De la gestion de bases de données à la gestion de grands espaces de données

juillet 2012

Comité Bases de Données Avancées
Contact : **Christine.Collet@grenoble-inp.fr**

Ce document présente le résultat d'une réflexion des chercheurs en bases de données sur l'état des recherches dans ce domaine en France. Sans prétendre à l'exhaustivité, le comité Bases de Données Avancées souhaitait identifier les enjeux scientifiques et sociétaux de la gestion des données pour la prochaine décennie, et d'en préciser les contours à la lumière des évolutions scientifiques et des nouveaux usages qu'elles génèrent.
Ont plus particulièrement participé à la rédaction de ce document : B. Amnan, N. Bidoit, M. Boughanem, M. Bouzeghoub, Ch. Collet, A. Doucet, D. Gross-Amblard, J-M. Petit, M. Said Hacid et G. Vargas-Solar.

Le succès de l'économie numérique, la généralisation des équipements mobiles et le développement des applications associées (loisirs, savoir numérique, etc.) sont autant d'exemples incitant à poursuivre les efforts vers des améliorations significatives dans les domaines des sciences (observatoire virtuel en astronomie, cartographie du génome, physique et énergie), de la santé et du bien-être (aide aux personnes âgées et à l'handicap, chirurgie assistée) et des humanités (sciences sociales). Ces efforts concernent également le développement de nouveaux services pour la ville, le transport et la durabilité de l'environnement en général (efficacité énergétique, climat, pollution, alimentation). A cela s'ajoute l'avènement de politiques de type *open data* visant à rendre libres, accessibles et réutilisables des données publiques. L'émergence des réseaux sociaux, l'intelligence ambiante, le Web ou l'intelligence économique ont profondément remodelé notre rapport aux sciences de l'information et de la communication et à leurs applications.

Dans ce contexte bouillonnant, les données sont élaborées, commercialisées et consommées comme n'importe quels produits manufacturés. Elles sont au cœur de processus informatiques, souvent complexes, mobilisant un grand nombre de ressources (sources de données, applications, services) et nécessitant des procédés rigoureux. Les données sont au cœur de l'économie numérique et de la société du savoir. Elles représentent une matière première à forte valeur ajoutée, sans lesquelles rien ne pourraient se passer.

En effet, la production de données par les utilisateurs (''crowdsourcing''), et notamment le partage d'informations ambiantes au travers des objets ubiquitaires (capteurs et senseurs mobiles, caméras, microphones, appareils photos, lecteurs RFID, réseaux de capteurs sans fil, etc.) augmentent drastiquement les volumes de données à traiter. On parle de déluge de données qui n'est pas sans conséquence sur leur manipulation. Les données acquises, de façon massive ou à la volée, doivent être homogénéisées, enrichies, croisées, filtrées et agrégées pour constituer *in fine* des produits informationnels riches en sémantique et stratégiques pour l'analyse et la décision. En dehors des types de manipulation sur les déluges de données, se pose la question de leur entreposage, leur pérennisation pour les générations futures et aussi sur le droit à l'oubli ou à l'éternité numérique.

De nombreux problèmes autour de la gestion des données se posent donc, relevant aussi bien de la recherche applicative que de la recherche fondamentale. Les techniques de gestion de ces données seront issues d'une refondation profonde des architectures des réseaux, de la logique des bases de données, de l'algorithmique, voire même des règles d'interprétation de ces données.

Pour l'heure, les systèmes classiques de gestion de données (SGBD) constituent encore le socle fondamental du stockage, du traitement et de l'accès aux données. Si les SGBD restent indispensables à un grand nombre de centres de production de données, leurs fonctionnalités doivent

être revisitées pour les adapter au contexte actuel de distribution à grande échelle, d'autonomie des sources de données et d'hétérogénéité des types de données. Dans cette perspective, les fonctionnalités traditionnelles des SGBD ne constituent plus une offre de systèmes intégrés mais un ensemble de services disséminés et adaptés à chaque usage et à chaque contexte. Les nouveaux efforts de recherche doivent également tenir compte de deux dimensions : des formats d'objets multiples (images, son, vidéos) et une multiplicité des sources de données (illustrée par le Web, les capteurs, compteurs intelligents, satellites).

Ainsi, les actions de recherche s'orientent vers la définition et l'orchestration de services de données tenant compte de ces deux dimensions dans des architectures ouvertes, multi-échelles et multi-niveaux appelées Espace de données (''*DataSpaces*''). Parmi ces architectures, les entrepôts de données, les systèmes de médiation, les systèmes Pair-à-Pair (P2P) ont apporté des solutions originales et innovantes. Plus récemment, d'autres architectures, comme les grilles informatiques et l'informatique en nuages (''*Cloud Computing*''), concentrent un effort d'investigation important. Les grilles informatiques apparaissent comme des applications à la croisée de la maîtrise de plusieurs technologies dont notamment la virtualisation des réseaux, les technologies du Web sémantique ou encore les composants logiciels.

Le bénéfice du Cloud réside dans un stockage massif de données à un coût proportionnel aux besoins, dans l'élasticité de l'offre qui supporte des charges lourdes ou soudaines (sans acquisition d'infrastructures), et dans l'accès ubiquitaire. Toutefois, ces architectures introduisent aussi de nouveaux risques à considérer ou défis à relever (confidentialité, cohérence, pérennisation et préservation des données) qui, s'ajoutant aux performances, deviennent de nouveaux défis scientifiques à résoudre.

