Bilješke

- Shopovi dolaze iz mog csv fajla
  shop, grupacija kojoj pripada, država, valuta
  `amazon.co.uk,amazon,United Kingdom,GBP`
  - ETL dodaje ID

```
+--------+-------------+--------+----------------+----------+
| shopID | shopName    | group  | country        | currency |
+--------+-------------+--------+----------------+----------+
|      1 | amazon.co.uk | amazon | United Kingdom | GBP      |
```

- Datumi dolaze iz excel fajla

http://www.kimballgroup.com/wp-content/uploads/2014/03/Ch3-SampleDateDim.xls

  - ETL uzima samo dio, te excel datum prikaže kako spada
    ❣ no formula ne odgovara specki, ne znam zašto :/ (?)

```
+----------+------------+-------------+------------------+----------+------------+
| dateID   | full_date  | day_of_week | day_num_in_month | day_name | day_abbrev |
+----------+------------+-------------+------------------+----------+------------+
| 20150101 | 2015-01-01 |           4 |                1 | Thursday | Thu        |

--------------+------------------+-------+------------+--------------+
 weekday_flag | week_num_in_year | month | month_name | month_abbrev |
--------------+------------------+-------+------------+--------------+
 Weekday      |                1 |     1 | January    | Jan          |

---------+------+--------+-----------------------+
 quarter | year | yearmo | last_day_in_month_flag |
---------+------+--------+-----------------------+
       1 | 2015 | 201501 | Not Month End          |
```

- Konverzije dolaze s HNB sajta

http://www.hnb.hr/temeljne-funkcije/monetarna-politika/tecajna-lista/tecajna-lista

  - ETL preračuna sve na 1 jedinicu (HUF je recimo na 100)

```
+----------+----------+-------------+
| dateID   | currency | averageRate |
+----------+----------+-------------+
| 20150101 | EUR      |  7.65770800 |
```

- Customere generiram iz talendovih podataka > u bazu bankSource

```
+------------+----------+----------+
| customerID | name     | surname  |
+------------+----------+----------+
|          1 | Abraham  | Harrison |
```

  - ETL samo loada

- Transakcije generiram samo s idevima > u bazu bankSource
  - moj generator za cijene
  - moj generator za id datuma
    ❣ ne pokriva točno period, jer ide samo do 28 za veljaču odnosno 30 za ostale mjesece

```
+---------------+----------+------------+--------+--------+
| transactionID | dateID   | customerID | shopID | amount |
+---------------+----------+------------+--------+--------+
|             1 | 20150101 |          7 |      8 |  53.99 |
```

  - ETL vuče valutu iz shopIDa pa onda spaja datumID i valutu, da bi dobio tečaj na dan
    ★ valueHRK je vrijednost transakcije prema tečaju na dan transakcije

```
+---------------+----------+------------+--------+--------+----------+-----------+
| transactionID | dateID   | customerID | shopID | amount | currency | amountHRK |
+---------------+----------+------------+--------+--------+----------+-----------+
|             1 | 20150101 |          7 |      8 |  53.99 | USD      |    340.11 |
```

♬ Bilo bi super da se može transakcije mogu generirati potpuno dinamički - prema broju shopova i prema broju generiranih korisnika, da se u generatoru transakcija te vrijednosti povuku iz baze / fajlova, a ne da su hardkodirane
♬ Postoje context varijable no nisam uspjela prokljuviti da li bi ih se dalo iskoristiti za ovo te kako točno

(?) Convert excel dates into sql dates

Kod izvlačenja full_datea iz excel one tablice
Date(0,0,-1) + add excel values = ?!?! zašto ovo radi?!

dileme

OLAP, star shema

? što je to… jel mi treba ROLAP ili nešto drugo
★ Sad mi je jasno da su OLAP cubes za prikaz gustih podataka gdje imamo za svaku vrijednost svih dimenzija neku vrijednost, a star schema je bolje kad su podaci raspršeniji, kao što je ovdje slučaj

