

DIGIT~Bio~TECH



**ΛΟ5 ΕΠΙΣΤΗΜΟΝΙΚΟΙ ΠΟΡΟΙ
ΑΝΟΙΚΤΗΣ ΠΡΟΣΒΑΣΗΣ:
ΨΗΦΙΑΚΕΣ ΒΑΣΕΙΣ
ΔΕΔΟΜΕΝΩΝ**

Αρχάριο επίπεδο

ΣΥΓΓΡΑΦΕΑΣ:

FABIANO CHALHOUB & ZOI GEORGIU



Περιεχόμενα

| | |
|---------------------------------------------------------------|----|
| Επιστημονικοί πόροι ανοικτής πρόσβασης..... | 4 |
| Εισαγωγή στους πόρους «Ανοικτής Πρόσβασης»..... | 4 |
| Εισαγωγή στα δεδομένα..... | 5 |
| Τι είναι τα «δεδομένα»..... | 5 |
| Μια σύντομη ανασκόπηση των δεδομένων | 6 |
| Τα δεδομένα πριν από την εφεύρεση των υπολογιστών | 6 |
| Τα δεδομένα στην σύγχρονη εποχή | 7 |
| Κατανόηση του βασικού λεξιλογίου..... | 8 |
| Ορολογία..... | 8 |
| Τι είναι τα δεδομένα στην υπολογιστική εποχή..... | 8 |
| Τι είναι τα «μεταδεδομένα» | 10 |
| Τι είναι μια «Βάση δεδομένων»..... | 11 |
| Τι είναι οι «Πίνακες» σε μια βάση δεδομένων | 12 |
| Τι είναι οι «Στήλες» σε μια βάση δεδομένων | 12 |
| Τι είναι μια «Εγγραφή»..... | 12 |
| Τι είναι τα «Ευρετήρια»..... | 13 |
| Τι είναι ένα «Αντικείμενο» | 13 |
| Αδόμητα δεδομένα..... | 14 |
| Μεγάλα δεδομένα (Big data) | 15 |
| Αναλυτικά στοιχεία (Analytics)..... | 15 |
| Αποθήκευση..... | 16 |
| Βασική δομή μιας βάσης δεδομένων | 16 |
| Εισαγωγή..... | 16 |
| Επισκόπηση της αρχιτεκτονικής μιας βάσης δεδομένων..... | 18 |
| Κοινές ορολογίες των βάσεων δεδομένων..... | 20 |
| Ποια είναι η διαφορά μεταξύ των κύριων συστημάτων DBMS? | 21 |
| Βάσεις δεδομένων στον επιστημονικό κόσμο..... | 22 |



2019-1-BG01-KA203-062371

| | |
|------------------------------------------------------------------------|----|
| Εισαγωγή σε υπάρχουσες βάσεις δεδομένων αφιερωμένες στην επιστήμη..... | 22 |
| Τελικές σκέψεις..... | 26 |
| Βιβλιογραφικές αναφορές..... | 28 |



Επιστημονικοί πόροι ανοικτής πρόσβασης

ΕΙΣΑΓΩΓΗ ΣΤΟΥΣ ΠΟΡΟΥΣ «ΑΝΟΙΚΤΗΣ ΠΡΟΣΒΑΣΗΣ»

Η Ανοικτή πρόσβαση ή OA είναι ένα σύνολο αρχών και μια σειρά πρακτικών μέσω των οποίων τα αποτελέσματα της έρευνας διανέμονται διαδικτυακά, χωρίς κόστος ή άλλα εμπόδια πρόσβασης. Με αυστηρά καθορισμένη την ανοικτή πρόσβαση (σύμφωνα με τον ορισμό του 2001), ή δωρεάν ανοικτή πρόσβαση, τα εμπόδια στην αντιγραφή ή την επαναχρησιμοποίηση μειώνονται ή εξαλείφονται με την εφαρμογή ανοικτής άδειας για πνευματικά δικαιώματα.

Ο κύριος στόχος του κινήματος ανοικτής πρόσβασης είναι η "βιβλιογραφία έρευνας από ομοτίμους". Ιστορικά, αυτό επικεντρώθηκε κυρίως σε έντυπα ακαδημαϊκά περιοδικά. Ενώ τα συμβατικά (μη ανοιχτής πρόσβασης) περιοδικά καλύπτουν το κόστος έκδοσης μέσω διοδίων πρόσβασης, όπως συνδρομές, άδειες ιστότοπου ή χρεώσεις πληρωμής ανά προβολή, τα περιοδικά ανοικτής πρόσβασης χαρακτηρίζονται από μοντέλα χρηματοδότησης που δεν απαιτούν από τον αναγνώστη να πληρώσει για να διαβάσει περιεχόμενο του περιοδικού. Η ανοικτή πρόσβαση μπορεί να εφαρμοστεί σε όλες τις μορφές δημοσιευμένων ερευνητικών αποτελεσμάτων, συμπεριλαμβανομένων άρθρων ακαδημαϊκών περιοδικών που έχουν αξιολογηθεί και δεν έχουν αξιολογηθεί από ομοτίμους, άρθρων συνεδρίων, διπλωματικών εργασιών, κεφαλαίων βιβλίων, μονογραφιών και εικόνων.

Ωστόσο, όταν πρόκειται για τον ορισμό της "δωρεάν" πρόσβασης, πρέπει να διακρίνει το "δωρεάν" από το "libre".

Προκειμένου να αντικατοπτριστούν οι πραγματικές διαφορές στον βαθμό της ανοικτής πρόσβασης, η διάκριση μεταξύ δωρεάν ανοικτής πρόσβασης και δωρεάν ανοικτής πρόσβασης προστέθηκε το 2006 από τους Peter Suber και Stevan Harnad, δύο από τους συν-συντάκτες της αρχικής πρωτοβουλίας Ανοικτής Πρόσβασης στη Βουδαπέστη (BOAI) ορισμός της δημοσίευσης ανοικτής πρόσβασης. Η δωρεάν ανοικτή πρόσβαση αναφέρεται στη δωρεάν πρόσβαση στο διαδίκτυο και η ελεύθερη πρόσβαση στην ελεύθερη πρόσβαση στην ηλεκτρονική πρόσβαση δωρεάν, καθώς και ορισμένα πρόσθετα δικαιώματα επαναχρησιμοποίησης. Η ανοικτή πρόσβαση Libre ισοδυναμεί με τον ορισμό της ανοικτής πρόσβασης στο BOAI, τη δήλωση Bethesda σχετικά με τις εκδόσεις ανοικτής πρόσβασης και τη δήλωση του Βερολίνου για την ανοικτή πρόσβαση στη γνώση στις επιστήμες και τις ανθρωπιστικές επιστήμες. Τα δικαιώματα επαναχρησιμοποίησης του libre OA συχνά καθορίζονται από διάφορες ειδικές άδειες Creative Commons. σχεδόν όλα αυτά απαιτούν την απόδοση συγγραφής στους αρχικούς συγγραφείς.

Το έγγραφο που κυκλοφόρησε τον Φεβρουάριο του 2002 από το BOAI περιέχει τον ακόλουθο ευρέως διαδεδομένο ορισμό:



2019-1-BG01-KA203-062371

• Με τον όρο "ανοιχτή πρόσβαση" σε αυτήν τη βιβλιογραφία, εννοούμε τη δωρεάν διαθεσιμότητά της στο δημόσιο διαδίκτυο, επιτρέποντας σε όλους τους χρήστες να διαβάζουν, να κατεβάζουν, να αντιγράφουν, να διανέμουν, να εκτυπώνουν, να αναζητούν ή να συνδέουν τα πλήρη κείμενα αυτών των άρθρων, να τα ανιχνεύουν για ευρετηρίαση, να τα μεταβιβάσετε ως δεδομένα στο λογισμικό ή να τα χρησιμοποιήσετε για οποιονδήποτε άλλο νόμιμο σκοπό, χωρίς οικονομικούς, νομικούς ή τεχνικούς φραγμούς, εκτός από αυτούς που είναι αδιαχώριστοι από την απόκτηση πρόσβασης στο ίδιο το Διαδίκτυο. Ο μόνος περιορισμός για την αναπαραγωγή και διανομή και ο μόνος ρόλος για τα πνευματικά δικαιώματα σε αυτόν τον τομέα, θα πρέπει να είναι να δοθεί στους συγγραφείς ο έλεγχος της ακεραιότητας του έργου τους και του δικαιώματος να αναγνωρίζονται και να αναφέρονται σωστά.

Λαμβάνοντας υπόψη τις παραπάνω πληροφορίες, η χρήση επιστημονικών πόρων ανοιχτού κώδικα πρέπει να ακολουθεί τους κανόνες που χρησιμοποιούνται συνήθως. Η δημοσίευση επιστημονικών πόρων ανοιχτού κώδικα πρέπει επίσης να αναφέρει σαφώς εάν είναι δωρεάν ή δωρεάν και πρέπει να αποδοθούν στον αρχικό συγγραφέα.

Εισαγωγή στα δεδομένα

ΤΙ ΕΊΝΑΙ ΤΑ «ΔΕΔΟΜΕΝΑ»

Σύμφωνα με το λεξικό Merriam-Webster, υπάρχουν τρεις διαφορετικοί ορισμοί *δεδομένων*:

1. Πραγματικές πληροφορίες, όπως μετρήσεις ή στατιστικές, που χρησιμοποιούνται ως βάση για συλλογισμό, συζήτηση ή υπολογισμό
2. Πληροφορίες σε ψηφιακή μορφή που μπορούν να μεταδοθούν ή να υποβληθούν σε επεξεργασία
3. Οι πληροφορίες που παρέχονται από μια συσκευή ή όργανο ανίχνευσης που περιλαμβάνουν χρήσιμες και άσχετες ή περιττές πληροφορίες και πρέπει να υποβάλλονται σε επεξεργασία για να έχουν νόημα

Σε αυτό το έγγραφο θα καλύψουμε τους περισσότερους από τους τρεις ορισμούς.



2019-1-BG01-KA203-062371

ΜΙΑ ΣΥΝΤΟΜΗ ΑΝΑΣΚΟΠΗΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Από τότε που οι άνθρωποι άρχισαν να επικοινωνούν, βίωσαν την ανάγκη να διατηρήσουν πληροφορίες για μακροπρόθεσμο ορίζοντα. Η διατήρηση πληροφοριών ήταν απαραίτητη για τους προγόνους μας για να διασφαλίσουν την επιβίωσή τους. Η μετάδοση πληροφοριών από γενιά σε γενιά τους επέτρεψε να παρακολουθούν τους πιθανούς κινδύνους, αλλά και να καταγράφουν τα καλύτερα μέρη για τη συλλογή τροφίμων, τα καλύτερα σημεία για ψάρεμα, τα πιο ενδιαφέροντα ζώα για κυνήγι και πού να βρουν τα καλύτερα καταφύγια. Όλες αυτές οι πληροφορίες διαβιβάστηκαν προφορικά. Με την εξέλιξη της γνώσης και την εφεύρεση της γραφής, άρχισαν να αποθηκεύουν πληροφορίες σε ανεξίτηλα μέσα.

Χωρίς να υπεισέλθουμε σε λεπτομέρειες σχετικά με την εξέλιξη της αναπαράστασης της πληροφορίας, θα παρασχεθούν ορισμένα σημαντικά παραδείγματα που βοήθησαν στη δομή της σκέψης, τα οποία οδήγησαν στην ανακάλυψη των εργαλείων υπολογιστών που χρησιμοποιούμε καθημερινά.

ΤΑ ΔΕΔΟΜΕΝΑ ΠΡΙΝ ΑΠΟ ΤΗΝ ΕΦΕΥΡΕΣΗ ΤΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Καθώς εμφανίστηκαν οι ανθρώπινες κοινωνίες, τα συλλογικά κίνητρα για την ανάπτυξη της γραφής καθοδηγήθηκαν από πραγματιστικές ανάγκες. Αυτές περιλαμβάνουν την οργάνωση και τη διακυβέρνηση των κοινωνιών μέσω της διαμόρφωσης νομικών συστημάτων, συμβάσεων, ιδιοκτησιών, φορολογίας, εμπορικών συμφωνιών, συνθηκών, αρχείων απογραφής, τήρησης ιστορικού, διατήρησης πολιτισμού, παρακολούθησης επιστημονικών ανακαλύψεων, κωδικοποίησης γνώσεων μέσω αναλυτικών προγραμμάτων και καταλόγων κειμένων που είναι καλλιτεχνικά εξαιρετικά ή θεωρείται ότι περιέχουν θεμελιώδεις γνώσεις και πολλές άλλες ανάγκες.

Για παράδειγμα, γύρω στην 4η χιλιετία π.Χ., η πολυπλοκότητα του εμπορίου και της διοίκησης στη Μεσοποταμία ξεπέρασε την ανθρώπινη μνήμη και η γραφή έγινε μια πιο αξιόπιστη μέθοδος καταγραφής και παρουσίασης συναλλαγών σε μόνιμη μορφή.



Η σφηνοειδής γραφή ήταν ένα από τα πρώτα συστήματα γραφής, που εφευρέθηκε από τους Σουμερίους στην αρχαία Μεσοποταμία. Διακρίνεται από τα σφηνοειδή του σημάδια σε πήλινες πλάκες, κατασκευασμένα με τη βοήθεια ενός αμβλύ καλαμιού για μια γραφίδα, όπως αποδεικνύεται στην Εικόνα 1.

