

---

*Προηγμένα Πληροφοριακά Συστήματα*

*Ακαδημαϊκό Έτος 2016-2017*

---



ΠΑΝΕΠΙΣΤΗΜΙΟ  
ΠΑΤΡΩΝ  
UNIVERSITY OF PATRAS

**Ομάδα:**

- |                             |          |
|-----------------------------|----------|
| 1. Κανούτος Κωνσταντίνος    | ΑΜ: 5775 |
| 2. Καραχάλιος Αθανάσιος     | ΑΜ: 5784 |
| 3. Κυριακού Ανδρόνικος      | ΑΜ: 5806 |
| 4. Ντενέζος Παναγιώτης      | ΑΜ: 5853 |
| 5. Παρασκευόπουλος Γεώργιος | ΑΜ: 5874 |
| 6. Πλούμης Θωμάς            | ΑΜ: 5880 |

## A. Σχεδιασμός Data Warehouse

Αρχικά, αποφασίσαμε να παρακολουθήσουμε τα δεδομένα, σχετικά με τα ατυχήματα τις χρονιές 2009-2014. Για το λόγο αυτό, επιλέξαμε από τα αρχεία csv Accidents (DfTRoadSafety\_Accidents) για τις χρονιές που αναφέρθηκαν, ένα σύνολο στηλών τις οποίες και ομαδοποιήσαμε με βάση κοινά τους γνωρίσματα για να ορίσουμε διαστάσεις και να δημιουργήσουμε τους ακόλουθους πίνακες:

- Conditions
  - Light Conditions
  - Weather Conditions
  - Road Surface Conditions
- Date
  - Date
  - Day of Week
  - Time
- Location
  - 1st Road Class
  - 2nd Road Class
- Severity
  - Accident Severity
  - Number of Vehicles
  - Number of Casualties

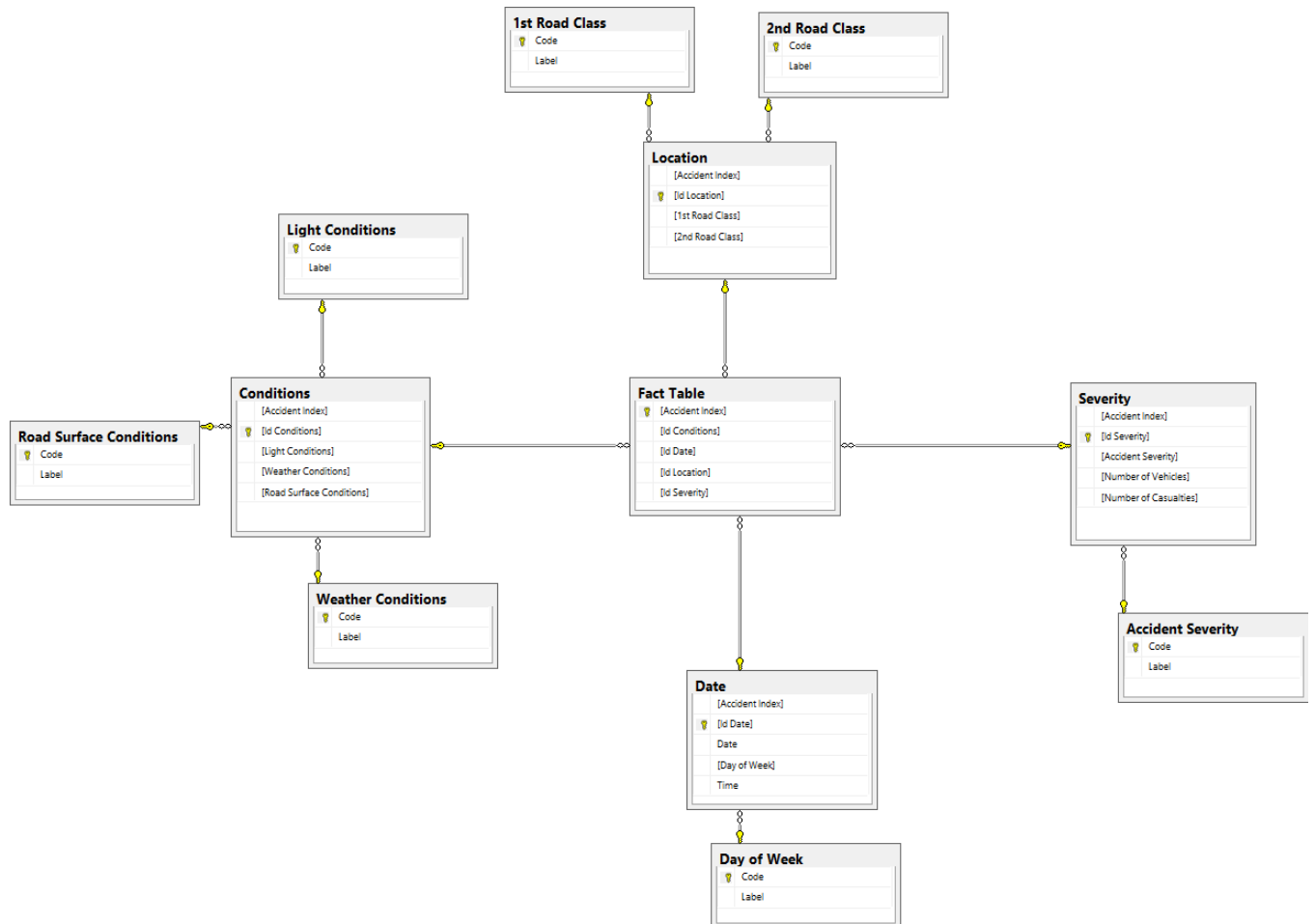
Σε καθέναν από τους παραπάνω πίνακες προσθέσαμε στο SQL Server μία στήλη με το Accident Index του ατυχήματος, καθώς και άλλη μια με ένα μοναδικό ID (Id Conditions, Id Date, Id Location και Id Severity αντίστοιχα) η οποία, ορισμένη ως auto increment, λειτουργεί ως αύξοντας αριθμός και αποτελεί το primary key για τους συγκεκριμένους πίνακες.

