

Big Data & Business Analytics



Φοιτητής: Παναγιώτης Μούχαλης

Σπάρτη 2015

Πνευματικά Δικαιώματα

Copyright © Παναγιώτης Μούχαλης, 2015

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

ΠΕΡΙΛΗΨΗ

Σε αυτήν την πτυχιακή εργασία θα αναλυθούν οι μοντέρνες τεχνολογίες βάσεων δεδομένων με την εμφάνιση των Big Data στον κλάδο των επιχειρήσεων και θα δούμε πως αυτά συμβάλουν άμεσα στην ανάπτυξη μιας εταιρίας σε επίπεδο τέτοιο ώστε να προηγείται του ανταγωνισμού.

Στη συνέχεια θα πούμε για τις βάσεις δεδομένων και θα αναφερθούν ήδη υπάρχοντες τεχνολογίες όπου θα καταγραφούν τα πλεονεκτήματα και τα μειονεκτήματα που προσφέρει η χρήση τους καθώς επίσης και νέες τεχνολογίες βάσεων δεδομένων που δεν χρησιμοποιούν SQL εντολές που λέγονται NoSQL Databases και χρησιμοποιούνται για αποθήκευση και διαχείριση πελώριου όγκου δεδομένων.

Επίσης θα αναλυθούν όλες οι απαραίτητες διεργασίες που χρειάζονται για την εξόρυξη, αποθήκευση και επεξεργασία των Big Data καθώς και οι τα νέα εργαλεία ανάλυσης που κάνουν για πρώτη φορά την εμφάνιση τους στο σύγχρονο εμπόριο, γνωστά ως Business Analytics. Θα δούμε πως αυτή η διαδικασία άμα γίνει με σωστά και προμελετημένα βήματα θα οδηγήσει τα αρμόδια στελέχη των επιχειρήσεων σε λήψη καλύτερων αποφάσεων που θα αποφέρουν περισσότερα κέρδη, μείωση του κόστους παραγωγής ή αύξηση της ποιότητας των προϊόντων.

Λέξεις Κλειδιά: Μεγάλος Όγκος Δεδομένων, Business Analytics, Εξόρυξη Δεδομένων, Σχεσιακή Βάση Δεδομένων, Επιχειρηματική Ευφυΐα, NoSQL Βάσεις Δεδομένων, MapReduce, Αναλυτής Δεδομένων

ABSTRACT

In this thesis the modern database technologies with the appearance of Big Data in the business sector will be analyzed and we will see how these contribute directly to the development of a company at a level that can precede the competition.

Then we will talk about databases and we will mention existing technologies where we will mention the advantages and disadvantages offered by their use as well as new database technologies which don't use SQL statements and they are called NoSQL Databases known to be used for managing immense data.

It will also be analyzed all the necessary processes need for extraction, storage and processing of Big Data and new analysis tools that make their first appearance in the modern trade, known as Business Analytics. We will see how this process when done correctly with deliberately steps will lead the responsible business managers to make better decisions that will bring more profit, reduce production costs or even increase quality.

Key Words: Big Data, Business Analytics, Data Mining, Relational Databases, Business Intelligence, NoSQL Data Bases, MapReduce, Data Scientist

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΕΡΙΛΗΨΗ.....	3
ABSTRACT.....	4
ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ	5
ΠΙΝΑΚΑΣ ΔΙΑΓΡΑΜΜΑΤΩΝ.....	7
ΕΙΣΑΓΩΓΗ.....	8
Κεφάλαιο1 -Big Data	9
1.1 Ορισμός.....	9
1.2 Παρουσίαση των 3Vs	11
1.3 Πως τα Big Data επηρεάζουν τις επιχειρήσεις.....	13
Κεφάλαιο 2-Εργαλεία των Big Data.....	14
2.1 Παράλληλες Βάσεις Δεδομένων	14
2.2 MapReduce.....	17
2.3 Hadoop	19
2.3.1 Πως το Hadoop διαφέρει από άλλες παλιότερες τεχνολογίες;.....	19
2.3.2 Λόγοι για τη χρήση του Hadoop	21
Κεφάλαιο 3 - NoSQL Databases	22
3.1 Ορισμός.....	22
3.2 Διαφορές NoSQL – SQL.....	22
3.3 Τύποι NoSQL Βάσεων Δεδομένων.....	25
Κεφάλαιο 4 - Παραδείγματα Βάσεων NoSQL	27
4.1 MongoDB.....	27
4.2 BigTable	29
4.3 Apache HBase	31
4.4 Apache Cassandra	33
Κεφάλαιο 5 - Τεχνολογίες Data - Warehousing Columnar Data - stores	37
5.1 Ορισμός.....	37
5.2 Διαφορές μεταξύ Αποθήκης Δεδομένων και άλλων Βάσεων Δεδομένων	38
5.3 Τι κάνουν οι Αποθήκες Δεδομένων	39
5.4 Οφέλη της Αποθήκης Δεδομένων	40
5.5 Διαδικασία ETL	41

Κεφάλαιο 6 - Εξόρυξη Γνώσης (Data Mining).....	42
6.1 Ορισμός Εξόρυξης Γνώσης.....	42
6.2 Εξόρυξη Δεδομένων και Ανεύρεση Γνώσης.....	42
6.3 Στόχοι της Εξόρυξης Δεδομένων	45
Κεφάλαιο 7 - Τομείς που εξελίχθηκαν με τη βοήθεια των Big Data.....	47
7.1 Αναλυτής Δεδομένων (Data Scientist).....	47
Κεφάλαιο 8 - Business Analytics	48
8.1 Ορισμός.....	48
8.2 Τα Business Analytics στις επιχειρήσεις.....	49
8.2.1 Marketing	50
8.2.2 Human Resources Management	51
8.2.3 Finance	55
8.2.4 Supply Chain Management	57
8.2.5 Production and Operations	58
8.2.6 Internet Analytics	59
Κεφάλαιο 9 - Επιχειρηματική Ευφυΐα (BI - Business Intelligence)	61
9.1 Ορισμός.....	61
9.2 Βήματα Δημιουργίας Business Intelligence Λύσεων.....	63
9.3 Business Intelligence Tools.....	65
9.3.1 OLAP Reports	65
9.3.2 KPI	66
Κεφάλαιο 10 - Μελλοντικά.....	67
Το μέλλον του ΣΔΒΔ	67
Βιβλιογραφία και Πηγές.....	68

ΠΙΝΑΚΑΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

Διάγραμμα 1-Απότομη Αύξηση των Big Data από το 2011	10
Διάγραμμα 2 - : Γραφική Αναπαράσταση των 3Vs.	11
Διάγραμμα 3 - Διαφορές ACID & BASE	24
Διάγραμμα 4 - : Η αρχιτεκτονική «Δαχτυλιδιού» του Cassandra.....	34
Διάγραμμα 5 - Διαδικασία ETL (Extract – Transform – Load).....	41
Διάγραμμα 6 - Διάγραμμα Ροής Διαδικασίας Business Intelligence.....	62

ΕΙΣΑΓΩΓΗ

Στην παρούσα πτυχιακή εργασία περιγράφονται οι νέες τεχνολογίες διαχείρισης δεδομένων, τα συστήματα τα οποία τα υποστηρίζουν και οι τεχνικές επεξεργασίας των δεδομένων αυτών. Στη συνέχεια θα δούμε τους τρόπους που μπορούμε να αξιοποιήσουμε τις πληροφορίες που προκύπτουν από την επεξεργασία των δεδομένων για τον καλύτερο βαθμό επίτευξης στόχων των επιχειρήσεων.

Ζούμε σε ένα σύγχρονο κόσμο που περιβάλλεται από νέες τεχνολογίες. Συνεπώς νέες ανάγκες έχουν εμφανιστεί τόσο σε ατομικό επίπεδο όσο και σε επιχειρησιακό επίπεδο. Οι επιχειρήσεις πλέον αποθηκεύουν όλα τους τα δεδομένα σε ψηφιακή μορφή. Έτσι όμως γεννιούνται νέες προκλήσεις όπως: πώς θα μπορέσουν οι επιχειρήσεις να διαχειριστούν τόσο μεγάλο όγκο πληροφορίας; Υπάρχουν τρόποι για να προκύψουν χρήσιμες πληροφορίες από την επεξεργασία όλων αυτών των δεδομένων;

Καθημερινά δημιουργείται μεγάλος όγκος δεδομένων. Παρακάτω θα αναφερθούν οι ανάγκες που οδήγησαν στην ανάπτυξη των Big Data και στην ανεύρεση εργαλείων ανάλυσης για την βέλτιστη επεξεργασία των δεδομένων με σκοπό να αποφέρουν χρήσιμα στοιχεία που μπορούν να οδηγήσουν μια επιχείρηση στην επιτυχία.

Έχουν πλέον αναπτυχθεί ποικιλόμορφες τεχνολογίες βάσεων δεδομένων που η χρήση τους διαφέρει ανάλογα με τις ανάγκες του κάθε οργανισμού. Παρακάτω θα αναλύσουμε τα πιο διαδεδομένα μοντέλα βάσεων δεδομένων.

Κεφάλαιο1 -Big Data

1.1 Ορισμός

Τα Big Data είναι ένας ευρύς όρος που αναφέρεται σε μεγάλου όγκου πολύπλοκα δεδομένα που παρουσιάζουν εκπληκτικά γρήγορη αύξηση όπου για την αποθήκευση και επεξεργασία τους κρίνεται απαραίτητη η έρευνα και ανάπτυξη νέων εργαλείων και μεθόδων. Ουσιαστικά τα Big Data είναι Data αλλά πολύ μεγαλύτερα. Οι παραδοσιακές τεχνικές και εφαρμογές των ημερών μας είναι πλέον ανεπαρκείς, διότι όσο περισσότερο αναπτύσσονται τα Big data τόσο περισσότερες προκλήσεις κάνουν την εμφάνιση τους στο παρασκήνιο. Μεταξύ των βασικών προκλήσεων είναι η συλλογή, αποθήκευση, επεξεργασία των δεδομένων καθώς και η κοινή χρήση, η ανάλυση και η απεικόνιση των αποτελεσμάτων.

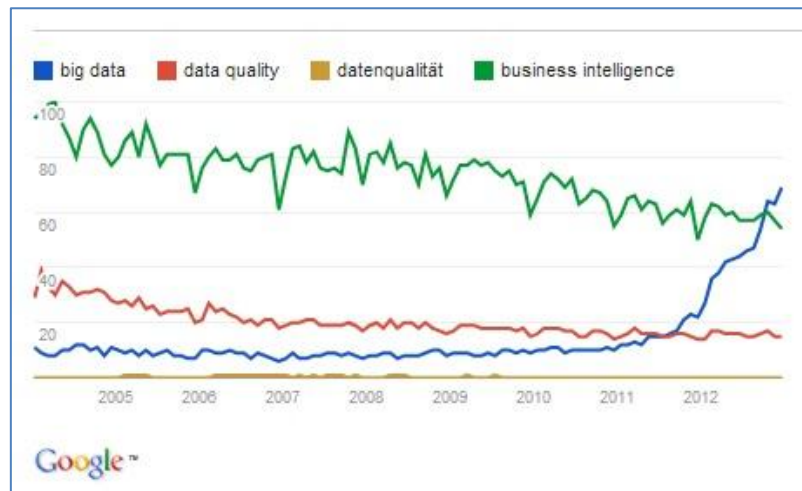
Οι παραπάνω διαδικασίες όταν εκτελούνται με ακρίβεια προσφέρουν έγκυρα στοιχεία, τα οποία οδηγούν στη λήψη κατάλληλων αποφάσεων. Τα Big data παίζουν πολύ σημαντικό ρόλο στη λειτουργικότητα των επιχειρήσεων ή των οργανισμών, διότι τα αποτελέσματα που προκύπτουν από την ορθή ανάλυση και αποσαφήνιση τους προσφέρουν ευκαιρίες στα ηγετικά μέλη για λήψη καλύτερων αποφάσεων που μπορούν να επιφέρουν μεγαλύτερη λειτουργική αποτελεσματικότητα και περισσότερα κέρδη ή μείωση του συνολικού κόστους.

Η ανάπτυξη της τεχνολογίας σε τομείς όπως εξόρυξη δεδομένων (Data Mining) και ανάλυση τους (Analytics), καθώς και οι τεράστιες υπολογιστικές και αποθηκευτικές ικανότητες των σύγχρονων υπολογιστικών συστημάτων κάνουν διαθέσιμο σε οργανισμούς και ιδιώτες ένα τεράστιο όγκο πληροφοριών. Ο ραγδαίος αυτός ρυθμός αύξησης των δεδομένων οφείλεται σε ένα πολύ μεγάλο βαθμό στη χρήση του διαδικτύου όπου καθημερινά δημιουργούνται ποικιλόμορφοι όγκοι δεδομένων.

Όπως βλέπουμε στο Διάγραμμα 1, με το πέρασμα του χρόνου τα Big data αναπτύσσονται ολοένα και περισσότερο. Κάθε μέρα άνθρωποι και μηχανές συμμετέχουν με ασταμάτητο ρυθμό στη δημιουργία τεραστίου όγκου δεδομένων.

Πιο συγκεκριμένα με έρευνα του USC Annenberg School of Communication and Journalism ξεδιπλώνεται αυτή η έκρηξη πληροφοριών με αποκαλυπτικούς πραγματικά αριθμούς. Το 2007 υπήρχαν στον κόσμο αποθηκευμένα 300 Exabyte. Από τα δεδομένα αυτά μόνο το 7% ήταν σε αναλογική μορφή, τα υπόλοιπα ήταν σε ψηφιακή μορφή ενώ το 2000

μόνο το 25% των δεδομένων ήταν σε ψηφιακή μορφή, το υπόλοιπο 75% ήταν αποθηκευμένο σε χαρτί, φιλμ, δίσκους βινυλίου, κασέτες, δισκέτες κλπ. Ήδη υπολογίζεται ότι ο όγκος της αποθηκευμένης πληροφορίας το 2013 ήταν 1.200 Exabyte από τα οποία μόνο το 2% θα ήταν σε αναλογική μορφή. Μια χιουμοριστική αλλά ταυτόχρονα παραστατική εικόνα του ασύλληπτου αυτού μεγέθους θα ήταν να φανταστούμε ότι μπορούμε να καλύψουμε όλη την έκταση των Η.Π.Α. σε 52 στρώματα βιβλίων.



Διάγραμμα 1-Απότομη Αύξηση των Big Data από το 2011

Επίσης σύμφωνα με μετρήσεις που έλαβαν χώρα το 2012, καθημερινά δημιουργούνται 2,5 Exabyte δεδομένων. Για παράδειγμα ας πάρουμε τη γνωστή σε όλους πλατφόρμα κοινωνικής δικτύωσης, το Facebook, το οποίο υπολογίζεται ότι καθημερινά παράγει 20 Terabyte νέων πληροφοριών σε συμπιεσμένη μορφή. Σύμφωνα με το νόμο του Moore κάθε 18 μήνες διπλασιάζονται τα ήδη υπάρχοντα δεδομένα στον κόσμο.

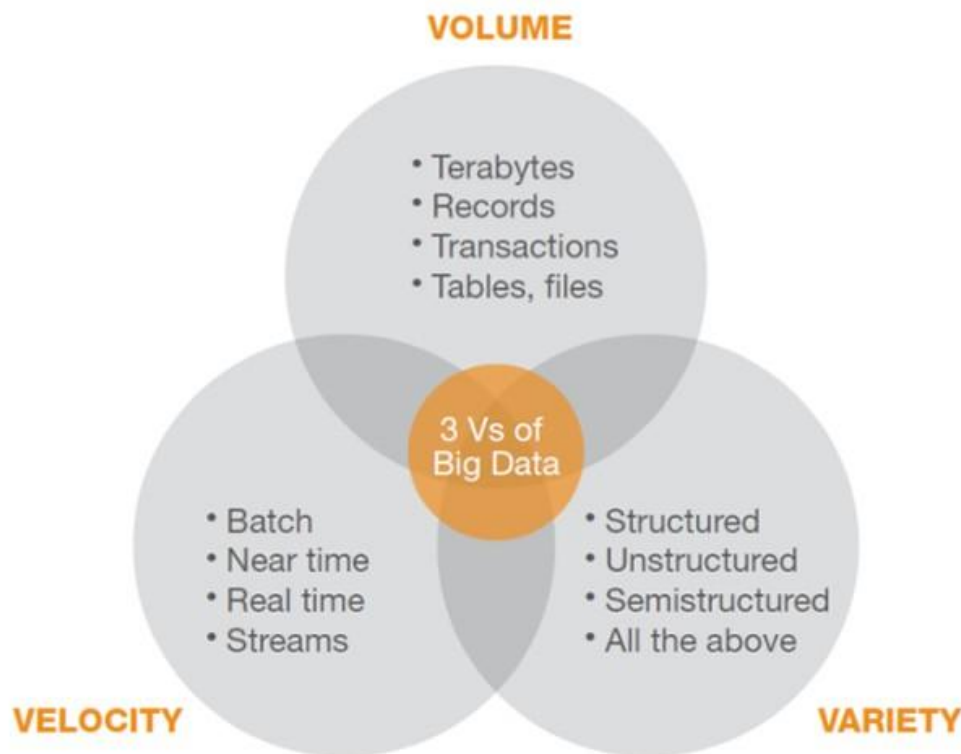
Πλέον με αυτούς τους όγκους δεδομένων η χρήση συστημάτων διαχείρισης σχεσιακών βάσεων δεδομένων οδηγεί σε αδιέξοδο αφού παρουσιάζεται συχνά δυσκολία στο χειρισμό των δεδομένων διότι έχει αυξηθεί σε αρκετά μεγάλο βαθμό η πολυπλοκότητα τους.

1.2 Παρουσίαση των 3Vs

Για την ορθότερη αντίληψη των Big Data και την καλύτερη αποσαφήνιση των προβλημάτων που προκύπτουν κατά το χειρισμό τους, δημιουργήθηκε ένας ευρύς όρος που περιλαμβάνει τα βασικά τους χαρακτηριστικά. Ο όρος αυτός είναι γνωστός ως τα '3Vs'.

Τα 3Vs των Big Data αποτελούνται από:

- **Volume** : Μεγάλος όγκος δεδομένων
- **Variety** : Μεγάλη ποικιλία δεδομένων
- **Velocity** : Μεγάλη ταχύτητα ανάλυσης των δεδομένων



Διάγραμμα 2 - : Γραφική Αναπαράσταση των 3Vs.

Αναλυτικότερα έχουμε:

1. **Ο Όγκος (Volume):** Ζούμε σε έναν κόσμο που θα μπορούσαμε να τον χαρακτηρίσουμε ψηφιακό. Καθημερινά στέλνονται e-mail, δημοσιεύονται κείμενα, φωτογραφίες, βίντεο, επικοινωνία μέσω SMS και με τον τρόπο αυτό γίνεται αντιληπτό πόσο ψηφιακός έχει γίνει ο κόσμος. Όμως δεν είμαστε οι μόνοι που δημιουργούμε πληροφορίες ή στέλνουμε δεδομένα από τη μία άκρη του κόσμου στην άλλη. Σήμερα εκατομμύρια ψηφιακές συσκευές δημιουργούν πληροφορίες. Αισθητήρες που συλλέγουν πληροφορίες ή και μηχανές που επικοινωνούν μεταξύ τους (Machine to machine).

Επίσης παράδειγμα μεγάλου όγκου μπορούν να θεωρηθούν εγγραφές ή και ολόκληροι πίνακες πληροφοριών, δεδομένα συναλλαγών, αρχεία κ.α.

2. **Η ποικιλία (Variety):** Τα δεδομένα σήμερα διακρίνονται από τη δομή τους σε:
 - Δομημένα με τη μορφή σχεσιακών δεδομένων αποθηκευμένων σε παραδοσιακές βάσεις,
 - Αδόμητων δεδομένων σε μορφή εικόνας, ήχου, video, email, κτλ.
 - επίσης υπάρχουν και τα Ημιδομημένα δεδομένα που παρουσιάζουν στοιχεία και από τις δυο προηγούμενες κατηγορίες.

