

# Adattárházak és üzleti intelligencia csoportmunka

2020/21 őszi félév

## Készítők:

Nagy Dávid

Scrum master, product owner

Lukács Andrea

Adattárház tervező

Balogh Bence

SQL fejlesztő

Bózsó Bence

ETL eszköz kezelő

Sallai András

Riportkészítő

# Tartalomjegyzék

<b>Feladatkiírás, Use case .....</b>	<b>3</b>
<b>Forrásadatok .....</b>	<b>4</b>
<b>Dimenzionális modell .....</b>	<b>5</b>
<b>3 ETL réteg adatbázisa: .....</b>	<b>7</b>
STAGE.....	7
Core / DW .....	7
View / DM .....	8
<b>ETL folyamat: .....</b>	<b>9</b>
EXTRACT .....	9
TRANSFORM.....	11
LOAD .....	16
<b>Riportkészítés .....</b>	<b>19</b>
<b>Scrum dokumentáció.....</b>	<b>23</b>

## Feladatkiírás, Use case

Az AdventureWorks nevű vállalat azügyben keresett fel minket, hogy hozzunk létre egy adattárházat az üzleteik produktivitásának maximalizálására. Közös egyeztetések sorozata után arra a közös megállapodásra jutottunk, hogy a cél eléréséhez fontos adatok a következők lesznek:

Elsősorban a dolgozók és produktivitásuk az, ami meghatározza az üzletek és a vállalat egészének sikerességét, ezekből kifolyólag a dolgozók és üzletek összes adatára szükségünk lesz.

A másik nagyon fontos szempont az adott eladások és azok részletei. Tehát az eddigi adatokból kiindulva meg lehet jósolni, hogy mely szállítási módszerek, fizetési lehetőségek a leghasználatosabbak, ami alapján lehet tervezni akár fejlesztéseket, korszerűsítéseket. Illetve a nem használt módszereket is meg lehet állapítani így könnyen át tudják majd csoportosítani fontosabb eddig akár elhanyagolt részekre az anyagi keretet.

Ehhez szorosan kapcsolódó adat a termékek és azok eladásának részletei. Az eladások száma, az átlagosan költött pénzértékek és az adott termék paraméterei alapján is lehet javítani az üzletek produktivitásán.

Valamint felmerült az is, hogy a vásárlókról is készíthetnénk elemzéseket, aminek a végén abban egyeztünk meg, hogy a vállalat szempontjából a vásárlók lakhelye a legfontosabb. A lakcímek alapján meg tudjuk nézni, hogy mely városok lakóitól jött a legtöbb rendelés, és ha nincs a környéken még üzlet akkor oda érdemes lehet terjeszkedni.

Vagyis a fentebb említett kérések mentén építettük fel a legideálisabb adattárházat a vállalat számára.

Az elvárások alapján ezekből a forrásadatokból dolgoztunk az adattárház megépítése során (AdventureWorks adatbázisból választottuk):



# Dimenzionális modell

A forrásadatokból pedig az alábbi ábrán látható sémát raktuk össze:

## Dimenzió táblák:

- **DimSalesPerson:** A cégnél dolgozó eladó dimenziója. Az üzlet szempontjából fontosnak vélt adatokat tárolunk az eladóról.  
[Név, Felvétel dátuma, Munkakör, Nem, Születési év, Terület]
- **DimAddress:** A cím dimenziója, ahova a terméket rendelték online rendelés esetén.  
[Város, Állam, Irányítószám]
- **DimStore:** A bolt dimenziója. A boltról tárolunk információt.  
[Bolt neve, Bolt tulajdonosának (eladó) azonosítója]
- **DimProduct:** A termékek dimenziója. A különböző eladó termékekről tárolunk információt.  
[Név, Szín, Ár, Méret...]
- **DimDate:** Dátum dimenzió. 17 évre előre tartalmazza minden nap dátumát.

## Ténytábla:

**FactSales:** Vásárlás tény táblája. Egy rekord valamely vásárlás/megrendelés egy adott termékét, tételét jelenti, darabszámmal, egységgel és egyéb adatokkal együtt.  
[Szállítás dátuma, Vásárolt darabszám, Egységár, Online vásárlás-e, Dimenziók azonosítói...]

DimSalesPerson		
DimSalesPersonID	int	PK
BusinessKey	int	
Name	varchar(64)	
HireDate	date	
JobTitle	varchar(64)	
Gender	char(1)	
BirthDate	date	
TerritoryName	varchar(64)	

DimStore		
DimStoreID	int	PK
BusinessKey	int	
Name	varchar(64)	
SalesPersonID	int	FK

FactSales		
OrderID	bigint	PK
DimDateID	int	FK1
ShipDate	date	FK2
DimAddressID	int	FK3
DimProductID	int	FK4
DimSalesPersonID	int	FK5
DimStoreID	int	FK6
OrderQty	int	
UnitPrice	money	
OnlineOrderFlag	int	

DimAddress		
DimAddressID	int	PK
BusinessKey	int	
City	varchar(32)	
StateProvince	varchar(64)	
PostalCode	varchar(24)	

DimDate		
DimDateID	int	PK
Date	date	
Day	int	
Month	int	
Quarter	int	
Year	int	

DimProduct		
DimProductID	int	PK
BusinessKey	int	
Name	varchar(64)	
ProductNumber	varchar(24)	
Color	varchar(16)	
ListPrice	money	
Size	varchar(4)	
Class	varchar(2)	
Style	varchar(2)	
DaysToManufacture	int	
SellStartDate	date	
SellEndDate	date	

## 3 ETL réteg adatbázisa:

### STAGE

Minden adat VARCHAR típusú és a forrásadatbázis megfelelő tábláinak összes oszlopát tartalmazza.

#### Táblák:

- Address
- Employee
- Product
- SalesOrderDetail
- SalesOrderHeader
- SalesPerson
- SalesTerritory
- StateProvince
- Store
- Person
- Customer

### Core / DW

Az adatok megfelelő típusúak és SCD Type 2-t is használunk. (ValidFrom, ValidTo)

#### Táblák:

- DWH\_Address
- DWH\_Employee
- DWH\_Product (SCD)
- DWH\_SalesOrderDetail
- DWH\_SalesOrderHeader
- DWH\_SalesPerson (SCD)
- DWH\_SalesTerritory (SCD)
- DWH\_StateProvince
- DWH\_Store (SCD)
- DWH\_Person
- DWH\_Customer

## View / DM

A Dimenzionális modell elkészítéséhez, adatok áttöltéséhez szükséges nézetek:

- DimAddressVW | BusinessKey: AddressID
- DimProductVW | BusinessKey: ProductID
- DimSalesPersonVW | BusinessKey: DWH\_SalesPerson.BusinessEntityID
- DimStoreVW | BusinessKey: DWH\_Store.BusinessEntityID
- FactSalesVW

A táblák tartalmazzák a LoadTimeStamp oszlopot, amely a rekord beszúrásának dátumát (Timestamp) jelenti, ezentúl a Surrogate Key is beszúrásra került.

