Ubiquité et confidentialité des données

Philippe Pucheral

Ubiquité et confidentialité des données

Philippe Pucheral

Laboratoire PRiSM, Université de Versailles, 78035 Versailles Cedex
& projet SMIS, INRIA Rocquencourt, 78153 Le Chesnay Cedex

1. Introduction

Depuis leur avènement dans les années 1960, de nombreux résultats théoriques et pratiques ont jalonné l'histoire des bases de données. Quel meilleur résumé de cette riche histoire que de citer les trois prix Turing de l'ACM ayant récompensé les avancées les plus fondamentales de ce domaine ? Charles Bachman a reçu le prix Turing 1973 (encadré par Dijkstra et Knuth) pour son travail de pionnier concernant la modélisation conceptuelle des bases de données et la réalisation du premier Système de Gestion de Bases de Données (SGBD) commercial, intitulé IDS (Integrated Data Store). Par un article d'anthologie publié en 1970, Ted Codd (Prix Turing 1981) a jeté les bases du modèle relationnel, apportant les fondements mathématiques manquants jusqu'alors au domaine des bases de données. Trente ans plus tard, ce modèle est toujours la référence et reste au cœur des SGBD les plus répandus tels que Oracle, DB2, SQLServer ou même Access. Ces fondements mathématiques ont permis des avancés significatives sur la conception de schéma, les langages de requêtes déclaratifs ou encore l'intégrité sémantique des données et ont ainsi largement contribué à simplifier l'usage et l'administration des bases de données. Pour la première fois, il devenait possible à un sujet non informaticien de créer, organiser, interroger, modifier et partager des données sans se soucier de leur représentation binaire. Victimes de leur succès, les SGBD ont alors été confrontés à une croissance quasi-exponentielle du volume des données à traiter, du nombre d'utilisateurs simultanés et du débit transactionnel escompté[1]. La communauté bases de données s'est donc penchée sur les stratégies d'indexation, d'optimisation de requêtes, d'exploitation du parallélisme et de gestion de transactions afin de traiter de façon efficace et cohérente de très grandes bases de données partagées par de multiples utilisateurs. Jim Gray a été récompensé par le prix Turing 1998 pour l'ensemble de sa contribution sur ces thèmes.

[1] Les bases de données en ligne de plusieurs téra-octets sont aujourd'hui monnaie courante et le volume des entrepôts de données et des banques de données évoluent d'ores et déjà vers le péta-octets de données. La performance des SGBD en terme de débit transactionnel se mesure elle en milliers de transactions par seconde selon les bancs d'essais publiés par le Transaction Processing Council (TPC).

Si cet héritage prestigieux est aujourd'hui largement exploité, force est de constater la profonde mutation qui s'amorce dans l'architecture des systèmes d'information et leur usage. L'information est aujourd'hui omniprésente et distribuée dans une multitude de sources de données autonomes et fortement hétérogènes (données tabulaires, fichiers structurés, documents XML semi-structurés, données multimédia telles que images, sons, vidéo). L'internet facilite le partage à grande échelle de ces sources de données et modifie par là-même les modes d'accès à l'information (nombre très élevé de clients et de ressources, diffusion sélective, abonnement et notification, échanges pair à pair, accès ubiquitaire via des terminaux mobiles …), l'exploitation de cette information (entrepôts de données, extraction de connaissances, grilles de données et de traitements) ainsi que les règles même du partage (préservation de la confidentialité, respect du droit de propriété).

Pour répondre à cette profonde mutation technologique et à l'évolution des usages qui l'accompagne, la communauté internationale en bases de données a ouvert de grands chantiers au cours de la dernière décennie. Parmi les plus représentatifs, et sans prétendre à l'exhaustivité, nous pouvons citer :

- la gestion de données semi-structurées (interrogation, mise à jour, diffusion, indexation, versionnement de documents XML) qui étend le champ d'application des SGBD jusqu'alors cantonnés à la gestion de données fortement structurées (typiquement des tables) ;
- la médiation de données permettant de construire une vision centralisée, cohérente et uniforme d'un ensemble de sources de données, en masquant leur distribution et leur hétérogénéité mais tout en respectant leur autonomie (par exemple pour fédérer des bases ou banques de données exprimées selon des conventions différentes et gérées par des organismes différents) ;
- la médiation de programmes permettant la construction de chaînes de traitement ou *workflows* en assemblant des programmes distribués (eg. bibliothèques de composants, programmes scientifiques, services Web) tout en capitalisant les ressources de calcul disponibles sur le réseau (ou grille) ;
- la gestion de données partagées en pair à pair (P2P), ajoutant à la problématique de médiation une dimension fortement décentralisée (absence de contrôle global et de méta-données globales) ;
- la conception de composants bases de données adaptables s'opposant à une vision monolithique des moteurs de SGBD et permettant d'intégrer des fonctions bases de données (stockage, recherche associative, cohérence, réplication …) dans de multiples environnements informatiques ;
- la gestion de données multimédia avec la définition de nouvelles méthodes d'indexation permettant d'accélérer les recherches sur le contenu (par exemple retrouver toutes les images semblables à un modèle) et la prise en compte de relations spatiales et temporelles (par exemple suivi de la dérive d'une nappe de pollution à partir de données issues de capteurs) ;

- le rafraîchissement et l'exploitation d'entrepôts de données (gigantesques bases de données destinées à l'analyse et à la prévision et tirant leurs informations de systèmes en ligne) alimentés par des sources disponibles sur l'internet ;
- l'extraction de connaissance à partir des données (ECD) permettant d'interpréter notamment le contenu de ces entrepôts et d'y découvrir des relations sémantiques cachées entre les données ;
- l'interrogation de données à large échelle intégrant le traitement de critères approximatifs (par exemple, recherche d'un billet d'avion « bon marché ») et la personnalisation des résultats grâce à l'acquisition par apprentissage de profils d'utilisateurs ;
- la gestion de la mobilité des données et des utilisateurs ;
- et enfin la sécurisation des systèmes d'information face à des attaques de plus en plus nombreuses et variées.

