DATA MANAGEMENT – ASSIGNMENT 1

MSC IN DATA SCIENCE – NATIONAL CENTRE FOR SCIENTIFIC RESEARCH "DEMOKRITOS"

1. DATA CREATION THROUGH PYTHON PARSER

The first step of this procedure is the import of our initial data to a python script and the appropriate processing in order to generate the data that will be loaded to our database in PgAdmin application.
Further down I will describe the reasoning behind the creation of every database table which came out from a respective pandas DataFrame:

a. Table 1 :"*dit2122_countries*"

I collected the total amount of values regarding *country_codes* and *countries* from the initial data* and I created a dataset consisting of the unique instances of these values assigning a unique *country_id* column that returns as the unique combination of the fields *country_codes* and *countries.* This table will keep the information about countries individually and the other tables containing information about countries will keep only the *country_id column* in order to access this info*.* This designing choice was made to avoid keeping the same information repeatedly in the most of our tables and to save resources.

b. Table 2 : "*dit2122_jurisdiction*"

I followed the same strategy regarding the information about *jurisdiction* and *jurisdiction_description* that was included in the panama_papers.nodes.entity.csv file and created the "dit2122_jurisdiction" table to keep such data individually. "dit2122_entity" is referring to this info having set as foreign key the column *jurisdiction_id* from table "dit2122_jurisdiction".

c. Table 3 : "*dit2122_provider*"

This table was created with the same reasoning as the "dit2122_jurisdiction" table and its relation with panama_papers.nodes.entity.csv file.

d. Table 4 : "*dit2122_source*"

The column sourceID was included to all of our initial data and therefore would be part of our database tables. Instead of keeping this information to multiple tables I kept it in a single table assigning the *table_source_id* column to the rest tables. This action was intended to make my database more efficient and lessen the disk storage needs.

e. **Table 5 : "*dit2122_valid_until*"**

The column valid_until was included to all of our initial data and therefore would be part of our database tables. Instead of keeping this information to multiple tables I kept it in a single table assigning the *table_valid_until_id* column to the rest tables. This action was intended to make my database more efficient and lessen the disk storage needs.

f. **Table 6 : "*dit2122_entity_status*"**

The status_id column was included in panama_papers.nodes.entity.csv file. Instead of keeping this column I created an individual table inserting an id for the unique values and assigned this status_id column as foreign key in the "dit2122_entity" table.

g. **Table 7 : "*dit2122_intermediary_status*"**

The status_id column was included in panama_papers.nodes.intermediary.csv file. Instead of keeping this column I created an individual table inserting an id for the unique values and assigned this status_id column as foreign key in the "dit2122_intermediary" table.

h. **Table 8 : "*dit2122_officer*"**

This specific table contains the column officer_id as primary key which refers to the column node_id of the file panama_papers.nodes.officer.csv. The columns name and note have taken their values from the same file. The columns country_id, table_source_id and table_valid_until_id have as values the primary key values of the dit2122_countries, dit2122_source and dit2122_valid_until respectively.

i. **Table 9 : "*dit2122_address*"**

This specific table contains the column address_id as primary key which refers to the column node_id of the file panama_papers.nodes.address.csv. The columns name and note have taken their values from the same file. The columns country_id, table_source_id and table_valid_until_id have as values the primary key values of the dit2122_countries, dit2122_source and dit2122_valid_until respectively.

j. **Table 10 : "*dit2122_entity*"**

This table contains the column entity_id as primary key which refers to the column node_id of the file panama_papers.nodes.entity.csv. The columns name,note,incorporation_date, inactivation_date, struck_off_date, closed_date, ibcruc, company_type have taken their values from the same file. The columns jurisdiction_id, country_id, status_id, providerid, table_source_id and table_valid_until have been assigned as foreign keys of the individual tables dit2122_jurisdiction,dit2122_countries, dit2122_entity_status, dit2122_provider,dit2122_source and dit2122_valid_until respectively.

k. Table 11 : "*dit2122_intermediary*"

This specific table contains the column intermediary_id as primary key which refers to the column node_id of the file panama_papers.nodes.intermediary.csv and the columns name and note as well. The columns country_id, status_id, table_source_id and table_valid_until_id have as values the primary key values of the dit2122_countries, dit2122_intermediary_status, dit2122_source and dit2122_valid_until respectively.

l. Table 12 : "*dit2122_officer_address*"

The first of the tables which created through the relations that are "described" in the panama_papers.edges.csv. Setting as conditions the value of column type and the range of values of start_id column I collected the compatible values for the columns. sourceID and valid_until columns are considered as no significant to be part of this table regardless that were part of edges file.

m. Table 13 : "*dit2122_officer_entity*"

This table was created in the same way as the above one. The conditions had to do with a different value of column type(="officer_of") in panama_papers.edges.csv file and the end_id values that should be below 11000000 in order to refer to entities. sourceID and valid_until columns are considered as no significant to be part of this table regardless that were part of edges file

n. Table 14 : "*dit2122_intermediary_entity*"

The third table is also a subset of panama_papers.edges.csv file with condition to the value of column type(="intermediary_of"). sourceID and valid_until columns are considered as no significant to be part of this table regardless that were part of edges file.

o. Table 12 : "*dit2122_entity_address*"

Respectively to the above tables with condition that column type of edges file to be equal to "registered_address" and start_id column values to be less than 11000000 in order to refer to entities.

*panama_papers.nodes.address.csv, panama_papers.nodes.entity.csv,
panama_papers.nodes.intermediary.csv, panama_papers.nodes.officer.csv, panama_papers.edges.csv