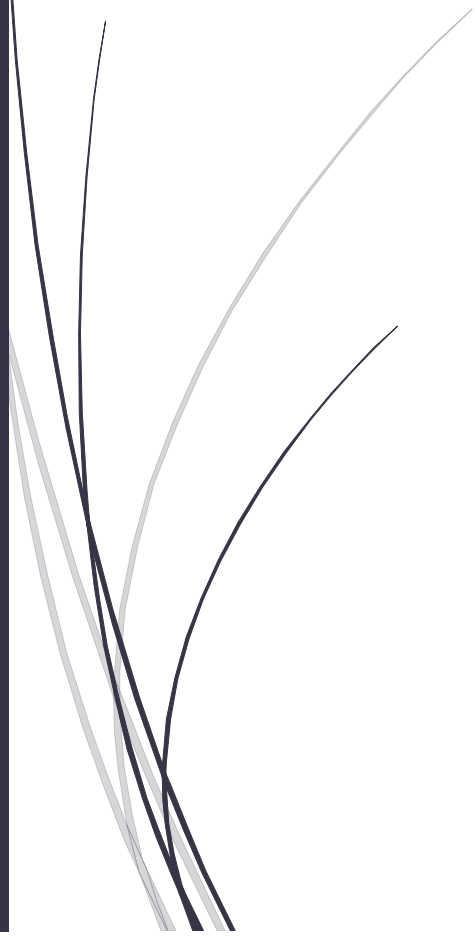
2020. 11. 21.

Projektmunka II.

Féléves feladat dokumentáció

Bózsó Bence



TARTALOMJEGYZÉK

BEVEZETÉS

A PROJEKT TÉMÁJA

A projekthez a *2019/2020/2* félévben a Projektmunka I. tárgyhoz készült beadandó munkámat fogom felhasználni, kibővíteni olyan módon, hogy az megfeleljen a feladat követelményeinek.

A PROJEKT ALAPJÁT ADÓ SZOLGÁLTATÁS MŰKÖDÉSE

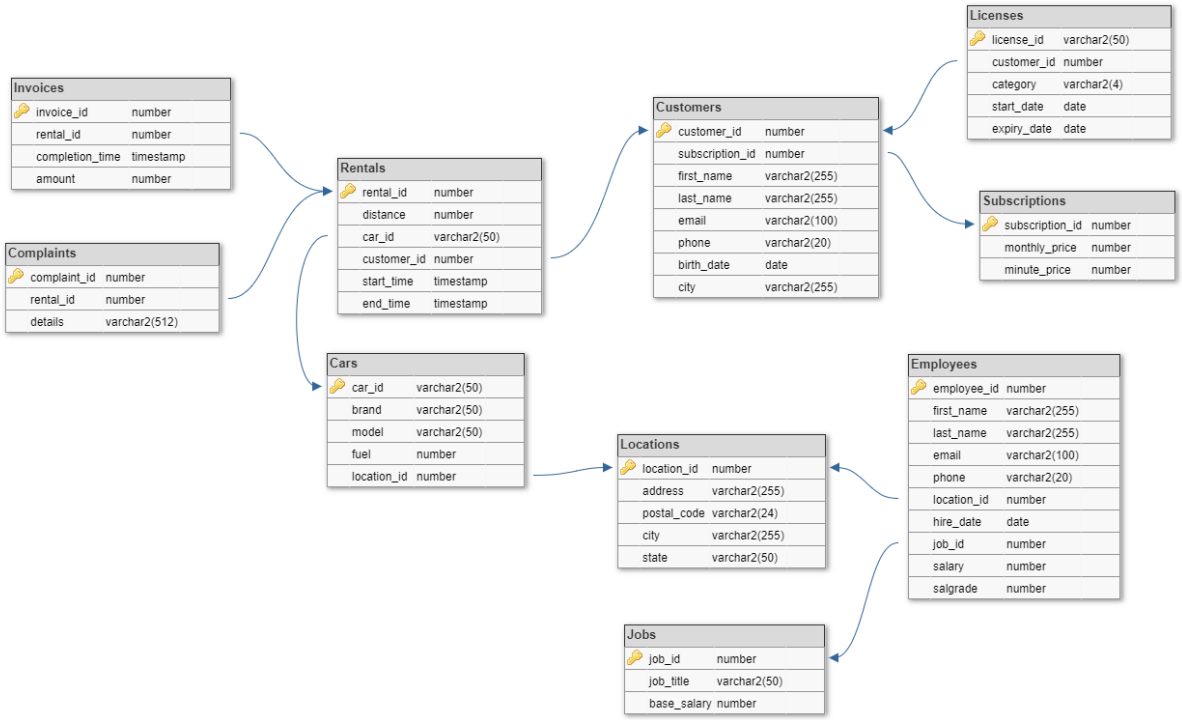
Egy mobiltelefon segítségével használható közösségi autóbérlő alkalmazás adatbázisát tervezem meg. A szolgáltatásra való regisztrációkor az alkalmazás rögzíti a felhasználó személyes adatait, valamint a jogosítványa adatait, mivel enélkül az autóbérlés nem lehetséges. Ezután a felhasználó kiválaszthatja a számára legoptimálisabb előfizetési konstrukciót (havidíj + alacsonyabb percdíj, vagy havidíj nélkül magasabb percdíj). Ezt követően a felhasználó számára megjelenik a szolgáltatási övezet térképe, amin szerepelnek az elérhető autók, valamint ezek adatai (típus, pozíció, töltöttség). Többféle típusú, méretű autó is rendelkezésre állhat, bizonyos típusoknál feláras is lehet annak a választása. A jármű kiválasztása után a felhasználó lefoglalhatja azt, a mobilalkalmazás segítségével nyithatja és használhatja. A bérlés végeztével az autót le kell parkolni a szolgáltatási övezeten belül, és az alkalmazásban leállítani a bérlést. Ha bármilyen problémát észlel az autóval (tisztasági probléma, törés stb.) ezt ekkor jelentheti, képekkel dokumentálhatja, ez a rendszerben rögzítésre kerül. A bérlés lezárását követően a rendszer elkészíti a számlát az előfizetése, bérlés időtartama, illetve esetlegesen az autó felára alapján, amit a felhasználónak ki kell egyenlítenie. A bérlés végétől az autó ismét elérhetővé válik a többi felhasználó számára is. A szolgáltatás több városban is működik, minden városban a megfelelő működtetésért felelős személyzettel. Az általam tervezett adatbázis a rendszer működéséhez szükséges legfontosabb adatokat tárolja.

