

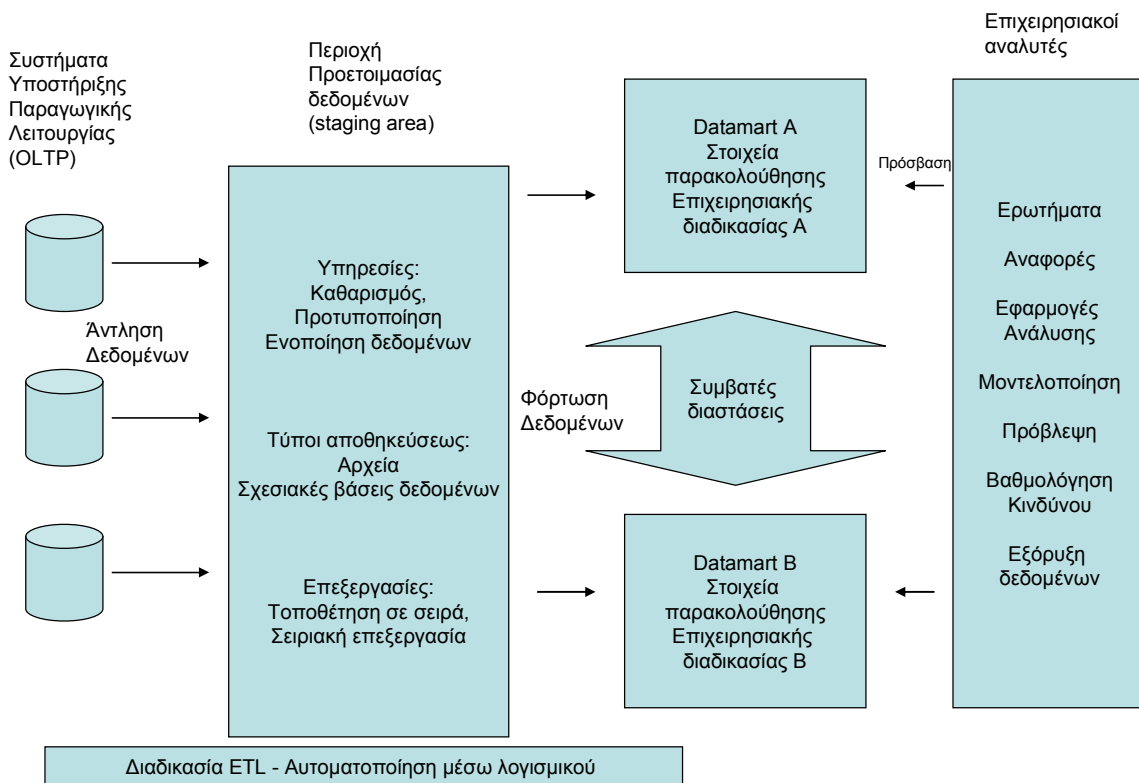
Αρχιτεκτονικές αποθηκών δεδομένων (data warehouse)

Ο σχεδιασμός συστημάτων υποστήριξης λήψης αποφάσεων, βασίζεται στην υλοποίηση υποδομών ‘αποθήκης πληροφορίας’ (γνωστά σαν Data Warehouse- DW).

Η μέθοδος υλοποίησης DW δεν έχει προτυποποιηθεί. Έχουν όμως επικρατήσει 2 βασικές προσεγγίσεις υλοποίησης που περιγράφονται παρακάτω.

Προσέγγιση Kimball

Η εικόνα 4 παρουσιάζει την αρχιτεκτονική υποδομής DW σύμφωνα με την προσέγγιση Kimball.



Εικόνα 1 – Προσέγγιση Kimball

Εντοπίζονται οι πηγές δεδομένων από τις οποίες θα αντλούνται συστηματικά δεδομένα για τον εμπλουτισμό της υποδομής DW. Αναλύεται η τεκμηρίωση των δομών δεδομένων ώστε βάσει αυτής να σχεδιαστεί ο τρόπος άντλησης των δεδομένων. Η ποιότητα της τεκμηρίωσης των δομών δεδομένων των πηγών αυτών επηρεάζει σημαντικά τον βαθμό δυσκολίας στον σχεδιασμό της άντλησης στοιχείων.

