ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΛΟΠΟΝΝΗΣΟΥ
ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Πτυχιακή Εργασία

Η διαχείριση μεγάλων δεδομένων με τις NoSQL ΒΔ και το SPARK

ΛΑΖΑΡΗ ΜΑΡΙΑ-ΕΛΕΝΗ

Επιβλέπων καθηγητής: Δρ. Ταμπακάς Βασίλειος

ΠΑΤΡΑ 2022

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή

Πάτρα,04/05/2023

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Ονοματεπώνυμο, Υπογραφή

2. Ονοματεπώνυμο, Υπογραφή

3. Ονοματεπώνυμο, Υπογραφή

Υπεύθυνη Δήλωση Φοιτητή

Βεβαιώνω ότι είμαι συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης έχω αναφέρει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επίσης βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά ειδικά για τη συγκεκριμένη εργασία.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Πανεπιστημίου Πελοποννήσου δεν υποδηλώνει απαραιτήτως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον καθηγητή κ. Ταμπακά Βασίλειο για την καθοδήγηση του στην εκπόνηση της παρούσας πτυχιακής εργασίας.

Περίληψη

Η μεγαλύτερη πρόκληση της επιστήμης των Δεδομένων αποτελεί η διαχείριση των δεδομένων, καθώς υπάρχει εκθετική αύξηση του όγκου τους εξαιτίας της αλματώδους τεχνολογικής προόδου. Η επιστήμη των υπολογιστών έχει αναπτύξει συστήματα τα οποία μπορούν να ανταποκριθούν στις απαιτήσεις της ταχείας παραγωγής, αποθήκευσης, εξόρυξης και ανάλυσης δεδομένων. Η έλλειψη αυτών των συστημάτων/εργαλείων θα δυσχέραινε την αξιοποίηση των Big Data για εξόρυξη νέων σημαντικών πληροφοριών. Η παρούσα πτυχιακή εργασία έχει στόχο την διερεύνηση των δυνατοτήτων και των χαρακτηριστικών της NoSQL, Apache Cassandra και του Spark για τα Big Data.

Πίνακας περιεχόμενων

Κεφάλαιο 1: Ψηφιακή εποχή

1.1 Εισαγωγή

Η επιστήμη των υπολογιστών έχει μικρή ιστορία συγκριτικά με τους υπόλοιπους επιστημονικούς κλάδους, ενώ ταυτόχρονα τους βοηθά σημαντικά μέσω των εργαλείων που έχουν αναπτυχθεί. Παρατηρώντας την ιστορία της ανθρωπότητας, αντιλαμβανόμαστε ότι ζούμε μέσα στην ψηφιακή επανάσταση. Αυτή η επανάσταση, όπως και η Βιομηχανική, επιφέρει αλλαγή στον τρόπο επικοινωνίας, εκπαίδευσης, εργασίας και γενικότερα σε όλο τον τρόπο ζωής. Καθημερινά δημιουργούνται και συλλέγονται διαφόρων ειδών δεδομένα (κείμενο, εικόνα, ήχος κτλ.), τα οποία είναι χρήσιμα για την αέναη ανάπτυξη της τεχνολογίας.

1.2 Σκοπός πτυχιακής

Σκοπός της πτυχιακής εργασίας είναι η διαχείριση μεγάλων δεδομένων (Big Data) με την αξιοποίηση όλων των δυνατοτήτων και των χαρακτηριστικών NoSQL βάσεων δεδομένων Apache Cassandra. Το εργαλείο που χρησιμοποιείται για την επεξεργασία BD είναι το Apache Spark.

Κεφάλαιο 2: Big Data

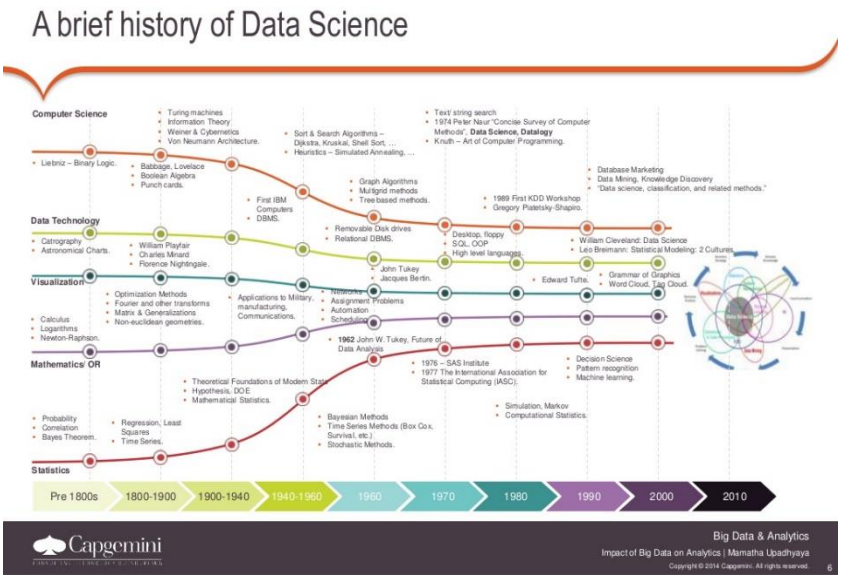
2.1: Τι είναι Big Data;

Η έννοια των Μεγάλων Δεδομένων εμφανίστηκε στα τέλη του 20ού αιώνα, αλλά στηρίχθηκε σε θεμέλια που τέθηκαν πριν την εμφάνιση των πρώτων υπολογιστών. Από την αρχαιότητα μέχρι και σήμερα υπάρχει ανάγκη για αποθήκευση και αξιοποίηση πληροφοριών. Τα τελευταία 20 χρόνια παρατηρούμε ραγδαία αύξηση στον όγκο των πληροφοριών που διαχειριζόμαστε, εξαιτίας της ραγδαίας εξέλιξης της τεχνολογίας και της εισβολής της στην ζωή μας.

Ιστορική αναδρομή

Από την αρχαιότητα μέχρι την εμφάνιση του ENIAC (1945)

Η ιστορία της ανάγκης των ανθρώπου να έχει αποθηκευμένα δεδομένα αρχίζει από την αρχαιότητα. Συγκεκριμένα, τον 18ο αιώνα π.Χ. βασικό μέλημα των ανθρώπων ήταν να καταγράφουν το απόθεμα τους και να ελέγχουν την εμπορική δραστηριότητά τους, ώστε να αυξάνουν το κέρδος τους μειώνοντας τα έξοδα. Το 2400 π.Χ. εφευρίσκεται το αριθμητήριο, το οποίο αποτελεί για αρκετούς επιστήμονες την αρχή των υπολογιστών. Αρκετά χρόνια αργότερα, το 300 π.Χ., χτίζεται το πρώτο αποθετήριο δεδομένων, η βιβλιοθήκη της Αλεξάνδρειας. Το πρώτο αρχέγονο υπολογιστικό σύστημα εμφανίζεται το 100 π.Χ., είναι ο μηχανισμός των Αντικυθήρων ένας αρχαίος αναλογικός υπολογιστής και όργανο για την παρατήρηση αστερισμών. Στη νεότερη ιστορία της Ευρώπης, καταγράφεται η προσπάθεια για στατιστική ανάλυση, με σκοπό να περιοριστεί η εξάπλωση της επιδημίας (πανώλη) από τον John Graunt 1663. Στις ΗΠΑ το 1881, ο Herman Hollerith θέλοντας να βοηθήσει την εργασία των απογραφέων δημιούργησε ένα μηχανικό πίνακα με τη χρήση διάτρητων καρτών, ώστε να ταξινομούνται ευκολότερα τα αποτελέσματα των μετρήσεων. Το 1926, ο εφευρέτης του εναλλασσόμενου ηλεκτρισμού Nikola Tesla σε συνέντευξη στο περιοδικό Colliers προέβλεψε ότι οι άνθρωποι στο μέλλον θα έχουν τη δυνατότητα πρόσβασης και ανάλυσης δεδομένων με τη χρήση μικρής συσκευής. Αυτή η μικρή συσκευή που αναφέρει μπορεί να παρομοιαστεί με τα σύγχρονα smart phones, διότι έχουν μικρό μέγεθος, αλλά μεγάλες υπολογιστικές δυνατότητες.



Εικόνα 1: Η ιστορία της επιστήμης των δεδομένων

Από τον ENIAC μέχρι σήμερα

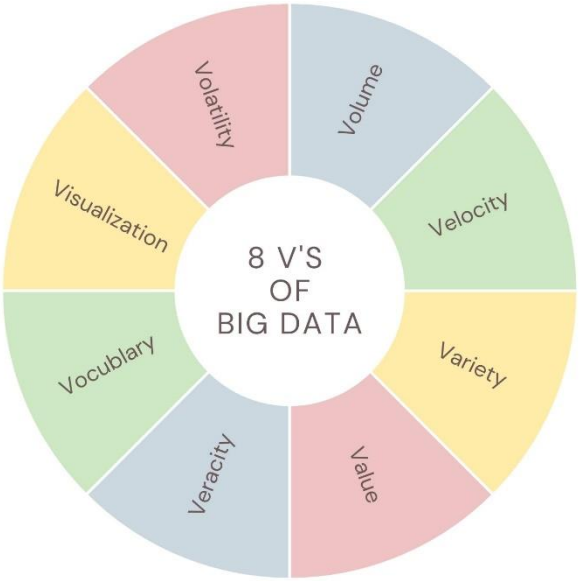
Από το 1945 και μετά η ιστορία των δεδομένων, αλλά και όλου του κλάδου της τεχνολογίας αλλάζει για πάντα με την δημιουργία του πρώτου σύγχρονου υπολογιστή ENIAC. Από το 1958-1968, η IBM αναπτύσσει τις έννοιες της επιχειρηματικής ευφυΐας και της ψηφιακού αποθετηρίου με μαγνητική ταινία, ώστε να αποθηκεύονται εκατομμύρια φορολογικές δηλώσεις και δακτυλικά αποτυπώματα των Αμερικάνων πολιτών. Οι επιστήμονες της IBM συνεχίζουν να πρωτοπορούν στην επιστήμη των δεδομένων, καθώς το 1976 ο Edgar F Codd προτείνει το σχεσιακό μοντέλο δεδομένων, με το οποίο ιεραρχείται ο τρόπος αποθήκευσης δεδομένων. Το μοντέλου το Codd αποτελεί ορόσημο στην ιστορία, διότι ξεκινάει η ευρέως χρήση των δεδομένων έξω από τους επιστημονικούς χώρους. Ο συγγραφέας Eric Larson ανέφερε τον όρο Big Data σε συνέντευξη του στο περιοδικό Harper. Ένα χρόνο αργότερα το 1990, ο John R.Mashley ορίζει την έννοια Μεγάλα Δεδομένα ονομάζεται το σύνολο των δεδομένων που δεν είναι αντιληπτά και διαχειρίσιμα από την κλασσική πληροφορική και τα εργαλεία του λογισμικού. Κατά την εικοσαετία 1990-2010, υπάρχει ραγδαία εξάπλωση στη χρήση πληροφοριακών συστημάτων, το οποίο σήμαινε και απότομη αύξηση των πληροφοριών που αποθηκεύονται. Ορόσημο της εικοσαετίας για τα Big Data είναι το 2009, με το ίδρυμα McKinsey να αναφέρει ότι η ανθρωπότητα βρίσκεται στις αρχές της 4ης βιομηχανικής επανάστασης. Η επανάσταση της ψηφιακής εποχής με αναπόσπαστο μέρος της τα Big Data, επιφέρει αλλαγές σε όλους τους τομείς της ζωής.

Δεν μπορεί να αποτυπωθεί με ακρίβεια ο ορισμός των Big Data, διότι υπάρχουν παντού,εξελίσσονται συνεχώς και είναι αυτά που ορίζουν το μέλλον, αξιοποιώντας τη γνώση του παρελθόντος. Ο πιο σύγχρονος ορισμός, αναφέρεται από την Wikipedia: <<Τα μεγάλα δεδομένα είναι το επιστημονικό πεδίο, όπου γίνεται επεξεργασία των τρόπων ανάλυσης και συστηματικής εξόρυξης πληροφοριών από σύνολα δεδομένων που είναι πολύπλοκα διασυνδεδεμένα και διαχειρίζονται δύσκολα με την χρήση του παραδοσιακού λογισμικού >>.

2.2: Χαρακτηριστικά Big Data

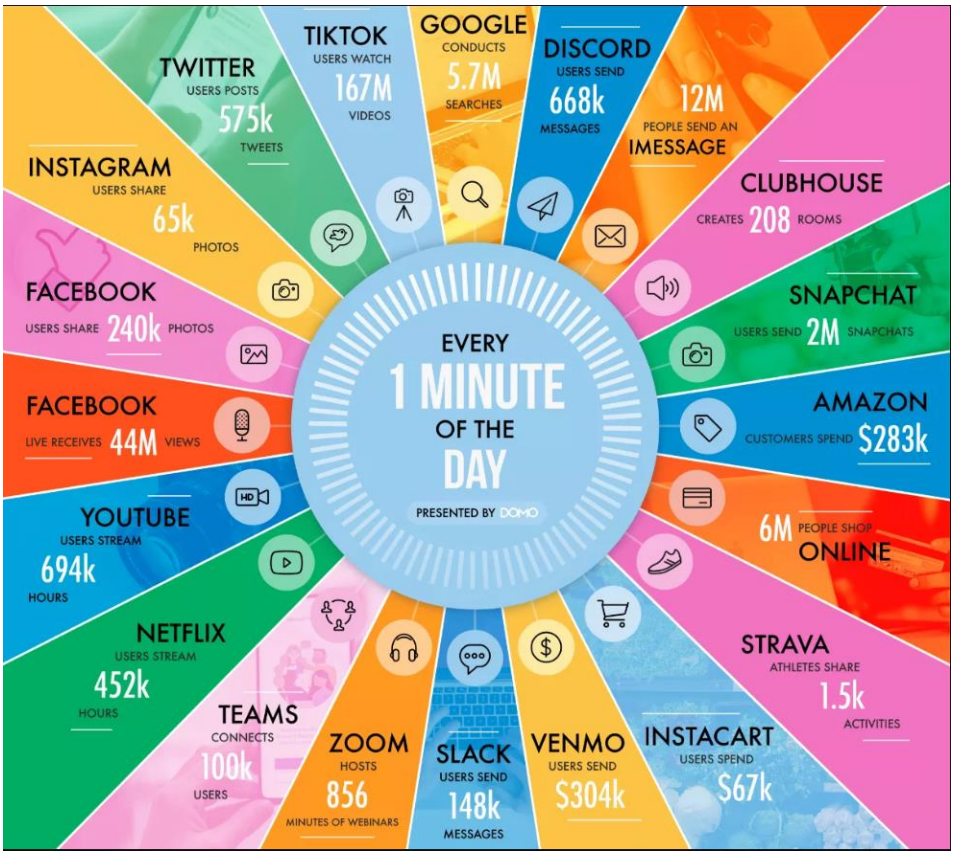
2.2.1: Τα 8V των BD

Στη διεθνή βιβλιογραφία, όλοι οι ορισμοί των Big Data συνοψίζονται σε τρεις βασικούς άξονες: όγκος(Volume), ταχύτητα (Velocity), ποικιλία (Variety). Ωστόσο, η ραγδαία ανάπτυξη της τεχνολογίας επέφερε την επέκταση των χαρακτηριστικών με στόχο την λεπτομερέστερη ακρίβεια. Ανάλογα με την χρήση και τον βαθμό χρήσης των Big Data, η κάθε εταιρία και ερευνητής θέτει το πλήθος των χαρακτηριστικών που τον εξυπηρετούν.



Πίνακας 1: 8 V's of Big Data

Όγκος (Volume): Τα τελευταία χρόνια παρατηρείται εκθετική αύξηση του πλήθους των δεδομένων που καλούμαστε να αποθηκεύσουμε. Στις αρχές του 21 αιώνα είχαν δημιουργηθεί 15.6 Exabytes (~17 εκατομμύρια GB). Είκοσι χρόνια αργότερα αυτός ο αριθμός εκατονταπλασιάστηκε. Το 2021 υπολογίστηκε ότι ο μέσος άνθρωπος παρήγαγε 50MB κάθε δευτερόλεπτο. Παρατηρώντας την παρακάτω εικόνα μπορεί να γίνει κατανοητό το μέγεθος των πληροφοριών που αποθηκεύονται και επεξεργάζονται κάθε λεπτό. Οι ερευνητές προβλέπουν ότι, το 2035 ο όγκος των δεδομένων θα είναι 2.142 Zetabytes. Γίνεται κατανοητό ότι υπάρχει επιτακτική ανάγκη για ανάπτυξη νέων εφαρμογών και επαναπροσδιορισμό των υπαρχόντων αρχιτεκτονικών, ώστε τα δεδομένα να εξάγονται (***Extract***) , να επεξεργάζονται (***Transform***) και να φορτώνονται (***Load***) σε ευρείας χρήσης εφαρμογές *ταχύτερα.*



Εικόνα 2: Τα δεδομένα που παράγονται κάθε λεπτό το 2021

Ταχύτητα (Velocity):Ο τεράστιος όγκος δεδομένων προήλθε από την αύξηση της ταχύτητας κατά την οποία παράγονται, διανέμονται και συλλέγονται τα δεδομένα. Τα σύγχρονα συστήματα χρησιμοποιούν ειδικές τεχνικές επεξεργασίας BD, ώστε να μπορέσουν να ανταποκριθούν στις ανάγκες του συστήματος χωρίς να χαθούν τα πολύτιμα δεδομένα. Όσο υψηλότερος είναι ο ρυθμός ταχύτητας, τόσο πιο γρήγορα τα δεδομένα επεξεργάζονται αυξάνοντας την αξίας τους. Οι δύο τύποι ταχύτητας που σχετίζονται με τα μεγάλα δεδομένα είναι η συχνότητα δημιουργίας και η συχνότητα χειρισμού, καταγραφής και αναφοράς [4] . Για τον έλεγχο της ταχύτητας ροής των δεδομένων έχουν αναπτυχθεί εφαρμογές που μπορούν να διαχειρίζονται μεγάλα πακέτα χρήσιμης πληροφορίας σε πραγματικό χρόνο, παραδείγματος χάριν Apache Spark .Παρόλα αυτά απαιτείται χρόνος και προσπάθεια για την δημιουργία αγωγών δεδομένων(data pipelines) για ομαλή διακίνηση των δεδομένων μέσα σε ένα σύστημα.

Ποικιλία (Variety):Τα μεγάλα δεδομένα αντλούν δεδομένα από διαφορετικές πηγές (αισθητήρες, Internet), ώστε να συνθέσουν τις χρήσιμες πληροφορίες που χρειάζεται ένα σύστημα. Η ραγδαία αύξηση της χρήσης αισθητήρων και του IoT, δημιούργησε νέες μορφές δεδομένων όπως δομημένα κείμενα, ήχος, εικόνα, βίντεο, email, καταγραφή διαδικτυακής κίνησης για website, διαγράμματα καρδιακών παλμών κ.ά.. Αυτοί οι τύποι δεδομένων είναι πολύ

διαφορετικοί σχέση με τις πρώτες DB όπου αποτελούνταν αποκλειστικά από πλειάδες και γραμμές. Αυτό σημαίνει ότι τις περισσότερες φορές είναι αδόμητα. Ένας τρόπος διαχείρισης της ποικιλομορφίας των δεδομένων είναι η δημιουργία οροσήμων σε κάθε στάδιο της επεξεργασίας τους. Τα ακατέργαστα δεδομένα αποθηκεύονται προσωρινά σε data lake και στην συνέχεια γίνεται μετατροπή τους σε συγκεκριμένους τύπους, ώστε να φορτωθούν σε RDBMS(σχεσιακή βάση δεδομένων).

