ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

των

ΑΛΕΞΙΟΥ ΑΝΝΑ (ΑΜ 2376)

ΤΡΕΜΟΠΟΥΛΟΥ ΗΛΕΚΤΡΑ (ΑΜ 2491)

Επιβλέπων : ΔΗΜΟΚΑΣ ΝΙΚΟΛΑΟΣ

Επίκουρος Καθηγητής

Καστοριά 3/10/2024

ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΜΑΚΕΔΟΝΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΤΕΧΝΙΚΕΣ ΑΝΑΛΥΣΗΣ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

των

ΑΛΕΞΙΟΥ ΑΝΝΑ (ΑΜ 2376)

ΤΡΕΜΟΠΟΥΛΟΥ ΗΛΕΚΤΡΑ (ΑΜ 2491)

Επιβλέπων : ΔΗΜΟΚΑΣ ΝΙΚΟΛΑΟΣ

Επίκουρος Καθηγητής

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 3/10/2024

.....	.....	.....
Ον/μο Μέλους	Ον/μο Μέλους	Ον/μο Μέλους
Ιδίοτητα Μέλους	Ιδίοτητα Μέλους	Ιδίοτητα Μέλους

Καστοριά 3/10/2024

Ευχαριστίες

Με το πέρας της πτυχιακής μας εργασίας, θα θέλαμε να ευχαριστήσουμε θερμά τον καθηγητή κ. Δημόκα Νικόλαο για την πολύτιμη βοήθεια και εμπιστοσύνη που μας έδειξε. Επιπλέον η επιμονή και η υποστήριξη του ήταν καθοριστική για την ολοκλήρωση της εργασίας μας.

Ωστόσο, ένα μεγάλο ευχαριστώ στις οικογένειες μας και σε όλα τα κοντινά μας πρόσωπα που με την επιμονή που έδειξαν μας βοήθησαν να πετύχουμε τους ακαδημαϊκούς μας στόχους.

Τέλος, να μην παραληφθεί η άψογη συνεργασία μας που με κόπο και κατανόηση έκανε το έργο μας πιο εύκολο.

Περίληψη

Από την εφεύρεση των ηλεκτρονικών υπολογιστών έως τη σημερινή ψηφιακή εποχή, ο πολλαπλασιασμός των δεδομένων έχει φτάσει σε πρωτοφανή επίπεδα, οδηγώντας στην εποχή των Μεγάλων Δεδομένων. Ο σημερινός παγκόσμιος πληθυσμός υπερβαίνει τα 8,1 δισεκατομμύρια ενώ σύμφωνα με την Forbes πάνω από 5,35 δισεκατομμύρια από αυτούς τους ανθρώπους είναι συνδεδεμένοι στο διαδίκτυο[1]. Οι εξελίξεις στις κινητές συσκευές, τους ψηφιακούς αισθητήρες, τις επικοινωνίες και την αποθήκευση απαίτησαν μέσα για τη συλλογή ψηφιακών πληροφοριών και την αξιοποίηση του τεράστιου αυτού όγκου δεδομένων στην εξαγωγή πολύτιμων πληροφοριών για τη λήψη αποφάσεων. Τα μεγάλα δεδομένα είναι ένας νέος όρος που προήλθε από την ανάγκη μεγάλων εταιρειών, όπως η Yahoo, η Google και το Facebook, να αναλύουν τη ροή αυτών των συνεχώς αυξανόμενων ποσοτήτων δεδομένων. Τις παραμέτρους του όρου αυτού επιχειρούμε να αναλύσουμε στην παρούσα εργασία.

Στην εισαγωγική ενότητα, εμβαθύνουμε στις θεμελιώδεις έννοιες των Μεγάλων Δεδομένων, ξεκινώντας από τον ορισμό τους. Εξετάζοντας το πολύπλευρο τοπίο των Μεγάλων Δεδομένων, στοχεύουμε να αναλύσουμε τις ρίζες τους , να εξηγήσουμε την ουσία τους και να διαφωτίσουμε τη σημασία και τις δυνατότητές τους σε πολλούς κλάδους της σημερινής εποχής. Τα μεγάλα δεδομένα αναφέρονται σε σύνολα δεδομένων που χαρακτηρίζονται από τον τεράστιο όγκο, την ταχύτητα, την ποικιλία και την ακεραιότητα τους. Ασχολούμαστε από τις απαρχές των Μεγάλων Δεδομένων έως την ταχεία τους επέκτασή που οφείλεται στην πρόοδο της τεχνολογίας. Εμβαθύνουμε στις θεμελιώδεις έννοιες και τα χαρακτηριστικά των Μεγάλων Δεδομένων θέτοντας προκλήσεις για τις παραδοσιακές τεχνικές αποθήκευσης επεξεργασίας και ανάλυσης τους. Στη συνέχεια, διερευνούμε τους διαφορετικούς τύπους και τα χαρακτηριστικά των Μεγάλων Δεδομένων, συμπεριλαμβανομένων των δομημένων, μη δομημένων και ημιδομημένων δεδομένων, καθένα από τα οποία απαιτεί εξειδικευμένη προσέγγιση για την ανάλυση και την αποθήκευση.

Ξεπερνώντας τη θεωρία, εμβαθύνουμε στις πρακτικές εφαρμογές των μεγάλων δεδομένων σε διάφορους τομείς όπως η υγεία, οι τηλεπικοινωνίες, τα μέσα κοινωνικής δικτύωσης. Εξετάζουμε τον τρόπο με τον οποίο οι οργανισμοί αξιοποιούν την ανάλυση των Μεγάλων Δεδομένων για τη βελτιστοποίηση των λειτουργιών και τη βελτίωση της εμπειρίας των πελατών. Ενώ αναφερόμαστε στην ευαισθησία με την οποία οφείλει κανείς να διαχειρίζεται τα δεδομένα αυτά προστατεύοντας την πνευματική ιδιοκτησία και τα προσωπικά δεδομένα.

Η συζήτηση επεκτείνεται στο πεδίο των βάσεων δεδομένων, όπου διερευνούμε την εξέλιξη της τεχνικής της αποθήκευσης δεδομένων και τους διαφορετικούς τύπους βάσεων που χρησιμοποιούνται στην ανάλυση μεγάλων δεδομένων επισημαίνοντας τα μοναδικά χαρακτηριστικά τους. Δίνουμε έμφαση στο σχεσιακό και μη σχεσιακό σύστημα διαχείρισης βάσεων και το ρόλο τους στην υποστήριξη της αποδοτικής επεξεργασίας μεγάλων δεδομένων. Ενώ εμβαθύνουμε στην αρχιτεκτονική και τον τρόπο λειτουργίας κάποιων σπουδαίων βάσεων δεδομένων που άλλαξαν την ιστορία της τεχνολογίας όπως η Apache Cassandra, η MongoDB, η Neo4j .Επιπλέον, εκμεταλλευόμενοι τα πλαίσια κάποιων σημαντικών λογισμικών στον τομέα όπως το Hadoop και το Elasticsearch, κατανοούμε καλύτερα την τεχνική της ανάλυσης των μεγάλων δεδομένων. Αυτά τα εργαλεία επιτρέπουν την κατανεμημένη επεξεργασία σε πραγματικό χρόνο και την προηγμένη ανάλυση, δίνοντας τη δυνατότητα στους οργανισμούς να ξεκλειδώσουν το πλήρες δυναμικό των Μεγάλων Δεδομένων. Στην καταληκτική ενότητα, παραθέτουμε ιδέες και ανησυχίες για το μέλλον της ανάλυσης Μεγάλων Δεδομένων, συζητώντας αναδυόμενες τάσεις όπως η τεχνητή νοημοσύνη και τη σημασία των μέτρων ασφαλείας για τη διασφάλιση της ιδιωτικότητας και της ακεραιότητας των δεδομένων.

Η παρούσα εργασία χρησιμεύει ως μια ολοκληρωμένη διερεύνηση της ανάλυσης μεγάλων δεδομένων, από τον ορισμό, την ιστορική εξέλιξη και τις εφαρμογές έως τις τεχνολογίες των βάσεων δεδομένων και των λογισμικών καθώς και τις μελλοντικές προοπτικές των Μεγάλων Δεδομένων

Λέξεις Κλειδιά: Μεγάλα Σύνολα δεδομένων, Ανοιχτά Δεδομένα, Βάσεις Δεδομένων

Abstract

The rise of smartphones, sensors, communication networks and storage has created a need to gather and analyze massive amounts of information to gain insights for better decision-making. This has led to the emergence of big data, a term used to describe the challenges faced by large organizations in handling the ever-growing stream of data. This paper aims to explore the concept of big data in detail.

The first section will provide a foundational understanding of big data, including its definition. We will then examine the various aspects of big data, exploring its roots, core characteristics, and its significance and potential applications across various fields today. Big data refers to extremely large and complex datasets. We will trace the development of big data, from its beginnings to its explosive growth driven by technological advancements. Finally, we will delve into the fundamental principles and defining features of big data, highlighting the challenges it poses to traditional data storage, processing, and analysis methods. We then explore the different types and characteristics of Big Data, including structured, unstructured and semi-structured data, each of which requires a specialized approach for analysis and storage. Moving beyond theory, we delve into the practical applications of Big Data in various domains such as healthcare, telecommunications, social media. We explore how organizations are leveraging Big Data analytics to optimize operations and improve customer experience. While we address the sensitivity with which one must manage this data by protecting intellectual property and personal data. Next, we'll dive into databases. We'll explore how data storage has changed and the different types of databases used for Big Data, explaining what makes each of them special. We'll focus on relational and non-relational databases and how they help process Big Data efficiently. While we delve into the architecture and mode of operation of some great databases that changed the history of technology such as Apache Cassandra, MongoDB, Neo4j. Furthermore, leveraging the functionalities of prominent software frameworks like Hadoop and Elasticsearch, we can gain a deeper understanding of big data analysis techniques. These tools enable real-time distributed processing and advanced analytics, giving the chance for organizations to unlock the full potential of Big Data.

To conclude, this paper presents a comprehensive examination of big data analytics. We embark on this exploration by establishing a foundational understanding of the term itself, then delve into its historical development and practical applications. Subsequently, we investigate the role of database and software technologies, culminating in a discussion on the anticipated future prospects of big data.

Πίνακας Περιεχομένων

Λίστα Εικόνων

Ακρώνυμα και Συντομογραφίες

IBM	International Business Machines Corporation
HTML	Hyper Text Markup Language
XML	Extensible Markup Language
GPS	Global Positioning System
NSA	National Security Agency
PII	Personal Identifiable Information
BI	Business Intelligence
IoT	Internet Of Things
JSON	JavaScript Object Notation
PSI	Public Sector Information
RDBMS	Relational Database Management System
SQL	Structured Query Language
NoSQL	Not only Structured Query Language
OLAP	Online Analytical Processing
AWS	Amazon Web Services
HDFS	Hadoop Distributed File System
API	Application Programming Interface
HA	High Availability
GCP	Google Cloud Platform
KMS	Key Management Service
ASF	Apache Software Foundation

IO	Input Output
MR	Map Reduce
DAG	Directed Acyclic Graph
REST	Representational State Transfer

1 ΕΙΣΑΓΩΓΗ

1.1 Ορισμός των Μεγάλων Δεδομένων

Οι περισσότερες εφευρέσεις σε όλη την ανθρωπότητα αφορούν την δημιουργία γλωσσών και λεξιλογίου. Με τα χρόνια η εξέλιξη στην εφεύρεση υπολογιστών είχε ως αντίκτυπο την παραγωγή της πληροφορίας και των δεδομένων. Τον τελευταίο καιρό πληθαίνουν όλο και περισσότερο οι συζητήσεις αλλά και τα σεμινάρια γύρω από τις τεχνικές και την ορολογία των big data. Τα big data ή αλλιώς «Μεγάλα Σύνολα Δεδομένων» έχουν αποκτήσει καθοριστικό ρόλο σε διάφορους τομείς της ζωής μας, καθώς και στην ανάπτυξη της ψηφιακής μας πληροφορίας. Ο ορισμός τους διαφέρει και προσεγγίζεται με διάφορες σημασίες από γνώστες του αντικειμένου. Πρόκειται για δεδομένα τόσο μεγάλου όγκου που η επεξεργασία αποθήκευσης και ανάλυση με τις παραδοσιακές μεθόδους καθίσταται σχεδόν αδύνατη. Αυτά τα σύνολα δεδομένων αποτελούνται συνήθως από δομημένα (structured), ημιδομημένα (semi-structured) και αδόμητα (unstructured) δεδομένα από διάφορες πηγές, όπως μέσα κοινωνικής δικτύωσης, αισθητήρες, ηλεκτρονικές συσκευές, συναλλαγές και άλλες ψηφιακές αλληλεπιδράσεις. Καθώς οι ανάγκες της εποχής αλλάζουν και η τεχνολογία εξελίσσεται ο όγκος των δεδομένων αυξάνεται δραματικά. Πολλές επιχειρήσεις στρέφονται στο ηλεκτρονικό εμπόριο, γεγονός που αυξάνει την παραγωγή δεδομένων. Με αυτόν τον τρόπο, τα Μεγάλα Σύνολα Δεδομένων στον κόσμο των επιχειρήσεων κυμαίνονται από μερικές δεκάδες terabytes έως και χιλιάδες petabytes. Για να γίνει πιο σαφές και κατανοητό το μέγεθος των μεγάλων δεδομένων παραθέτουμε τον παρακάτω πίνακα μονάδων μέτρησης δεδομένων [2].

Acronym	Description	Size
(B)	Byte	8 bits
(KB)	Kilobyte	1,024 bytes
(MB)	Megabyte	1,048,576 bytes
(GB)	Gigabyte	1,073,741,824 bytes
(TB)	Terabyte	1,099,511,627,776 bytes
(PB)	Petabyte	1,125,899,906,842,624 bytes
(EB)	Exabyte	1,152,921,504,606,846,976 bytes
(ZB)	Zettabyte	1,180,591,620,717,411,303,424 bytes
(YB)	Yottabyte	1,208,925,819,614,629,174,706,176 bytes

Εικόνα 1: πίνακας μονάδας μέτρησης μεγάλων δεδομένων

Γνωρίζουμε ότι στα συστήματα υπολογιστών, μια μονάδα δεδομένων με μήκος οκτώ δυαδικών ψηφίων είναι γνωστή ως byte. Το byte είναι η μονάδα που χρησιμοποιούν οι υπολογιστές για την αναπαράσταση ενός χαρακτήρα, όπως ένα γράμμα, ένας αριθμός ή ένα σύμβολο για παράδειγμα, "A", "3" ή "#". Η αποθήκευση στον υπολογιστή μετριέται γενικά σε πολλαπλάσια του byte.

Ένα zettabyte είναι μια τεράστια ποσότητα αποθήκευσης δεδομένων. Ισοδυναμεί με περίπου 10^{21} bytes. Για να το θέσουμε σε ένα πλαίσιο, αν κάθε χαρακτήρας στη μέση αγγλική λέξη (συμπεριλαμβανομένων των κενών) ήταν ένα μόνο byte, ένα zettabyte θα μπορούσε να χωρέσει περίπου 1 τρισεκατομμύριο αντίγραφα ολόκληρου του περιεχομένου της Wikipedia. Ωστόσο, τα όρια των δεδομένων θεωρούνται « Μεγάλα » ανάλογα με την εξέλιξη της επεξεργαστικής ισχύος των υπολογιστών καθώς και την αύξηση του διαθέσιμου αποθηκευτικού χώρου. Συνεπώς καταλαβαίνουμε ότι αυτά θα μεταρρυθμίζονται και θα αυξάνονται συνεχώς και σε τακτά χρονικά διαστήματα τα επόμενα χρόνια.

1.2 Ιστορική Αναδρομή

Είναι γεγονός πως με την πάροδο του χρόνου η τεχνολογία έχει αναπτυχθεί ραγδαία. Η ανάγκη όμως για την αποθήκευση της πληροφορίας δεν είναι κάτι καινούριο. Πιο συγκεκριμένα, οι προσπάθειες αποθήκευσης και διαχείρισης δεδομένων από τον άνθρωπο υπήρχαν εδώ και χιλιάδες χρόνια, πολύ πριν ανακαλυφθούν οι υπολογιστές. Τον τελευταίο αιώνα ωστόσο, η εφεύρεση ψηφιακών μέσων αποθήκευσης και η εμφάνιση του διαδικτύου είχαν καθοριστικό ρόλο στην εξέλιξη του τομέα των μεγάλων δεδομένων (big data). Η ιστορία των μεγάλων δεδομένων χωρίζεται συνοπτικά σε τρία κύρια στάδια: τη μετάβαση από τα megabyte στα gigabyte, την έκρηξη από τα gigabyte στα terabyte, και τελικά την μετάβαση από τα terabyte στα petabyte [3].

Πιο αναλυτικά, την δεκαετία του 1950, οι άνθρωποι για να μπορέσουν να αποθηκεύσουν αρχεία χρησιμοποιούσαν τις λεγόμενες «Διάτρητες κάρτες» και μαγνητικές ταινίες. Κατά τη διάρκεια αυτής της περιόδου, η επεξεργασία δεδομένων περιοριζόταν κυρίως σε δομημένα δεδομένα αποθηκευμένα σε παραδοσιακές βάσεις δεδομένων και ο όγκος των δεδομένων ήταν σχετικά μικρός σε σύγκριση με τα σύγχρονα πρότυπα. Η εξέλιξη για συσκευές μαζικής αποθήκευσης και αύξηση υπολογιστικής ισχύος προϋπόθετε συστήματα διαχείρισης δεδομένων. Την δεκαετία του 1960, δημιουργήθηκαν συστήματα διαχείρισης βάσεων δεδομένων για να αντικαταστήσουν τα μεμονωμένα αρχεία. Κύριος στόχος δημιουργίας τους, ήταν η οργάνωση και η διαχείριση των δεδομένων. Παράλληλα το 1964, ο Charles Bachman της General Electric, πρότεινε τον σχεδιασμό δικτυωτού μοντέλου δεδομένων, στο οποίο οι εγγραφές ήταν συνδεδεμένες μεταξύ τους σχηματίζοντας τεμνόμενα σύνολα δεδομένων. Κατά την περίοδο 1965, η εταιρία IBM (International Business Machines Corporation) σε συνεργασία με την διεύθυνση διαστήματος North American Aviaton ανέπτυξαν ιεραρχικό μοντέλο δεδομένων. Αυτό το μοντέλο παρουσιάζει τα δεδομένα ως δομές που μοιάζουν με δένδρα, οργανωμένες σε μια ιεραρχία εγγράφων. Ωστόσο, με την πάροδο του χρόνου το 1970, ο Edgar Codd, έδωσε τον ορισμό του σχεσιακού μοντέλου, το οποίο αναπτύχθηκε και κατέλαβε σημαντικό μερίδιο στην αγορά λόγω της ευελιξίας και της απλότητάς του σε σύγκριση με το ιεραρχικό μοντέλο.

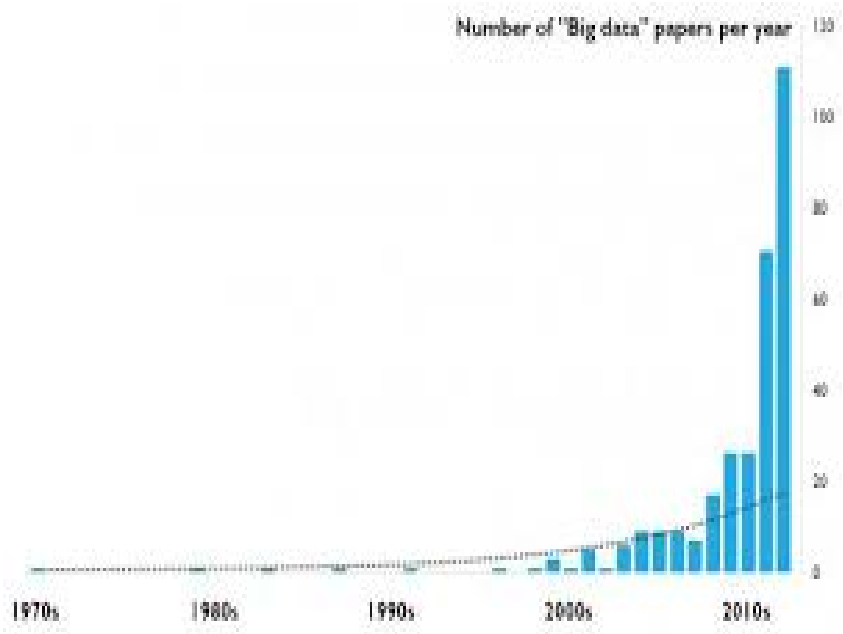
Την επερχόμενη δεκαετία ο P.P Chen περιέγραψε το μοντέλο οντοτήτων – σχέσεων (ER – Rentity Relationship model). Ένα μοντέλο που εξηγεί τα δεδομένα χρησιμοποιώντας γραφικά σύμβολα για να αποτυπώσει οντότητες, συσχετίσεις και γνωρίσματα. Το 1979, όλες οι εμπορικές υλοποιήσεις βάσεων δεδομένων βασίζονταν σε δικτυωτή ή ιεραρχική προσέγγιση. Επιπλέον, την ίδια χρονιά ιδρύθηκε η εταιρία Relational Software Incorporated και κυκλοφόρησε στην αγορά η σχεσιακή βάση ORACLE V.2. Κατά αυτά τα χρόνια, αναπτύχθηκαν σημαντικά προγράμματα σχεσιακών

συστημάτων όπως το INGRESS και το System R (IBM), καθώς και γλώσσες που συμπλήρωναν αυτά τα προγράμματα, όπως το SEQUEL, το QBE και το QUEL.

Στις δεκαετίες του 1980 και του 1990, η άνοδος της αποθήκευσης δεδομένων και των εργαλείων επιχειρηματικής ευφυΐας επέτρεψε στους οργανισμούς να συλλέγουν και να αναλύουν μεγαλύτερα σύνολα δεδομένων. Εκείνη την περίοδο ο Jim Gray, γνωστός ως ερευνητής και σχεδιαστής λογισμικού, συνέβαλε σε πολλά μεγάλα συστήματα επεξεργασίας βάσεων δεδομένων και συναλλαγών. Το System R της IBM ήταν ο προκάτοχος των σχεσιακών βάσεων δεδομένων SQL που έχουν γίνει πλέον πρότυπο σε όλο τον κόσμο για τη διαχείριση δομημένων δεδομένων. Η περίοδος αυτή καθορίστηκε και από άλλα σχεσιακά συστήματα και κατανεμημένες βάσεις όπως Oracle, Server και Informix.

Το 1990 η έλευση του διαδικτύου και ο πολλαπλασιασμός των διαδικτυακών υπηρεσιών οδήγησαν σε έκρηξη των δεδομένων που παράγονται από ιστότοπους, πλατφόρμες κοινωνικής δικτύωσης και άλλες ψηφιακές πηγές. Επιτεύχθηκε η επικοινωνία μεταξύ των χρηστών με βάσεις δεδομένων μέσω διαδικτύου (html, asp, xml), ενώ το 1991, με την βοήθεια του πρωτοκόλλου μεταφοράς (http) ο παγκόσμιος ιστός είναι πλέον το μέσο ανταλλαγής πληροφοριών. Το 1995, η αντικειμενοστραφής γλώσσα JAVA από την εταιρία Sun Microsystems γίνεται το βασικό εργαλείο της εποχής μετά την γλώσσα C. Καθιερώνεται στο διαδίκτυο και χρησιμοποιείται σε εφαρμογές μεσαίου μεγέθους. Αυτή η εποχή ήρθε αντιμέτωπη με την εμφάνιση μη δομημένων και ημιδομημένων τύπων δεδομένων, όπως κείμενο, εικόνες, βίντεο και δεδομένα αισθητήρων.

Την δεκαετία 1998, αναπτύχθηκε από τον Carlo Strozzi, ένα σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων με ονομασία NoSQL. Την ίδια χρονολογία ο Larry Page μαζί με τον Sergey Brin, ίδρυσαν την Google σαν ιδιωτική εταιρία. Το 1999, γίνεται πιο διαδεδομένος ο όρος Μεγάλα Σύνολα Δεδομένων μέσω ακαδημαϊκών άρθρων.



Εικόνα 2: Γράφημα δημοσιεύσεων σχετικά με τα Big Data ετησίως

Το 2001, προσδιορίζεται το μοντέλο «των 3V». Πρόκειται για τα τρία χαρακτηριστικά των μεγάλων δεδομένων (Volume, Velocity, Variety) τα οποία θα αναλύσουμε εκτενέστερα στην συνέχεια. Το 2006, πρωτοεμφανίστηκε και το Apache Hadoop από τους Doug Cutting και Mike Cafarella, μία συλλογή βοηθητικών προγραμμάτων λογισμικού ανοιχτού κώδικα παίζει καθοριστικό ρόλο στη δυνατότητα κατανεμημένης επεξεργασίας μεγάλων συνόλων δεδομένων σε δίκτυα υπολογιστικών πόρων.

Το 2008, οι επεξεργαστές κεντρικών μονάδων μπορούν πλέον να διαχειριστούν 9,25 Zettabytes πληροφοριών. Ο αριθμός συνδεδεμένων συσκευών στον παγκόσμιο δίκτυο ξεπερνά τον παγκόσμιο πληθυσμό. Το 2011, κάνει την εμφάνιση του από την IBM ο υπολογιστής Watson. Αναπτύχθηκε από ερευνητική ομάδα με σκοπό να απαντά σε ερωτήσεις που τίθενται σε φυσική γλώσσα. Έχει δυνατότητες σάρωσης και ανάλυσης έως 4 terabytes ενώ την ίδια χρονιά γράφεται ιστορία νικώντας για πρώτη φορά τον ανθρώπινο εγκέφαλο στο κουίζ show “Jeopardy”. Το 2012, σύμφωνα με την έρευνα που διενέργησε η IDC (Interactive Data Corporation) υπήρχαν 1,8 zettabytes (δηλαδή 1,8 τρισεκατομμύρια gigabytes) πληροφοριών που δημιουργήθηκαν μόνο το έτος 2011. Το ποσό αυτό διπλασιάζεται κάθε δύο χρόνια. Έτσι ο όγκος των πληροφοριών την επόμενη δεκαετία θα αυξηθεί κατά 50 φορές.

Το 2013, ξεκίνησε η δημοκρατικοποίηση των δεδομένων. Μέσω ασύρματων συνδέσεων WiFi και τηλέφωνα νέας γενιάς η παραγωγή δεδομένων είχε ταχύτατους ρυθμούς. Με τον καιρό ολοένα και περισσότεροι άνθρωποι είχαν πρόσβαση σε μεγάλους όγκους δημόσιων δεδομένων. Το 2015, είναι η πρώτη χρονιά όπου η σύνδεση στο διαδίκτυο γινόταν περισσότερο από τα κινητά τηλέφωνα παρά από τα computers ή

τα laptop. Αυτό είχε σαν αποτέλεσμα ο αριθμός των χρηστών διαδικτύου να ξεπεράσει το όριο από κάθε άλλη φορά. Την ίδια χρονιά έγινε γνωστή η είδηση ότι η Microsoft αγόρασε το Revolution Analytics. Οι γλώσσες προγραμματισμού ανοιχτού κώδικα είναι πλέον πιο δημοφιλής στον κόσμο για στατιστικούς υπολογισμούς και προγνωστικά αναλυτικά. Επιπλέον, πραγματοποιήθηκε έκθεση στον Λευκό Οίκο, σχετικά με τους κινδύνους και τις ευκαιρίες που δημιουργούν τα big data καθιστώντας τα πλέον κρατικό θέμα. Το συμπέρασμα ήταν πως ενώ η τεχνολογία προσφέρει δυνατότητες ανάπτυξης και αλλαγής, υπάρχουν ανησυχίες σχετικά με την προστασία ιδιωτικής ζωής και δεδομένων που θα πρέπει να αντιμετωπιστούν με νομοθεσία.

