



**SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



DEPARTMENT OF INFORMATION TECHNOLOGY

COURSE CODE: DJ19ITL503

DATE: 20-08-24

COURSE NAME: Data Warehousing and Mining

CLASS: I1-Batch-1

NAME: Anish Sharma

ROLL NO.: I011

LAB EXPERIMENT NO. 3

CO/LO: Apply ETL steps for a given dataset

AIM / OBJECTIVE: Executing ETL operations on Talend Tool

DESCRIPTION OF EXPERIMENT:

ETL (Extract, Transform, Load) is used to integrate data from various sources into a unified format for analysis. It extracts data, transforms it by cleansing and enriching, and loads it into data warehouses or other systems, enabling efficient reporting, business intelligence, and improved data quality.

ETL processes work by first extracting data from multiple sources, then transforming it through cleansing, aggregation, and formatting to ensure consistency and accuracy. Finally, the transformed data is loaded into a data warehouse or database. This sequence enables efficient data integration, analysis, and reporting for informed decision-making.

INPUT DATA / DATASET:

[customers.xlsx](#) file.

The dataset provided contains information about individuals and their associated details, presumably from a customer or client database. Here's a brief description of each column and what it represents:

1. **Id:** A unique identifier for each individual in the dataset.



**SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



2. **First_Name:** The individual's first name.
3. **Last_Name:** The individual's last name.
4. **Gender:** The gender of the individual (e.g., Male, Female).
5. **Age:** The age group of the individual, categorized into ranges such as "Under 18" or "25-34."
6. **Occupation:** The individual's job title or occupation.
7. **MaritalStatus_Out:** The individual's marital status (e.g., Single, Married, Divorced).
8. **Salary_Out:** The individual's salary range, given in financial brackets (e.g., "< 50,000" or "100,000-149,999").
9. **Address:** The street address of the individual.
10. **City:** The city where the individual resides.
11. **State:** The state or region where the individual resides.
12. **Zip:** The postal code for the individual's address.
13. **Phone:** The contact phone number of the individual.
14. **Email:** The email address of the individual.
15. **SubDate:** The subscription or registration date in DD-MMM-YYYY format.
16. **Number_of_rentals:** The total number of rentals or transactions associated with the individual.

PROCEDURE / ALGORITHM:

1. Open Talend Open Studio
2. Select data file for extraction process.
3. Perform transformation operations on the extracted file

TECHNOLOGY STACK USED:

Talend Open Studio is an open-source software suite for data integration, data quality, and data management. It provides a user-friendly graphical interface for designing ETL (Extract, Transform, Load) processes, allowing you to connect, transform, and manage data from various sources. Key features include:

- Graphical design interface
- Wide range of data source connectivity
- Data quality tools
- Extensible and open-source

It's used for tasks like data warehousing, ETL processes, and data migration.



**SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



OBSERVATIONS / DISCUSSION OF RESULT:

talend | Data Preparation

customers Preparation
customers

1 Change to title case on column
First_Name

2 Change to title case on column
Last_Name

Create new column

Preview Submit

Filters

Find in a column... Add filter

	Id Integer	First_Name First Name (text)	Last_Name Text	Gender Gender (text)	Age Text	Occupation Text	MaritalStatus_Ou Text
1	1	James	Butt	F	Under 18	K-12 Student	Single
2	2	Josephine	Darakjy	M	56+	Self-Employed	Married
3	3	Art	Venera	M	25-34	Scientist	Married
4	4	Lenna	Paprocki	M	45-49	Executive/Manager...	Divorced
5	5	Donette	Foller	M	25-34	Writer	
6	6	Simona	Morasca	F	50-55	Homemaker	Married
7	7	Mitsue	Tollner	M	35-44	Academic/Educator	Divorced
8	8	Leota	Dilliard	M	25-34	Programmer	
9	9	Sage	Wieser	M	25-34	Technical/Engineer	Divorced
10	10	Kris	Harrier	F	35-44	Academic/Educator	Divorced
11	11	Minna	Amigon	F	25-34	Academic/Educator	Divorced
12	12	Abel	Maclead	M	25-34	Programmer	Divorced
13	13	Kiley	Caldarera	M	45-49	Academic/Educator	
14	14	Graciela	Ruta	M	35-44	Other	Divorced
15	15	Cammy	Albares	M	25-34	Executive/Manager...	
16	16	Muttie	Poquette	F	35-44	Other	
17	17	Moaghan	Ganufi	M	50-55	Academic/Educator	Divorced

6040/6040

2 columns selected

Column Row Table

Filter

BOOLEAN

Negate value ...