Για κάθε κατηγορία οι διεργασίες εξόρυξης, αποθήκευσης και η ανάλυσης των δεδομένων διαφοροποιούνται.

3. **Η ταχύτητα (Velocity):** Πλέον η παραγωγή των δεδομένων γίνεται με πολύ μεγάλους ρυθμούς οπότε εμείς πρέπει να είμαστε σε θέση να τα αντιμετωπίσουμε έγκαιρα.

Σύγχρονες πηγές δεδομένων είναι:

- Τα δίκτυα
- Η ανθρώπινη αλληλεπίδραση σε συστήματα όπως τα social media
- Financial streams
- Σταθμοί μετεωρολογικών παρατηρήσεων
- Συστήματα γεωγραφικού εντοπισμού
- Άντληση δεδομένων από τις μηχανές

Καθημερινά οι Business Analysts έρχονται αντιμέτωποι με μεγάλου όγκου δεδομένα και κρίνεται απαραίτητο να μπορούν να τα επεξεργαστούν σε όσο το δυνατόν συντομότερο χρονικό και να αντλήσουν χρήσιμες πληροφορίες από αυτά. Η άμεση αντιμετώπιση μας στους γοργούς ρυθμούς των δεδομένων και η εξαγωγή των συμπερασμάτων σε πραγματικό χρόνο αποτελεί πρόκληση για τους σύγχρονους οργανισμούς. [1]

1.3 Πως τα Big Data επηρεάζουν τις επιχειρήσεις

Όσο περνάει ο καιρός, στον κόσμο των επιχειρήσεων γίνεται περισσότερο κατανοητή η σημασία των Big Data καθώς και η αξία που μπορούν να προσφέρουν στην ανάπτυξη μιας επιχείρησης. Αρκετοί οργανισμοί βρίσκονται στο πρώιμο στάδιο πειραματισμού και ελάχιστοι είναι αυτοί που έχουν σκεφτεί και αντιληφθεί την βαθύτερη επιρροή των Big Data στην υποδομή των πληροφοριών τους.

Σύμφωνα με τη διεθνή εταιρία οικονομικών συμβουλών (Consulting), Gartner έως το τέλος 2015 περίπου το 20% των διεθνών οργανισμών θα έχουν εστιάσει τη στρατηγική τους στη συλλογή και αποσαφήνιση των δεδομένων με την χρήση των κατάλληλων εφαρμογών. Επίσης στην έκθεση της Gartner που διεξήχθη το 2013 με τίτλο «Predicts 2013: Big Data and Information Infrastructure;» περιγράφεται λεπτομερώς πως τα Big Data θα επηρεάσουν σημαντικά τους οργανισμούς, τους πόρους, καθώς και την υποδομή των πληροφοριών. Συνεπώς δημιουργείται μια νέα υποχρέωση για τους ηγέτες του IT (Information Technology), να χρησιμοποιήσουν τις πληροφορίες της έρευνας αυτής, να προσαρμόσουν και να προγραμματίσουν τις αξιοσημείωτες αλλαγές που αναμένονται να ακολουθήσουν. Εξίσου σημαντικό κρίνεται οι IT Managers να μπορούν να χρησιμοποιήσουν την γνώση και την εμπειρία τους για να εξετάσουν τα εργαλεία ενοποίησης των δεδομένων καθώς επίσης και να αναπτύξουν νέα εργαλεία ώστε να μπορούν να εκμεταλλευτούν στο μέγιστο τα πλεονεκτήματα που προκύπτουν άφιξη των νέων τεχνολογιών.

Κεφάλαιο 2-Εργαλεία των Big Data

2.1 Παράλληλες Βάσεις Δεδομένων

Με τη ραγδαία αύξηση των Big Data, εμφανίστηκαν αρκετά προβλήματα όσον αφορά την αποθήκευση τους καθώς οι παραδοσιακές βάσεις δεδομένων δεν μπορούν να υποστηρίξουν τόσο μεγάλους όγκους δεδομένων. Τα μεγέθη των δεδομένων πλέον μετριούνται σε Terabytes και Petabytes. Επίσης η παράλληλη επεξεργασία των δεδομένων σε τέτοια συστήματα παρουσιάζει προβλήματα. Δημιουργήθηκε λοιπόν η ανάγκη για ανάπτυξη νέων συστημάτων βάσεων δεδομένων όπου βασικό στόχο έχουν τη βελτιστοποίηση της απόδοσης. Βάσεις δεδομένων που θα μπορούν να αποθηκεύσουν με ευκολία μεγάλες ποσότητες δεδομένων και θα παρέχουν αύξηση των διαθέσιμων πόρων ώστε να τους υποστηρίξουν την παράλληλη επεξεργασία των δεδομένων. [2]

Για την επίτευξη του στόχου αυτού το σύστημα πρέπει να χαρακτηρίζεται από τα εξής:

- Η πρόσβαση θα πρέπει να είναι επεκτάσιμη, το σύστημα αρχείων θα πρέπει να επιτρέπει να προστεθεί επιπλέον Hardware αν αυτό κριθεί απαραίτητο για την αύξηση της ικανότητας αποθήκευσης και των επιδόσεων.
- Το σύστημα αρχείων θα πρέπει να προσφέρει υψηλή απόδοση και επίσης να είναι σε θέση να εντοπίζει τα δεδομένα που μας ενδιαφέρουν επί των διανεμόμενων κόμβων σε πραγματικό χρόνο.
- Θα πρέπει να είναι αξιόπιστο και σε περίπτωση ανάκτησης των αρχικών δεδομένων θα πρέπει να είναι σε θέση να αναδημιουργήσει από τους κατανεμημένους κόμβους τα αρχικά δεδομένα με τρόπο πλήρη και χωρίς διαστρεβλώσεις.
- Θα πρέπει να έχει υψηλή διαθεσιμότητα, να χειρίζεται τις αποτυχίες και να ενσωματώνει μηχανισμούς για την παρακολούθηση και ανίχνευση σφαλμάτων, την ανοχή σφαλμάτων και την αυτόματη αποκατάσταση.

Μία καταναμημένη βάση δεδομένων (DDB) είναι μια συλλογή πολλαπλών λογικά διασυνδεδεμένων Βάσεων Δεδομένων που διανέμονται μέσω ενός δικτύου υπολογιστών. Ένα καταναμημένο σύστημα διαχείρισης Βάσης Δεδομένων (Database Management System – DBMS) είναι ένα σύστημα διαχείρισης που υλοποιείται σε έναν υπολογιστή με πολλούς επεξεργαστές. Το παράλληλο DBMS υλοποιεί την έννοια του οριζόντιου κατακερματισμού διανέμοντας τα μέρη ενός μεγάλου σχεσιακού πίνακα σε πολλούς κόμβους με δυνατότητα να υποβάλλονται σε επεξεργασία με παραλληλισμό. Αυτό απαιτεί διαμερισμένη εκτέλεση SQL λειτουργιών. Κάποιες βασικές λειτουργίες, όπως για παράδειγμα μία απλή SELECT, μπορεί να εκτελεστεί ανεξάρτητα σε όλους τους κόμβους. Οι πιο σύνθετες λειτουργίες εκτελούνται μέσω ενός αγωγού πολλαπλών λειτουργιών. Διαφορετικές αρχιτεκτονικές πολλαπλών παράλληλων συστημάτων, όπως είναι η κοινή μνήμη με κοινούς δίσκους ή τίποτα κοινό, καθορίζουν τις πιθανές στρατηγικές για την υλοποίηση ενός παράλληλου DBMS, όπου η κάθε μια με τα δικά της πλεονεκτήματα και μειονεκτήματα.

Με βάση τους παραπάνω ορισμούς, μπορούμε να συμπεράνουμε ότι τα συστήματα παράλληλων βάσεων δεδομένων βελτιώνουν τις επιδόσεις της επεξεργασίας των δεδομένων μέσω του παραλληλισμού του φόρτου, θέτοντας κατάλληλα ευρετήρια και επερωτήσεις στα δεδομένα. Στα καταναμημένα συστήματα βάσεων δεδομένων τα δεδομένα αποθηκεύονται σε διαφορετικά DBMS που μπορούν να λειτουργήσουν ανεξάρτητα. Επειδή τα συστήματα παράλληλων βάσεων δεδομένων μπορούν να διανέμουν δεδομένα για να αυξήσουν την απόδοση της αρχιτεκτονικής, υπάρχει μια λεπτή γραμμή που χωρίζει τις δύο έννοιες σε πραγματικές εφαρμογές.

Παρά τις διαφορές μεταξύ των παράλληλων και καταναμημένων DBMS, τα περισσότερα από τα πλεονεκτήματα του είναι κοινά:

- Τα αποθηκευμένα δεδομένα είναι σύμφωνα με ένα καλά καθορισμένο σχήμα και αυτό επικυρώνει τα δεδομένα και παρέχει την ακεραιότητα των δεδομένων.
- Τα δεδομένα είναι δομημένα σε ένα σχεσιακό μοντέλο γραμμών και στηλών.
- Τα ερωτήματα SQL είναι γρήγορα.
- Η γλώσσα ερωτημάτων SQL είναι ευέλικτη εύκολη να τη μάθει κανείς και να διαβαστεί και επιτρέπει στους προγραμματιστές να εφαρμόσουν σύνθετες εργασίες με ευκολία.

- Χρησιμοποιούν κατακερματισμό ή B-δενδρικά ευρετήρια για να επιταχύνουν την πρόσβαση στα δεδομένα.
- Μπορούν να επεξεργαστούν αποτελεσματικά σύνολα δεδομένων μεγέθους έως και δύο PetaBytes. Γνωστές εμπορικές παράλληλες βάσεις δεδομένων είναι Teradata, Aster Data, Netezza, DATAlegro, Vertica, Greenplum, IBMDB2 και OracleExaData έχουν αποδειχθεί επιτυχής, διότι επιτρέπουν την γραμμική κλιμάκωση.
- Το σύστημα μπορεί να διατηρήσει σταθερή απόδοση καθώς το μέγεθος της βάσης δεδομένων αυξάνεται με την προσθήκη περισσότερων κόμβων στο παράλληλο σύστημα.
- Επιτρέπουν την γραμμική επιτάχυνση, για μια βάση δεδομένων με ένα σταθερό μέγεθος, η απόδοση μπορεί να αυξηθεί με την προσθήκη περισσότερων συστατικών, όπως επεξεργαστές, μνήμη και τους δίσκους.
- Εφαρμόζουν εσωτερικά ερωτήματα, ενδο-ερωτήματα και ενδο-λειτουργία παραλληλισμού.
- Απαιτούν μειωμένη προσπάθεια εφαρμογής.
- Απαιτούν μειωμένη προσπάθεια διοίκησης
- Προσφέρουν υψηλή διαθεσιμότητα.

Στην αρχιτεκτονική μαζικής παράλληλης επεξεργασίας (Massively Parallel Processing - MPP) τα δεδομένα επεξεργάζονται από πολλαπλούς εξυπηρετητές ή κόμβους όπου ο καθένας από αυτούς διαθέτει δική του μονάδα επεξεργασίας και μνήμη RAM και επεξεργάζεται τα δεδομένα τοπικά.[3] [4]

Η MPP αρχιτεκτονική υλοποιείται ως συσκευή αποθήκευσης δεδομένων, μειώνει την προσπάθεια εφαρμογής λόγω του ότι το υλικό και το λογισμικό είναι προεγκατεστημένα και δοκιμασμένα για να λειτουργήσουν τη συσκευή, πριν την προμηθευτούμε. Επίσης, μειώνει την προσπάθεια διαχείρισης, καθώς έρχεται ως ανεξάρτητη συσκευή από το υπόλοιπο σύστημα. Οι συσκευές αποθήκευσης δεδομένων προσφέρουν υψηλή διαθεσιμότητα με ενσωματωμένες δυνατότητες ανακατεύθυνσης χρησιμοποιώντας πλεονασμό δεδομένων για κάθε δίσκο. Ιδανικά, κάθε μονάδα επεξεργασίας της συσκευής αποθήκης δεδομένων πρέπει να επεξεργάζεται την ίδια ποσότητα δεδομένων σε κάθε δεδομένη στιγμή. Για να επιτευχθεί αυτό, τα δεδομένα θα πρέπει να κατανέμονται ομοιόμορφα σε κάθε μονάδα επεξεργασίας. Έχοντας λοιπόν κάθε μονάδα επεξεργασίας να κάνει το ίδιο ποσό εργασίας εξασφαλίζεται ότι

όλες οι μονάδες επεξεργασίας θα τελειώσουν το έργο τους περίπου την ίδια χρονική στιγμή, ελαχιστοποιώντας τυχόν χρόνους αναμονής. Μια άλλη πτυχή που έχει σημαντικό αντίκτυπο στην απόδοση του ερωτήματος είναι ότι υπάρχουν μαζί όλα τα δεδομένα που σχετίζονται με την ίδια μονάδα επεξεργασίας. Με αυτόν τον τρόπο ο χρόνος που απαιτείται για τη μεταφορά δεδομένων μεταξύ των μονάδων επεξεργασίας εξαλείφεται. Ο τρόπος με τον οποίο τα δεδομένα κατανέμονται σε παράλληλους κόμβους της βάσης δεδομένων επηρεάζουν τη συνολική απόδοση. Αν και η ισχύς του παράλληλου DBMS δίδεται από τον αριθμό των κόμβων, μπορεί να είναι επίσης ένα μειονέκτημα. Για απλές ερωτήσεις ο πραγματικός χρόνος επεξεργασίας μπορεί να είναι πολύ μικρότερος από το χρόνο που απαιτείται για να ξεκινήσει η παράλληλη λειτουργία. Επίσης οι κόμβοι μπορούν να γίνουν hotspots ή εμπόδια καθώς καθυστερούν το όλο σύστημα.[10] [4]

2.2 MapReduce

Το μοντέλο MapReduce (MR) είναι ένα μοντέλο προγραμματισμού και μια σχετική εφαρμογή για την επεξεργασία και τη δημιουργία μεγάλων συνόλων δεδομένων. Είναι κατάλληλο για την ανάπτυξη εφαρμογών οι οποίες επεξεργάζονται γρήγορα και παράλληλα τεράστιες ποσότητες δεδομένων σε συστοιχίες (Clusters) υπολογιστών. Το μοντέλο αναπτύχθηκε από τον Jeffrey Dean και Sanjay Ghemawat στη Google. Το μοντέλο MapReduce στηρίζεται στη λειτουργία μιας συνάρτησης Map που χρησιμοποιεί ως αρχική διαδικασία ζεύγη κλειδιού-τιμής και μια συνάρτηση Reduce που συγχωνεύει όλες τις ενδιαμέσες τιμές του ίδιου κλειδιού. Το σύνολο των δεδομένων χωρίζεται σε μικρότερα υποσύνολα τα οποία υποβάλλονται σε επεξεργασία παράλληλα με μια μεγάλη συστάδα υπολογιστικών μηχανών. Η συνάρτηση Map, παίρνει ένα δεδομένο εισόδου και παράγει ένα σύνολο ενδιαμέσων υποσυνόλων. Η βιβλιοθήκη του μοντέλου MapReduce ομαδοποιεί όλα τα ενδιαμέσα υποσύνολα που συνδέονται με το ίδιο ενδιαμέσο κλειδί και να τα προωθεί στην συνάρτηση Reduce. Η συνάρτηση Reduce, δέχεται επίσης ένα ενδιαμέσο κλειδί και τα υποσύνολα που σχετίζονται με αυτό. Η συνάρτηση αυτή συγχωνεύει μαζί με αυτά τα υποσύνολα και το κλειδί για να σχηματίσουν ένα πιθανώς μικρότερο σύνολο τιμών. Κανονικά η συνάρτηση Reduce παράγει καμία ή μία τιμή εξόδου. Πολλές εργασίες γίνονται στον πραγματικό κόσμο με τη χρήση του μοντέλου MapReduce. Αυτό το μοντέλο χρησιμοποιείται για την υπηρεσία αναζήτησης στο διαδίκτυο, για τη διαλογή και την επεξεργασία των δεδομένων, για την εξόρυξη δεδομένων, για μηχανική μάθηση και για ένα

μεγάλο αριθμό από άλλα συστήματα. Το πλαίσιο του μοντέλου MapReduce διαχειρίζεται τον τρόπο με τον οποίο τα δεδομένα χωρίζονται μεταξύ των κόμβων και κατά πόσο τα ενδιάμεσα αποτελέσματα του ερωτήματος είναι καθολικά για το σύστημα.

Τα πλεονεκτήματα του μοντέλου MapReduce είναι:

- Το μοντέλο είναι εύκολο στη χρήση, ακόμη και για τους προγραμματιστές χωρίς εμπειρία σε παράλληλα και καταναεμημένα συστήματα.
- Παρέχει ανεξαρτησία αποθήκευσης του συστήματος, δεδομένου ότι δεν απαιτεί ιδιόκτητα συστήματα αρχείων δεδομένων ή προκαθορισμένα μοντέλα δεδομένων.
- Τα δεδομένα αποθηκεύονται σε αρχεία απλού κειμένου και το σύστημα δεν είναι υποχρεωμένο να συμμορφώνεται στα σχεσιακά συστημάτων δεδομένων ή οποιαδήποτε άλλη δομή.

Στην πραγματικότητα η αρχιτεκτονική MapReduce :

- Μπορεί να χρησιμοποιήσει δεδομένα που έχουν μια αυθαίρετη μορφή.
- Έχει ανοχή σε σφάλματα.
- Είναι διαθέσιμη για γλώσσες προγραμματισμού υψηλού επιπέδου.
- Παρέχει γλώσσα επερωτήσεων και επιτρέπει χειρισμό με μεγάλη ταχύτητα.

[15]

2.3 Hadoop

Το Hadoop είναι ένα framework ανοιχτού κώδικα που δημιουργήθηκε για να ενισχύσει την επεξεργασία των Big Data. Αναπτύσσεται από την εταιρία Apache Software Foundation και υποστηρίζει κατανεμημένη επεξεργασία μεγάλου όγκου δεδομένων. Η ανάπτυξή του ξεκίνησε το 2005 ως μια εναλλακτική επιλογή ανοιχτού κώδικα του Google MapReduce framework και του Google File System (GFS) στα οποία και βασίστηκε και έκτοτε αποτελεί ένα από τα καλύτερα κατανεμημένα συστήματα για batch processing.

Το Apache Hadoop framework αποτελείται από τα εξής τμήματα:

- **Hadoop Common:** Περιέχει βιβλιοθήκες και εργαλεία απαραίτητα για τη λειτουργία των υπόλοιπων τμημάτων.
- **Hadoop Distributed File System (HDFS):** Κατανεμημένο, κλιμακώσιμο (scalable) σύστημα αρχείων.
- **Hadoop YARN :** Πλατφόρμα υπεύθυνη για τη διαχείριση των υπολογιστικών πόρων και υπεύθυνη για τη χρονοδρομολόγηση των εργασιών του χρήστη. Ενσωματώθηκε στο Hadoop framework στη έκδοση του Hadoop 2.0.
- **Hadoop MapReduce:** Προγραμματιστικό μοντέλο για την επεξεργασία μεγάλου όγκου δεδομένων.

Ένας από τους κυριότερους λόγους της επιτυχίας αυτού του framework είναι το γεγονός ότι σχεδιάστηκε να τρέχει σε συστάδες από απλούς υπολογιστές (clusters) χαμηλού κόστους όπου όμως η συχνότητα βλαβών είναι υψηλή. Οι βλάβες αντιμετωπίζονται αυτόματα από το framework εξασφαλίζοντας τη σταθερότητα του συστήματος

Το Hadoop είναι προγραμματισμένο σε γλώσσα Java. Στην ουσία πρόκειται για μια συλλογή από Java processes που τρέχουν παράλληλα σε ένα σύνολο από μηχανές.

