Group Members:

- XXX XXX
- XXX XXX
- Répási Gergely
- XXX XXX
- XXX XXX

DATA WAREHOUSES & BUSINESS INTELLIGENCE GROUP PROJECT

Dokumentáció 2022.10.31.

TABLE OF CONTENTS

USE CASE

INTRODUCTORY

We would like to learn about information related to purchases, the relationship between merchants and products from different perspectives and over time, and the relationship between products and customers. In addition, we will cover basic data related to the delivery and return of individual products, which may have an impact on future sales.

QUESTIONS ASKED BY THE CUSTOMER:

1. How did the sales of each product group develop by season?
2. Which product subcategory received the most orders per active merchant ?
3. What was the turnover of each retailer between 2012 and 2014?
4. What was the average lead time (time from order placement to product arrival) per product, broken down by years?
5. How many orders and what value were placed per season and per delivery method in the last 3 years?
6. What was the percentage of rejected products between 2012 and 2014, broken down by product?

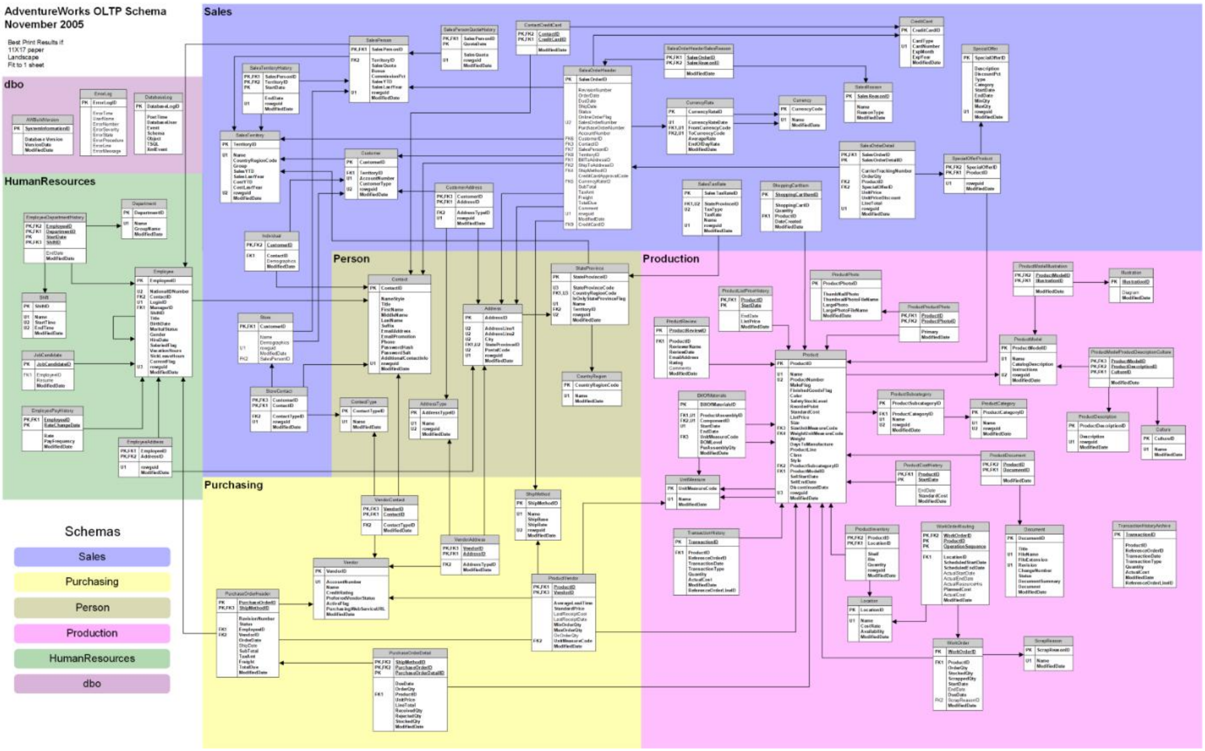
DATABASE LAYERS

SOURCE DATA

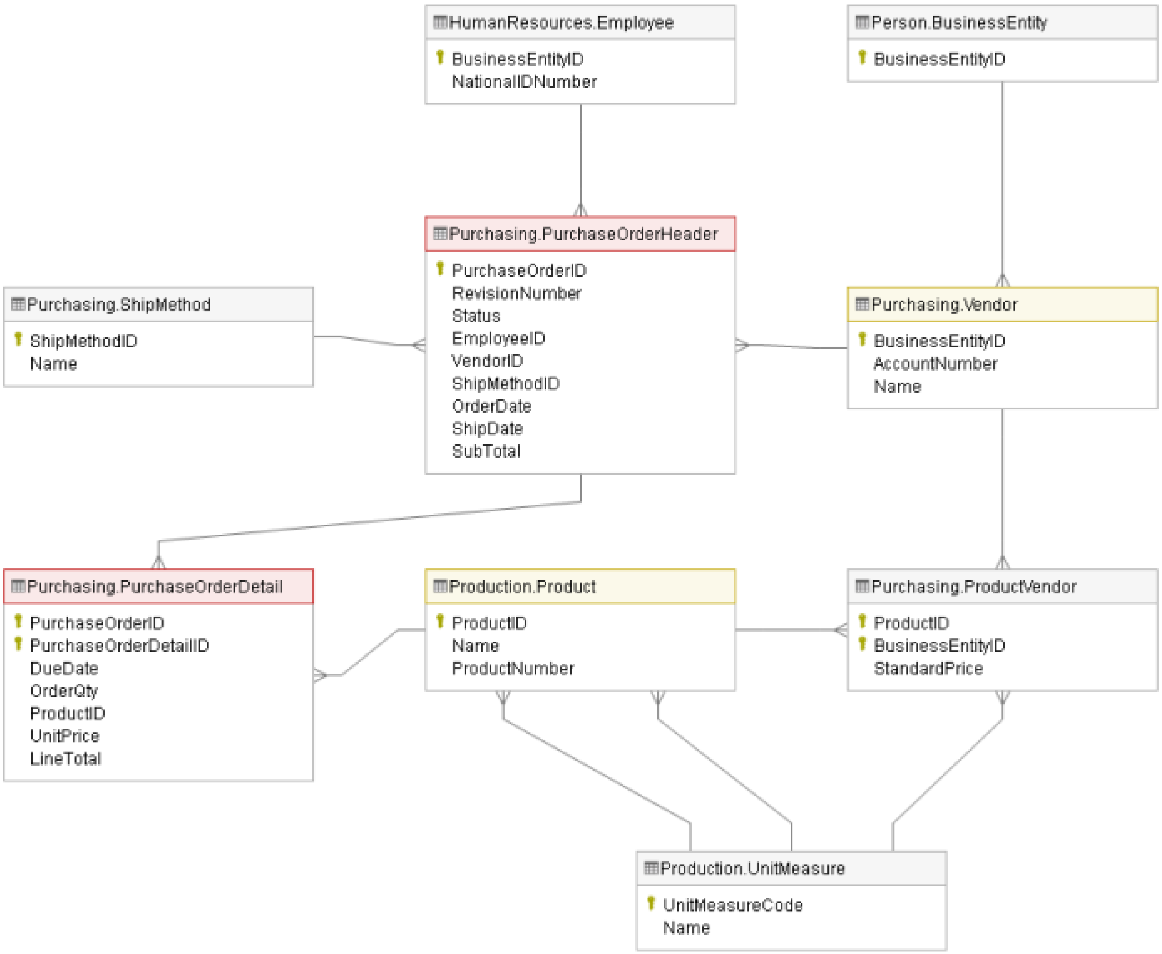
The AdventureWorks database as the source of the data.