COLUMNS

Concatenate columns ...

Delete column

Swap columns ...

CONVERSIONS

Chart Value Pattern Advanced

No chart for the current selection

Run



**SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



talend | Data Preparation | Talend Experience | Help | dhruv Nariani

customers Preparation

1 Change to title case on column First_Name

2 Change to title case on column Last_Name

3 Remove trailing and leading characters on column First_Name

4 Remove trailing and leading characters on column Last_Name

Create new column

Padding character

Whitespace

Preview

Submit

Filters

Find in a column...

Add filter

	Id	First_Name	Last_Name	Gender	Age	Occupation	MaritalStatus_Out
	Integer	First Name (text)	Text	Gender (text)	Text	Text	Text
1	1	James	Butt	F	Under 18	K-12 Student	Single
2	2	Josephine	Darakjy	M	56+	Self-Employed	Married
3	3	Art	Venerre	M	25-34	Scientist	Married
4	4	Lenna	Paprocki	M	45-49	Executive/Manager...	Divorced
5	5	Donette	Foller	M	25-34	Writer	
6	6	Simona	Morasca	F	58-55	Homemaker	Married
7	7	Mitsue	Tollner	M	35-44	Academic/Educator	Divorced
8	8	Leota	Dilliard	M	25-34	Programmer	
9	9	Sage	Wieser	M	25-34	Technical/Engineer	Divorced
10	10	Kris	Marrier	F	35-44	Academic/Educator	Divorced
11	11	Minna	Anigon	F	25-34	Academic/Educator	Divorced
12	12	Abel	MacLead	M	25-34	Programmer	Divorced
13	13	Kiley	Calderera	M	45-49	Academic/Educator	
14	14	Graciela	Ruta	M	35-44	Other	Divorced
15	15	Camy	Albares	M	25-34	Executive/Manager...	
16	16	Mattie	Poquette	F	35-44	Other	
17	17	Neaghan	Garufi	M	58-55	Academic/Educator	Divorced

6040/6040

2 columns selected

Column Row Table

Q Filter

BOOLEAN

Negate value ...

COLUMNS

Concatenate columns ...

Delete column

Swap columns ...

CONVERSIONS

Chart Value Pattern Advanced

No chart for the current selection

talend | Data Preparation | Talend Experience | Help | dhruv Nariani

customers Preparation

1 Change to title case on column First_Name

2 Change to title case on column Last_Name

3 Remove trailing and leading characters on column First_Name

4 Remove trailing and leading characters on column Last_Name

5 Search and replace on column Gender

6 Search and replace on column Gender

Create new column

Search for*

Replace with

Female

Overwrite entire cell

Preview

Submit

Filters

Find in a column...

Add filter

	Id	First_Name	Last_Name	Gender	Age	Occupation	MaritalStatus_Out
	Integer	First Name (text)	Text	Gender (text)	Text	Text	Text
1	1	James	Butt	Female	Under 18	K-12 Student	Single
2	2	Josephine	Darakjy	Male	56+	Self-Employed	Married
3	3	Art	Venerre	Male	25-34	Scientist	Married
4	4	Lenna	Paprocki	Male	45-49	Executive/Manager...	Divorced
5	5	Donette	Foller	Male	25-34	Writer	
6	6	Simona	Morasca	Female	58-55	Homemaker	Married
7	7	Mitsue	Tollner	Male	35-44	Academic/Educator	Divorced
8	8	Leota	Dilliard	Male	25-34	Programmer	
9	9	Sage	Wieser	Male	25-34	Technical/Engineer	Divorced
10	10	Kris	Marrier	Female	35-44	Academic/Educator	Divorced
11	11	Minna	Anigon	Female	25-34	Academic/Educator	Divorced
12	12	Abel	MacLead	Male	25-34	Programmer	Divorced
13	13	Kiley	Calderera	Male	45-49	Academic/Educator	
14	14	Graciela	Ruta	Male	35-44	Other	Divorced
15	15	Camy	Albares	Male	25-34	Executive/Manager...	
16	16	Mattie	Poquette	Female	35-44	Other	
17	17	Neaghan	Garufi	Male	58-55	Academic/Educator	Divorced

6040/6040

Gender

Column Row Table

Q Filter

SUGGESTIONS

Magic fill ...

Search and replace ...

Change to lower case ...

Change to upper case ...

BOOLEAN

Negate value ...