Parmi les services de gestion de données que ces architectures doivent offrir, le stockage et l'interrogation de masses de données constitueront l'offre de base, enrichie par des mécanismes adéquats d'indexation et de réplication. De nouveaux modèles de stockage, par exemple orientés colonnes ou fragments (à l'image de ''*BigTable*''), ouvrent des perspectives intéressantes de passage à l'échelle. Parallèlement, le traitement de requêtes bénéficie de nouvelles techniques de parallélisme (exemple de ''*MapReduce*''), permettant l'interrogation efficace de ces masses de données. Au delà du stockage et de l'accès, d'autres services, comme la confidentialité, la cohérence (relâchée) et la sécurité des données, restent un enjeu majeur dans ce domaine.

Ces services de gestion de données sont à replacer au cœur des systèmes d'information et de nouvelles architectures qui mettent en oeuvre des chaînes de traitement complexes : capture et saisie des données, stockage, production, structuration de données, extraction, intégration, analyse, et restitution des données. Les systèmes décisionnels et les *workflows* scientifiques sont des exemples de telles applications. La conception et la réalisation de ces systèmes d'information ne font plus appel aux seules techniques de bases de données mais aussi aux techniques de la recherche d'information, de la fouille de données et de l'apprentissage, aux méthodes statistiques, aux connaissances linguistiques, au raisonnement et à la visualisation. Il faut admettre que les frontières entre ces différents domaines ont tendance à s'estomper, préfigurant une large communauté d'intérêt centrée sur la gestion et l'exploitation intelligente des contenus et de leurs usages (au travers de services numériques). Pour autant, l'expérience acquise depuis bientôt 40 ans par la communauté bases de données est prépondérante pour relever les défis que pose la gestion des données de ce début de 21° siècle. Il est important de noter que ce domaine est essentiellement « boosté » par de grandes multinationales de la gestion de données (e.g. Oracle, IBM, Microsoft, SAP) et de petites entreprises innovantes sur des niches applicatives (e.g. Google il y a quelques années).

La suite de ce document donne une vision globale des principaux enjeux scientifiques de la recherche en bases de données selon quatre grands axes : la représentation des données (section 1), l'architecture des systèmes de gestion de données (section 2), l'accès aux données (section 3), et la sécurité et confidentialité des données et services (section 4). Il précise également le positionnement de ces recherches dans les grands enjeux sociétaux (section 5).

1- Représentation des données

Les données peuvent être de différents types : images, données temporelles, données spatiales, données de trajectoires, etc. Ce sont des données relatives au cœur de métier, à l'état des dispositifs d'acquisition des données, à leur localisation, leur utilisation, etc. Dans de nombreuses applications, il est important de croiser et d'utiliser conjointement plusieurs types de données. A titre d'exemple, les données LSST (*Large Synoptic Survey Telescope*) comportent plus de 9 types d'images différents et nécessitent plus de 75 tables de métadonnées, catalogues et alertes, avec des relations croisées entre elles.

L'exploitation des données elles-mêmes (leur contenu) et de tous les éléments de leur contexte (métadonnées, profil de l'utilisateur, localisation géographique, retour sur expérience) repose sur de multiples représentations des données et sur les méthodes de transformation et d'interprétation associées. La représentation des données est fortement dépendante :

- des données elles-mêmes, de part leur hétérogénéité (contenus, structure, propriétés), leur qualité (données bruitées, manquantes, erronées, simulées, évolutives, disponibles), leur volumétrie, leur mode de production (flux émis par des capteurs, données générées par des utilisateurs, etc.) ;
- de leur exploitation tout au long d'une chaîne de traitement qui va de leur capture jusqu'à leur visualisation, en passant par l'extraction d'information ou connaissance, leur interrogation, leur synthèse, leur agrégation, …leur stockage et pérennisation ;
- des infrastructures dans lesquelles elles sont déployées et des contraintes relatives à leur traitement (sécurité, confidentialité, vie privée, énergie, environnement).

Ceci conduit à une refondation des schémas de représentation des données et des traitements à différents niveaux d'abstraction :

1- au niveau physique :

   - le stockage de l'information numérique passe par des améliorations de modèles de représentation, de techniques de compression permettant de réduire la taille des données à préserver et des techniques d'indexation.

   - la mémoire constitue aujourd'hui encore un goulet d'étranglement pour les architectures. Les gains en performance sur la vitesse intrinsèque des composants sont relativement faibles. Les fabricants de mémoires jouent donc sur l'organisation logique et l'optimisation des accès.

2- au niveau logique :

   - la démarche d'acquisition et de traitement de données (en particulier filtrage et analyse) requiert une forte interaction entre les informations à capturer et leur environnement extérieur (contexte de capture ou d'usage des données, utilisateurs, autres données telles que localisation, heure de capture, etc.) et concerne plusieurs niveaux. En conséquence la représentation des données inclut des métadonnées qui donnent tout leur potentiel aux données.

   - l'interprétation temps-réel de données (à partir d'images ou de vidéos par exemple) et l'indexation automatique de contenus, particulièrement de contenus multimédias. Les représentations ou interprétations temps-réel des données sont utilisées par des algorithmes exécutables dans des conditions de ressources contraintes (puissance de calcul limitée, économie d'énergie, etc.).

   - l'agrégation de contenus ou plus généralement l'intégration des informations qui nécessitent une représentation partagée et des outils permettant de réduire le fossé sémantique entre les données, de découvrir et exploiter le sens dont elles sont porteuses. De nombreux modèles ont pu être développés ces dernières années et un effort en matière de normalisation a été fait avec le langage XML. Mais ce langage n'est pas suffisant pour permettre de représenter le « sens » des données. Des « outils » du web sémantique tels que RDF, *OWL, DAML+OIL, etc.*, ont été proposés.

