2. harjutustund (11.09)

Kava

- Õppeserveri andmebaaside lühitutvustus
  - Tarkvara ülevade, Microsoft SQL lühitutvustus
  - SQL päringukeele meeldetuletus
- Andmete ettevalmistamine andmelao koostamiseks
  - Denormaliseerimine.
  - Andmete puhastamine, kodeeritud väärtused ja abitabelid (*lookup tables*)
  - Võtmete ümberkodeerimine, surrogaatvõtmed.

Õppeserveri andmebaaside tutvustus

Õppeserverile ligipääs läbi *Remote Desktop Connection*i (mstsc.exe):

Serverinimi: ararat.ttu.ee
Kasutajanimi: intra\kasutaja ja psw

Andmebaasiserver (Microsoft SQL Server 2008 R2 Enterprise):

Serverinimi: IDBI\SQLDIMENSIONAL
Autentimine: Use Windows Authentication

Vahendid:

Microsoft SQL Server Management Studio (edaspidi *SSMS*)
SQL Server Business Intelligence Development Studio (edaspidi *BIDS*)

Andmebaasid:

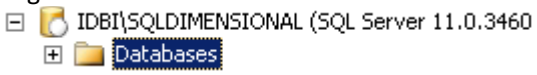
AdventureWorks2012 - sisaldab näidisandmeid jalgrattaid tootva ettevõtte kohta

Microsoft SQL töövahendite lühitutvustus

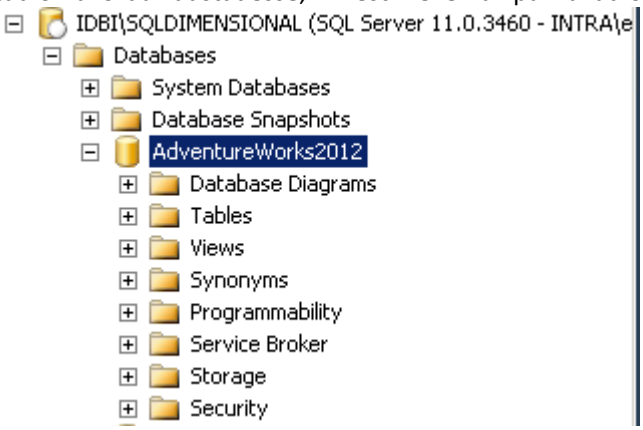
Päringute koostamiseks ja andmebaaside disainimiseks kasutame vahendit *SQL Server Management Studio* (asub ararat.ttu.ee töölaual).
Vahendi käivitamisel määrame andmebaasi, millega ühendus luuakse.

Andmebaasid

Nupp Connect loob ühenduse ning kuvab *Object Explorer* paani, milles on loetletud andmebaasid ning teised serveri objektid. Avame kausta *Databases* ning otsime üles demoandmebaasi **AdventureWorks2012**.



Andmebaasiobjektid on grupeeritud omakorda kaustadesse, millest meile huvipakkuvad on *Tables* ning *Views*.