2.3.1 Πως το Hadoop διαφέρει από άλλες παλιότερες τεχνολογίες;

Οι διαφορές του από τις άλλες τεχνολογίες εστιάζονται στα παρακάτω:

- ✓ Το Hadoop μπορεί να διαχειριστεί δεδομένα πολύ εύκολα. Είναι κάτι περισσότερο από απλά γρήγορο, περιέχει φθηνότερη βάση δεδομένων και εργαλείο Analytics. Σε αντίθεση με τις βάσεις δεδομένων, το Hadoop δεν απαιτεί τη δόμηση στα δεδομένα καθώς αυτά μπορεί να είναι αδόμητα και ακατέργαστα. Οι χρήστες μπορούν να

αποθηκεύουν τα δεδομένα τους στο framework χωρίς να υπάρχει ανάγκη για διαμόρφωση τους. Αντιθέτως οι σχεσιακές βάσεις δεδομένων απαιτούν τα δεδομένα να είναι δομημένα και με τη σωστή διαμόρφωση πριν από την αποθήκευσή τους.

- ✓ Το Hadoop έχει ένα απλοποιημένο μοντέλο προγραμματισμού. Το μοντέλο αυτό επιτρέπει στους χρήστες να γράφουν και να δοκιμάζουν σε σύντομο χρόνο κατακευματισμένα συστήματα. Εκτελεί υπολογισμούς σε μεγάλους όγκους δεδομένων συνήθως σε ένα κατακευματισμένο σύστημα όμως υπάρχει μεγάλη δυσκολία να γράφεις κατακευματισμένα συστήματα. Χρησιμοποιώντας όμως τεχνικές ευελιξίας προγραμματισμού, το Hadoop καθιστά πιο εύκολο το να γράψει κατακευματισμένα προγράμματα.
- ✓ Ένα άλλο πλεονέκτημα του Hadoop είναι πως διαχειρίζεται εύκολα. Εναλλακτικά υπολογιστικά συστήματα υψηλών επιδόσεων επιτρέπουν στα προγράμματα να τρέχουν σε μεγάλα σύνολα υπολογιστών, αλλά απαιτούν συνήθως άκαμπτη διαμόρφωση του προγράμματος καθώς επίσης χρειάζεται τα δεδομένα να είναι αποθηκευμένα σε συστήματα SAN (Storage Area Network). Οι χρονοπρογραμματιστές των Clusters απαιτούν πολύ προσεκτική διαχείριση και εφόσον η εκτέλεση προγραμμάτων έχει ευαισθησίες σε περιπτώσεις αποτυχίας των κόμβων, η διαχείριση του Hadoop cluster είναι πολύ πιο εύκολη. Το Hadoop χειρίζεται θέματα ελέγχου και αποτυχιών κόμβων αθόρυβα. Σε περίπτωση που ένας κόμβος τεθεί εκτός λειτουργίας, το Hadoop εξασφαλίζει τη συνέχεια των εργασιών σε κάποιον άλλο κόμβο και τα δεδομένα που είναι αποθηκευμένα στον κόμβο αυτό ανακτώνται από τους υπόλοιπους.
- ✓ Επίσης το Hadoop είναι ευκίνητο. Οι σχεσιακές βάσεις δεδομένων είναι καλές στην αποθήκευση και επεξεργασία των δεδομένων με προκαθορισμένα και άκαμπτα μοντέλα δεδομένων. Για τα αδόμητα δεδομένα, οι σχεσιακές βάσεις υστερούν σε ευελιξία και σε επεκτασιμότητα. Το Apache Hadoop έχει την ιδιότητα να επεξεργάζεται και να αναλύει με φθινό τρόπο μεγάλες ποσότητες δομημένων και αδόμητων δεδομένων μαζί. [6] [15]

2.3.2 Λόγοι για τη χρήση του Hadoop

- ✓ Είναι οικονομικό και αποδοτικό. Το Apache Hadoop ελέγχει τα κόστη αποθηκεύοντας δεδομένα ανά terabyte με πιο οικονομικό τρόπο από άλλες πλατφόρμες. Αντίθετα με τον ανταγωνισμό που χρεώνει 1000 έως 10.000 δολάρια ανά terabyte, το Hadoop προσφέρει επεξεργασία και αποθήκευση με 100 δολάρια ανά terabyte.
- ✓ Είναι ανεκτικό σε σφάλματα. Η ανοχή σε σφάλματα είναι ένα από τα σημαντικότερα πλεονεκτήματα της χρήσης του Hadoop. Ακόμα και όταν οι ανεξάρτητοι κόμβοι αντιμετωπίζουν μεγάλο ποσοστό αποτυχιών εκτελώντας διεργασίες σε ένα μεγάλο σύμπλεγμα υπολογιστών, τα δεδομένα αντιγράφονται σε ένα άλλο σύμπλεγμα έτσι ώστε να μπορούν εύκολα να ανακτηθούν στο δίσκο ή στον κόμβο.
- ✓ Είναι ευέλικτο. Ο ευέλικτος τρόπος που τα δεδομένα είναι αποθηκευμένα στο Apache Hadoop είναι ένα από τα μεγαλύτερα περιουσιακά του στοιχεία, επιτρέποντας στις επιχειρήσεις να μετατρέψουν τα δεδομένα αυτά σε χρήσιμη πληροφορία με έναν τρόπο που παλιότερα θεωρούνταν οικονομικά ασύμφορος διότι έπρεπε να αποθηκευτούν και να τεθούν σε επεξεργασία κάτω από το μοντέλο των παραδοσιακών βάσεων δεδομένων. Με το Hadoop οι χρήστες έχουν τη δυνατότητα να χρησιμοποιήσουν όλους τους τύπους δεδομένων είτε είναι δομημένα είτε είναι αδόμητα και να εξάγουν σημαντικές επιχειρηματικές ιδέες από όλο το σύνολο των δεδομένων που διαθέτει.
- ✓ Είναι επεκτάσιμο. Το Hadoop είναι μια εξαιρετικά επεκτάσιμη πλατφόρμα αποθήκευσης, επειδή μπορεί να αποθηκεύσει και να διανείμει μεγάλα σύνολα δεδομένων σε συστάδες εκατοντάδων φθηνών εξυπηρετητών που λειτουργούν παράλληλα. Το πρόβλημα με τα παραδοσιακά σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων είναι ότι δεν μπορούν να αναβαθμίσουν την επεξεργασία τεράστιων ποσοτήτων δεδομένων. [3]

Πολλές πασίγνωστες εταιρίες που χρησιμοποιούν το Hadoop είναι η Yahoo, Amazon, Facebook, Twitter κ.α.

Κεφάλαιο 3 - NoSQL Databases

3.1 Ορισμός

Οι σχεσιακές βάσεις δεδομένων αποτελούσαν για πολλά χρόνια την κυρίαρχη επιλογή για δικτυακές και επιχειρησιακές εφαρμογές. Λόγω της σχεδόν καθολικής επικράτησής τους μάλιστα, δημιουργήθηκαν ανά τα έτη πολλά εργαλεία που διευκολύνουν και εμπλουτίζουν τη χρήση τους. Το σχεσιακό σχήμα που υλοποιείται με τη χρήση δισδιάστατων πινάκων και κλειδιών, καλύπτει τις ανάγκες για ένα πολύ μεγάλο μέρος εφαρμογών μέχρι τώρα. Με τη ραγδαία ανάπτυξη των Big Data εμφανίστηκε η ανάγκη για πιο αποδοτικό και κατανεμημένο χειρισμό λόγω του ποικιλόμορφου και μεγάλου όγκου πληροφοριών και έτσι πολλοί οργανισμοί ή επιχειρήσεις έστρεψαν το ενδιαφέρον τους σε εναλλακτικούς τρόπους διαχείρισης και αποθήκευσης δεδομένων. Αυτό είχε σαν αποτέλεσμα τη δημιουργία πολλών συστημάτων βάσεων δεδομένων που δεν έχουν τα χαρακτηριστικά του σχεσιακού μοντέλου και πλέον κατέχουν μεγάλο μερίδιο της αγοράς, ιδίως σε διαδικτυακές εφαρμογές. Τα συστήματα αυτά έγιναν γνωστά ως «NoSQL Databases». Αν και ο όρος παραπέμπει σε «Όχι – SQL», θα ήταν ορθότερο να τον αποδώσουμε ως «Όχι μόνο SQL» ή «Μη – Σχεσιακή Βάση».

3.2 Διαφορές NoSQL – SQL

Τα σχεσιακά συστήματα βάσεων δεδομένων παρέχουν πολύ υψηλή απόδοση για αναζητήσεις βασισμένες στις σχέσεις μεταξύ των οντοτήτων, καθώς και για δεδομένα με αυστηρή δομή και ομοιομορφία. Μάλιστα η εύκολη στην κατανόηση μορφή τους και η πληθώρα εφαρμογών που διευκολύνουν τη διαχείριση τους, τα κάνει ιδιαίτερα προσιτά σε αρχάριους ή μη εξειδικευμένους χρήστες. Ωστόσο, τα δεδομένα μας πολλές φορές δεν έχουν αυστηρά καθορισμένη μορφή ή πιθανώς να μας ενδιαφέρει περισσότερο η οριζόντια κλιμάκωση όπως συμβαίνει στα κατανεμημένα συστήματα. Σε τέτοιες περιπτώσεις ενδείκνυται η χρήση NoSQL συστημάτων. Μία τέτοια βάση δεδομένων μπορεί να βρίσκεται διαμοιρασμένη σε μεγάλο αριθμό από επεξεργαστές, ελαφρύνοντας έτσι το φόρτο του καθενός ξεχωριστά, και δίνοντας τη δυνατότητα για διαχείριση τεραστίου όγκου δεδομένων. Ορισμένες NoSQL βάσεις μάλιστα, παρέχουν αυτόματο διαμοιρασμό των δεδομένων στους

κόμβους του συστήματος, ευνοώντας ακόμα περισσότερο τη χρήση κατακευματισμένων συστημάτων και μειώνοντας το φόρτο εργασίας του διαχειριστή. Κάτι τέτοιο είναι μεν εφικτό και στα RDBMS με χρήση clusters αλλά με πολύ μεγαλύτερο κόστος διαχείρισης, και με μικρότερη αποτελεσματικότητα.

Στα σχεσιακά συστήματα τα δεδομένα αποθηκεύονται σε πίνακες με αυστηρά καθορισμένη δομή. Οι στήλες ενός πίνακα καθορίζονται κατά τη δημιουργία του και δε μπορούν να μεταβάλλονται δυναμικά, ανάλογα με το στοιχείο που θέλουμε να εισάγουμε. Στις NoSQL όμως τα δεδομένα ανήκουν σε έγγραφα (από εκεί προέρχεται και ο όρος “Docubases” που χρησιμοποιείται συχνά από πολλούς) και κάθε βάση είναι μια συλλογή από αυτά τα έγγραφα. Έτσι κάθε έγγραφο μπορεί να έχει εντελώς ανεξάρτητη δομή από τα υπόλοιπα. Στην πραγματικότητα βέβαια, είναι προτιμότερο να υπάρχει μια στοιχειώδης δομή, ώστε να διευκολύνει τα ερωτήματα, αλλά ακόμα και έτσι αποφεύγεται η δέσμευση extra χώρου για null values που αποτελεί σημαντικό παράγοντα στις μέρες μας.

Ένα άλλο χαρακτηριστικό των RDBMS που ανατρέπεται από τις NoSQL είναι το μοντέλο ACID (Atomicity – Consistency – Isolation – Durability). Σε μια σχεσιακή βάση, κάθε συναλλαγή είτε θα ολοκληρωθεί επιτυχώς είτε θα αποτύχει πλήρως αφήνοντας τη βάση ανεπηρέαστη. Δεν υπάρχει μερική επιτυχία (Atomicity). Ακόμα, μόνο έγκυρα δεδομένα γράφονται στη βάση, φέρνοντάς την από μια έγκυρη κατάσταση στην επόμενη (Consistency), ενώ συναλλαγές που εκτελούνται ταυτόχρονα δεν επηρεάζουν η μια την άλλη, επιφέροντας έτσι ίδιο αποτέλεσμα με τη σειριακή τους εκτέλεση (Isolation). Τέλος, αφού ολοκληρωθεί με επιτυχία μια συναλλαγή, παραμένουν οι αλλαγές που έχει κάνει στη βάση δεδομένων, ακόμα και αν το σύστημα έχει πρόβλημα (Durability). Στις NoSQL βάσεις η συνέπεια και η απομόνωση θυσιάζονται για χάρη της διαθεσιμότητας, της ανοχής σε σφάλματα και της απόδοσης του συστήματος. Δημιουργείται έτσι το μοντέλο BASE (Basically Available, Soft State, Eventually Consistent) στο οποίο στηρίζονται οι NoSQL βάσεις. Σύμφωνα με αυτό, ένα σύστημα δουλεύει διαρκώς (Basically Available), χωρίς αναγκαστικά να είναι διαρκώς συνεπές (Soft State), αλλά θα είναι εντέλει σε κάποια γνωστή κατάσταση (Eventually Consistent). Οι συναλλαγές επομένως γίνονται πιο γρήγορα, αφού δεν υπόκεινται στους περιορισμούς του ACID, αλλά με μικρότερη αξιοπιστία και συνέπεια.

ACID vs. BASE

ACID

- ☐ Strong Consistency
- ☐ Isolation
- ☐ Focus on "commit"
- ☐ Nested transactions
- ☐ Less Availability
- ☐ Conservative (pessimistic)
- ☐ Difficult evolution (e.g. schema)

BASE

- ☐ Weak Consistency
- ☐ Availability first
- ☐ Best effort
- ☐ Approximated answers
- ☐ Aggressive (optimistic)
- ☐ Simpler!
- ☐ Faster
- ☐ Easier evolution

출처: Brewer

22

Διάγραμμα 3 - Διαφορές ACID & BASE

Στο παραπάνω διάγραμμα βλέπουμε καταγεγραμμένες τις διαφορές των ACID και Base μοντέλων διαχείρισης βάσεων δεδομένων. [5] [6]

3.3 Τύποι NoSQL Βάσεων Δεδομένων

Μεταξύ των διαφόρων συστημάτων διαχείρισης μη σχεσιακών βάσεων δεδομένων παρατηρούνται διαφορές ως προς τον τρόπο αποθήκευσης και οργάνωσης των δεδομένων τους. Χωρίζονται λοιπόν βάσει του μοντέλου δεδομένων τους σε τέσσερις κύριες κατηγορίες:

- Key-Value Stores
- Column Family Stores
- Document Databases
- Graph Databases

Στις Key-Value Stores κυριαρχεί η ιδέα της ύπαρξης ενός hash table όπου ο χρήστης αποθηκεύει δεδομένα-τιμές (Values) τα οποία δεικτοδοτούνται από κάποιο μοναδικό κλειδί. Δεν είναι απαραίτητη η ύπαρξη ενός αυστηρού σχήματος για τα δεδομένα που αποθηκεύονται, ενώ για την επίτευξη μεγαλύτερης κλιμακωσιμότητας και απόδοσης του συστήματος γίνεται χρήση μηχανισμών caching και δεν υποστηρίζονται συνενώσεις (Joins) και συναθροιστικές (Aggregate) λειτουργίες. Τα κυριότερα τέτοια συστήματα είναι τα Amazon Dynamo, Riak, Project Voldemort, Redis κ.α.

Στις Column Family Stores υπερισχύει η ιδέα της αποθήκευσης και επεξεργασίας δεδομένων κατά στήλη και όχι κατά γραμμή. Κύριος στόχος τους είναι να μπορούν να διαχειριστούν πολύ μεγάλο όγκο δεδομένων που βρίσκονται κατανεμημένα σε διάφορους εξυπηρετητές. Και εδώ γίνεται χρήση κλειδιών, που δείχνουν όμως σε μια ή περισσότερες Column Families. Τυπικά παραδείγματα τέτοιων συστημάτων είναι τα BigTable, Hbase, Accumulo κ.α.

Οι Document Databases μοιάζουν με τις Key-Value Stores αλλά υποστηρίζουν πιο πολύπλοκα δεδομένα. Τα δεδομένα αυτά δεν έχουν αυστηρή δομή και έτσι οι χρήστες έχουν την δυνατότητα να προσθέτουν και να αφαιρούν πεδία κατά βούληση. Όπως και στις Key-Value Stores, κάθε έγγραφο δεικτοδοτείται από ένα κλειδί με το οποίο μπορεί ο χρήστης να

το ανακτήσει, και επιπλέον στις περισσότερες Document Databases υπάρχει διαθέσιμο API ή Query Language ώστε να διευκολύνεται η ανάκτηση εγγράφων με βάση το περιεχόμενό τους. Οι πιο δημοφιλείς Document Databases είναι οι MongoDB, CouchDB, Cassandra κ.α.

Στις Graph Databases γίνεται χρήση κόμβων και ακμών για την αναπαράσταση και αποθήκευση των δεδομένων. Οι γράφοι δηλαδή αντικαθιστούν το κλασσικό μοντέλο των πινάκων. Το μοντέλο γραφημάτων που ακολουθείται ευνοεί την οριζόντια κλιμάκωση αφού μπορεί να χρησιμοποιηθεί παράλληλα σε πολλούς εξυπηρετητές. [10]

Κεφάλαιο 4 - Παραδείγματα Βάσεων NoSQL

4.1 MongoDB

Η MongoDB (από το «humongous» = τεράστιος) είναι η πιο διαδεδομένη Document Database και ίσως και η πιο διαδεδομένη NoSQL Database γενικότερα. Αναπτύχθηκε από την εταιρία 10gen στα πρότυπα του ανοιχτού κώδικα και δόθηκε για πρώτη φορά σε κυκλοφορία το 2009. Είναι υλοποιημένη στη γλώσσα προγραμματισμού C++, και κύριος στόχος της είναι η επίτευξη υψηλής ταχύτητας και κλιμακωσιμότητας. Χρησιμοποιείται στο back-end πολλών μεγάλων υπηρεσιών όπως Craigslist, eBay, Foursquare, SourceForge, κ.α., ακριβώς λόγω της πολύ υψηλής δυνατότητας για κλιμάκωση που προσφέρει. Ο προκαθορισμένος τρόπος για τη διαχείριση MongoDB βάσεων είναι μέσω του διαδραστικού JavaScript Shell που παρέχει. Επιπλέον, σε πολλές γλώσσες προγραμματισμού (C, C#, C++, Haskell, Java, JavaScript, Perl, PHP, Python και Ruby) έχουν αναπτυχθεί βιβλιοθήκες και Drivers για την υποστήριξή της.

Στη MongoDB ακολουθείται η νοοτροπία data as documents. Τα έγγραφα που αποθηκεύονται είναι στη μορφή BSON αντικειμένων, δηλαδή δυαδικά κωδικοποιημένα JSON αντικείμενα. Τα BSON είναι σχεδιασμένα έτσι ώστε να γίνεται εύκολα η διάσχιση και η ανάλυσή τους. Οι χρήστες εισάγουν τα δεδομένα τους σε μορφή JSON, τα οποία στη συνέχεια μετατρέπονται σε BSON ώστε να αποθηκευτούν στη βάση, και αντίστοιχα όταν παίρνουν κάποιο BSON αντικείμενο από τη βάση, το βλέπουν σε JSON μορφή. Δηλαδή, οι χρήστες δε βλέπουν καθόλου τη BSON μορφή, χρησιμοποιείται μόνο εσωτερικά στη βάση. Ένα JSON document είναι μηδέν ή περισσότερα ζεύγη κλειδιού – τιμής (key-value pairs). Σαν κλειδί αποθηκεύεται το όνομα του πεδίου ως συμβολοσειρά και σαν τιμή, μια τιμή ενός από τους υποστηριζόμενους JSON data types (Number, String, Boolean, Array, Object, Whitespace, null, ή JSON value) είτε κάποιος άλλος όπως ημερομηνία. Το μέγιστο μέγεθος ενός εγγράφου είναι 16MB, ενώ για μεγαλύτερα αρχεία χρησιμοποιείται το GridFS. Τα αντικείμενα αυτά είναι schemaless, που σημαίνει πως δεν υπάρχει κάποια ομαδοποίηση μεταξύ εγγράφων που περιέχουν ίδια κλειδιά, κάτι που συναντάμε στο σχεσιακό μοντέλο, όπου όλες οι εγγραφές ενός πίνακα έχουν την ίδια δομή. Αντιθέτως, έγγραφα με παρόμοια δομή που περιλαμβάνουν δεδομένα σχετικά με το ίδιο αντικείμενο, αλλά με διαφορετικά ζεύγη κλειδιού – τιμής, αποθηκεύονται μαζί, σε μια “συλλογή”.

