******

**Technical Assessment**

**Vince Iarusci**

**Friday May 15, 2020**

Contents

[Technical Assessment Requirements 3](#_Toc40433937)

[Overview 3](#_Toc40433938)

[Main Tasks 3](#_Toc40433939)

[Timeline 3](#_Toc40433940)

[Information 4](#_Toc40433941)

[Toolsets Used 5](#_Toc40433942)

[Database Structure 5](#_Toc40433943)

[Source Data 5](#_Toc40433944)

[Dataset 5](#_Toc40433945)

[Data Analysis and Profiling 5](#_Toc40433946)

[Tabular summary of Null/Blank Columns 5](#_Toc40433947)

[Stockcode Duplicate Descriptions 6](#_Toc40433948)

[Database Tables 6](#_Toc40433949)

[Staging Schema Tables 6](#_Toc40433950)

[ODS Schema Tables 8](#_Toc40433951)

[Entity Relationship Diagram 10](#_Toc40433952)

[Table Joins 10](#_Toc40433953)

[SQL Scripts 11](#_Toc40433954)

[ETL Processes 11](#_Toc40433955)

[Components 11](#_Toc40433956)

[Transformations 12](#_Toc40433957)

[Staging 12](#_Toc40433958)

[ODS - DIMS 12](#_Toc40433959)

[F\_Sales 14](#_Toc40433960)

[Jobs 15](#_Toc40433961)

[Reports and Dashboard 17](#_Toc40433962)

[Reports List 17](#_Toc40433963)

[Dashboard 17](#_Toc40433964)

# Technical Assessment Requirements

## Overview

One of the biggest European company focused on e-commerce would like to better understand their

customers purchase profile and be more successful on upcoming marketing campaigns that include but no

limited to market basket analysis. A dataset representing a sample of their B2B transactions was shared

with you (BI Engineer) and it was requested end-to-end analytical solution that will support the decision

making of SLT group.

## Main Tasks

The solution includes:

1) The dataset should be loaded on a RDBMS, you can use free versions available like MySQL.

2) The dataset has potentially missing or non-expected values based on the columns definition that will be

presented later. It means that a data cleansing process should be applied first, and these purged data must

be moved to a temporary data structure to be analyzed and manually fixed. This process will create a new

and filtered dataset that should be loaded on the same RDBMS.

Note: The data purge process(es) and the populated data structure are part of the solution and

should be part of the deliverable.

3) To better support data analysis, a dimensional data model (s) should be created and a set of ETL

processes developed to feed this model.

Note: The data model diagram and the ETL scripts are part of the solution and should be part of the

deliverable. The ETL scripts should be part of a job or similar functional object to support a one-shot

load considering dependencies. It means that a "job call" can load the dimensional data model with no

need to start ETLs one by one. It must be considered in this job the possibility to perform a full dataload

or incremental data-load as well.

4) The use of a BI Tool (Power BI, Tableau, ...) to build up reports and dashboards (Sales Book) is

mandatory (you can use free versions available) to give a self-service experience to the final user. The

expertise of the BI Engineer to design valuable data analysis is a key asset in this technical assessment

considering all the attributes belonging to the dataset, but additionally the SLT group would like to see

important charts like:

§ What time do people often purchase online?

§ How many items each customer buy?

§ Top 10 best sellers’ products?

§ Average transaction value (total revenue / number of transactions) Year over year.

Note: A high dollar amount could mean that shoppers are purchasing your more expensive

products or they’re buying larger quantities.

§ Basket Analysis including average size of basket and the set of common products purchased.

§ The frequency of cancelation (number of cancelled invoices) and average amount of cancelation. Is

there any common product associated with cancelations?

Note: Histograms, Time-series (day, month, quarter, year), maps / heat-maps visualizations and the use

dimensional filters are well appreciated technical features in analytical “books”. Tip: Data Analysis is

composed of a balanced set of reports and dashboards that creates a compelling Storytelling.

## Timeline

§ May.05 (noon): Send Technical assessment to the candidates (this document)

§ May.06 (noon): The candidate can send a suggested hour for a call in case of doubts or questions.

§ May.07 (afternoon): Call Marcos-Candidate (dismiss potential technical doubts).

