



ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Σχολή Τεχνολογίας

Τμήμα Ψηφιακών Συστημάτων

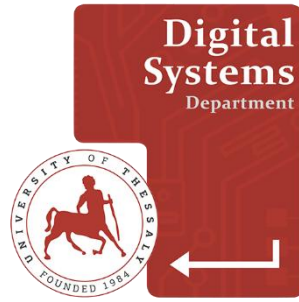
Ανάπτυξη Αποθήκης Δεδομένων Ανάλυσης Κλιματικών Φαινομέ- νων

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Χρήστος Αναστασίου (ΑΜ: 3121093)

Επιβλέπουσα: Γεωργία Γκαράνη, Καθηγήτρια

ΛΑΡΙΣΑ, Οκτώβριος 2025



UNIVERSITY OF THESSALY

School of Technology

Digital Systems Department

Development of a Data Warehouse for the Analysis of Climatic phenomena

BACHELOR THESIS

Christos Anastasiou (RN: 3121093)

Supervisor: *Georgia Garani, Professor*

LARISSA, October 2025

Υπεύθυνη Δήλωση περί Ακαδημαϊκής Δεοντολογίας και Πνευματικών Δικαιωμάτων

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, δηλώνω ρητά ότι η παρούσα πτυχιακή εργασία, καθώς και τα ηλεκτρονικά αρχεία και πηγαίοι κώδικες που αναπτύχθηκαν ή τροποποιήθηκαν στα πλαίσια αυτής της εργασίας, αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον, και πληρούν τους κανόνες της επιστημονικής παράθεσης. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο, αρχεία ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο/Η Δηλών/ούσα

(Υπογραφή)

Χρήστος Αναστασίου

Οκτώβριος 2025

Εγκρίνεται από την Επιτροπή Εξέτασης:

Επιβλέπων/πουσα	Ονοματεπώνυμο Επιβλέπουσα Βαθμίδα/ιδιότητα επιβλέποντα, Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Θεσσαλίας
Μέλος	Ονοματεπώνυμο Μέλους 1 Βαθμίδα/ιδιότητα μέλους 1, Τμήμα/Ιδρυμα μέλους 1
Μέλος	Ονοματεπώνυμο Μέλους 2 Βαθμίδα/ιδιότητα μέλους 2, Τμήμα/Ιδρυμα μέλους 2

Ημερομηνία έγκρισης: dd-mm-yyyy

Περίληψη

Η πτυχιακή εργασία θα επικεντρωθεί στη δημιουργία μιας αποθήκης δεδομένων που θα συγκεντρώνει και θα αναλύει δεδομένα από κλιματικά φαινόμενα, όπως βροχοπτώσεις, τυφώνες, ανέμους και πλημμύρες. Τα δεδομένα θα αντληθούν από μετεωρολογικές υπηρεσίες και ανοιχτές βάσεις δεδομένων, και θα ενσωματωθούν σε ένα ενιαίο σύστημα. Η αποθήκη δεδομένων θα επιτρέπει την ιστορική ανάλυση των φαινομένων και την εξαγωγή πολύτιμων πληροφοριών για τη συχνότητα, την ένταση και τη γεωγραφική κατανομή τους. Επιπλέον, θα υλοποιηθούν αναφορές και οπτικοποιήσεις με χρήση εργαλείων όπως το Power BI, παρέχοντας τη δυνατότητα εντοπισμού πρόβλεψης μελλοντικών κινδύνων. Το έργο αυτό μπορεί να συμβάλει στην καλύτερη προετοιμασία και διαχείριση φυσικών καταστροφών, υποστηρίζοντας τη λήψη αποφάσεων από αρμόδιες αρχές και φορείς πολιτικής προστασίας.

Abstract

The following thesis concerns the development of Data Warehouse for the analysis climate phenomena. Its purpose is to manage extreme weather events and to support forecast and rapid incidents that have been affecting the United States of America for decades.

For data collection, the NOAA Storms Events Database was used, covering the period from 2010 to 2024 across all U.S states. The data was organized using a Star Schema model, with a Fact Table and dimensions. This was followed by the ETL process (Extract, Transform, Load) which was used to clean and load the data into the warehouse. Using a business intelligence visualization tool, data analysis and dashboards were created, which contributed to understanding, drawing conclusions, and supporting decision making by authorities like Civil Protection.

Finally, the thesis presented and utilized the Business Intelligence and Data Warehouse tools, offering a practical and future-expandable model.

Ευχαριστίες

Θα ήθελα να εκφράσω τις ευχαριστίες μου στην επιβλέποντα καθηγήτρια κ. Γεωργία Γκαράνη για την αμέριστη προσοχή, την υπομονή και την υποστήριξη που εισέπραξα καθ' όλη την διάρκεια της εκπόνησης της εργασίας μου καθώς η συμβολή της ήταν καθοριστική.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου για την στήριξη τους και την πίστη τους σε εμένα.

Χρήστος Αναστασίου

Οκτώβριος 2025

Περιεχόμενα

ΠΕΡΙΛΗΨΗ	VII
ABSTRACT	IX
ΕΥΧΑΡΙΣΤΙΕΣ.....	XI
ΠΕΡΙΕΧΟΜΕΝΑ.....	XIII
1 ΕΙΣΑΓΩΓΗ	1
2 ΚΛΙΜΑΤΙΚΑ ΦΑΙΝΟΜΕΝΑ ΚΑΙ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ	3
2.1 ΤΥΠΟΙ ΚΛΙΜΑΤΙΚΩΝ ΦΑΙΝΟΜΕΝΩΝ ΚΑΙ ΕΠΙΠΤΩΣΕΙΣ	3
2.2 ΣΗΜΑΣΙΑ ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ ΚΛΙΜΑΤΙΚΩΝ ΦΑΙΝΟΜΕΝΩΝ.....	5
3 ΑΠΟΘΗΚΕΣ ΔΕΔΟΜΕΝΩΝ	7
3.1 ΟΡΙΣΜΟΣ ΚΑΙ ΣΚΟΠΟΣ ΑΠΟΘΗΚΩΝ ΔΕΔΟΜΕΝΩΝ.....	7
3.2 ΟΡΙΣΜΟΣ ΔΙΑΔΙΚΑΣΙΑΣ ETL (EXTRACT, TRANSFORM, LOAD)	10
3.2.1 <i>Extract</i>	12
3.2.2 <i>Transform</i>	13
3.2.3 <i>Load</i>	14
3.3 ΜΟΝΤΕΛΑ ΔΕΔΟΜΕΝΩΝ.....	15
3.3.1 <i>Μοντέλο αστέρα</i>	16
3.3.2 <i>Μοντέλο χιονονιφάδας</i>	17
3.4 ΔΙΑΦΟΡΕΣ ΑΠΟΘΗΚΩΝ ΔΕΔΟΜΕΝΩΝ, ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΛΙΜΝΩΝ ΔΕΔΟΜΕΝΩΝ	19
3.4.1 <i>Αποθήκες δεδομένων vs βάσεις δεδομένων</i>	22
3.4.2 <i>Αποθήκες δεδομένων vs λίμνες δεδομένων</i>	23
4 ΣΧΕΔΙΑΣΗ ΚΑΙ ΥΛΟΠΟΙΗΣΗ ΑΠΟΘΗΚΗΣ ΔΕΔΟΜΕΝΩΝ	25
4.1 ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΑΠΟΘΗΚΗΣ	25
4.2 ΔΗΜΙΟΥΡΓΙΑ ΛΟΓΙΚΟΥ ΣΧΗΜΑΤΟΣ.....	27
4.3 ΔΙΑΔΙΚΑΣΙΑ ETL	28
4.4 ΥΛΟΠΟΙΗΣΗ ΜΟΝΤΕΛΟΥ ΑΣΤΕΡΑ.....	30

4.4.1	Πίνακας Γεγονότων	30
4.4.2	Διαστάσεις.....	32
5	ΟΡΙΣΜΟΣ ΚΑΙ ΔΙΑΧΕΙΡΙΣΗ ΔΕΔΟΜΕΝΩΝ.....	35
5.1	ΔΗΜΙΟΥΡΓΙΑ ΚΩΔΙΚΑ SQL	35
5.2	ΕΡΩΤΗΜΑΤΑ.....	38
6	ΟΠΤΙΚΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ	41
7	ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΟΠΤΙΚΕΣ	50
7.1	ΑΞΙΟΛΟΓΗΣΗ ΕΡΓΑΛΕΙΩΝ ΠΟΥ ΧΡΗΣΙΜΟΠΟΙΗΘΗΚΑΝ.....	50
7.2	ΠΙΘΑΝΟ ΣΕΝΑΡΙΟ ΧΡΗΣΗΣ ΑΠΟ ΠΟΛΙΤΙΚΗ ΠΡΟΣΤΑΣΙΑ.....	51
7.3	ΒΑΣΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ.....	52
	ΒΙΒΛΙΟΓΡΑΦΙΑ.....	55
	ΠΑΡΑΡΤΗΜΑ Α: ΔΗΜΙΟΥΡΓΙΑ ΟΨΗΣ	59

1 Εισαγωγή

Η κλιματική αλλαγή και τα ακραία καιρικά φαινόμενα που συνοδεύονται από αυτή, όπως πλημμύρες, τυφώνες και πυρκαγιές, αποτελεί ένα από τα πιο σοβαρά θέματα στον πλανήτη και είναι μια σημαντική πρόκληση που αφορά κάθε άνθρωπο, καθώς επιφέρει συνέπειες καταστροφικές σε κοινωνικό, οικονομικό και φυσικά περιβαλλοντικό επίπεδο. Για αυτό είναι απαραίτητη η πρόβλεψη και τεχνολογική υποστήριξη ώστε να λαμβάνονται άμεσα μέτρα και αποφάσεις από τα αρμόδια μέσα.

Συγκεκριμένα, στις ΗΠΑ, που φημίζονται περισσότερο από οποιαδήποτε άλλη περιοχή στον κόσμο για τις καταστροφικές επιπτώσεις που έχει προκαλέσει η κλιματική αλλαγή, όπως για παράδειγμα στην Καλιφόρνια για δασικές αλλά και αστικές πυρκαγιές μεγάλης κλίμακας, στην Φλόριντα για τυφώνες, ανεμοστρόβιλους κλπ. και στο Τέξας για πολλαπλές πλημμύρες. Καθίσταται απαραίτητη η καλύτερη και πιο γρήγορη τεχνολογική υποστήριξη για την αποφυγή τραυματισμών, απώλεια περιουσίας και το πιο σημαντικό, αποτροπή θανάτων. Η παρούσα εργασία είναι επίκαιρη και σημαντική στον σύγχρονο κόσμο διότι μπορεί να χρησιμοποιηθεί από αρμόδιους φορείς όπως η πολιτική προστασία.

Ο σκοπός της πτυχιακής είναι η δημιουργία μιας ολοκληρωμένης αποθήκης δεδομένων για μια πλήρη και κατανοητή ανάλυση και οπτικοποίηση. Η μελέτη εστιάζει στις Ηνωμένες Πολιτείες Αμερικής με δεδομένα από την χρονική περίοδο 2009-2024. Η συλλογή τους πραγματοποιήθηκε από την βάση NOAA (National Oceanic and Atmospheric Administration) και συγκεκριμένα την NOAA Storm Events Database.

Τα βήματα για την δημιουργία της αποθήκης είναι ο σχεδιασμός της αρχιτεκτονικής με μοντελοποίηση star schema και η εφαρμογή της διαδικασίας ETL (Extract, Transform, Load) για την εκκαθάριση, μετασχηματισμό και φόρτωση στο σύστημα. Το τελικό κομμάτι είναι η οπτικοποίηση μέσω εργαλείου επιχειρηματικής ευφυΐας για την καλύτερη κατανόηση μέσω ενδεικτικών απεικονίσεων και αναλύσεων.

Σχετικά με την δομή της πτυχιακής απαρτίζεται από 6 κύρια κεφάλαια :

- Στο 2^ο κεφάλαιο επικεντρώνεται σε θεωρητικό επίπεδο καθώς έχει στόχο να αναδείξει την απαραίτητη ανάγκη ανάλυσης των κλιματικών φαινομένων, διάφορους

τύπους κλιματικών φαινομένων και τις επιπτώσεις τους και την πηγή συλλογής δεδομένων που θα αναλυθούν.

- Στο 3^ο κεφάλαιο αναλύεται σε γενικό επίπεδο η έννοια και ο σκοπός της αποθήκης δεδομένων, ο ορισμός της διαδικασίας ETL, τα σχηματικά μοντέλα star schema και snowflake schema και οι διαφορές των αποθηκών δεδομένων με τις βάσεις (database) και τις λίμνες δεδομένων (datalake).
- Στο 4^ο κεφάλαιο περιγράφεται η τεχνική σχεδιασμού και υλοποίησης της αποθήκης από την επιλογή των μεταβλητών που θα αναλυθούν και τη δημιουργία λογικού σχήματος με πίνακα γεγονότων (fact table), πίνακες διαστάσεων (dimensions) και σχέσεις μεταξύ τους μέχρι την υλοποίηση του star schema και την εφαρμογή της διαδικασίας ETL.
- Στο 5^ο κεφάλαιο γίνεται αναφορά των εργαλείων Power BI και SQL Server Management Studio (SSMS). Παρουσιάζεται η δημιουργία sql ερωτημάτων και η απεικόνιση dashboards και αναφορών ενώ επίσης προτείνεται ένα σενάριο χρήσης από αρμόδιο φορέα λήψης αποφάσεων και άλλων ειδικών. Τέλος, εξάγονται συμπεράσματα ανάλυσης και αξιολόγηση εργαλείων.
- Στο 6^ο και τελικό κεφάλαιο αναδεικνύονται τα τελικά συμπεράσματα και πιθανές μελλοντικές προεκτάσεις και βελτιώσεις της πτυχιακής εργασίας.

Η συγκεκριμένη εισαγωγή προσδίδει σαφήνεια ως προς την ανάλυση που θα επέλθει και αναδεικνύει επαρκώς την ακόλουθη δομή της εργασίας.

2 Κλιματικά φαινόμενα και ανάλυση δεδομένων

Αποτελεί αδιαμφισβήτητο γεγονός πως τα κλιματικά φαινόμενα και η κλιματική αλλαγή κατακλύζουν ολόκληρο τον πλανήτη με καταστροφικά φαινόμενα και τεράστιες κοινωνικοοικονομικές επιπτώσεις. Είναι απαραίτητο λοιπόν να υπάρχει ο κατάλληλος όγκος δεδομένων και ολοκληρωμένες αναλύσεις αυτών που θα προβλέπουν τέτοιου είδους φαινόμενα.

2.1 Τύποι κλιματικών φαινομένων και επιπτώσεις

Είναι λοιπόν κρίσιμα γεγονότα που σχετίζονται με την υδρολογική ατμοσφαιρική και θερμοκρασιακή μεταβολή του περιβάλλοντος. Κάποια από τα πιο σημαντικά όπως έχουν προαναφερθεί, είναι οι πλημμύρες, οι πυρκαγιές, οι χιονοθύελλες και οι καύσωνες. Τις τελευταίες δεκαετίες σε όλες τις περιοχές του πλανήτη τα γεγονότα έχουν αυξηθεί σε μεγάλο βαθμό σε συχνότητα αλλά και σε ένταση. Αυτό δυστυχώς οφείλεται κυρίως στην επίδραση που έχει ο άνθρωπος σήμερα στο περιβάλλον (Otto et al., 2022).

Από την άλλη μεριά, εκτός του σημαντικού αποτυπώματος που αφήνει ο άνθρωπος, υπάρχουν και άλλοι φυσικοί παράγοντες που επηρεάζουν το κλίμα. Σύμφωνα με τους Wohl, Pulwarty και Zhang (2019), η κλιματική αλλαγή δεν προβλέπεται εύκολα λόγω του μη σταθερού χρονικά κλίματος. Ως κλίμα ορίζεται ένα «σύνολο από ποσότητες» που αντιδρά με την ατμόσφαιρα, την υδρόσφαιρα και την κρυόσφαιρα με αποτέλεσμα να τις μεταβάλλει με το πέρασμα των χρόνων. Υπάρχουν διάφοροι φυσικοί μηχανισμοί οι οποίοι ευθύνονται για την κλιματική αλλαγή, όπως είναι τα αστρονομικά, ατμοσφαιρικά ή και γεωφυσικά φαινόμενα που επηρεάζουν άμεσα τη στρατόσφαιρα της Γης και μπορεί να έχουν ευθύνη για αλλαγές σε ηλιακή ακτινοβολία που συμβάλλουν στο λιώσιμο των πάγων, τροπικούς κυκλώνες που αποδεικνύονται καταστροφικοί και υποθαλάσσια ηφαιστεια που προκαλούν αλλαγή στη θερμοκρασία των ωκεανών.

Οι κλιματικοί κύκλοι είναι ένας σημαντικός παράγοντας και από τους πιο γνωστούς είναι το *El nino* και το *La Nina*. Οι ονομασίες τους είναι ισπανικοί όροι και προέρχονται από τις λέξεις το αγόρι και το κορίτσι αντίστοιχα. Ο πρώτος είναι ένα ωκεάνιο φαινόμενο

όπου τα κεντρικά και ανατολικά νερά του Ειρηνικού Ωκεανού κοντά στον Ισημερινό είναι πιο θερμά σε σχέση με τα υπόλοιπα άλλων περιοχών. Επιπρόσθετα, η εμφάνιση του φαινομένου γίνεται κάθε δύο με οχτώ χρόνια και κρατάει περίπου για ένα χρόνο. Η εμφάνιση του γίνεται κοντά στα Χριστούγεννα ενώ ευθύνεται για κλιματικά φαινόμενα όπως οι βροχοπτώσεις, οι ξηρασίες, οι πυρκαγιές, οι καύσωνες και οι τροπικοί κυκλώνες. Τέλος, το φαινόμενο συμβάλλει στο λιώσιμο των πάγων και φαίνεται να είναι ο κύριος λόγος πίσω από τον αφανισμό ενός ολόκληρου πολιτισμού, αυτού των Χμέρ. Το ωκεάνιο φαινόμενο La Nina, είναι το ψυχρότερο αντίστοιχο του El nino και έχει διάρκεια περίπου 5 μηνών. Και τα δύο αυτά φαινόμενα θεωρούνται μέρος της Νότιας ταλάντωσης του Ειρηνικού Ωκεανού (NOAA, 2024).

Άλλοι εξίσου σημαντικοί παράγοντες είναι οι ηλιακοί κύκλοι που αλλάζουν κάθε 11 χρόνια και επιφέρουν σημαντική αλλαγή της θερμοκρασίας, οι ηφαιστειακές εκρήξεις που μπορεί να προκαλέσουν τεράστιες φυσικές καταστροφές συμβάλλοντας παράλληλα στην ανάπτυξη φαινομένων όπως οι σεισμοί, οι ροές λάβας, οι πλημμύρες ακόμα και τσουνάμι. Τα κλιματικά φαινόμενα μπορούν να κατηγοριοποιηθούν σε ακραία και σταδιακά. Τα ακραία όπως πολλά από αυτά που έχουν ήδη αναφερθεί (πλημμύρες, πυρκαγιές), έχουν άμεσο αντίκτυπο στις διάφορες περιοχές του πλανήτη σε αντίθεση με τα σταδιακά που περιλαμβάνουν φαινόμενα όπως η αύξηση της θερμοκρασίας και η άνοδος της στάθμης της θάλασσας. Η διαφοροποίηση τους από τα πρώτα είναι ότι επηρεάζουν τον πλανήτη σε βάθος χρόνου μεταβάλλοντας το κλίμα και τις τάσεις χωρών και ηπείρων.

Όσο αναφορά την γεωγραφική τάση των φαινομένων, είναι φυσικό και αναμενόμενο σε κάθε περιοχή να είναι διαφορετική. Το παράδοξο σε αυτό είναι ότι χώρες λιγότερο ανεπτυγμένες επηρεάζονται περισσότερο από τις ανεπτυγμένες από ακραία καιρικά φαινόμενα λόγω γεωγραφικής κατανομής και έλλειψη αντιμετώπισης τους. Ενώ οι ανεπτυγμένες χώρες συμβάλλουν κατά πολύ περισσότερο σε φαινόμενα όπως αυτό του θερμοκηπίου και αλλάζουν έμμεσα το κλίμα της χώρας τους, λόγω μεγαλύτερης ανάπτυξης τους στον τομέα της Πολιτικής Προστασίας μπορούν να αντιμετωπίζουν πολύ καλύτερα ακραία φαινόμενα από τις χώρες-κράτη που δεν έχουν τα μέσα να τα διαχειριστούν, κάτι που επιφέρει καταστροφικές συνέπειες με περισσότερους νεκρούς, τραυματίες και υλικές καταστροφές. Μάλιστα, έχει παρατηρηθεί ότι υπάρχει αύξηση σε κλιματικά γεγονότα από 24% έως 48% στο 20% του φτωχότερου παγκόσμιου πληθυσμού ειδικά σε πιο φτωχές περιοχές (Camila Donatti et al. 2024). Επίσης οι ανεπτυγμένες χώρες χρειάζονται

μεγαλύτερες κλιματικές αλλαγές από τις αναπτυσσόμενες για να προκληθούν σοβαρές επιπτώσεις (Tschumi & Zscheimer, 2019).

2.2 Σημασία ανάλυσης δεδομένων κλιματικών φαινομένων

Η ανάλυση δεδομένων κλιματικών φαινομένων αποτελεί θεμελιώδη διαδικασία για την κατανόηση, την πρόβλεψη και την αντιμετώπιση των επιπτώσεων της κλιματικής αλλαγής. Η συμβολή της σε συνδυασμό με την εξέλιξη της τεχνολογίας θα βοηθήσει κυβερνήσεις και πολιτική προστασία να διαθέτουν τα κατάλληλα μέσα αντιμετώπισης.