### Táblák:

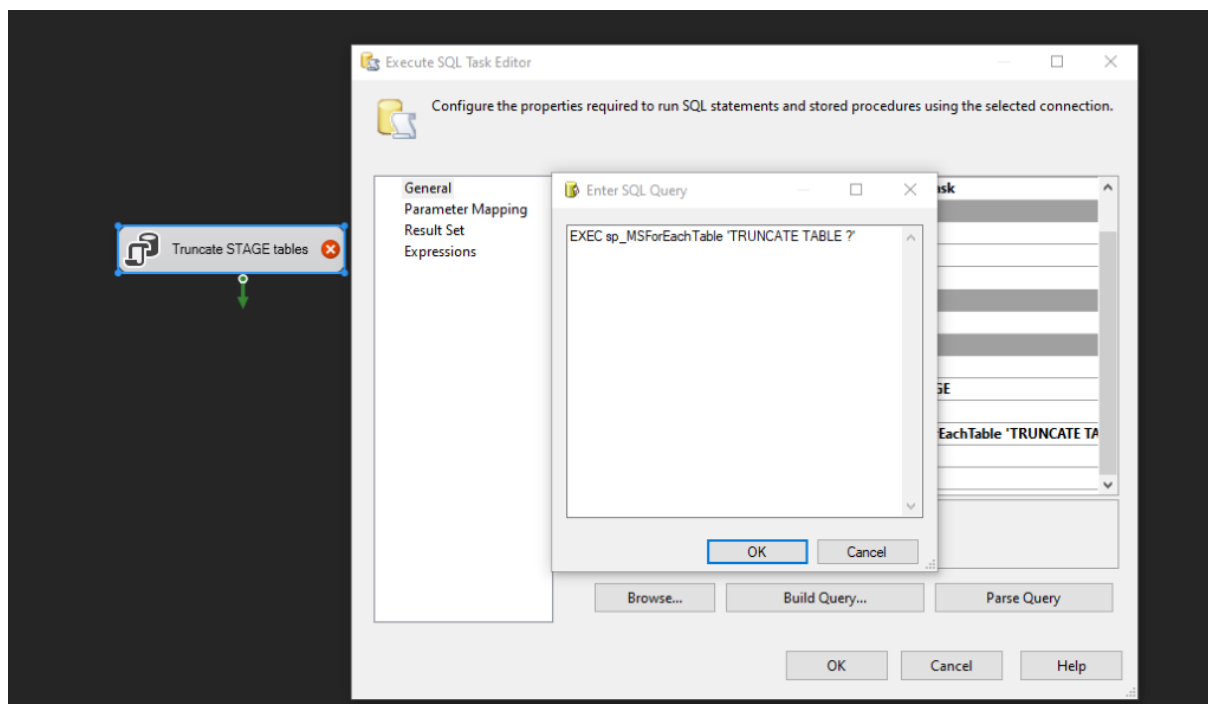
- FactSales
- DimStore
- DimSalesPerson
- DimProduct
- DimAddress
- DimDate



## ETL folyamat:

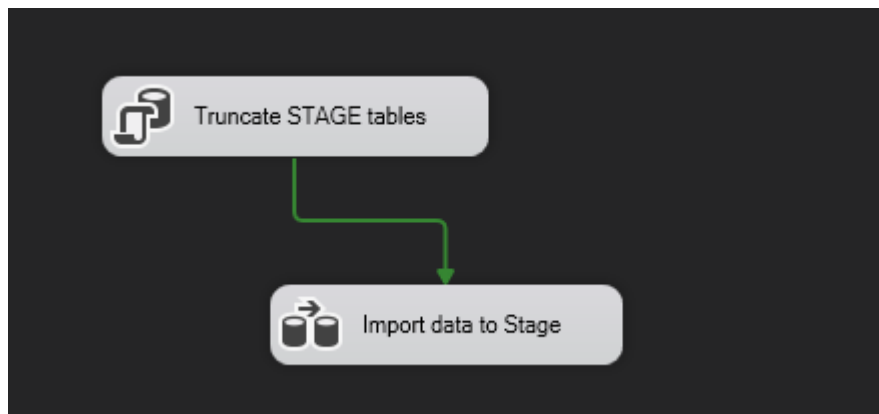
### EXTRACT

Az Staging rétegbe töltés előtt a táblák tartalmát üríteni kell, ezt a következő utasítással, egy Execute SQL task segítségével valósítottam meg:



Ez a művelet minden áttöltés előtt le fog futni, a Staging layerben csak egy áttöltés adatai lesznek megtalálhatóak.

Az ürítést követően megkezdődhet az adatok betöltése a .csv fájlokból, ezt egy Data Flow taskon belül fogjuk megvalósítani.



Az adatokat néhány mező (rowguid, beágyazott XML dokumentumok) kivételével, módosítás nélkül, VARCHAR típusként töltöttem be a Staging layerbe, Flat File Source és OLE DB Destination elemek felhasználásával.



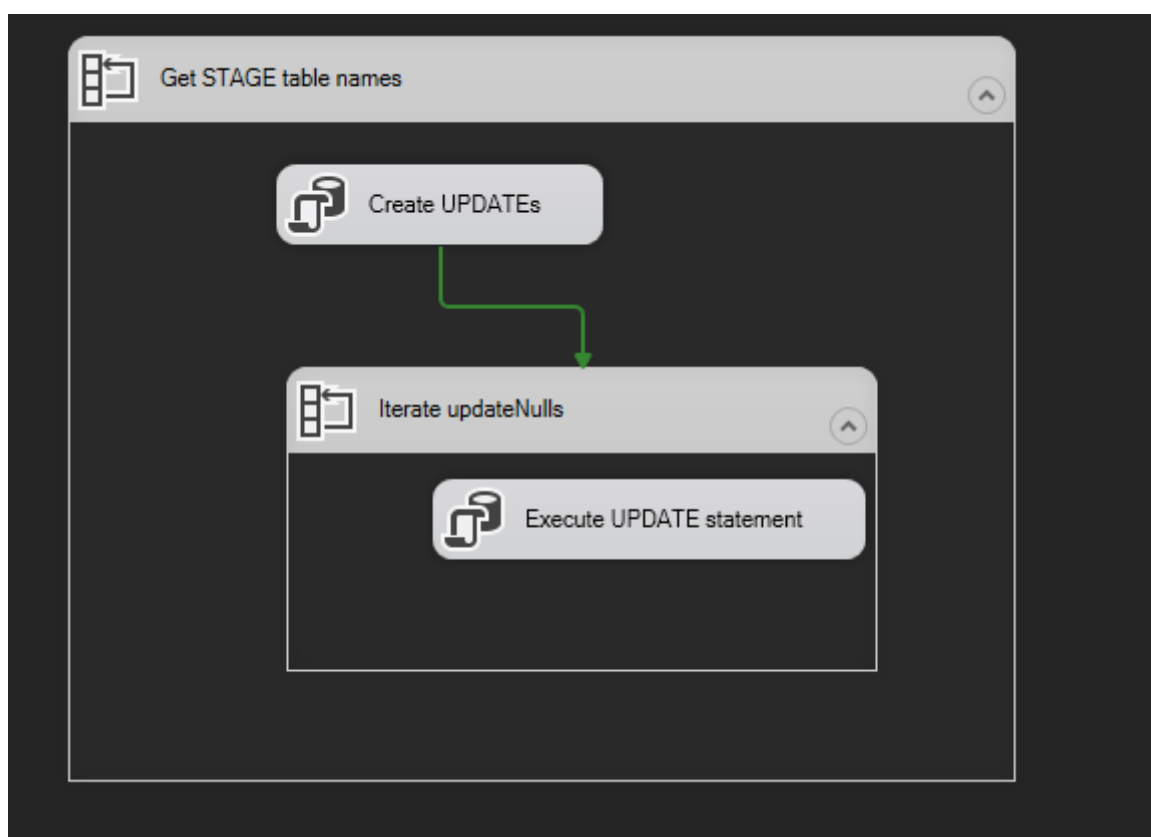
## TRANSFORM

A transzformálás előtt elkészítettem a mapping fájlt, ami alapján beállítom a megfelelő típusokat a DW réteghez és a Data Marthoz.