Il serait illusoire en un seul chapitre de vouloir embrasser l'ensemble de ces problématiques. Nous avons donc pris le parti, dans les deux sections suivantes, de concentrer le discours sur la gestion de la mobilité et la préservation de la confidentialité des données. Ce choix ne traduit pas une quelconque importance relative des problématiques, les hiérarchiser serait sans objet. Il est plutôt motivé par l'impact que devraient avoir ces deux problématiques sur la vie quotidienne des individus (émergence de l'intelligence ambiante) et sur leurs préoccupations immédiates (quelles menaces sur la vie privée).

2. Ubiquité et intelligence ambiante

2.1. *Position du problème*

La prolifération des calculateurs mobiles ultra-légers communicants a permis l'émergence d'une nouvelle forme d'accès à l'information appelée communément *informatique ubiquitaire*. La notion d'ubiquité traduit la capacité d'accéder à des données, de n'importe où, n'importe quand et à partir de n'importe quel terminal. Le domaine d'application est très vaste. Il peut s'agir d'applications personnelles dans lesquelles un utilisateur veut pouvoir accéder à tout moment à des données publiques (météo, trafic routier, cours de la bourse, informations touristiques …) ou privées (données bancaires, agenda, dossier médical, bookmarks …). Il peut également s'agir d'applications professionnelles dans lesquelles des employés itinérants doivent accéder et partager en permanence, et où qu'ils se trouvent, des données relatives à leur entreprise ou à une tâche exécutée en commun.

Du fait de la miniaturisation permanente des équipements, l'informatique ubiquitaire a rapidement donné corps à une forme plus diffuse de l'informatique que l'on nomme *intelligence ambiante*. Ce terme symbolise le fait que les objets avec lesquels nous interagissons quotidiennement deviennent intelligents (doté de capacité de calcul et de mémorisation), conscients de leur environnement (capteurs), capables

d'agir dessus (actionneurs) et capables de communiquer entre eux. L'état actuel de la technologie permet déjà de rendre réalisables des projets que l'on aurait qualifiés il y a peu de science-fiction. Par exemple, le projet « Aware Home » [AWA03] laisse entrevoir des commodes mémorisant le contenu de leurs tiroirs pour faciliter une recherche ultérieure, des équipements s'adaptant aux habitudes des usagers (éclairage, chauffage, arrosage …) et les conseillant (menu diététique composé à partir du contenu du réfrigérateur et de la consommation passée). Ainsi, les bases de données elles-mêmes deviennent ambiantes. D'autres exemples concernent la santé (biocapteurs intégrés dans les vêtements ou implantés dans le corps humain) ou encore les transports (capteurs dans les véhicules, aide à la navigation)[2]. Plus généralement, Imielinski et Nath décrivent, à l'occasion du *VLDB 10 years award* récompensant leurs travaux sur les bases de données mobiles, leur vision d'un espace où les données sont omniprésentes, générées par une multitude de capteurs, analysées par des serveurs qui restituent une information enrichie à destination d'utilisateurs humains, de programmes ou d'objets intelligents [ImN02].

Cependant, les besoins des applications décrites précédemment sont à mettre en opposition avec les contraintes fortes induites par l'environnement matériel et logiciel disponible : faible débit des réseaux hertziens, déconnexions fréquentes (volontaires ou non), faibles capacités des terminaux mobiles en termes d'affichage, d'autonomie électrique, de puissance de traitement et de stockage, inadéquation des intergiciels conçus jusqu'à présent pour interconnecter des clients et des serveurs fixes. Cette confrontation des besoins et des contraintes génère de multiples problèmes de recherche difficiles. Bien que la communauté bases de données s'intéresse au problème de la mobilité depuis une dizaine d'années, il apparaît de plus en plus clairement que nous n'en sommes qu'à la genèse de cette thématique et que le champ d'investigation reste immense. L'objectif de cette section est d'identifier quelques verrous scientifiques et technologiques majeurs dans ce domaine.

2.2. *Verrous scientifiques et technologiques*

Cette section aborde les problèmes liés à la gestion de données spatio-temporelles, à la gestion de données embarquées dans des calculateurs mobiles ultra-légers, aux nouveaux modes d'accès à l'information et à la gestion de la cohérence des traitements en environnement mobile[3].

---

[2] En 2000, le rapport "Internet du futur" du RNRT estimait qu'un citoyen habitant dans un pays développé utiliserait dans les prochaines années environ 80 processeurs quotidiennement par le biais de systèmes enfouis et/ou mobiles. Cette prédiction risque d'être largement dépassée si l'on considère qu'une Peugeot 607 contient à elle seule 37 processeurs.

[3] L'auteur tient à remercier les membres de l'AS CNRS « Mobilité/Accès aux données », co-auteurs du document collectif [ASM03], dont cette sous-section est inspirée.

