ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Πτυχιακή Εργασία

Τίτλος Πτυχιακής Εργασίας	Αυτόματη άντληση, μετασχηματισμός και απεικόνιση επιχειρηματικών δεδομένων με τεχνολογίες Επιχειρηματικής Ευφυΐας (Business Intelligence). Automatic Extraction, Transformation, and Visualization of Business Data with Business Intelligence technologies.
Ονοματεπώνυμο Φοιτητή	Νικόλαος Στυλιδιώτης
Πατρώνυμο	Ηλίας
Αριθμός Μητρώου	Π/19166
Επιβλέπων	Μαρία Βίβρου

Ημερομηνία Παράδοσης: Φεβρουάριος 2026

Copyright ©

Περίληψη

Στην παρούσα εργασία παρουσιάζεται η σχεδίαση και υλοποίηση ενός αυτοματοποιημένου συστήματος άντλησης, μετασχηματισμού και οπτικοποίησης επιχειρηματικών δεδομένων, με χρήση τεχνολογιών Επιχειρηματικής Ευφυΐας (Business Intelligence). Η κύρια πηγή δεδομένων είναι το δημόσιο αποθετήριο **SEC EDGAR**, ενώ η ροή δεδομένων υλοποιείται με την χρήση του εργαλείου ανοιχτού κώδικα ***Airbyte*** (μεταφορά δεδομένων), ***PostgreSQL*** (αποθήκευση δεδομένων), ***dbt*** (μετασχηματισμός-επεξεργασία δεδομένων) και ***Power BI*** (αναφορές και διαδραστικά dashboards). Εφαρμόζεται η μεθοδολογία «**Quarter Bucket**» για δίκαιες και σταθερές συγκρίσεις ανά ημερολογιακό τρίμηνο, αντιμετωπίζοντας τυχόν κενά και ασυγχρονίες δημοσιεύσεων. Τα αποτελέσματα δείχνουν σταθερότητα ταυτοτήτων (IDs), βελτίωση απόδοσης με incremental/merge, και αναφορές που ακολουθούν **IBCS** λογική ουδετερότητας για πολύ μικρές μεταβολές.

Λέξεις Κλειδιά: Business Intelligence, EDGAR, *Airbyte*, *dbt*, *Power BI*, ELT, Quarter Bucket, IBCS

Abstract

In this undergraduate thesis, the design and implementation of an automated system for the extraction, transformation, and visualization of business data is presented, using Business Intelligence (BI) technologies. The primary data source is the public Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval (SEC EDGAR) repository. In contrast, the overall data pipeline is implemented using open-source tools, specifically Airbyte (for data extraction and ingestion), PostgreSQL (for data storage), a data build tool (dbt) (for data transformation and processing), and Power BI (for reporting and interactive dashboards). To address issues related to temporal gaps, data asynchrony, and delayed financial disclosures, the “Quarter Bucket” methodology is applied, enabling fair and consistent comparisons across calendar quarters. The results of the study demonstrate the effectiveness of the proposed architecture in terms of identity stability (IDs), performance improvement through incremental/merge strategies, and the production of analytical reports that adhere to the International Business Communication Standards (IBCS), incorporating neutrality principles for the treatment of small variations.

Keywords: Business Intelligence, EDGAR, *Airbyte*, *dbt*, *Power BI*, ELT, Quarter Bucket, IBCS

Πίνακας Περιεχομένων

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Εισαγωγή

Αντικείμενο – Στόχοι

Η παρούσα εργασία παρουσιάζει το σχεδιασμό και την υλοποίηση ενός ολοκληρωμένου πλαισίου αυτοματοποιημένης άντλησης, μετασχηματισμού και οπτικοποίησης επιχειρηματικών δεδομένων, βασισμένο αποκλειστικά σε ανοιχτά δεδομένα της αμερικανικής κεφαλαιαγοράς και συγκεκριμένα, αναρτημένα στο δημόσιο αποθετήριο της Ηλεκτρονικής Συλλογής, Ανάλυσης και Ανάκτησης Δεδομένων της Επιτροπής Κεφαλαιαγοράς (Securities and Exchange Commission's Electronic Data Gathering, Analysis, and Retrieval (SEC EDGAR)). Το αντικείμενο της εργασίας εντάσσεται στο πεδίο της Επιχειρηματικής Ευφυΐας (Business Intelligence (BI)) και επικεντρώνεται στη δημιουργία ενός επαναχρησιμοποιήσιμου, end-to-end συστήματος τύπου εξαγωγής – φόρτωσης – μετατροπής (Extract – Load – Transform (ELT))→BI, κατάλληλου για εφαρμογή σε πραγματικά αναλυτικά περιβάλλοντα [1], [2].

Κεντρικός στόχος της εργασίας είναι η τεχνολογική υλοποίηση ενός αυτοματοποιημένου pipeline ανοικτών δεδομένων, το οποίο καλύπτει ολόκληρο τον κύκλο ζωής των δεδομένων, από την άντληση με διαδικτυακές Διεπαφές Προγραμματισμού Εφαρμογών (Application Programming Interfaces (APIs)) έως και την παραγωγή αναφορών και διαδραστικών dashboards. Ιδιαίτερη έμφαση δίνεται στη χρήση εργαλείων ανοικτού κώδικα και δωρεάν τεχνολογιών με σκοπό τη διαφάνεια, επεκτασιμότητα και δυνατότητα αναπαραγωγής της προτεινόμενης λύσης [1], [2]. Οι επιμέρους στόχοι της εργασίας συνοψίζονται ως εξής:

- Η αυτοματοποιημένη άντληση δεδομένων από το SEC EDGAR μέσω της ανάπτυξης custom connector στο εργαλείο Airbyte.
- Ο αξιόπιστος μετασχηματισμός και η μοντελοποίηση των δεδομένων με χρήση ενός data build tool (dbt), σύμφωνα με τη λογική ELT.
- Η ανάπτυξη αναφορών και dashboards στο Power BI, αξιοποιώντας τη μεθοδολογία Quarter Bucket για δίκαιες και συγκρίσιμες αναλύσεις ανά πολιτεία και τρίμηνο.
- Η εφαρμογή των αρχών των Διεθνών Προτύπων Επιχειρηματικής Επικοινωνίας (International Business Communication Standards (IBCS)), όπως η ουδέτερη ζώνη, η καθαρή σήμανση και η συνεπής κλίμακα, στη σχεδίαση των οπτικοποιήσεων.

Παρόμοιες σύγχρονες εργασίες

Στη διεθνή βιβλιογραφία έχουν παρουσιαστεί διάφορες προσεγγίσεις αξιοποίησης του EDGAR και του XBRL για ανάλυση χρηματοοικονομικών δεδομένων, όπως το OpenEDGAR framework για μαζική επεξεργασία υποβολών της SEC [3] ή μελέτες που εστιάζουν στη βελτίωση της ποιότητας και της χρησιμότητας των XBRL δεδομένων για αναλυτές και ερευνητές [4]. Η παρούσα εργασία διαφοροποιείται καθώς εστιάζει σε ένα πρακτικό, end-to-end BI εργαλείο σε επίπεδο πολιτείας, με έμφαση στην αυτοματοποίηση της ροής και στη μοντελοποίηση για *Power BI*.

Ανάγκη για εργασίες στο αντικείμενο

Η ύπαρξη αξιόπιστων και αυτοματοποιημένων ροών δεδομένων που συνδέουν ανοικτές πηγές, όπως το SEC EDGAR, με εργαλεία Επιχειρηματικής Ευφυΐας είναι ιδιαίτερα χρήσιμη τόσο για την ακαδημαϊκή έρευνα όσο και για την αγορά. Παρότι υπάρχουν εμπορικές λύσεις που παρέχουν χρηματοοικονομικά δεδομένα, αυτές συχνά δεν τεκμηριώνουν αναλυτικά ολόκληρο το τεχνικό pipeline:

API → αποθήκευση → μετασχηματισμοί → BI ενώ βασίζονται σε ιδιόκτητα εργαλεία και κλειστά συστήματα [4], [5].

Η παρούσα εργασία καλύπτει αυτό το κενό, παρουσιάζοντας ένα πλήρως τεκμηριωμένο παράδειγμα end-to-end ροής δεδομένων, με ελάχιστη εξάρτηση από ιδιόκτητα λογισμικά και αποκλειστική χρήση δωρεάν και ανοικτών εργαλείων.

Κίνητρο – Συνεισφορά

Οι αναλύσεις χρηματοοικονομικών δεδομένων που βασίζονται στο SEC EDGAR έρχονται συχνά αντιμέτωπες με ζητήματα ετεροχρονισμένων δημοσιεύσεων και ασυγχρονίας μεταξύ εταιρικών περιόδων [4], [5]. Η βασική συνεισφορά της εργασίας συνοψίζεται στα εξής:

- Ανάπτυξη pipeline ανοικτών δεδομένων από HTTP API έως και BI αναφορές, χωρίς τη χρήση ιδιόκτητων εργαλείων.
- Βελτίωση αποδοτικότητας με εφαρμογή επαυξητικών συγχωνεύσεων δεδομένων (incremental/merge upserts) στον πίνακα γεγονότων (fact table), διατηρώντας παράλληλα την ακεραιότητα των αναφορών (referential integrity).
- Εφαρμογή της μεθοδολογίας Quarter Bucket, επιτρέποντας ευαισθητοποιημένες σε σύνολα δεδομένων (dataset-aware) συγκρίσεις που μειώνουν χρονικά κενά και τυχόν αναλυτικές αδικίες.

Η επιλογή του SEC EDGAR και γενικότερα των ανοικτών δεδομένων (open data) της κεφαλαιαγοράς συνδέεται άμεσα με ζητήματα διακυβέρνησης δεδομένων (open data governance) και διαφάνειας. Η χρήση και δημοσιοποίηση τυποποιημένων οικονομικών καταστάσεων σε ανοικτά, μη ιδιόκτητα πρότυπα, όπως το XBRL, επιτρέπει τη συστηματική ανάλυση της χρηματοοικονομικής συμπεριφοράς των εισηγμένων εταιρειών σε ερευνητές και επαγγελματίες. Το εργαλείο που αναπτύσσεται σε αυτήν την εργασία αξιοποιεί αυτήν την υποδομή, αυτοματοποιώντας την άντληση, τον καθαρισμό και την απεικόνιση των δεδομένων, διευκολύνοντας τον εντοπισμό τάσεων, ανωμαλιών και μοτίβων σε επίπεδο αγοράς ή πολιτείας [5], [6].

Διάρθρωση

Στο Κεφάλαιο 2 συνοψίζεται το θεωρητικό υπόβαθρο, οι τεχνολογίες που χρησιμοποιούνται στην εργασία καθώς και άλλες επιστημονικές εργασίες σε παρόμοιο πεδίο και τυχόν ελλείψεις. Το Κεφάλαιο 3 περιγράφει τον σχεδιασμό της εφαρμογής, περισσότερες λεπτομέρειες-πληροφορίες και επίσης μερικά απαραίτητα UML διαγράμματα. Η αναλυτική υλοποίηση παρουσιάζεται στο Κεφάλαια 4 ενώ η λειτουργικότητά της και οι BI αναφορές αναπτύσσονται στο κεφάλαιο 5. Το κεφάλαιο 6 περιέχει το εγχειρίδιο χρήσης και στα κεφάλαια 7 και 8 γίνεται η αξιολόγηση απόδοσης και παρουσιάζονται οι περιορισμοί και πιθανή μελλοντική ανάπτυξη της εφαρμογής. Τέλος τα συμπεράσματα βρίσκονται στο Κεφάλαιο 9.

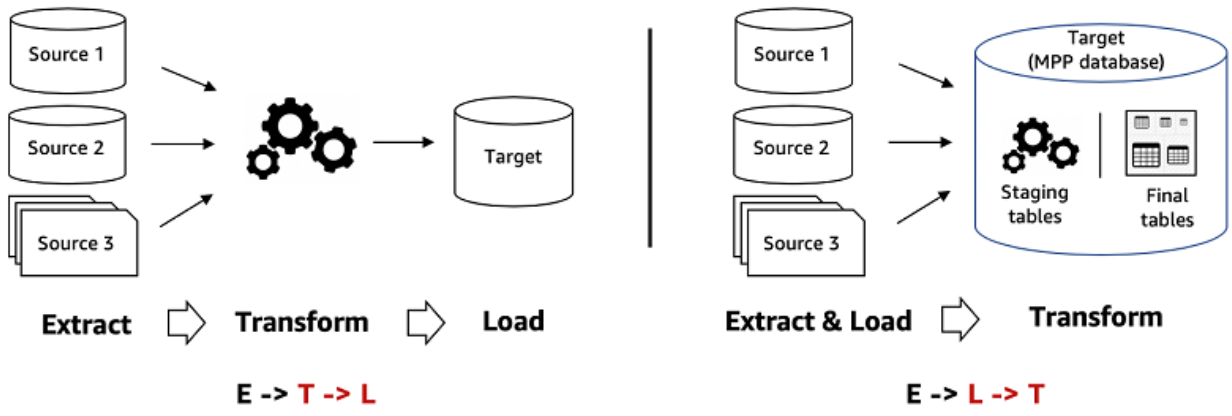
Ανασκόπηση Πεδίου

Επιχειρηματική Ευφυΐα (BI) & αρχιτεκτονικές ELT

Η Επιχειρηματική Ευφυΐα (*Business Intelligence - BI*) είναι ένα βασικό εργαλείο της σύγχρονης επιστήμης των δεδομένων, το οποίο ασχολείται με τη μετατροπή των ακατέργαστων δεδομένων σε χρήσιμη και αξιοποιήσιμη πληροφορία [7],[8]. Στόχος είναι η υποστήριξη των λήψεων τεκμηριωμένων επιχειρηματικών αποφάσεων μέσω της συλλογής, αποθήκευσης, επεξεργασίας, ανάλυσης και οπτικοποίησης των δεδομένων. Με αυτόν τον τρόπο, οι χρήστες μπορούν να κατανοήσουν της λειτουργίας τους, να εντοπίσουν πρότυπα και τάσεις και να καθορίσουν στρατηγικές που βασίζονται σε πραγματικά στοιχεία και όχι σε εκτιμήσεις [7], [8] .

Η BI διαδικασία συνήθως περιλαμβάνει την συγκέντρωση δεδομένων από πηγές, όπως λειτουργικά συστήματα, εφαρμογές διαχείρισης πελατειακών σχέσεων (*CRM*), συστήματα διαχείρισης επιχειρησιακών πόρων (*ERP*), καθώς και δεδομένα προερχόμενα από το διαδίκτυο ή τα μέσα κοινωνικής δικτύωσης. Τα δεδομένα αυτά αποθηκεύονται σε αποθήκες δεδομένων (*data warehouses*), όπου πραγματοποιείται μοντελοποίηση και καθαρισμός, ώστε να είναι αξιόπιστα και έτοιμα για ανάλυση. Τέλος, με την χρήση εργαλείων BI, δημιουργούνται *dashboards*, αναφορές και γραφικές απεικονίσεις που διευκολύνουν την κατανόηση της πληροφορίας [9].

Οι αρχιτεκτονικές Extract–Load–Transform (ELT) είναι σημαντικό στοιχείο των σύγχρονων υποδομών δεδομένων [10], [11]. Η προσέγγιση ELT διαφοροποιείται από το παραδοσιακό μοντέλο Extract–Transform–Load (ETL) ως προς τη σειρά εκτέλεσης των επιμέρους σταδίων. Συγκεκριμένα, στην αρχιτεκτονική ELT τα δεδομένα αρχικά εξάγονται από τις πηγές τους (Extract), στη συνέχεια φορτώνονται στην αποθήκη δεδομένων (Load) και τέλος υφίστανται μετασχηματισμούς (Transform) εντός της βάσης δεδομένων ή του data warehouse. Κατά αυτόν τον τρόπο αξιοποιείται η αυξημένη υπολογιστική ισχύς και οι εκτενείς δυνατότητες παραλληλισμού των σύγχρονων cloud πλατφορμών, καθιστώντας τη διαδικασία ταχύτερη, αποδοτικότερη και πιο ευέλικτη (Εικόνα 1) [10], [11], [12], [13].



Εικόνα 1 ETL vs ELT [14]

Αυτόματη άντληση, μετασχηματισμός και απεικόνιση επιχειρηματικών δεδομένων με τεχνολογίες Επιχειρηματικής Ευφυΐας.

Βασικό πλεονέκτημα της αρχιτεκτονικής ELT έγκειται στη μεταφορά της διαδικασίας μετασχηματισμού στα συστήματα αποθήκευσης δεδομένων. Έτσι μειώνεται η ανάγκη για εξωτερικά συστήματα ETL και περιορίζεται η πολυπλοκότητα της συνολικής αρχιτεκτονικής. Επίσης, η προσέγγιση αυτή διευκολύνει τη συντήρηση και τον έλεγχο των ροών δεδομένων, καθώς οι μετασχηματισμοί είναι διαφανείς και επαναλήψιμοι [10], [11]. Η φιλοσοφία αυτή ενσωματώνεται μέσω εργαλείων όπως το data build tool (dbt), το οποίο επιτρέπει τον ορισμό μετασχηματισμών σε μορφή κώδικα SQL, υποστηρίζοντας παράλληλα το version control, τον έλεγχο-testing και τη τεκμηρίωση [10], [11].

Εν κατακλείδι, το Business Intelligence (BI) αλλά και οι αρχιτεκτονικές ELT λειτουργούν συμπληρωματικά. Ο συνδυασμός τους μετατοπίζει το επίκεντρο από την απλή αποθήκευση δεδομένων στην ουσιαστική αξιοποίησή τους, επιτρέποντας τη ταχύτερη και πιο αξιόπιστη ανάλυση και παρέχοντας στους οργανισμούς τη γνώση που απαιτείται για τη λήψη στρατηγικών αποφάσεων και την διατήρηση της ανταγωνιστικότητας.

Ανοιχτά δεδομένα, SEC EDGAR και XBRL

Η έννοια των Ανοιχτών Δεδομένων (Open Data) αναφέρεται στη δημόσια διάθεση δεδομένων από κυβερνητικούς, θεσμικούς ή εταιρικούς φορείς, με τρόπο που επιτρέπει την ελεύθερη πρόσβαση, επαναχρησιμοποίηση και ανάλυσή τους [13], [15]. Στον χώρο της χρηματοοικονομικής πληροφόρησης, ένα από τα πλέον καθιερωμένα και σημαντικά συστήματα ανοιχτών δεδομένων είναι το σύστημα Ανάκτησης Δεδομένων της Επιτροπής Κεφαλαιαγοράς (Electronic Data Gathering, Analysis, and Retrieval (EDGAR)), το οποίο λειτουργεί υπό την εποπτεία της U.S. Securities and Exchange Commission (SEC) [16].

Το EDGAR συγκεντρώνει και δημοσιεύει υποβολές εταιρικών εκθέσεων όπως οι 10-K (ετήσιες αναφορές) και 10-Q (τριμηνιαίες αναφορές), καθώς και άλλες δηλώσεις που σχετίζονται με τη χρηματοοικονομική κατάσταση και τις εταιρικές δραστηριότητες των εισηγμένων οργανισμών στις Ηνωμένες Πολιτείες. Οι αναφορές αυτές είναι δημόσια διαθέσιμες και αποτελούν πολύτιμη πηγή δεδομένων για επενδυτές, ερευνητές ή αναλυτές, καθώς παρέχουν λεπτομερείς αναφορές στα οικονομικά αποτελέσματα, τις επενδύσεις, τις στρατηγικές και τους κινδύνους που αντιμετωπίζουν οι εταιρείες [4], [5], [6], [16].

Ένα σημαντικό χαρακτηριστικό/πλεονέκτημα του EDGAR είναι η χρήση του προτύπου eXtensible Business Reporting Language (XBRL), ενός διεθνούς, ανοιχτού προτύπου ανταλλαγής χρηματοοικονομικών δεδομένων που επιτρέπει ιδιαίτερα τη τυποποιημένη αναπαράσταση οικονομικών εννοιών. Το XBRL βασίζεται στη γλώσσα XML και έχει σχεδιαστεί ώστε να υποστηρίζει την ηλεκτρονική περιγραφή, ανταλλαγή, αλλά και κυρίως την αυτοματοποιημένη ανάλυση επιχειρηματικών πληροφοριών [16], [17], [18], [19]. Πρόκειται για ένα σημασιολογικό πλαίσιο που αποδίδει συγκεκριμένο νόημα σε κάθε οικονομικό στοιχείο. Κάθε έννοια ορίζεται μέσα σε μία ταξινομία (taxonomy), δηλαδή ένα οργανωμένο λεξικό εννοιών που καθορίζει τις σχέσεις και τη λογική σύνδεση μεταξύ τους.

Μέσω του XBRL, τα δεδομένα των εταιρικών υποβολών καθίστανται άμεσα επεξεργάσιμα από συστήματα ανάλυσης, διευκολύνοντας την αυτοματοποίηση της εξαγωγής των δεδομένων και τη συγκριτική μελέτη οικονομικών στοιχείων μεταξύ διαφορετικών οργανισμών και χρονικών περιόδων [16], [17], [18], [19].

Ωστόσο, η τριμηνιαία συχνότητα των υποβολών (π.χ. των 10-Q) σε συνδυασμό με τον ετεροχρονισμό στην ανάρτηση και δημοσίευση των δεδομένων, δημιουργεί σημαντικές προκλήσεις ως προς τη χρονική ευθυγράμμιση των αναφορών. Για το λόγο αυτό, στη παρούσα εργασία υιοθετείται η χρήση ενός μοντέλου “Quarter Bucket”, δηλαδή η ομαδοποίηση των δεδομένων ανά ημερολογιακό τρίμηνο. Το μοντέλο αυτό επιτρέπει τις συγκρίσεις εταιρικών επιδόσεων με ενιαίο ρυθμό αναφοράς (reporting cadence), βελτιώνοντας τη συνοχή των αναλύσεων και στοχεύοντας στην αποφυγή προβλημάτων που προκαλούνται από ασύγχρονα δεδομένα [18], [20].

BI Dashboards και πρότυπο IBCS

Τα BI dashboards αποτελούν τον βασικό μηχανισμό μέσω του οποίου οι χρήστες βλέπουν και ερμηνεύουν τα αποτελέσματα μίας ανάλυσης. Σε αντίθεση με τις παραδοσιακές, στατικές αναφορές που βασίζονταν κυρίως σε εκτενές κείμενο, τα σύγχρονα BI dashboards ζωντανεύουν τα δεδομένα αξιοποιώντας διαδραστικά γραφήματα, φίλτρα και οπτικά στοιχεία επιτρέποντας την άμεση κατανόηση των δεδομένων από τον χρήστη [21], [22]. Η βιβλιογραφία επισημαίνει ότι ο σχεδιασμός τους δεν είναι ουδέτερος, καθώς επιλογές όπως ο τρόπος επιλογής χρωμάτων, οι κλίμακες, οι τύποι διαγραμμάτων και η ομαδοποίηση των πληροφοριών επηρεάζουν άμεσα τη ταχύτητα και την ακρίβεια με την οποία ο χρήστης αντιλαμβάνεται τα μηνύματα που μεταφέρουν τα δεδομένα [23], [24], [25].

Για τον λόγο αυτό έχουν αναπτυχθεί διεθνή πρότυπα, με σημαντικότερο τα International Business Communication Standards (IBCS). Το IBCS προτείνει ένα σύνολο κανόνων και συμβάσεων για την ομοιόμορφη και σαφή παρουσίαση επιχειρηματικών αναφορών και dashboards [26], [27]. Ενδεικτικά, περιλαμβάνει:

- Τη συνεπή χρήση χρωμάτων για θετικές/αρνητικές μεταβολές και διαφορετικά σενάρια (π.χ. actual, plan, forecast).
- Τη διάκριση των τύπων γραφημάτων ανάλογα με το ζητούμενο (γραμμικά διαγράμματα για τάσεις, ραβδογραφήματα για συγκρίσεις μεγεθών, waterfall για ανάλυση γεφυρών κ.ά.).
- Τα σαφή και ευανάγνωστα labels, τίτλους και υποτίτλους που εξηγούν τι ακριβώς απεικονίζεται.
- Σταθερές κλίμακες στους άξονες, ώστε η οπτική σύγκριση ανάμεσα σε διαφορετικά διαγράμματα να είναι δίκαιη.
- Τον περιορισμό του “οπτικού θορύβου” (περιττές γραμμές πλέγματος, πολλαπλά χρώματα χωρίς λόγο, 3D εφέ κ.λπ.).

Στην παρούσα εργασία έχουν ενσωματωθεί αρκετές από τις βασικές αρχές του προτύπου IBCS πάνω στον σχεδιασμό των dashboards στο Power BI. Στα γραφήματα

χρησιμοποιείται ουδέτερη παλέτα χρωμάτων (π.χ. μαύρο/γκρι για τις καμπύλες του US average) και επίσης εμφανίζεται η επιλογή πράσινου/κόκκινου χρώματος μόνο για την ανάδειξη θετικών και αρνητικών αποκλίσεων. Η κλίμακα των αξόνων παραμένει σταθερή όταν συγκρίνονται δύο πολιτείες (State Comparison), ώστε ο χρήστης να μπορεί να κατανοήσει εύκολα ποια από τις δύο έχει μεγαλύτερη μεταβολή ή επίπεδο. Επιπλέον, έχει οριστεί μία ουδέτερη ζώνη (neutral zone) για τα ποσοστά μεταβολής (π.χ. |Δ%| < 0,01%), όπου οι πολύ μικρές αποκλίσεις αντιμετωπίζονται ως πρακτικά μηδενικές και δεν χρωματίζονται έντονα.

Τέλος, οι πίνακες κατάταξης (state ranks) και οι δείκτες coverage (%) ακολουθούν σταθερή μορφή ώστε ο χρήστης να μπορεί να τους διαβάζει με τον ίδιο τρόπο σε όλες τις σελίδες: πρώτα εμφανίζεται η τωρινή τιμή, στη συνέχεια η ποσοστιαία μεταβολή σε σχέση με το baseline και, όπου έχει νόημα, η θέση στην κατάταξη μεταξύ όλων των πολιτειών. Με αυτόν τον τρόπο, επιτυγχάνεται ένας βαθμός τυποποίησης που διευκολύνει τη σύγκριση μεταξύ διαφορετικών δεικτών και χρονικών περιόδων [26], [27].

Τεχνολογίες

Η επεξεργασία και αξιοποίηση των δεδομένων του EDGAR προϋποθέτουν την ύπαρξη μιας τεχνολογικής υποδομής που να υποστηρίζει τη ροή, τον μετασχηματισμό και την απεικόνιση των πληροφοριών. Στην εργασία χρησιμοποιούνται κάποιες σύγχρονες ανοιχτές πλατφόρμες και εργαλεία ανοιχτού κώδικα, καθένα από τα οποία επιτελεί συγκεκριμένο και αλληλοσυμπληρούμενο ρόλο [4], [5], [6], [16]:

Airbyte: Για την άντληση των δεδομένων επιλέχθηκε το Airbyte, καθώς είναι εργαλείο ανοικτού κώδικα και προσφέρει data movement. Αυτό επιτρέπει τη σύνδεση με εκατοντάδες πηγές δεδομένων μέσω connectors. Επίσης υποστηρίζει χρονοπρογραμματισμό, αυτοματοποίηση ροών εξαγωγής και φόρτωσης. Η χρήση του Airbyte καθιστά δυνατή τη συνεχή ενημέρωση των δεδομένων από το EDGAR στη βάση δεδομένων (postgreSQL). [25], [28], [29], [30]

***dbt* (data build tool)**: Το dbt χρησιμοποιήθηκε για τον μετασχηματισμό των δεδομένων, επιτρέποντας την εφαρμογή πρακτικών version control, *modularity* και *testing*. Επιτρέπει τη δημιουργία εξαρτήσεων μεταξύ μοντέλων, την εκτέλεση *incremental builds* και τη διαχείριση μετασχηματισμών με διαφάνεια και επαναληψιμότητα. Με τον τρόπο αυτό, διασφαλίζεται ότι τα δεδομένα του EDGAR υποβάλλονται σε αξιόπιστη και επαναλήψιμη προ-επεξεργασία πριν καταλήξουν στο στάδιο της ανάλυσης [20], [31].

Power BI: Το *Power BI* είναι ενα γνωστό *BI* εργαλείο της Microsoft, που παρέχει ισχυρές δυνατότητες ανάλυσης, οπτικοποίησης και διαδραστικής διερεύνησης δεδομένων. Μέσω των *DAX* υπολογισμών και του *Power Query*, επιτρέπει τη δημιουργία δυναμικών dashboards, προσφέροντας στους χρήστες τη δυνατότητα να παρακολουθούν εταιρικές τάσεις και να συγκρίνουν αποτελέσματα μεταξύ διαφορετικών περιόδων ή εταιρειών [21], [24], [25], [32].

IBCS (International Business Communication Standards): Οι αρχές του IBCS εφαρμόζονται για τη τυποποίηση των επιχειρησιακών αναφορών, διασφαλίζοντας ότι η οπτική παρουσίαση των δεδομένων είναι σαφής, συνεπής και κατανοητή. Οι οδηγίες αυτές περιλαμβάνουν έννοιες όπως η «ουδέτερη ζώνη» (*neutral zone*), η χρήση σαφών ετικετών

(*labels*) και ενιαίων κλιμάκων (*scales*), με σκοπό τη βελτίωση της αναγνωσιμότητας και της συγκρισιμότητας των αποτελεσμάτων. Η συμμόρφωση με τα πρότυπα IBCS ενισχύει τη διαφάνεια και την αξιοπιστία της πληροφόρησης. [26], [27].

Συνολικά, ο συνδυασμός των ανοιχτών δεδομένων του EDGAR με τις παραπάνω τεχνολογίες και τυποποιημένες αρχές οδηγεί στη δημιουργία ενός ολοκληρωμένου πλαισίου επιχειρησιακής ανάλυσης. Μέσα από τη σωστή συλλογή, τον αξιόπιστο μετασχηματισμό και την τυποποιημένη απεικόνιση των δεδομένων, είναι δυνατή η παραγωγή συγκρίσιμων χρηματοοικονομικών αναφορών.

Εμβάθυνση στην Αρχιτεκτονική του XBRL και στις Δομές Δεδομένων

Αν και στο προηγούμενο τμήμα αναφέρθηκε ο γενικός ορισμός του XBRL, για την πλήρη κατανόηση των προκλήσεων που ανακύπτουν κατά την άντληση δεδομένων από το SEC EDGAR απαιτείται εμβάθυνση στην εσωτερική αρχιτεκτονική του προτύπου. Το XBRL δεν αποτελεί απλώς μια παραλλαγή της μορφής XML, αλλά ένα πολύπλοκο σύστημα αναπαράστασης επιχειρηματικών δεδομένων που βασίζεται σε δύο θεμελιώδεις πυλώνες: τις Ταξινομίες (Taxonomies) και τα Έγγραφα Στιγμιότυπου (Instance Documents). Η κατανόηση αυτής της διάκρισης είναι κρίσιμη για την αιτιολόγηση των τεχνικών επιλογών που υιοθετήθηκαν κατά την υλοποίηση του connector [16], [17], [18], [19], [33].

Η Δομή των Ταξινομιών και τα Linkbases

Η ταξινομία XBRL λειτουργεί ουσιαστικά ως το "λεξικό" των δεδομένων. Σε αντίθεση με ένα απλό σύνολο ορισμών, οι έννοιες της ταξινομίας συνδέονται πολυδιάστατα μεταξύ τους μέσω βάσεων συνδέσμων (linkbases). Αυτή η σύνθετη δομή καθιστά την εξαγωγή δεδομένων από το EDGAR απαιτητική, καθώς προϋποθέτει τη σωστή αποκωδικοποίηση των σχέσεων μεταξύ των στοιχείων [34]. Τα κυριότερα linkbases που συναντώνται στις υποβολές U.S.-Generally Accepted Accounting Principles (GAAP) είναι:

- **Presentation Linkbase:** Καθορίζει την ιεραρχική δομή παρουσίασης των στοιχείων, όπως ακριβώς θα εμφανίζονταν σε μια εκτυπωμένη οικονομική κατάσταση (π.χ. το "Ενεργητικό" περιλαμβάνει το "Κυκλοφορούν Ενεργητικό"). Για την εργασία, αυτό το linkbase εξηγεί γιατί τα δεδομένα δεν έρχονται ως ένας επίπεδος πίνακας (flat table) αλλά ως ιεραρχικά δέντρα.
- **Calculation Linkbase:** Περιγράφει τις μαθηματικές σχέσεις μεταξύ των στοιχείων (π.χ. *Gross Profit = Revenues - Cost of Goods Sold*). Αυτό είναι ιδιαίτερα σημαντικό για τον έλεγχο ποιότητας των δεδομένων (data validation), καθώς επιτρέπει την επαλήθευση της συνέπειας των αριθμών που αντλούνται.
- **Definition Linkbase:** Ορίζει διαστάσεις (dimensions) και hypercubes. Στις σύγχρονες υποβολές, πολλά δεδομένα δεν είναι απλές τιμές, αλλά εξαρτώνται από διαστάσεις (π.χ. "Έσοδα ανά Γεωγραφικό Τομέα"). Η αγνόηση αυτού του linkbase κατά την άντληση θα οδηγούσε σε λανθασμένη συσσώρευση (aggregation) τιμών που ανήκουν σε διαφορετικά contexts.

- **Label Linkbase:** Συνδέει τις τεχνικές ονομασίες (π.χ. us-gaap:Assets) με αναγνώσιμες ετικέτες (π.χ. "Total Assets").

Το πρόβλημα των Επεκτάσεων (Taxonomy Extensions)

Μια ακόμα σημαντική πρόκληση στην ανάλυση δεδομένων του EDGAR, η οποία καθιστά απαραίτητη τη χρήση ευέλικτων διαδικασιών μετασχηματισμού (όπως το dbt), είναι η δυνατότητα των εταιρειών να δημιουργούν επεκτάσεις (extensions) [5], [15], [35]. Το πρότυπο US-GAAP επιτρέπει στις εταιρείες, όταν δεν καλύπτονται από τις τυπικές ετικέτες (tags), να ορίζουν δικές τους custom tags. Η πρακτική αυτή εισάγει σημαντική ετερογένεια στα δεδομένα [5], [15], [16], [18], [35].