AZ ADATOK FORRÁSA

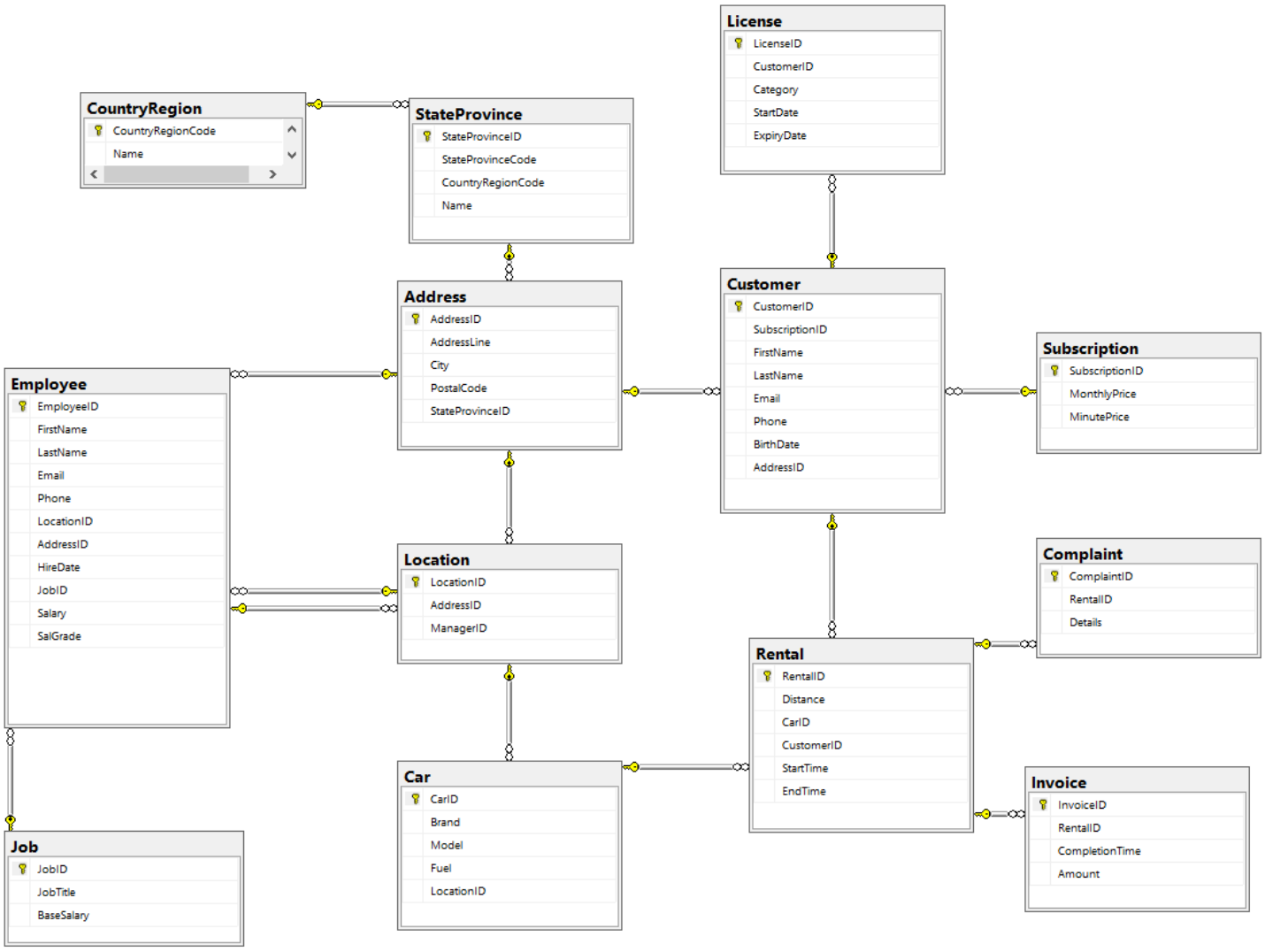
Mivel egy korábbi projektemet folytatom, ezért nem találtam hozzá megfelelő adathalmazt az interneten, így azokat magamnak generáltam le. Ehhez részben külső alkalmazást használtam, viszont az egyes táblák közti kapcsolat, vagy a szoftver mennyiségi korlátjai miatt ez a legtöbb esetben nem működött, így az adatok jelentős részét saját alkalmazással hoztam létre.

Az így korábban létrejött adatforrásokon további módosításokat kellett végeznem, hogy alkalmasabb legyen adattárház tervezésére, valamint látványosabb riportok készítésére. Ez főként a címek tárolását érintette, az eddigi gyakorlattal szemben már nem csak a város neve került tárolásra. A címek forrása az *AdventureWorks* adatbázis egy részhalmaza volt.