Modéliser et gérer efficacement les données spatio-temporelles

Les bases de localisation sont des bases de données particulières dont l'objectif est de retrouver rapidement la localisation d'un mobile (téléphone cellulaire, véhicule…) ou du moins de l'approximer avec une faible marge d'erreur. Cette approximation introduit les problèmes suivants [WXC98] :

- *Modélisation de la localisation :* alors que le déplacement d'un mobile est un événement continu, son reflet dans la base de données ne peut être que discret pour d'évidentes raisons de performance. Il faut donc pouvoir calculer les nouvelles positions d'un mobile via des fonctions d'approximation prenant en compte le temps. La modélisation de données mobiles a généré des approches jusqu'ici assez différentes qu'il faut rassembler dans un cadre unifié pour faciliter leur intégration dans un SGBD [GBE00].
- *Puissance d'expression du langage de requêtes :* le langage de requêtes doit être suffisamment puissant pour exprimer des recherches du type « trouver les hôtels les plus proches de la position que l'utilisateur aura atteinte dans une heure » . Il est donc nécessaire de manipuler à la fois les dimensions spatiale (points, lignes, régions, polygones) et temporelle. Par ailleurs, le langage de requêtes doit intégrer des modalités particulières (telles que sûrement, éventuellement …) pour tenir compte de l'incertitude liée à la prévision de la position future d'un objet.
- *Indexation des données :* compte tenu de la taille des bases de localisation et de la complexité d'évaluation des prédicats spatio-temporels, il est indispensable de définir des index ad-hoc permettant d'accélérer l'exécution des requêtes en minimisant le volume de données à extraire des disques. Un important travail reste à faire dans ce domaine, les index spatiaux traditionnels étant inopérants du fait de l'évolution continue des données indexées.

Maîtriser la gestion des données embarquées

La gestion de données embarquées dans des calculateurs spécialisés est désormais au cœur de multiples applications émergentes. Par exemple, les réseaux de capteurs utilisés pour collecter des informations environnementales (météorologie pollution, trafic routier) sont en train d'évoluer vers de véritables bases de données distribuées où chaque capteur devient un nœud *actif* du système; i.e., un micro-serveur de données interrogeable à distance. Les capacités locales de traitement permettent d'agréger, de trier ou filtrer les données et d'économiser ainsi du trafic réseau, particulièrement coûteux dans une infrastructure sans-fil [MFH03]. La gestion de données sécurisées sur cartes à puce (e.g., santé, assurance, téléphonie, etc.) est une autre motivation des traitements embarqués. Ces traitements peuvent être complexes et doivent être confinés dans la puce pour ne révéler aucune information confidentielle [PBV01]. Enfin, la gestion de données embarquées se justifie dans tous les contextes où des traitements doivent pouvoir être effectués en mode déconnecté. Ainsi, l'économie des coûts de communication, la confidentialité et le support des traitements déconnectés sont trois motivations importantes pour exécuter des traitements embarqués sur des calculateurs légers.

Concevoir des composants bases de données destinés à être embarqués sur des calculateurs mobiles ultra-légers (PDA, téléphones cellulaires, cartes à puce, capteurs, puces pour la domotique, l'automobile ou l'avionique) n'est cependant pas chose aisée. En effet, chaque calculateur est architecturé de sorte à satisfaire des propriétés précises (portabilité, autonomie électrique, sécurité, coût de production …) et ces architectures sont en permanente évolution pour couvrir les besoins de nouvelles applications. Il est donc tout aussi important de définir des moteurs bases de données adaptés à ces architectures spécialisées que d'exprimer des recommandations pour les concepteurs de ces architectures.

Jusqu'à présent, les éditeurs de SGBD ont abordé le problème de la gestion de données embarquées en proposant des versions allégées de leur SGBD (ex : Oracle 8i light, SQLServer pour Windows CE ou encore DB2 Everyplace). L'effort a porté sur l'empreinte mémoire et disque du SGBD ainsi que sur sa portabilité en termes de plates-formes matérielles et logicielles. Par contre, peu de travaux ont été entrepris pour adapter les techniques de gestion de données aux spécificités matérielles des calculateurs cibles. Ces techniques doivent pourtant être revues en profondeur sous l'influence de plusieurs paramètres matériels : la taille et la technologie des composants de mémoire stable (e.g., EEPROM, Flash, RAM alimentée et bientôt MEMS) qui déterminent les temps d'accès et d'écriture des données persistantes ; la taille de la mémoire de travail (RAM) utilisée comme mémoire tampon lors de la manipulation des données ; la puissance du processeur ; la bande passante et enfin l'autonomie électrique du calculateur. Les modèles de stockage de données, les stratégies d'évaluation de requêtes - locales à un calculateur ou distribuées entre plusieurs calculateurs - ainsi que les protocoles garantissant la cohérence et la tolérance aux pannes des données stockées sont tous fortement impactés par ces paramètres. La difficulté du problème vient du fait que ces paramètres ne peuvent être étudiés en isolation. Chaque type de calculateur exhibe un équilibre particulier de ressources qui en fait une équation unique à résoudre.

Les travaux académiques ont porté jusqu'à présent sur des architectures ponctuelles, notamment les réseaux de capteurs [BGS01, MFH03] et les cartes à puce [LeT97, ISO99, PBV01]. L'étude systématique de modèles de stockage et d'indexation, de stratégies d'évaluation de requêtes et de protocoles transactionnels pour des architectures fortement contraintes reste donc un challenge important pour les années à venir. Par ailleurs, des travaux de co-conception, mentionnés comme une des priorités du RNTL'03, pourraient permettre de spécifier le couple logiciel/matériel le mieux adapté au support d'applications embarquées ayant une forte composante de gestion de données.

