ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ

ΠΡΟΧΩΡΗΜΈΝΑ ΘΈΜΑΤΑ ΤΕΧΝΟΛΟΓΊΑΣ ΚΑΙ ΕΦΑΡΜΟΓΏΝ ΒΆΣΕΩΝ ΔΕΔΟΜΈΝΩΝ

ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΕΡΓΑΣΙΑ ΓΙΑ ΤΟ ΑΚΑΔΗΜΑΪΚΟ ΈΤΟΣ 2019-2020

ΟΜΑΔΑ ANDROU_SINT

ΑΝΔΡΟΥΤΣΟΠΟΥΛΟΣ ΓΕΩΡΓΙΟΣ, 2933

ΣΙΝΤΟΡΗΣ ΝΙΚΟΛΑΟΣ , 3071

ΤΕΛΙΚΗ ΑΝΑΦΟΡΑ

ΙΟΥΝΙΟΣ 2020

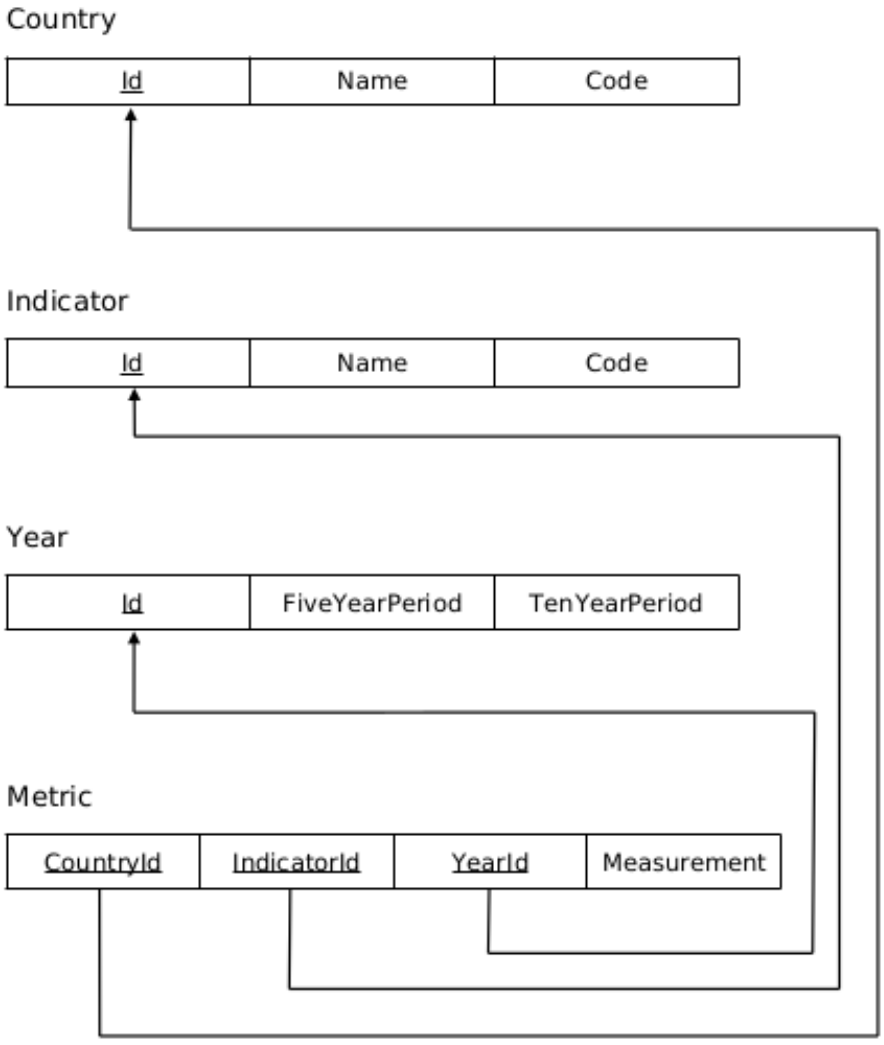
ΙΣΤΟΡΙΚΟ ΠΡΟΗΓΟΥΜΕΝΩΝ ΕΚΔΟΣΕΩΝ

Ημερομηνία	Έκδοση	Περιγραφή	Συγγραφέας
yyyy/mm/dd	x.x		

1 ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

1.1 ΣΧΕΣΙΑΚΌ ΣΧΉΜΑ ΣΕ ΛΟΓΙΚΌ ΕΠΊΠΕΔΟ

Database Name: WORLDMETRIC



Εντολές Δημιουργίας Σχήματος Βάσης Δεδομένων:

```
CREATE TABLE IF NOT EXISTS Country (Id INT, Name VARCHAR(60),
                                    Code VARCHAR(10), PRIMARY KEY(Id));
CREATE TABLE IF NOT EXISTS Indicator (Id INT, Name VARCHAR(255),
                                      Code VARCHAR(40), PRIMARY KEY(Id));
CREATE TABLE IF NOT EXISTS Year (Id INT, FiveYearPeriod VARCHAR(20),
                                 TenYearPeriod VARCHAR(20), PRIMARY KEY(Id));
CREATE TABLE IF NOT EXISTS Metric (CountryId INT, IndicatorId INT, YearId INT,
        Measurement DECIMAL(8,4), PRIMARY KEY(CountryId, IndicatorId, YearId),
        CONSTRAINT CountryId FOREIGN KEY(CountryId) REFERENCES Country(Id),
        CONSTRAINT IndicatorId FOREIGN KEY(IndicatorId) REFERENCES Indicator(Id),
        CONSTRAINT YearId FOREIGN KEY(YearId) REFERENCES Year(Id));
```

Παρατηρήσεις Επιλεγμένου Σχήματος:

1. Το μειονέκτημα αυτού του σχήματος είναι ότι για την δημιουργία κάθε γραφήματος απαιτούνται 3 joins μεταξύ και των τεσσάρων σχέσεων (Country, Indicator, Year, Metric).
2. Το πλεονέκτημα αυτού του σχήματος είναι ότι η εισαγωγή μετρήσεων για κάθε νέο έτος γίνεται πολύ εύκολα προσθέτοντας μία επιπλέον εγγραφή στη σχέση Year και όσων επιπλέον εγγραφών χρειάζονται στη σχέση Metric.
3. Στις σχέσεις **Country** και **Indicator** προστέθηκε ένα **integer** artificial key. Αυτό έγινε ώστε να είναι πιο εύκολη η δημιουργία και η χρήση των αντίστοιχων B+ tree indexes, σε αντίθεση με την δημιουργία τους απευθείας στα **string type** Code attributes. Στη σχέση Year χρησιμοποιήθηκε ως primary key η χρονολογία που ήταν ήδη **integer type.**
4. Οι παραπάνω εντολές δημιουργίας του σχήματος της Βάσης Δεδομένων, εκτελέστηκαν μέσω python script (source: create_db.py) με την χρήση της βιβλιοθήκης mysql.connector και συγκεκριμένα της συνάρτησης execute().

1.2 ΔΗΜΙΟΥΡΓΙΑ BACKUP

Η δημιουργία των backup files της βάσης δεδομένων: WORLDMETRIC γίνονται από το python script: backup_db.py.
Η εντολή με την οποία κάθε σχέση της ΒΔ αποθηκεύεται σε ένα csv αρχείο είναι της μορφής: **SELECT ... INTO OUTFILE '/tmp/*_data.csv'**. Όπου * το όνομα της κάθε σχέσης (country, indicator, year, metric).

1.3 ΑΛΛΕΣ ΔΙΑΧΕΙΡΙΣΤΙΚΕΣ ΛΕΙΤΟΥΡΓΙΕΣ ΤΗΣ ΒΔ

Λειτουργία Drop Database: μέσω του python script: drop_db.py, γίνεται η διαγραφή ολόκληρης της βάσης δεδομένων.

Λειτουργία Truncate Database: μέσω του python script: truncate_db.py, γίνεται η διαγραφή όλων των εγγραφών όλων των σχέσεων της βάσης δεδομένων, διατηρώντας όμως το σχήμα της.