A típuskonverzió kívül az alábbi módosításokat végezzük még el:

- AddressLine1 és AddressLine2 összevonása, ezekre ugyanis külön nem lesz szükségünk.
- SalesOrderNumber esetében a prefix eltávolítása, mert ezt számként fogjuk tárolni.

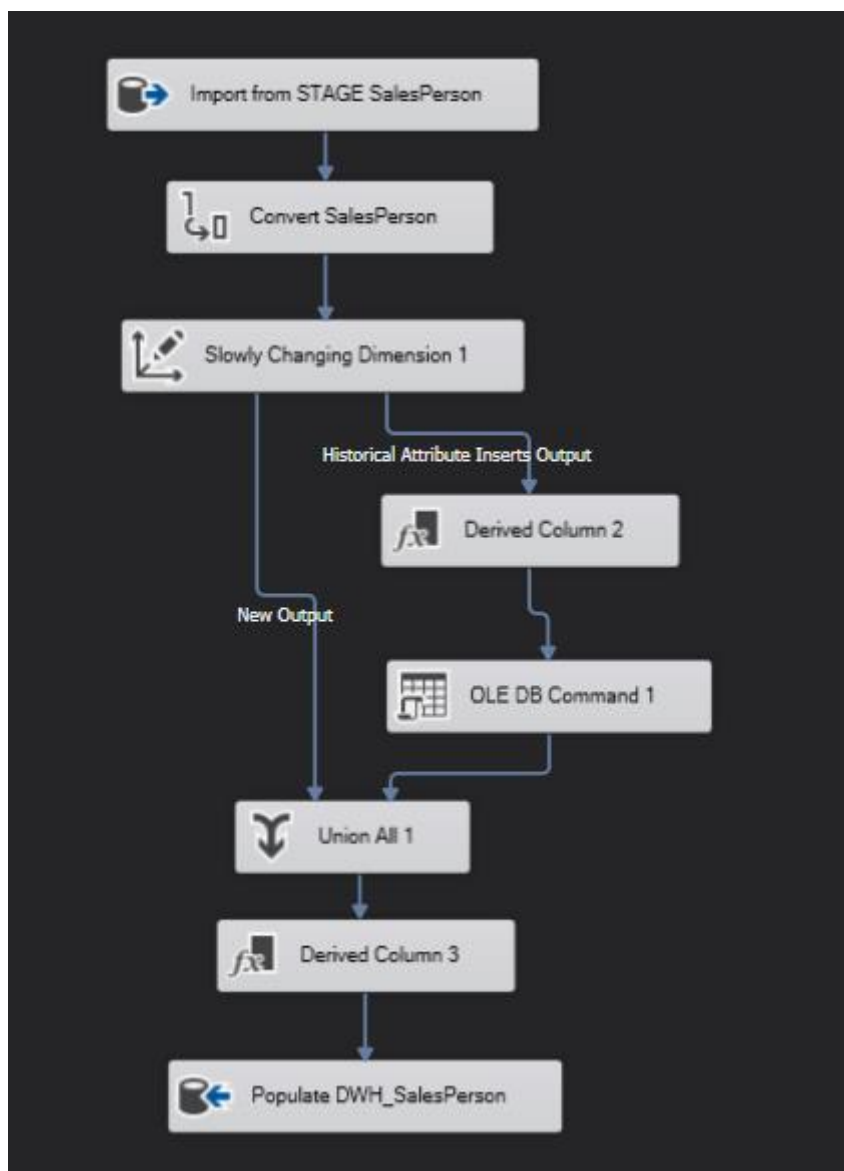
Az adatok áttöltése előtt a Staging rétegben a szöveges null értékeket kicseréltük null típusra, hogy később ebből ne legyenek problémák.



A külső foreach ciklusban lekérjük az adatbázisban található táblák nevét, ezen belül pedig UPDATE utasításokat generálunk az aktuális tábla minden oszlopára, ami ki fogja cserélni a szöveges NULL értéket NULL típusra. A belső foreach ciklusban ezeket az utasításokat hajtjuk végre egyesével, a tábla minden mezőjén, így végül nem maradnak szöveges NULL értékek az adatbázisban.

A transzformálás néhány kivételtől eltekintve minden esetben a következő lépésekből áll:

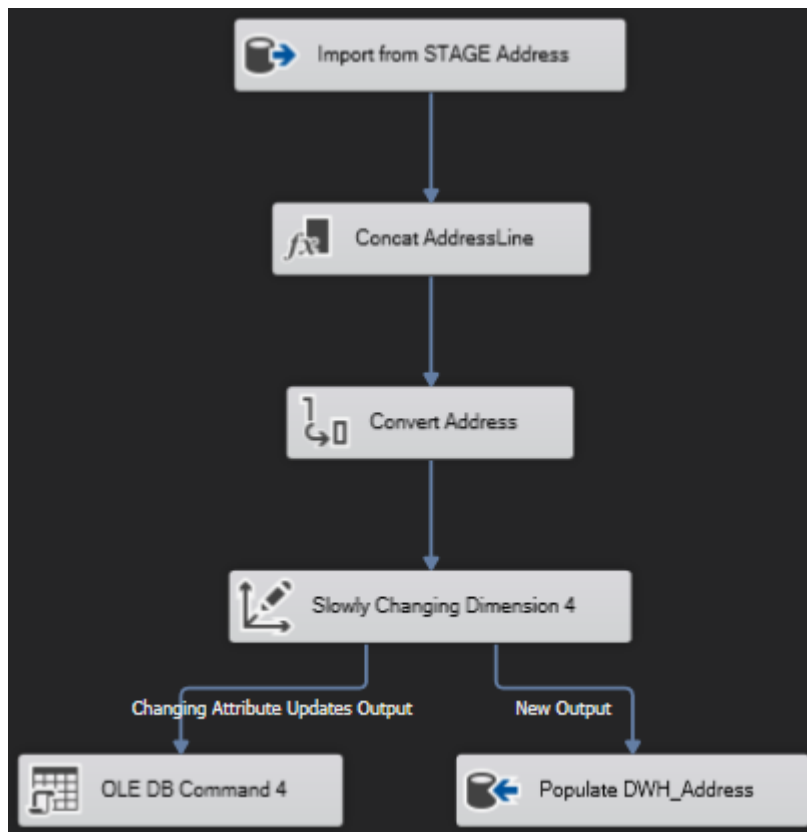
- Adatok beolvasása a STAGE adatbázis tábláiból,
- Típuskonverzió végrehajtása,
- SCD beállítása,
- Betöltés a DWH réteg tábláiba.



Az SCD úgy lett megvalósítva, hogy legtöbb esetben felülírjuk az új adatokkal a meglévő rekordokat, viszont néhány kiemelt tábla esetében, ahol fontos a historikus adatok elérhetősége, ott ezeket is eltároltuk, ezek pedig a következők:

- Product,
- SalesPerson,
- SalesTerritory,
- Store.

