ETL Project With Talend

Introduction

ETL stands for extract, transform, and load, is the process data engineers use to extract data from different sources, transform the data into a usable and trusted resource, and load that data into the systems end-users can access and use downstream to solve business problems.

ETL Process

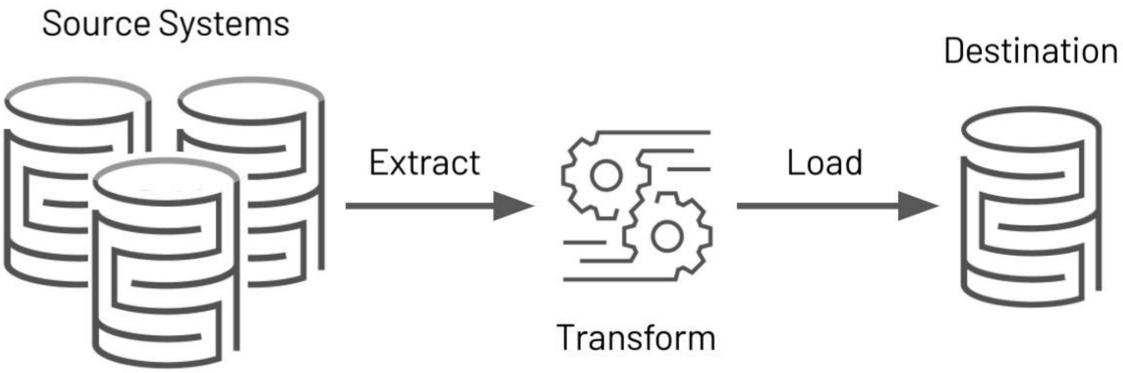


Image credits to: https://www.databricks.com/glossary/extract-transform-load

Extract:

- Reads data from multiple data sources and extracts required set of data.
- Recovers necessary data with optimum usage of resources.

Transform:

- Filtration, cleansing, and preparation of data extracted, with lookup tables.
- Authentication of records, refutation, and integration of data.
- Data to be sorted, filtered, cleared, standardized, translated, or verified for consistency.

Load:

- Writing data output, after transformation to a data warehouse
- Either physical insertion of record as a new row in database table or link processes for each record from the main source

Project Overview

In this ETL project, I extracted the data from 4 different data sources and did data transformation then uploaded it to AWS S3 storage in order to move it to Redshift later.

Data Sources	Staging Area	Final Destination
1- CSV files 2- Excel sheets 3- BigQuery 4- PostgreSQL database	Local PC	AWS S3 then AWS Redshift

Northwind is a database created by Microsoft for training and educational purpose.

Links to the data:
https://github.com/engindemirog/Northwind-Database-Script-for-Postgre-Sql/blob/master/script.sql

https://github.com/neo4j-contrib/northwind-neo4j/tree/master/data

https://docs.google.com/spreadsheets/d/1amQgBgIaUMVEj8gYKbvmIzuoA21ABDiLe0v1orZjjkg/edit#gid=1531710140

Related article this ETL project
https://medium.com/@kmsbmadhan/dimensional-modelling-visualization-of-northwind-database-beaac7fecb20

Objectives

There are many business drivers that can be driven from Northwind data warehouse as follows:

1. **Sales Reporting** to track sales by customer, employees, products, and suppliers to answer the following questions:

- what is our overall sales number?
- How much have we sold of each product?
- Which products are our best and worst sellers?
- Which of our clients order the most products? What do they order?
- How do our sales look when broken down by region?

2. **Request fulfillment Report** to track the order by how much time it has taken to get delivered to the customer and it can be analyzed to see if it can be improved.

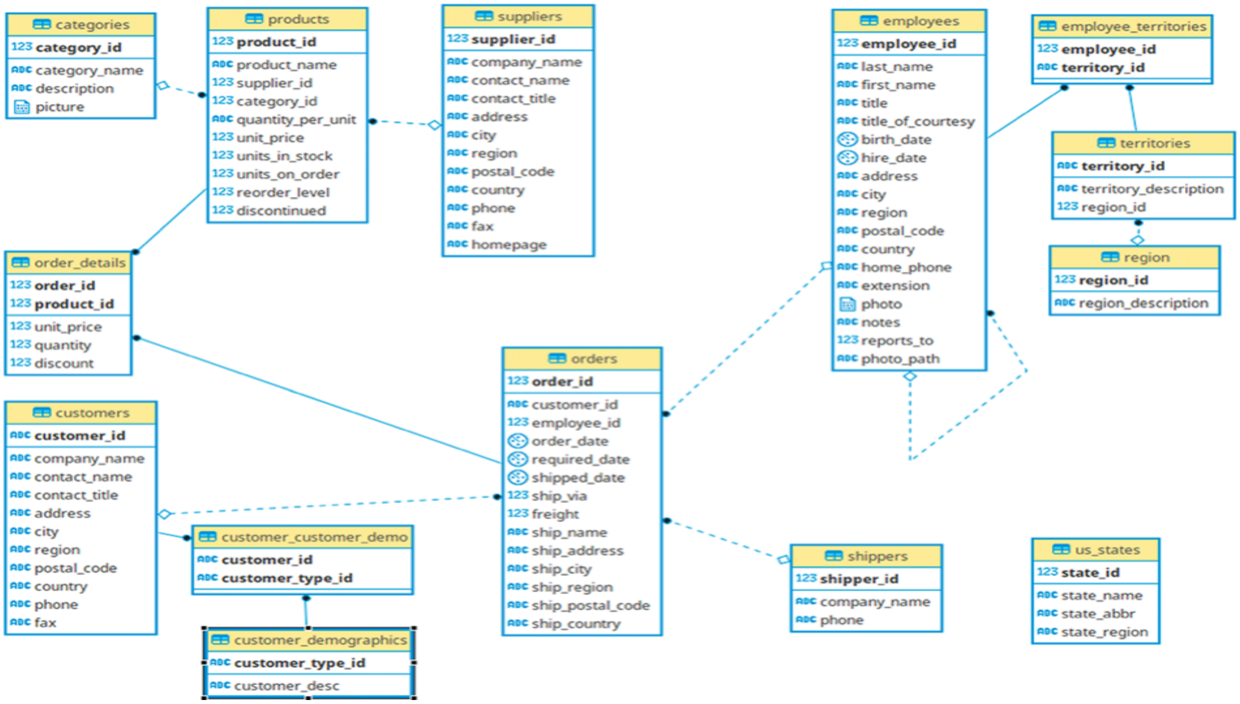
3. **Employee level reporting** to track the performance of the employees and see how it can be improved by either providing rewards to the best performers or giving training to the worst performers or both.