Tečajevi - conversion

? što s tečajevima, nije mi bilo jasno iz zadataka, ako ih trebam spremati, kako onda prikazati shemu, jer mi izgleda kao snowflake i dodatno isprepleteno, dakle nešto onda ne razumijem s dimenzijama
https://christianwade.wordpress.com/tag/currency-conversion/
★ uzimat ću ih na dan transakcije, jer to ima smisla, kasnije se može napraviti report i na bilo koji drugi dan

Čuvanje povijesti

? kako
★ Nema smisla, jer em radim initial load, em su transakcije koje se ne mijenjaju jednom kad su obračunate, kao ni tečajnica
Za neke stvari tipa stanje na skladištu bi imalo smisla pamtiti povijesne podatke tako da se npr svaki dan u DWH ubaci novo stanje, dakle ključ bi bili npr id robe i datum ili tako nekako

♬ Slowly changing dimensions SCD type 1, 2, 3
Type 1 - do not store changes at all
http://www.vikramtakkar.com/2013/03/implementing-scd-type-1-slowly-changing.html
Type 2 - two columns - start & end date. Primary key stay, suroggate key changes
http://www.vikramtakkar.com/2013/03/implementing-scd-slowly-changing.html
Type 3 - 1 additional column, for previous values
http://www.vikramtakkar.com/2013/03/implementing-slowly-changing-dimensions.html

talend

Talend ne oprašta probavanje, rename i slično, znamu se razletiti shema i jednostavnije je obrisati i napraviti nanovo nego skužiti gdje se ne poklapa i javlja warning iako sve izgleda isto

Tipovi kod bulkloada

Zezaju ga tipovu za bulkload, no riješi se sa dodavanjem convertType, koji u principu ne radi ništa?
https://www.talendforge.org/forum/viewtopic.php?id=24898

Decimal vs double

Izgleda da treba koristiti decimal za novac kad ide u bazu, no ne znam kakva je praksa za računanje s podacima u talendu, pa sam držala double 10:2
http://stackoverflow.com/questions/6831217/double-vs-decimal-in-mysql
http://code.rohitink.com/2013/06/12/mysql-integer-float-decimal-data-types-differences/

