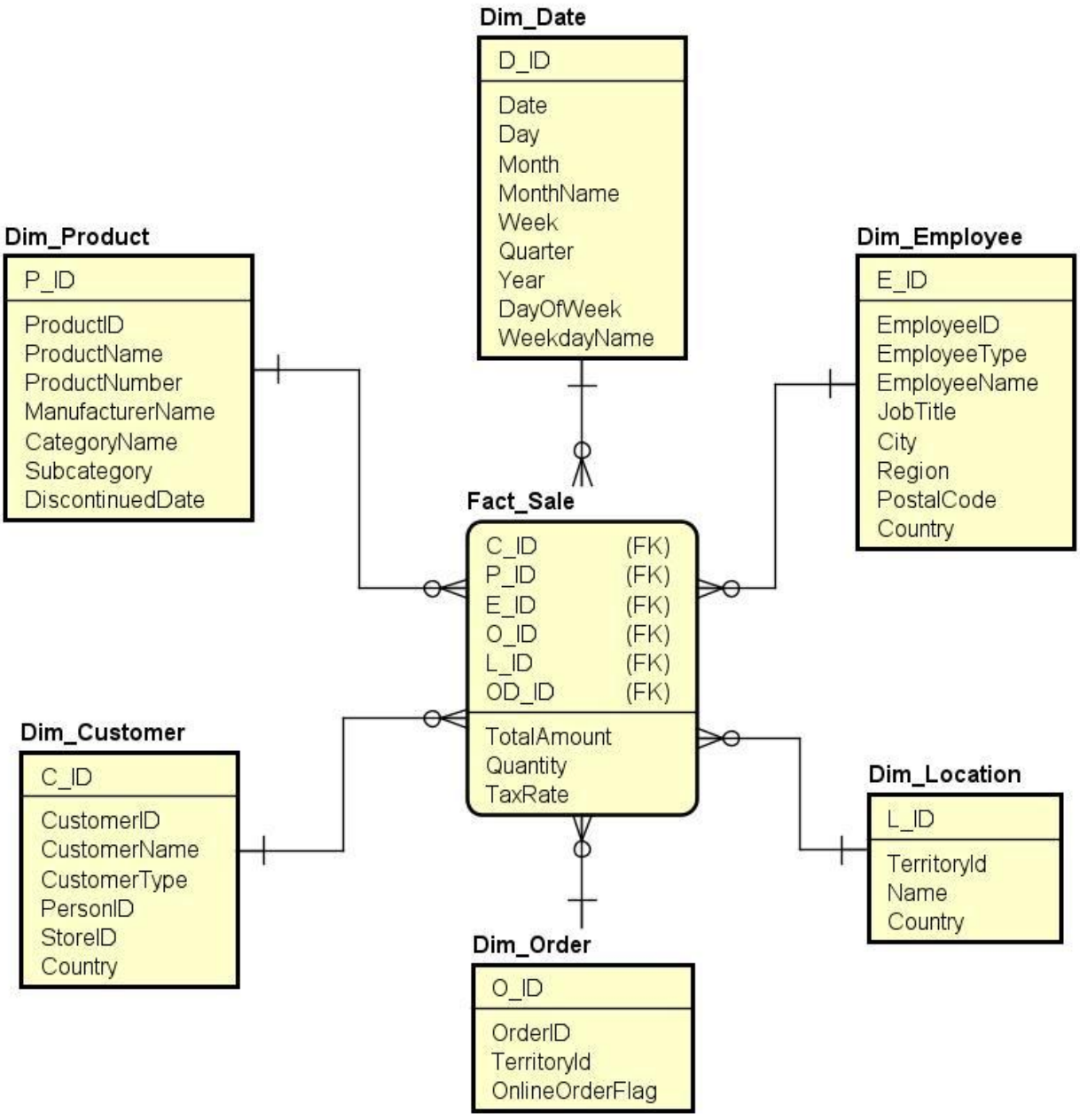
Revisions based on Peer Correction Sheet and problems we encountered during ETL

ER Diagram:



Updated ER Diagram showing the star schema

Dim_Order:

- Deleted Billing and Delivery Address and added TerritoryId instead
  Reason: Address contains the Street number, Street name, so, therefore, cannot group by it.

Dim_Customer

- Deleted Address
  Reason: Address contains Street number, Street name, so, therefore, cannot group by it.
- Deleted CompanyName because we will make use of CustomerName
  Reason: CustomerName is either first+last name if it is an individual customer or store name if it is a wholesale company order
- Deleted City, Region and PostalCode
  Reason: We don't have enough memory to run the query

Dim_Location

- Deleted City and Postal Code
  Reason: Results in a lot of unnecessary duplicates

Fact_Sale:

- Changed TotalAmountInEU to TotalAmount
  Reason: Apparently, we should assume that the sales will be comparable even though they have different currencies
  Also to make our life easier :))

- Added OrderDate
  Reason: In order to see how many orders have been made in relation to a personnel

SQL

- In the Customer table, the CustomerType data type changed to nchar(2) instead of nvarchar(2)

Reason: They has been made a mismatch between our source-target mappings and the implementation
- In Dim_Date we changed the data type of Quarter from nvarchar(2) to int.
  Reason: Following the example from northwind, it is just a number in the range [1,4]
- SQL code fixed so that it runs without errors.
- In the Employee table, we changed the length for the Country attribute from nvarchar(10) to nvarchar(50)
  Reason: This is the length stated in the AdventureWorks source database
- For the Employee table, the length of the Postal code changed from nvarchar(50) to nvarchar(15)
  Reason: That is the length stated in the AdventureWorks source database

Source target mappings

All changes mentioned above were changed in the source target mapping.