ADVENTUREWORKS SCHEMA

https://i0.wp.com/improveandrepeat.com/wp-content/uploads/2018/12/AdvWorksOLTPSchemaVisio.png

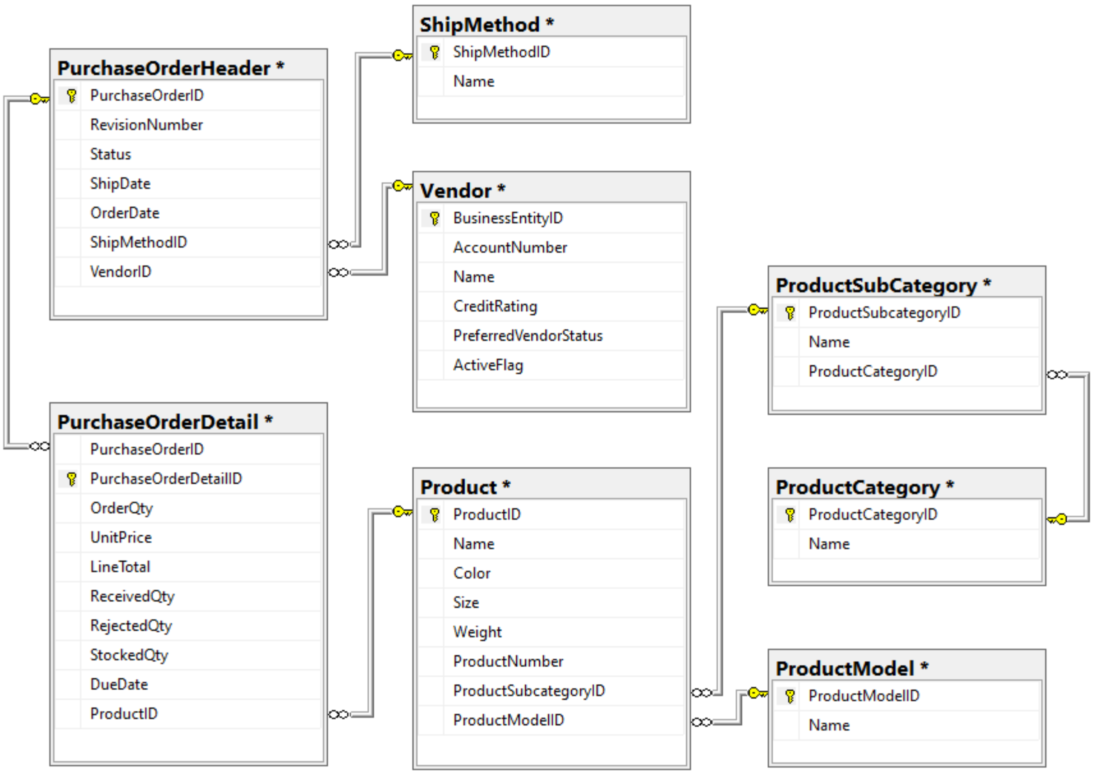


7. Purchasing



STAGING:

The source database tables are unchanged in structure, except that all data types are VARCHAR.



PURCHASEORDERHEADER

- PurchaseOrderID – Unique identifier of the order
- RevisionNumber - Number used to track changes to an order
- Status - Order status
- ShipDate - Estimated delivery time
- OrderDate - Order date
- ShipMethodID - Shipping method
- VendorID - Seller

PURCHASEORDERDETAIL

- PurchaseOrderID - Order ID
- PurchaseOrderDetailID - Order ID
- OrderQty – Ordered quantity
- UnitPrice – Selling unit price
- LineTotal - Total price per product
- ReceivedQty - Received quantity
- RejectedQty - Rejected quantity
- StockedQty - Stocked quantity
- DueDate - Expected time of arrival
- ProductID – Product ID

PRODUCT

- ProductID - Product ID
- Name - Product name
- Color - Product color
- Size - Product size
- Weight - Product weight
- ProductNumber - Product number
- ProductSubcategoryID - Product subcategory
- ProductModelID - Product model number

PRODUCTSUBCATEGORY

- ProductSubcategoryID - Product subcategory unique identifier
- Name - Product subcategory name
- ProductCategoryID - Product category foreign key

PRODUCT CATEGORY

- ProductCategoryID - Product category unique identifier
- Name - Product category

SHIPMETHOD

- ShipMethodID - Unique identifier for shipping method
- Name - Shipping company name

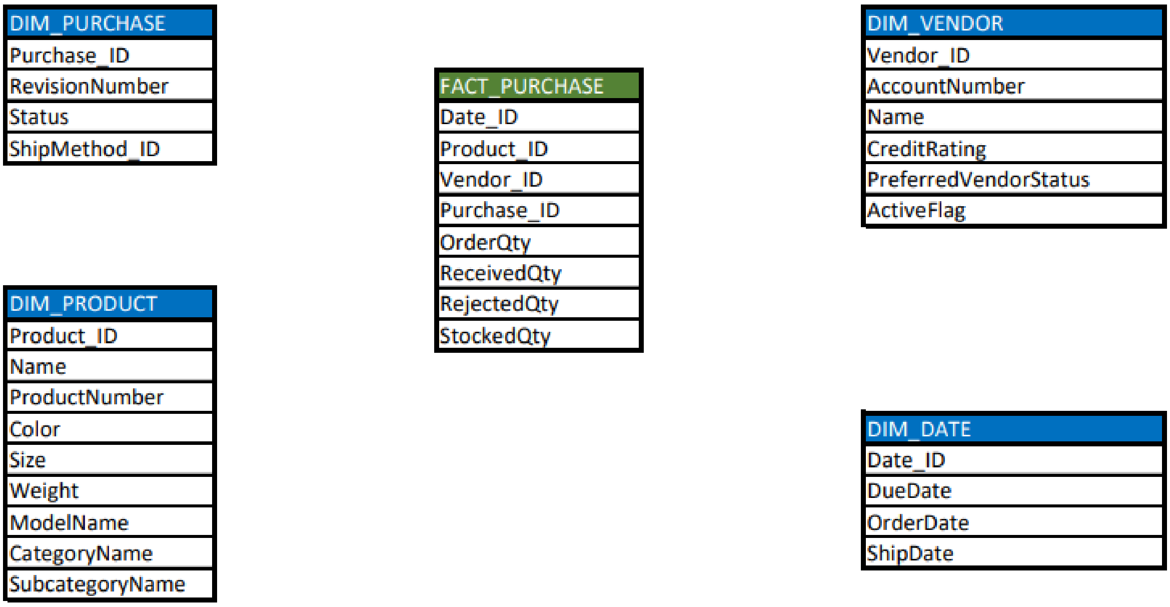
VENDOR

- BusinessEntityID - Seller's unique identifier
- AccountNumber - Seller user account number
- Name - Seller name
- CreditRating - Seller rating between 1 and 5
- PreferredVendorStatus - Among multiple vendors, if 1 then we use him, if 0 then we prefer to look for another
- ActiveFlag – 0 if the seller is not active and 1 if the seller is active (default: 1)

DATA WAREHOUSE:

It is structurally identical to the tables created in the Staging layer. The difference is that the data types are appropriate for the data, and each table has been supplemented with two date type columns: ValidFrom and ValidTo .

DIMENSIONAL MODEL:



DIM_PURCHASE
Purchase_ID
RevisionNumber
Status
ShipMethod_ID

FACT_PURCHASE
Date_ID
Product_ID
Vendor_ID
Purchase_ID
OrderQty
ReceivedQty
RejectedQty
StockedQty

DIM_VENDOR
Vendor_ID
AccountNumber
Name
CreditRating
PreferredVendorStatus
ActiveFlag

DIM_PRODUCT
Product_ID
Name
ProductNumber
Color
Size
Weight
ModelName
CategoryName
SubcategoryName

