Data Engineering 1: Term 1 Project Report

Naida Dzigal

November 19, 2022

1 Task description

The goal of this project is to tie together different pieces learnt during the course. This was an individual exercise, where students were invited to go beyond the SQL statements and techniques covered in the course. A particular emphasis was put on the naming, packaging, versioning, documenting and testing of this project.

Higher level requirements included:

1. **Operational Layer**: Create an operational data layer in MySQL. Import a relational data set of your choosing into your local instance. Find a data set that makes sense to be transformed into an analytical data layer for further analysis. In the ideal case, you may use the outcome of HW1.
2. **Analytics**: Create a short plan of what kind of analytics can be potentially executed on this data set. Plan how the analytical data layer, ETL, Data Mart would look like to support these analytics. (Remember the ProductSales example during class).
3. **Analytical Layer**: Design a normalized data structure using the operational layer. Create a table in MySQL for this structure.
4. **ETL Pipeline**: Create an ETL pipeline using Triggers, Stored procedures. Make sure to demonstrate every element of ETL (Extract, Transform, Load).
5. **Data Mart**: Create Views as data marts.

Each section below contains information for each of the above requirements and these are listed in the same order.

2 Data Set Importing

The data was downloaded using the World Bank APIs, accessed through the Postman software. Seven datasets for 266 countries (data from 1960 - 2021) were acquired for the following variables:

1. Population, indicator SP.POP.TOTL linked here.
2. CO2 emissions, indicator EN.ATM.CO2E.PC linked here.
3. GDP, PPP (current international $), indicator NY.GDP.MKTP.PP.CD linked here.
4. Electricity production from oil gas and coal sources, indicator EG.ELC.FOSL.ZS linked here.
5. Electricity production from nuclear sources, indicator EG.ELC.NUCL.ZS linked here.
6. Electricity production from renewable sources excluding hydroelectric, indicator EG.ELC.RNWX.ZS linked here.
7. Electricity production from hydroelectric sources, with the indicator EG.ELC.HYRO.ZS linked here.

Once the data was downloaded, it was imported into the MySQL Workbench, first by creating a schema named wdi, and then a separate table for each of the datasets above. The table were named: wdi_pop, wdi_co2_emissions, wdi_gdp, wdi_oilgascoal, wdi_nuclear, wdi_renewables and wdi_hydroelec. These tables were populated using csv files and it may be necessary to edit the file paths when reproducing this project. Each table was tested for import failures with the SELECT * FROM (table name) statement after importing, along with DESCRIBE (table name) to review the structure of imported tables. These tables made up the operational data layer (see Figure 1.)

