Vialis
GROUPE QUODAGIS

EBOOK

OBJETS, FLASH, BIG-DATA, NO-SQL, HADOOP, ETC.

Progrès ou cacophonie dans la gestion des données ?

Auteur Bertrand Quillévéré Vialis

Sous l'angle métier, comment évolue la production et l'usage des données ?

Les volumes de données produites par les entreprises, les particuliers et les objets connectés explosent littéralement. IDC estime cette croissance à environ +45% par an soit une multiplication par 40 en 10 ans. Nous atteindrions donc en 2020 35 Zo (zettaoctets soit 1021 soit un million de petaoctets). L'essentiel de ces volumes concerne des données non-structurées : (textes, images, vidéos, sons,...). Les données pèsent moins mais devraient toutefois augmenter avec les objets connectés qui en produiront plus que les personnes. Partant de 5 milliards actuellement, Gartner estime à 21 milliards le nombre d'objets connectés en 2020, dont 13 chez les particuliers et le reste en entreprise.

Par exemple, les sondes d'un Boeing 787 génèrent 20 To/h soit 3,6 Po par jour pour une flotte de 10 appareils. Le dossier médical moyen d'un patient AP-HP pèse 800 Go soit 800 Po par million de patients.

Les usages du « BigData » sont multiples et il en apparait de nouveaux en permanence. Marketing, Intelligence économique, Sciences et Médecine en particulier, Recherche, Sécurité.... Les données seront le principal carburant du « Machine learning » processus clef de l'intelligence artificielle. Dès 2011, le fameux Watson d'IBM a gagné le jeu télévisé Jeopardy en brassant à chaque question en 3 secondes 200 000 pages WEB dans un cluster Hadoop avec 90 serveurs sous Linux.

Toujours grâce aux travaux d'IBM, très actif sur le sujet avec Watson, la maternité de Toronto a pu, à partir d'historiques de paramètres biologiques de dizaines de milliers de nourrissons, prédire de manière fiable, avant n'importe quel pédiatre, quels bébés vont développer des infections néo-natales (sans pour autant dire pourquoi...). L'analyse des données des réseaux sociaux permet de suivre et prédire l'évolution des épidémies, les variations des prix ou du marché du travail...

Quelles entreprises sont réellement concernées ?

En première approche on pourrait penser que BigData et intelligence artificielle ne concernent que quelques niches métier spécifiques (Marketing, Finances, Juridique, Médecine,...). En approfondissant l'étude on réalise rapidement que tous les secteurs d'activité vont être rapidement concernés en grande partie du fait des objets connectés et des robots. Quasiment toutes les entreprises seront concernées par un mixage de données produites elles-mêmes par leurs clients et par d'autres entreprises avec de réels enjeux économiques par exemple pour les maintenances préventives (40 000 € de perte d'exploitation journalière pour notre exemple du Boeing 787).

Quoi de neuf dans les problématiques de stockage et de gestion de données ?

Les nouveautés découlent de la croissance des volumes, de la pertinence du stockage dans le Cloud et plus seulement dans l'entreprise. Dans un état type de la décennie 2000 les entreprises disposaient de deux types de stockage et de données :

1. Des données structurées stockées dans des baies de stockage fournies par les « Big Five » : (EMC, HP, IBM, HITACHI, NETAPP) accessibles via un SAN-FC et dédiés aux bases de données et aux systèmes de fichiers techniques des serveurs ainsi qu'à diverses formes de réplicas et de sauvegardes.

2. Des données non-structurées exploitées en mode fichier dans des NAS (NETAPP, EMC, MICROSOFT,...). Ces NAS stockent les fichiers bureautiques et collaboratifs en général pour des serveurs sous Windows (en CIFS) et les flux de données inter-applications UNIX (en NFS) ainsi que des fichiers de sauvegardes.

Ce paysage technique du stockage des données reste encore très présent en l'état en 2016. Toutefois la décennie en cours montre de nombreuses transformations en cours :

1. Les applications collaboratives structurées (Messagerie, Agenda, Visio, Téléphonie, Chat, Intranet,...) basculent dans le Cloud dans des offres intégrées, aujourd'hui matures et économiquement attractives (principalement Office 365 de Microsoft et son challenger Google Apps). Les données associées passent donc des NAS ou SAN internes vers le Cloud.