Η ενημέρωση των ανθρώπων ως προς τις επιπτώσεις που έχουν στο περιβάλλον βοήθησε στην εξέλιξη τέτοιων τεχνολογιών και εκτός από το γεγονός ότι τον έχει οδηγήσει σε περιορισμό δραστηριοτήτων επικίνδυνων προς το περιβάλλον, γίνεται προσπάθεια παγκοσμίως και για αλλαγές πολιτικής δεδομένων ώστε να παρθούν τα κατάλληλα μέτρα (Schmidt & Strauss, 2023). Οι αλλαγές του κλίματος προβλέπονται να αυξηθούν σε ένταση και συχνότητα τα επόμενα χρόνια επηρεάζοντας τομείς όπως:

- Καλλιέργεια
- Υγεία
- Οικονομία
- Περιβάλλον

Πιο συγκεκριμένα, η παραγωγή της τροφής από καλλιέργειες μπορεί να μειωθεί και σε κάποιες περιπτώσεις να μεταλλαχτεί εξαπλώνοντας ασθένειες και ιούς (Lobell & Gourdji, 2012).

Για την αντιμετώπιση των παραπάνω, θα αναλυθούν τρόποι που βασίζονται σε νέες τεχνολογίες όπως:

- GIS: Λογισμικό για διαχείριση, ανάλυση και απεικόνιση γεωγραφικών δεδομένων. Το εργαλείο αυτό συμβάλλει στην λήψη αποφάσεων σε πολλούς τομείς.
- Κλιματικά μοντέλα: Μοντέλα προσομοίωσης παρόμοιου κλίματος με αυτό της Γης. Βοηθά στην κατανόηση, σύγκριση και πρόβλεψη του κλίματος.
- Οικονομικά μοντέλα: Παρόμοια προσφέρουν προσομοίωση για την ανάδειξη οικονομικής συμπεριφοράς σε μια περίοδο χρόνου και φαινομένων.

Οι παραπάνω τρόποι προτείνονται στο άρθρο των Rawat, Kumar και Khati (2024). Επιπρόσθετα, οι Kamilaris, Kartakoullis και Prenafeta-Boldú (2017) πρότειναν την χρήση

μηχανικής μάθησης σε συνδυασμό με την τεχνητή νοημοσύνη για ανάλυση δεδομένων. Αυτό θα συνέβαλε στη λήψη καλύτερων αποφάσεων στον χώρο της καλλιέργειας. Όσο αναφορά την θάλασσα και τις επιπτώσεις που την κατακλύζουν, η συνεχής ανάπτυξη των big data και της τεχνητής νοημοσύνης, δημιουργούν πατήματα για την δημιουργία μοντέλων πρόβλεψης και ανάλυσης συχνότητας (Rawat, Kumar & Khati, 2024).

Συνοψίζοντας, είναι προφανής η ανάγκη για ύπαρξη κλιματικών δεδομένων. Η σημασία τους σε τομείς οικονομίας, γεωργίας και υγείας αποδεικνύεται τεράστια ενώ μπορούν να μετατραπούν σε κύρια εργαλεία που χρησιμοποιούνται για προσομοιώσεις, προβλέψεις και άλλες βασικές λειτουργίες με τις οποίες φορείς πολιτικής προστασίας μπορούν να λάβουν μέτρα. Επίσης, η ευαισθητοποίηση του ανθρώπου μπορεί να ενισχύσει την προσπάθεια για περισσότερες πολιτικές παρεμβάσεις. Ως εκ τούτου, ο συνδυασμός εφαρμογής δεδομένων, τεχνολογιών και πολιτικών πρωτοβουλιών αποτελεί ένα σημαντικό βήμα προς τη διαμόρφωση ενός υποσχόμενου μέλλοντος ενόψει των αυξανόμενων προκλήσεων της κλιματικής αλλαγής.

3 Αποθήκες Δεδομένων

Στις μέρες μας, η ανάγκη για αποθήκευση μεγάλου όγκου δεδομένων και η αξία εξαγωγή τους σε εμπεριστατωμένες αποφάσεις, είναι πολύ σημαντική και είναι απαραίτητο για επιχειρήσεις και κρατικούς μηχανισμούς να διαθέτουν τα μέσα για να την κάνουν. Σε αυτό το κεφάλαιο θα αναλυθεί αυτή η ανάγκη και θα δοθεί η έννοια της αποθήκης δεδομένων. Επιπροσθέτως, θα αναλυθεί η σημασία της διαδικασίας ETL και θα αναφερθούν τα σχηματικά μοντέλα star schema και snowflake schema. Τέλος, θα οριστούν οι απλές βάσεις δεδομένων και οι λίμνες δεδομένων και θα εστιάσουμε στις διαφορές τους με τις αποθήκες δεδομένων.

3.1 Ορισμός και σκοπός αποθηκών δεδομένων

Η αποθήκη δεδομένων μπορεί να οριστεί ως ένα κεντρικό αποθετήριο συλλογής πληροφοριών. Είναι μια βάση δεδομένων που χρησιμοποιείται για αναλύσεις και για την εξαγωγή συμπερασμάτων και αποφάσεων που βασίζονται σε πραγματικές πληροφορίες. Όπως αναφέρει ο Inmon, «η αποθήκη δεδομένων είναι το θεμέλιο όλων των συστημάτων υποστήριξης αποφάσεων για την επεξεργασία δεδομένων». Αυτό την καθιστά πολύ χρήσιμη και αναγκαία στον χώρο των επιχειρήσεων. Επίσης αναφέρεται ότι «είναι μια αντικειμενοστραφής, μη πτητική και χρονικά μεταβαλλόμενη συλλογή δεδομένων που υποστηρίζει αποφάσεις διοίκησης». Με την έννοια αντικειμενοστραφής, εννοείται ότι η αποθήκη στρέφεται γύρω από αντικείμενα, όπως είναι τα προϊόντα και οι πελάτες ή σε αυτή την περίπτωση γύρω από καιρικά φαινόμενα και άλλες μεταβλητές (Inmon, 2005, p. 29). Εάν δηλαδή δημιουργηθεί μια ανάλυση, αυτή θα γίνει με βάση τα αντικείμενα ή δεδομένα που διαθέτουμε. Παραδείγματος χάρη, στις αναλύσεις θα μπορούσαν να υλοποιηθούν ερωτήματα ως εξής:

- Πόσοι θάνατοι προκλήθηκαν από τυφώνες στις ΗΠΑ από το 2015 έως το 2024: Το ερώτημα στρέφεται με βάση τους θανάτους και τα φαινόμενα των τυφώνων μια χρονική περίοδο.
- Ποιος ο μέσος όρος πυρκαγιών στη Καλιφόρνια το διάστημα της δεκαετίας 2010–2020: Η ανάλυση είναι ξανά στηριγμένη γύρω από τα δεδομένα.

Με τον όρο μη πτητική, περιγράφεται ο περιορισμός του να μην αλλάζουν τα δεδομένα μετά την εισαγωγή τους στην αποθήκη. Αυτό αποσκοπεί στη σταθερή λειτουργία της ανάλυσης χωρίς αλλαγές των υπάρχοντων δεδομένων που μπορούν να αλλοιώσουν την ανάλυση. Τέλος, ο Inmon χαρακτηρίζει την αποθήκη ως χρονικά μεταβαλλόμενη για να υποστηρίζει σε διάφορες χρονικές στιγμές δεδομένα και να μην είναι στατική. Με αυτόν τον τρόπο, προσφέρεται η δυνατότητα εις βάθος αναλύσεων και μπορούμε να εντοπίζουμε την μεταβολή των δεδομένων σε διάρκεια ημερών, μηνών ή και ετών. Στην περίπτωση μου, αυτό θα αποδειχθεί ιδανικό όπως και σε πολλές άλλες περιπτώσεις που θα γίνει χρήση αναλύσεων με γνώμονα την σύγκριση ανάμεσα σε χρονολογικές περιόδους. Το αποθετήριο δεδομένων, διαθέτει πάντα στοιχεία χρόνου ως δεδομένα και κάθε περιβάλλον έχει τον δικό του χρονικό ορίζοντα. Ο χρονικός ορίζοντας είναι η διάρκεια των χρονικών στοιχείων μέσα σε μια αποθήκη δεδομένων. Είναι συνηθισμένο η χρονική περίοδος των στοιχείων αυτών να είναι αρκετά μεγάλη και να κυμαίνεται από ένα έτος έως και σε πολλά έτη και λόγω αυτού υπάρχουν τεράστιοι όγκοι δεδομένων που δεν συγκρίνονται σε μέγεθος με άλλα λειτουργικά συστήματα. Μάλιστα. Τα λειτουργικά συστήματα έχουν πολλές διαφορές σε σχέση με μια αποθήκη δεδομένων και είναι οι εξής:

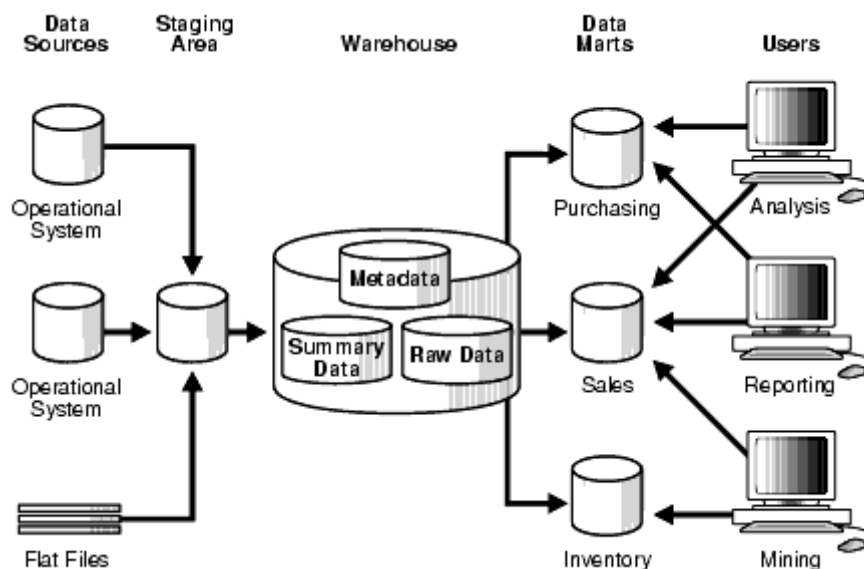
- Δεν χρειάζεται να περιέχουν χρονικά στοιχεία. Σε αντίθεση με τις αποθήκες που περιέχουν τουλάχιστον ένα.
- Ο χρονικός τους ορίζοντας κυμαίνεται από 60 έως και 90 μέρες. Για αυτόν τον λόγο το μέγεθος δεδομένων σε μια αποθήκη είναι πολύ μεγαλύτερο.
- Τα λειτουργικά δεδομένα είναι μη ενοποιημένα ενώ στην αποθήκη πρέπει πρώτα να ενοποιηθούν (Inmon, 2005, p. 18).

Η αρχιτεκτονική μιας αποθήκης δεδομένων σύμφωνα με Inmon, αποτελείται από 4 επίπεδα:

- Λειτουργικό
- Επίπεδο αποθήκης δεδομένων
- Επίπεδο υποσυνόλου αποθήκης δεδομένων ή Datamart
- Ατομικό

Το λειτουργικό διαθέτει τα αρχικά μη επεξεργασμένα δεδομένα και συγκεντρώνεται γύρω από τις εφαρμογές. Κύρια ευθύνη του είναι η επεξεργασία συναλλαγών που απαιτούν υψηλή απόδοση. Έπειτα, το επίπεδο αποθήκης δεδομένων έχει στην κατοχή του τα πλέον ολοκληρωμένα δεδομένα και μη επεξεργασμένα ιστορικά δεδομένα τα οποία δεν

μπορούν να τροποποιηθούν. Συνεχίζοντας, έχουμε το επίπεδο τμήματος ή αλλιώς Datamart που απαρτίζεται από σχεδόν εξολοκλήρου επεξεργασμένα αλλά όχι ολοκληρωμένα δεδομένα. Αυτό το επίπεδο δημιουργείται ανάλογα με τις απαιτήσεις των τελικών χρηστών. Το τελευταίο επίπεδο, είναι αυτό της ανάλυσης. Τα παραπάνω αναπαρίστανται αναλυτικά στην Εικόνα 1 με παράδειγμα εφαρμογής σε μια επιχείρηση. Αξίζει να σημειωθεί ότι και το μοντέλο αρχιτεκτονικής του Kimball είναι εξίσου δημοφιλές και ευρέως χρησιμοποιημένο με διαφορετικά πλεονεκτήματα και μειονεκτήματα από την αρχιτεκτονική του Inmon.



Εικόνα 1: Αρχιτεκτονική αποθήκης δεδομένων κατά Inmon

(<https://tdan.com/data-warehouse-design-inmon-versus-kimball/20300>)

Όσο αναφορά την δομή της **αποθήκης δεδομένων**, όπως περιγράφεται από τον Bill Inmon, αποτελείται από κάποια επίπεδα:

- Παλαιό επίπεδο λεπτομέρειας.
- Τρέχον επίπεδο λεπτομέρειας.
- Επίπεδο ελαφρώς συνοπτικών δεδομένων.
- Επίπεδο πολύ συνοπτικών δεδομένων.

Η μεγαλύτερη αλλαγή δεδομένων παίρνει μέρος στο βήμα μετάβασης από το λειτουργικό επίπεδο στο επίπεδο αποθήκης δεδομένων (Inmon, 2005, p.33).

Το πιο σημαντικό στοιχείο σχεδιασμού της αποθήκης δεδομένων είναι το θέμα της κοκκοποίησης. Η διαδικασία αυτή είναι αναγκαία γιατί καθορίζει την δομή των πινάκων γεγονότων (fact tables) και την απόδοση των ερωτημάτων sql. Σημαντική διαδικασία είναι

επίσης αυτή της διαχώρισης των δεδομένων σε μικρότερα μέρη καθώς έτσι είναι πιο εύκολη η διαχείριση τους σε σχέση με μια τεράστια μονάδα. Συνοψίζοντας, η κοκκοποίηση μετά την κατασκευή της αποθήκης, προσφέρει μεγάλα πλεονεκτήματα όπως η ευελιξία και η επαναχρησιμοποίηση δεδομένων (Inmon, 2005, pp. 41-42).

Σύμφωνα με τα παραπάνω, γίνεται αντιληπτό ότι μια αποθήκη δεδομένων είναι πολλά περισσότερα από απλώς μια συλλογή αποθήκευσης δεδομένων και είναι απαραίτητη σε πολλούς κλάδους της σύγχρονης εποχής όπως οι επιχειρήσεις, οι τομείς υγείας και οι μετεωρολογικοί σταθμοί που την χρησιμοποιούν ενεργά για σημαντικές αναλύσεις.

3.2 Ορισμός διαδικασίας ETL (Extract, Transform, Load)

Η διαδικασία ETL προέρχεται από τις λέξεις Extract, Transform και Load. Δηλαδή, εξαγωγή, μετατροπή και φόρτωση αντίστοιχα. Αποτελείται από εργαλεία που χρησιμοποιούνται πριν ενταχθούν τα δεδομένα σε μια αποθήκη δεδομένων. Είναι αναγκαία καθώς εξάγει δεδομένα από πηγές, έπειτα καθαρίζει, μετατρέπει και τελικά τα εισάγει στο αποθετήριο για ανάλυση (Panos, Simitsis & Skiadopoulos, 2002). Κάθε βήμα είναι εξίσου σημαντικό και απαιτεί προσοχή και λεπτομέρεια ώστε να επέλθει το καλύτερο δυνατό αποτέλεσμα για την αποθήκη και τις τελικές αναλύσεις. Τα τελευταία χρόνια, εκτός από την διαδικασία ETL, έχει ραγδαία άνοδο και η διαδικασία με παρόμοια ονομασία ELT. Η διαδικασία αυτή όπως φαίνεται και από την ονομασία της, φορτώνει τα δεδομένα απευθείας μετά την εξαγωγή τους και έπειτα έρχεται η φάση της μετατροπής. Με αυτόν τον τρόπο δημιουργούνται κάποια κύρια πλεονεκτήματα και μειονεκτήματα σε σχέση με την διαδικασία ETL που θα παρουσιαστούν στους παρακάτω πίνακες, Πίνακας 1 και Πίνακας 2.

ELT	
Πλεονεκτήματα	Μειονεκτήματα
Γρήγορη φόρτωση	Δύσκολο στην υλοποίηση
Πολλές επιλογές ανάλυσης	Μεγάλο κόστος υλοποίησης λόγω queries
Χαμηλές απαιτήσεις συντήρησης	

Πίνακας 1: Πλεονεκτήματα και Μειονεκτήματα ETL

ETL	
Πλεονεκτήματα	Μειονεκτήματα
Γρήγορη ανάλυση	Αρχικό κόστος
Ευρεία εφαρμογή	Έλλειψη ευελιξίας
Χαμηλές απαιτήσεις αποθήκευσης	Αργή επεξεργασία

ELT	
Πλεονεκτήματα	Μειονεκτήματα
Γρήγορη φόρτωση	Δύσκολο στην υλοποίηση
Πολλές επιλογές ανάλυσης	Μεγάλο κόστος υλοποίησης λόγω queries
Χαμηλές απαιτήσεις συντήρησης	

Πίνακας 2: Πλεονεκτήματα και Μειονεκτήματα ELT

Όπως διακρίνεται από τους πίνακες, η διαδικασία ETL δεν είναι το ίδιο ευέλικτη με την ELT όμως και η ELT έχει μεγαλύτερο κόστος και δυσκολότερο τρόπο υλοποίησης. Συμπερασματικά, η επιλογή μιας εκ των δύο διαδικασιών γίνεται ανάλογα με τα θέλω και τις ανάγκες του χρήστη καθώς δεν υπάρχει η τέλεια επιλογή. Όσο αναφορά τις επιχειρήσεις και τις εταιρείες, τότε για να διακρίνουν ποια από τις δύο διαδικασίες είναι πιο ωφέλιμη για εκείνες πρέπει για την ETL να γνωρίζουν ότι:

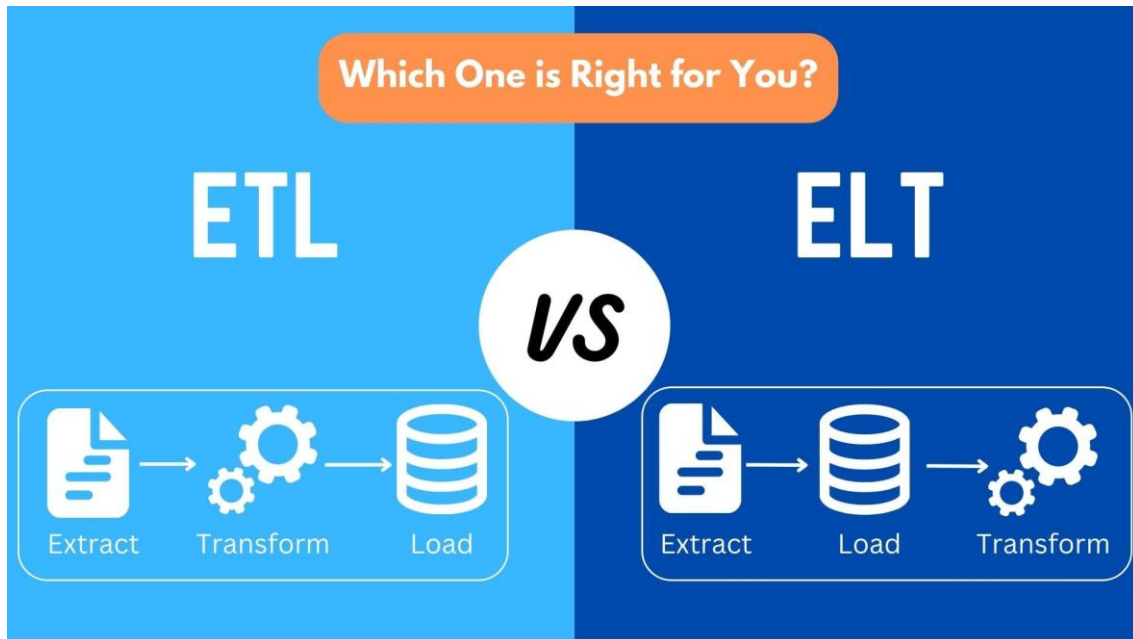
- Η εταιρεία θα είναι αναγκαίο να διαθέτει τον δικό της αποθηκευτικό χώρο.
- Το κόστος αποθήκευσης είναι ένα κρίσιμο σημείο που αξίζει προσοχή.
- Τα δεδομένα κυμαίνονται από μικρά έως μέτρια μεγέθη.
- Χρειάζονται τυπικές αναλύσεις.

Για την ELT να γνωρίζουν επίσης ότι:

- Θα ασχολούνται με πάρα πολλά δεδομένα.
- Οι πηγές δεδομένων αλλάζουν συνεχώς.
- Τα δεδομένα απαιτούν γρήγορη αποθήκευση.
- Είναι απαραίτητη η ανάλυση ιστορικών δεδομένων.

- Πρέπει να υπάρχουν όλα τα μέσα για την επεξεργασία (Young, 2025).

Στην Εικόνα 2:Αναπαράσταση των ETL, ELT διαδικασιών, φαίνεται η εικονική αναπαράσταση των δύο διαδικασιών.



Εικόνα 2:Αναπαράσταση των ETL, ELT διαδικασιών

(<https://www.kdnuggets.com/2023/03/etl-elt-one-right-data-pipeline.html>)

Συνοψίζοντας, στην συγκεκριμένη περίπτωση η διαδικασία ETL είναι η καλύτερη επιλογή λόγω των εργαλείων που χρησιμοποιώ και επειδή η ELT διαδικασία χρησιμοποιείται κυρίως σε cloud-based αποθήκες (Boyko & Chernenko, 2024). Όλες οι φάσεις είναι κρίσιμες και παρακάτω θα αναλυθούν ξεχωριστά ώστε να αναδείξουμε την σημασία τους.