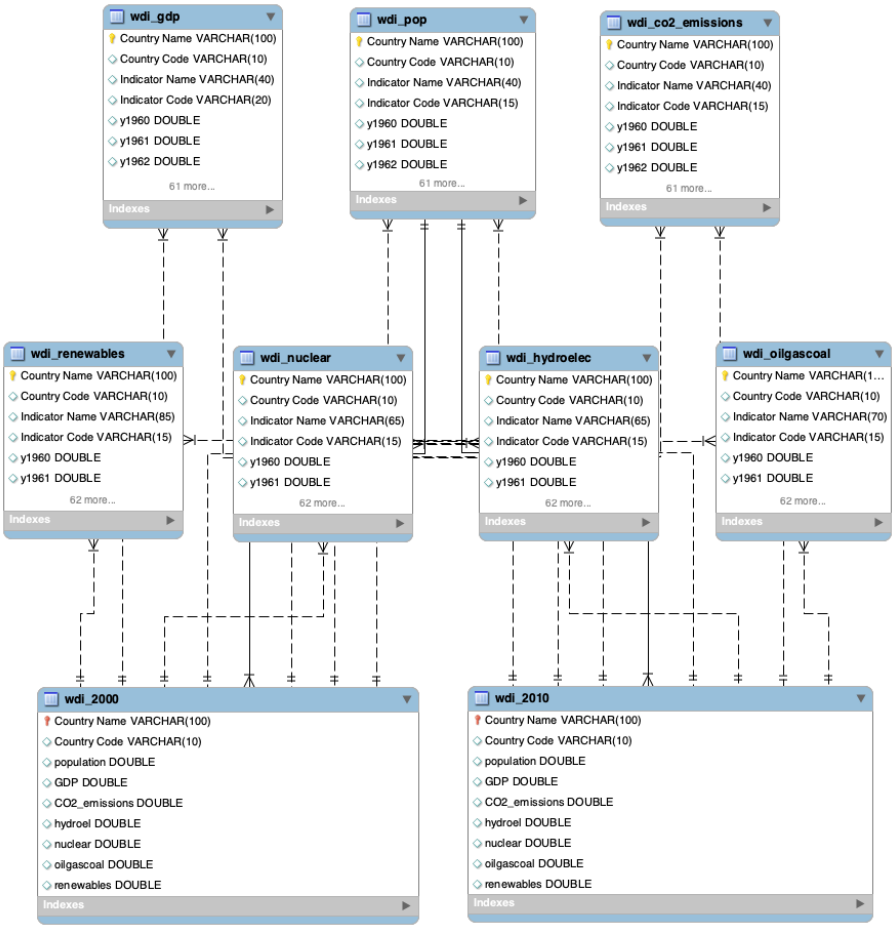


Figure 1: The EER diagram of operational and analytical layer database tables in MySQL. Note that not all fields are shown (e.g. there are hidden year fields that start at y1960 and go to y2021.

3 Analytical Layer

From the operational data layer, two tables were created using automated procedures and together made the analytical layer: wdi_2000 and wdi_2010. Both of these tables contained the following variables (dimensions) for the year 2000 (wdi_2000) or 2010 (wdi_2010):

1. Country Name,
2. Country Code,
3. oilgascoal (a value for percentage of electricity production from oil, gas and coal sources (% of total)),
4. nuclear (a value for percentage of nuclear electricity production (% of total)),
5. hydroel (a value for percentage of hydroelectric electricity production (% of total)),
6. renewables (a value for percentage of electricity production from renewables (% of total)),
7. GDP (GDP, PPP (current international $)),
8. population (population number) and
9. co2_emissions (a value for CO_2 emissions (metric tons per capita)).

Thus, the tables wdi_2000 and wdi_2010 consisted of 7 dimensions for each fact (i.e. country).

See Figures 1. and 2. for a depiction of how the analytical layer fits in the ETL pipeline and also the overall EER diagram. see Figure 3. for a depiction of the dimensions of each of the tables in the analytical layer. Note that both tables wdi_2000 and wdi_2010 have the same dimensions, but these correspond to different years.

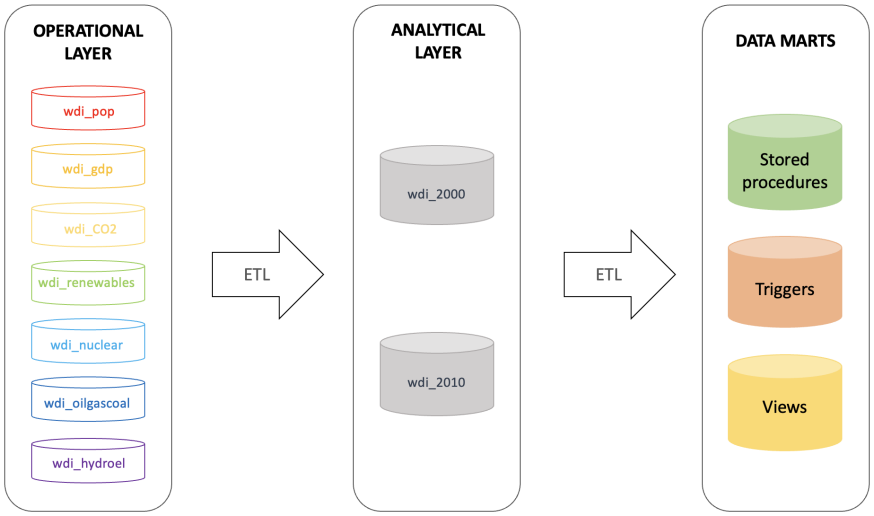


Figure 2: The ETL pipeline depicting the operational, analytical and data marts layers.