Το πλεονέκτημα των NoSQL που έχει να κάνει με τη μη ύπαρξη αυστηρά καθορισμένου σχήματος των δεδομένων, φαίνεται και στη MongoDB αφού υπάρχει απόλυτη ελευθερία ως προς τα πεδία, τη δομή και τους τύπους δεδομένων κάθε εγγράφου μιας συλλογής. Από την άλλη όμως, ένα πολύ βασικό στοιχείο είναι πως δεν υπάρχει δυνατότητα για joins μεταξύ συλλογών όπως στις σχεσιακές βάσεις 43 δεδομένων. Έτσι, ο σχεδιαστής καλείται να διαμορφώσει το κάθε έγγραφο ξεχωριστά, με τέτοιο τρόπο ώστε να ελαχιστοποιεί τον απαιτούμενο αποθηκευτικό χώρο αλλά και να διευκολύνει τις αναζητήσεις του, μοντελοποιώντας τα δεδομένα με γνώμονα τον τρόπο χρήσης τους και όχι τον τρόπο αποθήκευσης.

Επίσης σύμφωνα με την αρχιτεκτονική της βλέπουμε πως η MongoDB έχει ως κύριο στόχο την υποστήριξη εφαρμογών που τρέχουν σε κατανεμημένα συστήματα – clusters υπολογιστών. Έτσι, παίζει πολύ μεγάλο ρόλο η οριζόντια κλιμακωσιμότητα και για αυτό πραγματοποιείται αυτόματη κατανομή του φόρτου εργασίας (auto sharding) μεταξύ των κόμβων του cluster. [10]

4.2 BigTable

Το σύστημα διαχείρισης βάσεων δεδομένων BigTable αποτελεί ένα κατακευκαμένο σύστημα αποθήκευσης για τη διαχείριση μεγάλης ποσότητας ημιδομημένων δεδομένων . Ανήκει στις column-oriented βάσεις και σχεδιάστηκε και χρησιμοποιείται από την Google για την επεξεργασία τεραστίων όγκων δεδομένων (της τάξης των πολλών Terabyte) για τις γνωστές εφαρμογές της Google όπως είναι το Google Maps, το Google Earth, το Google Analytics, κ.α. Σχεδιάστηκε ώστε να καλύψει ένα εύρος απαιτήσεων από τις υπηρεσίες που το χρησιμοποιούν. Από αποθηκευτικό χώρο για τις πληροφορίες ιστοσελίδων ή δορυφορικές φωτογραφίες, μέχρι και εξυπηρέτηση αιτημάτων όπου ο χρόνος απόκρισης παίζει σημαντικό ρόλο όπως για παράδειγμα εφαρμογές πραγματικού χρόνου.

Κάποια από τα βασικά χαρακτηριστικά BigTable είναι:

- ❖ Μεγάλο εύρος εφαρμογών
- ❖ Batch processing εφαρμογές
- ❖ Εφαρμογές χαμηλής καθυστέρησης για χρήστες
- ❖ Κλιμακωσιμότητα
- ❖ Υψηλή απόδοση
- ❖ Υψηλή διαθεσιμότητα
- ❖ Δυνατότητα χρήσης σε συνδυασμό με MapReduce
- ❖ Εκτελείται σε μέσου κόστους υλικό

Το μοντέλο δεδομένων αυτό είναι στην ουσία ένας κατακευκαμένος πολυδιάστατος πίνακας. Διευθυνσιοδοτείται από το κλειδί γραμμής, το κλειδί στήλης και την χρονοσφραγίδα (timestamp) όπου λειτουργούν σαν συντεταγμένες. Κάθε κελί περιέχει ένα σύνολο bytes. (row, column, time) Value.

Χαρακτηριστικά των γραμμών (rows):

- ❖ Το κλειδί αποτελείται από ένα αλφαριθμητικό
- ❖ Οι ενέργειες πάνω σε μια γραμμή είναι ατομικές
- ❖ Λεξικογραφική ταξινόμηση με βάση τα κλειδιά

- ❖ Όλος ο πίνακας αποτελείται από εκατομμύρια λεξικογραφικά ταξινομημένες γραμμές.
- ❖ Το row key είναι το μόνο πεδίο που γίνεται indexed στον BigTable πίνακα, ενώ η αναζήτηση σε όλα τα άλλα πεδία γίνεται με full table scan.

Χαρακτηριστικά των στηλών (columns):

- ❖ Ομαδοποίηση σε column οικογένειες
- ❖ Σπάσιμο σε column οικογένειες ανάλογα με την εφαρμογή
- ❖ Μικρός αριθμός από column οικογένειες
- ❖ Άπειρος αριθμός από columns
- ❖ Ο έλεγχος πρόσβασης γίνεται με βάση τις column οικογένειες

Χαρακτηριστικά χρονοσφραγίδων (timestamps):

- ❖ Πολλαπλές εκδόσεις των ίδιων δεδομένων
- ❖ Πραγματικός χρόνος ή καθορισμένος από τον χρήστη
- ❖ Οι πιο πρόσφατες εκδόσεις είναι ευκολότερο προσβάσιμες

4.3 Apache HBase

Η Hbase είναι μια μη σχεσιακή διαδεδωμένη βάση δεδομένων που αναπτύχθηκε από την εταιρία Apache Software Foundation ως μια ανοιχτού κώδικα έκδοση της BigTable. Είναι γραμμένη σε Java και ανήκει στην οικογένεια των column-oriented συστημάτων. Τα column-oriented συστήματα αποθήκευσης έχουν κερδίσει την προσοχή των προγραμματιστών τα τελευταία χρόνια τόσο σε ερευνητικό όσο και σε βιομηχανικό επίπεδο, λόγω της υποστήριξης που προσφέρουν για μεγάλη κλιμακωσιμότητα των δεδομένων, αποδοτική πρόσβαση στα δεδομένα και υψηλή ανοχή σφαλμάτων. Τα δεδομένα οργανώνονται εσωτερικά στη βάση ως εμφωλευμένα ζεύγη κλειδιού-τιμής και παρουσιάζονται εξωτερικά στο χρήστη ως αραιοί πίνακες. Κάθε γραμμή σε αυτούς τους πίνακες αντιστοιχεί σε ένα σύνολο εμφωλευμένων ζευγών κλειδιού-τιμής που δεικτοδοτούνται σε ανώτερο επίπεδο από το ίδιο κλειδί (το κλειδί της γραμμής, “row key”). Σε δεύτερο επίπεδο το κλειδί είναι το “column family” και σε τρίτο το “column qualifier”). Κάθε στήλη σε μια γραμμή αντιστοιχεί στην τιμή (αποθηκευμένη ως πίνακας από bytes) που δεικτοδοτείται από το συνδυασμό του column family και του column qualifier. Στην Hbase τα δεδομένα αποθηκεύονται σε αραιούς πίνακες, δηλαδή πίνακες με γραμμές που έχουν μεταβλητό αριθμό στηλών. Κάθε γραμμή έχει ένα μοναδικό row key με βάση το οποίο ταξινομούνται αυτόματα. Οι στήλες ομαδοποιούνται σε οικογένειες στηλών. Τα δεδομένα της ίδιας οικογένειας στηλών αποθηκεύονται σε φυσικό επίπεδο, κοντά στο δίσκο ώστε να επιτυγχάνεται βελτίωση της απόδοσης των queries. Για την υποστήριξη πολλαπλών εκδόσεων των δεδομένων χρησιμοποιούνται timestamps. Αυτές ορίζονται άμεσα από το χρήστη όταν εισάγει τα δεδομένα, είτε αυτόματα από το σύστημα, και σε συνδυασμό με το row key και τη στήλη, προσδιορίζουν μοναδικά τα δεδομένα.

Στις column oriented βάσεις η επεκτασιμότητα επιτυγχάνεται με το rangepartitioning που γίνεται στα δεδομένα με βάση το row key, σε τμήματα ίσου μεγέθους που δε μοιράζονται καθόλου δεδομένα. Αυτά τα τμήματα (partitions) που προκύπτουν, αποστέλλονται σε διάφορους κατανεμημένους εξυπηρετητές. Όσο αυξάνεται ο όγκος των δεδομένων, τόσο περισσότερα partitions δημιουργούνται. Θεωρητικά λοιπόν, η ικανότητα αποθήκευσης των δεδομένων κλιμακώνεται τόσο όσο αυξάνονται οι διαθέσιμοι εξυπηρετητές

Η HBase χρησιμοποιεί μια master-slave τοπολογία. Τα partitions λοιπόν που εδώ ονομάζονται “regions” αποθηκεύονται στους slaves, οι οποίοι ονομάζονται “region servers”. Χρησιμοποιούνται πολλαπλοί masters ώστε να αποφεύγεται το πρόβλημα του single point of

failure. Όταν ένας region server τίθεται εκτός λειτουργίας, τα δεδομένα του ανακτώνται από το file system (HDFS) και αποστέλλονται σε άλλον εξυπηρετητή. Μια πιθανή συμφύρση που μπορεί να δημιουργηθεί όμως λόγω αυτής της αρχιτεκτονικής είναι όταν έχουμε πολλαπλά αιτήματα για δεδομένα στο ίδιο εύρος row key. Έτσι, όλα αυτά τα αιτήματα πάνε στον ίδιο εξυπηρετητή και πιθανόν να έχουμε υπερφόρτωση. Επί του παρόντος, η Hbase υποστηρίζει απλά queries με βάση το row key και τα timestamps, δεν υποστηρίζει όμως πολύπλοκα ερωτήματα, join λειτουργίες, συναλλαγές και δευτερεύοντες δείκτες. Μπορούμε επιπλέον να ανακτήσουμε και να 50 διατρέξουμε ένα σύνολο από στήλες, σε ένα δεδομένο εύρος γραμμών.

Αν και το query model φαίνεται αρκετά απλοϊκό και πιθανόν ανεπαρκές, αν οι πίνακες δομηθούν σωστά, επιτρέπουν την επίλυση αρκετών σύνθετων προβλημάτων, ενώ για ακόμα πιο πολύπλοκες λειτουργίες μπορούμε να χρησιμοποιήσουμε εργασίες map – reduce.[10] [5]

4.4 Apache Cassandra

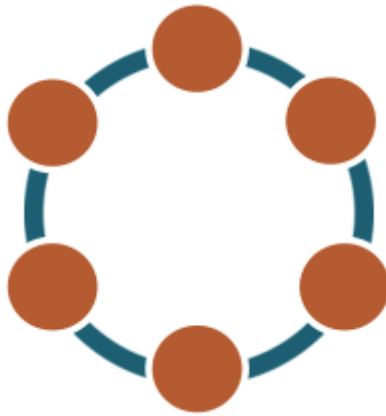
Το Apache Cassandra είναι ένα ανοιχτού κώδικα κατανεμημένο σύστημα διαχείρισης βάσεων δεδομένων σχεδιασμένο για να χειρίζεται μεγάλες ποσότητες δεδομένων σε πολλούς διακομιστές (Servers) ταυτόχρονα, παρέχοντας υψηλή διαθεσιμότητα χωρίς το ενδεχόμενο αποτυχίας. Το Cassandra προσφέρει ισχυρή υποστήριξη σε συστοιχίες υπολογιστών (Clusters) που εκτείνονται σε πολλαπλά κέντρα δεδομένων.

Επίσης, το Cassandra δίνει μεγάλη αξία στην απόδοση. Το 2012, στο πανεπιστήμιο του Toronto, ερευνητές που μελετούσαν στα συστήματα NoSQL κατέληξαν στο συμπέρασμα ότι στον τομέα της επεκτασιμότητας σε όλα τα πειράματα καλύτερο σύστημα αναδείχθηκε το Cassandra καθώς επιτύγχανε την υψηλότερη απόδοση για τον μέγιστο αριθμό των κόμβων.

Προσφέρει δυνατότητες που άλλες NoSQL βάσεις ή σχεσιακές βάσεις αδυνατούν να προσφέρουν όπως:

- Συνεχή διαθεσιμότητα
- Λειτουργική απλότητα
- Γραμμική κλίμακα επιδόσεων
- Εύκολη διανομή δεδομένων σε πολλαπλά κέντρα δεδομένων και σε ζώνες που συνδέονται σε Cloud.

Η αρχιτεκτονική του Cassandra είναι διαφορετική από άλλες γνωστές NoSQL βάσεις δεδομένων που χρησιμοποιούν το μοντέλο Master – Slave σύμφωνα με το οποίο μια συσκευή ή διαδικασία (γνωστή ως master) έχει τον πλήρη έλεγχο μιας ή περισσότερων άλλων συσκευών/διαδικασιών (γνωστών ως slaves). Σε μερικά συστήματα μια master συσκευή εκλέγεται μέσα από ένα σύνολο συσκευών όπου οι υπόλοιπες θα έχουν τον ρόλο slave. Μόλις καθιερωθεί η σχέση μεταξύ τους, η κατεύθυνση του ελέγχου θα είναι πάντα από την master στην/στις slave. Αντ' αυτού, το Cassandra χρησιμοποιεί ένα μοντέλο «δαχτυλίδι» (Ring) το οποίο έχει προτερήματα την κομψότητα, την εύκολη εγκατάσταση και συντήρηση. Σε αυτό το μοντέλο όλοι οι κόμβοι παίζουν ένα πανομοιότυπο ρόλο χωρίς να υπάρχει ο master και οι slaves και όλοι οι κόμβοι επικοινωνούν μεταξύ τους εξίσου.



Διάγραμμα 4 - : Η αρχιτεκτονική «Δαχτυλιδιού» του Cassandra

Η αρχιτεκτονική δομή Ring του Cassandra δηλώνει πως είναι σχεδιασμένο για να χειρίζεται μεγάλου όγκου δεδομένα, να εξυπηρετεί χιλιάδες χρήστες και να εκτελεί επίσης μεγάλο αριθμό λειτουργιών ανά δευτερόλεπτο ακόμα και σε πολλαπλά συστήματα δεδομένων. Χρειάζεται επίσης να αναφέρουμε πως για την εισαγωγή νέων κόμβων στο cluster, δεν χρειάζεται να απενεργοποιηθεί το σύστημα.

Βασικά Πλεονεκτήματα και Οφέλη του Apache Cassandra

Το Cassandra παρέχει ένα μεγάλο από λειτουργίες και οφέλη για όσους χρήστες επιθυμούν να το χρησιμοποιήσουν ως μια υποκείμενη βάση δεδομένων για σύγχρονες διαδικτυακές εφαρμογές. Συγκεκριμένα το Cassandra παρέχει:

- Τεράστια επεκτάσιμη αρχιτεκτονική: Χρησιμοποιεί masterless (χωρίς το μοντέλο master-slave) σχεδιασμό όπου όλοι οι κόμβοι είναι ίδιοι χωρίς διαφοροποιήσεις στους ρόλους.
- Active Everywhere σχεδιασμός: Δηλαδή υπάρχει η δυνατότητα ανάγνωσης και εγγραφής σε όλους τους κόμβους.
- Γραμμική κλιμάκωση απόδοσης: Υπάρχει η δυνατότητα να προστεθούν νέοι κόμβοι στο cluster χωρίς να διακοπούν ενεργές εργασίες ή λειτουργίες αφού δεν υπάρχει ανάγκη για απενεργοποίηση του συστήματος.

- Συνεχής διαθεσιμότητα: Εξαλείφει δεδομένα ή και λειτουργίες κόμβων που μπορεί να επιφέρουν σφάλματα και παρέχει σταθερή λειτουργία.
- Διαφανής ανίχνευση βλάβης και αποκατάσταση: Κόμβοι οι οποίοι δημιουργούν προβλήματα κατά την λειτουργία τους μπορούν εύκολα να αποκατασταθούν ή και να αντικατασταθούν.
- Ευέλικτο και δυναμικό μοντέλο δεδομένων: Υποστηρίζει σύγχρονους τύπους δεδομένων με τις λειτουργίες εγγραφής και ανάγνωσης να γίνονται σε πραγματικό χρόνο.
- Ισχυρή προστασία δεδομένων: Εμπλουτισμένο με ένα μοντέλο αποθήκευσης ιστορικού (Log) όπου εξασφαλίζει πως δεν θα υπάρξει απώλεια δεδομένων. Επίσης διαθέτει ενσωματωμένο σύστημα Backup & Restore (Διατήρηση αντιγράφου & Αποκατάσταση) που κρατάει τα δεδομένα ασφαλή.
- Δυνατότητα συντονισμού της συνοχής των δεδομένων: Παρέχει υποστήριξη για ισχυρή ή ενδεχόμενη συνοχή των δεδομένων σε ένα ευρέως κατανομημένο σύμπλεγμα υπολογιστών (Cluster).
- Αντιγραφή πολλαπλών κέντρων δεδομένων: Αποστολή των κέντρων δεδομένων σε Cloud με δυνατότητα εγγραφής & ανάγνωσης.
- Συμπίεση δεδομένων: Τα δεδομένα συμπιέζονται έως και 80% του αρχικού μεγέθους τους χωρίς αυτό να επιβαρύνει την απόδοση του συστήματος.
- CQL (Cassandra Query Language): Μια γλώσσα παρόμοιας λειτουργίας με την SQL η οποία κάνει την μεταφορά δεδομένων από μια σχεσιακή βάση δεδομένων σε μια μη-σχεσιακή πολύ εύκολη.

Σύμφωνα με την ιστορική αναδρομή, το Apache Cassandra αναπτύχθηκε στα εργαστήρια της Facebook με σκοπό να εμπλουτίσουν την διαδικασία αναζήτησης στον φάκελο εισερχομένων μηνυμάτων. Επικεφαλείς του project ήταν ο Avinash Lakshman (ένας από τους δυο εκδότες του Amazon Dynamo) και ο Prashant Malik. Πρωτοκυκλοφόρησε σαν έργο ανοιχτού κώδικα στο Google Code τον Ιούλιο του 2008 και το 2010 εξελίχθηκε σε κορυφαίου επιπέδου project για την εταιρία Apache.

Πλέον πολλές γνωστές εταιρίες κατάφεραν να αναπτύξουν με επιτυχία και να επωφεληθούν από το Apache Cassandra, κάποιες απ' τις οποίες είναι : η Apple, η Comcast, το eBay, το Instagram, το Netflix, η Spotify, κ.α. [16]

Κεφάλαιο 5 - Τεχνολογίες Data - Warehousing Columnar Data - stores

5.1 Ορισμός

Μια βάση δεδομένων μεγάλων διαστάσεων γνωστή ως αποθήκη δεδομένων, με προκαθορισμένες λειτουργίες, όπου ο χρήστης έχει μόνο read-only πρόσβαση, λέγεται Αποθήκη Δεδομένων (Data Warehouse). Μια αποθήκη δεδομένων μπορεί να δεχτεί δεδομένα από πολλά συστήματα διαφορετικών τύπων και τα αποθηκεύει σε ένα κοινό κεφάλαιο το οποίο ο χρήστης μπορεί να αναλύσει με μεγάλη ποικιλία τρόπων ώστε να αποκτήσει τις κατάλληλες πληροφορίες. [5]

Μια αποθήκη δεδομένων διατηρεί τα δεδομένα που αντλεί από τις βάσεις δεδομένων των πληροφοριακών συστημάτων μιας επιχείρησης αλλά και άλλες πηγές δεδομένων, όπως αρχεία της επιχείρησης ή δεδομένα που προέρχονται από εξωτερικές πηγές. Όλα αυτά τα δεδομένα αποθηκεύονται και οργανώνονται στην Αποθήκη Δεδομένων σε δομές, κατάλληλες να ικανοποιήσουν αποδοτικά τις απαιτήσεις των αναλυτών συστημάτων στήριξης αποφάσεων.