**Εικόνα1. Σφηνοειδής
γραφή**

Με την πάροδο του χρόνου, η ανάπτυξη της γνώσης, ο πολλαπλασιασμός των πληροφοριών, ο περιορισμός της ανθρώπινης μνήμης, η αναγκαιότητα γραφής και τήρησης αρχείων για τεράστιες



2019-1-BG01-KA203-062371

ποσότητες πληροφοριών έχει γίνει απαραίτητη. Ωστόσο, παρά την καταγραφή σχεδόν κάθε είδους πληροφοριών ή δεδομένων σε διάφορα μέσα, έγινε όλο και πιο περίπλοκο να ανακτηθούν με απλό τρόπο. Κάποιος έπρεπε να διαβάσει δεκάδες εκθέσεις και βιβλία για να μπορέσει να συνθέσει ένα θέμα.

ΤΑ ΔΕΔΟΜΕΝΑ ΣΤΗΝ ΣΥΓΧΡΟΝΗ ΕΠΟΧΗ

Σήμερα, η ποσότητα των δεδομένων που παράγονται κάθε χρόνο και διατηρούνται ψηφιακά, π.χ. λίστες υποχρεώσεων, συνταγές, υπενθυμίσεις, ημερολόγια, χάρτες, φωτογραφίες, e-mail, επιστημονικά δεδομένα, πολιτικές εκθέσεις, βίντεο κ.λπ. είναι τόσο εκθετικές που δημιουργεί την ανάγκη να δομήσουμε τον τρόπο με τον οποίο μπορούμε να ανακτήσουμε αυτά τα φαινομενικά μεγέθη.

Οι υπολογιστές κέρδισαν δημοτικότητα και έγιναν οικονομικά αποδοτικοί στη χρήση από ιδιώτες και ιδιωτικές εταιρείες στις αρχές της δεκαετίας του '80. Ωστόσο, η δεκαετία του '60 μπορεί να θεωρηθεί ως η νέα εποχή στον τομέα των βάσεων δεδομένων. Η εισαγωγή του όρου "βάση δεδομένων" συνέπεσε με τη διαθεσιμότητα αποθήκευσης άμεσης πρόσβασης ή DAS, από τα μέσα της δεκαετίας του '60 και μετά. Αυτή η νέα τεχνολογία αντιπροσώπευε μια αντίθεση με τις προηγούμενες κάρτες διάτρησης και τα συστήματα με βάση την ταινία, επιτρέποντας κοινή διαδραστική χρήση και όχι καθημερινή παρτίδα. Δύο κύρια μοντέλα δεδομένων αναπτύχθηκαν - το μοντέλο δικτύου "CODASYL" (Συνέδριο για τη γλώσσα του συστήματος δεδομένων) και το ιεραρχικό μοντέλο "IMS" (Σύστημα Διαχείρισης Πληροφοριών).

Η πρώτη γενιά συστημάτων βάσεων δεδομένων ήταν «πλοήγηση, σε αντίθεση με τη διαδοχική πρόσβαση λόγω των προηγούμενων τεχνολογιών που χρησιμοποιήθηκαν για την αποθήκευση δεδομένων, δηλαδή ταινίες και κάρτες διάτρησης. Οι εφαρμογές συνήθως είχαν πρόσβαση στα δεδομένα ακολουθώντας δείκτες από τη μία εγγραφή στην άλλη. Τα στοιχεία αποθήκευσης εξαρτώνταν από τον τύπο των δεδομένων που θα αποθηκευτούν.

Η προσθήκη ενός επιπλέον πεδίου σε μια βάση δεδομένων απαιτούσε επανεγγραφή του υποκείμενου σχεδίου πρόσβασης/τροποποίησης. Έμφαση δόθηκε στα αρχεία που θα επεξεργαστούν και όχι στη συνολική δομή του συστήματος. Ένας χρήστης θα πρέπει να γνωρίζει τη φυσική δομή της βάσης δεδομένων για να ζητήσει πληροφορίες. Μια βάση δεδομένων που αποδείχθηκε εμπορική επιτυχία ήταν το σύστημα «SABER» που χρησιμοποιήθηκε από την IBM για να βοηθήσει τις American Airlines να διαχειριστούν τα δεδομένα των κρατήσεων της. Αυτό το σύστημα εξακολουθεί να χρησιμοποιείται από τις μεγάλες ταξιδιωτικές υπηρεσίες για τα συστήματα κρατήσεων τους.

Στη σύγχρονη τεχνολογία πληροφοριών, υπήρχε πάντα σύγκυση μεταξύ των χρηστών μεταξύ βάσεων δεδομένων και διαδικτυακών μηχανών αναζήτησης στο διαδίκτυο στα οποία έχουν πρόσβαση τα προγράμματα περιήγησης. Μια βάση δεδομένων περιέχει συνήθως δομημένα δεδομένα, σε αντίθεση με τον Παγκόσμιο Ιστό (www), που συνήθως περιέχει μη δομημένα δεδομένα. Ακόμα κι αν η ανάκτηση



2019-1-BG01-KA203-062371

πληροφοριών τόσο από τις βάσεις δεδομένων όσο και από το "www" είναι απρόσκοπτη και μοιάζει παρόμοια, το περιεχόμενο και ο τρόπος με τον οποίο απευθύνονται τα ερωτήματα είναι εντελώς διαφορετικά. Τα δομημένα και μη δομημένα δεδομένα θα εξηγηθούν αργότερα σε αυτό το έγγραφο.

Κατανόηση του βασικού λεξιλογίου

ΟΡΟΛΟΓΙΑ

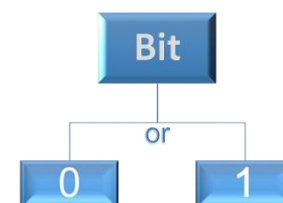
Όπως κάθε άλλη επιστήμη, η επιστήμη των υπολογιστών έχει τη δική της γλώσσα. Για να κατανοήσετε πλήρως τις πληροφορίες που θα παρέχονται σε αυτό το έγγραφο, είναι απαραίτητο να εξοικειωθείτε με το λεξιλόγιο που σχετίζεται με αυτό το θέμα.

Επιπλέον, η επικοινωνία με έναν DBA (Διαχειριστής βάσης δεδομένων) θα διευκολυνθεί. Όταν ένας Βιοχημικός θα πρέπει να εκφράσει τις ανάγκες του όσον αφορά τη δομή ή τη διαχείριση δεδομένων σε μια βάση δεδομένων, θα μπει στον πειρασμό να χρησιμοποιήσει τη δική του τεχνική γλώσσα. Στη συνέχεια, το DBA θα πρέπει να κατανοήσει το αίτημα και να το μετατρέψει σε γλώσσα υπολογιστή, η οποία θα είναι κατανοητή από τους βιοχημικούς.

Τι είναι τα δεδομένα στην υπολογιστική εποχή

Όπως αναφέρεται στην ενότητα 2.1, σύμφωνα με τον τομέα στον οποίο αναφέρεται, τα δεδομένα ενδέχεται να έχουν διαφορετική σημασία. Στην περίπτωση των υπολογιστών και των βάσεων δεδομένων, τα δεδομένα ορίζονται ως οποιαδήποτε ακολουθία ενός ή περισσότερων συμβόλων. Τα δεδομένα απαιτούν ερμηνεία για να γίνουν πληροφορίες. Στην τεχνολογία των πληροφοριών, το "bit" είναι η μικρότερη ποσότητα δεδομένων. Ένα κομμάτι είναι δυαδικό. Οι δυαδικοί αριθμοί είναι μια αναπαράσταση αριθμών που χρησιμοποιούν μόνο δύο ψηφία, 0 και 1 (Εικ. 2). Είναι ένα αριθμητικό σύστημα βάσης 2, δηλαδή:

- 0 0 0 1 = αριθμητική τιμή 2^0
- 0 0 1 0 = αριθμητική τιμή 2^1
- 0 1 0 0 = αριθμητική τιμή 2^2
- 1 0 0 0 = αριθμητική τιμή 2^3



Εικόνα 2. Ένα bit μπορεί να είναι 0 ή 1



2019-1-BG01-KA203-062371

Μια ακολουθία "bits" αποτελεί "Byte". Τα bytes αποτελούνται από πολλαπλάσιο των 4 bits (ένα byte των 4 bits ονομάζεται Nibble) όπως στο παραπάνω παράδειγμα. Σήμερα, το byte είναι μια μονάδα ψηφιακών πληροφοριών που συνηθέστερα αποτελείται από οκτώ bit. Ιστορικά, το byte ήταν ο αριθμός των bit που χρησιμοποιήθηκαν για την κωδικοποίηση ενός χαρακτήρα κειμένου σε έναν υπολογιστή. Με ένα byte οκτώ bits, ο μέγιστος δεκαδικός αριθμός είναι 256. Ιστορικά, το byte είναι επίσης η μονάδα πληροφοριών υπολογιστή ή χωρητικότητας αποθήκευσης δεδομένων που χρησιμοποιείται για τη μέτρηση της ποσότητας δεδομένων (Πίνακας 1).

Πίνακας 1

| Abbreviation | Unit | Value | Size in bytes |
|--------------|-----------|------------------------|-----------------------------------------|
| b | bit | 0 or 1 | 1/8 of a byte |
| B | bytes | 8 bits | 1 byte |
| KB | kilobytes | 10 ³ bytes | 1,000 bytes |
| MB | megabyte | 10 ⁶ bytes | 1,000,000 bytes |
| GB | gigabyte | 10 ⁹ bytes | 1,000,000,000 bytes |
| TB | terabyte | 10 ¹² bytes | 1,000,000,000,000 bytes |
| PB | petabyte | 10 ¹⁵ bytes | 1,000,000,000,000,000 bytes |
| EB | exabyte | 10 ¹⁸ bytes | 1,000,000,000,000,000,000 bytes |
| ZB | zettabyte | 10 ²¹ bytes | 1,000,000,000,000,000,000,000 bytes |
| YB | yottabyte | 10 ²⁴ bytes | 1,000,000,000,000,000,000,000,000 bytes |

