Data Warehousing Project (15 points)

INSTRUCTIONS

1. Download the project pack **retail_sales_project.zip** from IVLE Workbin (Files) under “Projects/Project 4/”.
2. **Deliverables:** A zipped file containing
   - i the finished retail_sales_project folder
   - ii tableau packaged workbook (.twbx format)
   - iii screenshots of the figures of Task 9 and 10
3. **IMPORTANT: The zipped file must be named as “Group(your group number).zip”. Naming violation will result in penalties.**
4. Submit your worksheet to the folder “Project 4 Submissions” in IVLE Workbin by **Sunday 18 November 18:00**.
5. After the deadline and until **Tuesday 20 November 18:00**, you can submit the deliverables to the folder “Pentaho Project Late Submissions” in IVLE Workbin (Files) (penalties apply).
6. Keep your source i.e the .ktr, .kjb files, the databases, tableau workspace, as you may be asked to demonstrate/reproduce the results you are reporting.

Datasets

Customer Information:- extract_customers.txt

Employee Information:- extract_employees.txt

Stores information:- extract_store.csv

Products information:- extract_product.txt

Point of sale data:- several extract_pos_details csv files and their corresponding extract_pos_headers text files.

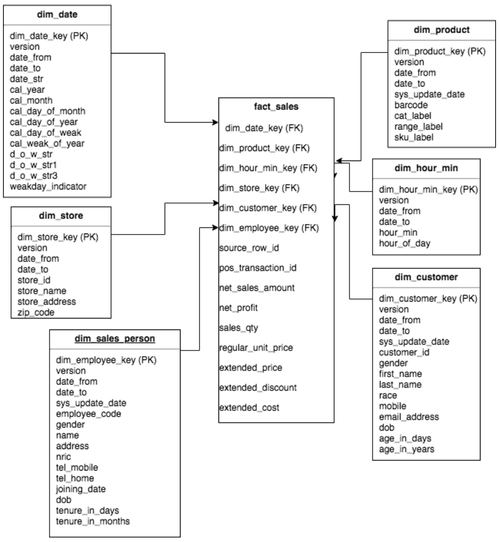
For detailed descriptions of the fields in the datasets consult class lecture slides.

Illustrated in the class

1. Creating product *dimension* from products raw data.
2. Creating partial sales *fact table* from point of sales details raw data.

Project Overview

Implement the following star schema in pentaho. Perform visual Analytics on the schema using Tableau.



1. There is a job file run_first_time.kjb in the project pack which generates the Date and Hour dimensions for you. You only run it once.
2. **You may edit/reuse the transformations for product dimension and sales fact tables for sales done in the class and uploaded in IVLE inside "retail_sales after class.zip" file.**

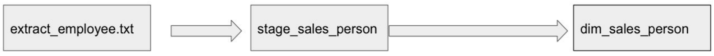
Task Description

Complete the following tasks in Spoon-

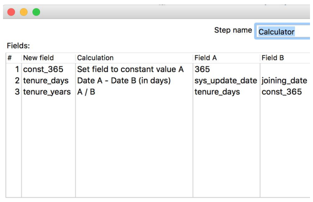
Task 1. Create transformations for generating sales_person dimension table (1.5 point)

i. Create **stage_sales_person.ktr**

ii. Create **dim_sales_person.ktr**



Tips: You may need to use calculator for calculating tenure of a sales person.

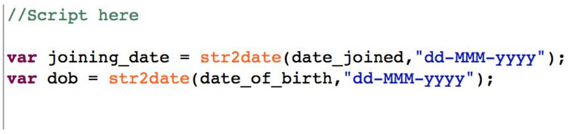


Step name Calculator

Fields:

#	New field	Calculation	Field A	Field B
1	const_365	Set field to constant value A	365	
2	tenure_days	Date A - Date B (in days)	sys_update_date	joining_date
3	tenure_years	A / B	tenure_days	const_365

You may need to use javascript step for transforming string date to Date type.



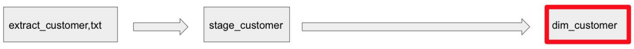
```
//Script here

var joining_date = str2date(date_joined,"dd-MMM-yyyy");
var dob = str2date(date_of_birth,"dd-MMM-yyyy");
```

Task 2. Create transformations for generating customer dimension table (1.5 point)

i. Create **stage_customer.ktr**

ii. Create **dim_customer.ktr**

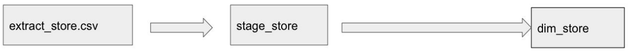


Tips/Hint: The same hint as sales_person dimension.