Επιπλέον, δημιουργήθηκε ένας ακόμα πίνακας, που χρησιμοποιείται ως Fact Table (πίνακας συμβάντων) και περιέχει σαν Primary Key το Accident Index και άλλες 4 στήλες με τα μοναδικά κλειδιά των παραπάνω 4 πινάκων (δηλωμένα ως ξένα κλειδιά). Μέσω αυτών των κλειδιών (στηλών) γίνεται η σύνδεσή του πίνακα συμβάντων με τους πίνακες διαστάσεων.

Τέλος, υλοποιήσαμε άλλους 7 πίνακες, οι οποίοι είναι επεξηγηματικοί των 4 Dimension Tables, δηλαδή αντιστοιχίζεται η αριθμητική τιμή της κάθε στήλης με την πραγματική της έννοια – επεξήγηση (Light Conditions, Weather Conditions, Road Surface Conditions επεξηγηματικοί του

Conditions - Day of Week του Date - 1st Road Class, 2nd Road Class του Location και Accident Severity του Severity).

Το αντίστοιχο διάγραμμα (Database Diagram) που δημιουργήθηκε στον SQL Server, παρουσιάζεται ακολούθως:

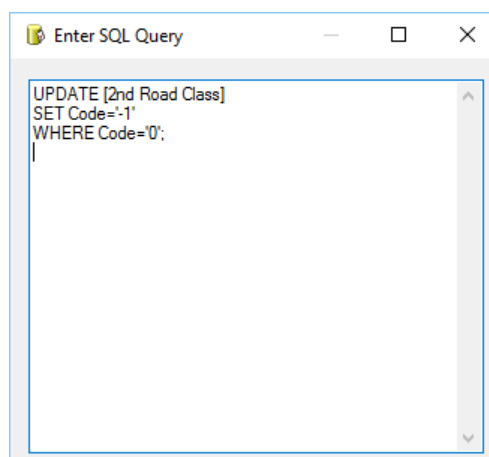


Εικόνα 1: Σχεδιάγραμμα Βάσης Δεδομένων που αναπτύχθηκε

## B. Διαδικασία ETL (Extract, Transform, Load)

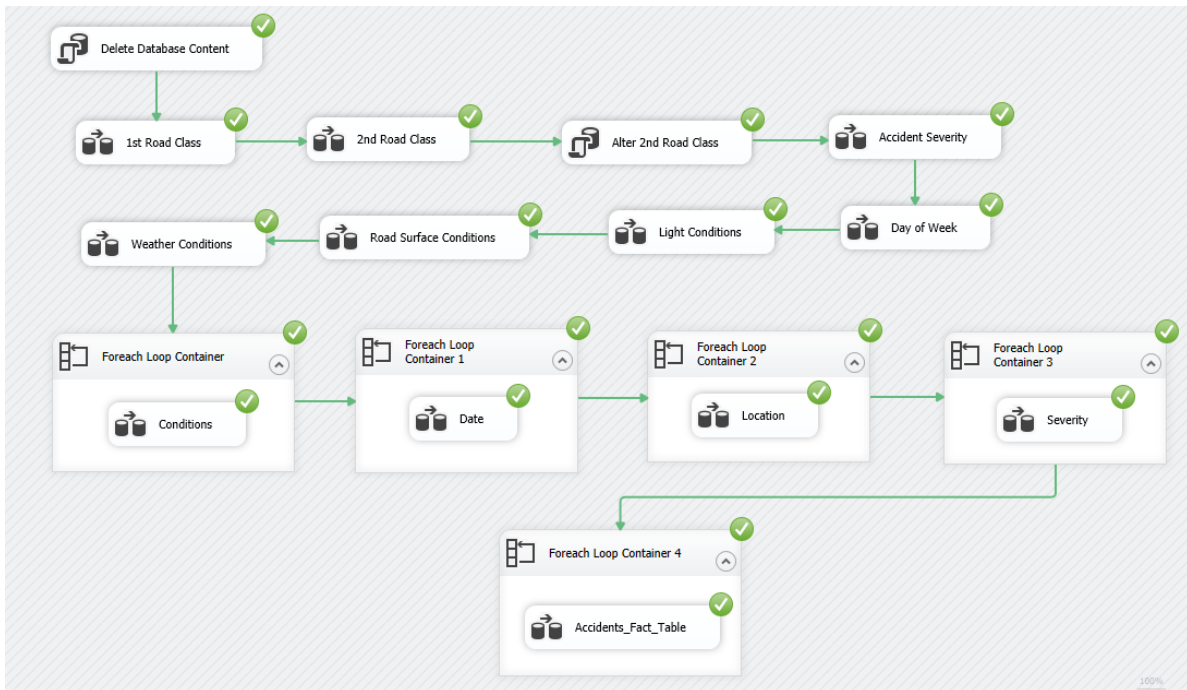
Όπως προαναφέρθηκε, αποφασίσαμε να ασχοληθούμε με δεδομένα από τα αρχεία που περιέγραφαν ατυχήματα και προκειμένου να επιτελέσουμε την ETL διαδικασία δημιουργήσαμε ένα Integration Services Project από το πακέτο SQL Server Data Tools του Microsoft Visual Studio.