Envisager de nouveaux modèles d'accès à l'information

Si le modèle client/serveur traditionnel est toujours possible en environnement mobile, d'autres modèles d'accès aux données peuvent être envisagés. L'objectif est de réduire la consommation de bande passante, de mieux supporter l'asynchronisme lié aux déconnexions, en acceptant éventuellement une dégradation de la latence.

Par exemple, la diffusion de données est particulièrement adaptée aux environnements mobiles. Un serveur diffuse périodiquement un ensemble de données à l'ensemble des nœuds du réseau, chacun filtrant ce flot en fonction de ses besoins propres. De nombreux travaux ont porté sur l'organisation du flot diffusé, son indexation et sa fréquence de diffusion [Aks98]. Le modèle *publication/souscription* offre une seconde alternative. Il s'agit de notifier automatiquement un client lorsqu'un événement satisfaisant un de ses abonnements est publié (émis) sur le réseau. Ce principe peut servir différents objectifs, et notamment le « rechargement de données » stockées sur un mobile. Par analogie au rechargement des batteries, il s'agit de rafraîchir les données présentes sur le mobile, sachant que plus la connexion sera longue, meilleur sera ce rafraîchissement. La charge optimale peut être déterminée grâce à un profil d'utilisateur définissant l'importance relative des données à recharger [CFZ01]. Enfin, des modèles basés sur une diffusion épidémique de l'information (chaque nœud recevant une information la retransmet de façon aléatoire à un certain nombre de ses voisins) peuvent être envisagés pour supporter le passage à l'échelle (très grand nombre de clients), augmenter la disponibilité des données (éviter les points uniques de défaillance) et s'adapter à un environnement mobile très dynamique (ex : réseaux ad-hoc). Si l'idée est intuitive, de nombreuses questions pratiques et théoriques restent posées par ce modèle [EuG02].

Quel que soit le modèle d'accès aux données, la résolution (ou qualité) des données a un impact considérable sur la consommation de bande passante. Cette résolution peut être représentée selon plusieurs dimensions : précision du contenu (échelle pour une carte, fidélité pour une séquence audio ou vidéo, résumé, texte seul ou avec graphiques pour un document) ; cohérence temporelle (degré de « fraîcheur » des données par rapport aux données réelles) ; cohérence sémantique (degré de respect des contraintes d'intégrité). Ces dimensions doivent pouvoir être combinées dans une stratégie d'adaptation propre à une application, à un utilisateur et au terminal utilisé.

Assurer la synchronisation et la cohérence des données

La gestion des traitements en mode déconnecté est intrinsèque à la mobilité, du moins tant qu'une connexion permanente ne peut être assurée (pour des raisons de couverture hertzienne, de coût financier ou encore de consommation électrique pour des équipements à faible autonomie). Les données nécessaires au traitement sont dupliquées sur le mobile, modifiées localement, puis renvoyées sur le réseau filaire à la prochaine connexion. Si le problème de cohérence est simple à résoudre en présence d'un seul utilisateur, il se complique fortement dès que des mises à jour concurrentes sont effectuées sur les mêmes données (e.g., équipe virtuelle). La présence de multiples écrivains impose de gérer la divergence entre les copies. Pour ce faire, les synchroniseurs de fichiers (Microsoft's Briefcase, Power Merge ...) propagent les mises à jour non conflictuelles et délèguent la résolution des conflits (écritures concurrentes sur le même fichier) aux utilisateurs. Les synchroniseurs de données (HotSync, ActiveSync ...) analysent plus finement le contenu des fichiers et

tiennent compte de la sémantique des opérations de mise à jour pour diminuer la probabilité de conflit (par exemple l'insertion simultanée de deux entrées différentes dans un même annuaire est non conflictuelle). Les systèmes de gestion de configuration (Diff3, rcsMerge, XMLDiffMerge) réalisent la fusion de versions de fichiers selon une approche encore différente. Cette prolifération d'outils et d'approches rend indispensable un travail de formalisation destiné à mieux caractériser les conflits et à définir précisément la correction d'une synchronisation.

Une piste intéressante a été ouverte dans le domaine du traitement collaboratif temps-réel [SuE98, VCF00]. Quand une même opération doit être exécutée sur différentes copies d'un même objet et que ces copies divergent du fait de mises à jour locales, le principe proposé consiste à réécrire cette opération par rapport à l'histoire locale de chacune de ces copies afin que le résultat soit équivalent partout. Ce principe, connu sous le nom de *transformées opérationnelles*, garantit trois propriétés fondamentales : (1) préservation de la causalité, (2) préservation de l'intention et (3) convergence des copies. Ce principe a été récemment adapté aux traitements asynchrones en mode déconnecté [IMO03].

