

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ**  
**ΤΜΗΜΑ ΜΗΧ. Η/Υ & ΠΛΗΡΟΦΟΡΙΚΗΣ**

**ΜΥΕ030/  
ΠΛΕ045**

**ΠΡΟΧ. ΘΕΜΑΤΑ ΤΕΧΝΟΛΟΓΙΑΣ & ΕΦΑΡΜΟΓΩΝ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ**  
**Π. Βασιλειάδης**

**ΑΝΟΙΞΗ 2020**

**ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΗ ΑΣΚΗΣΗ**

**Ημερομηνία Εξέτασης: Πέμπτη 21-05-2020**

Η προγραμματιστική άσκηση για το μάθημα είναι υποχρεωτική και αφορά τη σχεδίαση, υλοποίηση και ρύθμιση ενός ολοκληρωμένου πληροφοριακού συστήματος (κατασκευή βάσης δεδομένων, διαπροσωπεία, ρύθμιση λειτουργίας). Η εργαστηριακή άσκηση προσφέρει **3 μονάδες** στον τελικό βαθμό του μαθήματος. Φυσικά, πρέπει να πιάσετε τουλάχιστον τη βάση στην εργασία, όπως και στο διαγώνισμα. Σε περιπτώσεις εξαιρετικών εργασιών, η επίδοση επιβραβεύεται με bonus που μπορεί να φτάσει ως και μία μονάδα στον τελικό βαθμό.

**Οι προθεσμίες είναι ιερές.**

**Είναι υποχρεωτικό να υλοποιήσετε τουλάχιστον ένα σύστημα με σχεσιακό back-end και γραφική διαπροσωπεία + την τελική αναφορά (βλ. στο τέλος της εκφώνησης).**

Για φέτος, αποφάσισα ότι η έμφαση θα δοθεί στο πρόβλημα της *οπτικοποίησης δεδομένων*. Ο στόχος των τεχνικών οπτικοποίησης είναι να δώσουν στον χρήστη την πληροφορία με τρόπο που αναδεικνύει οπτικά ιδιότητες, τάσεις και πρότυπα που βρίσκονται κρυμμένα στα δεδομένα.

*Why bother? Κυρίως, γιατί ζούμε σε μια εποχή που έχουμε όλο και πιο πολλά δεδομένα γύρω μας, και γίνεται όλο και πιο δύσκολο να τα αξιοποιήσουμε, ρωτώντας τα. Οι απαντήσεις στις ερωτήσεις πλέον δεν αρκούν: στους χρήστες πρέπει να παρουσιάζονται και ενδιαφέρουσες ιδιότητες εντός των δεδομένων.*

Το project που θα κληθείτε να υλοποιήσετε στηρίζεται στα δεδομένα του οργανισμού World Bank. Στην τοποθεσία <http://data.worldbank.org/> θα βρείτε πλείστα όσα αρχεία για διάφορα είδη δεδομένων που χαρακτηρίζουν τον κόσμο μας τα τελευταία 50+ χρόνια. Τα αρχεία από μόνα τους, βέβαια, δεν προσφέρονται ούτε για την απάντηση ερωτήσεων, ούτε για διαδραστικές οπτικοποιήσεις. Ως εκ τούτου, *για να μπορούμε να απαντήσουμε ενδιαφέρουσες ερωτήσεις, πρέπει να οργανώσουμε τα δεδομένα σε μια βάση δεδομένων και να χτίσουμε μια εφαρμογή γύρω τους!*

### **Δεδομένα**

Τα δεδομένα της WorldBank μπορείτε να τα βρείτε στο διακτυακό της τόπο (<http://data.worldbank.org/>). Εμείς ενδιαφερόμαστε για τα δεδομένα που προσφέρονται ανά χώρα ( <http://data.worldbank.org/country> ).

