DATE	Purchase_VALUE	Pay_Method	Most_Purchased
	integer	integer	gender	us_state	first_name	integer	boolean	boolean	boolean	boolean	decimal	integer	date	decimal	integer	text
1	504308	53	Men	Kansas	Star	885	FALSE	FALSE	FALSE	FALSE	45.3	2	23/06/21	24.915	1	Fresh Fruits
2	504309	18	Men	Illinois	Star	656	FALSE	TRUE	TRUE	TRUE	36.2	3	18/12/21	2.850	2	Fresh Vegetables
3	504310	52	Men	New Mexico	Aurora	761	FALSE	TRUE	FALSE	FALSE	18.6	1	14/03/21	18.6	0	Dairy and Eggs
4	504311	29	Woman	Virginia	Star		FALSE	TRUE	FALSE	FALSE	54.1	5	25/10/21	43.28	1	Meat and Poultry
5	504312	21	Men	Connecticut	Star	685	FALSE	TRUE	FALSE	FALSE	56.9	1	14/09/21	56.9	1	Seafood
6	504313	55	Woman	Hawaii	Star	364	TRUE	FALSE	FALSE	FALSE		6	14/05/21	12.487	1	Bread and Baker
7	504314	17	Men	New Mexico	Star	654	FALSE	FALSE	FALSE	FALSE	38.7	6	09/01/21	2.456	0	Pantry Staples
8	504315	38	Men	Connecticut	Star	1011	FALSE	FALSE	FALSE	FALSE	6.1	7	28/03/21	6.581	3	Beverages
9	504316	51	Woman	Florida	Star	312	FALSE	TRUE	FALSE	FALSE	18	4	04/08/21	11.88	0	Snacks and Chips
10	504317	63	Men	Vermont	Star	828	FALSE	FALSE	FALSE	FALSE	19.2	4	00/10/21	11.904	3	Frozen Foods
11	504318	26	Woman	California	Star	1029	FALSE	FALSE	TRUE	TRUE	36.5	5	31/12/21	31.39	2	Breakfast Foods
12	504319	42	Men	Colorado	Star	479	TRUE	FALSE	FALSE	FALSE		4	22/11/21	4.34	3	Canned Goods
13	504320	40	Woman	Iowa	Star	645	FALSE	FALSE	FALSE	FALSE	14.7	2	02/08/21	2.54	3	Pasta and Grains
14	504321	19	Woman	South Carolina	Star	501	TRUE	FALSE	FALSE	FALSE	37.4	4	07/09/21	15.914	3	Cooking Oils and Vinegar
15	504322	38	Men	Colorado	Star	802	FALSE	FALSE	FALSE	FALSE	15.4	1	02/09/21	15.4	3	Baking Ingredients
16	504323	60	Woman	New York	Star	804	FALSE	FALSE	FALSE	FALSE	28.7	7	04/06/21	6.807	3	Grocery and Staples
17	504324	22	Woman	Maine	Star	931	FALSE	FALSE	FALSE	FALSE	39.7	3	22/02/21	30.969	2	International Flavors
18	504325	39	Men	New Mexico	Star	911	FALSE	TRUE	FALSE	FALSE	5.1	3	13/01/21	1.53	3	Organic Products
19	504326	21	Men	Maryland	Star	468	FALSE	FALSE	FALSE	FALSE	40.9	6	13/09/21	15.755	1	Gluten-Free Flours
20	504327	20	Men	Missouri	Star	714	FALSE	FALSE	FALSE	FALSE	36.4	6	16/01/21	14.56	2	Vegetables and Plantains
21	504328	54	Woman	Hawaii	Star	474	FALSE	FALSE	TRUE	FALSE	23.2	1	03/07/21	23.2	0	Ready-to-Eat Meals
22	504329	53	Woman	North Dakota	Star	691	TRUE	TRUE	FALSE	FALSE	26.3	7	22/12/21	15.517	3	Baby and Toddler
23	504330	22	Men	Ohio	Star	394	FALSE	FALSE	TRUE	FALSE	33.1	2	10/01/21	6.652	0	Pet Supplies
24	504331	32	Men	Nebraska	Star	1013	TRUE	FALSE	TRUE	TRUE	1.2	4	28/01/21	6.204	0	Cleaning Supplies
25	504332	36	Men	Montana	Star	105	FALSE	FALSE	FALSE	FALSE	41.3	6	16/04/21	24.387	1	Personal Care
26	504333	24	Men	Indiana	Star	710	FALSE	FALSE	FALSE	FALSE	25.5	3	13/05/21	4.88	0	Health and Wellness
27	504334	57	Men	Wisconsin	Star	604	FALSE	FALSE	TRUE	FALSE	34.1	5	07/09/21	18.873	3	Vitamins and Supplements
28	504335	27	Men	Alabama	Star	688	TRUE	FALSE	FALSE	FALSE	38.8	4	21/01/21	26.18	2	Home and Kitchen
29	504336	21	Woman	Alabama	Star	992	FALSE	FALSE	FALSE	FALSE	31.1	2	11/01/21	9.815	0	Office and School
30	504337	52	Men	Indiana	Star	311	FALSE	FALSE	TRUE	TRUE	48.1	6	18/10/21	34.151	3	Electronics Accessories
31	504338	52	Men	Nebraska	Star	283	TRUE	TRUE	FALSE	FALSE	42.7	2	04/02/21	22.204	1	Outdoor and Garden
32	504339	52	Men	Montana	Star	941	FALSE	FALSE	FALSE	FALSE	53.5	4	13/12/21	14.36	0	Home Decor
33	504340	55	Woman	Arkansas	Star		FALSE	FALSE	FALSE	FALSE	39	2	03/02/21	23.79	1	Clothing and Accessories
34	504341	53	Men	Missouri	Star	469	FALSE	FALSE	FALSE	FALSE	38.3	5	16/08/21	5.454	0	Shoes and Footwear
35	504342	17	Woman	Pennsylvania	Star	338	TRUE	FALSE	FALSE	FALSE	34.7	1	17/09/21	34.7	1	Beauty and Skincare
36	504343	40	Men	New Hampshire	Star	723	FALSE	FALSE	FALSE	FALSE	41.3	6	17/10/21	9.685	1	Hair Care