Το 2015 δεν συνέβησαν όμως μόνο αυτά. Η IBM ανακοίνωσε πως θα επενδύσει 3 δισεκατομμύρια δολάρια σε τεχνολογίες Internet τα επόμενα 4 χρόνια. Η καινοτόμος ιδέα πως όλες οι συσκευές -όχι μόνο οι υπολογιστές- μπορούν να μοιράζονται πληροφορίες και να αλληλοεπιδρούν μεταξύ τους αποτελεί μεγάλο πλεονέκτημα. Επιπλέον, τα big data παρουσίασαν νέες διευρυμένες υπηρεσίες. Για παράδειγμα, το Amazon Web Services ανακοίνωσε ότι οι πελάτες θα μπορούσαν σύντομα να επωφεληθούν από την μηχανική εκμάθηση μέσω του Amazon Machine Learning ενάντια των Azure της Microsoft και IBM Watson. Από την άλλη η Google ανακοίνωσε σημαντικές ενημερώσεις στο Big Query. Επιπρόσθετα, την χρονιά αυτή η κινέζικη υπηρεσία λιανικής και διαδικτύου Alibaba παρουσιάζει την Aliyum. Πρόκειται για μηχανή ανάλυσης cloud που σχεδιάστηκε ώστε οι επιχειρήσεις να πραγματοποιούν ανάλυση με γνώμονα την μηχανική μάθηση σε δικό της υλικό. Η Alibaba την αποκαλεί «πρώτη πλατφόρμα τεχνητής νοημοσύνης της Κίνας» και επεξεργάζεται 100 petabytes δεδομένων. Ενώ η διάδοση των πλατφορμών υπολογιστικού νέφους (cloud) και η διαθεσιμότητα αποθηκευτικών και υπολογιστικών πόρων κατά παραγγελία (on demand) επιτάχυνε περαιτέρω την υιοθέτηση τεχνολογιών μεγάλων δεδομένων.

Την χρονιά 2018, είναι μεγάλη η ζήτηση για μοντέλα οπτικοποίησης. Η εξέλιξη των νέων βελτιωμένων μοντέλων οπτικοποίησης αποτελεί ανάρπαστο μέρος της απόκτησης πληροφοριών από τα big data. Τέλος, το 2020, οι πλατφόρμες αυτοματισμού που έχουν σχεδιαστεί για την επεξεργασία μεγάλων δεδομένων, διευκολύνουν τη μετάβαση από την φάση της ανάλυσης έως και την παραγωγική φάση. Σήμερα, τα Μεγάλα Σύνολα Δεδομένων καλύπτουν ένα ευρύ φάσμα εφαρμογών και κλάδων. Ο τομέας συνεχίζει να εξελίσσεται με συνεχείς προόδους σε τομείς όπως η μηχανική μάθηση, η τεχνητή νοημοσύνη, η ανάλυση σε πραγματικό χρόνο και ο υπολογισμός ακμών. Στο μέλλον οι όγκοι δεδομένων θα συνεχίσουν να αυξάνονται όλο και περισσότερο. Κάποιες τάσεις μεγάλων δεδομένων περιλαμβάνουν νέες ιδέες, ενώ κάποιες άλλες συνδυάζουν ή συγχωνεύουν διαφορετικές τεχνολογίες υπολογιστών που βασίζονται στα Μεγάλα Σύνολα Δεδομένων. Χαρακτηριστικό παράδειγμα αποτελεί η μηχανική μάθηση που συγχωνεύεται με αναλυτικά στοιχεία και φωνητικές απαντήσεις, ενώ λειτουργεί σε πραγματικό χρόνο.

1.3 Πηγές και Τύποι Μεγάλων Συνόλων Δεδομένων

Στον τομέα της πληροφορικής, η δυνατότητες οργάνωσης και αποθήκευσης δεδομένων αποτελούν βασικές προϋποθέσεις ώστε τα δεδομένα να μπορούν να χρησιμοποιηθούν αποδοτικά. Πιο συγκεκριμένα, ένα μεγάλο ποσοστό δεδομένων δεν διαθέτει οριοθετημένη δομή, με αποτέλεσμα να αναπτυχθούν νέοι αλγόριθμοι ικανοί να διαχειριστούν δεδομένα σε ποικίλες μορφές που προέρχονται από διαφορετικές πηγές [2]. Για τον λόγο αυτό διαμορφώθηκαν οι εξής τύποι:

Δομημένα Δεδομένα (Structured Data): Δομημένα δεδομένα είναι πληροφορίες οργανωμένες και διαχειρίσιμες ως προς την ανάλυσή τους. Τα δεδομένα αυτά αποθηκεύονται σε μία κλασική σχεσιακή βάση δεδομένων και συνήθως αποτελούνται από στήλες και σειρές δεδομένων σε έναν ή περισσότερους συνδεδεμένους πίνακες. Για την ανάλυση δομημένων δεδομένων είναι απαραίτητη η δημιουργία κώδικα υπολογιστικής γλώσσας. Η γλώσσα HTML συνέβαλλε ώστε τα δεδομένα να μεταδίδονται με οργανωμένο τρόπο, με στόχο οι μηχανές αναζήτησης να μπορούν να εμφανίζουν το υλικό με ελκυστικό και προσιτό τρόπο. Ωστόσο, η καταγραφή των δομημένων πληροφοριών μπορεί να επιτευχθεί και χειροκίνητα. Χαρακτηριστικά παραδείγματα δομημένων δεδομένων αποτελούν τα αρχεία Excel, οι βάσεις δεδομένων SQL, οι φόρμες ερωτηματολόγιων.

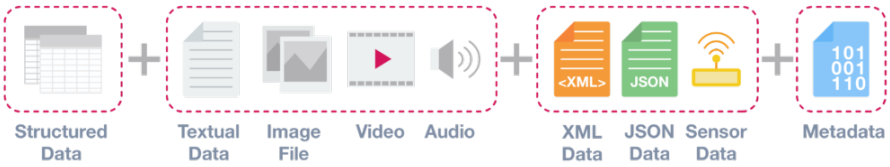
Ημι-δομημένα δεδομένα (Semi-structured Data): Τα ημιδομημένα δεδομένα είναι ένας τύπος δεδομένων που δεν συμμορφώνονται με τη δομή των παραδοσιακών σχεσιακών βάσεων δεδομένων, αλλά έχουν ορισμένες ιδιότητες που τα καθιστούν πιο δομημένα από τα αδόμητα δεδομένα. Επιπλέον αναφέρονται και ως αυτοπεριγραφόμενα δεδομένα (self- describing). Τα ημιδομημένα δεδομένα επιτρέπουν ευελιξία και δυναμικές αλλαγές στη δομή τους. Περιλαμβάνουν συνήθως ετικέτες που παρέχουν πρόσθετες πληροφορίες σχετικά με τη δομή ή το περιεχόμενο των δεδομένων. Αυτές οι ετικέτες βοηθούν στην οργάνωση και ερμηνεία των δεδομένων, αλλά δεν επιβάλλουν κάποιο συγκεκριμένο σχήμα. Η πληροφορία για το σχήμα βρίσκεται μαζί με τα δεδομένα καθώς δεν είναι γνωστό εκ των προτέρων ποιες ιδιότητες υπάρχουν σε κάθε αντικείμενο. Παραδείγματα ημι-δομημένων δεδομένων αποτελούν το μορφότυπο JSON (JavaScript Object Notation) η SWIFT, η HTML (Hypertext Markup Language) και η XML (extensible Markup Language).

Μη δομημένα δεδομένα (Unstructured Data): Τα μη δομημένα δεδομένα αναφέρονται σε δεδομένα που δεν διαθέτουν προκαθορισμένο μοντέλο ή κάποια συγκεκριμένη μορφή. Τα μη δομημένα δεδομένα λόγω του ότι δεν είναι οργανωμένα παρουσιάζουν ασάφειες και αοριστίες. Η ανάλυση τους λοιπόν είναι περίπλοκη και απαιτεί ξεχωριστές λύσεις. Βασικά τους χαρακτηριστικά όπως η έλλειψη σταθερού σχήματος, ο υψηλός όγκος και η δυσκολία ανάλυσης τους καθιστούν τα παραδοσιακά

συστήματα διαχείρισης σχεσιακών βάσεων δεδομένων ακατάλληλα για την αποτελεσματική επεξεργασία αυτών των δεδομένων. Παραδείγματα μη δομημένων δεδομένων αποτελούν έγγραφα, εικόνες, αρχεία ήχου, αρχεία βίντεο, αναρτήσεις στα μέσα κοινωνικής δικτύωσης, δεδομένα αισθητήρων και αρχεία εξυπηρέτησης πελατών.

Μεταδεδομένα (metadata): Τα μεταδεδομένα είναι δεδομένα τα οποία περιγράφουν άλλα δεδομένα. Πιο αναλυτικά, περιέχουν πρόσθετες πληροφορίες σχετικά με ένα σύνολο δεδομένων. Αποτελούν σημαντικό παράγοντα για την ανάλυση μεγάλων δεδομένων. Τα μεταδεδομένα μπορεί να είναι δομημένα ή μη δομημένα και να περιλαμβάνουν διάφορους τύπους πληροφοριών ανάλογα με το πλαίσιο. Για παράδειγμα, τα μεταδεδομένα περιέχουν πληροφορίες για ένα σύνολο φωτογραφιών δίνοντας ακριβή στοιχεία για την τοποθεσία και την ώρα. Άλλα χαρακτηριστικά παραδείγματα αποτελούν το όνομα ενός προϊόντος, το περιεχόμενο ενός σχολίου, οι βαθμολογίες [6].

Εν κατακλείδι, όταν υπάρχουν τόσες πολλές πληροφορίες σε διάφορες και ποικίλες μορφές γίνεται δύσκολη η διαχείριση των δεδομένων με παραδοσιακούς τρόπους. Για τον λόγο αυτό, οι επιστήμονες για να μπορέσουν να διαχειριστούν τον τύπο και την ιδιομορφία των δεδομένων οδηγούνται στην δημιουργία νέων πηγών δεδομένων με σκοπό την κάλυψη του τεράστιου όγκου πληροφοριών. Ο συνδυασμός ισχυρότερων συσκευών και παγκόσμιου δικτύου συμβάλλει σημαντικά στην σωστή και ελεγχόμενη χρήση δεδομένων.



Εικόνα 3: Τύποι Μεγάλων Δεδομένων (Πηγή: https://dataanalysis.substack.com/p/parsing-semi-structured-data-in-sql)

1.4 Χαρακτηριστικά Μεγάλων Συνόλων Δεδομένων

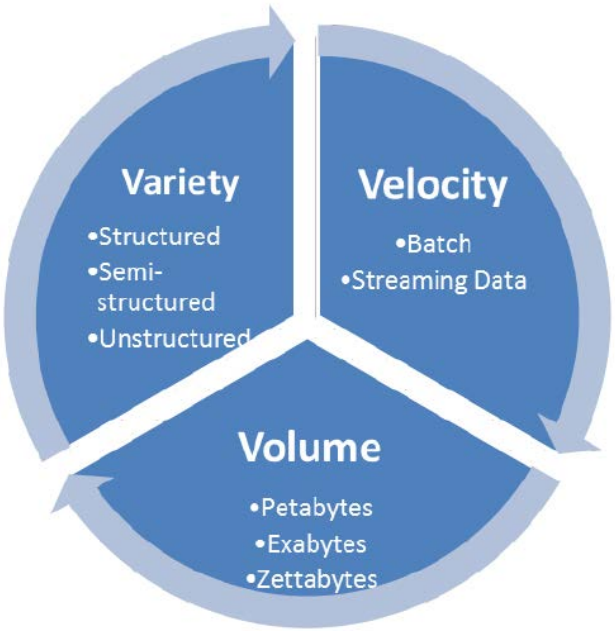
Όπως αναφέρθηκε και σε προηγούμενη ενότητα, το 2001 προσδιορίστηκε το πρότυπο 3V από τον αναλυτή της Gartner, Doug Laney. Σύμφωνα με την προσέγγιση αυτή τα πιο σημαντικά χαρακτηριστικά των μεγάλων δεδομένων είναι ο όγκος

(Volume), η ταχύτητα (Velocity), και η ποικιλομορφία (Variety). Τα 3V είναι οι κινητήριες διαστάσεις του Big Data Quantification [2] [4]. Πιο αναλυτικά:

Όγκος (Volume): Ο όγκος αναφέρεται στην ποσότητα δεδομένων που χρειάζεται να διαχειριστούμε. Με την πάροδο του χρόνου καθώς και την εξέλιξη της τεχνολογίας οι πληροφορίες σε αναλογική μορφή αντικαταστάθηκαν από πληροφορίες σε ψηφιακή μορφή. Έτσι, έχουμε να διαχειριστούμε τεράστιες ποσότητες δεδομένων σε διάφορες μορφές (εικόνες, βίντεο, ήχος) με κύρια και βασική πηγή τα μέσα κοινωνικής δικτύωσης. Πιο συγκεκριμένα, η αύξηση των δεδομένων στον παγκόσμιο ιστό, καθιστά απαραίτητη την δημιουργία νέων εφαρμογών και μεθόδων, διότι οι παραδοσιακοί τρόποι θεωρούνται ανεπαρκείς. Εκτιμάται ότι παράγονται περίπου 2.5 δισεκατομμύρια bytes δεδομένων καθημερινά. Ωστόσο, οι επιχειρήσεις για να μπορέσουν να καλύψουν τον τόσο τεράστιο όγκο δεδομένων χρησιμοποιούν Terabytes ή Petabytes. Με την βοήθεια εφαρμογών όπως η Hadoop και η Spark δίνεται η δυνατότητα να εξοικονομήσουμε χρόνο και να αναλύσουμε τον όγκο δεδομένων.

Ταχύτητα (Velocity): Η ταχύτητα αναφέρεται στην δυνατότητα γρήγορης διαχείρισης των νέων δεδομένων καθώς και στην εύρυθμη λειτουργία των υπαρχόντων. Τα Μεγάλα Σύνολα Δεδομένων εισρέουν με μεγάλη ταχύτητα και συχνά απαιτούν επεξεργασία και ανάλυση σε πραγματικό ή σχεδόν πραγματικό χρόνο. Ωστόσο η ταχύτητα αφορά επίσης τον ρυθμό αλλαγών, την σύνδεση συνόλων δεδομένων που έρχονται με διαφορετικές ταχύτητες στο σύστημα. Όταν η χρονική σχέση μεταξύ δύο ή περισσότερων συνόλων δεδομένων αλλάζει, τότε αλλάζουν όλα. Ο ρυθμός ανανέωσης των υπαρχόντων δεδομένων σχετίζεται με τον χρόνο που χρειάζεται ώστε να αντληθεί η πληροφορία από τα εισερχόμενα δεδομένα.

Ποικιλομορφία(Variety): Η ποικιλομορφία αναφέρεται στον αριθμό διαφορετικών δεδομένων. Αναλυτικότερα, η επεξεργασία, ο συνδυασμός και η αποθήκευση δεδομένων μέσω ποικίλων πηγών αντικατοπτρίζει το βασικό χαρακτηριστικό των big data. Στο πρώτο στάδιο, χρειάζεται οι πληροφορίες που αντλούμε από τις πηγές, να κατηγοριοποιηθούν ορθά ανάλογα με τον τύπο που βρίσκονται στις κατηγορίες των δομημένων (σχεσιακές βάσεις δεδομένων), των μη δομημένων (φωτογραφίες, βίντεο) και τέλος των ημιδομημένων (social media). Το δεύτερο στάδιο, προϋποθέτει οι πηγές δεδομένων να μην χρησιμοποιούν αυστηρή δόμηση διότι η ποικιλία και το εύρος είναι τέτοια με σκοπό να επηρεάζεται η μεταβλητότητα και η σημασιολογία του νοήματος.



Εικόνα 4: Απεικόνιση του μοντέλου 3V [2]

Με την πάροδο των χρόνων το αρχικό μοντέλο των 3V μετατράπηκε σε 7V. Τα χαρακτηριστικά που προστέθηκαν συνέβαλλαν ώστε οι επιστήμονες και οι ερευνητές να κατανοήσουν και να αντιληφθούν την πραγματική φύση και τις συνέπειες των μεγάλων δεδομένων. Ωστόσο, η εξέλιξη των χαρακτηριστικών των μεγάλων δεδομένων βοήθησε τις επιχειρήσεις να αντιμετωπίσουν εκατοντάδες χιλιάδες πηγές δεδομένων ροής που απαιτούν αναλύσεις σε πραγματικό χρόνο. Τα παραδοσιακά συστήματα διαχείρισης δεδομένων δεν είναι σε θέση να χειρίζονται τόσο μεγάλες ροές δεδομένων. Έτσι, οι μεγάλες τεχνολογίες δεδομένων επιτρέπουν στις επιχειρήσεις να δημιουργούν πληροφορίες σε πραγματικό χρόνο από υψηλούς όγκους δεδομένων. Τα νέα χαρακτηριστικά είναι: ακρίβεια (Veracity), αξία (Value), μεταβλητότητα (Variability) και οπτικοποίηση (Visualization).

Ακρίβεια (Veracity): Σύμφωνα με την IBM η ακρίβεια αποτελεί το τέταρτο κατά σειρά V και αντιπροσωπεύει πόσο ακριβές ή αληθές είναι ένα σύνολο δεδομένων. Πιο συγκεκριμένα, η ανάλυση, ο τύπος και η επεξεργασία των δεδομένων από όποια πηγή και αν προέρχονται, θα πρέπει να είναι αξιόπιστα. Η κατάργηση παραγόντων όπως ο θόρυβος, η μεροληψία ή οι ασυνέπειες είναι πτυχές που συντελούν στην βελτίωση της ακρίβειας των μεγάλων δεδομένων. Η διασφάλιση της ειλικρίνειας περιλαμβάνει την εφαρμογή διαδικασιών ελέγχου ποιότητας δεδομένων, τεχνικών καθαρισμού δεδομένων και μέτρων διασφάλισης ποιότητας για τη βελτίωση της αξιοπιστίας των δεδομένων.

Αξία (Value): Η αξία είναι η τιμή των δεδομένων που συλλέγονται. Πρόκειται για το πέμπτο κατά σειρά V και είναι ένα χαρακτηριστικό των μεγάλων δεδομένων που εισήγαγε η Oracle Corporation. Αναλυτικότερα, μία επιχείρηση συλλέγει ορισμένα big data που μπορεί να έχουν μικρή ή και καθόλου αξία στην λήψη των αποφάσεων ή στην βελτίωση των λειτουργιών. Ωστόσο, η επιχείρηση θα πρέπει να ελέγχει ακριβώς ποια δεδομένα συλλέγονται και κατά πόσο είναι πολύτιμα για την επιχείρηση. Όσα δεν μπορούν να αξιοποιηθούν αποτελούν καθυστέρηση στην επεξεργασία και εμποδίζουν την διαδικασία ανάλυσης δεδομένων. Συμπερασματικά, η αξία βρίσκεται στο όφελος που μπορεί να αποκομίσουν οι επιχειρήσεις από την ανάλυση και κατά συνέπεια από την αξία κάθε δεδομένου.

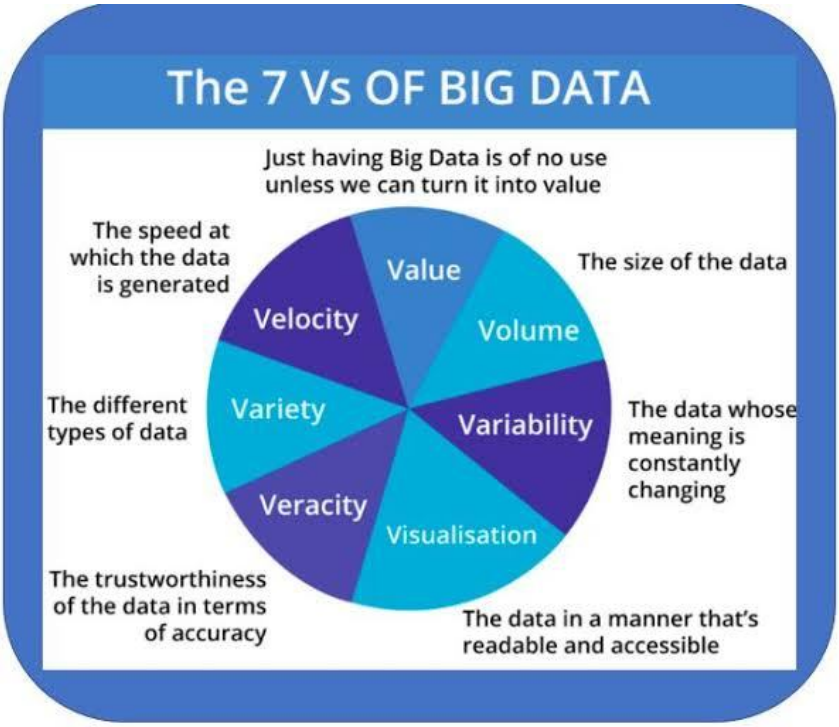
Κατά καιρούς, έχουν προταθεί περαιτέρω επεκτάσεις του πλαισίου, όπως δύο ακόμη Vs: μεταβλητότητα και οπτικοποίηση. Αυτές οι πρόσθετες διαστάσεις αφορούν τη μεταβλητότητα των μορφών δεδομένων και τη σημασία των τεχνικών οπτικοποίησης δεδομένων για την εξαγωγή συμπερασμάτων από τα Μεγάλα Σύνολα Δεδομένων [5] [7].

Μεταβλητότητα (Variability): Η μεταβλητότητα αφορά την διαφοροποίηση των δεδομένων σε διάφορες μορφές. Σε ορισμένα σύνολα δεδομένων, οι τιμές και τα χαρακτηριστικά των δεδομένων μπορεί να διαφέρουν σημαντικά με την πάροδο του χρόνου ή μεταξύ διαφορετικών πηγών. Παραδείγματα μεταβλητότητας στα δεδομένα είναι οι διακυμάνσεις στις τιμές των μετοχών, τα εποχιακά μοτίβα στη συμπεριφορά των καταναλωτών και οι αλλαγές στις ενδείξεις των αισθητήρων λόγω περιβαλλοντικών παραγόντων. Η αντιμετώπιση της μεταβλητότητας προϋποθέτει την ανάπτυξη ισχυρών συνδέσεων επεξεργασίας και ανάλυσης δεδομένων που μπορούν να προσαρμόζονται στις αλλαγές των χαρακτηριστικών και των κατανομών των δεδομένων.

Οπτικοποίηση (Visualization): Ο όρος της οπτικοποίησης αφορά την γραφική αναπαράσταση πληροφοριών και δεδομένων. Πιο αναλυτικά, χρησιμοποιώντας οπτικά στοιχεία όπως χάρτες και γραφήματα, παρέχεται ένας πιο προσβάσιμος τρόπος ώστε να γίνει πιο κατανοητή η επεξεργασία και ανάλυση των δεδομένων. Τα εργαλεία και η οπτικοποίηση δεδομένων συντελεί στην άριστη διαχείριση του τεράστιου όγκου πληροφοριών καθώς και στη λήψη αποφάσεων βάσει δεδομένων. Ο συνδυασμός των δεδομένων και των γραφημάτων είναι καθοριστικός για να πετύχει το μοντέλο της οπτικοποίησης. Με την εμφάνιση δια δραστικών εργαλείων και τεχνολογιών οπτικοποίησης, οι χρήστες μπορούν να εξερευνούν δυναμικά μεγάλα σύνολα δεδομένων, να εμβαθύνουν σε συγκεκριμένα σημεία και να αναλύουν δια δραστικά δεδομένα από διαφορετικές οπτικές γωνίες.

Παρατηρώντας μερικά από τα χαρακτηριστικά των μεγάλων δεδομένων είναι σημαντικό να αναφερθεί πως όταν αλλάζει ένα V αυξάνεται η πιθανότητα να αλλάξει και ένα άλλο. Δεν υπάρχουν καθολικά σημεία αναφοράς για τον όγκο, την

ποικιλομορφία και την ταχύτητα που καθορίζουν τα Μεγάλα Σύνολα Δεδομένων. Τα καθορισμένα όρια εξαρτώνται κυρίως από το μέγεθος, τον τομέα και την τοποθεσία της κάθε επιχείρησης και εξελίσσονται με την πάροδο του χρόνου. Οι επιχειρήσεις αναλογίζονται συνήθως την μελλοντική αξία που αναμένεται από τις μεγάλες τεχνολογίες δεδομένων σε σχέση με το κόστος εφαρμογής [2].



Εικόνα 5: Απεικόνιση του μοντέλου 7V [7]

1.5 Σπουδαιότητα Μεγάλων Δεδομένων

Η χρήση Μεγάλων Δεδομένων αποτελεί μια από τις μεγαλύτερες «επαναστάσεις» που έλαβαν χώρα τα τελευταία χρόνια σε επιχειρηματικό επίπεδο. Τα δεδομένα απαιτούν πρωταγωνιστικό ρόλο, καθώς μας δίνουν την δυνατότητα να πειραματιζόμαστε ταχύτερα και να ανακαλύπτουμε νέους δρόμους. Οργανισμοί και επιχειρήσεις από ποικίλους κλάδους αξιοποιούν δεδομένα και τα χρησιμοποιούν για να εντοπίσουν τις κατάλληλες ευκαιρίες. Αυτό έχει ως αποτέλεσμα να οδηγούμαστε σε έξυπνες επιχειρηματικές λύσεις , που κατά περίπτωση μπορούν να αποδειχθούν αποδοτικότερες και να φέρουν υψηλά κέρδη καθώς και μόνιμο πελατολόγιο. Η αξία των Μεγάλων Δεδομένων βασίζεται συνήθως στην βελτίωση των επιχειρηματικών διαδικασιών και των πρακτικών λήψεων αποφάσεων [6].Τα big data δημιουργούν αξία με διάφορους τρόπους οι οποίοι είναι οι εξής:

Δημιουργία Διαφάνειας (transparency): Με την σωστή χρήση και ανάλυση των δεδομένων από τις επιχειρήσεις, μπορούν οι οργανισμοί να αυξήσουν σημαντικά τα κέρδη τους. Τέτοιες ευκαιρίες μπορούν να δημιουργηθούν όταν παρατηρείται έλλειψη συμφωνίας κινήτρων για δημιουργία διαφάνειας δεδομένων. Πιο αναλυτικά, τα δεδομένα προσωπικού χαρακτήρα πρέπει να υποβάλλονται σε επεξεργασία με νόμιμο και διαφανή τρόπο, εξασφαλίζοντας την αντικειμενικότητα προς τα άτομα των οποίων τα δεδομένα προσωπικού χαρακτήρα υποβάλλονται σε επεξεργασία (νομιμότητα , αντικειμενικότητα και διαφάνεια). Για παράδειγμα, στον δημόσιο τομέα υπάρχουν περιπτώσεις όπου οι εργαζόμενοι σπαταλούν πολύτιμο χρόνο για να εντοπίσουν πληροφορίες από άλλες υπηρεσίες χρησιμοποιώντας μη ψηφιακά μέσα (τηλέφωνο , καταλόγους), και στην συνέχεια για να λάβουν τις πληροφορίες που αναζητούν επισκέπτονται την πηγή της πληροφορίας με φυσικά μέσα (οπτικοί δίσκοι). Έτσι , με την πάροδο του χρόνου και τη χρήση διαθέσιμων δικτύων αυτή η σπατάλη χρόνου μειώθηκε σημαντικά αναπτύσσοντας ευκολότερη και ταχύτερη την εύρεση της πληροφορίας.

Εντοπισμός Αναγκών , Μεταβλητότητας και Αύξηση της Απόδοσης: Η χρήση των δεδομένων αποτελεί μέρος του αντίκτυπου των τεχνολογιών πληροφορικής και επικοινωνιών. Αυτή την στιγμή βρισκόμαστε σε ένα σημείο καμπής καθώς ένας μεγάλος αριθμός τάσεων συγκλίνει λόγω της κλιμάκωσης και του εύρους των αλλαγών που έχουν επιφέρει. Ο μεγάλος όγκος πληροφοριών που συλλέγονται από τις επιχειρήσεις καθώς και η άνοδος των πληροφοριών στα κοινωνικά δίκτυα και στο διαδίκτυο θα φέρουν αύξηση των στοιχείων στο προβλεπόμενο μέλλον. Για παράδειγμα, στον δημόσιο τομέα δεδομένης της δημιουργίας και αποθήκευσης περισσότερων δεδομένων / συναλλαγών σε ψηφιακή μορφή συλλέγονται πιο ακριβή και λεπτομερή δεδομένα απόδοσης για τα πάντα (π.χ. αριθμός αναρρωτικής άδειας προσωπικού, συναλλαγές κτλ.). Έτσι , χρησιμοποιώντας τα στοιχεία και τους πειραματισμούς για την μεταβλητότητα τους , βελτιώνονται οι επιδόσεις. Επιπλέον είναι χρήσιμο να αναφερθεί πως τα Μεγάλα Σύνολα Δεδομένων επιτρέπουν στους οργανισμούς / επιχειρήσεις να δημιουργούν μικρότερα τμήματα και υπηρεσίες για την κάλυψη των αναγκών τους. Η προσέγγιση αυτή στο χώρο του μάρκετινγκ είναι γνωστή, όμως μπορεί να δημιουργήσει «επανάσταση» σε τομείς όπως το δημόσιο.