3- au niveau externe :

Il s'agit d'inventer de nouvelles façons de restituer les données, fondées sur la multi-représentation des données permettant de les adapter ou de les composer à la volée en fonction de la capacité d'affichage des supports (dans le cas d'applications mobiles), des droits d'accès, de la demande des utilisateurs ou de son contexte d'usage. La visualisation des données est aujourd'hui un axe de recherche majeur.

Cette vision classique dans la représentation des données selon trois niveaux est à revisiter compte tenu de la demande grandissante d'une gestion unifiée des informations (du disque à l'objet métier) au travers d'un référentiel unique comme dans les systèmes *NoSQL* émergeants. Ce référentiel est important pour améliorer en outre l'efficacité de l'organisation, de la production et de la gestion globale de la connaissance. En effet il permet de s'affranchir de tous les modèles nécessaires à la compréhension d'une information. Les axes de recherche peuvent concerner :

- le développement de modèles et de langages spécifiques à des domaines d'application, à des métiers, à des services, etc. ;
- le développement d'outils logiciels pour assister l'utilisateur dans la description des données et des domaines, en particulier pour les documents multimédias ;
- la représentation linguistique et sémantique des (modèles de) données, exploitant des approches pluridisciplinaires (sciences cognitives, psychologie, etc.);
- les techniques d'exploration des grandes masses de données, avec des règles de guidage intuitives et contextualisées ;
- les techniques et outils permettant de raisonner sur les modèles, pour assurer leur cohérence et en comprendre leur sémantique.

2- Architecture des systèmes de gestion de données

L'architecture ANSI/SPARC qui a caractérisé les systèmes de gestion de bases de données classiques (relationnels, objets, XML) a évolué de manière concomitante aux besoins applicatifs, aux volumes de données à gérer ainsi qu'aux évolutions des modèles de données. Ces évolutions ont abouti à une réflexion sur la manière dont les fonctions de gestion de données sont offertes aux applications : sous forme de composants/services à la fin des années 90, de réseaux P2P au début des années 2000 et de services bases de données ces toutes dernières années.

Les défis des prochaines infrastructures de gestion de données sont de faciliter l'exécution des algorithmes et des processus complexes utilisant de grosses masses de données (multimédia, graphes complexes avec des millions de nœuds). Les infrastructures doivent mettre à disposition des ressources de calcul et faciliter leur utilisation. Par conséquent, aujourd'hui, on constate que l'architecture d'un système de gestion de bases de données a évolué vers la notion « d'infrastructure à base de services ». Les applications adaptent et coordonnent les services pour effectuer la gestion, l'interrogation et l'exploitation efficace des données (analyse, prise de décisions, fouille). Les services sont déployés sur des architectures système/matériel comme la grille, les réseaux P2P et de capteurs, les systèmes embarqués, le *Web2.0*, et le *Cloud*.

Les grands thèmes de recherche sur les infrastructures de gestion de données sont corrélés aux :

- architectures (abstraites) pour la gestion de données sur les réseaux de capteurs, les dispositifs hétérogènes et les systèmes embarqués. Il s'agit de développer et de déployer des services pour la découverte et la localisation de ressources, l'accès élastique et transparent à de gros volumes de données réparties et dupliquées sur la planète, l'interrogation continue et efficace des flux de données.
- déploiements de fonctions bases de données sur des architectures systèmes, voire matériels. Il s'agit de décider sur quelles architectures déployer les services pour les exécuter et les coordonner.

Les problématiques explorées dans ces thèmes sont :

- le déploiement de fonctions de gestion de bases de données sur des architectures logicielles et matérielles : ces fonctions impliquent des processus qui peuvent demander des ressources de calcul importants. Elles doivent être revisitées et déployées sur des architectures comme la grille, et le *Cloud* mais aussi sur des réseaux de capteurs, et de dispositifs mobiles ou non, avec des capacités physiques différentes. Un accent particulier est mis sur les environnements de clusters avec des approches parallèles et distribuées *à la* MapReduce.
- la livraison de fonctions de gestion de bases de données (''*data management functions delivery*''). Il s'agit de trouver des infrastructures donnant accès à ces fonctions/services. Par exemple, la virtualisation et la para-virtualisation des services de stockage qui incluent la mise à disposition d'espaces de stockage avec la gestion transparente de ces ressources.
    - Gestion des espaces de stockage hautement distribués qui sont mis à disposition par différents fournisseurs.
    - Mise à disposition continue de données, ce qui implique la répartition et la duplication de données.
    - Transformation de données persistantes vers de nouveaux modèles de données et aussi vers de nouveaux supports matériels.

Les pistes de recherche actuelles sur les architectures pour la gestion de données se trouvent autour du *Cloud Computing*, Web2.0, réseaux de capteurs et de dispositifs. Les systèmes de gestion de données à base de services (inspirés des systèmes à composants) qui fournissent une gestion élastique, multi-échelle, dynamique, et efficace des services est une approche qui semble prometteuse et pertinente. Ces systèmes respectent les principes et méthodologies des *service-oriented architecture* (*SOA*) qui apportent en particulier une notion d'asynchronisme (également nommé EDA pour *Event-Driven Architecture*) pour répondre aux enjeux récurrents de flexibilité.