Chart Value Pattern Advanced

Row count*

Male

Female



**SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



talend | Data Preparation

customers Preparation

1 Change to title case on column First_Name

2 Change to title case on column Last_Name

3 Remove trailing and leading characters on column First_Name

4 Remove trailing and leading characters on column Last_Name

5 Search and replace on column Gender

6 Search and replace on column Gender

7 Change date format on column SubDate

Create new column

Current format*
I don't know, best guess

New format*
American standard with time

Preview Submit

Filters

Find in a column...

Add filter

		State US State Code (text)	Zip US Postal Code (text)	Phone US Phone (text)	Email Email (text)	SubDate Date	Number_of_ren... Integer
1	jeans	LA	78116	504-621-8927	jbuttl@gmail.com	3/17/2016 12:00 AM	541
2	in	MI	48116	810-292-9388	josephine_darakjy@	3/15/2013 12:00 AM	994
3	ort	NJ	8814	856-636-	art@venere	10/28/2007 12:00 AM	3
4	age	AK	99581	907-385-4412	lpa@rocki@hotmail...	11/24/2013 12:00 AM	586
5	in	OH	45811	513-578-1893	donette_foller@cox...	4/17/2012 12:00 AM	82
6	i	OH	44885	419-583-2484	simon@morasca.com	4/13/2016 12:00 AM	959
7		IL	60632	773-57	mitsue_tollner@yah...	6/7/2009 12:00 AM	323
8	ie	CA	95111	408	leota@hotmail.com	12/4/2008 12:00 AM	418
9	alls	SD	57185	605-414-2147	sage_wieser@cox.net	4/28/2013 12:00 AM	262
10	re	MD	21224	410-655-8723	kris@gmail.com	12/31/2012 12:00 AM	993
11	lie	PA	19443	215-8	mima_wigonyahoo...	7/21/2011 12:00 AM	377
12	Island	NY	11953	631-3	amaclead@gmail.com	9/3/2015 12:00 AM	567
13	jeles	CA	98034	310-498-5651	kiley_caldarera@ao...	3/8/2014 12:00 AM	493
14	y Falls	OH	44823	440-780-8425	gruta@cox.net	6/12/2013 12:00 AM	885
15		Texas	78845	956-537-6195	calbures@gmail.com	9/25/2011 12:00 AM	827
16		AZ	85813	602-277-4385	mattie@aol.com	12/1/2009 12:00 AM	121
17	iville	TN	37118	931-313-9635	meaghan@hotmail.c...	12/26/2015 12:00 AM	373

SubDate

Columns Row Table

Q Filter

SUGGESTIONS

Magic fill ...

Extract date parts ...

Change date format ...

Calculate time since ...

BOOLEAN

Negate value ...

Chart Value Pattern Advanced

Row count *

2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

Min 2007-01-01 Max 2017-01-01

talend | Data Preparation

customers Preparation

1 Change to title case on column First_Name

2 Change to title case on column Last_Name

3 Remove trailing and leading characters on column First_Name

4 Remove trailing and leading characters on column Last_Name

5 Search and replace on column Gender

6 Search and replace on column Gender

7 Change date format on column SubDate

8 Change semantic domain on column Zip

Filters

Find in a column...

Add filter

		State US State Code (text)	Zip US Postal Code (text)	Phone US Phone (text)	Email Email (text)	SubDate Date	Number_of_ren... Integer
1	jeans	LA	78116	504-621-8927	jbuttl@gmail.com	3/17/2016 12:00 AM	541
2	in	MI	48116	810-292-9388	josephine_darakjy@	3/15/2013 12:00 AM	994
3	ort	NJ	8814	856-636-	art@venere	10/28/2007 12:00 AM	3
4	age	AK	99581	907-385-4412	lpa@rocki@hotmail...	11/24/2013 12:00 AM	586
5	in	OH	45811	513-578-1893	donette_foller@cox...	4/17/2012 12:00 AM	82
6	i	OH	44885	419-583-2484	simon@morasca.com	4/13/2016 12:00 AM	959
7		IL	60632	773-57	mitsue_tollner@yah...	6/7/2009 12:00 AM	323
8	ie	CA	95111	408	leota@hotmail.com	12/4/2008 12:00 AM	418
9	alls	SD	57185	605-414-2147	sage_wieser@cox.net	4/28/2013 12:00 AM	262
10	re	MD	21224	410-655-8723	kris@gmail.com	12/31/2012 12:00 AM	993
11	lie	PA	19443	215-8	mima_wigonyahoo...	7/21/2011 12:00 AM	377
12	Island	NY	11953	631-3	amaclead@gmail.com	9/3/2015 12:00 AM	567
13	jeles	CA	98034	310-498-5651	kiley_caldarera@ao...	3/8/2014 12:00 AM	493
14	y Falls	OH	44823	440-780-8425	gruta@cox.net	6/12/2013 12:00 AM	885
15		Texas	78845	956-537-6195	calbures@gmail.com	9/25/2011 12:00 AM	827
16		AZ	85813	602-277-4385	mattie@aol.com	12/1/2009 12:00 AM	121
17	iville	TN	37118	931-313-9635	meaghan@hotmail.c...	12/26/2015 12:00 AM	373

Zip

Columns Row Table

Q Filter

SUGGESTIONS

Delete the rows with invalid cells

Fill invalid cells with value ...