4**. Order distribution & Product inventory analysis** to find orders distributed to customers across the world, track inventory, Order level, and Re-order level of the company for the betterment to answer important questions like

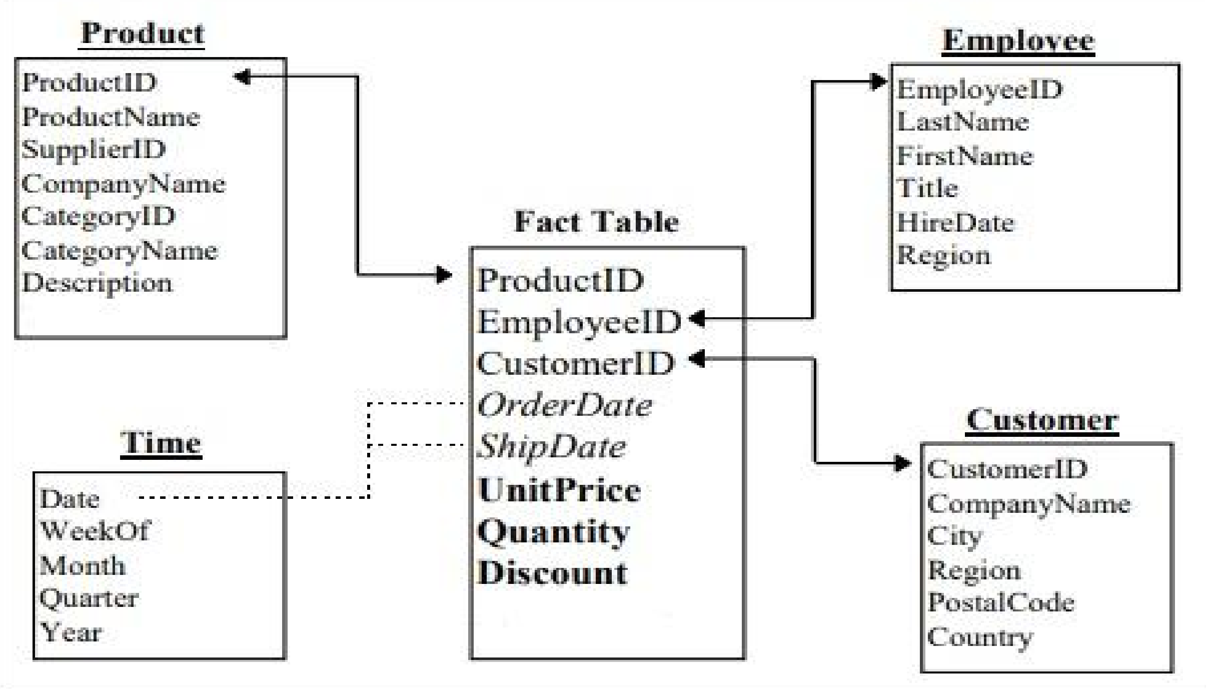
- What are the best-selling products, and do we need to store them more?
- What is the count of products left in the inventory?
- Are we going to run out of any products for delivery?
- What are the products that are going unsold and what can be done as improvement in selling or discontinuing them?
- Can we give discounts on unsold products to get attention in purchase?

Source:
https://medium.com/@kmsbmadhan/dimensional-modelling-visualization-of-northwind-database-beaac7fecb20

Northwind schema:



Target DWH schema:



Solving Some Errors when splitting the data on different data sources

Solve the problem of column header names when Upload the data to bigquery
https://medium.com/google-cloud/bigquery-create-table-from-google-sheets-causing-incorrect-column-names-string-field-0-134f6ecd3fc8

Import CSV file to PostgreSQL
https://www.neilwithdata.com/copy-permission-denied#:~:text=What%20the%20error%20means,the%20server%2C%20not%20the%20client.

Import data from Excel error solving
https://community.talend.com/s/feed/0D73p000004kIYICAM?language=en_US

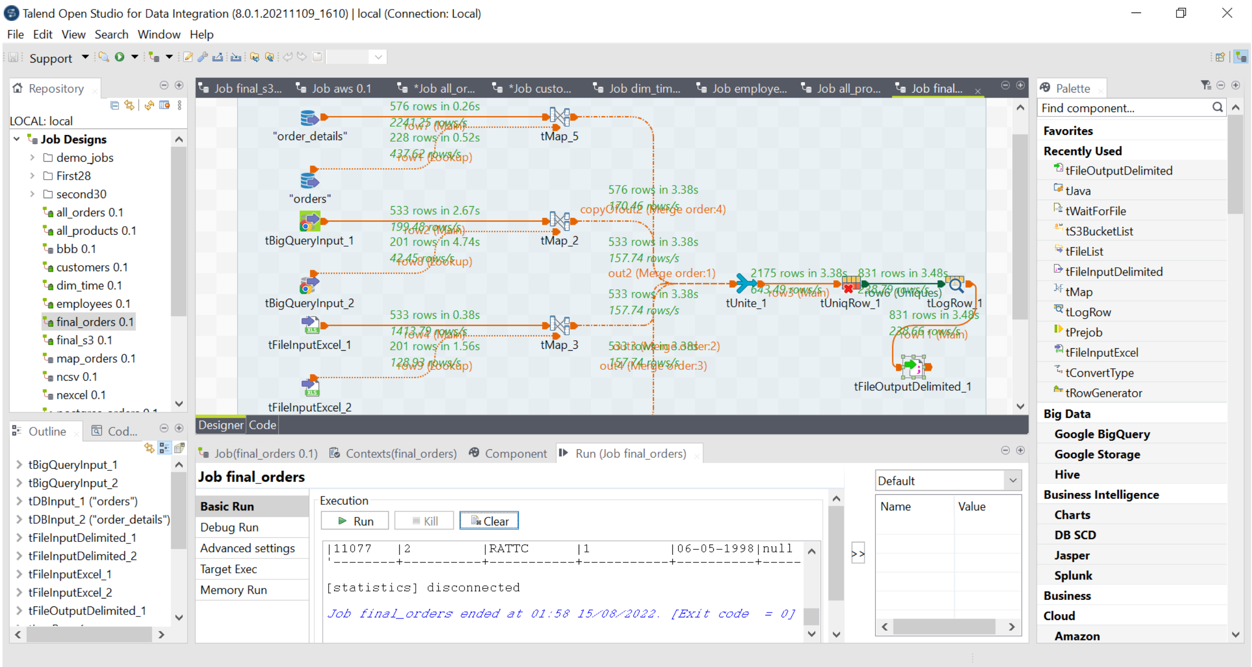
Transformation phase

- Concatenating first name and last name for employees and customers
- Parsing date from CSV files
- Transforming date to the used format
- Append unique rows from the 4 sources together
- Create date table

Load the data

- Upload the new tables to AWS S3 (Simple Storage Service) which is the very popular storage service of Amazon Web Services. It is widely used by customers and Talend provides out-of-the-box connectivity with S3.