Πίνακας 2. ASCII πίνακας

| BINARY BYTE OF 8 BIT | DECIMAL | HEXADECIMAL | CHARACTER | BINARY BYTE OF 8 BIT | DECIMAL | HEXADECIMAL | CHARACTER | BINARY BYTE OF 8 BIT | DECIMAL | HEXADECIMAL | CHARACTER | BINARY BYTE OF 8 BIT | DECIMAL | HEXADECIMAL | CHARACTER |
|----------------------|---------|-------------|-----------|----------------------|---------|-------------|-----------|----------------------|---------|-------------|-----------|----------------------|---------|-------------|-----------|
| 0 0 0 0 0 0 0 0 | 0 | 00 | NULL | 0 0 1 0 0 0 0 0 | 32 | 20 | SPACE | 0 1 0 0 0 0 0 0 | 64 | 41 | @ | 1 0 1 0 0 0 0 0 | 96 | 96 | ` |
| 0 0 0 0 0 0 0 1 | 1 | 01 | SOH | 0 0 1 0 0 0 0 1 | 33 | 21 | ! | 0 1 0 0 0 0 0 1 | 65 | 42 | A | 1 0 1 0 0 0 0 1 | 97 | 97 | a |
| 0 0 0 0 0 0 1 0 | 2 | 02 | STX | 0 0 1 0 0 0 1 0 | 34 | 22 | " | 0 1 0 0 0 0 1 0 | 66 | 43 | B | 1 0 1 0 0 0 1 0 | 98 | 98 | b |
| 0 0 0 0 0 0 1 1 | 3 | 03 | ETX | 0 0 1 0 0 0 1 1 | 35 | 23 | # | 0 1 0 0 0 0 1 1 | 67 | 44 | C | 1 0 1 0 0 0 1 1 | 99 | 99 | c |
| 0 0 0 0 0 1 0 0 | 4 | 04 | EDT | 0 0 1 0 0 1 0 0 | 36 | 24 | \$ | 0 1 0 0 0 1 0 0 | 68 | 45 | D | 1 0 1 0 0 1 0 0 | 100 | 9A | d |
| 0 0 0 0 0 1 0 1 | 5 | 05 | ENQ | 0 0 1 0 0 1 0 1 | 37 | 25 | % | 0 1 0 0 0 1 0 1 | 69 | 46 | E | 1 0 1 0 0 1 0 1 | 101 | 9B | e |
| 0 0 0 0 0 1 1 0 | 6 | 06 | ACK | 0 0 1 0 0 1 1 0 | 38 | 26 | & | 0 1 0 0 0 1 1 0 | 70 | 47 | F | 1 0 1 0 0 1 1 0 | 102 | 9C | f |
| 0 0 0 0 0 1 1 1 | 7 | 07 | BELL | 0 0 1 0 0 1 1 1 | 39 | 27 | ' | 0 1 0 0 0 1 1 1 | 71 | 48 | G | 1 0 1 0 0 1 1 1 | 103 | 9D | g |
| 0 0 0 0 1 0 0 0 | 8 | 08 | BS | 0 0 1 0 1 0 0 0 | 40 | 28 | (| 0 1 0 0 1 0 0 0 | 72 | 49 | H | 1 0 1 0 1 0 0 0 | 104 | 9E | h |
| 0 0 0 0 1 0 0 1 | 9 | 09 | TAB | 0 0 1 0 1 0 0 1 | 41 | 29 |) | 0 1 0 0 1 0 0 1 | 73 | 4A | I | 1 0 1 0 1 0 0 1 | 105 | 9F | i |
| 0 0 0 0 1 0 1 0 | 10 | 0A | LF | 0 0 1 0 1 0 1 0 | 42 | 2A | * | 0 1 0 0 1 0 1 0 | 74 | 4B | J | 1 0 1 0 1 0 1 0 | 106 | A0 | j |
| 0 0 0 0 1 0 1 1 | 11 | 0B | VT | 0 0 1 0 1 0 1 1 | 43 | 2B | + | 0 1 0 0 1 0 1 1 | 75 | 4C | K | 1 0 1 0 1 0 1 1 | 107 | A1 | k |
| 0 0 0 0 1 1 0 0 | 12 | 0C | FF | 0 0 1 0 1 1 0 0 | 44 | 2C | , | 0 1 0 0 1 1 0 0 | 76 | 4D | L | 1 0 1 0 1 1 0 0 | 108 | A2 | l |
| 0 0 0 0 1 1 0 1 | 13 | 0D | CR | 0 0 1 0 1 1 0 1 | 45 | 2D | - | 0 1 0 0 1 1 0 1 | 77 | 4E | M | 1 0 1 0 1 1 0 1 | 109 | A3 | m |
| 0 0 0 0 1 1 1 0 | 14 | 0E | SO | 0 0 1 0 1 1 1 0 | 46 | 2F | . | 0 1 0 0 1 1 1 0 | 78 | 4F | N | 1 0 1 0 1 1 1 0 | 110 | A4 | n |
| 0 0 0 0 1 1 1 1 | 15 | 0F | SI | 0 0 1 0 1 1 1 1 | 47 | 30 | / | 0 1 0 0 1 1 1 1 | 79 | 50 | O | 1 0 1 0 1 1 1 1 | 111 | A5 | o |
| 0 0 0 1 0 0 0 0 | 16 | 10 | DLE | 0 0 1 1 0 0 0 0 | 48 | 31 | 0 | 0 1 0 1 0 0 0 0 | 80 | 51 | P | 1 0 1 1 0 0 0 0 | 112 | A6 | p |
| 0 0 0 1 0 0 0 1 | 17 | 11 | DC1 | 0 0 1 1 0 0 0 1 | 49 | 32 | 1 | 0 1 0 1 0 0 0 1 | 81 | 52 | Q | 1 0 1 1 0 0 0 1 | 113 | A7 | q |
| 0 0 0 1 0 0 1 0 | 18 | 12 | DC2 | 0 0 1 1 0 0 1 0 | 50 | 33 | 2 | 0 1 0 1 0 0 1 0 | 82 | 53 | R | 1 0 1 1 0 0 1 0 | 114 | A8 | r |
| 0 0 0 1 0 0 1 1 | 19 | 13 | DC3 | 0 0 1 1 0 0 1 1 | 51 | 34 | 3 | 0 1 0 1 0 0 1 1 | 83 | 54 | S | 1 0 1 1 0 0 1 1 | 115 | A9 | s |
| 0 0 0 1 0 1 0 0 | 20 | 14 | DC4 | 0 0 1 1 0 1 0 0 | 52 | 35 | 4 | 0 1 0 1 0 1 0 0 | 84 | 55 | T | 1 0 1 1 0 1 0 0 | 116 | AA | t |
| 0 0 0 1 0 1 0 1 | 21 | 15 | NAK | 0 0 1 1 0 1 0 1 | 53 | 36 | 5 | 0 1 0 1 0 1 0 1 | 85 | 56 | U | 1 0 1 1 0 1 0 1 | 117 | AB | u |
| 0 0 0 1 0 1 1 0 | 22 | 16 | SYN | 0 0 1 1 0 1 1 0 | 54 | 37 | 6 | 0 1 0 1 0 1 1 0 | 86 | 57 | V | 1 0 1 1 0 1 1 0 | 118 | AC | v |
| 0 0 0 1 0 1 1 1 | 23 | 17 | ETB | 0 0 1 1 0 1 1 1 | 55 | 38 | 7 | 0 1 0 1 0 1 1 1 | 87 | 58 | W | 1 0 1 1 0 1 1 1 | 119 | AD | w |
| 0 0 0 1 1 0 0 0 | 24 | 18 | CAN | 0 0 1 1 1 0 0 0 | 56 | 39 | 8 | 0 1 0 1 1 0 0 0 | 88 | 59 | X | 1 0 1 1 1 0 0 0 | 120 | AE | x |
| 0 0 0 1 1 0 0 1 | 25 | 19 | EM | 0 0 1 1 1 0 0 1 | 57 | 3A | 9 | 0 1 0 1 1 0 0 1 | 89 | 5A | Y | 1 0 1 1 1 0 0 1 | 121 | AF | y |
| 0 0 0 1 1 0 1 0 | 26 | 1A | SUB | 0 0 1 1 1 0 1 0 | 58 | 3B | : | 0 1 0 1 1 0 1 0 | 90 | 5B | Z | 1 0 1 1 1 0 1 0 | 122 | B0 | z |
| 0 0 0 1 1 0 1 1 | 27 | 1B | FSC | 0 0 1 1 1 0 1 1 | 59 | 3C | ; | 0 1 0 1 1 0 1 1 | 91 | 5C | [| 1 0 1 1 1 0 1 1 | 123 | B1 | { |
| 0 0 0 1 1 1 0 0 | 28 | 1C | FS | 0 0 1 1 1 1 0 0 | 60 | 3D | < | 0 1 0 1 1 1 0 0 | 92 | 5D | \ | 1 0 1 1 1 1 0 0 | 124 | B2 | |
| 0 0 0 1 1 1 0 1 | 29 | 1D | GS | 0 0 1 1 1 1 0 1 | 61 | 3E | = | 0 1 0 1 1 1 0 1 | 93 | 5E |] | 1 0 1 1 1 1 0 1 | 125 | B3 | } |
| 0 0 0 1 1 1 1 0 | 30 | 1E | RS | 0 0 1 1 1 1 1 0 | 62 | 3F | > | 0 1 0 1 1 1 1 0 | 94 | 5F | ^ | 1 0 1 1 1 1 1 0 | 126 | B4 | ~ |
| 0 0 0 1 1 1 1 1 | 31 | 1F | US | 0 0 1 1 1 1 1 1 | 63 | 40 | ? | 0 1 0 1 1 1 1 1 | 95 | 60 | _ | 1 0 1 1 1 1 1 1 | 127 | B5 | DEL |

Ένα παράδειγμα χρήσης είναι ο πίνακας χαρακτήρων ASCII (American Standard Code for Information Interchange) που χρησιμοποιείται συνήθως για αλφαβητικούς χαρακτήρες (Πίνακας 2). Οι πρώτοι 32 χαρακτήρες ονομάζονται χαρακτήρες ελέγχου. Αρχικά, δεν σχεδιάστηκαν για να αντιπροσωπεύουν εκτυπώσιμες πληροφορίες, αλλά για τον έλεγχο συσκευών που χρησιμοποιούν κώδικα ASCII, όπως εκτυπωτές, ή για την παροχή μετα-πληροφοριών σχετικά με τις ροές δεδομένων, π.χ. αυτές που είναι αποθηκευμένες σε μαγνητική ταινία.



2019-1-BG01-KA203-062371

Τι είναι τα «μεταδεδομένα»

Τα μεταδεδομένα, ή, απλά, οι μετα-πληροφορίες, χρησιμοποιούνται για την αναφορά των δεδομένων σχετικά με τα δεδομένα. Η κατοχή δεδομένων δεν αρκεί για να τα θέσουμε απλά στο διαδίκτυο. Τα δεδομένα δεν μπορούν να χρησιμοποιηθούν έως ότου εξηγηθούν με τρόπο που μπορούν να επεξεργαστούν τόσο οι άνθρωποι όσο και οι υπολογιστές.



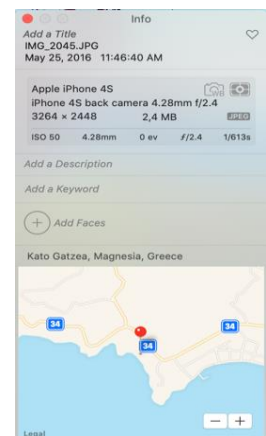
Εικόνα 3. Φωτογραφία στην Ελλάδα

Τα μεταδεδομένα μπορεί να υπονοούνται, να καθορίζονται ή να δίνονται. Περιλαμβάνει δεδομένα που σχετίζονται με φυσικά γεγονότα ή διαδικασίες και θα έχει επίσης ένα χρονικό συστατικό. Σε όλες σχεδόν τις περιπτώσεις αυτό το χρονικό στοιχείο υπονοείται. Μπορεί να είναι λίγο δύσκολο να το καταλάβουμε, ωστόσο, το ακόλουθο παράδειγμα θα δώσει μια σαφέστερη εξήγηση αυτού του όρου.

Φανταστείτε ότι ταξιδεύετε με το αγαπημένο σας smartphone σε κάποιο παράδεισο νησί. Αρχίζετε να τραβάτε φωτογραφίες (Εικ. 3) για να κρατήσετε ωραία αρχεία του ταξιδιού σας. Μια εβδομάδα αργότερα, το ταξίδι σας φτάνει στο τέλος του και πρέπει να επιστρέψετε στο σπίτι.

Επιστρέφοντας στο σπίτι, καλείτε τους καλύτερους φίλους σας για ένα πάρτι και θέλετε να μοιραστείτε μαζί τους τις ομορφιές που έχετε δει κατά τη διάρκεια του ταξιδιού σας. Αρχίζετε να εμφανίζετε τις εικόνες, αλλά δεν μπορείτε να θυμηθείτε ποια μέρα, τι ώρα και πού τραβήχτηκαν μερικές από αυτές. Αυτό είναι όπου τα μεταδεδομένα των εικόνων μπορούν να βοηθήσουν. Με λίγα λόγια, είναι η περιγραφή των δεδομένων. Σε αυτό το παράδειγμα, η εικόνα είναι τα δεδομένα και η περιγραφή της εικόνας είναι τα μεταδεδομένα (Εικ. 4).

Στη Βιοτεχνολογία, πρέπει να καταλάβουμε ότι τα μεταδεδομένα είναι μακράν πιο σημαντικά από τα δεδομένα. Είναι πολύ απλό να κατανοήσουμε τον λόγο για τον οποίο τα μεταδεδομένα είναι ένα κρίσιμο συστατικό που σχετίζεται άμεσα με τα δεδομένα. Φανταστείτε ένα πείραμα που θα οδηγήσει σε ένα συγκεκριμένο αποτέλεσμα. Αυτό το πείραμα, για να είναι έγκυρο, πρέπει να τεκμηριωθεί. Αυτή η τεκμηρίωση πρέπει να περιλαμβάνει όλες τις συνθήκες, υπό τις οποίες πραγματοποιήθηκε το πείραμα. Αυτό μπορεί να περιλαμβάνει την περιγραφή του είδους της πρώτης ύλης που χρησιμοποιήθηκε, την πηγή της, υπό ποιες συνθήκες συλλέχθηκε, τους τύπους μηχανών για την επεξεργασία του πειράματος, τη θερμοκρασία, την ημερομηνία, την ώρα κ.λπ. Για να είναι το



**Εικόνα 4.
Μεταδεδομένα
της φωτογραφίας**



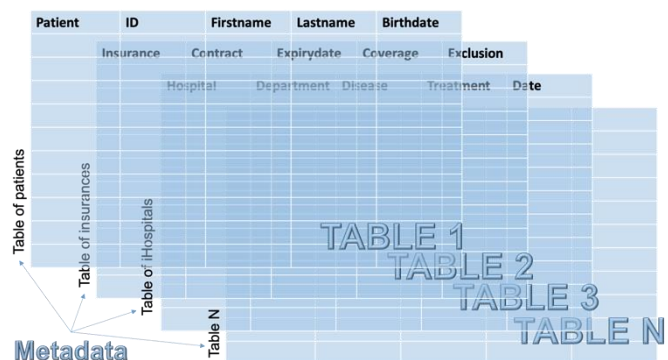
2019-1-BG01-KA203-062371

αποτέλεσμα αυτού του πειράματος συγκρίσιμο με άλλα αποτελέσματα παρόμοιων πειραμάτων, όλες οι συνθήκες πρέπει να είναι παρόμοιες. Τα ωμά δεδομένα χωρίς μεταδεδομένα είναι άχρηστα.

Η μεγαλύτερη πρόκληση στη βιοτεχνολογία, και σε οποιαδήποτε άλλη επιστήμη, είναι η τυποποίηση των μεταδεδομένων. Στις περισσότερες βάσεις δεδομένων της βιοτεχνολογίας, αυτό δεν τηρείται. Πρέπει κανείς να έχει απόλυτη συνείδηση αυτού του φαινομένου και να σέβεται πλήρως τα πρότυπα.