Οι αποθήκες δεδομένων έχουν προσελκύσει ευρύτατο ενδιαφέρον αφού μπορούν να αποτελέσουν λύση για μια πολύ μεγάλη ομάδα προβλημάτων, όπως ολοκλήρωση ετερογενών πηγών πληροφοριών και πλάγια αντιμετώπιση των γνωστών. Συγχρόνως υπάρχει τεράστια ποικιλία προϊόντων στρατηγικής των προμηθευτών, τεράστια ποικιλία εφαρμογών και των απαιτήσεων τους και οι προμηθευτές παραδέχονται ότι κάθε αποθήκη δεδομένων χρειάζεται ειδική κατασκευή και βελτιστοποίηση κατανόησης απαιτήσεων.

Χαρακτηριστικά μιας Αποθήκης Δεδομένων

- Subject Oriented (Θεματοστρεφή)

Οι Αποθήκες Δεδομένων είναι σχεδιασμένες ώστε να βοηθήσουν στην ανάλυση των δεδομένων. Για παράδειγμα, για να μάθουμε περισσότερα για τις πωλήσεις της εταιρίας μας, μπορούμε να δημιουργήσουμε μια Αποθήκη Δεδομένων που επικεντρώνεται στις πωλήσεις. Χρησιμοποιώντας αυτή τη ΑΔ, μπορούμε να μάθουμε πολλά χρήσιμα στοιχεία για την πορεία της εταιρίας μας όπως για παράδειγμα ποιός

ήταν ο καλύτερος πελάτης μας για κάποιο προϊόν φέτος. Αυτή η ικανότητα της Αποθήκης δεδομένων την κάνει προσανατολισμένη στο θέμα.

- Integrated (Ολοκληρωμένα)

Η ολοκλήρωση είναι στενά συνδεδεμένη με τον προσανατολισμό. Οι αποθήκες δεδομένων πρέπει να τοποθετήσουν τα δεδομένα από διαφορετικές πηγές σε ένα συνεχές ανομοιογενές σχήμα. Πρέπει να επιλύσουν προβλήματα όπως η ασυνέπεια μεταξύ των μονάδων αυτών. Όταν το επιτυγχάνουν αυτό τότε λέμε ότι ολοκληρώθηκαν.

- Nonvolatile (Σταθερότητα)

Σταθερότητα σημαίνει ότι από τη στιγμή που τα δεδομένα εισήχθησαν μια φορά στην αποθήκη δεδομένων δεν πρέπει να αλλάξουν. Αυτό είναι λογικό επειδή ο σκοπός μιας ΑΔ είναι να επιτρέψει την ανάλυση.

5.2 Διαφορές μεταξύ Αποθήκης Δεδομένων και άλλων Βάσεων Δεδομένων

Οι αποθήκες δεδομένων περιέχουν δεδομένα συγκεντρωμένα από πολλές πηγές, που συμπληρώνονται με συνοπτική πληροφορία και καλύπτουν μεγάλες χρονικές περιόδους. Όπως αναφέραμε, οι αποθήκες δεδομένων είναι και αυτές βάσεις όμως είναι πολύ μεγαλύτερες. Τα μεγέθη τους συνήθως κυμαίνονται από μερικά terabytes έως pentabytes. Τα συνήθη φορτία εργασιών περιλαμβάνουν απρόβλεπτα, αρκετά σύνθετα αιτήματα ενώ είναι σημαντικοί οι άμεσοι χρόνοι απόκρισης. Αυτά τα χαρακτηριστικά είναι που κάνουν την διαφορά των εφαρμογών αποθηκών δεδομένων από τις εφαρμογές απευθείας επεξεργασίας συναλλαγών και πρέπει για να επιτευχθούν αποτελέσματα ικανοποιητικά να χρησιμοποιηθεί διαφορετικός σχεδιασμός του DBMS καθώς και διαφορετικές τεχνικές υλοποίησης. Για πολύ μεγάλες Αποθήκες Δεδομένων απαιτείται ένα καταναεμημένο DBMS με καλή κλιμάκωση και υψηλή διαθεσιμότητα.

5.3 Τι κάνουν οι Αποθήκες Δεδομένων

Οι βασικές λειτουργίες-ιδιότητες μιας Αποθήκης Δεδομένων είναι:

- Χρησιμοποιείται συχνά ως βάση σε Συστήματα Στήριξης Αποφάσεων (DSS)
- Χρησιμοποιείται για συλλογή πληροφοριών.
- Συνδυάζει όλες τις πληροφορίες που κατέχει από διάφορες πηγές (OLTP, Spreadsheets, web, κείμενα κ.λπ.) τα ελέγχει για ακρίβεια και τα οργανώνει έτσι ώστε οι χρήστες να ανακαλύπτουν εύκολα αυτό που επιθυμούν.
- Συχνά οι Αποθήκες Δεδομένων τμηματοποιούνται σε άλλες μικρότερες εξειδικευμένες Αποθήκες Δεδομένων τις λεγόμενες Data Marts.

Από μια Αποθήκη Δεδομένων δεν διαγράφουμε ποτέ τίποτα και αυτό έχει ως αποτέλεσμα να συγκεντρώνεται ένας μεγάλος όγκος πληροφοριών που αποτελούν ιστορικά στοιχεία του οργανισμού - επιχείρησης που κατέχει την Αποθήκη Δεδομένων.

Οι Αποθήκες Data Marts είναι συχνά προτιμότερες καθώς έχουν μικρότερο κόστος υλοποίησης και χρειάζονται μικρότερο χρονικό διάστημα για να τις κατασκευάσουμε. Η αξία της Αποθήκης Δεδομένων είναι τελικά στην δυνατότητα ανάλυσης που παρέχει. Τα δεδομένα σε μια Αποθήκη Δεδομένων προσπελούνται και αναλύονται χρησιμοποιώντας διάφορα εργαλεία, συμπεριλαμβανομένων των μηχανών αιτημάτων OLAP, αλγορίθμων εξόρυξης δεδομένων, εργαλείων οπτικοποίησης πληροφορίας, στατιστικά πακέτα και γεννήτριες εκτυπώσεων.

Σκοπός των αποθηκών δεδομένων είναι η άμεση, αποδοτική και ακριβής απάντηση οποιαδήποτε ερώτησης σχετικά με τη λειτουργία του οργανισμού στον οποίο λειτουργεί. Συνεπώς, η σχεδίαση των Αποθηκών Δεδομένων έχει σαν στόχο την αποδοτική απάντηση των πολύπλοκων ερωτήσεων που θέτονται κατά την αναλυτική επεξεργασία δεδομένων από τις εφαρμογές στρατηγικού σχεδιασμού. [6]

5.4 Οφέλη της Αποθήκης Δεδομένων

Η επιτυχής υλοποίηση μιας αποθήκης δεδομένων μπορεί να προσφέρει σημαντικά οφέλη σε έναν οργανισμό όπως:

- Πιθανή υψηλή απόδοση της επένδυσης

Ένας αριθμός πρέπει να δεσμεύσει ένα τεράστιο αριθμό από πόρους ώστε να εξασφαλίσει την επιτυχή υλοποίηση μιας αποθήκης δεδομένων και το κόστος μπορεί να ποικίλει τρομερά, από 50.000 αγγλικές λίρες, έως πάνω από 10 εκατομμύρια αγγλικές λίρες, λόγω της ποικιλίας των διαθέσιμων τεχνικών λύσεων. Ωστόσο, μια έρευνα από την εταιρεία International Data Corporation (IDC) το 1996 έδειξε ότι η απόδοση της επένδυσης που πραγματοποιήθηκε σε μια αποθηκοποίηση δεδομένων έφτασε το 401% εντός μια περιόδου 3 ετών κατά μέσο όρο, με πάνω από το 90% που συμμετείχαν στην έρευνα να επιτυγχάνουν απόδοση της επένδυσης πάνω από 40%, οι μισές εταιρείες να επιτυγχάνουν πάνω από 160% και το ένα τέταρτο των εταιρειών να επιτυγχάνουν πάνω από 600%.

- Ανταγωνιστικό πλεονέκτημα

Η τεράστια απόδοση της επένδυσης των εταιρειών αυτών που έχουν υλοποιηθεί επιτυχώς. Μια αποθήκη δεδομένων αποτελεί μια απόδειξη του τεράστιου ανταγωνιστικού πλεονεκτήματος που συνοδεύει την τεχνολογία αυτή. Το ανταγωνιστικό πλεονέκτημα επιτυγχάνεται μέσα από τη δυνατότητα που παρέχεται στους λήπτες αποφάσεων να προσπελαίνουν δεδομένα τα οποία μπορούν να αποκαλύψουν πληροφορίες που προηγουμένως ήταν μη διαθέσιμες, άγνωστες και απροσπέλαστες σχετικά, για παράδειγμα, με πελάτες, τάσεις και απαιτήσεις.

- Αυξημένη παραγωγικότητα από τους λήπτες αποφάσεων της εταιρείας.

Μία αποθήκη δεδομένων αυξάνει την παραγωγικότητα των στελεχών μιας εταιρείας που είναι υπεύθυνα για τη λήψη αποφάσεων μέσα από την παροχή μιας ολοκληρωμένης βάσης δεδομένων από συνεπείς, θεματοστρεφείς, ιστορικές πληροφορίες. Μια αποθηκοποίηση δεδομένων ενοποιεί δεδομένα από πολλαπλά ασύμβατα συστήματα σε μια μορφή που παρέχει μια συνεπή εικόνα της εταιρείας. Με το να μετασχηματίζει τα δεδομένα σε χρήσιμες πληροφορίες, μια αποθήκη δεδομένων

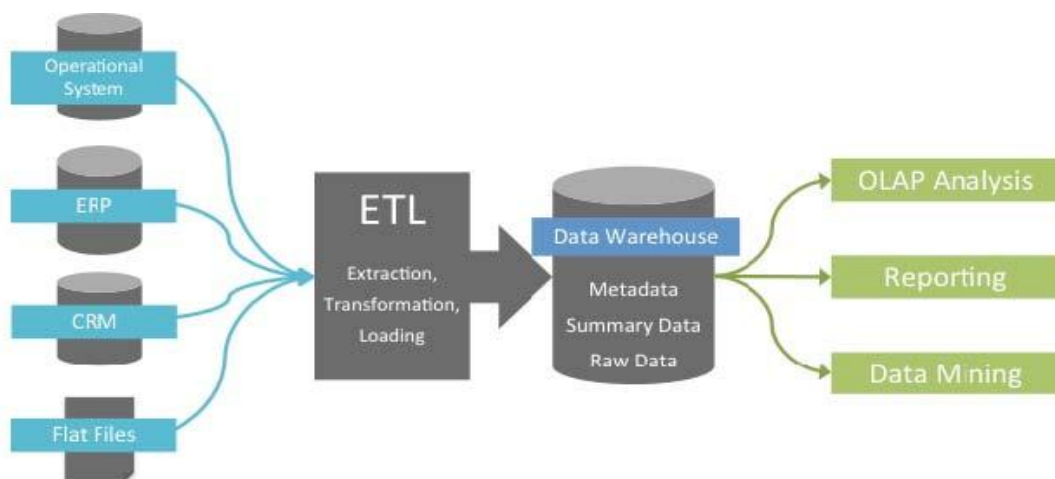
επιτρέπει στους διευθυντές της εταιρείας να επιτελούν πιο ουσιαστικές, σωστές και συνεπείς αναλύσεις.

5.5 Διαδικασία ETL

Λόγω της ύπαρξης διαφορετικών πηγών δεδομένων έχει αναπτυχθεί μια διαδικασία η οποία επεξεργάζεται τους διαφορετικούς τύπους δεδομένων και τα εισάγει στη βάση. Η διαδικασία αυτή είναι γνωστή ως ETL (Extract, Transform, Load).

- Η διαδικασία Extract εκτελεί εξόρυξη πληροφοριών από διαφορετικές πηγές δεδομένων
- Η διαδικασία Transform μετατρέπει τις πληροφορίες σε μια ενιαία και διορθώνει τυχόν σφάλματα που προκύπτουν.
- Τέλος η διαδικασία Load φορτώνει τα μορφοποιημένα δεδομένα στην αποθήκη δεδομένων.

Στο διάγραμμα 5, εμφανίζεται μια απλή αναπαράσταση της διαδικασίας ETL καθώς βλέπουμε πως λαμβάνει δεδομένα από διάφορες πηγές Πληροφοριακά Συστήματα Επιχειρήσεων ERP, Λογισμικά Διαχείρισης Πελατών CRM, Λειτουργικά Συστήματα κτλ, και με τις διαδικασίες Extract, Transform και Load τα μετατρέπει σε μια ομογενή μορφή και τα φορτώνει στην αποθήκη των δεδομένων. Τα δεδομένα έχουν μετατραπεί σε μια απλή μορφή όπου ο χρήστης μπορεί να χρησιμοποιήσει OLAP εργαλεία για την ανάλυση των δεδομένων καθώς και να εξορύξει δεδομένα.



Διάγραμμα 5 - Διαδικασία ETL (Extract – Transform – Load)

Κεφάλαιο 6 - Εξόρυξη Γνώσης (Data Mining)

6.1 Ορισμός Εξόρυξης Γνώσης

Η εξόρυξη γνώσης από δεδομένα (Data Mining) ή πιο απλά η εξόρυξη γνώσης είναι μια νέα δυναμική τεχνολογία που βοηθάει τις επιχειρήσεις να εστιάσουν στην σημαντική πληροφορία που βρίσκεται μέσα στις αποθήκες δεδομένων τους (Data Warehouses). Οι τεχνικές της είναι σε θέση να αναζητήσουν και να βρουν γρήγορα και λεπτομερειακά βάσεις δεδομένων για την αναζήτηση κρυμμένων προτύπων (Patterns). Έτσι λοιπόν μπορούμε να πούμε ότι η εξόρυξη γνώσης είναι μια διαδικασία εξαγωγής κρυμμένης πληροφορίας από μεγάλες βάσεις δεδομένων.

Η εξόρυξη γνώσης περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που μεταξύ άλλων περιλαμβάνουν την στατιστική ανάλυση (Statistical Analysis), τα δένδρα αποφάσεων (Decision Trees), τα νευρωνικά δίκτυα (Neural Networks), την εξαγωγή κανόνων (Rule Induction) και την γραφική οπτικοποίηση (Graphic Visualization). [13]

6.2 Εξόρυξη Δεδομένων και Ανεύρεση Γνώσης

Η εξόρυξη βοηθά τις σύγχρονες εταιρίες να εστιάζουν στα πιο σημαντικά στοιχεία από τις αποθήκες δεδομένων τους. Με άλλα λόγια είναι η διαδικασία εφαρμογής μεθόδων ανάλυσης σε μεγάλο όγκο δεδομένων. Ο χρήστης των εργαλείων εξόρυξης μπορεί να προβλέψει μελλοντικές συμπεριφορές και συνήθειες, ώστε οι εταιρίες να παίρνουν επιτυχημένες αποφάσεις. Συνειδητοποιούμε ότι οι τεχνικές εξόρυξης γνώσης αναπτύσσονται γρήγορα, δίχως αλλαγές στην υποδομή και με μοναδικό στόχο την αξιοποίηση των επεξεργασμένων δεδομένων.

Στη διεθνή βιβλιογραφία υπάρχει μια γενικότερη σύγχυση ανάμεσα στους όρους «Εξόρυξη Γνώσης» και «Ανεύρεση γνώσης στις βάσεις δεδομένων». Σε πολλές περιπτώσεις αξίζει να σημειωθεί ότι οι δυο αυτοί όροι ταυτίζονται, ενώ στην πραγματικότητα η εξόρυξη δεδομένων αποτελεί τμήμα της ανεύρεσης γνώσης, συγκροτώντας το πυρήνα αυτής (Zaiane, 1999). Προκειμένου λοιπόν να κατανοηθεί καλύτερα η εξόρυξη δεδομένων, θα γίνει μια σύντομη αναφορά στη διαδικασία της ανεύρεσης γνώσης.

Η ανεύρεση γνώσης είναι μια επαναληπτική διαδικασία που αποτελείται από μια σειρά βημάτων, τα οποία οδηγούν από τη συλλογή των δεδομένων στην ανακάλυψη και εξαγωγή χρήσιμης πληροφορίας από αυτά.

Τα βήματα από τα οποία αποτελείται η διαδικασία ανεύρεσης γνώσης είναι τα ακόλουθα:

- Καθαρισμός Δεδομένων (Data Cleaning) : Στο βήμα αυτό, αφαιρούνται από τη βάση δεδομένων αυτά που παράγουν θόρυβο, δηλαδή όλα εκείνα τα στοιχεία που μπορούν να επηρεάσουν ή και να διαστρεβλώσουν το αποτέλεσμα.
- Ενσωμάτωση Δεδομένων (Data Integration) : Σε αυτό το βήμα, τα δεδομένα που έχουν συλλεχθεί, πολλές φορές ανομοιογενή και από πολλές διαφορετικές πηγές, ενσωματώνονται σε μια κοινή βάση δεδομένων.
- Επιλογή Δεδομένων (Data Selection) : Από όλα εκείνα τα δεδομένα που έχουμε στη διάθεση μας, επιλέγονται προσεκτικά εκείνα που είναι σχετικά και χρήσιμα για την ανάλυση που θα ακολουθήσει.
- Τροποποίηση Δεδομένων (Data Transformation) : Τα δεδομένα που έχουμε επιλέξει δέχονται τις απαραίτητες τροποποιήσεις έτσι ώστε η μορφή τους να είναι κατάλληλη για την διαδικασία της εξόρυξης.
- Εξόρυξη Δεδομένων (Data Mining) : Είναι το σημαντικότερο από τα βήματα της διαδικασίας και αυτό γιατί στο συγκεκριμένο στάδιο, ποικίλες εξελιγμένες τεχνικές χρησιμοποιούνται για την εξαγωγή δυνητικά χρήσιμων προτύπων.
- Ανάλυση Προτύπων (Pattern Evaluation) : Στο βήμα αυτό αναγνωρίζονται χρήσιμα πρότυπα που αναπαριστούν γνώση, βάσει συγκεκριμένων μέτρων αξιολόγησης (Evaluation Measures).
- Αναπαράσταση Γνώσης (Knowledge Representation) : Στο τελικό αυτό στάδιο, η γνώση που έχει ανακαλυφθεί παρουσιάζεται στον χρήστη, βοηθώντας τον έτσι να κατανοήσει και να ερμηνεύσει τα αποτελέσματα της εξόρυξης δεδομένων.

Πολλές φορές κάποια από τα παραπάνω βήματα μπορούν να συνδυαστούν μεταξύ τους για το καλύτερο δυνατό αποτέλεσμα. Για παράδειγμα, τα βήματα του καθαρισμού και της ενσωμάτωσης των δεδομένων, μπορούν να υλοποιηθούν μαζί με στόχο την δημιουργία μιας αποθήκης δεδομένων. Με την ίδια λογική μπορούν να συνδυαστούν και τα βήματα της επιλογής και τροποποίησης των δεδομένων.