Az alábbi képen az adatbázis módosítás nélküli szerkezete látható:



A bővítést követően pedig így módosult a forrásadatbázis szerkezete, amit a projekt során használni fogok:



További változtatás, hogy a korábbi, Oracle 12c adatbázisban tárolt adatokat Microsoft SQL Server adatbázisba helyeztem át, ami a későbbi munkafolyamatokat hivatott megkönnyíteni.

USE-CASE MEGHATÁROZÁSA

Egy közösségi autómegosztással (e-carsharing) foglalkozó vállalat azért keresett fel, hogy tervezzek meg és valósítsak meg egy adattárházat, amelynek segítségével pontos képet kapnak a vállalat sikerességéről, hatékonyságáról, meghatározhatják azokat a területeket, ahol még fejlődésre van szükség.

A cég több szolgáltatási területen működik az Amerikai Egyesült Államok nyugati és déli részén, azonban ügyfeleik számottevő része másik országból származik, és csak az ideiglenes tartózkodás során használták a szolgáltatást. Fontos ismerni ezen ügyfelek arányát, valamint a tőlük származó bevételek megoszlását, ugyanis ennek függvényében tudnak módosítani a már meglévőkön, vagy új előfizetési konstrukciókat alkotni, ami elősegítheti az ügyfelek számának növekedését. Mindezek miatt a területi adatok rendkívül fontosak a vállalat számára.

Az előző ponthoz kapcsolódóan fontos ismerni a különböző előfizetési konstrukciók választásának arányát, ennek ismeretében lehetséges ezeket az igényeknek megfelelően fejleszteni.

A vállalat különös figyelmet fordít a reklamációk számának minimálisra csökkentésére, így fontos szempont, hogy melyik telephelyeken milyen arányban jelentenek problémát, megfigyelhetőek-e csak egyes telephelyeket érintő hibák.

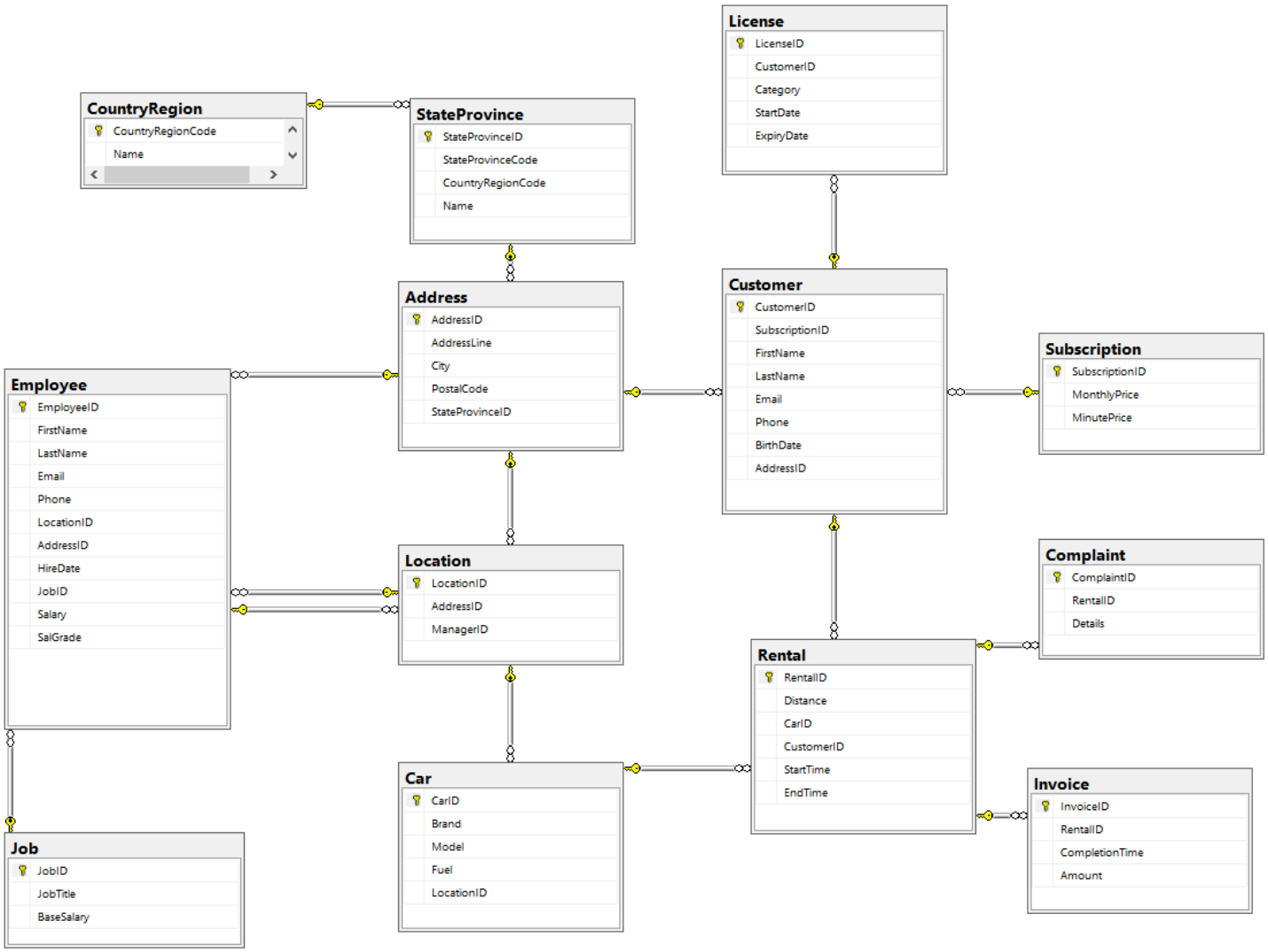
Fontos információ lehet az is, hogy a szolgáltatást használó ügyfelek életkor tekintetében hogyan oszlanak meg, melyek azok a korcsoportok, ahonnan további ügyfelek szerzése szükséges.