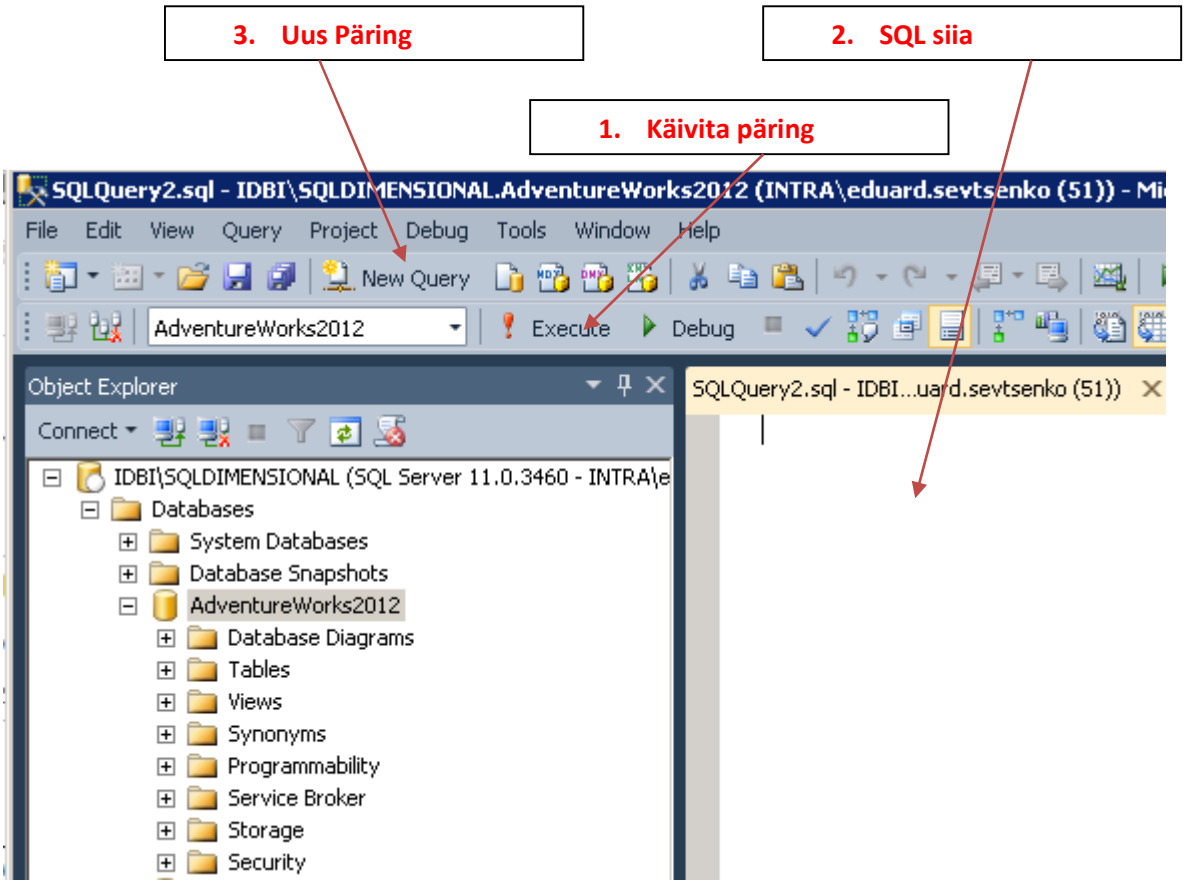


Tabelite loetelu saame, kui avame *Tables* kausta sisu. Tabelitega saab teha parema hiire kontekstimenüüst järgmisi operatsioone:

- Design -- avab tabeli disainivaates, kus saame muuta selle struktuuri (nt lisada väljasid).
- Select Top 1000 Rows -- koostab tabeli sisu kuvamiseks SELECT päringu.
- Edit Top 200 Rows -- avab tabeli kirjed muudetavas loendis, mille kaudu on mugav ilma SQLi oskamata andmetega manipuleerida.

Päringute käivitamine

Uue päringu loomiseks käivitage menüüst *File New Query With Current Connection* või klikkige tööriistaribal nupule New Query. Avaneb SQL redaktor. Päringu käivitamiseks käivitage menüüst *Query Execute* või klikkige tööriistaribal Execute nupule (F5).



Tabelid ja skeemid

MSSQLis kuulub tabel alati ühe skeemi (*schema*) koosseisu. Tabeli identifitseerimiseks tuleb näidata skeem ja tabeli nimi, nt. **Sales.Currenct** . Valuutakoodide küsimiseks näitame päringus mõlemad komponendid:

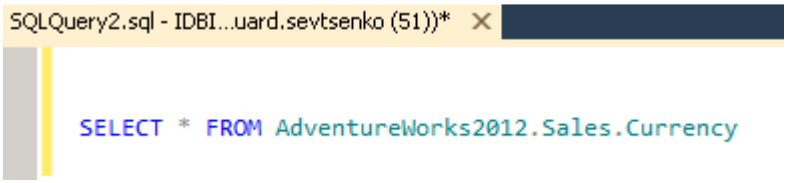
```
SELECT * FROM Sales.Currency
```

Vaikimisi skeemiks on **dbo**, mille võib päringutes ka ära jätta. Järgmised kaks päringut on samaväärsed

```
SELECT * FROM dbo.DatabaseLog
SELECT * FROM DatabaseLog
```

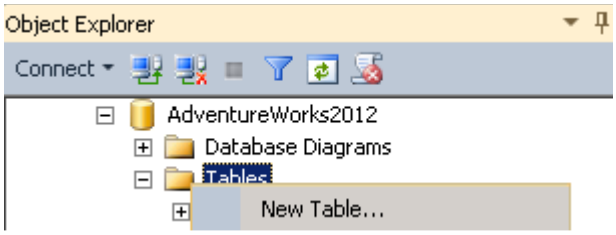
Enne päringu käivitamist määratakse alati tööandmebaas (SSMS-is Ctrl+U). Valitud tööandmebaasi näeme *Standard* tööriistaribal:

Kui soovime päringuid teha mitme andmebaasi vahel, nt. tööandmebaasiks on meie personaalne andmebaas, aga tahame andmeid küsida AdventureWorks baasist, näidatakse tabelinime ees ka andmebaasi nimi.