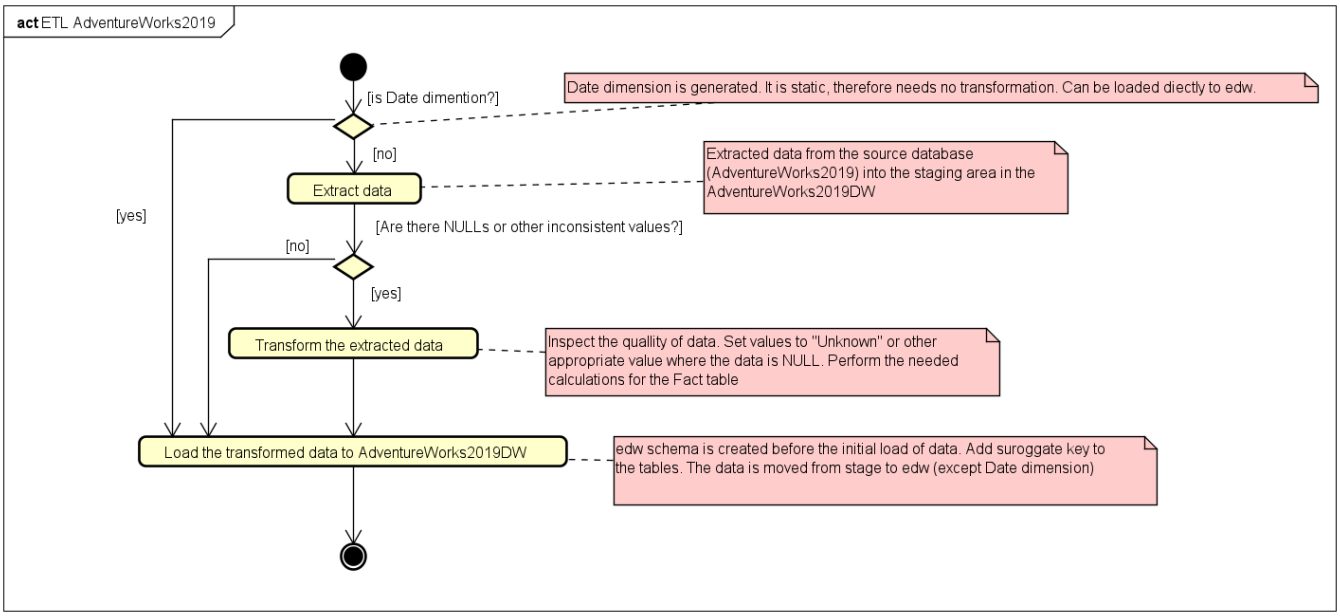
For Location, Customer table some attributes were deleted due to little memory to run the query (because of way too many duplicates).

General

Corrected typos

ETL

Before starting the ETL process we took the decision to update the implementation of the design in order to easier follow the process:
- Deleted the surrogate keys
- Deleted the date dimension table in the stage schema



Extract, transform and load the dimensional data

The first step is to extract the information from the source database to the stage schema in the DW. The stage tables allow NULL values, as the data will be inspected and transformed. The stage table also contains no primary or foreign keys. The Surrogate key is not part of the tables as well.

The next step is the inspection of the quality of data and transforming it where needed. For instance, in Dim_Employee if the name of the employee is NULL, the information is transformed to “UNKNOWN”. If a date is unknown, for instance, the discontinued date for a product, it is changed to “2100-01-01 00:00:00.000”. The transformation steps included also combining FirstName, LastName of Employees and Customers and deciding whether the CustomerName should be an individual's name or store name, depending on the value of CustomerType. Some tables did not need any transformations, for example, the Dim_Location table.

The final step is to create the schema “edw” and the tables there. Those tables do not allow NULLs and have all the necessary keys. The Surrogate keys are generated when a new line is inserted to the table.

Creation of the Date dimension table

The Date dimension table was created based on the previous design with some additional changes:

- Added Date field, containing the date in datetime data type.

- Added WeekdayName and Week. The reasoning is that this way the sales manager can have the data visualized better and in whatever time period she wishes.

Since Date Dimension is generated once and is static, there was no need to create it in the stage period and no transformation was needed. The date was loaded directly in the edw schema.

Extract, transform and load the facts

The first step is to extract from source db to the stage schema in DW. The stage table allows NULL values.

Next step the data was inspected and transformed. For example, the EmployeeId was NULL and it was transformed to -1.

In the final step, the edw Fact Sale table was created and the data loaded:

For that, we looked up the surrogate keys that were created by inserting records and generating an int identity surrogate key for all the dimension tables

By selecting the statements and writing joints to all the dimension tables we ensured that the correct surrogate keys were gotten

Lastly, the data was loaded in the edw fact sales table

Documentation

See appendix **A** for SQL code
See appendix **B** for revised source-target mappings