Εγκυρότητα (Veracity):Τα δεδομένα συχνά θεωρούνται επικαιροποιημένα και αξιόπιστα. Ωστόσο, στα πραγματικά σύνολα δεδομένων συχνά εμφανίζονται δεδομένα με ανακρίβεια και αβεβαιότητα ως προς την εγκυρότητά τους. Η εγκυρότητα των δεδομένων δεν έγκειται μόνο στην αληθοφάνεια, αλλά και στην ποιότητα των δεδομένων που πρόκειται να μετασχηματιστούν και να χρησιμοποιηθούν για την λήψη κρίσιμων αποφάσεων σε ένα σύστημα. Η ποιότητα των δεδομένων εξαρτάται από ορισμένους παράγοντες όπως: πηγή εξόρυξης, τεχνική συλλογής και τρόπος ανάλυσης κ.ά.. Δεδομένα χαμηλής εγκυρότητας, συνήθως περιέχουν υψηλό ποσοστό «θορύβων» και ανούσιων δεδομένων, τα οποία είναι ανώφελα προς διαχείριση. Από την άλλη πλευρά, τα δεδομένα υψηλής ακρίβειας περιέχουν πολλές εγγραφές, που είναι πολύτιμες για επεξεργασία, συμβάλλοντας σημαντικά στην επίτευξη των στόχων ενός συστήματος. Η βελτίωση της ποιότητας των δεδομένων, μπορεί να επιτευχθεί με την προσθήκη φίλτρων κατά την εξόρυξη τους. Η χρήση φίλτρων χρήζει ιδιαίτερη προσοχή, διότι μπορεί να αφαιρεθεί μεγάλο πλήθος «θορύβων», διπλότυπων εγγραφών, αλλά να καταστήσουν ένα σύστημα ασταθές.

Μεταβλητότητα (Volatility): Στην πάροδο του χρόνου το περιεχόμενο των αποθηκευμένων δεδομένων μεταβάλλεται εξαιτίας διαφόρων παραγόντων. Οι ερευνητές δίνουν έμφαση στον ρυθμό μεταβολής αυτών των τιμών, διότι οι παλαιότερες πληροφορίες ίσως να είναι λανθασμένες, ασαφείς και μη χρήσιμες. Το σημείο στο οποίο υπάρχει διχασμός απόψεων στην επιστήμη των δεδομένων αφορά τη διάρκεια ζωής των δεδομένων. Η διατήρηση και αξιοποίηση παλιωμένων δεδομένων ενέχει τους προαναφερθέντες κινδύνους, όμως η εξέλιξη ενός συστήματος βασίζεται σε παλαιότερες πληροφορίες. Το 2021 ο Sandvik όρισε το μοντέλο μεταβλητότητας. Το μοντέλο εξετάζει τις διαφορές μεταξύ του τρόπου διαχείρισης δεδομένων από τις συσκευές αποθήκευσης και την μέθοδο για την εξόρυξη τους.

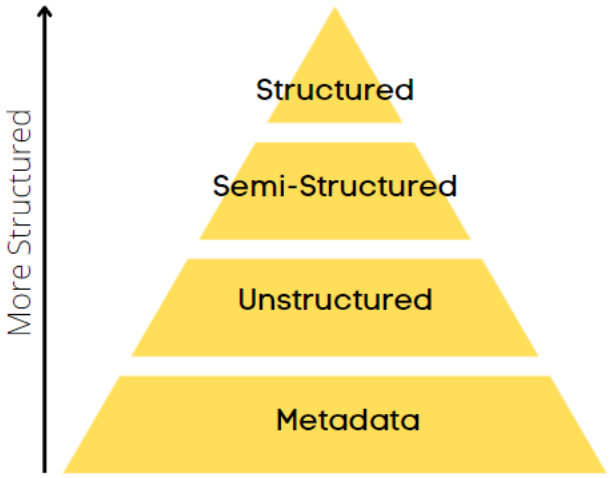
Αξία (Value):Πριν ξεκινήσει οποιαδήποτε διαδικασία μετασχηματισμού των δεδομένων είναι σημαντικό να εξεταστεί η αξία τους. Αναγκαία είναι η χρήση αποδοτικών τεχνικών ανάλυσης ώστε να εξαχθούν πρωτοπόρες γνώσεις για την ορθή λήψη αποφάσεων. Ο υπολογισμός της αξίας των δεδομένων είναι απαραίτητος για την ευημερία ενός συστήματος, διότι προκύπτει με την χρήση στατιστικών μοντέλων εμφανίζοντας όλους τους παράγοντες από τους οποίους επηρεάζεται αυτή η επιτυχία.

Οπτικοποίηση (Visualization):Τα μεγάλα δεδομένα περιέχουν εκατομμύριες ή και δισεκατομμύριες ακατέργαστες πληροφορίες, οι οποίες με την χρήση εργαλείων οπτικοποίησης κατανοούνται από τους χρήστες ενός συστήματος. Η οπτικοποίηση περιγράφει σχεδόν όλους του τύπους δεδομένων (αριθμοί, τριγωνομετρικές συναρτήσεις, στατιστικοί αλγόριθμοι) σε **οπτική μορφή** (εκθέσεις αναλυτικών στοιχείων, διαγράμματα). Τα γραφήματα αυτά διαφέρουν από τα τυπικά μαθηματικά διαγράμματα, καθώς περιέχουν συνθετότερες αναπαραστάσεις, όπως heatmaps, wordclouds κ.ά., αναδεικνύοντας απροσδόκητες συσχετίσεις και μοτίβα μεταξύ των δεδομένων, ώστε να ληφθούν αποτελεσματικότερες αποφάσεις προς όφελος ενός συστήματος.

Λεξιλόγιο (Vocabulary):Η επεξεργασία και η ανάλυση τεράστιων όγκων δεδομένων κάποιες φορές δημιουργεί αοριστίες συσχετίσεων, διότι δεν συνεπάγεται αιτιώδη και σημασιολογική συνάφεια. Η λέξη «λεξιλόγιο» στον κλάδο της επιστήμης των δεδομένων έχει δύο ερμηνείες. Η πρώτη ερμηνεία αναφέρεται στο ζήτημα της επικοινωνίας μεταξύ του παρόχου και του χρήστη ενός συστήματος καθώς και της γλώσσας που χρησιμοποιείται για να περιγραφεί το επιθυμητό αποτέλεσμα. Η δεύτερη ερμηνεία διακλαδίζεται στην σημασιολογική αναζήτηση και στις λειτουργίες μέσα σε ένα σημασιολογικό χώρο. Καθορίζονται με σαφήνεια οντολογίες που αντιπροσωπεύουν συγκεκριμένους ορισμούς αλλά και ταυτόχρονα είναι αλληλένδετες με άλλους όρους. Για παράδειγμα στον χώρο της τεχνητής νοημοσύνης ο όρος «παιδί» υποχρεωτικά συνεπάγεται ότι έχει «γονέα».

2.2.2: Δομή BD

Στην επιστήμη των υπολογιστών, μια δομή δεδομένων είναι ο ιδιαίτερος τρόπος οργάνωσης και αποθήκευσης δεδομένων σε έναν υπολογιστή ώστε να υπάρχει πρόσβαση και αποτελεσματική διαχείριση τους. Τα δεδομένα παράγονται από την αλληλεπίδραση ανθρώπου-υπολογιστή, την διεπαφή μηχανών χωρίς την ανθρώπινη παρέμβαση. Όλες αυτές οι αλληλεπιδράσεις δημιουργούν τέσσερις τύπους δομών.



Πίνακας 2: Πυραμίδα δομής τωνBig Data

Δομημένα δεδομένα (Structured data):Τα δομημένα δεδομένα αποτελούν την πιο «κλασική» μορφή αποθήκευσης δεδομένων, αφού τα πρώτα συστήματα διαχείρισης βάσεων δεδομένων μπορούσαν να αποθηκεύουν, να επεξεργάζονται και να έχουν πρόσβαση μόνο σε αυτά. Αποτελούνται από πλειάδες και στήλες, και τα δεδομένα εγγράφονται σύμφωνα με τις προκαθορισμένες συσχετίσεις του εκάστοτε συστήματος. Εξαιτίας της αυστηρής εξάρτησης από το μοντέλο/πρότυπο διαχείρισης που ακολουθούν, τα δομημένα δεδομένα είναι εξαιρετικά ισχυρά και μπορούν εύκολα να συλλεχθούν από διαφορετικές (κατανεμημένες ή και μη-κατανεμημένες) βάσεις δεδομένων και να αναλυθούν.

Ημιδομημένα δεδομένα(Semi-Structured data):Τα ημιδεδομένα δεδομένα, όπως αρχεία XML και JSON,διαθέτουν καθορισμένα και συνεπή χαρακτηριστικά, αλλά δεν δομούνται σύμφωνα με προεπιλεγμένες συσχετίσεις όπως τα δομημένα δεδομένα. Το γεγονός ότι δεν εξαρτώνται τόσο έντονα από πρότυπα διαχείρισης, δεν σημαίνει ότι είναι σημασιολογικά ανοργάνωτα και μη προσβάσιμα για επεξεργασία.

Αδόμητα δεδομένα(Unstructured data):Τα αδόμητα δεδομένα, όπως ήχος, εικόνες, email, δορυφορικά σήματα, ούτε διαθέτουν προκαθορισμένο μοντέλο δεδομένων ούτε οργανώνονται με αυστηρές συσχετίσεις. Αυτοί οι δύο παράγοντες προκαλούν παρατυπίες και ασάφειες κάνοντας τα αδόμητα δεδομένα δύσκολα ως προς την διαχείριση και ιδιαίτερα την ανάλυση τους. Η επιτακτική ανάγκη εξόρυξης πολύτιμων πληροφοριών από τα αδόμητα αποτελεί μία από τις βασικότερες αιτίες για την ταχεία ανάπτυξη των μεγάλων δεδομένων, καθώς αναπτύχθηκαν πολλές νέες τεχνολογίες και εργαλεία.

Μεταδεδομένα (Metadata):Τα μεταδεδομένα είναι ένα από τα σημαντικότερα στοιχεία για την ανάλυση των Big Data. Παρέχουν πρόσθετες πληροφορίες για ένα ήδη υπάρχον σύνολο δεδομένων, δηλαδή δεδομένα για τα δεδομένα. Το χαρακτηριστικότερο παράδειγμα χρήσης των μεταδεδομένων είναι οι αναρτημένες φωτογραφίες στα μέσα κοινωνικής δικτύωσης, οι οποίες αποτελούνται από αδόμητα δεδομένα -pixels-, δομημένα δεδομένα – σύντομο επεξηγηματικό κείμενο (λεζάντα)- και ημιδεδομένα δεδομένα-URL-. Τα μεταδεδομένα είναι η ημερομηνία και ώρα δημιουργία της ανάρτησης, το όνομα του δημιουργού. Τα μεταδεδομένα είναι ευμετάβλητα, επιτρέποντας την κατηγοριοποίηση και ανάλυση τους.

2.3: Αρχιτεκτονική των Big Data

Η αρχιτεκτονική των μεγάλων δεδομένων αποτελεί το πρωταρχικό σύστημα που χρησιμοποιείται για τη διαχείριση των BD, χρησιμοποιώντας τα κατάλληλα εργαλεία ανάλυσης δεδομένων ούτως ώστε να εξαχθούν ωφέλιμες πληροφορίες από τα πολύπλοκα δεδομένα. Στην αρχιτεκτονική καθορίζονται με σαφήνεια τα συστατικά στοιχεία, τα επίπεδα και τους αγωγούς ροής δεδομένων (data pipelines) που θα χρησιμοποιηθούν.

Πηγή δεδομένων(Data source):Ο σχεδιασμός της αρχιτεκτονικής εξαρτάται σε μεγάλο βαθμό από τις πηγές δεδομένων. Προτού εφαρμοστεί οποιαδήποτε αρχιτεκτονική μέθοδος, θα πρέπει οι πηγές δεδομένων να προσδιοριστούν και να κατηγοριοποιηθούν. Όπως προαναφέρθηκε τα δεδομένα προέρχονται από πολλές πηγές και σε διαφορετικές μορφές.

Αποθήκευση δεδομένων (Data storage):Η αποθήκευση δεδομένων αποτελεί το βασικότερο στοιχείο ή και επίπεδο της αρχιτεκτονικής των BD. Αυτό είναι το επίπεδο λήψης δεδομένων, που ενοποιεί δεδομένα από διάφορες πηγές, τα αποθηκεύει *ιδανικά* σε ένα κατανεμημένο χώρο αποθήκευσης και μετασχηματίζει τα αδόμητα και τα μεταδεδομένα σε μορφές δεδομένων σύμφωνες με τις απαιτήσεις ενός συστήματος. Τα structured data συνήθως αποθηκεύονται σε σχεσιακή DB, ενώ τα unstructured data μπορούν τοποθετηθούν σε μη-σχεσιακές (*NoSQL)*βάσεις, όπως Apache Cassandra.

Ομαδοποιημένη επεξεργασία (Batch processing):Τα σύνολα των δεδομένων είναι τεράστια, συχνά ως λύση βέλτιστης διαχείρισης απαιτείται ένα σύστημα ομαδοποιημένης επεξεργασίας. Το σύστημα αποτελείται από μακροχρόνιες διαδικασίες, οι οποίες φιλτράρουν, συγκεντρώνουν και επεξεργάζονται δεδομένα παλαιοτέρων αναλυτικών στοιχείων. Κατά την διάρκεια της ομαδοποιημένης επεξεργασίας, τα αρχεία, που δημιουργούνται από την ενοποίηση δεδομένων διαφόρων πηγών, διαβάζονται επεξεργάζονται και τα νέα αποτελέσματα τους εγγράφονται σε νέα αρχεία.

Επεξεργασία σε πραγματικό χρόνο (Real-time processing):Τα περισσότερα συστήματα επεξεργάζονται δεδομένα σε πραγματικό χρόνο και έχουν ανάγκη για την ύπαρξη κάποιου μηχανισμού απορροφώντας και αποθηκεύοντας μηνύματα για ομαλή ροή των δεδομένων. Η απορρόφηση των μηνυμάτων συνήθως πραγματοποιείται με την προσωρινή αποθήκευση μηνυμάτων, η οποία υποστηρίζει την βαθμιαία επεξεργασία, την αξιόπιστη παράδοση και την σημασιολογική αλληλουχία των πληροφοριών. Αφού καταγραφούν τα μηνύματα, τα δεδομένα πρέπει να φιλτραριστούν, να ενοποιηθούν και να προετοιμαστούν για ανάλυση. Αυτό το επίπεδο αρχιτεκτονικής εστιάζει στην κατηγοριοποίηση δεδομένων για ομαλή μετάβαση στα βαθύτερα στρώματα του συστήματος.

Ανάλυση και αναφορά δεδομένων(Data Analysis&reporting):Ο σκοπός των περισσότερων πλατφορμών διαχείρισης μεγάλων δεδομένων είναι η λήψη ορθών αποφάσεων, η οποία προέρχεται από την ανάλυση δεδομένων και τη δημιουργία αναφορών. Το επίπεδο ανάλυσης αλληλοεπιδρά με το επίπεδο αποθήκευσης θέτοντας καίρια ερωτήματα με στόχο την εξόρυξη πολύτιμων γνώσεων. Οι αναφορές αποτελούν την λογική διασύνδεση όλων των αποτελεσμάτων της επεξεργασίας και ανάλυσης δεδομένων, αναδεικνύοντας την βέλτιστη απόφαση. Τα τελευταία χρόνια, η ανάλυση δεδομένων και η δημιουργία αναφορών έχει εδραιωθεί εξαιτίας των εργαλείων που έχουν αναπτυχθεί, όπως Apache Spark.

Αυτοματοποίηση (Automation):Ο τρόπος διαχείρισης μεγάλων δεδομένων αποτελείται από επαναλαμβανόμενες λειτουργίες επεξεργασίας δεδομένων, που ακολουθούνται από μετασχηματισμό δεδομένων, μετακίνηση μεταξύ των πηγών και φόρτωση των επεξεργασμένων δεδομένων σε έναν χώρο αποθήκευσης αναλυτικών δεδομένων. Όλες αυτές οι ενσωματωμένες εργασίες πρέπει να αυτοματοποιηθούν χρησιμοποιώντας κάποιο εργαλείο συντονισμού όπως το Apache Oozzie ώστε να λαμβάνονται συνεχώς πολύτιμες πληροφορίες.

2.4: Εφαρμογές των Big Data;

Τα μεγάλα δεδομένα μπορούν να χρησιμοποιηθούν για την αποκάλυψη κρυφών μοτίβων και τάσεων, τα οποία είναι εξαιρετικά χρήσιμα για όλους τους τομείς. Ιδιαίτερα για τους οργανισμούς παρέχει την ευκαιρία σε εμπειρογνώμονες των επιχειρήσεων να ερωτούν και να κατανοούν τις πολύτιμες πληροφορίες σύμφωνα με τις επιχειρηματικές ανάγκες, ανεξάρτητα από τη δυσκολία και τον όγκο των δεδομένων, παρουσιάζοντας τα δεδομένα με κατανοητό τρόπο. Η ανάλυση αυτών των πληροφοριών μπορούν να εμφανιστούν νέες κατευθύνσεις για την λήψη αποφάσεων που θα αποφέρουν αποδοτική λειτουργία του οργανισμού, μεγάλα οικονομικά κέρδη και περισσότερους ευχαριστημένους πελάτες.

Εικόνα 3: Big Data Εφαρμογές

<u>Κυβέρνηση κράτους</u>: Τα μεγάλα δεδομένα για τις κυβερνήσεις των κρατών προέρχονται από διάφορες πηγές όπως κάμερες κυκλοφορίας/CCTV, δορυφόρους, αισθητήρες, email και τα μέσα κοινωνικής δικτύωσης καθώς και τις μυστικές υπηρεσίες. Το περιεχόμενο των πληροφοριών είναι πολύτιμο για την βέλτιστη διακυβέρνηση και διαχείριση του δημόσιου τομέα. Τα κράτη διαθέτουν διαδικτυακά αποθετήρια open-data, όπου δημοσιεύονται για χρήση από τους πολίτες. Τα μέλη μιας χώρας έχουν την δυνατότητα είτε να διαβάσουν λεπτομερείς εκθέσεις σχετικά με τον τρόπο διακυβέρνησης, είτε να επεξεργαστούν τα ακατέργαστα δεδομένα, οπτικοποιώντας τα στατιστικά με την χρήση διαγραμμάτων και τα αποτελέσματα να δημοσιευτούν, με απώτερο σκοπό την αύξηση της συμμετοχής για την λήψη αποφάσεων τους που τους αφορά. Επιπρόσθετα, παγκοσμίως όλες οι κυβερνήσεις επενδύουν ετησίως τεράστια χρηματικά ποσά σε συστήματα Τεχνητής Νοημοσύνης και υποκείμενες πλατφόρμες cloudκαι δεδομένων για την θωράκιση έναντι σε οποιοδήποτε κίνδυνο. Η επεξεργασία αυτών των δεδομένων αποτελεί υψίστης σημασίας αφού μπορεί να γίνει λεπτομερής πρόβλεψη για ακραία καιρικά φαινόμενα και να γίνει έγκυρη ενημέρωση των πολιτών, ώστε να προστατευτούν κατά την διάρκεια της κακοκαιρίας. Μεγάλο πλήθος εφαρμογών των Big Data υλοποιήθηκε για στρατιωτική χρήση, όπως και το Internet,ώστε να εντοπίσουν ασυνήθιστες ενέργειες στα σύνορα ενός κράτος, ενεργοποιώντας άμεσα τους μηχανισμούς άμυνας. Τέλος, η ανάλυση δεδομένων από τις κάμερες κυκλοφορίας, συμβάλλοντας στην διαχείριση του δημόσιου οδικού δικτύου και την μείωση των παραβάσεων του κώδικα οδικής ασφάλειας. Ταυτόχρονα, η λεπτομερής ανάλυση δεδομένων χρήσης των Μέσων Μαζικής Μεταφοράς, βοηθά στην βέλτιστη διαχείριση των πόρων (καύσιμα ΜΜΜ, πλήθος υπαλλήλων), κατανοώντας τις μέρες και ώρες αιχμής κάθε διαδρομής.

