

Εργαστήριο Σημασιολογικού Ιστού

Ενότητα 2: Εισαγωγή στην Οργάνωση των Σημασιολογικών Δεδομένων

Μ.Στεφανιδάκης

22-2-2018

Ποιο το κατάλληλο μοντέλο δεδομένων;

- ▶ Τα σημασιολογικά δεδομένα πρέπει να εκτεθούν “ώς έχουν” (raw)
 - ▶ Ποιο είναι το κατάλληλο μοντέλο οργάνωσης;
 - ▶ Και πώς θα εμπεριέχονται και τα μεταδεδομένα;
- ▶ Ας ξεκινήσουμε με ένα απλό μοντέλο: δεδομένα σε μορφή πίνακα (**tabular data**)
 - ▶ Η σημασιολογία των δεδομένων προκύπτει **έμμεσα από τη θέση τους** (γραμμή-στήλη)

Τύπος	Αριθμός
Λεωφορεία	58.519
ΙΧ	2.062.880
Οδοποιητικά	7.642
Εκχιονιστικά	6
Ποδήλατα	8.540.192

Η έμμεση σημασιολογία του πίνακα

- ▶ Κάθε γραμμή αντιστοιχεί σε μία βασική **οντότητα** (entity) δεδομένων
 - ▶ Έμμεση πληροφορία
- ▶ Κάθε στήλη αντιστοιχεί σε μια **ιδιότητα** (attribute)
 - ▶ Ρητή πληροφορία, αν υπάρχει περιγραφή στηλών

Οντότητα _i	→	Τύπος	Αριθμός	←	Ιδιότητα _j
		Λεωφορεία	58.519		
		ΙΧ	2.062.880		
		Οδοποιητικά	7.642		
		Εκχιονιστικά	6		
		Ποδήλατα	8.540.192		

Δοκιμάστε και εσείς!

- ▶ Διαλέξτε ένα ωρολόγιο πρόγραμμα στο τρέχον εξάμηνο
 - ▶ π.χ. του Η' εξαμήνου
- ▶ Προσπαθήστε να εκφράσετε την περιεχόμενη πληροφορία σε μορφή πίνακα
 - ▶ Για βοήθεια, σκεφτείτε πώς θα οργανώνατε την πληροφορία στο **σχεσιακό** μοντέλο
 - ▶ Θα καταλήξετε χονδρικά στους εξής σχεσιακούς πίνακες: Καθηγητής, Μάθημα, Αίθουσα, Εξάμηνο και Διάλεξη
 - ▶ Η Διάλεξη (ένα γεμάτο κελί του αρχικού ωρολογίου pdf) είναι η **κομβική οντότητα** που συνδέεται με όλες τις άλλες
 - ▶ Διαθέτει επίσης πληροφορία ώρας και ημέρας

Δοκιμάστε και εσείς!

- ▶ Χρησιμοποιήστε τη **Διάλεξη** ως **κομβική οντότητα πληροφορίας** του ωρολογίου προγράμματος
 - ▶ Κάθε γραμμή του πίνακα θα είναι εκφράζει μια τέτοια οντότητα (διάλεξη)
 - ▶ Και η υπόλοιπη πληροφορία θα τοποθετηθεί στις στήλες, ως ιδιότητες κάθε διάλεξης
- ▶ **Προσοχή!** το ζητούμενο **δεν είναι** να αναπαράγετε τον πίνακα του ωρολογίου προγράμματος ως έχει!

Μοντέλο και Μορφή Αποθήκευσης

- ▶ Η μορφή πίνακα είναι ένα μοντέλο οργάνωσης δεδομένων (**data model**)
 - ▶ Προσδιορίζει τον τρόπο δόμησης της πληροφορίας
- ▶ Η δομημένη πληροφορία όμως
 - ▶ Πρέπει να **αποθηκευτεί** ως ακολουθία bytes
 - ▶ Και να **μεταδοθεί** μεταξύ παραγωγού και καταναλωτή
- ▶ Συνεπώς, πέρα από το μοντέλο των δεδομένων, είναι απαραίτητο ένα μορφότυπο σειριοποίησης (**serialization format**) των δεδομένων

Η μορφή αποθήκευσης CSV

- ▶ Comma Separated Values
 - ▶ Ένα ..μη πρότυπο πρότυπο (τουλάχιστον μέχρι πρόσφατα)
 - ▶ Με πολλές “διαλέκτους” (σχεδόν κάθε εφαρμογή έχει τη δική της!)
 - ▶ Για εξαγωγή-εισαγωγή μεγάλων σετ δεδομένων σε μορφή πίνακα από-σε βάσεις δεδομένων
- ▶ Έλεγχος Ιδιοτήτων:
 - ▶ Ανοικτό πρότυπο: **NAI**
 - ▶ Χρήση στο Web: **NAI** (αν και όχι τόσο συχνά)
 - ▶ Ευκολία προγραμματισμού: **NAI** (βιβλιοθήκες για πολλές γλώσσες)
 - ▶ Ρητά μεταδεδομένα: **OXI** (προαιρετικά, ονόματα στηλών στην πρώτη γραμμή)

Δοκιμάστε και εσείς!