3.2.1 Extract

Η εξαγωγή είναι μια κρίσιμη διαδικασία από μόνη της διότι τα δεδομένα συλλέγονται από διάφορες εξωτερικές πηγές και είναι ετερογενή και ακατέργαστα. Υπάρχουν δύο τρόποι εξαγωγής, η άμεση και η αναβληθείσα εξαγωγή. Η άμεση γίνεται μέσω:

- Ενεργοποιητών βάσεων δεδομένων
- Αρχείων καταγραφής συναλλαγών
- Συστημάτων προέλευσης

Ο πρώτος τρόπος που αναφέρεται είναι αρκετά εύχρηστος καθώς επιβαρύνει σε μικρό βαθμό τους πόρους συστήματος ενώ ο δεύτερος είναι ο καλύτερος αφού δεν επιβαρύνει

καθόλου τα OLTP (Online Transaction Processing) συστήματα. Η τρίτη επιλογή δεν συνιστάται. Η αναβληθείσα υλοποιείται με δύο τρόπους. Χρονοσήμανση (timestamp) και με την τεχνική της σύγκρισης αρχείων, με την πρώτη να αποτελεί πιο γρήγορη διαδικασία. Η επιλογή της άμεσης ή της αναβληθείσας εξαγωγής, αποφασίζεται ανάλογα με ένα βασικό κριτήριο. Αν είναι αναγκαία η συνεχής ενημέρωση των δεδομένων (real-time data) τότε προτιμάται η άμεση, εάν όχι συνιστάται η αναβληθείσα (Munawar, 2021).

Εν κατακλείδι, η εξαγωγή δεδομένων είναι ένα πολύ σημαντικό βήμα στη διαδικασία ETL, καθώς μικρά σφάλματα, όπως ο ατελής καθαρισμός ή ποιοτικός έλεγχος, μπορούν να οδηγήσουν σε αναξιόπιστες αναλύσεις.

3.2.2 Transform

Η επόμενη φάση είναι αυτή του μετασχηματισμού (transform) και χωρίς αυτή η διαδικασία ETL δεν μπορεί να προχωρήσει. Αυτό συμβαίνει διότι στην πρώτη φάση (extract) εξάγονται δεδομένα που είναι ακατέργαστα και εάν δεν τεθούν υπό επεξεργασία και φορτωθούν ωμά (raw data) θα υπάρξουν τεράστια λάθη στις αναλύσεις και ένα μη κατανοητό αποτέλεσμα.

Τα πιο σημαντικά στοιχεία του μετασχηματισμού είναι η εκκαθάριση, η κανονικοποίηση, η επικύρωση και η διαμόρφωση των δεδομένων ενώ οι πιο συχνές λειτουργίες μετασχηματισμών, που συνήθως γίνεται στο Excel, διακρίνονται στον Πίνακα 3: Μετασχηματισμοί δεδομένων.

Πίνακας 3: Μετασχηματισμοί δεδομένων

Μετασχηματισμοί	
Λειτουργία	Περιγραφή
Ταξινόμηση δεδομένων	Τα δεδομένα μπορεί να μην εμφανίζονται σε ταξινομημένη δομή και να απαιτείται προσαρμογή.
Φιλτράρισμα	Συνήθως υπάρχουν πολλά άσχετα, ελλιπή και ακατέργαστα δεδομένα.
Συγκέντρωση δεδομένων	Με συναρτήσεις όπως η sum αναδεικνύονται χρήσιμα δεδομένα για την ανάλυση.

Η παραπάνω παράγραφος και ο Πίνακας 3, είναι εμπνευσμένα από το άρθρο του Geeks-forGeeks που αναφέρεται στην διαδικασία ETL και την φάση μετασχηματισμού (Geeks-forGeeks, 2025).

Οι παραπάνω λειτουργίες ανήκουν στους τέσσερις τύπους μετασχηματισμών δεδομένων οι οποίοι είναι οι ακόλουθοι:

- Κατασκευαστικός: Προσθέτει και αντιγράφει δεδομένα.
- Καταστροφικός: Διαγράφει δεδομένα. Το φιλτράρισμα αποτελεί λειτουργία του συγκεκριμένου τύπου.
- Δομικός: Αναδιοργανώνει τα δεδομένα. Η συγκέντρωση δεδομένων υπάγεται σε αυτή την κατηγορία.
- Αισθητικός: Κανονικοποίηση των δεδομένων όπως για παράδειγμα η ταξινόμηση δεδομένων σε κάποια στήλη του Excel που έχει ανακατεμένη διάταξη (Jaishree, 2025).

Κλείνοντας, ο μετασχηματισμός είναι ακόμη ένα κρίσιμο βήμα για την διασφάλιση, αξιοπιστία και συνέπεια των δεδομένων που θα φορτωθούν στο επόμενο βήμα-φάση (Load) στην αποθήκη.

3.2.3 Load

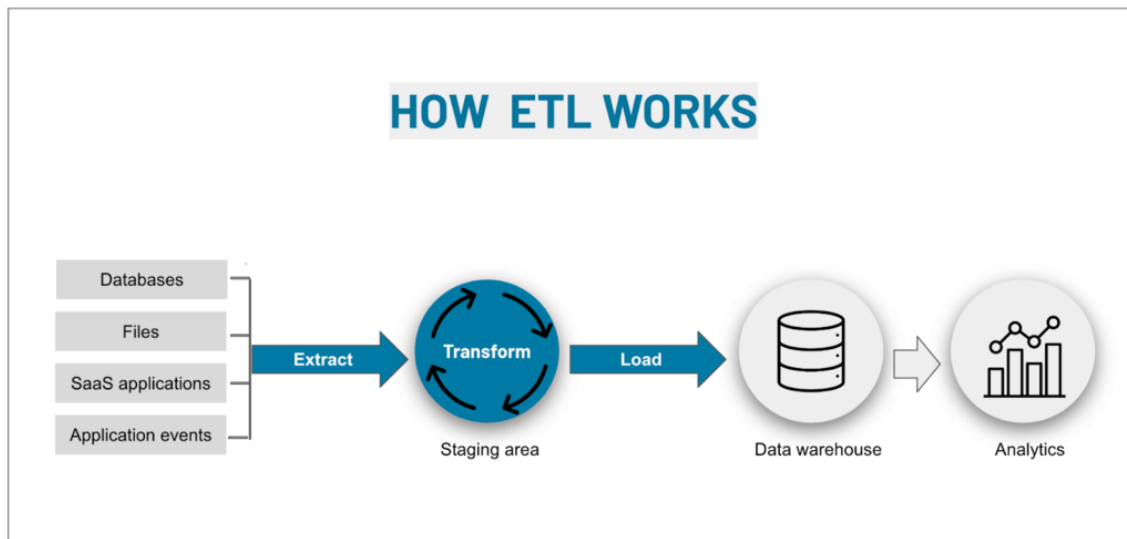
Η φόρτωση (Load) είναι το τελευταίο βήμα της διαδικασίας ETL. Σε αυτό, τα δεδομένα που έχουν πλέον τροποποιηθεί, φορτώνονται στην κεντρική αποθήκη δεδομένων.

Βέβαια, σύμφωνα με την AtScale (n.d), υπάρχουν προκλήσεις κατά την φόρτωση που χρήζουν αντιμετώπισης για την εύρυθμη λειτουργία του αποθετηρίου. Αρχικά, η απώλεια δεδομένων είναι εύκολο να συμβεί εάν δεν έχει προηγηθεί ολοκληρωτικός έλεγχος των δεδομένων. Επιπρόσθετα, είναι απαραίτητο να είναι γνωστή η προέλευση των δεδομένων και ποιος είναι ο τελικός προορισμός αυτών ώστε να αποφευχθεί πρόβλημα μορφοποίησης.

Υπάρχουν δύο μέθοδοι φόρτωσης, η πλήρη φόρτωση και η σταδιακή φόρτωση. Η πλήρης είναι η κύρια και όλα τα δεδομένα φορτώνονται μαζί στην αποθήκη δεδομένων. Στην σταδιακή φόρτωση, φορτώνεται η διαφορά των συστημάτων προορισμού και των πηγών ανά τακτικά χρονικά διαστήματα και αποθηκεύεται η τελευταία ημερομηνία εξαγωγής δεδομένων που πήρε μέρος ώστε να εισάγονται εγγραφές μετά από αυτήν την ημερομηνία για φόρτωση. Για μικρούς όγκους δεδομένων, η φόρτωση γίνεται με τον τρόπο που

ονομάζεται σταδιακή φόρτωση ροής ενώ για μεγάλους όγκους δεδομένων, εφαρμόζεται η σταδιακή φόρτωση παρτίδων (AWS, n.d).

Με λίγα λόγια, η φόρτωση αρχικά γίνεται με όλα τα δεδομένα να μεταφέρονται μέσα στην αποθήκη και στη συνέχεια για ενημέρωση ή διάφορες αλλαγές της αποθήκης, γίνονται κάποιες σταδιακές φορτώσεις. Η σύνοψη της διαδικασίας ETL (Extract, Transform, Load), παρουσιάζεται στην Εικόνα 3.



Εικόνα 3: Συνοπτική αναπαράσταση ETL διαδικασίας

(<https://blog.coupler.io/what-is-etl/>)

Μέσα από την περιγραφή του παραπάνω κεφαλαίου, συμπεραίνεται ότι κάθε στάδιο της διαδικασίας ETL είναι απαραίτητο και δεν πρέπει να παραληφθεί, αντίθετα είναι κύρια απαίτηση η προσοχή και υπομονή καθώς η γρήγορη προσπέραση της συγκεκριμένης διαδικασίας όχι μόνο δεν θα ωφελήσει πουθενά αλλά θα δημιουργήσει και σοβαρά λάθη σε πολλές πτυχές της δημιουργίας της αποθήκης δεδομένων και των τελικών αναλύσεων της.

3.3 Μοντέλα δεδομένων

Τα μοντέλα δομής δεδομένων αποτελούν σημαντικό κομμάτι για την ανάπτυξη της αποθήκης και την αποτελεσματική λειτουργία ερωτημάτων σε περιβάλλον SQL. Τα τρία πιο βασικά είναι τα εξής:

- Μοντέλο αστέρα: Αποτελείται από έναν βασικό πίνακα γεγονότων που συνδέεται με διαστάσεις πινάκων με την λειτουργία των ξένων κλειδιών.

- Μοντέλο χιονονιφάδας: Είναι η επέκταση του σχήματος αστέρα καθώς εκτός από τον πίνακα γεγονότων υπάρχουν κανονικοποιημένες διαστάσεις οι οποίες μπορούν επίσης να έχουν τους δικούς τους διαστασιακούς πίνακες.
- Μοντέλο γαλαξία: Θεωρείται ένα υβριδικό σχήμα που συνδυάζει τις δυνατότητες του μοντέλου του αστέρα με τις δυνατότητες του μοντέλου χιονονιφάδας. Είναι ένα εξαιρετικά σημαντικό σχήμα για επιχειρήσεις που προσπαθούν να καλύψουν πάνω από δύο επιχειρησιακές διαδικασίες ταυτόχρονα. Αυτό γιατί δίνει μεγάλη ελαστικότητα μέσω της ικανότητας να συνδέονται πολλοί πίνακες γεγονότων σε πολλές διαστάσεις πινάκων (Datapedia, 2023).

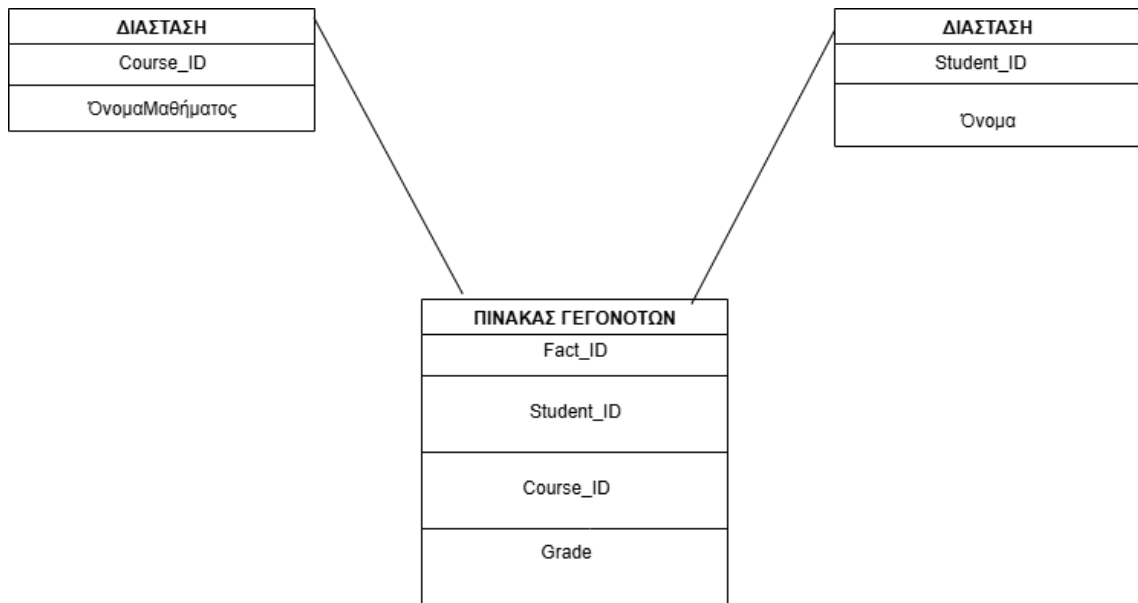
Παρακάτω θα εστιάσουμε στα σχήματα του αστέρα και της χιονονιφάδας αναδεικνύοντας πιθανά πλεονεκτήματα ή μειονεκτήματα και συγκρίσεις μεταξύ των δύο μοντέλων. Ο στόχος μέσα από αυτή την διαδικασία είναι να επέλθει ένα συμπέρασμα για το που μπορεί να πλεονεκτεί ή να υστερεί κάποιο από τα δύο και να δοθεί μια λεπτομερή ανάλυση η οποία θα αναδεικνύει για ποια χρήση είναι κατάλληλο κάθε σχήμα.

3.3.1 Μοντέλο αστέρα

Το μοντέλο του αστέρα θεωρείται το πιο γνωστό και όπως αναφέρθηκε, αποτελείται από τον κεντρικό πίνακα γεγονότων (fact table) με σύνδεση προς πολλαπλές διαστάσεις πινάκων (dimension tables) μέσω ξένων κλειδιών.

Η συσχέτιση του πίνακα γεγονότων με τις διαστάσεις είναι ένα προς πολλά. Οι διαστάσεις συνήθως περιέχουν περιγραφική πληροφορία και από ένα πρωτεύον κλειδί που είναι και το αναγνωριστικό του πίνακα.

Τα κύρια πλεονεκτήματα του συγκεκριμένου μοντέλου είναι η μείωση πινάκων στη βάση δεδομένων και ο μικρός αριθμός συσχετίσεων ανάμεσα στους πίνακες. Επιπρόσθετα, συμβάλλει στην απλότητα των ερωτημάτων μέσα στην αποθήκη καθώς οι συζεύξεις θα είναι επίσης λίγες (Beril, 2005). Αυτά προσφέρουν ένα απλό και κατανοητό σχήμα, όπως φαίνεται και στην Εικόνα 4, μειώνοντας την πολυπλοκότητα των ενεργειών για την δημιουργία της αποθήκης δεδομένων.



Εικόνα 4: Παράδειγμα σχήματος αστέρα

Από την άλλη μεριά, εάν οι διαστάσεις είναι δυναμικές και πολύπλοκες, θα προκύψουν και κάποια βασικά μειονεκτήματα που μπορούν να δυσκολέψουν την υλοποίηση της αποθήκης ή ακόμα και να αναγκάσουν σε αλλαγή άλλου σχηματικού μοντέλου, κάτι που είναι σημαντικό να αναγνωριστεί έγκαιρα για να μην σπαταληθεί άσκοπα ο χρόνος σε λάθος σχήμα. Τα κύρια μειονεκτήματα είναι τα εξής:

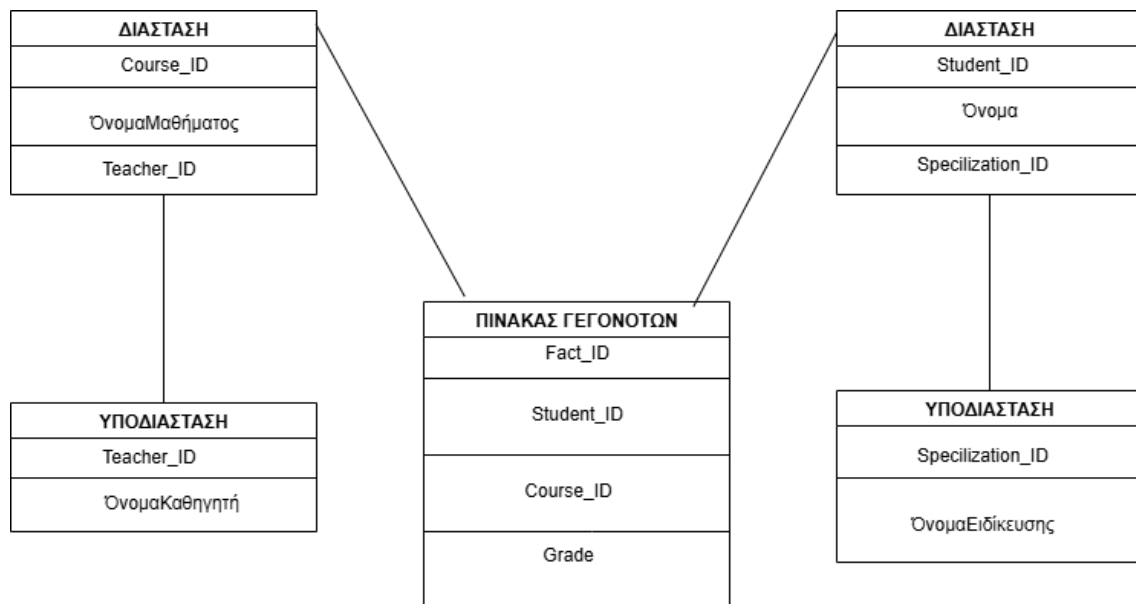
- Μπορεί τα δεδομένα να μην σταθούν συνεπή.
- Χρησιμοποιεί μη κανονικοποιημένα δεδομένα.
- Μη ικανό να υποστηρίξει πολλά προς πολλά συσχέτιση.
- Μπορεί να χρησιμοποιήσει περισσότερο χώρο από όσο πραγματικά χρειάζεται διότι οι πίνακες είναι μεγάλοι (Patairiya, 2024).

Συμπερασματικά, το μοντέλο αστέρα είναι το πιο απλό και κατανοητό και εάν οι συνθήκες δεν είναι μεγάλης πολυπλοκότητας, είναι το κατάλληλο για την ανάπτυξη της αποθήκης δεδομένων.

3.3.2 Μοντέλο χιονονιφάδας

Η ονομασία του συγκεκριμένου σχήματος προκύπτει, όπως φαίνεται και από την Εικόνα 5, από τους πρόσθετους πίνακες των διαστάσεων του πίνακα γεγονότων. Το σχήμα αποτελείται από κεντρικό πίνακα γεγονότων και διαστάσεις όπως και το σχήμα αστέρα με την διαφορά να είναι οι μικρότεροι πίνακες των διαστάσεων όπου ορίζουν μια δομή που

μοιάζει με αυτή μιας χιονονιφάδας. Ένας βασικός λόγος επιλογής του αντί του αστέρα, είναι η ικανότητα του να προσφέρει κανονικοποίηση των δεδομένων.



Εικόνα 5: Παράδειγμα μοντέλου χιονονιφάδας

Παρακάτω θα αναφερθούν τα πλεονεκτήματα και μειονεκτήματα του μοντέλου χιονονιφάδας, όπως αναφέρει ο Patairiya, για να δοθεί μια σαφής εικόνα των ικανοτήτων του.

Πλεονεκτήματα:

- Δυνατότητα υποστήριξης συσχετίσεων πολλά προς πολλά.
- Χρησιμοποιεί λιγότερο αποθηκευτικό χώρο λόγω κανονικοποίησης των δεδομένων.
- Μπορεί να δημιουργήσει πιο πολύπλοκα ερωτήματα.

Μειονεκτήματα:

- Πολύπλοκη δομή.
- Δυσκολία στην δημιουργία και κατανόηση του.
- Οι πολλαπλές διαστάσεις πινάκων περιπλέκουν την διαδικασία δημιουργίας ερωτημάτων.

Εκτός από την σχηματική δομή, υπάρχουν και άλλες διαφορές ανάμεσα στα δύο μοντέλα οι οποίες παρατίθενται στον Πίνακας 4:

Πίνακας 4: Διαφορές αστέρα-χιονονιφάδας

Μοντέλο του αστέρα	Μοντέλο χιονονιφάδας
Μοντέλο top-down	Μοντέλο Bottom-up
Λιγότερος χρόνος εκτέλεσης ερωτημάτων	Περισσότερος χρόνος εκτέλεσης ερωτημάτων
Χρησιμοποιεί περισσότερο χώρο	Χρησιμοποιεί λιγότερο χώρο
Μη κανονικοποίηση	Χρησιμοποιεί κανονικοποίηση και μη κανονικοποίηση
Χαμηλός βαθμός πολυπλοκότητας στα ερωτήματα	Υψηλός βαθμός πολυπλοκότητας ερωτημάτων
Μικρότερος αριθμός ξένων κλειδιών	Μεγάλος αριθμός ξένων κλειδιών
Εύκολο στην κατανόηση	Πιο δύσκολο να κατανοηθεί

Από τα παραπάνω, παρατηρείται ότι για έναν αρχάριο το μοντέλο αστέρα θα είναι ασφαλώς πιο προσιτό (αρκεί να μπορεί να χρησιμοποιηθεί κατάλληλα) και ότι θα είναι λιγότερο χρονοβόρα η διαδικασία. Από την άλλη, αν η πολυπλοκότητα είναι μεγάλου βαθμού, το μοντέλο της χιονονιφάδας θα φανεί πιο χρήσιμο, αρκεί να μπορεί ο χρήστης να αντιμετωπίσει τις δυσκολίες.

