Bases de données avancées

Modélisation des données et bases de données NoSQL

Survol

- Rôle de la base de données
- Modélisation des données
- Type des bases de données
- Comparaison des bases de données

Rôles de la base de données

- Traditionnellement
  - Enregistrer l'information
  - Accéder efficacement l'information
- Supporter les applications
  - Exigeantes en calcul – *compute-intensive*
  - Exigeantes en données – data-intensive

Rôles de la base de données

Application

- Serveur multimédia maison
  - Conserver la liste des film (et informations connexes)
  - Diffuse le film au travers d'une connexion réseau

Application exigeante en données

- Service de diffusion de film (Netflix)
  - Conserve la liste des film (et informations connexes)
  - Conserve les informations sur les membres
  - Calcule des scores de recommandations
  - Diffuse le film au travers de l'internet
  - Optimise l'utilisation de bande passante

Rôles de la base de données

- Besoins basés sur l'application
  - Fiable – *reliable*
    - *Fonctionne au niveau attendu malgré les erreurs humaines / matérielles*
  - Maintenable – *maintenable*
    - *Tous les intervenants peuvent travailler sur le système de façon productive*
  - Extensible – *scalable*
    - *Facile d'ajuster le système pour gérer le trafic futur anticipé*

Rôles de la base de données

- Serveur multimédia maison
  - Fiable
  - Maintenable
  - Extensible

- Fiable
  - Repartir le serveur à la main
- Maintenable
  - Une seule personne qui modifie le système
- Extensible
  - Les personnes de votre famille l'utiliseront

Rôles de la base de données

Design

- 1 ordinateur
  - Application web Flask
  - Base de données MySQL
  - Serveur de streaming Open Source
- 1 disque dur de backup

Rôles de la base de données

- Service de diffusion de film (Netflix)
  - Fiable
  - Maintenable
  - Extensible

- Fiable
  - Diffuser 24h/24, 7j/7
  - Via des connections internet non contrôlées
- Maintenable
  - Plusieurs équipes d'ingénieurs doivent ajouter des fonctionnalités
  - Plusieurs équipes opérationnelles corrigent des bugs
- Extensible
  - Catalogue qui croît à chaque semaine
  - Utilisable partout sur la planète
  - Peu importe la plateforme

Rôles de la base de données

Design

- Service web redondant
- Service de diffusion en direct
  - Optimisation géographique des données
  - Redondance dans les données
- Service de calcul de recommandation
  - Calcul dispendieux

- Chaque élément a des besoins différents en terme de base de données
- Une approche unique n'est peut être pas optimale
- Définition des besoins, des données et sélection de l'outil approprié

Rôles de la base de données

SQL

- ACID

NoSQL

- Pas ACID

Est-ce la seule métrique à considérer?

Rôles de la base de données

Propriétés ACID

- Atomicité
- Cohérence
- Isolation
- Durabilité

Cas d'utilisation

- Modification du profil d'utilisateur
- Sauvegarde des films préférés
- Sauvegarde d'une recommandation

Propriétés ACID – Ne sont pas les seules choses qui comptent

Rôles de la base de données

- Une application offre tacitement des garanties
  - Facile à atteindre dans le cours de BD 1
  - Cas plus complexe avec des système composites (vrai vie)
- Questions clés
  - Comment s'assurer que les données restent correctes, complètes et cohérentes
  - Comment offrir une bonne performance malgré des parties qui se dégradent

Modélisation des données

- Données relationnelles
  - Cas d'utilisation précis: données financières
  - Codd 1970
  - Relations / Tuples (table et enregistrements)
  - Hégémonie de l'empire SQL depuis les années 1980
- Hypothèses sur les données:
  - Toutes les données ont le même format
  - Denses
  - Inter-reliées
  - Centralisées

Modélisation des données

- Hypothèses non respectées par les données massives
  - Images
  - Textuelles
  - Données de senseurs
  - Données géospatiales
- Caractéristiques
  - Non-uniformes
  - Clairsemées
  - Non-structurées
  - Potentiellement distribuées

