OUTILS INTERNET II

Présentée par Mme Bensalem Hana

Université de Badji-Mokhtar Annaba

Promotion: MasterI ,GADM

Département d'informatique

CHAPITRE V
OUTILS DE MANIPULATION DE DONNÉES SUR INTERNET
(BASE DE DONNÉES &SERVEUR DE BDD)

BASES DE DONNÉES(BDD)

- une Base de Données est un ensemble structuré d'informations (données), centralisées ou réparties.
- Des utilisateurs et /ou des programmes peuvent manipuler les données de la base.
- La manipulation de base de données inclus: consultation, modification ou suppression de données existantes ou ajout de nouvelles données.
- La nature et l'étendue(champs et/ou table) de la manipulation sont définies selon les droits d'accès de chacun des utilisateurs.
- Les accès concurrents(en même temps par plusieurs postes) doit être géré.

LES TYPES DE BASES DE DONNÉES(1)

- **LES BASES DE DONNÉES RELATIONNELLES**:
  - elles stockent et fournissent des données reliant une dimension à une autre.
  - Les lignes des tables représentent ses données et les colonnes définissent les attributs .
  - les bases de données relationnelles ne sont pas conçues pour: le Big Data et l'ensemble de données trop large. Car cela posera un problème de performance.
  - L'API standard pour les bases de données relationnelles est **le Structured Query Language (SQL).**
- **LES BASES DE DONNÉES NOSQL (Not only SQL):**
  - émergées au début des années 2000 pour manipuler plus rapidement des données massive.
  - La technologie NoSQL consiste à distribuer le chargement des données sur plusieurs **hôtes** à mesure que le volume augmente.
  - Ces bases de données sont conçues pour gérer divers types de données, notamment des données non structurées, semi-structurées, et des données à grande échelle. Exemple: les bases de données orientées graph.

- Pour **analyser d'importantes quantités de données non structurées**, ou des données stockées sur plusieurs serveurs cloud virtuels, une BDD NoSQL est idéale.
- L'utilisation d'une solution relationnelle ou NoSQL dépend de la nature du workload et des données sous-jacentes(des enregistrements qui ont des structures variables ne se marient pas aux modèles relationnels.).

LES TYPES DE BASES DE DONNÉES(2)

- **LES BASES DE DONNÉES CLÉ/ VALEUR**:
  - elles contiennent une **clé** unique accompagnée d'un champ de données (la **valeur**).
  - elles ont le mérite d'être efficace pour les opérations de lecture et d'écriture,
  - elles sont très **flexibles** permettent de stocker rapidement de grandes quantités de données.
- **LES BASES DE DONNÉES GRAPH**:
  - Au sein de ces systèmes, les données sont stockées dans des **graphes**.
  - Les graphes contiennent des nœuds, des arêtes et des propriétés.
  - Les **bases de données graph** ont pour objectif de considérer la relation entre les données de la même façon que la donnée elle-même.
  - Car elle stocke la connexion des données, l'administrateur n'a donc plus besoin d'effectuer des jointures.

LES TYPES DE BASES DE DONNÉES(3)

- **LES BASES DE DONNÉES DE SÉRIES CHRONOLOGIQUES:**
  - Il s'agit d'une base qui stocke les données avec un **horodatage**.
  - Elles permettent de suivre l'évolution d'une valeur au cours du temps.
  - Elles connaissent la croissance la plus rapide en 2024.
- **LES BASES DE DONNÉES ORIENTÉES OBJET:**
  - Il s'agit d'un système qui présente ses données sous forme d'objets et de classes.
  - Et suit les règles de la programmation orientée objet.
  - L'**objet** est une entité réelle et la **classe** est une collection d'objets.
- Les **bases de données cloud** :
  - sont des bases de données installées dans des plateformes de **Cloud Computing**.
  - Les utilisateurs peuvent soit exécuter des bases de données sur le Cloud de manière indépendante ou acheter l'accès à un service de base de données, géré par un fournisseur **Cloud**.
  - Ces BDD offrent une disponibilité plus élevée.