Orders Table



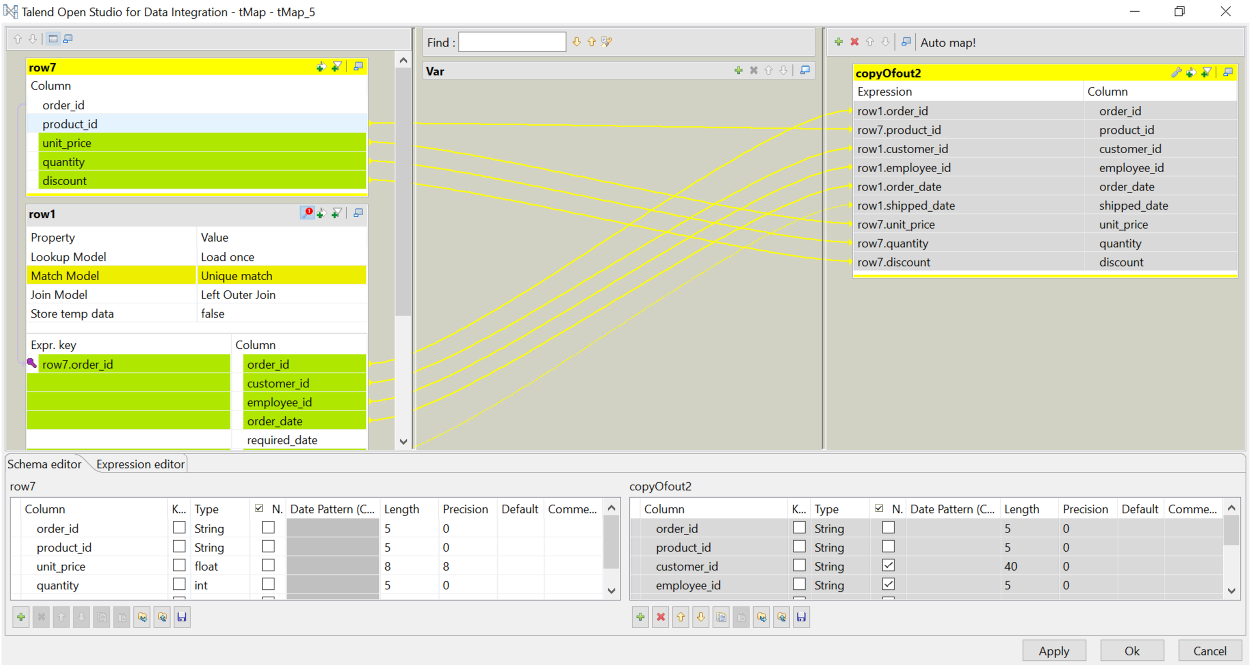
Talend Open Studio for Data Integration (8.0.1.20211109_1610) | local (Connection: Local)

Job final_orders

```
|11077    |2          |RATTC       |1           |06-05-1998|null
'---------+-----------+------------+------------+----------+-----

[statistics] disconnected

Job final_orders ended at 01:58 15/08/2022. [Exit code  = 0]
```



Talend Open Studio for Data Integration - tMap - tMap_5

row7

Column
order_id
product_id
unit_price
quantity
discount

row1

Property	Value
Lookup Model	Load once
Match Model	Unique match
Join Model	Left Outer Join
Store temp data	false

Expr. key	Column
row7.order_id	order_id
	customer_id
	employee_id
	order_date
	required_date

copyOfout2

Expression	Column
row1.order_id	order_id
row7.product_id	product_id
row1.customer_id	customer_id
row1.employee_id	employee_id
row1.order_date	order_date
row1.shipped_date	shipped_date
row7.unit_price	unit_price
row7.quantity	quantity
row7.discount	discount

Schema editor Expression editor

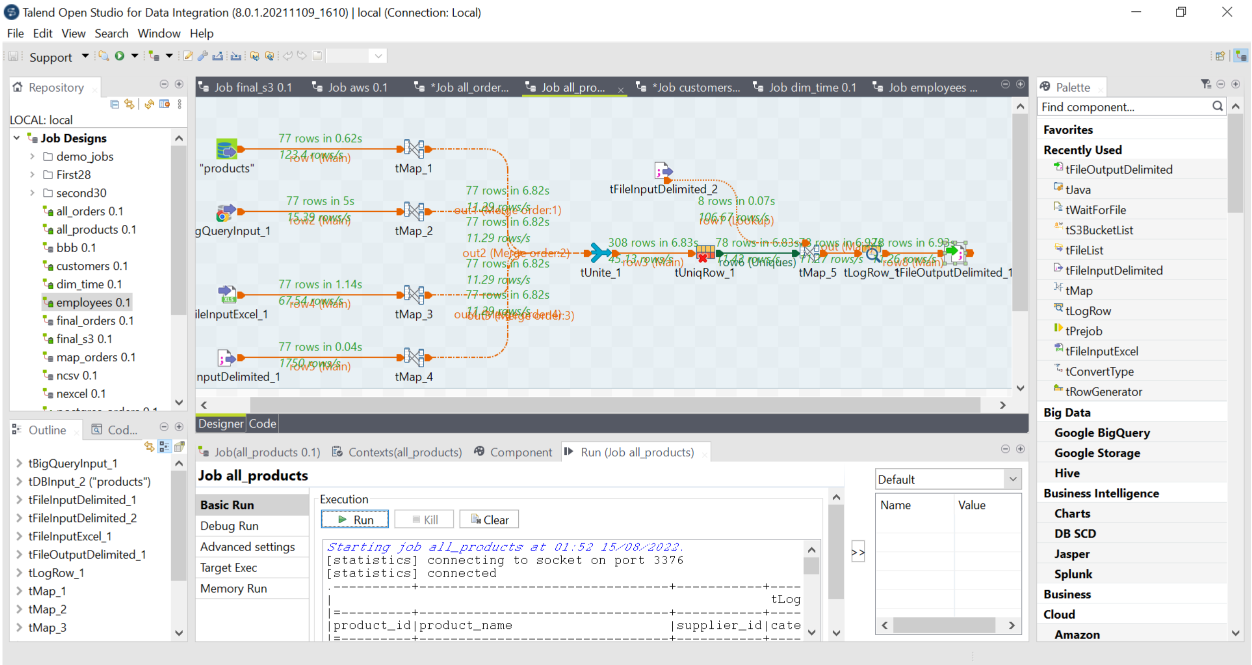
row7

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
order_id	□	String	□		5	0		
product_id	□	String	□		5	0		
unit_price	□	float	□		8	8		
quantity	□	int	□		5	0		

copyOfout2

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
order_id	□	String	□		5	0		
product_id	□	String	□		5	0		
customer_id	□	String	☑		40	0		
employee_id	□	String	☑		5	0		