Από τα παραπάνω λοιπόν συμπεραίνουμε ότι η εξόρυξη δεδομένων είναι μία διαδικασία κλειδί για την ανεύρεση γνώσης. Παρόλα αυτά, καταλαμβάνει μόνο ένα μικρό μέρος της όλης προσπάθειας, δεδομένου της πολυπλοκότητας της. Σε αυτό το σημείο αξίζει να σημειωθεί ότι ο χρήστης, εκμεταλλευόμενος την επαναληπτική μορφή της διαδικασίας ανεύρεσης γνώσης, έχει τη δυνατότητα να τροποποιήσει τα μέτρα αξιολόγησης, να τελειοποιήσει τη διαδικασία της εξόρυξης, να επιλέξει νέα δεδομένα, να τροποποιήσει περαιτέρω τα ήδη υπάρχοντα ή να ενσωματώσει στη βάση νέα από καινούργιες πηγές, με τελικό στόχο την εξαγωγή διαφορετικών και ακόμη πιο κατάλληλων αποτελεσμάτων.[13]

6.3 Στόχοι της Εξόρυξης Δεδομένων

Οι μέθοδοι εξόρυξης γνώσης στοχεύουν στην ανακάλυψη στοιχείων που θα είναι χρήσιμα για τους οργανισμούς και τις επιχειρήσεις. Πληροφορίες για τυποποιημένες μορφές όπως για παράδειγμα, ότι υπάρχουν πελάτες που θα ψωνίσουν περισσότερο από δυο φορές σε περίοδο εκπτώσεων ή προσφορών, ή είναι πιθανό να αγοράσουν τουλάχιστον μια φορά κατά την διάρκεια των εορταστικών ημερών, Πάσχα και Χριστουγέννων, μπορεί να αποτελέσουν καθοριστικούς παράγοντες για την λήψη αποφάσεων όσον αφορά τη λειτουργία μιας εμπορικής επιχείρησης. Αυτό συμβαίνει επειδή μπορεί να ληφθούν αποφάσεις σχετικά με το ωράριο, το ύψος και τη διάρκεια των εκπτώσεων, ακόμη και για την τοποθέτηση των προϊόντων μέσα στα καταστήματα.

Παράλληλα τέτοιου είδους πληροφορίες χρησιμοποιούνται για τον προγραμματισμό χρήσης πρόσθετων αποθηκευτικών χώρων ή και για τον σχεδιασμό διαφορετικών στρατηγικών μάρκετινγκ. Τα στελέχη της επιχείρησης, που είναι υπεύθυνα για την λήψη των αποφάσεων εκμεταλλεύονται τις δυνατότητες της εξόρυξης γνώσης και μετατρέπουν τις γνώσεις σε επιτυχή αποτελέσματα. Παρακάτω περιγράφονται και αναλύονται οι στόχοι της εξόρυξης δεδομένων.

Η εξόρυξη δεδομένων έχει λοιπόν σαν βασικούς της στόχους την εφαρμογή τεχνικών πρόβλεψης και συμπεριφοράς τάσεων, την αναγνώριση, την περιγραφή σε μεγάλες βάσεις δεδομένων, καθώς επίσης την ταξινόμηση και την βελτιστοποίηση των πόρων της. Ειδικότερα :

- Πρόβλεψη: Περιλαμβάνει την χρήση μερικών μεταβλητών ή χαρακτηριστικών μιας βάσης δεδομένων για την πρόβλεψη άγνωστων ή μελλοντικών τιμών χρήσιμων μεταβλητών. Με άλλα λόγια, οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων (Predictive Data Mining Tasks), προσπαθούν να κάνουν εκτιμήσεις βγάζοντας συμπεράσματα από τα διαθέσιμα δεδομένα. Η προσπάθεια πρόβλεψης μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις που να μεγιστοποιούν το κέρδος και να προλαμβάνουν δυσάρεστες καταστάσεις. Τα αποτελέσματα της εξόρυξης μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια συγκεκριμένη χρονική περίοδο, αλλά και αν το κλείσιμο μιας γραμμής παραγωγής θα είχε θετική επίδραση στις πωλήσεις.

Συγχρόνως σε επιστημονικό επίπεδο, η μελέτη παλαιότερων σεισμικών φαινομένων ίσως να οδηγούσε στην πρόβλεψη σεισμικής δραστηριότητας.

- Αναγνώριση: Σε αυτή τη φάση οι τυποποιημένες μορφές των δεδομένων χρησιμοποιούνται για να δείξουν την ύπαρξη μιας δραστηριότητας ή ενός γεγονότος.
- Περιγραφή: Είναι η διαδικασία η οποία επικεντρώνεται στην ανακάλυψη προτύπων και αναπαριστά τα δεδομένα μιας πολύπλοκης βάσης δεδομένων με όσο το δυνατό πιο κατανοητό και αξιοποιήσιμο τρόπο. Με άλλα λόγια, οι περιγραφικές διαδικασίες της εξόρυξης δεδομένων (Descriptive Data Mining Tasks) περιγράφουν τις γενικές ιδιότητες των υπαρχόντων διαθέσιμων δεδομένων.
- Ταξινόμηση: Σε αυτό το στάδιο έχουμε διαχωρισμό των στοιχείων, με αποτέλεσμα να προκύπτουν διαφορετικές κατηγορίες ή κλάσεις. Για παράδειγμα, οι πελάτες ενός σούπερ μάρκετ είναι δυνατόν να χωριστούν σε παρορμητικούς, πιστούς ή αλλιώς όπως θα λέγαμε κανονικούς, σπάνιους και σε φίλους των εκπτώσεων και προσφορών. Κατά την ανάλυση των πωλήσεων αυτή η κατηγοριοποίηση χρησιμοποιείται για να ληφθούν αποφάσεις, ώστε να προσελκυστούν περισσότεροι πελάτες ανεξαρτήτως κατηγορίας.
- Βελτιστοποίηση: Μεταξύ των άλλων σκοπός της εξόρυξης γνώσης είναι η βέλτιστη χρήση κάποιων πόρων κάτω από περιορισμούς. Τέτοιοι πόροι μπορεί να είναι ο χρόνος, ο χώρος, το χρήμα, και η μεγιστοποίηση κάποιων μεγεθών, όπως είναι τα κέρδη είτε οι πωλήσεις. Σε αυτή την περίπτωση η εξόρυξη γνώσης έχει κοινά σημεία με την επιχειρησιακή έρευνα. [13]

Κεφάλαιο 7 - Τομείς που εξελίχθηκαν με τη βοήθεια των Big Data

7.1 Αναλυτής Δεδομένων (Data Scientist)

Ας ξεκινήσουμε την δημιουργία νέου επαγγέλματος, αφού τα 2 τελευταία χρόνια έχει αναδειχθεί ένας νέος ρόλος στις εταιρίες και στους οργανισμούς, αυτός του Αναλυτή Δεδομένων κοινώς γνωστό ως Data Scientist.

Οι Data Scientists έχουν εξελιχθεί σε μία ανώτερη θέση στην εκτελεστική ιεραρχία μιας εταιρίας ή ενός οργανισμού διότι ο ρόλος τους συνδυάζει τα προσόντα ενός ικανού προγραμματιστή, στατιστικού και αφηγητή με σκοπό να ανακαλύψει τους «σβόλους χρυσού» που υπάρχουν κάτω από τους τεράστιους όγκους δεδομένων.

Προβλέπεται πως το επάγγελμα του Data Scientist θα γίνει σύντομα ένα από τα πλέον ελκυστικά. Τα δεδομένα πλέον είναι ευρέως διαθέσιμα όμως αυτό που σπανίζει είναι η ικανότητα εξόρυξης γνώσης από αυτά.

Υπάρχει μεγάλη αναγκαιότητα επαγγελματιών που προσφέρουν αυτές της υπηρεσίες και παράλληλα μεγάλη έλλειψη ανθρώπων σε αυτόν τον κλάδο.

Σύμφωνα με έρευνα που διεξήγαγε η συμβουλευτική εταιρία McKinsey, προβλέπεται ότι θα χρειαστούν 170.000 Data Scientists και ότι περίπου 1.500.000 Managers θα πρέπει να εκπαιδευτούν στη χρήση και ανάλυση των δεδομένων.

Υπάρχει μεγάλη συζήτηση για το ποιες είναι οι ικανότητες και γνώσεις που συνθέτουν το υπόβαθρο του Data Scientist. Καταρχάς θα πρέπει να έχει πολύ καλή και βαθιά γνώση των θεμάτων που αφορούν τη διαχείριση δεδομένων (μοντελοποίηση δεδομένων, γλώσσες ερωτημάτων και ανάκτησης δεδομένων, γλώσσα SQL και NOSQL, τεχνολογίες επεξεργασίας, αποθήκευσης κλπ). Επίσης χρειάζεται να έχει πολύ καλό μαθηματικό υπόβαθρο ώστε να μπορεί να κατανοήσει και να χειριστεί τις στατιστικές μεθόδους και τις τεχνικές Machine Learning Που απαιτούνται στην εξαγωγή προτύπων. Τέλος θα πρέπει να γνωρίζει κάποιες αρχές του σύγχρονου Management ώστε να ακολουθεί τις εξελίξεις του σημερινού τεχνολογικού κόσμου και να έχει μία αντίληψη πως μπορούν να χρησιμοποιηθούν τα δεδομένα σε μια εταιρία ή έναν οργανισμό ούτως ώστε οι εργασίες του πάνω στα δεδομένα να αποφέρουν τις πληροφορίες χρειάζεται μια επιχείρηση για να καλύψει τις ανάγκες τις.

Κεφάλαιο 8 - Business Analytics

8.1 Ορισμός

Με τον όρο Business Analytics εννοούμε τη χρήση των δεδομένων που συλλέγουμε από την διαδικασία Data Mining και την επεξεργασία τους με εργαλεία Analytics, στατιστικές μεθόδους και οικονομετρικές εφαρμογές για άντληση επιπλέον πληροφοριών που με σωστή διαχείριση μπορούν να αποφέρουν κέρδη ή μείωση κόστους σε έναν οργανισμό. Κύριος στόχος είναι η διαρκή διερεύνηση της παρελθοντικής πορείας μιας επιχείρησης και η ορθότερη εκτίμηση της μελλοντικής επίδοσης των τμημάτων της. Με τον τρόπο αυτό μπορούν να ληφθούν οι ορθότερες επιχειρηματικές αποφάσεις για την πορεία μιας επιχείρησης που βαίνει στην δύνη του ανταγωνισμού.

Εδώ γεννιέται η ανάγκη για δημιουργία ενός νέου επαγγέλματος στο χώρο των επιχειρήσεων, αυτό του Business Analyst. Οι Business Analysts είναι τα άτομα εκείνα που μπορούν να απαντήσουν σε ερωτήσεις όπως: Τι θα συμβεί αν πάρω αυτήν την απόφαση; Τι θα συμβεί αν συνεχιστούν αυτές οι τάσεις; Αν αλλάξω κάποιες μεταβλητές, πού θα έχουν αντίκτυπο αυτές οι αλλαγές στο μέλλον;

Η επιστήμη της ανάλυσης περιγράφει τον τρόπο με τον οποίο μια επιχείρηση θα φτάσει σε μια βέλτιστη απόφαση, βάσει της επεξεργασίας των υπαρχόντων δεδομένων. Χρησιμοποιώντας εφαρμογές μελέτης και στατιστικής ανάλυσης των στοιχείων, οι αναλυτές αποσκοπούν σε βελτίωση της επιχειρηματικής δραστηριότητας.

8.2 Τα Business Analytics στις επιχειρήσεις

Την σημερινή εποχή σχεδόν όλες οι επιχειρήσεις παρέχουν παρόμοια προϊόντα και υπηρεσίες, όμως το πεδίο των Analytics μπορεί να αποτελέσει στοιχείο διαφοροποίησης από τον ανταγωνισμό. Για τις περισσότερες επιχειρήσεις η διαδικασία ανάλυσης των δεδομένων είναι όμοια με τις άλλες, όμως ο τρόπος εφαρμογής και η ποιότητα της πληροφορίας που παράγεται διαφοροποιούνται από επιχείρηση σε επιχείρηση και μπορούν να οδηγήσουν σε νέα πρωτοφανή στοιχεία και καινοτομίες και αυτομάτως δημιουργείται ένα ισχυρό πλεονέκτημα για τους οργανισμούς που θα τα καταφέρουν. Οι επιχειρήσεις που ενστερνίζονται την σύγχρονη νοοτροπία τις μεθόδους ανάλυσης, έχουν πιο αποτελεσματική λειτουργία και κατά μέσο όρο τα στελέχη τους παίρνουν καλύτερες αποφάσεις. Σε αυτές τις μεθόδους υπάρχει μεγάλη δυσκολία να γίνει αντιγραφή και επίσης εφαρμόζονται σε πλήθος διαφορετικών καταστάσεων.

Τα Business Analytics δεν είναι απαραίτητο να αφορούν μόνο μεγάλες επιχειρήσεις. Τα τελευταία χρόνια η εξάπλωση τους έχει επηρεάσει και τις μικρές επιχειρήσεις καθώς έχουν γίνει πιο προσιτά οικονομικά και μπορούν να προσδώσουν το ανταγωνιστικό πλεονέκτημα στις επιχειρήσεις που το αναζητούν για να διακριθούν από τον ανταγωνισμό.

Επίσης τα Business Analytics δεν παρέχονται στο κάθε μέλος μιας επιχείρησης, αλλά σε συγκεκριμένες ομάδες ατόμων που έχουν το δικαίωμα λήψης αποφάσεων. Μιλάμε κυρίως για τα ανώτερα διοικητικά στελέχη, καθώς επίσης δίνεται πρόσβαση στις BA εφαρμογές και στα τμήματα:

- Marketing,
- HRM (Human Resources Management = Τμήμα Ανθρώπινου Δυναμικού)
- Finance
- Supply Chain Management
- Production & Operations
- Internet Analytics

Παρακάτω θα αναλύσουμε την συμβολή των Business Analytics σε κάθε σχετιζόμενο τμήμα μιας επιχείρησης:

8.2.1 Marketing

Ένα από τα σημαντικότερα τμήματα μιας επιχείρησης είναι αυτό του Marketing καθώς από την λειτουργία του εξαρτάται ο όγκος των πωλήσεων, ο τρόπος που θα προσεγγίσει τους καταναλωτές και συνεπώς το κέρδος που θα προκύψει. Η απόδοση του τμήματος του Marketing σύμφωνα με τα Business Analytics γίνεται με τα Marketing metrics. Τα Marketing metrics είναι οικονομετρικές αναλύσεις που επεξεργάζονται τα στοιχεία του τμήματος και βοηθούν στην τεκμηρίωση ερωτημάτων που μπορεί να προκύψουν.

Τα Marketing metrics εκτός των άλλων παρέχουν μια πολύ σημαντική δυνατότητα, την ανάλυση πωλήσεων. Τα αρμόδια στελέχη της επιχείρησης έχοντας στα χέρια τους την ανάλυση πωλήσεων, μπορούν να αντλήσουν μεγάλο αριθμό χρήσιμων πληροφοριών σχετικά με την εμπορική πορεία του τμήματος, να δουν ποιες κινήσεις και τακτικές πρέπει να τροποποιήσουν για αύξηση των πωλήσεων, να δουν τις μελλοντικές καταναλωτικές τάσεις των πελατών και να προσαρμόσουν τις μεθόδους τους στην σύγχρονη αγορά.

Η κάθε επιχείρηση γνωρίζει το σύνολο των πωλήσεων της, όμως τα BA δίνουν λεπτομερείς αναλύσεις για το μέρος των πωλήσεων που προέρχεται από:

- Ραδιόφωνο
- Τηλεόραση
- Φυλλάδια – Έντυπα
- Προσφορές
- Εκπτώσεις
- Πιστούς πελάτες

Με τόσες πληροφορίες που προσφέρουν τα Business Analytic metrics ο κάθε διευθυντής Marketing έχει στα χέρια του έγκυρα στοιχεία και μπορεί να αξιολογήσει ορθότερα την κάθε προωθητική ενέργεια ως προς το αποτέλεσμα που έφερε. Πιο συγκεκριμένα είναι πλέον

εύκολο να υπολογιστεί ο όρος ROI (Return on Investment) για την κάθε ενέργεια, δηλαδή ποίο ήταν το ακριβές κέρδος της επιχείρησης για την κάθε επένδυση που έκανε. Αυτό δίνει στην επιχείρηση το πλεονέκτημα να γνωρίζει αφενός ποιες ενέργειες της αποφέρουν περισσότερο κέρδος και αφετέρου ποιες ενέργειες να μην επαναλάβει διότι είχαν ασύμφορο οικονομικό αποτέλεσμα. Εκτός αυτού, μέσω της ανάλυσης προσδιορίζεται ο βαθμός στον οποίο χρειάζεται μια επιχείρηση να επενδύσει σε μια ενέργεια. Έτσι τα φαινόμενα της υποεπένδυσης και υπερεπένδυσης εξαλείφονται.

Εκτός των άλλων, τα BA metrics παρέχουν πληροφορίες για τους νέους πελάτες που προέκυψαν από προωθητικές ενέργειες, βοηθώντας τα τμήματα Marketing να εστιάσουν σε ενέργειες που θα φέρουν νέες ομάδες καταναλωτών.

Επιπλέον με την χρήση των BA metrics αναλύονται και οι κινήσεις των ανταγωνιστών. Συγκεκριμένα τα αρμόδια στελέχη έχουν την δυνατότητα να προσδιορίσουν τις κερδοφόρες ενέργειες των ανταγωνιστών και να βελτιώσουν την στρατηγική τους ώστε να δράσουν ισάξια απέναντι στον ανταγωνισμό.

Σε περίπτωση σύγκρισης μεταξύ μιας επιχείρησης και του ανταγωνισμού της τα Marketing metrics με τις λεπτομερείς αναλύσεις τους αποδίδουν έγκυρα αποτελέσματα και σε εμπορικό τομέα αλλά και σε κλαδικό δίνοντας έτσι την δυνατότητα μέτρησης της επίδοσης της επιχείρησης σε σχέση με τον ανταγωνισμό.

Η αξία των Business Analytics είναι ανεκτίμητη διότι παρέχει πολύ χρήσιμες πληροφορίες που βοηθούν στην καλύτερη πορεία μιας επιχείρησης. Τα στελέχη του Marketing μπορούν να πάρουν αποφάσεις έχοντας επίγνωση των επιπτώσεων που μπορεί να προκύψουν και επίσης έχουν πλεονέκτημα απέναντι στις επιχειρήσεις του ανταγωνισμού αφού έχουν την δυνατότητα να προβλέψουν τις μελλοντικές τάσεις της αγοράς.

8.2.2 Human Resources Management

Το Τμήμα Ανθρώπινου Δυναμικού διαχειρίζεται το προσωπικό μιας εταιρίας. Οι άνθρωποι είναι βασικό συστατικό από το οποίο απαρτίζεται κάθε οργανισμός και βασίζεται εκεί για την ύπαρξη του αλλά και για την αλληλεπίδραση με το περιβάλλον. Ο τρόπος με τον οποίον επιλέγονται τα μέλη μιας εταιρίας εκτελείται με ιδιαίτερη προσοχή ώστε να αποφεύγονται οι λάθος επιλογές και τα άτομα που θα ενταχθούν στη να έχουν βλέψεις για μελλοντική εξέλιξη. Τα Business Analytics μπορούν να επιφέρουν επιπλέον αποτελεσματικότητα στην διαχείριση του ανθρώπινου δυναμικού με την προϋπόθεση να συλλέγονται και να επεξεργάζονται

στοιχεία όπου σαν αποτέλεσμα θα έχουν χρήσιμες πληροφορίες που θα βελτιώσουν τις αποφάσεις που θα ληφθούν.

Η χρήση των ΒΑ στα τμήματα ανθρώπινου δυναμικού βοηθούν τους υπεύθυνους HR να κατανοήσουν:

- Πώς μια εταιρία μπορεί να προσελκύσει τους αποδοτικότερους εργαζομένους και αργότερα να είναι σε θέση να τους εξελίξει
- Τους λόγους για τους οποίους κάποιοι εργαζόμενοι εγκαταλείπουν το επάγγελμα τους
- Τις μεθόδους με τις οποίες θα εκπαιδευτούν τα νέα άτομα σε μια επιχείρηση ώστε η πορεία τους να είναι αποδοτική και να έχει τάσεις εξέλιξης
- Τους τρόπους που ένας διευθυντής τμήματος θα ενθαρρύνει τους υφισταμένους του να αυξήσουν την απόδοση της εργασίας τους και τα κριτήρια που θα αποφασίσει για τις δίκαιες μηνιαίες απολαβές τους
- Τους τρόπους που μια επιχείρηση μπορεί να επιβραβεύσει τους εργαζομένους της για την απόδοση της εργασίας.