Υγεία: Ο επιστημονικός τομέας της υγείας καθημερινά συλλέγει ένα τεράστιο πλήθος διαφόρων δεδομένων. Με την τεχνολογική πρόοδο πλέον τα δεδομένα δεν δημιουργούνται μόνο στους υγειονομικούς χώρους (νοσοκομεία, μικροβιολογικά εργαστήρια κ.τ.λ.), καθώς στα smart-watches υπάρχουν αισθητήρες, οι οποίοι μπορούν να καταγράψουν θερμοκρασία, καρδιακούς παλμούς, μυϊκούς σπασμούς και τις συνήθειες των χρηστών τους. Η λεπτομερή ανάλυση αυτών των δεδομένων εξυπηρετεί το ιατρικό προσωπικό, αφού μπορεί να κατανοήσει πλήρως την κατάσταση του κάθε ασθενή, εξατομικεύοντας την φαρμακευτική αγωγή. Ταυτόχρονα, τα δεδομένα κάθε ανθρώπου θα μπορούν να συσχετίζονται με τα ιατρικά δεδομένα της οικογενείας του, τα οποία παρέχουν πληροφορίες γονιδιακές παθήσεις που ίσως να εμφανιστούν στον μέλλον. Πέρα από την εξατομικευμένη θεραπεία, δημιουργούνται μοντέλα νοσημάτων, τα οποία θα μπορούν να προβλέπουν σοβαρές ασθένειες, όπως καρκίνος, AIDS, ακόμα και από ασθενείς που δεν εμφανίζουν κανένα σύμπτωμα. Η αναλυτική υγειονομικών δεδομένων απαλλάσσει την ταλαιπωρία των ασθενών από τις δοκιμές θεραπειών για την εύρεση της καταλληλότερη αγωγής, μειώνοντας τον χρόνο ίασης.

Εκπαίδευση: Ο κλάδος της εκπαίδευσης είναι πλημμυρισμένος από τεράστιο όγκο δεδομένων που σχετίζονται με εκπαιδευτικούς, εκπαιδευόμενους, μαθήματα, αποτελέσματα εξετάσεων κ.ά. Η ορθή μελέτη και ανάλυση αυτών των δεδομένων μπορεί να παρέχει πληροφορίες για βελτίωση αποτελεσματικότητας της λειτουργίας των εκπαιδευτικών ιδρυμάτων. Το διοικητικό προσωπικό των εκπαιδευτικών ιδρυμάτων θα έχουν πρόσβαση σε περισσότερες πληροφορίες, λαμβάνοντας αποφάσεις για τις μελλοντικές ανάγκες. Τα εκπαιδευτικά ιδρύματα αποτελούν ένας από τους πιο κοστοβόρους οργανισμούς και με την επεξεργασία των μεγάλων δεδομένων παρέχονται οι πληροφορίες για τα συνολικά έξοδα σε σχέση με τις ανάγκες τους. Τα αποτελέσματα του data analytics ίσως να οδηγήσει είτε σε καλύτερη αξιοποίηση ή την εμφάνιση ανάγκης για αύξηση των εσόδων είτε σε μείωση των σπαταλών του εκπαιδευτικού φορέα. Επιπρόσθετα, τα Big Data επιτρέπουν στο διδακτικό προσωπικό να προσαρμόζει τη δομή των μαθημάτων ανάλογα με τις ανάγκες των εκπαιδευόμενων και να εκσυγχρονίζει το περιεχόμενο της διδακτέας ύλης. Παράλληλα μπορούν να αξιοποιηθούν εργαλεία Μηχανικής Μάθησης, τα οποία στην αυτοματοποιημένη διαχείριση μάθησης. Τέλος, με την κατάλληλη ανάλυση των δεδομένων κάθε εκπαιδευόμενου, μπορεί να βοηθήσει στην κατανόηση της προόδου του και την εύρεση των αδυναμιών, των ενδιαφερόντων του, συμβάλλοντας στην ανάπτυξη στρατηγικών για εξατομικευμένη μάθηση από τους εκπαιδευτικούς.

Ενέργεια: Η ενέργεια είναι ένα πλεονέκτημα που δεν πρόκειται να διαρκέσει για πάντα. Πολλές τεχνολογικές εξελίξεις έχουν προωθήσει τους ανανεώσιμους και επαχρησιμοποιούμενους ενεργειακούς πόρους έναντι του πετρελαίου και του φυσικού αέριου, αλλά εξακολουθεί να υπάρχει η ανάγκη διατήρησης της περιορισμένης διαθέσιμης ενέργειας. Οι περισσότερες χώρες δεν είναι αυτόνομες στην παραγωγή ενέργειας. Η ανάλυση των μεγάλων δεδομένων μπορεί να χρησιμοποιηθεί για την δημιουργία νέων ευρημάτων σχετικά με τον τρόπο εκμετάλλευσης τους από τους δημόσιους οργανισμούς διαχείρισης και να προβλεφθεί η αποδοτικότητα τους σε ακραία καιρικά φαινόμενα. Το έξυπνο δίκτυο των σύγχρονων συστημάτων ηλεκτρικής ενέργειας

που κάνει αμφίδρομες τις ροές ενέργειας, επικοινωνίας και ελέγχου. Το δίκτυο ενεργειακής τροφοδοσίας χρησιμοποιείται για της παραγωγή ηλεκτρικής ενέργειας, διανομής και αξιοποίησης. Τέλος, οι κατασκευαστές δομών παραγωγής και διανομής ενέργειας έχουν πρόσβαση σε ιστορικά δεδομένα σε σχέση με τα ατυχήματα όπως διαρροή φυσικού αερίου σε δημόσιους δρόμους, έκρηξη εργοστασίων πυρηνικής ενέργειας, διαρροή καύσιμων από την εξόρυξη πετρελαίου. Αν αυτές οι πολύτιμες πληροφορίες διαχειριστούν καταλλήλως, θα μπορέσουν να ανοικοδομηθεί όλο ο απαραίτητος εξοπλισμός με σκοπό την ασφάλεια των εργαζομένων και του περιβάλλοντος. Οι τεχνικές ανάλυσης big data αποτελούν σημαντικό στοιχείο για την καλύτερη διαχείριση των διαθέσιμων ενεργειακών πόρων, τα οποία είναι ζωτικής σημασίας.

Γεωργία: Η γεωργία αποτελεί ένας από τους αρχαιότερους τομείς έρευνας και η φιλοσοφία των τρόπων καλλιέργειας μεταδίδεται από τη μια γενιά στην άλλη. Στην σύγχρονη εποχή υπάρχει επείγουσα ανάγκη να παραχθεί περισσότερη τροφή για τον αυξανόμενο πληθυσμό με λιγότερη καλλιεργήσιμη γη. Οι υπεύθυνοι του γεωργικού κλάδου αναζητούν βοήθεια στα τεχνολογικά επιτεύγματα- όπως IoT, ανάλυση μεγάλων δεδομένων και cloud computing–τα οποία τους παρέχουν την ικανότητα παρακολούθησης και πρόβλεψης φυσικών φαινομένων, συλλογή δεδομένων πραγματικού χρόνου που βρίσκονται στις καλλιεργήσιμες εκτάσεις και στα γεωργικά μηχανήματα. Τα Big Data παρέχουν στους αγρότες αναλυτικά δεδομένα σχετικά με τα μοτίβα βροχοπτώσεων, τις απαιτήσεις λιπασμάτων, την υγρασία του εδάφους κ.ά. που επικρατούν στις αγροτικές περιοχές, καθοδηγώντας τους στην επιλογή του καταλληλότερου φυτού και την αποδοτικότερη εκμετάλλευση της διαθέσιμης γης. Επιπρόσθετα, τα δεδομένα που εξάγονται από τους αισθητήρες του γεωργικού εξοπλισμού, αξιοποιούνται για την διαχείριση των εξόδων για καύσιμα, ηλεκτρική ενέργεια και για την εύρεση του χρήσιμων μηχανημάτων για κάθε συνθήκη. Τέλος, δυστυχώς η παραγωγή τροφίμων αρκετές φορές συσχετίζεται με ασθένειες και περιβαλλοντικές καταστροφές που επηρεάζουν εκατομμύρια ανθρώπους. Οι παγκόσμιοι περιβαλλοντικοί οργανισμοί αναλύουν τα δεδομένα και θέτουν αυστηρούς κανονισμούς για την χρήση χημικών φυτοφαρμάκων για την μείωση των μολύνσεων, την διατήρηση του οικοσυστήματος και την αύξηση της απόδοσης των καλλιεργειών.

Ναυτιλία: Στο ναυτιλιακό τομέα, από την αρχαιότητα, η σωστή παρακολούθηση του φορτίου διασφαλίζει την ασφάλεια του περιεχομένου των πλοίων. Τα μεγάλα δεδομένα χρησιμοποιούνται για την διαχείριση των αισθητήρων από τα πλοία εξυπηρετώντας την προγνωστική ανάλυση, η οποία είναι απαραίτητη για την αποφυγή καθυστερήσεων και τη βελτίωση της συνολικής λειτουργικής αποτελεσματικότητας του ναυτιλιακού κλάδου. Μέσω της επεξεργασίας των δεδομένων μπορούν να εξαχθούν οι βασικές αιτίες των απωλειών καραβιών και εμπορευμάτων εντός ή εκτός τερματικών σταθμών και τους τρόπους αποφυγής δαπανηρών προβλημάτων. Τα πλοία, όπως και τα αυτοκίνητα, μπορούν να λειτουργήσουν σε βέλτιστες ταχύτητες με μικρή κατανάλωση ενέργειας, όμως η αποδοτικότητα με την πάροδο του χρόνου μειώνεται εξαιτίας της φθοράς των μηχανικών μερών του κινητήρα. Επιπλέον, οι λιμενικοί χρειάζονται να γνωρίζουν πληροφορίες για την εκτιμώμενη ώρα άφιξης και το περιεχόμενο του

φορτίου. Οποιεσδήποτε αλλαγές στην ταχύτητα, στην ώρα αναχώρησης/άφιξης και στο βάρος του φορτίου παρακολουθούνται, αφού επηρεάζουν σημαντικά την κερδοφορία του κάθε εμπλεκόμενου οργανισμού. Η ανάλυση των δεδομένων επωφελή τους πλοιοκτήτες καθώς τους παρουσιάζονται αναλυτικά τα κόστη για καύσιμα και συντήρηση και οι περιβαλλοντικοί ρύποι, οδηγώντας τους στην λήψη αποφάσεων.

Αθλητισμός: Ο αθλητισμός πραγματοποιούσε στατιστική ανάλυση πολύ πριν εμφανιστεί ο επιστημονικός κλάδος της επιστήμης των δεδομένων. Εκατοντάδες εκατομμύρια αθλητικά δεδομένα παράγονται από διάφορες αθλητικές ομάδες. Οι εφαρμογές των Big Data αξιοποιούνται από τους αθλητικούς οργανισμούς και τις εταιρίες χρηματοδότησης των ομάδων, καθώς μπορούν να έχουν στην διάθεση τους τα στατιστικά στοιχεία για την απόδοση των παιχτών και να γνωρίζουν την κατάσταση της υγείας τους. Οι παίκτες κατανοούν τις αδυναμίες τους από τα δεδομένα της απόδοσής τους, τα οποία τους προσδιορίζουν τους στόχους της φυσικής κατάστασης για να έχουν με την μέγιστη απόδοση με την ελάχιστη κόπωση κατά την διάρκεια των αγώνων και την οικονομική τους αξία. Η μεγαλύτερη κερδοφορία στον αθλητικό τομέα εμφανίζεται από τα παιχνίδια στοιχημάτων. Οι θαυμαστές των αθλητικών αγώνων αναλύουν συνεχώς δεδομένα πραγματικού χρόνου, είτε πρόκειται για την παροχή ενημερώσεων play-by-play είτε για την συζήτηση προβλέψεων. Ο ιδρυτής της ιστοσελίδας «Advanced NFL Stars» Brian Burke [34], δηλώνει ότι χρησιμοποιώντας τα μεγάλα δεδομένα προπονητές και παίκτες μπορούν να προβλέψουν τα αποτελέσματα του αγώνα. Τέλος, αναπόσπαστο μέρος όλων των αθλητικών εκδηλώσεων αποτελούν οι επιβλέποντες/διαιτητές, οι οποίοι κατά την διάρκεια εκπαίδευσης τους έχουν πρόσβαση σε λάθη αθλητών και προβλήματα που δημιουργήθηκαν σε παλαιότερες αθλητικούς αγώνες. Αναπτύσσονται μοντέλα συσχέτισης παραβατικότητας και τρόπο επίλυσης για μη ομαλότερη διεξαγωγή των αθλητικών αγώνων.

2.6: Τι δεν προσφέρουν Big Data;

Τα δεδομένα που συλλέγονται μπορούν να ταξινομηθούν ή να παρουσιαστούν με την μορφή τυχαίων πληροφοριών. Εάν οι πληροφορίες είναι κατεστραμμένες και σημασιολογικά ανοργάνωτες, πολλοί χρήστες μπορεί να τις παρερμηνεύσουν εξάγοντας λανθασμένα συμπεράσματα. Παρά την σπουδαιότητα των μεγάλων δεδομένων, υπάρχουν μειονεκτήματα, τα οποία οι επιστήμονες προσπαθούν να εξαλείψουν.

Ασφάλεια(Security):Η αποθήκευση και η μεταφορά μεγάλων δεδομένων, ιδιαίτερα των ευαίσθητων αποτελεί ελκυστικό στόχο για κυβερνοεπιθέσεις. Στην έρευνα της AtScale το 2019, το 80% των ερωτηθέντων χρηστών του internet δήλωσαν πως αισθάνονται ανασφάλεια για τα προσωπικά τους δεδομένα που υπάρχουν στο internet. Με την αύξηση της παγκόσμιας πολιτικής κρίσης και τις περίπλοκες καταστάσεις μεταξύ των εθνών, τα δεδομένα που έχουν διαρρεύσει μπορούν να χρησιμοποιηθούν με δόλιο σκοπό. Επιπλέον, στον γενικότερο τομέα της επιστήμης

των υπολογιστών υπάρχει ανάγκη για συμμόρφωση με τους κυβερνητικούς κανονισμούς. Όλες οι πληροφορίες, που περιλαμβάνονται στα αποθήκες δεδομένων, πρέπει να διασφαλιστεί ότι πληρούν τα πρότυπα του κλάδου ή τις κρατικές απαιτήσεις κατά τον χειρισμό τους.

Οικονομικό Κόστος (Economic Cost):Τα μεγάλα δεδομένα μπορούν να εντοπίσουν τους αποτελεσματικότερους τρόπους για εξοικονόμηση επιχειρησιακών πόρων, αλλά ταυτόχρονα μπορούν να επιβαρύνουν οικονομικά την εταιρία/τον οργανισμό. Οι δαπάνες σχετίζονται με την εφαρμογή λογισμικού, τις τακτικές ενημερώσεις, την συντήρηση, την αποθήκη δεδομένων σε επίπεδο εξοπλισμού, την εκπαίδευση εργαζομένων στις νέες τεχνολογίες και στην πρόσληψη επιστημόνων δεδομένων για την διαχείριση των μεγάλων δεδομένων. Πολλά από τα εργαλεία διαχείρισης βασίζονται στην τεχνολογία ανοιχτού κώδικα, η οποία μειώνει το κόστος λογισμικού. Εξίσου κοστοβόρο είναι η ανάγκη σε υλικό εξοπλισμού για την δημιουργία αποθηκών δεδομένων(clusters,servers). Ορισμένοι οργανισμοί/επιχειρήσεις για να περιορίσουν αυτό το ζήτημα στρέφονται σε τεχνολογίες νέφους (cloud based technology) -όπως AWS- , αλλά αυτό δεν εξαλείφει εντελώς τα προβλήματα υποδομής. Δεν είναι ασυνήθιστο η διαχείριση των μεγάλων δεδομένων να υπερβαίνει τον προϋπολογισμό και τον απαιτούμενο χρόνο σε σχέση με άλλες διεργασίες ενός συστήματος.

Δυσκολία ενσωμάτωσης στα παλαιά συστήματα (Difficulty integrating legacy systems):Το λογισμικό που αναπτύσσεται θα πρέπει να είναι πάντα συμβατό με τις υπάρχοντες τεχνολογίες. Τα συστήματα που έχουν σχεδιαστεί πριν από 30~20 χρόνια δεν παρέχουν την δυνατότητα επέκτασης τους, μέσω της ενσωμάτωσης τους σε άλλα συστήματα. Τα παλαιού τύπου συστήματα τείνουν να είναι αργά, καθιστώντας τα ακατάλληλα για την διαχείριση του όγκου των μεγάλων δεδομένων. Ταυτόχρονα είναι ευάλωτα ως προς τις επιθέσεις κακόβουλων εισβολέων (hacker). Παρόλα αυτά, οι αποθηκευμένες πληροφορίες είναι πολύτιμες για την ανάλυση και τη λήψη αποφάσεων. Ο κλάδος των μεγάλων δεδομένων έχει εδραιωθεί σε όλους τους τομείς της τεχνολογίας τα τελευταία χρόνια, αναδεικνύοντας την ανάγκη για εκπαίδευση του υπάρχοντος προσωπικού που τις περισσότερες φορές δεν είναι εξοικειωμένο με τα νέα εργαλεία.

Κεφάλαιο 3: NoSQL

Τα Big Data έχουν υψηλές υπολογιστικές και αποθηκευτικές απαιτήσεις και τα περισσότερα δεδομένα που παράγονται δεν έχουν συγκεκριμένη μορφή και δομή. Τα παραδοσιακά μοντέλα διαχείρισης δεδομένων δεν καλύπτουν αυτές τις ανάγκες, δημιουργώντας διάφορα προβλήματα. Προκειμένου να καταπολεμήσουν αυτά τα προβλήματα, οι ερευνητές του κλάδου της επιστήμης των δεδομένων τροποποίησαν τις σχεσιακές βάσεις, αναπτύσσοντας τις μη-σχεσιακές βάσεις,NoSQL.

Εικόνα 4: NoSQL

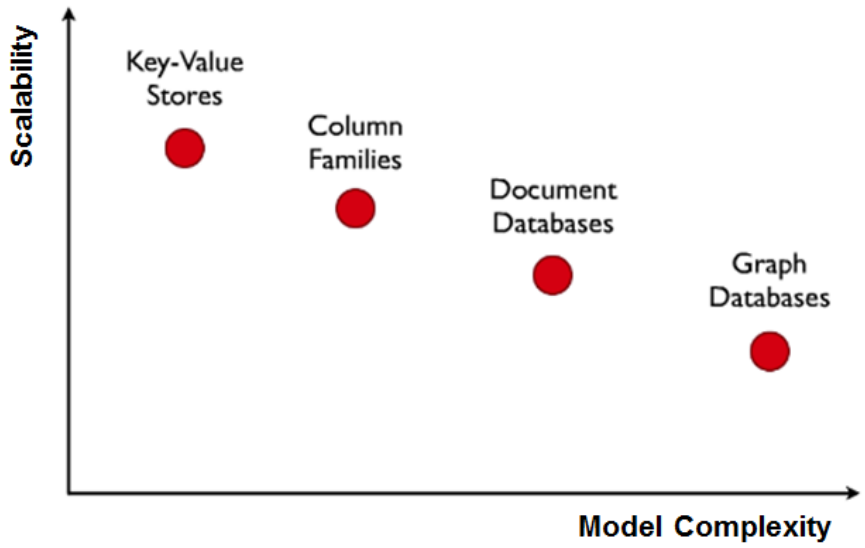
3.1: Τι είναι NoSQL;

Πολλοί επιστήμονες θεωρούν την NoSQL -NotOnlySQL- ως την σύγχρονη βάση δεδομένων, διότι προσαρμόζεται στις απαιτήσεις του Παγκόσμιου Ιστού και του Διαδικτύου των πραγμάτων. Οι NoSQL βάσεις δεδομένων είναι μια μη-σχεσιακή βάση δεδομένων, αποθηκεύοντας και αποκτώντας πρόσβαση σε δεδομένα με την χρήση κλειδιών-τιμών (***key-value***). Το κάθε δεδομένο αποθηκεύεται **ξεχωριστά** με ένα **μοναδικό** κλειδί, παρέχοντας ευελιξία στην αποθήκευση.