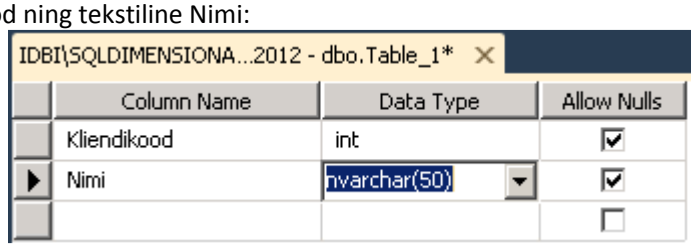


Tabelite defineerimine

Uute tabelite loomisel ja olemasolevate struktuuri muutmisel on mugav kasutada disaineri abi. Proovime luua oma AdventureWorks2012 andmebaasis uue tabeli. Selleks avame andmebaasis kausta *Tables* ning parema hiire kontekstimenüüst valime *New Table*.



Avaneb disainer, kus saame defineerida tabeli veerud ning nende andmetüübid. Loome lihtsa tabeli, mis koosneb kahest veerust: numbriline Kliendikood ning tekstiline Nimi:

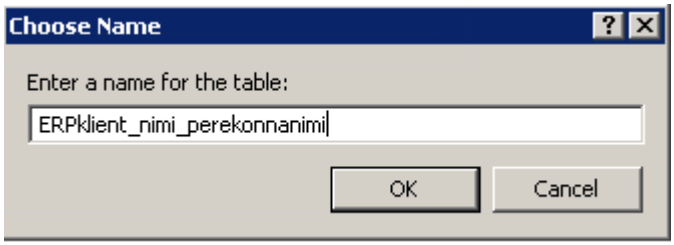


Andmetüüpideks on soovitav kasutada:

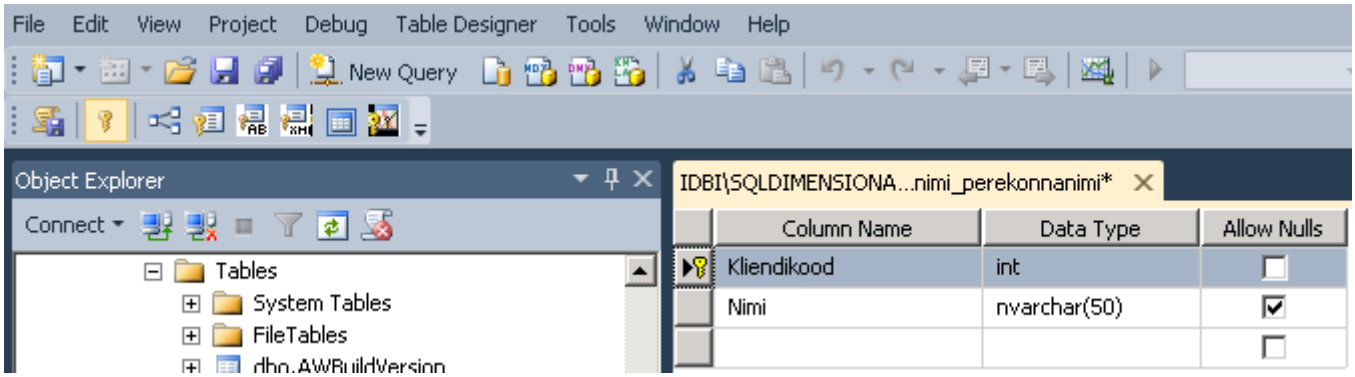
- INT - arvulised väärtused
- NVARCHAR(50) - tekstilised väärtused Unicodes (nt. 50 tähemärki)
- FLOAT - ujukomaarvud
- DECIMAL(28, 2) - täpselt salvestatavad väärtused (nt. 28 positsiooni , 2 kohta pärast koma)
- DATETIME - kuupäev, kellaaeg või mõlemad koos

Täpsemalt andmetüüpide kohta, vt. Help Index otsisõna *data types* data types [SQL Server].

Salvestame tabeli: käivitame *File* *Save Table1* või tööriistaribal nupp . Anname tabelile nimeks ERPklient_nimi_perekonnanimi



Eelnevalt ununes tabelis primaarvõti. Selle määramiseks valime veerud, mis peaksid koosnema primaarvõtme koosseisu ning klikime tööriistaribal võtmekese ikooni .



