Généralités sur les bases de données

Année scolaire : 2024-2025

Section : IG21/LA2

Matière : BDD - Support de cours N°1

Contexte général

Auteur : Isabelle Le Glaz

Donnée vs Information

Donnée *vs* Information ?

Il existe une différence subtile entre les données et les informations. Les données sont les faits ou les détails à partir desquels l'information est dérivée. Les données individuelles sont rarement utiles seules. Pour que les données deviennent des informations, elles doivent être mises en contexte.

- **Donnée** : Les données sont des faits bruts et non organisés qui doivent être traités. Les données peuvent être quelque chose de simple et d'apparemment aléatoire et inutile tant qu'elles ne sont pas organisées. Les données sont toujours interprétées, par un homme ou une machine, pour en tirer un sens. Les données sont donc dénuées de sens. Les données contiennent des chiffres, des énoncés et des caractères sous forme brute.

- **Information** : Lorsque des données sont traitées, organisées, structurées ou présentées dans un contexte donné afin de les rendre utiles, on les appelle des informations. L'information donne un sens et améliore la fiabilité des données. Elle contribue à réduire l'incertitude. Ainsi, lorsque les données sont transformées en informations, elles ne contiennent jamais de détails inutiles.

In fine, plus les données sont combinées, plus on peut en tirer des enseignements.

Source : https://blog.ostraca.fr/quel-est-la-difference-entre-informations-et-donnees/

Donnée *vs* Information ?

La notion de donnée n’est pas nouvelle, et n’a pas été introduite avec la numérisation / informatisation des sociétés humaines.

Quelques exemples historiques de manipulation et traitement de la donnée ?

Sauvegarder les données sur le temps long

Sauver les données sur le temps long

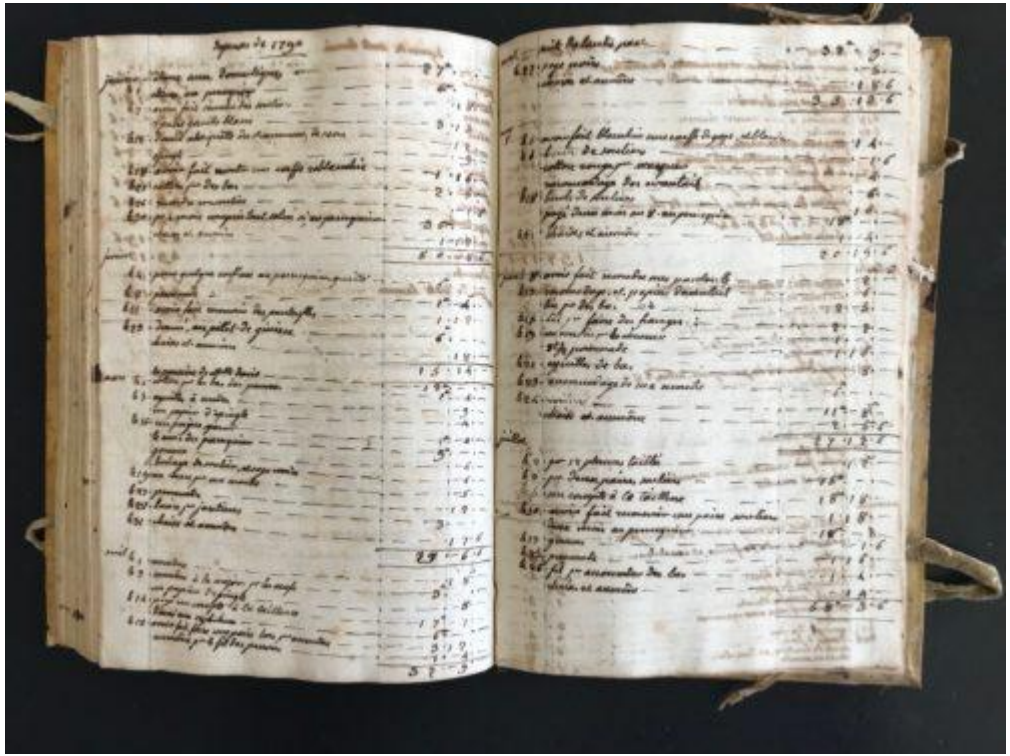
Un traitement informatique exploite en général des données existantes le temps “de vie” du processus informatique correspondant: on dit que ces données sont en “mémoire vive”, elles sont volatiles : dès que le programme se termine, ces données ne sont plus disponibles.

Pour garantir de retrouver certaines des informations (pas toutes, on verra pourquoi + loin), on doit sauvegarder les données sur le temps long. Cette sauvegarde permet de témoigner de ce que les données étaient au moment de leur sauvegarde, et de les exploiter pour poursuivre un travail ou une activité.