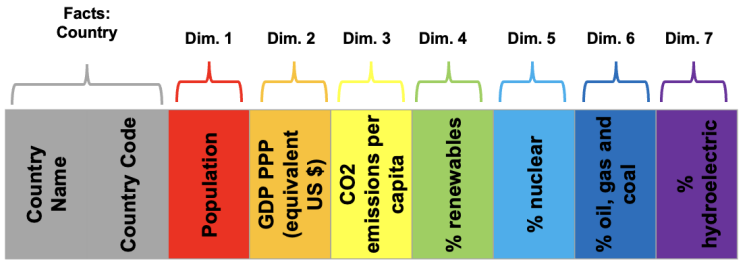


Figure 3: The dimensions for both tables wdi_2000 and wdi_2010.

4 Data Analytics

4.1 Analytics Plan

The aim of this project is to look at countries in central Europe and compare what sources their electricity production consists of and see whether countries with higher GDP also use more low CO2 technologies for their electric energy needs.

For this project, under low carbon emitting technologies we include renewables, hydroelectric and nuclear electricity production.

Ultimately, we want to see if the country with the highest GDP (normalised for population) is also the most responsible in its CO2 emissions (also normalised to per person to make a fair comparison).

We thus try to answer the following questions:

1. What is the GDP pP for Central European countries?
2. What is the CO2 emissions pP for each Central European country?
3. What is the percentage of oil, gas and coal electricity production in Central European countries in 2010?
4. What is the total percentage of low-CO2 emission technologies for electricity production (i.e. nuclear, renewables and hydorelectric) and what percentage of each Central European country's electricity production do they consist of (in the year 2010)?
5. How do those numbers compare to the answers from the first three questions?
6. Was the situation the same in the year 2000?

5 ETL Pipelines

The ETL pipeline in this project was twofold (see Figure 2.): one was used to create snapshots of the operational layer data for the years 2000 (and a second table for the year 2010) and the second consisted of an event, a trigger and several views as data marts. The first one was described in detail in the previous section.

As described above, the analytical layer consists of two snapshots of the operational layer data for the years 2000 and for the year 2010. These snapshots contain the same variables which are percentages of a country's electricity production per source (nuclear, hydroelectric, renewables, oil, gas and coal), as well as information such as the population of a country, its GDP (GDP, PPP current international $) and overall CO2 emissions (in metric tons per capita). The tables were made as part of an automated procedure that joined data from the operational level tables. We note that only one country was inconsistently named (Czechia vs. Czechoslovakia) in the datasets (out of 266 countries).

The second ETL pipeline was built by using MySQL Triggers and Events, and finally data marts with MySQL View to answer the questions from above.

5.1 Events and Triggers

An event was created for checking if there is new data for Austria in the 2000 table: CreateDataCheckAustria()(see Figure 4). This event was set to call Get2000DataByCountry('Austria') every 1 minute in the next 1 hour. The calls can be seen in the messages table (see Figure 6).

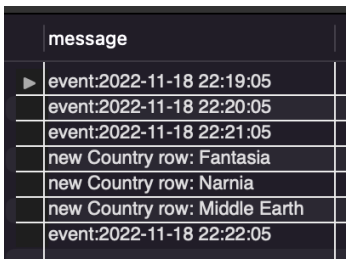
```
CREATE EVENT CreateDataCheckAustria
ON SCHEDULE EVERY 1 MINUTE
STARTS CURRENT_TIMESTAMP
ENDS CURRENT_TIMESTAMP + INTERVAL 1 HOUR
DO
    BEGIN
        INSERT INTO messages SELECT CONCAT('event:', NOW());
            CALL Get2000DataByCountry('Austria');
    END$$
```

Figure 4: An event CreateDAtaCheckAustria() calling Get2000DataByCountry('Austria') every 1 minute for 1 hour.