§ May.14: Due Date to deliver the solution (send the hyperlink)

§ May.15: Solution Presentation (1 hour). To be scheduled.

§ May.18: Candidates will have their solutions ranked, and results send to hiring manager.

§ Final Result (Expected Date): Until the end of May the selected candidate will be notified.

## Information

§ Each candidate will have 1-hour call-conference to present a detailed solution (May.15).

§ Each candidate should send a hyperlink where all deliverables can be accessed by WVC (May.14).

§ Deliverables: Database Scripts, Reports/Dashboards (pdfs) and a readme file with instructions or

orientation to the WVC-evaluator to support the solution understanding and evaluation.

# Toolsets Used

|  |  |
| --- | --- |
| **Stack** | **Tool** |
| Database | PostgreSQL 12 RDBMS |
| ETL | Pentaho Data Integration V9.0 |
| Reporting & Analysis | Microsoft Power BI Descktop V2.79 |
| Code Management | Github - (Repo: WVC) |

# Database Structure

## Source Data

### Dataset

File Name: Dataset.xlsx

Attributes Information:

**InvoiceNo**

Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction.

If this code starts with letter 'c', it indicates a cancellation.

**StockCode**

Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

**Description**

Product (item) name. Nominal.

**Quantity**

The quantities of each product (item) per transaction. Numeric.

**InvoiceDate**

Invoice Date and time. Numeric, the day and time when each transaction was generated.

**UnitPrice**

Unit price. Numeric, Product price per unit.

**CustomerID**

Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

**Country**

Country name. Nominal, the name of the country where each customer resides.

## Data Analysis and Profiling

### Tabular summary of Null/Blank Columns

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table Name** | **Column Name** | **Total Records** | **Invalid Records** | **% Null/Blank** |
| dataset | Invoiceno | 541909 | 0 | 0% |
|  | invoicedate | 541909 | 0 | 0% |
|  | customerid | 541909 | 135080 | 24.93% |
|  | country | 541909 | 0 | 0% |
|  | stockcode | 541909 | 0 | 0% |
|  | description | 541909 | 1454 | 0.27% |
|  | quantity | 541909 | 0 | 0% |
|  | unitprice | 541909 | 0 | 0% |

* 1454 null values for description are included in the customerid invalid records

### Stockcode Duplicate Descriptions

The following query was executed to identify duplicate stockcode descriptions:

SELECT stockcode, COUNT(description)

FROM ods.d\_product

GROUP BY stockcode

HAVING COUNT(stockcode) > 1

order by stockcode

Solution – added Unique Rows step to sort on the stockcode and select the first record in the dupcate list

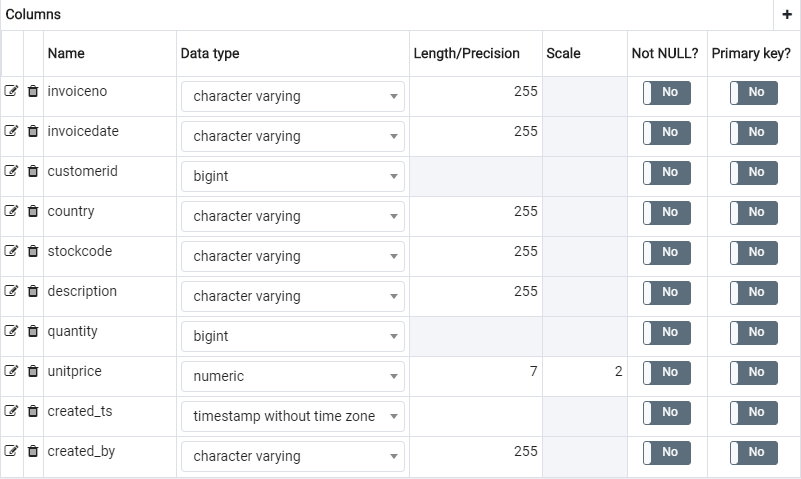
## Database Tables

Two schemas were created in the database to separate the staging tables and the operational data store (datawarehouse) tables. The schemas are Staging and ODS.