Kui saame salvestamisel vea *Saving changes not permitted...*, siis peame seadistama SSMS, et see lubaks tabelite muutmisel nende kustutamist-taasloomist. Jätame muudatused salvestamata ning avame Tools Options.

Otsime puust üles *Designers Table and Database Designers* ning eemaldame linnukese *Prevent Saving changes that require table re-creation.* Pärast parameetrite muutmist salvestame tabeli uuesti ja muudame parameetri hiljem tagasi.

SQL päringukeele meeldetuletus

Järgnevalt on toodud spikker mõnede SQL koodinäidetega, mida läheb tarvis käesoleva ning järgmiste harjutuste juures.

Andmete pärimine tabelist:

```
SELECT *
FROM Person.CountryRegion
```

Päring koos veeru ümbernimetamisega:

```
SELECT Name AS CountryName
FROM Person.CountryRegion
```

Päring tabelist koos kirjete arvu piiramisega:

```
SELECT TOP 100 *
FROM Person.CountryRegion
```

Päring teisest andmebaasist (tööandmebaas olgu sama andmebaasi):

```
SELECT *
FROM AdventureWorks2012.Person.CountryRegion
```

Kahe tabeli ühendamine (*join*):

```
SELECT C.Name, C.CountryRegionCode, S.Name, S.StateProvinceCode
FROM Person.CountryRegion C
       INNER JOIN Person.StateProvince S ON C.CountryRegionCode = S.CountryRegionCode
```

Päringu tulemuste tabelisse salvestamine (oma tabelisse Country_Nimi):

```
SELECT *
INTO Country_Eduard
FROM AdventureWorks.Person.CountryRegion
```

Tabeli kustutamine - kustutab tabelist kirjed ja tabeli struktuuri (kustutada ainult oma tabeli)

```
DROP TABLE Country_Eduard
```

Vaate loomine (salvestada oma nimega vwCountry_Nimi). Vaade on "salvestatud päring", kuhu saab talletada korduvkasutuseks sobivaid SQL-lauseid

```
CREATE VIEW vwCountry_Eduard AS
SELECT Name
FROM AdventureWorks.Person.CountryRegion
```

Vaate muutmiseks käivitage ALTER VIEW ... päring uue sisuga.

```
ALTER VIEW vwCountry_Eduard AS
SELECT CountryRegionCode, Name
FROM AdventureWorks.Person.CountryRegion
```

Vaate kustutamine – kuna vaade ise andmeid ei sisalda, siis kaob ainult päringu moodustanud SQL käsk

```
DROP VIEW vwCountry_Eduard
```

Grupeerimine ja kirjete arvu loendamine (kasulik andmete profileerimisel nt. unikaalsete väärtuste otsingul):

```
SELECT CountryRegionCode, COUNT(*) AS NumberOfStates
FROM Person.StateProvince
GROUP BY CountryRegionCode
```

Tingimuslaused väärtuste asendamiseks ja väljade kombineerimiseks:

```
SELECT
     CASE WHEN CountryRegionCode = 'EE' THEN 'Eesti'
          WHEN CountryRegionCode = 'LV' THEN 'Läti'
          ELSE Name END AS CountryRegionName
FROM Person.CountryRegion
```

Puuduva väärtuse (*NULL*) asendamine:

```
SELECT TOP 1000 AddressLine2,
       ISNULL(AddressLine2, '?') AS AddressLine2_Uus
FROM Person.Address
```

Tabeli loomine CREATE-lausega (luua oma nimega Region_Nimi)

```
CREATE TABLE Region_Eduard(
       ID INT PRIMARY KEY,
       Name NVARCHAR(50),
       Country NVARCHAR(3) )
```

Tabeli kirjete kuvamine

```
SELECT *
FROM Region_Eduard
       )
```

Tabelisse kirjete lisamine teise päringu tulemuste põhjal (lisage oma tabelisse):

```
INSERT INTO Region_Eduard (ID, Name, Country)
SELECT TerritoryID, Name, CountryRegionCode
FROM AdventureWorks.Sales.SalesTerritory
```

Tabelist kirjete kustutamine (tabel jääb alles):

```
DELETE FROM Region
```

Harjutus 2

Koostage päring, mis kuvab piirkondade lõikes müügitulemused: tellimuste arvu ning müügikäibe.