Οι NoSQL ανήκουν στην κατηγορία συστημάτων διαχείρισης DB που δεν ακολουθούν όλους του κανόνες των RDBMS, χωρίς να αξιοποιούν την SQL. Για την υποστήριξη των παλαιών συστημάτων που βασίζονται σε RDBMS και την ομαλή μετάβαση των συστημάτων από τις παραδοσιακές τεχνικές διαχείρισης στις μη-σχεσιακές, τα NoSQL συστήματα υποστηρίζουν γλώσσες ερωτημάτων που μοιάζουν με την SQL. Δεδομένου ότι για τις NRDBMS δεν είναι

απαραίτητη προϋπόθεση η ύπαρξη προκαθορισμένου σχεδίου (**schema database**), προσφέρεται η δυνατότητα για ταχεία κλιμάκωση διαχείρισης μεγάλων δεδομένων.

3.2: Είδη NoSQL βάσεων δεδομένων



Πίνακας 3: Σύγκριση των Μοντέλων δεδομένων NoSQL ως προς την πολυπλοκότητα και την κλιμάκωση

Κλειδί-τιμή(Key-Value stores):Οι βάσεις δεδομένων κλειδιού-τιμής αποτελούν την απλούστερη μορφή δεδομένων. Κάθε στοιχείο ταξινομείται με ένα μοναδικό σημασιολογικό ζεύγος key-value. Αυτός ο τύπος NoSQL DB έχει μια δομή δεδομένων λεξικού, που αποτελείται από ένα σύνολο αντικειμένων και αντιπροσωπεύουν πεδία δεδομένων. Η χρήση του κλειδιού μπορεί να παρομοιαστεί το primary key που υπάρχει στις RDBMS, ενώ η τιμή είναι μια σειρά απλών τύπων δεδομένων, όπως συμβολοσειρές. Η ανάκτηση των αποθηκευμένων δεδομένων υλοποιείται με την χρήση συγκεκριμένου κλειδιού και στην συνέχεια λαμβάνεται η τιμή που έχει εκχωρηθεί στο κλειδί. Χάρη στην απλότητα τους, εμφανίζουν μεγαλύτερη ευελιξία ως προς την αποθήκευση και ταχύτερη απόδοση, επιτρέποντας την οριζόντια κλιμάκωση μεγάλων ποσοτήτων δεδομένων. Ωστόσο, δεν είναι ιδανικό για ταυτόχρονη εξόρυξη δεδομένων. Οι περιπτώσεις χρήσης των βάσεων δεδομένων κλειδιού-τιμής περιλαμβάνουν καλάθια ηλεκτρονικών αγορών, προτιμήσεις χρηστών και προφίλ χρηστών σε όλες τις ηλεκτρονικές πλατφόρμες.

Ευρείας στήλης(Wide-Column families): Οι βάσεις δεδομένων ευρείας στήλης αποθηκεύουν τις πληροφορίες σε στήλες, ενώ στις RDBMS τα δεδομένα αποθηκεύονται σε σειρές και διαβάζονται πλειάδα προς πλειάδα. Οι NoSQL τύπου ευρείας στήλης έχουν σχεδιαστεί με σκοπό την αποτελεσματικότερη ανάγνωση δεδομένων και την ταχύτερη ανάκτηση τους. Κατά την διαδικασία εφαρμογής προγραμμάτων για ανάλυση των δεδομένων, δεν καταναλώνεται μνήμη

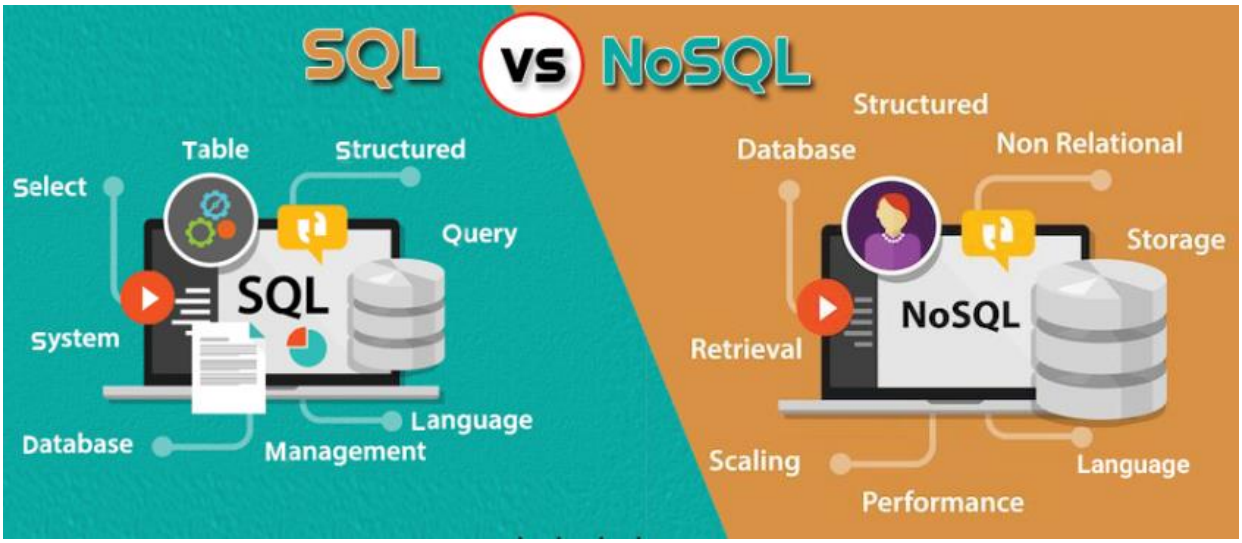
για τις μη-χρήσιμες και ανεπιθύμητες πληροφορίες. Επιπλέον, στα κατανεμημένα συστήματα οι στήλες που δεν υποβάλλονται συχνά σε ερωτήματα κατανέμονται σε διαφορετικούς και απομακρυσμένους κόμβους, ώστε να μην υπερφορτώνεται ο κεντρικός κόμβος (διακομιστής).Το Apache Cassandra είναι παράδειγμα βάσεων δεδομένων ανοιχτού κώδικα, ευρείας στήλης και είναι σχεδιασμένο να διαχειρίζεται μεγάλες ποσότητες δεδομένων σε πολλούς διακομιστές και ομαδοποίηση που εκτείνεται σε κέντρα πολλαπλών δεδομένων.

Εγγραφή (Document databases):Οι βάσεις δεδομένων εγγράφων αποθηκεύουν τα δεδομένα σε έγγραφα JSON, BSON ή XML. Τα έγγραφα είναι ένθετα και μπορούν να ταξινομηθούν για ταχεία αναζήτηση. Τα δεδομένα διατηρούνται ενωμένα όταν χρησιμοποιούνται σε εφαρμογές, μειώνοντας τον όγκο μετάφρασης που απαιτείται για την χρήση τους. Η βάση δεδομένων εγγράφων προσφέρει την ευελιξία ως προς το schema, καθώς δεν χρειάζεται να υπάρχει ομοιότητα στις μεταβλητές μεταξύ των εγγράφων. Παρόλα αυτά, η ανομοιότητα μεταξύ των εγγράφων μπορεί να δημιουργήσει μείζονα ζήτημα στις περιπτώσεις των πολύπλοκων συναλλαγών, προκαλώντας καταστροφή των πολύτιμων πληροφοριών.

Γραφήματα(Graph databases): Οι βάσεις δεδομένων γραφήματος εστιάζουν στην σχέση μεταξύ των στοιχείων δεδομένων. Κάθε στοιχείο αποθηκεύεται ως κόμβος, ακμή και ιδιότητα. Οι κόμβοι συνδέονται μεταξύ τους με τις ακμές, προσδιορίζοντας τις συσχετίσεις μεταξύ των δεδομένων. Μια βάση δεδομένων γραφημάτων είναι βελτιστοποιημένη για να καταγράφει και να αναζητά τις συνδέσεις μεταξύ των στοιχείων δεδομένων, ξεπερνώντας τις δυσκολίες από την χρήση του JOINσε πολλαπλούς πίνακες στην SQL. Ως αποτέλεσμα, όλες οι διεργασίες που πραγματοποιούνται στις βάσεις δεδομένων γραφήματος εκτελούνται παράλληλα.

3.3: NoSQL VS RDBMS

Οι NoSQL βασίζονται στις SQL βάσεις δεδομένων, αλλά διαφοροποιηθεί με την προσθήκη χαρακτηριστικών, τα οποία περιλαμβάνουν την έλλειψη schema database, την ομαδοποίηση δεδομένων και την υποστήριξη διπλότυπων.



Εικόνα 3: SQL VS NoSQL

Ακεραιότητα των δεδομένων (Data Integrity):Οι βάσεις δεδομένων SQL και NoSQL χρησιμοποιούν διαφορετικές προσεγγίσεις για την προστασία της ακεραιότητας των δεδομένων, καθώς δημιουργούνται, διαβάζονται, ενημερώνονται και διαγράφονται από διάφορους χρήστες και εφαρμογές. Οι SQL βάσεις δεδομένων βασίζονται στο μοντέλο ACID. Το μοντέλο βασίζεται σε τέσσερις βασικούς πυλώνες: Ατομικότητα(Atomicity), Συνοχή(Consistency), Απομόνωση(Isolation) και Ανθεκτικότητα(Durability). Κάθε διεργασία στον server: (α) εκτελείται μόνη της είτε επιτυχημένα είτε αποτυχημένα,(β) χωρίς να παραβιάζει τους περιορισμούς που έχουν ορισθεί για το σύνολο των δεδομένων και (γ) αποκρύπτεται από τις άλλες διεργασίες μέχρι να ολοκληρωθεί η εκτέλεσή της. Οι αλλαγές που θα προκύψουν στα δεδομένα διατηρούνται στον server. Αντίθετα, οι NoSQL βάσεις δεδομένων βασίζονται στο μοντέλο BASE. Το μοντέλο αποτελείται από τρεις βασικές ιδιότητες: Βασική Διαθεσιμότητα (Basic Availability), Μεταβλητότητα Συνοχής(Soft-State) και Ενδεχόμενη Συνέπεια(Eventual Consistency). Τα δεδομένα: (α) τις περισσότερες φορές είναι διαθέσιμα, (β) αναπαράγονται και διαδίδονται συνεχώς, προκαλώντας μερική ασυνέχεια στην ροή των δεδομένων για μικρό χρονικό διάστημα και (γ) η συνέπεια θα επανέλθει σε άγνωστη στιγμή. Ορισμένες εφαρμογές παρουσιάζουν ανοχή ως προς την ασυνέπεια των δεδομένων.

Σχηματισμός(Schema): Μία από τις σημαντικότερες διαφορές μεταξύ SQL και NoSQL βάσεων δεδομένων αποτελεί ο τρόπος σχεδιασμός μιας DB. Οι NoSQL βάσεις δεδομένων είτε δεν ακολουθούν κάποιο προκαθορισμένο σχεσιακό μοντέλο, είτε διαθέτουν δυναμικό σχεσιακό μοντέλο. Η έλλειψη ή χαλαρότητα στο schema, εξυπηρετεί τους τύπους δεδομένων στις NRDBMS. Οι NoSQL DB είναι εύκαμπτες ως προς τις σχεσιακές αλλαγές. Αντίθετα, οι SQL βάσεις δεδομένων λειτουργούν σύμφωνα με ένα αυστηρά προκαθορισμένο σχεσιακό μοντέλο. Το ίδιο αυστηρό και προκαθορισμένο schema υποχρεωτικά διατηρούν τα δεδομένα. Οι SQL DB

είναι άκαμπτες στις πιθανές αλλαγές του schema, απαιτώντας λεπτομερή ανάλυση στις ανάγκες ενός συστήματος πριν το σχεδιασμό της βάσης δεδομένων του.

Επεκτασιμότητα(Scalability): Σημαντική διαφορά αποτελεί ο τρόπος επέκτασης μιας βάσης δεδομένων, διότι τα δεδομένα αυξάνονται με εκθετικούς ρυθμούς. Οι περισσότερες SQL βάσεις δεδομένων, χρησιμοποιούν έναν μοναδικό server για την αποθήκευση κάθε DB, υποστηρίζοντας την κάθετη κλιμάκωση. Αυτή η κλιμάκωση επιβαρύνει είτε τον server με την αύξηση των διεργασιών που έχει να εκτελέσει, μειώνοντας την αποδοτικότητά του είτε οικονομικά τον οργανισμό/εταιρεία χρήσης της βάσης δεδομένων, με την προσθήκη ενός νέου server ή και αντικατάσταση του υπάρχοντος με μεγαλύτερες δυνατότητες. Αντίθετα, οι βάσεις δεδομένων NoSQL χρησιμοποιούν ένα σύστημα κατανεμημένων servers διαφορετικής τεχνολογίας ο καθένας, υποστηρίζοντας την οριζόντια κλιμάκωση. Αυτή η κλιμάκωση προσφέρει τη δυνατότητα διάσπασης των συνόλων δεδομένων σε μικρότερους servers, βελτιώνοντας την διαχείρισης μεγάλων ποσοτήτων δεδομένων.

Ανίχνευση απάτης και έλεγχος ταυτότητας: Η προστασία προσωπικών δεδομένων και η διασφάλιση πρόσβασης στις εφαρμογές μόνο πραγματικών χρηστών/πελατών αποτελεί την κορυφαία προτεραιότητα όλων των εταιρειών ανάπτυξης λογισμικού. Ιδιαιτέρως στις χρηματοοικονομικές και υγειονομικές υπηρεσίες. Οι hackers συνεχώς αναζητούν τρόπους να παραβιάσουν τις δικλίδες ασφαλείας, ώστε να μπορέσουν να έχουν πρόσβαση σε αυτές τις ευαίσθητες και πολύτιμες πληροφορίες. Για την ανίχνευση παραπλανητικών πράξεων, απαιτείται συνεχή ανάλυση δεδομένων σε πραγματικό χρόνο όλων των τύπων δεδομένων, ώστε να εντοπιστούν ασυνήθιστες ενέργειες των χρηστών του συστήματος. Αυτές οι ανωμαλίες μπορούν να ανιχνευθούν ακόμα και πριν συμβεί οποιαδήποτε απάτη. Ο συνδυασμός της αυτοματοποιημένης ανάλυσης σε πραγματικό χρόνο, των μεγάλων και συνεχώς αυξανόμενων συνόλων δεδομένων, των πολυάριθμων τύπων data, μαζί με την διεξαγωγή μοντέλων μηχανικής μάθησης και τεχνητής νοημοσύνης, καθιστούν τις NoSQL βάσεις ιδανικές για τον εντοπισμό απατών και πιστοποιήσεων ταυτότητας χρηστών.

3.4: ΕφαρμογέςNoSQL

Διαχείριση ηλεκτρονικού περιεχομένου (Content Management): Οι διαδικτυακές αγορές ξεπερνούν τις «φυσικές» πωλήσεις και το οπτικοακουστικό υλικό κυριαρχεί σε χιλιάδες ηλεκτρονικές αγορές και στις βιτρίνες των ιστοσελίδων. Οι εταιρίες διαδικτυακών πωλήσεων αξιοποιούν τα πολυμεσικά εργαλεία συμπεριλαμβανομένου του υλικού (φωτογραφία, βίντεο, κριτική του προϊόντος) που δημιουργείται από χρήστες στα μέσα κοινωνικής δικτύωσης, προσδίδοντας την ψευδαίσθηση της στιγμιαίας αλληλεπίδρασης με ένα προϊόν/υπηρεσία στους μελλοντικούς αγοραστές. Οι NoSQL document βάσεις προσφέρουν ένα ευέλικτο, ανοιχτού τύπου μοντέλο δεδομένων το οποίο είναι ιδανικό για την αποθήκευση ενός κράματος όλων των

τύπων δεδομένων. Επιπλέον, καθίσταται η δυνατότητα για συγκέντρωση δεδομένων που εξυπηρετούν πολλαπλές επιχειρηματικές εφαρμογές σε μια ενιαία βάση δεδομένων καταλόγου. Ενώ οι σχεσιακές βάσης δεδομένων διαχείρισης με τα σταθερά μοντέλα τείνει να έχει ως αποτέλεσμα τον πολλαπλασιασμό πολλαπλών, επικαλυπτόμενων καταλόγων για διαφορετικούς σκοπούς. Χαρακτηριστικό παράδειγμα του Content Management με NoSQL αποτελεί ο δημοσιογραφικός κολοσσός Forbes [44].Το Forbes ανέπτυξε ταχύτατα ένα προσαρμοσμένο σύστημα διαχείρισης περιεχομένου βασισμένο στο MongoDB μέσα σε λίγους μήνες παρέχοντας μεγαλύτερη ευελιξία με χαμηλότερο κόστος. Οι οικονομικοί πόροι της εταιρείας προέρχονται από τις προβολές των άρθρων, τις διαφημίσεις και την ενσωμάτωση του περιεχομένου συνεργατών και τον διαμοιρασμό (Share→clickstream) στα social media.

Εφαρμογές για κινητά (Mobile applications): Η χρήση κινητών τηλεφώνων και tablet ξεπέρασε τους ηλεκτρονικούς υπολογιστές ως η κορυφαία διαδικτυακή πλατφόρμα για αναζήτηση πληροφοριών, αγορές και προβολή περιεχομένου ιστοσελίδων. Ενδιαφέρον αποτελεί το γεγονός ότι το 90% των δεδομένων κινητής τηλεφωνίας εξυπηρετείται μέσω εφαρμογών και μόνο το 10% μέσω φυλλομετρητών (Browser apps) μια συντριπτική αλλαγή τα τελευταία χρόνια [Google statics]. Η ταχεία κλιμάκωση mobile-apps παγκοσμίως για την εξυπηρέτηση δεκάδων εκατομμυρίων χρηστών απαιτεί συχνά κατανεμημένες βάσεις δεδομένων, οι οποίες με την σειρά τους απαιτούν την υποστήριξη από τις τεχνολογίες NoSQL. Τα ευέλικτα μοντέλα των δεδομένων NoSQL έχουν αντοχή στους γρήγορους κύκλους ενημέρωσης εφαρμογών καλύτερα από τα μοντέλα σχεσιακών δεδομένων σε πολλές περιπτώσεις. Πλέον οι περισσότερες επιχειρήσεις επιθυμούν να αυξήσουν τα έσοδα τους από το περιεχόμενο ιστότοπων, χρησιμοποιώντας NoSQL data stores για τις εφαρμογές τους. Η mobile εφαρμογή The Weather Channel αποτελεί το βέλτιστο παράδειγμα χρήση NoSQL βάσεων [44]. Η συγκεκριμένη εφαρμογή είναι προ-εγκατεστημένη στις περισσότερες συσκευές κινητής τηλεφωνίας και tablet και χρησιμοποιείται κυρίως παθητικά από τους χρήστες καταναλώνοντας ελάχιστους ενεργειακούς πόρους. Το The Weather Channel διαχειρίζεται εκατομμύρια αιτήματα ανά λεπτό, ενώ παράλληλα επεξεργάζεται δεδομένα χρηστών και καταφέρνει να υλοποιήσει ενημερώσεις καιρού από δεκάδες χιλιάδες τοποθεσίες σε όλο τον κόσμο.