### Staging Schema Tables

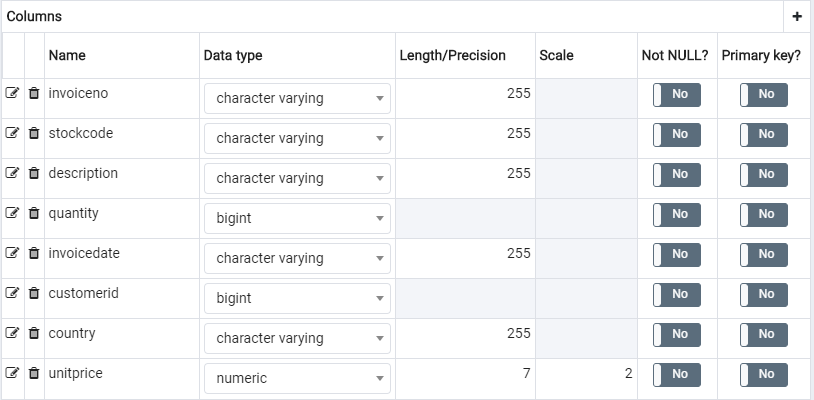
#### Staging.Dataset

Staging table used to source the dimension and fact tables in the ODS. The source data has been cleansed prior to loading of the table.



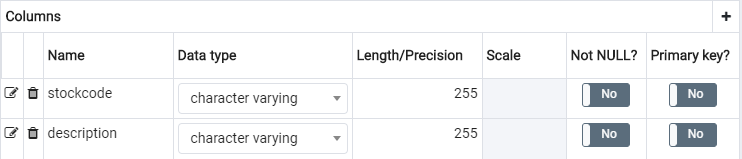
#### Rejected Records

This table contains the rejected records that have bad or missing data. The data in this table is not used in the reporting and analysis objects.



#### Rejected Stockcode Duplicates

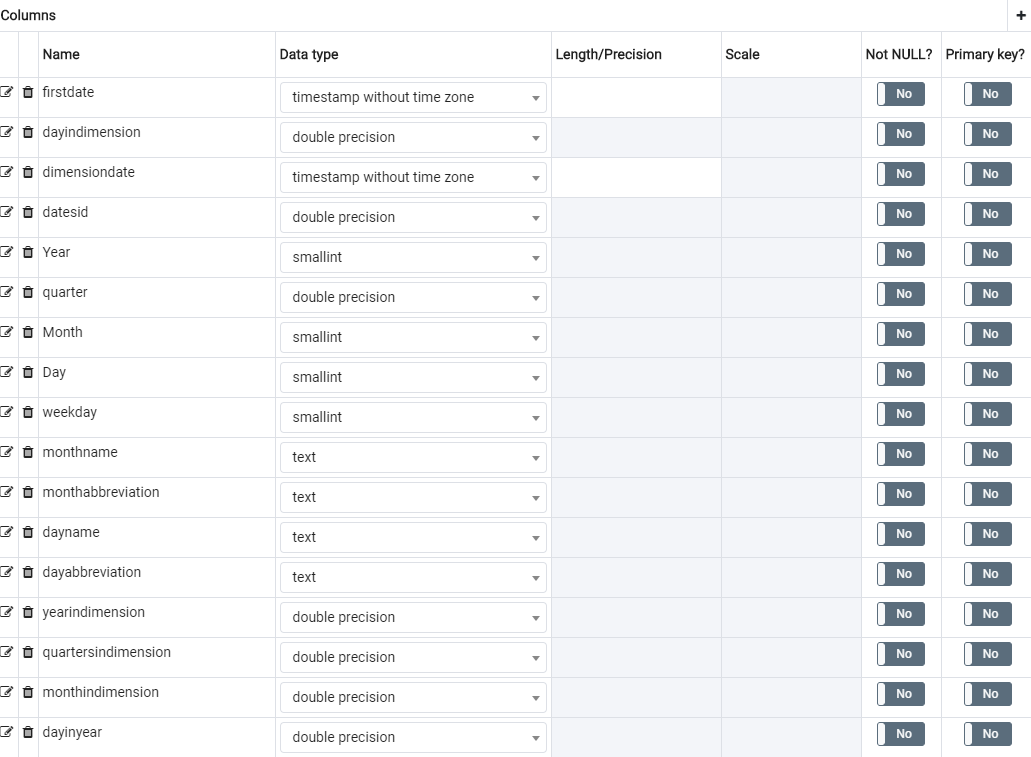
During the analysis of the dataset, it was found that there are stockcodes that have multiple descriptions. These duplicates were causing an issue with the D\_Product dimension and the uniqueness of the records was compromised and causing a many-to-many relationship between the dimension and sales fact. To fix the issue, the first record was used in the sort order and all duplicates were removed to the output file.