Mindezeken kívül szeretnék tudni azt is, hogy melyek a legnépszerűbb autók, melyekből érdemes többet vagy kevesebbet beszerezni, melyikre érdemes kedvezményt adni a jobb kihasználtságuk érdekében.

Az itt felsorolt adatok időbelisége kiemelten fontos, szükséges legalább éves, de lehetőség szerint negyedéves bontásban ezeket az információkat megjeleníteni.

FORRÁSADATOK

A forrásadatok egy Microsoft SQL Server adatbázisban vannak tárolva, melynek szerkezete a következő módon néz ki:



Az adatok legnagyobb részét a Rental, Invoice, és a Complaint táblák tartalmazzák, körülbelül 1,4 millió és több, mint 100.000 rekorddal, ezekből a táblákból fognak származni a tényadataink.

DIMENZIONÁLIS MODELL

A forrásadatok és a use-case alapján az ebben a fejezetben látható dimenziókat és tényeket különböztetjük meg.

DIMENZIÓK

DATE

A dátumokat tartalmazza napi bontásban, több évre előre.

CUSTOMER

Az egy bérléshez tartozó ügyfél üzleti szempontból releváns fontosabb adatait tartalmazza: életkora, lakhelye, az általa használt előfizetési konstrukció.

CAR

Egy autó lényegesebb adatait tartalmazza: márka, modell, elhelyezkedés.

LOCATION

Az a telephely, amelynek városában az adott bérlés megtörtént.

TÉNYTÁBLÁK

RENTAL

Egy bérlés adatait tartalmazza, egy rekord egy bérlést tartalmaz, tényadatok a megtett távolság, a bérlés kezdete, vége, történt-e reklamáció, valamint a fizetett összeg.

CSILLAGSÉMA

Az itt felsorolt dimenziókból és ténytáblából a következő csillagsémát hoztam létre:

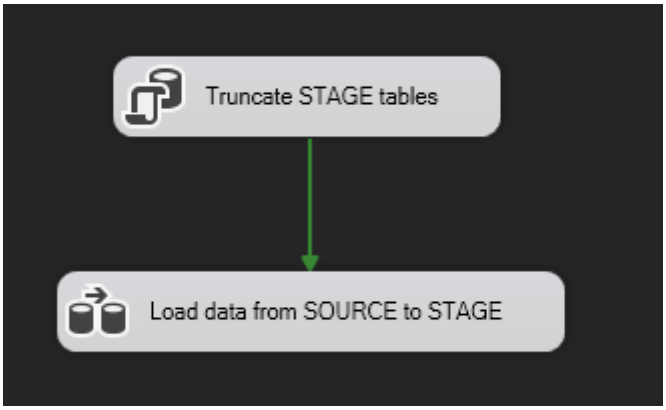
ETL FOLYAMATOK

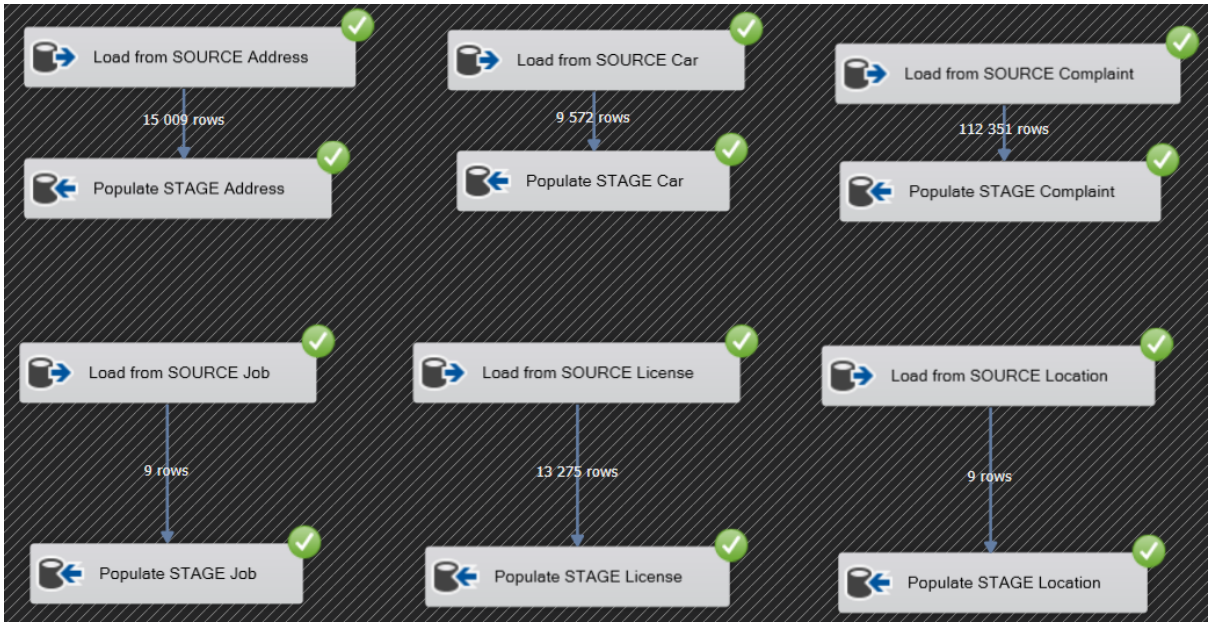
A következőkben a dimenzionális modell tábláiba való adatbetöltéshez szükséges ETL folyamatokat, valamint az egyes rétegek adatbázisait ismertetem.