Αντικατάσταση ή Υποστήριξη της Ανθρώπινης Λήψης Αποφάσεων Με Αυτοματοποιημένους Αλγόριθμους: Αναβαθμισμένα και εξελιγμένα analytics μπορούν να βελτιώσουν τη λήψη αποφάσεων με σκοπό την ανάδειξη πολύτιμων πληροφοριών. Πιο συγκεκριμένα, ως πρώτη ύλη τα Μεγάλα Σύνολα Δεδομένων, έχουν καταφέρει να περνούν μέσα από συστήματα υπολογιστών και τεχνολογιών μέσω αλγορίθμων, και να αναλύονται ώστε να δίνουν το βέλτιστο αποτέλεσμα για την λειτουργία διάφορων φορέων. Με τον τρόπο αυτό ελαχιστοποιούνται οι κίνδυνοι , καθώς μέσω αλγόριθμων εντοπίζονται πολύτιμες πληροφορίες που σε διαφορετική περίπτωση θα παρέμεναν κρυμμένες. Χαρακτηριστικά παραδείγματα αποτελούν οι κατασκευαστικές εταιρίες

όπου μέσω των μεγάλων δεδομένων επιτυγχάνεται η παραγωγή , η βελτιστοποίηση αποδοτικότητας και η μείωση της σπατάλης. Παράλληλα, ορισμένοι οργανισμοί λαμβάνουν πιο αποτελεσματικές αποφάσεις αναλύοντας ολόκληρα σύνολα δεδομένων από πελάτες και εργαζομένους.

Καινοτομία Με Νέα Επιχειρηματικά Μοντέλα, Προϊόντα, Υπηρεσίες: Ένα επιχειρηματικό μοντέλο περιγράφει την λογική για το πως ένας οργανισμός θα δημιουργήσει, θα διανέμει και θα αποκτήσει αξία. Καθορίζει την βιωσιμότητα του οργανισμού και προσφέρει αναλυτικές πληροφορίες σχετικά με το μονοπάτι που θα πρέπει να ακολουθήσει η επιχείρηση για να πετύχει. Τα Μεγάλα Σύνολα Δεδομένων αποτελούν αναγκαίο παράγοντα και βοηθάνε τις επιχειρήσεις να αναπτύξουν νέα προϊόντα και υπηρεσίες, καθώς και το κατάλληλο επιχειρηματικό μοντέλο. Στο λιανεμπόριο, μέσω των υπηρεσιών είναι εφικτή η σύγκριση τιμών και έτσι οι καταναλωτές έχουν πλήρη εικόνα των τιμών σε πραγματικό χρόνο. Επιπρόσθετα, οι βιομηχανίες με την χρήση προϊόντων και υπηρεσιών από την ανάλυση των μεγάλων δεδομένων βελτιώνουν σε σημαντικό βαθμό την ανάπτυξη της νέας γενιάς των προϊόντων τους καθώς και την δημιουργία καινοτόμων after sales υπηρεσιών. Στόχος είναι σε πραγματικό χρόνο να εντοπίζονται και να εμφανίζονται δεδομένα ώστε να δημιουργηθεί ένα καινούργιο σύνολο – mobile υπηρεσιών από πλοήγηση μέχρι και ανίχνευση ανθρώπων.

1.6 Περιπτώσεις Χρήσης

Με τον όρο δεδομένα μεγάλης κλίμακας, αναφερόμαστε σε δεδομένα μεγαλύτερα από τα παραδοσιακά δεδομένα αλλά όχι με την κυριολεκτική έννοια του όρου. Πιο αναλυτικά, όπως είδαμε και σε προηγούμενη ενότητα, τα δεδομένα δεν χαρακτηρίζονται μόνο από το μέγεθος τους αλλά και από διαφόρων τύπου αρχεία. Μπορεί να είναι δομημένα, ημι-δομημένα και αδόμητα. Το εύρος ποσότητας δεδομένων βρίσκεται κατανεμημένο σε δίκτυα υπολογιστών προκειμένου να επιτευχθεί η διαδικασία ανάλυσης. Κρίνεται σκόπιμο η ανάλυση να εκτελείται παράλληλα σε πολλούς πυρήνες ή υπολογιστές. Τα δεδομένα μεγάλης κλίμακας με την χρήση επιχειρηματικών μοντέλων, προϊόντων και υπηρεσιών επιτυγχάνουν την ανάλυση των στοιχείων για την πρόβλεψη, την γνώση, την ενημέρωση και την επίτευξη καλύτερων και πιο χρήσιμων αποτελεσμάτων. Η χρήση του διαδικτύου, τα μέσα κοινωνικής δικτύωσης, οι ιστοσελίδες ενημερωτικού περιεχομένου, η χρήση ηλεκτρονικών αγορών κτλ. αυξάνει τον όγκο των δεδομένων. Ως αποτέλεσμα οι εταιρίες που αναπτύσσονταν ταχύτατα και υποστήριζαν αυτόν τον τεράστιο όγκο θέλησαν να δημιουργήσουν έναν τρόπο να αναλύσουν και να ταξινομήσουν τα

δεδομένα ώστε να εισάγουν ένα καινοτόμο μοντέλο, εφόσον υπήρχαν αρκετά δεδομένα που δημιουργούν ξεχωριστές προτιμήσεις στον άνθρωπο. Αυτή η εμπορική κίνηση οδήγησε στη χρήση ενός εξελιγμένου μοντέλου και ένα νέο κατανεμημένο σύστημα αρχείων, το Hadoop, με στόχο την επεξεργασία και τον υπολογισμό σε χιλιάδες μεμονωμένους υπολογιστές. Πρόκειται για μια επιχείρηση ανοιχτού κώδικα που χρησιμοποιείται ευρέως για την αποτελεσματική επίβλεψη πολλών δεδομένων από πολλούς οργανισμούς. Είναι χρήσιμο διότι έχει την ικανότητα να αποθηκεύει και να επεξεργάζεται σύνολα δεδομένων μεγέθους που μπορεί να είναι αδύνατον να αποθηκευτούν σε μια συσκευή αποθήκευσης όπως για παράδειγμα σε έναν σκληρό δίσκο. Μέσω της μεγάλης κλίμακας δεδομένων και ακολουθώντας τις σωστές πολιτικές προϋποθέσεις της εκάστοτε επιχείρησης θα εφαρμοστεί ένα χρήσιμο εργαλείο για την ανάπτυξη σχεδίων που έχει ως γνώμονα : την μελλοντική παραγωγικότητα, την ικανοποίηση του πελάτη, τον ανταγωνιστικό σχεδιασμό, την έρευνα και την καινοτομία των προϊόντων[6].

2 Χρησιμότητα των Μεγάλων Συνόλων Δεδομένων

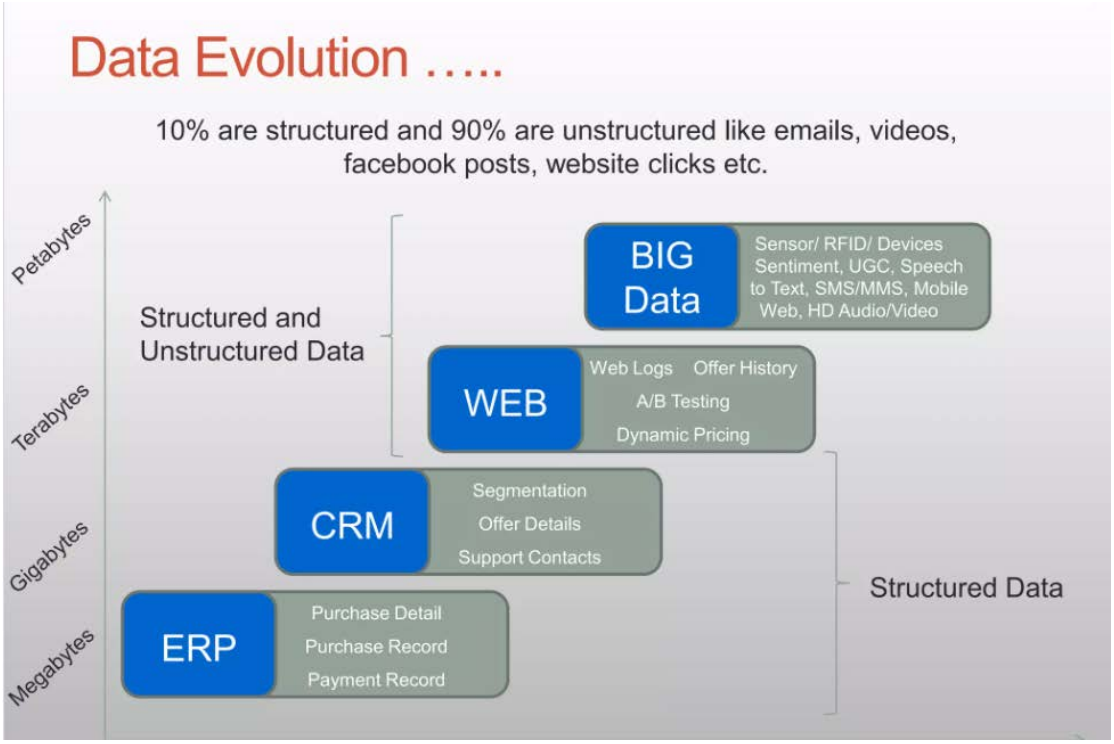
2.1 Εφαρμογές

Καθώς η τεχνολογία γύρω από τα Μεγάλα Σύνολα Δεδομένων αυξάνεται ολοένα και περισσότερο, πολλές επιχειρήσεις και οργανισμοί τα αξιοποιούν για να βελτιώσουν και να ενισχύσουν τα κέρδη τους. Πιο αναλυτικά, εταιρίες όπως για παράδειγμα η T-mobile κάνοντας χρήση των Big Data μείωσε σημαντικά κατά 50% τα έξοδά της. Startup εταιρίες έχουν βελτιώσει περιπτώσεις χρηματοοικονομικών, παιδείας, μεταφορών [6]. Κάθε περίπτωση αντιμετωπίστηκε χάρη στα big data. Άλλα παραδείγματα επιχειρήσεων είναι η Dell όπου έχει ενσωματώσει όλες τις περιπτώσεις των big data σε ένα format. Κάθε περίπτωση εταιριών που επωφελήθηκε είναι διαφορετική και μπορεί να κατηγοριοποιηθεί ανάλογα:

Analytics συμπεριφοράς: Πρόκειται για τα δεδομένα συμπεριφοράς των καταναλωτών. Μέσω της ανάλυσης προτιμήσεων του χρήστη και την ανάλογη προβολή και διαφήμιση οι οργανισμοί μπορούν να έχουν άνοδο των κερδών τους. Για παράδειγμα, η Time Warner Cable κρατά αρχεία συμπεριφοράς των πελατών για το πόσο δραστήριοι είναι και ποιες ώρες της ημέρας ώστε να βελτιώνει συνεχώς τις υπηρεσίες της.

Προβλεπόμενη Υποστήριξη: Το Διαδίκτυο και τα Big Data είναι ένα ισχυρό εργαλείο για ζητήματα συντήρησης και διάφορες υπολειτουργίες. Σημαντικό παράδειγμα αποτελεί η Southwest Airlines όπου χρησιμοποιεί δεδομένα από αισθητήρες με σκοπό να προλαβαίνει πιθανά προβλήματα ασφαλείας στα αεροπλάνα. Έτσι οποιοδήποτε πρόβλημα μπορεί να αντιμετωπιστεί γρηγορότερα και να μην αλλάξει το πρόγραμμα των πτήσεων.

Επιχειρησιακές εφαρμογές: Μέσω των επιχειρησιακών εφαρμογών (business applications) μεγάλα συστήματα χρησιμοποιούνται από δεδομένα για ευκολότερη απογραφή και βελτιστοποίηση στην συνεργασία τους με προμηθευτές. Με αυτόν τον τρόπο, περιορίζεται το χάσμα μεταξύ προσφοράς και ζήτησης, ενώ η χρηματοδότηση επιχειρήσεων που έχει στραφεί στα Big Data έχει αποκτήσει άλλη υπόσταση.



Εικόνα 6: Εφαρμογές των Big Data ανάλογα την δομή τους.
(Πηγή : https://www.cloudera.com/)

2.1.1 Υγεία

Τα δεδομένα που έχουν να κάνουν με ασθενής δημιουργούνται και πληθαίνουν ολοένα και περισσότερο. Ας αποσαφηνίσουμε όμως πρώτα τον όρο δεδομένα υγείας. Πιο αναλυτικά, με τον όρο δεδομένα υγείας περιγράφουμε τις πληροφορίες που παράγονται μέσω της υγειονομικής περίθαλψης, συμπεριλαμβανομένων των μητρώων ασθενειών , των δεδομένων κλινικών μελετών και των ηλεκτρονικών φακέλων υγείας. Η κύρια χρήση των δεδομένων υγείας σχετίζεται για την λήψη αποφάσεων σχετικά με την φροντίδα του ατόμου από το οποίο συλλέχτηκαν. Επιπλέον τα δεδομένα αυτά μπορούν να χρησιμοποιηθούν και σε δευτερογενές επίπεδο. Πιο συγκεκριμένα, με την χρήση συγκεντρωτικών δεδομένων υγείας από πηγές σε επίπεδο πληθυσμού όπως μητρώα υγείας, ασφαλιστικό φορέα υγείας μπορεί να υπάρξει βελτίωση στην ανάπτυξη φαρμάκων, την παρακολούθηση της ασφάλειας καθώς και την έρευνα και χάραξη της πολιτικής. Στην εποχή που ζούμε λόγω της πανδημίας του Covid-19 χαρακτηριστικό παράδειγμα αποτελεί η ανάλυση των κλινικών δεδομένων που μπορούν να βοηθήσουν στον έγκαιρο εντοπισμό νέων επιδημιών. Σύμφωνα με την πρόσφατη Έκθεση για την Δευτερογενή Χρήση Δεδομένων Υγείας οι περισσότερες χώρες στην Ευρωπαϊκή Ένωση συμπεριλαμβανομένου και την Ελλάδας, χρειάζεται να βελτιώσουν σημαντικά την

ετοιμότητα τους στον τομέα της δευτερογενούς χρήσης δεδομένων. Τα δεδομένα υγείας μπορούν να καταστήσουν τα συστήματα υγείας πιο βιώσιμα και να ενισχύσουν την επιστήμη και την καινοτομία προς όφελος των ασθενών. Σημαντικό παράδειγμα αποτελεί η Eurostat. Πρόκειται για την στατιστική υπηρεσία της Ευρωπαϊκής Ένωσης και είναι η κύρια πηγή συγκρίσιμων δεδομένων υγείας διοικητικά , όπως στατιστικές για τις αιτίες θανάτου και δεδομένα που παρέχουν οι συμμετέχοντες στο πλαίσιο της Ευρωπαϊκής Έρευνας βάση ερωτηματολογίου ή μέσω του Ευρωπαϊκού μέτρου ελάχιστης υγείας της έρευνας (EE –SILC). Τα δεδομένα συνήθως αφορούν θέματα για τις δαπάνες υγειονομικής περίθαλψης, αιτίες θανάτου, υγεία και ασφάλεια στην εργασία καθώς και δραστηριότητες υγειονομικής περίθαλψης [7].

2.1.2 Βιομηχανία

Με την αξιοποίηση των Big Data οι επιχειρήσεις σε όλο τον κόσμο έρχονται αντιμέτωπες με νέες προκλήσεις. Η βιομηχανική παραγωγή βρίσκεται στους κλάδους που χρησιμοποιεί τις μεγαλύτερες επενδύσεις σε big data και BDA (Big Data Analysis). Πιο συγκεκριμένα, ζούμε στην εποχή της 4ης Βιομηχανικής Επανάστασης , αυτό έχει ως αποτέλεσμα η ανθρώπινη κρίση και εμπειρία να έχουν κυρίαρχο ρόλο στα κριτήρια των επιχειρηματικών επιλογών. Ωστόσο, τα σύγχρονα εργαλεία ανάλυσης μεγάλου όγκου δεδομένων είναι ιδιαίτερα διαδεδομένα και απαραίτητα για την αύξηση της παραγωγικότητας, της αποδοτικότητας και της καινοτομίας. Η αγορά κινείται με εντυπωσιακούς ρυθμούς με αποτέλεσμα οι αρμόδιοι των βιομηχανικών εγκαταστάσεων να ανακαλύπτουν ολοένα και περισσότερες πληροφορίες ώστε να βελτιώσουν τις διαδικασίες και να αυξήσουν τα κέρδη τους. Για παράδειγμα, μέσω της τεχνολογίας BDA (Big Data Analysis), το λιανεμπόριο δίνει την δυνατότητα στις επιχειρήσεις να βελτιώσουν τις διαδικασίες και να κατασκευάσουν νέες προτάσεις στους καταναλωτές με βάση τα παλαιότερα ιστορικά των αγορών τους. Έτσι συλλέγονται νέες πληροφορίες μέσω συναλλαγών που γίνονται με τραπεζικές συναλλαγές, με την χρήση εγγραφής χρήστη καθώς και με διευθύνσεις IP. Τα δεδομένα συλλέγονται και αναλύονται ώστε να παρθούν σημαντικές και στρατηγικές αποφάσεις. Σε αυτό το σημείο να αναφερθεί, πως μέσω των big data και των αναλυτικών στοιχείων που προκύπτουν οι βιομηχανικές μονάδες παραγωγής είναι σε θέση να προγραμματίζουν προληπτικές συντηρήσεις. Με τον τρόπο αυτό θα αποφύγουν τυχόν δαπανηρές βλάβες του εξοπλισμού όπως και διακοπή της παραγωγής. Συμπερασματικά, στις μελέτες που έχουν υλοποιηθεί οι βλάβες του παραγωγικού εξοπλισμού μπορούν να μειωθούν έως και 26% και να μειώσουν τον προγραμματισμένο χρόνο διακοπής κατά ένα τέταρτο, χάρη στα BDA (Big Data Analysis). Εκτός από τα παραπάνω η βιομηχανική παραγωγή στηρίζεται στη χρήση εργαλείων για την βελτιστοποίηση της παραγωγής και την δημιουργία κερδών παραγωγικότητας και απόδοσης [8] [9]. Τα εργαλεία είναι τα εξής:

1)Αποθήκευση Δεδομένων: Ως κυρίαρχο βήμα για την λειτουργία των big data, την συλλογή και την αποθήκευση πληροφοριών.

2)Εργαλεία Εκκαθάρισης Δεδομένων: Τυποποίηση των δεδομένων σε διάφορες μορφές (δομημένα, αδόμητα) ώστε να χρησιμοποιούνται από ποικίλες εφαρμογές και συστήματα.

3)Εργαλεία Δημιουργίας Προφίλ: Συλλογή πληροφοριών έως το επίπεδο των μεταδεδομένων για να υπάρχει διαφάνεια στις λειτουργίες παραγωγής.

4)Εργαλεία Εξόρυξης Δεδομένων: Γρήγορη πρόσβαση σε πληροφορίες με σκοπό να λαμβάνονται πληροφορίες από τους παραγωγούς όταν τις χρειάζονται.

5)Εργαλεία Χαρτογράφησης Δεδομένων: Βοηθούν στον εντοπισμό εξαρτήσεων και πιθανών σημείων συμφόρησης, καθώς βελτιώνουν την ασφάλεια και εντοπίζονται εύκολα οι κίνδυνοι.

6)Ανάλυση Δεδομένων: Τα Analytics είναι σε θέση να βοηθούν τους βιομηχανικούς παραγωγούς μέσω πληροφοριών που προκύπτουν σε αστοχίες εξοπλισμού, έλλειψη εφοδιασμού κ.α. Μέσω αυτών υπάρχει καλύτερη λήψη αποφάσεων.

7)Οπτικοποίηση Δεδομένων: Μέσω γραφημάτων και πινάκων οι βιομηχανικοί παραγωγοί κατανοούν τα δεδομένα και προβαίνουν στις απαραίτητες αλλαγές.

8)Παρακολούθηση Δεδομένων: Επιτρέπουν στους κατασκευαστές την διασφάλιση ποιότητας , την απόδοση εξοπλισμού και τον αποτελεσματικό έλεγχο της παραγωγής.

Κλείνοντας, μέσω αυτών των εργαλείων η βιομηχανική παραγωγή είναι πιο αναβαθμισμένη προσφέροντας στην πλευρά των επιχειρήσεων κέρδη παραγωγικότητας και απόδοσης καθώς και ενίσχυση σχέσεων με πελάτες, συνεργάτες και προμηθευτές. Η ενσωμάτωση των big data στη βιομηχανική παραγωγή είναι σημαντικό «όπλο» για τις επιχειρήσεις.

2.1.3 Κυβέρνηση

Με την πάροδο των χρόνων και την ανάπτυξη της τεχνολογίας οι κυβερνήσεις πρέπει να διασφαλίσουν την προστασία των πολιτών τους από απειλές και προκλήσεις στον τομέα της δημόσιας ασφάλειας. Για να αντιμετωπιστούν τέτοιου είδους προκλήσεις είναι αναγκαίο τα αρμόδια θεσμικά όργανα και οι οργανισμοί να μπορούν να αντλούν τις δικές τους πληροφορίες. Κυβερνητικές υπηρεσίες για να μπορέσουν να ανταποκριθούν σε προγραμματισμένες εγκληματικές πράξεις ανέλυσαν δεδομένα. Οι πλατφόρμες Big Data όπως για παράδειγμα το PRISM, είναι εργαλείο της Κυβέρνησης των ΗΠΑ, και μπορεί να παρακολουθεί ηλεκτρονικές επικοινωνίες από πολίτες χωρών

εκτός των ΗΠΑ αλλά και τις επικοινωνίες Αμερικανών πολιτών. Με τη χρήση Μεγάλων Δεδομένων η κάθε κυβέρνηση επιτυγχάνει τα ακόλουθα. Αρχικά την αποτροπή επιθέσεων παρακολουθώντας μέσω ειδικών συστημάτων τις συμπεριφορές των υπόπτων ατόμων. Την Ανταλλαγή αξιόπιστων πληροφοριών μεταξύ όλων των τμημάτων. Την πρόσβαση, συλλογή και ανάλυση προτύπων συμπεριφοράς στα σχέδια των υπόπτων με σκοπό την πρόβλεψη των μελλοντικών τους κινήσεων. Εν κατακλείδι, οι δυνατότητες της απόκτησης Big Data από πολλαπλές πηγές για την επίλυση δύσκολων προβλημάτων είναι σε θέση να αποτρέψουν την εμφάνιση όλων των κυβερνητικών απειλών [9] [8] .

2.1.4 Διαδίκτυο των Πραγμάτων

Με τον όρο «Διαδίκτυο των Πραγμάτων» ή «IoT» (Internet of Things) αναφερόμαστε στο δίκτυο οποιουδήποτε μηχανισμού ή συσκευής που διαθέτει ενσωματωμένους αισθητήρες και λογισμικό με στόχο την ανταλλαγή δεδομένων και πληροφοριών μέσω διαδικτύου. Πιο συγκεκριμένα, συσκευές όπως φώτα, κινητά τηλέφωνα, ρολόγια κ.α. ανήκουν στο Διαδίκτυο των Πραγμάτων διότι έχουν την ικανότητα να συνομιλούν και να συνεργάζονται ορθά για την άμεση ανταλλαγή πληροφοριών μέσω Internet. Η τεχνολογία αυτή σε συνδυασμό με τα συστήματα ρομποτικής και αυτοματισμού συγκεντρώνουν και αναλύουν εύκολα και γρήγορα πληροφορίες για την διευκόλυνση και εξυπηρέτηση της ανθρώπινης δραστηριότητας. Η χρήση των έξυπνων συσκευών σε πρόσφατη έρευνα που πραγματοποιήθηκε σε παραγωγή εργοστασίου έδειξε την μείωση των κυβικών αποβλήτων κατά 35%. Αυτό συνέβη διότι η χρήση αισθητήρων και η συλλογή δεδομένων αυξάνει την απόδοση της παραγωγής διευκολύνοντας την χρήση των μηχανημάτων και των υλικών που σπαταλούνται, καθώς και της ενέργειας που καταναλώνεται. Από την άλλη πλευρά, κάθε συσκευή που μεταφέρει πληροφορίες στο δημόσιο διαδίκτυο υπάρχει πιθανότητα να κινδυνεύει από κλοπή πληροφοριών. Μέχρι στιγμής δεν υπάρχουν διεθνή πρότυπα ασφάλειας λόγω του ότι η τεχνολογία αυτή βρίσκεται ακόμη σε πρώιμο στάδιο. Έτσι κάθε κατασκευαστής εφαρμόζει δική του μέθοδο κρυπτογράφησης. Οι έξυπνες συσκευές εισέρχονται ολοένα και περισσότερο στην ζωή μας συλλέγοντας σημαντικές προσωπικές πληροφορίες διευκολύνοντας την καθημερινότητα μας [7] .

2.1.5 Μέσα Ενημέρωσης και Διασκέδασης

Στην εποχή που διανύουμε τα Μεγάλα Σύνολα Δεδομένων έχουν εισχωρήσει στην καθημερινότητα μας και είναι απαραίτητα. Παρέχουν έναν αρκετά μεγάλο πλούτο πληροφόρησης στους οργανισμούς και τις επιχειρήσεις στον κλάδο της ενημέρωσης και της διασκέδασης, δίνοντας πληροφορίες για τις προτιμήσεις των ανθρώπων. Μέσω της άντλησης πληροφοριών από τα κοινωνικά δίκτυα οι εταιρίες προβάλλουν και διαφημίζουν περιεχόμενα βλέποντας τις επιθυμίες των καταναλωτών προγραμματίζοντας κατάλληλα τις προβολές τους. Έτσι ο κλάδος της ενημέρωσης και της διασκέδασης γίνεται πιο προσιτός και απομακρύνονται τα παραδοσιακά μέτρα των εφημερίδων και των περιοδικών. Οι πλατφόρμες κοινωνικών μέσων, ενημέρωσης και

διασκέδασης προσφέρουν μορφές δεδομένων με δυνατότητες ήχων, κειμένων, βίντεο, εικόνων και γεωγραφικών τοποθεσιών. Χάρη σε αυτά τα δεδομένα για παράδειγμα οι ειδικοί είναι σε θέση να προβλέπουν καιρικά φαινόμενα και μέσω της ενημέρωσης να μοιράζεται δημόσια το περιεχόμενο με χαμηλό κόστος και συνεχή παρουσία [9].

2.1.6 Τηλεπικοινωνίες

Οι εταιρείες τηλεπικοινωνιών αξιοποιούν τα Μεγάλα Σύνολα Δεδομένων για τη βελτιστοποίηση του δικτύου, την πρόβλεψη της αποχώρησης των πελατών, το εξατομικευμένο μάρκετινγκ και τη βελτίωση της ποιότητας των υπηρεσιών. Η ανάλυση λεπτομερών αρχείων κλήσεων, δεδομένων κίνησης δικτύου, δημογραφικών στοιχείων πελατών και αλληλεπιδράσεων στα μέσα κοινωνικής δικτύωσης βοηθά τις εταιρείες τηλεπικοινωνιών να βελτιώσουν την απόδοση του δικτύου, να στοχεύσουν τις προσπάθειες μάρκετινγκ και να βελτιώσουν τη διατήρηση των πελατών.