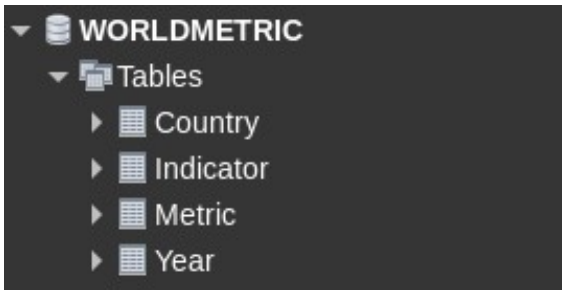
1.4 ΡΥΘΜΙΣΗ ΤΩΝ ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ DBMS

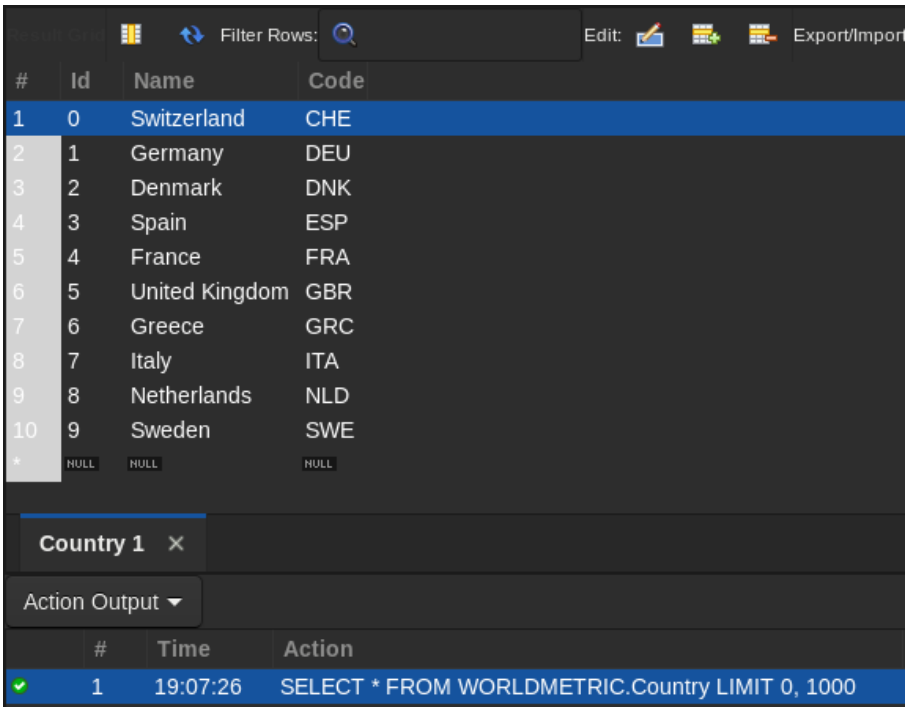
Ρυθμίσεις για την χρήση της εντολής: **LOAD DATA INFILE ...** :

(Τα παρακάτω βήματα επιτρέπουν την φόρτωση δεδομένων στην ΒΔ μέσω της εντολής LOAD DATA INFILE, τα οποία μπορεί να βρίσκονται οπουδήποτε στο σύστημά μας.)

1. Έλεγχος της παραμέτρου: secure_file_priv με την εντολή:
   SHOW variables LIKE “secure_file_priv”;
   Αν η τιμή του είναι μη κενή τότε κάνε τα ακόλουθα.
2. Βρες το configuration file my.cnf που βρίσκεται στον κατάλογο (/etc/mysql/).
   Προσοχή αυτό το path μπορεί να διαφέρει σε κάποια συστήματα.
3. Στο τέλος του αρχείου αυτού πρόσθεσε τα εξής:
   [mysqld] // Αν δεν υπάρχει ήδη.
   secure_file_priv = “”
4. Επανεκκίνησε τον demon της mysql που τρέχει στο σύστημα. Η εντολή επανεκκίνησης του mysql deamon, επίσης διαφέρει σε κάθε σύστημα, σε ubuntu distros είναι η εξής: “sudo /etc/init.d/mysql restart”.

1.5 SCREENSHOTS ΤΗΣ ΒΔ ΜΕΣΑ ΑΠΟ ΤΟ WORKBENCH





Query 2 | Indicator

Filter Rows: | Edit: | Export/Impor

#	Id	Name	Code
1	0	Merchandise exports by the reporti…	TX.VAL.MRCH.RS.ZS
2	1	Merchandise imports by the reporti…	TM.VAL.MRCH.RS.ZS
3	2	Urban population (% of total popula…	SP.URB.TOTL.IN.ZS
4	3	Population, male (% of total populat…	SP.POP.TOTL.MA.ZS
5	4	Population, female (% of total popul…	SP.POP.TOTL.FE.ZS
6	5	Age dependency ratio, young (% of …	SP.POP.DPND.YG
7	6	Military expenditure (% of GDP)	MS.MIL.XPND.GD.ZS
8	7	Population in the largest city (% of u…	EN.URB.LCTY.UR.ZS
9	8	Population ages 25-29, male (% of …	SP.POP.2529.MA.5Y
10	9	Population ages 25-29, female (% o…	SP.POP.2529.FE.5Y
*	NULL	NULL	NULL