Συμπεραίνουμε ότι κάθε μοντέλο είναι σημαντικό και παρά το γεγονός ότι υπάρχουν μειονεκτήματα και διάφορες δυσκολίες του καθενός, εάν επιλεγθεί το κατάλληλο και υπό τις σωστές συνθήκες, τότε το αποτέλεσμα θα είναι τουλάχιστον ικανοποιητικό ως προς την κατασκευή του σχηματικού μοντέλου, τα ερωτήματα και την τελική ανάπτυξη της αποθήκης.

3.4 Διαφορές αποθηκών δεδομένων, βάσεων δεδομένων και λίμνών δεδομένων

Σε αυτήν την ενότητα θα αναλυθούν εκτός της αποθήκης δεδομένων, δύο ακόμη βασικές έννοιες για την αποθήκευση δεδομένων. Η βάση δεδομένων και η λίμνη δεδομένων. Θα εστιάσουμε στην ανάδειξη των κύριων διαφορών και λειτουργιών ανάμεσα σε μια αποθήκη δεδομένων και σε μια κλασσική βάση δεδομένων αλλά και στις διαφορές αποθήκης δεδομένων με μια λίμνη δεδομένων. Με αυτόν τον τρόπο ανάλυσης θα δοθεί έμφαση στις

αποθήκες και θα προκύψουν συμπεράσματα για το πότε είναι κατάλληλη κάθε μια από τις τρείς περιπτώσεις.

Ορισμός βάσης δεδομένων:

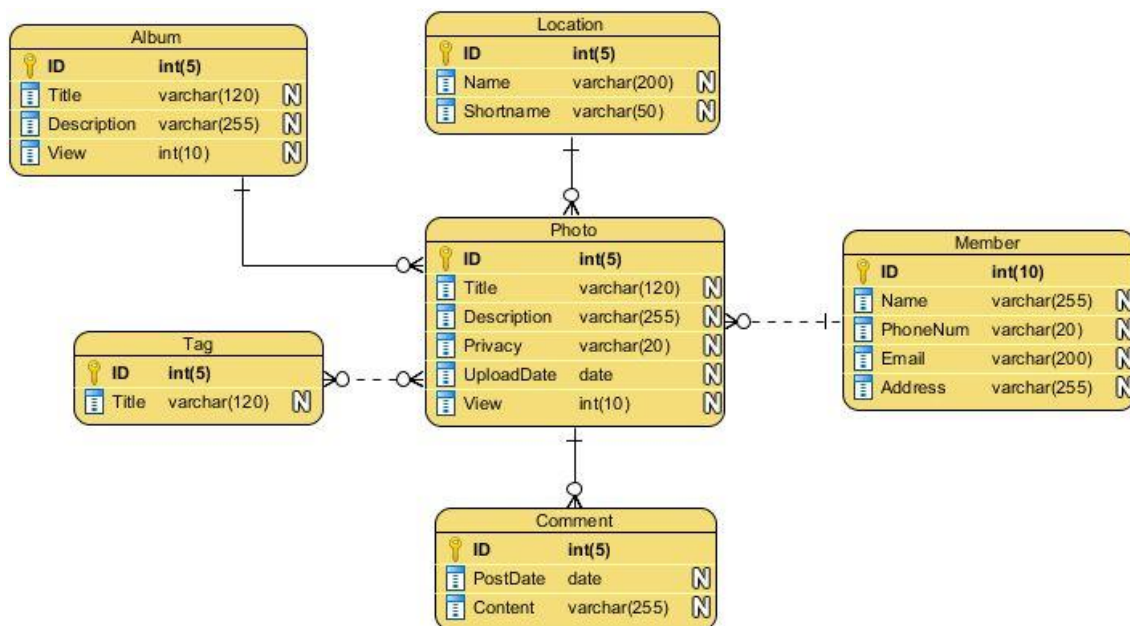
Πριν από τις βάσεις δεδομένων υπήρχαν πληροφοριακά συστήματα που αποθήκευαν τα δεδομένα. Αυτό άλλαξε με την πάροδο του χρόνου καθώς ο όγκος των δεδομένων αυξανόταν με ραγδαία ταχύτητα. Έτσι, η ανάγκη για μεγαλύτερο αποθηκευτικό χώρο συνέβαλε στην δημιουργία των πρώτων βάσεων δεδομένων (Elmasri & Navathe, 2016).

Σύμφωνα με το Oracle, ως βάση δεδομένων ορίζεται η οργανωμένη συλλογή δεδομένων ή πληροφορίας η οποία θεωρείται δομημένη και μπορεί να αποθηκεύεται σε ένα ηλεκτρονικό υπολογιστικό σύστημα.

Επίσης, αναφέρει ότι συστήματα βάσεων δεδομένων ονομάζονται όσα περιέχουν τα δεδομένα και το DBMS (Database Management system ή σύστημα Διαχείρισης Βάσης Δεδομένων). Τα DBMS είναι τα συστήματα που παρέχουν στους τελικούς χρήστες δυνατότητες συλλογής και επεξεργασίας της πληροφορίας, Δημοφιλή συστήματα είναι το MySQL, η Oracle και το SSMS (SQL Server Management Studio) το οποίο και χρησιμοποιούμε. Τέλος, κάνει αναφορά στο ότι οι βάσεις δεδομένων διακρίνονται και σε διάφορους τύπους:

- Σχεσιακή (Relational).
- Αντικειμενοστραφής.
- NoSQL.
- Αποθήκη δεδομένων κ.α.

Για την μοντελοποίηση των δεδομένων, εφαρμόζεται το μοντέλο οντοτήτων-συσχετίσεων (Entity-Relationship Model ή ER). Ως οντότητα ορίζεται το αντικείμενο που εκδηλώνεται στο περιβάλλον και ως συσχέτιση ορίζεται η σχέση που δημιουργείται με αυτό το αντικείμενο μέσα στην βάση δεδομένων. Όπως και στα μοντέλα μιας αποθήκης δεδομένων έτσι και εδώ χρησιμοποιούνται συσχετίσεις οι οποίες μπορεί να είναι 1:1, 1:M (1 to Many ή 1 προς πολλά), M:M (Many to Many ή πολλά προς πολλά) (Lucidchart, n.d). Στην Εικόνα 6, γίνεται αναπαράσταση των προαναφερόμενων για το διάγραμμα οντοτήτων-συσχετίσεων. Εφαρμογές όπως η draw.io είναι κατάλληλες για τη δημιουργία ενός λογικού διαγράμματος που θα χρησιμοποιηθεί για την σχεσιακή βάση δεδομένων.



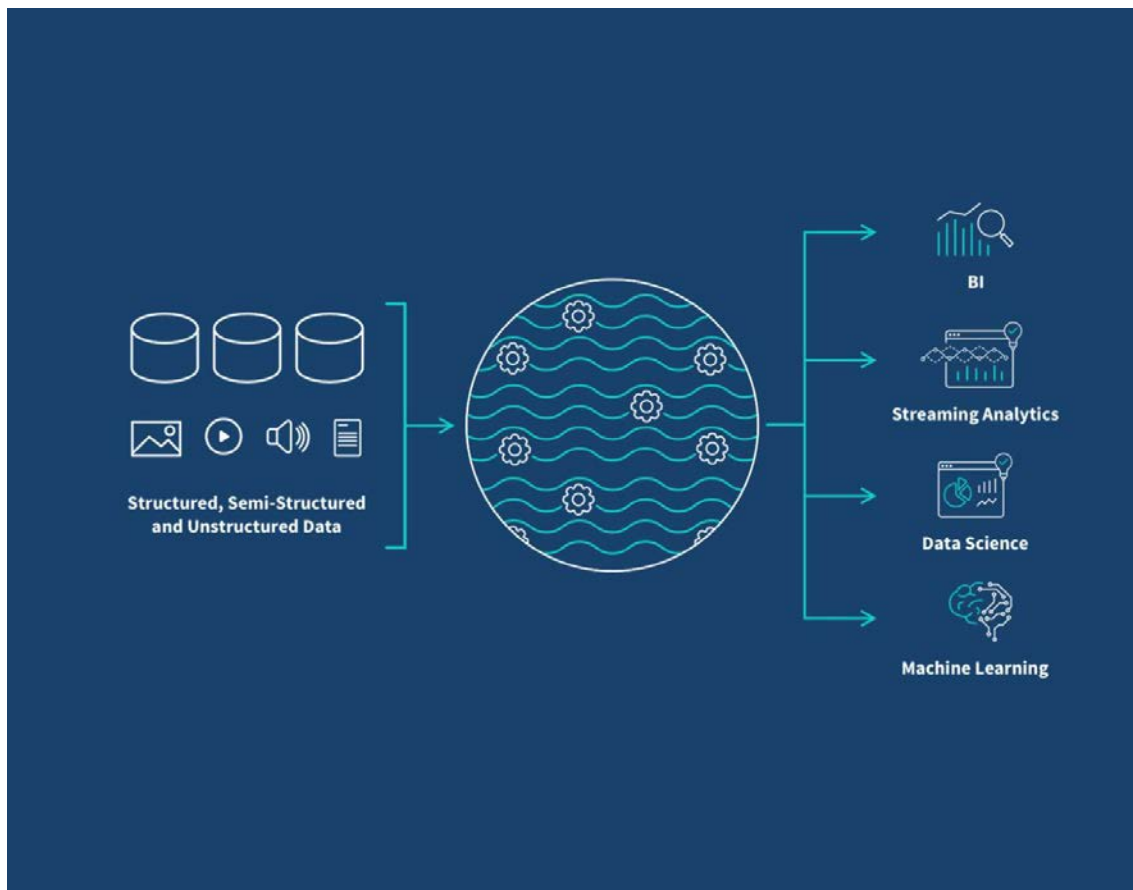
Εικόνα 6: E-R Diagram παράδειγμα

(<https://www.pinterest.com/pin/614459942880927462/>)

Ορισμός λίμνης δεδομένων:

Όπως οι αποθήκες δεδομένων έτσι και οι λίμνες δεδομένων, δημιουργήθηκαν γιατί η ανάγκη για αποθήκη μεγάλων όγκων δεδομένων αυξήθηκε ραγδαία. Οι λίμνες δεδομένων ή αλλιώς data lakes, παρέχουν μεγάλους αποθηκευτικούς χώρους που βοηθάνε τους χρήστες να αποθηκεύουν τα δεδομένα τους σε όποια μορφή και αν βρίσκονται. Είτε δηλαδή είναι ακατέργαστα είτε δομημένα ή μη δομημένα. Η ονομασία δόθηκε από το γεγονός ότι τα δεδομένα αποθηκεύονται διάσπαρτα και είναι μη συσχετισμένα μεταξύ τους μέσα σε μια αποθήκη χωρίς να υπόκεινται σε κάποια επεξεργασία. Υπάρχει και η περίπτωση τα δεδομένα μιας λίμνης να μεταφερθούν με την λεγόμενη ροή προς τα κάτω (bottom down) σε μια αποθήκη δεδομένων ώστε να επεξεργαστούν πριν την τελική χρήση (Nambiar & Mundra, 2022).

Μια αντιπροσωπευτική εικόνα είναι η Εικόνα 7, όπου αναπαρίσταται μια λίμνη ωμών δεδομένων και στην συνέχεια μπορούν να γίνουν οπτικοποιήσεις, αναλύσεις κ.ά. για διάφορες επιχειρησιακές μελέτες.



Εικόνα 7: Data Lake

(<https://www.qlik.com/us/data-lake>)

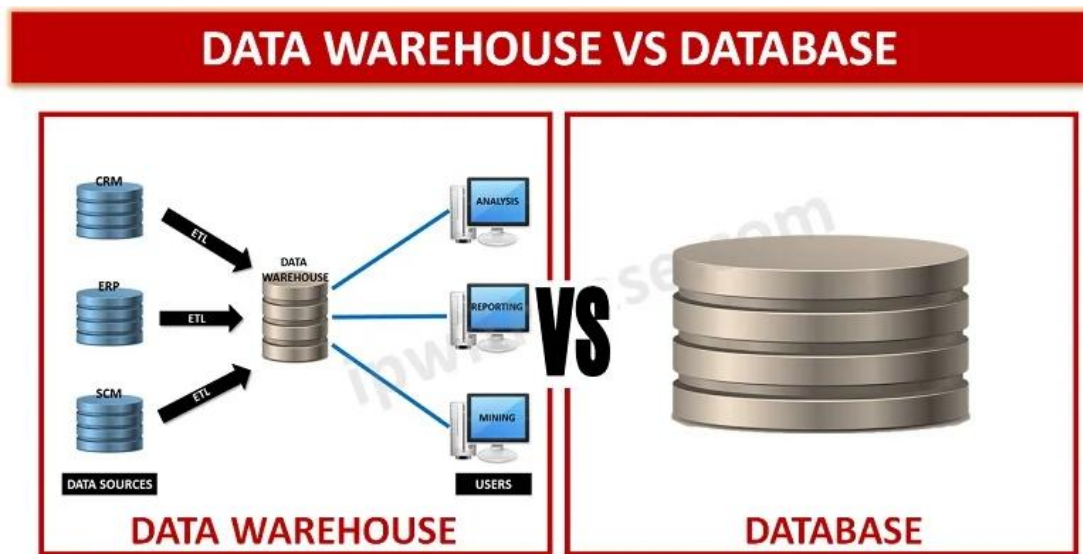
Οι δύο αυτές περιπτώσεις και η αποθήκη δεδομένων, είναι τρεις εξαιρετικά διαφορετικοί τρόποι για αποθήκευση δεδομένων και αυτό καθιστά πολύτιμη την ανάδειξη των διαφορών τους.

3.4.1 Αποθήκες δεδομένων vs βάσεις δεδομένων

Οι αποθήκες και οι βάσεις δεδομένων είναι αναγκαία εργαλεία αποθήκευσης για επιχειρήσεις και οργανισμούς καθώς η ανάγκη για αποθηκευτικούς χώρους μεγάλων ποσοτήτων πληροφοριών είναι υψηλή. Στην πράξη όμως είναι δύο έννοιες με πολύ διαφορετικά χαρακτηριστικά και λειτουργίες.

Αναλυτικά, μια βασική διαφορά που ενδιαφέρει σημαντικά τις εταιρείες για να ξεχωρίσουν ποια είναι η κατάλληλη επιλογή για τα θέλω τους, είναι η εξής. Μια αποθήκη δεδομένων μπορεί να αποθηκεύσει μικρό όγκο δεδομένων σε σχέση με μια βάση και μόνο για έναν συγκεκριμένο σκοπό. Αντίθετα, οι βάσεις δεδομένων μπορούν να φιλοξενήσουν μεγάλους όγκους πληροφορίας χωρίς συγκεκριμένο σκοπό. Αξίζει να αναφερθεί επίσης

ότι τα δεδομένα σε μια βάση αποθηκεύονται με την λογική να παραμείνουν εκεί για μια μακροχρόνια περίοδο ενώ τα αποθετήρια είναι κατάλληλα για ενεργή και βραχυχρόνια ανάλυση. Αυτό γιατί ο σκοπός της αποθήκης είναι η ενεργή ανάλυση των πληροφοριών και όχι οι πλήρεις μακροχρόνιες αναφορές. Τέλος, τα δεδομένα συλλογής για την αποθήκη δεδομένων χαρακτηρίζονται ως αντικειμενοστραφή γιατί η οργάνωση είναι αναγκαίο να γίνεται με βάση τα αντικείμενα ενδιαφέροντος. Από την άλλη μεριά, οι βάσεις συλλέγουν δεδομένα προσανατολισμένα προς την εφαρμογή (application-oriented) για τις αναφορές τους (Coursera Staff, 2025). Στην Εικόνα 8 αναπαρίσταται σχηματικά η αποθήκη και η βάση δεδομένων.



Εικόνα 8: Αποθήκη δεδομένων - Βάση δεδομένων

(<https://networkinterview.com/data-warehouse-vs-database-know-the-difference/>)

3.4.2 Αποθήκες δεδομένων vs λίμνες δεδομένων

Μπορεί και οι δύο έννοιες να είναι κοινές ως προς τη γενική λειτουργία τους, δηλαδή την αποθήκευση δεδομένων, αλλά στη πραγματικότητα διαφέρουν κατά πολύ σε πολλές πτυχές τους. Οι διαφορές τους θα αναλυθούν στον Πίνακα 5. Ο παρακάτω Πίνακας 5, είναι εμπνευσμένος από τα λεγόμενα των Nambiar και Mundra (2022) που παρατίθεται στην βιβλιογραφία.

Πίνακας 5: Διαφορές μεταξύ αποθήκης και λίμνης δεδομένων

Διαφορές		
Ως προς	Αποθήκη δεδομένων	Λίμνη δεδομένων
Χώρο	Οικονομικό σε χώρο	Ιδανικό για μεγάλες χωρητικότητες
Σκοπό χρήσης	Συγκεκριμένο σκοπό	Ο σκοπός δεν χρειάζεται να είναι συγκεκριμένος
Επίπεδο δυσκολίας	Εύκολο στη χρήση	Δυσκολία ανάλυσης
Ευελιξία	Δυσκολία και μεγάλο κόστος στην αναδιαμόρφωση	Μεγάλη ευελιξία λόγω μη δομημένης κατασκευής
Ασφάλεια	Λόγω της αυστηρής αρχιτεκτονικής είναι πολύ ασφαλής	Λόγω της πολύ ανοιχτής και εκτεθειμένης δομής της αρχιτεκτονικής, η ασφάλεια είναι χαμηλότερου επιπέδου

Επιπρόσθετα, σύμφωνα με το Bryteflow, οι λίμνες δεδομένων εντός μιας οργάνωσης χρησιμοποιούν την διαδικασία ELT (Extract, Load, Transform) και όχι την ETL (Extract, Transform, Load) όπως συνηθίζεται στις περισσότερες αποθήκες δεδομένων. Τέλος, σε μια λίμνη δεδομένων οι χρήστες είναι συνήθως επιστήμονες δεδομένων καθώς τα δεδομένα που θα αναλύσουν χρησιμοποιούνται για την παροχή μοντέλων τεχνητής νοημοσύνης και μηχανικής μάθησης. Αντίθετα, στις αποθήκες δεδομένων οι χρήστες είναι συνήθως αναλυτές επιχείρησης που θα εκμεταλλευτούν τα παρεχόμενα δεδομένα για επιχειρησιακές μελέτες και αναλύσεις.

Συνοψίζοντας, αποθήκες, βάσεις και λίμνες δεδομένων αποτελούν κρίσιμα εργαλεία για όλων των ειδών οργανισμών και επιχειρήσεων στον σύγχρονο κόσμο της τεχνολογίας. Η επιλογή εργαλείου κρίνεται από τον ίδιο τον οργανισμό ώστε να ωφεληθεί πλήρως από τις λειτουργίες και τα πλεονεκτήματα που θα προσφέρει εκείνο.

4 Σχεδίαση και υλοποίηση αποθήκης δεδομένων

Το κομμάτι της σχεδίασης και υλοποίησης της αποθήκης δεδομένων που θέλουμε να αναπτύξουμε είναι το πιο σημαντικό καθώς προϋποθέτει μεγάλη προσοχή και λεπτομέρεια. Αυτό γιατί η διαδικασία αποτελείται από απαραίτητα βήματα όπως η δημιουργία του λογικού σχήματος, έπειτα του star schema και η εφαρμογή της διαδικασίας ETL.

4.1 Επιλογή μεταβλητών αποθήκης

Η επιλογή μεταβλητών είναι ένα βασικό βήμα για την ανάλυση και είναι ιδιαίτερα αναγκαίο να δοθεί έμφαση στο να ξεκαθαριστούν οι ουσιαστικές και χρήσιμες πληροφορίες. Είναι το πρώτο βήμα και χρήζει ιδιαίτερης προσοχής καθώς με αυτές τις μεταβλητές θα πορευτώ στη συνέχεια με διαφορές προσαρμογές και επεξεργασίες που θα υποστούν μέχρι και την τελική ανάλυση. Οι μεταβλητές και τα δεδομένα συλλέχθηκαν από την NOAA Storm Events Database (NOAA National Centers for Environmental Information, 2025). Περιέχουν χρονικές, γεωγραφικές και άλλων πιο ειδικών ειδών μεταβλητών όπως μέτρησης έντασης φαινομένων και ονομαστικές για τις διάφορες πολιτείες και περιοχές της Αμερικής.

Οι στόχοι μέσα από αυτές τις επιλογές είναι:

- Η εύκολη ομαδοποίηση των δεδομένων
- Η δυνατότητα χωρικών συγκρίσεων
- Η χρήσιμη και κατανοητή απεικόνιση των κλιματικών φαινομένων μέσω Power BI
- Η δυνατότητα συγκρίσεων με βάση χρονικές περιόδους και ούτω καθεξής

Παρακάτω θα αναφερθούν οι διάφοροι τύποι μεταβλητών που επιλέχθηκαν. Αρχικά, έχουμε τις μεταβλητές μοναδικής αναγνώρισης EPISODE_ID και EVENT_ID.

1. EPISODE_ID: Είναι ένα μοναδικό αναγνωριστικό για μια μόνο καταγραφή κλιματικού φαινομένου.

2. **EVENT_ID**: Μοναδικό για ένα συγκεκριμένο γεγονός που συμβαίνει εντός επεισοδίου.

Έπειτα, χρησιμοποιούνται χρονικές μεταβλητές όπως **YEAR**, **MONTH_NAME**, **CZ_TIMEZONE** και μια μεταβλητή που συνδυάζει ημερομηνία και χρόνο, η **BEGIN** και η **END_DATETIME**.