Αρχικά, δημιουργήσαμε ένα Connection Manager, ο οποίος συνέδεσε τον server που φιλοξενούσε τη βάση μας με το project του Visual Studio. Έπειτα, εφόσον επιλέξαμε να εισάγουμε δεδομένα από τα αρχεία Accidents έπρεπε να δημιουργήσουμε κάποια επαναληπτική δομή η οποία να μας εισάγει δεδομένα από όλα τα αρχεία για τις χρονολογίες που επιλέξαμε (2009-2014). Πριν γίνει αυτό, όμως, και δεδομένου του σχήματος Data Warehouse (χιονονιφάδα) έπρεπε να εισάγουμε δεδομένα πρώτα στους πίνακες οι οποίοι περιείχαν τις επεξηγήσεις και ύστερα στους πίνακες διαστάσεων (δεδομένης της σχέσης primary keys με foreign keys). Κατά την εισαγωγή των δεδομένων στον πίνακα Location παρατηρήσαμε ότι υπήρξε μια παραβίαση των δυνατών τιμών που μπορούσε να πάρει το πεδίο. Αφού αναλύσαμε τόσο τα δεδομένα των ατυχημάτων όσο και τα δεδομένα του ευρετηρίου (αρχείο csv Road-Accident-Safety-Data-Guide), εντοπίσαμε ότι στα μεν δεδομένα η επεξήγηση «Not at junction or within 20 metres» της στήλης 2nd Road αντιστοιχίζεται με -1 ενώ στο ευρετήριο αντιστοιχίζεται με 0. Έτσι, καθώς οι πίνακες επεξηγήσεων γεμίζουν από το ευρετήριο και οι πίνακες διαστάσεων γεμίζουν από τα πραγματικά δεδομένα, υπήρχε σύγκρουση τιμών, η οποία επιλύθηκε με την αλλαγή των πιθανών τιμών που δέχεται ο πίνακας 2nd Road (Alter 2nd Road Class Task). Παρατίθεται ο μικρού μεγέθους κώδικας που υλοποιήθηκε ως SQL Execution Task στα πλαίσια του data flow για να γίνει η παραπάνω μετατροπή.