Indicator 1

Action Output

#	Time	Action
1	19:10:07	SELECT * FROM WORLDMETRIC.Indicator LIMIT 0, 1000

Year

Filter Rows: | Edit: | Export/Impor

#	Id	FiveYearPeriod	TenYearPeriod
1	1960	1960-1964	1960-1969
2	1961	1960-1964	1960-1969
3	1962	1960-1964	1960-1969
4	1963	1960-1964	1960-1969
5	1964	1960-1964	1960-1969
6	1965	1965-1969	1960-1969
7	1966	1965-1969	1960-1969
8	1967	1965-1969	1960-1969
9	1968	1965-1969	1960-1969
10	1969	1965-1969	1960-1969
11	1970	1970-1974	1970-1979
12	1971	1970-1974	1970-1979

Year 1

Action Output

#	Time	Action
1	19:12:46	SELECT * FROM WORLDMETRIC.Year LIMIT 0, 1000

Metric

Filter Rows: Edit: Export/Import

#	CountryId	IndicatorId	YearId	Measurement
1	0	0	1960	6.7522
2	0	0	1961	6.1537
3	0	0	1962	5.6544
4	0	0	1963	5.9738
5	0	0	1964	5.6643
6	0	0	1965	5.4608
7	0	0	1966	5.4918
8	0	0	1967	5.2199
9	0	0	1968	5.4074
10	0	0	1969	5.6266
11	0	0	1970	6.1094
12	0	0	1971	5.7399

Metric 1

Action Output

#	Time	Action
1	19:14:54	SELECT * FROM WORLDMETRIC.Metric LIMIT 0, 1000

2 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΛΟΓΙΣΜΙΚΟΎ

2.1 ΑΡΧΙΤΕΚΤΟΝΙΚΗ ΚΑΙ ΔΟΜΗ ETL

2.1.1 ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

Τα αρχικά δεδομένα είναι 10 csv files (1 ανά χώρα) τα οποία σε κάθε γραμμή περιέχουν κωδικούς και ονόματα (χωρών και δεικτών), αλλά και τιμές μετρικών για κάθε έτος από το 1960 έως το 2019.
Τα αρχεία αυτά επιλέχθηκαν από τον ιστότοπο: https://www.worldbank.org.
Τα αρχεία αυτά βρίσκονται στον κατάλογο: data/original_data/.
Δύο είναι τα script τα οποία αναλαμβάνουν την προ-επεξεργασία των δεδομένων.

Script **csv_parser.py**:

- Το script αυτό αναλαμβάνει το πρώτο πέρασμα των αρχικών αρχείων και την εύρεση ενός συνόλου δεικτών που πληρούν κάποια κριτήρια.
- Συγκεκριμένα:

  Το script διατρέχει όλα τα raw αρχεία (1 αρχείο ανά χώρα). Για κάθε ένα από αυτά κρατάει σε μία λίστα τους κωδικούς των δεικτών (indicatorCodes) για τους οποίους υπάρχουν μετρήσεις για τουλάχιστον 59 έτη για την συγκεκριμένη χώρα. Επιπλέον διατηρείται και μία global λίστα, με την τομή όλων αυτών των παραπάνω λιστών, η οποία ανανεώνεται μετά το πέρας της προσπέλασης του κάθε αρχείου. Άρα στο τέλος προκύπτει μία λίστα με όλους τους κωδικούς των δεικτών, για τους οποίους υπάρχουν σε όλες τις χώρες μετρήσεις για τουλάχιστον 59 έτη. Από αυτή την λίστα επιλέγουμε 10 δείκτες οι οποίοι θα δοθούν στο script csv_writer, ώστε να φτιάξει τα τελικά επεξεργασμένα αρχεία φόρτωσης της ΒΔ.

Script **csv_writer.py**:

- Το script αυτό αναλαμβάνει τη δημιουργία των τελικών-επεξεργασμένων αρχείων από τα οποία θα φορτωθεί η ΒΔ.