Για παράδειγμα, ενώ η πλειοψηφία των εταιρειών χρησιμοποιεί το tag «us-gaap:Revenues», μια εταιρεία μπορεί να δημιουργήσει ένα δικό της, custom tag, όπως παραδείγματος χάριν «my-company:RevenueFromSpecialServices». Στην παρούσα εργασία, η διαχείριση αυτής της πολυπλοκότητας αντιμετωπίστηκε στο στάδιο του Staging, όπου έγινε η τυποποίηση των βασικών μετρικών, αγνοώντας τα εξαιρετικά εξειδικευμένα custom tags που θα εισήγαγαν θόρυβο (noise) στην ανάλυση συγκριτικής αξιολόγησης (benchmarking).

Διακυβέρνηση Δεδομένων και Ιχνηλασιμότητα σε περιβάλλοντα ELT

Η μετάβαση από το παραδοσιακό ETL στο σύγχρονο ELT δεν είναι απλώς μια αλλαγή στη σειρά των τεχνικών βημάτων, αλλά αντικατοπτρίζει μια ευρύτερη αλλαγή φιλοσοφίας ως προς τη διαχείριση δεδομένων [10], [11], [13], [30], [36]. Στο πλαίσιο της παρούσας εργασίας, η επιλογή του ELT είναι καθοριστική, καθώς επηρεάζει άμεσα τόσο την αξιοπιστία των οικονομικών αποτελεσμάτων όσο και τη δυνατότητα ιχνηλάτησης της προέλευσή τους (data lineage) [10], [11], [13], [30], [36], [37].

Η σημασία του Data Lineage (Ιχνηλασιμότητα)

Στην ανάλυση χρηματοοικονομικών δεδομένων, η δυνατότητα να απαντηθεί το ερώτημα "από πού προήλθε αυτός ο αριθμός;" είναι κρίσιμη. Σε παλαιότερα συστήματα ETL, οι μετασχηματισμοί γίνονταν συχνά εντός "μαύρων κουτιών" (ιδιόκτητα σενάρια «proprietary scripts» ή Εργαλεία Βασισμένα σε Γραφική Διεπαφή Χρήστη (Graphical User Interface (GUI-based tools)) πριν τα δεδομένα φτάσουν στη βάση, γεγονός που καθιστούσε δύσκολο τον εντοπισμό σφαλμάτων (π.χ. φταίει η πηγή ή ο μετασχηματισμός) [10], [13], [37], [38].

Με την προσέγγιση ELT που υλοποιήθηκε μέσω του *dbt*:

- **Τα Raw Data είναι απαραβίαστα:** Τα δεδομένα φορτώνονται στην *PostgreSQL* ακριβώς όπως παρελήφθησαν από το EDGAR API. Αυτό δημιουργεί ένα μόνιμο audit trail. Οποιοδήποτε λάθος στους υπολογισμούς μπορεί να διορθωθεί με αλλαγή του κώδικα SQL και επαναϋπολογισμό, χωρίς να χρειάζεται εκ νέου άντληση από την πηγή.
- **Ο Γράφος Εξαρτήσεων (DAG):** Το *dbt* δημιουργεί αυτόματα έναν Κατευθυνόμενο Ακυκλικό Γράφο (Directed Acyclic Graph - DAG), ο οποίος χαρτογραφεί τη ροή των

δεδομένων. Κάθε πίνακας (model) στο Data Warehouse γνωρίζει ακριβώς από ποιους πίνακες εξαρτάται. Αυτή η διαφάνεια είναι απαραίτητη για τη διασφάλιση της ποιότητας των δεδομένων (Data Quality Assurance) [17], [34], [39].

Αντιμετώπιση της Ολίσθησης Σχήματος (Schema Drift)

Στα ανοιχτά δεδομένα παρατηρείται συχνά το φαινόμενο της ολίσθησης σχήματος (schema drift), δηλαδή η απρόβλεπτη μεταβολή της δομής των δεδομένων από την πηγή (π.χ. η SEC προσθέτει ένα νέο πεδίο στο JSON output). Στην κλασική προσέγγιση ETL, μια τέτοια αλλαγή θα "έσπαγε" – οδηγούσε σε αστοχία του pipeline κατά την εξαγωγή. Στην προτεινόμενη αρχιτεκτονική ELT, το Airbyte εμφανίζει ανθεκτικότητα σε τέτοιες αλλαγές, προσαρμόζοντας αυτόματα τους πίνακες raw και μεταθέτοντας την διαχείριση της αλλαγής στο επίπεδο του dbt. Αυτό επιτρέπει στο σύστημα να συνεχίσει να λειτουργεί και να συλλέγει δεδομένα, ακόμα και αν χρειάζεται μελλοντική προσαρμογή στα downstream μοντέλα, εξασφαλίζοντας υψηλότερη διαθεσιμότητα [10], [13], [37], [38], [40], [41].

Ο Ρόλος του Open Source Λογισμικού

Η επιλογή εργαλείων ανοιχτού κώδικα (Airbyte, dbt, Python, PostgreSQL) για την υλοποίηση αυτής της εργασίας υπαγορεύτηκε αποκλειστικά από λόγους κόστους, αλλά εκφράζει μία σαφή και μεθοδολογική στάση υπέρ της αναπαραγωγιμότητας (reproducibility). Σε αντίθεση με εμπορικές λύσεις "μαύρου κουτιού", η ανοιχτή αρχιτεκτονική επιτρέπει [38], [42]:

- **Επιθεώρηση του κώδικα:** Οποιοσδήποτε μπορεί να ελέγξει πώς ακριβώς υπολογίστηκε, για παράδειγμα, ο δείκτης *Debt-to-Equity* και ποιες παραδοχές έγιναν για τις τιμές που έλειπαν (null handling).
- **Επεκτασιμότητα:** Η κοινότητα μπορεί να προσθέσει νέους connectors ή νέα μοντέλα *dbt* για επιπλέον μετρικές, χωρίς να εξαρτάται από τον οδικό χάρτη (roadmap) μιας εμπορικής εταιρείας λογισμικού.

Συνεπώς, το θεωρητικό πλαίσιο της εργασίας δεν περιορίζεται μόνο στις τεχνολογίες καθαυτές, αλλά επεκτείνεται στο πώς ο συνδυασμός τους (Modern Data Stack) υπηρετεί τις αρχές της διαφάνειας, της ελεγξιμότητας και της ανοιχτής πρόσβασης στην πληροφορία.

Ελλείψεις – Σημείο τομής της παρούσας εργασίας

Αν και η διεθνής βιβλιογραφία έχει ασχοληθεί εκτενώς με μεμονωμένες πτυχές του προβλήματος, όπως η αξιολόγηση της ποιότητας των XBRL ταξινομιών ή η ανάπτυξη εξειδικευμένων εργαλείων για την ανάκτηση αρχείων από το σύστημα EDGAR, παρατηρήθηκε ένα σημαντικό κενό στην ενοποίηση αυτών των διαδικασιών. Συγκεκριμένα, απουσιάζουν από τη βιβλιογραφία εφαρμοσμένες μελέτες που να παρουσιάζουν μια πλήρη, αυτοματοποιημένη αλυσίδα αξίας δεδομένων (end-to-end pipeline) — από την αρχική άντληση (ingest) και τον μετασχηματισμό (transform) έως την τελική οπτικοποίηση (visualize) — βασισμένη αποκλειστικά σε σύγχρονα εργαλεία ανοιχτού κώδικα.

Επιπλέον, οι περισσότερες υπάρχουσες αναλυτικές προσεγγίσεις περιορίζονται σε επίπεδο μεμονωμένης εταιρείας ή συνολικής αγοράς, παραβλέποντας συχνά τη δυναμική της ανάλυσης σε γεωγραφικό επίπεδο πολιτείας (state-level analytics), η οποία μπορεί να προσφέρει πολύτιμα μακροοικονομικά και συγκριτικά συμπεράσματα.

Στόχος της παρούσας εργασίας είναι να αποτελέσει το σημείο τομής σε αυτές τις ελλείψεις, γεφυρώνοντας το χάσμα μεταξύ θεωρητικής προσέγγισης και τεχνικής υλοποίησης. Πιο συγκεκριμένα, προτείνει μία πλήρως αναπαραγώγιμη αρχιτεκτονική ELT→BI, η οποία όχι μόνο αυτοματοποιεί τη ροή της πληροφορίας αλλά εισάγει και τη μεθοδολογία «Quarter Bucket». Η μεθοδολογία αυτή προτείνεται ως λύση στο χρόνιο πρόβλημα της χρονικής ασυμμετρίας των εταιρικών αναφορών, επιτρέποντας για πρώτη φορά δίκαιες και αξιόπιστες συγκρίσεις σε σταθερή χρονική βάση [38], [43].

Ηθική, Εμπιστοσύνη και ο Ρόλος του Χρήστη στα Συστήματα Αποφάσεων

Όταν σχεδιάζουμε συστήματα Επιχειρηματικής Ευφυΐας (BI), συχνά εστιάζουμε στο τεχνικό κομμάτι: πώς θα τρέξουν γρήγορα τα δεδομένα και πώς θα φτιάξουμε αποδοτικά γραφήματα. Όμως, η σύγχρονη βιβλιογραφία τονίζει ότι η τεχνική αρτιότητα από μόνη της δεν αρκεί, απαιτείται η εδραίωση της εμπιστοσύνης (trust) του χρήστη προς το σύστημα.

Ειδικά τώρα που τα συστήματα αυτά αρχίζουν να χρησιμοποιούν Τεχνητή Νοημοσύνη (AI), το θέμα της εμπιστοσύνης γίνεται κρίσιμο. Σύμφωνα με πρόσφατες έρευνες [44], η εμπιστοσύνη δεν είναι ίδια για όλους τους χρήστες. Κάθε ενδιαφερόμενος (stakeholder) αντιδρά διαφορετικά ανάλογα με την εμπειρία και τον ρόλο του. Για την αντιμετώπιση αυτής της πολυμορφίας, προτείνεται το μοντέλο VIRTSI, το οποίο αναλύει πώς παράγοντες όπως η οπτικοποίηση και η διαφάνεια επηρεάζουν την εμπιστοσύνη στη λήψη αποφάσεων [44].

Επιπλέον, υπάρχει μια λεπτή ισορροπία ανάμεσα στο να αφήνουμε το σύστημα να λειτουργεί αυτόνομα και στο να τηρούμε τους κανόνες ηθικής. Όπως επισημαίνεται, υπάρχει μια ένταση (tension) μεταξύ της αυτονομίας των συστημάτων λογισμικού που ενδυναμώνονται με AI και των ανθρωποκεντρικών απαιτήσεων [45]. Οι προδιαγραφές του λογισμικού πρέπει να σέβονται τον χρήστη, διασφαλίζοντας ότι η τεχνολογία υπηρετεί τις ανθρώπινες αξίες και δεν λειτουργεί ως "μαύρο κουτί".

Σε αυτό το πλαίσιο, η σχέση μεταξύ Τεχνητής Νοημοσύνης και Εμπειρίας Χρήστη (UX) πρέπει να είναι αμφίδρομη (reciprocal). Η AI μπορεί να βελτιώσει το UX μέσω προσωποποίησης, αλλά και το σωστό UX είναι απαραίτητο για να γίνει η AI αποδεκτή και

κατανοητή [46]. Η ανάγκη αυτή για κατανόηση οδηγεί στην απαίτηση για Επεξηγήσιμη Τεχνητή Νοημοσύνη (Explainable AI - XAI), ειδικά σε πολύπλοκες αρχιτεκτονικές μικρουπηρεσιών (microservices), ώστε οι αποφάσεις του συστήματος να είναι αξιόπιστες και διαφανείς [47].

Ανάλυση και Σχεδιασμός

Απαιτήσεις συστήματος

Η ανάλυση απαιτήσεων του συστήματος αποτελεί ένα κρίσιμο στάδιο για το σχεδιασμό και την υλοποίηση του προτεινόμενου συστήματος Επιχειρηματικής Ευφυΐας (BI). Στο πλαίσιο της παρούσας εργασίας, οι απαιτήσεις διακρίνονται σε λειτουργικές και μη λειτουργικές, ώστε να καλύπτονται τόσο οι επιχειρησιακές ανάγκες όσο και τα ποιοτικά χαρακτηριστικά του συστήματος [25], [35], [38].

Λειτουργικές απαιτήσεις:

- Αυτόματη άντληση δεδομένων από EDGAR API.
- Αποθήκευση σε σχεσιακή βάση και δυνατότητα ιστορικότητας.
- Μετασχηματισμοί σε επίπεδο αποθήκης (ELT).
- Διαδραστικά dashboards ανά πολιτεία, metric, παράθυρο χρόνου.

Μη λειτουργικές απαιτήσεις:

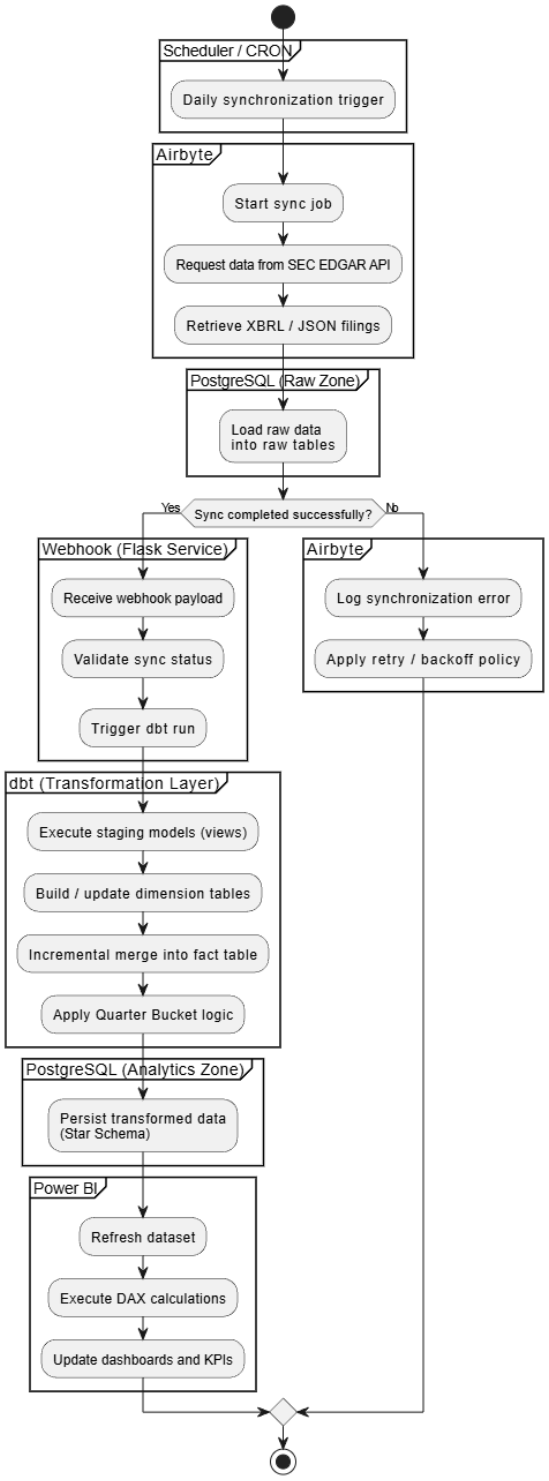
- Χρόνος πλήρους sync ~1 ώρες για ~550k εγγραφές.
- Αξιοπιστία (hash IDs, referential integrity).
- Δυνατότητα επεκτασιμότητας σε περισσότερα metrics/εταιρίες.

Επισκόπηση ροής

Η συνολική ροή δεδομένων βασίζεται σε μία πλήρως αυτοματοποιημένη αρχιτεκτονική εξαγωγής, φόρτωσης και μετασχηματισμού (Extract-Load-Transform (ELT)) [10], [13], [48].

Η διαδικασία ξεκινά από την εξαγωγή (Extract) των δεδομένων μέσω ενός custom Airbyte connector. Ο connector αυτός συνδέεται απευθείας με το API του συστήματος EDGAR και αντλεί τις σχετικές υποβολές εταιρικών δεδομένων, συμπεριλαμβανομένων XBRL concepts και στοιχείων υποβολών (submissions). Στη συνέχεια τα δεδομένα φορτώνονται (Load) σε πίνακες PostgreSQL του επιπέδου raw, όπου αποθηκεύονται χωρίς καμία τροποποίηση ή προκαταρκτική επεξεργασία, διατηρώντας πλήρως την αρχική τους μορφή. Η ροή του Airbyte είναι προγραμματισμένη να εκτελείται σε ημερήσια βάση. Μετά από κάθε επιτυχή συγχρονισμό (sync) ενεργοποιείται αυτόματα ένα Flask webhook, το οποίο εκκινεί την εκτέλεση του dbt. Μέσω του dbt πραγματοποιούνται οι μετασχηματισμοί (Transform) των δεδομένων στα επίπεδα σκηνικής παρουσίασης (staging), διάστασης (dimension) και γεγονότος (fact), ολοκληρώνοντας τον κύκλο ELT με αυτοματοποιημένο και επαναλήψιμο τρόπο [10], [13], [13], [35], [48]. Τα τελικά επεξεργασμένα δεδομένα αξιοποιούνται από το Power BI για τη δημιουργία αναφορών και διαδραστικών dashboards. Μία συνοπτική απεικόνιση της αρχιτεκτονικής ροής διαφαίνεται στο παρακάτω σχήμα:

EDGAR → *Airbyte* (custom source) → *PostgreSQL* (raw) → *dbt* (staging/dim/fact) → *Power BI*.



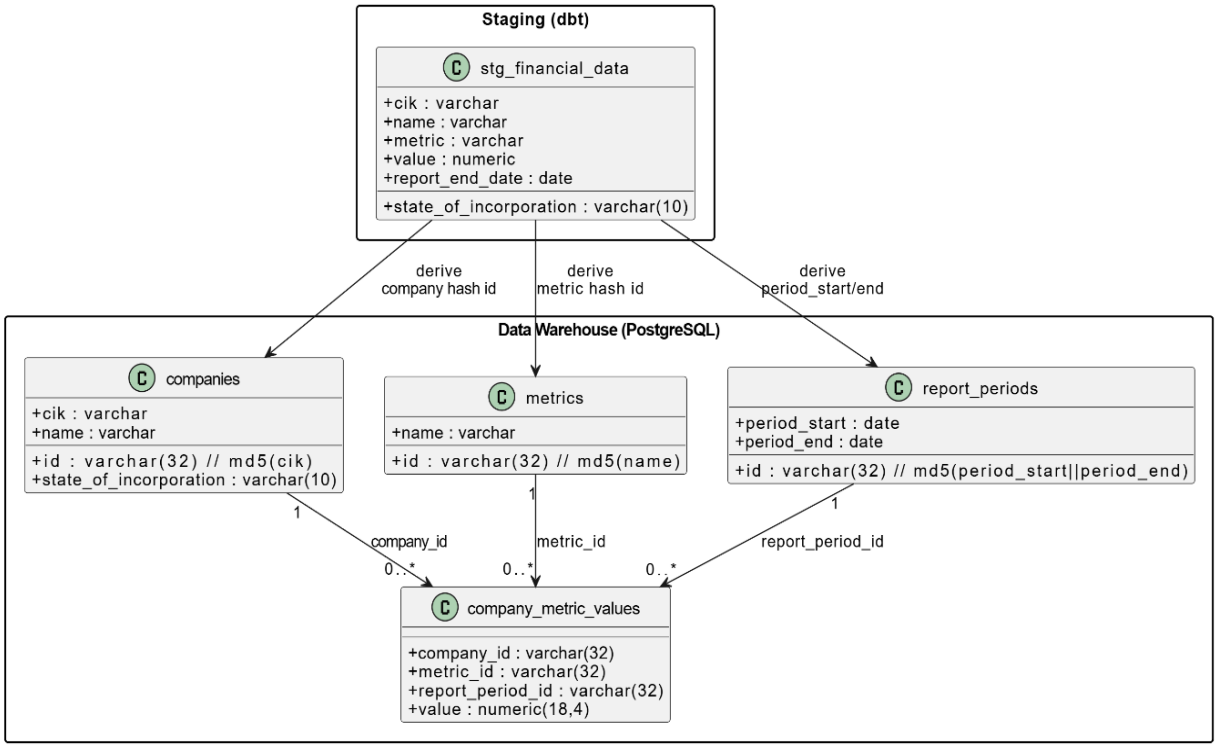
Εικόνα 2 Λογική αρχιτεκτονική της ροής ELT (→ Airbyte → PostgreSQL → dbt → Power BI)

Μοντελοποίηση (Star Schema)

Για τη μοντελοποίηση των δεδομένων επιλέχθηκε η αρχιτεκτονική του Αστεροειδούς Σχήματος (Star Schema), μία από τις επικρατέστερες προσεγγίσεις στη σχεδίαση αποθηκών δεδομένων για αναλυτικά συστήματα και εφαρμογές BI. Η προσέγγιση αυτή διαχωρίζει τα δεδομένα σε πίνακες γεγονότων (fact tables) και πίνακες διαστάσεων (dimension tables), επιτυγχάνοντας απλό σχήμα συσχετίσεων, βελτιωμένη απόδοση στα αναλυτικά ερωτήματα και ευκολία χρήσης σε εργαλεία BI [49], [50].

- **Companies**: Ο πίνακας αυτός περιέχει τις δημόσιες πληροφορίες κάθε εταιρείας. Το πεδίο id υπολογίζεται με MD5 hash του CIK. Η πολιτεία (state_of_incorporation) καθορίζεται βάσει της πιο συχνής μη-NULL τιμής. Η διαδικασία αυτή γίνεται για να αποφευχθούν τυχόν διπλοτιμές στο πεδίο cik απο τις εταιρικές υποβολές.
- **Metrics**: Ο πίνακας περιλαμβάνει όλους τους χρηματοοικονομικούς βασικούς δείκτες- μετρικές, με τυποποιημένα ονόματα (μέσω UPPER(TRIM(metric))). Το id δημιουργείται επίσης μέσω MD5 hash του ονόματος.
- **Report Periods**: Ο πίνακας περιγράφει τις χρονικές περιόδους αναφοράς.
- **Company Metric Values**: Αυτός είναι ο κύριος πίνακας γεγονότων που συνδέει τις τρεις διαστάσεις. Η τιμή value αποθηκεύεται ως NUMERIC(18,4) για διατήρηση ακρίβειας στα οικονομικά μεγέθη.

Η συγκεκριμένη αρχιτεκτονική επιτρέπει τη δημιουργία εύκολων joins και γρήγορων αναφορών στο επίπεδο του BI (*Power BI*, DAX).



Εικόνα 3 Λογικό σχήμα αστέρα (Star Schema) της αποθήκης δεδομένων στη PostgreSQL, με πίνακες διαστάσεων και fact.

IDs & incremental

Κατά τον σχεδιασμό της αποθήκης δεδομένων, κρίσιμο ζήτημα αποτέλεσε η σταθερότητα των αναγνωριστικών (IDs) και η ορθή διαχείριση των επαναλαμβανόμενων συγχρονισμών δεδομένων. Η αρχική προσέγγιση με συναρτήσεις κατάταξης τύπου «dense_rank()» κρίθηκε ακατάλληλη, καθώς οδηγούσε σε μεταβαλλόμενα Ids μεταξύ διαφορετικών φορτώσεων, προκαλώντας αστάθεια στις σχέσεις μεταξύ πινάκων. Για τον λόγω αυτό, υιοθετήθηκε η χρήση hash-based IDs (MD5), τα οποία παράγονται από σταθερά επιχειρησιακά κλειδιά (π.χ. CIK, metric name) [25], [35], [48], [50].

Η επιλογή αυτή σταθεροποιεί τις σχέσεις μεταξύ διαστάσεων και fact table, αποτρέπει σφάλματα ξένων κλειδιών (foreign key (FK) errors) και επιτρέπει ασφαλή επαναφόρτωση δεδομένων. Τα μοντέλα dbt υλοποιούνται ως αυξητικά (incremental) μοντέλα με στρατηγική συνένωσης (merge) χρησιμοποιώντας ορισμένο «unique_key», αποφεύγοντας τυχόν διπλότυπες εγγραφές με παράλληλη ενημέρωση μόνο σε νέες ή τροποποιημένες εγγραφές και διατήρηση της ιστορικότητας και της ακεραιότητας δεδομένων [25], [35], [48], [50].

Μεθοδολογία

Για τη παραγωγή αξιόπιστων και συγκρίσιμων αναφορών στο επίπεδο BI, εφαρμόζεται η μεθοδολογία Quarter Bucket, η οποία αντιμετωπίζει το πρόβλημα της χρονικής ασυμμετρίας στις εταιρικές δημοσιεύσεις [26], [27], [51], [52]. Ορίζονται ως:

- AsOf: το τελευταίο διαθέσιμο ημερολογιακό τρίμηνο (Quarter) στο dataset και
- MonthsBack ∈ {0, 3, 6, 12, 24, 36, 48}: που καθορίζουν το χρονικό βάθος ανάλυσης

Για κάθε συνδυασμό Company/Metrics επιλέγεται:

- η τελευταία διαθέσιμη τιμή εντός του αντίστοιχου ημερολογιακού τριμήνου (last-in-quarter).
- Παράλληλα, στο επίπεδο πολιτείας:
- υπολογίζεται ο μη σταθμισμένος (unweighted) μέσος όρος των εταιριών που διαθέτουν τιμή στο εξεταζόμενο τρίμηνο
- οι πολιτείες χωρίς διαθέσιμη τιμή αγνοούνται και
- η κατάταξη πραγματοποιείται με dense ranking

Για την οπτική παρουσίαση των διαφορών εφαρμόζεται neutral zone ±0,01 %, σύμφωνα με τις αρχές των IBCS, ώστε να αποφεύγεται η υπερερμηνεία αμελητέων μεταβολών.

Παράδειγμα ευθυγράμμισης Quarter Bucket

Για να γίνει πιο σαφής η λειτουργία της λογικής Quarter Bucket, ακολουθεί ένα απλοποιημένο παράδειγμα. Θεωρούμε τρεις εταιρείες που είναι ενσωματωμένες στην ίδια

πολιτεία (CA) και αναφέρουν την τιμή ενός δείκτη (π.χ. Net Income) σε διαφορετικές ημερομηνίες μέσα στο ίδιο ημερολογιακό τρίμηνο.

Εταιρεία	Πολιτεία	report_end_date	Τιμή net income	Υπολογισμένο QuarterStart	Επιλεγμένη τιμή στο Quarter
A	CA	2024-01-31	-643000	2024-01-01 (Q1 2024)	-643000
B	CA	2024-02-15	660000	2024-01-01 (Q1 2024)	660000
C	CA	2024-03-20	1423000	2024-01-01 (Q1 2024)	1423000

Πίνακας 1 Παράδειγμα ευθυγράμμισης Quarter Bucket

Στο επίπεδο εταιρείας, για κάθε συνδυασμό (Εταιρεία, Metric, QuarterStart) επιλέγεται η τελευταία διαθέσιμη τιμή μέσα στο τρίμηνο (Last-in-Quarter).

Στο συγκεκριμένο παράδειγμα:

- για την εταιρεία A, η επιλεγμένη τιμή είναι -643000 (31/01)
- για την εταιρεία B, η τιμή 660000 (15/02)
- για την εταιρεία C, η τιμή 1423000 (20/03)

Στη συνέχεια, στο επίπεδο πολιτείας δημιουργείται μία εγγραφή στο StateQuarter για το ζεύγος (CA, Q1 2024). Η τιμή avg_net_income για το τρίμηνο αυτό προκύπτει ως μη σταθμισμένος μέσος όρος των επιλεγμένων τιμών των εταιρειών:

$$\text{avg\_revenues(CA,Q1 2024)}=\frac{-643000+660000+1423000}{3}=480000$$

Αν, για παράδειγμα, το προηγούμενο αντίστοιχο τρίμηνο (Q1 2023) είχε avg_net_income = 390000, τότε:

$$\Delta=480000-390000=90000 \text{ και}$$

$$\Delta\%=\frac{480000-390000}{390000}\approx 23\%$$

Με αυτόν τον τρόπο, όλες οι εταιρείες της πολιτείας ευθυγραμμίζονται στο ίδιο χρονικό “bucket” (ημερολογιακό τρίμηνο) ανεξάρτητα από την ακριβή ημερομηνία δημοσίευσης, και οι συγκρίσεις γίνονται σε σταθερή χρονική βάση τόσο μεταξύ πολιτειών όσο και στο χρόνο.

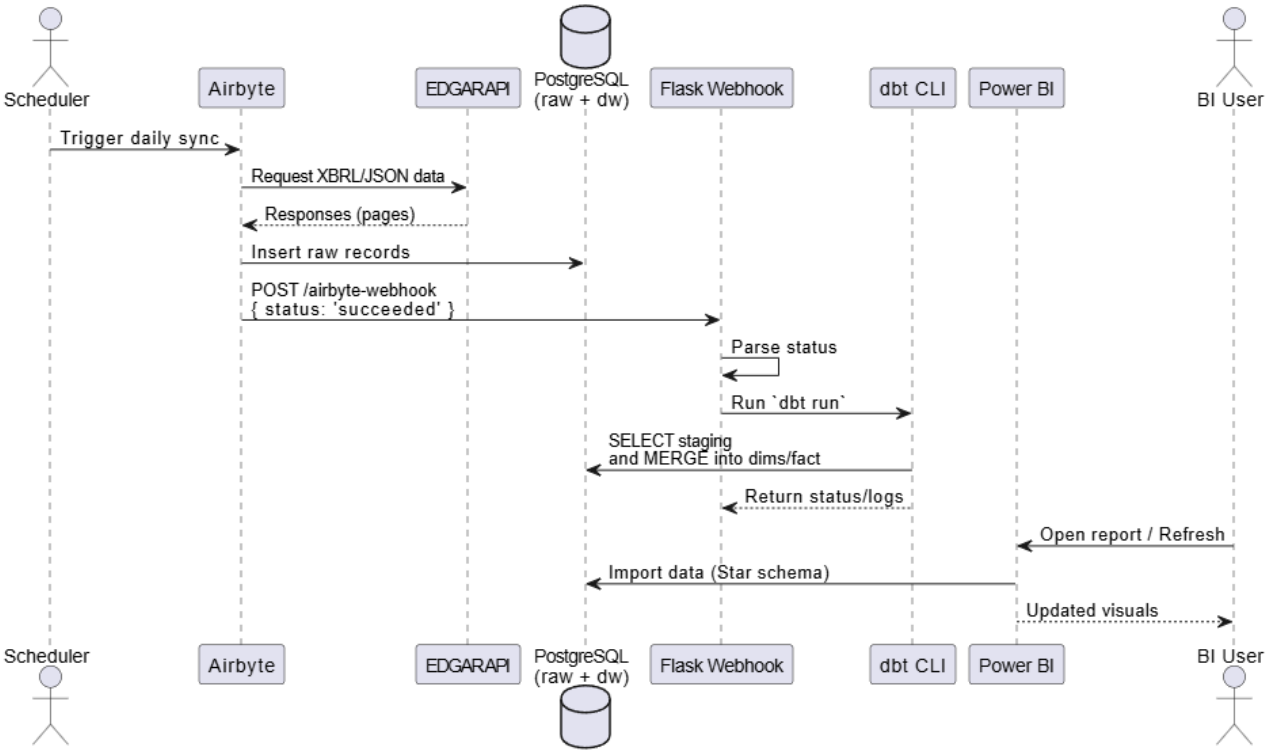
UML Διαγράμματα Συστήματος

Για την σαφή αποτύπωση της λειτουργίας του συστήματος και των αλληλεπιδράσεων μεταξύ των υποσυστημάτων του pipeline, παρουσιάζονται τα αντίστοιχα διαγράμματα UML. Τα διαγράμματα αυτά περιγράφουν [53], [54]:

- τη ροή εκτέλεσης της διαδικασίας συγχρονισμού (*Airbyte* → Webhook → *dbt*)
- τη χρονική ακολουθία ανταλλαγής μηνυμάτων μεταξύ των επιμέρους components
- την ολοκλήρωση της ροής δεδομένων μέχρι το επίπεδο BI (*Power BI* Desktop)

Οι παρακάτω αναπαραστάσεις βοηθούν στην κατανόηση της λειτουργικής λογικής, της σειράς εκτέλεσης και της αρχιτεκτονικής συνοχής του συστήματος.

Εικόνα 4 Διάγραμμα Ακολουθίας (Sequence Diagram) της ροής Airbyte → → dbt → Power BI.



Το διάγραμμα ακολουθίας αποτυπώνει τη χρονική σειρά των ενεργειών που πραγματοποιούνται κατά την εκτέλεση ενός πλήρους κύκλου ενημέρωσης δεδομένων (sync cycle) [53], [54], [55].

Η διαδικασία ξεκινά από τον Scheduler, ο οποίος ενεργοποιεί το καθημερινό συγχρονισμό του *Airbyte*. Το *Airbyte* επικοινωνεί με το EDGAR API, από όπου αντλεί δεδομένα σε μορφή XBRL/JSON και τα καταχωρεί στον πίνακα raw της *PostgreSQL*.

Με την ολοκλήρωση του συγχρονισμού, το *Airbyte* αποστέλλει ένα αίτημα στο Flask Webhook, το οποίο ελέγχει το περιεχόμενο του payload και εκκινεί την εντολή *dbt* run. Το *dbt* προχωρά σε μετασχηματισμούς staging, δημιουργία/ενημέρωση διαστάσεων και συγχώνευση («incremental merge») στο fact table.

Όταν η διαδικασία ολοκληρωθεί, τα ενημερωμένα δεδομένα είναι άμεσα διαθέσιμα για ανάκτηση από το *Power BI* Desktop, όπου ο χρήστης BI μπορεί να προβεί σε refresh και

ανάλυση. Το sequence diagram καταδεικνύει με ακρίβεια τη σειρά και τις εξαρτήσεις των επιμέρους βημάτων της ροής ELT.

Εικόνα 5 Διάγραμμα Δραστηριότητας (Activity Diagram) του ημερήσιου κύκλου συγχρονισμού δεδομένων.