1. **YEAR**: Είναι το έτος που το φαινόμενο παίρνει μέρος
2. **MONTH_NAME**: Ο μήνας του έτους που ξεκινάει το φαινόμενο
3. **CZ_TIMEZONE**: Η χρονική ζώνη της Αμερικής του φαινομένου
4. **BEGIN DATETIME**: Η ακριβής ώρα και ημερομηνία έναρξης του φαινομένου/γεγονότος
5. **END DATETIME**: Η ακριβής ώρα και ημερομηνία τερματισμού του φαινομένου/γεγονότος

Επιπρόσθετα, απαραίτητη είναι η χρήση γεωγραφικών μεταβλητών που δείχνουν τις συντεταγμένες και τους τόπους που παίρνουν μέρος τα φαινόμενα.

1. **BEGIN LOCATION**
2. **END LOCATION**
3. **BEGIN LATITUDE**
4. **END LATITUDE**
5. **BEGIN LONGTITUDE**
6. **END LONGTITUDE**

Αναγκαίες είναι και οι μεταβλητές απώλειας περιουσίας ή καλλιέργειας και οι μεταβλητές που δείχνουν αριθμό θανάτων και τραυματισμών που προκάλεσε κάποιο κλιματικό φαινόμενο.

1. **DAMAGE_PROPERTY**: Η ζημιά περιουσίας σε δολάρια
2. **DAMAGE_CROPS**: Η ζημιά καλλιέργειας σε δολάρια
3. **INJURIES_DIRECT**: Οι τραυματισμοί κατά την διάρκεια ενός γεγονότος
4. **DEATHS_DIRECT**: Οι απώλειες ανθρώπων κατά την διάρκεια ενός γεγονότος

Τέλος, μεγάλης σημασίας και ενδιαφέροντος, θεωρούνται και οι άλλων ειδών μεταβλητές όπως οι ονομαστικές των πολιτειών και των περιοχών (**STATE**, **COUNTY**), της έντασης των φαινομένων (**MAGNITUDE**, **MAGNITUDE_TYPE**), του τύπου φαινομένου (**EVENT_TYPE**) και μιας περιγραφικής μεταβλητής (**EPISODE_NARRATIVE**) που επέλεξα.

1. STATE: Όνομα πολιτείας.
2. COUNTY: Όνομα περιοχής της πολιτείας.
3. MAGNITUDE: Ένταση φαινομένου.
4. MAGNITUDE_TYPE: Τύπος έντασης
5. EVENT_TYPE: Τύπος γεγονότος όπως πλημμύρα, πυρκαγιά, χιονοθύελλα ή καταιγίδα.
6. EPISODE_NARRATIVE: Περιγραφή συγκεκριμένου επεισοδίου με κείμενο που αναλύει αξιοσημείωτα περιστατικά (π.χ. μεγάλη κατολίσθηση, πλημμύρα παρέσυρε δεκάδες αυτοκίνητα κλπ.) που πήραν μέρος εντός του επεισοδίου.

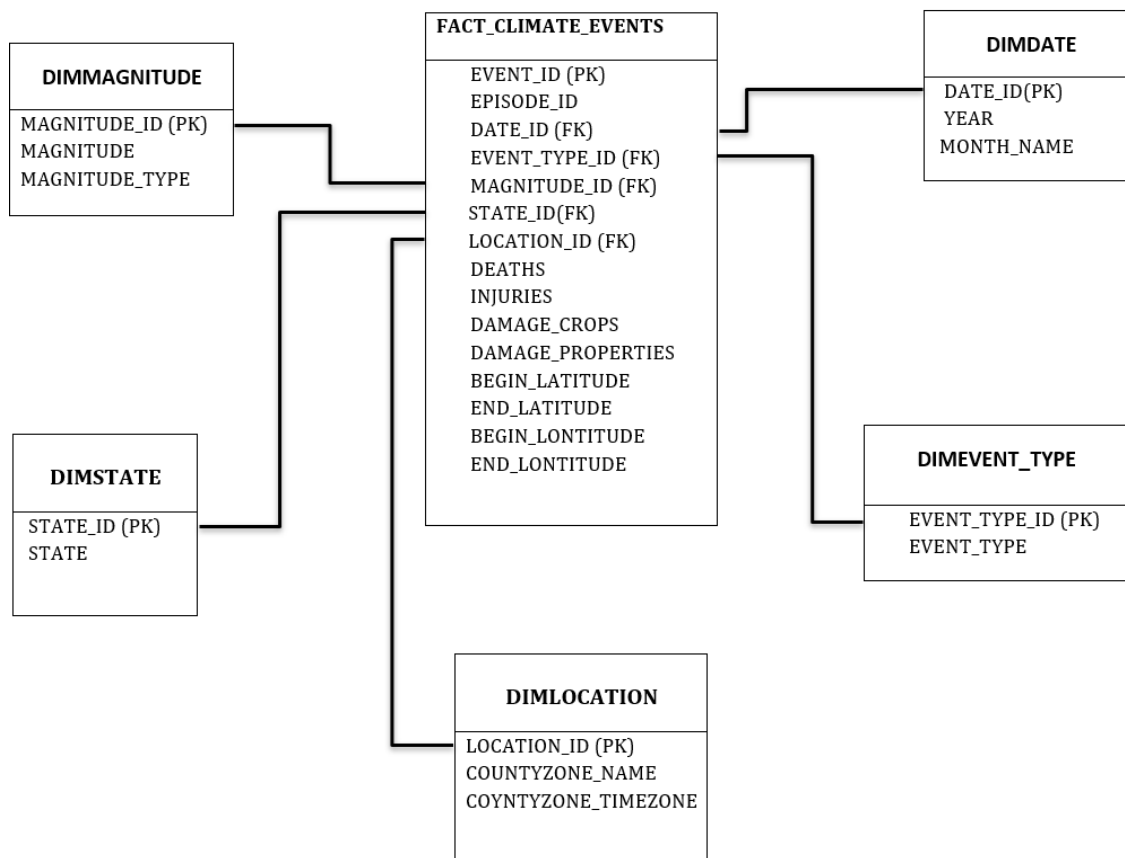
Συμπερασματικά, η επιλογή μεταβλητών για την αποθήκη δεδομένων στοχεύει στην πλήρη ανάλυση και ανάδειξη χρήσιμων πληροφοριών. Για παράδειγμα, οι γεωγραφικές μεταβλητές είναι με τέτοιο τρόπο επιλεγμένες που μπορούν να χρησιμοποιηθούν κατάλληλα για χαρτογράφηση φαινομένων ή παρακολούθηση περιοχών ευάλωτων σε καιρικούς κινδύνους. Κάθε κατηγορία όπως και αυτή των γεωγραφικών, συμβάλλουν σε μεγάλο βαθμό στην ευελιξία του συστήματος και όπως είναι φυσικό και στην δημιουργία του λογικού σχήματος (logical schema) που είναι το επόμενο βήμα.

4.2 Δημιουργία Λογικού Σχήματος

Επόμενο εξίσου σημαντικό στάδιο είναι αυτό της υλοποίησης του λογικού σχήματος. Είναι απαραίτητο να σχηματιστεί καθώς βοηθάει στην κατανόηση των δεδομένων που διαθέτουμε και επομένως συμβάλλει στη δημιουργία της βάσης δεδομένων και του star schema μέσα σε αυτή. Το λογικό σχήμα είναι μια αναπαράσταση, στη συγκεκριμένη περίπτωση ενός star schema με πίνακα γεγονότων (fact table) και πίνακες διαστάσεων (dimension tables). Το μοντέλο αστέρα είναι ιδανικό για την περίπτωση και σε αντίθεση με το snowflake schema δεν περιπλέκει την ανάλυση και τα δεδομένα καθώς είναι εύκολο και γρήγορο στον σχεδιασμό του (Kimball & Ross, 2013). Αποτελείται συνοπτικά από τα παρακάτω:

- Έναν πίνακα γεγονότων που περιέχει ως ξένα κλειδιά τα πρωτεύον κλειδιά των υπολοίπων πινάκων διαστάσεων, το πρωτεύον κλειδί του, στη συγκεκριμένη περίπτωση το EVENT_ID και διάφορα σημαντικά μεγέθη.
- Έξι πίνακες διαστάσεων: χρόνου, τοποθεσίας, τύπου γεγονότος, επεισοδίου, ημερομηνίας και έντασης με περιεχόμενο αντίστοιχο των ονομασιών τους.

Στην Εικόνα 9 φαίνεται το αρχικό λογικό σχήμα:



Εικόνα 9: Λογικό σχήμα

Επομένως, το λογικό σχήμα δημιουργείται κυρίως για λόγους κατανόησης ώστε η εφαρμογή του μοντέλου του αστέρα για την αποθήκη δεδομένων, να είναι πιο εύκολη.

4.3 Διαδικασία ETL

Αρχικά, η πηγή των δεδομένων Storm Events, λόγω των πολλών ετών στα οποία παίρνει μέρος η ανάλυση, προσέφερε 1.023.186 εγγραφές οι οποίες υπέστη καθαρισμό μέσω του εργαλείου Microsoft Excel. Στην Εικόνα 10 φαίνεται ένα στιγμιότυπο από τον επίσημο ιστότοπο NOAA Storm Events Database

Storm Events Database

Data Access

- Search
- Bulk Data Download (CSV)
- Storm Data Publication

Documentation

- Database Details
- Version History
- Storm Data FAQ
- NOAA's NWS Documentation
- Tornado EF Scale

External Resources

- NOAA's SPC Reports
- NOAA's SPC WCM Page
- NOAA's NWS Damage Assessment Toolkit
- NOAA's Tsunami Database
- ESRI/FEMA Civil Air Patrol Images
- SHELDUS
- USDA Cause of Loss Data

Storm Events Database

Search Results for All Counties in Alabama

Event Types: Flood

51 events were reported between 04/01/2009 and 04/30/2011 (760 days)

Summary Info:

Number of County/Zone areas affected:	20
Number of Days with Event:	22
Number of Days with Event and Death:	0
Number of Days with Event and Death or Injury:	0
Number of Days with Event and Property Damage:	12
Number of Days with Event and Crop Damage:	0
Number of Event Types reported:	1

Column Definitions:
'Mag': Magnitude, 'Dth': Deaths, 'Inj': Injuries, 'PrD': Property Damage, 'CrD': Crop Damage

Click on **Location** below to display details.
Available Event Types have changed over time. Please refer to the [Database Details](#) for more information.

Sort By: Date/Time (Oldest) ▼

Data Export: (current results)

 [CSV Download](#) / [Documentation](#)

Location	County/Zone	St.	Date	Time	T.Z.	Type	Mag	Dth	Inj	PrD	CrD
Totals:								0	0	3.880M	0.00K
FLETCHER CHAPEL	MADISON CO.	AL	04/02/2009	19:30	CST-6	Flood		0	0	0.00K	0.00K
SPEAKE	LAWRENCE CO.	AL	04/02/2009	21:30	CST-6	Flood		0	0	0.00K	0.00K
WATERLOO	LAUDERDALE CO.	AL	05/01/2009	19:31	CST-6	Flood		0	0	0.00K	0.00K
DECATUR	MORGAN CO.	AL	05/01/2009	20:30	CST-6	Flood		0	0	0.00K	0.00K
WHEELER LAKE NORTH	LIMESTONE CO.	AL	05/01/2009	22:00	CST-6	Flood		0	0	0.00K	0.00K
OLEANDER	MARSHALL CO.	AL	05/06/2009	07:15	CST-6	Flood		0	0	5.00K	0.00K
DOUGLAS	MARSHALL CO.	AL	05/06/2009	09:00	CST-6	Flood		0	0	5.00K	0.00K

Εικόνα 10: Στιγμιότυπο αναζήτησης στον ιστότοπο της NOAA Storm Events Database

Κύριες διαδικασίες που απαιτούσαν αρκετό χρόνο είναι οι εξής:

- **DAMAGE PROPERTIES, DAMAGE CROPS:** Οι απώλειες περιουσίας και καλλιέργειας αρχικά είχαν την μορφή κειμένου (π.χ. "15K", "2.5M", 1B) και μετατράπηκαν σε αριθμούς που παραπέμπουν σε χρηματικά ποσά με νόμισμα το δολάριο. Για παράδειγμα το K έγινε επί 1.000, το M επί 1.000.000 και το B επί 1.000.000.000 καθώς με το K, M, και B αναπαριστούσαν χιλιάδες, εκατομμύρια και δισεκατομμύρια σε δολάρια αντίστοιχα. Στη συνέχεια, αφαιρέθηκαν τα σύμβολα του δολαρίου και τα κόμματα για την ανάγκη της ανάλυσης.
- **MAGNITUDE:** Το πεδίο έντασης των φαινομένων, περιείχε εγγραφές επίσης σε μορφή κειμένου (π.χ. "MG" "EG") που έπρεπε να μετατραπούν με βάση διάφορες κλίμακες έντασης όπως η Richter για σεισμούς ή η κλίμακα Advanced Fujita για την μέτρηση έντασης ανεμοστρόβιλων.
- **BEGIN/END LATITUDE, BEGIN/END LONGTITUDE:** Στην συγκεκριμένη περίπτωση, οι γραμμές αυτών των στηλών στο Excel, εμφανίζονται σε πρώτο στάδιο με λανθασμένη μορφή που δεν αντιστοιχεί σε συντεταγμένες εντός ορίων των ΗΠΑ (δηλαδή 15 έως 75 μοίρες γεωγραφικό πλάτος και -180 έως -60 μοίρες

γεωγραφικό μήκος). Οι μετατροπές που πήραν μέρος (με βάση τον αριθμό ψηφίων) για να επιτευχθεί αυτό είναι οι εξής:

- ο <100: Διατήρηση ως έχει
- ο 3 ψηφία: Διαίρεση με το 10
- ο 4 ψηφία: Διαίρεση με το 100
- ο 5+ ψηφία: Διαίρεση με το 10.000

Με αυτήν την τακτική, δημιουργήθηκε η ανάγκη για ένα νέο πεδίο με έγκυρες συντεταγμένες που είναι εντός ορίων των ΗΠΑ με όνομα **COORDS_VALID**.

4.4 Υλοποίηση μοντέλου αστέρα

Η χρήση του σχήματος αστέρα ήταν ένα κρίσιμο βήμα στην κατασκευή της αποθήκης δεδομένων, καθώς καθόρισε τον γενικό σχεδιασμό για την αποτελεσματική αποθήκευση και αναζήτηση κλιματικών δεδομένων. Ενώ το σχήμα αστέρα είχε σχεδιαστεί αρχικά με βάση το λογικό σχήμα, η πραγματική εφαρμογή του απαιτούσε πολλές σχεδιαστικές τροποποιήσεις, για την επίτευξη της μέγιστης απόδοσης.

Το σχήμα αστέρα αποτελείται από έναν κεντρικό πίνακα δεδομένων, ο οποίος περιέχει τις βασικές μετρήσεις για κάθε καιρικό φαινόμενο (π.χ. πλημμύρες, τυφώνες, πυρκαγιές και χιονοθύελλες), ο οποίος περιβάλλεται από πίνακες διαστάσεων με δεδομένα όπως τοποθεσία, ώρα, τύπος φαινομένου και ένταση.

Ο σχεδιασμός του σχήματος αρχικά καθορίστηκε με βάση τη δομή των στηλών του συνόλου δεδομένων NOAA Storm Events. Ωστόσο, μετά την εισαγωγή και τον έλεγχο των ακατέργαστων δεδομένων, προέκυψαν διάφορα προβλήματα. Μεταβλητές ήταν περιττές ή είχαν ακατάλληλη μορφοποίηση. Πριν από την τελική καθιέρωση του σχήματος, έπρεπε να πραγματοποιηθεί καθαρισμός, κανονικοποίηση και μετατροπή των δεδομένων. Επιπρόσθετα, πραγματοποιήθηκαν αλλαγές σε τύπους δεδομένων πεδίων. Αντίστοιχα, σε άλλα πεδία χρειάστηκε να μειωθεί ο χώρος καθώς δεν συνέβαλλε στην ανάγκη για εξοικονόμηση χώρου της αποθήκης.

4.4.1 Πίνακας Γεγονότων

Το κύριο στοιχείο του star schema είναι ο πίνακας γεγονότων του ή αλλιώς fact table. Σε αυτόν, αποθηκεύονται οι ποσοτικές πληροφορίες σχετικά με τα καιρικά φαινόμενα, όπως είναι ο αριθμός θανάτων ή τραυματισμών και οι οικονομικές απώλειες που προκλήθηκαν στις ανθρώπινες περιουσίες και στις καλλιέργειες.

Τα χαρακτηριστικά είναι τα εξής:

- Τεχνητό πρωτεύον κλειδί ClimateEventKey
- Ξένο κλειδί EventTypeKey
- Ξένο κλειδί MagnitudeKey
- Ξένο κλειδί StateKey
- Ξένο κλειδί DateKey
- Ξένο κλειδί LocationKey

Τα παραπάνω είναι τα ξένα κλειδιά του πίνακα γεγονότων από όλες της διαστάσεις μαζί με το πρωτεύον κλειδί του πίνακα γεγονότων.

Συνεχίζοντας με τα μετρικά (ίδια με logical schema):

- Deaths
- BeginLocation
- EndLocation
- Injuries
- DamageProperties
- DamageCrops
- BeginLatitude
- EndLatitude
- BeginLongitude
- EndLongitude

Τεχνητές μεταβλητές (φτιαγμένες στον SQL Server) για περαιτέρω ανάλυση:

- TotalDamage
- TotalCasualties
- ImpactLevel
- EventDurationInHours
- HasValidCoordinates

Συνοψίζοντας, ο συγκεκριμένος πίνακας γεγονότων έχει διαμόρφωση που τον καθιστά ικανό για εξειδικευμένα ερωτήματα και αναλύσεις και σε συνδυασμό με τους πίνακες διαστάσεων που τον απαρτίζουν, αποτελεί την κύρια έμπνευση για την μοντελοποίηση του μοντέλου του αστερά και κύριο σημείο έναρξης για την δημιουργία της αποθήκης δεδομένων.

4.4.2 Διαστάσεις

Συνεχίζοντας, έχουμε τις διαστάσεις πινάκων ή αλλιώς dimension tables. Οι πίνακες αυτοί, σε ένα σχήμα αποθήκης δεδομένων διαστάσεων παρέχουν το περιγραφικό πλαίσιο για την εξέταση των αριθμητικών δεδομένων στον πίνακα γεγονότων. Οι πίνακες διαστάσεων περιέχουν τα χαρακτηριστικά που παρέχουν απαντήσεις στις ερωτήσεις "ποιος, τι, πότε, πού και πώς" σχετικά με τα γεγονότα ή τα φαινόμενα που μετριοούνται. Αυτά τα χαρακτηριστικά συνήθως επιτρέπουν το φιλτράρισμα, την ομαδοποίηση και την ανάλυση προς τα κάτω (Kimball & Ross, 2013).

Για τους σκοπούς αυτής της αποθήκης δεδομένων, η οποία θα ασχοληθεί με την ανάλυση των περιστατικών κακοκαιρίας στις Ηνωμένες Πολιτείες για μια δεκαπενταετία, η επιλογή και ο σχεδιασμός των πινάκων διαστάσεων επηρεάστηκαν από τις αρχές της διαστατικής μοντελοποίησης. Σύμφωνα με τους Kimball και Ross (2013), κάθε ένας από τους πίνακες διαστάσεων πρέπει να παρέχει καλή και κατανοητή δομή για την ανάλυση των γεγονότων ώστε οι χρήστες να μπορούν να εξερευνούν αποτελεσματικά μέσα σε τεράστιες ποσότητες δεδομένων. Οι διαστάσεις είναι οι εξής:

Διάσταση ημερομηνίας (Dim_Date):

Η διάσταση ημερομηνίας προσθέτει χρονικό πλαίσιο στην αποθήκη δεδομένων. Επιτρέπει την προβολή φαινομένων σε διάφορες χρονικές διαστάσεις όπως έτος, μήνας κλπ. Τα πεδία που καλύπτονται σε αυτή τη διάσταση είναι τα εξής:

- Τεχνητό κλειδί ημερομηνίας (πρωτεύον κλειδί)
- Year
- Month
- FullDate
- Quarter για τετράμηνη περίοδο
- MonthName
- MonthDay
- DayofWeek
- DayName
- IsWeekend (για σαββατοκύριακο)
- Season (για εποχιακή ανάλυση)

Αυτά τα χαρακτηριστικά επιτρέπουν στους χρήστες να εντοπίζουν τις διαχρονικές τάσεις αλλά και την συχνότητα και σοβαρότητα των φαινομένων.

Διάσταση τοποθεσίας (Dim_Location):

Η διάσταση τοποθεσίας αποθηκεύει χωρικές πληροφορίες σχετικά με την προέλευση και την περιοχή επίδρασης κάθε καιρικού φαινομένου. Τα δεδομένα για κάθε συμβάν εμφανίζονται επίσης με τη μορφή συντεταγμένων γεωγραφικού πλάτους και γεωγραφικού μήκους. Τα πεδία είναι:

- Τεχνητό κλειδί τοποθεσίας (πρωτεύον κλειδί)
- Beg/EndLat (από πίνακα γεγονότων)
- Beg/EndLon (από πίνακα γεγονότων)
- IsValidCoordinates (για έγκυρες μεταβλητές)
- LocationType (για τύπο τοποθεσίας)

Αυτή η δομή παρέχει φιλτράρισμα ειδικό για γεωγραφική και χωρική ανάλυση με την οποία οι αναλυτές μπορούν να ομαδοποιήσουν συμβάντα.