Τι είναι μια «Βάση δεδομένων»

Σε γενικές γραμμές, μια βάση δεδομένων ορίζεται ως μια συλλογή στοιχείων δεδομένων, όπως τηλεφωνικοί κατάλογοι, τιμοκατάλογοι, λίστες αποθεμάτων, διευθύνσεις πελατών κ.λπ. Ωστόσο, από τεχνική άποψη, μια βάση δεδομένων αναφέρεται ως «μια αυτο-περιγραφόμενη συλλογή ολοκληρωμένων ρεκόρ». Υπνοεί τεχνολογία υπολογιστών, συμπληρωμένη με μια συγκεκριμένη γλώσσα υπολογιστών, όπως η SQL (Structured Query Language).



Μια βάση δεδομένων αποτελείται από πολλούς πίνακες (Εικ. 5) και από δεδομένα και μεταδεδομένα. Τα μεταδεδομένα είναι τα δεδομένα που περιγράφουν τη δομή των δεδομένων μέσα σε μια βάση δεδομένων. Εάν γνωρίζετε πώς είναι τακτοποιημένα τα δεδομένα σας, τότε μπορείτε να τα ανακτήσετε. Δεδομένου ότι η βάση δεδομένων περιέχει μια περιγραφή της δικής της δομής, αναφέρεται ως *αυτο-περιγραφόμενη*. Η βάση δεδομένων είναι ενσωματωμένη επειδή δεν περιλαμβάνει μόνο στοιχεία δεδομένων αλλά και τις σχέσεις μεταξύ τους.

Εικόνα 5. Μερική δομή της βάσης δεδομένων

Η βάση δεδομένων αποθηκεύει μεταδεδομένα σε μια περιοχή που ονομάζεται λεξικό δεδομένων, η οποία περιγράφει πίνακες, στήλες, ευρετήρια, περιορισμούς και άλλα στοιχεία που αποτελούν τη βάση δεδομένων.

Επειδή ένα επίπεδο αρχείο αρχείων, δηλαδή το "Spreadsheet" δεν έχει μεταδεδομένα, οι εφαρμογές που έχουν γραφτεί για να λειτουργούν με επίπεδα αρχεία πρέπει να περιέχουν το ισοδύναμο των μεταδεδομένων ως μέρος του προγράμματος εφαρμογής.



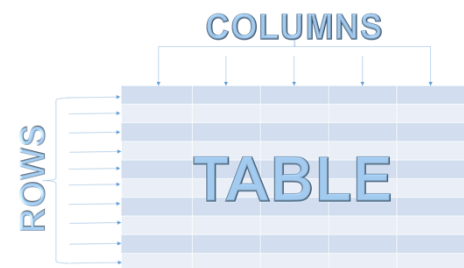
2019-1-BG01-KA203-062371

Τι είναι οι «Πίνακες» σε μια βάση δεδομένων

Ένας πίνακας είναι μια συλλογή σχετικών δεδομένων που διατηρούνται σε μορφή πίνακα που αποτελείται από στήλες και σειρές μέσα σε μια βάση δεδομένων. Μοιάζει με υπολογιστικό φύλλο (Εικ. 6).

Τι είναι οι «Στήλες» σε μια βάση δεδομένων

Μια στήλη είναι ένα σύνολο τιμών δεδομένων, όλες ενός τύπου, σε έναν πίνακα. Οι στήλες ορίζουν τα δεδομένα σε έναν πίνακα. Οι περισσότερες βάσεις δεδομένων επιτρέπουν στις στήλες να περιέχουν πολύπλοκα δεδομένα, όπως εικόνες, ολόκληρα έγγραφα ή ακόμη και βίντεο κλιπ. Επομένως, μια στήλη που επιτρέπει τιμές δεδομένων ενός τύπου δεν σημαίνει απαραίτητα ότι έχει μόνο απλές τιμές κειμένου. Ορισμένες βάσεις δεδομένων προχωρούν ακόμη περισσότερο και επιτρέπουν την αποθήκευση των δεδομένων ως αρχείο στο λειτουργικό σύστημα, ενώ τα δεδομένα της στήλης περιέχουν μόνο δείκτη ή σύνδεσμο προς το πραγματικό αρχείο. Αυτό γίνεται με σκοπό τη διατήρηση του συνολικού μεγέθους της βάσης δεδομένων - ένα μικρότερο μέγεθος βάσης δεδομένων σημαίνει λιγότερο χρόνο που απαιτείται για τη δημιουργία αντιγράφων ασφαλείας και λιγότερο χρόνο που απαιτείται για την αναζήτηση δεδομένων μέσα στη βάση δεδομένων.



Εικόνα 6. Πίνακας με στήλες και σειρές

Σε έναν πίνακα, σε κάθε στήλη εκχωρείται τυπικά ένας τύπος δεδομένων και άλλοι περιορισμοί, οι οποίοι καθορίζουν τον τύπο της τιμής που μπορεί να αποθηκευτεί σε αυτήν τη στήλη. Για παράδειγμα, μια στήλη ενδέχεται να δέχεται διευθύνσεις ηλεκτρονικού ταχυδρομείου και μια άλλη μπορεί να δέχεται αριθμούς τηλεφώνου με περιορισμό 10 ψηφίων.

Τι είναι μια «Εγγραφή»

Μια εγγραφή είναι μια αναπαράσταση ενός φυσικού ή εννοιολογικού αντικειμένου. Πείτε, για παράδειγμα, ότι θέλετε να παρακολουθείτε τους πελάτες μιας επιχείρησης. Εκχωρείτε μια εγγραφή για κάθε πελάτη. Κάθε εγγραφή έχει πολλά χαρακτηριστικά, όπως όνομα, διεύθυνση και αριθμό τηλεφώνου. Μεμονωμένα ονόματα, διευθύνσεις και ούτω καθεξής είναι τα δεδομένα.

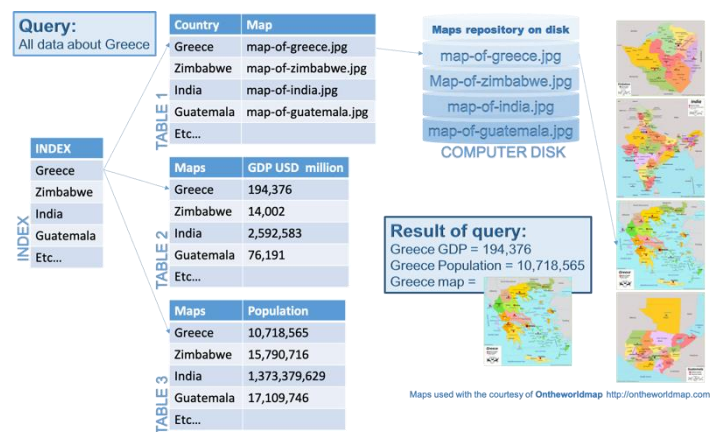


2019-1-BG01-KA203-062371

Τι είναι τα «Ευρετήρια»

Τα δομημένα δεδομένα αποθηκεύονται με τη μορφή εγγραφών σε μια βάση δεδομένων. Κάθε εγγραφή έχει ένα βασικό πεδίο, το οποίο το βοηθά να αναγνωρίζεται μοναδικά, δηλαδή το αναγνωριστικό ενός ασθενούς. Κανένας άλλος ασθενής δεν μπορεί να έχει τον ίδιο αριθμό ταυτότητας, αλλά άλλος ασθενής μπορεί να έχει το ίδιο όνομα και επώνυμο.

Η ευρετηρίαση μιας βάσης δεδομένων είναι μια τεχνική για την αποτελεσματική ανάκτηση εγγραφών από τα αρχεία της βάσης δεδομένων, με βάση ορισμένα χαρακτηριστικά στα οποία έχει εκτελεστεί η ευρετηρίαση. Για να το κάνουμε απλό, η ευρετηρίαση στα συστήματα βάσεων δεδομένων είναι παρόμοια με αυτήν που συνήθως βλέπουμε στα βιβλία. Στην αρχή ή στο τέλος ενός βιβλίου, μπορεί να βρεθεί ένα ευρετήριο (το οποίο διαφέρει από έναν πίνακα περιεχομένων), το οποίο παρέχει όλους τους αριθμούς σελίδων για ένα συγκεκριμένο θέμα. Για παράδειγμα, ένας Άτλας μπορεί να χωριστεί σε κεφάλαια που περιέχουν χάρτες, κεφάλαια που περιέχουν δεδομένα για τον πληθυσμό και κεφάλαια αφιερωμένα σε δεδομένα παραγωγής ή γεωργίας χωρών. Εάν ψάχνετε για μια συγκεκριμένη χώρα και θα θέλατε να έχετε μια επισκόπηση όλων των δεδομένων που αφορούν τη συγκεκριμένη χώρα, το ευρετήριο μπορεί να είναι πολύ χρήσιμο καθώς θα σας δείχνει τη σελίδα που σχετίζεται με τη συγκεκριμένη χώρα σε κάθε κεφάλαιο (Εικ. 7).



Εικόνα 7. Παράδειγμα ευρετηρίων

Τι είναι ένα «Αντικείμενο»

Στην επιστήμη των υπολογιστών, ένα αντικείμενο μπορεί να είναι μια μεταβλητή, μια δομή δεδομένων, μια συνάρτηση ή μια μέθοδος και, ως εκ τούτου, είναι μια τιμή στη μνήμη που αναφέρεται από ένα αναγνωριστικό. Στο σχεσιακό μοντέλο διαχείρισης βάσης δεδομένων, ένα αντικείμενο μπορεί να είναι ένας πίνακας ή στήλη, ή μια συσχέτιση μεταξύ δεδομένων και οντότητας βάσης δεδομένων, όπως η σχέση ηλικίας ενός ατόμου με ένα συγκεκριμένο άτομο.



2019-1-BG01-KA203-062371

Δομημένα δεδομένα

Σύμφωνα με το SNIA (Storage Networking Industry Association), τα δομημένα δεδομένα ορίζονται ως:

"Δεδομένα που οργανώνονται και διαμορφώνονται με γνωστό και σταθερό τρόπο. "

Η μορφή και η οργάνωση ορίζονται συνήθως σε ένα σχήμα. Ο όρος δομημένα δεδομένα συνήθως νοείται ως δεδομένα που παράγονται και διατηρούνται από βάσεις δεδομένων και επιχειρηματικές εφαρμογές.

Απαιτούνται τρεις προϋποθέσεις για να περιγραφούν τα δεδομένα ως δομημένα:

- Πρέπει να συμμορφώνεται με ένα μοντέλο δεδομένων,
- Πρέπει να έχει μια καλά καθορισμένη δομή,
- Πρέπει να ακολουθεί μια συνεπή σειρά και να είναι εύκολα προσβάσιμη και να χρησιμοποιείται από ένα άτομο ή ένα πρόγραμμα υπολογιστή.

Τα δομημένα δεδομένα συνήθως αποθηκεύονται σε καλά καθορισμένα σχήματα όπως οι βάσεις δεδομένων. Γενικά είναι πίνακας με στήλες και σειρές που καθορίζουν σαφώς τα χαρακτηριστικά του (Εικ. 8).

Η SQL (γλώσσα δομημένου ερωτήματος) χρησιμοποιείται συχνά για τη διαχείριση δομημένων δεδομένων που είναι αποθηκευμένα σε βάσεις δεδομένων.

Αδόμητα δεδομένα

Οι πληροφορίες που δεν είναι οργανωμένες σε ένα προκαθορισμένο μοντέλο ονομάζονται μη δομημένα δεδομένα ή μη δομημένες πληροφορίες. Στην επιστήμη των υπολογιστών, αρχεία όπως αρχεία κειμένου, φωτογραφίες, αρχεία βίντεο, αρχεία ήχου και παρουσιάσεις θεωρούνται αρχεία χωρίς δομή. Συνήθως, ένα αρχείο PDF περιέχει μη δομημένα δεδομένα (Εικ. 9).

Υπολογίζεται ότι το 80 έως 90% των παγκόσμιων συνολικών απούλοποιημένων δεδομένων είναι αδόμητο. Οι συνηθισμένοι αλγόριθμοι ερωτημάτων δεν είναι σε θέση να εξαγάγουν απλά και

Structured data

| First Name | Last Name | Citizenship | City | Etc... |
|------------|-------------|-------------|----------|--------|
| Fyodor | Dostoevsky | Russian | Moscow | ... |
| Albert | Camus | French | Mondovi | ... |
| Dante | Alighieri | Italian | Florence | ... |
| Nikos | Kazantzakis | Greek | Kandiye | ... |

Εικόνα 8. Δομημένα δεδομένα

Unstructured data

Some famous writers like M. Fyodor Dostoevsky, who is Russian origin, born in Moscow wrote the famous book *Crime and Punishment*, are worldwide known. We can also cite M. Dante Alighieri who was born in Florence, Italy, and wrote the non less renown book *The Divine Comedy*. Other countries like Greece have great authors. The citizen of Kandiye in Crete M. Nikos Kazantzakis wrote the internationally known book *Zorba The Greek* which was adapted for the cinema in 1965 and starring Anthony Queen. M. Albert Camus, born in the town of Mondovi in Algeria and French citizen whose book *The Plague* was edited in 1947 unfortunately died at the age of 46 in a car accident.

Εικόνα 9. Αρχείο PDF



2019-1-BG01-KA203-062371

αποτελεσματικά τις απαιτούμενες πληροφορίες από ένα μη δομημένο αρχείο, όπως στο παράδειγμα του σχήματος 9. Οι ίδιες πληροφορίες που περιέχονται στο σχήμα 9 μπορούν εύκολα να ανακτηθούν με ένα ερώτημα. Ωστόσο, σήμερα, διατίθενται μη δομημένα εργαλεία ανάλυσης δεδομένων που υποστηρίζονται από τεχνητή νοημοσύνη (AI), τα οποία δημιουργήθηκαν ειδικά για πρόσβαση στις διαθέσιμες πληροφορίες από μη δομημένα δεδομένα (βλ. 3.1.12 Analytics).