2. Les données des utilisateurs sous forme de fichiers de natures diverses stockées dans des partages de NAS CIFS d'entreprises migrent elles aussi vers le Cloud soit en accompagnement des suites collaboratives (OneDrive de Microsoft, Google Drive) soit dans des offres spécifiques du Cloud comme Dropbox. Les deux bénéfices majeurs de ce mouvement sont la disponibilité en mobilité et télétravail de l'accès à ces données et la simplification de la solution de sauvegarde et retour à des versions antérieures qui deviennent implicite dans les services du Cloud.

3. La migration d'une partie des applications de l'entreprise en mode SaaS : CRM, RH, ERP,... qui s'accélère et entraine le plus souvent le transfert de la totalité des données associées vers le Cloud.

4. Pour les applications restant dans l'entreprise le stockage évolue fortement par adoption du stockage Full FLASH (pour les bases de données structurées) ou hybride FLASH-Capacitif (pour le reste) en accès SAN, avec de nouveaux acteurs qui viennent concurrencer les Big-Five : PURESTORAGE, VIOLIN, ...

5. L'intégration du stockage dans des architectures hyperconvergées (NUTANIX, VMware,...) qui simplifient fortement la gestion d'ensemble et réduisent le coût de possession. Ces architectures font disparaitre le réseau dédié SAN et les baies de stockage traditionnelles.

6. L'apparition pour certaines entreprises et certains métiers spécifiques d'une production de données nouvelles (objets connectés ou extension de la numérisation du Métier) structurées ou non structurées, en grand volume, pousse les entreprises à s'équiper de solutions de stockage Objet « Scale-Out » soit dans leur Datacenter avec des solutions de type « Appliance » dédiées (EMC ISILON, DDN) ou avec des solutions logicielles sur des architectures X86 banalisées, (SCALITY, CLOUDIAN, CLEVERSAFE,...), soit directement dans le Cloud.

Par rapport au NAS d'entreprise traditionnel, ces nouvelles solutions sont nettement mieux adaptées à des objets de toutes tailles en très grande quantité et avec de très grands volumes (depuis quelques centaines de TO à plusieurs centaines de PO). Elles s'appuient sur une évolutivité horizontale très forte prévue dès la conception. Avec les mouvements 1 et 2 mentionnés ci-dessus (collaboratif et partage de fichiers) l'usage résiduel des NAS traditionnels pourrait être absorbé à terme par ces nouvelles solutions.

Le coût global d'acquisition se situe dans une fourchette de 0,15 à 0,5 € par Go utile en fonction de différents critères (volume total, performances requises, etc.) à comparer avec des solutions NAS traditionnelles qui se situent plutôt dans une fourchette de 0,5 à 1,5 € par GO utile soit un facteur 3 environ.

7. Des clusters dédiés Hadoop surgissent également dans la période actuelle. Ils pourraient toutefois se consolider dans un stockage Objet banalisé ou pour de petites volumétries dans des architectures hyperconvergées.

Pour résumer, une entreprise avec une IT « à la pointe » aurait à ce jour son environnement collaboratif et une partie de ses applications standard en SaaS dans le Cloud, le reste des applications virtualisées à 100% dans une architecture hyperconvergée X86 en « stretched cluster » dual-site en Datacenter neutre ou en Cloud public et enfin un stockage Objet capacitif « low cost » en Datacenter neutre ou dans le Cloud.

Quels sont les différenciateurs entre une gestion de données structurées ou non-structurées ?

La taille des entités élémentaires est le premier différenciateur. Un objet de type image ou vidéo est nettement plus gros qu'une ligne d'une table articles par exemple. Cette taille impacte donc la volumétrie globale à gérer. Par exemple un fournisseur de VOD (Vidéo à la Demande) aura besoin d'un système permettant à des milliers d'utilisateurs d'accéder simultanément à un grand nombre de gros fichiers. Les systèmes de stockage Objet modernes découpent chaque objet en un grand nombre de morceaux pour optimiser les performances et apporter une tolérance aux pannes par redondance. Plus besoin de sauvegarde pour ces solutions à très grand volume. La protection est assurée par réplication ou mieux encore par répartition redondante sur 3 sites distants.

L'autre différenciateur majeur concerne les opérations qui seront réalisées sur les données. Elles sont basiques pour les données non structurées : on stocke et on consulte et plus rarement on efface. Pour les données structurées elles sont plus élaborées : prendre la commande et le paiement d'un client e-commerce, déclencher sa livraison et décrémenter le stock du produit qu'il a commandé, le tout vu comme une **opération unique cohérente**. On parle alors de transaction. Un système de gestion de données structurées se devra d'être « transactionnel » alors qu'un gestionnaire de stockage d'objets pourra s'en dispenser.