Διάσταση τύπου γεγονότος (Dim_Event_Type):

Αυτή η διάσταση συγκεντρώνει τους τύπους μετεωρολογικών γεγονότων που παρατηρήθηκαν στα δεδομένα. Τα χαρακτηριστικά είναι τα εξής:

- Τεχνητό κλειδί τύπου γεγονότος (πρωτεύον κλειδί)
- EventType (π.χ. πλημμύρα, τυφώνας, πυρκαγιά, χιονοθύελλα)
- EventTypeCode
- EventTypeCategory

Αυτές οι ταξινομήσεις βελτιώνουν την ποιότητα των δεδομένων και υποστηρίζουν διαδικασίες λήψης αποφάσεων βάσει συμβάντων.

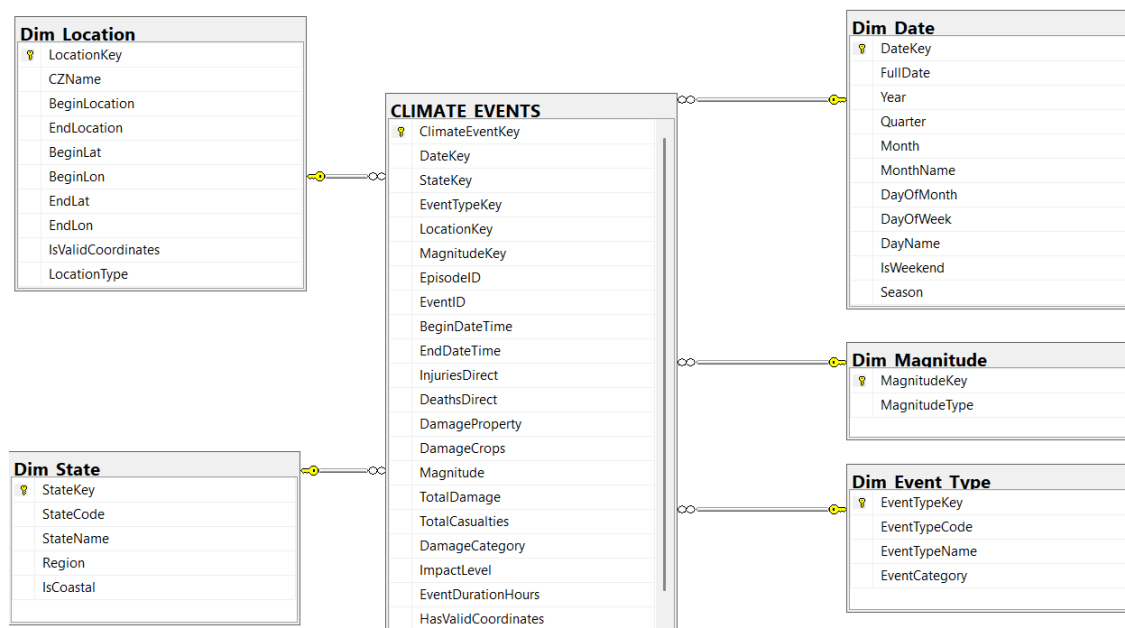
Διάσταση Πολιτείας (Dim_State):

- Τεχνητό αναγνωριστικό (πρωτεύον)
- StateName
- StateCode
- Region (για περιοχές)
- Is Coastal (για θαλάσσιες περιοχές)

Διάσταση έντασης (Dim_Magnitude):

- Τεχνητό κλειδί MagnitudeKey
- MagnitudeType

Το star schema που δημιουργείται με βάση τα παραπάνω φαίνεται στην Εικόνα 11:



Εικόνα 11: Star schema

Συνολικά, οι διαστάσεις, ο πίνακας γεγονότων και τα χαρακτηριστικά των δύο προηγούμενων, συμπεραίνουμε ότι είναι αναπόσπαστο κομμάτι της διαδικασίας που χρησιμοποιείται για να δώσει στην ανάλυση βάθος, ταχύτητα και ευελιξία.

5 Ορισμός και Διαχείριση Δεδομένων

Το συγκεκριμένο κεφάλαιο αποτελεί την συγγραφή κώδικα SQL στο γνωστό περιβάλλον SQL Server Management Studio (SSMS). Θα αναφερθεί η δημιουργία της βάση δεδομένων από τις διαστάσεις μέχρι και τον πίνακα γεγονότων καθώς και κάποια ερωτήματα και τα αποτελέσματα τους εντός του περιβάλλοντος. Η διαδικασία αυτή απαιτεί λεπτομερή προσοχή καθώς τα δεδομένα πρέπει να τροποποιηθούν σωστά.

5.1 Δημιουργία κώδικα SQL

Επόμενο βήμα, η δημιουργία της αποθήκης δεδομένων στο περιβάλλον SQL. Αφού επέλθει η διαδικασία ETL, το νέο καθαρισμένο Excel αρχείο θα εισαχθεί στο SSMS και θα πραγματοποιηθεί η απαραίτητη συγγραφή SQL κώδικα.

Αφού πρώτα δημιουργηθεί η βάση δεδομένων στο σύστημα με όνομα ClimateEvents (CREATE DATABASE ClimateEvents), το αρχείο Excel θα μετατραπεί σε csv (comma separated values) και θα εισαχθεί ως Flat File στην δημιουργημένη βάση. Στη συνέχεια, πρέπει να παραχθεί ο απαιτούμενος κώδικας για την δημιουργία των διαστάσεων. Παρακάτω παρατίθεται ο κώδικας δημιουργίας της διάστασης Dim_Date.

Δημιουργία πίνακα διάστασης ημερομηνίας:

```
CREATE TABLE Dim_Date (  
DateKey INT IDENTITY(1,1) PRIMARY KEY,  
FullDate DATE NOT NULL UNIQUE,  
Year SMALLINT NOT NULL,  
Quarter TINYINT NOT NULL,  
Month TINYINT NOT NULL,  
MonthName VARCHAR(20) NOT NULL,  
DayOfMonth TINYINT NOT NULL,  
DayOfWeek TINYINT NOT NULL,  
DayName VARCHAR(20) NOT NULL,  
IsWeekend BIT NOT NULL,
```

Season VARCHAR(15) NOT NULL

);

Όμοια, δημιουργούνται και οι υπόλοιπες διαστάσεις και επόμενο βήμα είναι η δημιουργία του πίνακα γεγονότων.

Δημιουργία κεντρικού πίνακα γεγονότων:

```
CREATE TABLE CLIMATE_EVENTS (  
    ClimateEventKey BIGINT IDENTITY(1,1) PRIMARY KEY,  
    DateKey INT NOT NULL,  
    StateKey INT NOT NULL,  
    EventTypeKey INT NOT NULL,  
    LocationKey INT NOT NULL,  
    MagnitudeKey INT,  
    EpisodeID INT NOT NULL,  
    EventID INT NOT NULL,  
    BeginDateTime DATETIME,  
    EndDateTime DATETIME,  
    InjuriesDirect SMALLINT DEFAULT 0,  
    DeathsDirect SMALLINT DEFAULT 0,  
    DamageProperty DECIMAL(15,2) DEFAULT 0,  
    DamageCrops DECIMAL(15,2) DEFAULT 0,  
    Magnitude DECIMAL(10,2),  
    TotalDamage DECIMAL(15,2),  
    TotalCasualties SMALLINT,  
    DamageCategory VARCHAR(50),  
    ImpactLevel VARCHAR(50),  
    EventDurationHours INT,  
    HasValidCoordinates BIT DEFAULT 0,  
    CreatedDate DATETIME DEFAULT GETDATE()  
);
```

Οι εντολές εισαγωγής εγγραφών στους πίνακες είναι η επόμενη κίνηση για την δημιουργία της αποθήκης δεδομένων.

Εισαγωγή εγγραφών στον πίνακα Dim_Date:

```
INSERT INTO Dim_Date (FullDate, Year, Quarter, Month, MonthName, DayOf-  
Month, DayOfWeek, DayName, IsWeekend, Season)
```

```

SELECT
    DATEADD(DAY, n, @StartDate) AS FullDate,
    YEAR(DATEADD(DAY, n, @StartDate)) AS Year,
    (MONTH(DATEADD(DAY, n, @StartDate)) - 1) / 3 + 1 AS Quarter,
    MONTH(DATEADD(DAY, n, @StartDate)) AS Month,
    CASE MONTH(DATEADD(DAY, n, @StartDate))
        WHEN 1 THEN 'January'
        WHEN 2 THEN 'February'
        WHEN 3 THEN 'March'
        WHEN 4 THEN 'April'
        WHEN 5 THEN 'May'
        WHEN 6 THEN 'June'
        WHEN 7 THEN 'July'
        WHEN 8 THEN 'August'
        WHEN 9 THEN 'September'
        WHEN 10 THEN 'October'
        WHEN 11 THEN 'November'
        WHEN 12 THEN 'December'
    END AS MonthName,
    DAY(DATEADD(DAY, n, @StartDate)) AS DayOfMonth,
    DATEPART(WEEKDAY, DATEADD(DAY, n, @StartDate)) AS DayOfWeek,
    CASE DATEPART(WEEKDAY, DATEADD(DAY, n, @StartDate))
        WHEN 1 THEN 'Sunday'
        WHEN 2 THEN 'Monday'
        WHEN 3 THEN 'Tuesday'
        WHEN 4 THEN 'Wednesday'
        WHEN 5 THEN 'Thursday'
        WHEN 6 THEN 'Friday'
        WHEN 7 THEN 'Saturday'
    END AS DayName,
    CASE WHEN DATEPART(WEEKDAY, DATEADD(DAY, n, @StartDate)) IN (1, 7)
    THEN 1 ELSE 0 END AS IsWeekend,
    CASE
        WHEN MONTH(DATEADD(DAY, n, @StartDate)) IN (12, 1, 2) THEN 'Win-
ter'

```

```

        WHEN MONTH(DATEADD(DAY, n, @StartDate)) IN (3, 4, 5) THEN
'Spring'
        WHEN MONTH(DATEADD(DAY, n, @StartDate)) IN (6, 7, 8) THEN 'Sum-
mer'
        ELSE 'Fall'
    END AS Season
FROM Numbers
WHERE DATEADD(DAY, n, @StartDate) <= @EndDate;

```

Indexes (Ευρετήρια):

```

CREATE NONCLUSTERED INDEX IX_Date_Year ON Dim_Date(Year);
CREATE NONCLUSTERED INDEX IX_Date_FullDate ON Dim_Date(FullDate);
CREATE NONCLUSTERED INDEX IX_State_Code ON Dim_State(StateCode);
CREATE NONCLUSTERED INDEX IX_State_Name ON Dim_State(StateName);

CREATE      NONCLUSTERED      INDEX      IX_EventType_Category      ON
Dim_Event_Type(EventCategory);
CREATE NONCLUSTERED INDEX IX_Location_Coordinates ON Dim_Location(Begin-
Lat, BeginLon);
CREATE NONCLUSTERED INDEX IX_Climate_Date ON CLIMATE_EVENTS(DateKey);
CREATE NONCLUSTERED INDEX IX_Climate_State ON CLIMATE_EVENTS(StateKey);
CREATE      NONCLUSTERED      INDEX      IX_Climate_EventType      ON
CLIMATE_EVENTS(EventTypeKey);
CREATE      NONCLUSTERED      INDEX      IX_Climate_TotalDamage      ON
CLIMATE_EVENTS(TotalDamage);
CREATE      NONCLUSTERED      INDEX      IX_Climate_ImpactLevel      ON
CLIMATE_EVENTS(ImpactLevel);
CREATE      NONCLUSTERED      INDEX      IX_Climate_DamageCategory      ON
CLIMATE_EVENTS(DamageCategory);
GO

```

5.2 Ερωτήματα

Μετά την εκτέλεση του παραπάνω κώδικα, η βάση έχει ολοκληρωθεί και μπορεί να ξεκινήσει η δοκιμή ερωτημάτων SQL που θα αναδείξουν τα δεδομένα μέσα στο περιβάλλον του SSMS.

Ερώτημα 1: Οι δέκα πολιτείες με την μεγαλύτερη συνολική ζημιά/απώλεια

```
SELECT TOP 10 s.StateName,SUM(ce.TotalDamage) AS TotalDamage
FROM vw_ClimateEvents_Analysis ce
JOIN Dim_State s ON ce.StateName = s.StateName
GROUP BY s.StateName
ORDER BY SUM(ce.TotalDamage) DESC;
```

Ερώτημα 2: Κατανομή γεγονότων ανά τύπο φαινομένου και εποχή

```
SELECT
    EventTypeName, Season, COUNT(EventID) AS ΑριθμόςΓεγονότων
FROM vw_ClimateEvents_Analysis
GROUP BY EventTypeName, Season
ORDER BY EventTypeName, Season;
```

Ερώτημα 3: Αριθμός γεγονότων σε παράκτιες και εσωτερικές περιοχές

```
SELECT
    IsCoastal, COUNT(EventID) AS EventCount
FROM vw_ClimateEvents_Analysis
GROUP BY IsCoastal;
```

Ερώτημα 4: Συνολική ζημιά/απώλεια περιουσίας ανά έτος

```
SELECT
    Year, SUM(DamageProperty) AS ΣυνολικήΖημιάΠεριουσίας
from vw_ClimateEvents_Analysis
GROUP BY Year
ORDER BY Year;
```

Ερώτημα 5: Αριθμός Γεγονότων ανά εποχή με την πάροδο του χρόνου

```
SELECT Year, Season, COUNT(EventID) AS EventsCount
FROM vw_ClimateEvents_Analysis
WHERE Year BETWEEN 2009 AND 2024
GROUP BY Year, Season
ORDER BY Year, Season;
```

Ερώτημα 6: Θάνατοι και τραυματισμοί ανά χρόνο και αριθμό γεγονότων

```
SELECT Year,SUM(TotalCasualties) AS TotalCasualties,COUNT(EventID) AS
EventCount
FROM vw_ClimateEvents_Analysis
WHERE Year BETWEEN 2009 AND 2024
GROUP BY Year
ORDER BY Year;
```

Ερώτημα 7: Μεγάλες καταστροφές ανά χρόνο (απώλεια περιουσίας και καλλιέργειας)

```
SELECT Year,SUM(TotalDamage) AS TotalDamage
FROM vw_ClimateEvents_Analysis
WHERE Year BETWEEN 2009 AND 2024
GROUP BY Year
ORDER BY Year;
```

Ερώτημα 8: Θνησιμότητα κλιματικών γεγονότων

```
SELECT EventTypeName, SUM(DeathsDirect) AS TotalDeaths
FROM vw_ClimateEvents_Analysis
```

```
GROUP BY EventTypeName
HAVING SUM(DeathsDirect) > 0
ORDER BY TotalDeaths DESC;
```

Ερώτημα 9: Υψηλή συχνότητα κλιματικών φαινομένων (ανάλυση για φαινόμενα με πάνω από 20 χιλιάδες γεγονότα)

```
SELECT EventTypeName, COUNT(EventID) AS EventCount
FROM vw_ClimateEvents_Analysis
GROUP BY EventTypeName
HAVING COUNT(EventID) > 20000
ORDER BY EventCount DESC;
```

Ερώτημα 10: Ώρες διάρκειας γεγονότος ανά πολιτεία

```
SELECT StateName, AVG(EventDurationHours) AS AvgEventDurationHours,
MAX(EventDurationHours) AS MaxEventDurationHours,
MIN(EventDurationHours) AS MinEventDurationHours
FROM vw_ClimateEvents_Analysis
GROUP BY StateName
ORDER BY AvgEventDurationHours DESC;
```

Ερώτημα 11: Τα 5 πιο ζημιογόνα φαινόμενα

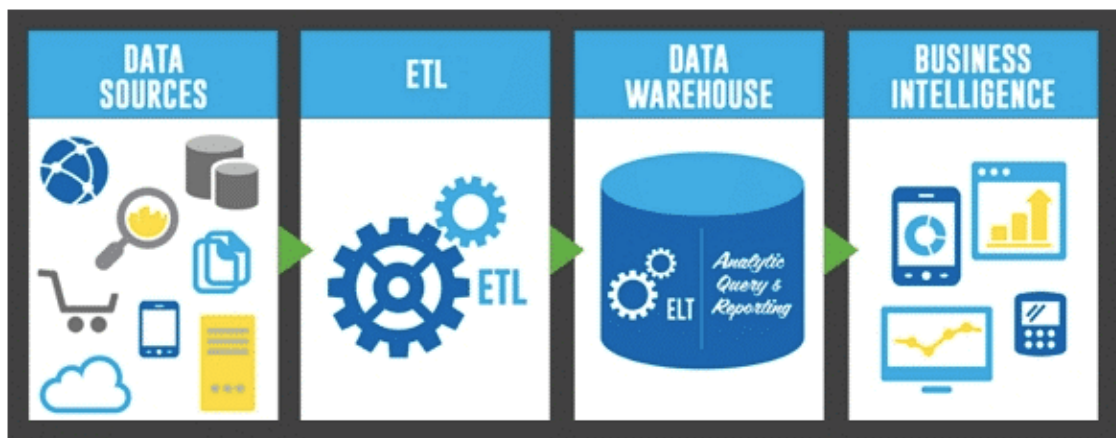
```
SELECT TOP 5 EpisodeID, EventTypeName, SUM(TotalDamage) AS TotalEpisodeDamage
FROM vw_ClimateEvents_Analysis
GROUP BY EpisodeID, EventTypeName
ORDER BY TotalEpisodeDamage DESC;
```

Τα παραπάνω ερωτήματα είναι πλήρως λειτουργικά στο περιβάλλον του SSMS και αναδεικνύουν σημαντικές πληροφορίες για δεδομένα όπως αριθμός γεγονότων ανά μήνα, τύπους γεγονότων, θανάτους τραυματισμούς και άλλα πεδία συνδυάζοντας τα σε σύνθετα ερωτήματα.

6 Οπτικοποίηση Δεδομένων

Με την δημιουργία του απαραίτητου κώδικα SQL και ουσιαστικά της αποθήκης δεδομένων, μένει το τελικό βήμα που είναι η οπτικοποίηση της με την βοήθεια του εργαλείου Power BI.

Το Power BI είναι ένα μέσο επιχειρηματικής ευφυΐας της Microsoft όπου δίνει την δυνατότητα σύνδεσης με άλλες πηγές δεδομένων. Μία από αυτές είναι και το SQL Server όπου η σύνδεση πραγματοποιείται εύκολα μέσα από το Power BI με μοναδική προϋπόθεση να δοθεί το όνομα του διακομιστή και το όνομα της βάσης δεδομένων. Η εισαγωγή των δεδομένων παίρνει μόλις λίγα λεπτά ή και δευτερόλεπτα ανάλογα με τον όγκο δεδομένων που φορτώνονται. Στην Εικόνα 12, απεικονίζεται συνοπτικά ολόκληρη η διαδικασία δημιουργίας του αποθετηρίου δεδομένων από την επιλογή πηγών έως και την οπτικοποίηση.

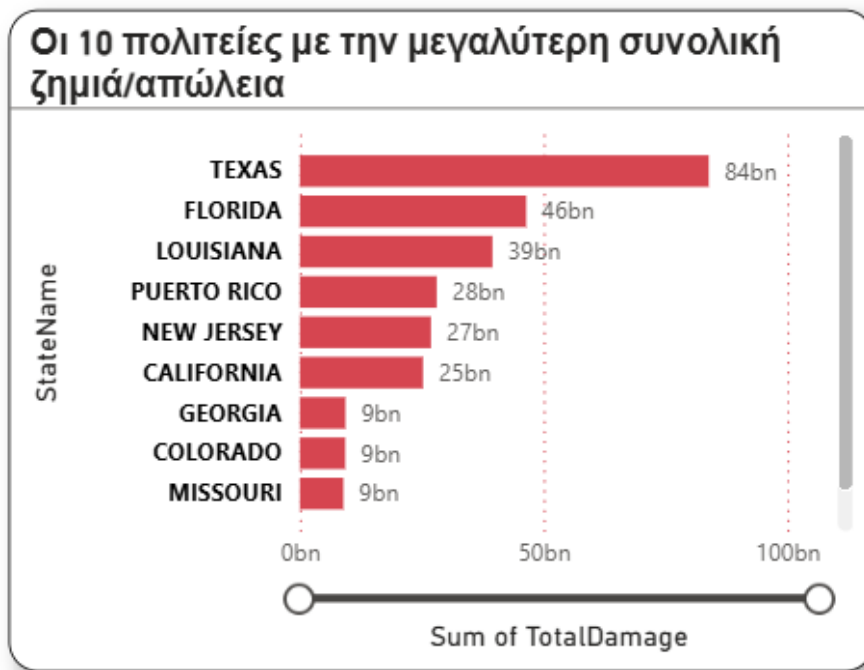


Εικόνα 12: Διαδικασία ανάπτυξης αποθήκης δεδομένων.

(<https://airbyte.com/data-engineering-resources/etl-in-business-intelligence>)

Εκτός από την εύκολη σύνδεση, προσφέρει και λειτουργίες όπως το Model View για την ανάδειξη μοντελοποιημένων δεδομένων, το Table View που δείχνει τα δεδομένα σε μορφή πίνακα, το Power Query και DAX κώδικας για μέτρα μέσα στο Power BI. Τα dashboards είναι η κύρια λειτουργία και για αυτό παρακάτω θα αναδείξουμε κάποιες σημαντικές οπτικοποιήσεις που πραγματοποιήθηκαν.