DIM_DATE
Date_ID
DueDate
OrderDate
ShipDate

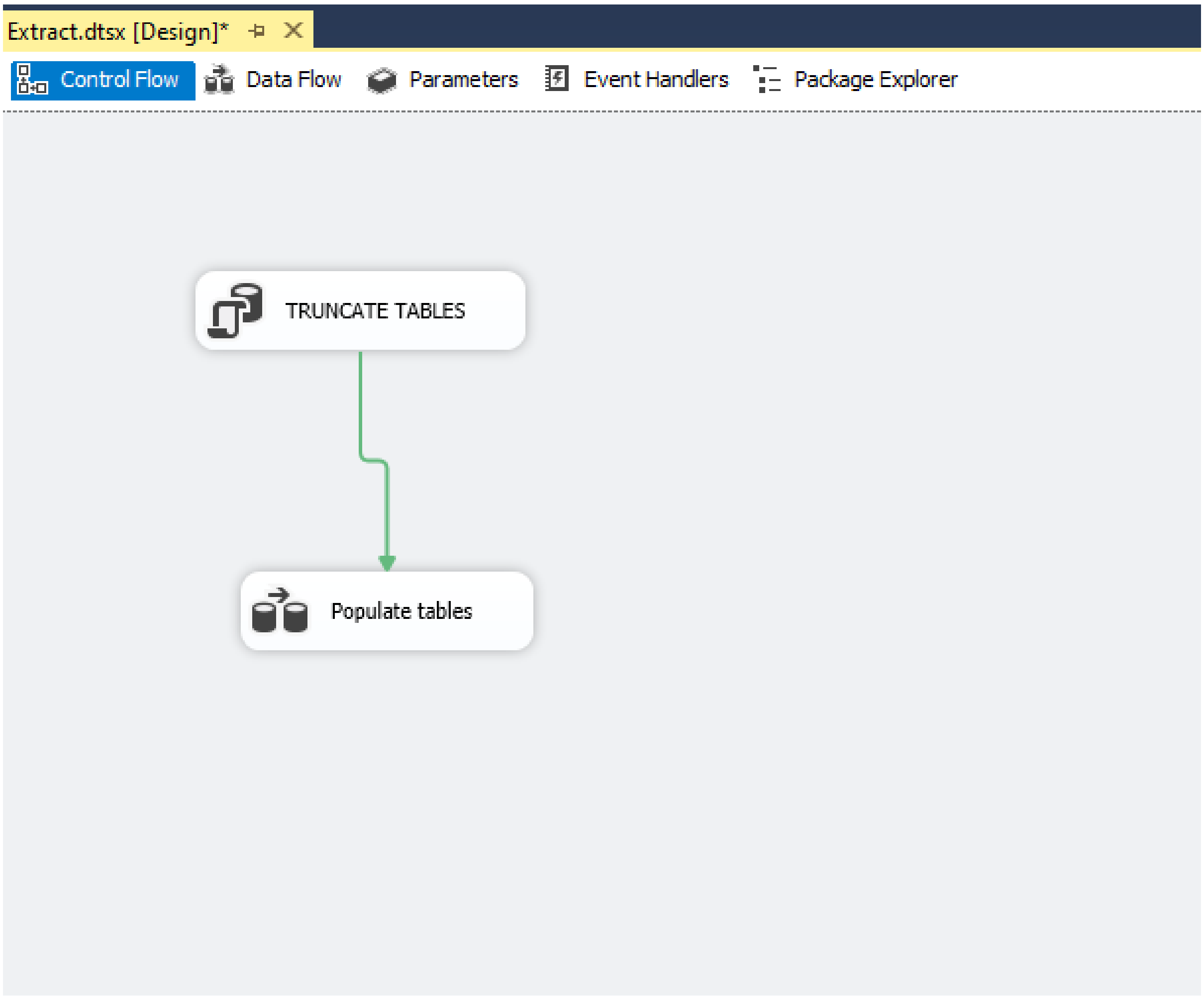
BOARDS

- FACT_PURCHASE: Purchase data
- DIM_PURCHASE: Purchase dimension
- DIM_PRODUCT: Purchased products
- DIM_VENDOR: Supplier data
- DIM_DATE: Date dimension

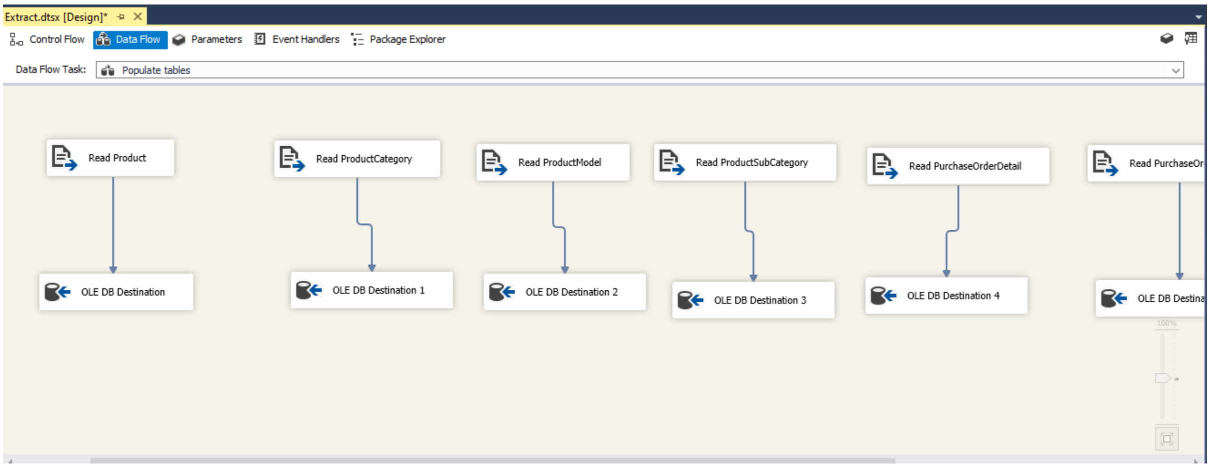
THIS PROCESS

EXTRACTION PROCESS

The first part of the ETL process is Extract , in which the Staging layer is populated with data. All data is imported with the VARCHAR(256) data type. No data conversion is performed in this phase, the goal is to achieve the fastest possible transfer.

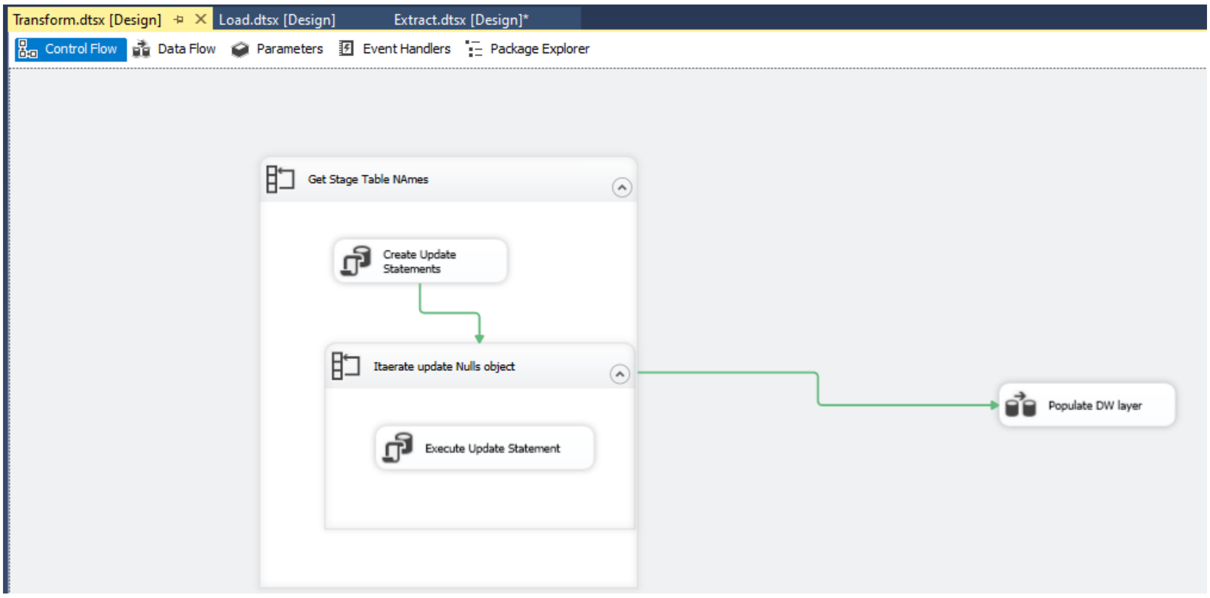


The Extract part was implemented using an Execute SQL Task and a Data Flow Task . The Execute SQL Task is responsible for emptying the Staging layer tables, and after this is successfully completed, the data is loaded from the CSV files.