Μεγάλα δεδομένα (Big data)

Σύμφωνα με το SNIA (Storage Networking Industry Association), τα *μεγάλα δεδομένα* ορίζονται ως:

"Ένας χαρακτηρισμός συνόλων δεδομένων που είναι πολύ μεγάλα για να υποστούν αποτελεσματική επεξεργασία στο σύνολό τους από τις πιο ισχυρές τυπικές διαθέσιμες υπολογιστικές πλατφόρμες."

Με άλλα λόγια, τα Big Data αναφέρονται σε τεράστιες ποσότητες δομημένων ή μη δομημένων δεδομένων που δεν μπορούν να υποβληθούν σε επεξεργασία από το συνηθισμένο λογισμικό ως παραδοσιακή γλώσσα ερωτήματος βάσης δεδομένων ή οποιοδήποτε άλλο είδος λήψης μηχανών.

Υπάρχει σύγχυση σχετικά με την τρέχουσα χρήση των όρων Big Data και Analytics. Τα Big Data είναι οι πληροφορίες, ενώ το Analytics είναι ο τρόπος εξαγωγής των επιθυμητών πληροφοριών από τεράστιες ποσότητες διαθέσιμων πληροφοριών.

Αναλυτικά στοιχεία (Analytics)

Στην τεχνολογία υπολογιστών, το Analytics είναι μια μέθοδος εξαγωγής αξίας από μεγάλα δεδομένα.

Στον τομέα της υγειονομικής περίθαλψης, η Big Data Analytics έχει οδηγήσει σε πολλές βελτιώσεις παρέχοντας εξατομικευμένη ιατρική και προγνωστικά αναλυτικά στοιχεία. Καθώς ο όγκος των δεδομένων αυξάνεται δραματικά, οι παραδοσιακές βάσεις δεδομένων και οι μηχανές αναζήτησης δεν θα είναι σε θέση να χειριστούν και να ανακτήσουν συγκεκριμένες πληροφορίες. Τα δεδομένα των ασθενών δημιουργούνται από μαγνητική τομογραφία, ακτίνες X, μηχανήματα αιματολογικών εξετάσεων, αισθητήρες παρακολούθησης και πολλές άλλες πηγές συμπλεγμάτων δεδομένων προς επεξεργασία. Εκτενείς πληροφορίες στον τομέα της υγειονομικής περίθαλψης είναι πλέον σε ηλεκτρονική μορφή. ταιριάζει κάτω από τη μεγάλη ομπρέλα δεδομένων καθώς τα περισσότερα είναι αδόμητα και δύσκολα στη χρήση.



2019-1-BG01-KA203-062371

Τα μεγάλα δεδομένα στην έρευνα για την υγεία είναι ιδιαίτερα ελπιδοφόρα όσον αφορά τη διερευνητική βιοϊατρική έρευνα, καθώς η ανάλυση βάσει δεδομένων μπορεί να προχωρήσει πιο γρήγορα από την έρευνα που βασίζεται σε υποθέσεις. Στη συνέχεια, οι τάσεις που παρατηρούνται στην ανάλυση δεδομένων μπορούν να δοκιμαστούν στην παραδοσιακή, βασισμένη σε υποθέσεις, βιολογική έρευνα και τελικά κλινική έρευνα.

Αποθήκευση

Ένα αποθετήριο δεδομένων ή μια αποθήκη δεδομένων είναι ένα κεντρικό μέρος για την αποθήκευση και τη διατήρηση δεδομένων. Ένα αποθετήριο δεδομένων μπορεί να αποτελείται από ένα ή περισσότερα δομημένα αρχεία δεδομένων, όπως βάσεις δεδομένων ή μη δομημένα αρχεία δεδομένων, τα οποία μπορούν να διανεμηθούν σε ένα δίκτυο και να διατηρηθούν μακροπρόθεσμα.

Βασική δομή μιας βάσης δεδομένων

Αυτό το τμήμα είναι αφιερωμένο στην επισκόπηση των κύριων δομικών στοιχείων που αποτελούν μια βάση δεδομένων.

ΕΙΣΑΓΩΓΗ

Από την εφεύρεση των υπολογιστών, η ποσότητα των δεδομένων που αποθηκεύονται και διαχειρίζονται ηλεκτρονικά έχει αυξηθεί δραστικά. Εκτιμάται ότι η ποσότητα των δεδομένων θα φτάσει τα 175 zettabytes (10²¹ Bytes) έως το 2025, αυξάνοντας από μερικά petabytes (10¹⁵ Bytes) το έτος 2000. Ένας κοινός τρόπος απλοποίησης της ζωής των χρηστών και αξιοποίησης των πόρων τους στο έπακρο είναι αποθήκευση και ανάκτηση του πιο αποτελεσματικά. Για παράδειγμα, ενώ ένα επίπεδο αρχείο λειτουργεί πολύ καλά για την αποθήκευση των προσωπικών σας δεδομένων, όπως ένα βιβλίο διευθύνσεων ή κάποιες συνταγές, δεν είναι τόσο κατάλληλο για την αποθήκευση ενός τηλεφωνικού καταλόγου πόλης ή, πιο συγκεκριμένα, των γονιδιωματικών δεδομένων στο πεδίο της βιοτεχνολογίας. Επιπλέον, εάν θέλετε να αποθηκεύσετε πολλά γονιδιωματικά είδη αξίας δεδομένων, είναι πολύ δύσκολο να αναζητήσετε και να ανακτήσετε δεδομένα από ένα επίπεδο αρχείο. Οι βάσεις δεδομένων προσφέρουν μια λύση σε αυτό το πρόβλημα κάνοντας την αποθήκευση, το χειρισμό και την ανάκτηση δεδομένων πολύ πιο εύκολη.



2019-1-BG01-KA203-062371

Το λογισμικό που χρησιμοποιείται για τη διαχείριση μιας βάσης δεδομένων ονομάζεται σύστημα διαχείρισης βάσεων δεδομένων (DBMS). Αυτό το εξειδικευμένο λογισμικό λειτουργεί ενδιάμεσα για να βοηθήσει τους τελικούς χρήστες να έχουν πρόσβαση στη βάση δεδομένων. Συνήθως, οι χρήστες δεν αλληλοεπιδρούν άμεσα με μια βάση δεδομένων, επειδή αυτό μπορεί να οδηγήσει σε αποδιοργάνωση της. Αντ' αυτού, χρησιμοποιούν ένα DBMS που διαβάζει δεδομένα από ή γράφει δεδομένα στη βάση δεδομένων.

Η αυξανόμενη πολυπλοκότητα μεγάλων ποσοτήτων δεδομένων απαιτούσε από ορισμένες εταιρείες να χρησιμοποιούν εργαλεία διαχείρισης δεδομένων με βάση το σχεσιακό μοντέλο, όπως το κλασικό RDMBS. Το RDBMS σημαίνει Σύστημα Διαχείρισης Σχεσιακών Βάσεων Δεδομένων. Παρ' όλα αυτά, μεγάλες εταιρείες Διαδικτύου, όπως η Google, η Yahoo και η Amazon ή όλα τα δημοφιλή Social Media, αντιμετώπισαν η κάθε μία πρόκληση στην αντιμετώπιση τεράστιων ποσοτήτων δεδομένων σε πραγματικό χρόνο, κάτι που οι συμβατικές λύσεις RDBMS δεν μπορούσαν να αντιμετωπίσουν. Αυτό εξηγεί την αυξανόμενη δημοτικότητα των συστημάτων βάσεων δεδομένων NoSQL που ξεπήδησαν παράλληλα.

Τα συστήματα NoSQL είναι κατανεμημένες, μη σχεσιακές βάσεις δεδομένων σχεδιασμένες για αποθήκευση δεδομένων μεγάλης κλίμακας και για μαζικά παράλληλη επεξεργασία δεδομένων υψηλής απόδοσης σε μεγάλο αριθμό διακομιστών βασικών προϊόντων. Προέκυψαν από την ανάγκη για ευκινησία, απόδοση και κλίμακα και μπορούν να υποστηρίξουν ένα ευρύ φάσμα περιπτώσεων χρήσης, συμπεριλαμβανομένων διερευνητικών και προγνωστικών αναλυτικών στοιχείων σε πραγματικό χρόνο. Χτισμένες από κορυφαίες εταιρείες Διαδικτύου για να συμβαδίζουν με τον κατακλυσμό δεδομένων, οι βάσεις δεδομένων NoSQL κλιμακώνονται οριζόντια και έχουν σχεδιαστεί για να κλιμακώνονται σε εκατοντάδες εκατομμύρια, ακόμη και δισεκατομμύρια χρήστες που εκτελούν ενημερώσεις καθώς και ανάγνωση.

Μερικές από τις κοινές εφαρμογές των βάσεων δεδομένων NoSQL είναι τα κοινωνικά μέσα, οι μεγάλης κλίμακας πάροχοι ηλεκτρονικού ταχυδρομείου και τα κυβερνητικά συστήματα υγειονομικής περίθαλψης.

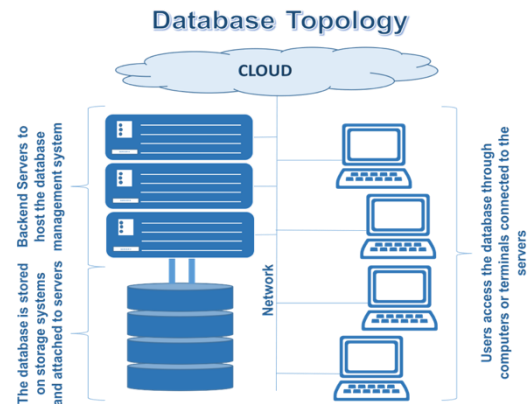
Συνήθως, μια κοινωνική εφαρμογή μπορεί να κλιμακωθεί από μηδέν σε εκατομμύρια χρήστες σε λίγες εβδομάδες και για να διαχειριστεί καλύτερα αυτήν την ανάπτυξη, χρειάζεται κάποιος DB που μπορεί να διαχειριστεί τεράστιο αριθμό χρηστών και δεδομένων, αλλά μπορεί επίσης εύκολα να κλιμακωθεί οριζόντια.

Σε αυτό το μάθημα, θα εστιάσουμε μόνο στο DBMS και το RDBMS. Αυτά είναι τα δύο είδη βάσεων δεδομένων που χρησιμοποιούνται συνήθως στον κόσμο της βιοτεχνολογίας μέχρι σήμερα.



ΕΠΙΣΚΟΠΗΣΗ ΤΗΣ ΑΡΧΙΤΕΚΤΟΝΙΚΗΣ ΜΙΑΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Οι βάσεις δεδομένων μπορούν να αποθηκεύσουν κάθε είδους πληροφορίες, από αριθμούς και κείμενο, έως email, περιεχόμενο στο διαδίκτυο, αρχεία τηλεφώνου, βιολογικά, γεωγραφικά δεδομένα κ.λπ. Οι βάσεις δεδομένων ταξινομούνται επίσημα ανάλογα με τον τρόπο που αποθηκεύουν αυτά τα δεδομένα. Σχετικές βάσεις δεδομένων αποθηκεύουν δεδομένα σε πίνακες. Οι αντικειμενοστραφείς βάσεις δεδομένων αποθηκεύουν δεδομένα σε κατηγορίες αντικειμένων και υποκατηγορίες. Θα επικεντρωθούμε στις σχεσιακές βάσεις δεδομένων, καθώς χρησιμοποιούνται συχνότερα. Ωστόσο, οι περισσότερες από τις βασικές τοπολογίες βάσεων δεδομένων πρέπει να έχουν διακομιστές backend για να φιλοξενήσουν το σύστημα διαχείρισης βάσεων δεδομένων, ένα σύστημα αποθήκευσης προσαρτημένο στους διακομιστές για την αποθήκευση της δομής και των δεδομένων της βάσης δεδομένων και, φυσικά, υπολογιστές, φορητούς υπολογιστές, επιτραπέζιους υπολογιστές ή τερματικά ως διεπαφή που επιτρέπει στους χρήστες να έχουν πρόσβαση στη βάση δεδομένων, στο σύστημα διαχείρισης και στο περιεχόμενό της. Απαιτείται επίσης ένα δίκτυο για ανταλλαγή μεταξύ όλων των στοιχείων υλικού και ένα συνημμένο Cloud που επιτρέπει στους απομακρυσμένους χρήστες να έχουν πρόσβαση στη βάση δεδομένων. Η Εικόνα 10 συνοψίζει με απλό τρόπο το ελάχιστο που απαιτείται για τη λειτουργία μιας βάσης δεδομένων.