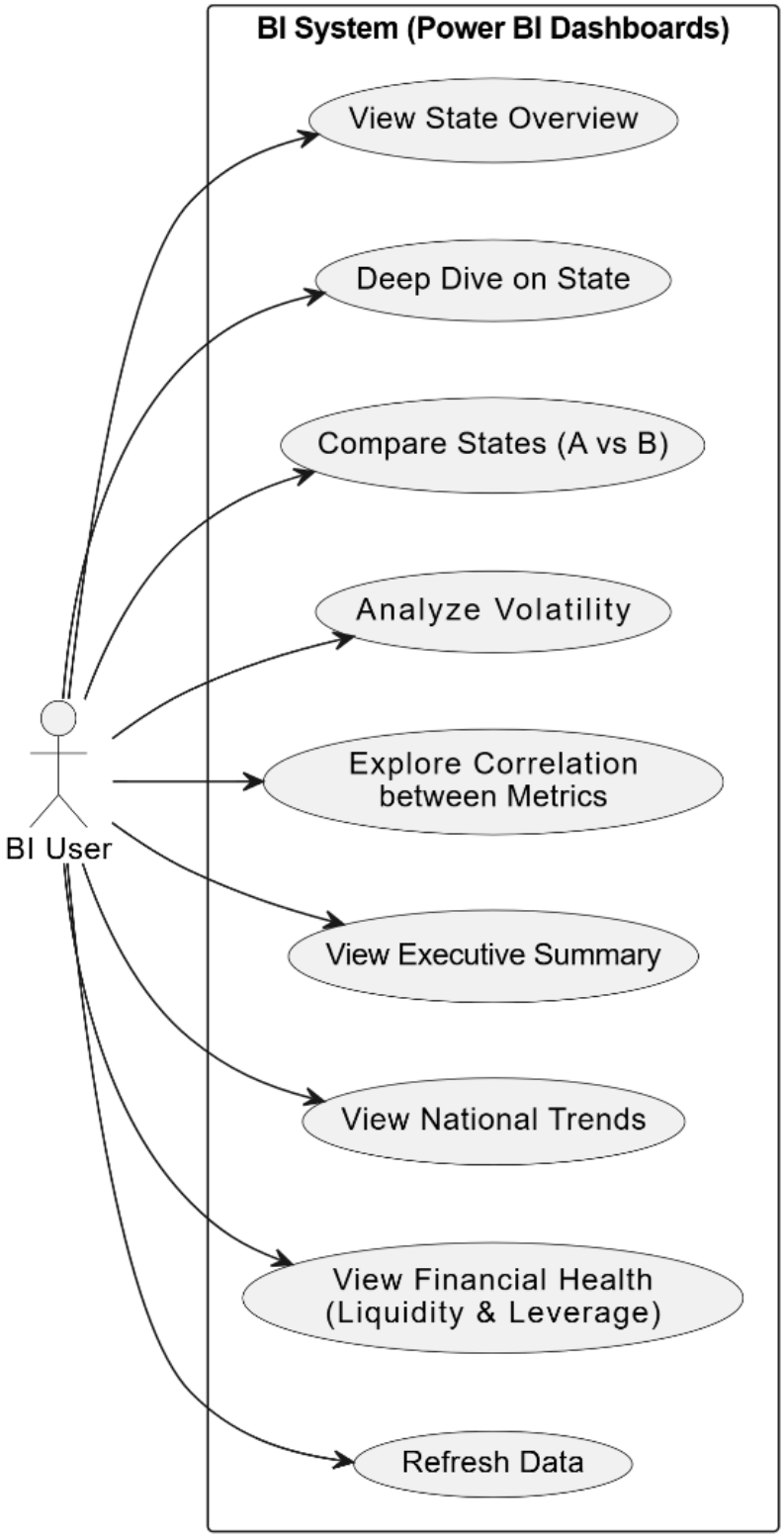
Το activity diagram παρουσιάζει τη λογική ροής της διαδικασίας συγχρονισμού δεδομένων, από την ενεργοποίηση του daily CRON έως τη διαθεσιμότητα των ενημερωμένων δεδομένων στο επίπεδο BI [53], [54], [56].

Η δραστηριότητα ξεκινά με την έναρξη του συγχρονισμού από το *Airbyte*, το οποίο αναλαμβάνει τη διαδικασία άντλησης δεδομένων EDGAR και αποθήκευσης σε raw μορφή στην *PostgreSQL*. Στη συνέχεια γίνεται έλεγχος επιτυχίας: σε περίπτωση αποτυχίας, το σύστημα καταγράφει το αντίστοιχο μήνυμα σφάλματος.

Σε επιτυχή εκτέλεση, το *Airbyte* ενεργοποιεί το Flask webhook, το οποίο αναλύει το status και εκτελεί την εντολή *dbt* run. Το *dbt* υλοποιεί όλους τους μετασχηματισμούς (staging, διαστάσεις, fact με incremental merge) και παράγει το τελικό schema προς χρήση.

Στο τέλος, είτε η διαδικασία ολοκληρωθεί επιτυχώς είτε εντοπιστεί σφάλμα, αποθηκεύονται αντίστοιχα logs. Με επιτυχή ολοκλήρωση, τα δεδομένα είναι πλέον διαθέσιμα στο *Power BI*, όπου ο αναλυτής μπορεί να πραγματοποιήσει ενημέρωση (refresh) και να προχωρήσει σε αναλυτική διερεύνηση.

Εικόνα 6 Διάγραμμα Use Case του συστήματος BI και των βασικών λειτουργιών για τον χρήστη.



Το διάγραμμα Use Case αποτυπώνει τις βασικές λειτουργίες που προσφέρει το σύστημα BI στον τελικό χρήστη [53], [54], [57]. Ο ρόλος BI User μπορεί να:

- εξετάσει συνοπτική εικόνα ανά πολιτεία (State Overview)
- εμβαθύνει σε μία συγκεκριμένη πολιτεία (State Deep Dive)
- συγκρίνει δύο πολιτείες μεταξύ τους (State Comparison A vs B)
- αναλύσει τη μεταβλητότητα των μετρικών (Volatility)
- μελετήσει σχέσεις μεταξύ μετρικών (Correlation Explorer)
- δει συνοπτικούς δείκτες και τάσεις (Executive Summary, National Trends)
- αξιολογήσει ρευστότητα και μόχλευση (Financial Health)
- και να εκκινήσει ενημέρωση δεδομένων (Refresh Data)

Το διάγραμμα δείχνει ότι όλες οι λειτουργίες είναι προσβάσιμες από έναν ενιαίο τύπο χρήστη, ο οποίος επικεντρώνεται στην ανάλυση και όχι στη διαχείριση της υποδομής.

Υλοποίηση Εφαρμογής

Ingestion – Custom Connector

Ο Custom source στο *Airbyte*, έχει αναπτυχθεί με την χρήση Python SDK και αντλεί CIKs/metrics κατευθείαν από τα XBRL company concepts [4], [17]. Η επιλογή να γραφτεί custom κώδικας αντί να χρησιμοποιηθεί ένας έτοιμος HTTP connector ήταν απαραίτητη λόγω της ιδιαιτερότητας του SEC EDGAR API, το οποίο απαιτεί δυναμική διαχείριση των URLs ανά εταιρεία (CIK) και αυστηρό έλεγχο ρυθμού κλήσεων (Rate Limiting). Κατα την χρήση του, υποστηρίζονται retries/backoff καθώς και δηλωτικός User-Agent που είναι απαιτούμενος απο το SEC. Η εκτέλεση γίνεται σε WSL/Ubuntu 24.04.

Ακολουθεί η ανάλυση των κρίσιμων τμημάτων του κώδικα, ο οποίος υλοποιεί την κλάση SECDataStream κληρονομώντας από την HttpStream του *Airbyte*.

A. Διαχείριση Rate Limiting και Ανθεκτικότητα (Resilience)

Ένα συμαντικό πρόβλημα που υπάρχει στα Web APIs είναι το σφάλμα 429 Too Many Requests. Στον κώδικα υλοποιήθηκε μηχανισμός Exponential Backoff with Jitter. Όπως φαίνεται στο παρακάτω απόσπασμα, όταν ο server απαντήσει με 429, το σύστημα περιμένει για ένα αυξανόμενο χρονικό διάστημα ($2^{attempt} + random$) πριν ξαναδοκιμάσει. Αυτό εξασφαλίζει ότι το pipeline δεν θα αποτύχει σε περιόδους φόρτου του server της SEC [51], [58].

```
def safe_request(self, url: str, headers: dict, max_retries=5) ->
requests.Response:
    for attempt in range(max_retries):
        response = requests.get(url, headers=headers)

        if response.status_code == 200:
            return response

        elif response.status_code == 429:
            # Υπολογισμός χρόνου αναμονής: 2^attempt + τυχαίος θόρυβος
            wait_time = 2 ** attempt + random.uniform(0, 1)
            logger.warning(f"⚠️ API 429: Retry σε {wait_time:.2f} sec")
            time.sleep(wait_time)

        else:
            logger.error(f"❌ Αποτυχία ({response.status_code}) για
{url}")
            break
    return None
```

Κώδικας διαχείρισης

Β. Δυναμική Χαρτογράφηση Μετρικών

Για χαρτογράφηση επιχειρηματικών εννοιών (π.χ. revenues) χρησιμοποιήθηκε η μέθοδος *financial_data_paths*. Αυτή τις χαρτογραφεί στα αντίστοιχα JSON endpoints του XBRL API. Η προσέγγιση αυτή κάνει τον κώδικα πιο εύκολα επεκτάσιμο (η προσθήκη μιας νέας μετρικής απαιτεί απλώς μια νέα γραμμή στο λεξικό (dictionary), χωρίς αλλαγή της κύριας λογικής ροής).

```
def financial_data_paths(self, cik: str) -> dict:
    return {
        "revenues": f"api/xbrl/companyconcept/CIK{cik}/us-
gaap/Revenues.json",
        "net_income": f"api/xbrl/companyconcept/CIK{cik}/us-
gaap/NetIncomeLoss.json",
        "gross_profit": f"api/xbrl/companyconcept/CIK{cik}/us-
gaap/GrossProfit.json",
        "current_assets": f"api/xbrl/companyconcept/CIK{cik}/us-
gaap/AssetsCurrent.json",
        "long_term_debt": f"api/xbrl/companyconcept/CIK{cik}/us-
gaap/LongTermDebt.json",
         ... (επιπλέον μετρικές όπως StockholdersEquity, OperatingExpenses
κ.λπ.)
    }
```

Λεξικό χαρτογράφησης μετρικών XBRL

Γ. Generator Pattern για Διαχείριση Μνήμης

Λόγω του μεγάλου όγκου των δεδομένων (εκατοντάδες χιλιάδες εγγραφές), η φόρτωση όλων των αποτελεσμάτων στη μνήμη RAM οδηγούσε συχνά σε κατάρρευση της εφαρμογής. Στη μέθοδο read_records, χρησιμοποιείται η εντολή yield αντί για return.

Αυτό μετατρέπει τη συνάρτηση σε generator, επιτρέποντας στο *Airbyte* να επεξεργάζεται και να αποθηκεύει τις εγγραφές μία-μία (streaming) καθώς αυτές κατεβαίνουν από το API.

```
def read_records(self, stream_state: dict = None, **kwargs) ->
Iterable[Mapping[str, Any]]:

    for index, cik in enumerate(self.ciks):

        financial_paths = self.financial_data_paths(cik)

        for metric, endpoint in financial_paths.items():
            url = self.url_base + endpoint
            response = self.safe_request(url, self.request_headers())

             ... (κώδικας parsing και φιλτραρίσματος δεδομένων) ...

             Δημιουργία και εκπομπή εγγραφής
            record = {
                "cik": cik,
                "metric": metric,
                "value": value,
                "report_end_date": report_date,
                 ... (υπόλοιπα πεδία)
```

```
            }

            yield record
```

Υλοποίηση Generator Pattern (read_records)

Raw & Staging

Στο επίπεδο raw αποθηκεύονται τα αρχικά δεδομένα ακριβώς όπως αντλήθηκαν από την πηγή SEC EDGAR, χωρίς τροποποίηση [39]. Περιλαμβάνονται συγκεκριμένα τα πεδία:

- cik, name,
- state_of_incorporation,
- metric, value και
- report_end_date απο κάθε υποβολή.

Ακολουθεί το επίπεδο staging, όπου εφαρμόζονται οι πρώτοι μετασχηματισμοί και καθαρισμοί δεδομένων. Σε αυτό το στάδιο πραγματοποιείται τυποποίηση του κειμένου (με χρήση συναρτήσεων όπως UPPER() και TRIM()), ώστε να εξασφαλίζεται η ομοιομορφία στις ονομασίες εταιρειών και μετρικών. Επίσης, υπολογίζονται δύο νέα πεδία, το period_start και το period_end (= report_end_date).Το period_start ορίζεται ως τρεις μήνες πριν την ημερομηνία αναφοράς (report_end_date - 3 μήνες), καθορίζοντας έτσι το χρονικό διάστημα κάθε αναφοράς.

Στρατηγική Staging και Data Cleaning

Στο επίπεδο του Staging, η βασική αρχιτεκτονική επιλογή ήταν η δημιουργία "λεπτών" όψεων (thin views) αντί για φυσικούς πίνακες. Η επιλογή του materialized='view' στο dbt εξασφαλίζει ότι τα δεδομένα στο staging είναι πάντα μια άμεση αντανάκλαση των raw δεδομένων χωρίς να καταλαμβάνουν επιπλέον χώρο αποθήκευσης [39], [59]. Σε αυτό το στάδιο, αντιμετωπίστηκαν συγκεκριμένες προκλήσεις ποιότητας δεδομένων (Data Quality):

- **Cast Data Types**: Η ρητή μετατροπή των πεδίων (π.χ. value::numeric, report_end_date::date) είναι κρίσιμη διότι το JSON format που επιστρέφει το API συχνά ερμηνεύεται ως απλό κείμενο (string) από τη βάση. Χωρίς αυτό το βήμα, οι αριθμητικές πράξεις στα επόμενα στάδια θα αποτύγχαναν.
- **Handling Nulls**: Εφαρμόστηκε φιλτράρισμα WHERE value IS NOT NULL νωρίς στη ροή. Αυτό μειώνει τον όγκο των δεδομένων που περνούν στην βάση δεδομένων, αφαιρώντας θόρυβο από εγγραφές που υπάρχουν στο XBRL taxonomy αλλά δεν έχουν συμπληρωθεί από την εκάστοτε εταιρεία.

Αρχιτεκτονική Διαστάσεων και Surrogate Keys

Για τους πίνακες διαστάσεων (*Dimensions*), η χρήση φυσικών κλειδιών (Natural Keys) όπως το CIK ή το Metric Name ήταν ανεπαρκής για τη διασφάλιση της μοναδικότητας και της απόδοσης των joins. Αντ' αυτού, χρησιμοποιήθηκε η πρακτική των Surrogate Keys μέσω κρυπτογραφικού κατακερματισμού (Hashing) [39], [60]. Η χρήση της συνάρτησης MD5()

προσφέρει τρία βασικά πλεονεκτήματα σε σχέση με την παραδοσιακή αρίθμηση (AUTO_INCREMENT ή DENSE_RANK):

1. **Ντετερμινισμός**: Το ίδιο input (π.χ. "APPLE INC") παράγει πάντα το ίδιο hash ID, ανεξάρτητα από το πότε τρέχει το pipeline. Αυτό επιτρέπει την επανεκτέλεση (re-run) του dbt χωρίς να "σπάνε" οι σχέσεις ξένων κλειδιών.
2. **Ανεξαρτησία**: Τα κλειδιά μπορούν να παραχθούν παράλληλα χωρίς να χρειάζεται "κλείδωμα" ενός κεντρικού μετρητή sequence.
3. **Ιχνηλασιμότητα**: Είναι εύκολο να επαληθευτεί αν μια εγγραφή στο Fact table αντιστοιχεί στη σωστή διάσταση κάνοντας απλά hash τα πεδία αναφοράς.

Βελτιστοποίηση του Fact Table με Incremental Strategy

Ο πίνακας γεγονότων company_metric_values είναι ο μεγαλύτερος πίνακας του συστήματος. Η πλήρης ανακατασκευή του (full refresh) σε κάθε εκτέλεση θα ήταν χρονοβόρα και με μεγάλο υπολογιστικό κόστος. Γι' αυτό, αξιοποιήθηκε η λειτουργία incremental του dbt με στρατηγική merge [61]. Ο αλγόριθμος λειτουργεί ως εξής:

1. Το dbt ελέγχει αν ο πίνακας υπάρχει ήδη στη βάση.
2. Αν υπάρχει, εντοπίζει την τελευταία ημερομηνία ενημέρωσης (_airbyte_emitted_at) ή το μέγιστο χρονικό σημείο αναφοράς.
3. Δημιουργεί έναν προσωρινό πίνακα μόνο με τις εγγραφές που είναι νεότερες από αυτό το σημείο.
4. Εκτελεί μια εντολή MERGE (upsert), η οποία είτε ενημερώνει τις υπάρχουσες εγγραφές (αν έχουν αλλάξει τιμές) είτε εισάγει τις νέες. Η χρήση του unique_key (που αποτελείται από το composite key των ids εταιρείας, μετρικής και περιόδου) διασφαλίζει την ακεραιότητα των δεδομένων (idempotency), αποτρέποντας τη δημιουργία διπλότυπων εγγραφών ακόμα και αν το pipeline τρέξει πολλές φορές την ίδια μέρα.

Orchestration – Webhook → *dbt*

Η επικοινωνία μεταξύ Airbyte και dbt βασίζεται στο πρωτόκολλο HTTP POST. Η διαδικασία αναλύεται στα εξής τεχνικά βήματα [20], [29], [31], [60]:

1. ***Payload Reception***: Το Airbyte, κατά την ολοκλήρωση του Job, κατασκευάζει ένα JSON αντικείμενο που περιέχει μεταδεδομένα για την κατάσταση του συγχρονισμού (π.χ. status, attempt_count, bytes_synced).
2. ***Status Verification***: Ο Flask server δεν εκτελεί τυφλά το dbt. Πραγματοποιεί έλεγχο (parsing) του πεδίου status. Μόνο αν η τιμή είναι SUCCEEDED, προχωρά στην επόμενη ενέργεια. Αυτό αποτρέπει την εκτέλεση μετασχηματισμών πάνω σε ελλιπή ή διεφθαρμένα δεδομένα σε περίπτωση αποτυχίας της άντλησης.

3. ***Subprocess Execution***: Η κλήση του dbt γίνεται μέσω της βιβλιοθήκης subprocess της Python. Αυτό επιτρέπει στον server να εκτελεί εντολές κελύφους (shell commands) σαν να ήταν τοπικός χρήστης.
4. ***Logging & Monitoring***: Το αποτέλεσμα της εκτέλεσης καταγράφεται σε αρχεία καταγραφής (logs). Έτσι, αν αποτύχει ένας μετασχηματισμός SQL, ο διαχειριστής μπορεί να ανατρέξει στο log του Flask και να δει το ακριβές μήνυμα λάθους της βάσης δεδομένων, χωρίς να χρειάζεται να μπει στην κονσόλα του συστήματος.

Υλοποίηση μετασχηματισμών με *dbt*

Η λογική στρώσης μετασχηματισμού (transform) της αποθήκης δεδομένων υλοποιείται με χρήση του εργαλείου *dbt* (data build tool). Στόχος είναι τα δεδομένα που αντλούνται ακατέργαστα από τον custom connector του *Airbyte* να μετασχηματίζονται σε ένα σταθερό σχήμα αστέρα, κατάλληλο για χρήση από το *Power BI* [29], [31], [32], [50].

Το *dbt* project οργανώνεται σε τέσσερις βασικές κατηγορίες μοντέλων:

1. Staging μοντέλο:
    - stg_financial_data
2. Διαστασιακά μοντέλα:
    - companies
    - metrics
    - report_periods
3. Πίνακας γεγονότων (fact):
    - company_metric_values

Staging: stg_financial_data

Το μοντέλο αυτό διαβάζει απευθείας από τον raw πίνακα που δημιουργεί το *Airbyte* (_531_sec_data_stream_combined) και πραγματοποιεί τους ελάχιστους απαραίτητους μετασχηματισμούς ώστε τα δεδομένα να γίνουν συνεπή και αξιοποιήσιμα από τα επόμενα στάδια [39], [59].

Στο στάδιο αυτό:

- Γίνονται μετονομασίες πεδίων σε πιο σταθερά ονόματα (π.χ. cik, name, metric, value, state_of_incorporation, report_end_date).
- Υπολογίζονται τα χρονικά πεδία:
    - period_end = report_end_date.

- period_start = report_end_date - 3 μήνες ώστε κάθε εγγραφή να συνδέεται με ένα συγκεκριμένο τρίμηνο αναφοράς.

➢ Γίνεται τυποποίηση τύπων δεδομένων (ημερομηνίες σε date, αριθμητικά σε numeric).

➢ Φιλτράρονται ή διορθώνονται εμφανώς εσφαλμένες ημερομηνίες (π.χ. υπερβολικά μελλοντικά έτη).

Το staging μοντέλο λειτουργεί ως «γέφυρα» ανάμεσα στο semi-structured schema που παράγει ο connector και στο λογικό μοντέλο της αποθήκης. Επίσης απομονώνει αλλαγές στο upstream schema, ώστε να επηρεάζονται όσο το δυνατόν λιγότερα downstream μοντέλα.

```sql
{{ config(materialized='view') }}

-- Strict staging view using only columns that exist in your raw table.
-- Assumes raw columns: cik, name, state_of_incorporation, metric, value, report_end_date.
with src as (
  select
    cik::text                                   as cik,
    upper(trim(name))                           as name,
    state_of_incorporation,
    upper(trim(metric))                         as metric,
    value,
    (report_end_date::date - interval '3 month') as period_start,
    report_end_date::date                       as period_end
  from {{ source('public', '_531_sec_data_stream_combined') }}
  where value is not null
)
select * from src
```

Εικόνα 7 .sql

Διαστάσεις: companies, metrics, report_periods

Οι διαστάσεις υλοποιούνται ως κεντρικοί reference πίνακες με σταθερά Ids, ώστε να υποστηρίζεται αναφορική ακεραιότητα και incremental ενημέρωση [39], [59], [61].

➢ **companies**

- Το πεδίο id δεν είναι απλό αύξοντα κλειδί αλλά hash της μορφής md5(cik). Με αυτόν τον τρόπο η τιμή είναι:
  - σταθερή ανά εταιρία.
  - ανεξάρτητη από τη σειρά εισαγωγής.
  - κατάλληλη για αναπαραγωγή σε άλλα περιβάλλοντα.
- Για κάθε CIK επιλέγεται ντετερμινιστικά ένα name (π.χ. το αλφαβητικά πρώτο) ώστε να αποφεύγονται διπλοεγγραφές.

- Το state_of_incorporation προκύπτει ως η συχνότερη μη-NULL τιμή που παρατηρείται στο raw staging, ώστε να μειωθεί ο αριθμός των κενών.

➢ **metrics**

- Το id υπολογίζεται ως md5(name), όπου name είναι κανονικοποιημένο (trim, upper).
- Η προσέγγιση αυτή εξασφαλίζει ότι η ίδια μετρική (π.χ. NET_INCOME) παίρνει το ίδιο ID σε κάθε run, ακόμη κι αν αλλάξει η σειρά εμφάνισης στο staging.

➢ **report_periods**

- Κάθε περίοδος ταυτοποιείται από το ζεύγος (period_start, period_end) και αποκτά ID md5(period_start || period_end).
- Η χρήση hash προστατεύει από «μετακίνηση» Ids όταν εισάγονται νέα τρίμηνα, σε αντίθεση με προσεγγίσεις τύπου dense_rank() που είχαν δοκιμαστεί αρχικά και δημιούργησαν προβλήματα συνέπειας.

Στο αρχικό σχεδιασμό τα Ids δημιουργούνταν με dense_rank() πάνω στα δεδομένα κάθε run. Αυτό όμως οδηγούσε σε αλλαγή των Ids όταν γίνονταν εισαγωγές νέων εγγραφών, προκαλώντας σφάλματα στις ξένες κλείδες του fact πίνακα. Η μετάβαση σε hash Ids έλυσε οριστικά αυτό το ζήτημα.

Fact: company_metric_values

Ο πίνακας company_metric_values αποτελεί τον πυρήνα του σχήματος αστέρα [50].

Για κάθε συνδυασμό:

- εταιρίας (company_id),
- μετρικής (metric_id),
- περιόδου (report_period_id),

αποθηκεύεται μία αριθμητική τιμή value τύπου numeric(18,4).

Το μοντέλο έχει ρυθμιστεί ως incremental με στρατηγική merge. Αυτό σημαίνει ότι:

- σε κάθε εκτέλεση του *dbt* γίνεται upsert των εγγραφών (update υπαρχουσών + insert νέων)
- δεν χρειάζεται πλήρης διαγραφή/επαναφόρτωση του πίνακα
- ο χρόνος εκτέλεσης παραμένει διαχειρίσιμος, ακόμη και όταν ο όγκος των δεδομένων μεγαλώσει.

Οι joins με τις διαστάσεις γίνονται με βάση τα hash Ids που υπολογίζονται στο staging, εξασφαλίζοντας ότι οι σχέσεις παραμένουν συνεπείς ακόμη και αν το raw layer αλλάξει.

Μοντελοποίηση και προεπεξεργασία στο *Power BI* (Power Query)

Στο επίπεδο του Power BI υλοποιείται το σημασιολογικό επίπεδο (semantic layer) που χρησιμοποιούν τα dashboards. Αντί οι σύνθετοι υπολογισμοί να γίνονται εξ ολοκλήρου σε DAX πάνω στο fact table, ένα σημαντικό μέρος της προεπεξεργασίας μεταφέρεται στο Power Query, ώστε [21], [25], [32], [39], [51]:

- να βελτιωθεί η απόδοση.
- να υπάρχει ξεκάθαρη και επαναχρησιμοποιήσιμη λογική προετοιμασίας δεδομένων.
- να στηριχθεί η μεθοδολογία Quarter Bucket.

Η βασική ροή στο Power Query δομείται σε τρία επίπεδα: Query1, CompanyQuarter και StateQuarter.

Query1 – Row-level δεδομένα

Η πρώτη ερώτηση (Query1) εισάγει από την *PostgreSQL* τα δεδομένα σε επίπεδο εταιρίας–μετρικής–περιόδου (ουσιαστικά join του fact με τις διαστάσεις). Στο στάδιο αυτό:

- Γίνεται καθαρισμός και τυποποίηση πεδίων (company_id, metric_name, state_of_incorporation, value, period_start, period_end).
- Υπολογίζεται το πεδίο QuarterStart ως η πρώτη ημέρα του τριμήνου στο οποίο ανήκει η ημερομηνία αναφοράς (π.χ. με Date.StartOfQuarter([period_end])).
- Ορίζονται αυστηρά οι τύποι δεδομένων (ημερομηνίες, decimal αριθμοί).
- Εφαρμόζονται φίλτρα ποιότητας (π.χ. χρονικό εύρος από συγκεκριμένο έτος και μετά, απομάκρυνση εξόφθαλμα λανθασμένων ημερομηνιών).

Το QuarterStart λειτουργεί ως βασικός χρονικός άξονας, πάνω στον οποίο θα βασιστούν όλα τα visuals.

CompanyQuarter – Τελευταία τιμή ανά εταιρία και τρίμηνο

Στη συνέχεια, από το Query1 προκύπτει ένα ενδιάμεσο ερώτημα CompanyQuarter που συμπυκνώνει τα δεδομένα ώστε κάθε γραμμή να αντιστοιχεί σε μία εταιρία, έναν δείκτη, ένα τρίμηνο. Σε αυτό το βήμα:

- Οι μετρικές pivot-άρονται σε στήλες (π.χ. revenues, gross_profit, net_income, current_assets, current_liabilities κ.λπ.).
- Για κάθε (εταιρία, μετρική, QuarterStart) διατηρείται η τελευταία διαθέσιμη τιμή μέσα στο τρίμηνο (Last-in-Quarter). Αυτό αντιμετωπίζει την ασύγχρονη δημοσίευση των EDGAR filings.
- Υπολογίζονται παράγωγες μετρικές σε επίπεδο εταιρίας, όπως:
    - gross_margin = gross_profit / revenues
    - net_margin = net_income / revenues

- working_capital = current_assets – current_liabilities
- current_ratio = current_assets / current_liabilities
- debt_to_equity = long_term_debt / stockholders_equity
- roe = net_income / stockholders_equity

- Χρησιμοποιούνται προστατευτικοί έλεγχοι (π.χ. try … otherwise null, έλεγχος για μηδενικό ή αρνητικό παρονομαστή) ώστε να αποφεύγονται ακραίες ή λανθασμένες τιμές.

Το αποτέλεσμα είναι ένα “wide” dataset σε granularité εταιρία × τρίμηνο, πάνω στο οποίο μπορεί να γίνει περαιτέρω συγκέντρωση ανά πολιτεία.

StateQuarter – Μ.Ο. ανά πολιτεία και τρίμηνο

Το τρίτο επίπεδο, StateQuarter, αποτελεί και τη βασική πηγή για τα περισσότερα dashboards. Δημιουργείται με:

- ομαδοποίηση των δεδομένων του CompanyQuarter κατά state_of_incorporation και QuarterStart,
- υπολογισμό:
  - πλήθους εταιρειών (company_count) που έχουν διαθέσιμη τιμή στο συγκεκριμένο τρίμηνο,
  - μέσων όρων των βασικών μεγεθών (avg_revenues, avg_gross_profit, avg_net_income, avg_current_assets, κ.λπ.) ως απλούς (unweighted) μέσους ανά εταιρία,
- παραγωγή δευτερογενών δεικτών πάνω στα averages, όπως:
  - avg_gross_margin = avg_gross_profit / avg_revenues
  - avg_net_margin = avg_net_income / avg_revenues
  - avg_working_capital = avg_current_assets – avg_current_liabilities
  - avg_current_ratio = avg_current_assets / avg_current_liabilities
  - avg_debt_to_equity = avg_long_term_debt / avg_stockholders_equity
  - avg_roe = avg_net_income / avg_stockholders_equity.

Με αυτόν τον τρόπο, το Power BI δουλεύει πάνω σε ένα λεπτό αλλά πολύ “ελαφρύ” fact σε επίπεδο state×quarter, αντί να επανυπολογίζει κάθε φορά τα aggregates από τα raw επίπεδα. Αυτό βελτιώνει σημαντικά την απόδοση και επιτρέπει την ευέλικτη υλοποίηση DAX measures (Quarter Bucket, ranks, volatility, correlations) χωρίς να ξεπερνιούνται τα resource limits [21], [25], [32], [39], [51].

Περιβάλλον & εκδόσεις

Υλικοτεχνική Υποδομή (Hardware): Για την εκτέλεση των σεναρίων χρήσης και τη φιλοξενία των containerized υπηρεσιών, χρησιμοποιήθηκε σταθμός εργασίας (workstation) με τις ακόλουθες τεχνικές προδιαγραφές [39]:

- **Επεξεργαστική Ισχύς**: Κεντρική Μονάδα Επεξεργασίας (CPU) με συχνότητα λειτουργίας 2.30 GHz, η οποία κρίθηκε επαρκής για τη διαχείριση των παράλληλων νημάτων (threads) του *Airbyte* και των υπολογισμών του *dbt*.
- **Μνήμη**: Μνήμη τυχαίας προσπέλασης (RAM) χωρητικότητας 16 GB, επιτρέποντας την άνετη λειτουργία του Docker engine και της βάσης δεδομένων στη μνήμη.
- **Αποθήκευση**: Μονάδα δίσκου SSD χωρητικότητας 500 GB, η οποία εξασφάλισε υψηλές ταχύτητες εγγραφής/ανάγνωσης (I/O) κατά τη διαδικασία του staging και της μεταφοράς δεδομένων.

Λογισμικό και Εκδόσεις (Software Stack) Το περιβάλλον λογισμικού δομήθηκε πάνω στο λειτουργικό σύστημα Windows 10. Ωστόσο, για την προσομοίωση ενός UNIX-based περιβάλλοντος παραγωγής και τη βέλτιστη εκτέλεση των open-source εργαλείων, αξιοποιήθηκε το υποσύστημα WSL (Windows Subsystem for Linux) με διανομή Ubuntu 24.04.

Η τεχνολογική στοίβα περιλαμβάνει τις εξής συγκεκριμένες εκδόσεις:

- **Python 3.12.8**: Χρησιμοποιήθηκε για την ανάπτυξη του custom connector και των βοηθητικών scripts ορχήστρωσης.
- **Docker Desktop 4.41.2**: Παρείχε το απαραίτητο περιβάλλον containerization για την απομόνωση και εκτέλεση των υπηρεσιών του *Airbyte*.
- ***PostgreSQL* 16.4**: Λειτούργησε ως το κεντρικό σύστημα διαχείρισης βάσης δεδομένων (RDBMS) για την αποθήκη δεδομένων (Data Warehouse).
- ***Airbyte* 0.63.13**: Αξιοποιήθηκε ως η πλατφόρμα Data Integration για την άντληση των δεδομένων.
- ***dbt* 0.38.28**: Χρησιμοποιήθηκε για τη μοντελοποίηση και τον μετασχηματισμό των δεδομένων (Transformation layer).
- ***Power BI* Desktop 2.147.1085.0**: Αποτέλεσε το εργαλείο τελικής απεικόνισης και δημιουργίας των αναφορών (Reporting layer).

Προβλήματα και δυσκολίες κατά την υλοποίηση

Κατά την ανάπτυξη της εφαρμογής προέκυψαν ορισμένες τεχνικές και σχεδιαστικές δυσκολίες, οι οποίες οδήγησαν σε επανασχεδιασμό τμημάτων της λύσης και σε βελτιώσεις της αρχιτεκτονικής.

Αυτόματη άντληση από EDGAR και υλοποίηση custom connector

Αρχικά, η μεγαλύτερη πρόκληση ήταν ο σχεδιασμός της αυτόματης άντλησης από το EDGAR. Το επίσημο API απαιτεί σωστό χειρισμό:

- Pagination
- rate limiting
- κατάλληλο User-Agent
- επιλογή των σωστών endpoints (company facts, filings, κ.λπ.)

Η απόφαση να χρησιμοποιηθεί το Python Source Connector του *Airbyte* σε περιβάλλον WSL/Ubuntu σήμαινε ότι έπρεπε να υλοποιηθεί εξ αρχής η λογική κλήσεων και mapping των XBRL πεδίων σε ένα ενιαίο stream. Χρειάστηκε αρκετός πειραματισμός μέχρι να βρεθεί η σωστή ισορροπία ανάμεσα στο μέγεθος κάθε κλήσης, στην ταχύτητα και στον όγκο των δεδομένων (περίπου 550.000 εγγραφές για ~5.400 CIKs σε πλήρη sync).