Οι παραδοσιακοί αναλυτές HR παρουσιάζουν μια σειρά από μετρήσιμα στοιχεία και φαίνονται παρακάτω:

1) Ανάλυση της συμπεριφοράς των εργαζομένων

- σε επίπεδο ικανοποίησης με την εργασία τους
- σε επίπεδο ικανοποίησης από τον προϊστάμενο τους

2) Κύκλος ζωής εργασιακής σχέσης

- κόστος ανά πρόσληψη (διαφήμιση, αγγελίες)
- κόστος αναπλήρωσης θέσης (λήξη σύμβασης, ανανέωση σύμβασης, νέα πρόσληψη)

- βαθμός αποχώρησης (το σύνολο των εργαζομένων που αποχωρούν από την επιχείρηση)
- χρόνος αναπλήρωσης (χρόνος από την δημιουργία νέας θέσης εργασίας μέχρι την πρόσληψη)
- διάρκεια εργασίας (για κάθε τμήμα ή και θέση)

3) Διαδικασία πρόσληψης

- περίοδος που παρουσιάστηκαν κενές θέσεις
- προσδιορισμός απόδοσης νέων προσλήψεων

4) Διατήρηση εργαζομένων

- ποσοστό εργαζομένων που δηλώνουν παραίτηση
- ποιοτικά στοιχεία εργαζομένων
- οικονομικές υποχρεώσεις ως προς το προσωπικό

5) Εκπαίδευση και ανάπτυξη

- ευκαιρίες εκμάθησης (το σύνολο των εργαζομένων που νιώθουν ικανοποιημένοι από την εκπαίδευση τους)
- ευκαιρίες για νέες προσλήψεις (το σύνολο των εργαζομένων που θεωρούν πως ο βασικός λόγος που προσεγγίζουν μία εταιρία είναι οι ευκαιρίες εκμάθησης που προσφέρουν)

Αυτά είναι τα βασικά μέτρα που χρησιμοποιούν τα τμήματα HR για την αξιολόγηση της λειτουργίας τους. Σήμερα με την έλευση των Business Analytics, μπορούμε να τα χρησιμοποιήσουμε και να έχουμε μεγαλύτερη και ορθότερη πληροφόρηση που προέρχεται από ανάλυση συσχετίσεων και την επεξεργασία ποιοτικών και ποσοτικών στοιχείων.

Τα BA και συγκεκριμένα τα HR metrics βασίζονται καθαρά στα στατιστικά στοιχεία που προέρχονται από όλα τα πεδία της εταιρίας που αφορούν το τμήμα ανθρώπινου δυναμικού και επεξεργάζονται τα δεδομένα με κατάλληλο τρόπο για την άντληση πολύτιμων συμπερασμάτων.

Παρακάτω παρουσιάζονται οι βασικές δυνατότητες που προσφέρουν τα BA HR metrics:

- Προσέλκυση των καλύτερων εργαζόμενων με χαμηλό κόστος
- Ανάλυση της αποτελεσματικότητας της διαδικασίας πρόσληψης
- Προσφέρεται δυνατότητα προβλέψεων για τις συνέπειες πιθανών αλλαγών
- Καλύτερη κατανόηση των λόγων που κάποιοι εργαζόμενοι αποχωρούν
- Ανάλυση της διάρκειας εργασίας των εργαζομένων
- Καλύτερη αξιολόγηση απόδοσης των εργαζομένων και ευκαιρίες ανάπτυξης
- Ανάλυση των ικανοτήτων και του ταλέντου του κάθε εργαζομένου μεμονωμένα
- Βελτιστοποίηση σχέσης απόδοσης και μισθού
- Ορθός υπολογισμός του μισθού
- Διασφάλιση δίκαιης αμοιβής μεταξύ των διαφορετικών ρόλων της επιχείρησης
- Καλύτερος έλεγχος αδειών έκτακτων και μη, με πλήρη ανάλυση ιστορικού

Επίσης τα BA HR metrics μπορούν να διαχωριστούν σε τρεις βασικές κατηγορίες όπου κάθε μια κατηγορία είναι σημαντική και ο συνδυασμός τους μπορεί να μας δώσει την συνολική εικόνα του κάθε HR τμήματος. Οι κατηγορίες είναι οι εξής:

1. Παρελθούσα απόδοση
2. Σημερινή κατάσταση
3. Μελλοντική προοπτική

Έτσι λοιπόν βλέπουμε πως τα Business Analytics βοηθούν στην λήψη καλύτερων αποφάσεων όσον αφορά τον τρόπο λειτουργίας του Τμήματος Ανθρώπινου Δυναμικού. Η σωστή λειτουργία του τμήματος HR συντελεί σημαντικά στην επίτευξη των στόχων ενός

οργανισμού και με την δημιουργία σημαντικών δεικτών και ορθών μετρήσεων ο κάθε οργανισμός έχει την δυνατότητα να αξιολογήσει αν η κάθε κίνηση που κάνει, μπορεί να εξυπηρετήσει τα συμφέροντά του.

8.2.3 Finance

Την σημερινή εποχή στον κόσμο των επιχειρήσεων θεωρείται πλέον δεδομένη η ύπαρξη τεραστίου όγκου διαθέσιμης πληροφορίας. Οι προκλήσεις παρουσιάζονται στον τρόπο με τον οποίο τα αρμόδια στελέχη θα διαχειριστούν και θα επεξεργαστούν τις πληροφορίες αυτές ώστε να γίνουν πιο κατανοητές και να φέρουν σαν αποτέλεσμα αξιοποιήσιμη γνώση που θα οδηγήσει στην βελτίωση της απόδοσης ενός οργανισμού.

Δεδομένου αυτού του όγκου πληροφορίας, ένας οικονομικός διευθυντής δεν μπορεί μόνος του να έχει συνολική εικόνα με την χρηματοοικονομική κατάσταση μιας εταιρίας. Επιπλέον είναι μεγάλος ο κίνδυνος για δημιουργία λανθασμένης εικόνας που εύκολα μπορεί να οδηγήσει σε λανθασμένες αποφάσεις. Το σύγχρονο χρηματοοικονομικό περιβάλλον έχει φτάσει σε ένα πολύπλοκο επίπεδο και οι αναπάντεχες εξελίξεις στην παγκόσμια οικονομία δυσκολεύουν σε μεγάλο βαθμό το έργο των οικονομικών τμημάτων των επιχειρήσεων.

Μέσα σε αυτό το χάος χρειάστηκε ένα νέο εργαλείο που θα διευκόλυνε την λειτουργία των οικονομικών τμημάτων ώστε να ανταπεξέλθουν στις χρηματοοικονομικές αλλαγές. Αυτό είναι τα Financial Analytics όπου μιλάμε για προγράμματα που έχουν στόχο να απλοποιήσουν τα προβλήματα που παρουσιάζονται στα τμήματα των χρηματοοικονομικών και συγχρόνως να προσφέρουν διευκολύνσεις σε καθημερινές διαδικασίες. Έχουν την ικανότητα να παρουσιάζουν σε πραγματικό χρόνο μια καλύτερη οπτική εικόνα της οικονομικής κατάστασης του οργανισμού και υποστηρίζουν την λήψη αποφάσεων.

Βοηθούν τα αρμόδια στελέχη να κατανοήσουν:

- ❖ Τους πιο επικερδείς πελάτες μιας εταιρίας
- ❖ Ποια προϊόντα επιστρέφουν το περισσότερο κέρδος
- ❖ Ποιος είναι ο ορθότερος οδηγός προς το κέρδος
- ❖ Τους χρηματοοικονομικούς κινδύνους που αντιμετωπίζει μια εταιρία

Τα Financial Analytics θεωρούνται αναγκαία εργαλεία για να μπορέσει μια επιχείρηση να σταθεί επάξια σε ένα πλήρως ανταγωνιστικό περιβάλλον με βάση τον οικονομικό παράγοντα. Με την χρήση τους βοηθούν μια επιχείρηση να χειρίζεται μεγάλου όγκου πληροφορία και να αναπαριστά αποτελέσματα σε αναφορές, γραφήματα, στατιστικά διαγράμματα σε πραγματικό χρόνο. Επίσης υπάρχει η δυνατότητα για ανάλυση ιστορικών κινήσεων αλλά και πρόβλεψη μελλοντικών τάσεων με νέους παράγοντες που δεν υπήρχαν. Η διαδικασία του προϋπολογισμού γίνεται με μεγαλύτερη ακρίβεια και οι προβλέψεις γίνονται με πιο αποτελεσματικό τρόπο χάρη στα Financial Analytics. Σε περιπτώσεις που οι ληφθείσες αποφάσεις θεωρούνται επίφοβες, οι προσομοιώσεις που λαμβάνουν χώρα είναι ακριβείς και τα διοικητικά στελέχη έχουν καλύτερη εικόνα για τις ενδεχόμενες συνέπειες που μπορεί να προκύψουν.

Εκτός των άλλων τα Financial Analytics προσφέρουν:

- ❖ Έλεγχο κόστους
- ❖ Ανάλυση της κερδοφορίας την εταιρίας, αλλά και ανά τμήμα
- ❖ Αποτελεσματική συνεργασία του τμήματος χρηματοοικονομικών με τα υπόλοιπα τμήματα της επιχείρησης
- ❖ Πρόβλεψη των κινδύνων
- ❖ Καλύτερη Κατανόηση των προβλημάτων που σαν αποτέλεσμα έχει τον καλύτερο έλεγχο τους
- ❖ Βελτιστοποίησης της πολιτικής λειτουργίας για αύξηση της κερδοφορίας

8.2.4 Supply Chain Management

Μιλάμε για το Τμήμα Εφοδιαστικής Αλυσίδας, το οποίο περιλαμβάνει όλες τις ενέργειες που χρειάζεται ώστε ένα προϊόν να φτάσει από την παραγωγή στον τελικό καταναλωτή. Το τμήμα αυτό χαρακτηρίζεται ως σύνθετο καθώς εκεί εκτελούνται οι διαδικασίες παραγωγής, παραλαβής πρώτων υλών, αποθήκευσης και διανομής των προϊόντων στα σημεία πώλησης ή κατευθείαν στους καταναλωτές. Για την λήψη σωστών αποφάσεων δεν αρκεί μόνο η εμπειρία των υπευθύνων του τμήματος. Είναι απαραίτητη και η συλλογή και επεξεργασία στοιχείων ώστε να παραχθούν χρήσιμες πληροφορίες που θα διευκολύνουν τα αρμόδια στελέχη να:

- Προσδιορίσουν το ιδανικό επίπεδο παραγωγής
- Καθορίσουν το ιδανικό επίπεδο αποθεμάτων
- Καθιερώσουν τα ιδανικά δρομολόγια για την αποτελεσματικότερη διανομή των εμπορευμάτων

Με την χρήση των Business Analytics στο Supply Chain Management οι επιχειρήσεις μπορούν να κατανοήσουν καλύτερα την ζήτηση που υπάρχει στην αγορά και μπορούν να προσαρμόσουν με τέτοιο τρόπο την λειτουργία τους ώστε να ανταπεξέρχονται επάξια σε κάθε πρόκληση που παρουσιάζεται. Γνωρίζονται ακριβώς τι ζητάει η αγορά, οι εταιρίες μπορούν να καθορίσουν την κατάλληλη ποσότητα παραγωγής, την καλύτερη διαχείριση των αποθεμάτων και τους πιο αποδοτικούς τρόπους διανομής.

Πιο συγκεκριμένα τα ΒΑ προσφέρουν:

- Διαχείριση αποθεμάτων και έλεγχο αποθηκευτικών χώρων
- Έγκυρα στοιχεία απογραφής
- Έλεγχο ποιότητας των λειτουργιών παραγωγής, αποθήκευσης και διανομής
- Καλύτερη πληροφόρηση όσον αφορά την παραγγελία πρώτων υλών
- Πρόβλεψη αναγκών και τάσεων που πρόκειται να προκύψουν στην αγορά
- Σχεδίαση καλύτερου κυκλώματος διανομής λαμβάνοντας υπόψη μεταφορικά μέσα, όγκο εμπορευμάτων, ημερομηνία παράδοσης των παραγγελιών κτλ.

- Έλεγχος ποιότητας της συνολικής λειτουργίας του τμήματος
- Μείωση των χειριστικών λαθών λόγω αυτοματοποίησης διαδικασιών
- Πλήρη έλεγχο της διαδικασίας παραγωγής
- Καλύτερη εξυπηρέτηση των πελατών επιχείρησης

Η εποχή όπου μόνο με την εμπειρία των υπευθύνων είχαμε μια ικανοποιητική λειτουργία του τμήματος της εφοδιαστικής αλυσίδας πέρασε. Πλέον τα εργαλεία των ΒΑ είναι το κλειδί για την εξέλιξη της λειτουργίας του τμήματος. Με τα πλεονεκτήματα που προσφέρουν βοηθούν τις επιχειρήσεις να αναλύσουν τις ανάγκες τις αγορές και να τις ικανοποιήσουν. Αυτό συντελεί σημαντικά στην επίτευξη των στόχων των επιχειρήσεων και αυτομάτως αποκτούν πλεονέκτημα απέναντι στον ανταγωνισμό.

Τα Supply Chain Management metrics παρέχουν πολλά οφέλη στην λειτουργία του τμήματος που οδηγούν στην αύξηση του κέρδους. Για να επιτευχθεί αυτό χρειάζεται να αξιοποιηθούν με κατάλληλο τρόπο όλοι οι διαθέσιμοι πόροι του τμήματος.

8.2.5 Production and Operations

Ένα εξίσου σημαντικό τμήμα μιας εταιρίας είναι αυτό της παραγωγής. Η λειτουργία του είναι πολύπλοκη καθώς αποτελείται από μεγάλο αριθμό βημάτων όπου πρώτα πρέπει να ολοκληρωθεί το ένα βήμα και μετά να ξεκινήσει το επόμενο. Χρειάζεται αρκετή προσοχή στην εκτέλεση της διαδικασίας και συνεπώς γεννιέται η ανάγκη για λεπτομερή έλεγχο των ενεργειών ώστε να διασφαλιστεί η σωστή λειτουργία του τμήματος.

Το βασικό απαιτούμενο για τον καλύτερο έλεγχο είναι η συλλογή δεδομένων. Τα ΒΑ με την μορφή των Production & Operations metrics αναλύουν τις πληροφορίες που συλλέγονται και δίνουν μια συνολική εικόνα λειτουργίας από την οποία τα αρμόδια στελέχη μπορούν να ενημερωθούν πλήρως για:

- Τις διάφορες αδυναμίες που εντοπίζονται στις διαδικασίες του τμήματος και τις διαδικασίες που χρειάζονται βελτίωση
- Τον βαθμό αποτελεσματικότητας της υπάρχουσας παραγωγικής διαδικασίας
- Την πιθανότητα μείωσης της παραγωγής ελαττωματικών προϊόντων
- Την πιθανότητα μείωσης του κόστους παραγωγής χωρίς να επηρεαστεί αρνητικά η ποιότητα των παραγόμενων προϊόντων

Η συμβολή των Production & Operation metrics είναι πολύτιμη αφού η αναλύσεις που προσφέρουν βοηθούν στον εντοπισμό λαθών που δεν παρατηρήθηκαν στους αρχικούς ελέγχους. Δεδομένα συλλέγονται συνεχώς από όλα τα σημεία της παραγωγικής διαδικασίας και οδηγούν στον προσδιορισμό των προβλημάτων που προκύπτουν. Τα προβλήματα αυτά τεκμηριώνονται και επιλύονται με ταχύτατους ρυθμούς και οι ενέργειες αυτές έχουν σαν αποτέλεσμα την αύξηση της απόδοσης παραγωγή και μείωσης της παραγωγής ελαττωματικών προϊόντων. Επίσης διεξάγονται επιπλέον έλεγχοι κατά την έξοδο των νέων προϊόντων από την παραγωγή και ο αρμόδιος υπεύθυνος αποφασίζει αν το προϊόν εγκρίνεται ή απορρίπτεται.

Τέλος με την χρήση των ΒΑ στο τμήμα παραγωγής επιτυγχάνεται ελαχιστοποίηση του χρόνου παραγωγής, εντοπίζονται τυχόν αδυναμίες στις διαδικασίες και προτείνονται τρόποι τροποποιήσεις που συντελούν στην πιο αποτελεσματική διαδικασία παραγωγής.

8.2.6 Internet Analytics

Σήμερα σχεδόν κάθε εταιρία διαθέτει δική της ιστοσελίδα παρουσιάζοντας στους καταναλωτές τα προϊόντα ή τις υπηρεσίες που προσφέρει. Αυτός ο σύγχρονος τρόπος προώθησης παρέχει μεγάλο αριθμό χρήσιμων πληροφοριών προς την επιχείρηση. Για να μπορέσει η κάθε εταιρία να αξιοποιήσει αυτές τις πληροφορίες και να έχει ένα μέτρο απόδοσης της ιστοσελίδας της χρησιμοποιεί συγκεκριμένα εργαλεία, γνωστά ως Internet Analytics. Είναι εύκολα στην χρήση τους χωρίς απαίτηση εξειδικευμένων γνώσεων. Μας παρουσιάζουν λεπτομερή στοιχεία σχετικά με:

- Την καταμέτρηση της επισκεψιμότητας της σελίδας
- Την καταμέτρηση των χρηστών που επισκέφθηκαν την σελίδα μόνο μια φορά
- Τις συνολικές σελίδες ανά επισκέπτη
- Τις πηγές επισκεψιμότητας
- Τον χρόνο παραμονής στην σελίδα

Χρησιμοποιώντας τα Internet Analytics metrics τα αρμόδια στελέχη μπορούν να πληροφορηθούν τους τρόπους της επισκεψιμότητας της σελίδας ανά χρήστη. Δηλαδή πόσοι χρήστες πληκτρολόγησαν το site της ιστοσελίδας, πόσοι βρήκαν την ιστοσελίδα από το Google ή κάποια άλλη μηχανή αναζήτησης και πόσοι πλοηγήθηκαν σε αυτήν πατώντας απλά σε ένα link. Βλέποντας τα αποτελέσματα γνωρίζουν με ποιους τρόπους γνωστοποιείται η ύπαρξη της ιστοσελίδας στους καταναλωτές και είναι σε θέση να δουν τα αδύναμα σημεία και να τα ενισχύσουν. Επίσης δίνουν σημαντικές πληροφορίες για τις δραστηριότητες και τα γούστα των χρηστών καθώς επίσης ενημερώνουν για την συμπεριφορά τους κατά την πλοήγηση τους στις ιστοσελίδες.

Μια σημαντική λειτουργία των Internet Analytics είναι ότι αναλύουν τις λέξεις-κλειδιά στις μηχανές αναζήτησης για να εμφανίζεται η επιθυμητή ιστοσελίδα στις πρώτες θέσεις των αποτελεσμάτων. Η λειτουργία αυτή ονομάζεται Βελτιστοποίηση στις Μηχανές Αναζήτησης (Search Engine Optimizations – SEO).

Εκτός των άλλων, τα Internet Analytics metrics είναι πολύ σημαντικά στα ηλεκτρονικά καταστήματα E-shops μιας και ο βασικός τους στόχος είναι να προσεγγίζουν τον κάθε επισκέπτη ώστε να γίνεται και καταναλωτής. Αποτελούν πολύτιμο εργαλείο για τους εμπορικούς διευθυντές καθώς βοηθούν στην διερεύνηση αποτελεσματικότερων μεθόδων promotion και marketing για την αύξηση των πωλήσεων σε ένα E-shop.