Clear the cells with invalid values

Standardize value (fuzzy matching) ...

Magic fill ...

BOOLEAN

Chart Value Pattern Advanced

Row count *

2007 2008 2009 2010 2011 2012 2013 2014 2015 2016

Min 2007-01-01 Max 2017-01-01



**SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



talend | Data Preparation

customers Preparation

1 Change to title case on column First_Name

2 Change to title case on column Last_Name

3 Remove trailing and leading characters on column First_Name

4 Remove trailing and leading characters on column Last_Name

5 Search and replace on column Gender

6 Search and replace on column Gender

7 Change date format on column SubDate

8 Change semantic domain on column Zip

9 Delete the rows with empty cells on column Email

10 Delete the rows with invalid cells on column Email

Filters

Find in a column...

Add filter

	State	Zip	Phone	Email	SubDate	Number_of_re...
	US State Code (text)	US Postal Code (text)	US Phone (text)	Email (text)	Date	Integer
1	LA	70116	504-621-8927	jbutt@gmail.com	3/17/2016 12:00 AM	541
2	MI	48116	810-292-9388	josephine_darakjy@...	3/15/2013 12:00 AM	994
3	AK	99501	907-385-4412	lpaprocki@hotmail...	11/24/2013 12:00 AM	586
4	OH	45011	513-578-1893	donette_foller@cox...	4/17/2012 12:00 AM	82
5	OH	44805	419-583-2484	simonamorasca@...	4/13/2016 12:00 AM	959
6	IL	60632	773-57	mitsue_tollner@yahoo...	6/7/2009 12:00 AM	323
7	CA	95111	408	leota@hotmail.com	12/4/2008 12:00 AM	418
8	SD	57105	605-414-2147	sage_wieser@cox.net	4/20/2013 12:00 AM	262
9	MD	21224	410-655-8723	kris@gmail.com	12/31/2012 12:00 AM	993
10	NY	11953	631-3	amaclead@gmail.com	9/3/2015 12:00 AM	567
11	CA	90034	310-498-5651	kiley_caldarera@aol...	3/8/2014 12:00 AM	493
12	OH	44023	440-788-8425	grutal@cox.net	6/12/2013 12:00 AM	885
13	Texas	78045	956-537-6195	calbares@gmail.com	6/25/2011 12:00 AM	827
14	AZ	85013	602-277-4385	mattie@aol.com	12/1/2009 12:00 AM	121
15	TN	37110	931-313-9635	meaghan@hotmail.com	12/26/2015 12:00 AM	373
16	MI	53207	414-661-9598	gladys_rim@rim.org	9/8/2011 12:00 AM	353
17	MI	48180	313-288-7937	yukl_whobrey@aol.c...	9/17/2012 12:00 AM	221

Email

Column Row Table

Filter

SUGGESTIONS

Magic fill ...

Extract email parts

Mask data (obfuscation) ...

Search and replace ...

Change to upper case ...

BOOLEAN

Chart Value Pattern Advanced

Row count

0 0.75 1.5 2.25 3

flou@cox.net

verna@cox.net

thresad@gmail.com

verna@gmail.com

waltzoo@hotmail.com

deandre@yahoo.com

robert@gmail.com

etayne@aol.com

mattie@cox.net

rub@aol.com

talend | Data Preparation

customers Preparation

1 Change to title case on column First_Name

2 Change to title case on column Last_Name

3 Remove trailing and leading characters on column First_Name

4 Remove trailing and leading characters on column Last_Name

5 Search and replace on column Gender

6 Search and replace on column Gender

7 Change date format on column SubDate

8 Change semantic domain on column Zip

9 Delete the rows with empty cells on column Email

10 Delete the rows with invalid cells on column Email

11 Fill empty cells with value on column MaritalStatus_Out

Filters

Find in a column...