Ορχήστρωση *Airbyte*–*dbt* και webhook

Ένα πρακτικό εμπόδιο ήταν ότι οι νεότερες εκδόσεις του *Airbyte* δεν προσφέρουν πλέον άμεση ενσωμάτωση με *dbt* Cloud ή *dbt* transformations, όπως συνέβαινε σε παλαιότερες υλοποιήσεις. Αυτό δημιούργησε την ανάγκη για έναν εξωτερικό μηχανισμό ορχήστρωσης.

Η λύση που επιλέχθηκε ήταν η ανάπτυξη ενός Flask webhook service, το οποίο:

- δέχεται POST από το *Airbyte* μετά την ολοκλήρωση ενός sync
- διαβάζει το status από το payload
- σε περίπτωση επιτυχίας εκτελεί το τοπικό *dbt* run μέσω subprocess.run
- καταγράφει stdout/stderr για έλεγχο λαθών

Η υλοποίηση αυτή, αν και απλή, απαίτησε δοκιμές για το σωστό χειρισμό χαρακτήρων, paths σε Windows/WSL και περιβάλλοντος εκτέλεσης, καθώς και για την αποφυγή ταυτόχρονων εκτελέσεων (race conditions) σε περίπτωση πολλαπλών syncs.

Σχεδιασμός Ids και αναφορικής ακεραιότητας στο *dbt*

Στην πρώτη έκδοση του *dbt* project τα Ids των πινάκων companies, metrics και report_periods δημιουργούνταν μέσω συναρτήσεων dense_rank() πάνω στα δεδομένα κάθε run. Αυτό είχε ως αποτέλεσμα:

- όταν γίνονταν νέες εγγραφές ή όταν άλλαζε η σειρά των δεδομένων
- οι τιμές των Ids μπορούσαν να μετακινηθούν
- με συνέπεια να σπάνε οι ξένες κλείδες του company_metric_values και να εμφανίζονται σφάλματα κατά το incremental load

Η λύση ήταν η μετάβαση σε σταθερά hash Ids (md5 των business keys). Με αυτόν τον τρόπο τα Ids είναι ντετερμινιστικά, δεν εξαρτώνται από τη σειρά των δεδομένων και είναι πλήρως αναπαραγώγιμα σε οποιοδήποτε περιβάλλον.

Ποιότητα δεδομένων και state_of_incorporation

Άλλη δυσκολία αφορούσε την πληρότητα και συνέπεια του πεδίου state_of_incorporation. Στα raw δεδομένα υπήρχαν πολλές εγγραφές με NULL ή αντικρουόμενες τιμές για την ίδια εταιρία. Αυτό δημιουργούσε προβλήματα τόσο στο *dbt* (κατά τη δημιουργία διαστάσεων) όσο και στο *Power BI* (κατανομή εταιριών ανά πολιτεία).

- Η προσέγγιση που υιοθετήθηκε ήταν:
- στον πίνακα companies να επιλέγεται ως state_of_incorporation η συχνότερη μη-NULL τιμή ανά CIK
- να αγνοούνται μεμονωμένες τιμές θορύβου (noisy values
- να υπάρχει η δυνατότητα μελλοντικής αναθεώρησης του κανόνα, αν εντοπιστούν καλύτερες πηγές

Απόδοση και σχεδιασμός Quarter Bucket στο *Power BI*

Στα πρώτα prototypes, η λογική σύγκρισης χρονικών περιόδων επιχειρήθηκε να υλοποιηθεί αποκλειστικά σε DAX πάνω στον αρχικό πίνακα με γραμμές ανά εταιρία–ημερομηνία. Αυτό οδήγησε σε:

- μεγάλα query plans
- συχνά resource errors («the query has exceeded the available resources»)
- καθυστερήσεις στην απόκριση των visuals

Η λύση ήταν να μετατοπιστεί η πολυπλοκότητα στο Power Query, με δημιουργία του StateQuarter table και υπολογισμό των μέσων όρων ανά πολιτεία και τρίμηνο εκ των προτέρων. Στη συνέχεια, η μεθοδολογία Quarter Bucket υλοποιήθηκε με DAX πάνω σε αυτό το συμπυκνωμένο dataset. Η αλλαγή αυτή βελτίωσε ουσιαστικά:

- τον χρόνο φόρτωσης
- τη σταθερότητα των αναφορών
- και την καθαρότητα της μεθοδολογίας

Λειτουργικότητα της Εφαρμογής

Στο κεφάλαιο αυτό περιγράφεται η λειτουργία της εφαρμογής από την πλευρά του developer. Δηλαδή πώς συνδέεται η ροή ενημέρωσης δεδομένων με τα measures και τα dashboards, ποια λογική ακολουθείται για τις συγκρίσεις στον χρόνο και με ποιον τρόπο υλοποιούνται οι βασικοί δείκτες απόδοσης (KPIs).

Συνολική λειτουργική εικόνα

Η εφαρμογή υλοποιεί ένα ολοκληρωμένο pipeline ELT → BI, το οποίο, από πλευράς λειτουργικότητας, μπορεί να περιγραφεί ως εξής:

- Σε τακτική βάση (daily CRON), ενεργοποιείται ένας κύκλος συγχρονισμού δεδομένων
- Ο συγχρονισμός αντλεί νέες εταιρικές υποβολές από το SEC EDGAR και ενημερώνει τους πίνακες της βάσης δεδομένων
- Με την επιτυχή ολοκλήρωση του συγχρονισμού ενεργοποιείται αυτόματα ο μηχανισμός μετασχηματισμού (*dbt*), ο οποίος ενημερώνει τα μοντέλα staging, dimensions και fact
- Το *Power BI* συνδέεται στο ενημερωμένο schema και επιτρέπει στον χρήστη να αλληλεπιδράσει με τα dashboards, επιλέγοντας μετρικές, πολιτείες και χρονικά παράθυρα σύγκρισης

Από την οπτική του τελικού χρήστη, η εφαρμογή εμφανίζεται ως ένα σύνολο θεματικών σελίδων, οι οποίες μοιράζονται την ίδια μεθοδολογία για τον χρόνο (Quarter Bucket) και τα ίδια βασικά φίλτρα (slicers). Ο χρήστης δεν χρειάζεται να γνωρίζει τις λεπτομέρειες του backend. Η λειτουργικότητα είναι συγκεντρωμένη στο επίπεδο των *Power BI* αναφορών.

Μετρικές και KPIs — Ορισμοί & Ερμηνεία

Πριν από την ανάλυση των επιμέρους σελίδων, είναι χρήσιμο να συνοψιστούν τα βασικά συνθετικά KPIs που εμφανίζονται στα dashboards. Τα περισσότερα από αυτά δεν είναι πρωτογενή λογιστικά μεγέθη, αλλά παράγωγοι δείκτες που βασίζονται στα StateQuarter averages και σε συγκρίσεις μεταξύ τριμήνων.

KPI	Τύπος / Ορισμός (σε λογικό επίπεδο)	Ερμηνεία
US Average @ qEff	Μέσος όρος των state-level τιμών στο επιλεγμένο metric, στο αποτελεσματικό τρίμηνο qEff.	• Αποτελεί το «benchmark» των Η.Π.Α. στο συγκεκριμένο metric. • Υψηλότερη τιμή σημαίνει ότι, κατά μέσο όρο, οι πολιτείες εμφανίζουν μεγαλύτερο επίπεδο στο δείκτη.
US Δ	US Average(qEff) – US Average(qBaseline)	• Απόλυτη μεταβολή του μέσου όρου Η.Π.Α. ανάμεσα στο τρέχον τρίμηνο και στο baseline

		που ορίζεται από το Months back. • Θετική τιμή = βελτίωση και • Αρνητική τιμή = επιδείνωση.
US Δ%	$\frac{US\ Average(qEff) - US\ Average(qBaseline)}{US\ Average(qBaseline)}$	• Ποσοστιαία μεταβολή του μέσου όρου Η.Π.Α. • Είναι ο πιο άμεσος δείκτης αν η τάση σε εθνικό επίπεδο είναι ανοδική ή καθοδική.
State Value @ qEff	Τιμή του επιλεγμένου metric για συγκεκριμένη πολιτεία στο qEff (StateQuarter average).	• Επίπεδο της πολιτείας στο τρέχον τρίμηνο. • Συγκρίνεται με τον US Average ή με άλλη πολιτεία.
State Δ	$State\ Value(qEff) - State\ Value(qBaseline)$	• Απόλυτη μεταβολή της πολιτείας σε σχέση με το baseline. • Αξιοποιείται κυρίως για μεγέθη σε “μονάδες χρήματος” ή δείκτες χωρίς φυσικό κάτω όριο.
State Δ%	$\frac{State\ Value(qEff) - State\ Value(qBaseline)}{State\ Value(qBaseline)}$	• Ποσοστιαία μεταβολή της πολιτείας. • Θετικές τιμές υποδηλώνουν τη βελτίωση (ή αύξηση) και αρνητικές τιμές την επιδείνωση. • Στις απεικονίσεις χρησιμοποιείται πράσινο για θετικό Δ% και κόκκινο για αρνητικό Δ%. • Μικρές τιμές εντός neutral zone (±0,01%) θεωρούνται ουδέτερες.
State Rank	Θέση της πολιτείας σε κατάταξη RANKX μεταξύ όλων των πολιτειών, με βάση το State Value @ qEff ή το Δ%.	• Για δείκτες “όσο μεγαλύτερο τόσο καλύτερο” (π.χ. Gross Profit, ROE): κατάταξη φθίνουσα • Για δείκτες “όσο μικρότερο τόσο καλύτερο” (π.χ. Debt-to-Equity, Volatility): αντιστροφή κατεύθυνσης. • Δείχνει τη σχετική θέση της πολιτείας σε σχέση με τις υπόλοιπες. • Χαμηλότερος αριθμός (Rank=1) σημαίνει καλύτερη επίδοση.
Coverage %	Αριθμός εταιρειών της πολιτείας με διαθέσιμη τιμή στο qEff ως προς το σύνολο εταιρειών της πολιτείας με ιστορικά δεδομένα.	• Δείχνει τη πληρότητα δεδομένων για την πολιτεία στο δεδομένο τρίμηνο.

		• Χαμηλές τιμές σημαίνουν ότι τα συμπεράσματα θα πρέπει να διαβάζονται με επιφύλαξη.
Volatility	Τυπική απόκλιση των state-level τιμών του metric για N τελευταία quarters (μετρημένη στο StateQuarter).	• Μετρά τη σταθερότητα ή αστάθεια του δείκτη σε κάθε πολιτεία. • Χαμηλή μεταβλητότητα σημαίνει σταθερή συμπεριφορά. • Υψηλή μεταβλητότητα σημαίνει έντονα σκαμπανεβάσματα. • Σε σελίδες όπως η “Volatility” και η “Executive Summary” οι πολιτείες κατατάσσονται από την πιο σταθερή προς την πιο ασταθή.
Correlation	Συντελεστής συσχέτισης Pearson μεταξύ δύο μετρικών (Metric A και Metric B), υπολογισμένος πάνω στα state-level averages των τελευταίων N quarters.	• Δείχνει κατά πόσο “παράλληλα” κινούνται δύο δείκτες σε επίπεδο πολιτειών. • Τιμές κοντά στο +1 σημαίνουν ισχυρή θετική συσχέτιση. • Τιμές κοντά στο −1 δηλώνουν ισχυρή αρνητική συσχέτιση. • Τιμές γύρω στο 0 ασθενή ή μη γραμμική σχέση.

Πίνακας 2 Μετρικές και KPIs — Ορισμοί & Ερμηνεία

Οι χρηματοοικονομικές μετρικές (metrics) που υπολογίζονται στην άντληση και συνολικά στην υλοποίηση του έργου διακρίνονται σε δύο κατηγορίες: βασικές (primary) και παράγωγες (derived). Οι βασικές μετρικές περιλαμβάνουν κρίσιμους δείκτες απόδοσης, όπως τα:

- **Revenues** (έσοδα): Τα έσοδα αποτελούν το σύνολο των πωλήσεων μιας εταιρείας πριν αφαιρεθούν τα κόστη και οι δαπάνες.
- **Gross Profit** (μικτό κέρδος): Ορίζεται ως: Gross Profit = $\frac{Revenues}{Cost\ of\ Goods\ Sold}$ και αντανακλά την παραγωγική αποτελεσματικότητα των επιχειρήσεων.
- **Operating Expenses** (λειτουργικά έξοδα): Περιλαμβάνουν γενικά & διοικητικά έξοδα, marketing, R&D κ.λπ.
  Δείχνει το βάρος λειτουργικού κόστους των εταιρειών.
- **Net Income** (καθαρά κέρδη): Ορίζεται ως: Net Income= Revenues – Expenses και είναι ο πιο άμεσος δείκτης συνολικής κερδοφορίας.
- **Current Assets** (κυκλοφορούν ενεργητικό)
- **Current Liabilities** (βραχυπρόθεσμες υποχρεώσεις): Τα κυκλοφορούντα στοιχεία καθορίζουν τη βραχυπρόθεσμη οικονομική ευχέρεια της εταιρείας.
- **Accounts Payable** (υποχρεώσεις προς προμηθευτές)

- **Long-term Debt** (μακροπρόθεσμο χρέος): Εμφανίζει τη δομή κεφαλαίου και το επίπεδο μόχλευσης.
- **Equity** (ίδια κεφάλαια): Είναι η καθαρή λογιστική αξία που ανήκει στους μετόχους. Αποτελεί θεμελιώδη μέτρηση για τη μελέτη της οικονομικής ευρωστίας.

Οι μεταβλητές αυτές αποτελούν βάση για κάθε χρηματοοικονομικής ανάλυσης, καθώς επίσης αποτυπώνουν τη λειτουργική δομή των επιχειρήσεων.

Οι παράγωγες μετρικές προκύπτουν απο τις βασικές, μέσω υπολογισμών και περιλαμβάνουν δείκτες όπως:

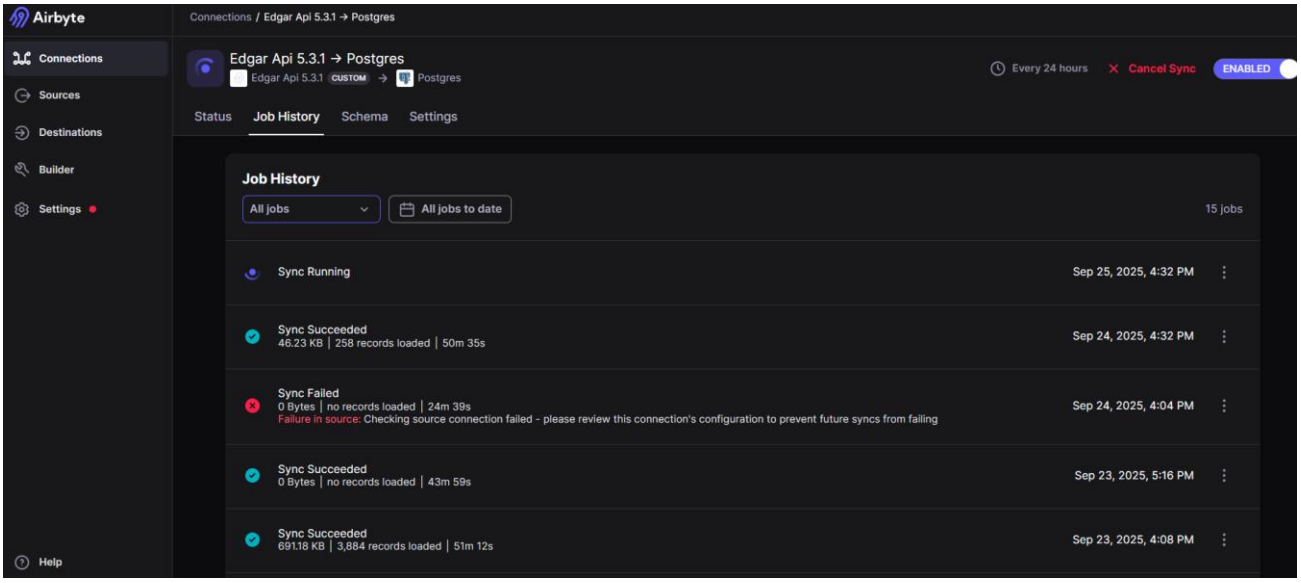
- **Gross Margin** (περιθώριο μικτού κέρδους): Ορίζεται ως: Gross Margin = $\frac{Gross\ Profit}{Revenues}$ και μετρά το περιθώριο κέρδους ανά μονάδα πωλήσεων.
- **Net Margin** (καθαρό περιθώριο): Ορίζεται ως: Net Margin = $\frac{Net\ Income}{Revenues}$ και αποτυπώνει την καθαρή αποδοτικότητα.
- **Working Capital** (κεφάλαιο κίνησης): Ορίζεται ως: Working Capital = Current Assets – Current Liabilities και δείχνει τη βραχυπρόθεσμη χρηματοοικονομική ισορροπία.
- **Current Ratio** (δείκτης κυκλοφορούντος ενεργητικού): Ορίζεται ως: Current Ratio = $\frac{Current\ Assets}{\text{Current Liabilities}}$ και εκφράζει τη δυνατότητα κάλυψης των άμεσων υποχρεώσεων.
- **Debt-to-Equity Ratio** (δείκτης δανειακής μόχλευσης): Ορίζεται ως: Debt/Equity =$\frac{Long\ Term\ Debt}{Stockholder's\ Equity}$. Μετρά τη μόχλευση. Τιμές >1 υποδεικνύουν υψηλότερο δανεισμό από ίδια κεφάλαια.
- **Return on Equity – ROE** (απόδοση ιδίων κεφαλαίων). Ορίζεται ως: Roe =$\frac{Net\ Income}{Stockholder's\ Equity}$ και είναι ένας από τους σημαντικότερους δείκτες απόδοσης για τους επενδυτές.

Αυτοί οι δείκτες δίνουν μια πιο σύνθετη εικόνα της αποδοτικότητας, ρευστότητας και χρηματοοικονομικής σταθερότητας των εταιρειών.

Η διαδικασία υπολογισμού των παραγώγων μετρικών πραγματοποιείται εντός του *Power BI*, μέσω DAX (Data Analysis Expressions). Με αυτόν τον τρόπο, το warehouse διατηρείται ελαφρύ και χρησιμοποιείται απλά για την αποθήκευση των πρωτογενών δεδομένων, ενώ η λογική των υπολογισμών υλοποιείται στο επίπεδο της παρουσίασης. Με αυτόν τον τρόπο υπάρχει μεγαλύτερη ευελιξία και δυνατότητα επαναχρησιμοποίησης των δεικτών.

Developer view – Στιγμιότυπα περιβάλλοντος

Για πληρότητα, στην ενότητα αυτή παρουσιάζονται ενδεικτικά στιγμιότυπα από τα εργαλεία που χρησιμοποιήθηκαν στην υλοποίηση, από την οπτική του αναπτυξιακά υπεύθυνου (developer).



Εικόνα 8 Airbyte connection & run history

Η εικόνα δείχνει τη ρύθμιση της σύνδεσης Airbyte (→ PostgreSQL) και το ιστορικό των ημερήσιων συγχρονισμών. Από εδώ ο διαχειριστής μπορεί να ελέγχει την κατάσταση των runs, τα όγκων (volumes) δεδομένων και τυχόν σφάλματα. Ακόμη, η Εικόνα 3 παρουσιάζει το σχήμα της βάσης στο pgAdmin, με τους πίνακες του μοντέλου αστέρα. Ο πίνακας λειτουργεί ως fact πλαισιωμένος από τους απαραίτητους πίνακες διαστάσεων (dimensions), πάνω στις οποίες «πατάει» το Power BI. Ο πίνακας _531_sec_data_stream_combined είναι τα raw δεδομένα απο το Airbyte. Επίσης, φαίνεται και το View που εκτελείται από το dbt.

Εικόνα 9 Σχήμα Βάσης στο pgAdmin

Λειτουργία ενημέρωσης δεδομένων (Refresh Lifecycle)

Η διαδικασία ενημέρωσης δεδομένων, όπως περιγράφεται στα UML διαγράμματα και στην υλοποίηση, έχει άμεση επίδραση στη λειτουργικότητα των αναφορών.

Σε λειτουργικό επίπεδο:

- Ο συγχρονισμός του *Airbyte* εκτελείται σύμφωνα με ένα ημερήσιο πρόγραμμα (daily CRON).
- Μετά την επιτυχή ολοκλήρωση, το *Airbyte* αποστέλλει webhook προς τη Flask υπηρεσία, η οποία εκκινεί το *dbt* run.
- Το *dbt* ενημερώνει τους πίνακες companies, metrics, report_periods και company_metric_values, χρησιμοποιώντας incremental / merge λογική, ώστε να μην απαιτείται πλήρες rebuild.
- Όταν ολοκληρωθεί ο κύκλος ELT, τα δεδομένα είναι διαθέσιμα για νέο Refresh στο *Power BI*.

Η λειτουργικότητα αυτή έχει δύο σημαντικές συνέπειες:

- Ο χρήστης μπορεί να εκτελεί refresh στην αναφορά του *Power BI* γνωρίζοντας ότι η βάση δεδομένων είναι ανανεωμένη.
- Η εφαρμογή μπορεί να επεκταθεί σε μεγαλύτερους όγκους δεδομένων, χωρίς ο χρόνος ενημέρωσης να γίνεται απαγορευτικός, καθώς η incremental προσέγγιση περιορίζει τις αλλαγές στα νέα ή τροποποιημένα records.

Κοινή διαδραστική λογική

Όλες οι σελίδες της εφαρμογής μοιράζονται μια κοινή λογική, η οποία βασίζεται σε συγκεκριμένους slicers και παραμέτρους:

- Metric: επιλογή οικονομικού μεγέθους (π.χ. Revenues, Gross Profit, Current Assets, Net Income, Gross Margin, Debt-to-Equity, ROE κ.λπ.).
- MonthsBack: ορίζει το χρονικό παράθυρο σύγκρισης μεταξύ τρέχοντος και παρελθόντος τριμήνου. Χρησιμοποιούνται τιμές όπως 0, 3, 6, 12, 24, 36, 48 μήνες.
- State / State A / State B: επιλογή μιας ή δύο πολιτειών για εξέταση ή σύγκριση.
- Quarter Window: ορίζει ποια τρίμηνα εμφανίζονται στα γραφήματα (π.χ. τα τελευταία N quarters).

Η διαδραστικότητα αυτή επιτρέπει στον χρήστη:

- να επιλέγει τη μετρική που τον ενδιαφέρει,
- να καθορίζει πόσο «πίσω» στο χρόνο θέλει να κοιτάξει,
- να εστιάζει είτε σε συγκεκριμένη πολιτεία είτε στη συνολική εικόνα των Η.Π.Α.

Σημαντικό στοιχείο είναι ότι τα slicers λειτουργούν πάνω σε ήδη συγκεντρωμένα δεδομένα επιπέδου State×Quarter, γεγονός που εξασφαλίζει άμεση απόκριση των γραφικών.

Μεθοδολογία Quarter Bucket και βασικά KPIs

Κεντρική ιδέα της λειτουργικότητας της εφαρμογής αποτελεί η μεθοδολογία Quarter Bucket. Αντί οι συγκρίσεις να γίνονται σε επίπεδο ημερομηνίας (η οποία διαφέρει από εταιρία σε εταιρία), όλες οι αναφορές ευθυγραμμίζουν τα δεδομένα σε ημερολογιακά τρίμηνα.

Λειτουργικά, για κάθε επιλογή του χρήστη:

- ορίζεται ένα AsOf Quarter (το πιο πρόσφατο διαθέσιμο τρίμηνο με δεδομένα στο dataset)
- υπολογίζεται ένα Past Quarter, με βάση την παράμετρο MonthsBack
- επιλέγεται για κάθε πολιτεία η τελευταία διαθέσιμη εταιρική τιμή μέσα στο τρίμηνο (Last-in-Quarter), με δυνατότητα fallback στην πιο πρόσφατη τιμή ≤ τέλος τριμήνου
- υπολογίζονται οι δείκτες:

- Τρέχουσα τιμή (Current)
- Διαφορά (Δ = Current − Past)
- Ποσοστιαία διαφορά (Δ% = Δ / Past)
- Κατάταξη (Rank) της πολιτείας μεταξύ όσων έχουν τιμή στο ίδιο quarter
- Coverage % (ποσοστό εταιριών που έχουν διαθέσιμη τιμή στο συγκεκριμένο τρίμηνο)
- Volatility (τυπική απόκλιση ή συντελεστής μεταβλητότητας σε παράθυρο N τριμήνων)
- Correlation % μεταξύ δύο μετρικών (σε συγκεκριμένο χρονικό παράθυρο).

Επιπλέον, εφαρμόζεται μια neutral zone για πολύ μικρές ποσοστιαίες μεταβολές (π.χ. |Δ%| < 0,01%), οι οποίες εμφανίζονται ως ουδέτερες, ώστε να μην δίνεται υπερβολική σημασία σε αμελητέες διαφορές.

Σελίδες και λειτουργικότητα

State Overview

Σκοπός: Η συνολική- συγκριτική εικόνα της οικονομικής κατάστασης όλων των πολιτειών των Ηνωμένων Πολιτειών για ένα επιλεγμένο οικονομικό μέγεθος (metric).Στόχος είναι η παρουσίαση της πανοραμικής μεταβολής (Δ και Δ%) του επιλεγμένου δείκτη σε σχέση με ένα προηγούμενο χρονικό σημείο (MonthsBack).

Κύρια στοιχεία: Slicers επιλογής metric και monthBack, Shape Map των πολιτειών των Η.Π.Α. χρωματικά κωδικοποιημένες με βάση τη μεταβολή Δ%, KPIs με US Average (μέση τιμή του δείκτη για όλες τις πολιτείες στο επιλεγμένο διάστημα) και US Δ και Δ% (απόλυτη και ποσοστιαία μεταβολή για το επιλεγμένο διάστημα), πίνακας Κατάταξης που περιλαμβάνει κάθε πολιτεία, την τελευταία διαθέσιμη τιμή, τη μεταβολή Δ%, και τον αριθμό εταιριών που συνεισφέρουν στο δείγμα, Συνοπτικοί δείκτες που δείχνουν τις πολιτείες με τη μεγαλύτερη θετική και αρνητική μεταβολή αντίστοιχα.

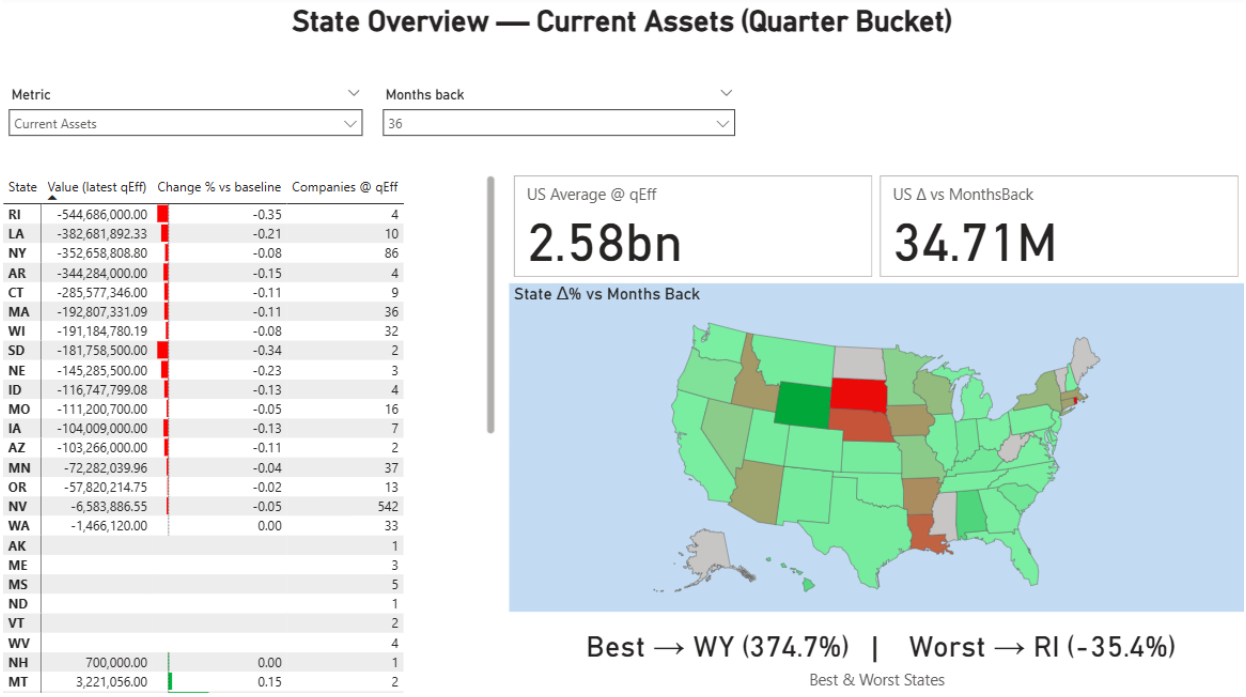
Μεθοδολογία: Όλα τα μεγέθη βασίζονται σε δεδομένα που έχουν ευθυγραμμιστεί χρονικά μέσω της λογικής Quarter Bucket, εξασφαλίζοντας ότι κάθε μέτρηση αφορά το ίδιο ημερολογιακό τρίμηνο για όλες τις πολιτείες.
Η μεταβολή (Δ και Δ%) υπολογίζεται συγκρίνοντας το τελευταίο διαθέσιμο τρίμηνο στο dataset, και το τρίμηνο που αντιστοιχεί στο διάστημα MonthsBack μηνών πριν.Τα Δ% υπολογίζονται ως:

$$\Delta\% = \frac{\text{CurrentQuarter} - \text{PastQuarter}}{\text{PastQuarter}}$$

όπου το PastQuarter καθορίζεται δυναμικά με βάση το επιλεγμένο παράθυρο χρόνου.

Ερμηνεία:Η σελίδα επιτρέπει την ταχεία αναγνώριση πολιτειών που βελτιώνονται ή υποχωρούν σε σχέση με τον εθνικό μέσο όρο. Οι πράσινες περιοχές υποδηλώνουν θετική μεταβολή, ενώ οι κόκκινες αρνητική. Ο πίνακας κατάταξης, σε συνδυασμό με τα KPIs και

τους συνοπτικούς δείκτες, διευκολύνει τον εντοπισμό ακραίων περιπτώσεων (outperformers / underperformers) και γεωγραφικών προτύπων.



State	Value (latest qEff)	Change % vs baseline	Companies @ qEff
RI	-544,686,000.00	-0.35	4
LA	-382,681,892.33	-0.21	10
NY	-352,658,808.80	-0.08	86
AR	-344,284,000.00	-0.15	4
CT	-285,577,346.00	-0.11	9
MA	-192,807,331.09	-0.11	36
WI	-191,184,780.19	-0.08	32
SD	-181,758,500.00	-0.34	2
NE	-145,285,500.00	-0.23	3
ID	-116,747,799.08	-0.13	4
MO	-111,200,700.00	-0.05	16
IA	-104,009,000.00	-0.13	7
AZ	-103,266,000.00	-0.11	2
MN	-72,282,039.96	-0.04	37
OR	-57,820,214.75	-0.02	13
NV	-6,583,886.55	-0.05	542
WA	-1,466,120.00	0.00	33
AK			1
ME			3
MS			5
ND			1
VT			2
WV			4
NH	700,000.00	0.00	1
MT	3,221,056.00	0.15	2

Εικόνα 10 State Overview

State Detail

Σκοπός:Η σελίδα State Detail εστιάζει σε μία επιλεγμένη πολιτεία (State) και παρέχει λεπτομερή εικόνα για την εξέλιξη ενός metric στο χρόνο, σε σύγκριση με τον εθνικό μέσο όρο (US Average). Επιπλέον, παρουσιάζει την κάλυψη (Coverage %) και τη θέση της πολιτείας στην κατάταξη (Rank) στο τρέχον ή στο πιο πρόσφατο διαθέσιμο τρίμηνο.

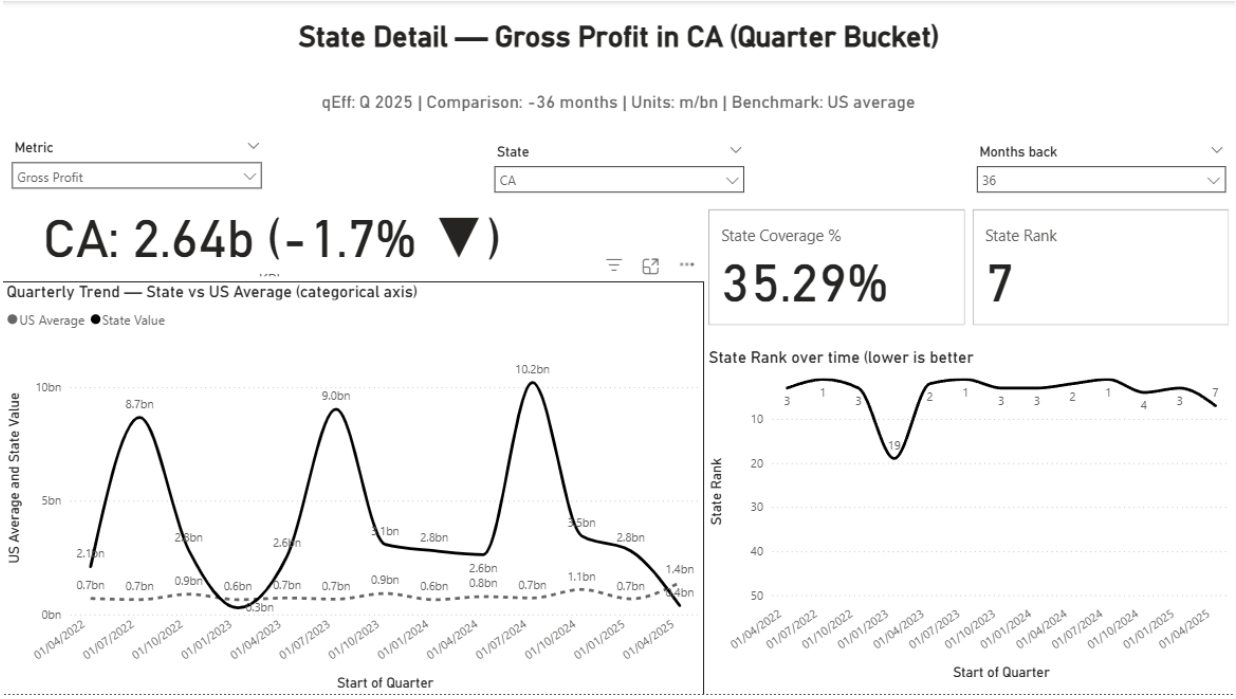
Κύρια στοιχεία:

- Slicers: επιλογή State, Metric και MonthsBack.
- Γραμμή State vs US: line chart που εμφανίζει την πορεία της επιλεγμένης πολιτείας και του US average ανά quarter, στο επιλεγμένο χρονικό παράθυρο.
- Coverage %: δείκτης που παρουσιάζει το ποσοστό των εταιριών της πολιτείας που έχουν διαθέσιμη τιμή για τη μετρική στο τρέχον quarter.
- Rank over time: γράφημα που δείχνει την εξέλιξη της κατάταξης της πολιτείας ανά τρίμηνο.
- Κάρτες KPIs: τρέχουσα τιμή, Δ και Δ% έναντι Past Quarter, καθώς και Current Rank.