Τέλος για την καλύτερη αξιοποίηση και απόδοση των ΒΑ πρέπει να γίνουν κτήμα της λειτουργίας των τμημάτων της επιχείρησης και η εφαρμογή τους να υποστηρίζεται από την ανώτερη διοίκηση. [7]

Κεφάλαιο 9 - Επιχειρηματική Ευφυΐα (BI - Business Intelligence)

9.1 Ορισμός

Ένας οργανισμός για να έχει την ικανότητα να λαμβάνει αποτελεσματικές αποφάσεις, πρέπει να εκτελεί μια σειρά από βήματα όπου το πρώτο είναι να τεθούν στόχοι που οδηγούν στην υλοποίηση της στρατηγικής που ακολουθεί ο οργανισμός. Αφού οι στόχοι καθοριστούν, το επόμενο βήμα είναι η όσο το δυνατόν καλύτερη πληροφόρηση των στελεχών για να εξετάσουν την εγκυρότητα των στόχων και να αποφασίσουν αν μπορούν να επιτευχθούν. Για τον οργανισμό λοιπόν είναι πολύ σημαντικό να υπάρχουν ορθές πληροφορίες διαθέσιμες την κατάλληλη στιγμή στα στελέχη που θα λάβουν τις αποφάσεις. Συνεπώς η δυσκολία που υπάρχει σε αυτόν καθώς και στους υπόλοιπους οργανισμούς είναι η παροχή της ορθής πληροφορίας στα στελέχη την κατάλληλη στιγμή.

Επομένως θα μπορούσαμε να πούμε πως η Επιχειρηματική Ευφυΐα (Business Intelligence) είναι η διαδικασία παροχής της έγκυρης και χρήσιμης πληροφορίας στα κατάλληλα στελέχη του οργανισμού ώστε να λάβουν την αποτελεσματικότερη δυνατή απόφαση, εντός του χρονικού περιθωρίου, για την βέλτιστη πορεία ενός οργανισμού.

Η Επιχειρηματική Ευφυΐα δεν διακρίνεται μόνο από αριθμούς σε αναφορές και διαγράμματα. Τα δεδομένα στις αναφορές μπορούν να μας δείξουν στοιχεία για την πορεία της επιχείρησης όπως για παράδειγμα ανάλυση πωλήσεων, ισοζύγια αγορών, εβδομαδιαίο τζίρο, κτλ, όμως δεν είναι επιχειρηματική ευφυΐα. Χρειάζεται πρώτα να τεθούν σε επεξεργασία ώστε να έλθουν σε μια απλούστερη μορφή, απόλυτα κατανοητή και να χορηγηθούν έγκαιρα στα αρμόδια στελέχη του οργανισμού που λαμβάνουν αποφάσεις και χρειάζονται τις πληροφορίες αυτές.

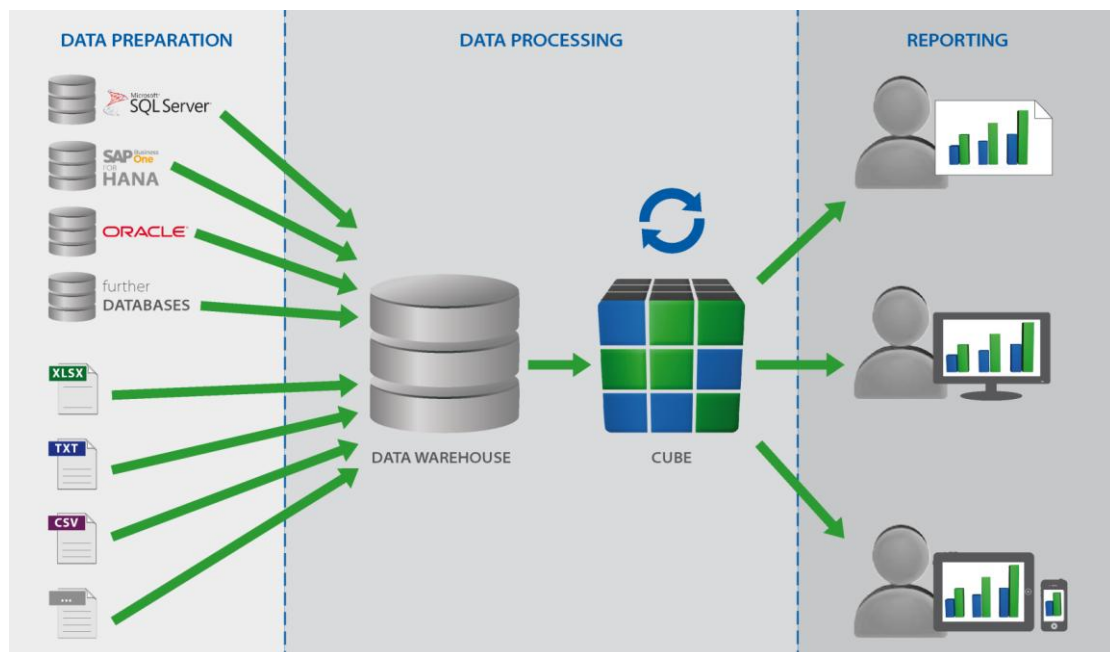
Ένας άλλος ορισμός της Επιχειρηματικής Ευφυΐας σύμφωνα με την Gartner είναι:

«Ένα μεγάλο σύνολο εφαρμογών και τεχνολογιών για την συγκέντρωση, αποθήκευση, ανάλυση, κοινοποίηση και παροχή πρόσβασης σε δεδομένα που βοηθούν τις επιχειρήσεις να λαμβάνουν καλύτερες επιχειρηματικές αποφάσεις.»

Σύμφωνα με τον παραπάνω ορισμό, η Επιχειρηματική Ευφυΐα είναι μια σειρά από μεθόδους που εκτελούνται για την αποθήκευση και παρουσίαση ορθών πληροφοριών που το κάθε αρμόδιο στέλεχος ενός οργανισμού μπορεί εύκολα και γρήγορα να ενημερώνεται και να

παίρνει απαντήσεις στα ερωτήματά του ώστε να λάβει τις καλύτερες αποφάσεις για τις μελλοντικές κινήσεις του οργανισμού. Επίσης οφείλουμε να τονίσουμε πως οι καλύτερες αποφάσεις που λαμβάνονται δεν αφορούν πάντα το εμπορικό κομμάτι, όπως για παράδειγμα αύξηση των πωλήσεων, άρα αύξηση και του κέρδους. Η Επιχειρηματική Ευφυΐα δεν αποσκοπεί μόνο στην κερδοφορία, αλλά και στην ποιότητα των αποφάσεων που λαμβάνονται από τα ανώτερα κλιμάκια των οργανισμών με την βοήθεια των BI εργαλείων που του παρέχονται.

Παρακάτω, το Διάγραμμα 6 παρουσιάζει την πορεία των δεδομένων τα οποία ανακτώνται μέσω εργαλείων Data Mining και αποθηκεύονται σε σχεσιακές ή μη σχεσιακές βάσεις δεδομένων. Στο επόμενο βήμα περνάνε από κατάλληλη επεξεργασία με την χρήση εργαλείων όπως OLAP εφαρμογές, Συστήματα Διαχείρισης Κύβων, αναφορές κ.α. Στη συνέχεια τα αποτελέσματα της επεξεργασίας των δεδομένων διαμοιράζονται σε ποικίλες μορφές (Αναφορές MS Excel, Παρουσιάσεις MS PowerPoint, εφαρμογές παρουσιάσεων για smart συσκευές) στα αρμόδια τμήματα της επιχείρησης.



Διάγραμμα 6 - Διάγραμμα Ροής Διαδικασίας Business Intelligence

9.2 Βήματα Δημιουργίας Business Intelligence Λύσεων

Η διαδικασία δημιουργίας μιας λύσης BI από την πλευρά του Business Analyst παρουσιάζεται ως εξής:

1. Καθορισμός συγκεκριμένης στρατηγικής σκέψης για τον οργανισμό όπου να γίνεται σαφές το τι πρέπει να μάθουν οι υπεύθυνοι από τα δεδομένα που χρησιμοποιούν.
2. Δημιουργία λίστας με τα δεδομένα που παράγονται συμπεριλαμβάνοντας στοιχεία πωλήσεων, προμηθευτών, πελατών, λογιστικές κινήσεις, μισθοδοσία, δραστηριότητα στο διαδίκτυο, κτλ.
3. Είναι απαραίτητο να γίνουν συζητήσεις με το προσωπικό για να εντοπιστούν αν υπάρχουν επιπλέον δεδομένα που μπορεί να χρειαστούν τα αρμόδια στελέχη ώστε να ελέγχουν με καλύτερο τρόπο την πορεία του οργανισμού.
4. Κατηγοριοποίηση των δεδομένων σύμφωνα με τον τύπο τους ώστε να είναι εύκολο από τον αναλυτή να αντιληφθεί τη λειτουργία του οργανισμού.
5. Ο αναλυτής πρέπει να ελέγχει για την ορθή καταγραφή των πληροφοριών. Για τον καλύτερο έλεγχο είναι αναγκαία η συνεργασία με τους διοικούντες του οργανισμού ώστε να επιτευχθεί η πληρότητα των δεδομένων.
6. Όταν ο αναλυτής αντιληφθεί πλήρως τα δεδομένα, έρχεται σε επικοινωνία με τους υπεύθυνους των τμημάτων για να καταγράψει τα δεδομένα στις κατάλληλες μετρικές. Δηλαδή χρειάζεται να δημιουργηθούν οι κατάλληλοι δείκτες απόδοσης (KPIs – Key Performance Indicators).
7. Σχεδίαση μιας κάρτας καταγραφής των στόχων (Scorecard) όπου θα περιλαμβάνει όλους τους δείκτες απόδοσης και θα πρέπει να αναθεωρείται από τον αναλυτή σε συνεργασία με τους υπεύθυνους των τμημάτων του οργανισμού. Αλλαγή νοοτροπίας της λειτουργίας των τμημάτων του οργανισμού και προσαρμογή της λειτουργίας σύμφωνα με τους νέους δείκτες απόδοσης που καθιερώθηκαν.

8. Αφού λοιπόν καθιερωθεί η κάρτα επιδόσεων, θα γίνουν οι διορθωτικές κινήσεις σε θέματα που μπορεί να προκύψουν και θα καθοριστούν πρότυπα αναφορών που θα χρειαστούν στο να διαμοιράσουν την αξία της πληροφορίας σε όλο τον οργανισμό.
9. Σχεδιασμός του συστήματος BI που θα χρησιμοποιεί ο αναλυτής ώστε να υπάρχει αυτόματη ροή των δεδομένων που χρειάζεται για να λειτουργήσει.
10. Χρειάζεται να δημιουργήσει κάρτες απόδοσης που θα χρησιμοποιούνται από το σύστημα και να αυτοματοποιήσει τον τρόπο που εξάγονται οι αναφορές με τα αποτελέσματα των δεικτών απόδοσης.

Το τελευταίο βήμα δημιουργεί το BI σύστημα που με την λειτουργία του θα υποστηρίζει την αποτελεσματικότητα των τμημάτων του οργανισμού και θα μεγιστοποιήσει τις ευκαιρίες ανάπτυξης του.

9.3 Business Intelligence Tools

9.3.1 OLAP Reports

Οι αναφορές Πολυδιάστατης Αναλυτικής Επεξεργασίας Δεδομένων (Online Analytical Processing – OLAP) είναι μια προσέγγιση απάντησης πολυδιάστατων αναλυτικών ερωτημάτων σε σύντομο χρονικό διάστημα. Τα OLAP είναι μέρος της ευρύτερης κατηγορίας των επιχειρηματικών πληροφοριών, τα οποία περιλαμβάνουν επίσης συστήματα σχεσιακών βάσεων δεδομένων, σύνταξη εκθέσεων και εξόρυξη δεδομένων. [8] Ο χρήστης με την χρήση εργαλείων OLAP μπορεί με μεγάλη ευκολία να προβάλει τα δεδομένα του από πολλές διαφορετικές οπτικές γωνίες, να εμφανίσει στοιχεία που τον ενδιαφέρουν όπως επίσης και να αποκρύψει ασήμαντα δεδομένα και μέσω της επεξεργασίας πολλαπλών επιπέδων να καταλήξει σε σημαντικά συμπεράσματα που θα τον βοηθήσουν στην λήψη καλύτερων αποφάσεων.

Για να μπορούν οι OLAP αναφορές να υποστηρίζουν πολυδιάστατη επεξεργασία δεδομένων, πρέπει να συνδέονται με πολυδιάστατες βάσεις δεδομένων. Μια απλή σχεσιακή βάση θεωρείται δισδιάστατη, ενώ για πολυδιάστατη βάση δεδομένων κάθε χαρακτηριστικό αποτελεί και μια ξεχωριστή διάσταση. Τα λογισμικά OLAP εντοπίζουν τη διατομή των διαστάσεων και την εμφανίζουν στην οθόνη. Με απλά λόγια χρησιμοποιώντας τις αναφορές OLAP μπορούμε σε μια οθόνη να εμφανίσουμε πολλές διαφορετικές διαστάσεις αναλύοντας για παράδειγμα πόση ποσότητα από ποια προϊόντα πωλήθηκαν σε ποια περιοχή για ποια χρονική περίοδο. Επίσης διαστάσεις – χαρακτηριστικά όπως χρονικά διαστήματα μπορούν να χωριστούν σε υποκατηγορίες για την διευκόλυνση της ανάλυσης μας.

Οι OLAP βάσεις δεν είναι τόσο μεγάλες όσο οι αποθήκες δεδομένων, αφού για την υλοποίηση μιας ανάλυσης δεν χρειάζονται όλα τα διαθέσιμα δεδομένα. Με την χρήση Επικοινωνίας Ανοικτών Βάσεων Δεδομένων (Open Database Connectivity – ODBC) μπορούμε με απλά βήματα να εισάγουμε δεδομένα από μια υπάρχουσα βάση και να δημιουργήσουμε για OLAP βάση για να κάνουμε τις αναλύσεις μας. [9]

9.3.2 KPI

Ο όρος KPI προέρχεται από τις λέξεις Key Performance Indicators (Καίριοι Δείκτες Απόδοσης) και κάποιες φορές αναφέρεται ως Καίριοι Δείκτες Επιτυχίας (Key Success Indicators, KSIs) και χρησιμοποιείται σε μια επιχείρηση για την καταμέτρηση της επίδοσης της ομαδάς αλλά και του κάθε εργαζομένου ξεχωριστά όσον αφορά την επίτευξη στόχων που έχουν τεθεί σε ποσοτικό, επιχειρησιακό και στρατηγικό επίπεδο.

Αρμόδιοι για την χρήση τους είναι τα ανώτερα κλιμάκια όπου χρειάζεται να αξιολογούν τις ενέργειες του κάθε τμήματος της επιχείρησης και να προσδιορίσουν τον βαθμό της επιτυχίας τους. Με τον τρόπο αυτό προσδιορίζεται η επιχειρηματική επιτυχία που ερμηνεύεται ως η πρόοδος μιας εταιρίας βάσει της επίτευξης των προκαθορισμένων στόχων της. Η αξιολόγηση συντελεί στην καλύτερη αποσαφήνιση των προβλημάτων που προκύπτουν από τις επιχειρησιακές λειτουργίες.

Για να φτάσει όμως μια επιχείρηση στο σημείο να επιτεύξει τους στόχους της, πρέπει να περάσει από ένα άλλο βασικό στάδιο πρώτα. Το στάδιο της στοχοθεσίας. Επίσης οι διαφορετικές τεχνικές επίτευξης στόχων κάθε εταιρίας, η αποσαφήνιση των προβλημάτων και ο προσδιορισμός της σημαντικότητας του στόχου καθώς και η επανασχεδίαση της στρατηγικής όταν προσεγγίζεται η επίτευξη του στόχου, καθορίζουν σημαντικά τους Καίριους Δείκτες Απόδοσης που επιλέγει η κάθε εταιρία. Για την επιλογή του καταλληλότερου Δείκτη, παίζει πολύ σημαντικό ρόλο η μετρησιμότητα του διότι όπως συνηθίζεται ότι δεν μπορεί να μετρηθεί δύσκολα αξιολογείται.

Ενδεικτικά παραδείγματα δεικτών:

- Για ένα πλυντήριο αυτοκινήτων, ο μέσος όρος αυτοκινήτων που πλύθηκαν ανά ημέρα
- Για μια επιχείρηση διανομής, ο μέσος όρων παραγγελιών που διανέμει σε μια ημέρα
- Για ένα πολυκατάστημα, ο αριθμός των κινητών τηλεφώνων που πούλησε κατά τη διάρκεια μιας εκπτώσεων ή μιας προσφοράς.

Κεφάλαιο 10 - Μελλοντικά

Το μέλλον του ΣΔΒΔ

Τα συστήματα διαχείρισης βάσεων δεδομένων της αγοράς συνεχίζουν να αυξάνουν ολοένα και περισσότερο την τάση να χρησιμοποιούν τεχνολογίες NoSQL για την αποθήκευση και διαχείριση των δεδομένων τους. Φυσικά, η «κλασική» SQL πούλησε ένα τεράστιο ποσό των λογισμικών συστημάτων διαφορετικής πολυπλοκότητας με τους γίγαντες της αγοράς, όπως η Microsoft και η Oracle. Αλλά με βεβαιότητα μπορούμε να υποθέσουμε ότι τα συστήματα NoSQL τα επόμενα χρόνια θα κινητοποιήσουν σημαντικά ηγέτες να ασχοληθούν με την ανάπτυξη σύνθετων πληροφοριακών συστημάτων.

Τα Big Data θα αλλάξουν τις επιχειρήσεις και οι επιχειρήσεις με τη σειρά τους θα αλλάξουν την κοινωνία. Η ελπίδα βέβαια, είναι ότι τα οφέλη θα αντισταθμίσουν τα μειονεκτήματα. Ο κόσμος των Big Data είναι καινούριος και ως κοινωνία δεν είμαστε καλοί στον χειρισμό όλων των δεδομένων που μπορούμε να συλλέξουμε τώρα. Η τεχνολογία θα συνεχίσει να μας εκπλήσσει.

Βιβλιογραφία και Πηγές

- [1] http://www.sas.com/el_gr/insights/big-data/what-is-big-data.html
- [2] http://delab.csd.auth.gr/courses/c_dbimpl/parallel.pdf
- [3] <https://www.mapr.com/products/apache-hadoop>
- [4] <https://dwarehouse.wordpress.com/2012/12/28/introduction-to-massively-parallel-processing-mpp-database/>
- [5] https://dspace.lib.uom.gr/bitstream/2159/13821/1/Iosifidis_Msc2010.pdf
- [6] <http://83.212.168.57/jspui/bitstream/123456789/2841/1/022006x03x156.pdf>
- [7] http://www.businessanalytics.gr/index.php?option=com_content&view=section&layout=blog&id=4&Itemid=123
- [8] https://en.wikipedia.org/wiki/Online_analytical_processing
- [9] <http://searchdatamanagement.techtarget.com/definition/OLAP>
- [10] <http://artemis-new.cslab.ece.ntua.gr:8080/jspui/bitstream/123456789/7129/1/DT2014-0265.pdf>
- [11] <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>
- [12] www.businessanalytics.gr
- [13] Σταυλιώτης Ε. Γεράσιμος (2009): Εξόρυξη Δεδομένων και Αναγνώριση προτύπων σε κατηγορικά δεδομένα μέσω συσταδοποίησης, Ελληνικό Στατιστικό Ινστιτούτο
- [14] Dean, Jeffrey και Sanjay Ghemawat: MapReduce: simplified data processing on large clusters. Commun. ACM, 51(1):107–113, 2008, ISSN 0001-0782
- [15] Hadoop MapReduce. <http://hadoop.apache.org/mapreduce/>.
- [16] Cassandra - A Decentralized Structured Storage System.pdf Avinash Lakshman-Facebook, Prashant Malik-Facebook