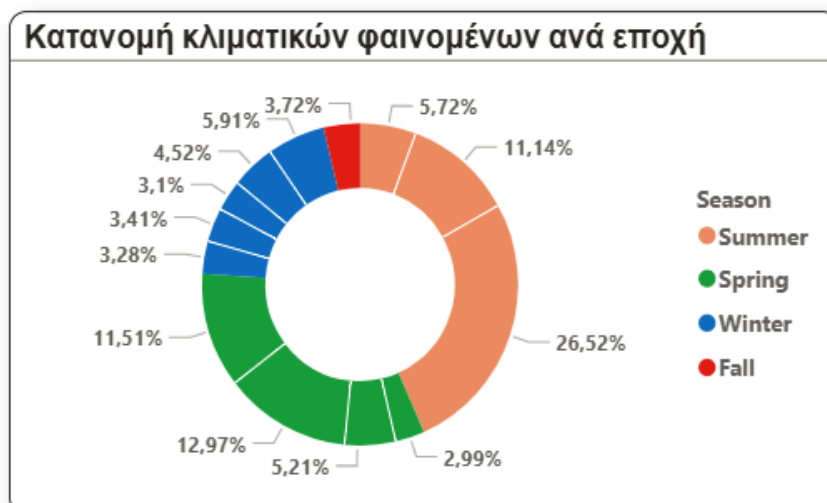
Ερώτημα 1: Οι δέκα πολιτείες με την μεγαλύτερη συνολική ζημιά/απώλεια



Εικόνα 13: Οπτικοποιημένο αποτέλεσμα ερωτήματος 1

Το συγκεκριμένο γράφημα δείχνει την συνολική απώλεια περιουσίας και καλλιέργειας προσθετικά σε δισεκατομμύρια δολάρια για την χρονική περίοδο 2009 έως 2024. Με μεγάλη διαφορά τα δεδομένα αναδεικνύουν το TEXAS ως την πιο επηρεασμένη πολιτεία. Αυτό είναι αναμενόμενο και γνωστό στις ΗΠΑ καθώς το Τέξας αντιμετωπίζει κάθε χρόνο περισσότερο από κάθε άλλη πολιτεία τις καταστροφικές συνέπειες της κλιματικής αλλαγής. Δεύτερη είναι η Φλόριντα σχεδόν με διπλάσια διαφορά από την πρώτη. Ακολουθούν μεγάλες απώλειες και σε Λουϊζιάνα, Πουέρτο Ρίκο (ευρύτερη περιοχή των ΗΠΑ, όχι πολιτεία), Νιου Τζέρσεϋ και φυσικά Καλιφόρνια.

Ερώτημα 2: Αριθμός κλιματικών φαινομένων ανά εποχή



Εικόνα 14: Οπτικοποιημένο αποτέλεσμα ερωτήματος 2

Όπως φαίνεται και από το γράφημα, έχουμε τις τέσσερις εποχές του χρόνου χωρισμένες σε 4 χρώματα να αναδεικνύουν τα πιο σημαντικά κλιματικά φαινόμενα που έλαβαν χώρα σε κάθε εποχή. Με το κόκκινο δείχνουμε το φθινόπωρο, με το μπλε τον χειμώνα, με το πράσινο την άνοιξη και με το πορτοκαλί το καλοκαίρι.

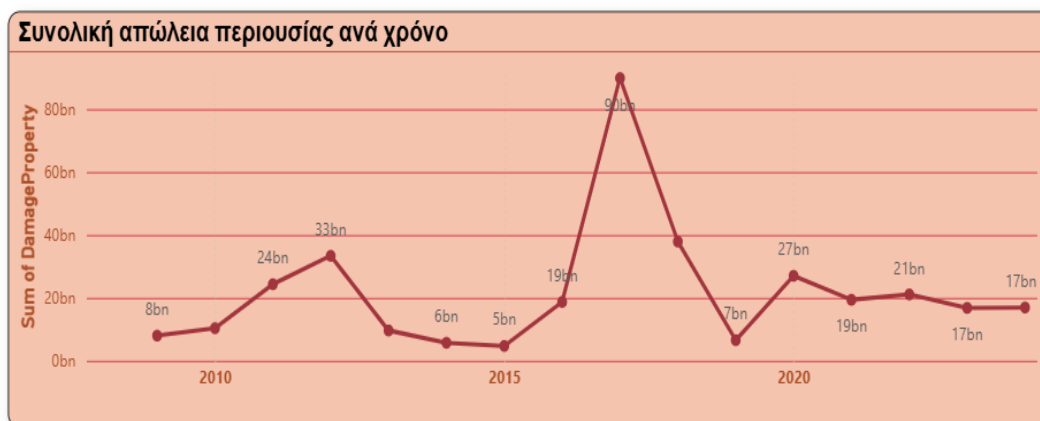
Ερώτημα 3: Αριθμός γεγονότων σε παράκτιες και εσωτερικές περιοχές



Εικόνα 15: Οπτικοποιημένο αποτέλεσμα ερωτήματος 3

Η Εικόνα 16 οπτικοποιεί την κατανομή των γεγονότων σε παράκτιες και εσωτερικές περιοχές. Η μεταβλητή IsCoastal σημαίνει ότι η περιοχή είναι παράκτια όταν True, δηλαδή όταν το χρώμα είναι μπλε και όταν είναι False, η περιοχή είναι εσωτερική χωρίς να βρέχεται από θάλασσα με πράσινο χρώμα. Παρατηρείτε ότι τα γεγονότα εκτός παράκτιων περιοχών, είναι περισσότερα καθώς καλύπτουν περίπου το 60% των γεγονότων.

Ερώτημα 4: Απώλεια περιουσίας ανά χρόνο



Εικόνα 16: Οπτικοποιημένο αποτέλεσμα ερωτήματος 4

Παραπάνω παρουσιάζεται ένα Line Chart. Μέσα από αυτό, φαίνεται ανάλογα με την δραστηριότητα της γραμμής, η ζημιά σε δισεκατομμύρια. Με το μάτι καταλαβαίνουμε ότι η χρονιά 2017 παρουσιάζει την μεγαλύτερη απώλεια περιουσίας. Επίσης η μικρότερη απώλεια περιουσίας συνέβη το 2015. Παρατηρούμε ότι σε μόλις δύο χρόνια (2015-2017), η ζημιά από πολύ μικρή γίνεται καταστροφική. Το συγκεκριμένο παράδειγμα, δείχνει με στατιστικά την μεγάλη δύναμη της φύσης και της κλιματικής αλλαγής.

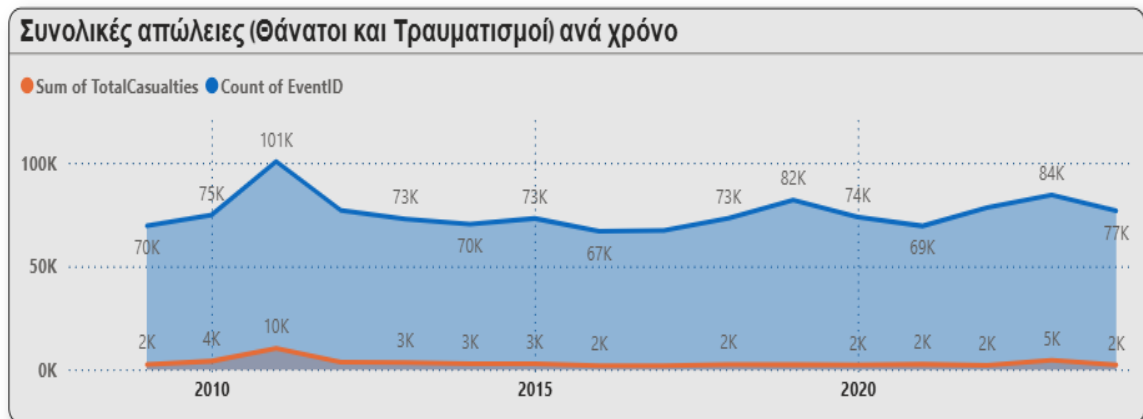
Ερώτημα 5: Αριθμός Γεγονότων ανά εποχή με την πάροδο του χρόνου



Εικόνα 17: Οπτικοποιημένο αποτέλεσμα ερωτήματος 5

Η παραπάνω οπτικοποίηση δείχνει με το πέρασμα του χρόνου την αυξομειώση των γεγονότων σε σχέση με την εποχή. Παρατηρείτε ότι το φθινόπωρο τα γεγονότα έχουν αξιοσημείωτα σταθερό ρυθμό χωρίς μεγάλες αυξομειώσεις ενώ και ο χειμώνας ακολουθεί παρόμοιο μοτίβο.

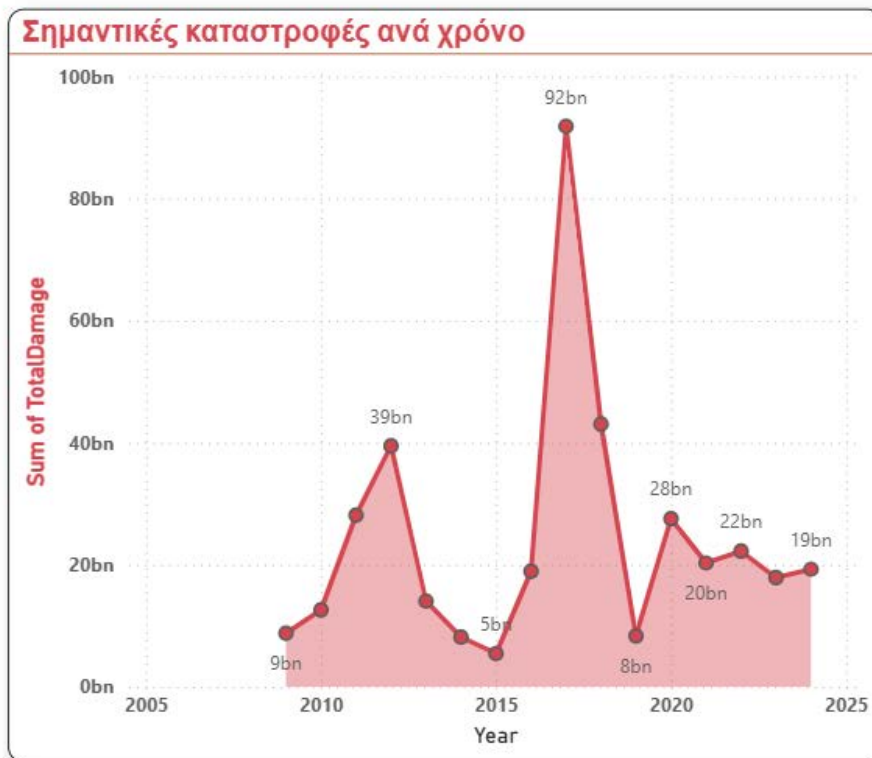
Ερώτημα 6: Θάνατοι και τραυματισμοί ανά χρόνο και αριθμό γεγονότων



Εικόνα 18: Οπτικοποιημένο αποτέλεσμα ερωτήματος 6

Το γράφημα παρουσιάζει με την μπλε γραμμή των αριθμό γεγονότων και με την πορτοκαλί γραμμή το σύνολο τραυματισμών και θανάτων. Με το παραπάνω γράφημα φαίνεται η αναλογία γεγονότων και θανάτων/τραυματισμών. Για παράδειγμα το 2015, τα γεγονότα ανέρχονται στα 73 χιλιάδες ενώ οι θάνατοι μαζί με τους τραυματισμούς στους 3 χιλιάδες.

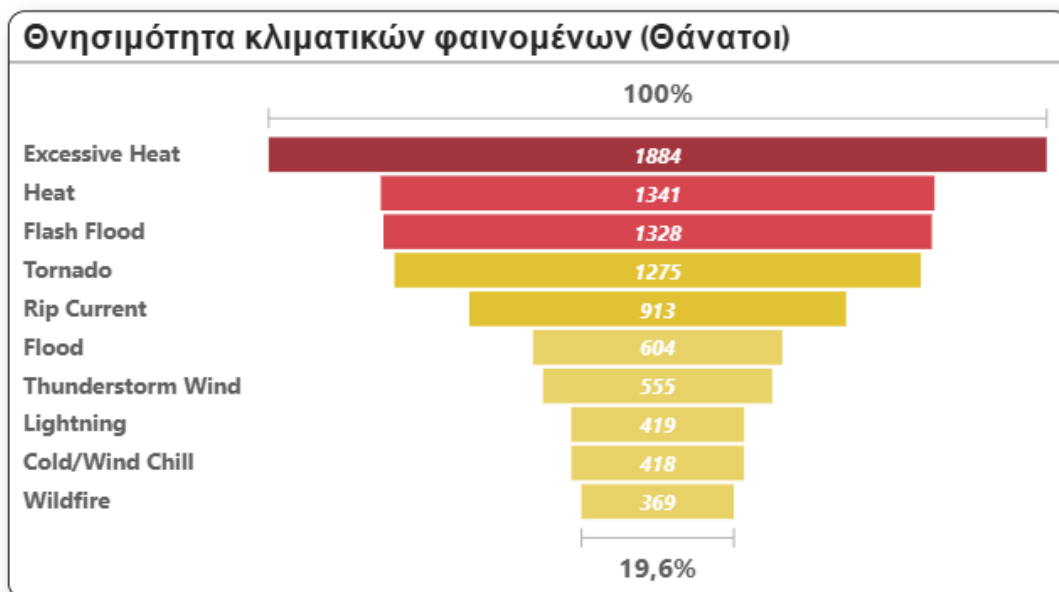
Ερώτημα 7: Μεγάλες καταστροφές ανά χρόνο (απώλεια περιουσίας και καλλιέργειας)



Εικόνα 19: Οπτικοποιημένο αποτέλεσμα ερωτήματος 7

Στο γράφημα 7, παρουσιάζεται η καταστροφή από κλιματικά φαινόμενα (η οποία μετριέται σε απώλεια περιουσίας και καλλιέργειας) σε δισεκατομμύρια με την πάροδο του χρόνου. Άξιο αναφοράς σημείο είναι ότι το 2015 σημειώνεται η μικρότερη ζημιά και δύο χρόνια μετά, το 2017 σημειώνεται η μεγαλύτερη ζημιά με τεράστια διαφορά από τα υπόλοιπα έτη.

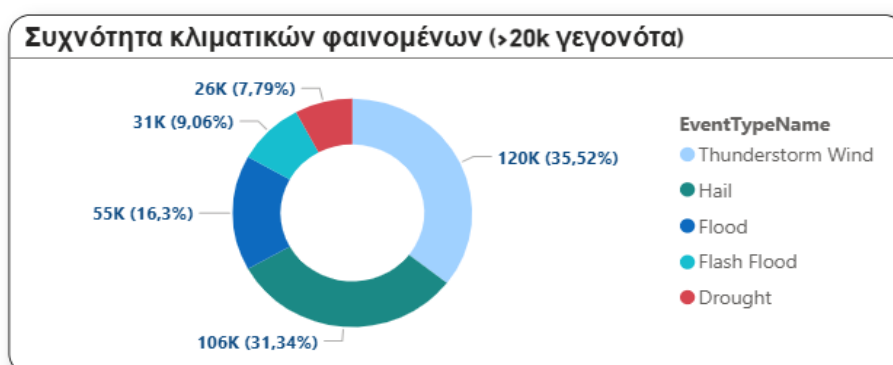
Ερώτημα 8: Θνησιμότητα κλιματικών γεγονότων



Εικόνα 20: Οπτικοποιημένο αποτέλεσμα ερωτήματος 8

Το γράφημα 8 αναδεικνύει τους τύπους γεγονότων που προκάλεσαν τους περισσότερους θανάτους. Τα πέντε πρώτα κλιματικά φαινόμενα παρουσιάζουν μεγάλη θνησιμότητα καθώς οι θάνατοι κυμαίνονται από 913 (ελάχιστη τιμή) από θαλάσσια ρεύματα έως και 1884 (μέγιστη τιμή) από υπερβολικό καύσωνα.

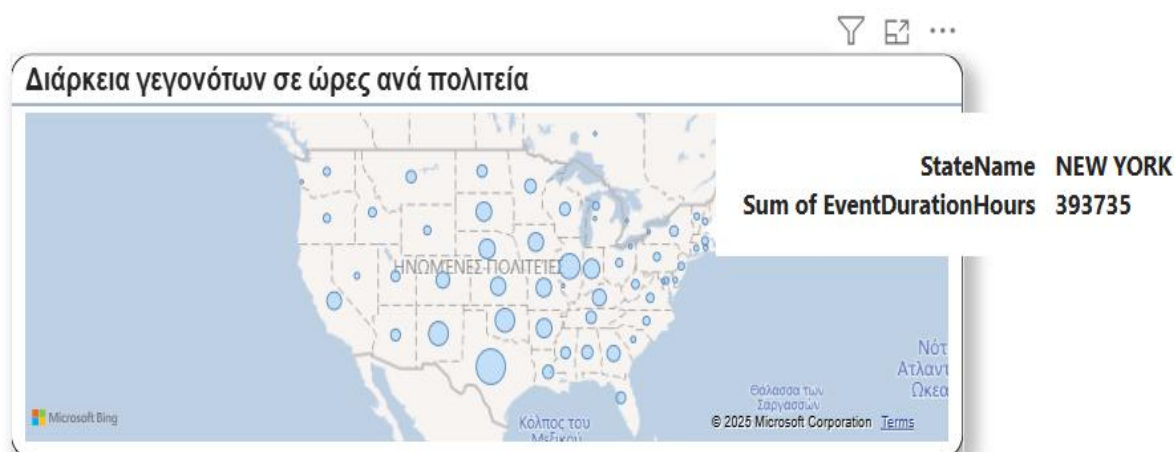
Ερώτημα 9: Υψηλή συχνότητα κλιματικών φαινομένων (ανάλυση για φαινόμενα με πάνω από 20 χιλιάδες γεγονότα)



Εικόνα 21: Οπτικοποιημένο αποτέλεσμα ερωτήματος 9

Το συγκεκριμένο σχήμα οπτικοποιεί δεδομένα κλιματικών φαινομένων που ξεπερνούν τα 20 χιλιάδες γεγονότα για την ανάγκη της ανάλυσης και καλύτερης ανάδειξης για το συγκεκριμένο γράφημα. Με το μέτρο αυτό, μένουν 5 γεγονότα και το πιο συχνό όλων είναι η καταιγίδα κεραυνών με περίπου 120 χιλιάδες καταγεγραμμένα γεγονότα.

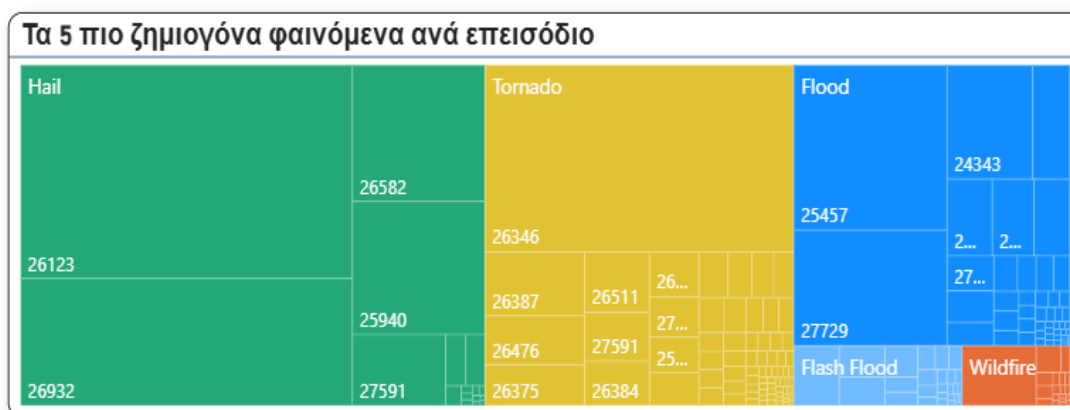
Ερώτημα 10: Ώρες διάρκειας γεγονότος ανά πολιτεία



Εικόνα 22: Οπτικοποιημένο αποτέλεσμα ερωτήματος 10

Ο παραπάνω χάρτης, αποσκοπεί στο να δείξει τις πολιτείες της Αμερικής και τις συνολικές ώρες διάρκειας γεγονότων σε αυτές. Πατώντας πάνω σε μια πολιτεία, (π.χ. New York) φαίνεται η συνολική διάρκεια σε ώρες στην συγκεκριμένη πολιτεία. Στο συγκεκριμένο παράδειγμα η Νέα Υόρκη επηρεάστηκε συνολικά 393.735 ώρες.

Ερώτημα 11: Τα 5 πιο ζημιογόνα φαινόμενα ανά επεισόδιο



Εικόνα 23: Οπτικοποιημένο αποτέλεσμα ερωτήματος 11

Εκτός από γεγονότα, υπάρχουν και επεισόδια στα δεδομένα τα οποία είναι ουσιαστικά ομαδοποιήσεις γεγονότων που έγιναν την ίδια χρονική περίοδο οπότε μπορούν να συμπεριληφθούν ως ένα επεισόδιο. Έτσι, στο παραπάνω παράδειγμα παρουσιάζονται σε ένα Tree map τα 5 πιο ζημιογόνα φαινόμενα ανά επεισόδιο. Ο μεγαλύτερος χώρος στο σχήμα καταβάλλεται από επεισόδια με καταιγίδα και αυτό σημαίνει ότι η καταιγίδα είναι το πιο ζημιογόνο φαινόμενο στα επεισόδια που έχουν καταγραφεί στην βάση δεδομένων.

Σημείωση: Τα επεισόδια στην αποθήκη δεδομένων ανέρχονται περίπου στα 160.000, ενώ τα συνολικά γεγονότα πάνω από 1.200.000 χιλιάδες. Αυτό σημαίνει ότι το συγκεκριμένο γράφημα δεν δείχνει τα πιο καταστροφικά φαινόμενα γενικά, αλλά τα πιο καταστροφικά που πήραν μέρος σε ένα επεισόδιο.

7 Συμπεράσματα και Προοπτικές

Στο τελευταίο κεφάλαιο, θα γίνει μια σύνοψη και αξιολόγηση των εργαλείων που χρησιμοποιήθηκαν, θα αναλυθεί το πώς μπορεί η εργασία να ενσωματωθεί ως ένα σενάριο χρήσης για την Πολιτική Προστασία και θα αναδειχθούν κύρια συμπεράσματα και πιθανές επεκτάσεις της ανάλυσης.