Certaines applications ont un besoin de cohérence plus fort que celui assuré par les trois propriétés précédentes et imposent le respect des propriétés transactionnelles ACID (Atomicité, Cohérence, Isolation, Durabilité) [BHG87]. La mobilité a cependant un impact majeur sur la mise en œuvre de ces propriétés, principalement du fait des déconnexions et de leur caractère non borné dans le temps. Ainsi, les protocoles de sérialisation traditionnels à base de verrouillage sont inapplicables (impossibilité de borner la détention d'un verrou), au même titre que les protocoles de validation atomique en deux phases (toute déconnexion conduisant à un abandon de transaction). Les recherches sur ce thème ont été particulièrement actives au cours de la dernière décennie, si bien qu'il est impossible de vouloir les résumer ici. Le lecteur intéressé pourra se référer utilement à [PiB99, SRA01]. Dans la majorité des travaux existants, les propriétés ACID ont été étudiées séparément les unes des autres ou bien ont été étudiées dans un contexte applicatif précis. Il reste donc un champ d'investigation important autour de la définition de modèles de transactions mobiles généraux et fédérateurs.

3. Confidentialité des données

3.1. *Position du problème*

En rendant l'information plus facilement accessible et en multipliant les moyens de son acquisition, informatique ubiquitaire et intelligence ambiante induisent de nouveaux comportements mais également de nouveaux risques au regard de la confidentialité. Pour s'en persuader, il suffit de considérer la gigantesque base d'informations qu'il est possible de construire sur les individus en croisant les données historiques accumulées par l'ensemble des objets constituant un environnement intelligent (habitudes alimentaires, état de santé, contenu des tiroirs,

heures d'arrivée à la maison, déplacements, sites web visités …). L'individu n'a plus conscience de la façon dont les données sont acquises, diffusées puis traitées et même les informations les plus anodines peuvent être sujettes à interprétation (ex : une mauvaise alimentation est un facteur de risque pour une compagnie d'assurance). Ce constat est confirmé par une étude de IBM-Harris selon laquelle 94% des citoyens américains considèrent qu'ils ont perdu tout contrôle sur l'utilisation des informations les concernant [IBM00]. Par ailleurs, d'autres phénomènes menacent la privacité tels que les célèbres systèmes ECHELON (NSA) et CARNIVORE (FBI), ou encore CAPPS-II (Computer Assisted Passenger Pre-Screening System) permettant aux compagnies aériennes américaines d'effectuer des croisements de bases de données commerciales et gouvernementales sous couvert de protection contre le terrorisme. Comme le souligne Ron Rivest, la révolution digitale inverse les défauts : "Ce qui était autrefois difficile à copier devient facile à dupliquer, ce qui était oublié devient mémorisé à jamais et ce qui était privé devient public".

Devant cet état de fait, il convient de se poser la question de la définition de la privacité du point de vue légal. Alan Westin, un des grands pionniers des travaux sur la privacité en informatique, donnait en 1967 la définition suivante : "le désir des gens de choisir en toute liberté dans quelles circonstances et jusqu'à quel point ils souhaitent dévoiler des informations sur eux-mêmes, leurs habitudes et leur comportement". Les points de vue pour assurer ce principe divergent selon les états. Alors que les Etats-Unis privilégient l'autorégulation en comptant sur l'effet de réciprocité ("watching the watchers"), l'Europe se dote d'un arsenal législatif plus contraignant. Ainsi, la directive européenne du 24 octobre 1995 relative à la protection des données personnelles précise les données dont la collecte est strictement interdite (par exemple opinions politiques, philosophiques ou religieuses) et fixe les obligations incombant aux collecteurs de données et les droits des personnes concernées (accès, opposition, rectification). La France est également, et depuis longtemps, active dans le domaine de la protection de la vie privée avec la loi du 6 janvier 1978 et l'établissement de la CNIL (Commission Nationale de l'Informatique et des Libertés).

Le problème n'est cependant pas simplement législatif. Encore faut-il être capable de mettre au point les outils permettant de faire respecter les règles édictées. Dans le volet IST (Information Society Technologies) du FP6 (6^{th} Framework Programme), l'Union Européenne considère comme un enjeu social et économique majeur le développement des technologies assurant la privacité et la protection de la propriété et des droits individuels. La suite de cette section présente quelques verrous scientifiques et technologiques majeurs relatifs à ce problème du point de vue des bases de données.

3.2. *Verrous scientifiques et technologiques*

La protection de la confidentialité des données nécessite de sécuriser l'intégralité de la chaîne reliant un utilisateur aux données auxquelles il a le droit d'accéder sur

disque. Les menaces sont multiples : usurpation d'identité, attaque des communications, détournement de droits, attaque sur l'empreinte disque de la base de données ou encore détournement d'informations détenues légalement.

L'usurpation d'identité et la sécurisation des communications ne sont pas des problèmes spécifiques aux bases de données et de nombreux protocoles sont aujourd'hui éprouvés pour y répondre. Ainsi, tout utilisateur doit être identifié (il doit dire qui il est) et authentifié (il doit le prouver). L'authentification peut se faire par simple mot de passe ou par des techniques matérielles (cartes à puce, token card …) et biométriques [Ora99]. La sécurisation des communications fait quant à elle appel à des techniques de chiffrement garantissant la confidentialité du contenu échangé, à des techniques de hachage assurant l'intégrité du message et à des techniques de signature assurant la non-répudiation des messages [Sch96]. La suite de cette section se concentre sur les menaces plus spécifiques aux bases de données.