- Andmed paiknevad baasi **AdventureWorks2012** tabelites **Sales.SalesOrderHeader** (müügitellimuste päised), **Sales.Customer** (kliendid) ning **Sales.SalesTerritory** (riigid-territooriumid).
- Riikide koodid tuleb teisendada piirkonna nimetuseks vastavalt järgmisele tabelile:

US	United States
CA	Canada
FR	European Union
DE	European Union
GB	European Union

- Tulemus salvestage andmebaasi vaatena **vwOrders_Nimi_Perekonnanimi**, mida on mugav käivitada järgmise SQL lausega:

```
SELECT * FROM vwOrders_Eduard
```

- Õige vastus "näeb välja" järgmine:

```
SELECT * FROM vwOrders
```

Results | Messages

	Region	SalesTotal	Orders
1	Canada	21501812.4574	4067
2	European Union	24582915.5599	8514
3	United States	82425341.2779	12041
4	Unknown	12197515.5294	6843

- Piirkondade teisendusreeglid on soovitatav salvestada tabelina oma nimega andmebaasis (st. ärge kasutage CASE WHEN teisendust).
  - Kui riigi kohta puudub teisendusreegel, määratakse piirkonnaks "Unknown".
  - Lisage teisendustabelisse reegel AU Australia and Oceania ning kontrollige, kas päringust kadus liigitus "Unknown".

Andmete ettevalmistamine andmelao koostamiseks

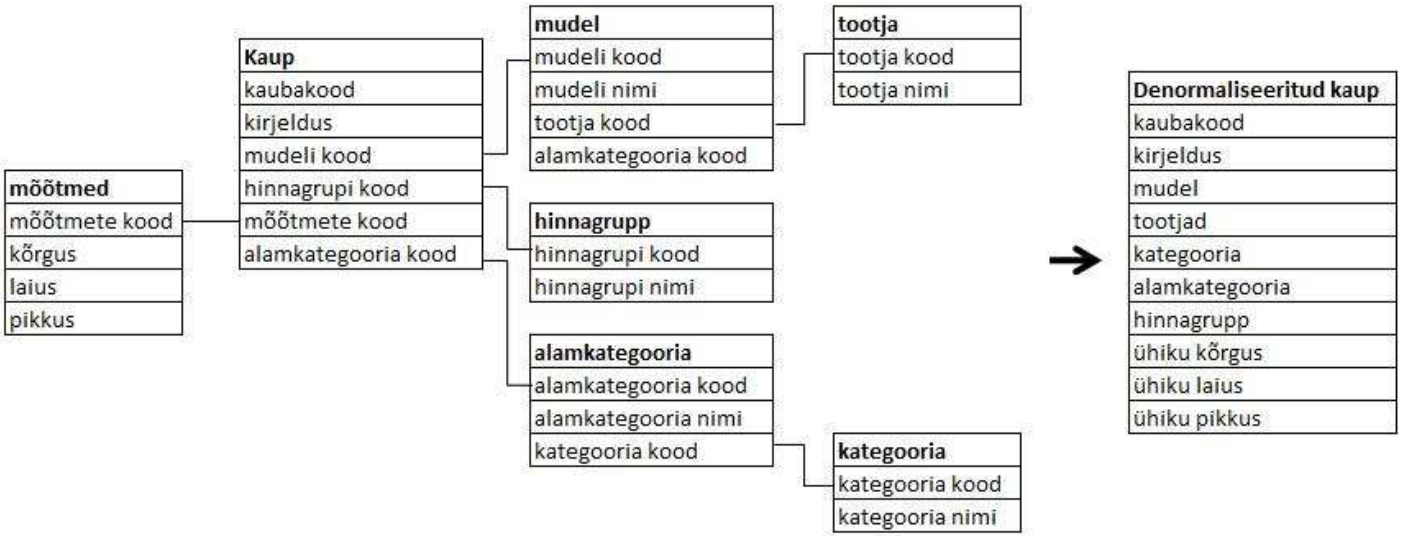
Denormaliseerimine

Denormaliseerimine on relatsioonilise andmebaasi normaalkuju tahtlik rikkumine, mida kasutatakse sageli nt. andmeladude disainimisel. Denormaliseerimise abil üritatakse saavutada järgmisi eesmärke:

- Andmete pärimisel tabelite ühendamiste (JOIN) arvu vähendamine – andmete lugemise kiirus.
- Andmebaasiskeemi lihtsustamine –meta-andmed on mõistetavad ka ilma diagrammideta.