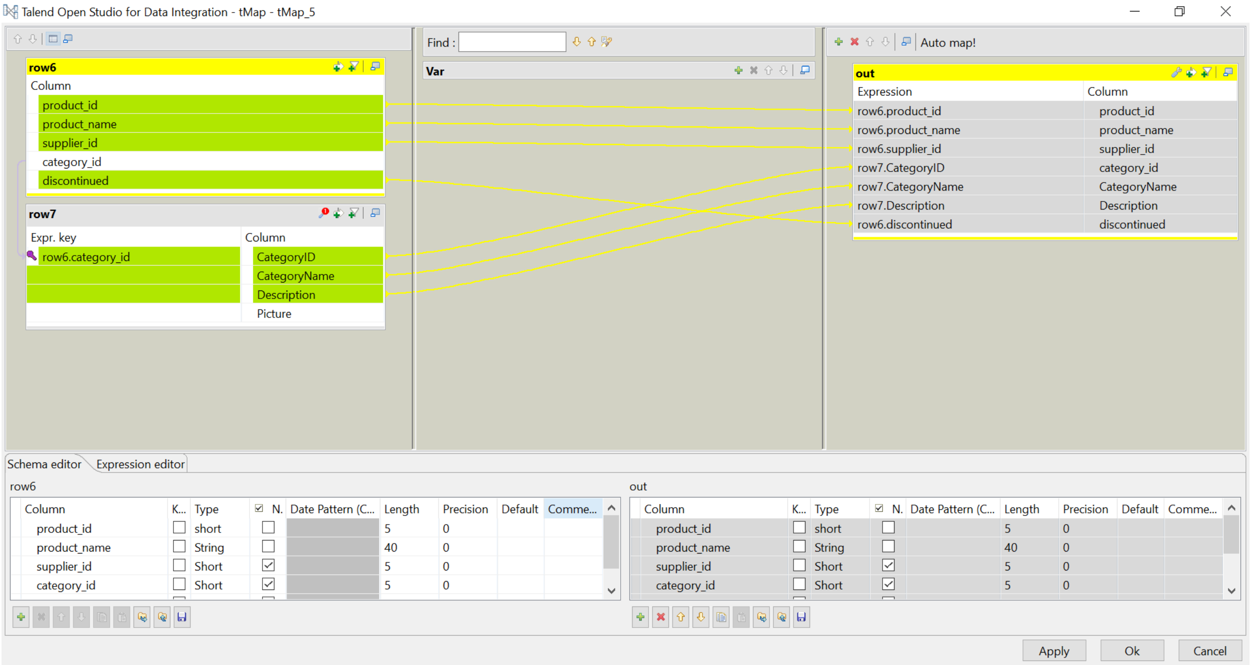
Products Table



Talend Open Studio for Data Integration (8.0.1.20211109_1610) | local (Connection: Local)

Job all_products

```
Starting job all_products at 01:52 15/08/2022.
[statistics] connecting to socket on port 3376
[statistics] connected
.----------+-----------------------------------+-----------+----
|                                                          tLog
|=---------+-----------------------------------+-----------+----
|product_id|product_name                       |supplier_id|cate
|=---------+-----------------------------------+-----------+----
```



Talend Open Studio for Data Integration - tMap - tMap_5

row6

Column
product_id
product_name
supplier_id
category_id
discontinued

row7

Expr. key	Column
row6.category_id	CategoryID
	CategoryName
	Description
	Picture

out

Expression	Column
row6.product_id	product_id
row6.product_name	product_name
row6.supplier_id	supplier_id
row7.CategoryID	category_id
row7.CategoryName	CategoryName
row7.Description	Description
row6.discontinued	discontinued

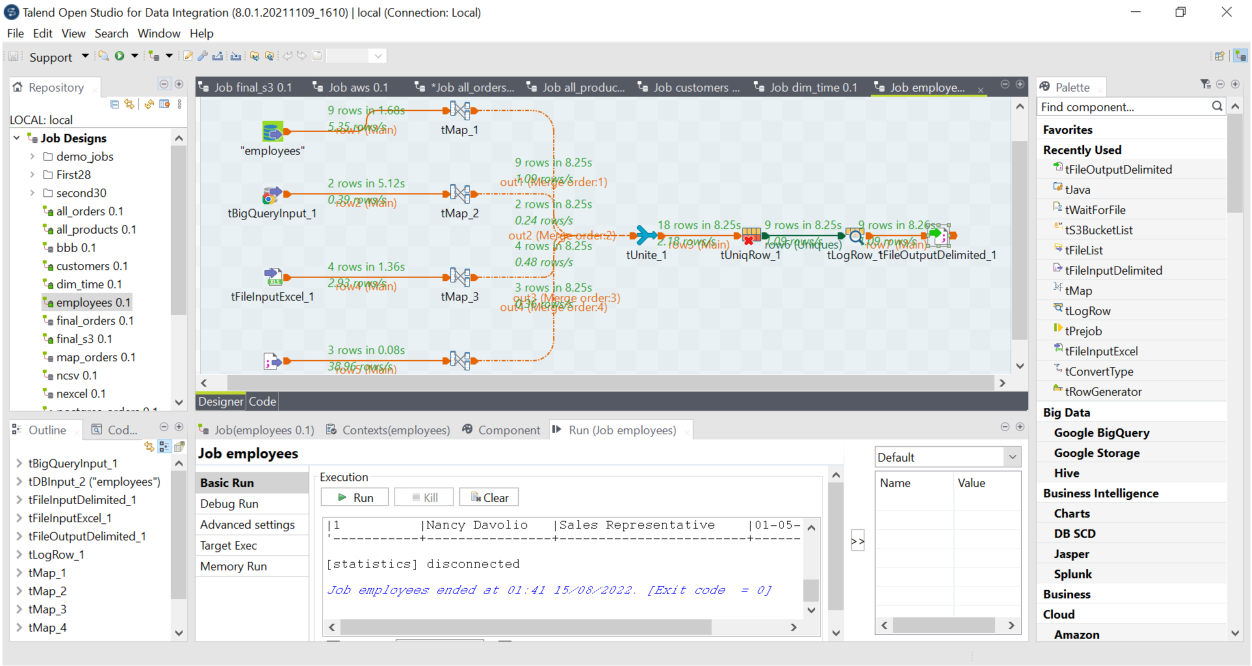
row6

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
product_id	□	short	□		5	0		
product_name	□	String	□		40	0		
supplier_id	□	Short	☑		5	0		
category_id	□	Short	☑		5	0		