datumi

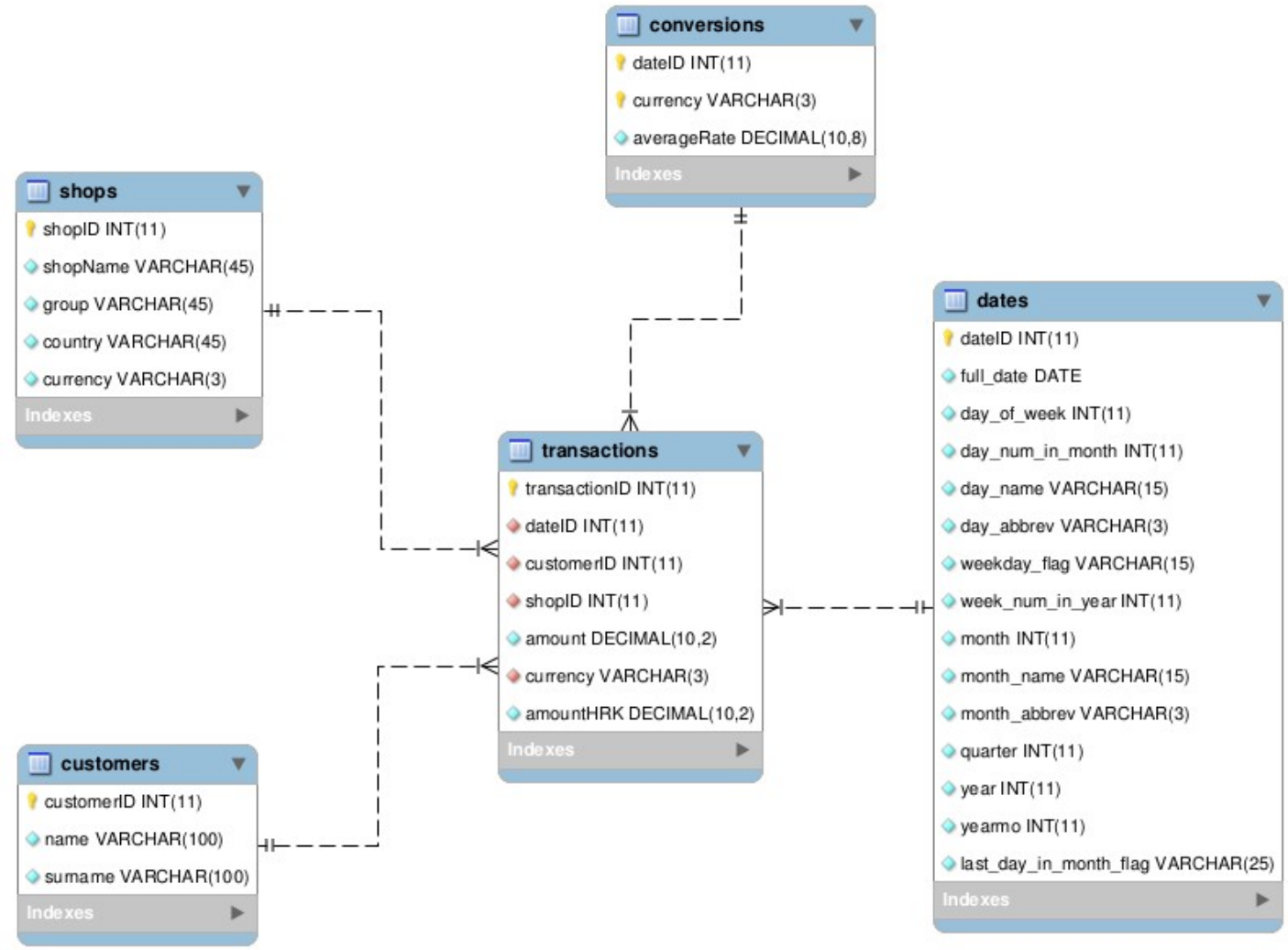
Pri učenju sam se jako puno zezala s datumima u Talendu da ih natjeram da se međusobno konvertiraju, xls, sql i slično, ovako se svi svode na intove :D

Bulk load datuma

Kod bulk loada fajla s datumima treba paziti da su yyyy-MM-dd formatiratni jer SQL jedino tako prepoznaje da je nešto datum!

DWH

ER shema



Reporti

Mjesečni izvadak prometa po računu za klijenta, odabire se mjesec i klijent

Mjesec i godina te klijent direktno kroz inicijalni query

	A	B	C	D	E
1	full_date	shopName	amountHRK	transactionID	
2	01.01.2016	ebay.com	168,02	8299	
3	01.01.2016	ikea/hu	1,99	8317	
4	02.01.2016	ekupi.si	106,82	8341	
5	04.01.2016	amazon.de	328,31	8378	
6	04.01.2016	ikea/se	77,78	8380	
7	04.01.2016	ikea/hu	0,33	8385	
8	06.01.2016	ebay.de	110,69	8425	
9	07.01.2016	ikea/us	425,4	8435	
10	07.01.2016	amazon.co.uk	853,26	8450	
11	08.01.2016	ikea/se	58,13	8475	
12	10.01.2016	amazon.co.uk	962,03	8525	
13	10.01.2016	amazon.co.uk	715,59	8528	
14	13.01.2016	ikea/hu	1,88	8609	
15	13.01.2016	ikea/se	20,88	8610	

Izvještaj na kojem se vide navike trošenja klijenta (na što klijent troši najviše novaca, npr u kojim dućanima, po mjesecima)

Za grupacije, za jednu godinu po mjesecima, sortirano po grupacijama i iznosu unutar svake

ima dosta dućana pa ispada preveliko za ilustraciju, zato grupacije; klijent te dani za godinu direktno kroz inicijalni query