Τα δεδομένα μπορείτε να τα κατεβάσετε σε μορφή xls, csv, xml. Οι δείκτες που καταγράφονται στα αρχεία αυτά είναι περίπου 1300 δείκτες με καταγραφές από το 1960 ως και σήμερα (ανάλογα με τη χώρα βέβαια). Κάθε αρχείο xls έχει και τη μεταπληροφορία για τους δείκτες αυτούς. Όπως είναι η συνήθης πρακτική, όταν έχουμε τέτοιο όγκο μεταπληροφορίας, την οργανώνουμε σε *κατηγορίες/υποκατηγορίες*. Δέστε για παράδειγμα ένα απόκομμα από το αρχείο <http://data.worldbank.org/country/greece> στο Σχήμα 1.

	A	B	C	D	E	F	G	H	I	J	K	
1	Data Source	World Development Indicators										
2												
3	Country Name	Country	Indicator Name	Indicator Code	1961	1962	1963	1964	1965	1966	1967	1968
4	Greece	GRC	Agricultural machinery, tractors	AG.AGR.TRAC.NO	22630	24530	28500	33500	39318	44774	50857	
5	Greece	GRC	Fertilizer consumption (% of fertilizer production)	AG.CON.FERT.PT.ZS								
6	Greece	GRC	Fertilizer consumption (kilograms per hectare of arable land)	AG.CON.FERT.ZS								
7	Greece	GRC	Agricultural land (sq. km)	AG.LND.AGRI.K2	89100	89020	90210	89910	86780	90900	91130	
8	Greece	GRC	Agricultural land (% of land area)	AG.LND.AGRI.ZS	69,1233514	69,0612878	69,9844841	69,7517455	67,3235066	70,5197828	70,6982157	70,8411130
9	Greece	GRC	Arable land (hectares)	AG.LND.ARBL.HA	2794000	2863000	3057000	3001000	2991000	2995000	3020000	
10	Greece	GRC	Arable land (hectares per person)	AG.LND.ARBL.HA.PC	0,33269628	0,33888743	0,36051123	0,35262617	0,34981094	0,34770389	0,34776248	0,34776248
11	Greece	GRC	Arable land (% of land area)	AG.LND.ARBL.ZS	21,6757176	22,2110163	23,716059	23,2816137	23,2040341	23,2350659	23,4290147	23,4290147
12	Greece	GRC	Land under cereal production (hectares)	AG.LND.CREL.HA	1772952	1758547	1644087	1766280	1779520	1716990	1690617	
13	Greece	GRC	Permanent cropland (% of land area)	AG.LND.CROP.ZS	7,02870442	6,50892164	6,50892164	6,57098526	6,69511249	6,64080683	6,62529092	6,62529092
14	Greece	GRC	Land area where elevation is below 5 meters (% of total land area)	AG.LND.ELSM.ZS								
15	Greece	GRC	Forest area (sq. km)	AG.LND.FRST.K2								
16	Greece	GRC	Forest area (% of land area)	AG.LND.FRST.ZS								
17	Greece	GRC	Agricultural irrigated land (% of total agricultural land)	AG.LND.IRIG.ZS								
18	Greece	GRC	Average precipitation in depth (mm per year)	AG.LND.PRCP.MM		652						652
19	Greece	GRC	Land area (sq. km)	AG.LND.TOTL.K2	128900	128900	128900	128900	128900	128900	128900	128900
20	Greece	GRC	Agricultural machinery, tractors per 100 sq. km of arable land	AG.LND.TRAC.ZS	80,9949893	85,6793573	93,2286555	111,629457	131,454363	149,495826	168,400662	171,629457
21	Greece	GRC	Cereal production (metric tons)	AG.PRO.CREL.MT	2243876	2426843	2122537	2874641	2940922	3131459	3296848	
22	Greece	GRC	Crop production index (2004-2006 = 100)	AG.PRO.CROP.XD	53,94	43,22	49,95	48,35	52,31	54,32	55,64	
23	Greece	GRC	Food production index (2004-2006 = 100)	AG.PRO.FOOD.XD	54,72	45,48	51,38	50,45	54,98	57,99	59,36	
24	Greece	GRC	Livestock production index (2004-2006 = 100)	AG.PRO.LVSK.XD	41,35	45,64	48,75	50,17	53,33	58,16	60,94	
25	Greece	GRC	Surface area (sq. km)	AG.SRF.TOTL.K2	131960	131960	131960	131960	131960	131960	131960	131960
26	Greece	GRC	Cereal yield (kg per hectare)	AG.YLD.CREL.KG	1265,616	1380,027	1291,013	1627,511	1652,649	1823,807	1950,086	1950,086
27	Greece	GRC	(%) Benefits held by 1st 20% population - All Social Safety Nets	allsa.bi_q1								
28	Greece	GRC	(%) Program participation - All Social Safety Nets	allsa.cov_pop								
29	Greece	GRC	(%) Generosity of All Social Safety Nets	allsa.gen_pop								
30	Greece	GRC	(%) Benefits held by 1st 20% population - All Social Insurance	allsa.bi_q1								

