oracle install

https://oraclecrud.wordpress.com/2015/07/03/howto-install-fedora-22-oracle-xe-11gr2-apex-5-ords-3-0-on-a-virtualbox-vm/
https://fedorahosted.org/spacewalk/wiki/OracleXeSetup

create oracle user, dir to hold oracle sw, put oracle into sudoers
install java sdk (8), use alternatives, export java home
install oracle db, xe

download talend (latest, prethodni treba javu 7), unzip
download XULrunner, unzip, polinkaj sa ini fajlom iz talenda
startaj sa .sh skriptom

demo tutorial

https://www.talendforge.org/tutorials/tutorial.php?language=english&idTuto=14

RC = right click
DC - double click

intro

> It reads a delimited file and displays the data in the console.

prvo import zip jobova, job design > import
onda otvoriti jednog (RC), pa označiti svaku kućicu pa RC i settings i vidjeti podatke ako je baza (uname, pass i to), postaviti tablicu, postaviti schemu, dobro radi kad se označi sync columns
i onda run jobs, i ispisat će status broja redaka koje je obavilo
rezultat je imati u bazi tablice customers i state

onda napraviti novi job, demo, i tu sad ide pravi zadatak

metadata > file delimited > create - povezati sa csv fajlom customers, staviti 'set heading row as column name', u schemi postaviti id da je key i da nije nullable (checkbox), ime customers_file
dodati component > logs > tLogRow
povezujemo ih t.d. customers_file RC pa row > main i onda otići do tLogRow pa se povežu
pod run staviti označeno statistics
kad se job runna, ćemo vidjeti ispis

tMap

> It reads a delimited file, transforms its data, and displays the result in the console.

dodati na ekran components > processing > tMap
postojeću konekciju (row1) pomaknuti tako da početak ide od tMap (prema tLogRow), a dodati novu iz customers_file prema tMap (row2)
DC tMap, otvara se editor, piknuti AutoMap, pa će se sami povezati
modificiramo konekcije

prvo, row2.lastname u row1 table -> klik na njega, pa na točkice i otvara se expression builder -

mijenjamo name u upercase (ctrlX postojećeg inputa, pa category>stringhandling pa functions>upcase pa DC pa ctrlV naziva umjesto hello pa ok
drugo, row1table pa row2.id edit, pa dolje u expression editor samo upisati *5 na kraj

pa ok u tMap prozoru pa save i run job

join

> It creates a join between two inputs and populates a database with the aggregated data.

metadata > dbconnection, create new, poveži s bazom, daj podatke, mydb je ime
mydb RC pa retrieve schema, use defaults (name filter)
select customers and states tables
check schema, tu se može modificirati no sad nećemo, finish

metadata > db con > mydb > table schemas > drag & drop states, i biraj input component (tMysqlInput)

RC states pa postavke, component view, kliknuti … next to query i otvara se sql builder wizard označi modify query using graphical editor (kao kalendarić sličica), ostaviti označeno samo postal i state

klik na running man ikonu, dolje lijevo bude rezultat, klik ok

u job designeru povezati states sa tMap (pisat će row3 lookup)
obrisati tLogRow
i ubaciti customers tablicu iz db dijela, ovaj put s MySqlOutput
podesi postavke bazi, postavi drop table if exists and create
spoj tMap sa customers, nazovi to out1

DC tMap za editor
Click the **row1** table and click **[x]** to delete it.
In the **row2** table, select all the columns except *states* and drag them to the **out1** table.
Now select the *states* column and drag it to the *Postal* column of the **row3** table to create the join.
In the **row3** table, select the *State* column and drag it at the end of the **out1** table.
In the **out1** table, select the *row3.State* column. Go to the **out1** table in the **Schema editor** area, and set it length to *14* characters.
Click **OK**

prije runa kaže traces box, no ne vidim to…

export job script

job > RC > build job

banka zadatak

Talend projekt - ines java

Neki podaci su generirani s Talend alatom - izvor kupnji i tablica kupaca
Customera 100, shopova 12, 1.1.2015 - 10.3.2016, transakcija 10000, za više je jako puno kupnji u istom danu za istog customera, ovako je raštrkanije