Τα δεδομένα που αντλούνται φορτώνονται στην ‘περιοχή προετοιμασίας’ (staging area), είτε σε μορφή απλών αρχείων δεδομένων, είτε σε βάση δεδομένων.

Η διαδικασία άντλησης-καθαρισμού-μετατροπής-φόρτωσης δεδομένων, γνωστή και σαν διαδικασία ETL (extraction-transformation-loading), εκτελείται στην ‘περιοχή προετοιμασίας’.

Η διαδικασία άντλησης απαιτεί τον καθορισμό των πινάκων-πεδίων από τα οποία θα λαμβάνονται δεδομένα (όπως προαναφέρθηκε αναλύεται η τεκμηρίωση των δομών δεδομένων των πηγών), και στο πλαίσιο αυτής καθορίζεται:

- ο η συχνότητα με την οποία τα στοιχεία αυτά αντλούνται
- ο η μεθοδολογία – τεχνολογία άντλησης
- ο το αρχείο ή η βάση δεδομένων της ‘περιοχής προετοιμασίας’ στην οποία αρχικά φορτώνονται

Επιπλέον εκτιμάται ο όγκος των δεδομένων που θα αντλούνται ώστε να γίνεται προγραμματισμός χωρητικότητας (capacity planning). Αναπτύσσονται πίνακες εκτίμησης του όγκου (γνωστοί και ως volumetric sheets), όπως είναι ο ακόλουθος:

Πεδίο πηγής	Συχνότητα άντλησης Δεδομένων	Εκτιμώμενος όγκος	Κανόνες Προτυποποίησης και μετατροπής	Πεδίο φόρτωσης στο DW

Η διαδικασία καθαρισμού δεδομένων επιδιώκει την βελτίωση της ποιότητας των στοιχείων. Σχετικές τεχνικές περιγράφονται στο Μέρος Γ: «Ποιότητα Πληροφορίας».

Ο σχεδιασμός και η αυτοματοποιημένη υλοποίηση της διαδικασίας ETL αποτελεί συχνά μεγάλο μέρος της ανθρωποπροσπάθειας ενός έργου ανάπτυξης υποδομής DW (διεθνείς στατιστικές εκτιμούν ότι υπερβαίνει το 70% της ανθρωποπροσπάθειας).

Η ανάπτυξη ‘περιοχής προετοιμασίας’ (staging area) μπορεί να υλοποιείται σε ξεχωριστό υλικό (staging server), προσθέτοντας κάποια πολυπλοκότητα και κόστος στην υλοποίηση. Έχει όμως αρκετά πλεονεκτήματα όπως:

- ο Απομόνωση των ‘ακατέργαστων’ δεδομένων που αντλούνται από τις πηγές, από τα επεξεργασμένα δεδομένα που τελικά αξιοποιούνται κατά την επιχειρησιακή ανάλυση
- ο Επιφέρει πρόσθετη ασφάλεια και ποιότητα στην διαδικασία, δεδομένου ότι οι επιχειρησιακοί αναλυτές δεν έχουν πρόσβαση στην περιοχή αυτή
- ο Επιτρέπει τον διαμοιρασμό φόρτου, δεδομένου ότι διαφορετικά συστήματα εκτελούν την διαδικασία ‘προετοιμασίας δεδομένων’ από αυτά που χρησιμοποιούνται για την λειτουργία του DW και την παραγωγή ‘κύβων’ πληροφορίας.
- ο Επιτρέπει την ύπαρξη ανεξάρτητης κεντρικής αποθήκης μεταδεδομένων (central metadata repository) που τηρεί τεκμηρίωση για το σύνολο των εμπλεκόμενων συστημάτων: παραγωγικά συστήματα, μηχανισμός ETL, data warehouse, προκαθορισμένοι κύβοι πληροφορίας και αναφορές.