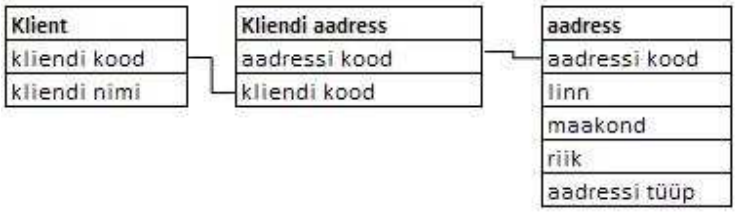
Millal on mõistlik tabelid denormaliseerida?

- N:1 suhte korral: võib denormaliseerida kõik välisvõtmetega määratud kirjete atribuudid. Vt. Joonis 1.



Joonis 1 - normaliseeritud ja denormaliseeritud "kaup"

- 1:N suhte korral: andmete kohta kehtivad kindlad piirangud (*constraints*), mida tõenäoliselt ei rikuta. Nt. Klient – Kliendi aadress mustri (Joonis 2) puhul on teada, et igal kliendil on ainult üks põhiaadress ja ainult üks elukoht.



Joonis 2 – "klient-kliendi aadress" muster

Harjutus 3 (0,5p)

Kuidas kirjeldada aega (kuupäevad või kellaajad) denormaliseeritud kujul? Näiteks olgu meil järgmine tabel:

Kuupäev
25.12.2011
26.12.2011
27.12.2011
28.12.2011
29.12.2011
30.12.2011
31.12.2011
1.01.2012
2.01.2012

Koostage Excelis kuupäevatabel ning lisage sinna veerud, mille abil saab mugavalt teada:

- nädalapäeva
- nädala numbri
- kuu nimetuse

Milliseid kuupäeva kohta käivaid tunnuseid peaks see tabel veel sisaldama, kui soovime seda kasutada nt:

- müükide ja kampaaniate analüüsis
- palgaarvestuses

Harjutus 4 (0,5p)

Soovime andmeaita luua kaupluse dimensiooni, mis sisaldaks andmeid kaupluse enda kohta (nimi), kaupluse omaniku kohta ning seal opereeriva müügimehe kohta. Hetkel on need infokillud tabelites laiali ning andmete koondamiseks tuleks tabel **Sales.Store** denormaliseerida.

- Koostage SQL-päring, mis ühendab vastavad tabelid. Seosed leiate **AdventureWorks** andmebaasi diagrammilt.
- Päringu põhjal looge oma isiklikku vaade **vwStores**_Nimi_Perekonnanimi

Kuna eesmärk on koostada lähteandmed dimensiooni tarbeks, siis päringu tulemuses peaks olema igat kauplust (st. *CustomerID* väärtust) 1 kord. Samas seostabel **Sales.StoreContact** sisaldab ühe kaupluse (*CustomerID*) kohta mitut kirjet.

Vihjed:

- Omaniku leidmiseks tuleb liikuda: **Sales.Store** -> **Person.BusinessEntityContact** -> **Person.Person**.
  - Kauplusel võib olla mitut tüüpi kontaktisikuid. Tüüpide kirjeldused leiate **Person.ContactType**.
- Müügimehe puhul on tegemist ettevõtte töötajaga (**HumanResources.Employee**), kuid kõikide isikute nimed leiab tabelist **Person.Person**.
  - Tabelite ühendamiseks tuleks liikuda: **Sales.Store** -> **Sales.SalesPerson** -> **Person.Person**
- Väärtuste unikaalsust (mittekorduvust) saab lihtsalt kontrollida grupeerides tulemused vastava veeru järgi ning rakendades filtri pärast grupeerimist. Näide:

```
SELECT CountryRegionCode, COUNT(*) AS kordusi
FROM Sales.SalesTerritory
GROUP BY CountryRegionCode
HAVING COUNT(*) > 1
```

- Lõpptulemuses tuleb kontrollida, kas *CustomerID* on unikaalne.