Az Address tábla esetében az AddressLine1 és AddressLine2 oszlopokat egyesítettük, ehhez Derived Column elemet használtunk, a következő beállításokkal:



Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

- Variables and Parameters
- Columns

- Mathematical Functions
- String Functions
- Date/Time Functions
- NULL Functions
- Type Casts
- Operators

Description:

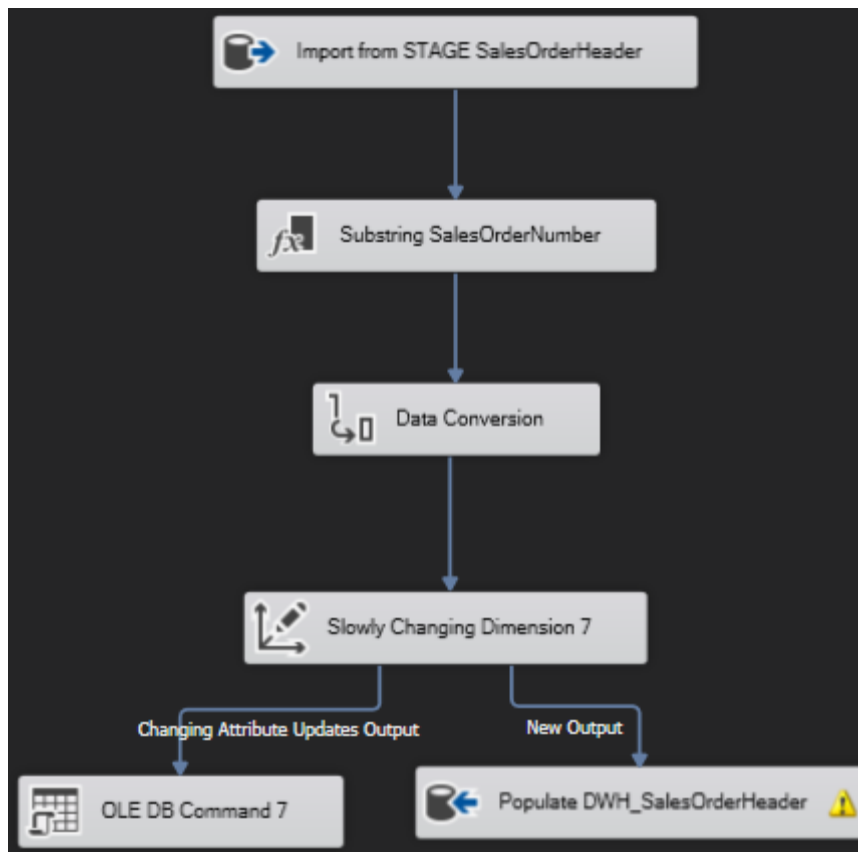
Derived Column Name	Derived Column	Expression	Data Type	Length	Precision	Scale	Code Page
AddressLine	<add as new column>	ISNULL(AddressLine2) ? AddressLine2 + " " + AddressLine1 : AddressLine1	Unicode string [DT_WS...	513			

Configure Error Output...

OK Cancel Help

Fontos volt ellenőrizni, hogy nem NULL érték-e az AddressLine2, mert a NULL értékkel végzett műveletek mindig NULL eredményt adnak, ami akadályozta volna a probléma megoldását.

A SalesOrderHeader tábla esetében a SalesOrderNumber oszlop számként kerül eltárolásra, így a prefixet el kell távolítani, ehhez szintén egy Derived Column elemet használtunk, a következő módon:



Derived Column Transformation Editor

Specify the expressions used to create new column values, and indicate whether the values update existing columns or populate new columns.

- Variables and Parameters
- Columns

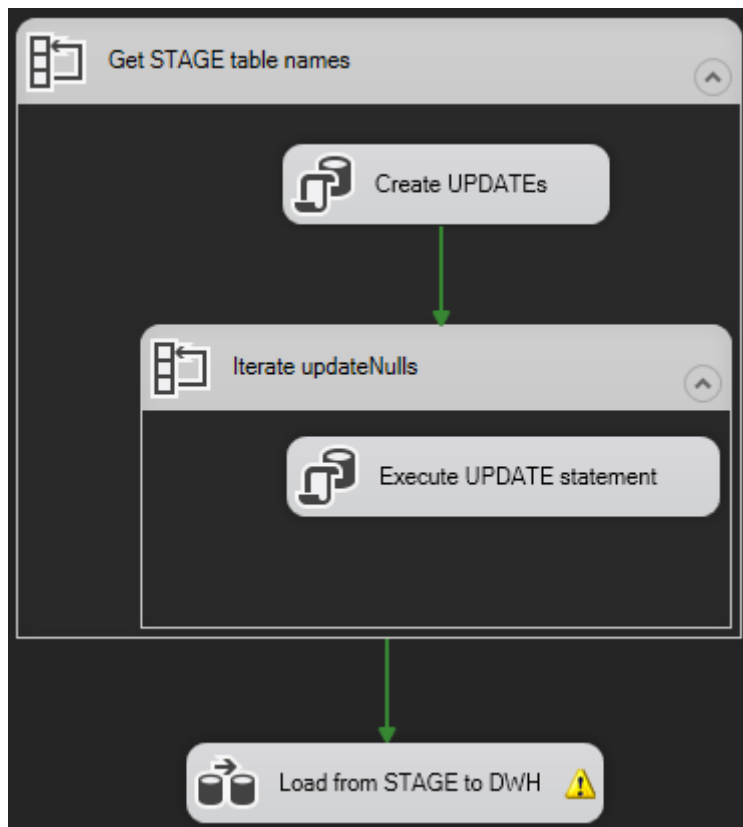
- Mathematical Functions
- String Functions
- Date/Time Functions
- NULL Functions
- Type Casts
- Operators

Description:

Derived Column Name	Derived Column	Expression	Data Type	Length	Precision	Scale	Code Page
SalesOrderNumber	Replace 'SalesOrderNu...	SUBSTRING(SalesOrderNumber,3,LEN(SalesOrderNumber) - 2)	string [DT_STR]	256			1252 (A

Configure Error Output...

OK Cancel Help

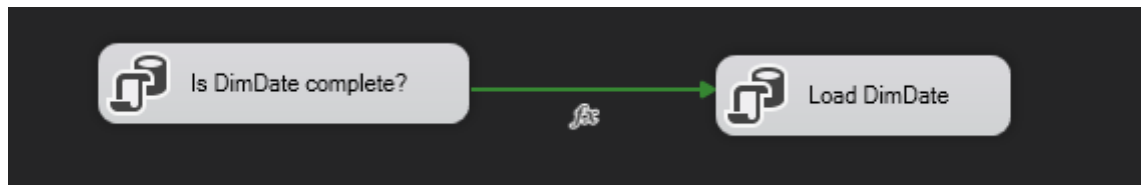


Összegzőképp tehát a Transform réteg a következőképp néz ki:

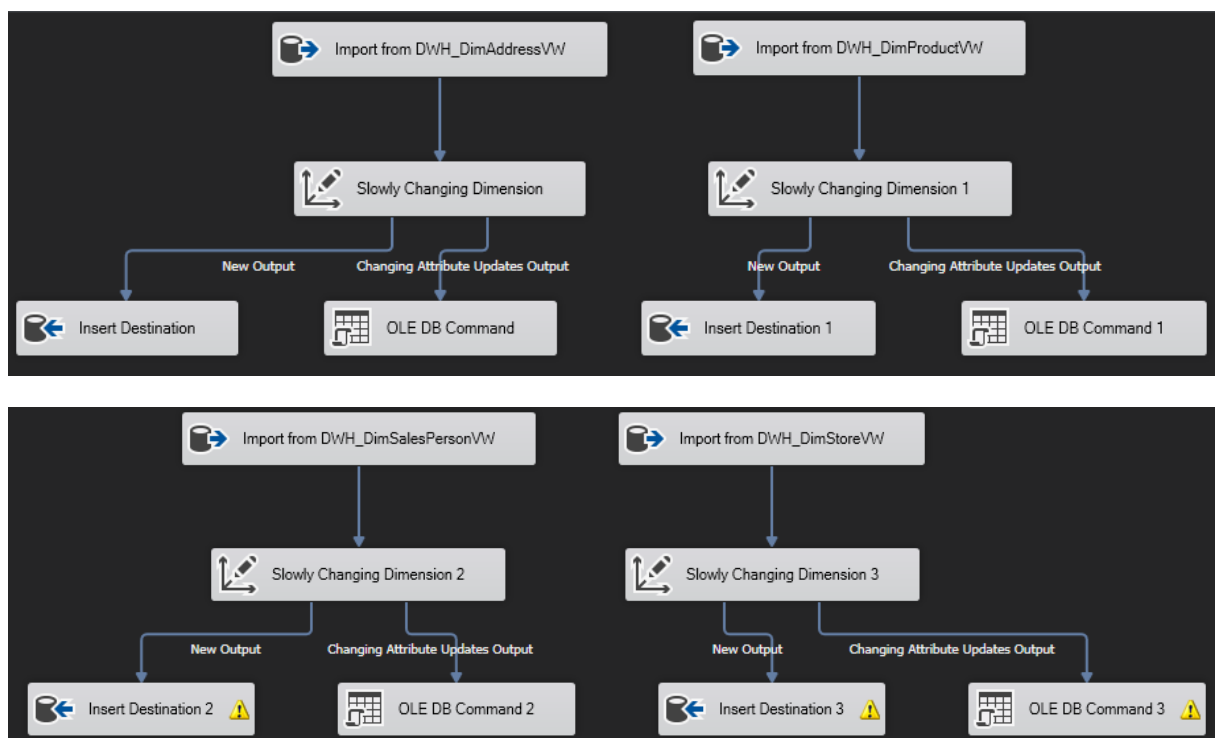


## LOAD

A LOAD rétegben először ellenőrizzük, hogy a dátum dimenzió fel van-e töltve adatokkal, és ha nincs, akkor egy tárolt eljárás segítségével legeneráljuk az adatokat és feltöltjük.



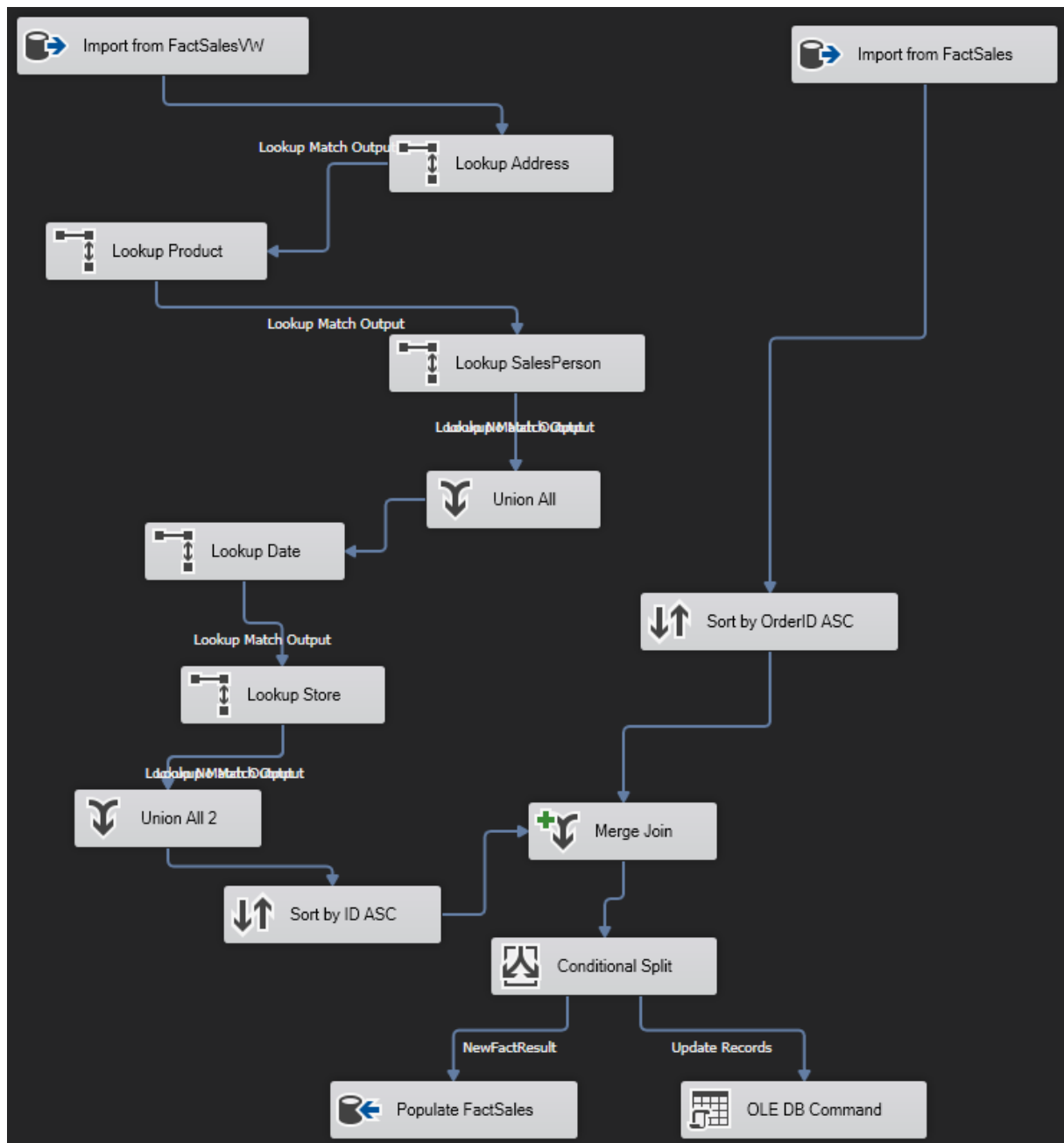
Ezzel párhuzamosan feltöltjük a többi dimenziót az adattárházunkból:



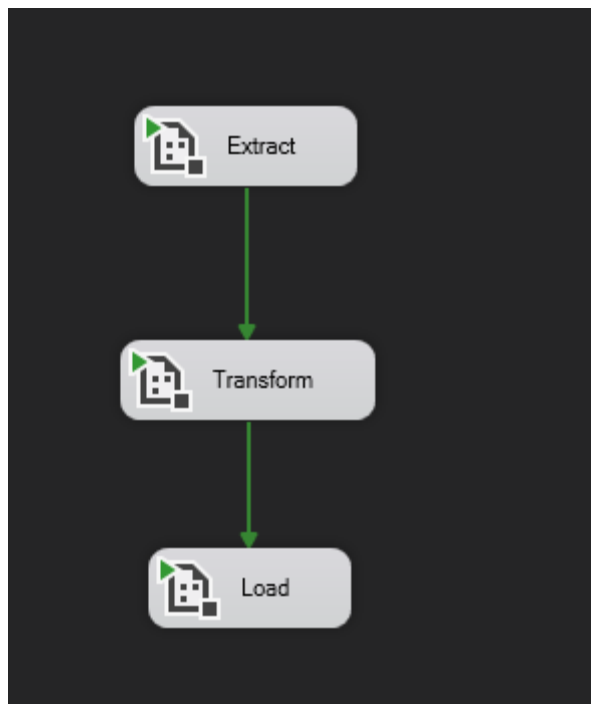
Az dimenziókban az adattípusok ugyanazok, mint az adattárházban, ezért itt nem végzünk típuskonverziót. Itt is állítunk be SCD-t, historizálás nélkül, mert itt mindig a legfrissebb adatokat szeretnénk látni.