ARCHITECTURE DE DÉPLOIEMENT DE BDD

- **LES BASES DE DONNÉES CENTRALISÉES**:
  - les données sont stockées à un seul endroit ou elles peuvent être modifiées.
  - La **base de données centralisée** est principalement utilisée par des entreprises et des organisations .
- **LES BASES DE DONNÉES DISTRIBUÉES**:
  - est un ensemble de BDD connectées entre elles et localisées à différents endroits.
  - Comme les données sont accessibles par différents **réseaux**, les BDD doivent être plus sécurisée qu'une base de données centralisée.

LES BASES DE DONNÉES EMBARQUÉES:

- Une base de données embarquée est un SGBD plus léger, car il est réduit sous la forme de **composants logiciel**.
- Ce système de base de données est lié dynamiquement avec un logiciel.
- Généralement, cette base de données est composée d'un **fichier unique** dont le format est identique quel que soit l'ordinateur utilisé.

- **LES BASES DE DONNÉES SPATIALES:**
  - Utilisées par les **systèmes d'information géographiques** et les **outils de conception assistée par ordinateur**.
  - elles stockent des données géométriques comme des points, des lignes, des coordonnées, des volumes ou encore des dimensions.

SGBD & LANGAGES

- Trois langages doivent être distinguer selon leur rôle:

SQL dispose de ces couches

  - La description des données (type, longueur, etc)est réalisées dans un langage de définition des données (LDD);exemple: CREATE de SQL .

  - La manipulation de données(ajout, suppression,etc) se fait par le biais d'un langage de manipulation de données(LMD) exemple: INSERT de SQL.

  - Le contrôle de données est décrit en utilisant un langage de contrôle des données(LCD).exemple: les contraintes d'integrité de SQL tel que la clause DEFAULT (qui définit la valeur par défaut).

```php
<?php
$servername = "localhost";
$username = "username";
$password = "password";

// Create connection
$conn = new mysqli($servername, $username,
$password);
// Check connection
if ($conn->connect_error) {
  die("Connection failed: " . $conn-
>connect_error);
}

// Create database
$sql = "CREATE DATABASE myDB";
if ($conn->query($sql) === TRUE) {
  echo "Database created successfully";
} else {
  echo "Error creating database: " . $conn-
>error;
}

$conn->close();
?>
```

CREATION DE BDD

```php
<?php
$servername = "localhost";
$username = "username";
$password = "password";

// Create connection
$conn = new mysqli($servername, $username, $password);
// Check connection
if ($conn->connect_error) {
  die("Connection failed: " . $conn->connect_error);
}

// Create database
$sql = "CREATE DATABASE myDB";
if ($conn->query($sql) === TRUE) {
  echo "Database created successfully";
} else {
  echo "Error creating database: " . $conn->error;
}

$conn->close();
?>
```

```
"CREATE TABLE MyGuests (
id INT(6) UNSIGNED AUTO_INCREMENT PRIMARY KEY,
firstname VARCHAR(30) NOT NULL,
lastname VARCHAR(30) NOT NULL,
email VARCHAR(50),
reg_date TIMESTAMP DEFAULT CURRENT_TIMESTAMP ON UPDATE
CURRENT_TIMESTAMP
)";
```

CREATION DE TABLE

Requête sql:

CREATE TABLE table1(nom1 type1(taille),etc)

Table1,nom1: des nom définit par l'utilisateur

Type1 peut être:

VARCHAR, INT,DECIMAL,DATE,TEXT…

REQUETTES SQL(INSERT)

- Syntaxe:
- INSERT INTO `nametable` (`namefield1`, `namefield2`,…) VALUES ([value-1],[value-2],…);
- Exemple (sous php):

```
$sql = "INSERT INTO MyGuests (firstname, lastname, email) VALUES ('John', 'Doe', 'john@example.com')";

if ($conn->query($sql) === TRUE) {
  echo "New record created successfully";}

  else {  echo "Error: " . $sql . "<br>" . $conn->error;}
```