out

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
product_id	□	short	□		5	0		
product_name	□	String	□		40	0		
supplier_id	□	Short	☑		5	0		
category_id	□	Short	☑		5	0		

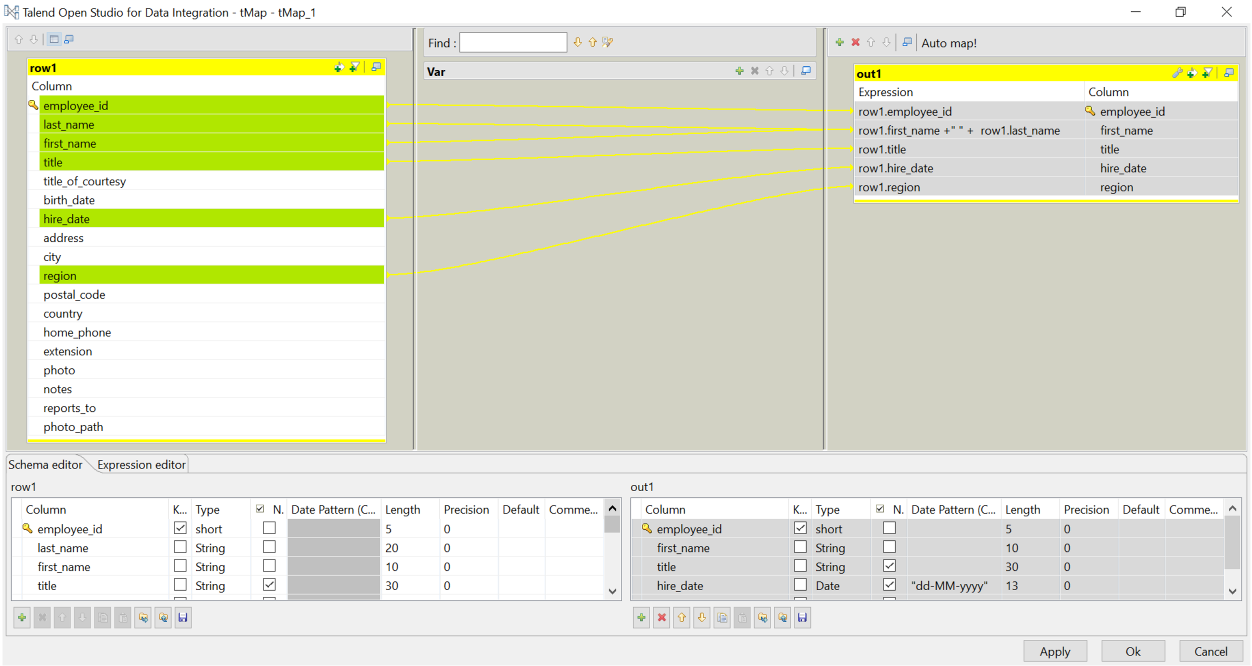
Employees Table



```
|1          |Nancy Davolio    |Sales Representative    |01-05-
'-----------+-----------------+------------------------+------

[statistics] disconnected

Job employees ended at 01:41 15/08/2022. [Exit code  = 0]
```



row1
Column
employee_id
last_name
first_name
title
title_of_courtesy
birth_date
hire_date
address
city
region
postal_code
country
home_phone
extension
photo
notes
reports_to
photo_path

out1

Expression	Column
row1.employee_id	employee_id
row1.first_name + " " + row1.last_name	first_name
row1.title	title
row1.hire_date	hire_date
row1.region	region

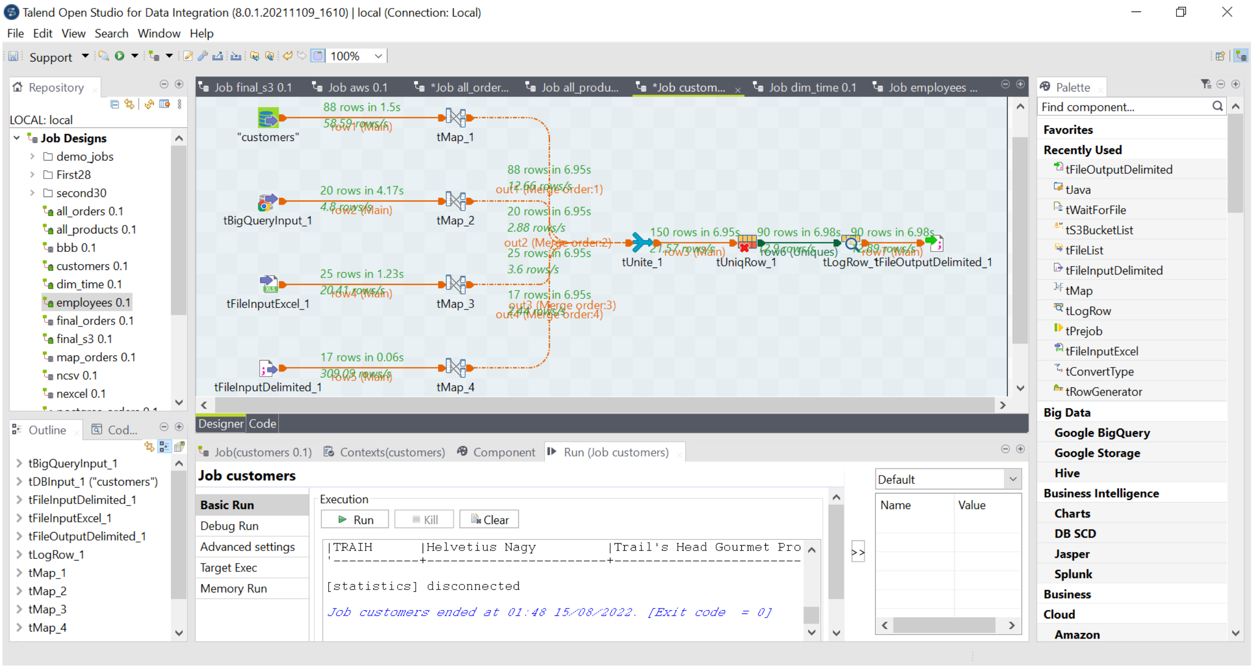
row1

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
employee_id	☑	short	□		5	0		
last_name	□	String	□		20	0		
first_name	□	String	□		10	0		
title	□	String	☑		30	0		