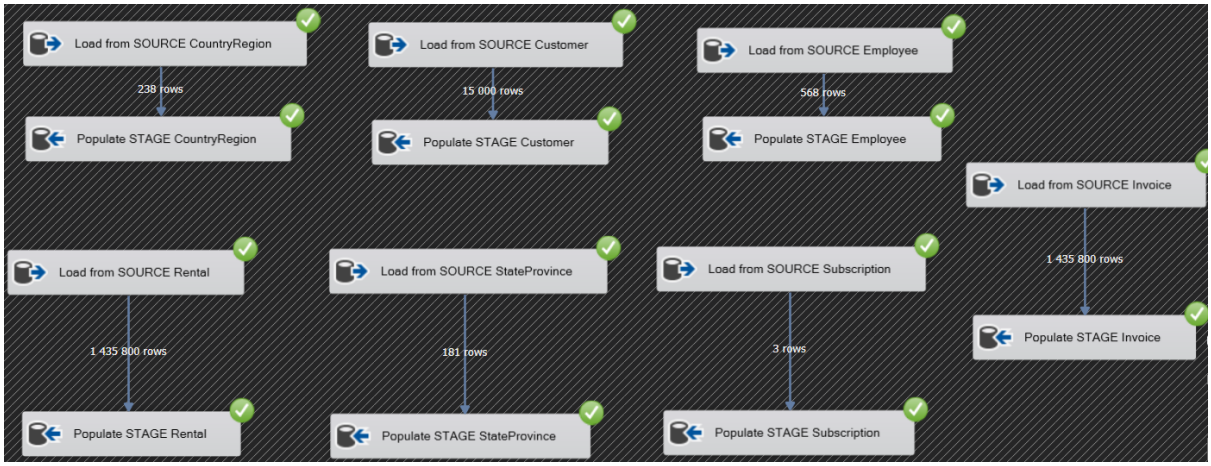
EXTRACT

Ebben a rétegben a forrásadatokat típuskonverzió nélkül, szöveges formában töltöm be a STAGE nevű adatbázis tábláiba annak érdekében, hogy a forrásrendszer a lehető legkevesebb ideig legyen az áttöltéssel terhelve, minél kevésbé akadályozza a szolgáltatások hatékony működését.

A forrásadatokat adatbázisból érem el, *OLE DB Source* elem segítségével, a betöltést megelőzően pedig kiürítem a STAGE adatbázis tábláit.







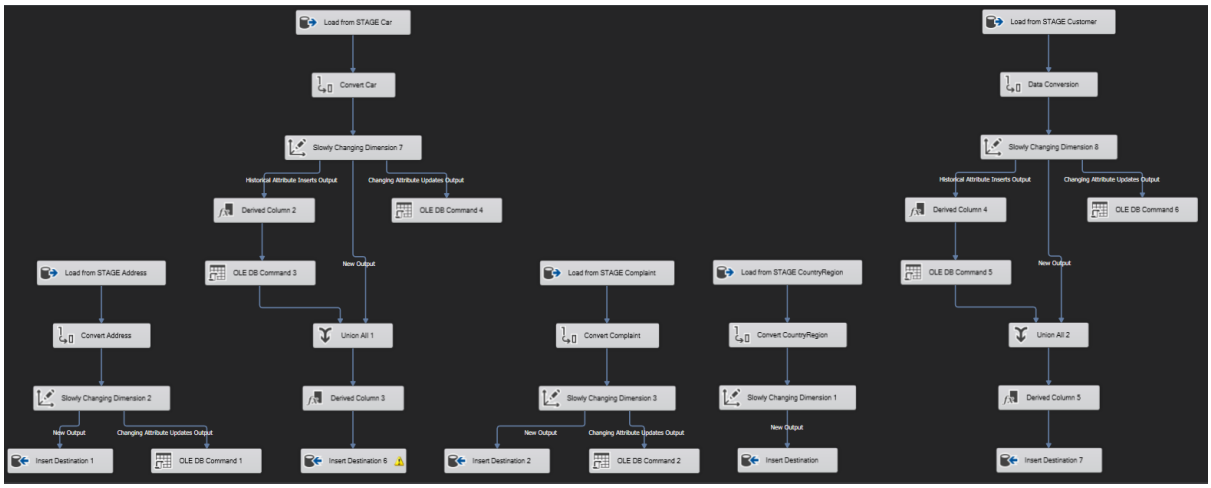
Mivel az adatbázisban nincs beállítva semmilyen megszorítás, ezek a folyamatok egymással párhuzamosan futnak, ezzel csökkentve a folyamat időigényét.

TRANSFORM

Ebben a rétegben a STAGE adattábláinak tartalmát áttöltöm az adattárházba, itt történik a megfelelő adattípusra konvertálás is, valamint a historizálás is. A típusok nagyrészt a forrásadatbázisnak megfelelően lesznek megválasztva, ettől csak a pénzösszegeket tartalmazó oszlopok esetén térek el, és ennek megfelelően *MONEY* típust állítok be.