Εμπλουτισμός ψηφιακής εμπειρίας (Digital Customer Experience): Μια συναρπαστική διαφοροποιημένη ψηφιακή εμπειρία βασιζόμενη στις δυνατότητες υψηλής έντασης δεδομένων, κρίσιμες για το χρόνο, όπως η εξατομίκευση, η διαχείριση προφίλ χρήστη και μια ενοποιημένη άποψη του πελάτη σε όλα τα σημεία επαφής με την εταιρεία που παρέχει ένα προϊόν ή μια ηλεκτρονική υπηρεσία. Πολλά δημογραφικά, συμπεριφοριστικά και υλικοτεχνικά δεδομένα προέρχονται από διαδικτυακά clickstreams, δημιουργώντας ένα φόρτο εργασίας πολλών schema εγγραφής που δυσκολεύουν την λειτουργία των σχεσιακών βάσεων διαχείρισης δεδομένων. Μια κατανεμημένη βάση δεδομένων NoSQL μπορεί να κλιμακωθεί με χαμηλό οικονομικό κόστος, να διαχειριστεί ένα συνεχώς αυξανόμενο αριθμό χαρακτηριστικών με λιγότερη διοικητική ταλαιπωρία και χωρίς να υπάρχουν χρονικές καθυστερήσεις. Ο παγκόσμιος πάροχος πολυμεσικών υπηρεσιών Comcast χρησιμοποιεί μια πλατφόρμα Couch base NoSQL για να

προσφέρει θετική εμπειρία υποστήριξης πελατών σε πολλαπλούς τομείς δραστηριότητας. Η πλατφόρμα καταγράφει δεδομένα από άπειρους αριθμούς καναλιών αλληλεπίδρασης και τα συσχετίζει με τους λογαριασμούς και την κατάσταση της υπηρεσίας μεμονωμένων πελατών, εμφανίζοντας ανάγκη για συνεχή επεκτασιμότητα και για ανθεκτικότητα.

3.5: Τι δεν προσφέρουν;

Οι NoSQL βάσεις δεδομένων αναπτύχθηκαν στις αρχές του 21ου αιώνα και είναι αρκετά νεότερες σε σχέση με τις SQL. Οι μη σχεσιακές βάσεις δεδομένων αποτελούν μία από τις μεγαλύτερες καινοτομίες του κλάδου της επιστήμης των δεδομένων, προσφέροντας ευελιξία ως προς την επεκτασιμότητα. Δυστυχώς, ακόμα δεν υπάρχει ολοκληρωμένη βιβλιογραφική τεκμηρίωση και παρατηρούνται αρκετά σημεία που χρήζουν βελτίωσης ώστε να χρησιμοποιούνται ευρέως [19].

Μη ύπαρξη τυποποιημένης γλώσσας(No standardized language): Δεν υπάρχει τυπική γλώσσα για την διενέργεια ερωτημάτων NoSQL. Η σύνταξη των ερωτημάτων διαφοροποιείται ανάλογα με τον τύπο των δεδομένων και δυσκολεύει την ευρεία χρήση της. Η ευελιξία των NoSQL ερωτημάτων είναι λιγότερη σε σχέση με την αποθήκευση τους. Συνήθως, δεν μπορούν να επιβάλουν ή και να εγγυηθούν την μοναδικότητα των κλειδιών εντός της βάσεις δεδομένων, όπως τα σχεσιακά συστήματα διαχείρισης δεδομένων. Η αξιοποίηση NRDBMS καθίσταται αναποτελεσματική σε εφαρμογές, όπου προβλέπεται αυστηρά η χρήση μοναδικών κλειδιών και τιμών.

Ασυνέπεια ανάκτησης δεδομένων(Data retrieval inconsistency): Οι NoSQL βάσεις δεδομένων αξιοποιούν τους κατανεμημένους servers, για ταχεία διαθεσιμότητα. Ταυτόχρονα, δυσχεραίνεται η συνέπεια των δεδομένων, αυξάνοντας τις πιθανότητες για ανολοκλήρωτα αποτελέσματα από τα ερωτήματα που δέχεται η βάση δεδομένων. Επίσης, ίσως να επιστρέφονται διαφορετικά αποτελέσματα ανάλογα των server που θα είναι διαθέσιμος για ανταπόκριση του ερωτήματος. Όπως προαναφέρθηκε, οι NoSQL βασίζονται στο πρότυποBASE, όπου περιέχει **ενδεχόμενη συνέπεια**. Όμως, οι SQL θεμελιώνονται με το μοντέλο ACID, στο οποίο τα δεδομένα πρέπει να είναι έγκυρα και συνεπή κατά την διάρκεια των διεργασιών στις βάσεις δεδομένων. Η ασυνέπεια, στις εφαρμογές εμφανίζεται για μερικά νανοδευτερόλεπτα και δεν είναι αντιληπτή από τον χρήστη.

Ασφάλεια (Security): Όλες οι βάσεις δεδομένων διατηρούν αντίγραφα όλων των δεδομένων και των διεργασιών που έχουν συμβεί, ακόμα και αν κάποιος χρήστης/διαχειριστής/προγραμματιστής τα έχει χειροκίνητα διαγράψει. Δυστυχώς, οι NoSQL βάσεις δεδομένων δεν επιτρέπουν την δημιουργία αντιγράφων. Τα τελευταία χρόνια, τα

εργαλεία NoSQL παρέχουν κάποια εργαλεία για την δημιουργία αντιγράφων ασφαλείας, αλλά δεν είναι αρκετά ώριμα για να εξασφαλισθεί η καταλληλότερη και η ασφαλέστερη διαχείριση της βάσης δεδομένων.

Κεφάλαιο 4: Apache Cassandra

Εικόνα 5: Λογότυπο Apache Cassandra

4.1 Τι είναι το Apache Cassandra;

Το εργαλείο Apache Cassandra ανήκει στην οικογένεια Ευρειών Στηλών των NoSQL βάσεων δεδομένων ανοιχτού κώδικα. Έχει σχεδιαστεί για τον χειρισμό Big Data που κατανέμονται σε πολλούς servers. Με την χρήση του Cassandra διασφαλίζεται η διαρκής διαθεσιμότητα ενός συστήματος και ελάχιστα σημεία αποτυχίας, εξαιτίας της peer-to-peer αρχιτεκτονικής, της οριζόντιας επεκτασιμότητας, της ανοχής σε σφάλματα και τους ταχύτατους χρόνους απόκρισης.

Το Cassandra αναπτύχθηκε το 2008 από τον Avinash Lakshman που εργαζόταν στην εταιρία Metaverse (πρώην Facebook), με σκοπό η πλατφόρμα κοινωνικής δικτύωσης Facebook να μπορεί να είναι πάντα διαθέσιμη και να επεκτείνεται χωρίς να προκαλεί πρόβλημα στο προ-υπάρχον σύστημα. Το 2009 ο Lakshman σε συνεργασία με το Prashant Malik εκδώσαν επιστημονικό άρθρο με τις αρχές του Cassandra. Το Μάρτιο 2009, το εργαλείο Cassandra εξαγοράστηκε και περιλαμβάνεται στα project του ιδρύματος Apache. Μέχρι το 2022 το Cassandra, χρησιμοποιείται από μεγάλο πλήθος εταιριών όπως: Cern, Wood Hole Oceanography Institution, IBM, Metaverse, Netflix κτλ.

4.2 Χαρακτηριστικά

Το εργαλείο Apache Cassandra εμπεριέχει τα βασικά χαρακτηριστικά των NoSQL βάσεων δεδομένων. Όπως όλες οι NoSQL DB είτε δεν ακολουθούν κάποιο προκαθορισμένο schema, είτε διαθέτουν σχεσιακό τμήμα. Το Cassandra βασιζόμενο στις βασικές αρχές των μη σχεσιακών βάσεων δεδομένων, έχει αναπτύξει μηχανισμούς οι οποίοι επεκτείνουν τις δυνατότητες αυτών των αρχών.

Αποκέντρωση (Decentralization): Όπως προαναφέρθηκε το Cassandra είναι κατανεμημένο σύστημα, οι κόμβοι του έχουν τους ίδιους ρόλους. Κάθε κόμβος εμπεριέχει διαφορετικά

δεδομένα, αφού τα σύνολα των δεδομένων κατανέμονται σε όλο το σύμπλεγμα κόμβων. Με την αποκέντρωση του Cassandra διευκολύνεται η λειτουργία του συστήματος και δεν εμφανίζονται σημεία αποτυχίας εκτέλεσης διεργασιών αυξάνοντας την αποδοτικότητά του.

Επεκτασιμότητα (Scalability):Στις NoSQL βάσεις δεδομένων, το σύστημα επεκτείνεται οριζόντια αυξομειώνοντας την κλίμακα του προσθέτοντας ή αφαιρώντας servers. Το λογισμικό του Cassandra περιλαμβάνει έναν εσωτερικό μηχανισμό για να διατηρεί τα δεδομένα του συγχρονισμένα και όταν αντιληφθεί την ύπαρξη ενός νέου server αυτομάτως τροποποιεί το υπάρχον σύστημα για βέλτιστη λειτουργικότητα.

Ανοχή σε σφάλματα (Fault-tolerant): Τα συστήματα πραγματικού χρόνου πρέπει να είναι συνεχώς διαθέσιμα προς χρήση ακόμα και αν παρουσιάζουν σφάλματα. Το Cassandra μπορεί να αντικαταστήσει τους κόμβους που εμφανίζονται σφάλματα, χωρίς να υπάρξει κάποιο χρονικό διάστημα μη λειτουργίας του συστήματος. Εξαιτίας της κατανεμημένης φύσης του εργαλείου αναπαράγει τα δεδομένα του αποτυχημένου κόμβου σε άλλους κόμβους, αποφεύγοντας την διακοπή της λειτουργίας του και ταυτόχρονα διατηρείται η αποδοτικότητά του.

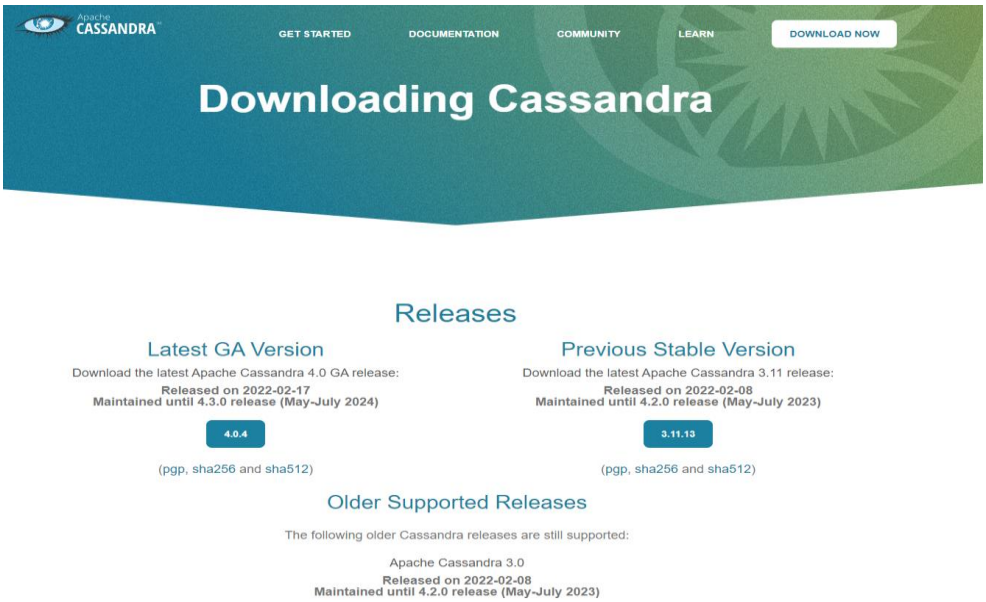
Ρυθμιζόμενη Συνέπεια (Tunable consistency): Οι NoSQL βάσεις δεδομένων, όπως αναφέρθηκε, βασίζονται στο μοντέλο BASE για την διασφάλιση της ακεραιότητας των δεδομένων τους. Ο τελευταίος βασικός πυλώνας αυτού του μοντέλου αφορά στην Ενδεχόμενη Συνέπεια. Ο όρος συνέπεια αναφέρεται ότι κατά την διαδικασία του «ανάγνωσης» μιας πληροφορίας, θα πρέπει να επιστρέφεται η πιο πρόσφατη τιμή που εγγράφθηκε για αυτή την πληροφορία. Το εργαλείο Cassandra επιτρέπει τον προσδιορισμό του μέτρου της συνέπειας, που είναι αναγκαίο για το κάθε σύστημα, εξισορροπώντας το με το επίπεδο διαθεσιμότητας. Μπορούν να προσδιοριστούν τρία επίπεδα συνέπειας: **αυστηρή** (strict), **αιτιώδης** (casual) και **αδύναμη** (weak). Η αυστηρή συνέπεια απαιτεί από το σύστημα σε κάθε «ανάγνωση» να επιστρέφεται πάντα η πιο πρόσφατη «εγγραφή». Αυτό σημαίνει ότι οι κόμβοι θα πρέπει να είναι απόλυτα συγχρονισμένοι για να ανταποκριθούν στις απαιτήσεις του συστήματος. Η αιτιώδης συνέπεια εφαρμόζεται στις περιπτώσεις που υπάρχει ανάγκη για σημασιολογική ακολουθία εγγραφών, ώστε να αναγνωσθούν σε λογική σειρά. Στην αδύναμη συνέπεια όλες οι πιο πρόσφατες εγγραφές θα διαδοθούν σε όλο το κατανεμημένο σύστημα με σχετικά μεγάλη χρονική καθυστέρηση.

Cassandra Query Language (CQL): Αρχικά το εργαλείο Cassandra απαιτούσε την χρήση ενός API, το οποίο εκτελούσε Εισαγωγή, Λήψη και Διαγραφή των δεδομένων από την βάση. Όσο η χρήση του Cassandra αυξανόταν, δημιουργήθηκε η γλώσσα CQL. Η CQL είναι γλώσσα ερωτημάτων, όπως και η SQL, χωρίς την δυνατότητα σύνδεσης πολλών πινάκων. Σε κάθε κόμβο του συστήματος ορίζεται μια σταθερά σημασιολογική λίστα key-value. Η σύνταξη της λίστας απαιτεί ιδιαίτερη προσοχή, καθώς η CQL είναι case insensitive.

4.3Εγκατάσταση σε Windows10

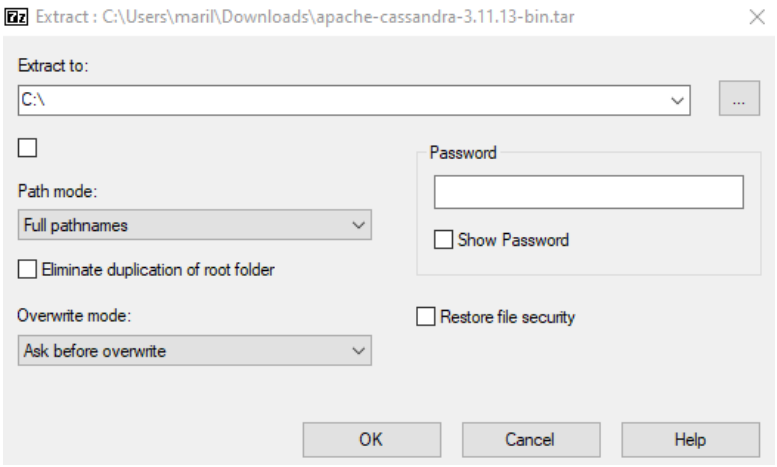
Το εργαλείο Apache Cassandra απαιτεί 2 βασικές προϋποθέσεις για την εγκατάσταση του: Java 8 και Python 2.7.

Βήμα 1: Κατεβάζουμε στον υπολογιστή την έκδοση Cassandra 3.11.18 ή οποιαδήποτε έκδοση προτείνει το website ως σταθερή.



Εικόνα 6:Επιλογή και αποθήκευση κατάλληλου λογισμικού από το επίσημο website

Βήμα 2: Τοποθέτηση του αρχείου στον root φάκελο του υπολογιστή



Εικόνα 7: Επιλογή root φακέλου

Βήμα 3: Διόρθωση σφαλμάτων που παρουσιάζονται.

Για τον έλεγχο της ορθής εγκατάστασης του Apache Cassandra, πρέπει να εκτελεσθεί η εντολή cassandraστο φάκελο C:\apache-cassandra-3.11.13\bin , χρησιμοποιώντας το Command Prompt. Σε κάθε υπολογιστή διαφέρουν τα σφάλματα που θα προκύψουν.

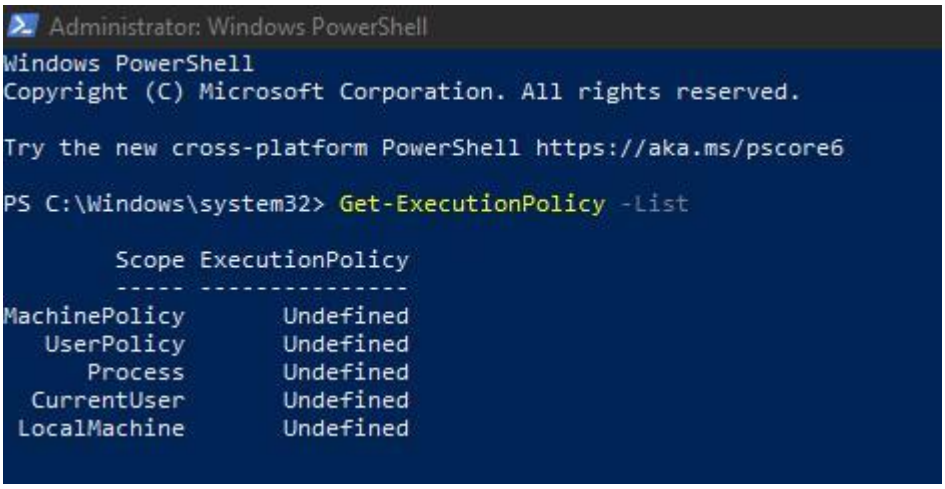
1. Τροποποίηση δικαιωμάτων με την χρήση του PowerShell

```
:\apache-cassandra-3.11.13\bin>cassandra
ARNING! Powershell script execution unavailable.
  Please use 'powershell Set-ExecutionPolicy Unrestricted'
  on this user-account to run cassandra with fully featured
  functionality on this platform.
tarting with legacy startup options
tarting Cassandra Server
NFO  [main] 2022-05-17 14:48:24,666 YamlConfigurationLoader.java:93 - Configuration location: file:/C:/apache-cassandra
3.11.13/conf/cassandra.yaml
NFO  [main] 2022-05-17 14:48:25,570 Config.java:555 - Node configuration:[allocate_tokens_for_keyspace=null; allow_extr
_insecure_udfs=false; allow_insecure_udfs=false; authenticator=AllowAllAuthenticator; authorizer=AllowAllAuthorizer; au
o_bootstrap=true; auto_snapshot=true; back_pressure_enabled=false; back_pressure_strategy=org.apache.cassandra.net.Rate
asedBackPressure{high_ratio=0.9, factor=5, flow=FAST}; batch_size_fail_threshold_in_kb=50; batch_size_warn_threshold_in
kb=5; batchlog_replay_throttle_in_kb=1024; broadcast_address=null; broadcast_rpc_address=null; buffer_pool_use_heap_if_
xhausted=true; cache_load_timeout_seconds=30; cas_contention_timeout_in_ms=1000; cdc_enabled=false; cdc_free_space_chec
_interval_ms=250; cdc_raw_directory=null; cdc_total_space_in_mb=0; check_for_duplicate_rows_during_compaction=true; che
k_for_duplicate_rows_during_reads=true; client_encryption_options=<REDACTED>; cluster_name=Test Cluster; column_index_c
che_size_in_kb=2; column_index_size_in_kb=64; commit_failure_policy=stop; commitlog_compression=null; commitlog_directo
y=null; commitlog_max_compression_buffers_in_pool=3; commitlog_periodic_queue_size=-1; commitlog_segment_size_in_mb=32;
commitlog_sync=periodic; commitlog_sync_batch_window_in_ms=NaN; commitlog_sync_period_in_ms=10000; commitlog_total_spac
_in_mb=null; compaction_large_partition_warning_threshold_mb=100; compaction_throughput_mb_per_sec=16; concurrent_compa
tors=null; concurrent_counter_writes=32; concurrent_materialized_view_writes=32; concurrent_reads=32; concurrent_replic
tes=null; concurrent_writes=32; counter_cache_keys_to_save=2147483647; counter_cache_save_period=7200; counter_cache_si
e_in_mb=null; counter_write_request_timeout_in_ms=5000; credentials_cache_max_entries=1000; credentials_update_interval
```

Εικόνα 8: Εντοπισμός του 1ου σφάλματος

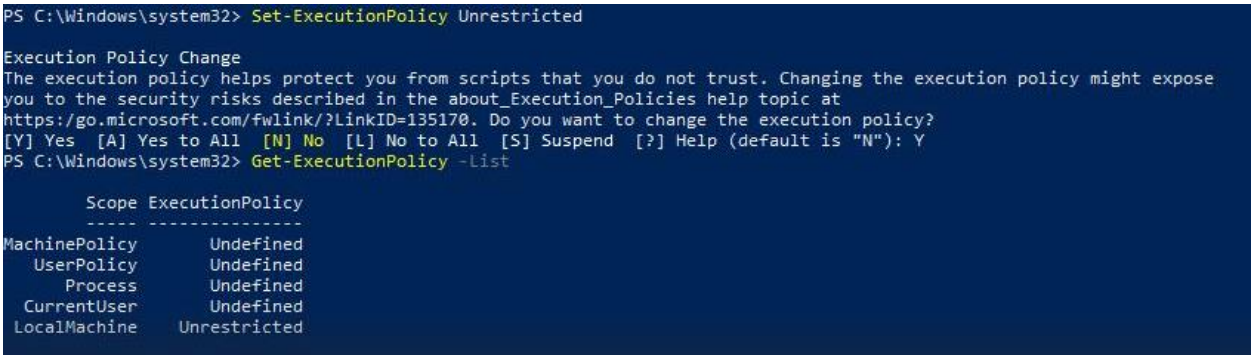
Εντολές επίλυσης:

a) ΣτοWindows PowerShell και εκτελώντας την εντολή Get-ExecutionPolicy -List , εμφανίζονται τα δικαιώματα των βασικών μεταβλητών του λειτουργικού συστήματοςWindows.
b) Τα δικαιώματα στην μεταβλητή LocalMachineδεν επιτρέπουν την δημιουργία κόμβων.