Mieux définir et contrôler les droits d'accès

La gestion des droits d'accès dans les SGBD suit traditionnellement un des trois modèles DAC, MAC ou RBAC [BPS96]. Dans le modèle DAC (Discretionary Access Control), le créateur d'un objet se voit affecter tous les droits sur cet objet et peut transmettre tout ou partie de ses droits à d'autres utilisateurs, et ce de façon récursive. Le modèle RBAC (Role-Based Access Control) affecte des droits à des rôles et un utilisateur peut être autorisé à jouer des rôles différents au cours de sessions différentes. Enfin, le modèle MAC (Mandatory Access Control) fixe des niveaux hiérarchiques de sécurité aux données (public, confidentiel, secret …) et des niveaux d'habilitation aux utilisateurs.

Si ces modèles sont bien établis, leur mise en œuvre s'accorde mal avec la complexité d'organisations largement distribuées. En effet, ces modèles supposent généralement une administration centralisée des droits. Ceci pose la question de la confiance nécessaire dans le personnel et les applications responsables de cette administration ainsi que la question de la cohérence et de la dynamicité de cette administration. Un travail préliminaire, mené autour d'un modèle décentralisé intitulé ORBAC (ORganization Based Access Control) [KBB03], traite de ce dernier point. Il est clair qu'un important travail de recherche doit être mené pour faire évoluer les modèles de droits et/ou en définir de nouveaux afin de mieux prendre en compte les environnements fortement décentralisés qui se développent via l'internet (ex : systèmes pair à pair).

Par ailleurs, les modèles actuels ne permettent pas d'intégrer la diversité croissante des informations à protéger ni des moyens d'accès à cette information. Par exemple, si la sémantique des politiques de contrôle d'accès est bien établie pour des données tabulaires et structurées, elle l'est beaucoup moins pour des données arborescentes et semi-structurées telles que des documents XML, même si des travaux ont été initiés dans ce domaine [BCF00, GaB01]. La définition de droits d'accès portant sur le contenu de données multimédia (images, vidéo) est vierge alors

que les besoins apparaissent (ex : anonymisation d'une séquence de vidéo-surveillance). D'autre part, les modèles actuels sont adaptés à des interactions de type client-serveur ou *pull*. Aujourd'hui se développent de plus en plus d'interactions de type diffusion ou *push*. Lorsqu'une même information est diffusée à destination d'utilisateurs ayant des droits différents, le contrôle de droits se trouve relégué au niveau du client. Il faut de ce fait trouver un moyen de sécuriser ce contrôle, par des techniques de chiffrement et/ou par l'utilisation de composants matériels sécurisés. Les interactions peuvent également être de type pair à pair et dynamiques (équipe virtuelle ouverte à de nouveaux membres), introduisant une nouvelle complexité dans l'attribution des droits. Enfin, du fait de la puissance des outils d'extraction de connaissance à partir des données (ECD), il devient de plus en plus difficile de se prémunir contre des tentatives d'inférence d'informations prohibées par recoupement d'informations autorisées. Ces nouvelles dimensions du problème ouvrent un champ d'investigation considérable.

Interroger et partager des données chiffrées

Selon le FBI, les attaques sur les bases de données sont de plus en plus fréquentes et coûteuses (estimées à plus de 100 milliards de dollars/an aux USA) et sont internes dans près de la moitié des cas [FBI02]. De ce fait, comment protéger une base de données des attaques dirigées par un intrus sur l'empreinte disque de la base de données ? De même, comment faire confiance à un administrateur de bases de données (DBA) ayant tous les droits pour administrer des données confidentielles ? Dans les deux cas, les droits d'accès contrôlés par le SGBD sont inopérants. Ce problème se pose avec d'autant plus d'acuité du fait d'un recours croissant à des hébergeurs de données (Database Service Providers, Web-hosting companies) par des individus ou des PME désireuses d'externaliser la gestion de leur système d'information.

Diverses techniques de chiffrement ont été proposées pour se prémunir des attaques précédentes. Par exemple, Oracle propose un mécanisme permettant de chiffrer les données stockées sur disque, celles-ci étant déchiffrées dynamiquement par le serveur au moment de l'évaluation des requêtes [Ora99]. Ce mécanisme est efficace contre un intrus attaquant l'empreinte disque de la base de données. Par contre, il est inopérant contre un DBA car celui-ci a le pouvoir de l'altérer et d'intercepter les clés ou les données en clair lors du déchiffrement. Une première solution à ce problème consiste à modifier le noyau du SGBD pour interdire certaines actions du DBA [HeW01]. Une seconde solution est de distinguer les fonctions d'administration de données des fonctions d'administration de la sécurité et d'isoler ces dernières dans un serveur séparé [Mat00]. Ces différentes solutions souffrent toutes du fait que les données sont déchiffrées sur le serveur et permettent diverses attaques du DBA. L'alternative est de réaliser le déchiffrement des données sur le client. Ceci implique de décomposer toute requête en une sous-partie évaluable sur le serveur directement sur les données chiffrées et une seconde partie évaluable uniquement sur le client après déchiffrement (e.g., prédicats d'inégalité, agrégation

…) [HIL02]. Le problème se complique fortement dès que les données sont partagées par plusieurs clients ayant des droits d'accès différents [BoP02].