2.1.7 Μεταφορές και εφοδιαστική αλυσίδα

Η ανάλυση μεγάλων δεδομένων χρησιμοποιείται στις μεταφορές και τα logistics για τη βελτιστοποίηση διαδρομών, τη διαχείριση στόλου, την πρόβλεψη ζήτησης και την ορατότητα της εφοδιαστικής αλυσίδας. Τα δεδομένα GPS, τα δεδομένα τηλεματικής, η πρόγνωση καιρού και τα δεδομένα κυκλοφορίας αναλύονται για τη βελτιστοποίηση των λειτουργιών της εφοδιαστικής αλυσίδας, τη μείωση του κόστους μεταφοράς και τη βελτιστοποίηση της παράδοσης.

2.2 Μεγάλα Σύνολα Δεδομένων στον Κόσμο

Με τις πληροφορίες που συλλέγουμε από την IBM, η πληροφορία που έχει παραχθεί από το ανθρώπινο είδος δεκάδες χιλιάδες χρόνια πριν, αποτελεί το 10% του συνόλου. Ωστόσο, το 90% των data που υπήρχαν στον κόσμο το 2013 είχαν δημιουργηθεί δύο χρόνια πριν. Στις μέρες μας, τα δεδομένα παράγονται με ιλιγγιώδη ταχύτητα και πλέον τα σύνολα δεδομένων σε αναλογική μορφή είναι ελάχιστα. Αναλυτικότερα, η πρόκληση της εποχής σε επιχειρηματικό και τεχνολογικό επίπεδο αφορά την αποτελεσματική συλλογή και ανάλυση δεδομένων που βρίσκεται στην διάθεση καταναλωτών και εταιριών. Αναζητήσεις στον παγκόσμιο ιστότοπο, βίντεο που ανεβαίνουν στο YouTube, συναλλαγές στο Amazon, μυστικές καταγραφές τηλεφωνικών κλήσεων από την NSA, ηλεκτρονικές εφαρμογές που διαθέτουν εταιρίες και πολίτες και υπάρχουν μέσω του διαδικτύου αποτελούν έναν τεράστιο όγκο δεδομένων που η ανθρωπότητα αρκετά χρόνια πριν δεν μπορούσε να διαχειριστεί. Δεν υπήρχε η κατάλληλη τεχνολογία για αποθήκευση και δημιουργία πληροφορίας. Εταιρίες γνωστές σε παγκόσμιο επίπεδο όπως η Amazon, η e-Bay και η Google έχουν την δυνατότητα να

επενδύουν τεράστια ποσά στην αναβάθμιση των τεχνολογιών τους για μεγαλύτερη ακρίβεια στα δεδομένα τους με καινοτόμες υπηρεσίες και προϊόντα για τους καταναλωτές τους. Σε έρευνα της Microsoft παρατηρήθηκε πως η αξία της αγοράς από τα Μεγάλα Σύνολα Δεδομένων διαμορφώνεται για τα επόμενα τρία χρόνια στα $1.6 τρις. Επιπλέον στον δυτικό κόσμο, σύνολα δεδομένων γίνονται ψηφιακά και προσφέρονται ελεύθερα για την ανάπτυξη νέων υπηρεσιών σε κλάδους όπως το real estate, εμπόριο κ.α. Τέλος, τα Μεγάλα Σύνολα Δεδομένων ωφέλησαν ολόκληρο τον κόσμο με τα εξαιρετικά εργαλεία analytics δίνοντας κερδοφόρες λύσεις στον τομέα του digital marketing ικανοποιώντας τις ανάγκες των καταναλωτών με τρόπους που στο παρελθόν δεν ήταν δυνατοί [8] .

2.3 Μεγάλα Σύνολα Δεδομένων και Νέες Θέσεις Εργασίας

Ο κόσμος της πληροφορίας καθώς και η ορθή ανάλυση των δεδομένων είναι ένας κλάδος που αναπτύσσεται γρήγορα. Η συλλογή, η κατηγοριοποίηση, η ανάλυση, ο σωστός χειρισμός και η μετάφραση στην εκάστοτε γλώσσα κάνει τα δεδομένα να αποκτούν δύναμη. Η απλή πληροφορία μετατρέπεται σε χρήσιμα δεδομένα και έτσι οι εταιρίες οδηγούνται στην πρόσληψη έμπειρου προσωπικού που εξειδικεύεται στον τομέα της ανάλυσης τους. Μέχρι το 2017, η έκρηξη και ο όγκος των δεδομένων είχε δημιουργήσει 3,75 εκατομμύρια καινούργιες θέσεις εργασίας καθώς το κέρδος είχε ξεπεράσει τα 50 δισεκατομμύρια δολάρια. Αναλυτικότερα, δημιουργήθηκε η ανάγκη για επιστήμονες δεδομένων οι οποίοι πραγματεύονται την επιστήμη της πληροφόρησης. Ουσιαστικά πρόκειται για ένα γνωστικό πεδίο που αφορά την συλλογή, την ανάλυση και την διαχείριση των πληροφοριών. Ενσωματώνει στοιχεία από ποικίλους κλάδους όπως πληροφορική, γνωσιακή επιστήμη, βιβλιοθηκονομία κ.α. Σύμφωνα με έρευνα από την Accenture, το 73% των εταιριών δαπανούν περίπου το 1/5 του προϋπολογισμού τους για την ανάλυση δεδομένων. Παράλληλα, οι εταιρίες ψάχνουν ικανούς επιστήμονες με εμπειρία και γνώση στο αντικείμενο ώστε να ανταπεξέλθουν στις απαιτήσεις της επιστήμης των δεδομένων. Το πρόβλημα είναι πως δεν υπάρχουν αρκετοί. Δεξιότητες που σχετίζονται με τα Big Data βρίσκονται σε ζήτηση όπως για παράδειγμα η εμπειρία σε Hadoop και Java, γνώση NoSQL και Map Reduce και περιβάλλοντος Linux , καλή χρήση προγραμματιστικών γλωσσών όπως Python και C++. Ο ρόλος και οι ευκαιρίες που έχουν διαμορφώσει οι επαγγελματίες της πληροφόρησης οδηγεί σε ένα πρότυπο αναγκών και εκπαίδευσης για την κατάρτιση τους. Τέλος, η τεχνολογία και η πληροφορική απαιτούν συνεχή ενημέρωση διότι οι εξελίξεις αυξάνονται προκαλώντας πρόκληση στους επαγγελματίες της πληροφόρησης [8] .

2.4 Ανοιχτά Δεδομένα

Τα Ανοιχτά Δεδομένα είναι τα δεδομένα που ελεύθερα μπορούν να χρησιμοποιηθούν, να επαναχρησιμοποιηθούν να αναδιανεμηθούν από κάθε χρήστη. Αυτά τα δεδομένα πρέπει να είναι διαθέσιμα αυτούσια, να διαθέτουν ένα συνετό κόστος αναπαραγωγής και να είναι διαθέσιμα κατά προτίμηση για λήψη από το Διαδίκτυο. Οφείλουν να είναι διαθέσιμα σε αναγνώσιμη μορφή και υπό όρους που επιτρέπουν την αναδιανομή και την επαναχρησιμοποίηση τους, συμπεριλαμβανομένης και της ανάμειξης τους με άλλα σύνολα δεδομένων. Για παράδειγμα , δεν τίθενται περιορισμοί για « μη-εμπορική χρήση » ή περιορισμοί για χρήση συγκεκριμένου σκοπού (π.χ. μόνο στην εκπαίδευση). Ουσιαστικά πρόκειται για δεδομένα χωρίς περιορισμούς όπως πνευματικά δικαιώματα, αναφορές σε πατέντες ή άλλους μηχανισμούς ελέγχου με κύριο στόχο την επίτευξη της διαλειτουργικότητας [2] [6] .

Η διαλειτουργικότητα είναι η δυνατότητα διαφορετικών συστημάτων να συνεργάζονται με αποτέλεσμα τη δόμηση πολύπλοκων και μεγαλύτερων συστημάτων. Ένα χαρακτηριστικό παράδειγμα που δίνεται στη βιβλιογραφία είναι η διάσημη ιστορία του Πύργου της Βαβέλ που αποτελεί την απόδειξη, όπου η αδυναμία επικοινωνίας - διαλειτουργικότητας οδήγησε στην αποτυχία της οικοδόμησής του. Ο πυρήνας της «κοινής ωφέλειας» που εντοπίζεται σε δεδομένα και κώδικα έγκειται στο γεγονός ότι ένα τμήμα ανοικτού υλικού που περιέχουν μπορεί να αναμειχθεί με άλλο ανοικτό υλικό, με στόχο την εξέλιξη καλύτερων προϊόντων και υπηρεσιών. Η έννοια λοιπόν των ανοιχτών δεδομένων προωθεί τη διαφάνεια, τη συνεργασία και την καινοτομία, καθιστώντας τα δεδομένα προσβάσιμα για ανάλυση, έρευνα και εφαρμογή σε διάφορους τομείς [6] .

2.4.1 Χαρακτηριστικά Ανοιχτών Δεδομένων

Προσβασιμότητα: Η πρόσβαση σε ανοιχτά δεδομένα γίνεται συνήθως μέσω διαδικτυακών πλατφόρμων ή ειδικών αποθετηρίων. Οι κυβερνήσεις, οι οργανισμοί και τα ιδρύματα συχνά είναι υποχρεωμένοι να δημοσιεύουν ανοικτά τα δεδομένα στους ιστότοπους τους ή μέσω ειδικών πυλών προσβάσιμα προς όλους.

Αδειοδότηση: Τα ανοικτά δεδομένα δημοσιεύονται με “ανοικτές άδειες” που επιτρέπουν την επαναχρησιμοποίηση και την αναδιανομή τους. Συνήθως ανοικτές άδειες για δεδομένα που προωθούν την ελεύθερη διακίνηση των έργων περιλαμβάνουν άδειες Creative Commons (CC) και άδειες Open Data Commons < OPEN DATA >, οι οποίες επιτρέπουν στους χρήστες να χρησιμοποιούν, να τροποποιούν και να διανέμουν ελεύθερα τα δεδομένα, αναφέροντας παράλληλα την πηγή περιορίζοντας τα δικαιώματα πνευματικής ιδιοκτησίας που εκμεταλλεύονται οι δημιουργοί των εν λόγω έργων.

Μορφές: Τα ανοικτά δεδομένα θα πρέπει να παρέχονται σε μη ιδιόκτητους και αναγνώσιμους από κοινά μηχανήματα, ώστε να διευκολύνεται η εύκολη πρόσβαση και επαναχρησιμοποίηση τους. Κάποιες από τις πιο συνηθισμένες μορφές για ανοικτά δεδομένα αποτελούν τα CSV (Comma-Separated Values), JSON (JavaScript Object Notation), XML (eXtensible Markup Language) και RDF (Resource Description Framework).

Διαφάνεια: Τα ανοικτά δεδομένα προωθούν τη διαφάνεια και τη λογοδοσία καθιστώντας τα κυβερνητικά δεδομένα, τα δημόσια αρχεία και τα ερευνητικά πορίσματα να είναι προσβάσιμα στους πολίτες, τους δημοσιογράφους, τους ερευνητές και τους υπεύθυνους χάραξης πολιτικής. Αυτό ουσιαστικά συμβάλλει στη βελτίωση της εμπιστοσύνης, στην προώθηση της συμμετοχής των πολιτών και στη λήψη τεκμηριωμένων αποφάσεων.

Τα ανοιχτά δεδομένα περιλαμβάνουν ένα ευρύ φάσμα από διάφορες πηγές, συμπεριλαμβανομένων κυβερνητικών δεδομένων (π.χ. δεδομένα απογραφής, δεδομένα καιρού, δεδομένα μεταφορών), δεδομένα επιστημονικής έρευνας, εκπαιδευτικούς πόρους και δεδομένα που προέρχονται από το πλήθος (π.χ. OpenStreetMap). Οι ερευνητές, οι επιχειρηματίες, οι προγραμματιστές και οι πολίτες μπορούν να έχουν πρόσβαση και να αναλύουν αυτά τα δεδομένα για να αναπτύξουν νέες εφαρμογές, εργαλεία και γνώσεις που αντιμετωπίζουν κοινωνικές προκλήσεις και προωθούν την οικονομική ανάπτυξη [6] [10] .

Εικόνα 7: Λογότυπα βασικών Creative Commons αδειών
(Πηγή : https://vecpho.com/creative-commons-license/)

2.4.2 Ανοιχτά Δημόσια Δεδομένα

Στη συζήτηση σχετικά με τα δεδομένα χρησιμοποιούνται συχνά οι όροι Ανοιχτά, Δημόσια,Κυβερνητικά. Η διευκρίνιση των παραπάνω όρων καθώς και του πλαισίου

εντός του οποίου χρησιμοποιούνται είναι σημαντική για την κατανόηση του θέματος που εξετάζουμε [6] .

Δημόσια Δεδομένα

Με τον όρο Ανοιχτά Δημόσια Δεδομένα (Open Public Data) γίνεται αναφορά σε « *δεδομένα ή σύνολα δεδομένων που αφορούν το συλλογικό γίγνεσθαι και για τα οποία υφίσταται μία συνειδητή και συνεπής πολιτική η οποία επιτρέπει την ελεύθερη διάθεση και επαναχρησιμοποίηση τους»* [6] .

Ο όρος "Δημόσια Δεδομένα" αναφέρεται σε πληροφορίες και δεδομένα που σχετίζονται με τον δημόσιο τομέα, ανεξαρτήτως εάν προέρχονται από ιδιωτικές ή δημόσιες πηγές.. Είναι σημαντικό να σημειωθεί ότι τα δημόσια δεδομένα δεν περιορίζονται αποκλειστικά σε δεδομένα που σχετίζονται με διοικητικές, κρατικές ή πολιτειακές αρμοδιότητες. Μπορεί επίσης να περιλαμβάνουν δεδομένα που παράγονται από επιχειρήσεις και ιδιωτικές πηγές, εφόσον χρηματοδοτήθηκαν από κρατικούς πόρους για την παραγωγή τους. Αν αυτά προέρχονται από τον ιδιωτικό τομέα, οφείλουν να συμβάλλουν στο κοινωνικό σύνολο.

Κυβερνητικά Δεδομένα

Ο ορισμός των Δημοσίων Δεδομένων είναι ευρύτερος αυτού των Κυβερνητικών Δεδομένων. Με τον όρο Κυβερνητικά Δεδομένα εννοούμε «*το υποσύνολο εκείνο των δημοσίων δεδομένων και πληροφοριών που παράγονται ή συλλέγονται από την κυβέρνηση ή κυβερνητικά ελεγχόμενους οργανισμούς (Δημόσιοι Οργανισμοί και Επιχειρήσεις)* » [6] .

Παρακάτω αναπαρίσταται το οικοσύστημα των Δημόσιων Δεδομένων. Είναι εμφανές ότι τα Ανοιχτά Κυβερνητικά Δεδομένα είναι η τομή των Ανοιχτών Δημόσιων Δεδομένων και των Κυβερνητικών Δεδομένων, τα οποία αποτελούν με την σειρά τους το υποσύνολο των Δημόσιων Δεδομένων.

Εικόνα 8: Το οικοσύστημα των Δημόσιων Δεδομένων [6]

2.4.3 Διαφορές Ανοιχτών και Μεγάλων Συνόλων Δεδομένων και Κίνδυνοι

Υπάρχουν ορισμένες σημαντικές διαφορές μεταξύ των Big Data και των Ανοικτών Δεδομένων, οι οποίες συνδέονται με την ευρύτερη έννοια της διαφάνειας και της δημοκρατίας στην διακυβέρνηση. Συνήθως, τα Big Data που δεν είναι ανοικτά, δεν αντικατοπτρίζουν τις δημοκρατικές αρχές. Αυτά τα σύνολα δεδομένων περιλαμβάνουν διάφορες κατηγορίες, όπως τα δεδομένα που συλλέγονται για την εθνική ασφάλεια από τις εθνικές υπηρεσίες πληροφοριών, ή τα δεδομένα δραστηριότητας που καταγράφονται από μεγάλες εταιρείες λιανικού εμπορίου για τους πελάτες τους. Αυτού του είδους τα Μεγάλα Σύνολα Δεδομένων αποτελούν σημαντικό πλεονέκτημα για τους ανθρώπους ή τις οργανώσεις που τα χρησιμοποιούν, αλλά ταυτόχρονα μπορούν να έχουν επιπτώσεις στα προσωπικά δικαιώματα των άλλων. Η διαφοροποίηση αυτή και το είδος των μεγάλων δεδομένων τίθεται ως ένα πιο τα πιο αμφιλεγόμενα ζητήματα της εποχής μας [8] [6].

Συνήθως η σημασία των δεδομένων συνδέεται άμεσα από το μέγεθος τους, συνεπώς δεν οφείλουν πάντα τα Ανοιχτά Δεδομένα να είναι Μεγάλα. Οι μεγάλες ποσότητες δεδομένων, όταν γίνονται προσβάσιμες στο ευρύ κοινό, μπορούν να έχουν σημαντικό αντίκτυπο. Για παράδειγμα, τα δεδομένα από την τοπική αυτοδιοίκηση μπορούν να βοηθήσουν τους πολίτες να επιλέξουν την κατάλληλη υγειονομική περίθαλψη, να αξιολογήσουν την ποιότητα των τοπικών υπηρεσιών, να συμμετέχουν στη διαμόρφωση του τοπικού προϋπολογισμού, ή ακόμα να συνεισφέρουν στη δημιουργία εφαρμογών που ευνοούν την καθημερινή κίνηση των ανθρώπων μέσω των μέσων μαζικής μεταφοράς.

Τα Big open data δεν περιορίζονται μόνο στα δεδομένα που παρέχει η κυβέρνηση. Πολλοί επιστήμονες μοιράζονται την έρευνά τους σε διάφορους τομείς όπως η γονιδιωματική και η αστρονομία μέσω ενός νέου, συνεργατικού μοντέλου έρευνας. Άλλοι ερευνητές χρησιμοποιούν τα Μεγάλα Σύνολα Δεδομένων που συλλέγονται από τα μέσα κοινωνικής δικτύωσης - τα οποία συχνά είναι ανοικτά για το κοινό - για να αναλύσουν τις τάσεις της αγοράς και τις απόψεις του κοινού.

Ωστόσο, αυτό είναι ιδιαίτερα ισχυρό όταν η κυβέρνηση μετατρέπει τα Big Data σε ανοικτά δεδομένα. Οι κρατικές υπηρεσίες διαθέτουν τους αναγκαίους πόρους και την τεχνογνωσία για να συγκεντρώσουν μεγάλες ποσότητες δεδομένων και μπορούν να αποκομίσουν σημαντικά οικονομικά οφέλη από το άνοιγμα αυτών των συνόλων δεδομένων. Το «άνοιγμα» αυτών των δεδομένων ωστόσο μπορεί να φανεί ακόμα πιο επικίνδυνο καθώς είναι ικανό να προκαλέσει:

Ανησυχίες σχετικά με την προστασία της ιδιωτικής ζωής: Το άνοιγμα μεγάλου όγκου δεδομένων στο κοινό εγείρει ανησυχίες για την προστασία της ιδιωτικής ζωής, ιδίως

εάν τα δεδομένα περιέχουν πληροφορίες προσωπικής ταυτοποίησης (PII) ή ευαίσθητες πληροφορίες για άτομα. Ακόμα και ανωνυμοποιημένα δεδομένα μπορούν μερικές φορές να ταυτοποιηθούν εκ νέου μέσω διασταύρωσης με άλλα σύνολα δεδομένων, θέτοντας σε κίνδυνο την ιδιωτική ζωή των ατόμων.

Παραβιάσεις δεδομένων και κίνδυνοι ασφάλειας: Η δημοσιοποίηση των μεγάλων δεδομένων αυξάνει τον κίνδυνο παραβίασης δεδομένων και μη εξουσιοδοτημένης πρόσβασης. Οι κακόβουλοι φορείς ενδέχεται να εκμεταλλευτούν τα τρωτά σημεία των ανοικτών συστημάτων δεδομένων για να κλέψουν ευαίσθητες πληροφορίες, να διαταράξουν τις υπηρεσίες ή να εξαπολύσουν επιθέσεις στον κυβερνοχώρο, οδηγώντας σε οικονομικές απώλειες και βλάβη της φήμης.

Κατάχρηση και διακρίσεις: Ανοιχτά δεδομένα μπορούν να χρησιμοποιηθούν καταχρηστικά για διακρίσεις εις βάρος ατόμων ή ομάδων με γνώμονα χαρακτηριστικά όπως η εθνικότητα, το φύλο ή η κοινωνικοοικονομική κατάσταση. Οι προκατειλημμένοι αλγόριθμοι και οι διαδικασίες λήψης αποφάσεων που τροφοδοτούνται από τα ανοικτά δεδομένα μπορούν να διαιωνίσουν τις ανισότητες και να ενισχύσουν τις κοινωνικές αδικίες.

Παρερμηνεία και παραπληροφόρηση: Μεγάλοι όγκοι σύνθετων δεδομένων μπορεί να παρερμηνευθούν ή να παραποιηθούν, οδηγώντας σε παραπληροφόρηση και σύγχυση του κοινού. Χωρίς το κατάλληλο πλαίσιο και την κατάλληλη κατανόηση, τα ανοικτά δεδομένα μπορούν να χρησιμοποιηθούν καταχρηστικά για την υποστήριξη ψευδών αφηγήσεων ή παραπλανητικών συμπερασμάτων.

Για να μετριαστούν αυτοί οι κίνδυνοι, είναι απαραίτητο να εφαρμοστούν ισχυρά πλαίσια διακυβέρνησης δεδομένων, τεχνικές διατήρησης της ιδιωτικής ζωής και μέτρα διαφάνειας. Οι πρωτοβουλίες για τα ανοικτά δεδομένα θα πρέπει να θέτουν ως προτεραιότητα την προστασία των δεδομένων, τη συγκατάθεση μετά από ενημέρωση, ώστε να διασφαλίζεται ότι τα οφέλη των ανοικτών δεδομένων μεγιστοποιούνται, ελαχιστοποιώντας παράλληλα τις πιθανές βλάβες. Επιπλέον, τα ενδιαφερόμενα μέρη πρέπει να συμμετέχουν σε συνεχή διάλογο και συνεργασία για την αντιμετώπιση των αναδυόμενων προκλήσεων και την οικοδόμηση εμπιστοσύνης στα συστήματα ανοικτών δεδομένων.

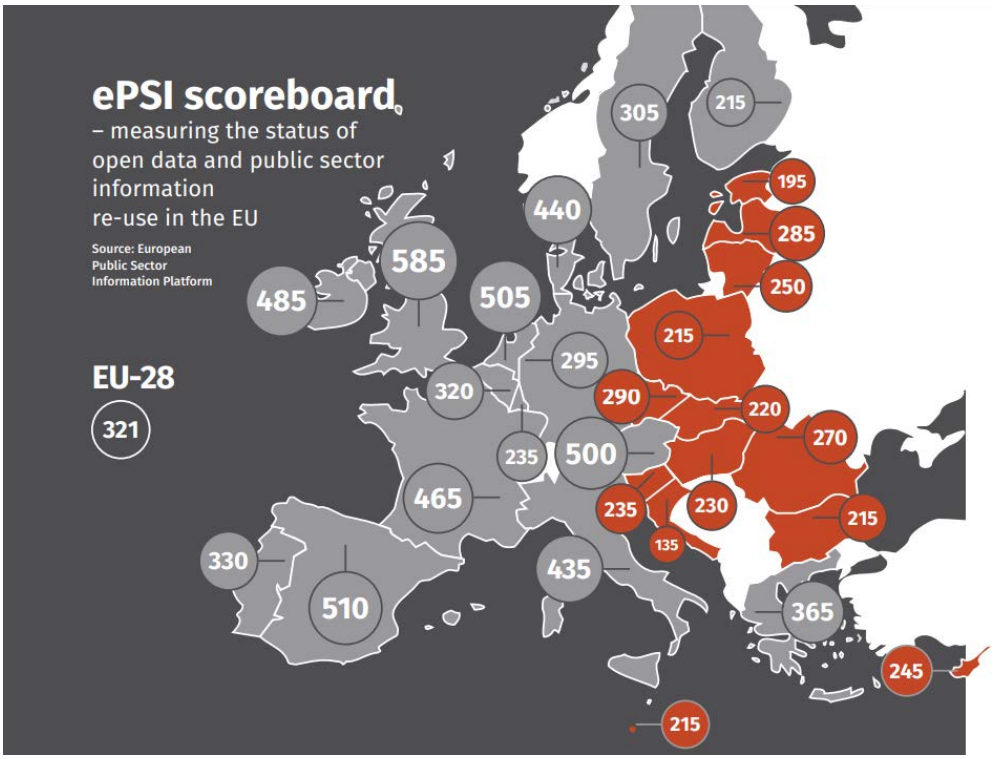
Προσαρμόζοντας τις αρχές των Ανοικτών Δεδομένων στα Big Data, μπορούμε να συμβάλουμε στην αντιμετώπιση ορισμένων από τα πιο δύσκολα θέματα που έχουν προκαλέσει τα μαζικά δεδομένα.

2.5 Τα Δεδομένα στην Ευρώπη

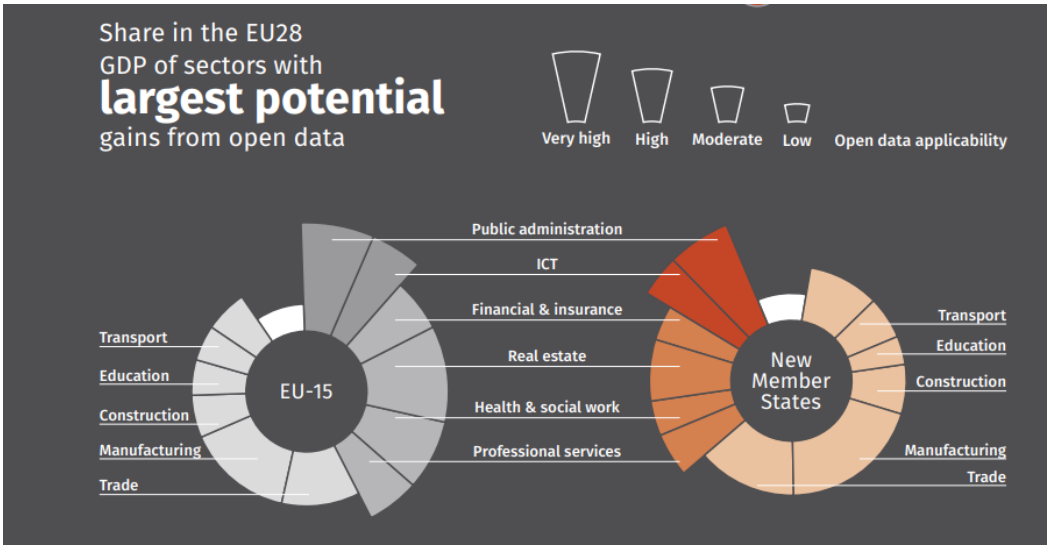
Με τη συνεχή αύξηση του όγκου των δεδομένων και την αύξηση του όγκου των ανοικτών δεδομένων η Ευρωπαϊκή Ένωση έχει στρέψει την προσοχή της στην διαχείριση την εκμετάλλευση αλλά και την προστασία αυτών.

Η τυποποίηση σε διάφορα επίπεδα όπως τα σχήματα μεταδεδομένων, οι μορφές αναπαράστασης δεδομένων και οι όροι αδειοδότησης των ανοικτών δεδομένων είναι απαραίτητη για να καταστεί δυνατή η διαλειτουργικότητα με γενικό στόχο την προώθηση της καινοτομίας. Αυτό αναφέρεται σε όλους τους τύπους πολύγλωσσων δεδομένων, συμπεριλαμβανομένων τόσο δομημένων όσο και μη δομημένων δεδομένων από διαφορετικούς τομείς, όπως, στατιστικά δεδομένα, μετεωρολογικά δεδομένα, πληροφορίες του δημόσιου τομέα (PSI) και ερευνητικά.

Σύμφωνα με την εικόνα 9 απεικονίζονται στο χάρτη της Ευρώπης οι μετρήσεις με το μεγαλύτερο πιθανό κέρδος από επαναξιοποίηση των Ανοιχτών Δεδομένων και στην εικόνα 10 σε γράφημα πίτας το ποσοστό του ΑΕΠ της Ε.Ε που επενδύεται στον κάθε τομέα. Οι 15 χώρες της Ε.Ε του χάρτη είναι με γκρι χρώμα και με πορτοκαλί τα νέα μέλη. Το μέγεθος κάθε κομματιού της “πίτας” αναπαριστά τη συχνότητα εφαρμογής των Ανοιχτών Δεδομένων στους αντίστοιχους κλάδους. Αξίζει να σημειωθούν ότι ο κλάδος της Δημόσιας Διοίκησης υπερτερεί και στα δύο γραφήματα [8] .