7.1 Αξιολόγηση εργαλείων που χρησιμοποιήθηκαν

Τα εργαλεία θα αξιολογηθούν με βάση τις ικανότητες, τους περιορισμούς και την ευκολία χρήσης τους. Τα εργαλεία είναι τα εξής:

- NOAA Storm Events Database: Χρησιμοποιήθηκε για την χρήση ανάκτησης δεδομένων. Διαθέτει ότι χρειάζεται κάποιος για να κάνει μια ολοκληρωμένη ανάλυση με βάση τα κλιματικά φαινόμενα στις ΗΠΑ. Οι εγγραφές από το 2000 και έπειτα, διαθέτουν τεράστιο όγκο πληροφοριών. Φυσικά, δεκαετίες του 1960 ή του 1970 διαθέτουν μη επαρκή δεδομένα για εις βάθος αναλύσεις λόγω περιορισμένων πόρων. Η λειτουργία Bulk Download, όπως περιγράφεται στο σύστημα, με βοήθησε πολύ στο να κατεβάσω μαζικά αρχεία που χρειάζομαι για την αποθήκη μου. Χωρίς αυτό, θα χρειαζόμουν περισσότερο χρόνο για να περάσω τα δεδομένα μου στο Excel. Η βάση δεδομένων του NOAA, είναι εύκολη στην χρήση με ικανότητα για φιλτράρισμα και ο όγκος που διαθέτει δεν καθιστά απαραίτητη την πρόσθεση δεδομένων από άλλες πηγές.
- Excel: Το εργαλείο αυτό χρησιμοποιήθηκε περισσότερο από οποιοδήποτε άλλο για την εργασία καθώς οι διαδικασίες αποδείχθηκαν αρκετά χρονοβόρες για εμένα. Αν και σε γενικές γραμμές το Excel είναι αρκετά φιλικό προς τον χρήστη, οι διαδικασίες μπορούν να γίνουν πολύπλοκες και να απατούν μια καλή γνώση με το αντικείμενο όταν τα δεδομένα είναι πολλά. Ενδεικτικά, σε κάποια σημεία της ETL, υπήρχε η ανάγκη για προχωρημένες λειτουργίες όπως αυτή της συγγραφής κώδικα M για να δημιουργηθούν αυτοματισμοί που θα άλλαζαν μαζικά δεδομένα αντί χειροκίνητα.

Παρά το γεγονός ότι υπήρχαν δυσκολίες όσο αναφορά τον χρόνο που έπρεπε να διαθέσω και την εμπειρία μου σε προχωρημένες λειτουργίες, το Excel είναι η κατάλληλη εφαρμογή για να εφαρμόσει κάποιος καθαρισμό στα δεδομένα του.

- SQL Server Management Studio: Η εξοικείωση που είχα αναπτύξει με το εργαλείο τα προηγούμενα έτη σπουδών μου, συνέβαλε σε μεγάλο βαθμό στο να αναπτύξω την αποθήκη δεδομένων με μεγαλύτερη ευκολία. Συγκεκριμένα, γνωρίζοντας ήδη το περιβάλλον και έχοντας βασική γνώση σε δημιουργία SQL ερωτημάτων και υλοποίηση star schema's, δεν συνάντησα ιδιαίτερη δυσκολία και θεώρησα αρκετά αξιόλογο και λειτουργικό το περιβάλλον του.
- Power BI: Τελικό στάδιο του έργου με φιλικό προς τον χρήστη περιβάλλον και εξαιρετικές ικανότητες οπτικοποίησης και τρόπους ανάδειξης των δεδομένων. Διαθέτει τεράστια ποικιλία γραφημάτων, διαγραμμάτων και λειτουργιών όπως χαρτογράφηση και διαδραστικά dashboards με σχετικά εύκολες διαδικασίες και φιλτραρίσματα. Η εξοικείωση του χρήστη με τα παραπάνω δεν είναι δύσκολη και μπορεί να γίνει κατά την διάρκεια της ανάλυσης. Συνοψίζοντας, με κατάλληλη χρήση, το σύστημα δεν θα απογοητεύσει και θα δώσει τα κατάλληλα εφόδια για μια κατανοητή και ενδιαφέρουσα τελική οπτικοποίηση.

7.2 Πιθανό σενάριο χρήσης από Πολιτική Προστασία

Σύμφωνα με προηγούμενα λεγόμενα, είναι γεγονός πως η συγκεκριμένη πτυχιακή εργασία μπορεί να προσφέρει σημαντική βοήθεια σε φορείς της Πολιτικής Προστασίας.

Τα ερωτήματα SQL σε συνδυασμό με την οπτικοποίηση στο Power BI και συνολικά η αποθήκη δεδομένων αποτελεί κύριο εργαλείο που μπορεί να χρησιμοποιήσει κάποιος αρμόδιος για την ανάλυση συμβάντων που πήραν μέρος σε μια περιοχή.

Συγκεκριμένα, η ευρύτερη περιοχή της Θεσσαλίας την τελευταία πενταετία έχει υποστεί πρωτόγνωρες καταστροφές από πλημμύρες με τεράστιες συνέπειες σε περιουσίες, καλλιέργειες, ακόμη και θανάτους. Είναι απαραίτητη λοιπόν, η άμεση επέμβαση της Πολιτικής Προστασίας σε συνδυασμό με εργαλεία όπως το συγκεκριμένο και δεδομένα επαρκή για να επέλθει συμπέρασμα που θα καθοδηγήσει σε σημαντικές λήψεις αποφάσεων βελτιώνοντας σε μεγάλο βαθμό την προσαρμοστικότητα της περιοχής σε τέτοιου είδους και έντασης κλιματικά φαινόμενα. Αναλυτικότερα, το αποθετήριο δεδομένων θα συλλέξει επαρκή δεδομένα και πληροφορίες σχετικά με τα γεγονότα της περιοχής όπως οι σταγόνες νερού που έπεσαν μια συγκεκριμένη χρονική περίοδο των φαινομένων (π.χ. πρώτη ώρα ή πρώτη μέρα των φαινομένων). Άλλο σημαντικό στοιχείο για την ανάλυση είναι η αρχική τοποθεσία των φαινομένων, δηλαδή σε ποιο σημείο ξεκίνησαν. Αυτή η πληροφορία, με την βοήθεια σύνθετων SQL ερωτημάτων και του Power BI για την οπτικοποίηση,

μπορεί να συμβάλλει στο να βρεθούν οι ευάλωτες περιοχές της Θεσσαλίας που συγκεντρώνουν πιο εύκολα μεγάλες ποσότητες νερού μέσα από dashboards που θα αναδεικνύουν πολύτιμες πληροφορίες για την κατανόηση των πλημμυρικών φαινομένων. Παραδείγματος χάρη με το παρακάτω ερώτημα (σύμφωνα πάντα με τα δικά μας δεδομένα):

```
SELECT EVENT_ID, BEGIN_LOCATION, STATE, EVENT_TYPE, MONTH NAME, YEAR,
BEGIN_DATETIME
FROM FactClimateEvents
WHERE EVENT_TYPE= 'Flood'
And STATE= 'California'
AND MONTH_NAME= 'September'
AND YEAR= '2023'.
```

Τέλος, μεταβλητές απώλειας περιουσίας, καλλιέργειας και τραυματισμών ή θανάτων που υπάρχουν και στην υπάρχουσα ανάλυση, είναι πολύ σημαντικά στοιχεία για την στατιστική ανάλυση και διάφορες συγκρίσεις.

Η οπτικοποίηση τέτοιων δεδομένων θα αποδειχτεί μεγάλης σημασίας καθώς η Πολιτική Προστασία θα κατανοήσει καλύτερα τις ανάγκες της περιοχής και μπορεί να πάρει αποφάσεις όπως ο καθαρισμός φερτών υλικών από ποτάμια που εμφανίζονται ως αρχές πλημμυρικών φαινομένων και το σκάψιμο μεγάλων λάκκων κοντά σε ευάλωτα σημεία που θα αποθηκεύουν προσωρινά και θα αντισταθμίζουν την μεγάλη ποσότητα νερού που εμφανίζεται. Έτσι θα αποτρέπονται τα περισσότερα περιστατικά πλημμυρικών επεισοδίων προστατεύοντας τις πόλεις και τα χωράφια τους.

Συνοψίζοντας, η εργασία εκτός από ακαδημαϊκό έργο μπορεί να προσφέρει και πρακτικό σε πραγματικές καταστάσεις κλιματικών επεισοδίων.

7.3 Βασικά συμπεράσματα και μελλοντικές προεκτάσεις

Όπως αναλύθηκε στα παραπάνω κεφάλαια, η πτυχιακή μου εργασία εστίασε στην ανάπτυξη μιας πλήρη λειτουργικής αποθήκης δεδομένων για ανάλυση κλιματικών φαινομένων στις Ηνωμένες Πολιτείες της Αμερικής, το εύρος ετών 2009-2024.

Τα δεδομένα συλλέχθηκαν από την βάση της NOAA που ονομάζεται Storm Events Database και στη συνέχεια καθαρίστηκαν και μοντελοποιήθηκαν μέσω γνωστών εφαρμογών και διαδικασιών που αναλύθηκαν εκτενώς σε προηγούμενα κεφάλαια. Τα δεδομένα

πριν εισέλθουν στο SSMS, έχουν υποστεί καθαρισμό ο οποίος έγινε με την εφαρμογή της διαδικασίας ETL και αργότερα η μοντελοποίηση μέσω SSMS με το μοντέλο του αστέρα. Τα βήματα αυτά ήταν από τα πιο σημαντικά για την δημιουργία της αποθήκης δεδομένων καθώς απαιτούσαν χρόνο και ιδιαίτερη προσοχή. Τέλος με τη χρήση του Power BI δημιουργήθηκαν οπτικοποιήσεις των δεδομένων μέσω διαδραστικών αναλύσεων για να ολοκληρωθεί πλήρως η αποθήκη δεδομένων.

Στο πλαίσιο εκπόνησης της εργασίας, συνάντησα διάφορες χρονοβόρες διαδικασίες και δυσκολίες κυρίως στην διαδικασία εκκαθάρισης των δεδομένων (ETL) μέσα στο Excel και στο SSMS. Αρχικά, η ιδιαίτερα μεγάλη πρόκληση που αντιμετώπισα ήταν η επεξεργασία των τεράστιων όγκων ακατέργαστων δεδομένων που επέλεξα μέσα από την βάση. Εκεί συνάντησα ελλιπείς εγγραφές σε αρκετές γραμμές του Excel και λανθασμένες μορφοποιήσεις οι οποίες αντιμετωπίστηκαν με αρκετές μετατροπές, φίλτραρίσματα, λειτουργίες για προχωρημένους που προϋποθέτανε κώδικα Μ και αλλαγές τύπων των τιμών. Η πτυχιακή εργασία ολοκληρώθηκε επιτυχώς όμως εάν μου δινόταν η ευκαιρία να την αρχίσω από την αρχή σήμερα, ίσως να έκανα κάποια πράγματα διαφορετικά. Για παράδειγμα, με την εμπειρία που έχω αποκτήσει στις διάφορες εφαρμογές και περιβάλλοντα που χρησιμοποιήθηκαν, θα ήμουν πολύ πιο γρήγορος στην ανάπτυξη της αποθήκης και θα γνώριζα εξαρχής ποια σημεία να προσέξω και να τελειοποιήσω.

Παρόλα αυτά, μέσα από τις δυσκολίες και την συνεχή ενασχόληση με το αντικείμενο της πτυχιακής μου μέσα στη χρονιά, μπορώ να πω με σιγουριά πως άξιζε τον κόπο καθώς απέκτησα σημαντικές γνώσεις και εμπειρίες. Έμαθα ποικιλία λειτουργιών του Excel, ασχολήθηκα πολλές ώρες εφαρμόζοντας την διαδικασία ETL, εμβάθυνα σε σημαντικό βαθμό στη γλώσσα SQL, εξοικειώθηκα περισσότερο με εργαλεία οπτικοποίησης και διάβασα σημαντικές πληροφορίες σχετικά με τις αποθήκες δεδομένων μέσω ηλεκτρονικών πόρων όπως άρθρα, ιστοσελίδες και ηλεκτρονικά βιβλία που μου παρείχε το διαδίκτυο.

Στις δύο προηγούμενες παραγράφους αναφέρθηκα στις δυσκολίες και στις γνώσεις που μου προσφέρθηκαν. Μέσω αυτών, το ενδιαφέρον μου για ενασχόληση με το αντικείμενο μεγάλωσε. Η επαφή με εργαλεία όπως το SSMS και το Power BI μου δημιούργησαν περιέργεια και ενθουσιασμό λόγω των αμέτρητων δυνατοτήτων που προσφέρουν.

Συνοψίζοντας, η ενασχόληση αυτή με βοήθησε στο να κινηθώ στοχευμένα στον επαγγελματικό προσανατολισμό που θέλω να επιλέξω σε αυτό το στάδιο της ζωής μου.

Όσο αναφορά μελλοντικές προεκτάσεις του συγκεκριμένου έργου, υπάρχουν δυνατότητες επέκτασης και βελτίωσης του μοντέλου που μπορούν να υλοποιηθούν είτε από εμένα

είτε από κάποιον άλλον φοιτητή. Μια από αυτές είναι η πρόσθεση περισσότερων εξιδεικευμένων μεταβλητών που μπορούν να προσφέρουν ειδικές και ενδιαφέρουσες αναλύσεις μετά την ενσωμάτωσή τους στην αποθήκη. Για παράδειγμα, μεταβλητές όπως πραγματικές αποζημιώσεις για την απώλεια ανθρώπινης περιουσίας (damage_property) ή έναν κοινωνικοοικονομικό δείκτη που θα δείχνει την κοινωνική και οικονομική κατάσταση της οποιαδήποτε πληγείσας περιοχής στις ΗΠΑ συγκριτικά με το πόσο επηρεάστηκε από τα κλιματικά φαινόμενα μιας συγκεκριμένης περιόδου. Τέλος, η αρχιτεκτονική είναι επίσης επεκτάσιμη για πηγές δεδομένων όπως δορυφορικές εικόνες, προγράμματα που δείχνουν καιρό σε πραγματικό χρόνο και υποστηρίζει advanced analytics όπως είναι η μηχανική μάθηση (machine learning) και αναλύσεις ομαδοποιήσεων (clustering analysis).

Συμπερασματικά, τα οφέλη από την πτυχιακή εργασία ήταν πολύ μεγάλα καθώς κατάφερα να συλλέξω μια ποικιλία γνώσεων σε όλη την διάρκεια εκπόνησης της σε πρακτικό και ερευνητικό επίπεδο. Εκτός από αυτό, μέσω της διαδικασίας της έμαθα και να χρησιμοποιώ γενικές γνώσεις που μου προσφέρθηκαν στο πρόγραμμα σπουδών και να τις υλοποιώ σε ένα πρακτικό μοντέλο όπως αυτό της αποθήκης δεδομένων.

Βιβλιογραφία

- AtScale. What is Data Loading? Ανακτήθηκε 23 Αυγούστου 2025 από: <https://www.atscale.com/glossary/data-loading/>
- AWS. What is ETL (Extract Transform Load)? Ανακτήθηκε 23 Αυγούστου 2025 από: <https://aws.amazon.com/what-is/etl/>
- Başaran, B. P. (2005, Ιανουάριος). A comparison of data warehouse design models (Μεταπτυχιακή διατριβή, Τμήμα Μηχανικών Η/Υ, Atilim University). Ανακτήθηκε 23 Αυγούστου 2025 από: <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=73d6ac070a5413aed770a053c531d09f52aa565d>
- Boyko, N. I., & Chernenko, A. V. (2024). Modern approaches to data storage: Comparison of relational and cloud data warehouses using ETL and ELT methods. Ανακτήθηκε 22 Αυγούστου 2025 από: https://journals.urau.ua/vestnikpgtu_tech/article/view/310669
- BryteFlow. Data Lake vs Data Warehouse for Enterprise Data Integration. Ανακτήθηκε 23 Αυγούστου 2025 από: <https://bryteflow.com/data-lake-vs-data-warehouse-for-enterprise-data-integration>
- Coursera Staff. (2025, Ιούλιος 16). Data Warehouse vs. Database: What's the Difference? Ανακτήθηκε 23 Αυγούστου 2025 από: <https://www.coursera.org/articles/data-warehouse-vs-database>
- Donatti, C. I., Nicholas, K., Fedele, G., Delforge, D., Speybroeck, N., Moraga, P., Blatter, J., Below, R., & Zvoleff, A. (2024, Μάιος 2). Global hotspots of climate-related disasters. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://doi.org/10.1016/j.ijdr.2024.104488>
- Elmasri, R., & Navathe, S. B. (2016). Fundamentals of Database Systems (7th ed.). Pearson.
- GeeksforGeeks. (2025, Ιούλιος 19). ETL Process in Data Warehouse. Ανακτήθηκε 25 Αυγούστου 2025 από: <https://www.geeksforgeeks.org/dbms/etl-process-in-data-warehouse/>

- Inmon, W. (2005). Building the Data Warehouse (4th ed.). Wiley.
- Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://doi.org/10.1016/j.compag.2017.09.037>
- Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit (3rd ed.). Wiley.
- Lobell, D., & Gourdji, S. (2012, Οκτώβριος 10). The Influence of Climate Change on Global Crop Productivity. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://doi.org/10.1104/pp.112.208298>
- Lucidchart. What is an Entity Relationship Diagram (ERD)? Ανακτήθηκε 23 Αυγούστου 2025 από: <https://www.lucidchart.com/pages/er-diagrams>
- Munawar. (2021). Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development. 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI). IEEE. <https://doi.org/10.1109/ICCSAI53272.2021.9609770>
- Nambiar, A., & Mundra, D. (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. Big Data and Cognitive Computing, 6(4), 132. Ανακτήθηκε 23 Αυγούστου 2025 από: <https://doi.org/10.3390/bdcc6040132>
- NOAA. (2024, Ιούνιος 16). What are El Niño and La Niña? Ανακτήθηκε 22 Αυγούστου 2025 από: <https://oceanservice.noaa.gov/facts/ninonina.html>
- NOAA National Centers for Environmental Information (NCEI). (2025). Storm Events Database. NOAA. <https://www.ncdc.noaa.gov/stormevents/>
- Oracle. (2020, Νοέμβριος 24). What is a Database? Ανακτήθηκε 23 Αυγούστου 2025 από: <https://www.oracle.com/database/what-is-database/>
- Otto, F., Minnerop, P., Raju, E., Harrington, L. J., Stuart-Smith, R. F., Boyd, E., James, R., Jones, R., & Laut, K. C. (2022, Ιούνιος 23). Causality and the fate of climate litigation: The role of the social superstructure narrative. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://doi.org/10.1111/1758-5899.13113>
- Patairya, D. K. (2024, Νοέμβριος 7). Star Schema vs. Snowflake Schema Explained. Ανακτήθηκε 23 Αυγούστου 2025 από: <https://builtin.com/articles/star-schema-vs-snowflake-schema>
- Rawat, A., Kumar, D., & Khatri, B. S. (2023, Οκτώβριος 31). A review on climate change impacts, models, and its consequences on different sectors: A systematic approach.

- Journal of Water and Climate Change, 15(1), 104–126. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://doi.org/10.2166/wcc.2023.536>
- Schmidt, W., & Strauss, B. (2023, Δεκέμβριος 1). The Power of Data to Fight Climate Destruction. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://time.com/6341252/climate-destruction-data-cop28/>
- TheDatapedia. (2023, Δεκέμβριος 24). Star, Snowflake and Galaxy Schema in Datawarehouse with examples. Ανακτήθηκε 23 Αυγούστου 2025 από: <https://thedatape-dia.com/2023/12/24/star-snowflake-and-galaxy-schema-in-datawarehouse-with-examples/>
- Tomar, J. (2025, Ιανουάριος 31). Data Transformation: Types, Process, Benefits & Definition. Ανακτήθηκε 25 Αυγούστου 2025 από: <https://www.guvi.in/blog/data-trans-formation-types-and-process/>
- Tschumi, E., & Zscheisler, J. (2019, Δεκέμβριος 4). Countrywide climate features during recorded climate-related disasters. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://link.springer.com/article/10.1007/s10584-019-02556-w>
- Vassiliadis, P., Simitsis, A., & Skiadopoulos, S. (2002). Conceptual modeling for ETL processes. In Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP (DOLAP '02) (pp. 14–21). ACM. <https://doi.org/10.1145/583890.583893>
- Wohl, E. E., Pulwarty, R. S., & Zhang, J. Y. (2000, Οκτώβριος 10). Assessing climate impacts. Ανακτήθηκε 22 Αυγούστου 2025 από: <https://doi.org/10.1073/pnas.97.21.11141>
- Yung, Z. (2025, Ιούνιος 16). What Is ETL and How It Works? Ανακτήθηκε 22 Αυγούστου 2025 από: <https://blog.coupler.io/what-is-etl/>

Παράρτημα Α: Δημιουργία όψης

```
CREATE VIEW vw_ClimateEvents_Analysis AS
```

```
SELECT
```

```
    ce.ClimateEventKey,  
    ce.EpisodeID,  
    ce.EventID,  
    ce.BeginDateTime,  
    ce.EndDateTime,  
    ce.EventDurationHours,  
    ce.InjuriesDirect,  
    ce.DeathsDirect,  
    ce.DamageProperty,  
    ce.DamageCrops,  
    ce.TotalDamage,  
    ce.TotalCasualties,  
    ce.DamageCategory,  
    ce.ImpactLevel,  
    d.FullDate,  
    d.Year,  
    d.Month,  
    d.MonthName,  
    d.Season,  
    s.StateCode,  
    s.StateName,  
    s.Region,  
    s.IsCoastal,  
    et.EventTypeName,  
    et.EventCategory,  
    l.CZName,  
    l.BeginLat,  
    l.BeginLon,
```

```
    l.LocationType
FROM CLIMATE_EVENTS ce
JOIN Dim_Date d ON ce.DateKey = d.DateKey
JOIN Dim_State s ON ce.StateKey = s.StateKey
JOIN Dim_Event_Type et ON ce.EventTypeKey = et.EventTypeKey
JOIN Dim_Location l ON ce.LocationKey = l.LocationKey
```