REQUETES SQL(SELECT)

- Syntaxe:
- SELECT * FROM `nametable` WHERE namefield1=‘ value-1’& namefield2= ‘value-2’;
- Exemple:

```
$sql = "SELECT id, firstname, lastname FROM MyGuests WHERE lastname='Doe'";

$result = $conn->query($sql);

if ($result->num_rows > 0) {  // output data of each row
  while($row = $result->fetch_assoc()) {
    echo "id: " . $row["id"]. " - Name: " . $row["firstname"]. " " . $row["lastname"]. "<br>";
  }}

else {  echo "0 results";}
```

REQUETES SQL(DELETE)

- Suppression d'un élément de la table
- DELETE FROM `namtable` WHERE condition

```
$sql = "DELETE FROM MyGuests WHERE id=3";
```

REQUETES SQL(UPDATE)

- Modifier à tout les coups:
  - UPDATE `namtabl` SET `namfield1`=[value-1],` namfield1 `=[value-2],… WHERE 1

- Ne modifier que si l'élément rempli une condition, après le WHERE, ou plusieurs modifications.
  - $sql = "UPDATE MyGuests SET lastname='Doe' WHERE id=2";

LES SYSTÈMES DE GESTION DE DONNÉES (SGBD)

- Ses fonctions:
  - La création de la base de données
  - L'interaction avec la base de données
- Ses composants:
  - De manipulation de données,
  - De description des données qui permettent la :
    - Construction de la structure de base
    - Usage des types de données
    - Définition des relations entre les données
    - Définition des contraintes d'intégrités
  - De partage des ressources(gestion des accès concurrent),
  - etc.
- Les SGBD peuvent être sous forme de composant logiciel, de serveur, de logiciel applicatif ou d'environnement de programmation.

SERVEUR DE BASES DE DONNÉES

- Le serveur est à l'écoute des requêtes provenant des clients. Il y répond en faisant appel aux fonctions d'un S.G.B.D, c'est à dire qu'il tourne tout le temps et scrute s'il y a des requêtes à traiter. Le client et le serveur dialogue de la manière suivante:
- **Dans un type de dialogue à session :**
  - La fonction serveur reçoit la demande de connexion, le S.G.B.D. vérifie les droits et crée le contexte
  - la fonction serveur reçoit les requêtes et les transmet au S.G.B.D.
  - le S.G.B.D. exécute les requêtes, transmet le résultat à la fonction serveur .
- **Dans un type de dialogue par messages:** toutes les demandes ainsi que les réponses sont stockées dans une file d'attente FIFO de messages.

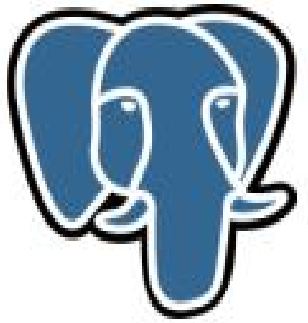
ORACLE

- **Oracle Database** :
  - est un système de gestion de base de données relationnelle et relationnel-objet, crée par Oracle dans les années 70.
  - Elle est la première **database** conçue pour le grid computing.
  - Le **grid computing** entreprise est la technique la plus flexible et rentable pour gérer les systèmes informatiques et les applicatifs.
- License :commercial
- AVANTAGES DE ORACLE
  - Bonne capacité de sauvegarde et de récupération des données
  - Régulièrement mis à jour
  - Grande **portabilité**
  - Gère facilement plusieurs bases de données au sein d'une même **transaction**
  - La base de données la plus populaire selon le classement DB-Engines
- INCONVÉNIENTS DE ORACLE
  - Le prix
  - Un système difficile à maîtriser