Next, a trigger as part of the ETL pipeline was written for monitoring the source table wdi_pop (see Figure 5). The trigger was set up so that if there is a change in the table wdi_pop, this will propagate to the analytical layer tables (such as wdi_2000). This was tested on 3 fictional countries inserted into this table, which were then printed by a select statement afterwards. A count(*) was also run to verify that 3 more countries were added to the tables. These changes could be viewed as well in the message table (see Figure 6).

```
CREATE TRIGGER after_insert
AFTER INSERT ON wdi_pop
FOR EACH ROW
BEGIN
    -- log the order number of the newley inserted order
    INSERT INTO messages
    SELECT CONCAT('new Country row: ', NEW.`Country Name`);
    -- archive the order and assosiated table entries to wdi_2000
    INSERT INTO wdi_2000 (`Country Name`)
    SELECT `Country Name`
    FROM   wdi_pop
    WHERE `Country Name` = NEW.`Country Name`;
```

Figure 5: A trigger created for insertions in the wdi_pop table.



message
event:2022-11-18 22:19:05
event:2022-11-18 22:20:05
event:2022-11-18 22:21:05
new Country row: Fantasia
new Country row: Narnia
new Country row: Middle Earth
event:2022-11-18 22:22:05

Figure 6: A list of results in the message table: events of the CreateDataCheckAustria() as well as the messages triggered by adding the countries Fantasia, Narnia and Middle Earth to wdi_pop.

6 Views as Data Marts

The core of this project is in this section: answering the analytical questions from the short data plan above by implementing views as data marts. The SQL syntax was tested in the file 3_analytics.sql , but the views were implemented in the file 5_views_as_data_marts.sql . These are listed in the subsections below.

6.1 GDP normalised for population

The analytical question we try to answer using this view is: What is the GDP pP for Central European countries? The SQL syntax as well as the results table are listed in Figure 7. We see from the results table in Figure 7 that Switzerland has the highest GDP per capita.

```
CREATE VIEW `Central_Europe_GDP_perPop_2010` AS
    SELECT `Country Name`, GDP/population AS `GDP normalised for population in US dollar equivalent
    FROM wdi_2010
    WHERE `Country Name` = 'Austria'
    OR `Country Name` = 'Hungary'
    OR `Country Name` = 'Czechia'
    OR `Country Name` = 'Slovak Republic'
    OR `Country Name` = 'Germany'
    OR `Country Name` = 'Liechtenstein'
    OR `Country Name` = 'Switzerland'
    OR `Country Name` = 'Slovenia'
    ORDER BY GDP/population DESC;

SELECT * FROM Central_Europe_GDP_perPop_2010;
```

100% 40:21

Result Grid Filter Rows: Search Export:

Country Name	GDP normalised for population in US dollar e...
Switzerland	54858.44807583615
Austria	42009.43852774385
Germany	38952.69460210037
Czechia	27881.967997227624
Slovenia	27826.872343755265
Slovak Republic	25302.23271603534
Hungary	21751.83383928097
Liechtenstein	NULL

Figure 7: Central European Countries and the GDP per capita for the year 2010 (in equivalent US dollars). No data was available for Liechtenstein.

6.2 CO2 emissions per capita

The analytical question we try to answer using this view is: What are the CO2 emissions per capita for each Central European country? The SQL syntax as well as the results table are shown in Figure 8.