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

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

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

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/

Kalendar - date

> Izvor > [fajl] xls
http://www.kimballgroup.com/wp-content/uploads/2014/03/Ch3-SampleDateDim.xls
> ETL uzima id, datum, dan u tjednu (broj), mjesec (broj), godinu, kvartal (broj)

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

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

Convert excel dates into sql dates

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

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/

> izvor > [fajl] Tečajna lista, xls
http://www.hnb.hr/temeljne-funkcije/monetarna-politika/tecajna-lista/tecajna-lista
> ETL uzima datum, valutu, srednji tečaj (kao float), dodijeli ID svakom zapisu
ID zato da se može jasnije povezati sa transakcijama, ključ je u stvari datum i valuta, no ne znam to

izvesti da ne zakompliciram shemu - datum kao shareani ključ mi je weird

Dućan - shop

Izvor > [fajl] csv
ETL > dodaje id

Kupci - customer

izvor> [generiran] generirano kroz Talend
ETL > samo puni

Kupnje - transactions

> Izvor > [generiran] u originalnoj valuti
> ETL radi konverziju u HRK
> report može raditi dodatne konverzije za određeni dan i određenu valutu

Bilo bi super da se može generirati potpuno dinamički - prema broju shopova i prema broju generiranih korisnika, da se u generatoru transakcija te vrijednosti povuku, a ne da su hardkodirane

Report mjesečni izvadak

Po customer idu i po yearmonth vrijednosti

Report trošenje

Po customer idu, aggregiranje po shopu i godini (jer ima dosta shopova), sortirano po godinama pa po vrijednostima od najveće k najmanjoj

Čuvanje povijesti

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

Notes

MariaDB [bank]> select month(datum), sum(iznos), ducan from transakcije where klijentId=1 group by ducan, month(datum);

```
/* DDL for the date dimension */
create table Date_Dimension (
date_key smallint not null,
full_date smalldatetime,
day_of_week tinyint,
day_num_in_month tinyint,
```

day_num_overall smallint,
day_name varchar(9),
day_abbrev char(3),
weekday_flag char(1),
week_num_in_year tinyint,
week_num_overall smallint,
week_begin_date smalldatetime,
week_begin_date_key smallint,
month tinyint,
month_num_overall smallint,
month_name varchar(9),
month_abbrev char(3),
quarter tinyint,
year smallint,
yearmo int,
fiscal_month tinyint,
fiscal_quarter tinyint,
fiscal_year smallint,
last_day_in_month_flag char(1),
same_day_year_ago_date smalldatetime,
primary key (date_key))

https://www.talend.com/blog/2015/12/07/talend-%E2%80%9Cjob-design-patterns%E2%80%9D-and-best-practices

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

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

Tmap vs…
http://www.vikramtakkar.com/2013/04/tmap-vs-tjoin-talend-open-studio.html
http://www.vikramtakkar.com/2013/09/difference-between-tmap-and-tfilterrow.html
tJoin only unique join, one lookup flow, only exact match on keys

http://www.etladvisors.com/2012/11/26/using-variables-in-the-tmap-component/

http://www.vikramtakkar.com/2013/07/fetch-last-record-from-fileflow-in.html
Using **tFileRowCount** component we can find the row count of the file and then set the following value of COUNT global variable of component **tFileRowCount** to the header part of the **tFileInputDelimited** component.
((Integer)globalMap.get("tFileRowCount_1_COUNT"))-1

http://www.vikramtakkar.com/2013/03/sharing-database-connection-with-child.html
1. To share the DB connection, provide the database credential and click on "**Use or register a shared DB connection**" in **tMYSQLConnection** component and provide the name to this shared connection in the "**Shared DB Connection Name**" text box.
2. Now, if you want to use the same connection in the child box, click on "**Use or register a shared DB connection**" in **tMYSQLConnection** component and provide the same name of the shared connection in the "**Shared DB Connection Name**" text box in the Child box.