MYSQL

- MySQL est un **SGBD** (Système de Gestion de Base de Données) relationnelle, crée par MySQL AB en 1995. Appréciée des professionnels et des particuliers, le plus utilisée au monde.
- Licence : Licence publique générale GNU version 2 (GPLv2) : système de gestion de base de données open-source
- AVANTAGES DE MYSQL
  - Facile à utiliser
  - De bonnes **performances**
  - Plusieurs fonctionnalités pour sécuriser ses données
  - Open-source
- INCONVÉNIENTS DE MYSQL
  - Difficilement **scalable:**les performances du système se détériorent à partir d'un certain volume de données(la montée en charge).

POSTGRESQL



- PostgreSQL est un **SGBD relationnelle** et objet créé par le groupe PostgreSQL en 1996. Il s'agit d'un outil libre, non contrôlé par une entreprise, mais par une communauté mondiale de développeurs et d'organisations.
- Licence : Licence PostgreSQL
- **Avantages de PostgreSQL**
  - **Open-source**
  - Facile à utiliser
  - Possède un type de données défini par l'utilisateur
- **Inconvénients de PostgreSQL**
  - Documentation extensible en anglais seulement
  - Vitesse de lecture relativement faible

MICROSOFT SQL SERVER

- **Microsoft SQL Server**, abrégé MSSQL est un **SGBD relationnel** crée par Microsoft en 1989. Cet outil se démarque de la concurrence grâce à un large choix d'options offertes selon la version choisie.
- Licence : Licence propriétaire et EULA
- **Avantages de SQL Server**
  - Bonne **sécurité** des données
  - Facile à installer et à configurer
  - De nombreux outils pour gérer l'ensemble des tâches en entreprise
- **Inconvénients de SQL SERVER**
  - Le prix
  - Le manque de **compatibilité** avec des produits ne provenant pas de Microsoft
  - Besoin de machines performantes.

MONGODB

- Mongo DB est un **SGBD orienté documents** pouvant être répartis sur plusieurs ordinateurs sans schéma prédéfini des données. MongoDB a été créé en 2009 par MongoDB, Inc. Elle est reconnue pour sa haute **scalabilité** et **accessibilité**.
- Licence : Server Side Public License
- **Avantages de MongoDB**
  - Facile à installer
  - De très bonnes performances
  - Prise en charge des requêtes ad hoc
  - Base de données **évolutive** horizontalement
- **Inconvénients de MongoDB**
  - L'imbrication des documents est limitée
  - Ne supporte pas les jointures
  - Augmente l'utilisation de la **mémoire** inutilement

SQLITE

- **SQLite** est une librairie en C qui intègre un SGBD relationnelle de hautes performances. Elle a été créée en 2000 par Richard Hipp. SQLite est le **moteur de base de données** le plus utilisé au monde, elle est utilisée par de nombreuses entreprises opérant dans le secteur des nouvelles technologies comme Firefox, Apple ou Skype.
- Licence : Domaine public
- **Avantages de SQLITE:**
  - Léger
  - De bonnes **performances**
  - Aucune installation requise
  - Facile à utiliser
  - **Open-source**
- **Inconvénients DE SQLITE**
  - Difficilement scalable, il peut ne pas être adapté à des charges de travail à haute concurrence
  - Manque de fonctionnalités multi-utilisateur
  - La taille des **bases de données** est limitée à 2 Go dans la plupart des cas

DATASET

- est une collection de données, organisée en un format structuré.
- Peut inclure différents types de données :numérique, texte, image, etc.
- Utilisé dans la recherche, l'analyse de données et les projets d'apprentissage machine.
- Peut être créé de différent sources: expérimentation, enquête ou sondage ou même une base de donnée existante.
- Peut être partagées en privé ou en public(pour reproduire ou valider d'autre résultats de recherche)