- Συγκεκριμένα:

  Το script αποτελείται από 4 συναρτήσεις (μία για κάθε σχέση του σχήματος της ΒΔ). Καθεμία εξ' αυτών είναι υπεύθυνη για την δημιουργία του csv αρχείου από το οποίο θα φορτωθούν τα δεδομένα της σχέσης της. Οι συναρτήσεις που δημιουργούν τα αρχεία φόρτωσης των σχέσεων **County, Indicator, Metric** χρησιμοποιούν ως είσοδο τα αρχικά raw αρχεία, καθώς και την λίστα των επιλεγμένων δεικτών και κρατάνε μόνο τις γραμμές που αφορούν τους δείκτες αυτούς (filtering). Στην συνέχεια οι γραμμές αυτές γράφονται στα αντίστοιχα τελικά αρχεία σύμφωνα με το σχήμα της κάθε σχέσης. Όσον αφορά την συνάρτηση για την σχέση **Year**, δημιουργεί ένα csv αρχείο με μία γραμμή για κάθε έτος. Σε κάθε μία από αυτές τις γραμμές υπάρχει το αντίστοιχο έτος καθώς και η πενταετία, δεκαετία στην οποία ανήκει αυτό.

2.1.2 ΦΟΡΤΩΣΗ ΔΕΔΟΜΕΝΩΝ

Δύο είναι τα script τα οποία αναλαμβάνουν την φόρτωση της βάσης από τα επεξεργασμένα (ή backup) csv αρχεία.

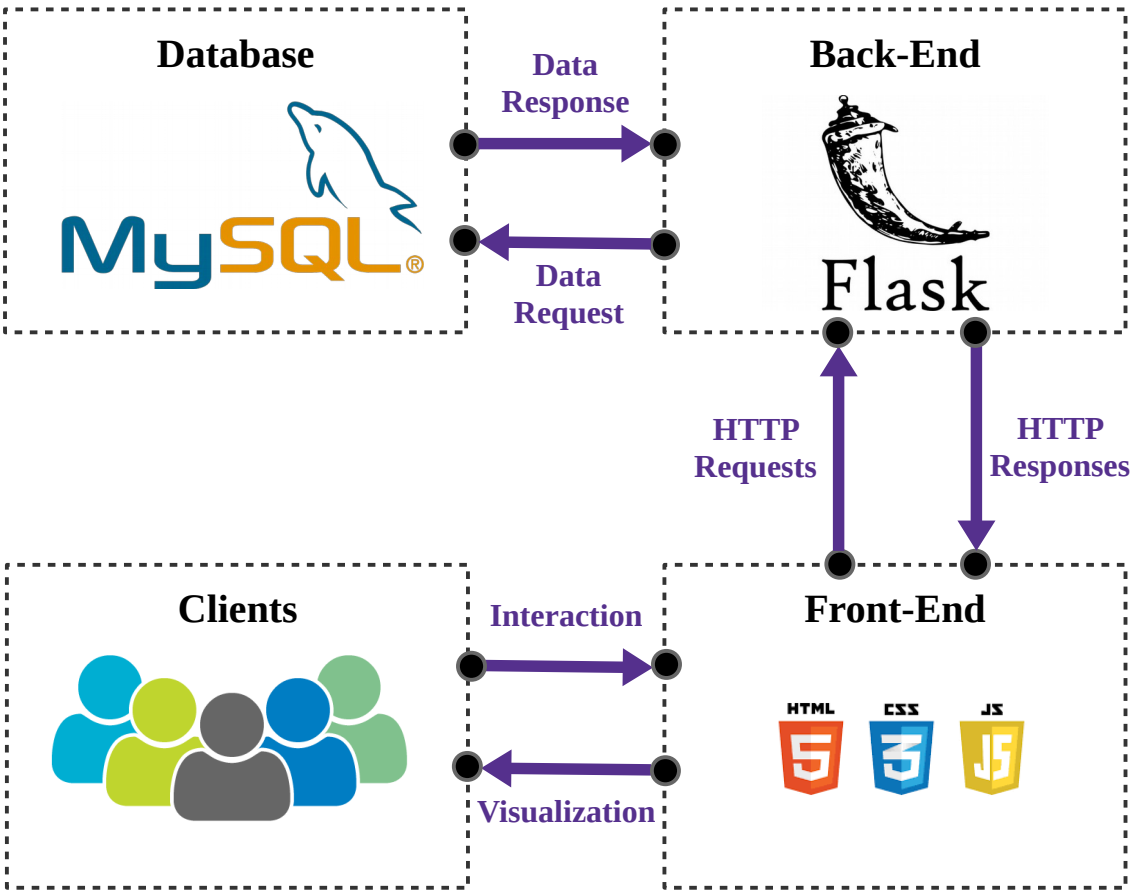
Script **load_orig_db.py**:

- Το script αυτό αναλαμβάνει το φόρτωμα της ΒΔ από τα επεξεργασμένα αρχεία που έδωσε η csv_writer ως έξοδο. Για το φόρτωμα τους χρησιμοποιείται η εντολή **LOAD DATA INFILE ...** με τις κατάλληλες παραμέτρους.