### ODS Schema Tables

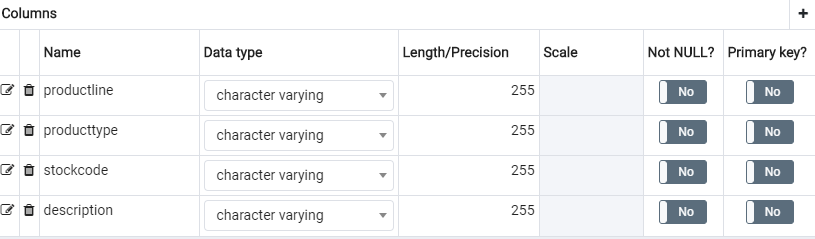
#### D\_Date

Dimension to store time attributes



#### D\_Product

Dimension to store stockcode attributes. During the data analysis, the findings showed that the stockcode values contained logic that was used to create the product hierarchy. The first character was used to define the product line and the first 3 characters was used to create the product type groupings. The actual value was used for any stockcodes that do not start with a numeric character.



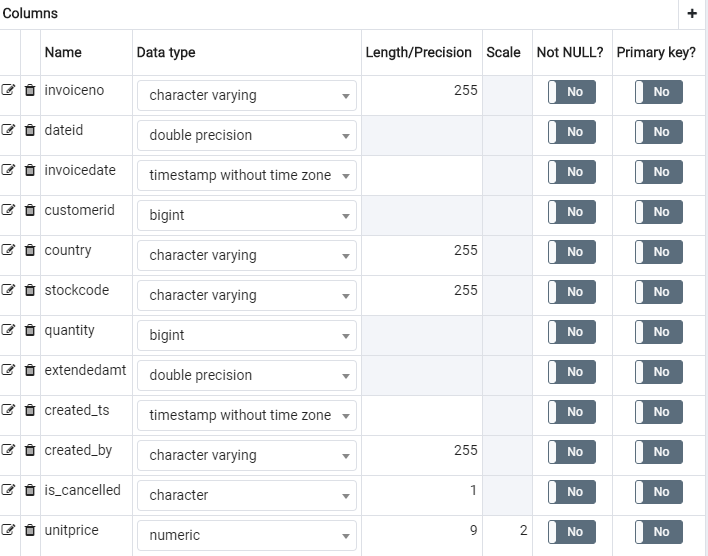
#### D\_Countries

Dimension to store the country and region attributes. Note that the separate Regions\_Countries.xlxs source spreadsheet was created to define the region groupings.



#### F\_Sales

Fact table to store the measures for the sales model.



## Entity Relationship Diagram

The data model is designed in a star schema structure with the following tables…

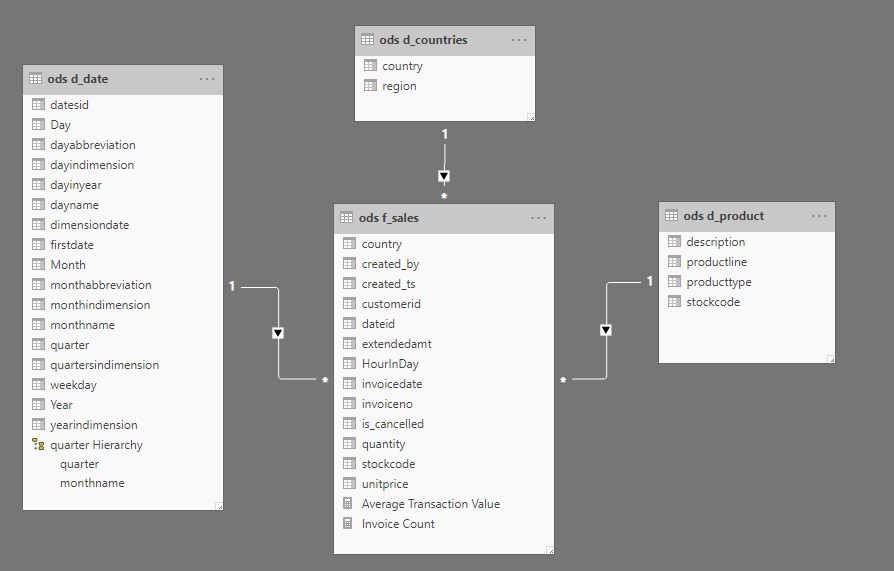
Dimensions:

* D\_Date
* D\_Product
* D\_Countries

Facts:

* F\_Sales

ERD Diagram



## Table Joins

|  |  |  |
| --- | --- | --- |
| **Table.Column** | **Join** | **Table** |
| ods.d\_date.dateid | One to Many (1..n) | ods.f\_sales.datesid |
| ods.d\_product.stockcode | One to Many (1..n) | ods.f\_sales.stockcode |
| ods.d\_countries.country | One to Many (1..n) | ods.f\_sales.country |

## SQL Scripts

Source and target SQL scripts to create the data flows have been saved to the **Scripts** folder in the WVC Github repo.

Script Files:

Staging

* Staging\_Dataset\_Output.sql
* Rejected\_Records\_Output.sql

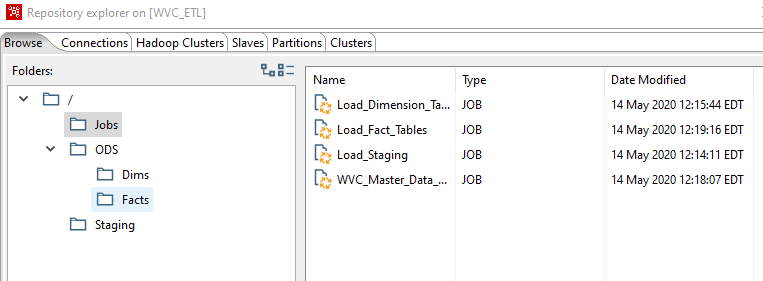
ODS

* D\_Date\_Output.sql
* D\_Product\_Input.sql
* D\_Product\_Output.sql
* Rejected\_Stockcode\_Duplicates.sql
* D\_Countries\_Output.sql
* F\_Sales\_Input.sql
* F\_Sales\_Output.sql

# ETL Processes

## Components

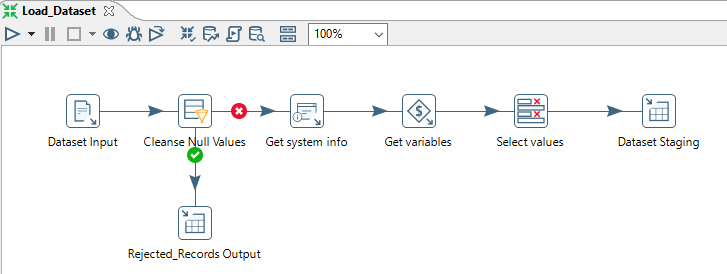
All jobs and transformations are executed using the Pentaho Data Integration tool. The components for the data flows are saved under the following structure:



## Transformations

### Staging

#### Load Dataset

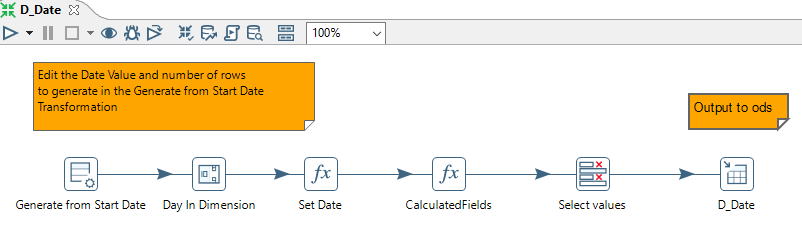


There are 6 steps used in this transformation.