Μεθοδολογία: Η «άγκυρα» χρόνου είναι το AsOf Quarter, δηλαδή το τελευταίο διαθέσιμο quarter στο StateQuarter. Αν για την επιλεγμένη πολιτεία λείπει τιμή στο AsOf (π.χ. δεν υπάρχουν πρόσφατες υποβολές), τότε υπολογίζεται ένα Effective Quarter (qEff) ως το πιο πρόσφατο quarter ≤ AsOf στο οποίο υπάρχει διαθέσιμη τιμή, το Rank της πολιτείας υπολογίζεται μόνο μεταξύ των πολιτειών που έχουν επίσης τιμή στο qEff, ώστε να

αποφεύγονται άδικες συγκρίσεις και κενές τιμές. Για τη γραφική απεικόνιση, οι καμπύλες State και US «γεμίζουν» με fallback στις τελευταίες διαθέσιμες τιμές μέχρι κάθε σημείο, ώστε η εικόνα να είναι συνεχής ακόμη κι αν λείπουν ενδιάμεσα τρίμηνα.

Ερμηνεία: Ο χρήστης μπορεί να διαπιστώσει αν η πολιτεία κινείται πάνω ή κάτω από τον εθνικό μέσο όρο, αν βελτιώνει ή χειροτερεύει τη θέση της στην κατάταξη, καθώς και πόσο «πυκνή» είναι η κάλυψη δεδομένων της (Coverage). Η fallback λογική στο qEff εξασφαλίζει ότι η κατάταξη είναι πάντα ερμηνεύσιμη και δεν αφήνει κενά λόγω ασύγχρονων υποβολών.



Εικόνα 11 State Detail

State Comparison (A vs B)

Σκοπός: Η σελίδα State Comparison επιτρέπει την άμεση και δίκαιη σύγκριση δύο πολιτειών (State A και State B) ως προς ένα επιλεγμένο metric, στο ίδιο χρονικό πλαίσιο.

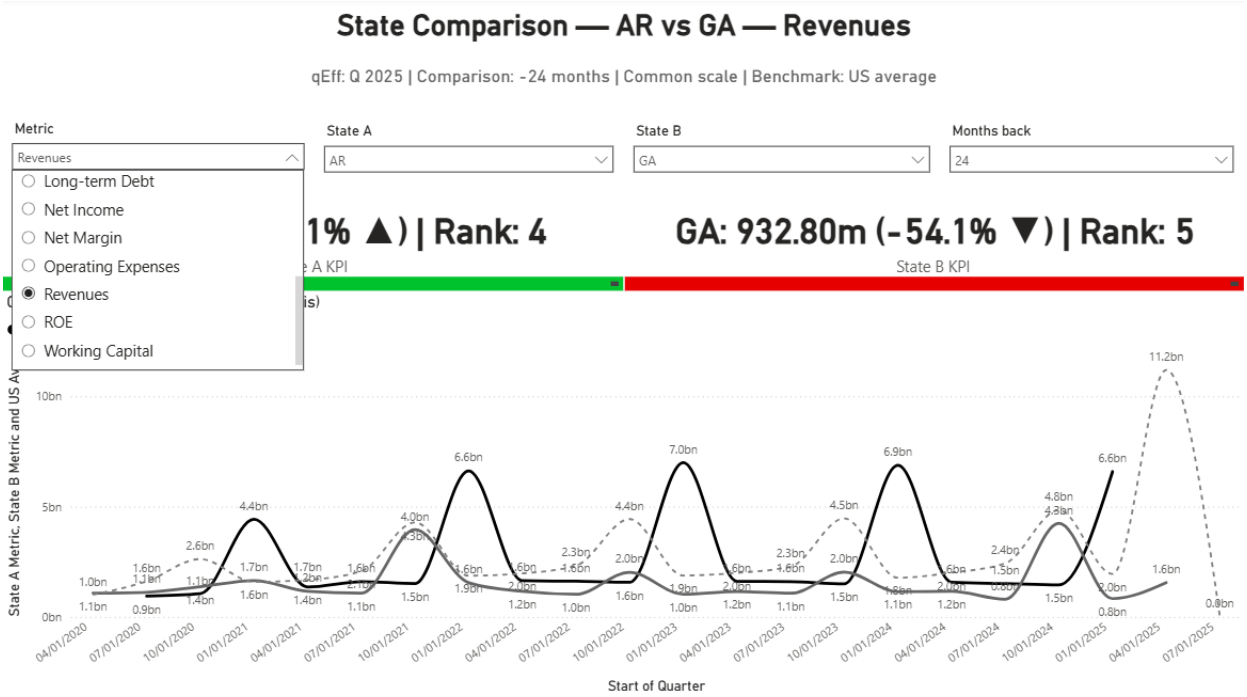
Κύρια στοιχεία:

- Slicers: State A, State B, Metric, MonthsBack.
- Κάρτες KPIs: τρέχουσες τιμές, Δ και Δ% για κάθε πολιτεία, καθώς και τα αντίστοιχα ranks.
- Γραφήματα τάσης: κοινά line charts για State A, State B και US Average, με κοινή κλίμακα στον άξονα τιμών.
- Συγκριτικοί δείκτες: ενδείξεις που τονίζουν ποια πολιτεία προηγείται στο qEff και ποια παρουσιάζει μεγαλύτερη βελτίωση ή επιδείνωση.

Μεθοδολογία: Όλες οι συγκρίσεις γίνονται με βάση τη λογική Quarter Bucket, ώστε State A και State B να αξιολογούνται στο ίδιο τρίμηνο (qEff) και στο ίδιο Past Quarter σύμφωνα με το MonthsBack. Οι άξονες τιμών είναι κοινής κλίμακας, ώστε η οπτική σύγκριση να είναι δίκαιη και να μην αλλοιώνεται από διαφορετικά ranges.

Τα ranks υπολογίζονται για κάθε πολιτεία σε σχέση με όλες τις πολιτείες που έχουν διαθέσιμη τιμή στο ίδιο quarter, χρησιμοποιώντας dense ranking.

Ερμηνεία: Η σελίδα υποστηρίζει σενάρια όπως: «Ποια πολιτεία έχει σήμερα καλύτερη επίδοση στο metric;», «Ποια βελτιώθηκε περισσότερο τα τελευταία 12 ή 24 μήνες;», «Πώς τοποθετούνται οι δύο πολιτείες σε σχέση με τον εθνικό μέσο όρο;». Η κοινή κλίμακα trends και η χρήση Quarter Bucket κάνουν τη σύγκριση πιο διαφανή και πιο σταθερή στο χρόνο.



Εικόνα 12 State Comparison

Volatility

Σκοπός: Η σελίδα Volatility εξετάζει τη σχέση ανάμεσα στο μέσο επίπεδο μιας μετρικής και στη σταθερότητα/μεταβλητότητά της ανά πολιτεία, συνήθως στα τελευταία 8 quarters. Στόχος είναι να αποτυπωθεί ποια πολιτεία εμφανίζει υψηλές τιμές με χαμηλή μεταβλητότητα και ποια παρουσιάζει έντονες διακυμάνσεις.

Κύρια στοιχεία:

- Scatter plot με:
  - άξονα x = μέση τιμή (average level) της μετρικής ανά πολιτεία
  - άξονα y = τυπική απόκλιση (std-dev) ή σχετικός δείκτης μεταβλητότητας
  - μέγεθος φυσαλίδας ανάλογο με τον αριθμό εταιριών (coverage)
- Tooltips με πληροφορίες για:
  - μέση τιμή
  - μεταβλητότητα

- coverage (πλήθος εταιριών)

- Προαιρετικές γραμμές αναφοράς (μέσος όρος level/volatility) που χωρίζουν το διάγραμμα σε τεταρτημόρια.

Μεθοδολογία: Για κάθε πολιτεία υπολογίζονται, σε παράθυρο N τελευταίων quarters:

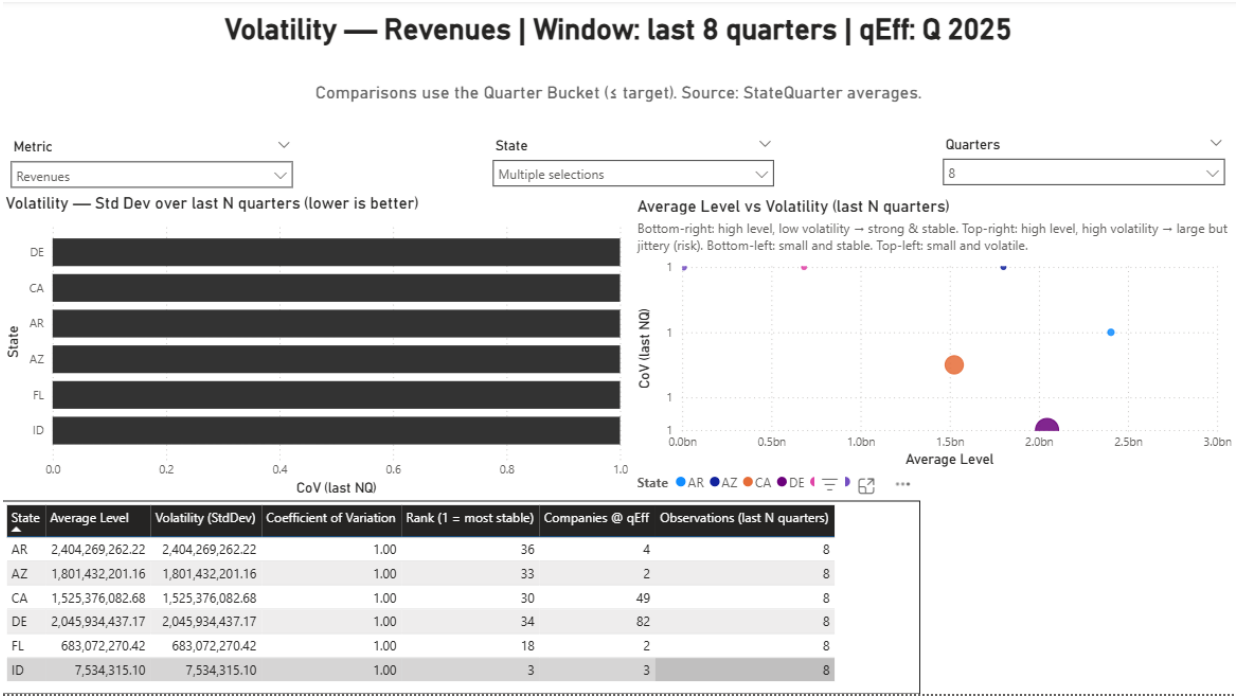
- ο μέσος όρος της επιλεγμένης μετρικής (unweighted μέσος StateQuarter),
- η τυπική απόκλιση ή ένας σχετικός δείκτης μεταβλητότητας (π.χ. std-dev / mean).

Οι υπολογισμοί γίνονται πάνω στο επίπεδο State×Quarter, με βάση τα επικυρωμένα StateQuarter averages. Έτσι, η ανάλυση δεν επηρεάζεται από εταιρικές ιδιαιτερότητες, αλλά αποτυπώνει τη συμπεριφορά της πολιτείας ως συνόλου.

Ερμηνεία:

- Κάτω–δεξιά: υψηλό επίπεδο και χαμηλή μεταβλητότητα (ισχυρή και σταθερή κατάσταση).
- Πάνω–δεξιά: υψηλό επίπεδο αλλά υψηλή μεταβλητότητα (δυναμική αλλά ασταθής εικόνα).
- Κάτω–αριστερά: χαμηλό επίπεδο με χαμηλή μεταβλητότητα (σταθερά χαμηλές τιμές).
- Πάνω–αριστερά: χαμηλό επίπεδο και υψηλή μεταβλητότητα (αδύναμη και ασταθής κατάσταση).

Η σελίδα βοηθά στον εντοπισμό πολιτειών που συνδυάζουν υψηλές τιμές με σταθερότητα (ιδανική κατάσταση) ή παρουσιάζουν σημαντικό ρίσκο λόγω μεταβλητότητας.



State	Average Level	Volatility (StdDev)	Coefficient of Variation	Rank (1 = most stable)	Companies @ qEff	Observations (last N quarters)
AR	2,404,269,262.22	2,404,269,262.22	1.00	36	4	8
AZ	1,801,432,201.16	1,801,432,201.16	1.00	33	2	8
CA	1,525,376,082.68	1,525,376,082.68	1.00	30	49	8
DE	2,045,934,437.17	2,045,934,437.17	1.00	34	82	8
FL	683,072,270.42	683,072,270.42	1.00	18	2	8
ID	7,534,315.10	7,534,315.10	1.00	3	3	8

Εικόνα 13 Volatility

Correlation Explorer

Σκοπός: Η σελίδα Correlation Explorer εξετάζει τη συσχέτιση (correlation) μεταξύ δύο οικονομικών μετρικών (Metric A και Metric B) σε επίπεδο πολιτείας, χρησιμοποιώντας κοινό χρονικό πλαίσιο (Quarter Bucket). Στόχος είναι να διαπιστωθεί κατά πόσο δύο δείκτες κινούνται παράλληλα ή ανεξάρτητα μεταξύ τους.

Κύρια στοιχεία:

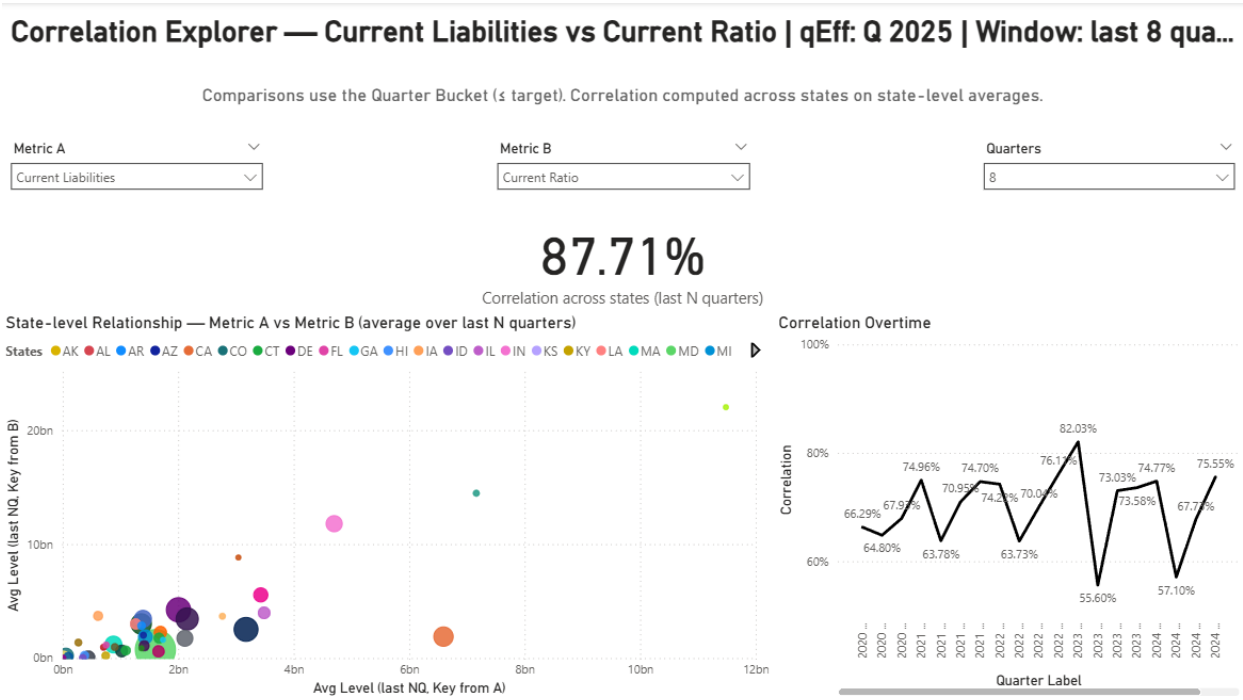
- Slicers: επιλογή Metric A, Metric B, χρονικού παραθύρου.
- Συνολικός δείκτης συσχέτισης (correlation %) για το επιλεγμένο διάστημα.
- Scatter plot με κάθε πολιτεία ως σημείο (x = μέση τιμή Metric A, y = μέση τιμή Metric B).
- Γράφημα εξέλιξης συσχέτισης ανά quarter, που δείχνει πώς μεταβάλλεται η σχέση των δύο δεικτών στο χρόνο.

Μεθοδολογία: Ο υπολογισμός της συσχέτισης βασίζεται στα StateQuarter averages των δύο μετρικών για τα τελευταία N quarters. Για κάθε πολιτεία χρησιμοποιείται ένα συνεπές δείγμα χρονικών σημείων, ώστε:

- να διασφαλιστεί ότι η συσχέτιση αντανακλά την πραγματική κοινή κίνηση στο χρόνο.
- να ελαχιστοποιούνται τα κενά δεδομένων μέσω της Quarter Bucket λογικής.

Η συνολική συσχέτιση (π.χ. Pearson correlation) υπολογίζεται πάνω σε αυτό το dataset, ενώ τα γραφήματα επιτρέπουν την ανάλυση τόσο ανά πολιτεία όσο και σε εθνικό επίπεδο.

Ερμηνεία: Υψηλή τιμή συσχέτισης (π.χ. 0,8 ή 80%) σημαίνει ότι, σε γενικές γραμμές, όταν αυξάνεται η Metric A αυξάνεται και η Metric B (ή αντίστοιχα για μείωση). Χαμηλή ή αρνητική συσχέτιση υποδηλώνει ανεξάρτητη ή αντίθετη κίνηση. Ο συνδυασμός scatter plot και γραφήματος εξέλιξης βοηθά στη διάκριση δομικών σχέσεων από παροδικά φαινόμενα.



Εικόνα 14 Correlation Explorer

Executive Summary

Σκοπός: Η σελίδα Executive Summary συγκεντρώνει σε μία οθόνη τα βασικά συμπεράσματα για:

- τη μεταβολή του επιλεγμένου metric σε εθνικό επίπεδο (US Δ, US Δ%),
- τις πολιτείες με την καλύτερη και τη χειρότερη επίδοση,
- τις πολιτείες με υψηλή μεταβλητότητα.

Απευθύνεται σε χρήστες που θέλουν μία γρήγορη συνοπτική εικόνα χωρίς να πλοηγηθούν σε λεπτομερείς σελίδες.

Κύρια στοιχεία:

- KPIs: US Δ και US Δ% για το επιλεγμένο χρονικό διάστημα.
- Best/Worst state: πολιτεία με τη μεγαλύτερη θετική και μεγαλύτερη αρνητική μεταβολή.

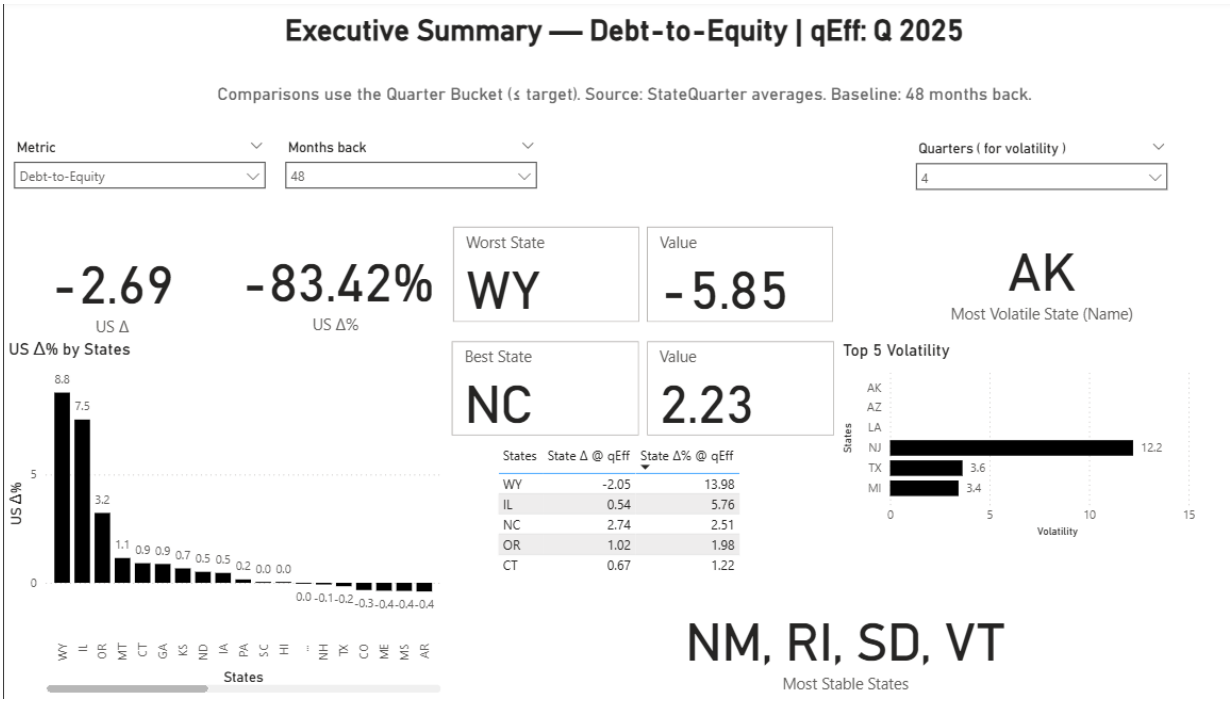
- Top 5 Volatile states: λίστα πολιτειών με υψηλότερη μεταβλητότητα στο επιλεγμένο παράθυρο.
- Πίνακας με State, Δ%, Volatility και πιθανές ταξινομήσεις (π.χ. κατά Δ% ή volatility).

Μεθοδολογία: Η σελίδα βασίζεται σε Quarter Bucket averages με baseline τυπικά MonthsBack = 48 (τετραετία), ώστε οι μεταβολές να αποτυπώνουν μεσοπρόθεσμη τάση και όχι μόνο βραχυπρόθεσμο θόρυβο. Για κάθε πολιτεία υπολογίζονται:

- Δ και Δ% μεταξύ Current Quarter και Past Quarter (48 μήνες πίσω),
- δείκτες Volatility σε παράθυρο N quarters.

Οι τιμές αντλούνται από το StateQuarter και τις παραγώγες μετρικές του.

Ερμηνεία: Η σελίδα λειτουργεί σαν "ταμπλό" που αναδεικνύει ποια κατεύθυνση ακολουθεί το metric συνολικά στις Η.Π.Α. και ποιες πολιτείες ξεχωρίζουν θετικά ή αρνητικά. Ο συνδυασμός Δ%, Volatility και Best/Worst επιτρέπει μια γρήγορη αλλά ουσιαστική πρώτη εικόνα.



Εικόνα 15 Executive Summary

National Trends

Σκοπός: Η σελίδα National Trends εστιάζει στην εθνική τάση ενός metric ανά τρίμηνο, προβάλλοντας την πορεία του US average στο χρόνο και τις αντίστοιχες ποσοστιαίες μεταβολές ανά quarter.

Κύρια στοιχεία:

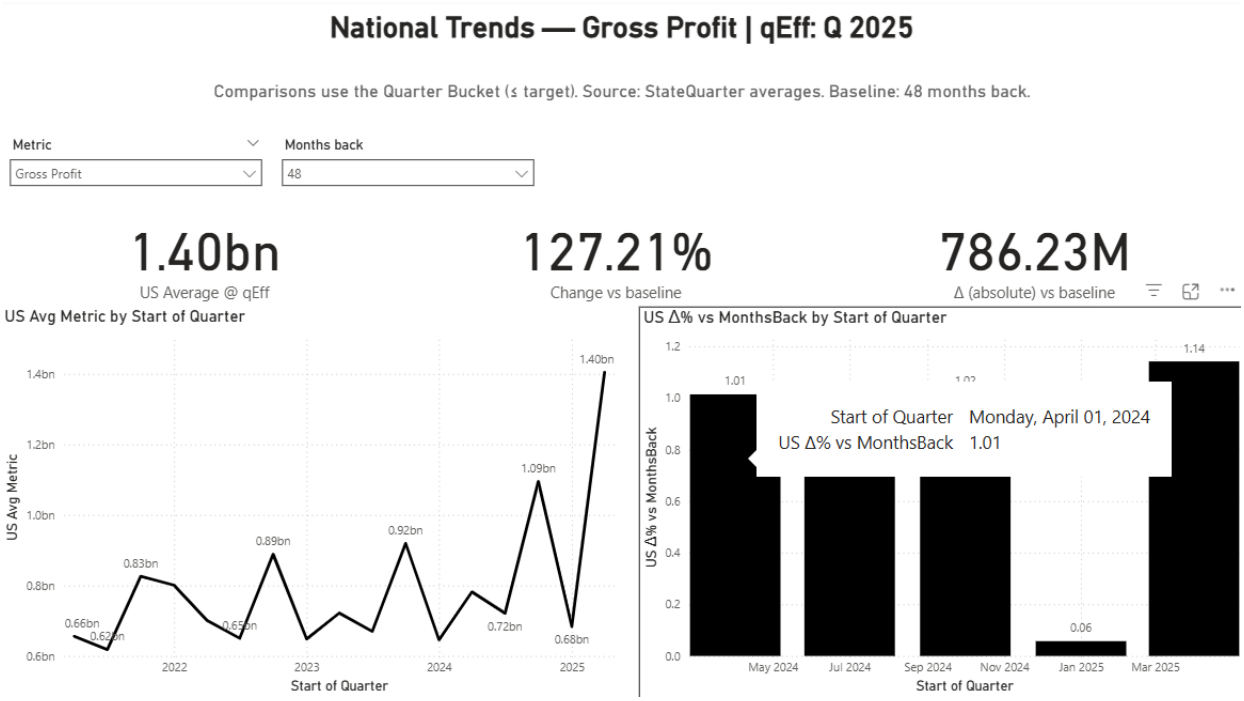
- Line chart με US Average ανά quarter (QuarterStart).

- Bar chart με Δ% ανά quarter σε σχέση με προηγούμενο quarter ή με baseline.
- KPIs που συνοψίζουν την συνολική μεταβολή στο επιλεγμένο χρονικό παράθυρο (π.χ. 48 μήνες).

Μεθοδολογία: Η σελίδα συγκρίνει τις μέσες τιμές των US quarters σε χρονικό παράθυρο (π.χ. 48 μηνών), χρησιμοποιώντας την ίδια λογική Quarter Bucket. Οι μεταβολές υπολογίζονται με συνεπές cadence (quarter-to-quarter ή quarter vs baseline), ώστε ο χρήστης να βλέπει:

- πότε η τάση είναι ανοδική/καθοδική,
- πότε υπάρχουν περίοδοι στασιμότητας ή έντονης μεταβολής.

Ερμηνεία: Η σελίδα δίνει μία καθαρή εικόνα για το αν η επιλεγμένη μετρική βελτιώνεται, χειροτερεύει ή παραμένει σταθερή σε εθνικό επίπεδο, και βοηθά στην τοποθέτηση των παρατηρήσεων σε state-level (άλλες σελίδες) μέσα σε ένα γενικότερο μακροοικονομικό πλαίσιο.



Εικόνα 16 National Trends

Financial Health — Liquidity & Leverage

Σκοπός: Η σελίδα Financial Health – Liquidity & Leverage αποτυπώνει τη ρευστότητα (liquidity), τη μόχλευση (leverage) και γενικότερα την κεφαλαιακή ισχύ των πολιτειών. Στόχος είναι να εντοπιστούν πολιτείες με ισορροπημένη ή μη κεφαλαιακή δομή, συνδυάζοντας δείκτες όπως το Current Ratio και το Debt-to-Equity.

Κύρια στοιχεία:

- KPIs:

  - US Avg Current Ratio,
  - US Avg Debt-to-Equity στο qEff.
- Scatter plot:
  - x = Avg Current Ratio,
  - y = Avg Debt-to-Equity,
  - μέγεθος σημείου ανάλογο με τον αριθμό εταιριών (coverage).
- Bar chart: Avg Debt-to-Equity ανά πολιτεία.
- Line chart: US Avg Current Ratio ανά quarter, για την παρακολούθηση τάσης.
- KPI “Most Leveraged State”: πολιτεία με το υψηλότερο Debt-to-Equity.

Μεθοδολογία: Όλες οι μετρικές προκύπτουν από τα StateQuarter averages.

Συγκεκριμένα:

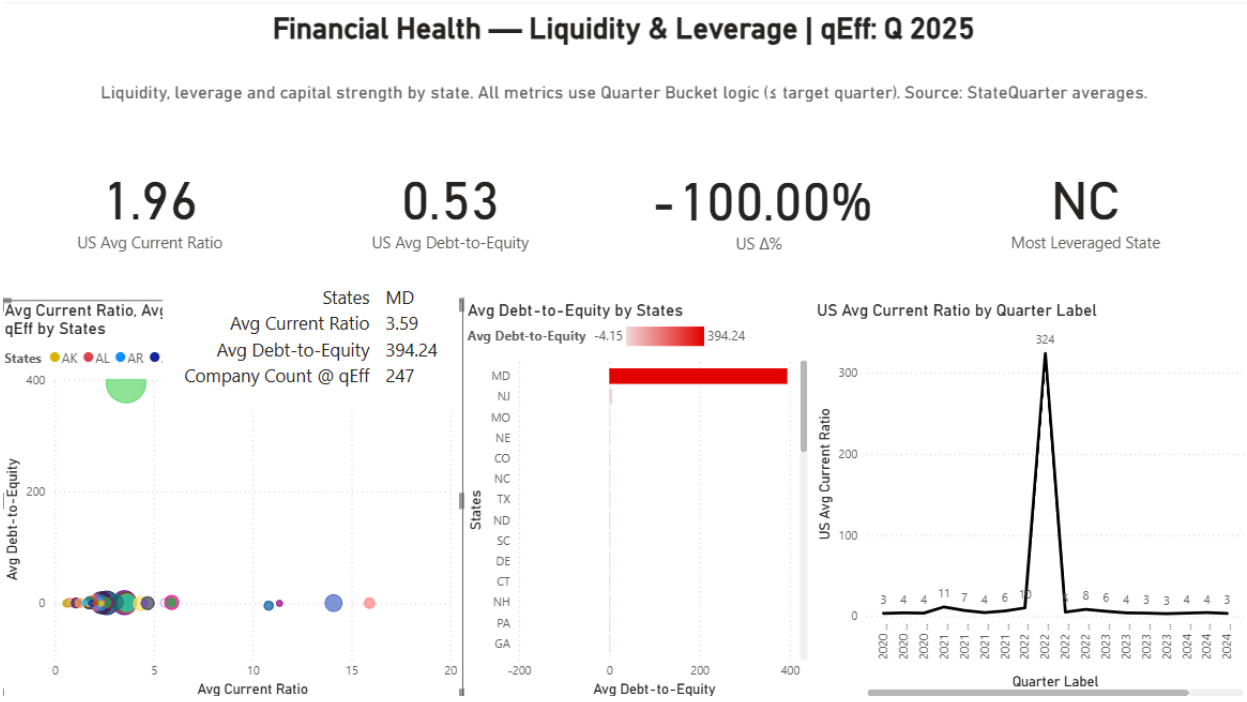
- ο δείκτης Debt-to-Equity υπολογίζεται ως avg_long_term_debt / avg_stockholders_equity,
- το Current Ratio υπολογίζεται ως avg_current_assets / avg_current_liabilities.

Οι υπολογισμοί γίνονται σε επίπεδο πολιτείας, πάνω σε unweighted averages των εταιριών που δραστηριοποιούνται στην κάθε πολιτεία, χρησιμοποιώντας τη λογική Quarter Bucket για χρονική ευθυγράμμιση.

Ερμηνεία: Μεγαλύτερες τιμές Debt-to-Equity υποδηλώνουν υψηλότερη χρηματοοικονομική μόχλευση, δηλαδή μεγαλύτερη εξάρτηση από δανεισμό. Υψηλότερες τιμές Current Ratio δείχνουν αυξημένη ικανότητα κάλυψης βραχυπρόθεσμων υποχρεώσεων και, συνεπώς, καλύτερη ρευστότητα.

Η οπτική συνδυαστική απεικόνιση (scatter + bar + line) επιτρέπει τον εντοπισμό:

- πολιτειών με υγιή ισορροπία ρευστότητας και μόχλευσης,
- πολιτειών που ενδεχομένως είναι υπερβολικά leveraged,
- και πολιτειών με χαμηλή ρευστότητα που μπορεί να υποδηλώνει αυξημένο κίνδυνο.



Εικόνα 17 Financial Health

Μελέτες Περίπτωσης (Case Studies) και Σενάρια Ανάλυσης

Η αξία ενός συστήματος Επιχειρηματικής Ευφυΐας (BI) δεν κρίνεται μόνο από την τεχνική του αρτιότητα, αλλά κυρίως από την ικανότητά του να απαντά σε πραγματικά ερωτήματα. Στο κεφάλαιο αυτό παρουσιάζονται τρία διακριτά σενάρια χρήσης (use cases), τα οποία αναδεικνύουν πώς η εφαρμογή μπορεί να χρησιμοποιηθεί για την εξαγωγή οικονομικών συμπερασμάτων. Μέσω αυτών των σεναρίων.