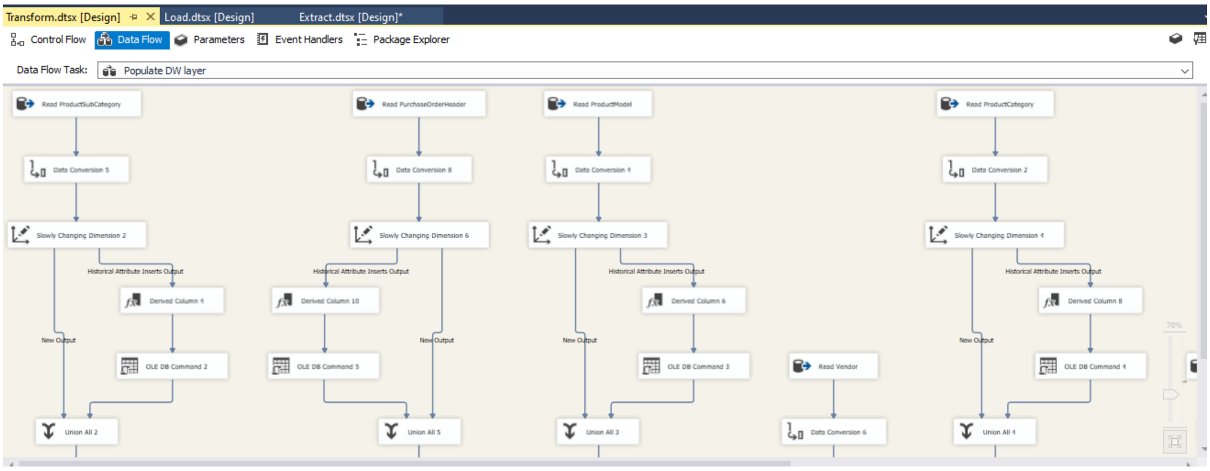


TRANSFORMATION PROCESS

the Transform layer is to convert the Staging content to the appropriate data type and transfer this data to the DW layer.



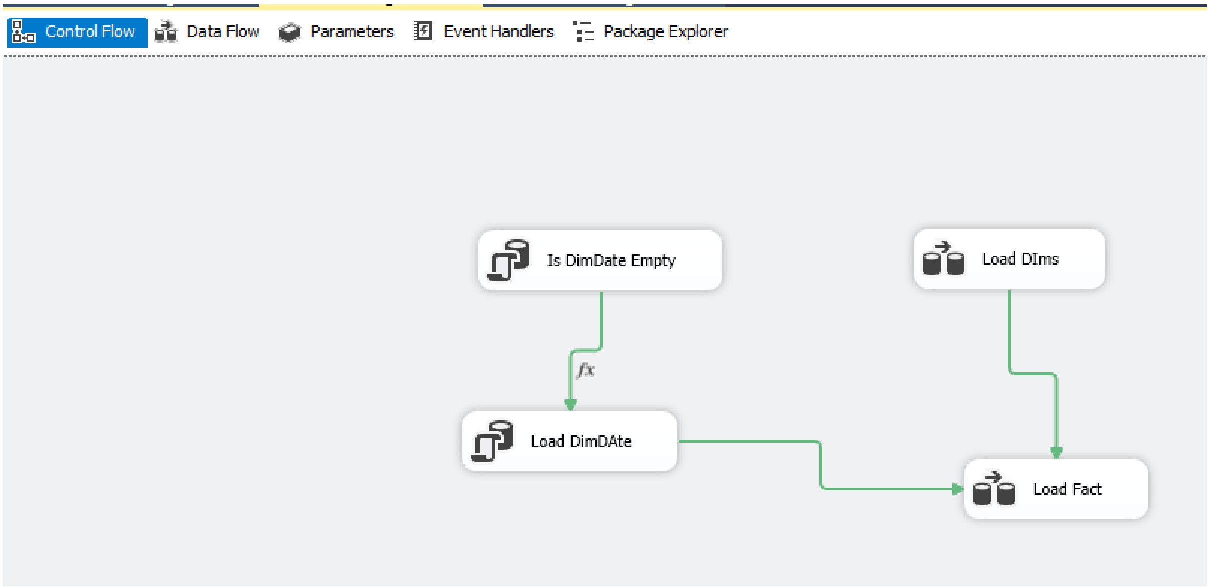
To implement it, you need a Foreach Loop Container , which reads the table names from the Staging layer. To properly handle null values, an Execute SQL task , Foreach , is required within the loop. Loop and an Execute SQL Task that converts VARCHAR NULL values to actual values. After this has successfully run, we can load the data into the DW layer using a Data Flow Task .


Data Flow Task: Populate DW layer

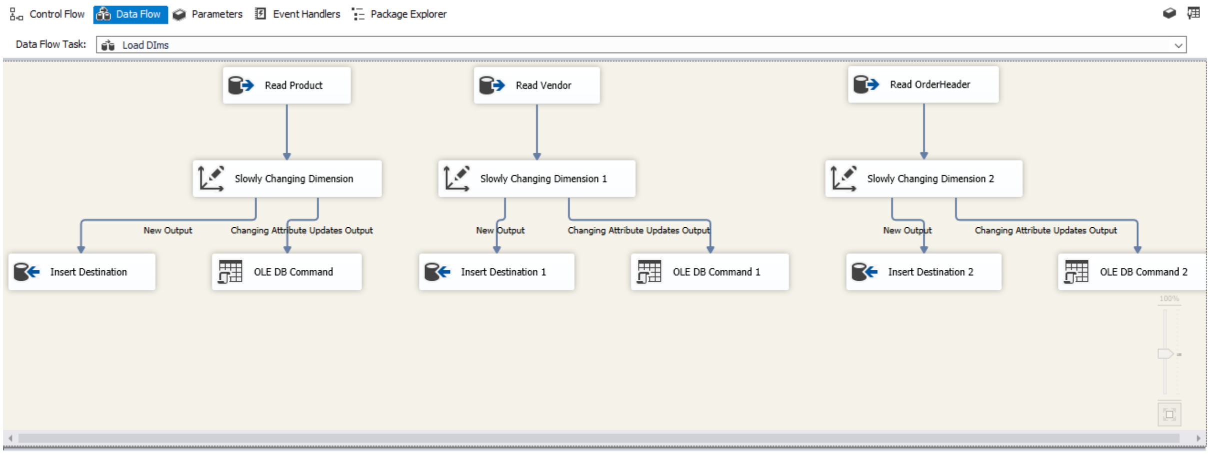
The data is read from the Staging layer and then converted to the appropriate type using the Data Conversion task . To avoid duplicate data in the DW layer, we load the data using Slowly Changing Dimension.

LOADING PROCESS

Load layer is used to transfer data from the DW layer to the DM , here we use the previously defined data types, and we create a dimension, date and fact table corresponding to the star schema. In the first step, we check whether the date table contains values. If not, we continue the process by filling it. The last two steps are solved with an Execute SQL Task each. The first counts how many pieces of data there are in the Dimension table, if this returns a zero value, the second Execute SQL Task is run .