Modélisation des données

- Mouvement "Pas seulement SQL" - Not only SQL #NoSQL
  - Ne cherche pas à remplacer SQL
  - Persistence Polyglotte
  - Couvrir les cas d'utilisation non relationnel
    - Haute extensibilité et niveau de performance très élevées
    - Solutions ouvertes et gratuites
    - Résoudre des requêtes difficiles à effectuer
    - Besoin de modèles de données expressif

Modélisation des données

- Revisite un problème des années 1970
  - Modélisation non relationnelle
    - Document (hierarchical model)
    - Graphe (network model)

Modélisation des données - Documents

- Collection de clé-valeurs

```
{
    "Nom":"Jean-Thomas",
    "Prenom":"Baillargeon",
    "DateFavorite":isoDate("1926-08-13"),
    "MetsFavorits":["Fajitas", "Pokébol", "Pizza"]
    "Taille":172.72
}
```

- Modèle expressif et dynamique
- Correspondance avec les objets du domaine
- Document JSON / Dictionnaire Python
- Structure atomique - efficacité de lecture en mémoire

Modélisation des données - Documents

- Relations one-to-many représentées par une liste
  - Efficacité de rechercher dans des listes imbriquées
- Utilité des relations many-to-one
  - Orthographe cohérent
  - Enlever l'ambiguïté
  - Facilité de mise à jour
  - Support multilingue
  - Aide certaines recherches

Débat normalisation vs dénormalisation : on en reparlera

Modélisation des données - Documents

Aspects principaux

- Simplicité du code
  - Oui : pas de schéma, pas de shredding
  - Non: Cohérence validée du côté applicatif
  - Non: Relation many-to-many
- Flexibilité du schéma
  - schémat-on-read: la structure est validée à la lecture (runtime)
  - Aucun contrôle sur la source des données (Une table par "source" n'est pas pratique)
- Localité du document
  - Une seule lecture
- Atomicité des écritures
  - Tout le document est réécrit si la mise à jour est majeure

Modélisation des données

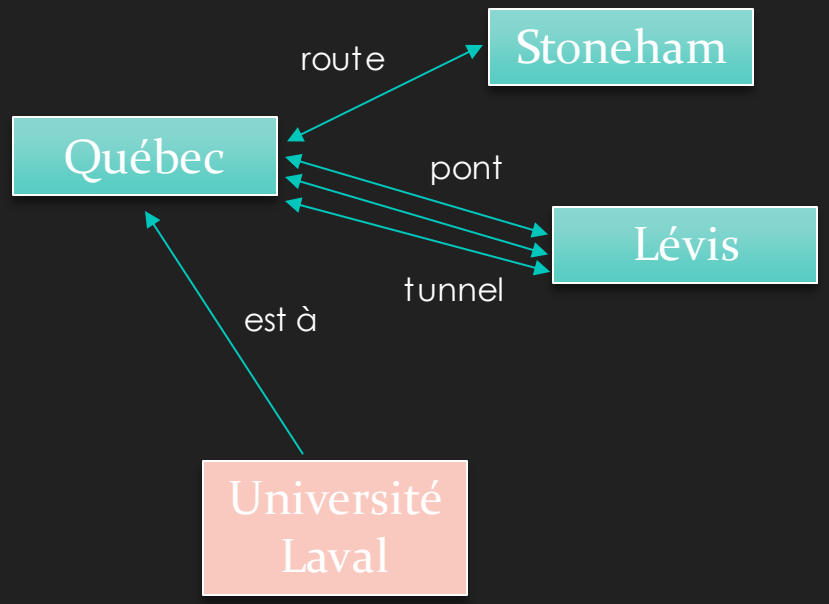
Facilité à gérer les relations many-to-many

Document **Relationnel** **Graphe**

|- Pas vraiment ---------- Un peu ouais ---------- Oui! -|

Modélisation des données - Graphe

- Sommet (entité / nœud) et côtés (relations)
- Idéal pour les objets simples avec des relations complexes
- Exemples de données
  - Réseau social
    - Sommet : Personnes / Pages, côtés : ami / Likes
  - Réseau roulier
    - Sommet : Intersection, côtés : routes
- Exemples de requêtes
  - Trouver les amis des amis
  - Trouver le chemin le plus court pour aller d'un nœud à l'autre