Εικόνα 10. Αρχιτεκτονική υλικού για μια βάση δεδομένων

Ένας άλλος βασικός τρόπος για να το περιγράψουμε, είναι να δείξουμε την αρχιτεκτονική τριών επιπέδων μιας βάσης δεδομένων. Είναι μια εικονική προβολή των απαραίτητων επιπέδων για να λειτουργήσει σωστά μια βάση δεδομένων. Η Εικόνα 11 παρουσιάζει την αρχιτεκτονική προβολής τριών επιπέδων. Ονομάζεται μοντέλο ANSI-SPARC. Παρ' όλα αυτά, παρά το γεγονός ότι αυτό το μοντέλο δεν έγινε ποτέ τυπικό πρότυπο, παρουσιάζει την ιδέα της λογικής ανεξαρτησίας δεδομένων που έχει υιοθετηθεί ευρέως.



2019-1-BG01-KA203-062371

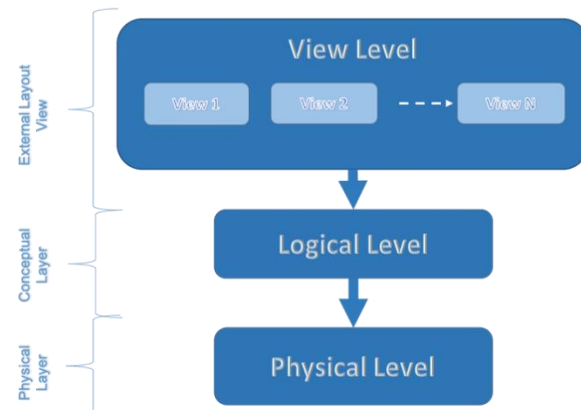
Οι πληροφορίες που αποθηκεύονται μέσα σε μια σχεσιακή βάση δεδομένων περιέχονται σε πίνακες. Αυτοί οι πίνακες αποτελούνται από σειρές δεδομένων και κάθε γραμμή περιέχει πεδία ή στήλες. Σε έναν καλά σχεδιασμένο ορισμό βάσης δεδομένων, που ονομάζεται σχήμα, μόνο παρόμοια δεδομένα αποθηκεύονται σε κάθε πίνακα και οι διπλές στήλες διατηρούνται στο ελάχιστο. Οι προγραμματιστές μπορούν να συνδέσουν ή να ενώσουν δεδομένα από δύο πίνακες για να συνδέσουν διαφορετικούς τύπους πληροφοριών μεταξύ τους.

Τα ευρετήρια μπορούν να δημιουργηθούν σε πεδία στον πίνακα βάσεων δεδομένων για να διευκολυνθεί η ανάκτηση δεδομένων από το DBMS. Τα ευρετήρια διαμορφώνονται συνήθως για στήλες που αναζητούνται συχνά, όπως το όνομα ενός ατόμου ή μια τιμή ημερομηνίας. Το μειονέκτημα της χρήσης ευρετηρίων είναι ότι καταλαμβάνουν χώρο στο δίσκο αποθήκευσης και μπορούν να επιβραδύνουν τα πράγματα, αν διατηρηθούν πάρα πολλά από αυτά, επειδή κάθε φορά που ενημερώνεται μια σειρά στη βάση δεδομένων, πρέπει επίσης να ενημερώνεται το ευρετήριο.

Οι περισσότερες βάσεις δεδομένων υποστηρίζουν Structured Query Language (SQL), μια τυπική γλώσσα για αλληλεπίδραση με πληροφορίες που περιέχονται σε μια βάση δεδομένων. Το SQL επιτρέπει στους χρήστες και τις εφαρμογές να αλληλοεπιδρούν με συγκεκριμένα υποσύνολα δεδομένων από έναν ή περισσότερους πίνακες χρησιμοποιώντας διάφορες προτάσεις ως SELECT, INSERT, UPDATE και DELETE.

Οι σχεσιακές βάσεις δεδομένων παρέχουν επίσης μια πολυεπίπεδη προσέγγιση στην αποθήκευση, επιτρέποντας τον ορισμό του τι αντικείμενα βάσης δεδομένων βρίσκονται σε συγκεκριμένα αρχεία δεδομένων και πού αυτά τα αρχεία δεδομένων τοποθετούνται στη δομή αρχείων του λειτουργικού συστήματος. Εκτός από τη διαχείριση της φυσικής θέσης αποθήκευσης αντικειμένων βάσης δεδομένων, πολλά συστήματα βάσεων δεδομένων δίνουν κάποιο έλεγχο στον τρόπο αποθήκευσης των δεδομένων στα αρχεία δεδομένων.

Three-Level Architecture of a Database



Εικόνα 11. Αρχιτεκτονική 3 επιπέδων



ΚΟΙΝΕΣ ΟΡΟΛΟΓΙΕΣ ΤΩΝ ΒΑΣΕΩΝ ΔΕΔΟΜΕΝΩΝ

Ορισμένοι όροι βάσης δεδομένων προέρχονται από τρόπους με τους οποίους οι βάσεις δεδομένων αυτοματοποιούν ενέργειες εγγραφής. Οι προγραμματιστές βάσεων δεδομένων συχνά αυτοματοποιούν τη γραφή σε ορισμένα πεδία ή άλλους πίνακες, όπως η εγγραφή ενός αντιγράφου της σειράς που εισάγεται - μαζί με μια χρονική σήμανση ή ένα όνομα χρήστη - σε έναν πίνακα ιστορικού ή ελέγχου. Τα περισσότερα συστήματα DBMS παρέχουν διάφορους τρόπους για την αυτόματη διαχείριση ενεργειών εγγραφής βάσης δεδομένων.

Οι ενεργοποιητές βάσης δεδομένων είναι η πιο συνηθισμένη μέθοδος ανάληψης ενεργειών στα δεδομένα καθώς γράφονται στη βάση δεδομένων. Οι ενεργοποιητές συνήθως σχετίζονται με έναν συγκεκριμένο πίνακα και διαμορφώνονται ώστε να εκτελούνται σε ένα συγκεκριμένο σημείο κατά τη διάρκεια μιας συγκεκριμένης ενέργειας εγγραφής, όπως πριν ή μετά από μια ενημέρωση ή μετά την εισαγωγή μιας σειράς. Οι ενεργοποιητές μπορούν να χρησιμοποιηθούν για τη μορφοποίηση δεδομένων, τη συμπλήρωση μιας στήλης με δεδομένα που προέρχονται από υπάρχουσες πληροφορίες ή ακόμα και την εγγραφή σε έναν άλλο πίνακα με βάση τη γραμμή που εισάγεται ή ενημερώνεται.

Μια αποθηκευμένη διαδικασία είναι ένας άλλος τρόπος αλληλεπίδρασης με μια σχεσιακή βάση δεδομένων. Οι αποθηκευμένες διαδικασίες είναι πιο περίπλοκες από τις ενεργοποιήσεις και δεν συνδέονται με έναν συγκεκριμένο πίνακα. Συνήθως δημιουργούνται από προγραμματιστή, χρησιμοποιούν συνδυασμό SQL και γλώσσας προγραμματισμού, όπως Java ή SQL (ανάλογα με την πλατφόρμα της βάσης δεδομένων). Οι αποθηκευμένες διαδικασίες παρέχουν στους προγραμματιστές πολύ έλεγχο στον τρόπο επικύρωσης ή μασάζ των δεδομένων από μια εφαρμογή. Μια αποθηκευμένη διαδικασία θα μπορούσε να χρησιμοποιηθεί για τη διαχείριση του τρόπου σύνδεσης ενός χρήστη σε μια εφαρμογή. Η διαδικασία μπορεί πρώτα να επικυρώσει το όνομα χρήστη και τον κωδικό πρόσβασης και, στη συνέχεια, να καταγράψει την επιτυχία ή την αποτυχία της προσπάθειας σε έναν άλλο πίνακα, μαζί με άλλες πληροφορίες, συμπεριλαμβανομένου του ονόματος του υπολογιστή και μιας χρονικής σήμανσης. Θα μπορούσε ακόμη και να σταλεί μια ειδοποίηση στον χρήστη που θα τον ενημερώνει ότι ο κωδικός πρόσβασης έχει λήξει και πρέπει να αλλάξει.

Οι συναρτήσεις είναι απλούστερες από την αποθηκευμένη διαδικασία και μερικές φορές μπορούν να χρησιμοποιηθούν ακόμη και μέσα από ερωτήματα SQL. Οι συναρτήσεις χρησιμοποιούνται συνήθως σε μια βάση δεδομένων για την εκτέλεση ενός συνόλου ενεργειών που επιστρέφουν μία ή περισσότερες τιμές, όπως ο υπολογισμός του αθροίσματος μιας στήλης για γραμμές που ταιριάζουν με μια συγκεκριμένη συνθήκη. Ενώ αυτές οι ενέργειες μπορούν να εκτελεστούν χρησιμοποιώντας SQL, η ενσωμάτωσή τους σε μια συνάρτηση μπορεί να τις κάνει πιο εύχρηστες σε άλλους κώδικες. Τόσο οι λειτουργίες όσο και οι αποθηκευμένες διαδικασίες μπορούν να εκτελέσουν κοινές ενέργειες με βελτιωμένο και συνεπή τρόπο, διευκολύνοντας τον φόρτο εργασίας για τους διαχειριστές και τους προγραμματιστές βάσεων δεδομένων.



2019-1-BG01-KA203-062371

ΠΟΙΑ ΕΙΝΑΙ Η ΔΙΑΦΟΡΑ ΜΕΤΑΞΥ ΤΩΝ ΚΥΡΙΩΝ ΣΥΣΤΗΜΑΤΩΝ DBMS?

Το DBMS γενικά καθορίζεται από αυτό που χρειάζονται οι εφαρμογές χρήστη για να υποστηρίξουν. Τούτου λεχθέντος, εδώ είναι μια σύντομη σύγκριση των τριών πιο ευρέως χρησιμοποιούμενων πλατφορμών.

Ο Microsoft SQL Server χρησιμοποιείται ευρέως σε εταιρικές εφαρμογές και ενσωματώνεται εύκολα με άλλα εργαλεία της Microsoft. Ο Microsoft SQL Server 2019 Express είναι η τελευταία έκδοση της δωρεάν προσφοράς της Microsoft και συχνά συνοδεύεται από εφαρμογές που χρησιμοποιούν SQL Server.

Το MySQL ήταν το αγαπημένο για προγραμματιστές ανοιχτού κώδικα για το μεγαλύτερο μέρος δύο δεκαετιών. Συχνά χρησιμοποιείται ως back-end για blog ανοιχτού κώδικα ή συστήματα διαχείρισης περιεχομένου, η MySQL έχει μια τεράστια εγκατεστημένη βάση σε όλο τον κόσμο. Το 2008, η MySQL AB εξαγοράστηκε από την Sun Microsystems, η οποία εξαγοράστηκε η ίδια από την Oracle Corp. το 2009, φέρνοντας την MySQL υπό την ομπρέλα ενός από τους μεγαλύτερους ανταγωνιστές της. Ωστόσο, η MySQL Community Edition παραμένει δωρεάν και υποστηρίζεται καλά από την κοινότητα. Το MySQL είναι διαθέσιμο για πολλά λειτουργικά συστήματα, όπως Linux, UNIX, Mac OS X και Windows.

Η βάση δεδομένων Oracle θεωρείται από πολλούς ως το πρότυπο σε πλατφόρμες βάσεων δεδομένων σε επίπεδο επιχείρησης και υποστηρίζει πολυάριθμες εταιρικές εφαρμογές. Το Oracle Database Express Edition διατίθεται δωρεάν και διατίθεται επίσης δωρεάν (αν και δεν είναι τεχνικά δωρεάν λογισμικό), καθιστώντας το μια άλλη δημοφιλή επιλογή για προγραμματιστές ή χομπίστες σε Windows ή Linux.

Τώρα που μάθατε τους βασικούς όρους και έννοιες της βάσης δεδομένων, είστε πολύ πιο κοντά στο να μιλάτε την ίδια γλώσσα με τους προγραμματιστές βάσεων δεδομένων του οργανισμού σας.



Βάσεις δεδομένων στον επιστημονικό κόσμο

Αυτό το μέρος ασχολείται με τα βασικά των βάσεων δεδομένων που χρησιμοποιούνται στον επιστημονικό κόσμο

ΕΙΣΑΓΩΓΗ ΣΕ ΥΠΑΡΧΟΥΣΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΑΦΙΕΡΩΜΕΝΕΣ ΣΤΗΝ ΕΠΙΣΤΗΜΗ

Αυτή η ενότητα είναι αφιερωμένη στην επισκόπηση των πιο κοινών βάσεων δεδομένων ανοικτής πρόσβασης που χρησιμοποιούνται στην επιστήμη.

Οι συνεχείς εξελίξεις στους τομείς της βιοτεχνολογίας και της τεχνολογίας των πληροφοριών έχουν οδηγήσει στην εκθετική αύξηση των δεδομένων. Μελέτες που διεξήχθησαν από ερευνητές στο Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EMBL-EBI) έδειξαν ότι αυτή η αύξηση των πληροφοριών διπλασιάζεται περίπου κάθε χρόνο. Αυτές οι εκτεταμένες ποσότητες δεδομένων αποθηκεύονται, οργανώνονται και ενημερώνονται συνεχώς σε επιστημονικές βάσεις δεδομένων, όπου είναι άμεσα διαθέσιμες για επιστήμονες, συμπεριλαμβανομένων βιολόγων και βιοπληροφορικών, για ερευνητικούς σκοπούς. Οι πληροφορίες που είναι διαθέσιμες σε βιολογικές βάσεις δεδομένων λαμβάνονται από μια σειρά επιστημονικών πεδίων, συμπεριλαμβανομένων των μεταβολικών, της έκφρασης του γονιδίου της μικροσυστοιχίας και της πρωτεομικής. Εκτός από την αποθήκευση, την οργάνωση και την κοινή χρήση τεράστιου όγκου δεδομένων, ο κύριος στόχος των βιολογικών βάσεων δεδομένων είναι να προσφέρουν διεπαφές προγραμματισμού εφαρμογών Ιστού (API) για υπολογιστές για ανταλλαγή και ενσωμάτωση δεδομένων από πολλούς διαφορετικούς πόρους βάσεων δεδομένων μέσω αυτοματοποιημένης μεθόδου.