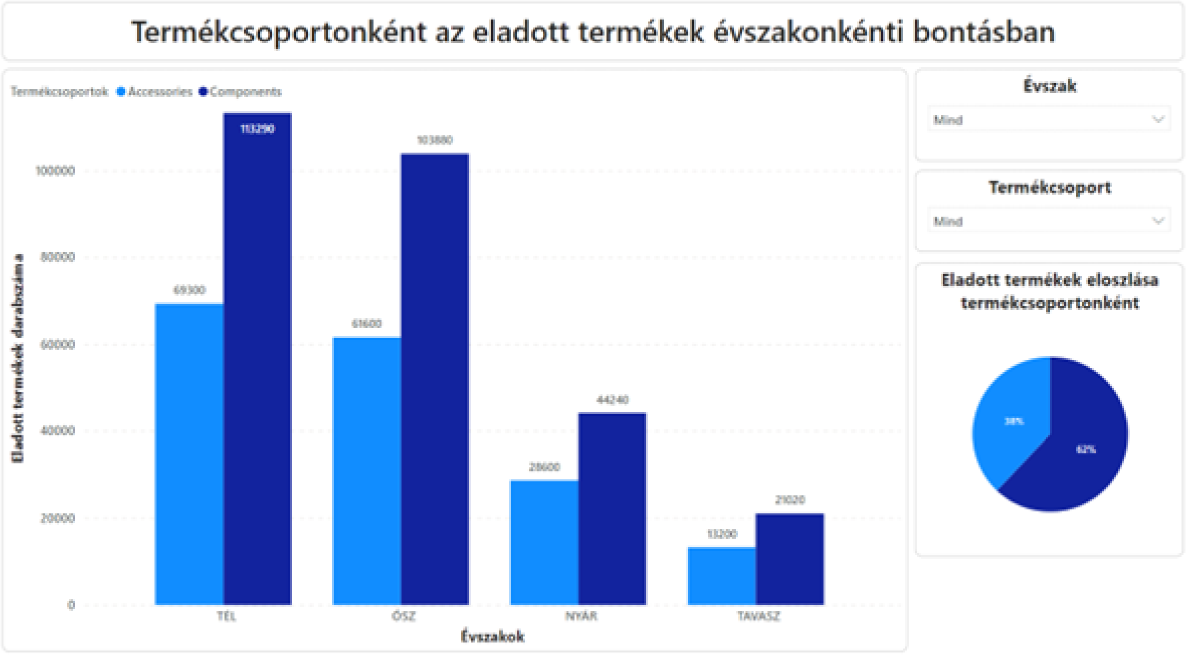
To populate the dimension tables, we use views created in the DW layer, from which we read the data into the dimension tables. To populate the tables, we also use Slowly Changing We use Dimension to avoid duplicate data.



After successfully loading the dimension tables, we can also load the fact table, since this is where we use the data from each dimension. We can connect the tables using Lookup and load them into the FactPurchase table using the SCD we have already used several times .

REPORT

1. HOW DID THE SALES OF EACH PRODUCT GROUP DEVELOP BY SEASON?



Termékcsoportonként az eladott termékek évszakonkénti bontásban

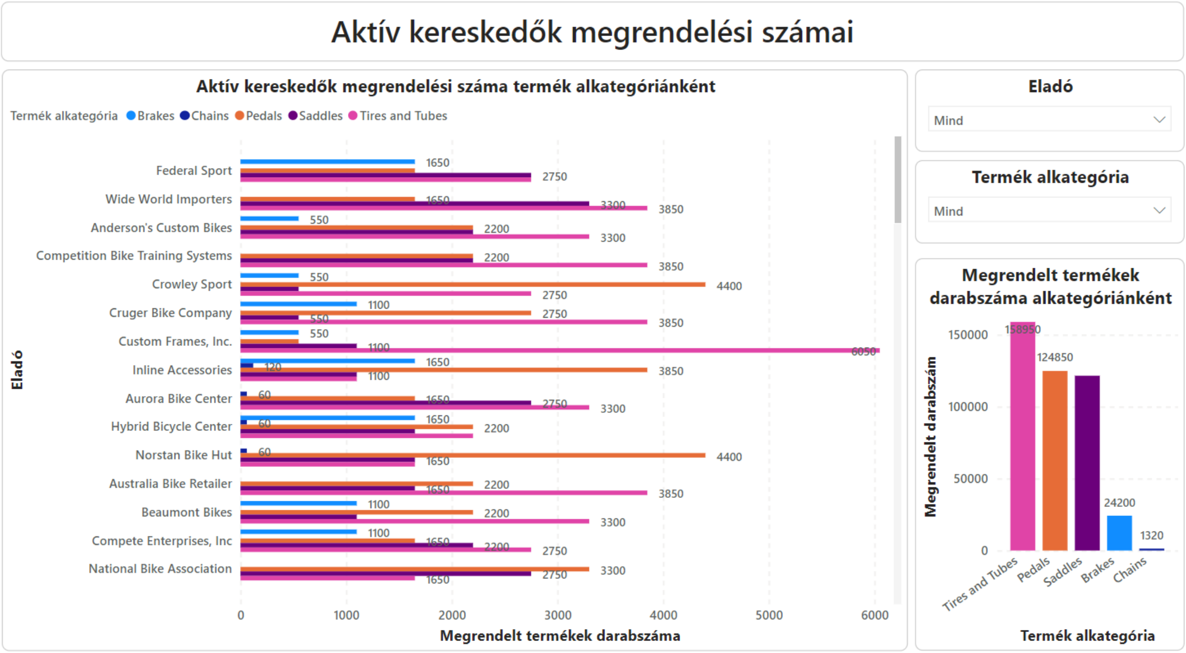
Évszak

Mind

Termékcsoport

Mind

2. WHICH PRODUCT SUBCATEGORY RECEIVED THE MOST ORDERS PER ACTIVE MERCHANT?



Aktív kereskedők megrendelési számai

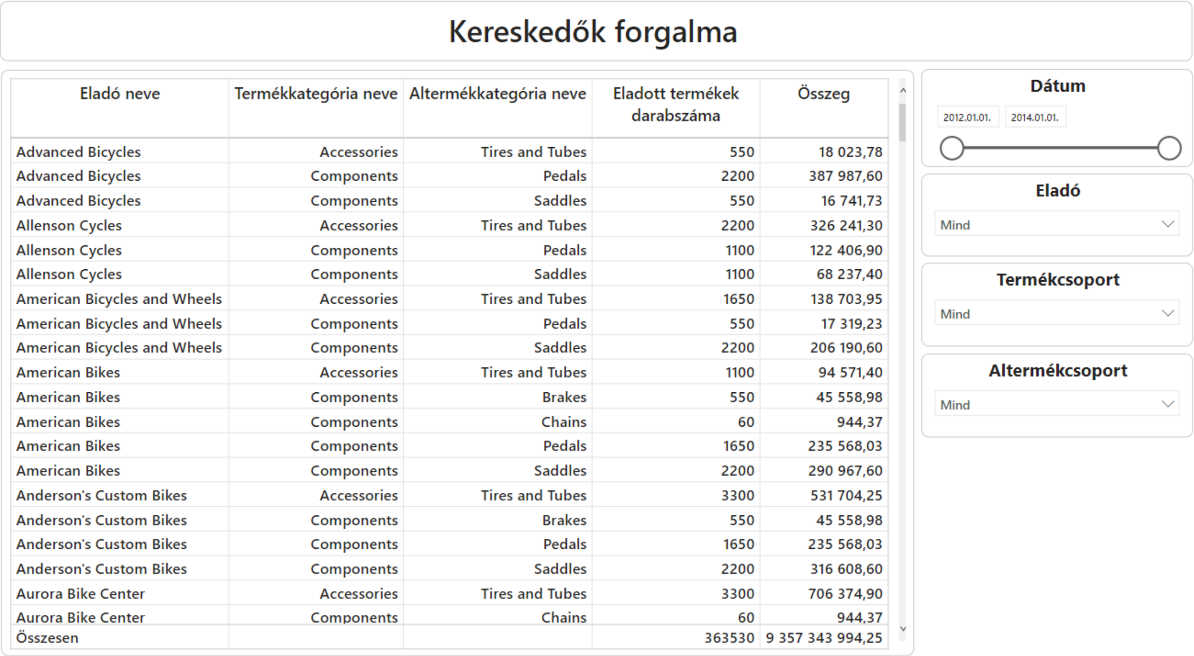
Eladó

Mind

Termék alkategória

Mind

3. WHAT WAS THE TURNOVER OF EACH TRADER BETWEEN 2012 AND 2014?



Kereskedők forgalma

Eladó neve	Termékkategória neve	Altermékkategória neve	Eladott termékek darabszáma	Összeg
Advanced Bicycles	Accessories	Tires and Tubes	550	18 023,78
Advanced Bicycles	Components	Pedals	2200	387 987,60
Advanced Bicycles	Components	Saddles	550	16 741,73
Allenson Cycles	Accessories	Tires and Tubes	2200	326 241,30
Allenson Cycles	Components	Pedals	1100	122 406,90
Allenson Cycles	Components	Saddles	1100	68 237,40
American Bicycles and Wheels	Accessories	Tires and Tubes	1650	138 703,95
American Bicycles and Wheels	Components	Pedals	550	17 319,23
American Bicycles and Wheels	Components	Saddles	2200	206 190,60
American Bikes	Accessories	Tires and Tubes	1100	94 571,40
American Bikes	Components	Brakes	550	45 558,98
American Bikes	Components	Chains	60	944,37
American Bikes	Components	Pedals	1650	235 568,03
American Bikes	Components	Saddles	2200	290 967,60
Anderson's Custom Bikes	Accessories	Tires and Tubes	3300	531 704,25
Anderson's Custom Bikes	Components	Brakes	550	45 558,98
Anderson's Custom Bikes	Components	Pedals	1650	235 568,03
Anderson's Custom Bikes	Components	Saddles	2200	316 608,60
Aurora Bike Center	Accessories	Tires and Tubes	3300	706 374,90
Aurora Bike Center	Components	Chains	60	944,37
Összesen			363530	9 357 343 994,25

Dátum

2012.01.01. 2014.01.01.