Task 3. Create transformations for generating store dimension table. (1 point)

i. Create **stage_store.ktr**

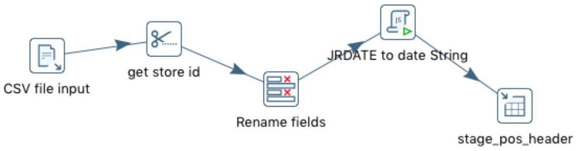
ii. Create **dim_store.ktr**



Task 4. Join pos_header table with pos_details table to store salesperson, store and customer information in staging table ‘stage_sales’. (3 points)

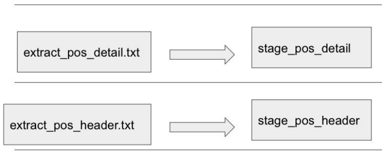
i. Create **stage_pos_header.ktr** to create stage_pos_header table from header data.

Tips/Hint: You will need to rename a few fields (e.g. JRCUST, JRCASHIER, TRX_ID) , get store_id by cutting first 5 characters from string JRSTORE.

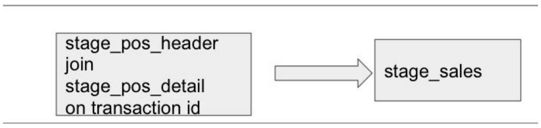


ii. Create **stage_pos_detail.ktr** to create stage_pos_details table from details data.

Tips/Hint: We created stage_sales.ktr in the class to process extract_pos_details.txt. We do to same here, with few exceptions - the calculator now will need to calculate unit profit, net_profit etc; the output table is not stage_sales but stage_pos_details.



iii. Create **stage_sales.ktr** which will join pos_details and pos_headers on transaction id and write it to stage_sales table.

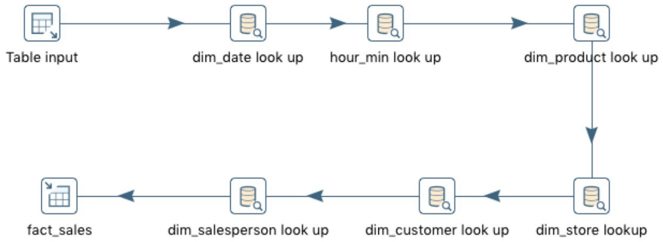


Tips: Make sure you have the natural keys of all the dimensions (e.g. store_id for store dimension, customer_id for customer dimension etc.) in the stage_sales table. You will need them to look up the foreign keys in fact table.

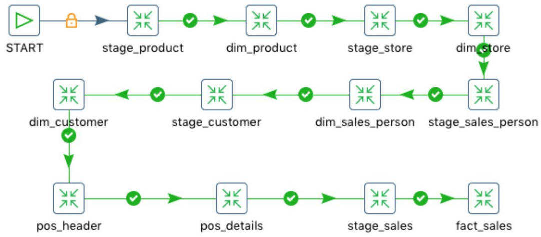
You may write sql join queries manually in a 'table input' Step.

Task 5. Create **fact_sales.ktr** for generating fact table (1.5 point)

Tips/Hint: You will need to add look up for all the dimension for foreign keys.



Task 6. Create an ETL job **"main job.kjb"** to add all the transformations required to populate the data mart with data. (1.5 point)



Tips/Hint: You may use (if needed) truncate_all_tables.ktr inside one_time folder to truncate all the tables from your data mart. You may want to truncate before you run this job since fact_sales table do not have a primary key)

Task 7. Develop a Job to be run daily as new files are dropped into the data folder. (2 points)

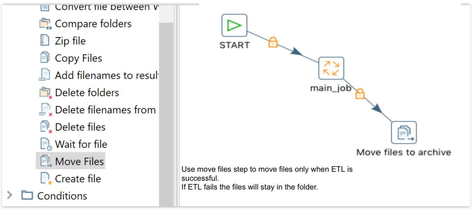
i. There are daily data files (pos_header and pos_detail) given to you inside data folder (**Important: extract_pos_details.txt and extract_pos_header.txt is not part of the daily data files) . Assuming that each file will be deposited in the data folder everyday at 6am, you are to **modify stage_pos_header.ktr and stage_pos_details.ktr** so that they process files incrementally to update the data warehouse for the users to query at 8am.