Un important travail d'investigation reste à mener autour du chiffrement des bases de données et de leur exploitation (interrogation, mises à jour, partage). Compte tenu du volume de données chiffrées et de la persistance des données, il est nécessaire de définir des algorithmes de chiffrement résistant aux attaques statistiques. Il est par ailleurs intéressant d'étudier des algorithmes de chiffrement préservant certaines propriétés arithmétiques sur les données chiffrées afin d'élargir la portée des requêtes traitables par un serveur directement sur les données chiffrées. Enfin, le partage de données chiffrées reste un problème majeur. Lorsqu'un droit ne s'exprime que par la détention d'une clé de déchiffrement, la gestion des droits devient statique (le retrait d'un droit est problématique) et de granularité grossière (il est possible de limiter l'accès à un objet physique mais pas à un objet dérivé d'un calcul). La puissance d'expression des droits autorisée par les SGBD s'en trouve fortement altérée. Des solutions préliminaires ont été proposées en utilisant des composants matériels sécurisés sur le client et demandent à être étendues et validées.

Contrôler l'utilisation d'informations détenues légalement

Tout prestataire de services Internet détient légalement des informations sur ses clients (identité, adresse, informations bancaires, statistiques sur l'utilisation du service offert) et est tenu de respecter une charte de confidentialité relative à l'utilisation de ces informations. Cependant, de nombreuses violations de chartes de confidentialité sont régulièrement et depuis longtemps relevées [Ols99].

Le concept de *SGBD Hippocratique*, à savoir de SGBD donnant l'assurance du respect d'un serment de confidentialité, a été introduit dans [AKS02]. Un tel SGBD se doit de respecter un ensemble de principes fondateurs parmi lesquels : préciser l'objectif d'utilisation de chaque donnée collectée sur un utilisateur et recueillir l'assentiment de l'utilisateur sur cet objectif, ne stocker que l'information strictement nécessaire à l'atteinte de cet objectif et uniquement pendant le laps de temps strictement nécessaire, ne pas divulguer cette information à des tiers sans autorisation préalable de l'utilisateur, donner à l'utilisateur la possibilité de consulter les informations qui le concernent et enfin offrir des outils permettant à un tiers de contrôler que ces principes sont bien respectés. Le standard P3P (Platform for Privacy Preferences) promu par le W3C [Mar02] est également un moyen donné aux utilisateurs pour mieux contrôler les informations que les sites Web accumulent sur eux.

Si les préceptes d'un SGBD Hippocratique ont été énoncés, il reste à mettre en œuvre de tels systèmes. Ceci nécessite la définition de nouveaux noyaux de SGBD capables d'intégrer ces préceptes et offrant des outils d'audits non falsifiables.

4. Conclusion

Sans nul doute, informatique ubiquitaire et intelligence ambiante vont profondément modifier l'organisation et les usages des systèmes d'information dans la prochaine décennie. Aux Etats-Unis, le DARPA développe une vision tendant à faire disparaître la frontière entre le monde réel (objet physique) et sa représentation abstraite (objet binaire accessible et actionnable). L'Union Européenne partage cette vision et l'inscrit dans la liste de ses objectifs stratégiques, en tenant compte également des forces de l'industrie européenne notamment en termes de communications mobiles, de micro-électronique et de systèmes embarqués. De son côté, la communauté bases de données française s'est mobilisée depuis quelques années sur l'étude de la mobilité et est bien en phase avec les objectifs poursuivis outre-Atlantique, comme l'attestent deux rapports prospectifs [JoG01, ASM03].

Un des corollaires du développement de l'intelligence ambiante est la protection de la confidentialité des données. Cet objectif s'est imposé en quelques années comme un challenge social et économique majeur et commence à mobiliser des efforts considérables au niveau international. Au-delà des verrous scientifiques et technologiques identifiés dans ce chapitre, l'absence de confiance dans les serveurs qui gèrent les données génèrent de nouveaux problèmes connexes. Par exemple, comment certifier l'authenticité des résultats émis par un serveur à une requête ? Comment vérifier la complétude de ces mêmes résultats ? Comment garantir que l'information contenue dans un serveur n'y est pas entrée en violation des droits d'auteurs ou de la propriété intellectuelle ? Comment faire en sorte qu'un serveur ne puisse pas garder trace de chaque recherche effectuée par un individu ? Le domaine d'investigation est ici très large. On peut imaginer que la communauté bases de données française, encore peu mobilisée sur ces thèmes, saura s'appuyer sur d'autres communautés partageant un vif intérêt pour les problèmes de sécurité (notamment en cryptologie, réseaux, systèmes distribués) et sur des acteurs industriels européens occupant des positions dominantes dans quelques secteurs liés à la sécurité, pour développer une recherche de niveau international sur ce sujet passionnant.

5. Références

[AHI03] The Aware Home Research Initiative – Georgia Institute of Technology - http://www.cc.gatech.edu/fce/ahri/

[Aks98] Aksoy D. et al., “Research in Data Broadcast and Data Dissemination”, Proc. Int. Conf. On Advanced Multimedia Processing, Osaka, 1998.

[AKS02] Agrawal R., Kiernan J., Srikant R., Xu Y., “Hippocratic Databases”, Int. Conf. on Very Large Data Bases (VLDB), 2002.

[ASM03] AS CNRS Mobilité/Accès aux données, "Mobilité et bases de données : état de l’art et perspectives", Chronique *Technique et Science Informatiques (TSI)*, Vol. 22, n°3 et n°4, 2003. Version anglaise "Mobile Databases: a Report on Open Issues and Research Directions", ACM Sigmod Record, à paraître.

[BCF00] Bertino E., Castano S., Ferrari E., Mesiti M.,”Specifying and Enforcing Access Control Policies for XML Document Sources”, WWW journal vol 3 number 3. Baltzer Science Publisher, 2000.