Eladó

Mind

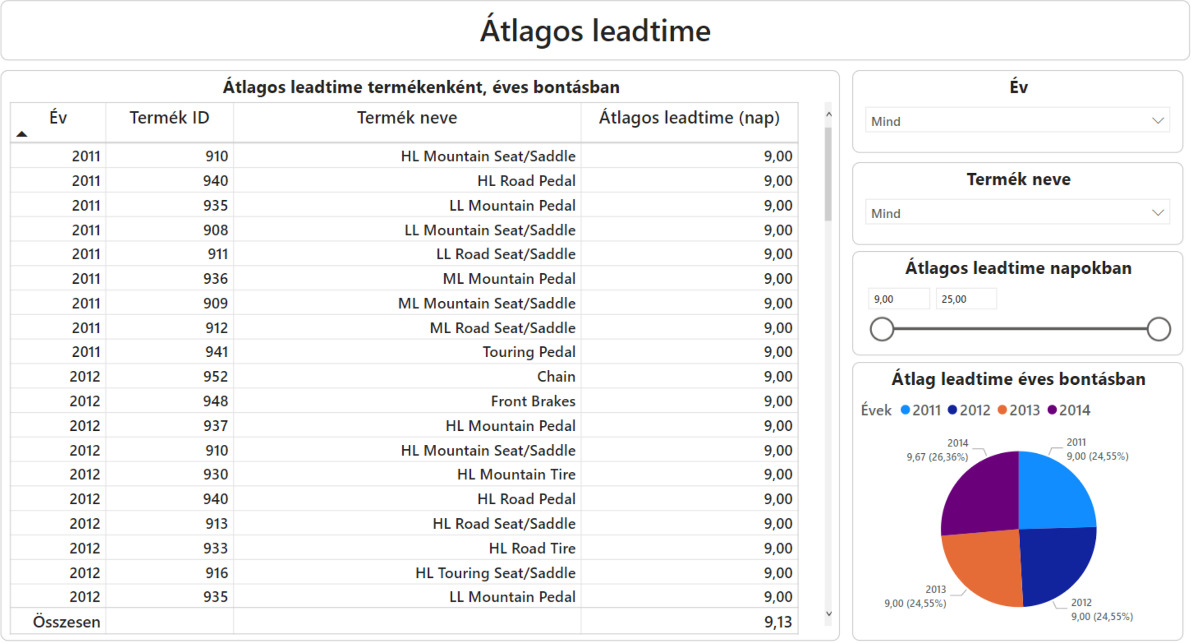
Termékcsoport

Mind

Altermékcsoport

Mind

4. WHAT WAS THE AVERAGE LEAD TIME (TIME FROM ORDER PLACEMENT TO PRODUCT ARRIVAL) PER PRODUCT, BROKEN DOWN BY YEARS?



Átlagos leadtime

Átlagos leadtime termékenként, éves bontásban

Év	Termék ID	Termék neve	Átlagos leadtime (nap)
2011	910	HL Mountain Seat/Saddle	9,00
2011	940	HL Road Pedal	9,00
2011	935	LL Mountain Pedal	9,00
2011	908	LL Mountain Seat/Saddle	9,00
2011	911	LL Road Seat/Saddle	9,00
2011	936	ML Mountain Pedal	9,00
2011	909	ML Mountain Seat/Saddle	9,00
2011	912	ML Road Seat/Saddle	9,00
2011	941	Touring Pedal	9,00
2012	952	Chain	9,00
2012	948	Front Brakes	9,00
2012	937	HL Mountain Pedal	9,00
2012	910	HL Mountain Seat/Saddle	9,00
2012	930	HL Mountain Tire	9,00
2012	940	HL Road Pedal	9,00
2012	913	HL Road Seat/Saddle	9,00
2012	933	HL Road Tire	9,00
2012	916	HL Touring Seat/Saddle	9,00
2012	935	LL Mountain Pedal	9,00
Összesen			9,13