Une caractéristique clef du *Cloud Computing* est son élasticité; les utilisateurs peuvent facilement ajuster les ressources informatiques nécessaires offrant une variété de CPU, I / O, performance de la mémoire et de prix. Ceci induit diverses possibilités de planification des tâches des applications en charge de la gestion des données et qui utilisent généralement des clusters statiques. L'utilisation d'une plate-forme de *Cloud* conduit à considérer les caractéristiques des données et des tâches permettant aux fonctions de gestion de données de se déployer sur les nœuds avec de meilleurs performances I/O ou CPU. Evidement, ce déploiement doit être envisagé comme dynamique puisque le nombre de nœuds d'un nuage n'est pas statique. Par conséquent, toute fonction de gestion de données doit être conçue comme un système adaptable devant tenir compte des contraintes de coût économique (et pas seulement algorithmique) et d'énergie dépensée pour s'exécuter.

Par ailleurs, d'autres aspects comme la robustesse – qui vise à offrir une performance stable et prévisible à tout moment - est à considérer. Il s'agit d'un problème difficile dans le cas de systèmes classiques mais qui s'amplifie avec les requêtes/calculs dans les environnements ambiants, en particulier dans le *Cloud*, où les « défaillances » de nœuds et la variabilité des performances sont courantes (par exemple avec le nombre de nœuds utilisés, les défaillances ne sont plus considérées comme des exceptions, mais comme une norme). Ceci repose le problème du monitoring des processus de gestion de données et de leur adaptation à la volée aux caractéristiques de l'environnement matériel et des contraintes de performances. Il s'agit globalement de traiter des problèmes de réplication de services et de données pour assurer une disponibilité continue des données tout en respectant la sécurité, le contrôle d'accès, la propriété des données, l'exploitation vs la confidentialité des données issues des services/fournisseurs divers et bien évidement de la tolérance aux pannes.

3- Accès aux données

Les nouvelles infrastructures distribuées à large échelle (grille, P2P, *SOA*, *Cloud*) couvrent un large spectre de configurations d'accès aux données allant de l'accès aux données stockées sur des disques locaux jusqu'à l'accès continu à des flux de données générés simultanément par des milliers

de capteurs mobiles. Elles ont en même temps montré les limites des technologies existantes et positionné l'accès aux données au cœur des recherches en bases de données, mais également au cœur des recherches dans d'autres disciplines comme l'apprentissage, la fouille de données et les réseaux.

Les recherches sur l'accès aux données peuvent être organisées selon trois grands axes: (a) les langages des requêtes (et leurs modèles de données associés), (b) la compilation et l'optimisation de requêtes et (c) leur évaluation.

Langages

L'approche traditionnelle qui consiste à fournir un langage logique produisant une réponse exacte et complète (par rapport à une expression logique bien définie et une BD fermée) à une requête ensembliste rencontre vite ses limites, en particulier dans les environnements ubiquitaires et dynamiques (où l'utilisateur ne connaît ni la description des sources de données ni le nombre de sources). Aujourd'hui, les requêtes des utilisateurs sont imprécises (critères flous), incomplètes et souvent constituées uniquement de quelques mots clés à l'image des requêtes de recherche d'information sur le Web.

Ce type de requête spécifie une intention et non un besoin précis ; leur interprétation nécessite alors un modèle spécifique tenant compte du caractère approximatif de la requête, de la nécessité de relâcher certains critères ou de les interpréter de façon flexible pour éviter des réponses vides ou pléthoriques.

Les requêtes à base de mots clés permettent aujourd'hui, grâce à des modèles d'interprétation et d'évaluation puissants (indexation, clusterisation, PageRank), une liste de documents ordonnée selon le degré de pertinence. Chaque document proposé est supposé répondre plus ou moins bien à la requête de l'utilisateur. Mais si la réponse est éclatée dans plusieurs documents, il revient à l'utilisateur d'en faire la synthèse ou l'agrégation. De nouveaux modèles de requêtes doivent permettre de produire directement cette synthèse sous la forme d'un document couvrant au mieux l'intention de la requête. Avec la diversité des objets recherchés (documents XML, workflows, ressources RDF, ….), les requêtes à structure de graphes deviennent de plus en plus courantes, en particulier dans le domaine scientifique (recherche de structure de molécules, découvertes de patterns, …). Les langages d'expression de ces requêtes sont peu étudiés, tant dans leur représentation abstraite que dans leur interface graphique.

Les recherches sur les langages doivent donc tenir compte des nouveaux types de requêtes : requêtes à mots clés, requêtes logiques à prédicats flexibles, requêtes à structure de graphes, requêtes avec préférences, requêtes agrégatives… La question d'un langage polymorphe permettant l'expression de toutes ces requêtes se pose. Les recherches s'orientent vers des sous-langages spécialisés dont les fondements théoriques permettent de mieux comprendre la structure et la sémantique des requêtes, leur méthode d'interprétation et les problèmes liés à leur calculabilité (complexité). Elles sont à rapprocher des orientations de recherche dans les langages de programmation avec les DSL (domain specific languages).

Les problèmes de recherche suivants doivent particulièrement être investigués au niveau langage :

- L'intégration de critères flexibles et de préférences dans l'expression des requêtes,
- La prise en compte du contexte de l'utilisateur ou de l'application dans l'interprétation de la requête,
- La reformulation des requêtes tenant compte d'une part du profil et du contexte de l'utilisateur, et d'autre part, de la méconnaissance des sources de données cibles,
- L'expression de requêtes agrégatives permettant l'élaboration de résultats à partir de fragments d'informations.