out1

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
employee_id	☑	short	□		5	0		
first_name	□	String	□		10	0		
title	□	String	☑		30	0		
hire_date	□	Date	☑	"dd-MM-yyyy"	13	0		

Customers Table

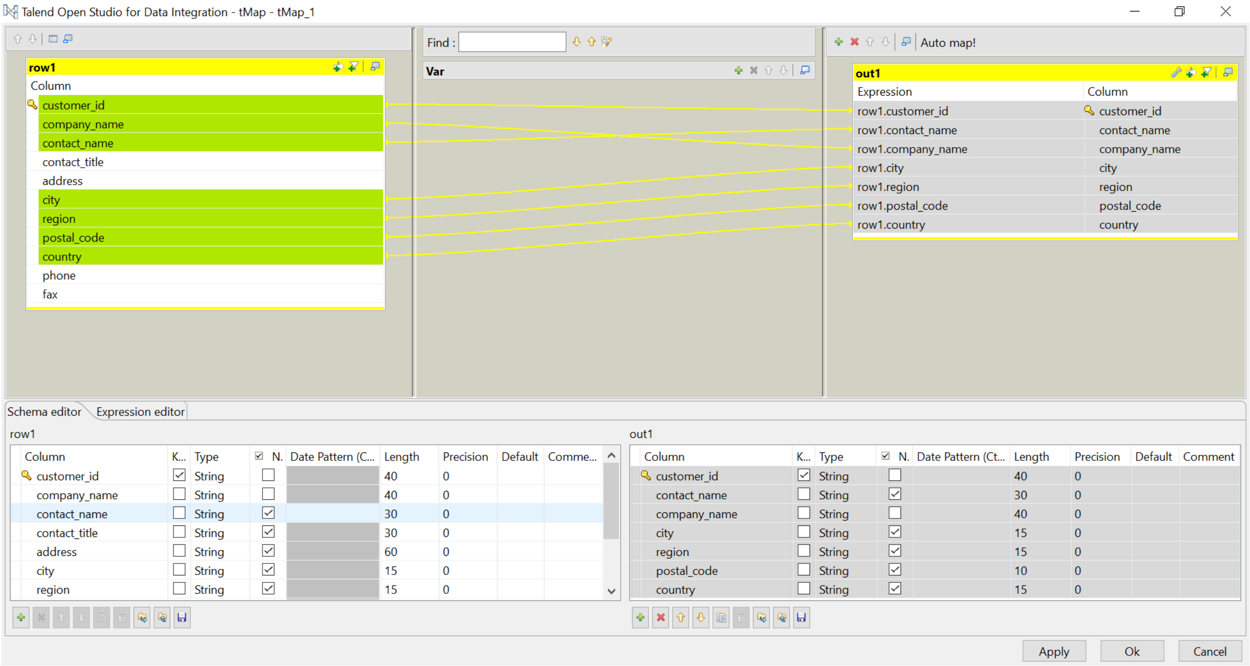


Job customers

```
|TRAIH        |Helvetius Nagy            |Trail's Head Gourmet Pro
'------------+------------------------+------------------------

[statistics] disconnected

Job customers ended at 01:48 15/08/2022. [Exit code  = 0]
```



row1

Column
customer_id
company_name
contact_name
contact_title
address
city
region
postal_code
country
phone
fax

out1

Expression	Column
row1.customer_id	customer_id
row1.contact_name	contact_name
row1.company_name	company_name
row1.city	city
row1.region	region
row1.postal_code	postal_code
row1.country	country

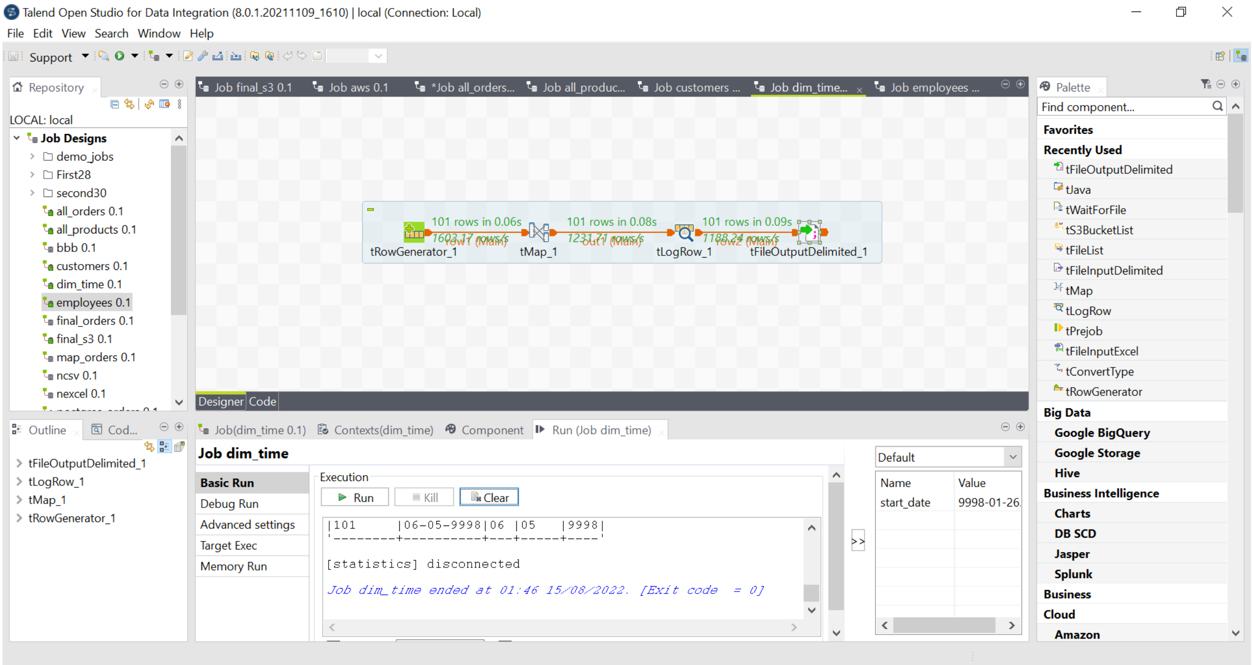
row1

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
customer_id	☑	String	□		40	0		
company_name	□	String	□		40	0		
contact_name	□	String	☑		30	0		
contact_title	□	String	☑		30	0		
address	□	String	☑		60	0		
city	□	String	☑		15	0		
region	□	String	☑		15	0		