Στην περιοχή προετοιμασίας γίνονται διάφορες μορφές επεξεργασίας των ακατέργαστων δεδομένων όπως:

- ο Προτυποποίηση δεδομένων: μετατροπή δεδομένων στην προτυποποιημένη μορφή εφόσον απαιτείται

- Τοποθέτηση των δεδομένων σε διαφορετική σειρά
- Ενοποίηση ομοειδών κατηγοριών δεδομένων από διαφορετικές πηγές, αφού προτυποποιηθούν
- Παραγωγή υπολογιζόμενων μετρήσεων (calculated facts)
- Διαχείριση ‘υποκατάστατων κλειδιών’ (surrogate keys) που αντικαθιστούν κλειδιά που χρησιμοποιούνται στα παραγωγικά συστήματα πληροφορικής
- Εισαγωγή τιμών αναφοράς (default values) όπου απαιτούνται και δεν υπάρχουν
- παραγωγή συγκεντρωτικών στοιχείων όπου προβλέπεται
- τροποποίηση δεδομένων σύμφωνα με την τεχνολογία της αποθήκης (αλλαγή βάσης δεδομένων και λειτουργικού συστήματος)

Μόλις σχεδιαστεί η διαδικασία ETL, επιδιώκεται η ανάπτυξη λογισμικού που αυτοματοποιεί την διαδικασία αυτή, δεδομένου ότι εκτελείται περιοδικά για την ενημέρωση του data warehouse.

Η ανάπτυξη ‘Data mart’ επί επιλεγμένων επιχειρησιακών διαδικασιών, βασίζεται στην μοντελοποίηση δεδομένων, βάσει των μοντέλων διαστάσεων (dimensional modeling) που περιγράφονται στην ενότητα 2.

Η προσέγγιση ‘Kimball’ δέχεται κριτική στα ακόλουθα σημεία:

- Κατά πόσο είναι εφικτή η διασύνδεση των διαφόρων data mart, ειδικά όταν χαρακτηρίζονται από διαφορετικό επίπεδο καταγραφής της πληροφορίας (βαθμός λεπτομέρειας -βλέπε ενότητα 2). Η διασύνδεση διαφορετικών data mart είναι ιδιαίτερα σημαντική δεδομένου ότι επιτρέπει την συνδυαστική επεξεργασία-ανάλυση στοιχείων που ανήκουν σε διαφορετικές επιχειρησιακές διαδικασίες (ανάλυση γνωστή ως drill across). Το θέμα αυτό δεν έχει απαντηθεί με ικανοποιητικό τρόπο.
- Το κανονικοποιημένο σχήμα βάσεων δεδομένων αποτυπώνει καλύτερα περίπλοκες σχέσεις οντοτήτων, σε σχέση με αποκανονικοποιημένα σχήματα. Η πρόταση Kimball προβλέπει αποκλειστική χρήση αποκανονικοποιημένων μοντέλων διαστάσεων.

Προσέγγιση Inmon για την ανάπτυξη DW

Η προσέγγιση Inmon επιδιώκει την εξαρχής συνολική θεώρηση του Οργανισμού και των πληροφοριακών του αναγκών.

Πρώτο βήμα υλοποίησης πρέπει να είναι ο σχεδιασμός συνολικού μοντέλου δεδομένων του Οργανισμού που αποτυπώνει με πληρότητα και λεπτομέρεια τον τρόπο αξιοποίησης της πληροφορίας από τον Οργανισμό (‘atomic data model’). Το μοντέλο αυτό δεν είναι το μοντέλο υφιστάμενων βάσεων δεδομένων του Οργανισμού αλλά ένα νοητό μοντέλο (abstract model) του τρόπου αξιοποίησης της πληροφορίας από τον Οργανισμό.

Βάσει αυτού σχεδιάζεται η δομή δεδομένων της μοναδικής αποθήκης δεδομένων του Οργανισμού (σύμφωνα με την ορολογία της προσέγγισης: Enterprise Data warehouse).

Από το σύνολο των παραγωγικών συστημάτων αντλούνται στοιχεία και μέσω διαδικασίας ETL, ενημερώνεται η μοναδική αποθήκη δεδομένων του Οργανισμού.