Οι βιολογικές βάσεις δεδομένων μπορούν να οριστούν ως συλλογές δεδομένων, οι οποίες είναι δομημένες με τέτοιο τρόπο ώστε να είναι εύκολο να εξερευνησετε, να χειριστείτε και να ενημερώσετε το περιεχόμενό τους. Παραδείγματα τέτοιων βάσεων δεδομένων παρουσιάζονται στην Εικόνα 12. Το 1972, δημιουργήθηκε η πρώτη βάση δεδομένων πρωτεϊνικής δομής, γνωστή ως Τράπεζα Δεδομένων Πρωτεϊνών (PDB). Αυτή η βάση δεδομένων περιείχε αρχικά μόνο 10 καταχωρήσεις, η οποία τώρα έχει επεκταθεί και περιέχει περισσότερες από 10.000 καταχωρήσεις, υποδηλώνοντας την ταχεία ανάπτυξη βιολογικών δεδομένων. Μια βιολογική βάση δεδομένων μπορεί να περιέχει διάφορους τύπους δεδομένων, συμπεριλαμβανομένων αλληλουχιών πρωτεϊνών, περιγραφών κειμένου, χαρακτηριστικών και δεδομένων πίνακα. Γενικά, μπορούν να χωριστούν σε πρωτογενείς, δευτερεύουσες και σύνθετες



Εικόνα 12. Παραδείγματα ονομάτων βάσεων δεδομένων στην βιοτεχνολογία



2019-1-BG01-KA203-062371

βάσεις δεδομένων. Οι πρωτογενείς βάσεις δεδομένων περιλαμβάνουν δεδομένα μόνο για την ακολουθία ή τη δομή, ενώ οι δευτερεύουσες βάσεις δεδομένων περιλαμβάνουν δεδομένα που προέρχονται από την κύρια βάση δεδομένων. Δεδομένα, όπως η διατηρημένη αλληλουχία και τα υπολείμματα ενεργών θέσεων των οικογενειών πρωτεϊνών, μπορούν να βρεθούν σε βάσεις δεδομένων δευτερεύουσας δομής. Επιπλέον, καταχωρήσεις του ΠΣΠ, που είναι μια κύρια βάση δεδομένων, μπορούν να βρεθούν σε βάσεις δεδομένων δευτερεύουσας δομής, αποθηκευμένες με οργανωμένο τρόπο.

Σε γενικές γραμμές, οι βιολογικές βάσεις δεδομένων μπορούν να κατηγοριοποιηθούν σε βάσεις δεδομένων ακολουθίας, δομής και διαδρομής:

- Βάσεις δεδομένων ακολουθίας: Οι πιο συχνά χρησιμοποιούμενες βιολογικές βάσεις δεδομένων. Αυτές περιλαμβάνουν βάσεις δεδομένων πρωτεϊνών και νουκλεοτιδικών αλληλουχιών, οι οποίες περιέχουν αποτελέσματα υγρού εργαστηρίου και αποτελούν την κύρια πηγή πειραματικών αποτελεσμάτων. Τα GenBank και EMBL είναι παραδείγματα βάσεων δεδομένων ακολουθίας.
- Βάσεις δεδομένων δομής: Αυτές οι βάσεις δεδομένων περιέχουν πληροφορίες σχετικά με τη δομή της πρωτεΐνης και τις μοριακές αλληλεπιδράσεις. Το PDB είναι ένα παράδειγμα δομής βάσης δεδομένων.
- Βάσεις δεδομένων διαδρομής: Αυτές οι βάσεις δεδομένων βασίζονται σε δεδομένα που προέρχονται από τη συγκριτική μελέτη των μεταβολικών οδών. Η Kyoto Encyclopedia of Genes and Genomes (KEGG) και η Biocyc είναι δύο ενδεικτικές βάσεις δεδομένων.

Μια τυπική αναζήτηση σε μια βάση δεδομένων αλληλουχιών νουκλεοτιδίων μπορεί, για παράδειγμα, να δημιουργήσει δεδομένα σχετικά με την επιστημονική ονομασία του οργανισμού προέλευσης από τον οποίο απομονώθηκε, όνομα επαφής, αλληλουχία εισόδου με λεπτομέρειες τύπου μορίου και, συχνά, βιβλιογραφικές αναφορές που σχετίζονται με αλληλουχία.

Ορισμένα εργαλεία έχουν αναπτυχθεί για να διευκολύνουν τους επιστήμονες στην επεξεργασία και ανάκτηση δεδομένων από βιολογικές βάσεις δεδομένων. Αυτά τα εργαλεία, τα οποία ονομάζονται εργαλεία βιοπληροφορικής, είναι προγράμματα λογισμικού που δημιουργήθηκαν για την εξαγωγή σημαντικών δεδομένων από τον τεράστιο αριθμό βιολογικών βάσεων δεδομένων και για τη διεξαγωγή ακολουθίας ή δομικής ανάλυσης. Τα εργαλεία βιοπληροφορικής χρησιμοποιούνται για τη λήψη δεδομένων από βάσεις δεδομένων γονιδιωματικής αλληλουχίας και για την οπτικοποίηση, ανάλυση και ανάκτηση ημερομηνίας από πρωτεϊμικές βάσεις δεδομένων. Αυτά τα εργαλεία χωρίζονται σε μεγάλο βαθμό σε:

- Εργαλεία ομολογίας και ομοιότητας: Αυτά τα εργαλεία χρησιμοποιούνται για τον εντοπισμό ομοιότητας μεταξύ των αλληλουχιών άγνωστων δομικών και λειτουργικών αλληλουχιών, των οποίων η λειτουργία και η δομή είναι ήδη γνωστές.



2019-1-BG01-KA203-062371

- Εργαλεία ανάλυσης λειτουργίας πρωτεΐνης: Προγράμματα που εφαρμόζονται για τη σύγκριση μιας αλληλουχίας πρωτεΐνης με μια δευτερεύουσα (ή παράγωγη) πρωτεΐνη, τα οποία επιτρέπουν την εκτίμηση της βιοχημικής λειτουργίας μιας πρωτεΐνης ερωτήματος.
- Εργαλεία δομικής ανάλυσης: Αυτά τα εργαλεία επιτρέπουν τη σύγκριση δομών με τις γνωστές βάσεις δεδομένων δομής και τη δημιουργία της δομής 2D/3D μιας πρωτεΐνης.
- Εργαλεία ανάλυσης αλληλουχίας: Προγράμματα που χρησιμοποιούνται για την επιπρόσθετη, πιο ολοκληρωμένη αξιολόγηση μιας ακολουθίας ερωτήματος, που περιλαμβάνει εξελικτική ανάλυση και προσδιορισμό μεταλλάξεων.

Οι βιολογικές βάσεις δεδομένων μπορούν επίσης να κατηγοριοποιηθούν, με βάση το εύρος της κάλυψης δεδομένων, σε:

- Ολοκληρωμένες βάσεις δεδομένων: Αυτές οι βάσεις δεδομένων περιλαμβάνουν διάφορους τύπους δεδομένων από διάφορα είδη. Παραδείγματα ολοκληρωμένων βάσεων δεδομένων είναι η GenBank και η EMBL.
- Εξειδικευμένες βάσεις δεδομένων: Αυτές οι βάσεις δεδομένων περιλαμβάνουν συγκεκριμένους τύπους δεδομένων ή δεδομένα από συγκεκριμένους οργανισμούς. Ένα παράδειγμα εξειδικευμένων βάσεων δεδομένων είναι το WormBase, το οποίο περιέχει πληροφορίες σχετικά με τη βιολογία νηματωδών και τη γονιδιοματική.

Σε σχέση με το επίπεδο βιοαπόδοσης, το οποίο ορίζεται ως η δραστηριότητα οργάνωσης, επίδειξης και διάθεσης βιολογικών πληροφοριών άμεσα σε ανθρώπους και υπολογιστές, οι βιολογικές βάσεις δεδομένων ταξινομούνται ως πρωτογενείς και δευτερεύουσες ή παράγωγες βάσεις δεδομένων. Οι πρωτογενείς βάσεις δεδομένων αποτελούνται από ακατέργαστα δεδομένα ως αποθετήριο αρχείων, ενώ δευτερεύουσες ή παράγωγες βάσεις δεδομένων αποτελούν επιμελημένες πληροφορίες ως προστιθέμενη αξία. Όσον αφορά τη μέθοδο που χρησιμοποιείται για την επιμέλεια των δεδομένων, οι βιολογικές βάσεις δεδομένων μπορούν να ταξινομηθούν περαιτέρω ως βάσεις δεδομένων με επιμέλεια εμπειρογνομόνων ή βάσεις δεδομένων με επιμέλεια της κοινότητας, οι οποίες επιμελούνται με συνεργατικό τρόπο από πολυάριθμους ερευνητές.

Επιπλέον κατηγοριοποίηση βιολογικών βάσεων δεδομένων μπορεί επίσης να γίνει με βάση τον τύπο δεδομένων. Οι τύποι δεδομένων που ταξινομούν ανάλογα τις βάσεις δεδομένων περιλαμβάνουν DNA, RNA, πρωτεΐνη, έκφραση, οδό, ασθένεια, ονοματολογία, βιβλιογραφία και πρότυπο και οντολογία. Μερικές από τις πιο σημαντικές και ευρέως χρησιμοποιούμενες βιολογικές βάσεις δεδομένων είναι οι ακόλουθες: GenBank, το UCSC Genome Browser και Ensembl, οι οποίες είναι βάσεις δεδομένων/πύλες ακολουθίας. WormBase και The Arabidopsis Information Resource (TAIR), οι οποίες είναι πρότυπα βάσεων δεδομένων οργανισμών. και το PDB, Online Mendelian Inheritance in Man (OMIM), MetaCyc και KEGG, τα οποία χαρακτηρίζονται ως βάσεις δεδομένων που δεν βασίζονται στην ακολουθία.



2019-1-BG01-KA203-062371

Η χειραγώγηση δεδομένων αποτελεί ουσιαστικό μέρος της πειραματικής διαδικασίας όλων των μελετών, ανεξάρτητα από την κλίμακα τους. Η διαδικτυακή διαθεσιμότητα βιολογικών δεδομένων σε συνδυασμό με το μειωμένο κόστος των αυτοματοποιημένων αλληλουχιών γονιδιώματος επέτρεψαν στα μικρά εργαστήρια βιολογίας να γίνουν γεννήτριες μεγάλων δεδομένων. Ακόμα κι αν ένα εργαστήριο δεν είναι εξοπλισμένο με τέτοια όργανα, μπορεί να γίνει χρήστης μεγάλων δεδομένων αποκτώντας πρόσβαση σε δημόσια αποθετήρια που περιέχουν βιολογικά δεδομένα, όπως το Εθνικό Κέντρο Πληροφοριών Βιοτεχνολογίας των ΗΠΑ στην Bethesda. Ένα μεγάλο μέρος της κατασκευής στη βιολογία των μεγάλων δεδομένων είναι εικονικό, βασισμένο σε υπολογιστικό νέφος, στο οποίο τα δεδομένα και το λογισμικό βρίσκονται σε τεράστια, εκτός κέντρου κέντρα στα οποία είναι προσβάσιμα κατόπιν αιτήματος. Επομένως, δεν είναι απαραίτητο οι χρήστες να αγοράζουν το δικό τους υλικό. Το σύστημα υπολογιστικού νέφους επιτρέπει στους πιθανούς χρήστες να δημιουργούν εικονικούς χώρους για δεδομένα, λογισμικό και αποτελέσματα που είναι ελεύθερα προσβάσιμα από όλους ή να διατηρούν τους χώρους κλειδωμένους πίσω από ένα τείχος προστασίας που επιτρέπει την πρόσβαση σε μια επιλεγμένη ομάδα συνεργατών.

Η χρήση βιολογικών βάσεων δεδομένων μπορεί να είναι επωφελής σε διάφορους τομείς έρευνας. Για παράδειγμα, οι βάσεις δεδομένων μπορούν να βοηθήσουν τον πειραματικό σχεδιασμό επιτρέποντας την αυτόματη ανάλυση και εύκολη επεξεργασία πειραματικών δεδομένων και καθιστώντας την εξέταση των πειραματικών αποτελεσμάτων απλή και γρήγορη. Η ανακάλυψη ναρκωτικών είναι ένας άλλος τομέας που μπορεί να απλοποιηθεί με τη χρήση βάσεων δεδομένων. Σε αυτόν τον συγκεκριμένο τομέα, οι βάσεις δεδομένων μπορούν να σαρωθούν προκειμένου να βρεθούν νέοι υποψήφιοι για φάρμακα εκπαιδύοντας έναν ταξινομητή σε ένα σύνολο δεδομένων όπου έχουν εντοπιστεί λειτουργικά και μη λειτουργικά φάρμακα. Επιπλέον, τεχνικές μηχανικής μάθησης μπορούν να εφαρμοστούν για τον σχεδιασμό εικονικών αναλύσεων που είναι σε θέση να προσδιορίσουν πολλά υποσχόμενα νέα φάρμακα, τα οποία μπορούν στη συνέχεια να αναλυθούν σε εργαστηριακό περιβάλλον. Και το πιο σημαντικό, μπορούν να πραγματοποιηθούν νέα επιστημονικά πειράματα και να προκύψουν νέα αποτελέσματα αναλύοντας υπάρχοντα σύνολα δεδομένων.