A dimenziótáblák után feltöltjük a ténytáblát is, ehhez készítettünk egy nézetet az adattárházban, amiben szerepelnek azok az adatok, amikre szükségünk lesz.





Végül csináltunk egy új package-t, ami lefuttatja a teljes ETL folyamatot:



# Riportkészítés

## LEGKEVESEBB ÉS LEGTÖBB BEVÉTELT HOZÓ TERMÉKEK (TOP5) (AMELYEKNÉL VAN ELADÁS):

Least sold product:

TotalOrderCount	TotalPrice	Name
10	\$162,72	LL Road Seat/Saddle
90	\$513	Mountain Bike Socks, L
4	\$800,208	LL Touring Frame - Blue, 58
8	\$1 198,99	LL Mountain Frame - 2 Black, 40
91	\$1 480,75	LL Touring Seat/Saddle

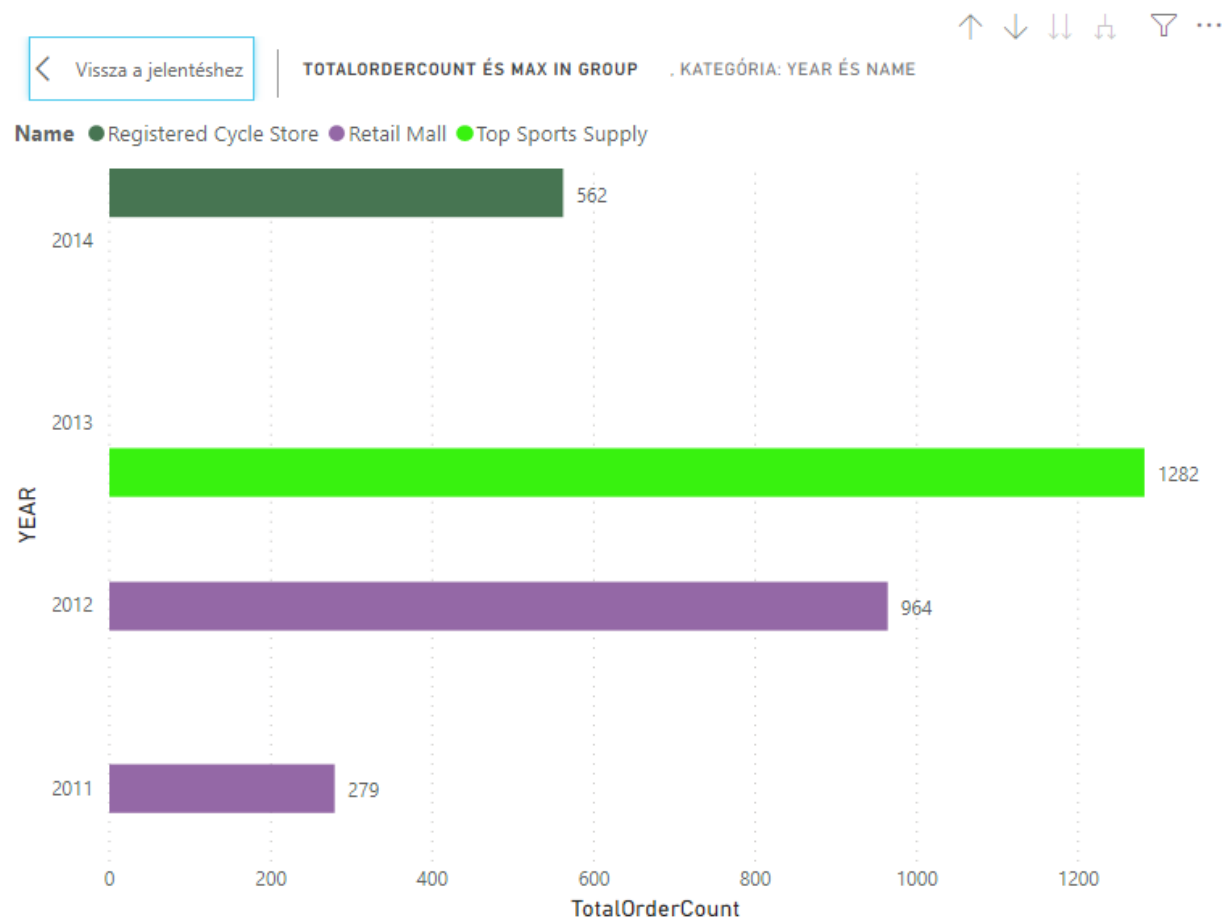
Most sold product:

TotalOrderCount	TotalPrice	Name
2977	\$4 406 151,2662	Mountain-200 Black, 38
2664	\$4 014 067,7999	Mountain-200 Black, 42
2394	\$3 696 486,4726	Mountain-200 Silver, 38
2234	\$3 441 292,5443	Mountain-200 Silver, 42
2216	\$3 436 090,7946	Mountain-200 Silver, 46