	A	B	C	D
1	group	month_name	amountHRK	
2	amazon	January	5443,2	
3	amazon	October	5142,9	
4	amazon	November	4452,99	
5	amazon	March	4239,39	
6	amazon	February	3999,69	
7	amazon	June	3850,9	
8	amazon	December	3149,42	
9	amazon	September	2834,19	
10	amazon	August	2600,51	
11	amazon	May	2418,11	
12	amazon	July	2187,99	
13	amazon	April	1160,57	
14	ebay	February	5303,09	
15	ebay	January	5096,34	
16	ebay	August	4651,92	
17	ebay	July	4518,81	
18	ebay	September	4026,49	
19	ebay	December	4004,48	
20	ebay	April	3941,21	
21	ebay	June	3926,41	
22	ebay	November	3400,76	
23	ebay	October	2549,68	

	A	B	C	D
1	shopName	year	amountHRK	
2	ebay.co.uk	2015	20497,89	
3	amazon.co.uk	2015	16020,7	
4	ebay.de	2015	15457,78	
5	amazon.de	2015	15398,99	
6	ikea/us	2015	13108	
7	amazon.com	2015	12189,61	
8	ekupi.si	2015	11617,86	
9	ebay.com	2015	9801,79	
10	ebay.com	2016	4135,91	
11	ebay.co.uk	2016	2985,61	
12	amazon.co.uk	2016	2912,13	
13	amazon.com	2016	2787,75	
14	ekupi.si	2016	2492,45	
15	amazon.de	2016	2414,96	
16	ebay.de	2016	1580,81	
17	ekupi.hr	2015	1326,85	
18	ikea/hr	2015	1299,75	
19	ikea/se	2015	1120,14	
20	ikea/hr	2016	649,42	
21	ekupi.hr	2016	271,36	
22	ikea/se	2016	226,41	
23	ikea/us	2016	181,74	
24	ikea/hu	2015	35,52	
25	ikea/hu	2016	5,43	
26				

Za dućane, po godinama, sortirano po iznosu da se vidi gdje je najviše kupovao i koje godine

Klijent direktno kroz inicijalni query

Materijali

ideje

BEST PRACTICES:

http://www.vikramtakkar.com/2014/10/talend-data-integration-development.html

✔ 1. Talend workspace path should not contain any spaces.

? 2. Never forget to perform Null Handling.
✔ 3. Create Repository Metadata for DB connections and retrieve database table schema for DB tables.
✔ 4. Use Repository Schema for Files/DB and DB connections.
? 5. Create Database connection using t<Vendor>Connection component and use this connection in the Job. Do not make new connection with every component.
6. Always close the connection to database using t<Vendor>Close component.
7. Create a Repository Document corresponding to every Talend job including revision history.
✔ 8. Provide Sub Job title for every sub job to describe the sub job purpose/objective.
? 9. Avoid Hard Coding in Talend Job component. Instead use Talend context variables.
10. Create Context Groups in Repository
11. Use Talend.properties file to provide the values to context variables using tContextLoad.
12. Create Variables in tMap and use the variables to assign the values to target fields.
13. Create user routines/functions for common transformation and validation.
✔ 14. Develop Talend job iteratively.
✔ 15. Always Exit Talend open studio before shutting down the PC.
✔ 16. Always rename Main Flows in Talend Job to meaningful names.
17. Always design Talend jobs by keeping performance in mind.

PERFORMANCE optimization tips

http://www.vikramtakkar.com/2014/05/talend-job-design-performance-tuning.html

✔ 1.Remove Unnecessary fields/columns ASAP using tFilterColumns component.
✔ 2. Remove Unnecessary data/records ASAP using tFilterRows component.
✔ 3. Use Select Query to retrieve data from database
✔ 4. Use Database Bulk components
? 5. Store on Disk Option - tSortRow, tFilterRow, tMap, tAggregateRow, tHashOutput - use memory
6. Allocating more memory to the Jobs
7. Parallelism
8. Use Talend ELT Components when required
9. Use SAX parser over Dom4J whenever required
10. Index Database Table columns
✔ 11. Split Talend Job to smaller Subjobs