Év

Mind

Termék neve

Mind

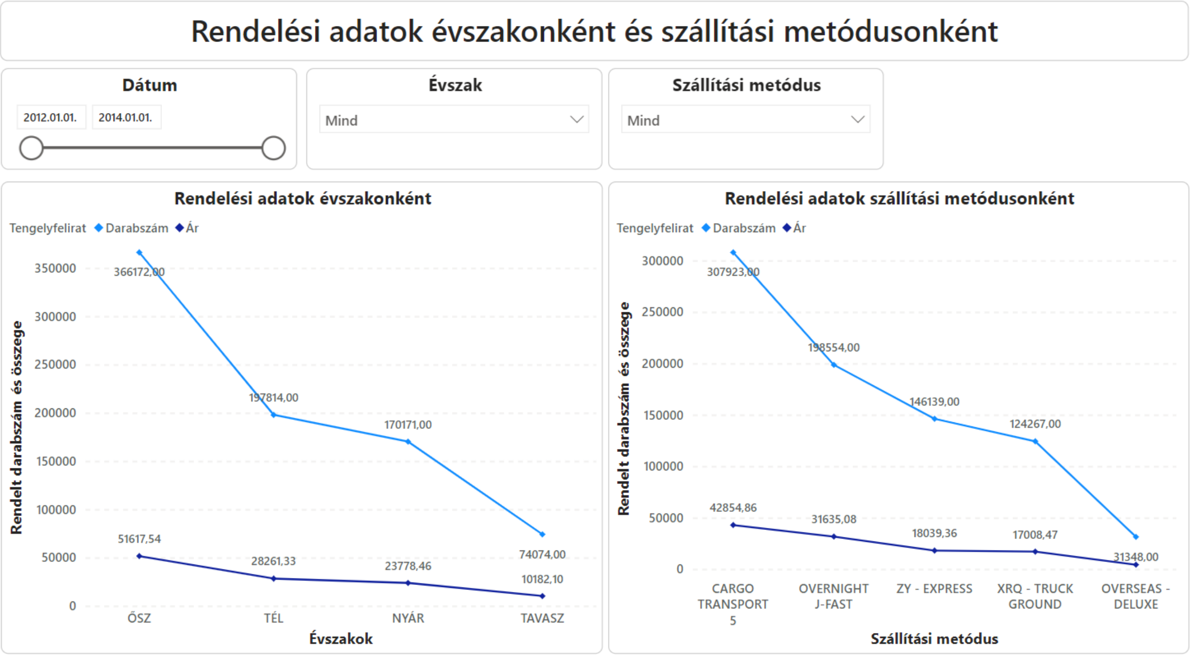
Átlagos leadtime napokban

9,00 25,00

Átlag leadtime éves bontásban

Évek ● 2011 ● 2012 ● 2013 ● 2014

5. HOW MANY ORDERS AND WHAT IS THEIR VALUE PER SEASON AND DELIVERY METHOD IN THE LAST 3 YEARS?



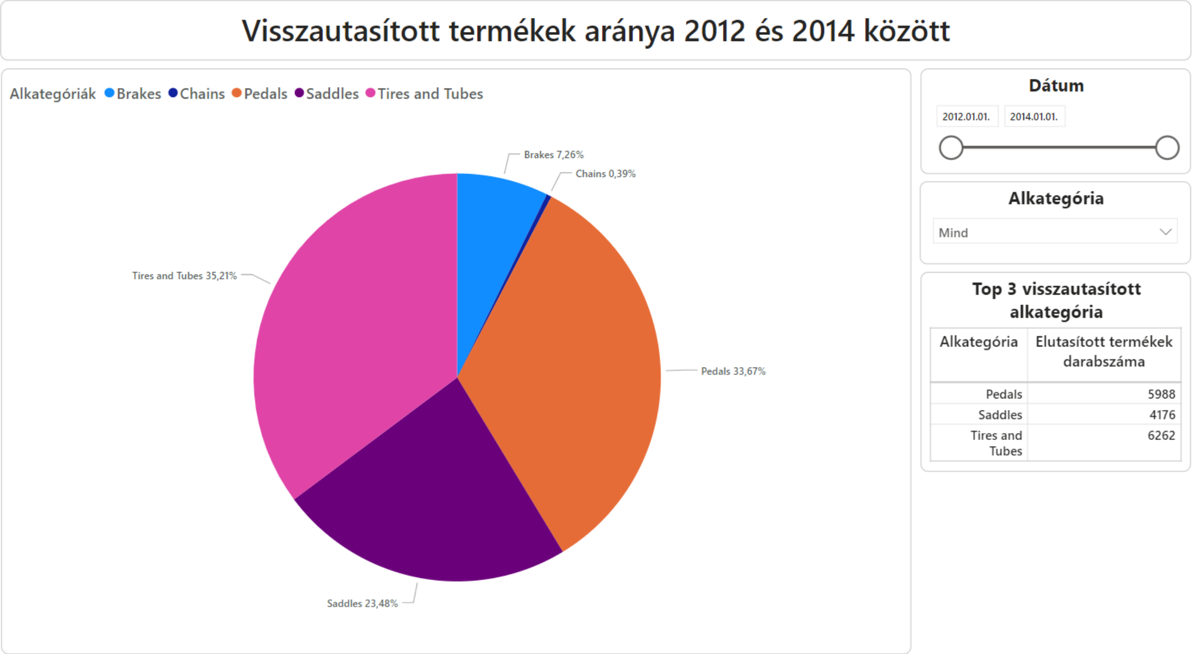
Rendelési adatok évszakonként és szállítási metódusonként

Dátum: 2012.01.01. – 2014.01.01.

Évszak: Mind

Szállítási metódus: Mind

6. WHAT WAS THE PERCENTAGE OF REJECTED PRODUCTS BETWEEN 2012 AND 2014, BROKEN DOWN BY PRODUCT?



Visszautasított termékek aránya 2012 és 2014 között

Dátum: 2012.01.01. – 2014.01.01.

Alkategória: Mind

Top 3 visszautasított alkategória

Alkategória	Elutasított termékek darabszáma
Pedals	5988
Saddles	4176
Tires and Tubes	6262

SCRUM DOCUMENTATION

TABLE OF CONTENTS

ROLES:

- XXX XXX – Data Warehouse Designer
- XXX XXX – SQL Developer
- XXX XXX – ETL Device Manager
- XXX XXX – Report Developer
- Gergely Répási – Scrum master, product owner

PRODUCT BACKLOG:

1. use on a subset of the data case formulation, which can be solved by building a data warehouse. Use Presentation of questions to be answered during the case .
1. The use specification of a data model suitable for the case, definition and presentation of the necessary dimensional model (fact table and dimension tables)
2. Creating the database for the 3 ETL layers
3. Preparing and documenting data loading procedures (ETL)
4. Report development, visualization and documentation (using Power BI)

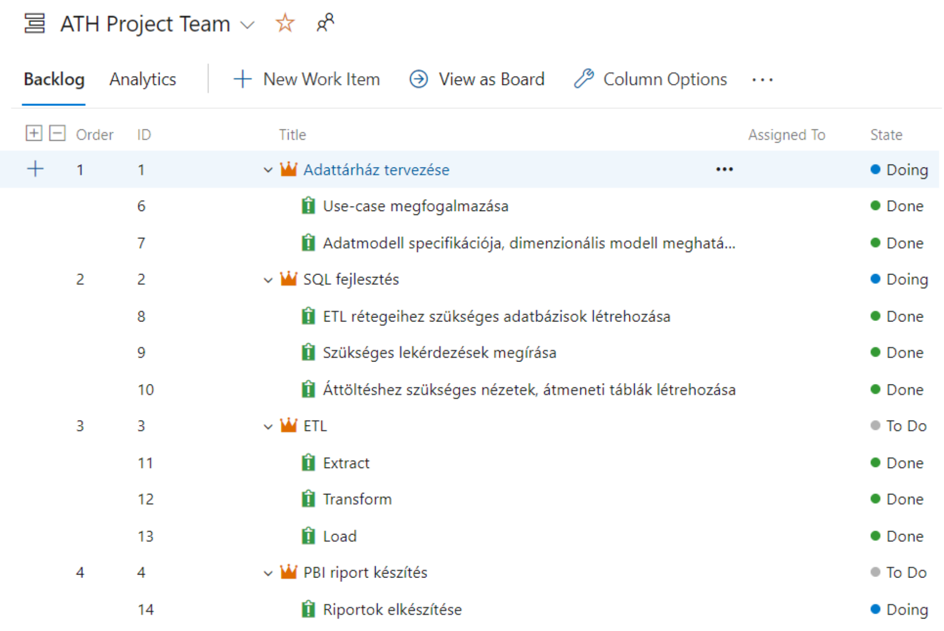
SPRINT PLANNING MEETING REPORT:

we would like to use some kind of project management tool during development to continuously track the status of the project, and then we jointly decided that this would be Azure. Let it be DevOps . We chose this because we found it easy to use and transparent, and also because there was someone on the team who was already familiar with this tool, so it was much easier to learn how to use it.

the tasks in the Product Backlog and divided them into smaller subtasks, which were then included in the Sprint Backlog .

SPRINT BACKLOG:

The image below shows the different Tasks of the Product Backlog. allocated per developer .



Order	ID	Title	Assigned To	State
1	1	Adattárház tervezése		● Doing
	6	Use-case megfogalmazása		● Done
	7	Adatmodell specifikációja, dimenzionális modell meghatá...		● Done
2	2	SQL fejlesztés		● Doing
	8	ETL rétegeihez szükséges adatbázisok létrehozása		● Done
	9	Szükséges lekérdezések megírása		● Done
	10	Áttöltéshez szükséges nézetek, átmeneti táblák létrehozása		● Done
3	3	ETL		● To Do
	11	Extract		● Done
	12	Transform		● Done
	13	Load		● Done
4	4	PBI riport készítés		● To Do
	14	Riportok elkészítése		● Doing

DAILY SCRUM REPORT:

WEEK 1 :

- ATH designer:
  - He figured out the use - case and wrote the necessary questions, 6 in total. Based on the questions, he defined the individual tables and fields of the data model.
  - You will create the fact and dimension tables of the dimensional model with each field.
  - a dimensional model is not entirely clear.
- SQL developer:
  - You will create the Staging and Core layers.

WEEK 2:

- ATH designer:
  - The dimensional model is complete.
- SQL developer:
  - Staging and Core layer.
  - In the next step, you will create a Data Mart layer and the queries and views.
- ETL developer:
  - It will prepare the Extract and Load processes.

WEEK 3:

- SQL developer:
  - There is also layer 3, or queries and views.
  - Converts NVARCHAR fields to VARCHAR .
- ETL developer:
  - There were issues with Unicode character encoding, which prevented the Transform processes from running. The NVARCHAR data type caused the error.
  - It will make all three layers.

WEEK 4:

- SQL developer:
  - Fixed data types in all layers.
- ETL developer:
  - All layers of the ETL are completed and running flawlessly.
- Report developer:
  - It will create reports based on the data you enter.