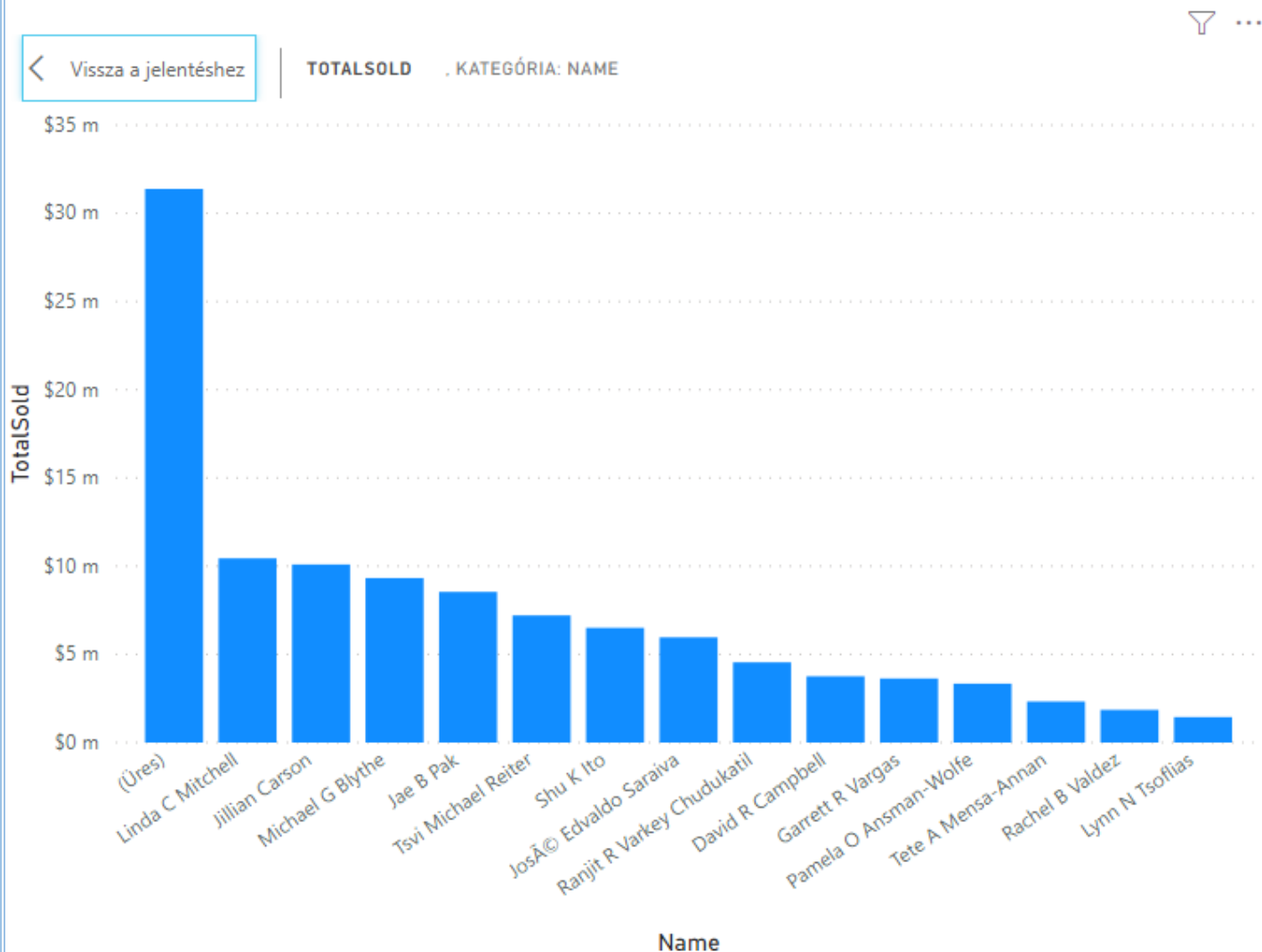
## VÁROSONKÉNT ELADOTT TERMÉKEK MENNYISÉGE:



**AZ EGYES ÉVEKBEN, MELYEK AZOK A BOLTOK, AMELYEKNEK A LEGNAGYOBB KÉSZLETRAKTÁRRÁ VOLT SZÜKSÉGÜK, AZAZ MELY BOLTOK ADTÁK EL A LEGTÖBB DARAB TERMÉKET ÉVENKÉNT:**



## JELENÍTÜK MEG AZ EGYES ÜZLETVEZETŐK/ELADÓK (SALESPERSON) PRODUKTIVITÁSÁT!



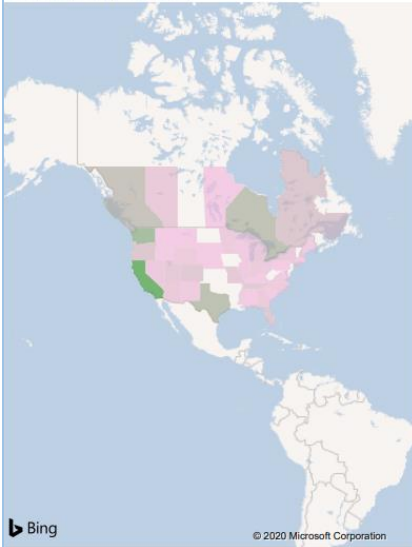
*Az üres oszlopban az online rendelt termékek összértéke van! Innen látszik, hogy mennyire fontos a cég számára az online webshop!*

Dimenzionális és számított táblák (mezők), amiket használtunk a riportkészítéshez:

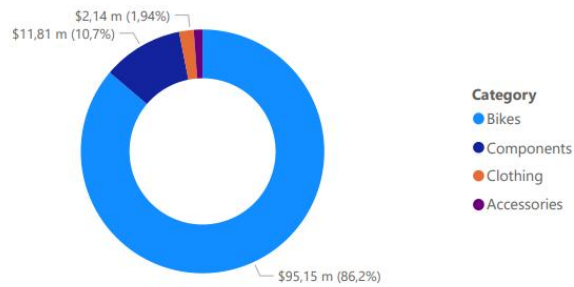
- ✓ CityByOrders
- ✓ DimAddress
- ✓ DimDate
- ✓ DimProduct
- ✓ DimSalesPerson
- ✓ DimStore
- ✓ FactSales
- ✓ SalesByProduct
- ✓ SalesPersonBySales
- ✓ StoreBySales

## DASHBOARD

Total Sales by location



Total Sales by product category



**\$23,13 m!**  
Cél: 43,37 m (-46.66%)

Continent	2011	2012	2013	2014	Összesen
North America	\$10 186 359,35	\$25 827 030,12	\$29 664 162,19	\$14 072 296,57	<b>\$79 749 848,22</b>
Europe	\$789 825,04	\$3 417 078,28	\$9 454 049,29	\$6 279 218,48	<b>\$19 940 171,09</b>
Pacific	\$1 442 128,50	\$2 142 819,33	\$4 098 371,41	\$3 000 550,77	<b>\$10 683 870,01</b>
Összesen	<b>\$12 418 312,88</b>	<b>\$31 386 927,73</b>	<b>\$43 216 582,88</b>	<b>\$23 352 065,82</b>	<b>\$110 373 889,31</b>

Select Year and Quarter

- ☒ 2011
- ☒ 2012
- ☒ 2013
- ☒ 2014

**\$110,37 m**

Total Sales

**13**

Count of Salesperson

**699**

Count of Stores

**504**

Count of Products

## Scrum dokumentáció

A Scrum keretrendszer megvalósításához a Jira ügy- és projektkövető szoftvert használtuk, Confluence-szel együtt.

A dokumentációt online készítettem Confluence segítségével és sajnos a PDF-eket nem 100%-osan generálta le (pár helyen az ékezetes betűk hiányoznak).

# Sprint Planning - oenikAdattarhaz Team Work

01 Oct 2020

Jelenlévk:

- Balogh Bence
- Bózsó Bence
- Lukács Andrea Eszter
- Nagy Dávid

## Sprint planning checklist

Elkészület	Meeting	Follow up
<input checked="" type="checkbox"/> Csapat tagjaival kapcsolat felvétele	<input checked="" type="checkbox"/> Bemutatkozás	<input type="checkbox"/> Epic létrehozása a taskokhoz
<input checked="" type="checkbox"/> Csapat kommunikációs platformjának kialakítása (Discord)	<input checked="" type="checkbox"/> Jelenlét felmérése	<input type="checkbox"/> Kapcsolat felvétele a hiányzó csapattaggal
<input checked="" type="checkbox"/> Meeting időpontjának meghatározása	<input checked="" type="checkbox"/> Jira+Scrum gyorstalpaló a csapat tagjainak	<input type="checkbox"/> Új kommunikációs platform meghatározása
	<input checked="" type="checkbox"/> Csapatmunka áttekintése (Taskok meghatározása)	
	<input checked="" type="checkbox"/> Szerepkörök kiosztása	
	<input checked="" type="checkbox"/> Sprint indítása	

## Sprint team members

Name	Role
Nagy Dávid	Scrum Master
Lukács Andrea Eszter	Adattárház tervező
Balogh Bence	SQL fejlesztő
Bózsó Bence	ETL eszköz kezelő
Sallai András	Riportkészítő

## Részletek

<b>Sprint téma</b>	Üzleti produktivitás
<b>Kezdés dátuma</b>	01 Oct 2020
<b>Zárás dátuma (tervezett)</b>	29 Oct 2020
<b>Kapacitás</b>	5 f
<b>Napok száma</b>	20

## Összegzés

A Sprint Planning során kiválasztotta mindenki a neki megfelelő szerepkört és elindítottuk a sprintet. Ezután rögtön el is kezdtünk gondolkodni azon, hogy melyik adatbázisból használjunk adatokat. Mivel sokkal több adat volt az AdventureWorks adatbázisban, így arra esett a választás. A meeting végére pedig meghatároztuk a sprintünknek a témáját (Üzleti produktivitás), ami első kapcsolatban fog állni a use case során megválaszolandó kérdésekkel is.