Εναλλακτικά και εάν τα παραγωγικά συστήματα πληροφορικής δεν είναι ολοκληρωμένα μεταξύ τους (π.χ. δεν έχουν κοινή βάση δεδομένων), προτείνεται επιπλέον η υλοποίηση ολοκληρωμένης βάσης δεδομένων λειτουργίας σε επίπεδο Οργανισμού (γνωστό ως Operational Data Store- ODS). Η ολοκληρωμένη βάση δεδομένων (ODS) λειτουργίας έχει διττό ρόλο:

- ο Εξυπηρετεί παραγωγικές διαδικασίες
- ο Υποστηρίζει την λήψη αποφάσεων,

και είναι εκείνη που τροφοδοτεί το Enterprise Datawarehouse, εφόσον υλοποιηθεί. Κατά αυτό τον τρόπο αποφεύγεται η ύπαρξη πολλαπλών διαδικασιών ETL (άντλησης-καθαρισμού-μετατροπής-φόρτωσης), από κάθε διαφορετική πηγή δεδομένων.

Σύμφωνα με την προσέγγιση Inmon, τα data mart πρέπει να τροφοδοτούνται αποκλειστικά από την μοναδική αποθήκη δεδομένων του Οργανισμού (Enterprise Data warehouse), ώστε να διασφαλίζεται η μοναδική εκδοχή της αλήθειας. Η προσέγγιση αυτή δημιουργίας data mart είναι γνωστή στην διεθνή βιβλιογραφία και σαν εξαρτημένα data mart ('dependent data mart'), δεδομένου ότι αυτά εξαρτώνται από τη μοναδική αποθήκη δεδομένων.

Σύμφωνα με τον Inmon, η ανάπτυξη 'ανεξάρτητων' data mart (που αντλούν στοιχεία απευθείας από τα παραγωγικά συστήματα) δημιουργεί νησίδες πληροφορίας (information silos) που δεν διασυνδέονται, δεν διασφαλίζουν την μοναδική εκδοχή της αλήθειας και δεν επιτρέπουν την συνδυαστική ανάλυση της πληροφορίας.

Επιπλέον ισχυρίζεται ότι, η έλλειψη της μοναδικής αποθήκης δεδομένων του Οργανισμού αυξάνει την πολυπλοκότητα σε μεγάλους Οργανισμούς όπου υπάρχουν πολλά παραγωγικά συστήματα και χτίζονται πολλά data mart, ειδικά στην περιοχή προετοιμασίας (staging area)(αυτή η άποψη ελέγχεται δεδομένου ότι θεωρεί δεδομένο ότι τα data mart ανήκουν αποκλειστικά σε Τμήματα του Οργανισμού και είναι ασύνδετα).

Η προσέγγιση Inmon έχει δεχτεί κριτική που εστιάζεται στα ακόλουθα :

- ο Δεν είναι σαφής ο τρόπος σχεδιασμού του 'data model' του Οργανισμού, δεδομένου ότι ο τρόπος που ο Οργανισμός αξιοποιεί την πληροφορία πολύ συχνά δεν είναι ούτε τυποποιημένος, ούτε τεκμηριωμένος. Θεωρείται δηλαδή ότι η έννοια του 'data model' του Οργανισμού είναι θεωρητική και ασαφής, συνεπώς είναι δύσκολο στην πράξη να σχεδιαστεί ένα τέτοιο μοντέλο που να ανταποκρίνεται επιτυχημένα στην πραγματικότητα (στον πραγματικό τρόπο αξιοποίησης της πληροφορίας στον Οργανισμό σε βάθος χρόνου). Το επιχείρημα ενισχύεται από το γεγονός ότι οι ανάγκες πληροφόρησης ενός Οργανισμού μεταβάλλονται δυναμικά, δεδομένου ότι ο κάθε Οργανισμός δραστηριοποιείται σε δυναμικά μεταβαλλόμενο περιβάλλον. Η θεώρηση της 'μονομιάς' ανάπτυξης του μοντέλου πληροφόρησης δεν συμβαδίζει με την γενικότερη τάση στην πληροφορική για σταδιακή επένδυση και συστηματική αναθεώρηση του αποτελέσματος αυτής.
- ο Είναι υψηλού κινδύνου διότι απαιτεί μεγάλες επενδύσεις σε χρόνο και χρήμα για τον σχεδιασμό του data model και την ανάπτυξη μοναδικής αποθήκης δεδομένων (Enterprise Datawarehouse-EDW) που καλύπτει το σύνολο των αναγκών του Οργανισμού και σχεδιάζεται βάσει του data model. Βέβαια ο

Ημμοη απαντά ότι προτείνει επαναληπτική (iterative) προσέγγιση ανάπτυξης του EDW, βάσει βέβαια του data model, αντιμετωπίζοντας κάθε φορά ένα μέρος των θεματικών περιοχών (subject areas) του Οργανισμού.