Tips/hint: Read about "Get File Names" Step from Spoon documentation. Learn how to use wildcard/ regEx (regular expression) for including certain csv/text files having similar naming pattern).

ii. Data files after being processed are to be moved to archive folder (using PDI tool)
Create a job named **'rundaily.kjb'** which will invoke "main job.kjb" and move headers and details file to the folder named "archived".

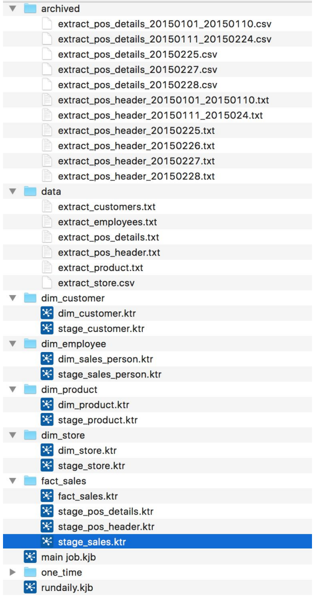
Tips/Hint:

Create a 'super' job that calls main_job



Use move files step to move files only when ETL is successful.
If ETL fails the files will stay in the folder.

Upon successful completion of all the tasks until this point, your retail_sales_project folder must look like the following. This folder is one of the deliverables for the project.



```
▼ archived
    extract_pos_details_20150101_20150110.csv
    extract_pos_details_20150111_20150224.csv
    extract_pos_details_20150225.csv
    extract_pos_details_20150227.csv
    extract_pos_details_20150228.csv
    extract_pos_header_20150101_20150110.txt
    extract_pos_header_20150111_2015024.txt
    extract_pos_header_20150225.txt
    extract_pos_header_20150226.txt
    extract_pos_header_20150227.txt
    extract_pos_header_20150228.txt
▼ data
    extract_customers.txt
    extract_employees.txt
    extract_pos_details.txt
    extract_pos_header.txt
    extract_product.txt
    extract_store.csv
▼ dim_customer
    dim_customer.ktr
    stage_customer.ktr
▼ dim_employee
    dim_sales_person.ktr
    stage_sales_person.ktr
▼ dim_product
    dim_product.ktr
    stage_product.ktr
▼ dim_store
    dim_store.ktr
    stage_store.ktr
▼ fact_sales
    fact_sales.ktr
    stage_pos_details.ktr
    stage_pos_header.ktr
    stage_sales.ktr
  main job.kjb
▶ one_time
  rundaily.kjb
```

Evaluation criteria: Completeness of the star schema.

Complete the following tasks in Tableau-

Task 8. Load the dimension tables and the fact table from postgres by successfully connecting to the database from tableau. (Use the extract mode for connection).

(1 point)

Task 9. The retail owner would like to know about net sales amount of different product categories. Draw a figure illustrating the net sales amount by product categories. Add your findings as a caption.

(1 point)

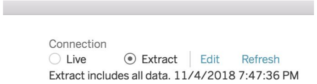
Bonus +1 point: Implement hierarchy of category>range> sku labels so that the owner can drill down by category, range, and further by sku labels.

Task 10. The retail owner would like to know about overall sales in his shops by location. Draw a map illustrating net sales amount by the shop locations, so that he can learn about the underselling shops. Add your findings as a caption.

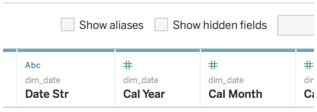
(1 point)

Note:

1. Make sure your tableau connection is in extract mode all the time.



Connection
○ Live ● Extract | Edit Refresh
Extract includes all data. 11/4/2018 7:47:36 PM



□ Show aliases □ Show hidden fields

2. We are not going to evaluate on the aesthetics of the visualization. However the viz should include proper title, axis labels, captions etc. We are going to penalize if you miss any of those basics (in general tableau takes care of those things, hence the stringency). The goal is to have a viz that is clear, readable and meet those basic requirements.
3. For each viz, you must add a *Caption* mentioning your findings e.g. for Task 9, one may write "Graphics cards products are the least selling among all categories whereas facial skin care products are the best selling." as caption.

*****Important:** For task 9 and 10, you must capture a screenshot/screenshots of your figure, rename them as task 9 and task 10. Moreover, you must save your work as a tableau packaged workbook, which is a **.twbx** file. Provide the figures and the packaged Work Book inside the deliverable zip file.