Σχήμα 1. Απόκομμα αρχείου με δεδομένα μιας χώρας

## Στόχος

Ο τελικός σκοπός σας ως ομάδες είναι να μπορέσετε να υλοποιήσετε μια εφαρμογή οπτικής εξαγωγής συμπερασμάτων η οποία θα αξιοποιεί δεδομένα που θα έχουν ενσωματωθεί σε μια βάση δεδομένων. Κάθε ομάδα θα αναλάβει να κατεβάσει δεδομένα για ένα σύνολο χωρών και να ασχοληθεί με ένα υποσύνολο δεικτών που υπάρχουν στα δεδομένα αυτά. Η ανάθεση θα γίνει μετά τη συγκρότηση ομάδων (βλ. στο τέλος για χρονοπρογραμματισμό εργασιών).

Το project έχει τρεις φάσεις: (α) setup & προεπεξεργασία DBMS και δεδομένων, (β) σχεδίαση και φόρτωση δεδομένων και (γ) ανάπτυξη εφαρμογής.

## ΦΑΣΗ Ι: αρχική οργάνωση

Κάθε ομάδα πρέπει να προβεί στις παρακάτω ενέργειες:

1. Στήσιμο της βάσης και ενός γραφικού εργαλείου διαχείρισης (π.χ., MySQL & MySQL Workbench) στο μηχανήμά σας.
2. Download το κομμάτι των δεδομένων που της αναλογεί – τα αντίστοιχα αρχεία δηλαδή.
3. Δημιουργήστε το σχήμα της βάσης για τα δεδομένα που σας αναλογούν – όπως θα συζητήσουμε στο μάθημα (βλ. υποδείξεις στο παρακάτω παράδειγμα). Χρησιμοποιήστε InnoDB τύπο αποθήκευσης στη MySQL.
4. Δημιουργία scripts που μετατρέπουν τα εισερχόμενα αρχεία σε αρχεία φόρτωσης δεδομένων – αρχεία δηλαδή, στα οποία τα δεδομένα είναι έτοιμα προς φόρτωση
5. Δημιουργία scripts φόρτωσης των αρχείων φόρτωσης (π.χ., δείτε την εντολή LOAD DATA INFILE στη MySQL)
6. Φόρτωση των αρχείων και εξαγωγή backup της βάσης

**Σχεδίαση.** Το αρχείο έχει δεδομένα σε ένα συγκεκριμένο format. Η απεικόνισή του σε ένα σχεσιακό σχήμα δεν είναι μονόδρομος. Εδώ, στο προαναφερθέν αρχείο παίρνω ένα μικρό υποσύνολο από στήλες και γραμμές:

Country Name	Country Code	Indicator Name	Indicator Code	1961	...	2001
Greece	GRC	Agricultural machinery, tractors	AG.AGR.TRAC.NO	22630	...	254527
...	...	...	...	...	...	...
Greece	GRC	Agricultural land (sq. km)	AG.LND.AGRI.K2	89100	...	85020
...	...	...	...	...	...	...