Optimisation

L'optimisation est une étape importante dans le processus d'évaluation de requêtes car elle détermine son plan d'exécution. Ce plan présente la meilleure distribution de calcul, d'utilisation de mémoire, de combinaison des opérateurs sur les données et les meilleurs algorithmes pour chaque opérateur. Les

objectifs d'optimisation sont classiquement représentés au travers d'une fonction de coût fondée sur des critères de temps d'exécution d'opérateurs, de services, de communication, sur des statistiques sur les données, voire des heuristiques.

L'interrogation de données massivement distribuées (et cachées) fait que le temps ne peut plus être le seul critère d'optimisation. D'autres paramètres comme l'énergie consommée par les dispositifs, l'espace mémoire nécessaire à l'exécution d'une opération/service, le coût financier ou économique (lié à l'utilisation de certains types de réseaux ou à certains services payants) doivent également être pris en compte afin de trouver le meilleur compromis. Des pistes de recherche concernent par exemple la définition de nouveaux modèles de coûts et la définition d'algorithmes permettant de réduire la consommation d'énergie. Par ailleurs dans le cadre de requêtes d'analyse où il s'agit d'évaluer des requêtes par association/agrégation de fragments d'information, il est possible d'envisager la mise en place de techniques de mémoire associative et/ou filtrage intelligent pour optimiser les calculs.

L'optimisation doit ainsi satisfaire une variété d'objectifs et de contraintes et tenir compte également des spécificités de l'architecture sous-jacente. En particulier, pour les architectures de grilles ou de *Cloud*, les travaux de recherche concernent l'optimisation des *workflows* ou graphes d'activités sur des données. Les difficultés sont liées à la définition de la meilleure coordination d'activités selon les critères de qualité de services et la possibilité de considérer des réponses approchées de manière à vérifier les contraintes économiques et de temps.

On retrouve des problèmes similaires pour les requêtes dites continues sur des flux de données. La phase d'optimisation des requêtes continues, quant à elle, vise plus spécifiquement à :

- échantillonner le flot de données de manière à borner les résultats des différentes requêtes et réduire les coûts de communication ;
- voir comment agréger/résumer les données / réponses de telle sorte qu'il soit possible d'obtenir des réponses approximatives avec une précision bornée ;
- placer les opérateurs au bon niveau du réseau de dispositifs pour l'exécution et ordonnancer les flux (ou parties de flots) de manière efficace.

Le modèle d'optimisation classique des bases de données est basé sur une connaissance a priori des opérateurs qui seront exécutés. Dans les approches *NoSQL*, à base de mots clés, cette hypothèse est mise à mal : il n'est pas toujours possible de disposer des informations sur les données et les opérateurs, en particulier dans l'approche orientée services. De nouvelles approches basées sur les statistiques, l'échantillonnage de données, les profils des services, les coûts moyens, et de manière plus large la surveillance du processus d'exécution des requêtes, permettent d'en extraire des connaissances utiles à l'optimisation.

Evaluation

Du point de l'exécution des requêtes, les nouveaux environnements distribués dynamiques (e.g. introduisent également de nouveaux défis. Certains travaux ajoutent des structures d'indexation à des architectures P2P pour localiser efficacement des données intéressantes et/ou améliorer l'expressivité des langages de requêtes. Ces systèmes reposent sur un schéma global et souvent des organisations logiques de réseaux prédéfinies.

Les recherches portent essentiellement sur la distribution, le cache des données et l'implantation d'opérateurs et d'algorithmes adaptés. Il s'agit de proposer des algèbres spécialisées et hybrides (autorisant les traitements sur données statiques et flux) qui intègrent des aspects d'approximation et de distribution. L'implantation des opérateurs tenant compte de nouveaux supports (processeurs, stockage, protocoles réseau) et modes de programmation (*MapReduce*, ...)

Les stratégies d'évaluation qui doivent subir des changements profonds sont corrélées:

- à l'optimisation « Cross-layer » dès lors que le réseau fait partie des conditions d'évaluation des requêtes ;
- à la construction de réponses exhaustives vs de réponses approximatives/partielles ;

- aux modèles d'exécution de composition de services (de requêtes) manipulant des données provenant de plusieurs fournisseurs ;
- aux algorithmes exécutables dans des conditions de ressources contraintes (puissance de calcul limitée, économie d'énergie, etc.). Ces algorithmes s'appuient sur des représentations / interprétations temps-réel des données.

4- Sécurité et confidentialité des données et services

L'enjeu de la sécurité et de la confidentialité des données a été identifié depuis de nombreuses années (création en 1986 de l'IFIP WG 11.3 Working Group on Data and Application Security and Privacy, en 1998 du journal ACM transactions on Information and System Security – TISSEC, apparition de la thématique à la conférence française Bases de données Avancées – BDA – en 1996). De nombreuses solutions adaptées à la gestion de données ont ainsi été proposées, selon l'évolution des systèmes et des usages. Parmi ces évolutions récentes, on peut citer :

