Project Assignment - Part 1

Anna Monreale, Cristiano Landi

October 30, 2024

Introduction

In **Part 1** of the project, you are required to create and populate a database starting from different files and perform the required assignments. Additionally, you are required to solve some problems on the database you created using SQL Server Integration Services (SSIS) with computation on the client side (i.e., minimize the use of any SQL command in the nodes, try using native SSIS nodes only[1]). In the following, you can find a set of incremental assignments, each with a brief description of what you are required to produce and which tools you can use for the task. Look at the section with your group id to find the assignments you need to do. Deliver your project with all the required packages in a single .zip folder named LDS_GroupID_part1.zip.

Common to all groups

Build the datawarehouse

The project is about traffic incidents in Chicago; the data are a simplified version of this dataset available on Kaggle. It aims to simulate a decision support system for an assurance company.

Attached to this document, you can find 3 distinct files: **Crashes.csv**, **People.csv**, and **Vehicles.csv**.

- **Crashes.csv** contains the main body of data: a table with data about road incidents between January 2014 and January 2019 in Chicago, USA. The same table also includes information about the causes of the incident, some road properties, and some information about the reported injuries.
- **People.csv** contains information about the people involved in the incidents, including their sex, age, city of residency, etc.
- **Vehicles.csv** contains information about the vehicle(s) involved in the incidents. This file includes information about the vehicle and the information collected by the police after the incident.

[1] SQL command are executed only on the server side! SSIS nodes instead automatically distribute the operation between the server and the client

***Assignment 1**: data understanding*

Understand the data you are working with. How do the 3 files relate to each other? Are there missing values? Can the missing values be recovered/filled easily? Can you integrate additional data (hierarchical GeoHash/Uber H3/Google S2 encoding for spatial data, properties of the road, additional weather conditions, etc.) from external sources with a reasonable effort?

ONLY for this assignment, you can use any software/package you want!

FROM NOW ON, YOU CAN NOT USE ANY PANDAS (AND PANDAS-like) PACKAGES

***Assignment 2**: data cleaning*

Given the information collected in the previous assignment, address the problem related to the missing data (if any) and integrate the additional data (if any).

***Assignment 3**: DW Schema*

It's time to switch from operational databases to analysis-oriented data warehouses. To design the DW schema, since we're simulating an insurance company, the fact table should capture details about the damage costs reimbursement each client incurs for each crash event. You can use the data warehouse schema in Figure 1 as a reference. Next, create the data warehouse tables on the server lds.di.unipi.it. You must use the database named *Group_ID_DB* (example: Group_01_DB) as specified in the credentials' email.
Please note that the DW schema in Figure 1 is just a suggestion; you can modify it as you prefer. You can use both Python and SQL Server Management Studio to create the DW.

***Assignment 4**: Data preparation*

Write a Python program that splits the data into different files, one for each table in the data warehouse proposed in the previous step.

***Assignment 5**: Data uploading with python*

Write a Python program that populates the database *Group_ID_DB* according to the schema relations with all the data you prepared in Assignment 4.
Please note that this operation could take a while, so design the code accordingly!

***Assignment 6**: Data uploading*

Duplicate each table without the records, renaming them as TABLENAME_SSIS. Then, create a SSIS project that populates the new set of tables TABLENAME_SSIS with 10% of the data you prepared in Assignment 4.

At this point, you should have two fact tables along with their corresponding (duplicated) dimension tables. From now on, perform all assignments using the fact table that contains all the data[2].