Add filter

	Last_Name	Gender	Age	Occupation	MaritalStatus_Out	Salary_Out	Address
	Text	Gender (text)	Text	Text	Text	Text	Address Line (text)
1	Butt	Female	Under 18	K-12 Student	Single	0	6649 N Blue C
2	Darakjy	Male	56+	Self-Employed	Married	100,000-149,999	4 B Blue Ridg
3	Paprocki	Male	45-49	Executive/Manager...	Divorced	150,000-199,999	639 Main St
4	Foller	Male	25-34	Writer	No Data	50,000-99,999	34 Center St
5	Morasca	Female	50-55	Homemaker	Married	100,000-149,999	3 McAuley Dr
6	Tollner	Male	35-44	Academic/Educator	Divorced	100,000-149,999	7 Eads St
7	Dilliard	Male	25-34	Programmer	No Data	100,000-149,999	7 W Jackson E
8	Wieser	Male	25-34	Technical/Engineer	Divorced	150,000-199,999	5 Boston Ave
9	Harrrier	Female	35-44	Academic/Educator	Divorced	< 50,000	228 Runamuck
10	Maclead	Male	25-34	Programmer	Divorced	< 50,000	37275 St Rt 1
11	Caldarera	Male	45-49	Academic/Educator	No Data	150,000-199,999	25 E 75th St
12	Ruta	Male	35-44	Other	Divorced	< 50,000	98 Connectio...
13	Albares	Male	25-34	Executive/Manager...	No Data	> 200,000	56 E Morsehead
14	Poquette	Female	35-44	Other	No Data	< 50,000	73 State Road
15	Garufi	Male	50-55	Academic/Educator	Divorced	> 200,000	60734 E Carri
16	Rjm	Female	18-24	Clerical/Admin	No Data	50,000-99,999	322 New Horiz
17	Whobrey	Male	Under 18	K-12 Student	Single	0	1 State Route...

MaritalStatus_Out

Column Row Table

Filter

SUGGESTIONS

Magic fill ...

Search and replace ...

Change to lower case ...

Change to upper case ...

BOOLEAN

Negate value ...

Chart Value Pattern Advanced

Row count

0 350 700 1050 1400

Single

Married

Divorced

No Data



**SHRI VILEPARLE KELAVANI MANDAL'S
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING**
(Autonomous College Affiliated to the University of Mumbai)
NAAC ACCREDITED with "A" GRADE (CGPA : 3.18)



The screenshot displays the Talend Data Preparation interface. On the left, a list of tasks is visible, including removing trailing characters, searching and replacing, changing date formats, changing semantic domains, deleting rows with empty or invalid cells, and filling empty cells. The main workspace shows a table of customer data with columns: Gender, Age, Occupation, MaritalStatus_Out, Salary_Out, Address, and City. A filter is applied to the 'Age' column, showing 'Under 18'. The table contains 18 rows of data. On the right, a 'SUGGESTIONS' panel offers various actions like 'Magic fill', 'Search and replace', and 'Change to upper case'. Below this, a 'Row count' chart shows the distribution of data across different age groups.

Gender	Age	Occupation	MaritalStatus_Out	Salary_Out	Address	City
Female	Under 18	K-12 Student	Single	0	6649 N Blue Gum St	Ne
Male	Under 18	K-12 Student	Single	0	1 State Route 27	Ta
Female	Under 18	K-12 Student	Single	0	4486 W O St #1	Ne
Female	Under 18	K-12 Student	Single	0	2737 Pistorio Rd #.	Lc
Female	Under 18	K-12 Student	Single	0	461 Prospect Pl #3.	Eu
Female	Under 18	K-12 Student	Single	0	64 5th Ave #1153	Mc
Female	Under 18	K-12 Student	Single	0	9390 S Howell Ave	Al
Male	Under 18	K-12 Student	Single	0	749 W 18th St #45	Se
Female	Under 18	K-12 Student	Single	0	486 Main St	Sc
Female	Under 18	K-12 Student	Single	0	78 Mechanic St	Nc
Male	Under 18	K-12 Student	Single	0	8772 Old County Rd.	Ke
Female	Under 18	K-12 Student	Single	0	48 Cambridge Ave	Mc
Male	Under 18	K-12 Student	Single	0	9 Hwy	Pr
Male	Under 18	K-12 Student	Single	0	4 Bayhill Dr	Lt
Female	Under 18	K-12 Student	Single	0	85624 Butler St	Al
Female	Under 18	K-12 Student	Single	0	98 Veronica Ave	Ps
Male	Under 18	Other	Single	0	59 Highway 1 #2954	La

CONCLUSION:

In this experiment, we learned how to apply ETL steps for a given dataset

Website References:

1. <https://www.talend.com/products/data-preparation/>
2. <https://help.talend.com/r/en-US/7.3/data-preparation-getting-starte>