Τα προβλήματα που έχουμε είναι:

- Έχουμε πολλές χώρες με τις οποίες θα ασχοληθούμε

- Έχουμε πολλά χρόνια, για τα οποία επιπλέον, σας ζητείται υποχρεωτικά να τα οργανώσετε σε 5ετίες, 10ετίες, 20ετίες
- Έχουμε πολλούς δείκτες που μας απασχολούν (και μάλιστα, με ιεραρχίες)

Υπάρχει η σχεδιαστική δυνατότητα, να διατηρήσετε τη δομή του αρχείου σε ένα πίνακα (still, think: θα έχετε πολλά αρχεία, ένα ανά χώρα). Σίγουρα, η δομή δεν είναι κανονικοποιημένη (γιατί?) και άρα χρειάζεται επεμβάσεις ώστε να είναι σε 3NF. Η σχεδιαστική λύση αυτή έχει πλεονεκτήματα αλλά, ως συνήθως, δεν είναι δωρεάν – κάτι πληρώνουμε και κάτι κερδίζουμε.

Υπάρχουν σχεδιαστικές λύσεις ώστε να έχετε την πληροφορία ανά χώρα και έτος με τη χρήση lookur πινάκων. Προσέξτε πώς οι χώρες αποθηκεύονται σε ένα lookur πίνακα. Πώς πρέπει να αποθηκεύσω την κυρίως ειπείν πληροφορία, τότε? Προσέξτε ότι, αντί για τιμές μόνο (Albania, Algeria ...), ο πίνακας Countries έχει μέσα (i) numeric primary key, (ii) country code (which could act as a candidate key, but we chose to use an artificial key – why?), (iii) το όνομα, φυσικά, καθώς και άλλες πληροφορίες (αυτοσχεδιάζω στα επιπλέον πεδία):

01	Albania	ALB	Europe	Tirana	...
02	Algeria	ALG	Africa	Algiers	...
...					

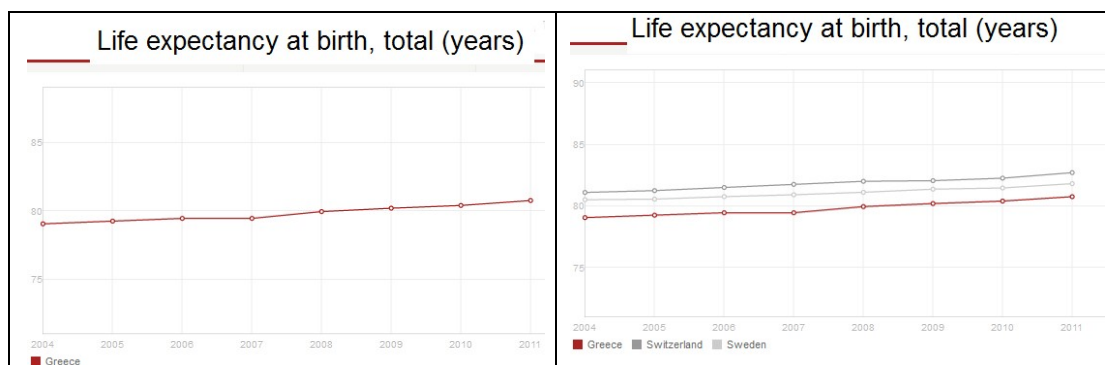
Προσέξτε επίσης πως το αρχείο εισόδου θα αποθηκευθεί πλέον σε ένα (ή πολλούς?) fact πίνακα(ες) με ένα foreign keys σε κάθε lookur πίνακα που το αφορά. Η σχεδιαστική λύση αυτή έχει και αυτή πλεονεκτήματα (ποια?) και, ως συνήθως, δεν είναι δωρεάν (με πρώτο εμφανές κόστος ότι τα δεδομένα θέλουν ευρύτερους μετασχηματισμούς).