- Õige vastus "näeb välja" järgmine:

```
Select *
FROM vwStores_Eduard Order by CustomerID
```

100 %

Results | Messages

	CustomerID	StoreName	OwnerName	OwnerSurname	SalesPersonName	SalesPersonSurname
1	292	Next-Door Bike Store	Gustavo	Achong	Tsvi	Reiter
2	294	Professional Sales and Service	Catherine	Abel	Linda	Mitchell
3	296	Riders Company	Kim	Abercrombie	Jillian	Carson
4	298	The Bike Mechanics	Humberto	Acevedo	Michael	Blythe
5	300	Nationwide Supply	Pilar	Ackerman	Lynn	Tsoflias
6	302	Area Bike Accessories	Frances	Adams	Shu	Ito
7	304	Bicycle Accessories and Kits	Margaret	Smith	David	Campbell
8	306	Clamps & Brackets Co.	Carla	Adams	Michael	Blythe
9	316	Fun Toys and Bikes	Robert	Ahlering	Shu	Ito
10	318	Great Bikes	François	Ferrier	David	Campbell
11	320	Metropolitan Sales and Rental	Kim	Akers	Michael	Blythe
12	322	Irregulars Outlet	Lili	Alameda	Rachel	Valdez
13	324	Valley Toy Store	Amy	Alberts	José	Saraiva
14	326	Worthwhile Activity Store	Anna	Albright	Tsvi	Reiter

Andmete puhastamine

- Lähte-andmebaasidest pärinevad andmed ei pruugi olla alati ühtlase kvaliteediga. Tüüp-probleemid:
  - Osadel kirjetel ei ole kõik atribuudid väärtustatud (NULL-väärtused).
  - Kui lähte-andmebaas ei ole normaliseeritud 3.normaalkujuni, siis tüüpiliselt on vead korduvates väärtustes (Joonis 3).

Klient	Linn
Peeter	Tallinn
Maie	tallinn
Marju	TALLINN
Kalle	tallinn
Kaspar	TALlinn
Vello	tallin
Helen	Tartu
Pille	Tartu

Joonis 3 - ebapiisava normaliseerimise tõttu tekkinud vead

  - Osad atribuudid võivad lähtetabelis olla ainult kodeeritud kujul. Nt. sugu: "M" / "F". Andmete esitamisel lõppkasutajale on mõistlik koodid asendada kirjeldavate nimetustega ("mees" / "naine"). Tüüpiliselt on andmebaasides kodeeritud kujul erinevad klassifikaatorid, mille nimetusi ei pruugi samas andmebaasis alati leiduda.
  - Lähteandmebaasi halva disaini korral tüüp-probleemiks välisvõtmete puudumine (*referential integrity*). Tulemuseks on seostamata kirjed (*nonmatched foreign key*) tehingutabelites.

- Lahendused:
  - Puuduva väärtuse haldus: NULL-väärtuse asendamine kokkulepitud väärtustega. Nt. "(määramata)", "(unknown)", "(--)" vm.
  - Kui lähte-andmebaasis ei ole võimalik andmeid parandada (puudub kirjutamisõigus vm.), siis vigased väärtused kogutakse otsingutabelitesse (*lookup tables*) – vt. Joonis 4.

Create

Linn	Korrektne_linn
Tallin	Tallinn
talin	Tallinn
tallinn	Tallinn
talinn	Tallinn
TLN	Tallinn
TL	Tallinn
TRT	Tartu
Tartu	Tartu
Trtu	Tartu

Joonis 4 - vigaste väärtuste leidmine otsingutabelist

  - Otsingutabeleid saab kasutada ka kodeeritud väärtustele nimede leidmisel.
  - Seostamata välisvõtmete puhul saab kasutada kokkulepitud väärtuseid, nt. "(Vigane)".
  - *SQL Server Integration Services* andmete laadmise paketis on komponent *Fuzzy Grouping*, millega saab sarnaseid tekstilisi väärtuiseid grupeerida.

Harjutus 5 (1p)

Andmebaasis on tabel **Aadress**, mis sisaldab klientide aadressiinfot. Andmed on sisestatud käsitsi ning võivad olla vigased. Koostage SQL vahenditega päringud/vaated, millega saaks aadressiandmed korrektseks.
Vihjeks: looge seostabel, mis "tõlgib" vigased väärtused korrektseteks. Salvestage seostabel kujul CorrectValues_Nimi_Perekonnnanimi

Salvestage vastus oma andmebaasis vaadetena, mida saab pärida kujul:
SELECT * FROM vwCleanAddresses_Nimi_Perekonnanimi

- Õige vastus "näeb välja" järgmine:

```
SELECT *
FROM vwCleanAddress_Eduard
```

100 %

Results | Messages

	Name	City	CorrectCity
1	Helen	Tartu	Tartu
2	Kalle	tallinn	Tallinn
3	Kaspar	TALinn	Tallinn
4	Krete	NULL	Unknown
5	Maie	tallinn	Tallinn
6	Marju	TALLINN	Tallinn
7	Peeter	Tallinn	Tallinn
8	Pille	TARTU	Tartu
9	Vello	tallinn	Tallinn