http://www.vikramtakkar.com/2013/05/how-to-pass-data-from-child-to-parent.html
1. In the child Job, connect the output to **tBufferOutput** component.
2. Now in the parent job (main Job). Right click tRunJob and select **Copy child Job schema**. This will copy the child job schema to main job.
3. Now right click tRunJob, Select Row and connect it to tLogRow to see if the data from child job is retrieved. Now you can use this data in your main job as per your logic.

http://www.vikramtakkar.com/2013/05/example-to-execute-multiple-sql-queries.html

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

http://www.vikramtakkar.com/2013/01/understand-context-variables-with.html
http://www.vikramtakkar.com/2013/01/understand-context-variables-with_20.html

http://www.vikramtakkar.com/2013/01/understand-context-variables-with_26.html

http://www.vikramtakkar.com/2013/03/how-to-convert-or-cast-string-to-date.html

Do kud sam stigla kopajući
http://www.vikramtakkar.com/search/label/Talend?updated-max=2013-03-09T08:40:00-08:00&max-results=20&start=39&by-date=false

Source.transactions

Column	Db Column	Key	DB Typ	Type	Nullab	Date Pa	Leng	Prec	De	Cor
transactionID	transactionID	✓	INT	int						
dateID	dateID		INT	int						
customerID	customerID		INT	int						
shopID	shopID		INT	int						
amount	amount		DECIMA	doub			10	4		

Source.customers

Column	Db Column	Key	DB Typ	Type	Nullab	Date Pa	Leng	Prec	De	Cor
customerID	customerID	✓	INT	int						
name	name		VARCHA	String			100			
surname	surname		VARCHA	String			100			

Etl.customers

Column	Db Column	Key	DB Typ	Type	Nullab	Date Pa	Leng	Prec	De	Cor
customerID	customerID	✓	INT	int						
name	name		VARCHA	String			100			
surname	surname		VARCHA	String			100			

Etl.transactions

Column	Db Column	Key	DB Typ	Type	Nullab	Date Pa	Leng	Prec	De	Cor
transactionID	transactionID	✓	INT	int						
dateID	dateID		INT	int						
customerID	customerID		INT	int						
shopID	shopID		INT	int						
amount	amount		DECIMA	doub						
currency	currency		VARCHA	Strin			3			
amountHRK	amountHRK		DECIMA	doub			10	4		

Etl.conversions

Column	Db Column	Key	DB Typ	Type	Nullab	Date Pa	Leng	Prec	De	Cor
conversionID	conversionID	✓	INT	int						
dateID	dateID		INT	int						
currency	currency		VARCHA	Strin			3			
averageRate	averageRate		DECIMA	doub			10	8		

Etl.shops

Column	Db Column	Key	DB Typ	Type	Nullab	Date Pa	Leng	Prec	De	Cor
shopID	shopID	✓	INT	int						
shopName	shopName		VARCHA	Strin			45			
group	group		VARCHA	Strin			45			
country	country		VARCHA	Strin			45			
currency	currency		VARCHA	Strin			3			

Etl.date

Column	Db Column	Key	DB Typ	Type	Nullab	Date Pa	Leng	Prec	De	Cor
date_key	date_key	✓	INT	int	☐					
full_date	full_date	☐	DATE	Date	☐	"dd.MM.y				
day_of_week	day_of_week	☐	INT	int	☐					
day_num_in_	day_num_in_m	☐	INT	int	☐					
day_name	day_name	☐	VARCHA	String	☐		15			
day_abbrev	day_abbrev	☐	VARCHA	String	☐		3			
weekday_flag	weekday_flag	☐	VARCHA	String	☐		15			
week_num_in	week_num_in_y	☐	INT	int	☐					
month	month	☐	INT	int	☐					
month_name	month_name	☐	VARCHA	String	☐		15			
month_abbre	month_abbrev	☐	VARCHA	String	☐		3			
quarter	quarter	☐	INT	int	☐					
year	year	☐	INT	int	☐					
yearmo	yearmo	☐	INT	int	☐					
last_day_in_r	last_day_in_mo	☐	VARCHA	String	☐		25			