*Στην σχεδίαση που θα κάνετε, σκεφτείτε τι θα πράξετε και αιτιολογήστε γιατί και θα τα συζητήσουμε στο μάθημα.*

## ΦΑΣΕΙΣ II και III: υλοποίηση εφαρμογής

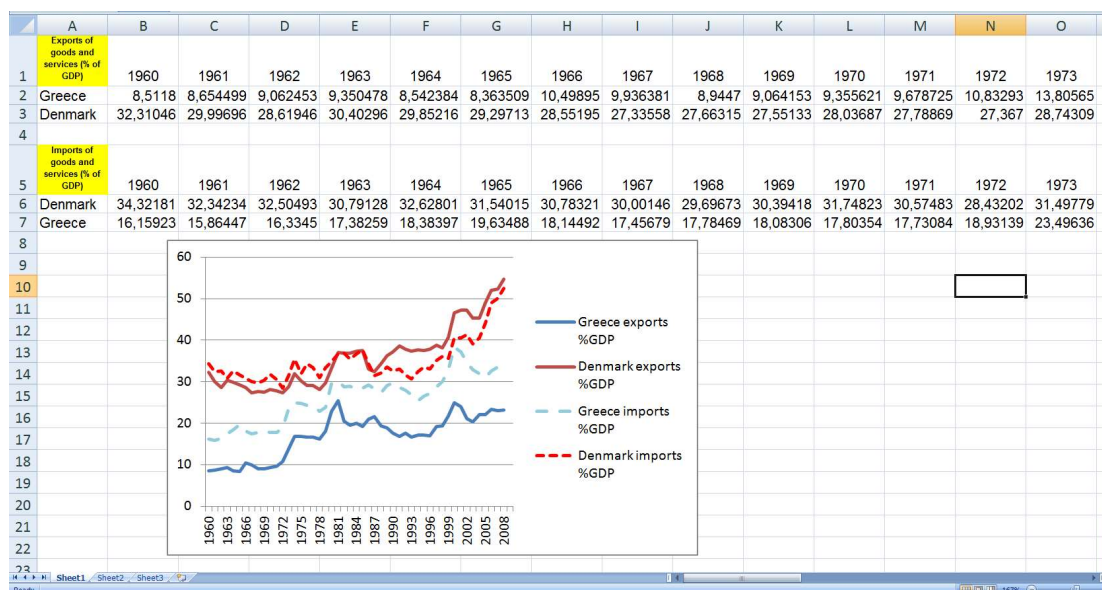
Στη φάση II θα φτιάξετε κάποια γρήγορα prototypes από τις ερωτήσεις και τις οπτικοποιήσεις που απαιτούνται (όπως θα δείτε παρακάτω). Στην φάση III θα αξιοποιήσετε πλήρως τα δεδομένα για την εξαγωγή συμπερασμάτων και θα εμπλουτίσετε την εφαρμογή σας με την πλήρη γκάμα από ερωτήσεις και οπτικοποιήσεις που ζητούνται.

**Timelines / trendlines.** Αν θέλουμε να δείξουμε την εξέλιξη ενός ή περισσότερων δεικτών στο χρόνο, το πιο συχνά χρησιμοποιούμενο μέσο είναι οι timelines. Ο χρόνος απεικονίζεται στον άξονα των x και το μετρούμενο μέγεθος στον άξονα των y. Αν αντί για χρόνο έχουμε άλλο ποσό στον άξονα των x (π.χ., ο πληθυσμός μιας χώρας, η έκτασή της κλπ) τότε εμπίπτουμε στη γενικότερη κατηγορία των trendlines.