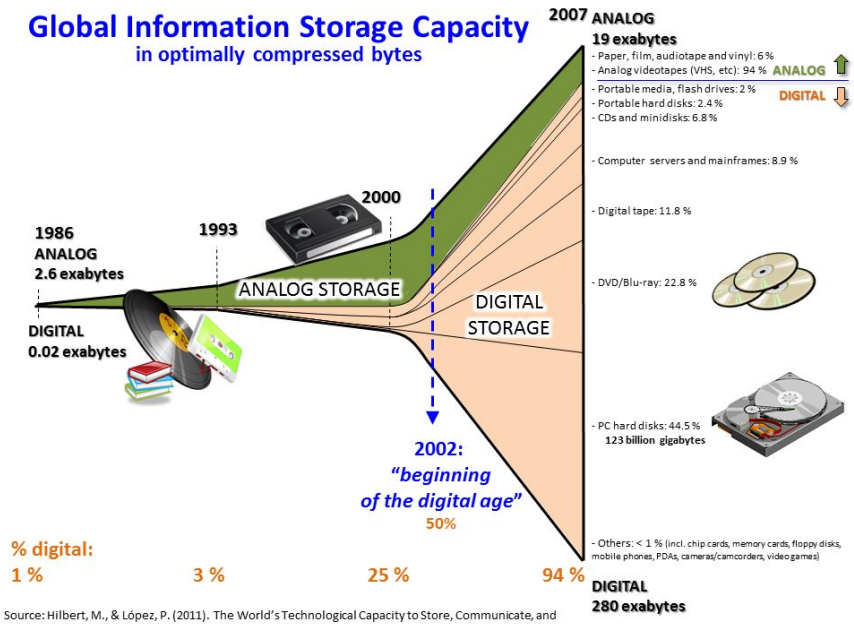
Comme pour la notion de données, la sauvegarde des données n’est pas une affaire récente. De tout temps, ceux qui manipulaient des données ont souhaité en garde la trace : cette trace est tout autant une preuve, qu’un socle de travail pour poursuivre une activité.

Comment sauvegardait-on les données autrefois, avant l’ère de l’informatique ?

Sauver les données sur le temps long



https://photos.app.goo.gl/wEEJiSiEqXxUdhyf7



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Le big bang du big data

Volume annuel de données numériques créées à l'échelle mondiale depuis 2010, en zettaoctets *

* Prévisions de 2020 à 2035. Un zettaoctet équivaut à mille milliards de gigaoctets.
Source : Statista Digital Economy Compass 2019

statista

Sauver les données sur le temps long

Wikipedia est notre ami :

Histoire de l’informatique

https://fr.wikipedia.org/wiki/Histoire_de_l%27informatique

Stockage des données :

https://fr.wikipedia.org/wiki/Stockage_d%27information

Sauver les données sur le temps long

La “dématérialisation “ (absence de papier) et le stockage “en ligne” (dans le cloud), ne nous affranchissentt pas pour autant de support matériel : le stockage de données reste attaché à des ressources matérielles (qu’il soit en local ou quelque part sur la planète).

Ce stockage a toujours une durée de vie limitée , et il faut penser au besoin de pérenniser cette sauvegarde en changeant régulièrement de support pour en garantir l’accès physique.

Mais l’accès physique n’est pas le seul écueil : savez-vous lire ceci ?

Exploiter les données sur le temps long

Exploiter les données sur le temps long

Même si les données restent accessibles physiquement, encore faut-il disposer des logiciels et des ordinateurs pour les exploiter.

Par exemple : comment exploiter des sauvegardes datant de 10, 20 ou 30 ans, quand les logiciels les ayant produits n’existent plus, ou ne sont plus exploitables sur les ordinateurs actuels ?

Exemples ?

Vers une vie longue de l'exploitation des données

Avant les années 1970, les données sont stockées sous forme de fichiers. Chaque application (ou du moins son équipe de développement) décide du format de ses fichiers; le stockage est de type binaire, afin de gagner de la place (l'octet de stockage coûtait trop cher pour exploiter des fichiers textes/ascii).

A partir des années 70, pour répondre aux besoins d'ouverture, d'inter-opérabilité et de longévité des données, sont apparues les techniques de moteur de bases de données :

- Les données ne sont plus gérées directement pas les applications, mais par des applications qui rendent ce service : les serveurs de bases de données. Ce serveur repose sur un moteur (ensemble de logiciels) spécialisé dans l'accès aux données. Il assure aussi le respect de la sécurité et de la confidentialité des données.

- Les données sont modélisées pour être exploitées avec un minimum de standard,par exemple avec un langage universel de requête, le SQL, dans un modèle relationnel; ce modèle est "lisible" par tout informaticien, sans avoir besoin d'analyser le code applicatif l'exploitant.