1. The Dataset Input step loads the data from the Dataset.csv source file. All fields are left and right trimmed and data types and formatting are defined.
2. Next, the customerids that have a null value are removed from the data flow and loaded into the Rejected\_records table in the staging schema.
3. The Get System Info step populates the current system date for the Created\_TS column.
4. The Get variables step generates the user name for the Created\_By column.
5. Define the data types and metadata in the Select Values step
6. The Dataset Staging step defines the connection settings for the output table. The table is set to truncate on load.

### ODS - DIMS

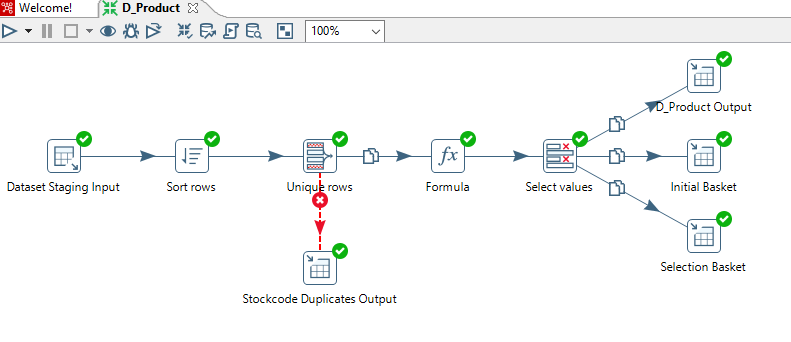
#### D\_Date



There are 6 steps used in this transformation.

1. Using a “Generate Rows” step, set the first day of the dimension to iterate an arbitrary number of times. In the sample, 5475 lines are created, which means there are 5475 days in the dimension or 15 years. Start date is set to 2010-01-01.
2. Next, use an “Add Sequence” step to create a counter for each row.
3. Use a “Formula” step to create each date in the dimension, using the counter from the previous step.
4. Create the rest of the date dimension elements using another “Formula” step.
5. Define the data types and metadata in the Select Values step
6. Finally, output the dimension table to an Excel file. You can change this to any other output type that you would like.

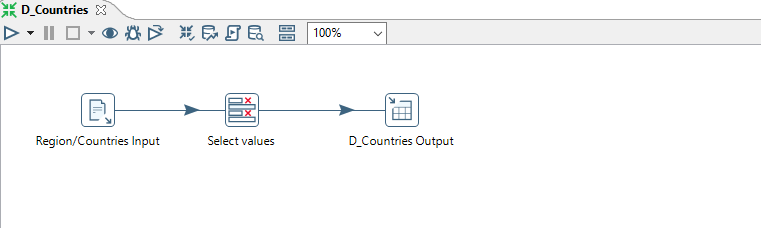
#### D\_Product



There are 6 steps used in this transformation.

1. The Dataset Staging Input step loads the data from the Staging.Dataset table.
2. The Sort Rows step is required to sort the data on the stockcode column. This is required to sequence the rows that are duplicated due to multiple description.
3. The Unique Rows step keeps the first row in the previous step’s sort order and passes those records to the Formula step. The rejected duplicates are loaded into the Staging.rejected\_stockcode\_duplicates table.
4. The Formula step is used to define the productline and producttype groupings for the dimension.
5. Define the data types and metadata in the Select Values step
6. The D\_Product Output step defines the connection settings for the output table. The table is set to truncate on load. The Initial Basket and Selection Basket are stand alone output tables that are used for the Basket Analysis reports. These tables are joined virtually at the report level.

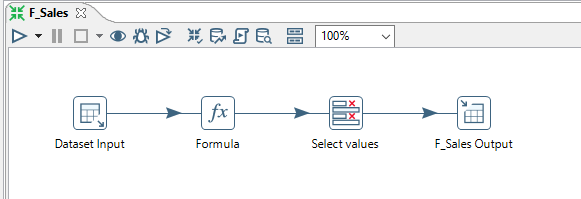
#### D\_Countries



There are 3 steps used in this transformation.

1. The Regions/Countries Input step loads the data from the Regions\_Countries.csv source file.
2. Define the data types and metadata in the Select Values step
3. The D\_Countries Output step defines the connection settings for the output table. The table is set to truncate on load.