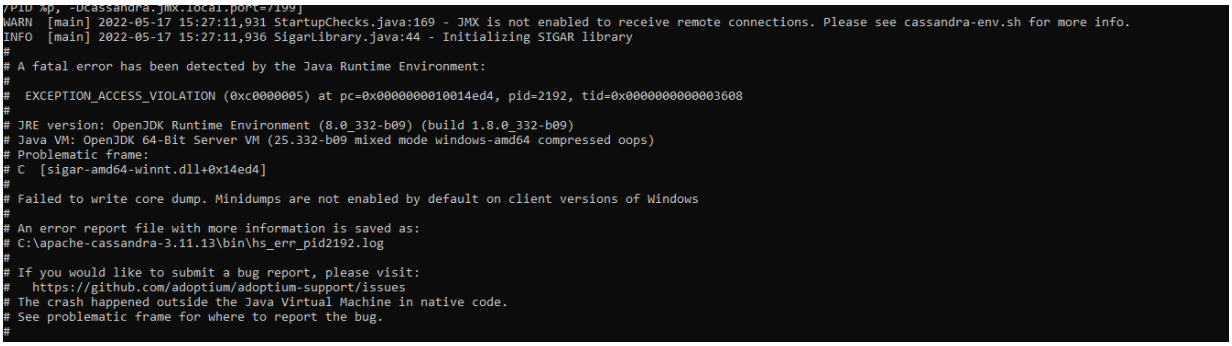
Εικόνα 9: Default δικαιώματα

c) Εκτελώντας την εντολή Set- ExecutionPolicyUnrestricted, τροποποιώντας τα δικαιώματα του υπολογιστή.



Εικόνα 10: Τροποποίηση δικαιώματος της μεταβλητής LocalMachine

2. Bypass SIGAR CHECK



Εικόνα 11: Εντοπισμός 2ου σφάλματος

Σφάλμα: sigar-amd64-winnt.dll+0x14ed4

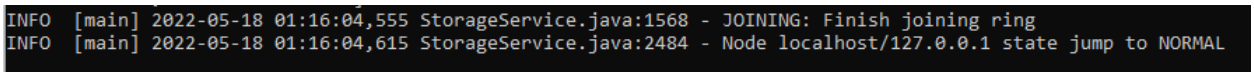
Το πακέτο sigar-amd64-winnt.dll δεν εμπεριέχεται στα βασικά πακέτα της Java 8 και είναι απαραίτητο η προσθήκη του για την λειτουργία του ApacheCassandra.

Επίλυση:

a. Αποθήκευση του συμπιεσμένου αρχείου hyperic-sigar https://sourceforge.net/projects/sigar/files/sigar/1.6/hyperic-sigar-1.6.4.zip/download
b. Στον φάκελο sigar-bin εντοπίζουμε το πακέτο sigar-amd64-winnt.dll και το μεταφέρουμε στον binφάκελο της Java.

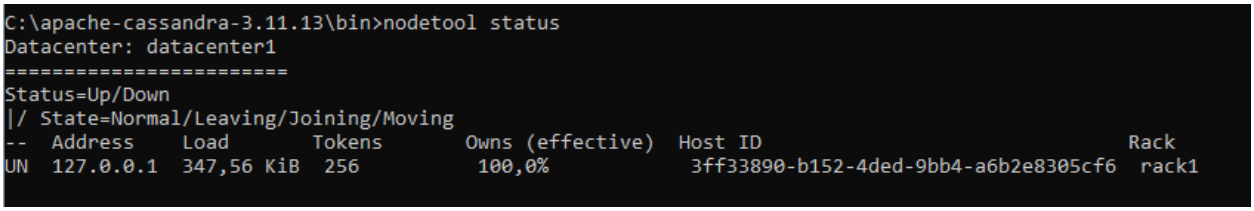
Βήμα 4:Έλεγχος εγκατάστασης

- Εκτέλεση της εντολής cassandra μέχρι να διορθωθούν όλα τα σφάλματα.
- Όταν εμφανισθεί το μήνυμα: <<Nodelocalhost/127.0.0.1 statejumptoNORMAL>>, τότε το εργαλείο έχει εγκατασταθεί επιτυχώς.



Εικόνα 12: Επιτυχής εγκατάσταση του ApacheCassandra

- Έλεγχος για ορθή επικοινωνία των κόμβων με την χρήση και δεύτερου CommandPrompt.
- Στο δεύτερο CommandPrompt πρέπει να εκτελεσθεί η εντολή nodetool status στο φάκελο C:\apache-cassandra-3.11.13\bin



Εικόνα 12:Επικοινωνία κόμβων

Κεφάλαιο 5: Apache Spark

Εικόνα14: Apache Spark

5.1: Τι είναι Apache Spark;

Το Apache Spark ανήκει στο ίδρυμα Apache και μπορεί να θεωρηθεί ως η λογική συνέχεια της μη σχεσιακής βάσης δεδομένων Apache Cassandra. Το Spark είναι πλατφόρμα ανοιχτού κώδικα, για κατανεμημένα συστήματα επεξεργασίας υψηλού φόρτου εργασίας Big Data. Μπορεί να εκτελεί ταχύτατα τις διεργασίες διαχείρισης δεδομένων, διανέμοντας τες στο δίκτυο εμβέλειας των διακομιστών του συστήματος. Η κατανομή των εργασιών πραγματοποιείται είτε αυτόματα στους υπολογιστές του είτε σε συνεργασία με άλλα κατανεμημένα συστήματα, όπως AWS. Οι λειτουργικές δυνατότητες του Spark, το κάνουν ιδανικό για την εφαρμογές Big Data και Machine Learning (Μηχανική Μάθηση), οι οποίες έχουν υψηλές και απαιτητικές ανάγκες για υπολογιστική ισχύ.

Το Apache Sparkαποτελεί την βελτιωμένη εκδοχή του Hadoop MapReduce.Το Spark βασίζεται στην λειτουργία του MapReduce να διαχωρίζει τις εργασίες επεξεργασίας μεγάλων δεδομένων σε μικρότερες και τις διανέμει σε όλους τους κόμβους του κατανεμημένου συστήματος, μειώνοντας το συνολικό χρόνο εκτέλεσης. Επιπλέον χρησιμοποιεί την προσωρινή αποθήκευση στην μνήμη(Random Access Memory), ώστε να αποκρίνεται άμεσα σε συχνά ερωτήματα που δέχεται η βάση δεδομένων, τα οποία δεν απαιτούν μεγάλη υπολογιστική ισχύ. Το Spark είναι ταχύτερο σε σχέση με το Hadoop MapReduce, διότι έχει μειώσει την πολυπλοκότητα του προγραμματισμού με την χρήση των API. Πολλοί διαχειριστές NRDMS χαρακτηρίζουν το Spark ως πολυεργαλείο, καθώς μπορεί να εκτελέσει κατανεμημένα ερωτήματα με την γλώσσα SQL,να δημιουργήσει data pipelines, να απορροφήσει δεδομένα σε μια βάση δεδομένων, να εκτελέσει αλγόριθμους Μηχανικής Μάθησης και να κατασκευάσει γραφήματα και ροές δεδομένων.

Ιστορική αναδρομή

Το Apache Sparkδημιουργήθηκε το 2009 στο Ερευνητικό Εργαστήριο Αλγόριθμων, Μηχανών και Ανθρώπων (**AMPLAB**) του Πανεπιστημίου της Καλιφόρνιας στο Berkeley από τους Matei Zaharia, Mosharaf Chowdhurry, Michael Franklin,Scott Shenker και Ion Stoica.Οι ερευνητές του εργαστηρίου παρατήρησαν ότι το εργαλείο Hadoop MapReduce αντιμετώπιζε δυσκολία στη διαχείριση επαναλαμβανόμενων υπολογιστικών διεργασιών, οι οποίες απαιτούν ταχεία επεξεργασία εφαρμόζοντας τεχνικές παράλληλου προγραμματισμού και κοινή δεδομένων χωρίς χρονική καθυστέρηση. Στις 3 Οκτωβρίου 2010 εκδόθηκε η 1^{η}έκδοση του Spark στο GitHub, η οποία ήταν δέκα φορές ταχύτερο από το MapReduceκαι οι πρώτοι χρήστες του ήταν ερευνητικά εργαστήρια του Πανεπιστημίου της Καλιφόρνιας.

Η πρώτη έκδοση του Spark υποστήριζε μόνο batch εφαρμογές, αλλά σύντομα αποσαφηνίστηκε η ουσιαστική χρήση του. Η διαδραστική αλληλεπίδραση χρήστη με τις βάσεις δεδομένων και τα ad-hoc ερωτήματα είναι τα κυριότερα χαρακτηριστικά που έκαναν το εργαλείο επιτυχημένο. Μετά από τις αρχικές εκδόσεις, έγινε σαφές ότι το εργαλείο θα επέκτεινε τις δυνατότητες του με την προσθήκη εξειδικευμένων βιβλιοθηκών, όπου οι Data Scientists θα μπορούν να επιλέξουν τα κατάλληλα plugins ανάλογα τις απαιτήσεις του κάθε συστήματος. Οι δημοφιλέστερες βιβλιοθήκες είναι MLlib για μηχανική μάθηση,Streaming για έλεγχο των ροών δεδομένων, Spark SQLκαι Graphx για επεξεργασία γραφημάτων.

5.2: Χαρακτηριστικά Apache Spark

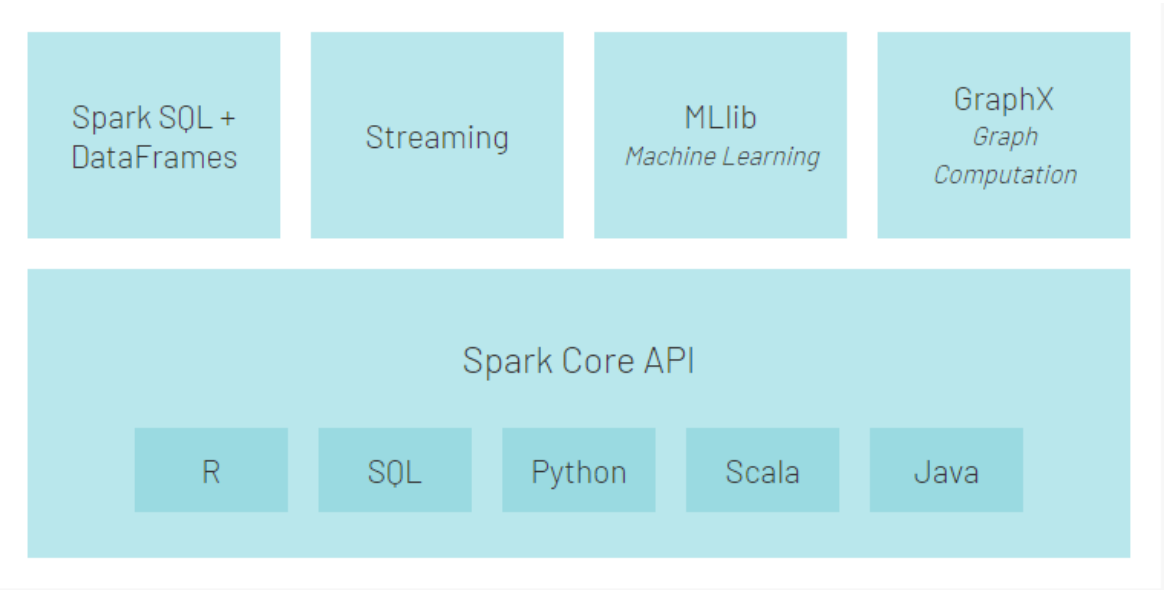
Ταχύτητα (Speed): Το Spark έχει σχεδιαστεί με σκοπό να αυξήσει την ταχύτητα εκτέλεσης υπολογιστικών εργασιών λειτουργώντας τόσο στη μνήμη όσο και στο δίσκο. Η εσωτερική του εφαρμογή επωφελείται την τεράστια πρόοδο της βιομηχανίας του hardware με την βελτίωση τιμής και απόδοσης των CPUκαι της μνήμης. Οι σύγχρονοι severs διαθέτουν εκατοντάδες πυρήνες και το λειτουργικό τους σύστημα βασίζεται στο Unix, εκμεταλλευόμενο την πολυνηματική (multithreading) και παράλληλη επεξεργασία με χαμηλό οικονομικό κόστος. Επιπλέον, το Spark αποκρίνεται στα ερωτήματα που δέχεται ο server με την δημιουργία υπολογιστικού κατευθυνόμενου ακυκλικού γραφήματος (DAG). Το γράφημα έχει σχεδιαστεί με πολλαπλά στάδια για αποτελεσματικότερη διανομή των εργασιών στους κόμβους, αναλύοντας τες που εκτελούνται στο σύμπλεγμα του συστήματος. Τέλος, η μηχανή φυσικής εκτέλεσης χρησιμοποιεί τη δημιουργία μαζικού και συμπαγούς κώδικα για την εκτέλεση[30].

Αρθρωτότητα (Modularity): Το Spark επιτυγχάνει την απλότητα αφαιρώντας την πολυπλοκότητα του προγραμματισμού και παρέχοντας μια θεμελιώδη και απλή λογική δομής δεδομένων. Στο Ανθεκτικό Κατανεμημένο Σύνολο Δεδομένων (**R**esilient **D**istributed **D**ataset) βασίζονται όλες οι άλλες αποδομήσεις δομημένων δεδομένων υψηλότερου επιπέδου, όπως Data Frames και Datasets. Επιπλέον, το Spark παρέχει ένα σύνολο μετασχηματισμών και ενεργειών ως λειτουργίες και ένα απλοποιημένο μοντέλο προγραμματισμού, που μπορεί να

χρησιμοποιείται για την δημιουργία εφαρμογών Big Data στις υποστηριζόμενες γλώσσες: SQL, Java, Python,R και Scala. Όλες οι δυνατότητες του είναι προσβάσιμες Spark μέσω API. Τα API είναι τεκμηριωμένα και δομημένα με περίτεχνο τρόπο καθιστώντας την αλληλεπίδραση των Data Scientists και τους developers των εφαρμογών με τις NRDBMS εύχρηστη. Επιστήμονες των δεδομένων και προγραμματιστές χρησιμοποιούν στο μέγιστο των δυνατοτήτων τα κλιμακούμενα εργαλεία και την αξιόπιστη απόδοση και ταχύτητα χωρίς πολλές αναζητήσεις και λεπτομέρειες.

5.3: Spark Ecosystem

Το προσφέρει ενοποιημένες βιβλιοθήκες με καλά τεκμηριωμένα API που περιλαμβάνουν τις ακόλουθες ενότητες ως βασικά στοιχεία: SparkSQL, Spark Structured Streaming, Spark MLlib και Graphx, συνδυάζοντας όλους τους φόρτους εργασίας που εκτελούνται σε έναν διακομιστή.



Εικόνα 13:Apache Spark Ecosystem

Spark Core: Όλες οι λειτουργίες του εργαλείου Spark βασίζονται στο Spark Core, καθιστώντας το ως το θεμελιώδες στοιχείο για παράλληλη και κατανεμημένη επεξεργασία των μεγάλων δεδομένων. Το Core είναι υπεύθυνο για όλες τις βασικές λειτουργίες Εισόδων/Εξόδων (I/O), τον προγραμματισμό και την παρακολούθηση των εργασιών στο σύμπλεγμα κόμβων του συστήματος. Επίσης, διατηρεί όλα τα στοιχεία που συσχετίζονται με το συγχρονισμό των διεργασιών, την δικτύωση με τα διάφορα συστήματα αποθήκευσης που συνεργάζεται, την ανάκτηση σφαλμάτων και την αποτελεσματική διαχείριση μνήμης. Τέλος, το Spark Core αξιοποιεί μια ειδική ανθεκτική δομή δεδομένων **R**esilient **D**istributed **D**atasets, τα οποία επαναχρησιμοποιούν τα δεδομένα κατανεμημένων υπολογιστικών συστημάτων. Τα RDDείναι ανθεκτικές/αμετάβλητες και κατακερματισμένες συλλογές εγγράφων και μπορούν να περιέχουν οποιοδήποτε γλώσσα αντικειμενοστραφή προγραμματισμού (Python,Java,Scala) ή και

αντικείμενα κλάσης που ορίζονται από τον χρήστη ενός συστήματος. Τα RDD υλοποιούνται είτε με τον μετασχηματισμό των υπαρχόντων RDD είτε με τη μεταφόρτωση ενός εξωτερικού dataset από μια σταθερή αποθήκευση όπως HBase, HDFS.

Spark SQL(Shark): Το Spark SQL αποτελεί το κατανεμημένο πλαίσιο για δομημένη επεξεργασία δεδομένων, βασιζόμενο στον πυρήνα του Spark. Οι προγραμματιστές του Shark έχουν την δυνατότητα να αξιοποιήσουν την ισχύ των δηλωτικών ερωτημάτων και τη βελτιστοποιημένη αποθήκευση εκτελώντας ερωτήματα τύπου SQL, τα οποία υπάρχουν σε RDD και σε άλλες εξωτερικές πηγές, χωρίς να εξαρτάται από το API ή την γλώσσα που χρησιμοποιείται για εκτέλεση υπολογιστικών εργασιών. Επίσης, επιτρέπει την διαδραστική και αναλυτική εφαρμογή τόσο σε data pipelines όσο και σε ιστορικά δεδομένα. Οι διαδραστικές εφαρμογές εξυπηρετούν τους χρήστες του συστήματος, παρέχοντας την δυνατότητα εκτέλεσης, εξαγωγής, μετασχηματισμού και φόρτωσης λειτουργιών στα δεδομένα, τα οποία προέρχονται από αρχεία JSON και σε συνεχεία να διεξάγουν ad-hoc ερωτήματα. Τέλος, το Shark διευκολύνει τη διαδικασία εξαγωγής και σύμπτυξης διαφόρων datasets, ώστε να είναι έτοιμα για εφαρμογή μηχανικής μάθησης.