- ▶ **Βήμα 1^ο:** Αποθηκεύστε τον πίνακα που φτιάξατε προηγουμένως σε μορφή csv
 - ▶ μέσω της εφαρμογής spreadsheet
- ▶ **Βήμα 2^ο:** Γράψτε πρόγραμμα Python που διαβάζει το αρχείο csv και μπορεί να απαντήσει σε ερωτήματα όπως:
 - ▶ Τι διδάσκεται την ώρα/μέρα X στην αίθουσα Y;
 - ▶ Ποιες μέρες διδάσκει ο X το μάθημα Y;
 - ▶ Ποιος διδάσκει την ημέρα X στην αίθουσα Y;

Στην επόμενη σελίδα: ανάγνωση CSV μέσω Python

Python και ανάγνωση αρχείου CSV

```
import csv

# open csv file for reading
ifp = open('test.csv','r',newline='',encoding='utf-8')
# create csv reader object
ir = csv.reader(ifp) # defaults to excel 'dialect'

# read first row (headers)
hdrow = next(ir)

# iterate over table rows in csv file
for row in ir:
    # each row is a list of strings
    # (table column values for this row)

    # do something with each row here...

ifp.close()
```

Προσθήκη ρητού αναγνωριστικού (id)

- ▶ Οικείο σχήμα από τις σχεσιακές βάσεις...
- ▶ Κάθε βασική οντότητα διαθέτει μοναδικό αναγνωριστικό
 - ▶ Θεωρήστε προς το παρόν ότι αυτός είναι ο μοναδικός πίνακας στον κόσμο!

id	Τύπος	Αριθμός
1	Λεωφορεία	58.519
2	ΙΧ	2.062.880
3	Οδοποιητικά	7.642
4	Εκχιονιστικά	6
5	Ποδήλατα	8.540.192

Το μετα-μοντέλο ΕΑV

- ▶ Μια νέα μεταμόρφωση: το μοντέλο **Entity-Attribute-Value**
 - ▶ Μορφή **τριάδας** (triple): (**Οντότητα**, **Ιδιότητα**, **Τιμή**)
 - ▶ Η Οντότητα συμβολίζεται με το ρητό αναγνωριστικό της
 - ▶ Ως Ιδιότητες μπαίνουν οι (ρητές ή εννοούμενες) επικεφαλίδες των στηλών
 - ▶ Ως Τιμές χρησιμοποιούνται τα περιεχόμενα των κελιών στις διασταυρώσεις γραμμών-στηλών
 - ▶ Όλα τα μεταδεδομένα (ιδιότητες) δηλώνονται ρητά
 - ▶ Ο “εφιάλτης” του σχεσιακού μοντέλου!!!
 - ▶ Πλήρης απο-κανονικοποίηση (denormalization)

Παράδειγμα μετασχηματισμού

- ▶ Έστω ο πίνακας (δείχνεται μια γραμμή μόνο)

id	Engine	Weight(kg)	Tracks(mm)	Boom(m)
...
38rb	V8	60000	850	18
...

Δεδομένα κατά το μοντέλο EAV

- ▶ Η μία αυτή γραμμή παράγει από μόνη της τις εξής τριάδες

Entity	Attribute	Value
38rb	Engine	V8
38rb	Weight(kg)	60000
38rb	Tracks(mm)	850
38rb	Boom(m)	18

- ▶ Κάθε άλλη γραμμή του αρχικού πίνακα θα μετασχηματιστεί επίσης στις αντίστοιχες τριάδες!

Δοκιμάστε και εσείς!

- ▶ Φτιάξτε πρόγραμμα Python
 - ▶ Διαβάστε το csv αρχείο σας με το ωρολόγιο πρόγραμμα
 - ▶ Για κάθε μία γραμμή αποθηκεύστε σε ένα νέο csv τις τριάδες που παράγονται από τη γραμμή αυτή
 - ▶ Σύμφωνα με το μοντέλο EAV

```
import csv
```

```
# open file for csv writing  
ofp = open('out.csv','w',newline='',encoding='utf-8')  
# create csv writer (default format)  
ow = csv.writer(ofp)  
  
# write a row of values in csv file  
ow.writerow([ent,attr,val])  
  
# close output file  
ofp.close()
```