Εικόνα 9: Εφαρμογή Ανοιχτών Δεδομένων ανά Ευρωπαϊκή χώρα[8].



Εικόνα 10: Έσοδα από τη χρήση Ανοιχτών Δεδομένων ανά τομέα [8].

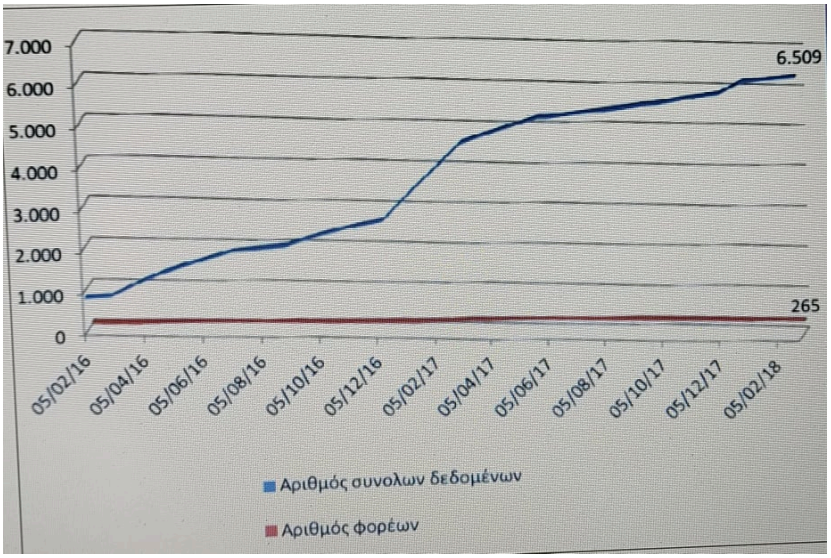
Η Ευρωπαϊκή Επιτροπή διαθέτει μια πύλη ανοιχτών δεδομένων για κάθε τύπο πληροφοριών που βρίσκονται στην κατοχή της Επιτροπής και των άλλων θεσμικών οργάνων και οργανισμών της Ε.Ε . Από τον Δεκέμβριο του 2012 βρίσκεται σε λειτουργία αυτή η ανοιχτή πύλη δεδομένων της Ε.Ε. Η Πύλη Δημόσιων Δεδομένων της ΕΕ: https://open-data.europa.eu/el/data, είναι το ενιαίο σημείο πρόσβασης στα συνεχώς αυξανόμενα και ποικίλα δεδομένα που παράγουν οι οργανισμοί και τα όργανα της Ευρωπαϊκής Ένωσης. Αυτά τα δεδομένα μπορούν να χρησιμοποιηθούν, να επαναχρησιμοποιηθούν, να διαδοθούν μέσω συνδέσμων και να αναδημοσιευθούν για εμπορικούς ή μη σκοπούς, δωρεάν.

2.6 Τα Ανοιχτά Δεδομένα στην Ελλάδα

Η Ελλάδα βρίσκεται να υστερεί στην αξιοποίηση ανοικτών δεδομένων , καθώς στοιχεία του Βαρόμετρου Ανοιχτών Δεδομένων που επεξεργάσθηκε το Ίδρυμα Οικονομικών και Βιομηχανικών Ερευνών (ΙΟΒΕ) δείχνουν πως η χώρα μας υστερεί στον τομέα αυτό των υπολοίπων ευρωπαϊκών κρατών. Από τις 77 χώρες του Βαρόμετρου η Ελλάδα κατατάσσεται στην 27^{η} , υπερισχύει απλώς από τις περισσότερες χώρες της Μέσης Ανατολής, της Κεντρικής Ασίας και της Αφρικής. Το ζητούμενο είναι πώς τα Ανοιχτά Δεδομένα θα μπορούσαν να αξιοποιηθούν κατάλληλα ώστε να συμβάλουν σε μια επιχειρηματική επιτυχία ή μια αποτελεσματική πολιτική όπως συμβαίνει σε αρκετές χώρες του εξωτερικού. Το βασικό προαπαιτούμενο για την ανάπτυξη νέων πολιτικών και μιας επιτυχημένης επιχειρηματικότητας είναι το να δημοσιοποιεί κανείς υψηλής ποιότητας ανοικτά δεδομένα σε ψηφιακή μορφή . Στις μέρες μας, μόνο ένα μικρό ποσοστό δημόσιων οργανισμών κοινοποιεί τα δεδομένα της σε μια μορφή που να

μπορεί να χρησιμοποιηθεί από εξωτερικά συστήματα με στόχο την οικοδόμηση ουσιαστικών υπηρεσιών δεδομένων για τους τελικούς χρήστες. Κυρίως αυτό συμβαίνει λόγω του ότι οι κυβερνήσεις δεν είναι πλήρως πεπεισμένες και ενημερωμένες για τον οικονομικό και κοινωνικό αντίκτυπο των υπηρεσιών που θα δημιουργηθούν με βάση τα ανοικτά δεδομένα [6] .

Σύμφωνα με τον κύριο ελληνικό διαδικτυακό τόπο για Μεγάλα Σύνολα Δεδομένων data.gov.gr εκεί δεν κοινοποιούνται έγγραφα, αλλά σύνολα δεδομένων (data sets) ή σύνδεσμοι προς τους δικτυακούς τόπους των φορέων, όπου αυτά τηρούνται. Στα στοιχεία που αντλήθηκαν με βάση την παραπάνω ιστοσελίδα και απεικονίζονται παρακάτω στην εικόνα 11 , διαπιστώθηκε μια εντυπωσιακή αύξηση κατά την περίοδο Φεβρουάριος 2016 – Οκτώβριος 2018, τόσο στον αριθμό των εγγεγραμμένων φορέων στο σύστημα, όσο και στον αριθμό των διαθέσιμων συνόλων δεδομένων. Ιδιαίτερα ο αριθμός των εγγεγραμμένων φορέων αυξήθηκε από 70 φορείς τον Φεβρουάριο του 2016, σε 159 φορείς τον Ιανουάριο του 2017, για να φθάσει σε 329 φορείς τον Οκτώβριο του 2018. Επιπλέον, τα διαθέσιμα σύνολα δεδομένων, αυξήθηκαν από 847 τον Φεβρουάριο του 2016, σε 3.622 τον Ιανουάριο του 2017, για να φθάσουν σε 8.594 τον Οκτώβριο του 2018.



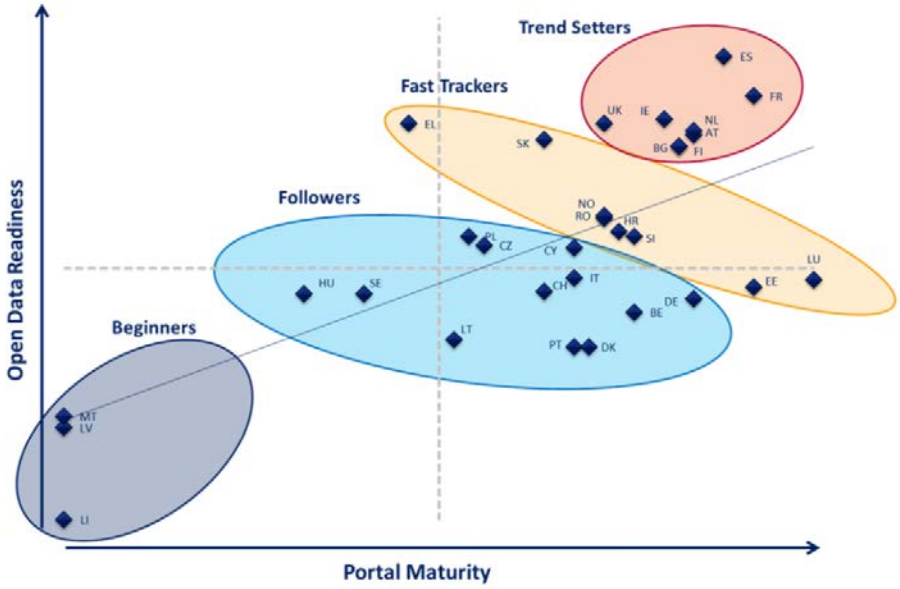
Εικόνα 11: Διαχρονική εξέλιξη αριθμού φορέων και data sets
(Πηγή : https://www.data.gov.gr/)

Στην έκθεση του European Data Portal , καταγράφεται ο βαθμός ωριμότητας των Ανοιχτών Δεδομένων στην Ευρωπαϊκή Ένωση, εντάσσει την Ελλάδα στην κατηγορία των Fast-trackers, δηλαδή στην ομάδα των χωρών που έχουν επιταχύνει σημαντικά την πορεία τους στην αξιοποίηση των Ανοιχτών Δεδομένων, διαμορφώνοντας ουσιαστικά

μια πύλη δεδομένων, ωστόσο αντιμετωπίζοντας ακόμη έναν αριθμό ελλείψεων εικόνα 9. Έχει ελαφρώς αυξηθεί η χρήση των Ανοιχτών Δεδομένων στην Ελλάδα σύμφωνα με όσα αναφέρει η έκθεση επισημαίνει όμως ότι λιγότερο από το 25% αυτών των δεδομένων ενημερώνεται αυτοματοποιημένα ενώ έχει μειωθεί σημαντικά και ο αριθμός των σχετικών οικονομικών ερευνών.

Ο Σύνδεσμος Ελλήνων Βιομηχάνων και Βιοτεχνών (ΣΕΒ), υπολογίζει ότι στην επόμενη 5ετία, το μέγεθος της αξίας των Ανοιχτών Δεδομένων στην Ελλάδα θα πλησιάσει τα 3,2 δις ευρώ και θα αναπτυχθούν 1.000 νέες θέσεις εργασίας στα Ανοικτά Δεδομένα. Στις δημόσιες υπηρεσίες θα εξοικονομηθούν περίπου 30 εκατ. ευρώ, ενώ το συνολικό έμμεσο και άμεσο όφελος από αγορά ανοικτών δεδομένων, νέες υπηρεσίες , εξοικονομήσεις, βελτίωση ποιότητας ζωής και τα λοιπά θα πλησιάσει τα 12 δις ευρώ. Ασφαλώς υπάρχει μεγάλο περιθώριο για βελτίωση, δεδομένου ότι οι εγγεγραμμένοι φορείς στην Εθνική Πύλη Ανοικτών Δεδομένων, είναι 265 από τους 1.730 και έχουν αναρτήσει 6.500 δεδομένα εκ των οποίων όμως μόνο το 16% είναι πλήρως αυτόματα αναγνώσιμα. Ακόμα, παρατηρείται απουσία τυποποίησης και έλλειψη γενικότερης στρατηγικής, απουσία διασύνδεσης μητρώων και πληροφοριακών συστημάτων για τα ανοικτά δεδομένα.

Το Υπουργείο Διοικητικής Ανασυγκρότησης έχοντας εντοπίσει αρκετά από τα προβλήματα αυτά, επιζητά μακροπρόθεσμα την περαιτέρω αύξηση του αριθμού των εγγεγραμμένων φορέων ,αύξηση του αριθμού των αναρτημένων συνόλων δεδομένων, την βελτίωση της ποιότητας των συνόλων δεδομένων που αναρτώνται, την προώθηση συνεργασιών μεταξύ του Υπουργείου Διοικητικής Ανασυγκρότησης με συλλογικούς φορείς που εντάσσονται στο δημόσιο και στον ιδιωτικό τομέα καθώς και με ευρωπαϊκούς οργανισμούς (ESPON EGTC, Ευρωπαϊκός Οικονομικός Χώρος, Expertise France, κα.) για την μελέτη αλλά και υλοποίηση σχετικών δράσεων και εφαρμογών.



Εικόνα 12: Βαθμός ωριμότητας των Ανοιχτών Δεδομένων των χωρών της Ε.Ε (Πηγή : https://wayback.archive-it.org/12090/*/https:/ec.europa.eu/digital-single-market/en/blog/european-countries-are-reaping-benefits-open-data)

3 Βάσεις Δεδομένων και Μεγάλα Σύνολα Δεδομένων

Τα Μεγάλα Σύνολα Δεδομένων και οι βάσεις δεδομένων είναι στενά συνυφασμένες έννοιες στο πεδίο της διαχείρισης και της ανάλυσης δεδομένων. Οι βάσεις δεδομένων χρησιμεύουν ως θεμελιώδη στοιχεία για την αποθήκευση, την οργάνωση και τη διαχείριση δομημένων δεδομένων, ενώ τα Μεγάλα Σύνολα Δεδομένων περιλαμβάνουν μεγάλα και πολύπλοκα σύνολα δεδομένων που μπορεί να μην ταιριάζουν απόλυτα.

3.1 Παλαιότερες Βάσεις Δεδομένων

Αρχικά, κάθε αρχείο αποθηκευόταν χειροκίνητα, αλλά η έλευση της τεχνολογίας οδήγησε σε δραστικές αλλαγές με την πάροδο των ετών. Για να διευκολυνθεί η διατήρηση των δεδομένων δημιουργήθηκαν βάσεις δεδομένων.

Υπάρχουν διάφοροι τρόποι να κατηγοριοποιήσει κανείς τις βάσεις δεδομένων βάση διαφορετικών κριτηρίων. Για παράδειγμα κάποιοι τρόποι κατηγοριοποίησης των βάσεων δεδομένων είναι με βάση το μοντέλο δεδομένων τους, την δομή δεδομένων που αποθηκεύουν, την επεκτασιμότητα τους (οριζόντια ή κάθετη) , τις περιπτώσεις χρήσης τους (πραγματικού χρόνου ή μη, αναλυτική , συναλλακτική κλπ.), το μοντέλο ανάπτυξης (on premises , cloud , hybrid) κα. Εμείς θα αναλύσουμε κάποιες από αυτές σύμφωνα με το μοντέλο δεδομένων τους [11] [12] .

Βάσεις δεδομένων με αξιοσημείωτα μοντέλα δεδομένων στην εξέλιξη του χρόνου είναι:

3.1.1 Ιεραρχικές βάσεις δεδομένων

Οι Ιεραρχικές βάσεις δεδομένων (Hierarchical Databases) εμφανίστηκαν στα τέλη της δεκαετίας του 1950 και απέκτησαν δημοτικότητα τη δεκαετία του 1960. Το σύστημα διαχείρισης πληροφοριών (IMS) της IBM είναι ένα από τα πρώτα παραδείγματα, που παρουσιάστηκε στα μέσα της δεκαετίας του 1960. Όπως σε κάθε ιεραρχία, η βάση δεδομένων ακολουθεί την εξέλιξη των δεδομένων που κατηγοριοποιούνται σε βαθμίδες ή επίπεδα, όπου τα δεδομένα κατηγοριοποιούνται με βάση ένα κοινό σημείο σύνδεσης. Ως αποτέλεσμα, δύο οντότητες δεδομένων θα βρίσκονται σε χαμηλότερη βαθμίδα και το κοινό σημείο θα αναλάβει υψηλότερη βαθμίδα. Μια άλλη οπτική απεικονίζει τα δεδομένα που οργανώνονται σε μια σχέση

γονέα-παιδιού, η οποία κατά την προσθήκη πολλαπλών στοιχείων δεδομένων θα μοιάζει με δέντρο. Σημαντικό είναι ότι λόγω μιας τέτοιας δομής, οι ιεραρχικές βάσεις δεδομένων δεν είναι εύκολα διαχειρίσιμες - η προσθήκη στοιχείων δεδομένων απαιτεί μια μακρά διάσχιση της βάσης δεδομένων. Οι ιεραρχικές βάσεις δεδομένων δεν χρησιμοποιούνται για εφαρμογές μεγάλων δεδομένων λόγω διαφόρων περιορισμών και περιορισμών που τις καθιστούν λιγότερο κατάλληλες για τον χειρισμό της κλίμακας, της ποικιλίας και της πολυπλοκότητας των μεγάλων δεδομένων καθώς αποτελούν πλέον αρκετά παρωχημένη μέθοδο.

3.1.2 Βάσεις δεδομένων δικτύου

Οι δικτυακές βάσεις (Network Databases) δεδομένων εμφανίστηκαν στα τέλη της δεκαετίας του 1960 και στις αρχές της δεκαετίας του 1970 ως εξέλιξη των ιεραρχικών βάσεων δεδομένων. Αναπτύχθηκαν για να αντιμετωπίσουν ορισμένους από τους περιορισμούς των ιεραρχικών βάσεων δεδομένων, ιδίως όσον αφορά τις σχέσεις δεδομένων και την πλοήγηση. Με απλά λόγια, μια δικτυακή βάση δεδομένων είναι μια ιεραρχική βάση δεδομένων, αλλά με μια σημαντική αλλαγή. Οι εγγραφές-παιδιά έχουν την ελευθερία να συσχετίζονται με πολλαπλές εγγραφές-γονείς. Ως αποτέλεσμα, παρατηρείται ένα δίκτυο ή δίχτυ αρχείων βάσης δεδομένων που συνδέονται με πολλαπλά νήματα. Αν και οι δικτυακές βάσεις δεδομένων μπορεί να εξακολουθούν να χρησιμοποιούνται σε ορισμένα παλαιά συστήματα ή εξειδικευμένες εφαρμογές όπου επικρατούν ιεραρχικές δομές δεδομένων, συνήθως δεν θεωρούνται η βέλτιστη επιλογή για εφαρμογές μεγάλων δεδομένων λόγω των περιορισμών τους όσον αφορά και πάλι την επεκτασιμότητα, την ευελιξία και την υποστήριξη μη δομημένων δεδομένων.

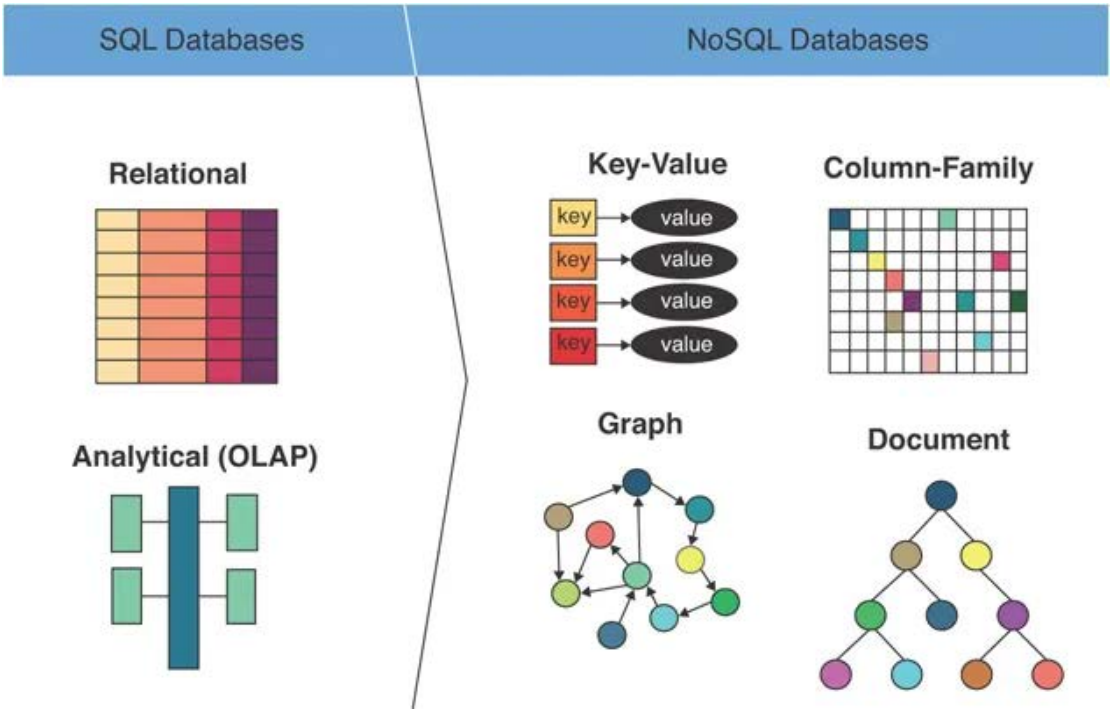
Θα επιμείνουμε ωστόσο στις επόμενες δυο βασικές κατηγορίες αφού μας ενδιαφέρει η χρήση τους ως προς την ανάλυση και την διαχείριση Μεγάλων Δεδομένων.

3.2 Σχεσιακές Βάσεις Δεδομένων και SQL

3.2.1 Σχεσιακό σύστημα διαχείρισης βάσεων δεδομένων

Μια βάση δεδομένων ποικίλλει από ένα απλό έγγραφο κειμένου έως πολύ πιο πολύπλοκο σύνολο πληροφοριών σε βάσεις δεδομένων. Αυτές οι βάσεις δεδομένων πρέπει να βελτιώνονται περιοδικά για την αφαίρεση κάθε είδους περιττού, ασυνεπούς ή “βρώμικου” δεδομένου ώστε να αποδίδουν αποτελεσματικά. Η πιο συνηθισμένη αντίληψη για την αποθήκευση αυτών των δεδομένων είναι μέσω σχεσιακών μοντέλων. Η δομημένη γλώσσα ερωτήσεων (SQL) εξάγει τα σχετικά δεδομένα από τη “δεξαμενή” της βάσης δεδομένων. Οι σχεσιακές βάσεις δεδομένων είναι ο πιο κοινός τύπος βάσης

δεδομένων λόγω της απλότητάς του. Σε ένα RDBMS (Relational Database Management System) τα δεδομένα χωρίζονται σε διάφορους πίνακες στους οποίους μπορεί να γίνει πρόσβαση σύμφωνα με τις απαιτήσεις χωρίς να γίνονται αλλαγές στον πίνακα. Λειτουργίες όπως η ένωση, η συνάθροιση, η πρόσθεση, η δημιουργία, η ανάκτηση και διαγραφή εκτελούνται εύκολα στις σχεσιακές βάσεις δεδομένων και είναι επίσης πολύ εύκολη η επέκταση ή η τροποποίηση υφιστάμενων πινάκων. Οι σχεσιακές βάσεις δεδομένων αναπτύχθηκαν στις αρχές της δεκαετίας του 1970 από τον Edgar F. Codd, ο οποίος εισήγαγε το σχεσιακό μοντέλο δεδομένων. Το πρώτο εμπορικά διαθέσιμο σύστημα διαχείρισης σχεσιακών βάσεων δεδομένων (RDBMS) ήταν η Oracle, που κυκλοφόρησε το 1979 [13] [12] .



Εικόνα 13: Κατηγορίες βάσεων δεδομένων [13]

3.2.2 Σχεσιακές αναλυτικές βάσεις δεδομένων OLAP

Ορισμένες αναλυτικές βάσεις δεδομένων OLAP βασίζονται σε σχεσιακά συστήματα διαχείρισης βάσεων δεδομένων και αποθηκεύουν δεδομένα σε σχεσιακούς πίνακες (RDBMS). Αυτές οι βάσεις δεδομένων χρησιμοποιούν ένα πολυδιάστατο σχήμα γνωστό ως σχήμα αστέρα ή σχήμα χιονονιφάδας, όπου οι πίνακες γεγονότων που περιέχουν αριθμητικά μέτρα περιβάλλονται από πίνακες διαστάσεων που αντιπροσωπεύουν διάφορα χαρακτηριστικά ή διαστάσεις.

Οι βάσεις δεδομένων OLAP είναι βελτιστοποιημένες για την ανάλυση και την υποβολή ερωτημάτων σε πολυδιάστατα δεδομένα, επιτρέποντας στους χρήστες να τεμαχίζουν, να κάνουν drill-down και να αναδιπλώνουν δεδομένα κατά μήκος διαφόρων διαστάσεων για να αποκτήσουν πληροφορίες σχετικά με τις επιχειρηματικές επιδόσεις, τις τάσεις και τα πρότυπα. Οι αναλυτικές βάσεις δεδομένων OLAP είναι προσαρμοσμένες για την υποστήριξη αναλυτικών ερωτημάτων και εργασιών αναφοράς και όχι για την επεξεργασία συναλλαγών. Ενώ τόσο οι σχεσιακές όσο και οι αναλυτικές βάσεις δεδομένων εξυπηρετούν σημαντικούς ρόλους στη διαχείριση και ανάλυση δεδομένων, είναι βελτιστοποιημένες για διαφορετικές περιπτώσεις χρήσης και διαφορετικά πρότυπα ερωτημάτων. Οι σχεσιακές βάσεις δεδομένων επικεντρώνονται στη διαχείριση δεδομένων σε επίπεδο συναλλαγών, ενώ οι αναλυτικές βάσεις δεδομένων επικεντρώνονται στην αναλυτική αναζήτηση και την πολυδιάστατη ανάλυση μεγάλων συνόλων δεδομένων.

Δεν χρησιμοποιούν όλες οι σχεσιακές βάσεις δεδομένων την SQL (Structured Query Language) ως γλώσσα ερωτημάτων, αλλά η SQL είναι η πιο κοινή και ευρέως χρησιμοποιούμενη γλώσσα ερωτημάτων για σχεσιακές βάσεις δεδομένων συνεπώς οι σχεσιακές βάσεις συμπίπτουν συχνά με τον όρο SQL Databases. Οι βάσεις δεδομένων SQL χρησιμοποιούνται συνήθως για εφαρμογές μεγάλων δεδομένων, ιδίως σε σενάρια όπου τα δεδομένα είναι δομημένα και σχεσιακής φύσης [14] .

3.2.3 Χρησιμότητα σχεσιακών βάσεων στα Μεγάλα Σύνολα Δεδομένων

Οι λόγοι για τα οποίους οι βάσεις δεδομένων SQL χρησιμοποιούνται για Μεγάλα Σύνολα Δεδομένων οφείλεται στα παρακάτω χαρακτηριστικά:

Δομή δεδομένων: Οι βάσεις δεδομένων SQL υπερέχουν στη διαχείριση δομημένων δεδομένων οργανωμένων σε πίνακες με γραμμές και στήλες. Πολλές εφαρμογές μεγάλων δεδομένων ασχολούνται με δομημένα δεδομένα, όπως δεδομένα συναλλαγών, αρχεία πελατών και οικονομικά δεδομένα, τα οποία είναι κατάλληλα για αποθήκευση και ανάλυση σε βάσεις δεδομένων SQL.

Ακεραιότητα δεδομένων: Οι βάσεις δεδομένων SQL επιβάλλουν την ακεραιότητα των δεδομένων μέσω περιορισμών, όπως τα “πρωτεύοντα κλειδιά”(primary keys) , τα ξένα κλειδιά (foreign keys), και οι περιορισμοί ελέγχου, διασφαλίζοντας τη συνέπεια και την ακρίβεια των δεδομένων. Αυτό είναι ζωτικής σημασίας για τη διατήρηση της ποιότητας των δεδομένων, ιδίως σε εφαρμογές μεγάλου όγκου και κρίσιμων εφαρμογών μεγάλων δεδομένων.

Ιδιότητες ACID: Οι βάσεις δεδομένων SQL υποστηρίζουν τις ιδιότητες ACID (Atomicity, Consistency, Isolation, Durability), εξασφαλίζοντας τη συναλλακτική συνέπεια και αξιοπιστία. Αυτό είναι σημαντικό για εφαρμογές που απαιτούν αυστηρές εγγυήσεις σχετικά με την ορθότητα και την αξιοπιστία των συναλλαγών δεδομένων, όπως τα χρηματοπιστωτικά συστήματα και οι πλατφόρμες ηλεκτρονικού εμπορίου.