We see from the above results table that Hungary and Liechtenstein have the lowest CO2 emissions per capita, and Germany and Austria the worst (about double!).

```sql
CREATE VIEW `Central_Europe_CO2_2010` AS
    SELECT `Country Name`, CO2_emissions AS `CO2 emissions (metric tons per capita)`
    FROM wdi_2010
    WHERE `Country Name` = 'Austria'
    OR `Country Name` = 'Hungary'
    OR `Country Name` = 'Czechia'
    OR `Country Name` = 'Slovak Republic'
    OR `Country Name` = 'Germany'
    OR `Country Name` = 'Liechtenstein'
    OR `Country Name` = 'Switzerland'
    OR `Country Name` = 'Slovenia'
    ORDER BY CO2_emissions DESC;

SELECT * FROM Central_Europe_CO2_2010;
-- here we see Germany and Austria are the highest CO2 emitter per capita (in 2010)
```

100% 1:39

Result Grid | Filter Rows: Search | Export:

Country Name	CO2 emissions (metric tons per ca...
Germany	9.4533997219536
Austria	8.36501490116616
Slovenia	7.70288522992293
Slovak Republic	6.57154288347648
Switzerland	5.7777028569243
Liechtenstein	5.27836419646111
Hungary	4.7879890944366

Figure 8: Central European Countries and the CO2 emissions per capita for the year 2010 (in cubic metric tons).

6.3 Electricity production from oil, gas and coal sources (% total)

The analytical question we try to answer using this view is: What is the percentage of oil, gas and coal electricity production in Central European countries in 2010? The SQL syntax as well as the results table are shown in Figure 9.

```sql
CREATE VIEW `Central_Europe_oilgascoal_2010` AS
    SELECT `Country Name`,
    oilgascoal AS `Electricity production from oil, gas and coal (% of total)`
    FROM wdi_2010
    WHERE `Country Name` = 'Austria'
    OR `Country Name` = 'Czechia'
    OR `Country Name` = 'Hungary'
    OR `Country Name` = 'Slovenia'
    OR `Country Name` = 'Germany'
    OR `Country Name` = 'Liechtenstein'
    OR `Country Name` = 'Switzerland'
    OR `Country Name` = 'Slovakia'
    ORDER BY oilgascoal DESC;

SELECT * FROM Central_Europe_oilgascoal_2010;
```

100% 68:58

Result Grid | Filter Rows: Search | Export:

Country Name	Electricity production from oil, gas and coal (...
Germany	59.4572466855947
Hungary	49.3377217628642
Slovenia	35.9520147646878
Austria	32.8686666470398
Switzerland	1.65929873433053
Czechia	NULL
Liechtenstein	NULL

Figure 9: Central European Countries and the percentage of total electricity production from oil, gas and coal sources for the year 2010. No data was available for Czechia or Liechtenstein.

We see from the above results table that Hungary and Germany have the highest percentages of electricity production from oil, gas and coal sources for the year 2010. We note that Switzerland has the lowest (under 2%).

6.4 Electricity production from low carbon (green) sources (% total)

The analytical question we try to answer using this view is: What is the total percentage of low-CO2 emission technologies for electricity production (i.e. nuclear, renewables and hydroelectric)

```sql
CREATE VIEW `Central_Europe_green_2010` AS
    SELECT `Country Name`, SUM(hydroel + nuclear + renewables) AS `El. production from low carbon t
    FROM wdi_2010
    WHERE `Country Name` = 'Austria'
    OR `Country Name` = 'Hungary'
    OR `Country Name` = 'Czechia'
    OR `Country Name` = 'Slovak Republic'
    OR `Country Name` = 'Germany'
    OR `Country Name` = 'Liechtenstein'
    OR `Country Name` = 'Switzerland'
    OR `Country Name` = 'Slovenia'
    GROUP BY `Country Name`
    ORDER BY SUM(hydroel + nuclear + renewables) DESC;
```

100% 1:75

Result Grid Filter Rows: Search Export:

Country Name	El. production from low carbon tech. in % from total
Switzerland	96.60570459637862
Slovak Republic	74.69414506262738
Austria	66.21132275443813
Slovenia	64.01722546908638
Hungary	50.25554574402611
Germany	39.15921753383032
Czechia	NULL

Figure 10: Central European Countries and the percentage electricity production from low carbon technologies for the year 2010.

and what percentage of each Central European country's electricity production do they consist of (in the year 2010)? The SQL syntax as well as the results table are shown in Figure 10.