- La complexité croissante des systèmes : actuellement, n'importe quel téléphone mobile muni d'un navigateur web embarque de fait un composant de gestion de données (de type *SQLLight* par exemple). Il faut alors s'assurer de la sûreté de fonctionnement de ce composant complexe, en présence de données personnelles ou de transactions financières. D'une façon générale, les nouvelles architectures à base de services rendent nécessaire la mise en place de politiques de sécurité au niveau de la donnée et de sa sémantique ainsi qu'au niveau des services ou des traitements applicatifs.
- une distribution toujours plus large : le recours aux services en P2P ou en nuage expose les utilisateurs à de nouvelles menaces, les attaques pouvant provenir du fournisseur de services lui-même ou des utilisateurs avec lesquels les ressources sont partagées.
- la présence de données plus sensibles : les propositions techniques de gestion de données personnelles (agenda partagé), médicales (dossier médical électronique), ou sociales (réseaux sociaux, *microblogging*) incitent les utilisateurs à dévoiler toujours plus d'informations sensibles. Il existe de plus un marché en croissance exponentielle pour cette information, qui est utilisée par les entreprises par exemple pour le ciblage publicitaire ou le profilage de leurs employés, ce qui incite à son exploitation irraisonnée, sans l'accord éclairé des utilisateurs.
- la cohabitation avec des contraintes juridiques : la fluidité de l'échange de données entre en collision avec de nombreux systèmes de régulation (droit à la vie privée, respect des données médicales, droits d'auteur, propriété intellectuelle ou industrielle). Cette régulation est à la fois de nature technique et juridique : les concepts manipulés n'étant pas nécessairement compatibles ou correctement modélisés.

Notons que la combinaison de mécanismes de sécurité peut révéler des contradictions. Par exemple, un mécanisme de contrôle d'accès pourrait interdire l'accès à une certaine ressource pour un utilisateur tandis qu'un autre laisse explicitement un tel accès. La détection automatique et éventuellement la génération d'explications des origines de telles contradictions nécessite au préalable l'intégration de politiques de sécurité et de divers modèles d'autorisation qui peuvent coexister (e .g. un mécanisme de contrôle d'accès avec un mécanisme de négociation de confiance basé sur des permissions/qualifications).

Pour prendre en compte ces évolutions dans une approche globale de la gestion de données et de services, les problématiques de recherche à traiter sont les suivantes :

- Identification et intégration de concepts juridiques pertinents dans les modèles et les processus de développement des systèmes de gestion de données.

  Par exemple le traitement de données médicales est encadré par de nombreuses contraintes. De même, les dispositifs de protection de la propriété intellectuelle sont complexes. Il est nécessaire de pouvoir traduire fidèlement ces contraintes par des langages expressifs de contrôle d'accès et d'usage. Ce travail doit être réalisé en collaboration avec des spécialistes du droit des technologies

de l'information. Il est nécessaire de prendre en compte de façon anticipée les contraintes et outils juridiques de la sécurité dans le développement des systèmes (approche dite *privacy-by-design).*

- Adaptation des procédés cryptographiques aux modalités de traitement de données massives.

  Les primitives cryptographiques classiques sont déjà largement utilisées dans des composants de gestion de données pour différents buts (intégrité des données, confidentialité, anonymisation et protection de la vie privée, protection de la propriété intellectuelle, traçabilité des données). Il faut d'une part améliorer les performances de ces approches (approches matérielles, nouveaux protocoles), et d'autre part dépasser une approche « composant par composant » pour une intégration globale de ces primitives dans une gestion complète des données (compatibilité entre primitives cryptographiques au sein d'un système, sans fuite d'information). Un fort compromis entre sécurité et performance doit être réalisé.

- Certification de la gestion de données de bout en bout.

  Il s'agit de proposer des méthodologies de développement de bout en bout permettant d'obtenir des systèmes de gestion de données et dérivés (P2P, DHT, etc.) certifiés. De la même façon que des processeurs ou des compilateurs certifiés apparaissent, il est crucial d'aboutir à la réalisation d'un système de gestion de données complètement certifiés (au moyen par exemple d'un assistant de preuve comme Coq).

Comme indiqué en introduction de cette section, il convient de prendre en compte dans ces problématiques la façon dont les données sont distribuées, et sur quel support matériel elles évoluent : les systèmes embarqués (capteurs, téléphones étendus, objets mobiles, domotique, serveur de données personnel et/ou en mémoire FLASH), les ordinateurs personnels, les nœuds d'un réseau P2P, les serveurs centralisés, membres d'une grille de calcul ou organisés en nuage (services Web, *Software as a Service* -- SaaS -- et *Platform as a Service* -- PaaS). Chaque type de distribution et de support matériel pose en effet des contraintes spécifiques du point de vue juridique et technique. Par ailleurs quand on combine des mécanismes de sécurité, on doit vérifier que le résultat de la combinaison satisfait les exigences du système dans sa globalité. Ceci nécessite la caractérisation formelle de la notion de satisfiabilité. Des mécanismes de propagation de mises à jour et d'explication de leurs impacts sont nécessaires.

5- Données et enjeux sociétaux

Les générateurs et exploiteurs des espaces de données relèvent de divers domaines : commerce et affaires (systèmes d'information d'entreprise, banques, transactions commerciales, systèmes de réservation, réseaux, …), gouvernements et organisations (lois, réglementations, standards, infrastructures, ….), loisirs (musique, vidéo, jeux, réseaux sociaux, …), sciences fondamentales (astronomie, physique et énergie, génome, …), santé (dossier médical, sécurité sociale,…) environnement (climat, développement durable, pollution, alimentation,…), humanités et sciences sociales (numérisation du savoir, données archéologiques, …). La suite de cette section précise le positionnement des recherches en gestion de données dans certains de ces domaines.