Γλώσσα ερωτημάτων : Η SQL παρέχει μια ισχυρή και τυποποιημένη γλώσσα ερωτημάτων για την αλληλεπίδραση με σχεσιακές βάσεις δεδομένων. Το συντακτικό της SQL επιτρέπει στους χρήστες να εκφράζουν σύνθετα ερωτήματα και λειτουργίες με συνοπτικό και διαισθητικό τρόπο. Μάλιστα οι αναλυτικές βάσεις δεδομένων ενσωματώνονται συχνά με εργαλεία επιχειρηματικής ευφυΐας (BI) και πλατφόρμες αναφοράς για τη διευκόλυνση της οπτικοποίησης δεδομένων, της δημιουργίας πινάκων και της ad-hoc ανάλυσης. Οι χρήστες μπορούν να αξιοποιήσουν αυτά τα εργαλεία για να δημιουργήσουν δια δραστικές αναφορές, διαγράμματα και πίνακες οργάνων με βάση τα δεδομένα που είναι αποθηκευμένα στη βάση δεδομένων

Ώριμο οικοσύστημα: Οι βάσεις δεδομένων SQL διαθέτουν ένα ώριμο οικοσύστημα με ένα ευρύ φάσμα εργαλείων, βιβλιοθηκών και πλαισίων για τη διαχείριση δεδομένων, την ανάλυση και την επιχειρηματική ευφυΐα. Αυτό διευκολύνει τους οργανισμούς να ενσωματώσουν τις βάσεις δεδομένων SQL στις υπάρχουσες ροές εργασίας και υποδομές τους για επεξεργασία και ανάλυση μεγάλων δεδομένων.

Οι βάσεις αυτές ήταν οι πρώτες που υιοθετήθηκαν ευρέως για τη διαχείριση και ανάλυση μεγάλων όγκων δομημένων δεδομένων σε διάφορους κλάδους και περιπτώσεις χρήσης και κυριαρχούσαν έως ότου εμφανιστούν οι NoSQL.

Παραδείγματα σχεσιακών βάσεων δεδομένων που χρησιμοποιούν SQL περιλαμβάνουν: Oracle Database, MySQL, Microsoft SQL , Server PostgreSQL , SQLite, IBM Db2

Ωστόσο, υπάρχουν επίσης σχεσιακές βάσεις δεδομένων που χρησιμοποιούν εξειδικευμένες ή ιδιόκτητες γλώσσες ερωτημάτων ή διαλέκτους αντί της τυπικής SQL. Για παράδειγμα: Η IBM Informix χρησιμοποιεί μια διάλεκτο της SQL που ονομάζεται Informix-SQL. Η Teradata χρησιμοποιεί μια γλώσσα ερωτημάτων βασισμένη στην SQL που ονομάζεται Teradata SQL. Η Microsoft Access χρησιμοποιεί μια διάλεκτο της SQL γνωστή ως Access SQL. Το Progress OpenEdge χρησιμοποιεί μια ιδιόκτητη γλώσσα ερωτημάτων που ονομάζεται ABL (Advanced Business Language) [13] [15] .

3.3 NoSQL Βάσεις Δεδομένων

3.3.1 Από SQL σε NoSQL

Η μεγαλύτερη πρόκληση με τα συνεχώς αυξανόμενα Μεγάλα Σύνολα Δεδομένων είναι η ανομοιομορφία τους. Λόγω αυτού του προβλήματος, τα τελευταία χρόνια, απαιτείται μια μη σχεσιακή βάση δεδομένων που να ανταποκρίνεται στις αυξανόμενες ανάγκες της βιομηχανίας και ταυτόχρονα να είναι ιδιαίτερα αποδοτική. Έτσι προέκυψαν οι βάσεις δεδομένων NoSQL, οι οποίες είναι εξαιρετικά κλιμακούμενες, αποδοτικές και μπορούν να αποθηκεύουν μεγάλο όγκο δεδομένων. Αν και τα RDBMS είναι σε θέση να διαχειρίζονται και τα τρία είδη δεδομένων, δηλαδή δομημένα, ημιδομημένα και αδόμητα, απαιτείται εργασία και συμβιβασμοί για την επίτευξη αποτελεσματικής αποθήκευσης αδόμητων και ημιδομημένων δεδομένων. Τα RDBMS αποθηκεύουν τα δομημένα δεδομένα ως έχουν, επειδή βρίσκονται ήδη στην απαιτούμενη μορφή. Όμως, η αποθήκευση των ημιδομημένων δεδομένων ενέχει ορισμένες πολυπλοκότητες. Τα ημιδομημένα δεδομένα πρέπει να μετατραπούν σε σχεσιακά δεδομένα πριν από την αποθήκευση.

Ως εκ τούτου αυτή η ανομοιομορφία των δεδομένων, οδήγησε στη δημιουργία των βάσεων δεδομένων NoSQL που συχνά αναφέρονται ως βάσεις δεδομένων "Not Only SQL". Οι βάσεις δεδομένων NoSQL είναι μια κατηγορία βάσεων δεδομένων που αποκλίνουν από το παραδοσιακό μοντέλο σχεσιακών βάσεων δεδομένων και τις επιλέγουμε ανάλογα με χαρακτηριστικά όπως η επεκτασιμότητα, η διαθεσιμότητα και η ανοχή σε σφάλματα. Δεν ακολουθούν τη γενική προσέγγιση πίνακα/γραμμή/στήλη που εφαρμόζεται από όλα τα RDBMS. Οι NoSQL ονομάζονται κυρίως κατανεμημένες ή μη σχεσιακές βάσεις δεδομένων. Οι βάσεις δεδομένων NoSQL έχουν σχεδιαστεί για να χειρίζονται μεγάλους όγκους αδόμητων ή ημιδομημένων δεδομένων και παρέχουν μεγαλύτερη ευελιξία και επεκτασιμότητα από τις παραδοσιακές σχεσιακές βάσεις δεδομένων. Τα χαρακτηριστικά των NoSQL βάσεων δεδομένων είναι το μη σχεσιακό μοντέλο δεδομένων που ακολουθούν, η επεκτασιμότητα, η ευελιξία και η υψηλή τους απόδοση. Παρακάτω θα αναφερθούμε σε κάποιες σημαντικές βάσεις δεδομένων που εμπίπτουν στην κατηγορία NoSQL Databases [13] [15] .

3.3.2 Χρησιμότητα NoSQL Βάσεων Δεδομένων Μεγάλα Σύνολα Δεδομένων

Οι βάσεις δεδομένων NoSQL λοιπόν έχουν πολύ κρίσιμο ρόλο στην ανάλυση μεγάλων δεδομένων, κυρίως λόγω των μοναδικών δυνατοτήτων τους που είναι προσαρμοσμένες στη διαχείριση μεγάλων, ποικίλων και συνεχώς εξελισσόμενων συνόλων δεδομένων. Αυτές οι βάσεις δεδομένων είναι ειδικά σχεδιασμένες για κατανεμημένα συστήματα ή περιβάλλοντα που βασίζονται στο νέφος, όπου οι παραδοσιακές σχεσιακές βάσεις δεδομένων μπορεί να δυσκολεύονται να

συμβαδίσουν. Πιο αναλυτικά τα χαρακτηριστικά που ξεχωρίζουν τις βάσεις αυτές για την επεξεργασία Μεγάλων Δεδομένων είναι:

Επεκτασιμότητα: Οι βάσεις δεδομένων NoSQL έχουν σχεδιαστεί για να κλιμακώνονται οριζόντια σε κατανεμημένες συστάδες υλικού βασικών προϊόντων, επιτρέποντάς τους να διαχειρίζονται μεγάλους όγκους δεδομένων και φόρτους εργασίας υψηλής απόδοσης. Αυτή η επεκτασιμότητα είναι απαραίτητη για την επεξεργασία μαζικών συνόλων δεδομένων σε εργασίες ανάλυσης μεγάλων δεδομένων, όπου οι παραδοσιακές σχεσιακές βάσεις δεδομένων μπορεί να δυσκολεύονται να συμβαδίσουν με την κλίμακα των δεδομένων.

Ευελιξία: Οι βάσεις δεδομένων NoSQL προσφέρουν ευέλικτα μοντέλα δεδομένων που μπορούν να φιλοξενήσουν ποικίλους τύπους και δομές δεδομένων, συμπεριλαμβανομένων των αδόμητων και ημιδομημένων δεδομένων. Αυτή η ευελιξία ενδείκνυται για την ανάλυση μεγάλων δεδομένων, όπου τα δεδομένα μπορεί να προέρχονται από διάφορες πηγές και να έχουν διαφορετικές μορφές, όπως κείμενο, εικόνες, δεδομένα αισθητήρων και ροές από μέσα κοινωνικής δικτύωσης.

Απόδοση: Οι βάσεις δεδομένων NoSQL είναι βελτιστοποιημένες για πρόσβαση και επεξεργασία δεδομένων υψηλής απόδοσης και χαμηλής καθυστέρησης. Αξιοποιούν κατανεμημένες αρχιτεκτονικές, προσωρινή αποθήκευση στη μνήμη και παράλληλη επεξεργασία για την επίτευξη γρήγορων λειτουργιών ανάγνωσης και εγγραφής, γεγονός που τις καθιστά ιδανικές για αναλύσεις σε πραγματικό χρόνο και δια δραστική εξερεύνηση δεδομένων σε ροές εργασίας ανάλυσης μεγάλων δεδομένων.

Schema-on-Read προσέγγιση: Οι βάσεις δεδομένων NoSQL χρησιμοποιούν μια προσέγγιση schema-on-read, επιτρέποντας την εισαγωγή και αποθήκευση δεδομένων χωρίς προκαθορισμένα σχήματα. Αυτή η ευελιξία σχήματος απλοποιεί τη διαδικασία εισαγωγής δεδομένων και προσαρμόζεται στις εξελισσόμενες απαιτήσεις δεδομένων στην ανάλυση μεγάλων δεδομένων, όπου οι δομές δεδομένων μπορεί να αλλάζουν με την πάροδο του χρόνου.

Οριζόντια επεκτασιμότητα: Οι βάσεις δεδομένων NoSQL προσφέρουν οριζόντια επεκτασιμότητα, επιτρέποντάς τους να επεκταθούν σε κατανεμημένες συστάδες προσθέτοντας περισσότερους κόμβους στη συστάδα. Οριζόντια κλιμάκωση σημαίνει αύξηση της χωρητικότητας ενός συστήματος με την προσθήκη πρόσθετων μηχανών (κόμβων), σε αντίθεση με την αύξηση των δυνατοτήτων των υφιστάμενων μηχανών. Για παράδειγμα, έστω ότι έχουμε μια εφαρμογή με μια βάση δεδομένων cloud που έχει φθάσει στα όρια του διακομιστή στον οποίο εκτελείται: μια μοναδική περίπτωση GCP 8 vCPU με 32 GB RAM. Για να κλιμακώσουμε οριζόντια αυτή τη βάση δεδομένων, θα μπορούσαμε να την κατατμήσουμε σε δύο επιπλέον κόμβους διακομιστή, ο καθένας με 8 vCPU και 32 GB RAM. Ενώ κάθε μηχάνημα ξεχωριστά έχει την ίδια χωρητικότητα με τον αρχικό μας διακομιστή, μπορούμε τώρα να κατανείμουμε τον φόρτο εργασίας μας

σε αυτούς τους τρεις κόμβους, γεγονός που με τη σειρά του θα μας επιτρέψει να χειριστούμε βαρύτερους φόρτους εργασίας περισσότερων μεγάλων δεδομένων

Ανάλυση σε πραγματικό χρόνο: Οι βάσεις δεδομένων NoSQL υποστηρίζουν την ανάλυση και επεξεργασία δεδομένων ροής σε πραγματικό χρόνο, επιτρέποντας στους χειριστές να αναλύουν και να αντλούν πληροφορίες από τα δεδομένα καθώς αυτά παράγονται ή εισάγονται. Αυτή η δυνατότητα επιτρέπει στις επιχειρήσεις να λαμβάνουν αποφάσεις βάσει δεδομένων σε πραγματικό χρόνο και να ανταποκρίνονται γρήγορα στις μεταβαλλόμενες συνθήκες της αγοράς ή στις αναδυόμενες τάσεις [13] [15] .

3.3.3 Παραδείγματα NoSQL Βάσεων Δεδομένων

3.3.3.1 Apache Cassandra Βάση Δεδομένων

Μια βάση δεδομένων βασικών αξιών, ή key-value store, είναι ένα πρότυπο αποθήκευσης δεδομένων σχεδιασμένο για την αποθήκευση, ανάκτηση και διαχείριση συσχετιστικών πινάκων και μια δομή δεδομένων πιο γνωστή σήμερα ως λεξικό ή πίνακας κατακερματισμού. Τα λεξικά περιέχουν μια συλλογή αντικειμένων ή εγγραφών, οι οποίες με τη σειρά τους έχουν πολλά διαφορετικά πεδία εντός τους, καθένα από τα οποία περιέχει δεδομένα. Αυτές οι εγγραφές αποθηκεύονται και ανακτώνται χρησιμοποιώντας ένα κλειδί που προσδιορίζει μοναδικά την εγγραφή και χρησιμοποιείται για την εύρεση των δεδομένων μέσα στη βάση δεδομένων.

Η Apache Cassandra είναι μια μαζικά κλιμακούμενη βάση δεδομένων NoSQL χαρακτηριστικό παράδειγμα μιας key value βάσης δεδομένων . Οι τεχνικές ρίζες της Cassandra βρίσκονται σε εταιρείες που αναγνωρίζονται για την ικανότητά τους να διαχειρίζονται αποτελεσματικά Μεγάλα Σύνολα Δεδομένων - Google, Amazon, και Facebook - με το Facebook να παραδίδει το Cassandra στο Apache Foundation το 2009. Χρησιμοποιείται σήμερα από πολυάριθμες σύγχρονες επιχειρήσεις για τη διαχείριση των κρίσιμων υποδομών δεδομένων τους. Η Cassandra είναι γνωστή ως η λύση στην οποία στρέφονται οι τεχνικοί επαγγελματίες όταν χρειάζονται μια βάση δεδομένων NoSQL που παρέχει υψηλές επιδόσεις σε τεράστια κλίμακα, η οποία δεν πέφτει ποτέ. Η αρχιτεκτονική της Cassandra συμβάλλει σε μεγάλο βαθμό στην ικανότητα της να κλιμακώνεται, να αποδίδει και να προσφέρει συνεχή διαθεσιμότητα. Η Cassandra χτίστηκε από την αρχή με την κατανόηση ότι μπορούν να συμβούν και συμβαίνουν αστοχίες υλικού και συστήματος. Αυτό μεταφράζεται στο ότι η Cassandra διαθέτει ένα διαφορετικό τρόπο διαχείρισης και προστασίας των μεγάλων δεδομένων από ένα παραδοσιακό RDBMS. Αντί να χρησιμοποιείται ένα παλαιό master-slave ή ένας χειροκίνητος και δύσκολα συντηρήσιμος σχεδιασμός sharded, το Cassandra έχει μια ομότιμη κατανεμημένη αρχιτεκτονική που είναι πολύ πιο κομψή και εύκολη στη

δημιουργία και τη συντήρηση. Στην Cassandra, όλοι οι κόμβοι είναι ίδιοι- δεν υπάρχει η έννοια του κύριου κόμβου, με όλους τους κόμβους επικοινωνούν μεταξύ τους μέσω ενός πρωτοκόλλου. Η αρχιτεκτονική της Cassandra που είναι κατασκευασμένη για κλίμακα σημαίνει ότι είναι ικανή να διαχειρίζεται petabytes πληροφοριών και χιλιάδες ταυτόχρονους χρήστες/λειτουργίες ανά δευτερόλεπτο (σε πολλαπλά κέντρα δεδομένων) με την ίδια ευκολία που μπορεί να διαχειρίζεται πολύ μικρότερες ποσότητες δεδομένων και χρηστών κίνησης [16] [17] [18] .

3.3.3.2 MongoDB Βάση Δεδομένων

Μια βάση δεδομένων προσανατολισμένη στα έγγραφα (document-oriented), ή αποθήκη εγγράφων (document store) , είναι ένα σύστημα αποθήκευσης δεδομένων σχεδιασμένο για την αποθήκευση, ανάκτηση και διαχείριση πληροφοριών προσανατολισμένων στα έγγραφα γνωστά και ως ημιδομημένα δεδομένα.Οι βάσεις δεδομένων προσανατολισμένες στα έγγραφα είναι μία από τις κύριες κατηγορίες των NoSQL βάσεων δεδομένων. Οι βάσεις δεδομένων XML είναι μια υποκατηγορία των βάσεων δεδομένων προσανατολισμένων στα έγγραφα που είναι βελτιστοποιημένες για να λειτουργούν με έγγραφα XML. Οι προσανατολισμένες στα έγγραφα βάσεις δεδομένων είναι εγγενώς μια υποκατηγορία του key-value store. Η διαφορά έγκειται στον τρόπο επεξεργασίας των δεδομένων- σε ένα κατάστημα κλειδιών-τιμών, τα δεδομένα θεωρούνται εγγενώς αδιαφανή για τη βάση δεδομένων, ενώ ένα σύστημα προσανατολισμένο στα έγγραφα βασίζεται στην εσωτερική δομή του εγγράφου προκειμένου να εξαχθούν μεταδεδομένα που χρησιμοποιεί η μηχανή για περαιτέρω βελτιστοποίηση. Αν και η διαφορά είναι συχνά αμελητέα λόγω των εργαλείων των συστημάτων, εννοιολογικά το document-store έχει σχεδιαστεί για να προσφέρει μια πλουσιότερη εμπειρία με σύγχρονες τεχνικές προγραμματισμού. Ένα χαρακτηριστικό παράδειγμα βάσης δεδομένων προσανατολισμένη στα έγγραφα είναι η MongoDB[16] [20] .

Η MongoDB είναι μια βάση δεδομένων προσανατολισμένη στα έγγραφα. Αποθηκεύει δεδομένα σε ευέλικτα έγγραφα τύπου JSON, γεγονός που την καθιστά μέρος της κατηγορίας βάσεων δεδομένων προσανατολισμένων στα έγγραφα στην οικογένεια των NoSQL βάσεων δεδομένων. Η MongoDB είναι μία από τις πιο τυπικές βάσεις δεδομένων NoSQL. Διαθέτει υψηλές επιδόσεις και επεκτασιμότητα της μεθόδου αποθήκευσης τιμών-κλειδιών και πλούσιες λειτουργίες επεξεργασίας δεδομένων. Τα Μεγάλα Σύνολα Δεδομένων περιλαμβάνουν έγγραφα, ηλεκτρονικό ταχυδρομείο, βίντεο, εικόνες και άλλους τύπους δεδομένων, τα οποία σε αντίθεση με τα αυστηρά δομημένα δεδομένα των σχεσιακών δεδομένων, συνήθως δεν βασίζονται σε μορφή γραμμής και στήλης και έχουν τεράστιο όγκο δεδομένων που απαιτεί πολύ αποθηκευτικό χώρο. Σε πολλές εφαρμογές, όπως οι μηχανές αναζήτησης, οι απαιτήσεις

τους σε συνέπεια και ακεραιότητα μπορούν να μειωθούν, αλλά χρειάζονται περισσότερη διαθεσιμότητα, επεκτασιμότητα και απόδοση. Η MongoDB έχει τα ακόλουθα χαρακτηριστικά: Το μοντέλο δεδομένων είναι βολικό στο σχεδιασμό. Υψηλές επιδόσεις. Χωρίς joins και χωρίς συναλλαγές κάνει την πρόσβαση γρήγορη, με πλήρη ευρετήρια σε ενσωματωμένα έγγραφα και πίνακες, ενημέρωση in-place και προαιρετικές ασύγχρονες εγγραφές. Υψηλή διαθεσιμότητα και ισχυρή ευρωστία. Εύκολη επεκτασιμότητα. Αυτόματο διαχωρισμό με διαμερισμό “με διατήρηση της τάξης” (order-preserving partitioning) το οποίο καθιστά την απόδοση της κατανεμημένης ανάγνωσης και εγγραφής γρήγορη και αποτελεσματική. Ενώ επίσης διαθέτει εύκολη οριζόντια επεκτασιμότητα με χαμηλού κόστος υλικό.

Εικόνα 14: Λογότυπα αναφερόμενων Βάσεων Δεδομένων

3.3.3.3 Apache HBase Βάση Δεδομένων

Το Column Family Store (Αποθήκευση Οικογένειας Στήλης) , επίσης γνωστό ως Wide Column Store (Αποθήκευση Ευρείας Στήλης), είναι ένα μοντέλο αποθήκευσης δεδομένων που οργανώνει τα δεδομένα με προσανατολισμό προς τις στήλες. Σε αντίθεση με την παραδοσιακή αποθήκευση με βάση τις γραμμές, όπου τα δεδομένα αποθηκεύονται σε γραμμές, το Column Family Store αποθηκεύει τα δεδομένα σε στήλες, επιτρέποντας την αποτελεσματική πρόσβαση και ανάκτηση συγκεκριμένων στηλών ή οικογενειών στηλών. Κάθε οικογένεια στηλών μπορεί να περιέχει πολλαπλές στήλες και κάθε στήλη μπορεί να έχει πολλαπλές εκδόσεις ή χρονοσφραγίδες. Αυτή η ευελιξία επιτρέπει την αποθήκευση μεγάλων ποσοτήτων δεδομένων με διαφορετικές δομές, καθιστώντας το κατάλληλο για εφαρμογές που απαιτούν υψηλή κλιμάκωση και

απόδοση όπως η διαχείριση μεγάλων δεδομένων. Σε ένα Column Family Store, τα δεδομένα αποθηκεύονται σε οικογένειες στηλών, οι οποίες είναι λογικές ομαδοποιήσεις σχετικών στηλών. Κάθε οικογένεια στηλών μπορεί να έχει διαφορετικό σχήμα ή δομή, επιτρέποντας την αποθήκευση ετερογενών δεδομένων. Μέσα σε κάθε οικογένεια στηλών, τα δεδομένα οργανώνονται περαιτέρω σε γραμμές ή κλειδιά. Οι στήλες εντός μιας οικογένειας στηλών αποθηκεύονται σε ξεχωριστά φυσικά αρχεία ή δομές δεδομένων, επιτρέποντας αποτελεσματικές λειτουργίες ανάγνωσης και εγγραφής σε συγκεκριμένες στήλες ή υποσύνολα στηλών. Αυτή η αποθήκευση με προσανατολισμό στις στήλες επιτρέπει καλύτερη συμπίεση και βελτιωμένη απόδοση για αναλυτικά ερωτήματα που περιλαμβάνουν αθροίσεις ή σαρώσεις σε πολλαπλές στήλες.

Μια τυπική περίπτωση column family βάσης δεδομένων είναι η Apache HBase. Η HBase είναι η συντομογραφία της βάσης δεδομένων Hadoop ως κλιμακούμενος χώρος αποθήκευσης μεγάλων δεδομένων και τρέχει πάνω στο λογισμικό Hadoop, στο οποίο θα αναφερθούμε εκτενώς παρακάτω. Είναι η βάση δεδομένων του Hadoop, που σημαίνει ότι έχει εξ ορισμού τα πλεονεκτήματα του κατανεμημένου συστήματος αρχείων και του μοντέλου MapReduce του Hadoop. Αναφέρεται ως βάση δεδομένων με στήλες, επειδή σε αντίθεση με μια σχεσιακή βάση δεδομένων που αποθηκεύει τα δεδομένα σε γραμμές, η HBase αποθηκεύει τα δεδομένα σε στήλες.

Η HBase επιτρέπει την ανάγνωση-εγγραφή με χαμηλή αναμενόμενη ταχύτητα πάνω από το HDFS(Hadoop Distributed File System). Οι πίνακες στην HBase αποθηκεύονται ως πολυδιάστατος αραιός(wide) χάρτης με γραμμές και στήλες που επιτρέπει την τυχαία πρόσβαση ανάγνωσης-εγγραφής σε πραγματικό χρόνο. Κάθε κελί έχει τη χρονοσφραγίδα και αναγνωρίζεται μοναδικά από τον πίνακα, τη γραμμή, την οικογένεια στηλών και τη χρονοσφραγίδα. Καθώς η HBase διαθέτει API πελάτη Java, οι πίνακες στην HBase μπορούν να χρησιμοποιηθούν ως στόχος εισόδου και εξόδου για εργασία MapReduce. Η HBase χρησιμοποιεί το Zookeeper το οποίο είναι ένα έργο ανοικτού κώδικα του Apache που χρησιμοποιείται ειδικά για τη διαχείριση μερικών αποτυχιών σε βάσεις δεδομένων. Επιπλέον, παρέχει επίσης τη συντήρηση των πληροφοριών διαμόρφωσης και τον κατανεμημένο συγχρονισμό. Το Zookeeper διαθέτει κόμβους οι οποίοι αντιπροσωπεύουν τους διακομιστές περιοχής οι οποίοι χρησιμοποιούνται για την παρακολούθηση των αποτυχιών και των τμημάτων δικτύου.

Όπως αναφέρθηκε προηγουμένως, η HBase είναι η βάση δεδομένων προσανατολισμένη στις στήλες. Το σχήμα πίνακα της HBase ορίζει μόνο τις οικογένειες στηλών με ζεύγη τιμών κλειδιών. Οι πίνακες είναι η συλλογή γραμμών, οι γραμμές είναι η συλλογή οικογενειών στηλών, μια οικογένεια στηλών είναι η συλλογή στηλών και αυτές οι στήλες είναι η συλλογή ζευγών κλειδιού-τιμής. Το κύριο πλεονέκτημα της βάσης δεδομένων με προσανατολισμό στις στήλες έναντι της βάσης δεδομένων με

προσανατολισμό στις γραμμές είναι ότι μπορεί να χρησιμοποιηθεί για τον τεράστιο όγκο δεδομένων που απαιτεί μια online αναλυτική επεξεργασία [21] [22] .

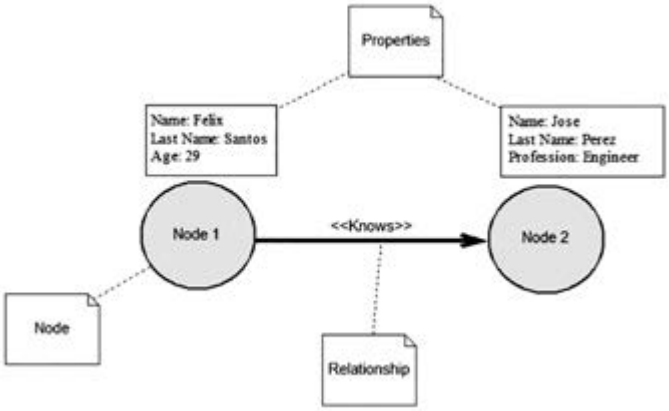
3.3.3.4 Neo4j Βάση Δεδομένων

Μια βάση δεδομένων γράφων ορίζεται ως μια εξειδικευμένη πλατφόρμα ενιαίου σκοπού για τη δημιουργία και τον χειρισμό γράφων. Οι γράφοι περιέχουν κόμβους, ακμές και ιδιότητες, τα οποία χρησιμοποιούνται για την αναπαράσταση και την αποθήκευση δεδομένων με τρόπο που οι σχεσιακές βάσεις δεδομένων δεν είναι εξοπλισμένες για να το κάνουν. Τα γραφήματα και οι βάσεις δεδομένων γραφημάτων παρέχουν μοντέλα γραφημάτων για την αναπαράσταση των σχέσεων στα δεδομένα. Η δύναμη των γράφων έγκειται στην ανάλυση, στις πληροφορίες που παρέχουν και στην ικανότητά τους να συνδέουν διαφορετικές πηγές δεδομένων. Όταν πρόκειται για την ανάλυση γράφων, οι αλγόριθμοι διερευνούν τις διαδρομές και την απόσταση μεταξύ των κορυφών, τη σημασία των κορυφών και την ομαδοποίηση των κορυφών. Για παράδειγμα, για τον προσδιορισμό της σημαντικότητας οι αλγόριθμοι συχνά εξετάζουν τις εισερχόμενες ακμές, τη σημασία των γειτονικών κορυφών και άλλους δείκτες.

Οι αλγόριθμοι γραφημάτων ειδικά σχεδιασμένες για την ανάλυση σχέσεων και συμπεριφορών μεταξύ δεδομένων σε γραφήματα καθιστούν δυνατή την κατανόηση πραγμάτων που είναι δύσκολο να γίνουν αντιληπτά με άλλες μεθόδους. Όταν πρόκειται για την ανάλυση γραφημάτων, οι αλγόριθμοι διερευνούν τις διαδρομές και την απόσταση μεταξύ των κορυφών, τη σημασία των κορυφών και την ομαδοποίηση των κορυφών. Οι αλγόριθμοι συχνά εξετάζουν τις εισερχόμενες ακμές, τη σημασία των γειτονικών κορυφών και άλλους δείκτες για να βοηθήσουν στον προσδιορισμό της σημασίας. Για παράδειγμα, οι αλγόριθμοι γράφων μπορούν να προσδιορίσουν ποιο άτομο ή στοιχείο συνδέεται περισσότερο με άλλα σε κοινωνικά δίκτυα ή επιχειρηματικές διαδικασίες.