Modélisation des données - Graphe

Aspects principaux

- Algorithmes classiques pour les chemins les plus courts
- Calcul rapide sur des attributs d'entités reliées à plusieurs degrés de séparations
- Données hétérogènes
  - Différents types de relations entre des entités de mêmes types
  - Différents types d'entités peuvent être reliées ensemble

Type de bases de données

- Base de données relationnelles
  - RDBMS – MySQL, Oracle …
- Base de données NoSQL
  - Orienté colonnes
  - Orienté clé-valeur
  - Orienté documents – Mongo DB
  - Orienté graphe – Neo4J

https://db-engines.com/en/ranking

Type de bases de données - Colonnes

Propriété

- Format très similaire aux bases de données relationnelles
- Concept de famille de colonne permettant de regrouper les données utilisées ensemble
- Données stockées en mémoire par famille de colonnes
- Rapidité à faire des requêtes d'analyses sur des champs spécifiques

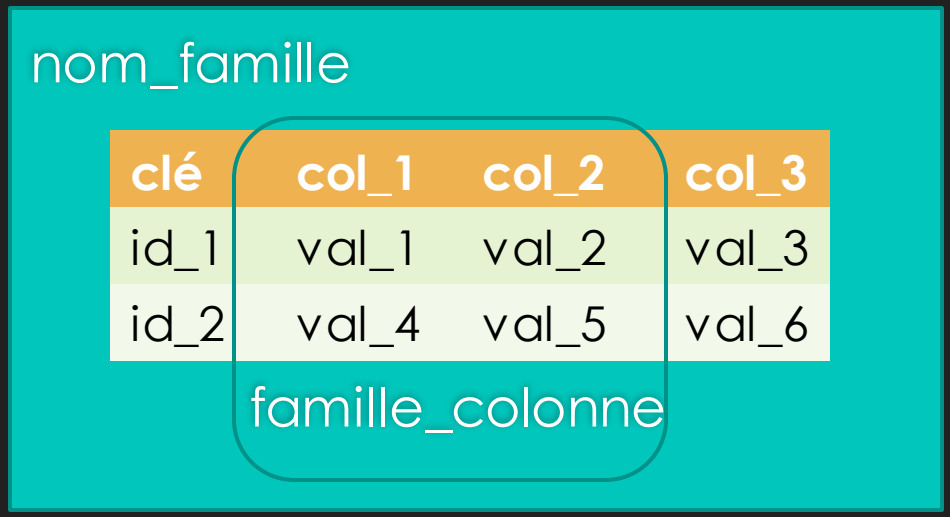
Structure

Une **famille** contient des **enregistrements**

Chaque **enregistrement** a une **clé unique**

Chaque **enregistrement** a des **colonnes**

Une **famille de colonne** a des **colonnes**



Type de bases de données - Clé Valeur

Propriété

- Les clés des enregistrements sont dans la mémoire vive de l'ordinateur
- Les valeurs sont quelconques (binaire, document …)
- Peuvent ou non être des bases de données "in-memory"
- Très efficaces
- Demande BEAUCOUP de mémoire vive

Structure

Un **Bucket** contient des **paires clé-valeur**

nom_bucket

```
string_1: [n'importe quoi]
```

```
string_2: [n'importe quoi]
```

Type de bases de données - Document

Propriété

- Contient des données sous formats documents (JSON)
- Contenu des collections plus ou moins structuré
  - Les documents n'ont pas nécessairement le même type

Structure

Une **collection** contient des **documents**

nom_collection

```
{
    "_id":string_1,
    "key_1": …,
    "key_2": …
}
```

```
{
    "_id":string_2,
    "key_2": …
}
```