[BGS01] Bonnet P., Gehrke J., Seshadri P., “Towards Sensor Database Systems”. Mobile Data Management 2001: 3-14.

[BHG87] Bernstein P. A., Hadzilacos V., Goodman N.,”Concurrency Control and Recovery in Database Systems”, book, Addison Wesley, 1987.

[BoP02] Bouganim L., Pucheral P., “Chip-Secured Data Access: Confidential Data on Untrusted Servers”, Int. Conf. on Very Large Data Bases (VLDB), 2002.

[BPS96] Baraani A., Pieprzyk J., Safavi-Naini R., “Security In Databases: A Survey Study”, TR-96-02, Department of Computer Science, The University of Wollongong, Wollongong, Australia, 1996.

[CFZ01] Cherniack M., Franklin M.J., Zdonik S., “Expressing User Profiles for Data Recharging”, IEEE Personal Communications, August 2001.

[EuG02] Eugster P, Guerraoui R., “Probabilistic Multicast”, IEEE Int. Conf. on Dependable Systems and Networks (DSN), 2002.

[FBI02] Computer Security Institute, "CSI/FBI Computer Crime and Security Survey", 2002. http://www.gocsi.com/forms/fbi/pdf.html

[GBE00] Guting R., Bohlen P., Erwig M., Jensen C.S., Lorentzos N., Schneider M., Vazirgiannis M., “A Foundation for Representing and Querying Moving Objects”, ACM Transactions on Database Systems (TODS), 2000.

[GaB01] Gabillon A., Bruno E., “Regulating Access to XML documents”, Fifteenth Annual IFIP Working Conference on Database Security, 2001.

[HeW01] He J., Wang M., “Cryptography and Relational Database Management Systems”, Int. Database Engineering and Application Symposium, 2001.

[HIL02] Hacigumus H., Iyer B.R., Li C., Mehrotra S., “Executing SQL over Encrypted Data in the Database-Service-Provider Model”, ACM Int. Conf. On Management of Data (SIGMOD), 2002.

[IBM00] IBM, Harris, “The IBM-Harris Multi-National Consumer Privacy Survey”, Privacy & American Business, Vol. 7, No. 6, 2000.

[ImN02] Imielinski T., Nath B., “Wireless Graffiti – Data, data everywhere”, Int. Conf. on Very Large Data Bases (VLDB), 2002.

[IMO03] Imine A., Molli P., Oster G., and Rusinowitch M., "Proving Correctness of Transformation Functions in Real-time Groupware", European Conf. on Computer-Supported Cooperative Work, 2003.

[ISO99] Int. Standardization Organization (ISO), "Integrated Circuit(s) Cards with Contacts – Part 7: Interindustry Commands for Structured Card Query Language-SCQL" , ISO/IEC 7816-7, 1999.

[JoG01] Joshi A., Goldin D., "Report of the Group Discussion on Pervasive Computing", http://itlab.uta.edu/idm01/FinalReports/pervasiveReport.pdf, 2001.

[KBB03] Kalam A., Baida R., Balbiani P., Benferhat S., Cuppens F., Deswarte Y., Miège A., Saurel C., Trouessin G., "Organization Based Access Control", Policy'2003, 2003.

[LeT97] Lecomte S., Trane P., "Failure Recovery Using Action Log for Smartcards Transaction Based System", *IEEE On Line Testing Workshop*, 1997.

[MFH03] Madden S., Franklin M., Hellerstein J., Hong W., "The Design of an Acquisitional Query Processor for Sensor Networks", ACM Int. Conf. On Management of Data (SIGMOD), 2003.

[Mar02] Marchiori M., editor, "The Platform for Privacy Preference 1.0 (P3P1.0) Specification", W3C proposed recommendation, 2002.

[Mat00] Mattsson U., "Secure.Data Functional Overview", Protegity Technical Paper TWP-0011, 2000. (http://www.protegrity.com/White_Papers.html)

[Ols99] Olsen S., "Top Web Sites Compromise Consumer Privacy", CNET News Archive, 1999.

[Ora99] Oracle Corp., "Database Security in Oracle8i", Oracle Documentation, 1999.

[PiB99] Pitoura E., Bhargava B., "Data Consistency in Intermittently Connected Distributed Systems", Transactions on Knowledge and Data Engineering (TKDE), 11(6), 1999.

[PBV01] Pucheral P., Bouganim L., Valduriez P., Bobineau C., "PicoDBMS: Scaling down Database Techniques for the Smartcard", Very Large Data Bases Journal (VLDBJ), 10(2-3), 2001.

[Sch96] Schneier B., "Applied Cryptography", book, 2nd Edition, John Wiley & Sons, 1996.

[SRA01] Serrano-Alvarado P., Roncancio C., Adiba M., "Analyzing Mobile Transactions Support for DBMS", Int. Workshop on Mobility in Databases and Distributed Systems (in DEXA), 2001.

[SuE98] Sun C., Ellis C.S., "Operational Transformation in Real-Time Group Editors", ACM Int. Conf. on Computer Supported Cooperative Work (CSCW), 1998.

[VCF00] Vidot N., Cart M., Ferrié J., Suleiman M., "Copies convergence in a distributed real-time collaborative environment", ACM Int. Conf. on Computer Supported Cooperative Work (CSCW), 2000.

[WXC98] Wolfson O., Xu B, Chamberlain S., Jiang L., "Moving Objects Databases: Issues and Solutions", Int. Conf. on Scientific and Statistical Database Management, 1998.