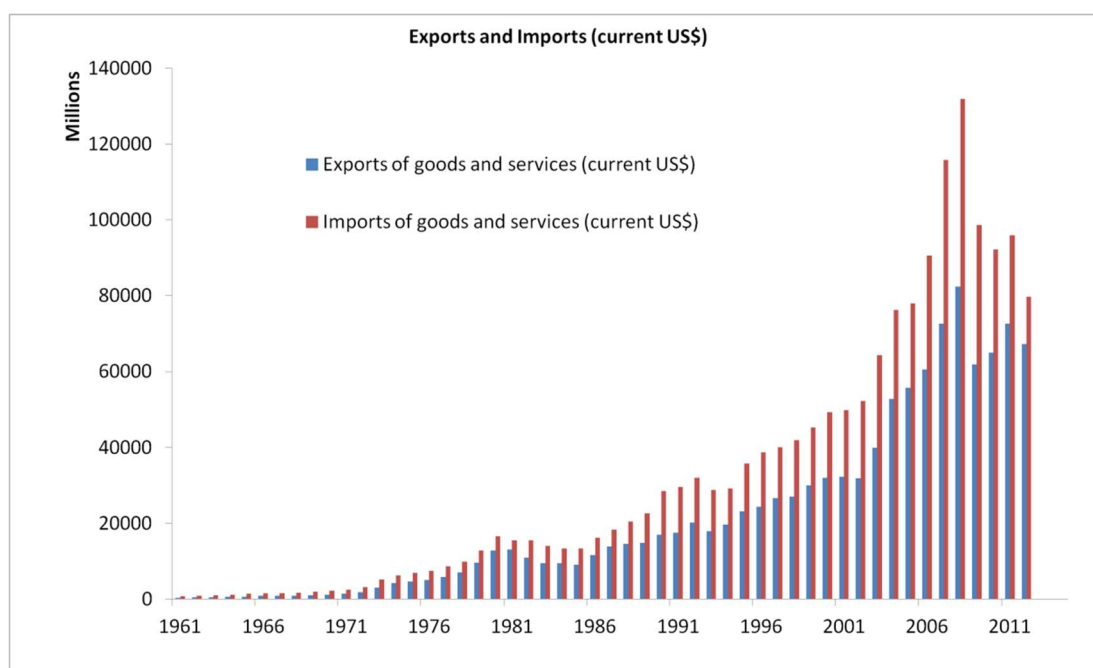


Οι υποκατηγορίες που μπορεί να έχουμε είναι:

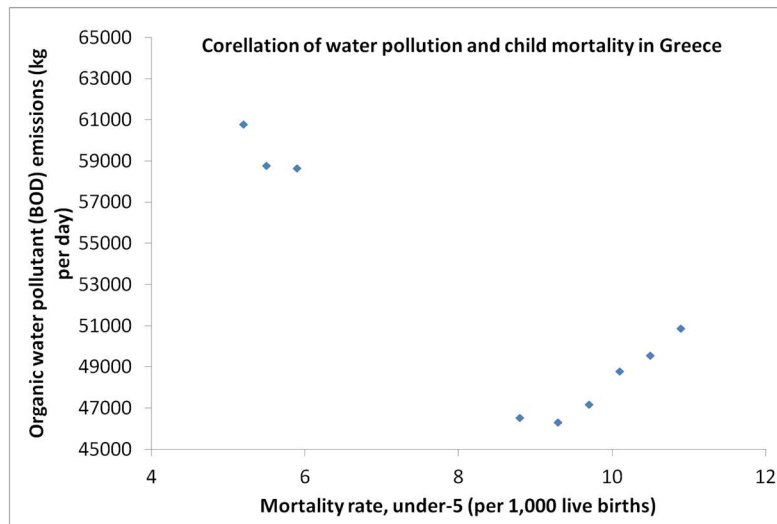
- Για k-το πλήθος χώρες, για ένα δείκτη, δείξτε πώς εξελίσσεται στο χρόνο (απλή περίπτωση: 1 χώρα)
- Για k<sub>c</sub> -το πλήθος χώρες και για k<sub>m</sub> -το πλήθος δείκτες, δείξτε πώς εξελίσσεται στο χρόνο ο καθένας (το παράδειγμα εδώ είναι από άλλο data set, αλλά ενδεικτικό του τι εννοούμε)



**Bar charts.** Η εν λόγω τεχνική χρησιμοποιείται για να συγκρίνει δύο ή περισσότερες μετρικές (y-axis) πάνω στις ίδιες τιμές του άξονα των x. Γενικά, μπορούμε να γενικεύσουμε το παραπάνω σε περισσότερες από 2 μετρικές, k-το πλήθος στη γενική περίπτωση, αλλά με μικρό k (σκεφθείτε πόσο άσχημο θα ήταν το διάγραμμα για παραπάνω των 2 δεικτών και 2 χωρών). Στην περίπτωση μας, μπορούμε πάλι να έχουμε ένα συνδυασμό από χώρες και δείκτες.