Az alábbi táblákon SCD Type 2-t használok, mert ezeknél fontos, hogy az egyes rekordok változásai nyomon követhetők legyenek:

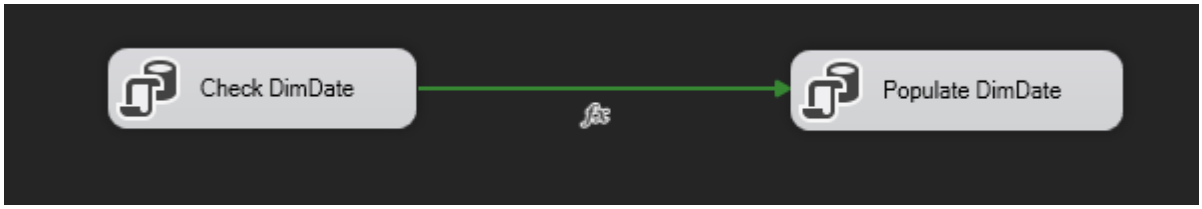
- Job,
- Employee,
- Location,
- Car,
- Subscription,
- Customer.



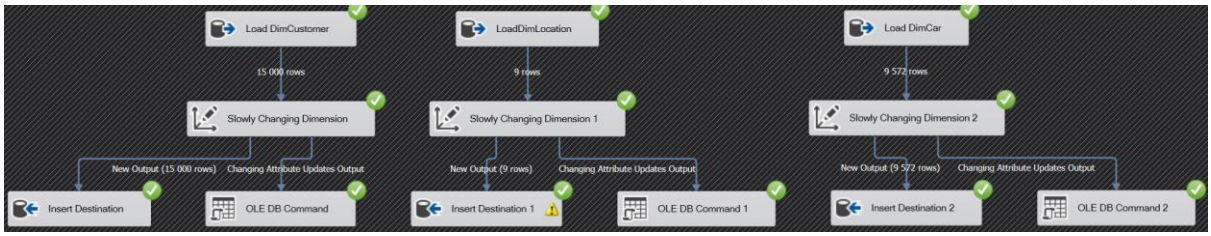
A kép a folyamat mérete miatt annak csak egy részhalmazát tartalmazza.

LOAD

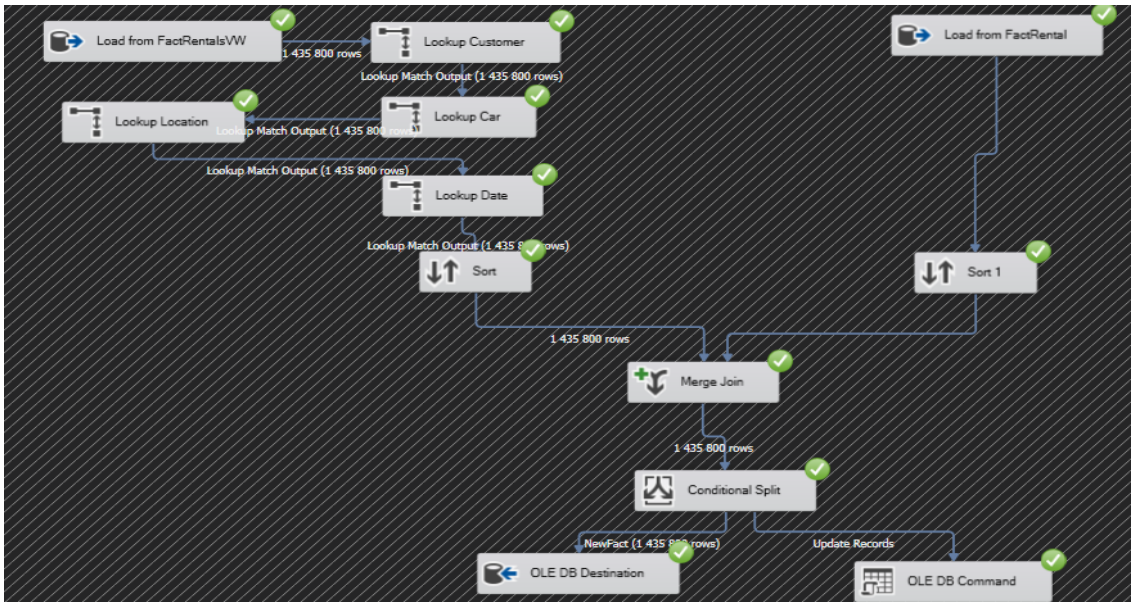
Itt történik meg a csillagséma feltöltése az adattárház alapján. Ehhez először létrehoztam a dimenziók és a ténytáblák feltöltéséhez szükséges nézeteket a DW adatbázisban. Ezt követően az SSIS-ben ellenőrzöm, hogy a dátum dimenzió fel van-e töltve a megfelelő adatokkal, és ha nincs, akkor egy tárolt eljárás segítségével ezt feltöltöm.



Ezután a nézetek alapján először feltöltöm a dimenziókat adatokkal:



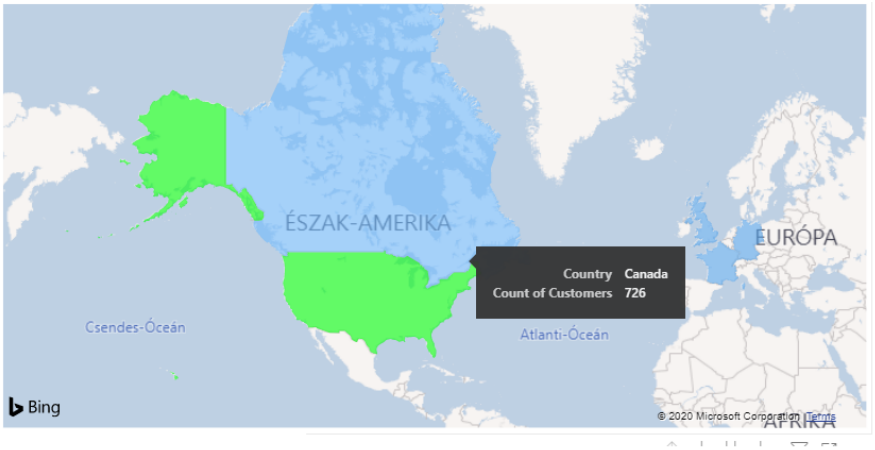
A dimenzió táblák feltöltését követően a ténytáblát is feltöltöm adatokkal. Ehhez először beolvasom az ideiglenes nézetből az adatokat, majd *Lookup* elemek segítségével minden tény rekordhoz megkeresem a hozzá tartozó dimenziók elsődleges kulcsát. Megvizsgálom, hogy a ténytábla már tartalmazza-e ezeket a rekordokat, ha nem, akkor beszúrom a ténytáblába, ha pedig igen, és csak módosítás történt, akkor *egy OLE DB Command* elem használatával lefuttatok egy *UPDATE* utasítást, ami frissíti az adott rekord oszlopait. A teljes folyamat az alábbi képen látható:



RIPORTKÉSZÍTÉS

A dimenzió és tényadatok feltöltését követően a use-case meghatározása során felmerült kérdésekre adok választ adatvizualizációk segítségével.

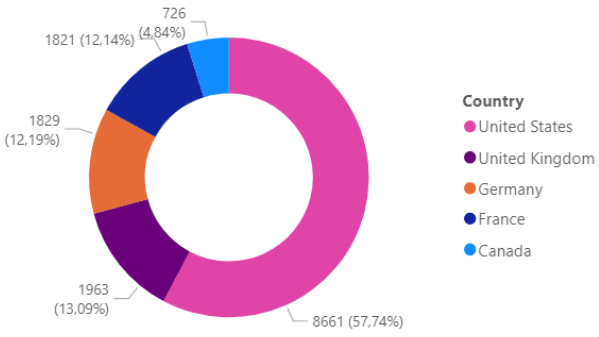
HOGYAN OSZLIK MEG AZ ÜGYFELEK SZÁMA TERÜLETENKÉNT?



Country	Count of Customers	Percent of Total Customers
Canada	**726**	**4,84%**
Alberta	12	0,08%
British Columbia	680	4,53%
Ontario	30	0,20%
Quebec	4	0,03%
France	**1821**	**12,14%**
Germany	**1829**	**12,19%**
United Kingdom	**1963**	**13,09%**
United States	**8661**	**57,74%**
Alabama	12	0,08%
Arizona	13	0,09%
California	4620	30,80%
Colorado	13	0,09%
Connecticut	10	0,07%
Florida	39	0,26%
Georgia	10	0,07%
Idaho	5	0,03%
Illinois	23	0,15%
Indiana	12	0,08%
Kentucky	4	0,03%
Maine	3	0,02%
Massachusetts	4	0,03%
Michigan	15	0,10%
Minnesota	4	0,03%
Mississippi	8	0,05%
Missouri	19	0,13%
Montana	3	0,02%
Nebraska	1	0,01%
Nevada	7	0,05%
New Hampshire	3	0,02%
New Mexico	4	0,03%
New York	14	0,09%
North Carolina	10	0,07%
Ohio	16	0,11%
Oregon	1083	7,22%
Összesen	**15000**	**100,00%**

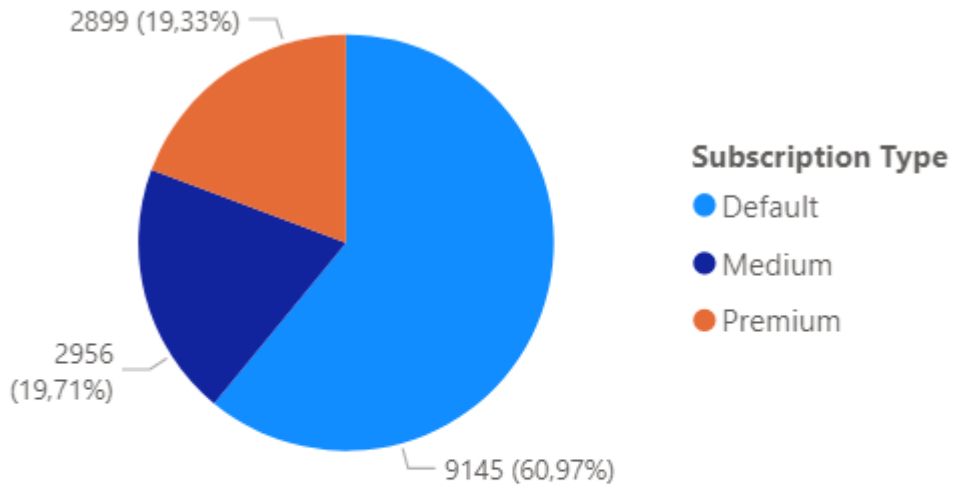
36

Average age of customers



A riport arra ad választ, hogy a cég ügyfelei hogyan oszlanak meg terület szerint. Az első képen térképen is jelölve vannak az országok, ahonnan vannak vásárlók, az ország színét az határozza meg, hogy hány ügyfél lakik ott. A színskála világoskék és világoszöld közötti, ahol világoszöld az az ország, ahol a legtöbb ügyfél van. Lehetőség van részletezni is az adatokat állam / tartomány, valamint ezen túl város szinten is. A jobb oldalt található táblázatból az is leolvasható, hogy az adott országban élő ügyfelek az összesnek mekkora részét alkotják.

MELYIK A LEGNÉPSZERŰBB ELŐFIZETÉSI FORMA?