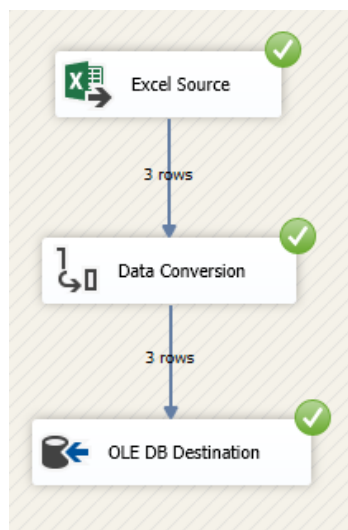
```
UPDATE [2nd Road Class]
SET Code=-1
WHERE Code=0;
```

Όσον αφορά τα tasks που παρουσιάζονται στην Εικόνα 2, αποτελούν ένα σύνολο από data flows τα οποία σκοπό έχουν να φορτώσουν δεδομένα στο data warehouse μας.



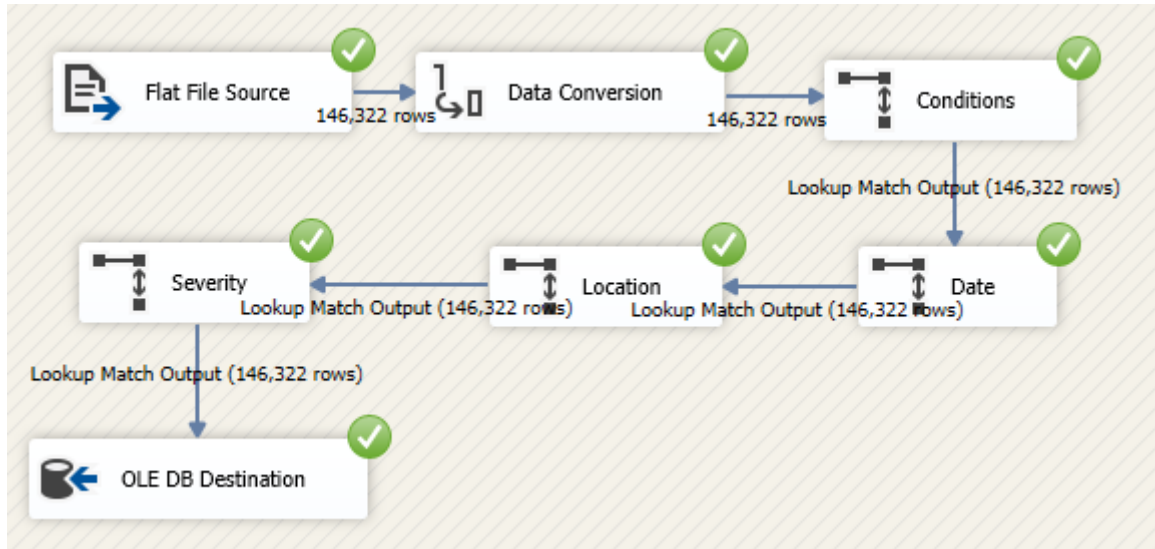
Εικόνα 2: Data Flow

Πιο συγκεκριμένα, επιλέχθηκαν για τους περιγραφικούς πίνακες οι σελίδες από το ευρετήριο (αρχείο csv Road-Accident-Safety-Data-Guide) και αφού έγιναν convert οι τύποι δεδομένων για να είναι συμβατοί με αυτούς που είχαν δηλωθεί στην βάση, φορτώθηκαν σε αυτή (εικόνα 3). Για τα υπόλοιπα δεδομένα που ουσιαστικά γεμίζουν τους πίνακες διατάσεων επιλέχθηκαν οι κατάλληλες στήλες από τα csv αρχεία και με την χρήση ενός foreach loop container, το οποίο παραμετροποιήθηκε κατάλληλα, φορτώθηκαν τα αντίστοιχα δεδομένα.