out1

Column	K...	Type	☑ N.	Date Pattern (Ct...	Length	Precision	Default	Comment
customer_id	☑	String	□		40	0		
contact_name	□	String	☑		30	0		
company_name	□	String	□		40	0		
city	□	String	☑		15	0		
region	□	String	☑		15	0		
postal_code	□	String	☑		10	0		
country	□	String	☑		15	0		

Dim time



Job dim_time

Basic Run | Debug Run | Advanced settings | Target Exec | Memory Run

Execution: ▶ Run | Kill | Clear

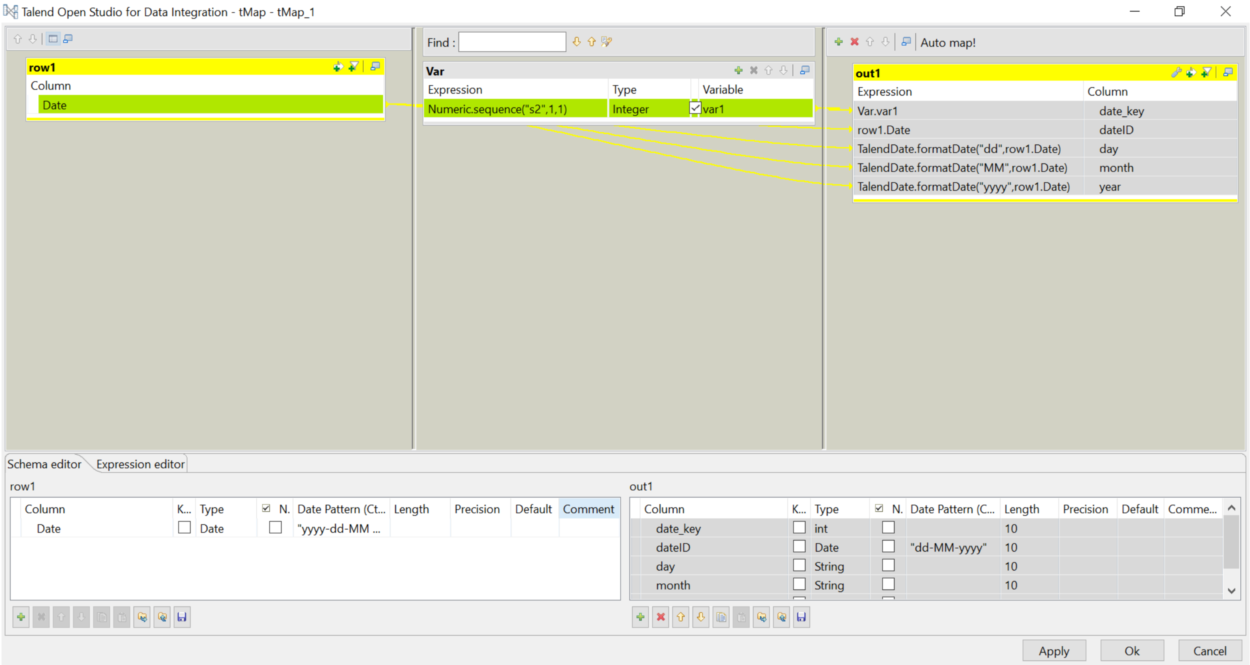
```
|101      |06-05-9998|06 |05   |9998|
'--------+----------+---+-----+----'

[statistics] disconnected

Job dim_time ended at 01:46 15/08/2022. [Exit code  = 0]
```

Default

Name	Value
start_date	9998-01-26.



row1

Column
Date

Var

Expression	Type	Variable
Numeric.sequence("s2",1,1)	Integer	☑ var1

out1

Expression	Column
Var.var1	date_key
row1.Date	dateID
TalendDate.formatDate("dd",row1.Date)	day
TalendDate.formatDate("MM",row1.Date)	month
TalendDate.formatDate("yyyy",row1.Date)	year

row1

Column	K...	Type	☑ N.	Date Pattern (Ct...	Length	Precision	Default	Comment
Date	□	Date	□	"yyyy-dd-MM ...				