5-1 : Données et Web

Le Web s'est considérablement enrichi de données sociales apportées par les utilisateurs (*crowdsourcing*), sous forme d'annotations, d'opinions, et d'interactions, d'expériences supportées par des plates-formes sociales. Cette information permet d'explorer le Web de façon personnalisée, afin de prendre en compte les préférences de chaque utilisateur et de lui proposer une expérience unique reflétant ses intérêts. L'augmentation de la quantité de métadonnées acquises auprès des utilisateurs ou de communautés permet d'effectuer des recommandations sans cesse plus précises, et d'augmenter ainsi la satisfaction des utilisateurs. Ces métadonnées sont aussi liées aux éléments de preuve sur les actions que les documents subissent et par quelles entités (provenance). Les documents sont en effet au cœur de nombreux processus pour lesquels des contraintes métier (liées aux usages des données), techniques et légales doivent êtres vérifiées.

Les technologies du web sémantique sont utilisées à grande échelle pour publier et relier des données sur le web. Elles doivent être complétées pour analyser et organiser les données du Web et du Web social, et de passer de l'un à l'autre au travers de projets de recherche sur l'apprentissage et la fouille de données à large échelle permettant de : structurer les collections, enrichir les documents, inférer des communautés d'utilisateurs, calculer les proximités entre entités, mettre en correspondance leurs vocabulaires et reconnaître des structures communes, voire prédire de nouveaux liens.

5-2 : Données et Services

Les infrastructures de communication, stockage et de calcul visent au déploiement de services innovants par exemple nomades et ubiquitaires, enrichis, personnalisés ou personnalisables, interactifs, permettant de distribuer du contenu. Les défis scientifiques et techniques concernent la prise en compte et l'adaptation au contexte d'utilisation de ces services, l'offre d'une composition dynamique de ces services, la virtualisation des ressources logicielles et matérielles pour une distribution et une disponibilité au plus grand nombre, le passage à l'échelle et la prise en compte de la mobilité. Ces défis relèvent principalement du domaine du Génie Logiciel et ne sont pas sans lien avec le domaine des bases de données de part la prise en compte de la dimension Données et non seulement contrôle dans les processus de déploiement, d'exploitation, d'administration, d'optimisation des ressources, de contrôle, et de programmation et d'algorithmiques innovantes.

La problématique des données est cruciale avec la mise à disposition de masses données hétérogènes et dynamiques nécessitant des placements adaptés, des recherches efficaces et une cohérence adaptée. Ainsi, le développement de plates-formes de services permettant un accès rapide à des données réparties, hétérogènes et multiformes, associées à des services composables dynamiquement est un enjeu majeur.

5-3 : Données, santé et bien-être

Avec l'intégration de l'informatique, de la robotique et des organismes vivants, nous sommes au début de l'ère de la cybernétique, des systèmes de systèmes extrêmement complexes, à la fois vivants, mécaniques, électroniques et informatiques. Nous avons ici des données à gérer à un niveau très spécifique et en environnements contraints. A l'opposé, la mise en place de systèmes d'information basés sur une intégration intelligente de flux de données et un accès performant à ces informations parfois incomplètes doit permettre de se préparer aux mutations démographiques à venir avec le vieillissement de la population, qui devraient susciter des adaptations et des offres renouvelées du côté de l'habitat, de l'aménagement et des services urbains.

5-4 : Données et simulations numériques

La simulation numérique désigne le procédé de représentation d'un phénomène physique. Elle exige une étape préalable, celle de la modélisation (traduction d'un phénomène en langage mathématique). On distingue trois types de simulation: *de conception* dans la mécanique des fluides, la science des matériaux, etc. ; *prédictive* (simulation des phénomènes dans le nucléaire, la météorologie…) et *comportementale* -- une approche multi-agents consistant à décrire le comportement de chaque entité qui s'adapte à la situation courante dans les trafics routiers, phénomènes biologiques et sociaux. Du point de vue des données, il s'agit d'offrir (i) des algorithmes robustes, performants et adaptés aux traitements dans des calculateurs de forte puissance (grilles, clusters, multi-cœurs spécialisés ou hybrides), (ii) des solutions de stockage, des systèmes d'exploitation ou des intergiciels distribués à grande échelle.

5-5 : Données et environnement

La surveillance de la qualité de l'environnement est le pilier principal des politiques environnementales. Tous les milieux environnementaux sont concernés : eau, air et sols. De même, ce thème couvre à la fois les problématiques de surveillance de la qualité de l'environnement, de prévention des risques naturels ou anthropiques, y compris pour la surveillance d'événements extrêmes. Les thématiques de recherche se focalisent sur la mise en place de stratégies de déploiement (au sol, en mer ou dans

l'espace), d'instrumentations à bas coûts, d'acquisition de données de masses et d'outils de validation, de gestion et d'interprétation de ces données. Il s'agit de déployer des systèmes d'instrumentation intégrés (capteurs, acquisition, interprétation) très innovants et couplant les systèmes d'observation spatiale ou aérienne et les mesures au sol. Du point de vue de la gestion des données il s'agit de :

- collecter, traiter et analyser en continu les informations environnementales, économiques, sociales et de santé ;
- évaluer les interactions entre environnement et développement économique et social pour être en mesure d'objectiver les progrès vers le développement durable ;
- réaliser des analyses et des études prospectives pour construire des scénarios d'avenir et guider les prises de décisions.

5-6 : Données et énergie

La réduction de l'empreinte environnementale des infrastructures de calcul et de communication est une préoccupation forte. Il s'agit d'optimiser la consommation d'énergie des objets communicants, des réseaux, des *''data centers''* et des calculateurs (*Green IT*).

Les systèmes de gestion des données sont évidemment concernés par deux grands aspects :

- le stockage des données (centres de données)
- prise en compte de la dimension énergie dans les fonctions de gestion des données.