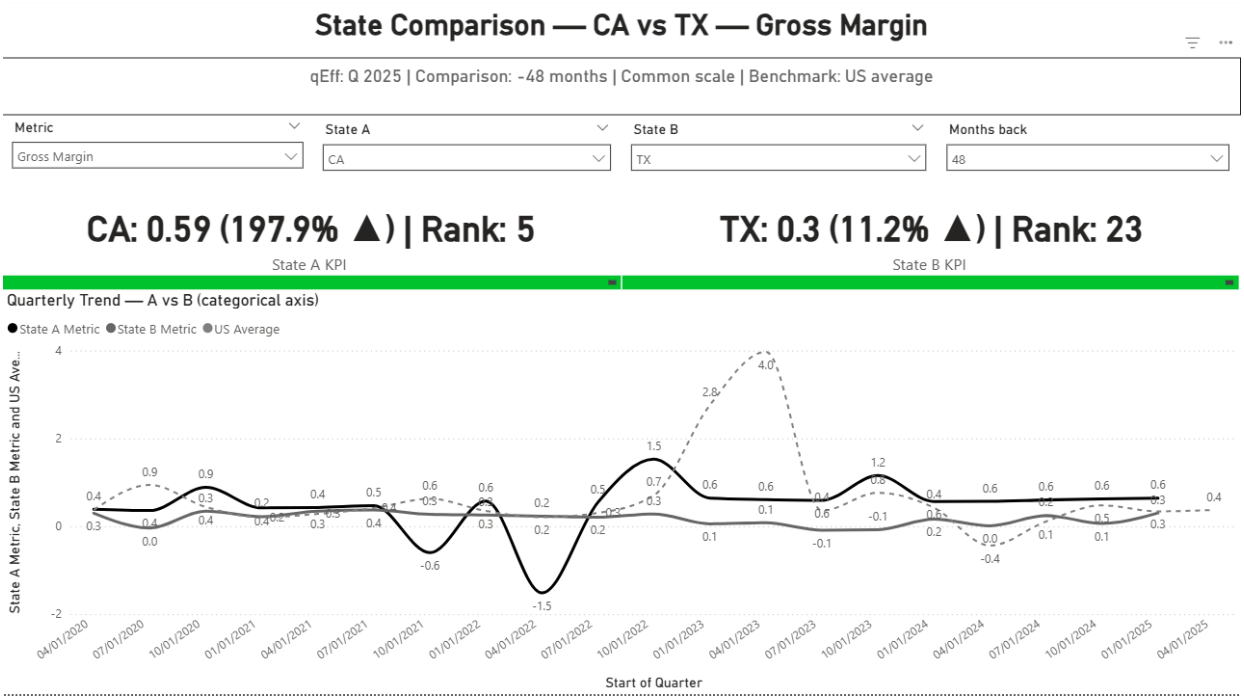
Σενάριο 1: Συγκριτική Ανάλυση Τομεακών Οικονομιών (Καλιφόρνια vs Τέξας)

Ένα ενδιαφέρον χαρακτηριστικό της αμερικανικής οικονομίας είναι η εξειδίκευση των πολιτειών σε διαφορετικούς κλάδους. Για παράδειγμα, η Καλιφόρνια (CA) είναι γνωστή ως έδρα τεχνολογικών κολοσσών (Silicon Valley), ενώ το Τέξας (TX) κυριαρχείται από εταιρείες ενέργειας και βιομηχανίας.

Χρησιμοποιώντας τη σελίδα "State Comparison (A vs B)", επιχειρούμε να συγκρίνουμε τις δύο αυτές πολιτείες ως προς το Gross Margin (Περιθώριο Μικτού Κέρδους).

Παρατηρήση:

- **Gross Margin**: Επιλέγοντας το *Gross Margin*, παρατηρούμε ότι η Καλιφόρνια εμφανίζει συστηματικά υψηλότερα ποσοστά σε σχέση με το Τέξας. Αυτό είναι αναμενόμενο και επιβεβαιώνει την εγκυρότητα των δεδομένων, καθώς οι εταιρείες λογισμικού έχουν παραδοσιακά πολύ χαμηλό κόστος πωληθέντων (COGS) και άρα υψηλό περιθώριο, σε αντίθεση με τις βιομηχανικές εταιρείες του TX που έχουν υψηλό λειτουργικό κόστος [62].



Εικόνα 18 Σύγκριση περιθωρίου κέρδους μεταξύ CA (Τεχνολογία) και TX (Βιομηχανία/Ενέργεια).

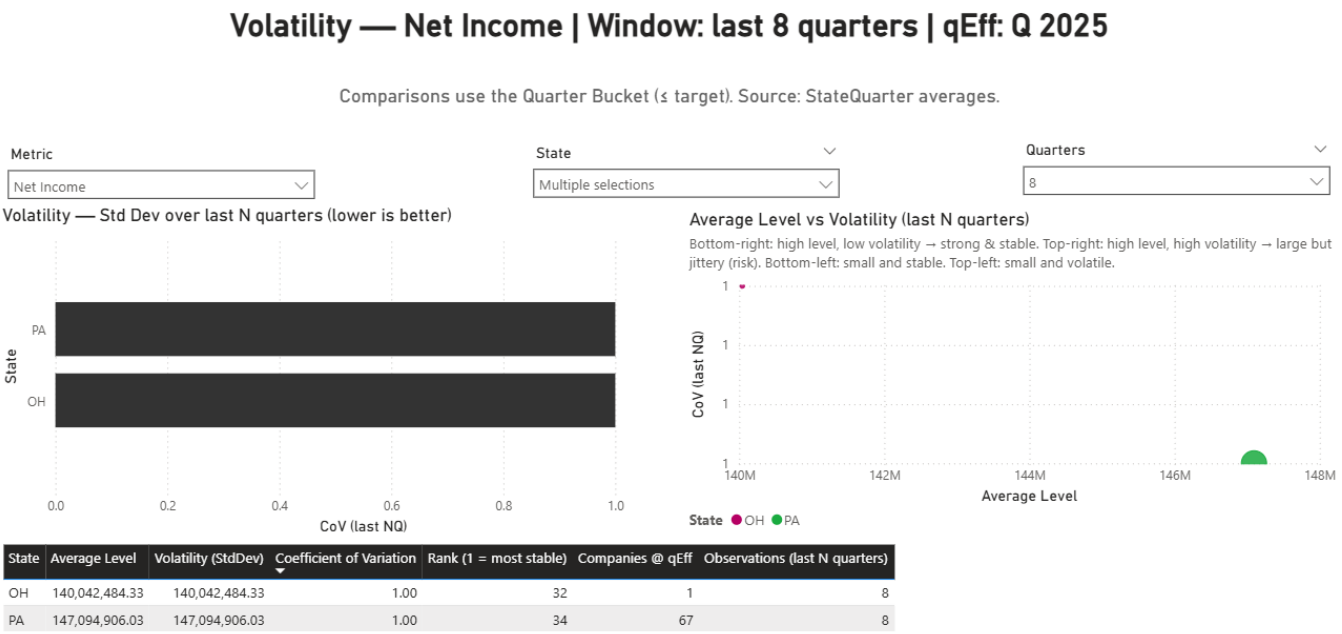
Σενάριο 2: Ανίχνευση Κινδύνου και Μεταβλητότητας (Risk Assessment)

Για έναν επενδυτή ή έναν αναλυτή ρίσκου, δεν έχει σημασία μόνο η απόδοση (π.χ. υψηλά έσοδα), αλλά και η σταθερότητα. Μια πολιτεία της οποίας οι εταιρείες εμφανίζουν έντονες διακυμάνσεις στα αποτελέσματά τους θεωρείται υψηλότερου ρίσκου [63].

Χρησιμοποιώντας τη σελίδα "Volatility", εξετάζουμε τη μετρική Net Income (Καθαρά Κέρδη) σε βάθος 8 τριμήνων.

Ανάλυση: Στο Scatter Plot της σελίδας, αναζητούμε πολιτείες που βρίσκονται στο άνω αριστερό ή άνω δεξί τεταρτημόριο (Υψηλό Volatility).

- Πολιτείες με μικρό αριθμό εταιρειών (χαμηλό coverage) τείνουν να εμφανίζουν ακραία μεταβλητότητα, καθώς μία και μόνο εταιρεία με κακό τρίμηνο μπορεί να επηρεάσει τον μέσο όρο.
- Αντιθέτως, μεγάλες οικονομίες όπως η Νέα Υόρκη (NY) εμφανίζονται συνήθως στο κάτω μέρος του γραφήματος (χαμηλό Volatility), υποδεικνύοντας μια πιο ώριμη και σταθερή αγορά.



Εικόνα 19 Ανάλυση μεταβλητότητας καθαρών κερδών. (Οχάιο vs Πενσυλβάνια)

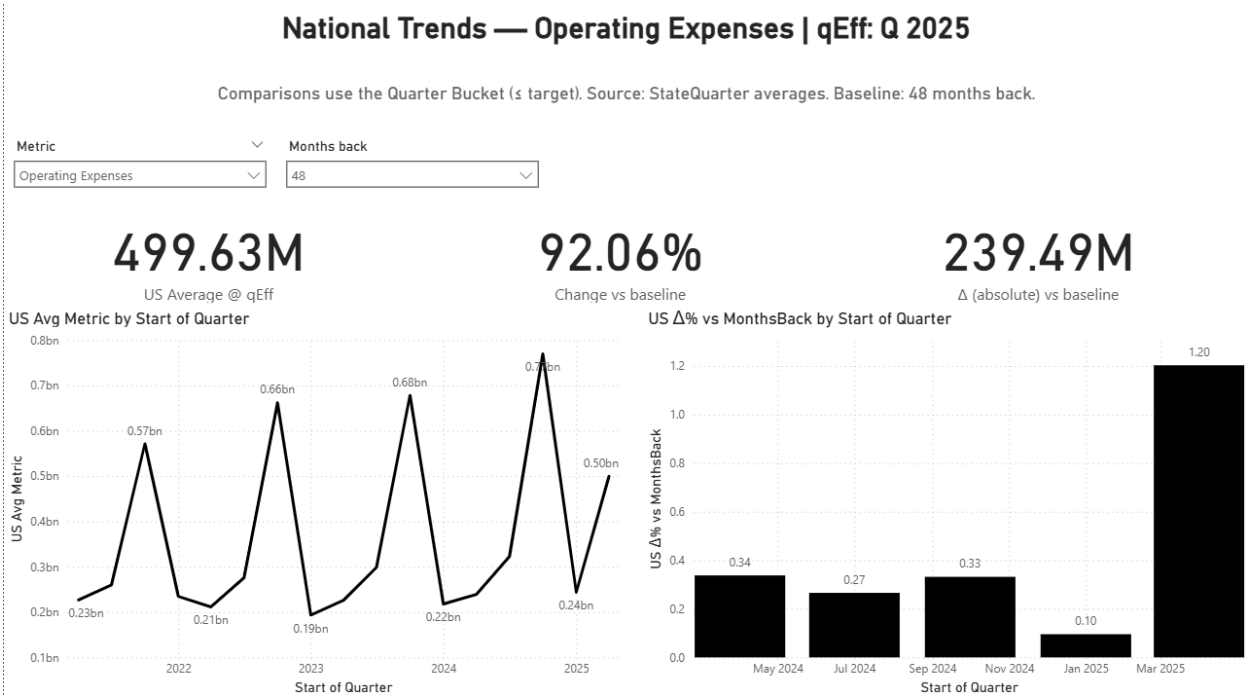
Το εργαλείο επιτρέπει στον χρήστη να απομονώσει γρήγορα τις "ασφαλείς" αγορές από τις "κερδοσκοπικές".

Σενάριο 3: Διαχρονική Εξέλιξη και η Επίδραση Εξωγενών Παραγόντων

Η σελίδα "National Trends" μας επιτρέπει να εξετάσουμε τη συνολική πορεία της αμερικανικής αγοράς. Ένα κλασικό σενάριο χρήσης είναι η μελέτη της επίδρασης μεγάλων γεγονότων (όπως η πανδημία ή ο πληθωρισμός) στους ισολογισμούς.

Εξετάζοντας το δείκτη Operating Expenses (Λειτουργικά Έξοδα) για την περίοδο 2020-2024:

- Παρατηρούμε την πτώση ή την επιβράδυνση της ανάπτυξης σε συγκεκριμένα τρίμηνα που ταυτίζονται με περιόδους ύφεσης.
- Μέσω του ραβδογράμματος της ποσοστιαίας μεταβολής (US Δ%), μπορούμε να εντοπίσουμε ακριβώς το τρίμηνο όπου ξεκίνησε η ανάκαμψη.



Εικόνα 20 Εθνική τάση λειτουργικών εξόδων. Η διαγραμματική απεικόνιση επιτρέπει την άμεση αναγνώριση των μακροοικονομικών κύκλων.

Τα σενάρια που αναλύθηκαν παραπάνω αποδεικνύουν στην πράξη πως η εφαρμογή δεν περιορίζεται στην απλή παρουσίαση οικονομικών δεικτών, αλλά λειτουργεί ως ένα ουσιαστικό εργαλείο λήψης αποφάσεων. Συνδυάζοντας την τεχνική επεξεργασία των δεδομένων (μέσω ELT και *dbt*) με ορθές πρακτικές οπτικοποίησης, το σύστημα επιτρέπει στον χρήστη να μετατρέψει γρήγορα τα ακατέργαστα δεδομένα σε χρήσιμη πληροφορία, κατανοώντας άμεσα την οικονομική εικόνα που κρύβεται πίσω από τους αριθμούς.

Εγχειρίδιο Χρήστη

Εκκίνηση της εφαρμογής

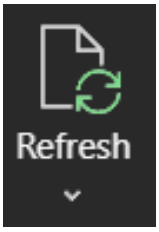
Η εφαρμογή διατίθεται ως αρχείο *Power BI* Desktop (.pbix), το οποίο ανοίγει τοπικά στον υπολογιστή του χρήστη.

Για να ξεκινήσει ο χρήστης:

1. Ανοίγει το ***Power BI* Desktop.**
2. Επιλέγει **File → Open** και φορτώνει το αρχείο της εφαρμογής US_States_QuarterBucket_EDGAR.pbix.
3. Μετά το άνοιγμα, εμφανίζεται η βασική σελίδα του report (*State Overview*).

Εφόσον τα δεδομένα στην βάση (*PostgreSQL*) έχουν ήδη ενημερωθεί από το pipeline, ο χρήστης μπορεί να εκτελέσει **Refresh**:

Από το μενού **Home → Refresh.**



Εικόνα 21 Εγχειρίδιο χρήσης

Με τον τρόπο αυτό, το *Power BI* φορτώνει τα πιο πρόσφατα διαθέσιμα δεδομένα από τη βάση και ενημερώνει όλες τις σελίδες του report.

Βασικά στοιχεία της διεπαφής

Το περιβάλλον της εφαρμογής ακολουθεί την τυπική λογική ενός *Power BI* report:

- **Σελίδες (tabs) (1*)** στο κάτω μέρος, μία για κάθε θεματική ενότητα:
  - State Overview
  - State Detail
  - State Comparison (A vs B)
  - Volatility
  - Correlation Explorer
  - Executive Summary
  - National Trends

- Financial Health – Liquidity & Leverage

➢ **Slicers (φίλτρα) (2*)** στην επάνω περιοχή κάθε σελίδας (κάτω απο τον τίτλο-υπότιτλο), για:

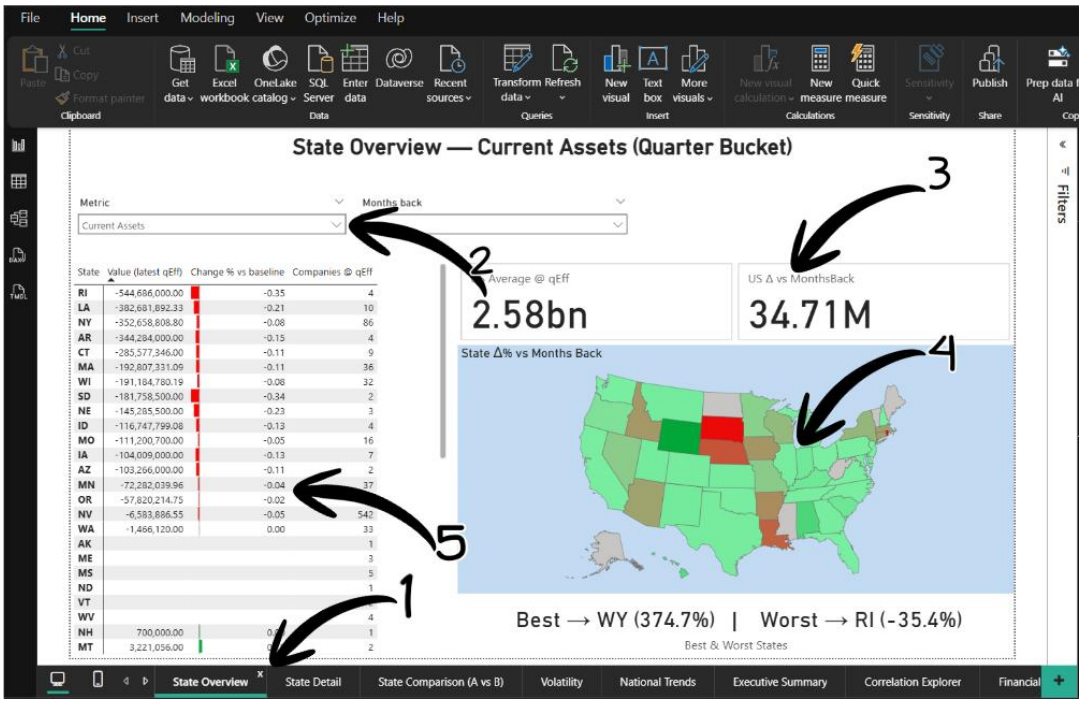
- επιλογή μετρικής (*Metric*)
- επιλογή παραθύρου σε μήνες (*MonthsBack*)
- επιλογή πολιτείας (*State*, *State A*, *State B*)
- όπου χρειάζεται, επιλογή χρονικού παραθύρου (π.χ. πόσα quarters να εμφανιστούν).

➢ **Κάρτες (KPIs) (3*)** με βασικούς αριθμούς (US Average, Δ, Δ%, Rank κ.λπ.).

➢ **Γραφήματα (4*)** (line charts, bar charts, scatter plots, shape maps) που απεικονίζουν τάσεις ή συγκρίσεις.

➢ **Πίνακες (5*)** με αναλυτικά στοιχεία ανά πολιτεία (τιμές, μεταβολές, αριθμό εταιριών κ.λπ.).

Τα labels και οι τίτλοι μέσα στο *Power BI* είναι στα **αγγλικά**.



Εικόνα 22 Εγχειρίδιο χρήσης

Κοινές αρχές χρήσης

Πριν παρουσιαστούν οι επιμέρους σελίδες, είναι χρήσιμο να τονιστούν μερικές κοινές αρχές:

➢ **Metric**
Ορίζει ποιο οικονομικό μέγεθος εξετάζεται (π.χ. Revenues, Net Income, Current

Ratio, Debt-to-Equity κ.λπ.). Με την αλλαγή του metric, ενημερώνονται όλα τα γραφήματα και οι δείκτες της σελίδας.

- **MonthsBack**
  Καθορίζει πόσους μήνες «πίσω» θα κοιτάξει η σύγκριση. Π.χ. 12 μήνες σημαίνει περίπου 4 τρίμηνα πίσω.

- **Quarter Bucket**
  Όλες οι συγκρίσεις γίνονται ανά ημερολογιακό τρίμηνο και όχι ανά ακριβή ημερομηνία. Ο χρήστης δεν χρειάζεται να κάνει κάποια ενέργεια γι' αυτό. Η λογική εφαρμόζεται αυτόματα στα measures. Αρκεί να γνωρίζει ότι "Current" και "Past" αναφέρονται σε τρίμηνα, όχι σε ημερομηνίες.

- **Drill-down / Tooltips**
  Σε πολλά γραφήματα, ο χρήστης μπορεί να:

  - περάσει τον δείκτη του ποντικιού πάνω από ένα σημείο (tooltip) για να δει αναλυτικά τις τιμές,
  - κάνει κλικ σε μια πολιτεία ή γραμμή, ώστε τα υπόλοιπα visuals της σελίδας να φιλτραριστούν ανάλογα.

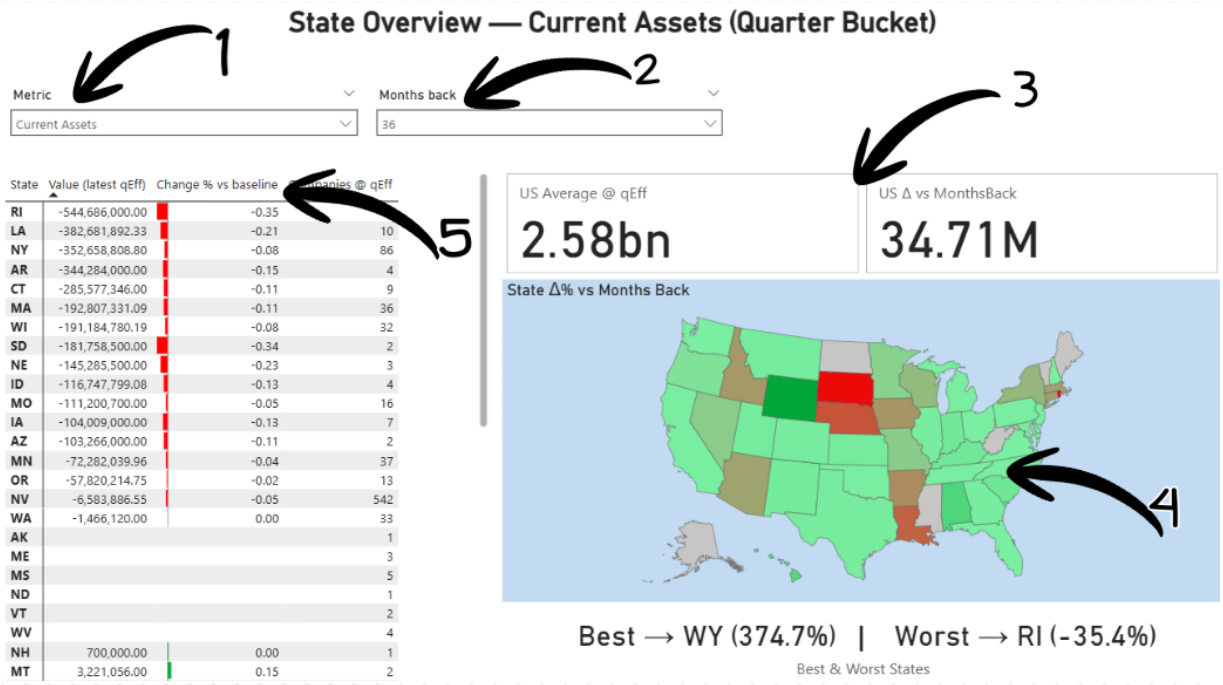
Σελίδα "State Overview"

Η σελίδα *State Overview* είναι το πρώτο σημείο εισόδου στην εφαρμογή και προσφέρει μια πανοραμική εικόνα όλων των πολιτειών για ένα συγκεκριμένο metric.

Βήματα χρήσης:

1. Επιλογή **Metric** από τον slicer (π.χ. Revenues ή Net Income).
2. Επιλογή **MonthsBack** (π.χ. 12 ή 24 μήνες) για το χρονικό διάστημα σύγκρισης.
3. Ανάγνωση των βασικών KPIs:
   - US Average (τρέχουσα μέση τιμή για όλες τις πολιτείες),
   - US Δ και US Δ% (μεταβολή σε σχέση με το Past Quarter).
4. Παρατήρηση του **χάρτη (Shape Map)**:
   - Οι πολιτείες με θετική Δ% συνήθως εμφανίζονται με πιο "θερμά" χρώματα (π.χ. πράσινο),
   - οι πολιτείες με αρνητική Δ% με "ψυχρότερα" (π.χ. κόκκινο).
5. Χρήση του **πίνακα** για να:
   - ταξινομηθούν οι πολιτείες κατά Δ%,
   - εντοπιστούν οι καλύτερες και χειρότερες επιδόσεις,
   - δει ο χρήστης πόσες εταιρίες συμμετέχουν σε κάθε πολιτεία.

Η σελίδα αυτή είναι ιδανική για ένα "γρήγορο σκανάρισμα" του χάρτη και του πίνακα, ώστε να αποφασίσει ο χρήστης σε ποιες πολιτείες ή metrics θέλει να εμβαθύνει στις επόμενες σελίδες.



State	Value (latest qEff)	Change % vs baseline	Companies @ qEff
RI	-544,686,000.00	-0.35	
LA	-382,681,892.33	-0.21	10
NY	-352,658,808.80	-0.08	86
AR	-344,284,000.00	-0.15	4
CT	-285,577,346.00	-0.11	9
MA	-192,807,331.09	-0.11	36
WI	-191,184,780.19	-0.08	32
SD	-181,758,500.00	-0.34	2
NE	-145,285,500.00	-0.23	3
ID	-116,747,799.08	-0.13	4
MO	-111,200,700.00	-0.05	16
IA	-104,009,000.00	-0.13	7
AZ	-103,266,000.00	-0.11	2
MN	-72,282,039.96	-0.04	37
OR	-57,820,214.75	-0.02	13
NV	-6,583,886.55	-0.05	542
WA	-1,466,120.00	0.00	33
AK			1
ME			3
MS			5
ND			1
VT			2
WV			4
NH	700,000.00	0.00	1
MT	3,221,056.00	0.15	2

Εικόνα 23 Εγχειρίδιο χρήσης

Σελίδα "State Detail"

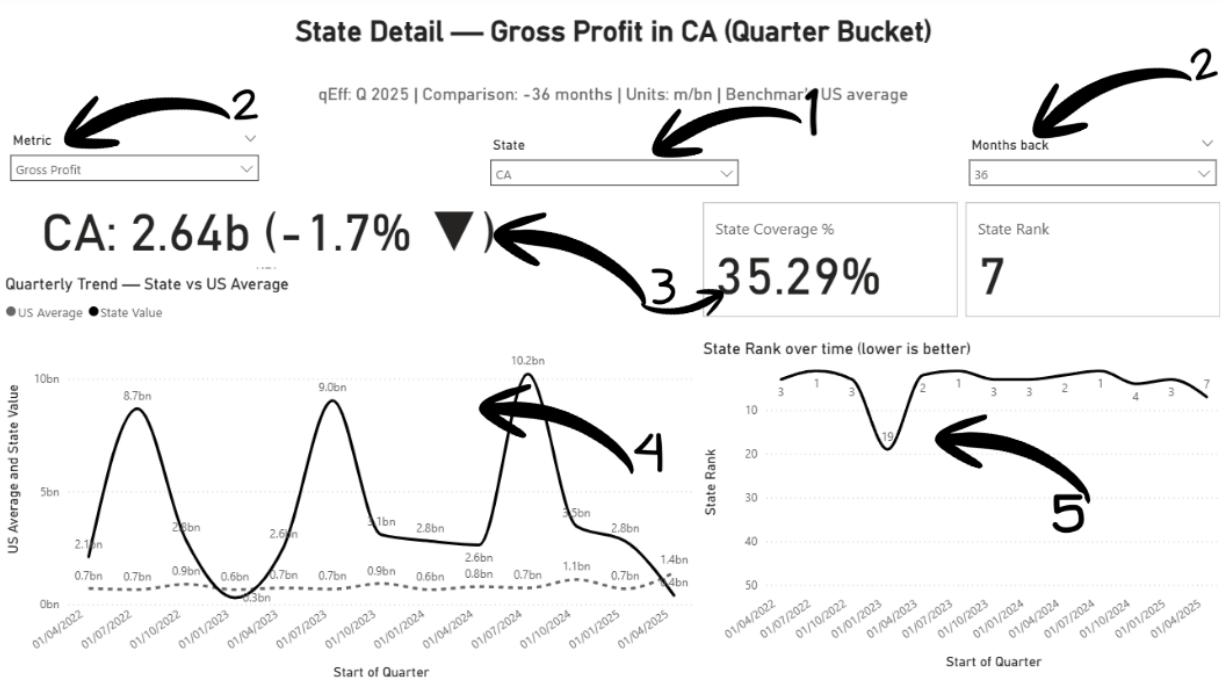
Η σελίδα State Detail επιτρέπει την αναλυτική εξέταση μιας συγκεκριμένης πολιτείας.

Βήματα χρήσης:

1. Επιλογή State από τον slicer (μία πολιτεία κάθε φορά)
2. Επιλογή Metric και MonthsBack
3. Παρατήρηση των KPIs:
   - τρέχουσα τιμή (Current value)
   - διαφορά (Δ)
   - ποσοστιαία διαφορά (Δ%)
   - τρέχουσα κατάταξη (Rank)
4. Ανάλυση του γραφήματος State vs US:
   - η γραμμή της πολιτείας δείχνει πώς εξελίσσεται το metric ανά quarter
   - η γραμμή του US Average δίνει το σημείο αναφοράς

5. Εξέταση του γραφήματος Rank over time:
    - ο άξονας δείχνει την θέση της πολιτείας ανά quarter (π.χ. 1η, 5η, 20η),
    - ο χρήστης μπορεί να δει αν η πολιτεία βελτιώνει ή χειροτερεύει τη θέση της διαχρονικά.

Η σελίδα State Detail είναι κατάλληλη όταν ο χρήστης θέλει να "προσγειωθεί" σε μία πολιτεία και να δει όχι μόνο τη σημερινή εικόνα, αλλά και την πορεία της.



Εικόνα 24 Εγχειρίδιο χρήσης

Σελίδα "State Comparison (A vs B)"

Η σελίδα State Comparison επιτρέπει τη σύγκριση δύο πολιτειών στο ίδιο metric.

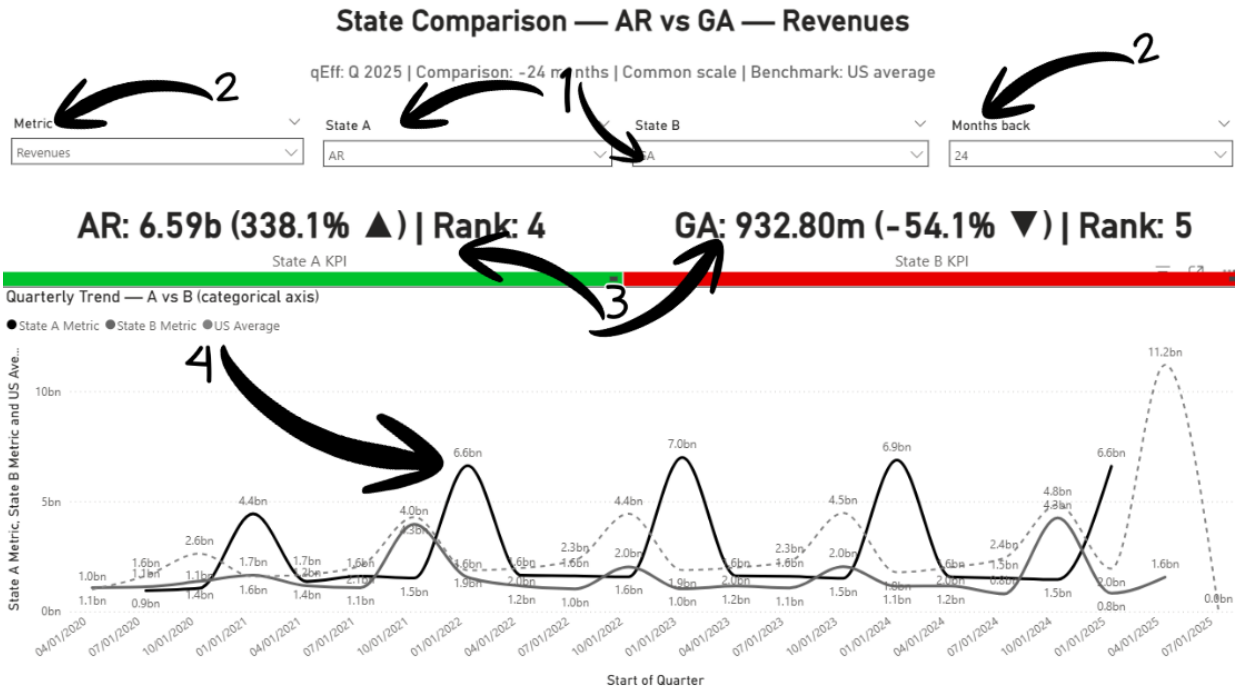
Βήματα χρήσης:

1. Επιλογή State A και State B από τους αντίστοιχους slicers.
2. Επιλογή Metric και MonthsBack.
3. Παρατήρηση των καρτών:
    - Current value, Δ και Δ% για State A
    - Current value, Δ και Δ% για State B
    - Rank κάθε πολιτείας
4. Ανάλυση των γραφημάτων τάσης:

- οι δύο πολιτείες απεικονίζονται στην ίδια κλίμακα,
- μπορεί να προστεθεί και η γραμμή US για αναφορά.

5. Ερμηνεία:
   - αν μία πολιτεία έχει σταθερά υψηλότερες τιμές,
   - αν μία πολιτεία βελτιώνεται ταχύτερα (πιο μεγάλη Δ%),
   - αν οι γραμμές συγκλίνουν ή αποκλίνουν στο χρόνο.

Η σελίδα είναι ιδιαίτερα χρήσιμη όταν ο χρήστης θέλει να συγκρίνει πολιτείες με παρόμοιο προφίλ (π.χ. δύο πολιτείες με σημαντική βιομηχανική δραστηριότητα).



Εικόνα 25 Εγχειρίδιο χρήσης

Σελίδα “Volatility – Level vs Stability”

Η σελίδα *Volatility* συνδυάζει την πληροφορία για το επίπεδο μιας μετρικής με το πόσο σταθερή ή ασταθής είναι στο χρόνο.

Βήματα χρήσης:

1. Επιλογή **Metric, State** (με πατημένο το ctrl μπορεί να επιλέξει πάνω απο μία πολιτεία) και **quarters** (π.χ. τελευταία 8 quarters).
2. Παρατήρηση του **scatter plot**:
   - ο οριζόντιος άξονας δείχνει τη μέση τιμή (average level),
   - ο κατακόρυφος άξονας δείχνει τη μεταβλητότητα (volatility),

- το μέγεθος του κύκλου συνήθως αντιστοιχεί στο πλήθος εταιριών της πολιτείας.