# Daily Scrum | 2020-10-08

## Dátum

08 Oct 2020

## Résztvev

- Nagy Dávid [Scrum Master]
- Balogh Bence [SQL fejleszt]
- Bózsó Bence [ETL eszköz kezel]
- Lukács Andrea Eszter [Adattárház tervez]
- Sallai András [Riportkészít]



## Célok

- Dimenzionális modell elkészítésének véglegesítése
- SQL fejleszt feladatainak konkrét meghatározása, áttekintése és elkezdése
- Sallai András szerepkörének kiosztása, Jira+Scrum gyorstalpaló

## Meeting topikok

Id	Téma	Eladó	Megjegyzések
10 perc	Jira+Scrum gyorstalpaló	Balogh Bence	<ul style="list-style-type: none"><li>Nagy Dávid 10 perccel késbb érkezett, ezért Balogh Bencét bízta meg ezzel kapcsolatban.</li></ul>
20 perc	Dimenzionális modell	Lukács Andrea	<ul style="list-style-type: none"><li>Közös beszélgetés, tervezés.</li></ul>
15 perc	SQL fejleszt feladatok áttekintése	Nagy Dávid	

## Döntések

- Új Kommunikációs platform: Teams
- Daily Scrum meetingek ezentúl csütörtökönként 20:00-tól tartjuk.

# Daily Scrum | 2020-10-15

## Dátum

15 Oct 2020

## Résztvev

- Nagy Dávid [Scrum Master]
- Balogh Bence [SQL fejleszt]
- Bózsó Bence [ETL eszköz kezel]
- Lukács Andrea Eszter [Adattárház tervez]

## Célok

- 3 ETL réteg adatbázisainak átnézése, javítása és dokumentálása
- ETL eszköz kezel feladatainak átbeszélése, határid meghatározása
- F dokumentáció létrehozása, elkezdése

## Meeting topikok

Id	Téma	Eladó	Megjegyzések
20 perc	3 ETL réteg	Balogh Bence	<ul style="list-style-type: none"><li>A forrásadatbázisban(AdventureWorks) több tábla és attribútum hiányzik vagy helytelen a neve a modellben, így csúszás következett be.</li><li>Feladat félreértelmezése miatt korrigálás és kiegészítés (nézetek) szükséges</li></ul>
15 perc	ETL eszköz kezel	Bózsó Bence	
10 perc	F dokumentáció	Lukács Andrea Eszter	

## Végrehajtandó feladatok

- ☒
- Lukács Andrea Eszter: F dokumentáció formájának létrehozása és a use case végleges megírása

# Daily Scrum | 2020-10-24

## Dátum

24 Oct 2020

## Résztvevők

- Nagy Dávid [Scrum Master]
- Bózsó Bence [ETL eszköz kezel]
- Lukács Andrea Eszter [Adattárház tervez]
- Sallai András [Riportkészít]

## Célok

- ETL folyamatok átnézése, dokumentáció ellenrzése
- Riportok megbeszélése, megtervezése
- Dokumentáció szerkezetének ellenrzése, javítása

## Meeting topikok

Id	Téma	Eladó	Megjegyzés
15 perc	ETL folyamatok	Bózsó Bence	<ul style="list-style-type: none"><li>• Felesleges és értelmetlen warning pár helyen, de nem okoz problémát.</li></ul>
15 perc	Riportkészítés	Sallai András	<ul style="list-style-type: none"><li>• Mivel az órán konkrét vizualizáció nem történt, így csúszás várható a sprint végét illeten.</li><li>• Kérdéseket írtunk fel, amikre a riportok segítségével adunk választ.</li></ul>
5 perc	Dokumentáció	Nagy Dávid	

## Végrehajtandó feladatok



@ Dávid Nagy

Dokumentáció teljes átnézése, javítása és hiányosságok pótlása.

## Retrospektív | 2020-10-31

Date	31 Oct 2020
Team	oenikadattarhaz2020
Participants	Nagy Dávid, Balogh Bence, Bózsó Bence, Lukács Andrea Eszter, Sallai András

### Háttér

A Retrospektív meeting az OE-NIK 2020 Adattárház csoportmunka sprintjéről ad önértékelést, visszatekintést. A sprint során a forrásadatbázis kiválasztása után egy dimenzionális modell elkészítését hajtotta végre a csapat a megfelelő rétegekkel, majd a létrejött modellen riportkészítés történt. A munka során minden lépésről dokumentáció történt.

A Daily stand-up meetingek nem naponta, hanem hetente történtek meg.

### Retrospektív

Kezdjük el!	Hagyjuk abba!	Folytassuk!
<ul style="list-style-type: none"><li>Új folyamatok, technikák kipróbálása</li><li>Gyakoribb, részletes és fókuszált meetingek</li><li>Feladatok párhuzamosítása Feladatkör felosztása több emberre, így jobban érezhetőbb lenne a közös csapatmunka. (pl.: ETL folyamatokat 3 fele is lehetne osztani így hamarabb meglenne)</li></ul>	<ul style="list-style-type: none"><li>Hosszas meetingek</li><li>Last minute munkavégzés</li></ul>	<ul style="list-style-type: none"><li>Eddig használt kommunikációs platformok</li><li>Jól szétválasztott szerepkörök</li><li>A meetingek hangulata és eredményessége, a munkamegosztás, egymás kisegítése.</li></ul>

### Új ismeretek

- ☒ Scrum működése, csapatmunka, meetingek
- ☒ Jira
- ☒ Új eszközök megismerése (SSIS, PowerBI)
- ☒ Hasznos elméleti ismeretek

NMBAdattarhaz

Next-gen software project

Roadmap

Backlog

Board

Code

Pages

Add item

Project settings

Projects / NMBAdattarhaz

# NMBAD Sprint

DN

BB

AL

AS

BB

TO DO

## Complete NMBAD Sprint

This sprint contains 7 completed issues.  
That's all of them - well done!

Complete sprint

Cancel

0 days remaining

Complete sprint

GROUP BY

None

DONE 7

ADATTÁRHÁZ TERVEZÉS:  
Dimenzionális modell elkészítése

ÜZLETI PRODUKTIVITÁS

NMBAD-4

AL

SQL FEJLESZTÉS:3 réteghez szükséges  
adatbázisok létrehozása

ÜZLETI PRODUKTIVITÁS

NMBAD-5

BB

SQL FEJLESZTÉS: Lekérdezések  
megírása

ÜZLETI PRODUKTIVITÁS

NMBAD-6

Quickstart