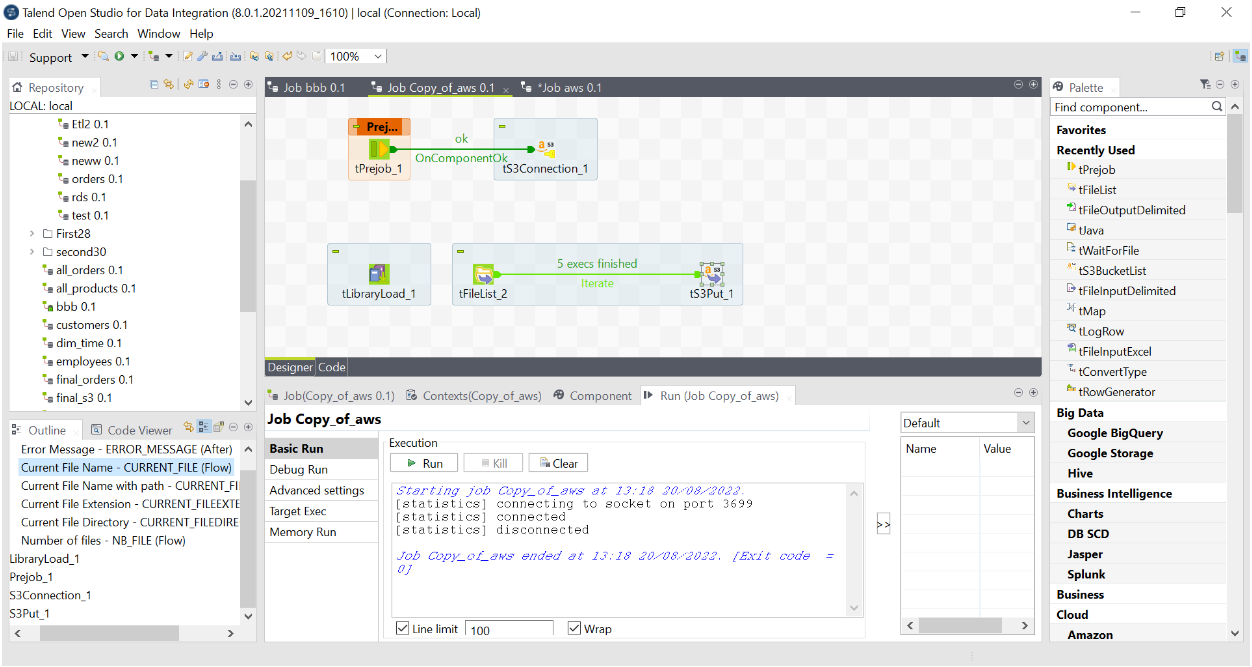
out1

Column	K...	Type	☑ N.	Date Pattern (C...	Length	Precision	Default	Comme...
date_key	□	int	□		10			
dateID	□	Date	□	"dd-MM-yyyy"	10			
day	□	String	□		10			
month	□	String	□		10			

Apply | Ok | Cancel

Upload the final tables to AWS S3

Iterate through the files in the data staging folder and upload them to AWS S3 storage in order to move them to Redshift data warehouse.

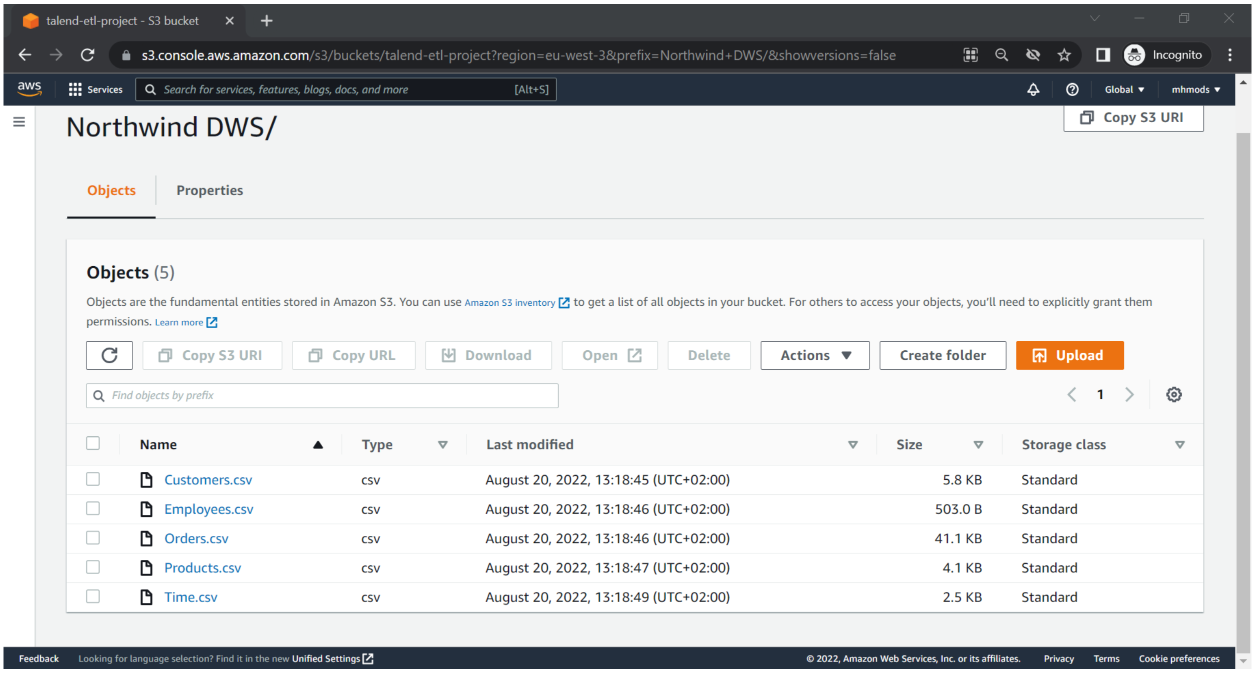


Job Copy_of_aws

```
Starting job Copy_of_aws at 13:18 20/08/2022.
[statistics] connecting to socket on port 3699
[statistics] connected
[statistics] disconnected

Job Copy_of_aws ended at 13:18 20/08/2022. [Exit code = 0]
```

☑ Line limit 100 ☑ Wrap



talend-etl-project - S3 bucket
s3.console.aws.amazon.com/s3/buckets/talend-etl-project?region=eu-west-3&prefix=Northwind+DWS/&showversions=false
Copy S3 URI
Objects
Properties
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix

Northwind DWS/

Objects (5)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Name	Type	Last modified	Size	Storage class
Customers.csv	csv	August 20, 2022, 13:18:45 (UTC+02:00)	5.8 KB	Standard
Employees.csv	csv	August 20, 2022, 13:18:46 (UTC+02:00)	503.0 B	Standard
Orders.csv	csv	August 20, 2022, 13:18:46 (UTC+02:00)	41.1 KB	Standard
Products.csv	csv	August 20, 2022, 13:18:47 (UTC+02:00)	4.1 KB	Standard
Time.csv	csv	August 20, 2022, 13:18:49 (UTC+02:00)	2.5 KB	Standard

Link for the method: https://www.youtube.com/watch?v=TqJAd6RQypU&t=4s

Courses

I have studied these data engineering courses before I made the ETL project.

University course for data warehouse
https://www.youtube.com/playlist?list=PLiJhHdYdI84DzwH47IZQWN6de1tOq8LFu

Big Data Engineering in Depth Course
https://www.youtube.com/playlist?list=PLxNoJq6k39G_m6DYjpz-V92DkaQEiXxkF

Talend course
https://www.youtube.com/playlist?list=PLOr008ImHvfan_fuDr5RVyexpeYJAp9FX

Talend ETL project (the previous project was inspired by it)
https://www.youtube.com/playlist?list=PLKdHo47jRFvf1VzST1RDYiAs8ItMFychE

data engineering podcast
https://www.youtube.com/playlist?list=PLiKvD85qG0l6pLvsQChJO6UBFLfMFfwOw

Thanks for Reading, and I will appreciate your feedback.
Kindly check my Previous projects here: https://linktr.ee/mhmod36

See you next project,
Mahmoud Sallam