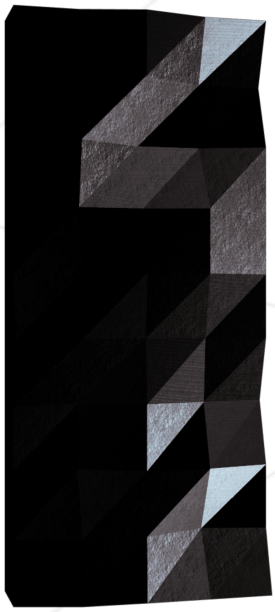


Training Piscine datascience - 1

Data Warehouse

Summary: Today, you will discover the creation of a Data Warehouse

Version: 1.1

Contents

Chapter I

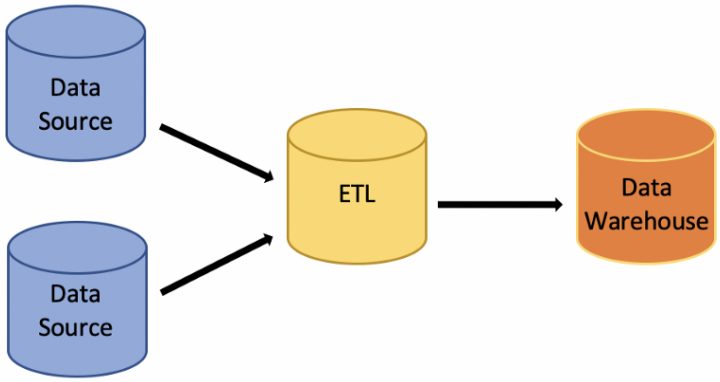
General rules

- You have to render your modules from a computer in the cluster either using a virtual machine:
    - You can choose the operating system to use for your virtual machine
    - Your virtual machine must have all the necessary software to realize your project. This software must be configured and installed.
- Or you can use the computer directly in case the tools are available.
    - Make sure you have the space on your session to install what you need for all the modules (use the goinfre if your campus has it)
    - You must have everything installed before the evaluations
- Your functions should not quit unexpectedly (segmentation fault, bus error, double free, etc) apart from undefined behaviors. If this happens, your project will be considered non functional and will receive a `0` during the evaluation.
- We encourage you to create test programs for your project even though this work **won't have to be submitted and won't be graded**. It will give you a chance to easily test your work and your peers' work. You will find those tests especially useful during your defence. Indeed, during defence, you are free to use your tests and/or the tests of the peer you are evaluating.
- Submit your work to your assigned git repository. Only the work in the git repository will be graded. If Deepthought is assigned to grade your work, it will be done after your peer-evaluations. If an error happens in any section of your work during Deepthought's grading, the evaluation will stop.
- By Odin, by Thor ! Use your brain !!!

Chapter II

Introduction

ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse.



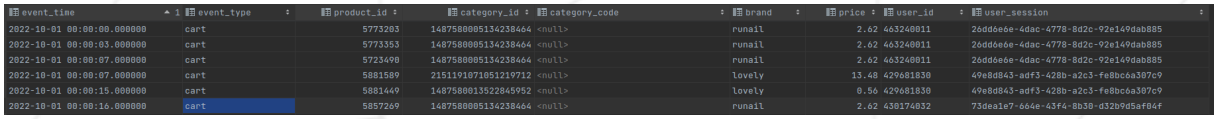
Be careful with this "piscine". Even if you manage to validate a module, you may be stuck later if you haven't cleaned up or stored your data properly.

Chapter III

Exercise 00

	Exercise 00
Exercice 00 : Show me your DB	
Turn-in directory : $ex00/$	
Files to turn in :	
Allowed functions : `pgadmin, Postico, dbeaver or what you want to see the db easily`	

- Find a way to see the db easily with a software
- The software chosen must be easy to file and to use for the search of an ID



event_time	event_type	product_id	category_id	category_code	brand	price	user_id	user_session
2022-10-01 00:00:00.000000	cart	5773203	1487580005134238464	<null>	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
2022-10-01 00:00:03.000000	cart	5773353	1487580005134238464	<null>	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
2022-10-01 00:00:07.000000	cart	5723490	1487580005134238464	<null>	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
2022-10-01 00:00:07.000000	cart	5881589	2151191071051219712	<null>	lovely	13.48	429681830	49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2022-10-01 00:00:15.000000	cart	5881449	1487580013522845952	<null>	lovely	0.56	429681830	49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2022-10-01 00:00:16.000000	cart	5857269	1487580005134238464	<null>	runail	2.62	430174032	73dea1e7-664e-43f4-8b30-d32b9d5af04f