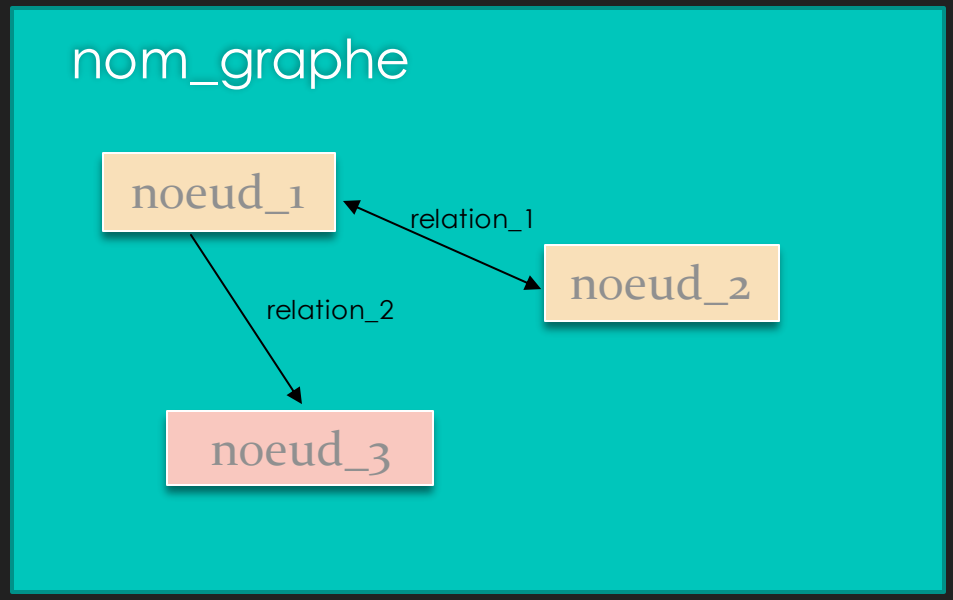
Type de bases de données - Graphe

Propriété

- Contient des données sous formats graphe attribué
- Très différent des autres BDs
  - L'unité de base n'est pas la paire clé-valeur
- Opération JOIN explicitement modélisée dans la base de données

Structure

Un **graphe** contient des **nœuds** reliés par des **relations**



Comparaison type base de données

- Cas d'utilisation – votre serveur multimédia en Nextflix !
  - Avec toutes les garanties que Netflix offre déjà
- Cas 1: Obtenir la liste des **films disponibles à l'écoute** ainsi que leur information connexe
  - Identifiant unique, titre, date de réalisation, acteurs, ...
- Cas 2: **Calculer des recommandations** basées sur les films aimés par ses amis
- Cas 3: Présenter les pochettes des recommandations **déjà calculées**.
- Cas 4: Faire le suivi des **factures**
- Cas 5: Afficher un tableau de bord qui contient des **statistiques dynamiques** sur les revenus **par pays et par type de film** préféré par les membres

Comparaison des bases de données

- Cas 1: Mise à l'échelle de votre serveur SQL
  - Problème 1: Latence impardonnable pour les clients hors Amérique.
  - Étude de mise en marché
    - [2016] 4669 serveurs à 245 endroits
    - Ping moyen par ville (Québec-Mtl : 4ms, Québec-Hanoi 250 ms)
  - Solution 1: Répliquer la bd SQL en Europe, en Afrique, en Asie et en Australie.
  - Effet de bord 1: Des scripts de synchronisations doivent être mis en place pour conserver la cohérence entre les bases de données.
  - Effet de bord 2 : Le service est inaccessible lorsque les bases de données se mettent à jour.
  - Problème 2: Vos clients sont partout sur la planète et assument que le service est disponible 24h/24
  - Solution 2: ….

Comparaison des bases de données

- À vous de jouer
- Pour chaque cas d'utilisation
  - Choisir le type de bd qui s'appliquer
  - Présenter un exemple de données

Conclusion

Rôle d'une BD

- Une BD supporte une application (et non l'inverse)
  - Choisir le bon type de BD par rapport au cas de figure
- Définir les besoins et les garanties de l'application
- Fiabilité, extensibilité, maintenabilité

Modèles de données

- Lien many-to-many (document – relationnel – graphe)

4 Type des BD NoSQL

- Document
- Clé Valeur
- Colonnes
- Graphe