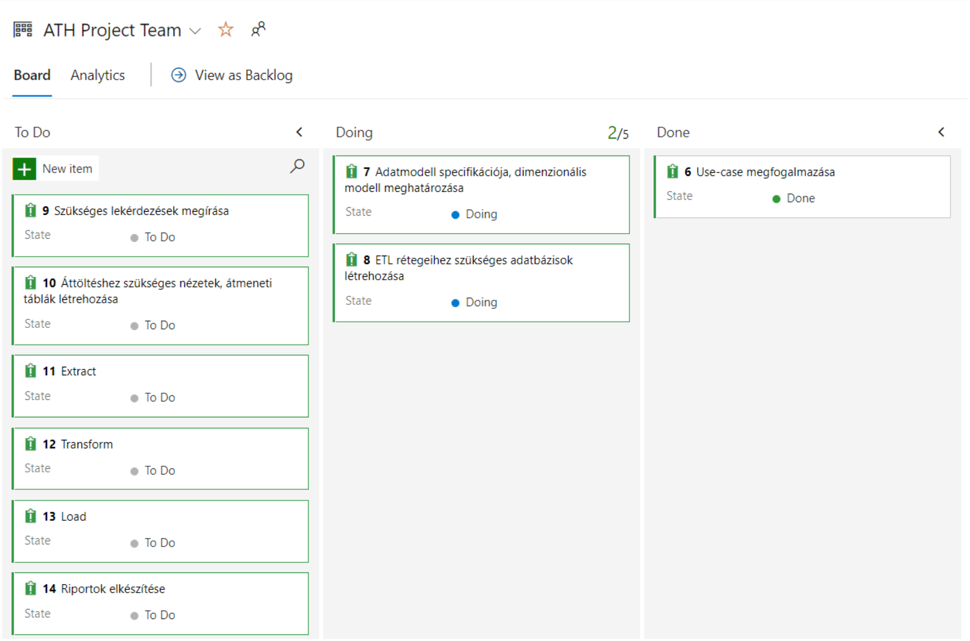
WEEK 5:

- Report developer:
  - The reports have been successfully completed.

SCRUM BOARD:

some of the main phases of the workflow on the Scrumboard.

Roughly the first week's status:



To Do

- 9 Szükséges lekérdezések megírása — State: To Do
- 10 Áttöltéshez szükséges nézetek, átmeneti táblák létrehozása — State: To Do
- 11 Extract — State: To Do
- 12 Transform — State: To Do
- 13 Load — State: To Do
- 14 Riportok elkészítése — State: To Do

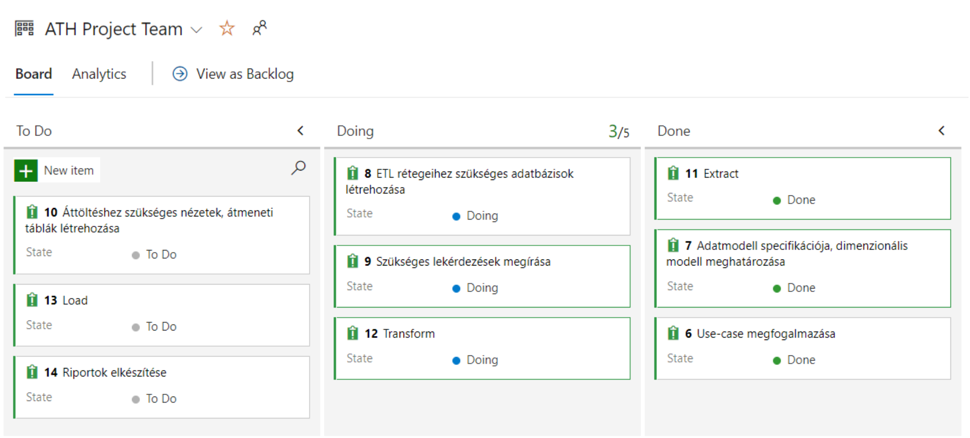
Doing 2/5

- 7 Adatmodell specifikációja, dimenzionális modell meghatározása — State: Doing
- 8 ETL rétegeihez szükséges adatbázisok létrehozása — State: Doing

Done

- 6 Use-case megfogalmazása — State: Done

Approaching week 3, we were about halfway through the Sprint Backlog tasks.



To Do

- 10 Áttöltéshez szükséges nézetek, átmeneti táblák létrehozása — State: To Do
- 13 Load — State: To Do
- 14 Riportok elkészítése — State: To Do

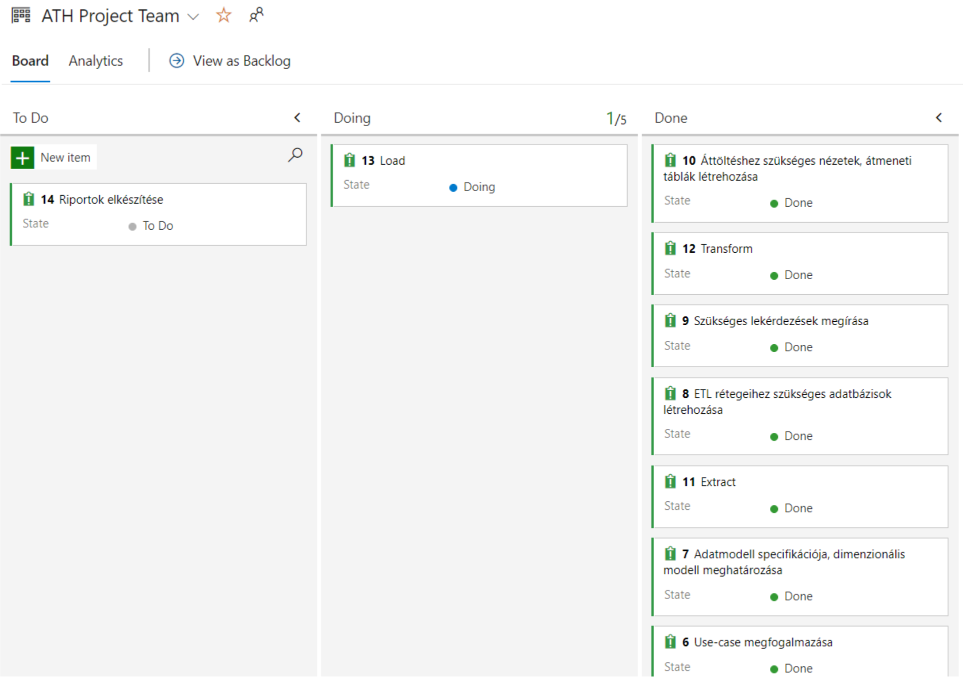
Doing 3/5

- 8 ETL rétegeihez szükséges adatbázisok létrehozása — State: Doing
- 9 Szükséges lekérdezések megírása — State: Doing
- 12 Transform — State: Doing

Done

- 11 Extract — State: Done
- 7 Adatmodell specifikációja, dimenzionális modell meghatározása — State: Done
- 6 Use-case megfogalmazása — State: Done

Towards the end of week 4, there were still a few tasks left, so it took almost 5 weeks to complete the project.



ATH Project Team

To Do

14 Riportok elkészítése
State ● To Do

Doing 1/5

13 Load
State ● Doing

Done

10 Áttöltéshez szükséges nézetek, átmeneti táblák létrehozása
State ● Done

12 Transform
State ● Done

9 Szükséges lekérdezések megírása
State ● Done

8 ETL rétegeihez szükséges adatbázisok létrehozása
State ● Done

11 Extract
State ● Done

7 Adatmodell specifikációja, dimenzionális modell meghatározása
State ● Done

6 Use-case megfogalmazása
State ● Done

SPRINT RETROSPECT MEETING REPORT:

It is recommended to use plain VARCHAR from the beginning instead of NVARCHAR data type because it causes problems when loading data.

When saving data to a CSV file, the decimal point used by the database may mess up the data, so it is recommended to use a different delimiter , e.g. semicolon, tab.

During an ETL process, it is only worth moving on to the next task if the current one is working properly and has been verified. If we do not maintain this and skip a faulty task, many retrospective fixes may be required.