Η Neo4j είναι μια βάση δεδομένων NoSQL που ανήκει στην κατηγορία των βάσεων δεδομένων γράφων και ακολουθεί τη μαθηματική θεωρία των δέντρων. Το Neo4j ακολουθεί μια δομή όπου οι κόμβοι αναπαρίστανται ως κορυφές και οι σχέσεις ως ακμές. Η οργάνωση αυτή εφαρμόζει εξελιγμένους αλγορίθμους και μαθηματικούς υπολογισμούς για την αποτελεσματική ανάκτηση δεδομένων. Διατηρεί επίσης τη χρήση μιας δυναμικής δομής, την ανάθεση τιμών μόνο όταν είναι απαραίτητο και έναν πιο ακριβή σχεδιασμό ευθυγραμμισμένο με τους επιχειρηματικούς κανόνες. Επιπλέον, η Neo4j εγγυάται τη συμπεριφορά ACID (Atomicity, Consistency, Isolation and Durability) και αυτό καθιστά τη Neo4j στις λίγες NoSQL βάσεις δεδομένων που υποστηρίζουν συναλλαγές. Το σχήμα 1 απεικονίζει τα συστατικά ενός γράφου, τον τρόπο κατανομής και την αποθήκευση των πληροφοριών. Ένα άλλο κύριο χαρακτηριστικό της Neo4j είναι η στιβαρή αρχιτεκτονική που υλοποιείται βάση της έννοιας της υψηλής διαθεσιμότητας

(HA). Η συστάδα master-slave είναι η πιο σημαντική θεώρηση στο μοντέλο της HA χωρίζει τη Neo4j σε δύο μέρη: την ίδια τη βάση δεδομένων και το συστατικό διαχείρισης της συστάδας. Σε αυτό το συστατικό, υπάρχει ένας μηχανισμός που παρέχει συνεχή συγχρονισμό όλων των περιπτώσεων και εξασφαλίζει ότι η εκλογή του master cluster είναι αυτόματη. Αυτή η διαδικασία επιτρέπει στην κύρια συστάδα να χειρίζεται όλες τις λειτουργίες εγγραφής και παρέχει έναν κεντρικό έλεγχο για την επίτευξη επεκτασιμότητας. Επιπλέον, όλα τα γραφήματα αντιγράφονται σε κάθε περίπτωση κάθε συστάδας και αυτό το χαρακτηριστικό παρέχει ασφάλεια για τη συνέχιση των εργασιών και της απόκρισης παρά τις πιθανές αποτυχίες σε ορισμένες συστάδες [23] .



Εικόνα 15: Απεικόνιση των συστατικών ενός γράφου.
(Πηγή: https://www.researchgate.net/figure/Graph-illustration_fig1_307180380)

3.3.3.5 Datastore/Firestore Cloud Βάση Δεδομένων

Μια βάση δεδομένων νέφους είναι μια υπηρεσία βάσεων δεδομένων που έχει δημιουργηθεί και είναι προσβάσιμη μέσω μιας πλατφόρμας υπολογιστικού νέφους. Εξυπηρετεί πολλές από τις ίδιες λειτουργίες με μια παραδοσιακή βάση δεδομένων με την πρόσθετη ευελιξία του υπολογιστικού νέφους. Οι χρήστες εγκαθιστούν λογισμικό σε μια υποδομή νέφους για την υλοποίηση της βάσης δεδομένων. Η διαχείριση δεδομένων δέσμευσης και εφαρμογών για τεράστια δίκτυα κινητών χρηστών ή απομακρυσμένων συσκευών μπορεί να αποτελέσει πρόκληση κλιμάκωσης και διαθεσιμότητας. Το πρόβλημα είναι ότι οι περισσότερες βάσεις δεδομένων απαιτούν οι ενημερώσεις να πραγματοποιούνται σε μια κεντρική "κύρια" βάση δεδομένων. Αυτό μπορεί να οδηγήσει σε συμφόρηση επιδόσεων και να εμποδίσει την εκτέλεση εφαρμογών εάν η σύνδεση με την κύρια βάση δεδομένων δεν είναι διαθέσιμη. Μια βάση δεδομένων cloud δίνει τη δυνατότητα στους οργανισμούς να προωθήσουν την πρόσβαση στη βάση δεδομένων στην πιο απομακρυσμένη άκρη του δικτύου για κινητές

συσκευές, απομακρυσμένες εγκαταστάσεις, αισθητήρες και αγαθά με δυνατότητα σύνδεσης στο διαδίκτυο. Αυτό συμβάλλει στη βελτίωση της επεκτασιμότητας και επιτρέπει στις εφαρμογές να συνεχίσουν να εκτελούνται ενώ βρίσκονται εκτός σύνδεσης. Οι βάσεις δεδομένων νέφους συλλέγουν, παραδίδουν, αναπαράγουν και προωθούν έως το τέλος όλα τα δεδομένα ενός οργανισμού χρησιμοποιώντας την έννοια του υβριδικού νέφους. Οι χρήστες δεν χρειάζεται πλέον να αναπτύσσουν το εξαρτώμενο ενδιάμεσο λογισμικό για να παραδίδουν τα αιτήματα βάσεων δεδομένων οπουδήποτε στον κόσμο. Μπορούν να συνδέουν τις εφαρμογές απευθείας με τη βάση δεδομένων τους. Οι υβριδικές βάσεις δεδομένων δημιουργούν ένα κατανεμημένο υβριδικό νέφος δεδομένων για αυξημένη απόδοση, εμβέλεια, κινητικότητα σε χρόνο διαθεσιμότητας και εξοικονόμηση κόστους [24] .

Το Datastore πλέον γνωστό ως Firestore στην πλατφόρμα Google Cloud Platform είναι μια πλήρως διαχειρίσιμη υπηρεσία βάσεων δεδομένων NoSQL που παρέχεται από την πλατφόρμα Google Cloud Platform (GCP). Το Firestore έχει σχεδιαστεί για να ενσωματώνεται με άλλες υπηρεσίες του GCP και να προσφέρει χαρακτηριστικά όπως αυτόματη κλιμάκωση, υψηλή διαθεσιμότητα και παγκόσμια αντιγραφή, ώστε να διασφαλίζει αξιόπιστη και αποδοτική αποθήκευση και συγχρονισμό δεδομένων για εφαρμογές που βασίζονται στο cloud.

4 Λογισμικό και Μεγάλα Σύνολα Δεδομένων

Οι βάσεις δεδομένων όπως αναφέραμε δεν αποτελούνται απλώς από έναν χώρο αποθήκευσης δεδομένων όπως συχνά προσδιορίζεται. Χωρίς την ανάπτυξη του απαραίτητου ειδικού λογισμικού δεν θα ήταν εφικτό να γίνει η χρήση μια βάσης \ ούτε και η εκμετάλλευση των δεδομένων της. Το λογισμικό κατέχει έναν καθοριστικό ρόλο στην διαχείριση των Μεγάλων Δεδομένων, παρέχοντας τα εργαλεία και τις πλατφόρμες που είναι απαραίτητες για τη συλλογή, αποθήκευση, επεξεργασία, ανάλυση και οπτικοποίηση τεράστιων ποσοτήτων δεδομένων. Το λογισμικό μεγάλων δεδομένων διευκολύνει τη συλλογή και την εισαγωγή δεδομένων από διάφορες πηγές, όπως αισθητήρες, συσκευές IoT, πλατφόρμες μέσων κοινωνικής δικτύωσης, διαδικτυακές εφαρμογές και εταιρικά συστήματα. Εργαλεία λογισμικού όπως ο Apache Kafka, το Flume και το NiFi χρησιμοποιούνται για τη ροή και την εισροή δεδομένων σε πραγματικό χρόνο, ενώ πλαίσια επεξεργασίας δέσμης όπως το Apache Hadoop επιτρέπουν την εισροή μεγάλου όγκου δεδομένων σε μαζική μορφή. Οι λύσεις που παρέχουν τα λογισμικά σε κλιμακούμενα και κατανεμημένα συστήματα αποθήκευσης δεδομένων τα καθιστούν ικανά να χειρίζονται αποτελεσματικά τεράστια σύνολα δεδομένων. Το λογισμικό μεγάλων δεδομένων επιτρέπει την επεξεργασία και ανάλυση μεγάλων συνόλων δεδομένων για την εξαγωγή πολύτιμων πληροφοριών και αξιοποιήσιμων πληροφοριών. Πλαίσια κατανεμημένου υπολογισμού όπως το Apache Spark και το Apache Lucene προσφέρουν δυνατότητες παράλληλης επεξεργασίας για επεξεργασία δέσμης και ροής, ενώ το Apache Hadoop παρέχει επεξεργασία δέσμης με βάση το MapReduce για ανάλυση δεδομένων μεγάλης κλίμακας. Οι βιβλιοθήκες και τα πλαίσια μηχανικής μάθησης, όπως το TensorFlow, το PyTorch και το scikit-learn, χρησιμοποιούνται για προγνωστική ανάλυση, ανίχνευση ανωμαλιών και αναγνώριση προτύπων σε εφαρμογές μεγάλων δεδομένων.

Εργαλεία λογισμικού παρέχουν δυνατότητες οπτικοποίησης και εξερεύνησης για την κατανόηση δυσνόητων δεδομένων και την αποτελεσματική επικοινωνία των αποφάσεων. Οι πλατφόρμες επιχειρησιακής ευφυΐας (BI), όπως το Tableau, το Power BI και το Qlik Sense, προσφέρουν αναφορές και οπτικοποιήσεις για την ανάλυση και την παρουσίαση δεδομένων. Εργαλεία εξερεύνησης δεδομένων, όπως το Apache Zeppelin και τα Jupyter Notebooks, επιτρέπουν στους επιστήμονες και αναλυτές δεδομένων να εξερευνούν και να απεικονίζουν δια δραστικά δεδομένα χρησιμοποιώντας γλώσσες προγραμματισμού όπως η Python, η R και η Scala. Το λογισμικό μεγάλων δεδομένων συμβάλλει στη διασφάλιση της διακυβέρνησης δεδομένων, της συμμόρφωσης και της ασφάλειας, εφαρμόζοντας ελέγχους πρόσβασης, κρυπτογράφηση και μηχανισμούς ελέγχου. Εργαλεία διακυβέρνησης δεδομένων όπως το Apache Ranger και το Apache

Atlas παρέχουν έλεγχο πρόσβασης βάσει πολιτικής και διαχείριση μεταδεδομένων για περιβάλλοντα μεγάλων δεδομένων, ενώ τεχνολογίες κρυπτογράφησης όπως το Apache Sentry και η υπηρεσία διαχείρισης κλειδιών (KMS) της AWS διασφαλίζουν ευαίσθητα δεδομένα. Παρακάτω θα αναφερθούμε εκτενώς σε κάποιες βασικές πλατφόρμες λογισμικού για Μεγάλα Σύνολα Δεδομένων [25] [26] .

4.1 MapReduce

Το MapReduce είναι ένα μοντέλο προγραμματισμού και μια εξειδικευμένη υλοποίηση για την επεξεργασία και τη δημιουργία μεγάλων συνόλων δεδομένων. Οι χρήστες καθορίζουν μια συνάρτηση map που επεξεργάζεται ένα ζεύγος κλειδιού/τιμής για να δημιουργήσει ένα σύνολο ενδιάμεσων ζευγών κλειδιού/τιμής και μια συνάρτηση reduce που συγχωνεύει όλες τις ενδιάμεσες τιμές που σχετίζονται με το ίδιο ενδιάμεσο κλειδί. Προγράμματα γραμμένα σε αυτό το λειτουργικό στυλ παραλληλοποιούνται αυτόματα και εκτελούνται σε μια μεγάλη συστάδα μηχανών βασικών προϊόντων. Το σύστημα χρόνου εκτέλεσης φροντίζει για τις λεπτομέρειες της κατάτμησης των δεδομένων εισόδου, του προγραμματισμού της εκτέλεσης του προγράμματος σε ένα σύνολο μηχανών, του χειρισμού των αποτυχιών των μηχανών και της διαχείρισης της απαιτούμενης επικοινωνίας μεταξύ των μηχανών. Αυτό επιτρέπει στους προγραμματιστές που δεν έχουν εμπειρία σε παράλληλα και κατανεμημένα συστήματα να αξιοποιούν εύκολα τους πόρους ενός μεγάλου κατανεμημένου συστήματος. Η υλοποίησή του MapReduce εκτελείται σε ένα μεγάλο σύμπλεγμα μηχανών commodity και είναι εξαιρετικά επεκτάσιμη. Ένας τυπικός υπολογισμός MapReduce επεξεργάζεται πολλά terabytes δεδομένων σε χιλιάδες μηχανές. Η πλειονότητα των προγραμματιστών βρίσκει το σύστημα εύκολο στη χρήση καθώς εκατοντάδες προγράμματα MapReduce έχουν υλοποιηθεί και πάνω από χίλιες εργασίες MapReduce αναρτιούνται στη Google καθημερινά [26] [27] .

4.1.1 Μοντέλο Προγραμματισμού

Ο αλγόριθμος λαμβάνει ένα σύνολο ζευγών κλειδιών/τιμών εισόδου και παράγει ένα σύνολο ζευγών κλειδιών/τιμών εξόδου. Ο χρήστης της βιβλιοθήκης MapReduce εκφράζει τους υπολογισμούς ως δύο συναρτήσεις: Map και Reduce. Η Map, γραμμένη από τον χρήστη, λαμβάνει ένα ζεύγος εισόδου και παράγει ένα σύνολο ενδιάμεσων ζευγών κλειδιού/τιμής. Η βιβλιοθήκη MapReduce ομαδοποιεί όλες τις ενδιάμεσες τιμές που σχετίζονται με το ίδιο ενδιάμεσο κλειδί “I” και τις περνάει στη συνάρτηση Reduce. Η συνάρτηση Reduce, επίσης γραμμένη από τον χρήστη, δέχεται ένα ενδιάμεσο κλειδί “I” και ένα σύνολο τιμών για αυτό το κλειδί. Ύστερα συγχωνεύει

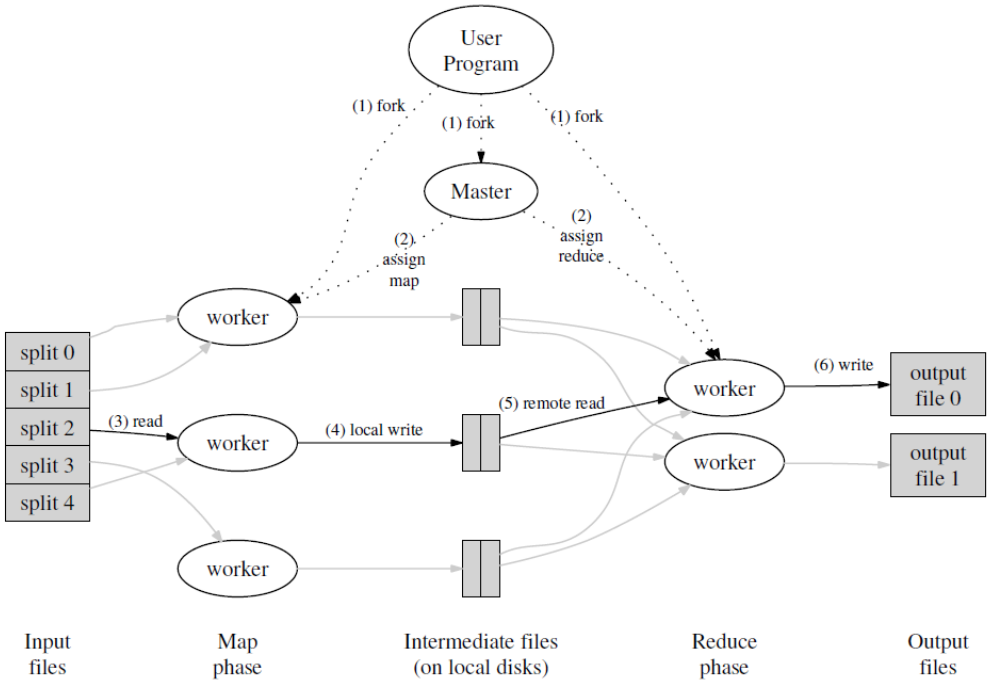
αυτές τις τιμές για να σχηματίσει ένα ενδεχομένως μικρότερο σύνολο τιμών. Συνήθως παράγεται μόνο μηδέν ή ένα ως τιμή εξόδου ανά κλήση της Reduce. Οι ενδιάμεσες τιμές παρέχονται στη συνάρτηση reduce του χρήστη μέσω ενός επαναλήπτη. Αυτό μας επιτρέπει να χειριζόμαστε λίστες τιμών που είναι πολύ μεγάλες για να χωρέσουν στη μνήμη[26] [27] .

4.1.2 Αρχιτεκτονική

Τα τμήματα εισόδου μπορούν να υποβληθούν σε παράλληλη επεξεργασία από διαφορετικές μηχανές. Οι κλήσεις Reduce κατανέμονται με διαμερισμό του ενδιάμεσου χώρου κλειδιών σε R κομμάτια χρησιμοποιώντας μια συνάρτηση διαμερισμού (π.χ. hash(key) mod R). Ο αριθμός των διαμερισμάτων (R) και η συνάρτηση διαμέρισης καθορίζονται από τον χρήστη. Το Σχήμα 1 δείχνει τη συνολική ροή μιας λειτουργίας MapReduce στην υλοποίησή μας. Όταν το πρόγραμμα του χρήστη καλεί τη λειτουργία MapReduce, λαμβάνει χώρα η ακόλουθη ακολουθία ενεργειών:

i) Η βιβλιοθήκη MapReduce στο πρόγραμμα του χρήστη πρώτα χωρίζει τα αρχεία εισόδου σε M μικρά κομμάτια τα οποία καθορίζονται από τον χρήστη, τυπικά των 16 megabytes έως 64 megabytes (MB) ανά κομμάτι.
ii) Δημιουργεί πολλά αντίγραφα του προγράμματος σε μια συστάδα μηχανών. Ένα από αυτά τα αντίγραφα είναι ο master, ενώ τα υπόλοιπα είναι οι εργάτες που λαμβάνουν εργασία από τον master
iii) Υπάρχουν M εργασίες map και R reduce εργασίες προς ανάθεση. Κάθε εργαζόμενος επιλέγεται για μια εργασία χαρτογράφησης ή μείωσης. Κάθε εργαζόμενος χαρτογράφησης διαβάζει τα δεδομένα εισόδου και μεταβιβάζει τα ενδιάμεσα αποτελέσματα στη μνήμη. Αναλύει ζεύγη κλειδιών/τιμών από τα δεδομένα εισόδου και μεταβιβάζει κάθε ζεύγος στη συνάρτηση Map που ορίζει ο χρήστης.
iv) Τα ενδιάμεσα ζεύγη κλειδιών/τιμών που παράγονται από τη συνάρτηση Map αποθηκεύονται στη μνήμη και διαιρούνται σε περιοχές R οι θέσεις των οποίων διαβιβάζονται στον master και εκείνος τις διαβιβάζει στους εργάτες μείωσης.
v) Οι εργάτες μείωσης διαβάζουν τα δεδομένα από τους τοπικούς δίσκους των εργατών χαρτογράφησης με χρήση απομακρυσμένων κλήσεων διαδικασιών. Όταν ένας reduce worker έχει διαβάσει όλα τα ενδιάμεσα δεδομένα, τα ταξινομεί με βάση τα ενδιάμεσα κλειδιά έτσι ώστε όλες οι εμφανίσεις του ίδιου κλειδιού να ομαδοποιούνται μαζί.
vi) Οι εργάτες μείωσης μεταβιβάζουν τα δεδομένα στη συνάρτηση Reduce του χρήστη για κάθε μοναδικό ενδιάμεσο κλειδί που συναντούν. Η έξοδος της συνάρτησης Reduce προστίθεται σε ένα τελικό αρχείο εξόδου για αυτό το τμήμα της μείωσης.

vii) Όταν όλες οι εργασίες χαρτογράφησης και μείωσης έχουν ολοκληρωθεί, ο master ξυπνάει το πρόγραμμα χρήστη.



Εικόνα 16 : Αρχιτεκτονική MapReduce [27]

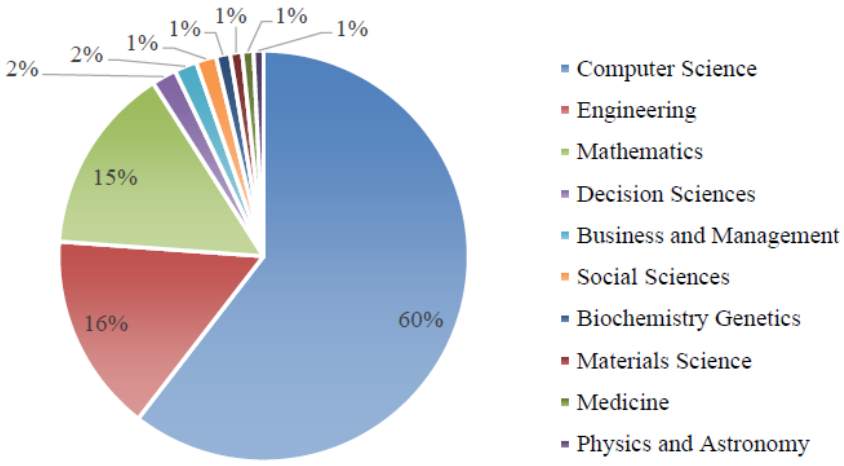
Μετά την επιτυχή ολοκλήρωση, η έξοδος της εκτέλεσης mapreduce είναι διαθέσιμη στα αρχεία εξόδου R. Αντιστοιχεί ένα ανά reduce task, με ονόματα αρχείων όπως καθορίζονται από τον χρήστη. Συνήθως, οι χρήστες δεν χρειάζεται να συνδυάσουν αυτά τα αρχεία εξόδου R σε ένα αρχείο - συχνά παίρνουν αυτά τα αρχεία ως είσοδο στην άλλη κλήση MapReduce, ή τα χρησιμοποιούν από άλλη κατανεμημένη εφαρμογή που είναι σε θέση να χειριστεί είσοδο που είναι κατανεμημένη σε πολλαπλά αρχεία [26] [27] .

4.1.3 Χρησιμότητα

Πολλές διαφορετικές υλοποιήσεις της διεπαφής MapReduce είναι δυνατές. Η σωστή επιλογή εξαρτάται από το περιβάλλον. Για παράδειγμα, μια υλοποίηση μπορεί να είναι κατάλληλη για μια μικρή μηχανή κοινής μνήμης, μια άλλη για έναν μεγάλο πολυεπεξεργαστή NUMA, και μια άλλη για μια ακόμη μεγαλύτερη συλλογή από δικτυωμένες μηχανές. Μία χαρακτηριστική υλοποίηση στοχεύει στο υπολογιστικό περιβάλλον που χρησιμοποιείται ευρέως στην Google.

Τα τελευταία χρόνια, το MapReduce έχει αναδειχθεί σε βασικό αλγόριθμο στο πεδίο της επεξεργασίας μεγάλων δεδομένων, βρίσκοντας ευρεία υιοθέτηση σε διάφορους κλάδους και εφαρμογές. Η ευελιξία, η επεκτασιμότητα και η αποδοτικότητά της την έχουν καταστήσει ιδανική για τον χειρισμό των πολύπλοκων εργασιών επεξεργασίας δεδομένων μεγάλης κλίμακας [26] [27] .

Παρακάτω φαίνονται οι κλάδοι στους οποίους έχει χρησιμοποιηθεί το Map Reduce τα τελευταία χρόνια σύμφωνα με το ποσοστό δημοσιεύσεων στις οποίες αναφέρεται.



Εικόνα 17 : Ερευνητικά πεδία με βάση το ποσοστό των δημοσιεύσεων [26]

4.2 Hadoop

Το Hadoop κυκλοφόρησε το 2006 και αποτελεί ένα έργο ανοικτού κώδικα του Apache Software Foundation (ASF) γραμμένο σε Java, το οποίο παρέχει οικονομικά αποδοτική, κλιμακούμενη υποδομή για την κατανεμημένη επεξεργασία μεγάλων συνόλων δεδομένων σε συστάδες. Η Yahoo! υιοθέτησε το Apache Hadoop το 2006 για να αντικαταστήσει την υποδομή που οδηγούσε την εφαρμογή WebMap, την τεχνολογία που δημιουργεί ένα γράφημα του γνωστού ιστού για να τροφοδοτήσει τη μηχανή αναζήτησης. Εκείνη την εποχή, ο γράφος του ιστού περιείχε περισσότερους από 100 δισεκατομμύρια κόμβους και 1 τρισεκατομμύριο ακμές. Η προηγούμενη υποδομή, που

ονομαζόταν "Dreadnaught", είχε φτάσει στα όρια της κλιμάκωσής της σε 800 μηχανές και απαιτούνταν μια σημαντική αλλαγή στην αρχιτεκτονική της για να συμβαδίζει με τους ρυθμούς του ιστού. Το Hadoop "εμπνεύστηκε" από το Google File System GFS και το περιβάλλον κατανεμημένου υπολογισμού MapReduce της Google. Αρχικά, ξεκίνησε ως ένα κατανεμημένο έργο μηχανής αναζήτησης Nutch και ονομάστηκε από τον προγραμματιστή Doug Cutting από το παιχνίδι-ελέφαντα του γιου του. Χρησιμοποιήθηκε με επιτυχία για την επεξεργασία προβλημάτων υψηλής κατανομής σε μεγάλο αριθμό συνόλων δεδομένων με τη χρήση κοινά διαθέσιμων διακομιστών σε μια πολύ μεγάλη συστάδα, όπου κάθε διακομιστής διαθέτει ένα σύνολο φθηνών εσωτερικών μονάδων δίσκων. Το τρέχον έργο Hadoop αποτελείται από τρεις κύριες ενότητες, δηλαδή το κατανεμημένο σύστημα αρχείων που ονομάζεται Hadoop Distributed File System, τη μηχανή MapReduce και τον Yet Another Resource Negotiator (YARN) [27][28] .

1) Κατανεμημένο σύστημα αρχείων Apache Hadoop

Το κατανεμημένο σύστημα αρχείων Apache Hadoop (HDFS) είναι ένα κατανεμημένο σύστημα αρχείων που διαχειρίζεται μεγάλα σύνολα δεδομένων και εκτελείται σε βασικό υλικό. Χρησιμοποιείται για την κλιμάκωση μιας συστάδας Apache Hadoop σε εκατοντάδες (ή και χιλιάδες) κόμβους. Ένα σημαντικό χαρακτηριστικό του Hadoop είναι η κατανομή των δεδομένων και των υπολογισμών σε πολλούς (χιλιάδες) κεντρικούς υπολογιστές και η παράλληλη εκτέλεση υπολογισμών εφαρμογών κοντά στα δεδομένα τους. Μία Hadoop συστάδα κλιμακώνει την υπολογιστική ικανότητα, την αποθηκευτική ικανότητα και το IO εύρος ζώνης προσθέτοντας απλά εξυπηρετητές κοινής χρήσης. Οι συστάδες Hadoop στη Yahoo! καλύπτουν 25 000 διακομιστές και αποθηκεύουν 25 petabytes δεδομένων εφαρμογών, με τη μεγαλύτερη συστάδα να αποτελείται από 3500 διακομιστές.