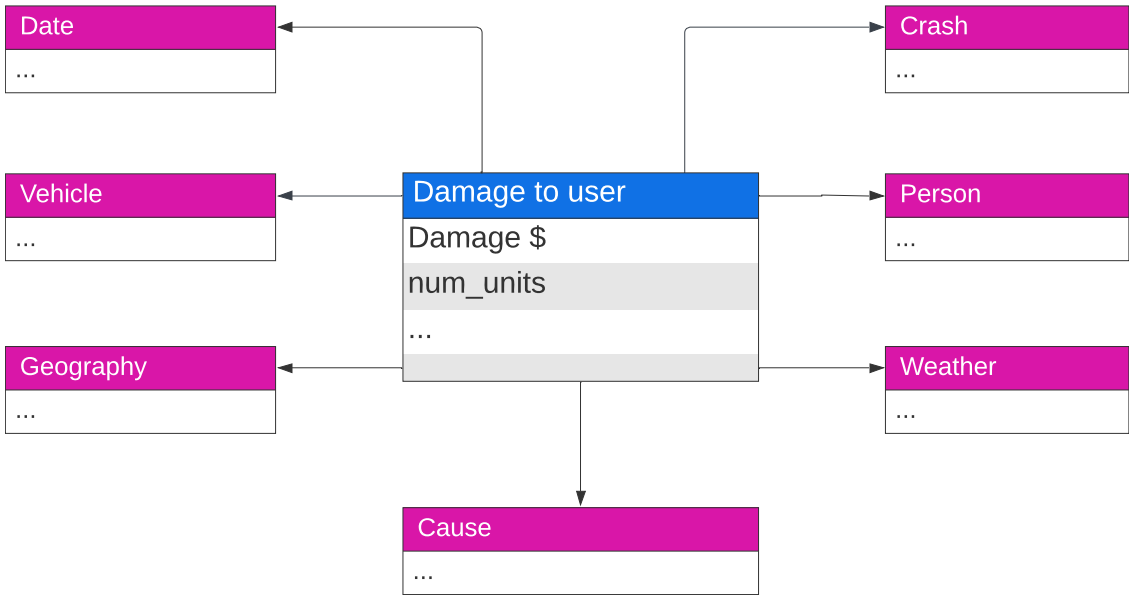

Figure 1: Datawarehouse schema of reference. Fact table in blue.

[2]We suggest using the smaller fact table to manually verify the correctness of your results

Groups from 0 to 13

Answer the following question using only Microsoft SQL Server Integration Services (SSIS).

Assignment 6a

For every year, show all participants ordered by the total number of crashes

Assignment 7a

For every police beat, compute the *day-night crash index*, defined as the ratio between the number of vehicles (NUM_UNITS) involved in an incident between 9 pm and 8 am, and the number of vehicles involved in an incident between 8 am and 9 pm

Assignment 8a

For each quarter, weather condition, and beat, show the average ratio of people under 21 years old to people over 21 years old involved in crashes

Assignment 9a

Based on the analytical results of the previous queries and the insights gathered during the data understanding phase, define an **interesting** query and answer it using SSIS.

Groups from 14 to 24

Answer the following question using only Microsoft SQL Server Integration Services (SSIS).

Assignment 6b

Show all participants ordered by the total damage costs for every vehicle type.

Assignment 7b

For each month, calculate the percentage of the total damage costs caused by incidents occurring between 9 pm and 8 am and incidents occurring between 8 am and 9 pm, with respect to the average total damage costs for all months within the same year

Assignment 8b

Show the total crash damage costs for each vehicle type and weather condition.

Assignment 9b

Based on the analytical results of the previous queries and the insights gathered during the data understanding phase, define an **interesting** query and answer it using SSIS.

Groups from 25 to 36

Answer the following question using only Microsoft SQL Server Integration Services (SSIS).

Assignment 6c

For every year, show the total crash damage costs for each incident

Assignment 7c

A beat is classified as *dangerous* if the number of crashes within that beat exceeds the average number of crashes across all other beats by more than 10%. List all the beats that meet this criterion.

Assignment 8c

For each beat, show the primary contributory cause to the crash ordered by the total crash damage costs in percentage.

Assignment 9c

Based on the analytical results of the previous queries and the insights gathered during the data understanding phase, define an **interesting** query and answer it using SSIS.

Delivery Instruction

When you want to deliver your first project, compress all the files, including the PDF report, and create a .zip file named LDS_GroupID.zip. Then, upload the file using the following Google form using your *studenti.unipi.it* email: https://forms.gle/y8tLrkjjebMuCCcW9. If you have problems using the form, email ALL teachers with the subject: **[LDS] PART1 Group_Id**. The zip file should contain all the files you created/used except the original dataset.

The project must be submitted by **December 5th** (inclusive). Projects uploaded after the deadline will not be accepted. Only one member of the group is required to upload the project.

The submission of the project is not final: if, during the second (final) part of the project, you discover any bugs or errors in the code from the first part, you will have the opportunity to resubmit the first part. The resubmission must include a detailed list of all the changes made compared to the version of the code previously submitted.