Voilà qui nous amène aux fameuses propriétés « ACID », de quoi s'agit-il ?

ACID signifie : Atomicité, Cohérence, Isolation, Durabilité. C'est en quelque sorte la table de la loi pour un système de gestion de données « transactionnel ». Une transaction étant définie par un ensemble d'opérations élémentaires (création, modification, suppression portant sur les données).

Atomicité : assure en toute circonstance et en particulier dans tous les cas de panne, qu'une transaction se fera soit complètement soit pas du tout : si une partie d'une transaction ne peut être faite, les données seront remises dans l'état où elles étaient avant le début de la transaction (« rollback »).

Cohérence : assure que chaque transaction amènera le système d'un état valide à un autre état valide. Tout changement à la base de données doit être valide selon un ensemble de règles d'intégrité (par exemple : pas de doublons ou existence d'un attribut dans une liste de référence, etc.).

Isolation : assure que toute transaction s'exécute comme si elle était la seule sur le système. Assure que l'exécution simultanée de transactions produit le même état que celui qui serait obtenu par l'exécution en série des transactions.

Durabilité : assure que lorsqu'une transaction a été confirmée (« commit » atomique), elle demeure enregistrée quel que soit le type de panne qui puisse se produire par la suite.

Qu'est-ce qu'un SGBD NoSQL ?

C'est un gestionnaire de données qui n'utilise pas le modèle relationnel et ne respecte pas **simultanément** les 4 règles « ACID » et permet un accès aux données autrement qu'en langage SQL : NoSQL signifie donc plutôt « Not Only SQL » et non pas « Non SQL ». Il y a donc un malentendu de départ sur ce terme. Il faudrait retenir plutôt « Non-relationnel » et surtout avec compromis sur les propriétés ACID.

Quand et pourquoi les SGBD No-SQL sont-ils apparus ?

Le terme NoSQL a été introduit en juin 2009, mais les produits répondant à cette définition sont en fait apparus début 2000. Ce sont les géants du web amenés à traiter des volumes de données très importants qui ont été confrontées aux limitations intrinsèques des SGBD relationnels traditionnels. Ces systèmes respectant les propriétés ACID, et généralement conçus pour fonctionner sur un seul serveur ou un petit nombre de serveurs en cluster leur posaient des contraintes d'évolutivité.

Par ailleurs les aspects de coûts et d'indépendance ont également beaucoup joué. De même que les géants du Web n'ont jamais adopté les produits des leaders du stockage (EMC, HP, IBM, HITACHI, NETAPP) ils n'ont pas non plus souhaité acheter des solutions bases de données ORACLE, MICROSOFT ou IBM.

Ces entreprises ont donc développé leurs propres systèmes de gestion de base de données fonctionnant sur des architectures matérielles distribuées et permettant de traiter de très grands volumes de données. Ces systèmes propriétaires : Google (BigTable), Amazon (Dynamo), Facebook (Cassandra puis HBase), SourceForge.net (MongoDB),... ont été les premiers précurseurs du modèle NoSQL.
Une communauté de développeurs de logiciels NoSQL s'est créée à partir de 2009. Les start-tups du WEB ont voulu s'affranchir des coûteux SGBDR commerciaux et ont copié les démarches de Google et Amazon.

Le modèle relationnel-SQL est-il donc périmé ou en déclin ? Les entreprises ont-elles de bonnes raisons de passer au NoSQL ?

Les entreprises ne sont pas dans les mêmes cas d'usage que les géants du WEB. Dans la quasi-totalité des cas, y compris les plus grandes entreprises, un SGBD Relationnel en cluster comme ORACLE ne sera pas à ses limites pour traiter l'accès aux données en utilisant un stockage FULL-FLASH si nécessaire. Les géants des TELECOM en sont la meilleure illustration, en brassant des quantités de données et des flux transactionnels extrêmement élevés avec des SGBD Relationnels.

