ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Δ.Π.Μ.Σ. ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

Εξόρυξη Γνώσης από Δεδομένα

ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ 2

[redacted]

Ιανουάριος 2023

Άσκηση 1

Και για τις δύο μεθόδους smoothing χρειάζεται να ταξινομηθούν τα δεδομένα (21, 25, 16, 25, 33, 19, 45, 25, 22, 35, 52, 36, 70, 20, 35, 22, 35, 25, 15, 35, 20, 30, 33, 13, 40, 46, 16) σε αύξουσα σειρά ως εξής:

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

Στην συνέχεια χωρίζονται σε 9 partitions για να έχουν μήκος 3 όπως αναφέρεται στην εκφώνηση.

Partition 1: 13, 15, 16

Partition 2: 16, 19, 20

Partition 3: 20, 21, 22

Partition 4: 22, 25, 25

Partition 5: 25, 25, 30

Partition 6: 33, 33, 35

Partition 7: 35, 35, 35

Partition 8: 36, 40, 45

Partition 9: 46, 52, 70

Ερώτημα a

Για την μέθοδο bin means υπολογίζουμε τον μέσο όρο κάθε partition και αντικαθιστούμε τις τιμές με αυτόν. Το αποτέλεσμα θα είναι:

Bin 1: 14, 14, 14

Bin 2: 18, 18, 18

Bin 3: 21, 21, 21

Bin 4: 24, 24, 24

Bin 5: 26, 26, 26

Bin 6: 33, 33, 33

Bin 7: 35, 35, 35

Bin 8: 40, 40, 40

Bin 9: 56, 56, 56

Σημείωση: Ο μέρος όρος κάποιων partitions είναι δεκαδικοί αλλά επειδή οι αριθμοί αυτοί αντιστοιχούν σε ηλικία, στρογγυλοποιήθηκαν προς τον μικρότερο ακέραιο.

Ερώτημα b

Για την μέθοδο bin boundaries αντικαθιστούμε τις τιμές της κοντινότερης τιμής ορίου. Το αποτέλεσμα θα είναι:

Bin 1: 13, 16, 16

Bin 2: 16, 20, 20

Bin 3: 20, 20, 22

Bin 4: 22, 25, 25

Bin 5: 25, 25, 30

Bin 6: 33, 33, 35

Bin 7: 35, 35, 35

Bin 8: 36, 36, 45

Bin 9: 46, 46, 70

Άσκηση 2

Ερώτημα a

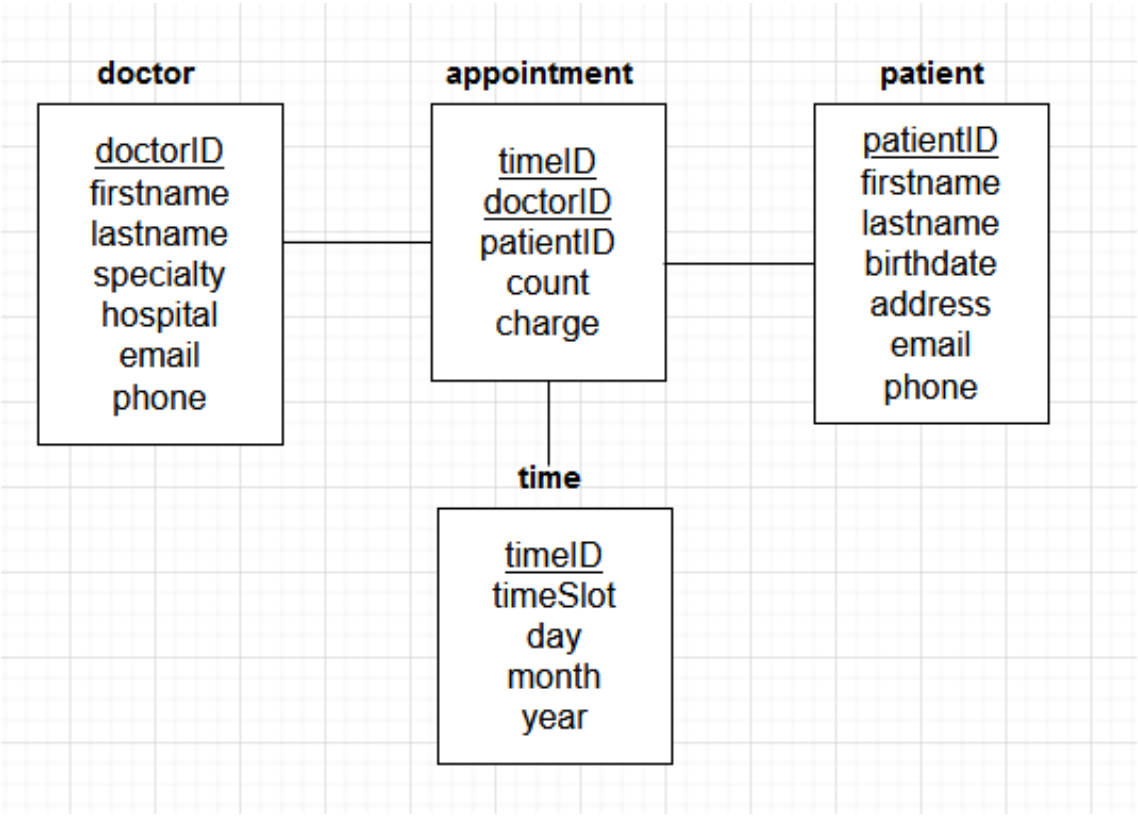
Στο σχήμα της Εικόνας 1 απεικονίζεται το star schema του data warehouse. Ο πίνακας appointment αποτελεί το fact table ενώ οι πίνακες doctor, patient και time είναι οι dimension tables. Τα υπογραμμισμένα πεδία αποτελούν πρωτεύοντα κλειδιά. Οι παραδοχές που έγιναν ήταν:

1. Ο πίνακας doctor έχει ως πεδία το όνομα του γιατρού (firstname, lastname), την ειδικότητα του (specialty), το νοσοκομείο που εργάζεται (hospital), το email και το τηλέφωνο του (phone). Το πρωτεύων κλειδί είναι το doctorID.

2. Ο πίνακας patient έχει ως πεδία το όνομα του ασθενή (firstname, lastname), την ημερομηνία γέννησης (birthdate), την διεύθυνση του (address), το email και το τηλέφωνο του (phone). Το πρωτεύων κλειδί είναι το patientID που θα μπορούσε να είναι το ΑΜΚΑ.

3. Ο πίνακας time έχει ως πεδία ένα time slot όπου δηλώνεται η ώρα έναρξης και ολοκλήρωσης ενός ραντεβού (timeSlot), η ημέρα (day), ο μήνας (month) και ο χρόνος (year) του ραντεβού. Το πρωτεύων κλειδί είναι το timeID.

4. Ο πίνακας appointment έχει ως πεδία το ID της ημερομηνίας και ώρας του ραντεβού (timeID), το ID του γιατρού (doctorID) και του ασθενή (patientID), έναν μετρητή επίσκεψης του ασθενή στον συγκεκριμένο ιατρό μια συγκεκριμένη μέρα (count) και την συνολική χρέωση του ασθενή ανά ημέρα (charge) έχοντας λάβει υπόψη και τον αριθμό των επισκέψεων. Το πρωτεύων κλειδί είναι ο συνδιασμός timeID και doctorID καθώς δεν γίνεται ένας γιατρός να έχει δύο ραντεβού την ίδια ώρα και έτσι προσδιορίζονται μοναδικά οι εγγραφές του πίνακα αυτού.



Εικόνα 1: Star schema

Ερώτημα b

Ξεκινώντας από το [day, doctor, patient] για την μετρική charge θα μπορούσαν να εκτελεστούν

οι παρακάτω διαδικασίες ώστε να ανακτηθεί το συνολικό εισόδημα κάθε γιατρού για το έτος 2021:

1. Roll-up on time (from timeID to year): Από το κάθε timeID μεταφερόμαστε στις χρεώσεις (charges) ανά χρόνο για κάθε ασθενή και γιατρό.

2. Slice for time=2021: Επιλογή συγκεκριμένης χρονιάς. Εξαφανίζεται η διάσταση του χρόνου (time) και μένουν οι χρεώσεις κάθε ασθενή και γιατρού για το έτος 2021.

3. Roll-up on patient: Δεν μας ενδιαφέρει η χρέωση ανά ασθενή πλέον αλλά ανά γιατρό. Οπότε εξαφανίζεται και η διάσταση του patient και μένει μόνο η χρέωση κάθε γιατρού για όλους τους ασθενείς το έτος 2021 που είναι και το ζητούμενο.

Δεν χρειάζεται να ελέγξουμε το [day, doctor, patient] για την μετρική count, αφού έχουμε υποθέσει ότι η μετρική charge έμπεριέχει την συνολική χρέωση του ασθενή ανά μέρα με βάση τις επισκέψεις.

Ερώτημα c

Υποθέτουμε ότι τα δεδομένα ήταν αποθηκευμένα σε μια σχεσιακή βάση δεδομένων με το σχήμα warehouse (day, month, year, doctor, hospital, patient, count, charge). Τότε, η ανάκτηση του συνολικού εισοδήματος κάθε γιατρού για το έτος 2021 θα γινόταν όπως φαίνεται παρακάτω, με δεδομένες τις υποθέσεις που έγιναν και στα προηγούμενα ερωτήματα.

```
SELECT doctor, SUM(charge)
FROM warehouse
WHERE year = 2021
GROUP BY doctor
```