Οι στόχοι του HDFS είναι:

i) Γρήγορη ανάκαμψη από βλάβες υλικού. Επειδή μια περίπτωση HDFS μπορεί να αποτελείται από χιλιάδες διακομιστές, η αποτυχία τουλάχιστον ενός διακομιστή είναι αναπόφευκτη. Το HDFS έχει κατασκευαστεί για να ανιχνεύει βλάβες και να ανακάμπτει αυτόματα και γρήγορα.
ii) Πρόσβαση σε δεδομένα ροής. Το HDFS προορίζεται περισσότερο για επεξεργασία παρτίδων παρά για δια δραστική χρήση, οπότε η έμφαση στο σχεδιασμό του δίνεται σε υψηλούς ρυθμούς απόδοσης δεδομένων, οι οποίοι εξυπηρετούν την πρόσβαση σε σύνολα δεδομένων με ροή.
iii) Φιλοξενία μεγάλων συνόλων δεδομένων. Το HDFS φιλοξενεί εφαρμογές που έχουν σύνολα δεδομένων τυπικά μεγέθους gigabytes έως terabytes.

iv) Φορητότητα. Για τη διευκόλυνση της υιοθέτησης, το HDFS έχει σχεδιαστεί ώστε να είναι φορητό σε πολλαπλές πλατφόρμες υλικού και συμβατό με διάφορα υποκείμενα λειτουργικά συστήματα.

Για να γίνει πιο κατανοητή η λειτουργία του HDFS θα παραθέσουμε το ακόλουθο παράδειγμα. Σκεφτείτε ένα αρχείο που περιλαμβάνει τους τηλεφωνικούς αριθμούς για όλους τους κατοίκους των Ηνωμένων Πολιτειών. Οι αριθμοί για τα άτομα με επώνυμο που αρχίζει με Α μπορεί να αποθηκεύονται στον διακομιστή 1, ο Β στον διακομιστή 2 κ.ο.κ. Με το Hadoop, κομμάτια αυτού του τηλεφωνικού καταλόγου θα αποθηκευτούν σε όλη τη συστάδα και για να ανακατασκευαστεί ολόκληρος ο τηλεφωνικός κατάλογος, το πρόγραμμά θα χρειαστεί τα μπλοκ από κάθε διακομιστή της συστάδας. Για να διασφαλιστεί η διαθεσιμότητα εάν και όταν ένας διακομιστής αποτύχει, το HDFS αναπαράγει αυτά τα μικρότερα κομμάτια σε δύο επιπλέον διακομιστές από προεπιλογή. (Ο πλεονασμός μπορεί να αυξηθεί ή να μειωθεί ανά αρχείο ή για ολόκληρο το περιβάλλον- για παράδειγμα, μια συστάδα ανάπτυξης Hadoop συνήθως δεν χρειάζεται πλεονασμό δεδομένων). Αυτός ο πλεονασμός προσφέρει πολλαπλά οφέλη, με πιο προφανές την υψηλότερη διαθεσιμότητα. Ο πλεονασμός επιτρέπει επίσης στη συστάδα Hadoop να χωρίζει τις εργασίες σε μικρότερα κομμάτια και να εκτελεί αυτές τις εργασίες σε όλους τους διακομιστές της συστάδας για καλύτερη κλιμάκωση. Τέλος, αποκτάται το πλεονέκτημα της τοπικότητας των δεδομένων, το οποίο είναι κρίσιμο όταν κανείς εργάζεται με μεγάλα σύνολα δεδομένων [27][28] .

2) Σχέση MapReduce και Hadoop

Το Hadoop και το MapReduce είναι στενά συνδεδεμένα στοιχεία στο σύστημα διαχείρισης και επεξεργασίας μεγάλων δεδομένων. Αναφερθήκαμε εκτενώς στο MapReduce και τη χρήση του ωστόσο για να γίνει κατανοητή η διαφορά με το Hadoop ας αναφέρουμε μια μικρή ανάλυση της σχέσης τους.

Το Hadoop είναι ένα πλαίσιο ανοικτού κώδικα που έχει σχεδιαστεί για την κατανεμημένη αποθήκευση και επεξεργασία μεγάλων συνόλων δεδομένων σε συστάδες υλικού διαφόρων χρήσεων. Το MapReduce είναι ένα μοντέλο προγραμματισμού και μια μηχανή επεξεργασίας ειδικά σχεδιασμένη για την παράλληλη επεξεργασία και ανάλυση μεγάλων συνόλων δεδομένων σε περιβάλλοντα κατανεμημένων υπολογιστών [26] .

Ενσωμάτωση στο πλαίσιο του Hadoop: Το MapReduce αποτελεί βασικό συστατικό του πλαισίου Hadoop και χρησιμεύει ως η προεπιλεγμένη μηχανή επεξεργασίας του για την επεξεργασία δεδομένων δέσμης. Το Hadoop αξιοποιεί το MapReduce για την εκτέλεση εργασιών επεξεργασίας δεδομένων σε κατανεμημένες συστάδες υπολογιστών, χρησιμοποιώντας το HDFS για κατανεμημένη αποθήκευση και το YARN για διαχείριση πόρων και χρονοπρογραμματισμό. Οι προγραμματιστές γράφουν

προγράμματα MapReduce για να εκφράσουν τη λογική της επεξεργασίας δεδομένων τους, η οποία στη συνέχεια εκτελείται από το πλαίσιο Hadoop σε όλους τους κόμβους της συστάδας. Συνοπτικά, το Hadoop και το MapReduce είναι αλληλένδετα στοιχεία στο οικοσύστημα του Hadoop, με το Hadoop να παρέχει την υποδομή και τα εργαλεία για κατανεμημένη αποθήκευση και επεξεργασία και το MapReduce να χρησιμεύει ως μοντέλο προγραμματισμού και μηχανή επεξεργασίας για παράλληλες εργασίες επεξεργασίας δεδομένων στο πλαίσιο του Hadoop. Μαζί, δίνουν τη δυνατότητα στους οργανισμούς να αποθηκεύουν, να επεξεργάζονται και να αναλύουν αποτελεσματικά μεγάλους όγκους δεδομένων σε κατανεμημένες συστάδες υπολογιστών, καθιστώντας την επεξεργασία μεγάλων δεδομένων προσιτή για τους προγραμματιστές και τις επιχειρήσεις[27] [29] .

3) Yet Another Resource Negotiator YARN

Οι εξελίξεις του Hadoop είχαν ένα πρόβλημα το 2011, το επίκεντρο του προβλήματος επισημάνθηκε από τον Eric Baldeschwieler, τον τότε διευθύνοντα σύμβουλο της Hortonworks, όταν το MapReduce παρουσίαζε δύο περιοχές αδυναμίας, η μία ήταν η επεκτασιμότητα και η δεύτερη η χρήση των πόρων. Ο στόχος του νέου πλαισίου το οποίο είχε τον τίτλο Yet Another Resource Negotiator (YARN) ένα πλαίσιο διαχείρισης πόρων και χρονοπρογραμματισμού εργασιών που εισήχθη στο Hadoop 2.x. Ένα τέτοιο λειτουργικό σύστημα στο Hadoop διασφαλίζει την επεκτασιμότητα, την απόδοση και τη χρήση των πόρων, γεγονός που είχε ως αποτέλεσμα να εφαρμοστεί για πρώτη φορά μια αρχιτεκτονική για το Διαδίκτυο των πραγμάτων (IoT). Η πιο σημαντική έννοια του YARN είναι η δυνατότητα υλοποίησης ενός παραδείγματος επεξεργασίας δεδομένων που ονομάζεται χαλαρή αξιολόγηση (lazy evaluation) και εξαιρετικά καθυστερημένη δέσμευση (extremely late binding) χαρακτηριστικό το οποίο αποτελεί το μέλλον της επεξεργασίας και διαχείρισης δεδομένων. Οι απαιτήσεις του YARN λοιπόν προέκυψαν από πρακτικές ανάγκες. Εκτός από τις σωληνώσεις εξαιρετικά μεγάλης κλίμακας (extremely large-scale pipelines) του Yahoo! Search, η κοινότητα του Apache Hadoop κλιμάκωνε την πλατφόρμα για όλο και μεγαλύτερες εργασίες MapReduce, οι επιστήμονες που βελτιστοποιούσαν την ανάλυση διαφημίσεων, το φιλτράρισμα ανεπιθύμητων μηνυμάτων και τη βελτιστοποίηση περιεχομένου οδήγησαν σε πολλές από τις πρώιμες απαιτήσεις του. Η αρχιτεκτονική του YARN ανταποκρίνεται σε πολλές μακροχρόνιες απαιτήσεις της μηχανικής, με βάση την εμπειρία από την εξέλιξη της πλατφόρμας MapReduce[30] .

Συνοψίζουμε εν συντομία ορισμένα έργα που είτε είναι εγγενή στο YARN είτε έχουν μεταφερθεί στην πλατφόρμα για να καταδείξουμε τη γενικότητα της αρχιτεκτονικής του και τη συμβολή του στη διαχείριση μεγάλων δεδομένων. Το Apache Hadoop MapReduce λειτουργεί ήδη πάνω στο YARN με σχεδόν το ίδιο σύνολο χαρακτηριστικών. Έχει δοκιμαστεί σε κλίμακα, ενώ τα υπόλοιπα έργα του οικοσυστήματος όπως τα Pig, Hive, Oozie κ.λπ. έχουν τροποποιηθεί ώστε να λειτουργούν στο MR πάνω στο YARN,

μαζί με τυποποιημένα benchmarks που αποδίδουν στο ίδιο ή και καλύτερο επίπεδο σε σύγκριση με το κλασικό Hadoop. Η κοινότητα του MapReduce έχει φροντίσει ώστε οι εφαρμογές που έχουν γραφτεί στην έκδοση 1.x να μπορούν να τρέξουν πάνω στο YARN με πλήρη δυαδική συμβατότητα (mapred APIs) ή απλά με επαναμεταγλώττιση (συμβατότητα με τον πηγαίο κώδικα για τα mapreduce APIs). Το Apache Tez είναι ένα έργο Apache που έχει ως στόχο να παρέχει ένα γενικό πλαίσιο εκτέλεσης κατευθυνόμενων ακυκλικών γραφημάτων (DAG). Ένας από τους στόχους του είναι να παρέχει μια συλλογή δομικών στοιχείων που μπορούν να συντίθενται σε ένα αυθαίρετο DAG συμπεριλαμβανομένου ενός απλού DAG 2 σταδίων για να διατηρηθεί η συμβατότητα με το MapReduce. Το Tez παρέχει στα συστήματα εκτέλεσης ερωτημάτων, όπως το Hive και το Pig, ένα πιο φυσικό μοντέλο για το σχέδιο εκτέλεσής τους, σε αντίθεση με την αναγκαστική μετατροπή αυτών των σχεδίων σε MapReduce. Το Spark είναι ένα ερευνητικό πρόγραμμα ανοικτού κώδικα από το UC Berkeley το οποίο στοχεύει στην εκμάθηση μηχανών και σε δια δραστικούς φόρτους εργασίας. Το Spark έχει πρόσφατα μεταφερθεί και αυτό στο YARN. Ενώ αξίζει να σημειωθούν τίτλοι όπως Dryad, Giraph και Storm που επίσης έχουν υιοθετήσει το YARN [27][30] .

4.3 ElasticSearch

Η συλλογή μεγάλων δεδομένων είναι ένα πρώτο ζήτημα αλλά η ανάλυση μεγάλων δεδομένων είναι ένα δεύτερο. Απαιτείται γνώση των μηχανών αναζήτησης επιχειρήσεων για να καταστεί το περιεχόμενο από διαφορετικές πηγές, όπως η βάσεις δεδομένων της επιχείρησης, τα μέσα κοινωνικής δικτύωσης, τα δεδομένα αισθητήρων κ.λπ., αναζητήσιμα από ένα καθορισμένο κοινό.

Το Elasticsearch είναι μια κατανεμημένη μηχανή αναζήτησης και ανάλυσης, πλήρους κειμένου. Είναι μια μηχανή αναζήτησης ανοικτού κώδικα η οποία είναι ικανή να χειρίζεται διάφορους τύπους δεδομένων, όπως αλφαβητικά, αριθμητικά, δομημένα καθώς και αδόμητα δεδομένα. Πρόκειται για ένα κατανεμημένο σύστημα αποθήκευσης εγγράφων που βασίζεται στο Lucene. Λόγω της κατανεμημένης φύσης του, λέγεται ότι είναι μηχανή αναζήτησης υψηλής διαθεσιμότητας και μπορεί επίσης να κλιμακωθεί εύκολα. Παρέχει REST API μια διεπαφή προγραμματισμού εφαρμογών (API) που συμμορφώνεται με τις αρχές σχεδιασμού του αρχιτεκτονικού στυλ μεταφοράς κατάστασης αναπαράστασης (REST), με βάση το JSON. Μας διευκολύνει να εκτελέσουμε ορισμένες πολύ πολύπλοκες συσσωρεύσεις δεδομένωνν, μερικές από τις οποίες δεν μπορούν να υποστηρίζονται από το Lucene που προαναφέραμε. Το Elasticsearch είναι κάτι περισσότερο από μια απλή μηχανή αναζήτησης, καθώς μπορεί να χρησιμοποιηθεί για μια ποικιλία άλλων εφαρμογών, όπως η ανάλυση, αποθήκευση εγγράφων και η αυτόματη υποβολή προτάσεων. Βάσεις δεδομένων όπως αυτές της MySQL που αποθηκεύουν δεδομένα μας βοηθούν να κάνουμε ερωτήματα πάνω σε αυτά. Σε αντίθεση με τη MySQL, η Elasticsearch είναι μια αποθήκη εγγράφων JSON που

χρησιμοποιεί μια μέθοδο ευρετηρίασης, που δημιουργεί ανεστραμμένα ευρετήρια για το κείμενο εισόδου, γεγονός που την καθιστά πολύ γρήγορη και σχεδόν σε πραγματικό χρόνο, μηχανή αναζήτησης παράγοντας αποτελέσματα μιας αναζήτησης μέσα σε λίγα χιλιοστά του δευτερολέπτου. Επίσης, η Elasticsearch αποτελεί μέρος της "ομάδας" ELK stack ή Elastic Stack η οποία αποτελείται επίσης από την Kibana και το Logstack. Η Kibana είναι λογισμικό που βοηθά στην οπτικοποίηση των δεδομένων και το Logstash στην συλλογή και την αποστολή δεδομένων στον επιθυμητό προορισμό. Ο συνδυασμός των τριών οδηγεί σε μια ιδανικά εξοπλισμένη ομάδα για την ανάλυση και την επεξεργασία μεγάλων δεδομένων μιας επιχείρησης[31] .

5 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Με βάση τις πρόσφατες τάσεις και τις εξελίξεις στον χειρισμό μεγάλων δεδομένων, αρκετές επεκτατικές παρεμβάσεις είναι πιθανό να διαμορφώσουν το μέλλον του τομέα για την καλύτερη αποθήκευση επεξεργασία και εκμετάλλευση τους. Ακολουθούν ορισμένοι πιθανοί τομείς ανάπτυξης στον τομέα των Μεγάλων Δεδομένων.

Τεχνητή νοημοσύνη και Μεγάλα Σύνολα Δεδομένων: Οι συνεχείς ραγδαίες εξελίξεις στη μηχανική μάθηση, την τεχνητή νοημοσύνη (AI) και τη βαθιά μάθηση αναμένεται να οδηγήσουν στην ανάπτυξη πιο εξελιγμένων τεχνικών ανάλυσης για την εξαγωγή συμπερασμάτων από τα Μεγάλα Σύνολα Δεδομένων. Αυτό περιλαμβάνει την προγνωστική ανάλυση, την ανίχνευση ανωμαλιών, την επεξεργασία φυσικής γλώσσας και τις εφαρμογές υπολογιστικής όρασης, επιτρέποντας στους οργανισμούς να αποκτούν βαθύτερες και πιο αξιοποιήσιμες γνώσεις από τα δεδομένα τους.

Επεξεργασία δεδομένων σε πραγματικό χρόνο: Αν και η τεχνολογία πραγματικού χρόνου βρίσκεται σε καλό επίπεδο, η ζήτηση για επεξεργασία και ανάλυση δεδομένων σε πραγματικό χρόνο είναι πιθανό να αυξηθεί, λόγω της ανάγκης για άμεσες γνώσεις και λήψη αποφάσεων σε διάφορους τομείς, όπως η χρηματοδότηση, η υγειονομική περίθαλψη, το IoT και η ασφάλεια στον κυβερνοχώρο. Τεχνολογίες όπως τα πλαίσια επεξεργασίας ροής, οι αρχιτεκτονικές που βασίζονται σε συμβάντα και οι βάσεις δεδομένων στη μνήμη θα καθορίσουν την εξέλιξη του τομέα στη δυνατότητα επεξεργασίας δεδομένων σε πραγματικό χρόνο.

Αύξηση όγκου δεδομένωνν και IoT: Με τον πολλαπλασιασμό των συσκευών και των αισθητήρων του Διαδικτύου των Πραγμάτων (IoT), η αύξηση του όγκου των μεγάλων δεδομένων θα οδηγήσει σε ολοένα και μεγαλύτερη έμφαση στις λύσεις υπολογισμού για την επεξεργασία και την ανάλυση δεδομένων πιο κοντά στην πηγή. Οι πλατφόρμες ανάλυσης άκρων (Edge analytics platforms) και οι αλγόριθμοι Τεχνητής Νοημοσύνης άκρων (edge AI algorithms) θα επιτρέψουν την επεξεργασία δεδομένων σε πραγματικό χρόνο, μειώνοντας την καθυστέρηση, τη χρήση εύρους ζώνης και εξαλείφοντας την εξάρτηση από κεντροποιήμενα κέντρα δεδομένων (centralized data centers).

Υιοθέτηση υβριδικών αρχιτεκτονικών: Οι οργανισμοί θα υιοθετούν ολοένα και περισσότερο υβριδικές και πολυ-νεφοϋπολογιστικές αρχιτεκτονικές για να αξιοποιήσουν την επεκτασιμότητα, την ευελιξία και τον πλεονασμό που προσφέρουν οι διάφοροι πάροχοι νέφους. Αυτό θα περιλαμβάνει την ενσωμάτωση υποδομών στις εγκαταστάσεις με περιβάλλοντα δημόσιου και ιδιωτικού cloud, καθώς και την

υιοθέτηση τεχνολογιών ενορχήστρωσης όπως το Kubernetes σύστημα ιδανικό για τη φορητότητα και τη διαχείριση του φόρτου εργασίας.

Απόρρητο και ασφάλεια δεδομένων: Με την αύξηση του όγκου των μεγάλων δεδομένων ωστόσο θα αυξηθούν και οι ανησυχίες σχετικά με το απόρρητο και την ασφάλεια τους. Είναι πασιφανές πως θα δοθεί αυξημένη έμφαση στην ανάπτυξη ισχυρής διακυβέρνησης δεδομένων, κρυπτογράφησης και τεχνικών ανωνυμοποίησης για την προστασία ευαίσθητων πληροφοριών. Οι αναλύσεις με γνώμονα τη διατήρηση της ιδιωτικής ζωής, θα φέρουν στο προσκήνιο έρευνα σε τομείς όπως η ομοσπονδιακή μάθηση (federated learning) και η διαφορική ιδιωτικότητα (differential privacy).

i. Federated learning καλείται η εκπαίδευση ενός κεντρικού μοντέλου σε αποκεντρωμένες συσκευές ή διακομιστές. Αντί να μεταφέρονται όλα τα δεδομένα σε μια κεντρική τοποθεσία, το μοντέλο εκπαιδεύεται τοπικά σε κάθε συσκευή και μόνο οι ενημερώσεις του μοντέλου μοιράζονται.
ii. Differential Privacy είναι ένας μαθηματικός τρόπος προστασίας των ατόμων όταν τα δεδομένα τους χρησιμοποιούνται σε σύνολα δεδομένων. Αυτές οι τεχνικές θα γίνουν πιο διαδεδομένες για την εξισορρόπηση της χρήσης των δεδομένων με τη διασφάλιση της ιδιωτικότητας.

Νομοθεσία χρήσης δεδομένων: Θα δοθεί μεγαλύτερη έμφαση στην ηθική χρήση δεδομένων και στις υπεύθυνες πρακτικές ΤΝ για τη διασφάλιση της δικαιοσύνης, της διαφάνειας και της λογοδοσίας στις διαδικασίες λήψης αποφάσεων βάσει δεδομένων. Αυτό περιλαμβάνει την εφαρμογή τεχνικών ανίχνευσης της μεροληψίας, την καθιέρωση κατευθυντήριων γραμμών για την ηθική ΤΝ και την προώθηση συνεργασιών μεταξύ επιστημόνων, ηθικολόγων και νομικών συμβούλων.

Συμπεραίνοντας λοιπόν τα Μεγάλα Σύνολα Δεδομένων έχουν αναδειχθεί σε μια ισχυρή δύναμη η οποία είναι ικανή να αναδιαμορφώνει βιομηχανίες, οικονομίες καθώς και τις κοινωνίες παγκοσμίως. Μέσα από τη διερεύνηση που κάναμε, εμβαθύναμε στο πολύπλευρο χώρο των Μεγάλων Δεδομένων, περιλαμβάνοντας τον ορισμό, την ιστορική εξέλιξη, τους διάφορους τύπους και τα βασικά χαρακτηριστικά τους. Η διαχείριση δεδομένων από τις απαρχές καταγραφής σε πάπυρους μέχρι τη σημερινή τους εξέχουσα θέση υπογραμμίζει τον καθοριστικό τους ρόλο στην προώθηση της εξέλιξης και την εξαγωγή συμπερασμάτων σε διάφορους τομείς. Όπως αναφέραμε, η εκθετική αύξηση του όγκου των δεδομένωνν, σε συνδυασμό με την πρόοδο της τεχνολογίας, άνοιξε το δρόμο για νέες τεχνολογίες στη συλλογή, αποθήκευση, επεξεργασία και ανάλυση δεδομένων. Η χρησιμότητα των μεγάλων δεδομένων εκτείνεται πέρα από τις τεχνικές εφαρμογές τους, καθορίζοντας διάφορους τομείς και κλάδους παγκοσμίως. Στον τομέα των βάσεων δεδομένων, η διχοτόμηση μεταξύ των παραδοσιακών σχεσιακών βάσεων δεδομένων και των βάσεων δεδομένων NoSQL αντικατοπτρίζει τις εξελισσόμενες ανάγκες της ανάλυσης μεγάλων δεδομένων. Στο

επίκεντρο της επεξεργασίας των Big Data βρίσκονται πλαίσια λογισμικού όπως αυτά που αναφέραμε, τα οποία δίνουν τη δυνατότητα στους οργανισμούς να αξιοποιήσουν τη δύναμη της κατανεμημένης πληροφορικής για πολύπλοκες αναλύσεις δεδομένων.

Στην ουσία, η σημασία των μεγάλων δεδομένων στην εποχή μας δεν μπορεί να υπερεκτιμηθεί. Η ικανότητά τους να προωθούν την καινοτομία, να ενημερώνουν για τη λήψη αποφάσεων και να ενδυναμώνουν τους οργανισμούς με αξιοποιήσιμες γνώσεις υπογραμμίζει την σημασία τους. Το μόνο σίγουρο είναι πως ο αντίκτυπός τους θα αυξηθεί, διαμορφώνοντας το μέλλον της κοινωνίας. Ευθύνη μας είναι να μεριμνήσουμε για την ασφαλή και ηθική διαμόρφωση αυτού, με στόχο το κοινό καλό. Η υιοθέτηση και η ασφαλής εκμετάλλευση των δυνατοτήτων των Μεγάλων Δεδομένων δεν είναι απλώς επιλογή, αλλά αναγκαιότητα για την διαδρομή μας, στον πολύπλοκο, εμπλουτιζόμενο συνεχώς με νέα δεδομένα, κόσμο μας.

6 Βιβλιογραφία

[1] Lexie Pelchen, (2024), Internet Usage Statistics, Forbes

[2] Isitor Emmanuel, Dr Clare Stanier,(2016), Defining Big Data

[3] Ibrar Yaqooba, Ibrahim Abaker Targio Hashema, Abdullah Gania, Salimah Mokhtara, Ejaz Ahmeda, Nor Badrul Anuara, Athanasios V. Vasilakos, (2016), Big data: From beginning to future

[4] Andrew McAfee and Erik Brynjolfsson,(2012),Big Data: The Management Revolution, Harvard Business Review

[5] N. H. Saeed,Laden Husamald,(2021),Big Data Characteristics (V's) in Industry, Iraqi Journal of Industrial Research

[6] Θοδωρής Παπαδόπουλος, Ανοιχτά Δημόσια Δεδομένα, Ε.Κ.Δ.Δ.Α | ΚΑ' ΕΚΠΑΙΔΕΥΤΙΚΗ ΣΕΙΡΑ

[7] Liyakathunisa, Saima Jabeen, Manimala S, and Hoda A. Elsayed, (2019), Data Science Algorithms and Techniques for Smart Healthcare using IoT and Big Data Analytics

[8] Sonia Buchholtz, Maciej Bukowski, Aleksander Śniegocki,(2014), Big and open data in Europe, Warsaw Institute for Economic Studies

[9] Jianqing Fan, Fang Hanand Han Liu, (2014), Challenges of Big Data analysis, National Science Review

[10] Μεχίλι Μαρία, Χριστοπούλου Ευσταθία, (2015), Τεχνικές Ανάλυσης Μεγάλων Δεδομένων,ΤΕΙ Δυτικής Ελλάδας, Τμήμα Διοίκηση Επιχειρήσεων

[11] Cornelia Győrödi Robert Gyorodi Alexandra Ștefan Bandici Livia,(2016), A Comparative Study of Databases with Different Methods of Internal Data Management

[12] Felix Gessert, Wolfram Wingerath, Steffen Friedrich, Norbert Ritter (2016), NoSQL database systems: a survey and decision guidance

[13] Sourav Mukherjee,(2019), The battle between NoSQL Databases and RDBMS

[14] Alfredo Cuzzocrea,Domenico Saccà,Jeffrey D. Ullman, (2013), Big Data: A Research Agenda

[15] Adity Gupta, Swati Tyagi, Nupur Panwar, Shelly Sachdeva Jaypee, Upaang Saxena,(2017), NoSQL Databases: Critical Analysis and Comparison

[16] Min Chen, Shiwen Mao, Yunhao Liu, (2014), Big Data: A Survey

[17] Avinash Lakshman,Prashant Malik,(2010), Cassandra A Decentralized Structured Storage System

[18] White Paper by Datastax corporation,(2013), Introduction to Apache Cassandra

[19] Yunhua Gu, Shu Shen, Jin Wang, Jeong-Uk Kim, (2015), Application of NoSQL Database MongoDB, International Conference on Consumer Electronics-Taiwan

[20] Hema Krishnan, M.Sudheep Elayidom, T.Santhanakrishnan,(2017), MongoDB a comparison with NoSQL databases, International Journal of Scientific & Engineering Research

[21] Hiren Patel,(2017), HBase: A NoSQL Database

[22] Rohit Reddy Nalla,(2015), A case study on Apache HBase, Department of Computer and Information Sciences SUNY Polytechnic Institute Utica, New York

[23] Félix Melchor Santos López, Eulogio Guillermo Santos De La Cruz,(2015) Literature review about Neo4j graph database as a feasible alternative for replacing RDBMS, Revista de la Facultad de Ingeniería Industrial
[24] Szymon DZIUBAK, (2023) REVIEW OF CLOUD DATABASE BENEFITS AND CHALLENGES
[25] Ibrahim Abaker Targio Hashem, Nor Badrul Anuar, Abdullah Gani, Ibrar Yaqoob, Feng Xia, Samee Ullah Khan,(2016), MapReduce: A bibliometric, review and open challenges
[26] Jeffrey Dean,Sanjay Ghemawat ,(2004), MapReduce: Simplified Data Processing on Large Clusters, Google Inc.
[27] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, (2010), The Hadoop Distributed File System, Yahoo!, Sunnyvale, California USA
[28] Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali,Muhammad Alam,Muhammad Shiraz and Abdullah Gani,(2014), Big Data: Survey, Technologies, Opportunities, and Challenges
[29] V. Sajwan, V. Yadav, Dr.M.Haider, (2015), The Hadoop Distributed File System: Architecture and Internals
[30] Vinod Kumar Vavilapalli Arun, Murthy Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, Bikas Saha, Carlo Curino, Owen O'Malley, Sanjay Radia, Benjamin Reed, Eric Baldeschwieler, (2013) Apache Hadoop YARN: Yet Another Resource Negotiator
[31] Nikita Kathare, O. Vinati Reddy, Dr. Vishalakshi Prabhu,(2021), Comprehensive Study of Elasticsearch, International Journal of Science and Research (IJSR)
[32] https://www.researchgate.net/figure/Graph-illustration_fig1_307180380