Subscription Type	Amount	Age	Average Amount Per Rental
Default	$269 711 294	36,06	$308,3141
Medium	$43 877 312	35,55	$154,8547
Premium	$18 642 290	36,11	$67,1405
Összesen	**$332 230 896**	**35,97**	**$231,3908**

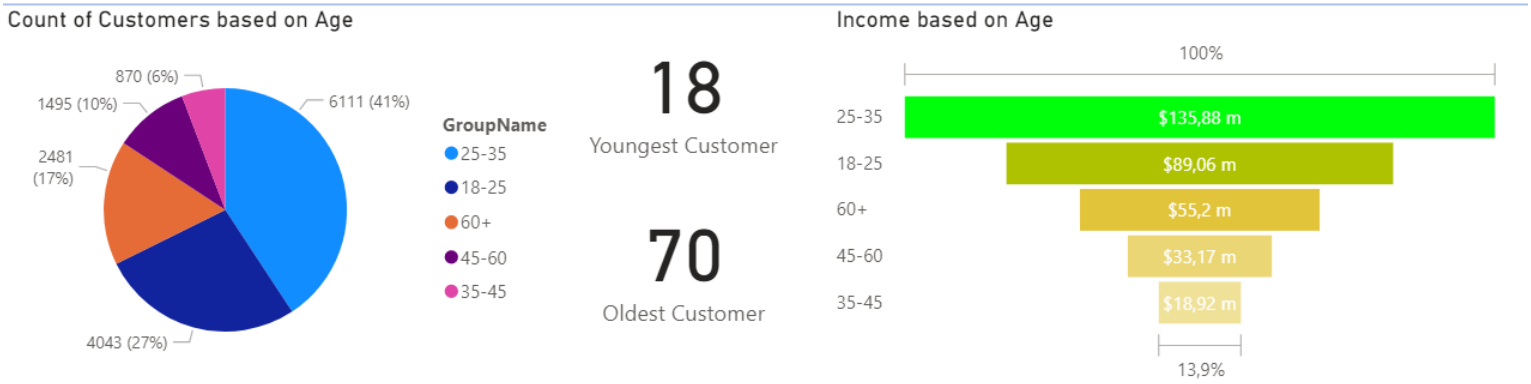
A vizualizáción az látható, hogy az alapértelmezett, havidíj nélküli, de magas percdíjjal rendelkező konstrukció a legnépszerűbb, ezt használja a legtöbb ügyfél, valamint ezzel arányosan a bevétel legnagyobb része is ilyen típusú bérlésekből származik. Az egyes előfizetési formákat használó ügyfelek átlagos életkorában nem figyelhető meg számottevő eltérés.

HOGYAN OSZLIK MEG A HIBAJELENTÉSEK SZÁMA A SZOLGÁLTATÁSI TERÜLETEN?

State	2017	2018	2019	2020	Összesen
Arizona	3803	3977	3792	672	**12244**
California	7882	7958	7833	1458	**25131**
Colorado	3811	3949	3887	682	**12329**
Nevada	3955	4034	3938	755	**12682**
New Mexico	3988	4004	3868	741	**12601**
Oregon	3770	3726	3745	723	**11964**
Texas	3975	3979	3982	726	**12662**
Washington	3994	4054	3946	744	**12738**
Összesen	**35178**	**35681**	**34991**	**6501**	**112351**

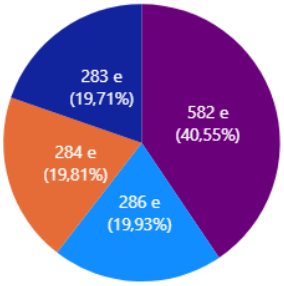
Az eddigi adatok alapján California államban jelentősen több panasz érkezik, azonban mivel ott két telephely is található, így összességében nem figyelhető meg jelentős eltérés az átlagtól.

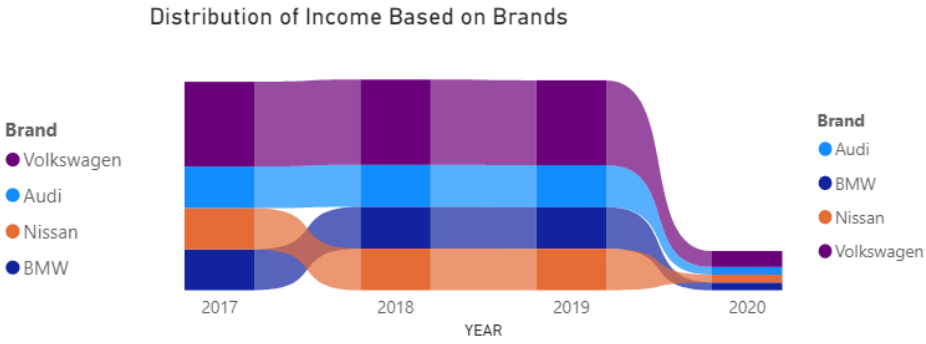
MELYIK KORCSOPORTOK KÖRÉBEN LEGNÉPSZERŰBB A SZOLGÁLTATÁS?



A riporton látható, hogy a 25-35 közti korosztályban a legnépszerűbb a szolgáltatás, innen érkezik a legtöbb darabszámú bérlés, valamint a legtöbb bevétel is innen származik.

MELYIK MÁRKÁK A LEGNÉPSZERŰBBEK?





Az első képen az látható, hogy a bérlések darabszámát tekintve a Volkswagen a legnépszerűbb márka a felhasználók körében. A második képen évekre lebontva láthatjuk az egyes márkák járműveiből származó bevételt.