Un autre objectif dans ce domaine est le *Green by IT* en facilitant, généralisant et rendant sûrs les différents modes de communication, autorisant le télétravail, les téléconférences, les services associés, la consultation médicale à distance, etc. évitant ainsi des déplacements. La gestion sécurisée des données et de leur confidentialité est au cœur de ces processus.

5-7 : Données ouvertes

L'ouverture des données publiques est « une vague inévitable et cette tendance va se poursuivre de façon générale sur la décennie » (F. Bancilhon, CEO Data Publica). Elle devrait permettre d'amplifier la mise à disposition, à encore plus d'utilisateurs et d'acteurs, d'un nombre grandissant de données enregistrées. Les données publiques ne représentent qu'une partie de cet écosystème des échanges de données. La plupart des secteurs d'entreprise produisent une grande variété de données, sans savoir toujours précisément comment elles fonctionnent, comment elles sont structurées, collectées, utilisées, ni ce qu'il est possible d'en faire. De plus se posent des questions sur la nature des données à partager et à proposer à la réutilisation dans la mesure où ces données donnent des indications sur les activités et l'identité du propriétaire.

Cette ouverture des données favorisant l'observation des systèmes et la prédiction des actions a un fort impact sur les citoyens, les entreprises et les organismes publics. Les recherches portent sur : - une meilleure disponibilité des données en orientant l'effort sur les jeux de données à fort potentiel politique, social ou économique, - une meilleure complétude et compréhension des données (représentation) couplées aux techniques élaborées de traitement et visualisation des *big* data, - et également sur une nouvelle façon de livrer les données selon des modèles économiques qui doivent encore être définis (Data as a Service).

5-8 : Données pérennes

Matière première, transformée, partagée, la donnée a néanmoins une durée de vie limitée, longue mais limitée comme le patrimoine informationnel des entreprises, les données personnelles (stockées dans les disques privés ou publiés sur le Web) ou les données publiques (fichiers sécu, police, …) ou parfois une durée illimitée comme les données représentant des connaissances scientifiques, des produits culturels, des connaissances archéologiques et environnementales ou encore sociétales (e.g. enquêtes, recensements).

En dehors des recherches pour le développement de mémoires et matériels pour le stockage et l'archivage de ces grands espaces de données, se pose la question de la caractérisation des données à pérenniser, de la définition des processus de collecte et/ou de compression de ces données, et de

la construction des systèmes d'archivage adéquats et pérennes ! Les défis concernent le chargement/la transformation, le stockage, la compression des données, les méthodes d'accès, et l'administration des archives. Ces problématiques recoupent celles de la gestion de l'Interopérabilité entre les sites d'archivage, de la distribution des archives et processus et des migrations des archives entre divers sites. On retrouve ici des problématiques de systèmes d'intégration de données et d'interopérabilité des systèmes. La dématérialisation des données posant comme nous l'avons vu des problèmes de sécurité, d'authenticité et d'intégrité ; ces problèmes restent vrais pour les données pérennes sans compter les coûts de leur préservation (conversion de formats, migration, pérennisation de logiciels). La pérennité des données inclut également la pérennité de leur provenance qui leur donne une valeur probante.

5-9 : Données scientifiques

Dans de nombreux domaines scientifiques, tels que la physique, l'astronomie, la biologie ou les sciences de l'environnement, l'évolution rapide des appareils et instruments scientifiques ainsi que le recours intensif à la simulation informatique ont conduit, ces dernières années, à une production importante de données. Les applications scientifiques modernes sont alors confrontées à de nouveaux problèmes qui sont liés essentiellement au stockage et à l'exploitation de ces données. Outre le volume croissant des données à manipuler, leur nature complexe (e.g., images, données incertaines, multi-échelles…), l'hétérogénéité de leurs formats ainsi que les traitements variés dont elles font l'objet constituent les principales sources des difficultés. Les problèmes posés sont tels que la gestion des données scientifiques est reconnue aujourd'hui comme étant un véritable goulot d'étranglement qui a pour effet de ralentir les recherches scientifiques, ces dernières s'appuyant de plus en plus sur l'analyse de données massives. Dans ce contexte, le rôle de l'informatique, comme un moyen direct qui permet d'améliorer le processus de découvertes en science est primordial. Ce constat a conduit les scientifiques de disciplines différentes à unir leurs efforts de réflexion pour faire émerger de nouveaux outils, approches et techniques de gestion et d'exploitation de ces gigantesques masses de données. C'est le cas par exemple des deux conférences XLDB (eXtremely Large Data Bases, http://www.xldb.org) et SciDB (Scientific Data Bases, http://www.scidb.org/).

Crédits :

- Challenges of *strategic interest to European society*, Axes recherche EIT ICT Labs,
- Ministère de l'Enseignement Supérieur et de la Recherche : Stratégie Nationale de Recherche et d'Innovation, Alliance Allistene, *Alliance des sciences et technologies du Numérique,*
- Agence National de la recherche (Défis sociétaux, économiques, environnementaux et Fiches programmes/appels projets),
- Ministère du redressement productif, Direction générale de la compétitivité, de l'industrie et des service : Programme Développement de l'économie du numérique des investissements d'avenir, Technologies clés 2010 et 2015.

http://eprints.soton.ac.uk/257648/1/The_Data_Deluge.pdf
http://www.nitrd.gov/pubs/200311_grand_challenges.pdf
http://www.lsst.org/lsst/ : *Large synoptic survey telescope*

http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/

http://www.cs.purdue.edu/homes/ake/pub/CommunityCyberInfrastructureEnabledDiscovery.pdf

http://www.emc.com/leadership/programs/digital-universe.htm