Chapter IV

Exercise 01

	Exercise 01
Exercice 01 : customers table	
Turn-in directory : $ex01/$	
Files to turn in : `customers_table.*`	
Allowed functions : `All`	

- You have to join all the data_202*_*** tables together in a table called "customers"

Chapter V

Exercise 02

	Exercise 02
Exercice 02 : remove duplicates	
Turn-in directory : $ex02/$	
Files to turn in : `remove_duplicates.*`	
Allowed functions : `All`	

- You must delete the duplicate rows in the "customers" table.

`Sometimes the server sends the same instruction with 1 second interval`

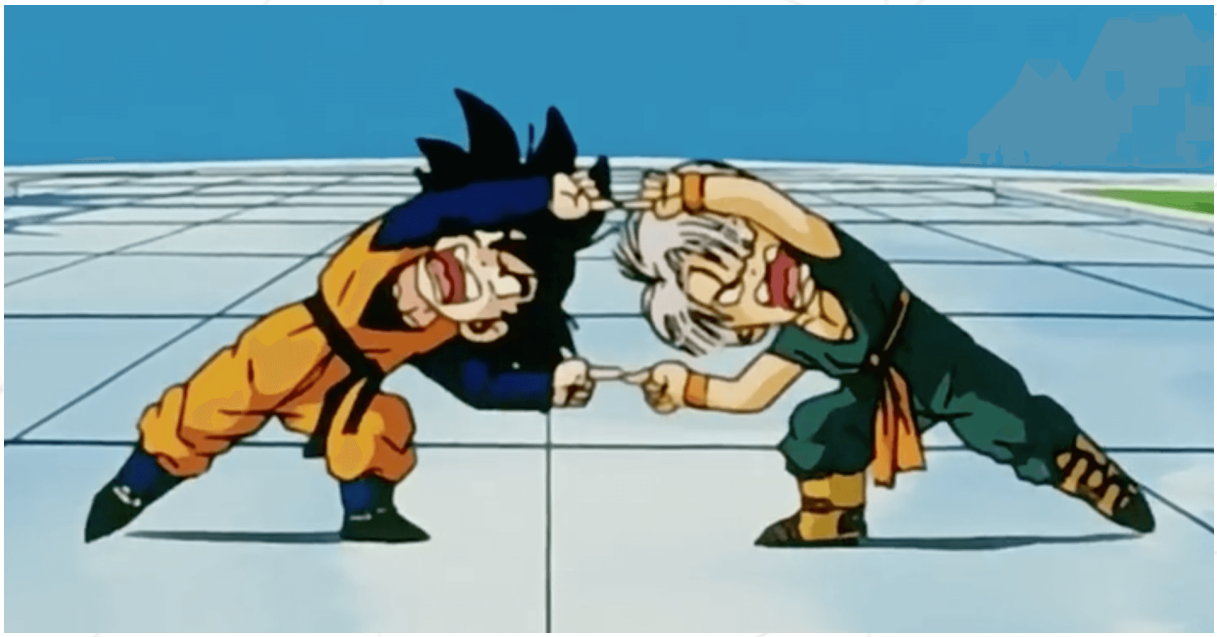
For exemple:

```
event_time          event_type      product_id
2022-10-01 00:00:32,remove_from_cart,5779403
2022-10-01 00:00:33,remove_from_cart,5779403
```

Chapter VI

Exercise 03

	Exercise 03
Exercice 03 : fusion	
Turn-in directory : $ex03/$	
Files to turn in : `fusion.*`	
Allowed functions : `All`	



- You must combine the "customers" tables with "items" in the "customers" table

Be careful not to lose any information

Chapter VII

Submission and peer-evaluation

Turn in your assignment in your `Git` repository as usual. Only the work inside your repository will be evaluated during the defense. Don't hesitate to double check the names of your folders and files to ensure they are correct.

`The evaluation process will happen on the computer of the evaluated group.`