DATASET LIBRE

- Selon le type de données:
- **DIVERS:**
  - GOOGLE DATASET SEARCH: lancé en 2018,est un moteur de recherche de données
  - KAGGLE: lancé en 2010,est une plateforme libre reconnus. Elle offre: une collaboration basée-cloud pour les data-scientists, des outils éducatifs pour l'enseignement de l'intelligence artificielle, …
- Business et finance:
  - Datahub.io :couvre une grande variété de topics mais se concentre sur l'économie et les finances
- Santé:
  - Global Health Observatory Data Repository: est un portail des statistics relié à la santé du globe.
- …

BDD OU DATASET

Par conséquence les fonctionnalités de chacune différent aussi

- BDD et Dataset se diffèrent par leur :
- Objectifs:
  - dataset utilisée pour l'analyse et la modélisation et son de petite taille relativement aux BDD
  - Une BDD est utilisée pour le stockage et la gestion des données, pour de longue durées
- Structure:
  - les formats de stockage des dataset : CSV,JSON,EXCEL…

OUTILS SUR INTERNET POUR LES DONNÉES MASSIVES

- Plusieurs solutions existent sur internet, chaque solution ou technologie dépend de:
- l'infrastructure matérielle en place
- des missions :entre la collecte, le traitement, le nettoyage, la clusterisation, l'analyse en temps réel et le machine learning,
- aussi plusieurs outils spécifiques qui peuvent s'intégrer ensemble afin de couvrir tous les besoins des Data Analysts et des Data Scientists.

OUTILS SUR INTERNET POUR LES DONNÉES MASSIVES(1)

- ***HADOOP:*** C'est une solution open source créée par Apache qui permet de traiter de très larges volumes de données grâce à un fonctionnement déporté sur serveur.
- utilise un système de fichiers distribué permettant une vitesse de traitement très importante grâce à des transferts élevés entre les noeuds d'un serveur.
- C'est un outil utilisé par les plus grandes entreprises de la technologie, comme Google ou Yahoo.

OUTILS SUR INTERNET POUR LES DONNÉES MASSIVES(2)

- ***CASSANDRA:*** Apache Cassandra est une technologie de gestion des bases de données distribuée NoSQL , toujours disponible et très flexible en terme d'adaptabilité et de scalabilité.
- utilisée par des grandes entreprises comme Facebook, Netflix, Twitter, Cisco ou eBay en raison de sa très haute vélocité déployable sur de multiples serveurs.
- prend en charge les données: structurées, non structurées ou semi-structurées
- gère particulièrement bien les changements dynamiques pour s'adapter aux évolutions des besoins.

OUTILS SUR INTERNET POUR LES DONNÉES MASSIVES(3)

- ***OPENREFINE:*** Initialement baptisé Freebase Gridworks avant d'être achetée par Google en 2010 (puis abandonnée en 2012), c'est une solution désormais open source conçue pour travailler avec des données non structurées et désorganisées.
- OpenRefine (ou GoogleRefine)est simple d'utilisation quelques clics suffisent pour transformer un jeu de données brutes en données exploitables, pertinentes et uniques.

OUTILS SUR INTERNET POUR LES DONNÉES MASSIVES(4)

- ***STORM:*** Storm est une autre solution open source qui permet de traiter des calculs complexes en temps réel.
- Technologie particulièrement résiliente et tolérante aux pannes, Storm peut monter en charge dynamiquement en ajoutant des serveurs selon les besoins.
- C'est une solution relativement simple à déployer.
- distribué et développé par l'Apache Software Foundation.

OUTILS SUR INTERNET POUR LES DONNÉES MASSIVES(5)

- ***RAPIDMINER:*** Rapidminer est une technologie et un environnement de travail qui fournit tous les outils pour analyser et préparer des données non structurées. À travers une interface soignée
- est utilisé pour des projets de machine learning, deep learning, text mining et d'analyses prédictives. Un outil régulièrement cité par Gartner et Forrester comme l'un des plus puissants en termes de traitement et d'analyse des données.
- Voir aussi: www.appvizer.fr/analytique/big-data
- …