Võtmete ümberkodeerimine

- Eesmärk : vähendada sõltuvust lähte-andmebaaside primaarvõtmetest.
  - Kui lähtetabeli primaarvõti on naturaalne võti (loodud reaalse maailma atribuutide põhjal), siis see võib muutuda.
  - Kui andmed pärinevad mitmest allikast (nt. kliendid ERP ja veebiportaali andmebaasidest), siis võib tekkida olukord, kus kirjete vahel on mittevastavused.
  - Lahenduseks on kunstlikult genereeritud primaarvõti e. surrogaatvõti (*surrogate key*), mis on garanteeritult unikaalne ja samatähenduslik.

ERP		Veebiportaal			Genereeritud võti
Kliendi kood	Nimi	CustomerID	Name	Kliendikood_IN_ERP	CustomerKey
4325	Peeter	X1357	Peeter	4325	1
262	Malle	X1062	Malle	262	2
		X0235	Kalle		3
52	Marju				4
		Y2135	Helen		5
24	Vello	Y3131	Vello	24	6

Joonis 5 – mitmest allikast pärinev klienditabel

- Võtmete genereerimine eeldab sobiva programmi või päringu kirjutamist (SQL funktsioonid ROW_NUMBER() vm). Kui andmete transformeerimise platvormina kasutada *Microsoft SQL Server Intergation Services*'t, siis on üks lahendus vabavaraline komponent: Kimball Method SSIS Slowly Changing Dimension Component.

Harjutus 6 - iseseisvaks nuputamiseks (1p)

Kuidas genereerida võtmeid nii, et igale lähtetabeli kirjele vastaks alati sama surrogaatvõti? Eeldame, et lähtetabelisse tekkib ajapikku kirjeid juurde.

ERP		Veebiportaal			Genereeritud võti
Kliendi kood	Nimi	CustomerID	Name	Kliendikood_IN_ERP	CustomerKey
4325	Peeter	X1357	Peeter	4325	1
262	Malle	X1062	Malle	262	2
941	Viktor				???
		X0235	Kalle		3
52	Marju				4
		Y2135	Helen		5
Z074	Mihkel	Z1065	Mihkel	Z074	???
		Z074	Pille		???
24	Vello	Y3131	Vello	24	6

Andmebaasis **AdventureWorks2012** on tabelid **ERP** ja **Veebiportaal**, mis sisaldavad erinevate lähtesüsteemide kliendiandmeid.

- Koostada tabel, mis seostab igale kliendile genereeritud kliendikoodi (*CustomerKey*). NB! Olemasolevaid tabeleid ei tohi muuta. Tabelis on ainult surrogaatvõtmed 1-6 (vt. joonist ülalt, valge taustaga kirjed).
- Koostage päring, mis ühendab EPR ja Veebiportaal tabelid (sisaldavad ka uusi kirjeid, märgitud roosaga) ning arvutab seostamata klientidele uued unikaalsed surrogaatvõtmed.
  - Vihjeks: kirjetele järjenumbrite genereerimiseks võite kasutada funktsiooni ROW_NUMBER(). Näide:

```
SELECT Name,
    ROW_NUMBER() OVER (ORDER BY Name)
FROM AdventureWorks2012.Sales.SalesTerritory
```

- Salvestage päring vaatena vwERPWebCustomer_Nimi_Perekonnnanimi ning testige, kas see töötab ka uute kirjete lisamisel ERP / Veebiportaal tabelisse.

Materjalid

- Video: Getting Started with T-SQL Queries using SQL Server Management Studio. http://www.youtube.com/watch?v=V4hTLja7jU8
- SQL Server 2008 How-To-Guide videos http://learningsqlserver.wordpress.com/2011/03/09/sql-server-2008-express-how-to-guide-videos

- The data warehouse toolkit : the complete guide to dimensional modeling, 2nd edition. *Ralph Kimball, Margy Ross*. Wiley, New York, 2002. Chapter 1.
- Is ER Modeling Hazardous to DSS? Ralph Kimball. http://www.kimballgroup.com/1995/10/01/is-er-modeling-hazardous-to-dss (07.09.2012).
- Surrogate Keys - Keep control over record identifiers by generating new keys for the data warehouse. Ralph Kimball. http://www.kimballgroup.com/1998/05/02/surrogate-keys (07.09.2012).
- Denormalization Guidelines. Craig S. Mullins. http://www.tdan.com/view-articles/4142 (07.09.2012).