**Scatter Plots.** Η εν λόγω τεχνική οπτικοποίησης συσχετίζει περισσότερες της μίας μετρικές και προσπαθεί να δείξει το βαθμό συσχέτισής τους. Για παράδειγμα, αν θέλουμε να δούμε πώς σχετίζεται η παιδική θνησιμότητα με την πρόσβαση σε «υψηλής ποιότητας» νερό για μια συγκεκριμένη χώρα, πρέπει να κάνουμε μια ερώτηση που να επιστρέφει για κάθε έτος το ποσοστό παιδικής θνησιμότητας και το ποσοστό πρόσβασης σε νερό υψηλής ποιότητας (κάθε εγγραφή του αποτελέσματος λέει έτος, παιδ. θνησ., μόλυνση ύδατος). Η συσχέτιση προκύπτει βάζοντας τις τιμές για τον ένα δείκτη στον ένα άξονα και τις τιμές για τον άλλο δείκτη στον άλλο άξονα.



## Οδηγίες προς ναυτιλλομένους

Στο τέλος της εργασίας, θέλουμε ο χρήστης να μπορεί να επιλέξει (α) χώρες, (β) δείκτες και (γ) χρονικό εύρος και να απεικονίζεται το αποτέλεσμα είτε ανά χρόνο, είτε ανά πενταετία, **κοκ**. Κατασκευάστε αρχικά από ένα τέτοιο γράφημα ανά περίπτωση (ξεκινήστε από τα πιο απλά), με fixed query πίσω του, και δοκιμάστε να οπτικοποιήσετε το αποτέλεσμα. Μετά, ΠΡΟΟΔΕΥΤΙΚΑ, προσθέστε τη δυνατότητα επιλογών για τα (α)-(γ), ώστε η ερώτηση να κατασκευάζεται δυναμικά.

Από τις πολύ συχνά χρησιμοποιούμενες **βιβλιοθήκες οπτικοποίησης** είναι οι d3.js (javascript) for web development και οι JavaFX (built-in Java) ή jfreechart (Java library) για Java. **Για φέτος θα χρησιμοποιήσουμε d3** (<http://d3js.org/>) (αξίζει να αφιερώσετε ώρα να περιηγηθείτε στα παραδείγματα του <https://github.com/mbostock/d3/wiki/Gallery> και του <http://bl.ocks.org/mbostock> τα οποία έχουν, όλα, και τον κώδικά τους μαζί).

Στη **φάση II**, θα χρειαστεί:

1. Στήσιμο του προγραμματιστικού περιβάλλοντος στο οποίο θα γίνει η ανάπτυξη
2. Στήσιμο του περιβάλλοντος στο οποίο θα στηθεί και θα τρέξει η εφαρμογή σας (ενδεχομένως το ίδιο).
3. Πειραματισμός με έτοιμα παραδείγματα από την τεκμηρίωση των τεχνολογιών που θα χρησιμοποιήσετε: φτιάξτε μικρά προγραμματάκια που να τρέχουν
4. Κατασκευή του πρώτου script που προσπελάζει τη βάση δεδομένων και (α) συνδέεται, (β) υποβάλει μια ερώτηση, (γ) διαχειρίζεται το αποτέλεσμα της
5. Κατασκευή του πρώτου script που οπτικοποιεί δεδομένα (όχι απαραίτητα αποτελέσματα ερωτήσεων σε βάση) με τον επιθυμητό τρόπο.
6. Προοδευτική σύνδεση των παραπάνω

Στη **φάση III**, θα πρέπει να προσθέσετε και ένα βαθμό διαδραστικότητας στο παραπάνω. Προσθέστε μενού επιλογής (ή άλλους τρόπους επιλογής) και χρησιμοποιήστε γραφικούς τρόπους αλληλεπίδρασης (π.χ., φόρμες και drop-down listboxes) ώστε να πάρετε από το χρήστη τι ακριβώς επιθυμεί να δει. Συνδέστε το κομμάτι αυτό με ερωτήσεις και οπτικοποιήσεις.