3. Παρατήρηση του **bar chart**:
   - Μέση τιμή της μεταβλητότητας των εταιρειών κάθε επιλεγμένης πολιτείας (καλύτερη τιμή =χαμηλότερη)
4. Χρήση των **tooltips**:
   - περνώντας το ποντίκι πάνω από μια πολιτεία, εμφανίζονται η ακριβής μέση τιμή, η τυπική απόκλιση και άλλα στοιχεία.
5. Ευρεση απο τον **πίνακα**:
   - Μεσος όρος τιμών των εταίριών για κάθε πολιτεία
   - Μεταβλητότητα των τιμών ανα πολιτεία
   - Συτελεστής μεταβλητότητας
   - Το rank κάθε επιλεγμένης πολιτείας (πιο ψηλά οι πιο σταθερές)
   - Τον αριθμό των εταιριών που προσμετρούνται στις μετρήσεις κάθε επιλεγμένης πολιτείας
   - Σε πόσα quarters απο τα επιλεγμένα υπάρχει τιμή για κάθε πολιτεία
6. Ερμηνεία:
   - πολιτείες κάτω–δεξιά έχουν υψηλό επίπεδο και χαμηλή μεταβλητότητα,
   - πολιτείες πάνω–δεξιά έχουν υψηλό επίπεδο αλλά αστάθεια,
   - πολιτείες πάνω–αριστερά συνήθως συνδυάζουν χαμηλές τιμές και υψηλή αστάθεια.

Η σελίδα βοηθά τον χρήστη να ξεχωρίσει πολιτείες που είναι όχι μόνο “καλές” σε επίπεδο metric, αλλά και σταθερές.

Σελίδα “Correlation Explorer”

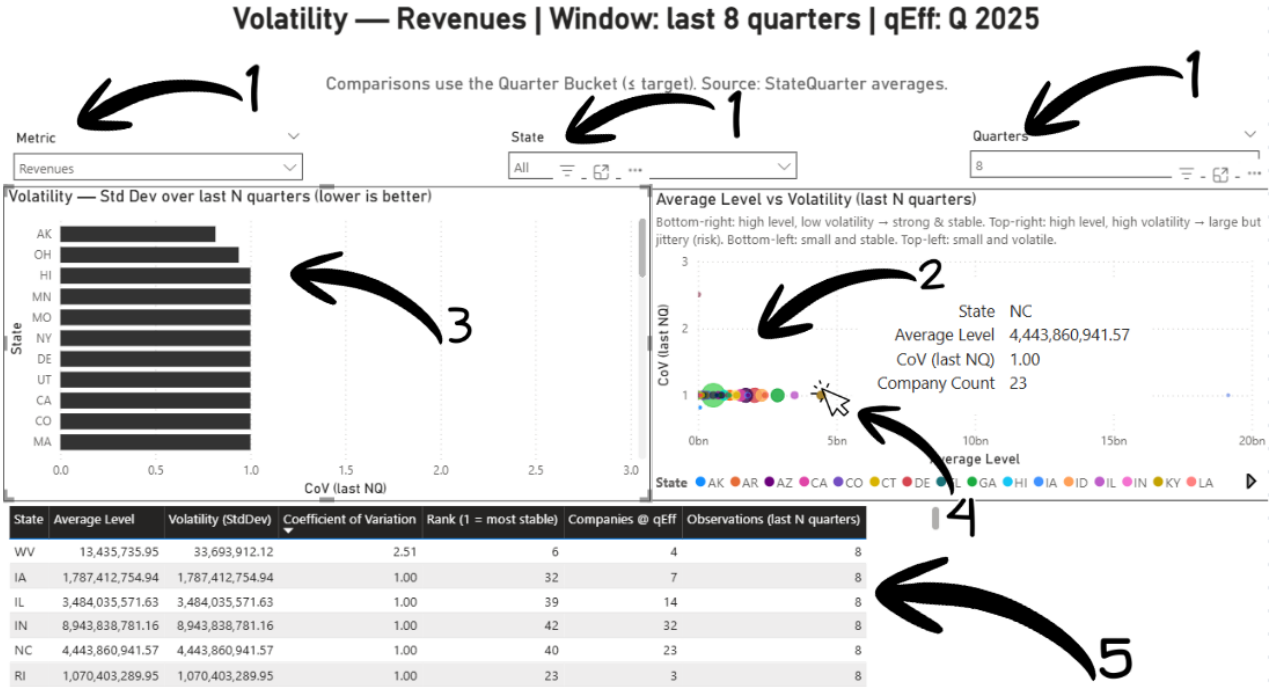
Η σελίδα *Correlation Explorer* επιτρέπει στον χρήστη να διερευνήσει αν δύο οικονομικά μεγέθη κινούνται μαζί ή ανεξάρτητα στις πολιτείες.

Βήματα χρήσης:

1. Επιλογή **Metric A** και **Metric B** από τους slicers.
2. Επιλογή χρονικού παραθύρου (π.χ. τελευταία 8 ή 12 quarters).
3. Παρατήρηση του συνολικού **Correlation %**:
   - υψηλή θετική τιμή (π.χ. 0,8) σημαίνει ότι οι δύο δείκτες συνήθως αυξάνονται/μειώνονται μαζί,
   - χαμηλή ή αρνητική τιμή υποδηλώνει ανεξάρτητη ή αντίθετη κίνηση.

4. Ανάλυση του **scatter plot**:
   - κάθε σημείο είναι μία πολιτεία με (x = Metric A, y = Metric B),
   - το σχήμα του νέφους δείχνει αν υπάρχει ή όχι μοτίβο συσχέτισης.
5. Εξέταση του γραφήματος εξέλιξης correlation ανά quarter, όπου είναι διαθέσιμο, για να φανεί αν η συσχέτιση είναι σταθερή ή αλλάζει στο χρόνο.

Η σελίδα αυτή ενδείκνυται για πιο “αναλυτικούς” χρήστες που θέλουν να δουν σχέσεις μεταξύ δεικτών, όπως π.χ. Revenues vs Net Income, ή Debt-to-Equity vs ROE.



State	Average Level	Volatility (StdDev)	Coefficient of Variation	Rank (1 = most stable)	Companies @ qEff	Observations (last N quarters)
WV	13,435,735.95	33,693,912.12	2.51	6	4	8
IA	1,787,412,754.94	1,787,412,754.94	1.00	32	7	8
IL	3,484,035,571.63	3,484,035,571.63	1.00	39	14	8
IN	8,943,838,781.16	8,943,838,781.16	1.00	42	32	8
NC	4,443,860,941.57	4,443,860,941.57	1.00	40	23	8
RI	1,070,403,289.95	1,070,403,289.95	1.00	23	3	8

Εικόνα 26 Εγχειρίδιο χρήσης

Σελίδα “Executive Summary”

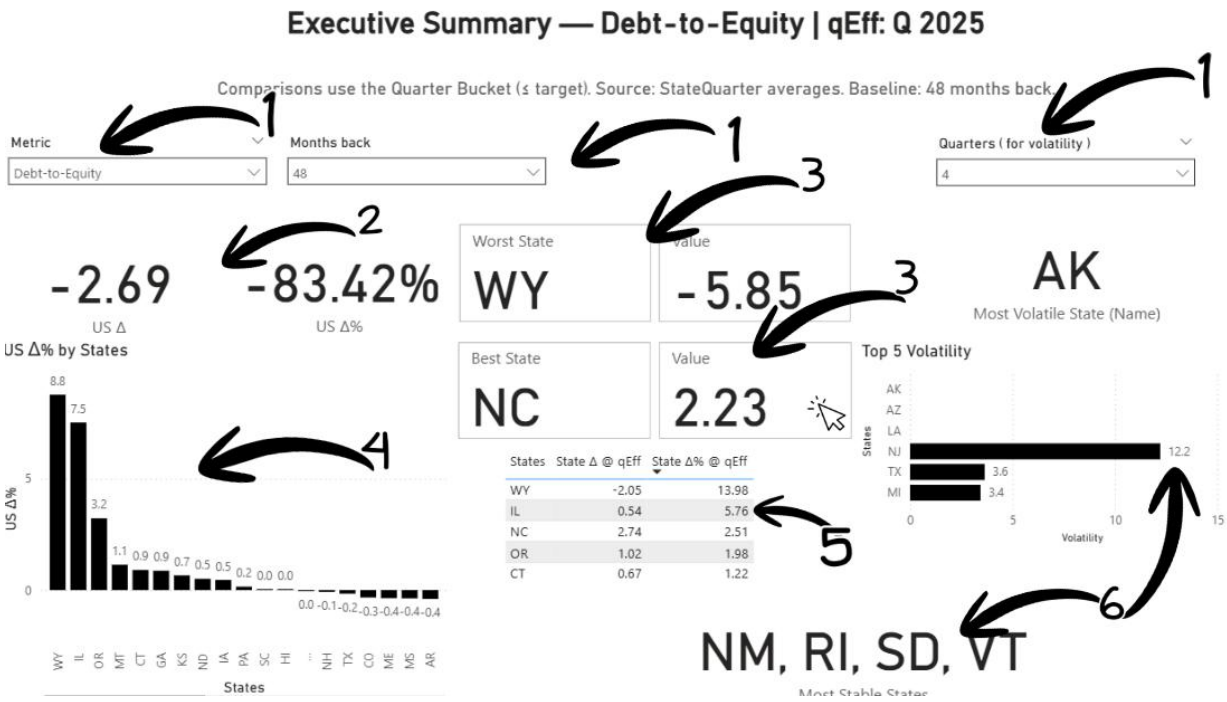
Η σελίδα *Executive Summary* απευθύνεται κυρίως σε χρήστες που θέλουν μία γρήγορη, συνοπτική εικόνα χωρίς λεπτομέρειες.

Βήματα χρήσης:

1. Επιλογή metric, MonthsBack και Quarters για το volatility.
2. Ανάγνωση των **US KPIs** (US Δ, US Δ%).
3. Εντοπισμός:
   - της πολιτείας με την καλύτερη επίδοση (Best state),
   - της πολιτείας με τη χειρότερη (Worst state).

4. Εξέταση της κατανομής US Δ% ανά πολιτεία
    - Το ραβδογράφημα *US Δ% by States* δείχνει για κάθε πολιτεία τη Δ% σε σχέση με το baseline.
    - Οι μπάρες είναι ταξινομημένες, ώστε στα αριστερά να βρίσκονται οι μεγαλύτερες θετικές μεταβολές και στα δεξιά οι πιο αρνητικές.
    - Με hover σε μία μπάρα εμφανίζονται αναλυτικά η πολιτεία και η ακριβής τιμή Δ%.
5. Χρήση του συνοπτικού πίνακα States – State Δ & State Δ% @ qEff
    - Ο πίνακας στον κεντρικό άξονα παρουσιάζει επιλεγμένες πολιτείες με:
    - *State Δ @ qEff*: απόλυτη μεταβολή,
    - *State Δ% @ qEff*: ποσοστιαία μεταβολή.
    - Ο χρήστης μπορεί να ταξινομήσει τις στήλες για να δει γρήγορα ποιες πολιτείες υπεραποδίδουν ή υστερούν.
6. Ανάλυση μεταβλητότητας (Volatility)
    - Η κάρτα *Most Volatile State (Name)* δείχνει την πολιτεία με τη μεγαλύτερη μεταβλητότητα στο επιλεγμένο metric, μέσα στο παράθυρο των N τελευταίων quarters.
    - Το γράφημα *Top 5 Volatility* παρουσιάζει τις 5 πιο ασταθείς πολιτείες, με το μήκος της μπάρας να αντιστοιχεί στο επίπεδο volatility (τυπική απόκλιση).
    - Στο κάτω μέρος της σελίδας εμφανίζεται λίστα *Most Stable States* (π.χ. “NM, RI, SD, VT”), που δείχνει τις πολιτείες με τη χαμηλότερη μεταβλητότητα.

Η σελίδα αυτή λειτουργεί σαν “αποτύπωμα” της τρέχουσας κατάστασης, δίνοντας τα βασικά σημεία χωρίς την ανάγκη λεπτομερούς πλοήγησης.



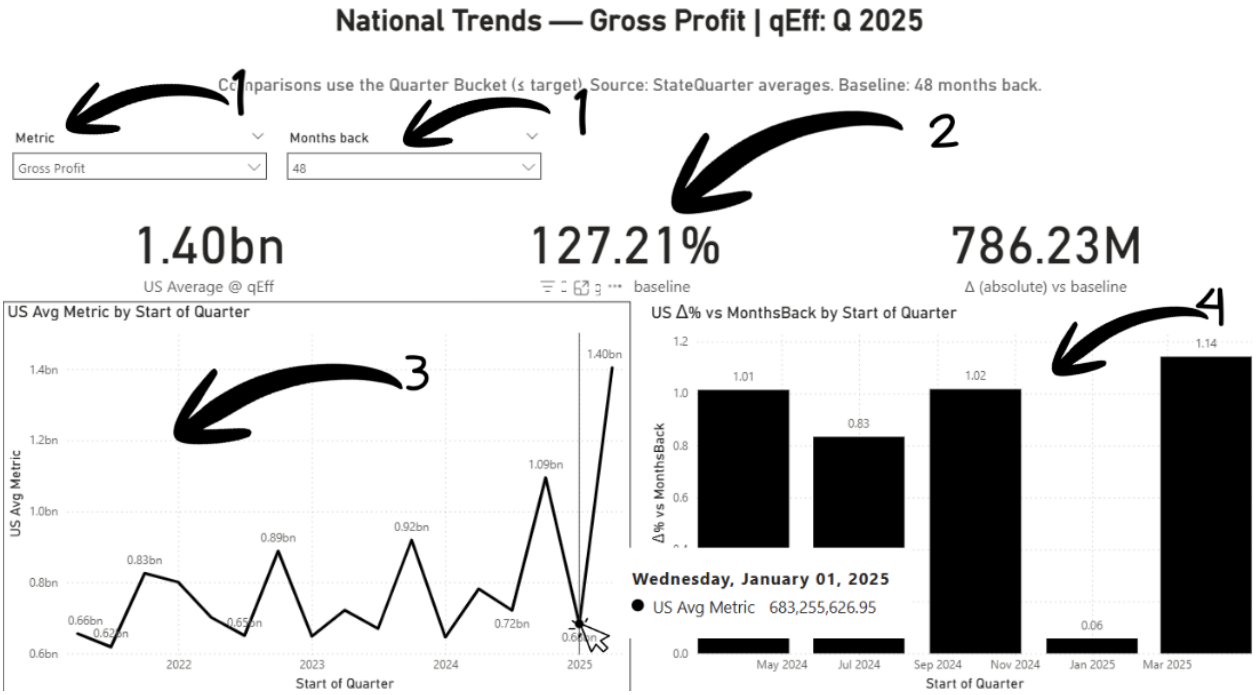
Εικόνα 27 Εγχειρίδιο χρήσης

Σελίδα “National Trends”

Η σελίδα *National Trends* εστιάζει στην χρονική πορεία της επιλεγμένης μετρικής στο σύνολο των Η.Π.Α.

Βήματα χρήσης:

1. Επιλογή metric και MonthsBack.
2. Παρατήρηση των καρτών:
   - Μέσος όρος της Αμερικής αυτή την στιγμή
   - Ποσοστιαία αλλαγή σε σχέση με την χρονική επιλογή του χρήστη
   - Η αλλαγή της τιμής σε σχέση με την χρονική επιλογή του χρήστη
3. Παρατήρηση του γραφήματος **US Average ανά quarter**:
   - αν η πορεία είναι ανοδική, καθοδική ή στάσιμη.
4. Εξέταση του **Δ% ανά quarter**:
   - το bar chart δείχνει σε ποια τρίμηνα υπήρξαν μεγαλύτερες αλλαγές.



Εικόνα 28 Εγχειρίδιο χρήσης

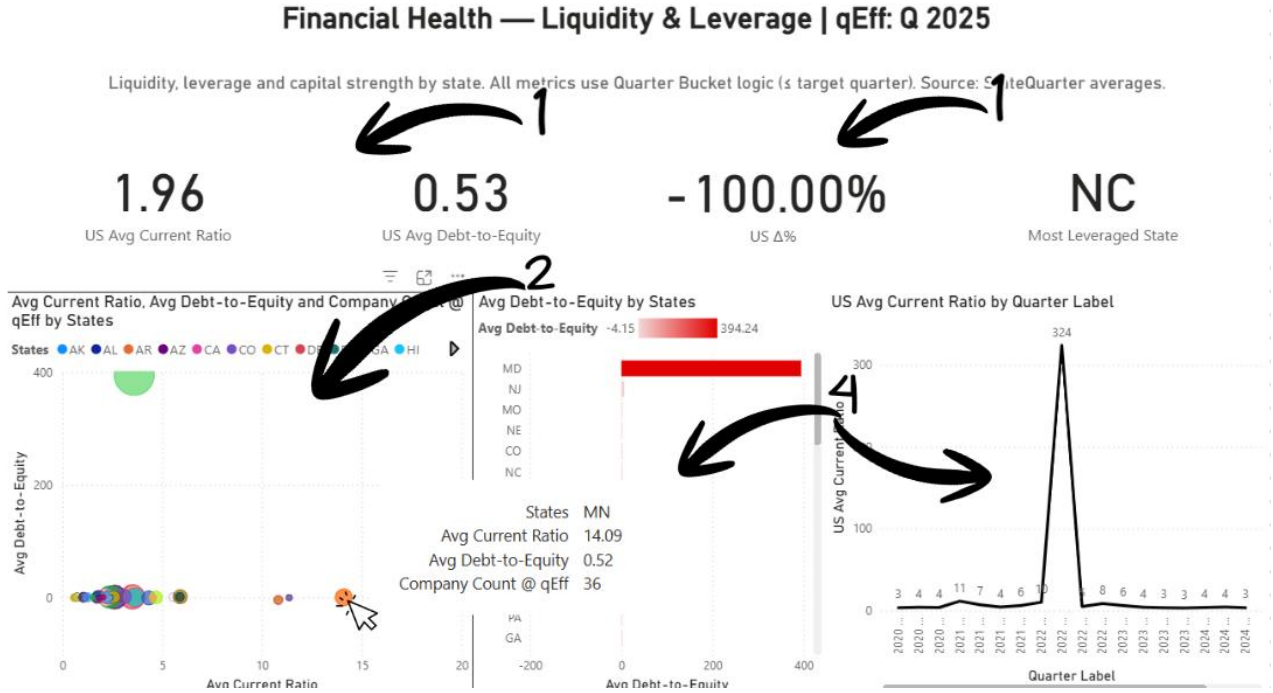
Σελίδα "Financial Health – Liquidity & Leverage"

Η σελίδα *Financial Health* συνδυάζει πληροφορίες για ρευστότητα και μόχλευση ανά πολιτεία.

Βήματα χρήσης:

1. Ανάγνωση των **US KPIs**:
   - US Avg Current Ratio,
   - US Avg Debt-to-Equity.
2. Παρατήρηση του **scatter plot**:
   - x = Current Ratio,
   - y = Debt-to-Equity,
   - μέγεθος κύκλου ≈ πλήθος εταιριών.
3. Εντοπισμός:
   - πολιτειών με χαμηλή ρευστότητα και υψηλή μόχλευση (δυνητικά πιο "ευάλωτες"),
   - πολιτειών με υψηλή ρευστότητα και ισορροπημένη μόχλευση.
4. Χρήση του **bar chart** για Debt-to-Equity ανά πολιτεία, και του **line chart** για την πορεία του US Average Current Ratio στο χρόνο.

Η σελίδα βοηθά χρήστες που ενδιαφέρονται για τη χρηματοοικονομική “υγεία” της κάθε πολιτείας, πέρα από απλές μετρικές κερδοφορίας.



Εικόνα 29 Εγχειρίδιο χρήσης

<u>Γενικές συστάσεις και περιορισμοί</u>

Τέλος, είναι χρήσιμο ο χρήστης να έχει υπόψη του τα εξής:

- Οι τιμές αφορούν μέσους όρους ανά εταιρία και πολιτεία (unweighted averages). Δεν πρόκειται για σταθμισμένους δείκτες με βάση το μέγεθος των εταιριών.
- Τα δεδομένα προέρχονται από το SEC EDGAR και είναι τόσο πλήρη όσο οι διαθέσιμες και σωστά καταχωρημένες υποβολές των εταιριών. Σε ορισμένες περιπτώσεις μπορεί να υπάρχουν κενά ή ασυνέχειες.
- Οι συγκρίσεις γίνονται σε ημερολογιακά τρίμηνα (Quarter Bucket). Αυτό παρέχει σταθερότητα και “δίκαιο” χρονικό πλαίσιο, αλλά σημαίνει ότι δεν βλέπουμε κινήσεις σε επίπεδο ημέρας ή μήνα.
- Η εφαρμογή λειτουργεί ως εργαλείο διερεύνησης (exploratory BI) και όχι ως σύστημα πρόβλεψης ή επενδυτικής σύστασης.

Με αυτά τα δεδομένα, το εγχειρίδιο χρήσης δίνει στον αναγνώστη/χρήστη τα απαραίτητα βήματα και τον τρόπο σκέψης για να αξιοποιήσει το dashboard με συνέπεια και επίγνωση των περιορισμών του.

Αξιολόγηση – Απόδοση

Η αξιολόγηση εστιάζει σε χρόνους εκτέλεσης, όγκους δεδομένων και λειτουργική σταθερότητα του pipeline. Τα μεγέθη εξαρτώνται από τη διαθεσιμότητα/ανανέωση δεδομένων στη SEC και ενδέχεται να μεταβάλλονται.

Χρόνοι εκτέλεσης (ενδεικτικοί)

Στάδιο	**Λειτουργία**	**Διάρκεια**
Airbyte	Full sync	μερικά λεπτά όταν υπάρχουν λίγες νέες εγγραφές· έως ~1 ώρα όταν έχει ανανεωθεί μεγάλο μέρος του dataset
dbt	Incremental run (staging/dim/fact, merge)	μερικά δευτερόλεπτα
Power BI	Refresh (Import) του μοντέλου	μερικά δευτερόλεπτα

Πίνακας 3 Χρόνοι εκτέλεσης

Σχόλιο: Η μετάβαση σε **hash IDs** και **incremental/merge** σταθεροποίησε τις σχέσεις και μείωσε δραστικά τον χρόνο ανακατασκευής, επιτρέποντας μικρά, ταχέα incremental builds.

Όγκοι και μεγέθη

Πίνακας/Οντότητα	**Μέγεθος**
Raw stream (_531_sec_data_stream_combined)	~550.000 γραμμές (μεταβλητό· αυξομειώνεται με νέες αναρτήσεις)
Companies (unique CIKs)	~5.400
Metrics (distinct)	βάσει λίστας base+derived του μοντέλου
Report Periods (quarters)	εξαρτάται από το χρονικό εύρος δεδομένων
Fact company_metric_values	αναλογικά με (#CIKs × #metrics × #quarters)

Πίνακας 4 Όγκοι και μεγέθη

Περιορισμοί και Μελλοντική Εργασία

Περιορισμοί της προσέγγισης

Παρά τα θετικά αποτελέσματα και τη σταθερότητα του pipeline, η υλοποίηση παρουσιάζει ορισμένους περιορισμούς. Πρώτον, η λογική Quarter Bucket βασίζεται σε ημερολογιακά τρίμηνα. Ωστόσο, αρκετές εταιρείες χρησιμοποιούν φορολογικά έτη που δεν ταυτίζονται πλήρως με τα calendar quarters. Αυτό σημαίνει ότι, ειδικά στα τρίμηνα μετάβασης χρήσης, οι αξίες που συγκρίνονται ενδέχεται να αντιστοιχούν σε ελαφρώς διαφορετικές λογιστικές περιόδους. Η επιλογή αυτή υιοθετήθηκε συνειδητά, καθώς τα ημερολογιακά τρίμηνα αποτελούν κοινό παρονομαστή για όλες τις εταιρείες, αλλά παραμένει ένας παράγοντας που πρέπει να λαμβάνεται υπόψη στην ερμηνεία των αποτελεσμάτων.

Δεύτερον, οι μέσες τιμές σε επίπεδο πολιτείας υπολογίζονται ως μη σταθμισμένοι μέσοι όροι (unweighted averages). Με άλλα λόγια, μια πολύ μεγάλη εταιρεία και μία μικρή συμμετέχουν με το ίδιο βάρος στον μέσο όρο μιας πολιτείας. Αυτό είναι συνεπές με την ιδέα της “τυπικής εταιρείας πολιτείας”, αλλά δεν αποτυπώνει την πραγματική συνεισφορά κάθε εταιρείας στο συνολικό οικονομικό μέγεθος της πολιτείας. Σε περιπτώσεις όπου ο χρήστης ενδιαφέρεται για macro εικόνα, ένας σταθμισμένος μέσος (π.χ. με βάρος τα Revenues) θα ήταν πιθανώς καταλληλότερος.

Επιπλέον, η κάλυψη των δεδομένων δεν είναι ομοιόμορφη σε όλες τις πολιτείες και σε όλα τα metrics. Παρότι υπολογίζεται και παρουσιάζεται το Coverage %, υπάρχουν περιπτώσεις όπου το πλήθος εταιρειών με διαθέσιμη τιμή σε ένα τρίμηνο είναι περιορισμένο. Σε τέτοιες περιπτώσεις, οι τιμές ενός δείκτη για συγκεκριμένη πολιτεία μπορεί να είναι ευαίσθητες σε outliers και η ερμηνεία τους πρέπει να γίνεται με προσοχή. Τέλος, η λύση επικεντρώνεται αποκλειστικά σε δημόσιες εταιρείες που αναφέρονται στο SEC EDGAR, και συνεπώς δεν καλύπτει μη εισηγμένες επιχειρήσεις ή άλλους κλάδους της οικονομίας.

Τέλος, είναι σημαντικό να τονιστεί ότι το σύστημα έχει σχεδιαστεί ως εργαλείο εξερευνητικής ανάλυσης (exploratory BI) και όχι ως πλατφόρμα πρόβλεψης ή παραγωγής επενδυτικών συστάσεων. Τα dashboards βοηθούν στον εντοπισμό τάσεων, αποκλίσεων και συσχετίσεων, αλλά δεν ενσωματώνουν μοντέλα πρόβλεψης κινδύνου ή αξιολόγησης αξίας εταιρειών. Οποιαδήποτε χρήση των αποτελεσμάτων πέρα από την περιγραφική ανάλυση θα πρέπει να γίνεται με ιδιαίτερη προσοχή και επιπλέον μεθοδολογική τεκμηρίωση.

Προτάσεις για μελλοντική εργασία

Οι παραπάνω περιορισμοί ανοίγουν τον δρόμο για σειρά από επεκτάσεις και βελτιώσεις σε μελλοντική εργασία. Μία πρώτη κατεύθυνση είναι η υποστήριξη σταθμισμένων δεικτών σε επίπεδο πολιτείας, όπου οι μέσες τιμές θα υπολογίζονται με βάρη ανάλογα με το μέγεθος των εταιρειών (π.χ. βάρος ίσο με τα Revenues ή την κεφαλαιοποίηση). Αυτό θα επέτρεπε την ύπαρξη δύο συμπληρωματικών οπτικών: της “τυπικής εταιρείας” και της “συνολικής οικονομικής ισχύος” μιας πολιτείας.

Μία δεύτερη κατεύθυνση αφορά την καλύτερη μοντελοποίηση του χρόνου. Θα μπορούσε να ενσωματωθεί υποστήριξη για fiscal calendars, ώστε οι εταιρείες που έχουν μη τυποποιημένες χρήσεις να χαρτογραφούνται σε “οικονομικά τρίμηνα” αντί για καθαρά

ημερολογιακά. Παράλληλα, θα μπορούσε να επεκταθεί η λογική Quarter Bucket σε άλλους χρονικούς ορίζοντες, όπως rolling 12-month windows ή half-year buckets, ανάλογα με τις ανάγκες του αναλυτή.

Τέλος, μπορεί να γίνει επέκταση του PowerBi dashboard με πρόσθετες σελίδες για πλουσιότερη πληροφόρηση του χρήστη. Για να επιτευχθεί αυτό το σενάριο θα είναι χρήσιμο και παραγωγικό η άντληση περισσότερων μερτήσεων απο το SEC EDGAR. Αυτό θα έχει σαν αποτέλεσμα και την δυνατότητα κατασκευής περιισότερων παράγωγων μετρικών.

Επέκταση με Generative AI και Κανονιστικές Προκλήσεις

Μια φυσική εξέλιξη του παρόντος συστήματος BI, πέραν της βελτιστοποίησης των διαδικασιών ETL, είναι η ενσωμάτωση τεχνολογιών Παραγωγικής Τεχνητής Νοημοσύνης (Generative AI), όπως τα Μεγάλα Γλωσσικά Μοντέλα (LLMs). Σύμφωνα με την βιβλιγραφία, οι τεχνολογίες αυτές μετασχηματίζουν ριζικά τη λήψη αποφάσεων, προσφέροντας νέες δυνατότητες στην ανάλυση και σύνθεση πληροφορίας [64].

Στόχος μιας τέτοιας επέκτασης θα ήταν η αυτόματη ερμηνεία των οπτικοποιημένων δεδομένων και η παροχή επεξηγήσεων σε φυσική γλώσσα προς τη διοίκηση. Ωστόσο, η υιοθέτηση τέτοιων τεχνολογιών συνοδεύεται από σοβαρές προκλήσεις. Η χρήση Generative AI εγείρει ηθικά διλήμματα και απαιτεί συμμόρφωση με αυστηρά κανονιστικά πλαίσια, ειδικά σε αυτόνομα συστήματα λογισμικού [65]. Υπάρχει ο κίνδυνος τα μοντέλα να οδηγήσουν σε ανακριβή συμπεράσματα (hallucinations), εάν δεν υπάρχει αυστηρός έλεγχος.

Για την αντιμετώπιση αυτών των κινδύνων, η διεθνής κοινότητα Μηχανικής Λογισμικού (Software Engineering) στρέφεται προς την ανάπτυξη προτύπων για "AI-Empowered Software Engineering" (AIESE). Όπως αναλύεται στα πρακτικά και τις εισαγωγές του συνεδρίου AIESE 2024, απαιτούνται νέες μεθοδολογίες που ενσωματώνουν την ηθική και την ασφάλεια στον πυρήνα της ανάπτυξης λογισμικού [66], [67].

Αυτό είναι ιδιαίτερα κρίσιμο σε Κυβερνοφυσικά Συστήματα (Cyber-Physical Systems) και έξυπνες εφαρμογές, όπου οι αποφάσεις του λογισμικού έχουν άμεσο αντίκτυπο στον πραγματικό κόσμο [68]. Η μελλοντική έρευνα, λοιπόν, πρέπει να εστιάσει στο πώς το pipeline που υλοποιήθηκε (Airbyte, dbt) μπορεί να τροφοδοτεί μοντέλα AI με τρόπο που να διασφαλίζει την ακεραιότητα, ακολουθώντας τις σύγχρονες εξελίξεις στα ευφυή πληροφοριακά συστήματα [69].

Συμπεράσματα

Μέσα από την υλοποίηση της εργασίας, φάνηκε ότι η μεγαλύτερη πρόκληση δεν ήταν η οπτικοποίηση, αλλά η διαχείριση της ασυμμετρίας των δεδομένων στο στάδιο του ELT. Παρόλα αυτά, αποδείχθηκε ότι είναι εφικτή η ανάπτυξη ενός πλήρως αυτοματοποιημένου, αναπαραγώγιμου και επεκτάσιμου συστήματος επιχειρηματικής ευφυΐας (End-to-End ELT→BI) με την αποκλειστική χρήση ανοικτών πηγών δεδομένων και λογισμικού ανοιχτού κώδικα. Η αξιοποίηση του δημόσιου αποθετηρίου SEC EDGAR και του προτύπου XBRL επέτρεψε την τυποποιημένη συλλογή ενός τεράστιου όγκου εταιρικών οικονομικών στοιχείων, αποδεικνύοντας την αξία των Open Data στη σύγχρονη ανάλυση.

Σε τεχνικό επίπεδο, η αρχιτεκτονική που υλοποιήθηκε μέσω του συνδυασμού *Airbyte* και *dbt* απέδειξε την στιβαρότητά της. Η στρατηγική υιοθέτηση των hash IDs για την παραγωγή κλειδιών, σε συνδυασμό με τη λογική incremental/merge, εξασφάλισε τη σταθερότητα του μοντέλου δεδομένων και τη διατήρηση της αναφορικής ακεραιότητας. Παράλληλα, βελτιστοποίησε δραστικά την απόδοση του συστήματος, επιτρέποντας την τακτική ανανέωση του Data Warehouse χωρίς την ανάγκη χρονοβόρων διαδικασιών πλήρους ανακατασκευής (full rebuilds).

Σε μεθοδολογικό επίπεδο, καθοριστική υπήρξε η εισαγωγή και εφαρμογή της μεθοδολογίας «Quarter Bucket». Η προσέγγιση αυτή έλυσε αποτελεσματικά το χρόνιο πρόβλημα της χρονικής ασυμμετρίας και των διαφορετικών οικονομικών χρήσεων (fiscal calendars) που χαρακτηρίζουν τις εταιρικές υποβολές. Μέσω της ενοποίησης των αναφορών σε ημερολογιακά τρίμηνα και της επιλογής της τελευταίας διαθέσιμης τιμής (Last-in-Quarter), κατέστησαν δυνατές οι δίκαιες και αξιόπιστες συγκρίσεις μεταξύ πολιτειών, προσφέροντας ανθεκτικότητα απέναντι σε κενά δεδομένων ή ετεροχρονισμένες δημοσιεύσεις.

Το τελικό αποτέλεσμα αποτυπώνεται στο περιβάλλον του *Power BI*, όπου οι σελίδες του dashboard προσφέρουν πολλαπλές και συμπληρωματικές οπτικές—από τη λεπτομερή ανάλυση ανά πολιτεία έως τη μελέτη της μεταβλητότητας και των συσχετίσεων. Συνολικά, η εργασία ανέδειξε ότι η τεχνολογική στοίβα *Airbyte → PostgreSQL → dbt → Power BI* αποτελεί μια βιώσιμη και παραγωγική λύση για τη διαχείριση μεγάλων συνόλων δεδομένων. Το σύστημα που παραδόθηκε είναι λειτουργικό και θέτει ισχυρές βάσεις για μελλοντική επέκταση, όπως η ενσωμάτωση σταθμισμένων (weighted) αναλύσεων, η αυστηρότερη τυποποίηση κατά IBCS και η μετάβαση σε υποδομές cloud.