Χωρίς την ύπαρξη βάσεων δεδομένων, η ανταλλαγή και η ενσωμάτωση μεγάλων ποσοτήτων δεδομένων θα ήταν ουσιαστικά αδύνατη. Αν και πολλοί επιστήμονες ζωής έχουν προηγμένες υπολογιστικές δεξιότητες, ένα μεγάλο ποσοστό δεν είναι εξοικειωμένοι με την ανάπτυξη ή την προσαρμογή του σχετικού λογισμικού. Παρ' όλα αυτά, η συμμετοχή των επιστημόνων της ζωής σε αυτή τη διαδικασία είναι ζωτικής σημασίας, καθώς μπορούν να παρέχουν ανατροφοδότηση στους ειδικούς της πληροφορικής με επίκεντρο τις διαφορετικές ανάγκες και προσεγγίσεις της επιστήμης. Η δυνατότητα πρόσβασης στα πραγματικά σύνολα δεδομένων που χρησιμοποιήθηκαν αρχικά σε μια συγκεκριμένη μελέτη παρέχει στους ερευνητές την ευκαιρία να αναπαράγουν και να επεκτείνουν τη μελέτη αυτή. Αυτός είναι ο λόγος για τον οποίο είναι σημαντικό τα δεδομένα να είναι ελεύθερα διαθέσιμα στους επιστήμονες ανά πάσα στιγμή χωρίς περιορισμούς, μια έννοια που υποστηρίζεται από την Open Science και πολλές σχετικές πρωτοβουλίες. Μία από αυτές τις πρωτοβουλίες είναι γνωστή ως ELIXIR, ένα έργο που σχεδιάστηκε για να βοηθήσει τους επιστήμονες σε όλη την Ευρώπη να διαφυλάξουν και να μοιραστούν τα δεδομένα τους και να ενισχύσουν τους τρέχοντες πόρους,



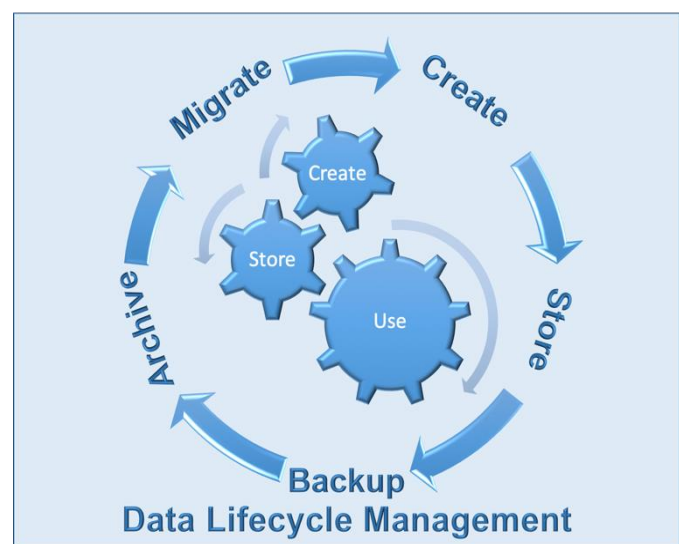
2019-1-BG01-KA203-062371

συμπεριλαμβανομένων των βάσεων δεδομένων και των υπολογιστικών εγκαταστάσεων, σε μεμονωμένες χώρες.

Παρόλο που η δημιουργία βιολογικών βάσεων δεδομένων έχει επιφέρει πολλά οφέλη, όπως η προώθηση της επιστημονικής ποιότητας παραγωγής που ενεργοποιείται με τη δικτύωση, εξακολουθούν να απαιτούν βελτίωση όσον αφορά τη βελτιστοποίηση της γνώσης. Είναι ζωτικής σημασίας η διαχείριση της διεπιστημονικής γνώσης με τέτοιο τρόπο που θα οδηγήσει σε αύξηση της ποιότητας και της ποσότητας της. Η ετερογένεια των δεδομένων είναι ένα άλλο κοινό ζήτημα που αντιμετωπίζει η ενσωμάτωση βιολογικών δεδομένων. Στον τομέα της βιολογίας, υπάρχουν διάφορες διαφορετικές μέθοδοι για την αναπαράσταση παρόμοιων δεδομένων. Αυτό περιπλέκει την ενσωμάτωση και την επεξεργασία δεδομένων, γεγονός που με τη σειρά του καθιστά δυσκολότερη την απόκτηση ενοποιημένων απόψεων αυτών των δεδομένων. Ένα παράδειγμα αυτού του προβλήματος είναι η χρήση διαφόρων εναλλακτικών ονομάτων όταν γίνεται αναφορά σε γονίδια, ανεξάρτητα από την ύπαρξη πλήρων κατευθυντήριων γραμμών που εκδόθηκαν το 1979 και προτείνουν την υιοθέτηση του προτύπου ονοματολογίας γονιδίων, οδηγώντας σε δυσκολίες στην ανταλλαγή δεδομένων. Η εφαρμογή προτύπων επιτρέπει την επαναχρησιμοποίηση των δεδομένων, ωστόσο, η απουσία τους προκαλεί σημαντική απώλεια της παραγωγικότητας και συμβάλλει στη μείωση των δεδομένων που είναι προσβάσιμα από τους ερευνητές. Ως εκ τούτου, είναι επιτακτική ανάγκη να βρεθεί μια λύση σε αυτό το θέμα, προκειμένου να εξλειφθούν οι προκλήσεις που αντιμετωπίζουν οι επιστήμονες όταν χρησιμοποιούν βιολογικές βάσεις δεδομένων για τη διεξαγωγή της έρευνάς τους.

ΤΕΛΙΚΕΣ ΣΚΕΨΕΙΣ

Η αντιμετώπιση δεδομένων συνεπάγεται μια δραστική πειθαρχία για να διατηρηθεί η μακροπρόθεσμη πρόσβαση στις αποθηκευμένες πληροφορίες. Η τεχνολογία εξελίσσεται, πράγμα που σημαίνει ότι το υλικό και το λογισμικό που χρησιμοποιούνται σήμερα δεν είναι το πρότυπο του αύριο. Αυτό σημαίνει ότι για να μπορέσουμε να διαβάσουμε όλα τα δεδομένα που γράφτηκαν σήμερα, θα πρέπει να εκτελέσουμε δύο διαφορετικά είδη μετακινήσεων. Μια λογική μετανάστευση και μια τεχνολογική μετανάστευση. Η λογική μετεγκατάσταση σχετίζεται με το είδος της μορφής στην οποία αποθηκεύονται τα δεδομένα. Η τεχνολογική μετάβαση σχετίζεται με το είδος του υλικού που χρησιμοποιείται. Για παράδειγμα, εάν





2019-1-BG01-KA203-062371

προσπαθήσετε να ανοίξετε ένα αρχείο Word που γράφτηκε το 1993 με την έκδοση του Word 6 με την τελευταία έκδοση του Word 2019, δεν θα λειτουργήσει. Αυτό το παράδειγμα δείχνει έλλειψη λογικής συμβατότητας. Για να αποφύγετε αυτό το ζήτημα και να διατηρήσετε μια ανερχόμενη συμβατότητα, το αρχείο θα έπρεπε να έχει μεταφερθεί μέχρι τότε στην πιο πρόσφατη έκδοση, προκειμένου να διατηρείται ενημερωμένο και ευανάγνωστο με τις πιο πρόσφατες εκδόσεις λογισμικού.

Το ίδιο ισχύει για το υλικό, δηλαδή διακομιστές, χώρο αποθήκευσης, δίκτυα κλπ ... Ένα άλλο παράδειγμα θα μπορούσε να είναι το είδος του διακομιστή και του λειτουργικού συστήματος που χρησιμοποιούνται για την εκτέλεση μιας βάσης δεδομένων. Σε περίπτωση που αποφασίσετε να αλλάξετε το υλικό σας και να μετακινηθείτε από, ας πούμε, τα Windows στο UNIX, θα χρειαστεί ένα διαφορετικό είδος υλικού για την εκτέλεση του UNIX και μια διαφορετική έκδοση της βάσης δεδομένων για εκτέλεση στο UNIX. Τα Windows εκτελούνται σε πλατφόρμες που βασίζονται στην Intel (και Intel όπως) και το Unix εκτελείται σε πλατφόρμες που βασίζονται σε SPARC, πράγμα που σημαίνει ότι θα πρέπει να μεταβείτε σε μια συμβατή έκδοση της βάσης δεδομένων με UNIX - SPARC.

Λαμβάνοντας υπόψη αυτή τη συνεχή εξέλιξη του υλικού, των λειτουργικών συστημάτων, του λογισμικού και των μορφών, η έγκαιρη εκτέλεση των κατάλληλων λογικών και τεχνολογικών μετακινήσεων θα μπορούσε να σας εξοικονομήσει πολύ χρόνο και προβλήματα.

Τέλος, είναι σημαντικό να συνεχίζετε να δημιουργείτε αντίγραφα ασφαλείας των δεδομένων σας. Μία φορά κάθε τρεις έως έξι μήνες, εκτελέστε μια δοκιμή επαναφοράς για να δείτε εάν είστε σε θέση να ανακτήσετε τα αντίγραφα ασφαλείας σας. Αυτό είναι κρίσιμο για δύο λόγους:

- Θα σας κρατά ενημέρους για τον τρόπο επαναφοράς των δεδομένων σας
- Είναι η καλύτερη μέθοδος δοκιμής για να διαπιστώσετε εάν δημιουργήθηκαν αντίγραφα ασφαλείας των δεδομένων σας



2019-1-BG01-KA203-062371

Βιβλιογραφικές αναφορές

Baxevanis AD, Bateman A. 2015. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics.*, 50(1):1.1.1-1.1.8.

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2014. GenBank. *Nucleic Acids Res.*, 42:D32–D37.

Brooksbank C, Bergman MT, Apweiler R, Birney E, Thornton J. 2014. The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res.*, 42:D18–D25.

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, 44(D1):D471-80.

Figueiredo MSN, Pereira AM. 2017. Managing knowledge – the importance of databases in the scientific production. *Procedia Manuf.*, 12:166–73.

Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ. 2014. WormBase 2014: new views of curated biology. *Nucleic Acids Res.*, 42:D789–D793.

Howe D, Costanzo M, Fey P, Gojobori T, Hannick L, Hide W, et al. 2008. Big data: The future of biocuration: Big data. *Nature.*, 455(7209):47–50.

Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. 2021. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, 49(D1): D545–51.

Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. 2019. The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform.*, 20(4):1085–93.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res.*, 12(6):996-1006.

Lapatas V, Stefanidakis M, Jimenez RC, Via A, Schneider MV. Data integration in biological research: an overview. *J Biol Res (Thessalon)*. 2015;22(1):9.

Marx V. 2013. Biology: The big challenges of big data: Biology. *Nature.*, 498(7453):255–60.

Nature Structural Biology 10, 980. 2003; doi: 10.1038/nsb1203-980



2019-1-BG01-KA203-062371

Oliveira AL. 2019. Biotechnology, big data and artificial intelligence. *Biotechnol J.*, 14(8):e1800613.

Razvi SRH, Rampogu S. 2016. Bioinformatics in the present day. *MOJ proteom bioinform [Internet].*, 3(1):11–2. Available from: <http://dx.doi.org/10.15406/mojpb.2016.03.00073>

Toomula N, Kumar A, Kumar D S, Bheemidi VS. 2012. Biological databases- integration of life science data. *J Comput Sci Syst Biol.*, 04(05):087-092. Available from: <http://dx.doi.org/10.4172/jcsb.1000081>

Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. 2020. Ensembl 2020. *Nucleic Acids Res.*, 48(D1): D682–8.

Zou D, Ma L, Yu J, Zhang Z. 2015. Biological databases for human research. *Genomics Proteomics Bioinformatics.*, 13(1):55–63.

Web sources:

https://en.wikipedia.org/wiki/Airline_reservations_system

<https://en.wikipedia.org/wiki/CODASYL>

https://en.wikipedia.org/wiki/Database_administrator

https://en.wikipedia.org/wiki/IBM_Information_Management_System

<http://www.redbooks.ibm.com/abstracts/sg245352.html>

https://en.wikipedia.org/wiki/Navigational_database

<https://en.wikipedia.org/wiki/SQL>

<https://omim.org/>

<https://www.ascii-code.com>

<https://www.budapestopenaccessinitiative.org/>

<https://www.merriam-webster.com/dictionary/data>



2019-1-BG01-KA203-062371

<https://www.snia.org/education/online-dictionary/term/big-data>

<https://www.snia.org/education/online-dictionary/term/structured-data>

www.arabidopsis.org/aboutarabidopsis.html



Project website: www.digit-biotech.eu

The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.