Εικόνα 3: Πρότυπο Data Flow Task (Accident Severity)

Τέλος, όσον αφορά τον fact table, όπως φαίνεται και στην εικόνα 4, φορτώθηκαν όλα τα πιθανά accident id's και με χρήση lookup tasks (ουσιαστικά join με βάση το accident id) αναζητήθηκαν τα keys των πινάκων διαστάσεων τα οποία και περάστηκαν στον τελικό πίνακα και αυτά επαναληπτικά με την σειρά τους.



Εικόνα 4: Data Flow Task Accidents\_Fact\_Table

## C. Δημιουργία Υπερκύβου OLAP

Για την δημιουργία του υπερκύβου δημιουργήθηκε ένα Analysis Services Project και αποκαταστάθηκε μια σύνδεση με τον SQL Analysis Server στον οποίο και αποθηκεύτηκε ο κύβος.

Πιο συγκεκριμένα, αφού δημιουργήθηκε ένα Data Source View, αναπτύχθηκαν οι διαστάσεις οι οποίες είχαν επιλεγεί (dimension), αλλά και κάποιες επιπλέον έτσι ώστε να μπορέσουμε να αναλύσουμε τα αποτελέσματα στην τελική οπτικοποίηση με βάση τα ονόματα των χαρακτηριστικών που αναλύουμε (label) και όχι τους κωδικούς (code) που έχουμε χρησιμοποιήσει για την εσωτερική ανάπτυξη του Data Warehouse.

Τέλος, δημιουργήθηκε ο υπερκύβος και έγινε deploy στον Analysis Server όπου και επεξεργάστηκε για να είναι σε θέση να δεχθεί τα ερωτήματα που παρουσιάζονται παρακάτω.

Ακολούθως απαντάται το ερώτημα: «πόσα ατυχήματα έχουν γίνει κατά τη διάρκεια της βετίας που έχουμε επιλέξει την ημέρα Σάββατο με κάτω από 10 θύματα και ενώ δεν υπάρχει φωτισμός» που υλοποιήθηκε στην καρτέλα Browse του κύβου μέσα από το εν λόγω project (Η απάντηση στο ερώτημα αυτό είναι 7203 ατυχήματα).

Dimension	Hierarchy	Operator	Filter Expression	Parameter
Date	Label	Equal	{ Saturday }	<input type="checkbox"/>
Severity	Number Of Casualties	Range (Inclusive)	1 : 10	<input type="checkbox"/> <input type="checkbox"/>
Conditions	Lightning Label	Equal	{ Darkness - no lighting }	<input type="checkbox"/>

Fact Table Count
7203

## D. Οπτικοποίηση Υπερκύβου OLAP

Για την οπτικοποίηση χρησιμοποιήθηκε το εργαλείο Powerview το οποίο είναι επέκταση του Microsoft Office Excel. Παρακάτω παρουσιάζονται τα αποτελέσματα μας.

### Παράδειγμα 1

Ερώτηση 1<sup>η</sup>: «Ποιος είναι το αριθμός των ατυχημάτων που συνέβησαν την ημέρα Κυριακή, εντός της βετίας που έχει επιλεγεί παραπάνω, όπου και ο πρώτος και ο δεύτερος δρόμος είναι τύπου A και επικρατούν συνθήκες απόλυτης συσκότισης.» (Η απάντηση είναι 103 ατυχήματα).

Example 1

Fact Table Count	Lightning Label	Label	1st Road Class - Label	2nd Road Class - Label
103	Darkness - no lighting	Sunday	A	A