We see from the above results table that Switzerland and the Slovak Republic have the highest percentages of electricity production from low carbon (green) sources for the year 2010. We note that Germany and Hungary have the lowest (under 2%). No data was available for Czechia.

6.5 Comparison for year 2010

The analytical question we try to answer using this view is: How do those numbers compare to the answers from the first two questions? The SQL syntax as well as the results table are shown in Figure 11. Since we are interested in knowing the most sustainable country, we order them in descending order according to the percentage of low carbon electricity production.

```sql
CREATE VIEW `Central_Europe_comparison_2010` AS
    SELECT `Country Name`,
    oilgascoal AS `El. production from oil, gas and coal sources (% of total)`,
    SUM(hydroel + nuclear + renewables) AS `El. production from low carbon tech. in % from total`,
    CO2_emissions AS `CO2 emissions (metric tons per capita)`,
    GDP/population AS GDP_pP
    FROM wdi_2010
    WHERE `Country Name` = 'Austria'
    OR `Country Name` = 'Hungary'
    OR `Country Name` = 'Czechia'
    OR `Country Name` = 'Slovak Republic'
    OR `Country Name` = 'Germany'
    OR `Country Name` = 'Liechtenstein'
    OR `Country Name` = 'Switzerland'
    OR `Country Name` = 'Slovenia'
    GROUP BY `Country Name`, oilgascoal, CO2_emissions, GDP/population
    ORDER BY SUM(hydroel + nuclear + renewables) DESC;
```

100% 13:101

Result Grid Filter Rows: Search Export:

Country Name	El. production from oil, gas...	El. production from low carbo...	CO2 emissions (metric t...	GDP_pP
Switzerland	1.65929873433053	96.60570459637862	5.7777028569243	54858.44807583615
Slovak Republic	25.0728226041363	74.69414506262738	6.57154288347648	25302.23271603534
Austria	32.8686666470398	66.21132275443813	8.36501490116616	42009.43852774385
Slovenia	35.9520147646878	64.01722546908638	7.70288522992293	27826.872343755265
Hungary	49.3377217628642	50.25554574402611	4.7879890944366	21751.83383928097
Germany	59.4572466855947	39.15921753383032	9.4533997219536	38952.69460210037
Czechia	NULL	NULL	10.7165940611452	27881.967997227624
Liechtenstein	NULL	NULL	5.27836419646111	NULL

Figure 11: Comparison of results for Central European Countries for the year 2010.

We see from the above results table that Switzerland has the highest green energy percentage, the lowest oil, gas and coal electrical energy production, the highest GDP and quite a low overall CO2 emission amount per capita. By comparison, Germany, the third highest GDP per capita country has the least sustainable electrical energy production with the highest CO2 emissions per capita. We conclude that indeed GDP per capita is no indication of overall CO2 emissions or how green their electricity production is.

We note that there was no complete data was available for Czechia or Liechtenstein.

6.6 Comparison for year 2000

The analytical question we try to answer using this view is: Was the situation the same in 2000? The SQL syntax as well as the results table are shown in Figure 12.

```
CREATE VIEW `Central_Europe_comparison_2010` AS
    SELECT `Country Name`,
85      oilgascoal AS `El. production from oil, gas and coal sources (% of total)`,
86      SUM(hydroel + nuclear + renewables) AS `El. production from low carbon tech. in % from total`,
87      CO2_emissions AS `CO2 emissions (metric tons per capita)`,
88      GDP/population AS GDP_pP
89      FROM wdi_2010
90      WHERE `Country Name` = 'Austria'
91      OR `Country Name` = 'Hungary'
92      OR `Country Name` = 'Czechia'
93      OR `Country Name` = 'Slovak Republic'
94      OR `Country Name` = 'Germany'
95      OR `Country Name` = 'Liechtenstein'
96      OR `Country Name` = 'Switzerland'
97      OR `Country Name` = 'Slovenia'
98      GROUP BY `Country Name`, oilgascoal, CO2_emissions, GDP/population
99      ORDER BY SUM(hydroel + nuclear + renewables) DESC;
```