## F\_Sales

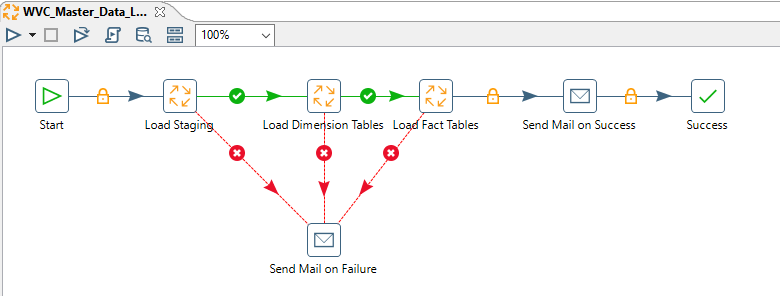


There are 4 steps used in this transformation.

1. The Dataset Input step loads the data from the staging.dataset source file.
2. The Formula step calculates the values for the new extendedamt, is\_cancelled and dateid fields.
3. Define the data types and metadata in the Select Values step
4. The F\_Sales Output step defines the connection settings for the output table. The table is set to truncate on load.

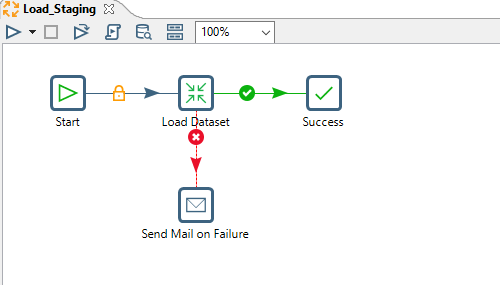
## Jobs

#### WVC\_Master\_Data\_Load



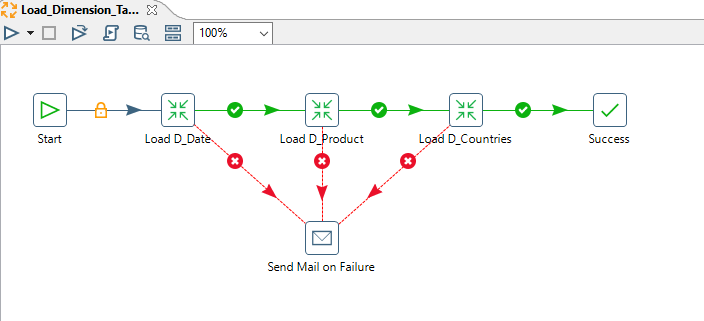
The WVC\_Master\_Data\_Load job is the primary “wrapper” job that runs and calls the Staging, Dimension and Fact Jobs. The order of the flow is important as some of the steps are dependent on the previous jobs steps. All jobs have a modular design so that the separate components can be run individually. The Master job also has an email step that notifies the administrator when the job completes or if there is a failure during the run process. The Master job can also be run from the command line using the Pentaho Kettle scripting tool. The command line script can be automated with the Microsoft Scheduler tool.

#### Load\_Staging



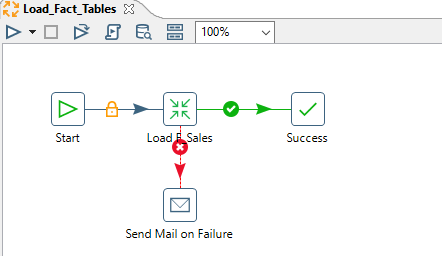
The Load\_Staging job calls the Load Dataset transformation. The job also has an email step that notifies the administrator in the case there is a failure during the run process.

#### Load\_Dimension\_Tables



The Load\_Dimension\_Tables job calls the Load D\_Date, Load D\_Product and Load D\_Countries transformations. The job also has an email step that notifies the administrator in the case there is a failure during the run process.

#### Load\_Fact\_Tables



The Load\_Fact\_Tables job calls the Load F\_Sales transformation. The job also has an email step that notifies the administrator in the case there’s a failure during the run process.

# Reports and Dashboard

Reports and the Dashboard are developed using Power BI. A model is created to connect to the Postgres database to access the dimension and fact tables. The relationships are defined in the model and transformations are created to calculate additional columns and measures required at the report level.

## Reports List

* Online Purchases by Hour
* Unit Sales by Customer
* Top 10 Customer Key Products
* Purchases by Country
* Average Transaction Value – Seasonal
* Cancelled Invoices and Discounts

## Dashboard

* Sales and Marketing