Filters

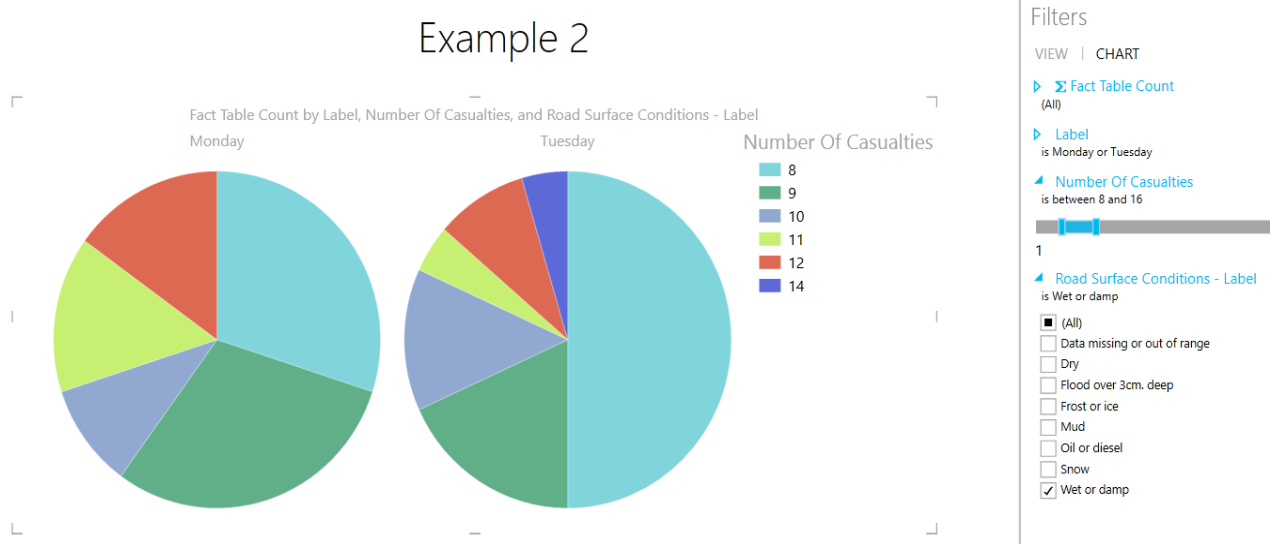
VIEW | TABLE

- 1st Road Class - Label  
is A
  - (All)
  - A
  - A(M)
  - B
  - C
  - Motorway
  - Unclassified
- 2nd Road Class - Label  
is A
  - Fact Table Count (All)
- Label  
is Sunday
- Lightning Label  
is Darkness - no lighting
  - (All)
  - Darkness - lighting unknown
  - Darkness - lights lit
  - Darkness - lights unlit
  - Darkness - no lighting
  - Data missing or out of range
  - Daylight



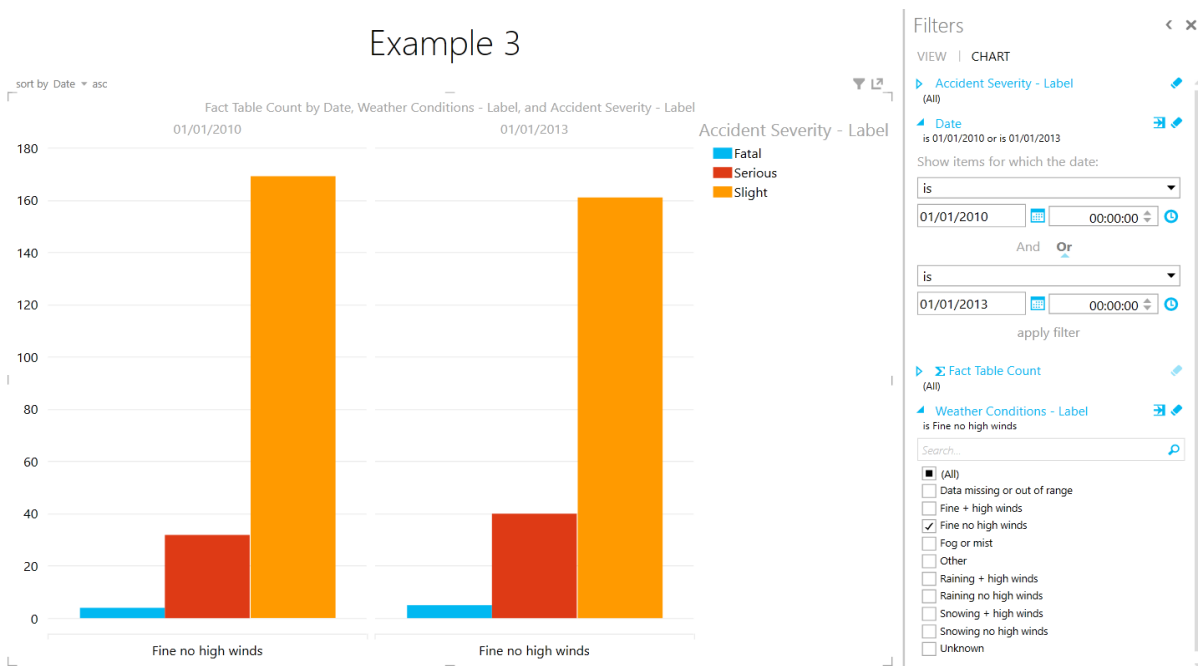
## Παράδειγμα 2

Ερώτηση 2: «Ποια είναι η κατανομή των θυμάτων για τις μέρες Δευτέρα και Τρίτη, εντός της δετίας που έχει επιλεγεί παραπάνω, όπου επικρατούν συνθήκες υγρασίας και ο αριθμός των θυμάτων κυμαίνεται μεταξύ των 8 και 16.» (Η απάντηση διαφαίνεται στα παρακάτω διαγράμματα-pies).