Μπορείτε να έχετε έτοιμες κάποιες **προκατασκευασμένες βοηθητικές όψεις** (views) ή να χρησιμοποιείτε **προσωρινές όψεις ανάλογα με το ερώτημα** (CREATE VIEW ... -- SELECT ... -- DROP VIEW ...) ώστε να κάνετε την προγραμματιστική δουλειά πιο εύκολη.

Μπορείτε να **αναλύσετε τις ερωτήσεις σας** και αν διαπιστώσετε ότι μπορεί να **παραμετροποιήσετε την κατασκευή τους** (π.χ., ανάλογα με το τι δίνει ο χρήστης, να προσαρμόζονται τα πεδία του SELECT / GROUP-BY / ...) και να φτιάξετε μεθόδους/συναρτήσεις που κατασκευάζουν την ερώτηση στη βάση με βάση τις παραμέτρους αυτές (με προφανές όφελος: write once, test a few times, safely use for ever)

## Χρονοδιάγραμμα

Στη συνέχεια παρατίθενται στάδια της ανάπτυξης, ενδιάμεσες προθεσμίες (milestones) και καταληκτικές ημερομηνίες ολοκλήρωσης (deadlines).

<b>[13/02]</b>	Εκφώνηση
<i>Κάντε την Φάση I και ότι μπορείτε από II</i>	Εκτέλεση των βημάτων της ΦΑΣΗΣ I Παραδοτέα: P1.1: Exported Σχήμα + workbench screenshot P1.2: Φορτωμένη βάση για τα backbone data backup P1.3: scripts + 1-page diagram for the transformation process
<b>[12/03]</b>	<b>Milestone: ΟΛΟΚΛΗΡΩΣΗ ΦΑΣΗΣ I</b>
<i>Λύστε το πώς θα δουλέψετε νωρίς</i>	Μία αρχική οπτικοποίηση με fixed queries ανά κατηγορία Στήσιμο προγραμματιστικού περιβάλλοντος/framework και κατανόησή τους <i>60% του χρόνου να τρέξει η πρώτη αναφορά, 20% του χρόνου να τρέξει η δεύτερη, the rest of the time for the rest</i>
<b>[26/03]</b>	<b>Milestone: Ενδιάμεσα στη φάση II</b>
<i>Αν έχετε μία αναφορά, οι άλλες είναι εύκολες</i>	Τουλάχιστον μία οπτικοποίηση με dynamically constructed queries Περιβάλλον αλληλεπίδρασης για τις επιλογές του χρήστη Παραδοτέα: P2.1: Application code containing the above P2.2. 1η εκδοχή της αναφοράς με την τρέχουσα εκδοχή των 3 πρώτων ενοτήτων
<b>[09/04]</b>	<b>Hard Deadline: ΟΛΟΚΛΗΡΩΣΗ ΦΑΣΗΣ II</b>
	Πλήρης υλοποίηση της εφαρμογής Παραδοτέα: <b>P3.1:</b> το <b>σύστημα</b> εν λειτουργία <b>P3.2:</b> <b>τελική αναφορά</b> εκτυπωμένη (όπως στο σχετικό πρότυπο που βρίσκεται αναρτημένο στο δικτυακό τόπο του μαθήματος) <b>P3.3:</b> <b>DVD</b> με τον κώδικα, τα scripts, τα δεδομένα (input, output, backups) και την τελική αναφορά
<b>[21/05]</b>	<b>Hard Deadline: ΟΛΟΚΛΗΡΩΣΗ ΦΑΣΗΣ III + ΕΠΙΔΕΙΞΗ @ 2020/05/28</b>

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!!