- Le volume de données ne cesse de croître, il est nécessaire de mettre en oeuvre des techniques améliorant la performance d'accès aux données : c'est la problématique résolue par les moteurs de bases de données (puissance algorithmique spécialisée), et la modélisation relationnelle (accès aux seules données nécessaires à un besoin, dans une requête)..

Un peu d’histoire ...

Le terme *database* (base de données) est apparu en 1964 pour désigner une collection d'informations partagées par différents utilisateurs d'un système d'informations militaire.

Les premières bases de données hiérarchiques sont apparues au début des années 1960. Les informations étaient découpées en deux niveaux de hiérarchie : un niveau contenait les informations qui sont identiques sur plusieurs enregistrements de la base de données. Le découpage a ensuite été étendu pour prendre la forme d'un diagramme en arbre.

En 1965, Charles Bachman conçoit l'architecture Ansi/Sparc encore utilisée de nos jours. En 1969, il créa le modèle de données réseau au sein du consortium CODASYL pour des applications informatiques pour lesquelles le modèle hiérarchique ne convient pas. Charles Bachman a reçu le prix Turing en 1973 pour ses « contributions exceptionnelles à la technologie des bases de données ».

En 1968, Dick Pick crée Pick, un système d'exploitation contenant un système de gestion de base de données « multivaluée » (SGBDR MV).

En 1970, Edgar F. Codd note dans sa thèse mathématiques sur l'algèbre relationnelle qu'un ensemble d'entités est comparable à une famille définissant une relation en mathématiques et que les jointures sont des produits cartésiens. Cette thèse est à l'origine des bases de données relationnelles. Edgar F. Codd a reçu le prix Turing en 1981.

Le modèle entité-association a été inventé par Peter Chen en 1975 ; il est destiné à clarifier l'organisation des données dans les bases de données relationnelles.

Un peu d’histoire ...

Dans le modèle relationnel, la relation désigne l'ensemble des informations d'une table, tandis que l'association, du modèle entité-association, désigne le lien logique qui existe entre deux tables contenant des informations connexes.

Les premières bases de données étaient calquées sur la présentation des cartes perforées : réparties en lignes et colonnes de largeur fixe. Une telle répartition permet difficilement de stocker des objets de programmation ; en particulier, elles ne permettent pas l'héritage entre les entités, caractéristique de la programmation orientée objet.

Apparues dans les années 1990, les bases de données objet-relationnel utilisent un modèle de données relationnel tout en permettant le stockage des objets. Dans ces bases de données les associations d'héritage des objets s'ajoutent aux associations entre les entités du modèle relationnel.

Base de données : Qu’est-ce donc ?

Une base de données est un « conteneur » stockant des données telles que des chiffres, des dates ou des mots, pouvant être retraités par des moyens informatiques pour produire une information ; par exemple, des chiffres et des noms assemblés et triés pour former un annuaire.

C'est la pièce centrale d'un système d'information ou d'un système de base de données (ou base de données tout court), qui régit la collecte, le stockage, le traitement/transformation et l'utilisation de données. Ce dispositif comporte souvent un logiciel moteur, des logiciels applicatifs, et un ensemble de règles relatives à l'accès et l'utilisation des informations.

Le système de gestion de base de données (SGBD) est une suite de programmes qui manipule la structure de la base de données et dirige l'accès aux données qui y sont stockées.

Une base de données est composée d'une collection de fichiers ; on y accède via le SGBD qui reçoit des demandes de manipulation du contenu et effectue les opérations nécessaires sur les fichiers. le SGBD cache la complexité des opérations et offre une vue synthétique sur le contenu. Le SGBD permet à plusieurs usagers de manipuler simultanément le contenu, et peut offrir différentes vues sur un même ensemble de données.

Base de données : Qu’est-ce donc ?

Le recours aux bases de données est une alternative au procédé classique de stockage de données, par lequel une application place des données dans des fichiers manipulés par l'application.

Il facilite le partage des informations, permet le contrôle automatique de la cohérence et de la redondance des informations, la limitation de l'accès aux informations et la production plus aisée des informations synthétiques à partir des renseignements bruts.

La base de données a de plus un effet fédérateur : dans une collectivité utilisant une base de données, l'administrateur de données (DA) organise le contenu de la base d'une manière bénéfique à l'ensemble de la collectivité, ce qui peut éviter des conflits dus à des intérêts divergents entre les membres de la collectivité.

Système de gestion de base de données (SGBD) : Qu'est-ce donc ?