Les raisons d'un abandon éventuel et partiel du relationnel en entreprise portent plutôt sur la question de l'agilité des développements, d'une volonté d'adopter l'Open Source, de baisser les coûts de licences ou d'adopter le Cloud. La culture d'une nouvelle génération de développeurs d'applications pèse aussi fortement. Ils préfèrent dépendre le moins possible des équipes infrastructures et des administrateurs de bases de données pour la gestion des données à minima dans les premières étapes d'un développement agile. Certaines entreprises créent des équipes dédiées aux nouvelles applications innovantes orientées « économie numérique ». Ces sortes de startups internes développent souvent dans le Cloud et dans 80% des cas gèrent les données en NoSQL.

Les SGBD Relationnels historiques sont subjectivement perçus comme des machineries lourdes et complexes conçues au siècle précédent dont les développeurs aiment bien s'affranchir.

Y aura-t-il cohabitation entre SQL et NoSQL ?

Oui et cela pour deux raisons. D'une part une approche raisonnée des choses poussera à utiliser malgré tout un SBGD Relationnel transactionnel quand cela est pertinent, d'autre part les nouvelles applications ne peuvent pas faire abstraction de l'héritage historique (Legacy). Les données au cœur des SI des entreprises sont de fait gérées par des SGBD Relationnels. Les nouveaux systèmes seront donc doublement hybrides : d'une part en mixant SQL et NoSQL et d'autre part en mixant hébergement interne et Cloud. La mixité des deux ne sera pas simple à mettre en musique : gare aux usines à gaz !

Un de nos clients a remplacé avec grand succès son CRM mainframe par un Cloud Hybride où l'applicatif tourne sur Amazon AWS en conservant la base de données Oracle dans son Datacenter hébergé.

N'y a-t-il pas également un mouvement d'adoption du relationnel en Open Source ?

En effet ce mouvement existe. Avec la réduction des coûts à presque tous les niveaux (serveurs, stockage, hyperviseurs, logiciels de sauvegarde, etc.) les utilisateurs voient la ligne licence/maintenance ORACLE ou Microsoft arrivée en haut de la liste de leurs principaux budgets. Au fil du temps ils trouvent le niveau de facturation exigé en maintenance par ces SGBD commerciaux de plus en plus illégitime.

Historiquement MySQL a porté le mouvement vers l'Open Source aussitôt remplacé par MARIADB après l'appropriation de MySQL par ORACLE. Historiquement l'adoption de MySQL était cantonnée à des usages périphériques non stratégiques et non critiques.

Aujourd'hui, le SGBD relationnel Open Source phare est désormais POSTGRESQL qui dans 90% des cas d'usage pourra remplacer ORACLE ou SQL-SERVER. Il présente désormais toutes les qualités requises pour une très large adoption. Les conditions de migration vers POSTGRESQL étant moins complexes que certains ne le disent.

Quels sont les avantages de flexibilité des SGBD NoSQL ?

La première étape de la création d'une base de données relationnelle est de définir son schéma ce qui crée une certaine rigidité dans le développement et implique d'avoir une bonne vision dès le départ ainsi que des évolutions et par ailleurs est contraignant pour la gestion des évolutions. Les systèmes NoSQL sans-schéma peuvent ignorer cette étape et gérer des données hétérogènes au fur et à mesure du besoin et de l'arrivée des données. Cette utilisation permet une grande flexibilité et des capacités d'adaptation en contrepartie d'une plus grande complexité d'intégration des données entre elles dans le développement applicatif.

Avec le NoSQL, les développeurs d'applications ne sont plus dépendants des équipes infrastructures et DBA pour la gestion des données. C'est pourquoi les développeurs tout particulièrement les plus jeunes adoptent volontiers le NoSQL.

Face à toutes ces opportunités de transformations possibles, comment aborder le sujet ?

Le plan de transformation stockage-gestion de données doit s'inscrire dans un plan de transformation plus vaste du SI. L'entreprise devra clarifier sa stratégie Cloud et construire sa feuille de route (RoadMap) vers un Cloud hybride dans un premier temps et cela en regard des enjeux métiers basés sur ses projets concrets de court terme et d'une vision à moyen-long terme.

Un minutieux travail d'analyse et de préparation du plan de transformation est nécessaire. Il doit faire appel à un accompagnement par de l'expertise externe nourrie des retours d'expériences d'autres utilisateurs plus avancés. Les aspects de transformation sociale dans l'entreprise doivent impérativement être pris en compte. En effet l'inéluctable transformation vers le Cloud impacte beaucoup plus fortement les ressources humaines de l'IT que les transformations du passé.