Πίνακας συντμήσεων & ακρωνυμίων

Ακρωνύμιο	Αγγλικός Όρος	Ελληνική Απόδοση / Επεξήγηση
API	Application Programming Interface	Διεπαφή Προγραμματισμού Εφαρμογών
BI	Business Intelligence	Επιχειρηματική Ευφυΐα
CIK	Central Index Key	Μοναδικός κωδικός αναγνώρισης εταιρείας (από τη SEC)
CRON	(Time-based Job Scheduler)	Χρονοπρογραμματιστής εργασιών σε συστήματα Unix
DAG	Directed Acyclic Graph	Κατευθυνόμενος Ακυκλικός Γράφος (ροή δεδομένων)
DAX	Data Analysis Expressions	Γλώσσα εκφράσεων ανάλυσης δεδομένων (Power BI)
dbt	data build tool	Εργαλείο μετασχηματισμού δεδομένων
DDL	Data Definition Language	Γλώσσα Ορισμού Δεδομένων (SQL εντολές δομής)
EDGAR	Electronic Data Gathering, Analysis, and Retrieval	Σύστημα ηλεκτρονικής συλλογής δεδομένων της SEC
ELT	Extract, Load, Transform	Εξαγωγή, Φόρτωση, Μετασχηματισμός
ETL	Extract, Transform, Load	Εξαγωγή, Μετασχηματισμός, Φόρτωση
GAAP	Generally Accepted Accounting Principles	Γενικά Αποδεκτές Λογιστικές Αρχές
HTTP	HyperText Transfer Protocol	Πρωτόκολλο Μεταφοράς Υπερκειμένου
IBCS	International Business Communication Standards	Διεθνή Πρότυπα Επιχειρηματικής Επικοινωνίας
JSON	JavaScript Object Notation	Μορφή ανταλλαγής δεδομένων κειμένου
KPI	Key Performance Indicator	Βασικός Δείκτης Απόδοσης
REST	Representational State Transfer	Αρχιτεκτονικό στυλ για διαδικτυακές υπηρεσίες
SEC	Securities and Exchange Commission	Επιτροπή Κεφαλαιαγοράς των Η.Π.Α.
SQL	Structured Query Language	Γλώσσα Δομημένων Ερωτημάτων
UML	Unified Modeling Language	Ενοποιημένη Γλώσσα Μοντελοποίησης
XBRL	eXtensible Business Reporting Language	Επεκτάσιμη Γλώσσα Επιχειρηματικών Αναφορών

Πίνακας ορολογίας

Όρος	Ορισμός
Connector	Λογισμικό που επιτρέπει τη σύνδεση και τη μεταφορά δεδομένων μεταξύ ενός συστήματος πηγής και ενός συστήματος προορισμού.
Data Lineage	Η καταγραφή της προέλευσης των δεδομένων και της διαδρομής που ακολουθούν μέσα στο σύστημα, συμπεριλαμβανομένων των μετασχηματισμών τους.
Data Warehouse	Κεντρικό αποθετήριο δεδομένων που συγκεντρώνει πληροφορίες από διαφορετικές πηγές, βελτιστοποιημένο για ανάλυση και αναφορές.
Fact Table	Ο κεντρικός πίνακας σε ένα σχήμα αστέρα (Star Schema) που περιέχει τις ποσοτικές μετρήσεις (metrics) και τα κλειδιά σύνδεσης.
Dimension Table	Πίνακας που περιέχει περιγραφικά χαρακτηριστικά (π.χ. όνομα εταιρείας, ημερομηνία) και συνδέεται με τον Fact Table.
Incremental Load	Η διαδικασία φόρτωσης μόνο των νέων ή τροποποιημένων εγγραφών, αντί για την πλήρη επαναφόρτωση όλων των δεδομένων.
Orchestration	Ο συντονισμός και η διαχείριση της εκτέλεσης των διαφόρων βημάτων μιας ροής δεδομένων (pipeline).
Quarter Bucket	Μεθοδολογία ομαδοποίησης εγγραφών με βάση το ημερολογιακό τρίμηνο, για την ευθυγράμμιση ασύγχρονων χρονικά δεδομένων.
Schema Drift	Το φαινόμενο όπου η δομή των δεδομένων στην πηγή αλλάζει απροειδοποίητα (π.χ. προσθήκη νέων πεδίων), προκαλώντας πιθανά σφάλματα στη ροή.
Staging Area	Ενδιάμεση περιοχή αποθήκευσης όπου τα δεδομένα καθαρίζονται και προετοιμάζονται πριν φορτωθούν στο τελικό Data Warehouse.
Star Schema	Μοντέλο σχεδίασης βάσης δεδομένων όπου ένας κεντρικός πίνακας γεγονότων συνδέεται με πολλούς πίνακες διαστάσεων, σχηματίζοντας ένα αστέρι.
Surrogate Key	Τεχνητό κλειδί (συνήθως ακέραιος ή hash) που χρησιμοποιείται ως μοναδικό αναγνωριστικό σε έναν πίνακα, αντί για το φυσικό κλειδί των δεδομένων.
Webhook	Μηχανισμός που επιτρέπει σε μια εφαρμογή να στέλνει αυτοματοποιημένα μηνύματα/δεδομένα σε μια άλλη εφαρμογή όταν συμβεί ένα συγκεκριμένο γεγονός.

Παραρτήματα

Παράρτημα Α: Οδηγός Αναπαραγωγής (*Airbyte → PostgreSQL → dbt → Power BI*)

Προαπαιτούμενα περιβάλλοντος: Windows 10, WSL/Ubuntu 24.04, Docker Desktop 4.41.2, Python 3.12.8, *PostgreSQL* 16.4, *Airbyte* 0.63.13, *dbt* 0.38.28, *Power BI* Desktop 2.147.1085.0. Hardware: CPU 2.30 GHz, RAM 16 GB, SSD 500 GB.

Βήμα 1 — *PostgreSQL* (Raw):

1) Δημιούργησε κενή βάση/σχήμα για το raw stream.
2) Κατέγραψε στοιχεία σύνδεσης (host/port/DB/user/password) για χρήση στο *Airbyte* & *Power BI*

Βήμα 2 — *Airbyte* (Ingestion):

1) Εκκίνηση *Airbyte* (Docker Desktop).
2) **Source:** Custom connector (Python SDK) για SEC EDGAR, με ορισμένα τα headers (User-Agent σύμφωνα με SEC) και backoff/retry.
3) **Destination:** *PostgreSQL* (τα στοιχεία του Βήματος 1).
4) **Connection:** Προγραμμάτισε **daily CRON**.
5) **Webhook on success:** Δήλωσε URL (Flask endpoint) ώστε σε «succeeded» sync να καλείται το *dbt* (επόμενο βήμα).

Βήμα 3 — *dbt* (Staging/Dim/Fact):

1) Ρύθμισε το προφίλ σύνδεσης προς τη βάση (ίδια διαπιστευτήρια).
2) Επιβεβαίωσε materializations: `staging` ως view, `dimensions` και `fact` ως **incremental/merge** με `unique_key`.
3) Βεβαιώσου ότι τα **IDs είναι hash-based** (md5) σε Companies/Metrics/ReportPeriods.
4) Δοκίμασε χειροκίνητα ένα `dbt run` για πλήρη build, και στη συνέχεια έλεγξε ότι το webhook εκτελεί incremental run μετά από sync.

Βήμα 4 — *Power BI* (Model & Reports):

1) **Σύνδεση:** Import προς τους πίνακες companies, metrics, report_periods, company_metric_values (και τυχόν υλικό για StateQuarter αν χρησιμοποιείται).
2) **Σχέσεις:** One-to-many από κάθε διάσταση προς το fact.
3) **Slicers:** Metric, MonthsBack, State.
4) **Μέθοδοι:** Εφάρμοσε **Quarter Bucket** (AsOf = τελευταίο διαθέσιμο quarter, MonthsBack = {0,3,6,12,24,36,48}), **Dense rank**, **US Average (unweighted)**, **neutral zone ±0.01%**, trends με fallback.
5) **Μονάδες/Labels:** όπως στις αναρτήσεις (m/bn κ.λπ.).

Βήμα 5 — Έλεγχοι/QA:

1) Συμφωνία πρώτων εγγραφών μεταξύ raw και staging.
2) Έλεγχος μοναδικοτήτων στα hash IDs.
3) Έλεγχος μερικών joins στο fact (company_id/metric_id/report_period_id).
4) Coverage KPIs σε 1–2 states και οπτικός έλεγχος State vs US ανά quarter.
5) Επιβεβαίωση refresh χρόνων (Import) και σταθερότητας σε βαριές σελίδες.

Παράρτημα Β: Δείγματα Δεδομένων και Σχήμα Βάσης

Στο παρόν παράρτημα παρατίθενται ενδεικτικά δείγματα των πρωτογενών δεδομένων (raw data) όπως αυτά αντλούνται από το API της SEC, καθώς και ο κώδικας δημιουργίας (DDL) του βασικού πίνακα αποθήκευσης στην *PostgreSQL*. Σκοπός είναι η τεκμηρίωση της δομής της πληροφορίας πριν αυτή υποστεί τους μετασχηματισμούς του *dbt*.

Δείγμα Raw JSON από SEC EDGAR API

Τα παρακάτω τμήματα κώδικα (snippet) αποτελούν αντιπροσωπευτικά δείγματα της απόκρισης που λαμβάνει ο connector από τα endpoint submissions και companyfacts της SEC αντίστοιχα. Πρόκειται για δεδομένα τύπου XBRL που έχουν μετατραπεί σε JSON format από την υπηρεσία της SEC.

https://data.sec.gov/submissions/CIK0000320193.json


```
{
  "cik": "0000320193",
  "entityType": "operating",
  "sic": "3571",
  "sicDescription": "Electronic Computers",
  "ownerOrg": "06 Technology",
  "insiderTransactionForOwnerExists": 0,
  "insiderTransactionForIssuerExists": 1,
  "name": "Apple Inc.",
  "tickers": [
    "AAPL"
  ],
  "exchanges": [
    "Nasdaq"
  ],
  "ein": "942404110",
  "lei": null,
  "description": "",
  "website": "",
  "investorWebsite": "",
  "category": "Large accelerated filer",
  "fiscalYearEnd": "0927",
  "stateOfIncorporation": "CA",
  "stateOfIncorporationDescription": "CA",
  "addresses": {
    "mailing": {
      "street1": "ONE APPLE PARK WAY",
      "street2": null,
      "city": "CUPERTINO",
      "stateOrCountry": "CA",
      "zipCode": "95014",
      "stateOrCountryDescription": "CA",
      "isForeignLocation": 0,
      "foreignStateTerritory": null,
      "country": null,
      "countryCode": null
    },
    "business": {
```

```
      "street1": "ONE APPLE PARK WAY",
      "street2": null,
      "city": "CUPERTINO",
      "stateOrCountry": "CA",
      "zipCode": "95014",
      "stateOrCountryDescription": "CA",
      "isForeignLocation": null,
      "foreignStateTerritory": null,
      "country": null,
      "countryCode": null
    }
  },
  "phone": "(408) 996-1010",
  "flags": "",
  "formerNames": [
    {
      "name": "APPLE INC",
      "from": "2007-01-10T00:00:00.000Z",
      "to": "2019-08-05T00:00:00.000Z"
    },
    {
      "name": "APPLE COMPUTER INC",
      "from": "1994-01-26T00:00:00.000Z",
      "to": "2007-01-04T00:00:00.000Z"
    },
    {
      "name": "APPLE COMPUTER INC/ FA",
      "from": "1997-07-28T00:00:00.000Z",
      "to": "1997-07-28T00:00:00.000Z"
    }
  ],
  "filings": {
    "recent": {
      "accessionNumber": [
        "0001462356-25-000012",
        "0001354457-25-001138",
        "0001631982-25-000011",
        "0000320193-25-000079",
        "0000320193-25-000077",
        "0002050912-25-000008",
        "0001631982-25-000009",
        "0001950047-25-008030",
        "0001214156-25-000011",
        "0001767094-25-000009".... }
```

https://data.sec.gov/api/xbrl/companyconcept/CIK0000320193/us-gaap/AccountsPayableCurrent.json


```
{
  "cik": 320193,
  "taxonomy": "us-gaap",
  "tag": "AccountsPayableCurrent",
  "label": "Accounts Payable, Current",
  "description": "Carrying value as of the balance sheet date of
liabilities incurred (and for which invoices have typically been received)
and payable to vendors for goods and services received that are used in an
entity's business. Used to reflect the current portion of the liabilities
(due within one year or within the normal operating cycle if longer).",
  "entityName": "Apple Inc.",
  "units": {
    "USD": [
      {
        "end": "2008-09-27",
        "val": 5520000000,
        "accn": "0001193125-09-214859",
        "fy": 2009,
        "fp": "FY",
        "form": "10-K",
        "filed": "2009-10-27"
      },
      {
        "end": "2008-09-27",
        "val": 5520000000,
        "accn": "0001193125-10-012091",
        "fy": 2009,
        "fp": "FY",
        "form": "10-K/A",
        "filed": "2010-01-25",
        "frame": "CY2008Q3I"
      }...}
```

Σχήμα Βάσης Δεδομένων (SQL DDL)

Ακολουθεί ο κώδικας SQL για τη δημιουργία του πίνακα _531_sec_data_stream_combined στην *PostgreSQL*. Αυτός ο πίνακας λειτουργεί ως ο υποδοχέας (raw landing table) για τα δεδομένα που στέλνει το *Airbyte*.

Σημαντικό είναι ότι, εκτός από τα επιχειρησιακά πεδία (cik, metric, value), το *Airbyte* προσθέτει αυτόματα και τρία τεχνικά πεδία (_airbyte_ab_id, _airbyte_emitted_at, _airbyte_normalized_at).

public._531_sec_data_stream_combined

```
-- Table: public._531_sec_data_stream_combined

-- DROP TABLE IF EXISTS public._531_sec_data_stream_combined;

CREATE TABLE IF NOT EXISTS public._531_sec_data_stream_combined
(
    cik character varying COLLATE pg_catalog."default",
    name character varying COLLATE pg_catalog."default",
    value numeric(38,9),
    metric character varying COLLATE pg_catalog."default",
    period_start date,
    report_end_date date,
    state_of_incorporation character varying COLLATE pg_catalog."default",
    _airbyte_raw_id character varying(36) COLLATE pg_catalog."default" NOT NULL,
    _airbyte_extracted_at timestamp with time zone NOT NULL,
    _airbyte_generation_id bigint,
    _airbyte_meta jsonb NOT NULL
)

TABLESPACE pg_default;

ALTER TABLE IF EXISTS public._531_sec_data_stream_combined
    OWNER to airbyte_user;
-- Index: _531_sec_data_stream_combined__airbyte_extracted_at_idx

-- DROP INDEX IF EXISTS
public._531_sec_data_stream_combined__airbyte_extracted_at_idx;

CREATE INDEX IF NOT EXISTS
_531_sec_data_stream_combined__airbyte_extracted_at_idx
    ON public._531_sec_data_stream_combined USING btree
    (_airbyte_extracted_at ASC NULLS LAST)
    TABLESPACE pg_default;
-- Index: _531_sec_data_stream_combined__airbyte_raw_id_idx

-- DROP INDEX IF EXISTS public._531_sec_data_stream_combined__airbyte_raw_id_idx;

CREATE INDEX IF NOT EXISTS _531_sec_data_stream_combined__airbyte_raw_id_idx
    ON public._531_sec_data_stream_combined USING btree
    (_airbyte_raw_id COLLATE pg_catalog."default" ASC NULLS LAST)
    TABLESPACE pg_default;
-- Index: _531_sec_data_stream_combined_cik_metric_report_end_date_re_idx

-- DROP INDEX IF EXISTS
public._531_sec_data_stream_combined_cik_metric_report_end_date_re_idx;

CREATE INDEX IF NOT EXISTS
_531_sec_data_stream_combined_cik_metric_report_end_date_re_idx
    ON public._531_sec_data_stream_combined USING btree
    (cik COLLATE pg_catalog."default" ASC NULLS LAST, metric COLLATE
```

```
pg_catalog."default" ASC NULLS LAST, report_end_date ASC NULLS LAST,
report_end_date ASC NULLS LAST, _airbyte_extracted_at ASC NULLS LAST)
    TABLESPACE pg_default;
```

Βιβλιογραφία

[1] D. Seenivasan, "ETL vs ELT: Choosing the right approach for your data warehouse," vol. 7, pp. 110–122, Feb. 2022, doi: 10.6084/m9.doione.IJRTI2202018.

[2] M. J. Bommarito, D. M. Katz, and E. M. Detterman, "OpenEDGAR: Open Source Software for SEC EDGAR Analysis," Jun. 13, 2018, *arXiv*: arXiv:1806.04973. doi: 10.48550/arXiv.1806.04973.

[3] R. Debreceny, S. Farewell, M. Piechocki, C. Felden, and A. Gräning, "Does it add up? Early evidence on the data quality of XBRL filings to the SEC," *J. Account. Public Policy*, vol. 29, no. 3, pp. 296–306, Jun. 2010, doi: 10.1016/j.jaccpubpol.2010.04.001.

[4] M. Bommarito, D. Katz, and E. Detterman, *OpenEDGAR: Open Source Software for SEC EDGAR Analysis*. 2018. doi: 10.48550/arXiv.1806.04973.

[5] J. Gerdes, "EDGAR-Analyzer: automating the analysis of corporate data contained in the SEC's EDGAR database," *Decis. Support Syst.*, vol. 35, no. 1, pp. 7–29, Apr. 2003, doi: 10.1016/S0167-9236(02)00096-9.

[6] F. W. Li and C. Sun, "Information acquisition and expected returns: Evidence from EDGAR search traffic," *J. Econ. Dyn. Control*, vol. 141, p. 104384, Aug. 2022, doi: 10.1016/j.jedc.2022.104384.

[7] R. K. Pagidi, R. Kshir-sagar, P. Kankanampati, E. Shrivastav, P. Goel, and O. Goel, "Leveraging Data Engineering Techniques for Enhanced Business Intelligence," *Univers. Res. Rep.*, vol. 9, pp. 561–581, Oct. 2022, doi: 10.36676/urr.v9.i4.1392.

[8] S. Chaudhuri, U. Dayal, and V. Narasayya, "An overview of business intelligence technology," *Commun. ACM*, vol. 54, no. 8, pp. 88–98, Aug. 2011, doi: 10.1145/1978542.1978562.

[9] M. C. Solano and J. C. Cruz, "Integrating Analytics in Enterprise Systems: A Systematic Literature Review of Impacts and Innovations," *Adm. Sci.*, vol. 14, no. 7, p. 138, Jul. 2024, doi: 10.3390/admsci14070138.

[10] D. Seenivasan, "ETL vs ELT: Choosing the right approach for your data warehouse," vol. 7, pp. 110–122, Feb. 2022, doi: 10.6084/m9.doione.IJRTI2202018.

[11] P. Vassiliadis, "A Survey of Extract-Transform-Load Technology.," *Int. J. Data Warehous. Min.*, vol. 5, pp. 1–27, Jul. 2009.

[12] R. Wrembel and C. Koncilia, *Data Warehouses and OLAP: Concepts, Architectures and Solutions*. IGI Global Scientific Publishing, 1 AD. Accessed: Jan. 11, 2026. [Online]. Available: https://www.igi-global.com/book/data-warehouses-olap/www.igi-global.com/book/data-warehouses-olap/235

[13] I. Eti-mfon, "Data Integration Strategies," Medium. Accessed: Jan. 11, 2026. [Online]. Available: https://medium.com/@etimfonime/data-integration-strategies-75770c100b17

[14] "ETL vs ELT - Difference Between Data-Processing Approaches - AWS," Amazon Web Services, Inc. Accessed: Jan. 02, 2026. [Online]. Available: https://aws.amazon.com/compare/the-difference-between-etl-and-elt/

[15] S. Sadiq and M. Indulska, "Open data: Quality over quantity," *Int. J. Inf. Manag.*, vol. 37, no. 3, pp. 150–154, Jun. 2017, doi: 10.1016/j.ijinfomgt.2017.01.003.

[16] R. Debreceny, S. Farewell, M. Piechocki, C. Felden, and A. Gräning, "Does it add up? Early evidence on the data quality of XBRL filings to the SEC," *J. Account. Public Policy*, vol. 29, pp. 296–306, Jun. 2010, doi: 10.1016/j.jaccpubpol.2010.04.001.

[17] J. Boritz and W. No, "The SEC's XBRL voluntary filing program on EDGAR: A case for quality assurance," *Curr. Issues Audit.*, vol. 2, pp. A36–A50, Dec. 2008, doi: 10.2308/ciia.2008.2.2.A36.

[18] R. Debreceny, C. Felden, B. Ochocki, and M. Piechocki, *XBRL for interactive data: Engineering the information value chain*. 2009, p. 214. doi: 10.1007/978-3-642-01437-6.

[19] S. O'Riain, E. Curry, and A. Harth, "XBRL and open data for global financial ecosystems: A linked data approach," *Int. J. Account. Inf. Syst.*, vol. 13, no. 2, pp. 141–162, Jun. 2012, doi: 10.1016/j.accinf.2012.02.002.

[20] D. Vuppu and M. Achanta, "Optimizing Cost and Performance in Cloud Data Lakes," *Int. J. Comput. Trends Technol.*, vol. 73, pp. 82–88, Jun. 2025, doi: 10.14445/22312803/IJCTT-V73I6P110.

[21] M. Maghsoudi and N. Nezafati, "Navigating the acceptance of implementing business intelligence in organizations: A system dynamics approach," *Telemat. Inform. Rep.*, vol. 11, p. 100070, Sep. 2023, doi: 10.1016/j.teler.2023.100070.

[22] A. Al-Sulaiti, M. Mansour, H. Al-Yafei, S. Aseel, M. Kucukvar, and N. Onat, *Using Data Analytics and Visualization Dashboard for Engineering, Procurement, and Construction Project's Performance Assessment*. 2021, p. 211. doi: 10.1109/ICIEA52957.2021.9436728.
[23] S. Few, "Information Dashboard Design : The Effective Visual Communication of Data / S. Few.," Jan. 2006.
[24] C. T. Gonçalves, M. J. A. Gonçalves, and M. I. Campante, "Developing Integrated Performance Dashboards Visualisations Using Power BI as a Platform," *Information*, vol. 14, no. 11, p. 614, Nov. 2023, doi: 10.3390/info14110614.
[25] R. Matheus, M. Janssen, and D. Maheshwari, "Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities," *Gov. Inf. Q.*, vol. 37, no. 3, p. 101284, Jul. 2020, doi: 10.1016/j.giq.2018.01.006.
[26] J. Faisst, R. Hichert, X. Subirats, and A. Assessors, "How to implement IBCS®? Concepts, Templates, Notation Manual and 11 Real-Life Examples of Application," *Rev. Comptab. Dir.*, vol. 31, pp. 77–131, Oct. 2020, [Online]. Available: https://accid.org/wp-content/uploads/2021/12/HOWTOI1.pdf
[27] "IBCS Standards 1.2 • IBCS - International Business Communication Standards," IBCS - International Business Communication Standards. Accessed: Jan. 11, 2026. [Online]. Available: https://www.ibcs.com/ibcs-standards-1-2/
[28] "PostgreSQL: Release Notes." Accessed: Jan. 11, 2026. [Online]. Available: https://www.postgresql.org/docs/release/16.11/
[29] "Airbyte Docs." Accessed: Jan. 11, 2026. [Online]. Available: https://docs.airbyte.com/
[30] Z. ALI, "AI-Ready Data Infrastructure: A Review of Zero-ETL, Declarative Pipelines, and Data Contracts in Modern Data Engineering," *Cogniz. J. Multidiscip. Stud.*, vol. 5, pp. 486–507, Aug. 2025, doi: 10.47760/cognizance.2025.v05i08.019.
[31] "dbt Developer Hub." Accessed: Jan. 11, 2026. [Online]. Available: https://docs.getdbt.com/
[32] JulCsc, "Power BI documentation - Power BI." Accessed: Jan. 11, 2026. [Online]. Available: https://learn.microsoft.com/en-us/power-bi/
[33] A. Ahmi and M. H. Mohd Nasir, "Examining the Trend of the Research on eXtensible Business Reporting Language (XBRL): A Bibliometric Review," vol. 5, pp. 1145–1167, Aug. 2019.
[34] S. Farewell, L. Hao, V. Kashyap, and R. Pinsker, "A Field Study Examining the Indian Ministry of Corporate Affairs XBRL Implementation," *J. Inf. Syst.*, vol. 31, Jan. 2016, doi: 10.2308/isys-51389.
[35] M. Arslan and C. Cruz, "Semantic Enrichment of Taxonomy for BI Applications using Multifaceted data sources through NLP techniques," *Procedia Comput. Sci.*, vol. 207, pp. 2424–2433, Jan. 2022, doi: 10.1016/j.procs.2022.09.533.
[36] S. Gill and P. Prasad, "The First Step to Business Intelligence: Ensuring Data Quality Through Rigorous ETL Processes," *Int. J. Trend Res. Dev. IJTRD*, vol. 7, no. 4, Aug. 2020, [Online]. Available: http://www.ijtrd.com/papers/IJTRD28919.pdf
[37] K. Jain, "Data Lineage in Modern Data Engineering," Feb. 2024.
[38] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi, and S. B. Yahia, "Data quality in ETL process: A preliminary study," *Procedia Comput. Sci.*, vol. 159, pp. 676–687, Jan. 2019, doi: 10.1016/j.procs.2019.09.223.
[39] W. Elouataoui and Y. Gahi, "Empirical Evaluation of Big Data Stacks: Performance and Design Analysis of Hadoop, Modern, and Cloud Architectures," *Big Data Cogn. Comput.*, vol. 10, no. 1, p. 7, Jan. 2026, doi: 10.3390/bdcc10010007.
[40] P. Adekola, "Handling Schema Evolution and Data Drift in Real-Time Analytics," Sep. 2025.
[41] R. Vagnino and C. Walker, "Schema drift: Relational concepts and conceptual change," *Cognition*, vol. 271, p. 106418, Jan. 2026, doi: 10.1016/j.cognition.2025.106418.
[42] H. Leo and A. James, "A Review of Open-Source ETL (Extract, Transform, Load) Tools for Data Integration," Sep. 2024.
[43] R. Ball and L. Shivakumar, "Earnings quality in UK private firms: comparative loss recognition timeliness," *J. Account. Econ.*, vol. 39, no. 1, pp. 83–128, Feb. 2005, doi: 10.1016/j.jacceco.2004.04.001.
[44] G. A. Tsihrintzis *et al.*, "Evaluating Stakeholder Decision-Making Trust and Efficiency in AI-Empowered Energy Software: A VIRTSI Model Approach Within an Agile Process," in *Artificial Intelligence-Empowered Software Engineering 2024*, M. Virvou, Y. Tanabe, and L. C. Jain, Eds., Cham: Springer

Nature Switzerland, 2025, pp. 380–399. doi: 10.1007/978-3-031-98410-5_24.
[45] M. Virvou, “Balancing Autonomy and Ethics in AI-Empowered Software Engineering by Addressing User-Centred Requirements Tension: Keynote,” in *Artificial Intelligence-Empowered Software Engineering 2024*, M. Virvou, Y. Tanabe, and L. C. Jain, Eds., Cham: Springer Nature Switzerland, 2025, pp. 99–126. doi: 10.1007/978-3-031-98410-5_6.
[46] M. Virvou, “Artificial Intelligence and User Experience in reciprocity: Contributions and state of the art,” *Intell. Decis. Technol.*, vol. 17, no. 1, pp. 73–125, Feb. 2023, doi: 10.3233/IDT-230092.
[47] D. P. Panagoulias, M. Virvou, and G. A. Tsihrintzis, “A microservices-based iterative development approach for usable, reliable and explainable A.I.-infused medical applications using R.U.P,” in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, Oct. 2022, pp. 1028–1035. doi: 10.1109/ICTAI56018.2022.00157.
[48] H. Foidl, V. Golendukhina, R. Ramler, and M. Felderer, “Data pipeline quality: Influencing factors, root causes of data-related issues, and processing problem areas for developers,” *J. Syst. Softw.*, vol. 207, p. 111855, Jan. 2024, doi: 10.1016/j.jss.2023.111855.
[49] A. Ribeiro, A. Silva, and A. R. da SIlva, “Data Modeling and Data Analytics: A Survey from a Big Data Perspective,” *J. Softw. Eng. Appl.*, vol. 8, no. 12, Dec. 2015, [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=1644581
[50] R. Lumbantoruan, E. Sibarani, M. Sitorus, A. Mindari, and S. Sinaga, “An Approach for Automatically Generate Star Schema from Natural Language,” *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 12, p. 501, Jun. 2014, doi: 10.12928/telkomnika.v12i2.63.
[51] R. K. Pagidi, R. Kshir-sagar, P. Kankanampati, E. Shrivastav, P. Goel, and O. Goel, “Leveraging Data Engineering Techniques for Enhanced Business Intelligence,” *Univers. Res. Rep.*, vol. 9, pp. 561–581, Oct. 2022, doi: 10.36676/urr.v9.i4.1392.
[52] N. around us, “Express to Impress: Leveraging IBCS Standards for Powerful Data Presentations,” Numbers around us. Accessed: Jan. 11, 2026. [Online]. Available: https://medium.com/number-around-us/express-to-impress-leveraging-ibcs-standards-for-powerful-data-presentations-3c3a269f0ec0
[53] G. Bergström *et al.*, “Evaluating the layout quality of UML class diagrams using machine learning,” *J. Syst. Softw.*, vol. 192, p. 111413, Oct. 2022, doi: 10.1016/j.jss.2022.111413.
[54] J. Tavares, Y. Costa, and T. Colanzi, *Classification of UML Diagrams to Support Software Engineering Education*. 2021, p. 107. doi: 10.1109/ASEW52652.2021.00030.
[55] R. Karampure, C. Y. Wang, and Y. Vashi, “UML sequence diagram to axiomatic design matrix conversion: a method for concept improvement for software in integrated systems,” *Procedia CIRP*, vol. 100, pp. 457–462, Jan. 2021, doi: 10.1016/j.procir.2021.05.104.
[56] R. Yasmina, A. Chaoui, and E. Kerkouche, “A Framework for Modeling and Analysis UML Activity Diagram using Graph Transformation,” *Procedia Comput. Sci.*, vol. 56, pp. 612–617, Dec. 2015, doi: 10.1016/j.procs.2015.07.261.
[57] M. Mornie *et al.*, “Visualisation of User Stories to UML use Case Diagram,” *J. Adv. Res. Appl. Sci. Eng. Technol.*, vol. 63, pp. 68–80, Jan. 2025, doi: 10.37934/araset.63.3.6880.
[58] B. Owen and C. Wang, “Optimization of ETL/ELT Pipelines in High-Volume Data Platforms,” Jun. 2024.
[59] W. Villegas, X. Palacios, and S. Luján-Mora, “A Business Intelligence Framework for Analyzing Educational Data,” *Sustainability*, vol. 12, p. 5745, Jul. 2020, doi: 10.3390/su12145745.
[60] “A complete guide to surrogate keys and why they matter,” dbt Labs. Accessed: Jan. 11, 2026. [Online]. Available: https://www.getdbt.com/blog/guide-to-surrogate-key
[61] K. Okoye, “IMPLEMENT INCREMENTAL DATA LOAD ETL WITH SQL (FOR LARGE DATASETS AND REAL WORLD SCENARIOS),” Medium. Accessed: Jan. 11, 2026. [Online]. Available: https://medium.com/@exceldispensing/data-integration-made-easy-with-this-simple-sql-function-e5b37edcde45
[62] M. Porter, “The Economic Performance of Regions,” *Reg. Stud.*, vol. 37, no. 6–7, pp. 549–578, Aug. 2003, doi: 10.1080/0034340032000108688.
[63] I. D. Dichev and V. W. Tang, “Earnings volatility and earnings predictability”.
[64] G. Phillips-Wren and M. Virvou, “Issues and trends in generative AI technologies for decision making,”

Intell. Decis. Technol., vol. 19, no. 2, pp. 574–584, Mar. 2025, doi: 10.1177/18724981251320551.
[65] G. A. Tsihrintzis *et al.*, "Artificial Intelligence-Empowered Autonomous Software – Moral Dilemmas, Ethics, Regulations, Challenges and Requirements: Interdisciplinary Panel Discussion," in *Artificial Intelligence-Empowered Software Engineering 2024*, M. Virvou, Y. Tanabe, and L. C. Jain, Eds., Cham: Springer Nature Switzerland, 2025, pp. 445–464. doi: 10.1007/978-3-031-98410-5_28.
[66] M. Virvou, G. A. Tsihrintzis, and Y. Tanabe, "Introduction to the 15th International Conference on AI-Empowered Software Engineering - AIESE 2024: Contributions, Emerging Themes, and Future Vision," in *Artificial Intelligence-Empowered Software Engineering 2024*, M. Virvou, Y. Tanabe, and L. C. Jain, Eds., Cham: Springer Nature Switzerland, 2025, pp. 58–70. doi: 10.1007/978-3-031-98410-5_2.
[67] M. Virvou, G. A. Tsihrintzis, and Y. Tanabe, "Introduction to the 15th International Conference on AI-Empowered Software Engineering - AIESE 2024: Contributions, Emerging Themes, and Future Vision," in *Artificial Intelligence-Empowered Software Engineering 2024*, M. Virvou, Y. Tanabe, and L. C. Jain, Eds., Cham: Springer Nature Switzerland, 2025, pp. 58–70. doi: 10.1007/978-3-031-98410-5_2.
[68] M. Virvou, G. A. Tsihrintzis, N. G. Bourbakis, and L. C. Jain, Eds., *Handbook on Artificial Intelligence-Empowered Applied Software Engineering: VOL.2: Smart Software Applications in Cyber-Physical Systems*, vol. 3. in Artificial Intelligence-Enhanced Software and Systems Engineering, vol. 3. Cham: Springer International Publishing, 2022. doi: 10.1007/978-3-031-07650-3.
[69] G. A. Tsihrintzis, M. Virvou, N. G. Bourbakis, and L. C. Jain, "Introduction to Advances in Information, Intelligence, Systems and Applications," in *Extended Selected Papers of the 14th International Conference on Information, Intelligence, Systems, and Applications*, N. Bourbakis, G. A. Tsihrintzis, M. Virvou, and L. C. Jain, Eds., Cham: Springer Nature Switzerland, 2024, pp. 1–9. doi: 10.1007/978-3-031-67426-6_1.