Catalyst: Από την πρώτη έκδοση του Apache Spark, μέλος του βασικού στοιχείου του Spark SQL αποτελεί το Catalyst. Το Catalyst είναι ένα πλαίσιο, το οποίο βελτιστοποιεί την ενσωματωμένη γλώσσα Scala, αυξάνοντας την απόδοση των δηλωτικών ερωτημάτων που γράφουν οι προγραμματιστές των Big Data συστημάτων. Με μεγάλη ευκολία επιτυγχάνεται ο καθορισμός σύνθετων σχεσιακών βελτιστοποιήσεων και ο μετασχηματισμός ερωτημάτων, αξιοποιώντας ισχυρές δομές προγραμματισμού. Τέλος, το Catalyst διευκολύνει την προσθήκη νέων κανόνων βελτιστοποίησης πηγών και τύπων δεδομένων, καθώς παρατηρείται ότι τα συστήματα μεγάλων δεδομένων αυξάνουν με ραγδαία ταχύτητα την εδραίωση τους.

Data Frame: Στην έκδοση 1.6.3 του Apache Spark (7Νοεμβρίου 2016), αφαιρέθηκε το Data Frame από Spark SQL, δημιουργώντας ένα νέο στοιχείο στο οικοσύστημα του εργαλείου. Στις παλαιότερες εκδόσεις αυτό το στοιχείο περιλαμβανόταν στο Schema του RDD. Πλέον το Data Frame API αποτελεί μια ξεχωριστή κατανεμημένη συλλογή δεδομένων, τα οποία ταξινομούνται σε στήλες και ενσωματώνονται με διαδικαστικό κώδικα και σχεσιακή επεξεργασία. Οι λειτουργίες αξιολογούνται με επιεική κριτήρια, παρέχοντας υποστήριξη για σχεσιακές βελτιώσεις και βελτιστοποίηση της συνολικής ροής εργασιών επεξεργασίας δεδομένων.

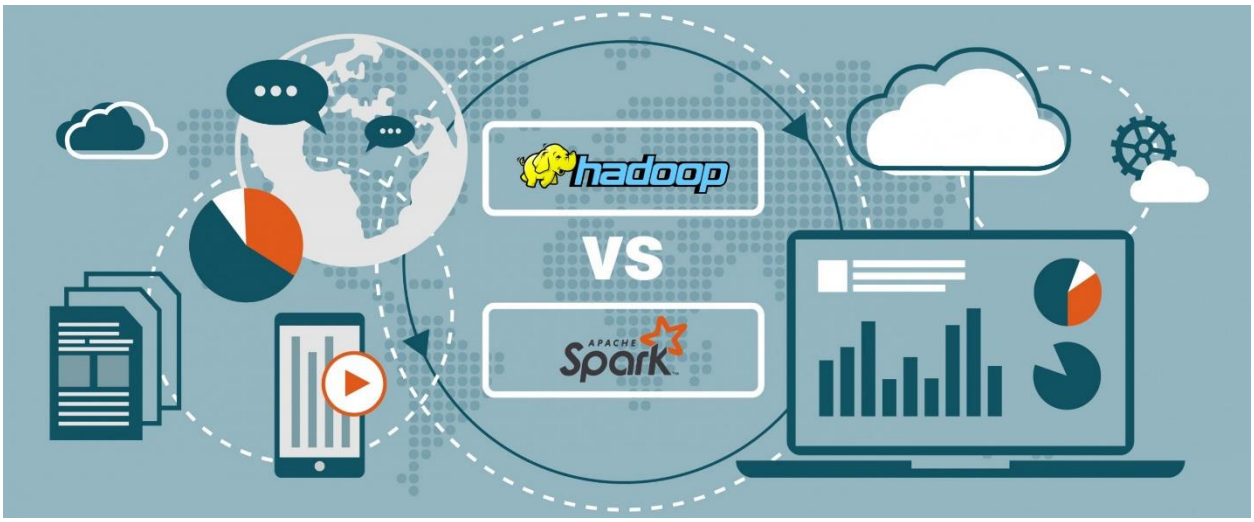
Spark Streaming: Από τα αρχεία καταγραφής έως τα δεδομένα αισθητήρων, οι προγραμματιστές των εφαρμογών αναγκάζονται να αντιμετωπίσουν ολοένα και περισσότερο το πρόβλημα διαχείρισης πραγματικού χρόνου ροής δεδομένων (data pipelines). Το Spark Streaming επιτυγχάνει την υψηλή απόδοση και την ανοχή σε σφάλματα κατά την διαδικασία ζωντανών ροών δεδομένων. Το Spark λειτουργεί αξιοποιώντας πολλούς και διαφορετικούς

αλγόριθμους, οι οποίοι λαμβάνουν τα δεδομένα σε ένα σύστημα αρχείων, μια βάση δεδομένων και σε πίνακα ελέγχου πραγματικού χρόνου. Η βασικότερη τεχνική που χρησιμοποιεί τοSpark Streaming είναι το Micro-batching, η οποία επιτρέπει σε μια διεργασία να αντιμετωπίσει έναdata pipeline ως ακολουθία μικρών πακέτων πληροφοριών. Ως εκ τούτου, είναι κατανοητό ότι το Spark Streaming λειτουργεί σε 3 φάσεις. Αρχικά, **συγκεντρώνει** όλες τις πηγές ροών δεδομένων από τα διασυνδεδεμένα API και τις υποδοχές TCP. Έπειτα τα συγκεντρωμένα δεδομένα **επεξεργάζονται** χρησιμοποιώντας πολύπλοκους αλγόριθμος εκφράζοντας συναρτήσεις υψηλού επιπέδου και πολύπλοκους μετασχηματισμούς. Τέλος, τα επεξεργασμένα δεδομένα προωθούνται και **αποθηκεύονται** σε βάσεις δεδομένων.

Spark MLlib: Όπως έχει προαναφερθεί, ο όγκος των δεδομένων αυξάνεται συνεχώς, και οι αλγόριθμοι μηχανικής μάθησης παρουσιάζουν ευκολία χρήσης σε ένα τεράστιο εύρος συστημάτων και τα αποτελέσματα τους αναδύουν μεγάλη ακρίβεια, προβάλλοντας την ανάγκη για την ύπαρξη βιβλιοθηκών ML. Το project Apache Spark εντόπισε αυτή την ανάγκη και πρόσθεσε στο βασικό οικοσύστημα του την βιβλιοθήκη MLlib. Το Spark MLlib αποτελεί μια επεκτάσιμη βιβλιοθήκη μηχανικής μάθησης, η οποία αναφέρεται στους αλγορίθμους υψηλής ποιότητας και ταχύτητας. Περιέχει βιβλιοθήκες ML που έχουν εφαρμογή σε διαφόρους αλγορίθμους, όπως ομαδοποίηση, ταξινόμηση και ταυτόχρονος καθαρισμός δεδομένων. Παρόλα αυτά το MLlib διατηρεί τους αρχέγονους χαμηλού επιπέδου αλγορίθμους, για λόγους συμβατότητας με τα παλαιά συστήματα και τις αναχρονισμένες τεχνικές μηχανικής μάθησης. Το χρονολογικό ορόσημο για την MLlib, αποτελεί η έκδοση 2.2.0 (17 Ιουλίου 2017) του Spark. Στην έκδοση 2.0,0 (26 Ιουλίου 2016) το API βασιζόταν σε RDD στο πακέτο spark.mllib εισήχθη σε λειτουργία συντήρησης. Αντίθετα από την έκδοση 2.2.0 και έπειτα το API της MLlib βασίζεται στο Data Frame, κάνοντας την φιλικότερη προς το χρήστη σχέση με το RDD. Τέλος, η MLlib χρησιμοποιεί το πακέτο γραμμικής άλγεβρας Breeze, αυξάνοντας την ταχύτητα αριθμητικών υπολογισμών.

Spark Graphx (Graph Computation): Η πιο πρόσφατη βιβλιοθήκη που προστέθηκε στο οικοσύστημα του Spark αποτελεί το Graphx (έκδοση 3.0.0, 18 Ιουνίου 2020). Το Graphx είναι API για παράλληλη εκτέλεση και αναπαράσταση υπολογιστικών αποτελεσμάτων σε γραφήματα. Ταυτόχρονα, αυτή η βιβλιοθήκη παρέχει την δυνατότητα για ομαδοποίηση, ανάλυση και εύρεση της βέλτιστης διαδρομής αναζήτησης γραφήματος, καθώς πραγματοποιεί λεπτομερής ανάλυση του δικτύου στα οποία αποθηκεύονται τα δεδομένα. Επιπλέον το Spark Graphx συμβάλει στην ανάκτηση πολύτιμων πληροφοριών από τα ασυνεπή δεδομένα, αξιοποιώντας την διαδικασία ELT και μειώνοντας τον χρόνο και το κόστος ανάλυσης δεδομένων. Καθώς το Spark είναι ικανό να αποθηκεύει πληροφορίες στη μνήμη και μπορεί να εκτελεί συνεχόμενα ερωτήματα γρήγορα, καθιστά εύκολο τον εντοπισμό των αλγορίθμων μηχανικής μάθησης που μπορούν να επαναχρησιμοποιηθούν για ένα συγκεκριμένο είδος δεδομένων. Τέλος, η βιβλιοθήκη Graphx αποτελεί το ιδανικότερο εργαλείο για αναπαράσταση δεδομένων πραγματικού χρόνου, αξιοποιώντας το Spark Streaming.

5.4: Hadoop Vs Spark



Εικόνα 16:Hadoop VS Spark

Το Apache Spark και Hadoop αποτελούν τα δύο από τα πιο σημαντικά κατανεμημένα συστήματα επεξεργασίας δεδομένων. Αμφότερα ανήκουν στα project του ομίλου Apache και συχνά χρησιμοποιούνται ταυτόχρονα, καθώς παρομοιάζουν αρκετές ομοιότητες. Παρόλες τις ομοιότητες κρίνεται απαραίτητο να κατανοηθούν τα χαρακτηριστικά και οι μέγιστες δυνατότητες που προσφέρει καθένα από τα εργαλεία.

Αρχιτεκτονική: Στο Hadoop, όλα τα δεδομένα χωρίζονται σε block που αναπαράγονται στις μονάδες δίσκου των διαφόρων servers σε ένα cluster, με το HDFS (Hadoop Distributed File System) να παρέχει υψηλά επίπεδα πλεονασμού και ανοχής σφαλμάτων. Οι εφαρμογές του Hadoop μπορούν στη συνέχεια να εκτελεστούν ως μία εργασία ή ως κατευθυνόμενο άκυκλο γράφημα που περιέχει πολλαπλές εργασίες. Στην πρώτη έκδοση του Hadoop (1 Απριλίου 2006), μια κεντρική υπηρεσία Job Tracker κατένειμε εργασίες σε κόμβους που θα μπορούσαν να εκτελούνται ανεξάρτητα ο ένας από τον άλλο και μια τοπική υπηρεσία Task Tracker διαχειριζόταν την εκτέλεση εργασιών από μεμονωμένους κόμβους. Από την έκδοση 2.0.0 (23 Μαΐου 2012) αυτές οι δύο υπηρεσίες αντικαταστάθηκαν από το υποσύστημα YARN. Το YARN περιέχει τρεις υπηρεσίες διαχείρισης διεργασιών Resource Manager, Node Manager και Application Manageroι οποίες λειτουργούν στο υπόβαθρο όλων των συστημάτων Hadoop.Ο Resource Manager λειτουργεί ως παγκόσμιος χρονοπρογραμματιστής εργασιών και κατανέμει τους διαθέσιμους υπολογιστικούς πόρους. Ο Node Manager είναι εγκατεστημένος σε κάθε κόμβο του συμπλέγματος, παρακολουθώντας την χρήση πόρων. Ο Application Manager υλοποιεί για κάθε εφαρμογή που διαπραγματεύεται τους πόρους από τον Resource Manager και συνεργάζεται με τον Node Manager για την εκτέλεση εργασιών επεξεργασίας δεδομένων. Ταυτόχρονα παρέχεται μίαabstract πηγή πόρων, όπου συγκρατούνται όλες οι πληροφορίες

σχετικά με τους πόρους που έχουν εκχωρηθεί σε διαφορετικούς κόμβους και εφαρμογές. Όπως αναφέρθηκε και στην αρχή το Spark αποτελεί την βελτιωμένη έκδοση του Hadoop.Η θεμελιώδης διαφορά μεταξύ Hadoop και Spark σχετίζεται με τον τρόπο οργάνωσης των δεδομένων για επεξεργασία. Στο Spark, η πρόσβαση στα δεδομένα υλοποιείται από εξωτερικά αποθετήρια αποθήκευσης όπως HDFS, cloud ή διάφορες databases και άλλους τύπους data stores. Παρόλο που οι περισσότερες επεξεργασίες γίνονται στη μνήμη, η πλατφόρμα μπορεί να διαμοιράσει δεδομένα στον δίσκο αξιοποιώντας όλες τις δυνατότητες του cluster.

Κλιμάκωση (Scalability): Τα συστήματα Hadoop μπορούν να κλιμακωθούν για να φιλοξενήσουν μεγαλύτερα σύνολα δεδομένων στα οποία γίνεται σταδιακή πρόσβαση, επειδή τα δεδομένα μπορούν να αποθηκευτούν και να υποβληθούν για επεξεργασία με την μικρή χρήση του δίσκου σε σχέση με την μνήμη. Το YARN επιτρέπει στα clusters (συμπλέγματα) να υποστηρίζουν δεκάδες χιλιάδες κόμβους συνδέοντας πολλά υποσυμπλέγματα που έχουν τους δικούς τους διαχειριστές πόρων. Απαραίτητη επένδυση για την επέκταση συστημάτων Hadoop αποτελεί το εργατικό δυναμικό, καθώς είναι αναγκαία η εγκατάσταση εσωτερικών εφαρμογών για την παροχή νέων κόμβων και την προσθήκη τους σε ένα σύμπλεγμα. Επιπρόσθετα, με το Hadoop, η αποθήκευση συγκεντρώνεται με υπολογιστικούς πόρους στους κόμβους των συμπλεγμάτων, γεγονός που μπορεί να δυσκολέψει την λειτουργία των εφαρμογών και των χρηστών εκτός του cluster να έχουν πρόσβαση στα δεδομένα. Παρόλα αυτά κάποια από τα ζητήματα επεκτασιμότητας μπορούν να επιλυθούν με τις υπηρεσίες Hadoop Cloud. Εν αντιθέσει, στο Spark ο χώρος αποθήκευσης και οι υπολογιστικές διαδικασίες διαχωρίζονται, γεγονός που μπορεί να διευκολύνει τις εφαρμογές και τους χρήστες να έχουν πρόσβαση στα δεδομένα από οπουδήποτε. Επιπλέον, το Spark περιλαμβάνει εργαλεία που μπορούν να βοηθήσουν του χρήστες να κλιμακώσουν δυναμικά τους κόμβους ανάλογα με τις απαιτήσεις των διεργασιών. Στις περισσότερες περιπτώσεις δεν χρειάζεται επένδυση σε εργατικό δυναμικό, για ανακατανομή κόμβων στα συμπλέγματα, καθώς για τα συστήματα του Spark αυτή η διαδικασία είναι αυτοματοποιημένη. Ωστόσο, για την κλιμάκωση των εφαρμογών στο Spark πρέπει να διασφαλιστεί ο διαμοιρασμός του φόρτου εργασίας μεταξύ των κόμβων για να μειωθεί ο διασκορπισμός μνήμης.

Ασφάλεια(Security): Το Hadoop παρέχει υψηλότερο επίπεδο ασφαλείας με λιγότερα εξωτερικά έξοδα για μακροπρόθεσμη διατήρηση δεδομένων. Το HDFS προσφέρει ολοκληρωμένη κρυπτογράφηση με ξεχωριστές ζώνες κρυπτογράφησης και ενσωματωμένη υπηρεσία διαχείρισης των κλειδιών κρυπτογράφησης. Επιπλέον, περιλαμβάνει ένα μοντέλο που βασίζεται σε δικαιώματα για την επιβολή ελέγχων πρόσβασης για αρχεία και καταλόγους με τη δυνατότητα δημιουργίας λιστών ελέγχου πρόσβασης που μπορούν να χρησιμοποιηθούν για την εφαρμογή ασφαλείας βάσει ρόλων και άλλων τύπων κανόνων για διαφορετικούς χρήστες ή ομάδες. Μπορεί επίσης να εκμεταλλευτεί τα σχετικά εργαλεία όπως το Apache Knox, μια πύλη που παρέχει RESTAPI υπηρεσίες ελέγχου ταυτότητας και διακομιστή μεσολάβησης για την επιβολή πολιτικών ασφαλείας στα συμπλέγματα Hadoop και το Apache Ranger, ένα κεντρικό πλαίσιο διαχείρισης ασφαλείας για περιβάλλοντα Hadoop[45]. Αντιθέτως, το Spark διαθέτει ένα

πιο περίπλοκο μοντέλο ασφαλείας που υποστηρίζει διαφορετικά επίπεδα ασφαλείας για διαφορετικούς τύπους ανάπτυξης. Χρησιμοποιεί μια κοινή μυστική προσέγγιση ελέγχου ταυτότητας για κλήσεις απομακρυσμένων διαδικασιών μεταξύ διεργασιών του Spark, με ειδικούς μηχανισμούς ανάπτυξης για την δημιουργία μυστικών κωδικών πρόσβασης. Σε ορισμένες περιπτώσεις, οι προστασίες ασφαλείας είναι περιορισμένες επειδή όλες οι εφαρμογές διαμοιράζονται κοινά απόρρητα αρχεία. Το μοντέλο Spark κατασκευάστηκε κυρίως για να επιβάλλει την ασφάλεια άνω του επιπέδου των ροών δεδομένων, οι οποίες είναι λιγότερο μόνιμες από τα δεδομένα που αποθηκεύονται για μεγάλες περιόδους, δημιουργώντας ανησυχία για ευρείας κλίμακας κυβερνοεπιθέσεις στις υποδομές του. Ο έλεγχος ταυτότητας και άλλα μέτρα ασφαλείας δεν είναι ενεργοποιημένα από προεπιλογή στο Spark. Ως εκ τούτου, με τον κατάλληλο συνδυασμό αλγορίθμου κρυπτογράφησης με πολιτικές διαχείρισης κλειδιών μπορεί να επιτευχθούν επαρκείς ζώνες προστασίας δεδομένων.

5.5: Τι δεν προσφέρει;

Το Apache Spark αποτελεί ένα από τα βασικότερα εργαλεία στην διαχείριση Μεγάλων Δεδομένων που χρησιμοποιείται ευρέως από τις βιομηχανίες, αλλά εκτός από το πλήθος των δυνατοτήτων όπως προαναφέρθηκαν, έχει ορισμένα μειονεκτήματα, όπως έλλειψη παροχής υποστήριξης σε πραγματικό χρόνο, διαχείριση αρχείων μικρού μεγέθους, κόστος κ.ά..

Έλλειψη υποστήριξης για επεξεργασία δεδομένων σε πραγματικό χρόνο: Το Spark δεν υποστηρίζει πλήρως την επεξεργασία ροής δεδομένων σε πραγματικό χρόνο. Στο Spark Streaming, η εισερχόμενη ροή δεδομένων πραγματικού χρόνου διαμοιράζεται σε προκαθορισμένα διαστήματα και σε ειδικές ανθεκτικές δομές δεδομένων. Τα RDD υποβάλλονται σε επεξεργασία χρησιμοποιώντας λειτουργίες όπως σύνδεση DB, MapReduce. Το αποτέλεσμα αυτών των διεργασιών επιστρέφεται σε Micro-Batch. Επομένως, δεν πρόκειται για επεξεργασία σε πραγματικό χρόνο, αλλά το Spark επεξεργάζεται δεδομένα **σχεδόν** σε πραγματικό χρόνο.

