Project Assignment - Part 1

Roberto Pellungrini, Anna Monreale

October 8, 2020

Introduction

In **Part 1** of the project you are required to create and populate a database starting from .csv files and perform different operations on it. In the following you can find a set of incremental assignments, each one with a brief description of what you are required to produce and what tools you can use for the task.

Build the datawarehouse

fact.csv is in the form *(Id,gpu_code,cpu_code,ram_code,time_code,geo_code,vendor_code ,sales_uds,sales_currency)* and it contains the main body of data: a fact table with sales data of three different product lines: cpu, gpu and ram. Note that fields *gpu_code, cpu_code, ram_code* indicate the type of product: only one can have a non null value in a given row, the other two are necessarily missing.

gpu.csv, **cpu.csv** and **ram.csv** are product dimensions. They can be linked to the main fact table with the respective primary keys. The primary keys in these three dimension tables are linked, respectively, to *gpu_code, cpu_code, ram_code* in the fact table.

geo.csv time.csv vendor.csv are dimensions. They can be linked to the main fact table with their primary keys.

The goal of the following assignments is to build in server apa.di.unipi.it the datawarehouse represented in Figure 1.

Assignment 0

Create the database schema in Figure 1 using SQL Server Management Studio in server apa.di.unipi.it. The name of the database must be *GroupIDHWMart* (example: Group01HWMart).

Assignment 1

Write a python program that splits **fact.csv** into three separate fact tables, one for each line of products: cpu, gpu and ram. Write your program using only basic python functionalities (the use of the pandas library is forbidden).

Assignment 2

Write a Python program that populates the database *GroupIDHWMart* with the three fact tables you obtained from the previous assignments and with the necessary linked dimensions described in the remaining .csv files. Note that in Time table the values of the attributes *day_of_week* and *quarter* should be derived from the date. Indicate each quarter with the values: Q1, Q2, Q3, Q4. While for each *day_of_week* indicate each day with its proper name ("Monday", "Tuesday", etc.).

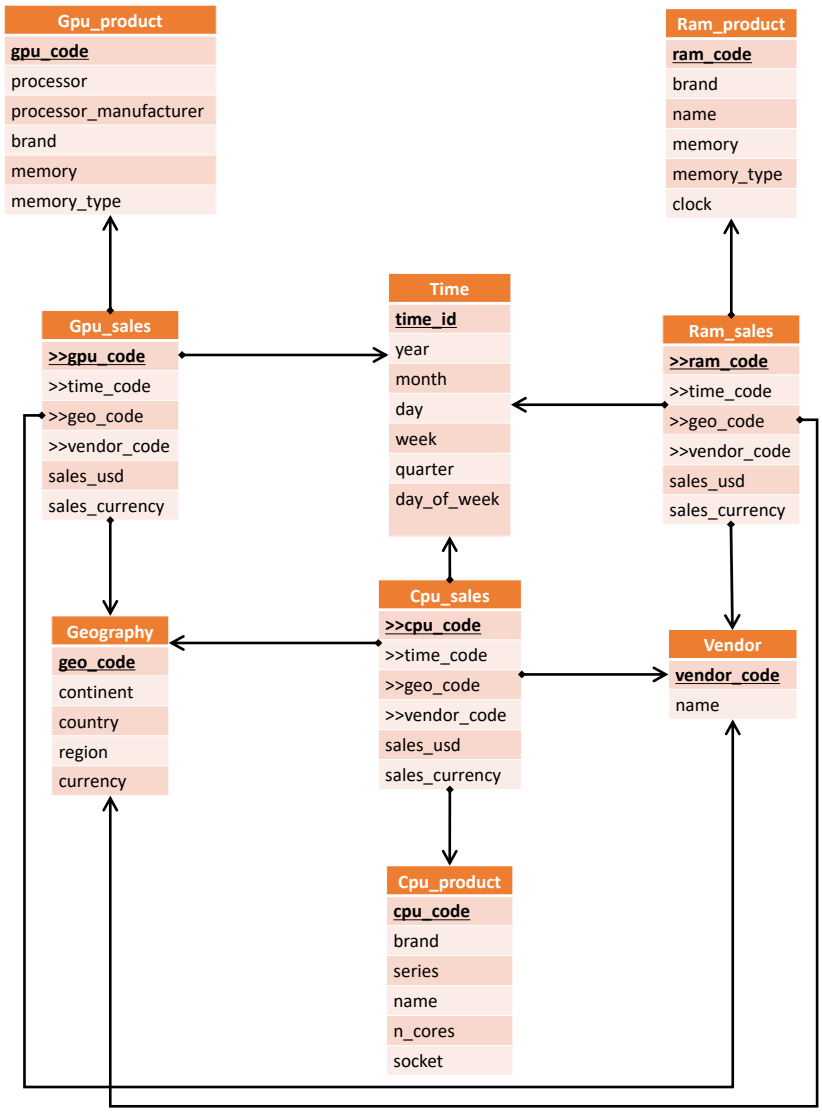

Figure 1: Datawarehouse schema of reference.