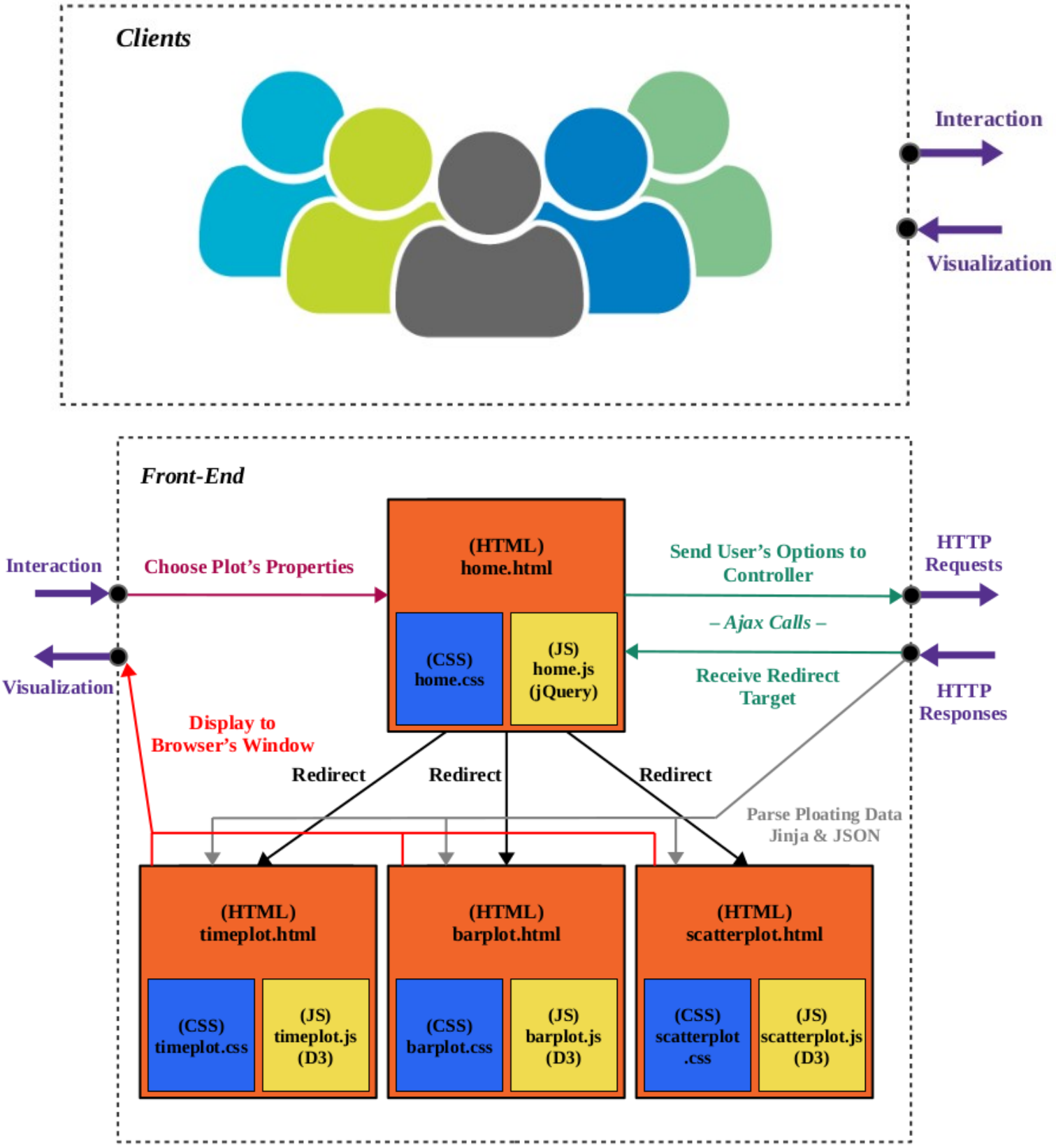
Script **load_backup_db.py**:

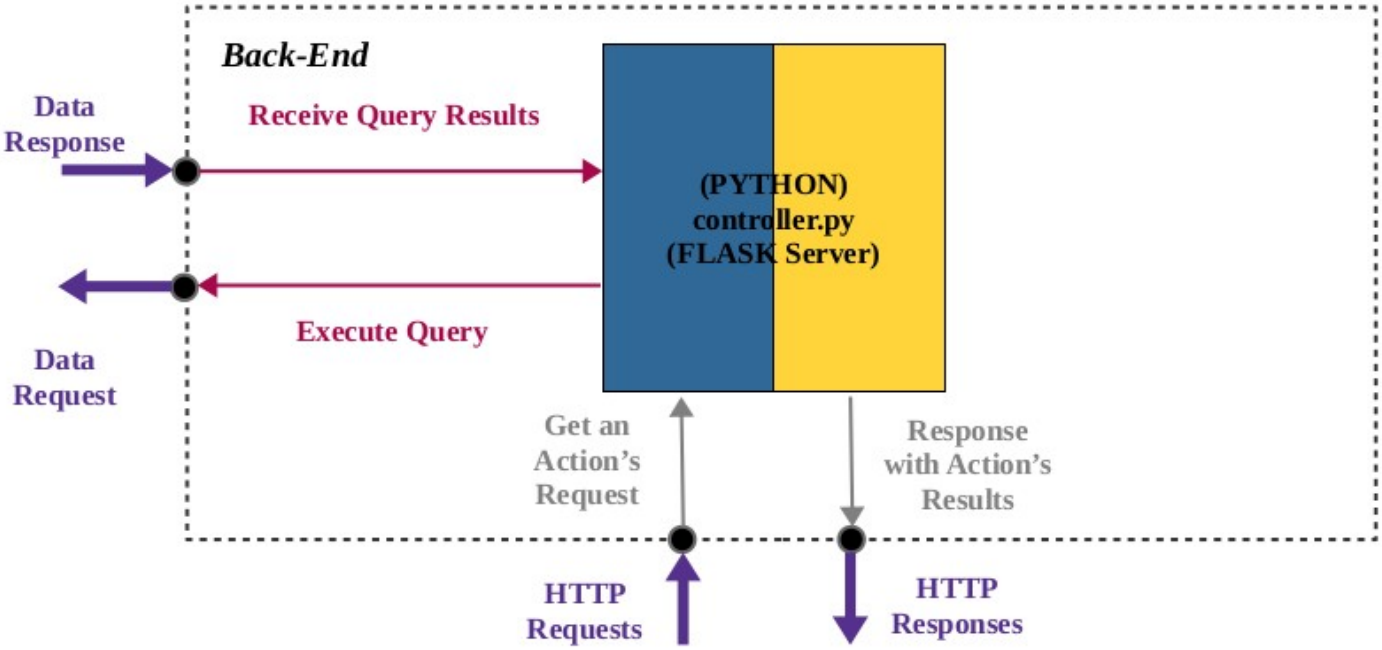
- Το script αυτό αναλαμβάνει το φόρτωμα της ΒΔ από τα backup αρχεία που έδωσε η backup_db ως έξοδο. Για το φόρτωμα τους χρησιμοποιείται η εντολή **LOAD DATA INFILE ...** με τις κατάλληλες παραμέτρους.