Διαχείριση αρχείων μικρού μεγέθους: Το Spark βασίζεται σε άλλες πλατφόρμες όπως το Hadoopή άλλη Cloud-based εφαρμογή για την διαχείριση αρχείων. Ακόμα και με τον συνδυασμό Hadoop και Spark διαπιστώνεται ότι υπάρχει πρόβλημα με την διαχείριση αρχείων μικρού μεγέθους. Το HDFS (**H**adoop **D**istributed **F**ile **S**ystem) παρέχει περιορισμένο αριθμό μεγάλου μεγέθους αρχείων αντί για μεγάλο αριθμό μικρού όγκου αρχείων. Επιπλέον, όταν στο σύστημα διαθέτει πολλά μικρά gzip δεδομένα, το Spark τα διατηρεί στο δίκτυο και τα αποσυμπιέζει με την προϋπόθεση ότι ολόκληρο το αρχείο βρίσκεται στο Spark Core. Δαπανιέται μεγάλο χρονικό διάστημα για την εγγραφή και αποσυμπίεση τους στο Core. Στο RDD κάθε αρχείο που προκύπτει, διαιρείται σε ένα μεγάλος πλήθος μικρότερων για αποτελεσματικότερη διαχείριση, απαιτώντας εκτενή αναζήτηση μέσα στους κόμβους του συστήματος.

Οικονομική επιβάρυνση: Το Spark όπως προαναφέρθηκε επεξεργάζεται τα δεδομένα απευθείας στην μνήμη (In-Memory), αλλά κάποιες φορές αποτελεί πρόβλημα. Για την In-Memory επεξεργασία απαιτείται μεγάλη οικονομική επένδυση. Η κατανάλωση μνήμης δεν αντιμετωπίζεται με user-friendly τρόπο. Το Spark απαιτεί πολλή μνήμη RAM, για να λειτουργεί στην μνήμη αυξάνοντας δραματικά το κόστος ενός συστήματος.

Έλλειψη αυτόματων βελτιώσεων: Δυστυχώς, το Apache Spark δεν διαθέτει αυτοματοποιημένη βελτίωση κώδικα. Η χειροκίνητη βελτιστοποίηση είναι επαρκής για συγκεκριμένα σύνολα δεδομένων. Ουσιαστικά, ο προγραμματιστής ορίζει ξεχωριστά κάθε partition, προσθέτοντας ως δεύτερη παράμετρο την μέθοδο της παραλληλοποίησης. Σε κάποιες περιπτώσεις είναι επιθυμητή η χειροκίνητη βελτιστοποίηση για ορθή κατάτμηση και προσωρινή αποθήκευση στο Spark.

5.6: Εγκατάσταση σε Windows 10

Το Apache Spark καθώς ανήκει στην κατηγορία των open-source projects, προϋποθέτει η λειτουργία του να πραγματοποιείται σε ελεύθερο λογισμικό Linux. Το Linux και τα Windows βασίζονται στο Unix. Ωστόσο, για την αξιοποίηση των χαρακτηριστικών των Linux στα Windows απαιτείται η ενεργοποίηση του υποσυστήματος WSL (Windows Subsystem for Linux). Το WSL βοηθά στην ενσωμάτωση και στην εκτέλεση εφαρμογών Linux εγγενώς στα Windows, παρέχοντας τη δυνατότητα άμεσης αλληλεπίδρασης, χωρίς να καταναλώνονται επιπλέον πόροι όπως θα γινόταν με την χρήση Virtual Machine. Εκτός από το WSL, χρειάζεται η εγκατάσταση εφαρμογής UBUNTU 20.04 LTS η οποία είναι το περιβάλλον διαχείρισης του Spark. Τέλος, καθώς το Spark βασίζεται στη τεχνολογία του Hadoop, στο εικονικό περιβάλλον που θα αναπτυχθεί υποχρεωτικά εμπεριέχονται όλα τα δομικά στοιχεία του.

Βήμα 1:

Στην εφαρμογή UBUNTU 20.04 LTS πρέπει να εγκατασταθεί το περιβάλλον Hadoop με την χρήση των εντολών[49].

tar –xvzf auto_hadoopuser_WSL2.tar.gz

rm –rfauto_hadoopuser_WSL2.tar.gz

/home/hadoopuser/hadoop/myScripts/install-cluster2.sh

```
(base) hadoopuser@DESKTOP-L0A1HS6:~$ /home/hadoopuser/hadoop/myScripts/install-cluster.sh
==========================
Δημιουργία Περιβάλλοντος.
==========================
------------------------------
Δημιουργία Group hadoopgroup.
------------------------------
[sudo] password for hadoopuser:
```

Εικόνα 17: Δημιουργία περιβάλλοντοςHadoop

Βήμα 2:

Το Spark Core API αξιοποιεί όλες τις γλώσσες αντικειμενοστραφή προγραμματισμού (Java, Python). Για την εκτέλεση προγραμμάτων γίνεται η επιλογή της γλώσσας Python.

condacreate –namepyspark_envpython=3.9

```
(base) hadoopuser@DESKTOP-L0A1HS6:~$ conda create --name pyspark_env python=3.9
Collecting package metadata (current_repodata.json): done
Solving environment: done


==> WARNING: A newer version of conda exists. <==
  current version: 4.12.0
  latest version: 4.13.0
```

Εικόνα 18: Δημιουργία περιβάλλοντος Python

Βήμα 3:

Για την ολοκληρωμένη χρήση του Pseudo Cluster που έχει δημιουργηθεί χρειάζεται να γίνει αλλαγή των δικαιωμάτων του χρήστημε την εντολή:

sudochown –R hadoopuser:hadoopgroup /home/hadoopuser/hadoop

Βήμα 4:

Ρύθμιση παραμέτρων, όπου αναφέρεται στη Java, Python, Hadoop, Yarn.

nano ~/.bashrc

```
# <<< conda initialize <<<

#------------------------------------------ SETTINGS -------------------------------------
# >>> conda activate conda env pyspark_env >>>
conda activate pyspark_env
# <<< conda activate conda env pyspark_env <<<
#------------------------------------------ JAVA -----------------------------------------
#-----------------------------------------------------------------------------------------
export JAVA_HOME='/usr/lib/jvm/java-8-openjdk-amd64'
export JRE_HOME='/usr/lib/jvm/java-8-openjdk-amd64/jre'
#------------------------------------------ HADOOP ---------------------------------------
#-----------------------------------------------------------------------------------------
export
HADOOP_CLASSPATH='/usr/lib/jvm/java-8-openjdk-amd64/lib/tools.jar:/home/hadoopuser/spark/yarn/spark-3.1.3-ya
rn-shuffle.jar'
export HADOOP_HOME='/home/hadoopuser/hadoop'
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$HADOOP_HOME/myScripts
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/home/hadoopuser/hadoop/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
#------------------------------------------ SPARK ----------------------------------------
#-----------------------------------------------------------------------------------------
export PYSPARK_PYTHON='/home/hadoopuser/miniconda3/envs/pyspark_env/bin/python3'
export PYSPARK_DRIVER_PYTHON='/home/hadoopuser/miniconda3/envs/pyspark_env/bin/python3'
export SPARK_HOME='/home/hadoopuser/spark'
export PATH=$PATH:$SPARK_HOME/bin:$SPARK_HOME/sbin:$SPARK_HOME/jars:/etc/default
export SPARK_CLASSPATH=$JAVA_HOME/lib/tools.jar:$SPARK_HOME/yarn/spark-3.1.3-yarn-shuffle.jar
export PYTHONPATH=$SPARK_HOME/python:$SPARK_HOME/python/lib/py4j-0.10.9-src.zip:$PYTHONPATH
export HBASE_OPTS="$HBASE_OPTS --add-opens java.base/sun.nio.ch=ALL-UNNAMED"
export SPARK_OPTS="--packages graphframes:graphframes:0.8.2-spark3.1-s_2.12"
#------------------------------------ Settings for Jupyter -------------------------------
alias jupyter-notebook='~/miniconda3/envs/pyspark_env/bin/jupyter-notebook --no-browser'
#------------------------------------ END SETTINGS ---------------------------------------
```

Εικόνα 19: Προσθήκη παραμέτρων στο bashrc

Βήμα 5:

Το WSL αντιλαμβάνεται σαν master κόμβο τον υπολογιστή που χρησιμοποιείται, ο οποίος δεν διαθέτει άλλους slaves κόμβους για την εκτέλεση των εφαρμογών Spark. Με την προσθήκη του αρχείου wsl.conf, ορίζεται ως hostname=master και δεν πραγματοποιείται αυτόματη δημιουργία hosts.

sudonano /etc/wsl.conf

```
# This file was automatically generated by WSL. To stop automatic generation of this file, add the following
entry to /etc/wsl.conf:
[network]
hostname = master
generateHosts = false
```

Εικόνα 20: Παράμετροι στο wsl.conf

Βήμα 6:

Τροποποίηση αρχείου host

sudo nano /etc/hosts

```
# This file was automatically generated by WSL. To stop automatic generation of this file, add the following
entry to /etc/wsl.conf:
# [network]
# generateHosts = false
127.0.1.1 master
0.0.0.0 worker
# The following lines are desirable for IPv6 capable hosts
::1 ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
```

Εικόνα 21: Παράμετροι στο host

Επεξήγηση:

- 127.0.1.1 master : ο τοπικός κόμβος (υπολογιστής δημιουργίας του εικονικού περιβάλλοντος)ορίζεται ως master του περιβάλλοντος
- 0.0.0.0 worker : ο απομακρυσμένος κόμβος, ο οποίος εκτελεί όλες τις διεργασίες

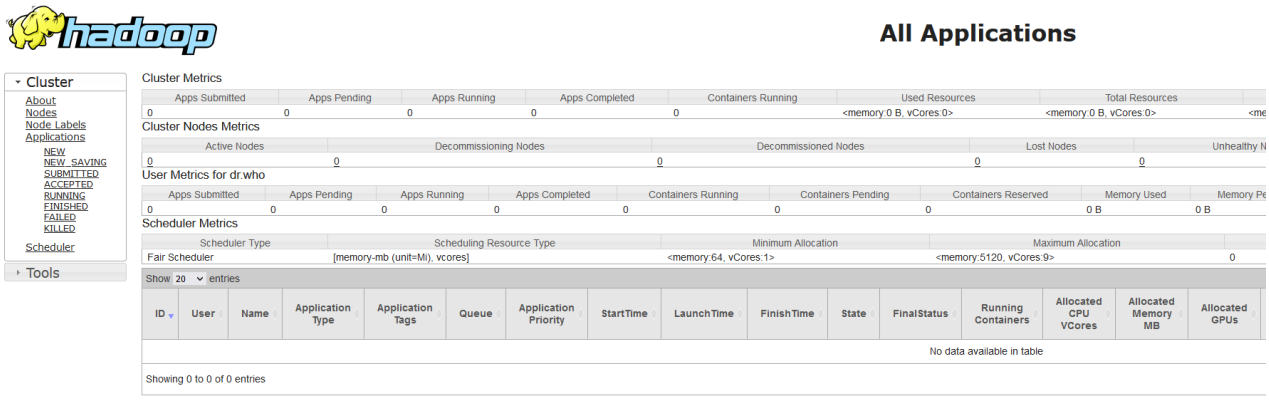
Βήμα 7:

Δημιουργία cluster με την εντολή

create-cluster.sh

Τέλος το cluster έχει δημιουργηθεί και παρέχεται η δυνατότητα ελέγχου της κατάστασής του

http://localhost:8088/cluster



Εικόνα 22: Κέντρο ελέγχου cluster

Συντομογραφίες

BD	Big Data
DB	Βάση δεδομένων
RDBMS	Σχεσιακή βάση δεδομένων
NoSQL	Not Only SQL
NRDBMS	Μη-Σχεσιακή βάση δεδομένων
RDD	Ανθεκτικό Κατανεμημένο Σύνολο Δεδομένων
ML	Machine Learning

Ευρετήριο Διαγραμμάτων

Ευρετήριο Εικόνων

Βιβλιογραφία

[1] Manning P. (2013): Big Data in History (Palagrave Pivot) 2013th Edition, Palgrave Pivot

[2] Zennaro Marco (December 2016): "Developing the ICU ecosystem to harness IoTs" Intro to Big Data https://www.itu.int/en/ITU-D/Regional-Presence/AsiaPacific/SiteAssets/Pages/Events/2016/Dec-2016-IoT/IoTtraining/BigData%20-Zennaro.pdf

[3] Marr Benard (Febuary 25,2015): A brief history of big data everyone should read https://www.weforum.org/agenda/2015/02/a-brief-history-of-big-data-everyone-should-read/

[4] Maria Cristina Mariani, Osei Kofi Tweneboah, Maria Pia Beccar-Varela (2021): Data Science in theory and practice 2nd Edition,Willey

[5] Indicate Team: What is Data Volumehttps://www.indicative.com/resource/volume-of-data/

[6] Indicate Team: What is Data Velocity https://www.indicative.com/resource/data-velocity/

[7] Indicate Team: What is Data Variety https://www.indicative.com/resource/data-variety/

[8] Indicate Team: What is Data Veracityhttps://www.indicative.com/resource/data-veracity/

[9] Smallcombe Mark (August 20,2020): The 7Vs of Big Datahttps://www.integrate.io/blog/7-vs-big-data/#volume

[10] Oracle: What is Big Data? https://www.oracle.com/a/ocom/docs/what-is-big-data-ebook-4421383.pdf

[11]Jacquelyn Bulao (May 2,2022): How much data is created every day in 2022? https://techjury.net/blog/how-much-data-is-created-every-day/#gref

[12]Jens-Petter Sandvik,Katin Franke, Habtanum Abie, Ande Arnes (2022):Quantifying data volatility of IoT forensics with examples from Contik OS, 9th Annual DFRWS Europe Conference

[13] Tableau: What Is Data Visualization?Definition,Examples,And Learning Resourceshttps://www.tableau.com/learn/articles/data-visualization#:~:text=Data%20visualization%20is%20the%20graphical,outliers%2C%20and%20patterns%20in%20data.

[14]Rock Content (June 1,2020): What is Big Data Visualization? https://rockcontent.com/blog/big-data-visualization/#:~:text=Big%20Data%20visualization%20describes%20data,easy%20to%20understand%20and%20interpret.

[15] Enterprice Big Data Framewort (January 9,2019): Three different data structureshttps://www.bigdataframework.org/data-types-structured-vs-unstructured-data/

[16]Cyntia Harvey (August 9,2018): Big Data Pros and Conshttps://www.datamation.com/big-data/big-data-pros-and-cons/

[17] Oracle: What is NoSQL?https://www.oracle.com/database/nosql/what-is-nosql/

[18]TechTerms (August 27,2013): NoSQLhttps://techterms.com/definition/nosql

[19]GeekforGeeks (July 7,2022): Types of NoSQL Databases https://www.geeksforgeeks.org/types-of-nosql-databases/

[20]IBM Cloud Education (August 6,2019): NoSQL Databases https://www.ibm.com/cloud/learn/nosql-databases

[21]MongoDB:Types of NoSQL Databases https://www.mongodb.com/scale/types-of-nosql-databases

[22]https://www.geeksforgeeks.org/features-of-cassandra/?ref=gcse

[23] Jeff Carpenter,Eben Hewitt (April 2020): Cassandra: The Definitive Guide, 3rd Edition, O'Reilly Media

[24]C.Y.Kan (December 23, 2014): Cassandra Data Modeling and Analysis, Packt Publishing

[25]GeekforGeeks(May 10, 2022): Features of Cassandrahttps://cassandra.apache.org/doc/latest/cassandra/cql/

[26]Apache Cassandra : The Cassandra Query Language (CQL) https://www.programsbuzz.com/article/apache-spark-history

[27]Shiksha Dahiya (July 21, 2021): Apache Spark History https://sql-stack.com/2018/12/07/apache-sparks-history/

[28]LFEIOCK (December 7, 2018):Apache Spark's Historyhttps://www.knowledgehut.com/tutorials/apache-spark-tutorial/apache-spark-evolution

[29]Jule S.Damji, Brooke Wenig, Tathagata Das, Denny Lee (July 16,2020) :Learning Spark Lightning-Fast Data Analytics 2nd Edition, O'Reilly Media

[30] James Peebles(July 11, 2022): Apache Spark Ecosystem and Spark Components https://www.projectpro.io/article/apache-spark-ecosystem-and-spark-components/219#mcetoc_1faffpsntj

[31]Big Data Solutions Blog(January 4,2019)**:** Limitations of Apache Sparkhttps://bigdatasolutionss.blogspot.com/2019/01/limitations-of-apache-spark.html

[32]Talend : Big Data in Government https://www.talend.com/resources/big-data-in-government/#:~:text=Big%20data%20in%20government%20is%20the%20influx%20of%20data%20from,and%20manage%20the%20public%20sector.

[33]Mashooque A. Memon1, Safeeullah Soomro2 , Awais K. Jumani3 and Muneer A. Kartio3, Big Data Analytics and Its Applications

[34]Ruth Dmonte, Asher Dmello (January 1,2017): Big Data in Sports (vol.6),International Journal of Engineering Research &Techonology

[35]Mallika Rangaia (December 26,2020):Applications of Big Data in Sports Industry https://www.analyticssteps.com/blogs/applications-big-data-sports-industry

[36]Maris Mirovic,Mario Milicevic, Ines Obradovic (March 2018) : Big Data in the Maritime Industry

[37]Sjoukje A.Osinga, Dilli Paudel, Spiros A. Mauzakitis, Ioannis N. Athanasiadis (January 2022) : Big data ina agriculture: Between opportunity and solution

[38]Market Trends (January7, 2021): The impact of Big Data in agriculture https://www.analyticsinsight.net/the-impact-of-big-data-in-agriculture/

[39] Nana Terra (January 27, 2021): Data Science in energy- what is your big data talent paln? https://www.airswift.com/blog/data-science-in-energy

[40] Ravi V Angadi, P.S Venkataramu, Suresh Babu Daram (January 17, 2020): Role of Big Data Analytics in Power system Application

[41] altexsoft (June 7, 2021): Hadoop Vs Spark: Main Big Data Tools Explaind https://www.altexsoft.com/blog/hadoop-vs-spark/

[42]Soumyaa Rawat (August 12, 2021): NoSQL Database: Characteristics, Types and Applications https://www.analyticssteps.com/blogs/nosql-database-characteristics-types-and-applications

[43]Robert Buda (August 27, 2021): Beyond "Fast and Simple":Top 5 Use Cases for NoSQL Database Technology https://www.budaconsulting.com/fast-and-simple-use-cases-for-nosql-database-technology/

[44] George Lawton (February 17, 2022): Hadoop vs Spark: An in-depth big data framework comparisonhttps://www.techtarget.com/searchdatamanagement/feature/Hadoop-vs-Spark-Comparing-the-two-big-data-frameworks

[45] Rajat Mark Grover, Ted Malaska, Jonathan Seidman, Gwen Shapira (July 2015): Hadoop Application Architectures, O'Reilly Media

[46]Douglas Eadline (June 2018) : Hadoop and Spark Fundamentals, Addison-Wesley Professional

[47]ScholarNest (February 2022): Spark programming in Python for Beginners with Apache Spark 3, Packt Publishing

[48] Ευάγγελος Πεφάνης (May 7, 2022): Διαδικτυακό σεμινάριο: Ενιαίο περιβάλλον ανάπτυξης εφαρμογών Python σε Pseudo Cluster με χρήση WSL2, Hadoop,Yarn-Spark.https://www.youtube.com/watch?v=GQTQm9N4DE4&t=1824s