Όπως γίνεται σαφές από τα προηγούμενα, υπάρχουν ριζικές διαφορές στις δυο προσεγγίσεις που περιγράφηκαν παραπάνω, και αντιπροσωπεύουν τις βασικές ‘σχολές σκέψης’, σχετικά με την ανάπτυξη υποδομών datawarehouse. Τα κοινά σημεία και οι διαφορές αναλύονται στην ακόλουθη ενότητα.

Κοινά σημεία και διαφορές επί των δυο προσεγγίσεων

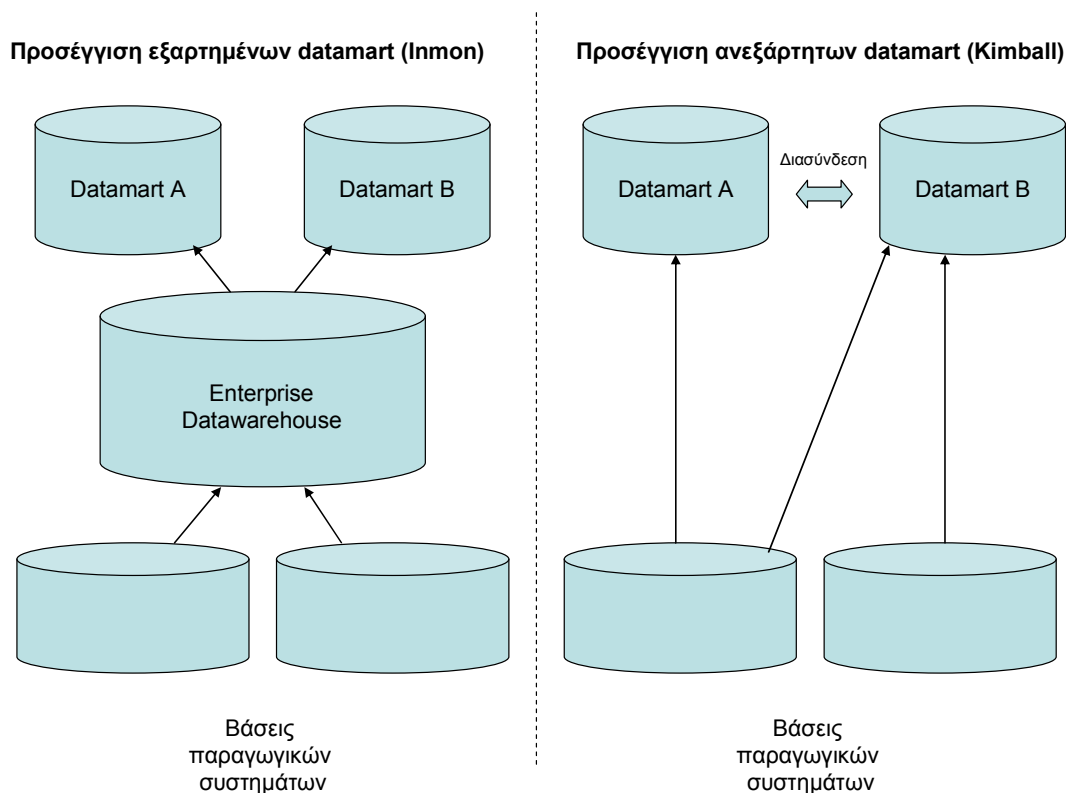
Οι δυο προσεγγίσεις έχουν τα ακόλουθα κοινά σημεία:

- Την χρήση ‘περιοχής προετοιμασίας’ (staging area), όταν η πολυπλοκότητα είναι μεγάλη (όγκος δεδομένων, πολυπλοκότητα διαδικασίας ETL)
- Την υλοποίηση ETL διαδικασιών: άντλησης-καθαρισμού-μετατροπής-φόρτωσης από τις πηγές δεδομένων και προσπάθεια αυτοματισμού αυτής μέσω κατάλληλου λογισμικού
- Την χρήση πολυδιάστατης ανάλυσης σε επίπεδο data mart βάσει μοντέλου διαστάσεων και εργαλείων (On-line Analytical Processing)
- Την επαναληπτική προσέγγιση στην υλοποίηση (iterative approach), βάσει όμως διαφορετικής μεθόδου σχεδιασμού και ανάπτυξης.

Στον ακόλουθο πίνακα παρατίθενται οι βασικές διαφορές των 2 προσεγγίσεων.

	Kimball	Inmon
Φιλοσοφίες ανάπτυξης Data warehouse	Βασισμένο στην παρακολούθηση επιχειρησιακών διαδικασιών του οργανισμού.	Βασισμένο στο 'μοντέλο δεδομένων' του Οργανισμού, όπως το ορίζει η προσέγγιση
	Απευθείας ανάπτυξη data mart σε επιλεγμένες επιχειρησιακές διαδικασίες. Σχεδόν αποκλειστική χρήση αποκανονικοποιημένων σχημάτων μοντέλων διαστάσεων.	Ανάπτυξη μοναδικής αποθήκης δεδομένων (Enterprise Datawarehouse) με κανονικοποιημένο σχήμα βάσης, προ της υλοποίησης data mart
Ορισμός data mart	Τηρεί αναλυτικά στοιχεία (granular data) στο μέγιστο δυνατό βαθμό, και αφορούν την παρακολούθηση επιλεγμένης επιχειρησιακής διαδικασίας. Αναπτύσσονται βάσει της μεθοδολογίας σχεδιασμού μοντέλων διαστάσεων (βλέπε ενότητα 2)	Συγκεντρωτικά στοιχεία (aggregate data) που αφορούν αποκλειστικά ένα τμήμα του Οργανισμού. Χτίζονται βάσει προεπιλεγμένων δεικτών απόδοσης (Key Performance Indicators)
	Ανεξάρτητη ανάπτυξη	Εξαρτημένα από μοναδική αποθήκη δεδομένων (Enterprise Datawarehouse)
	Πλήρης ιστορικότητα	Περιορισμένη ιστορικότητα
Σταδιακή ανάπτυξη του Datawarehouse	Σταδιακή ανάπτυξη νέων datamart επί επιλεγμένων διαδικασιών, που διασυνδέονται βάσει της αρχιτεκτονικής Datawarehouse Bus.	Αρχικός σχεδιασμός του συνολικού Enterprise Datawarehouse, βάσει του 'data model' και σταδιακή υλοποίηση θεματικών περιοχών αυτού.

Στην εικόνα 5 αποτυπώνεται η βασική διαφορά τεχνικής αρχιτεκτονικής μεταξύ των 2 προσεγγίσεων:



Εικόνα 2 – Βασική διαφορά τεχνικής αρχιτεκτονικής προσεγγίσεων Inmon - Kimball

Η διεθνής εμπειρία καταγράφει δυσκολία επιτυχημένης εφαρμογής της προσέγγισης Inmon. Από την άλλη μεριά, οι Οργανισμοί που έχουν αναπτύξει ανεξάρτητα, ασύμβατα-ασύνδετα data mart σε διάφορα τμήματα χωρίς κεντρικό συντονισμό, αντιμετωπίζουν την πρόκληση ενοποίησης αυτών (data mart consolidation) για να επιτύχουν τα σημαντικά οφέλη που προκύπτουν από την συνδυαστική ανάλυση στοιχείων. Συχνά η ενοποίηση αυτή απαιτεί τον ανασχεδιασμό μεγάλου μέρους του data warehouse. Η προσέγγιση Kimball, που λαμβάνει ολόενα και αυξανόμενη υποστήριξη, σε καμία περίπτωση δεν προτείνει ανεξάρτητα και ασύνδετα data mart.