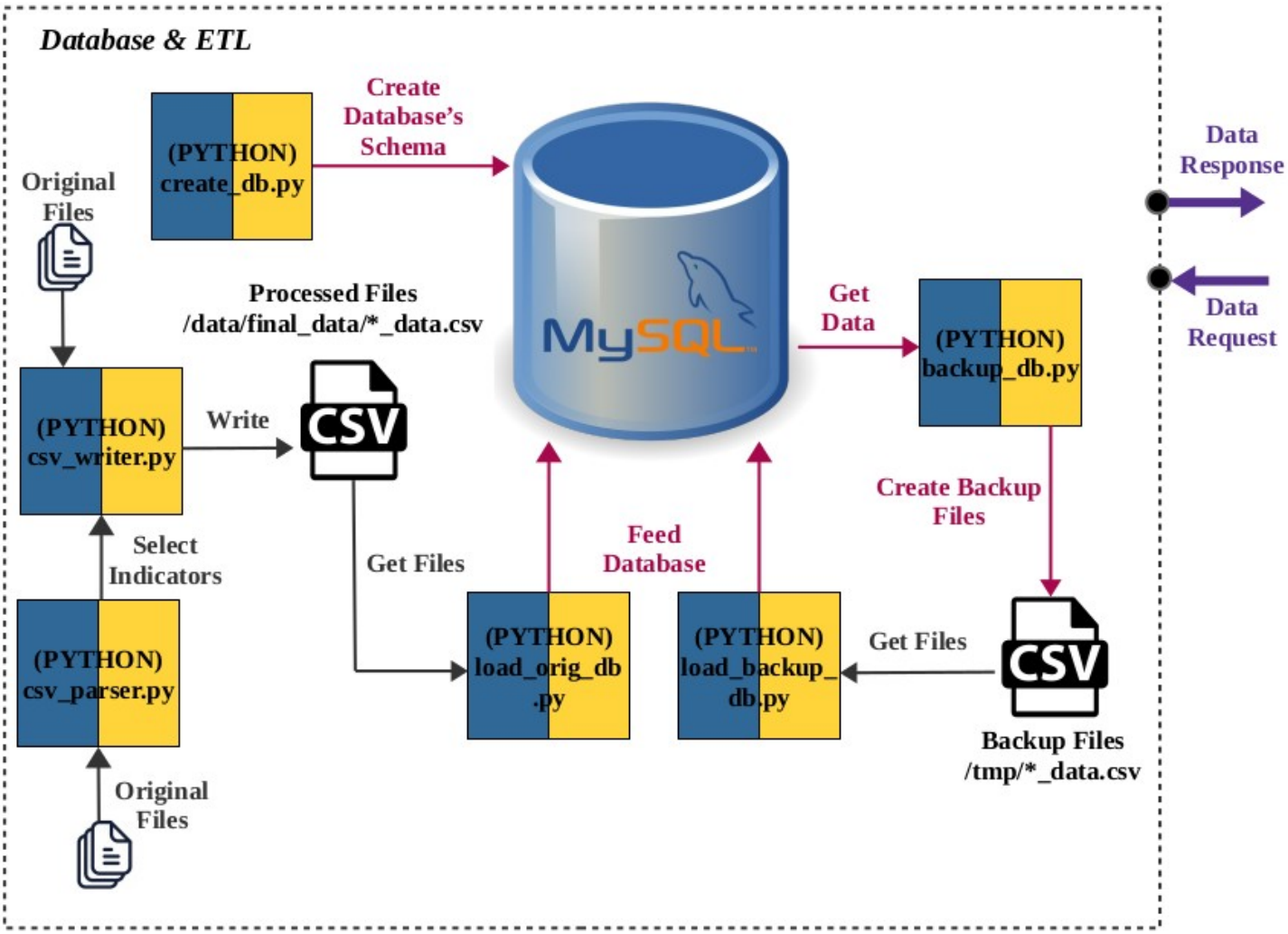
2.2 ΥΠΟΣΥΣΤΗΜΑΤΑ ΚΕΝΤΡΙΚΗΣ ΕΦΑΡΜΟΓΗΣ



2.3 ΔΟΜΗ ΥΠΟΣΥΣΤΗΜΑΤΩΝ ΚΕΝΤΡΙΚΗΣ ΕΦΑΡΜΟΓΗΣ




Back-End
Data Response
Receive Query Results
(PYTHON) controller.py (FLASK Server)
Execute Query
Data Request
Get an Action's Request
Response with Action's Results
HTTP Requests
HTTP Responses


Database & ETL
(PYTHON) create_db.py
Create Database's Schema
Original Files
Processed Files /data/final_data/*_data.csv
MySQL
Data Response
Data Request
Get Data
(PYTHON) backup_db.py
(PYTHON) csv_writer.py
Write
CSV
Create Backup Files
Select Indicators
Get Files
Feed Database
(PYTHON) csv_parser.py
(PYTHON) load_orig_db.py
(PYTHON) load_backup_db.py
Get Files
CSV
Backup Files /tmp/*_data.csv
Original Files

3 ΥΠΟΔΕΊΓΜΑΤΑ ΕΡΩΤΉΣΕΩΝ ΚΑΙ ΑΠΑΝΤΉΣΕΩΝ

Βλέπε σχετικό βίντεο.