100% 13:101

Result Grid Filter Rows: Search Export:

Country Name	El. production from oil, gas...	El. production from low carbo...	CO2 emissions (metric t...	GDP_pP
Switzerland	1.65929873433053	96.60570459637862	5.7777028569243	54858.44807583615
Slovak Republic	25.0728226041363	74.69414506262738	6.57154288347648	25302.23271603534
Austria	32.8686666470398	66.21132275443813	8.36501490116616	42009.43852774385
Slovenia	35.9520147646878	64.01722546908638	7.70288522992293	27826.872343755265
Hungary	49.3377217628642	50.25554574402611	4.7879890944366	21751.83383928097
Germany	59.4572466855947	39.15921753383032	9.4533997219536	38952.69460210037
Czechia	NULL	NULL	10.7165940611452	27881.967997227624
Liechtenstein	NULL	NULL	5.27836419646111	NULL

Figure 12: Comparison of results for Central European Countries for the year 2000.

We see from Figure 12 that indeed, Switzerland was the most green and highest GDP country with relatively low CO2 emissions in the year 2000 too.

To finalize our investigation, we can use one of the previously written procedures to get the data for Switzerland for both the year 2000 and 2010. This will give us a better idea of the green part of its electricity production. See Figures 13 and 14 for the SQL syntax and results tables.

```
143     -- and in year 2000 this was:
144 •   CALL Get2000DataByCountry('Switzerland');
```

100% 33:144

Result Grid Filter Rows: Search Export:

Country Name	Country Code	population	GDP	CO2_emissions	hydroel	nuclear	oilgascoal	renewables
Switzerland	CHE	7184250	26418423...	6.0841423948...	55.7044...	39.99455...	1.63934426...	1.29604984...

Figure 13: Switzerland data for the year 2000.

We note from figures 13. and 14. that hydroelectric and nuclear sources make over 90% of Switzerland's total electricity production. We see a slight increase from year 2000 to 2010 of (other) renewables electricity production too, which explains the slight drop in CO2 emissions per capita.

```
139     -- To see what the electricty production of Switzerland consisted of, we can run the
140     -- stored procedure from part 4 (4_ETL_triggs_stored_pro.sql):
141 •   CALL Get2010DataByCountry('Switzerland');
```

100% 25:141

Result Grid Filter Rows: Search Export:

Country Name	Country Code	population	GDP	CO2_emissions	hydroel	nuclear	oilgascoal	renewables
Switzerland	CHE	7824909	42926236…	5.7777028569243	54.59486…	39.876158…	1.65929873…	2.13468176…

Figure 14: Switzerland data for the year 2010.

7 Conclusion

The project aimed to look at countries in central Europe and compare what sources their electricity production consists of as well as see whether countries with higher GDP also use more low carbon technologies for their electrical energy needs. We wanted to ultimately conclude if the country with the highest GDP (normalised for population) is also the most responsible (regarding its CO2 emissions).

Our investigation concluded that indeed, Switzerland as the highest GDP per capita country also had the highest percentage of electricity production from low carbon technologies. It had almost the lowest CO2 emissions per capita from all Central European countries. Overall we confirmed our speculation, but note that it was not an overall trend: Germany for instance, the third largest GDP per capita country in this region, had the worst CO2 emissions per capita (almost double that of Switzerland) and the highest electricity production from oil, gas and coal sources.