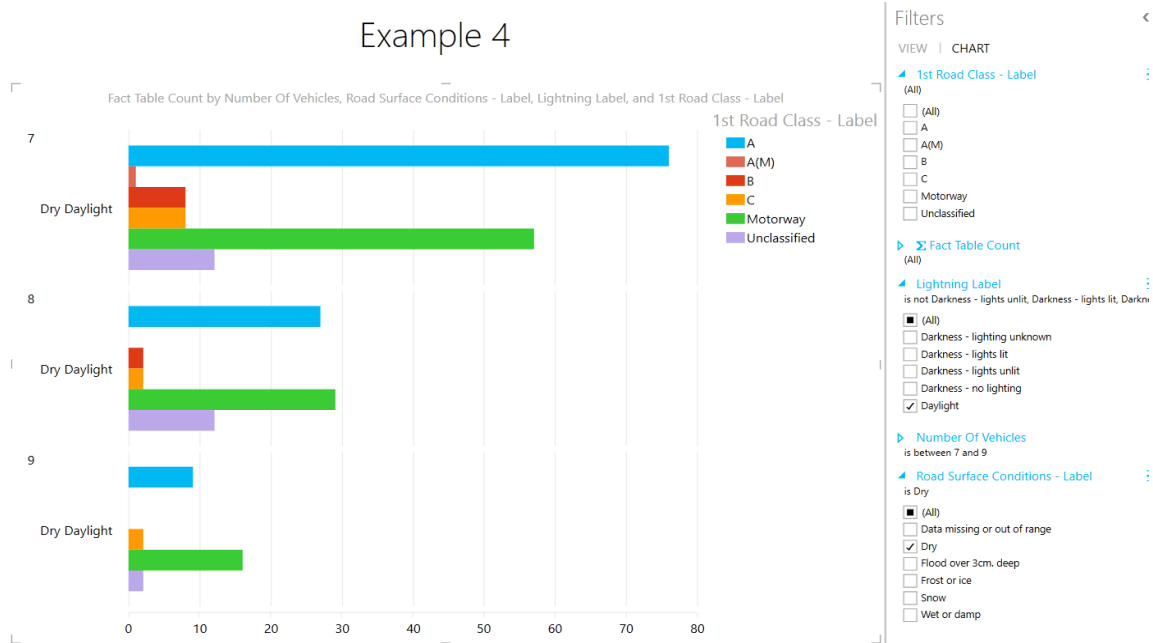
## Παράδειγμα 3

Ερώτηση 3: «Ποια είναι η κατανομή των ατυχημάτων ανάλογα με την σοβαρότητα ατυχήματος (fatal, serious, slight) για τις ημερομηνίες 1/01/2010 και 1/01/2013 αντίστοιχα, ενώ επικρατούσαν ήπιοι άνεμοι.» (Η απάντηση διαφαίνεται στα παρακάτω διαγράμματα-bars).



## Παράδειγμα 4

Ερώτηση 4: «Ποια είναι η κατανομή των ατυχημάτων με αριθμό οχημάτων 7, 8 και 9 αντίστοιχα για τους διάφορους τύπους του 1<sup>ου</sup> δρόμου κατά τη διάρκεια της μέρας και ενώ η επιφάνεια του δρόμου είναι στεγνή.» (Η απάντηση διαφαίνεται στα παρακάτω διαγράμματα-bars).



## Παράδειγμα 5

Ερώτηση 5: «Ποια είναι η κατανομή των θυμάτων ανάλογα με τον αριθμό των οχημάτων για την ημέρα Παρασκευή, εντός της δετίας που έχει επιλεγεί παραπάνω, δεδομένο ότι ο αριθμός των θυμάτων είναι μεγαλύτερος ή ίσος του 20.» (Η απάντηση διαφαίνεται στα παρακάτω διαγράμματα-pies).