Un système de gestion de base de données est un ensemble de logiciels qui manipulent le contenu des bases de données. Il sert à effectuer les opérations ordinaires telles que rechercher, ajouter ou supprimer des enregistrements (*Create*, *Read*, *Update*, *Delete* abrégé ***CRUD***), manipuler les index, créer ou copier des bases de données).

Les mécanismes du système de gestion de base de données visent à assurer la cohérence, la confidentialité et la pérennité du contenu des bases de données. Le logiciel refusera qu'un usager modifie ou supprime une information s'il n'y a pas été préalablement autorisé ; il refusera qu'un usager ajoute une information si celle-ci existe dans la base de données et fait l'objet d'une règle d'unicité ; il refusera également de stocker une information qui n'est pas conforme aux règles de cohérence telles que les règles d'intégrité référentielle dans les bases de données relationnelles.

Le système de gestion de base de données adapte automatiquement les index lors de chaque changement effectué sur une base de données et chaque opération est inscrite dans un journal contenu dans la base de données, ce qui permet d'annuler ou de terminer l'opération même en cas de crash informatique et ainsi garantir la cohérence du contenu de la base de données.

SGBD - R(Relationnel)

Une **base de données relationnelle** est une base de données où l'information est organisée dans des tableaux à deux dimensions appelés des relations ou tables, selon le modèle introduit par Edgar F. Codd en 1970. Selon ce modèle relationnel, une base de données consiste en une ou plusieurs relations. Les lignes de ces relations sont appelées des nuplets ou enregistrements. Les colonnes sont appelées des attributs.

Les logiciels qui permettent de créer, utiliser et maintenir des bases de données relationnelles sont des systèmes de gestion de base de données relationnels (SGBDR).

Pratiquement tous les systèmes relationnels utilisent le langage SQL pour interroger les bases de données. Ce langage permet de demander des opérations d'algèbre relationnelle telles que l'intersection, la sélection et la jointure.

Page wikipedia : https://fr.wikipedia.org/wiki/Base_de_donn%C3%A9es_relationnelle

Moteur de SGBD

Partie centrale du SGBD, le moteur de base de données effectue les opérations d'enregistrement et de récupération des données.

Selon le SGBD, la base de données peut être composées d'un ou de plusieurs fichiers. Le rôle du moteur est de manipuler ces fichiers.

Les Usages : l'OLTP

Le Transactionnel ou **OLTP** (**O**n**L**ine **T**ransactionnal **P**rocessing)

Principe d'informatique désignant l'ensemble des moyens mis à disposition pour développer des applications permettant de créer, mettre à jour et supprimer des données de façon ponctuelle, dans une transaction.

Le système transactionnel :

- Traite des milliers de lignes mais une transaction ne traite souvent que quelques lignes à la fois
- Chaque transaction contient un volume faible de données

Les usages : l'OLAP

La vision décisionnelle de l'entreprise est une réflexion du haut vers le bas. Tout intervenant est décideur dans l'entreprise, mais les décisions de départ partent toujours de la direction, suivant ses axes stratégiques de développement, pour être déclinées en actions et décisions aux niveaux du dessous. On part d'une définition globale de l'activité de l'entreprise, pour descendre vers le niveau de détail nécessaire et suffisant pour chaque niveau de prise de décision :

C'est le principe de l'**OLAP**

OLAP : **O**n**L**ine **A**nalytical **P**rocessing

Principe d'informatique désignant l'ensemble des moyens mis à disposition pour analyser les données de l'entreprise, présentées sous forme de mesures calculées à la croisée d'axes d'analyses hiérarchiques, au niveau de détail nécessaire et suffisant pour chaque métier

NB : Cette technique est étudiée dans le module à la carte "Décisionnel"

No SQL : Qu'est-ce donc ?

En informatique et en bases de données, **NoSQL** désigne une famille de systèmes de gestion de base de données (SGBD) qui s'écarte du paradigme classique des bases relationnelles. L'explicitation la plus populaire de l'acronyme est Not only SQL (« pas seulement SQL » en anglais)

L'architecture machine en clusters induit une structure logicielle distribuée fonctionnant avec des agrégats répartis sur différents serveurs permettant des accès et modifications concurrentes mais imposant également de remettre en cause de nombreux fondements de l'architecture SGBD relationnelle traditionnelle, notamment les propriétés ACID.

NB : Cette année, nous étudierons MongoDB dans le cadre d'un TP sur le NoSQL.

Pages Wikipedia :https://fr.wikipedia.org/wiki/NoSQL

Au programme de cette année

- Les traitements côté serveur : comment améliorer la performance des applications en exploitant des capacités d'un moteur de SGBDR
- Le No SQL : Pour quoi faire ? Découverte de MongoDB
- Les outils de performance et de sécurité des données : réplication et partitionnement des SGBDR