SAS Enterprise Miner

Enterprise Miner - WQD7005AltAsmnt1

File Edit View Actions Options Window Help

WQD7005AltAsmnt1

Data Sources

Diagrams

CaseStudy

Model Packages

Property Value

General

Node ID FIMPORT

Imported Data

Exported Data

Notes

Train

Variables

Import File

Maximum Rows to Import 1000000

Maximum Columns to Import 10000

Delimiter

Name Row Yes

Number of Rows to Skip 0

Guessing Rows 500

File Location Local

File Type csv

Advanced Advisor No

Return No

Score

Role Train

Report

Summarize No

Status

Create Time 1/7/24 9:28 AM

Run ID

General

Sample Explore Modify Model Assess Utility Credit Scoring HPDM Applications Text Mining Time Series

CaseStudy

File Import

Select My Computer if the data file you want to import is located on your local machine. Select SAS Server to import a data file located on your SAS workspace server.

My Computer

SAS Servers

D:\Download D\CustInfo PREPARATION.csv

Browse...

View File Import Types Preview OK Cancel

Diagram Log

Diagram CaseStudy opened

wie190006@siswa.um.edu.my as u62569858 Connected to SASApp - Logical Workspace Server (pdaws01-apse1.oda.sas.com)

Results - Node: File Import Diagram: CaseStudy

File Edit View Window

Output

40

41 Data Set Page Size 131072

42 Number of Data Set Pages 16

43 First Data Page 1

44 Max Obs per Page 818

45 Obs in First Data Page 792

46 Number of Data Set Repairs 0

47 Filename /home/u62569858/WQD7005AltAsmnt1/Workspaces/ENWS1/fimport_data.sas7bdat

48 Release Created 9.0401M7

49 Host Created Linux

50 Inode Number 224789474

51 Access Permission rw-r--r--

52 Owner Name u62569858

53 File Size 2MB

54 File Size (bytes) 2282224

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

98

99

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

161

162

163

164

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214

215

216

217

218

219

220

221

222

223

224

225

226

227

228

229

230

231

232

233

234

235

236

237

238

239

240

241

242

243

244

245

246

247

248

249

250

251

252

253

254

255

256

257

258

259

260

261

262

263

264

265

266

267

268

269

270

271

272

273

274

275

276

277

278

279

280

281

282

283

284

285

286

287

288

289

290

291

292

293

294

295

296

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

312

313

314

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

571

572

573

574

575

576

577

578

579

580

581

582

583

584

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

660

661

662

663

664

665

666

667

668

669

670

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702

703

704

705

706

707

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756

757

758

759

760

761

762

763

764

765

766

767

768

769

770

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863

864

865

866

867

868

869

870

871

872

873

874

875

876

877

878

879

880

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918

919

920

921

922

923

924

925

926

927

928

929

930

931

932

933

934

935

936

937

938

939

940

941

942

943

944

945

946

947

948

949

950

951

952

953

954

955

956

957

958

959

960

961

962

963

964

965

966

967

968

969

970

971

972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

