Archive, dé-duplication de données, compression : état de l'art, techniques et performances

Clément Ghnassia
Mathieu Leroux
Edgar Rodriguez

24 février 2012

Chapitre 1

Introduction

1.1 Analyse

La sauvegarde est, de manière générale, une opération qui consiste à copier des données, souvent importantes, de façon à les protéger en cas de perte ou d'altérations de celles-ci ou des systèmes et supports sur lesquels elles sont stockées. Cette opération est assez proche de l'archivage. Même si les procédés mis en oeuvre sont les mêmes, l'objectif in fine est lui différent : tandis que la sauvegarde consiste à se prémunir contre d'évenutelles altérations de données encore utilisées, l'archivage consiste lui à conserver des données obsolètes dans un but plus historique, c'est à dire qui ne sont plus utilisées, mais dont on pourrait avoir besoin d'exploiter dans le futur.

Aujourd'hui plus que jamais, la sauvegarde des données est une opération critique pour les particuliers, mais encore plus pour les entreprises. Bien qu'ils devraient avoir une place centrale et faire l'objet de toutes les attentions et de toutes les vérifications possibles, les processus de sauvegarde, et tout ce qu'ils englobent, sont souvent délaissés. On explique assez facilement pourquoi tant de négligeances. Tout d'abord il faut comprendre que la sauvegarde n'est utilisée qu'en cas de pertes de données, ce qui est normalement rare dans une entreprise. Il n y a donc aucune visibilité pour les utilisateurs dans leur quotidien. De plus, tous les moyens mis en oeuvre pour la sauvegarde impliquent des coûts qui ne seront jamais rentabilisés. Il faut aussi avouer que la sauvegarde, malgré son importance, n'est pas ce qu'il y a de plus attrayant dans le domaine de l'informatique. Il est logique qu'elle ne soit pas aussi développée et normalisée que d'autres éléments dans le même domaine.

De nos jours, beaucoup d'entreprises n'utilisent exclusivement plus que l'informatique comme moyen de communication et pour le sockage de documents. Des données critiques transitent donc sur le réseau et sont stockées dans des ordinateurs. Il est donc vital pour l'entreprise de s'assurer que ces données ne seront pas perdues ou altérées en toutes circonstances. C'est là que la sauvegarde, et tout ce que cela entraîne, prend toute son importance. A l'heure où le papier est de moins en moins utilisé, et ce au profit de l'informatique, il faudra veiller à ce que ces données soient protégées efficacement. La sauvegarde est une des réponse à cet enjeu. De plus la sauvegarde implique la restauration et l'élaboration de plans de reprise d'activité.

On s'attachera donc à comprendre la sauvegarde, d'un point de vue méthodologie, puis pragmatique, avec les besoins nécessaires, l'importance dans le choix des stratégies, l'optimisation de ces procédés, et les bonnes pratiques à adopter. Pour ce qui est de l‘'optimisation des sauvegardes et des restaurations, on fera le tour des technologies existantes, expliquant les principes de fonctionnement de la compression et de la déduplication, ainsi que leurs intérêts dans un contexte de sauvegarde. On verra aussi pourquoi elle doit évoluer de la même façon que l'informatique et les utilisations évoluent, comme le “ tout informatique “ ou encore l'augmentation d'ordinateurs portables dans les entreprises. Il s'agira donc de trouver des solutions adaptées ou des orientations à prendre. On verra aussi comment les techniques de sauvegarde doivent s'adapter à l'infrastucture informatique déjà existante. Etant un point critique

dans l'entreprise, et coûteux en terme d'investissement, il s'agira d'affiner au plus les stratégies de sauvegardes pour correspondre aux ressources et aux besoins.

Enfin, les données sauvegardées étant logiquement importantes, on s'attachera à protéger ces données aussi bien contre un évènement catastrophique, que contre un vol ou une exploitation de ces données. On verra où et comment stocker ces données, et les techniques qui permettront de ne les rendre exploitables que par les utilisateurs qui sont habilités à le faire, comme le chiffrement par exemple.

1.2 Problématique

Un nombre considérable de questions et de réflexions sont générées lors de l'élaboration et la mise en place de sauvegardes, aussi bien à un niveau méthodologique que pratique. Analyser ces différents aspects et pouvoir répondre aux questions permettront de trouver la solution ou la combinaison de solutions la plus adaptée en fonction du besoin.

1.2.1 Objectifs

L'aspect le plus important de la sauvegarde et le premier auquel on doit réfléchir est : quels sont les objectifs finaux de la sauvegarde ? De quoi veut-on se prémunir. Quels sont les scénarios catastrophe auquels la stratégie de sauvegarde doit répondre ? Autant de questions primordiales et qui seront spécifiques à chaque situation.

1.2.2 Contraintes

Au delà des objectifs, il s'agira d'élaborer les contraintes, qui jouent aussi un rôle déterminant dans l'élaboration d'une stratégie de sauvegarde. Si toutes les techniques de sauvegardes ne sont pas équivalentes en terme de sécurité et de performance, c'est aussi le cas au niveau des coûts. L'élaboration d'une stratégie de sauvegarde qui devra se calquer sur une infrastructure déjà existante est souvent un réel investissement, et n'est pas à négliger. Après les contraintes de coûts, il y a toutes les contraintes liées au contexte de sauvegardes, tel que la quantitée de données à sauvegarder, la fréquence des sauvegardes et la durée de rétention.

1.2.3 Optimisation

L'optimisation des sauvegardes est liées aux contraintes. En effet, comme les contraintes sont bien présentes et sont un réel obstacle, il conviendra d'optimiser au maximum ces sauvegardes, aussi bien en termes de coûts que de performances et d'espace disque (qui sont liés). Une partie sur l'optimisation sera développée et les technologies existantes seront présentées, tel que la compression et la déduplication.

1.2.4 Données

Cet aspect est lié aux objectifs et aux contraintes posés. Il faudra savoir finement quelles données doivent être sauvegardées. Sauvegarder des données inutiles engendrera un surcoût inutile, et ne pas sauvegarder des données importantes sera catastrophique lorsqu'on devra effectuer une restauration. De plus, les données à sauvegarder sont évidentes lorsqu'il s'agit des données des utilisateurs, mais il faudra faire attention aux données relatives au système qui pourraient être critiques. Une autre problématique liée au données et non négligeable est de pouvoir sauvegarder des données qui sont en cours d'utilisation et de modification.

1.2.5 Gestion

On doit aussi de poser la question sur la gestion de la stratégie de sauvegarde au quotidien, comme l'automatisation des tâches et la supervision. De plus en plus d'utilisateurs souhaitent aussi pouvoir effectuer eux même les opérations de sauvegarde et de restauration, sans avoir besoin d'avertir un administrateur. Des techniques permettant la délégation des tâches qui les concernent seront explorées.

1.2.6 Application

Il s'agira d'effectuer un comparatif sur les différentes solutions existantes permettant de faire des sauvegardes, de manière plus ou moins automatisée. En effet, un grand nombre de logiciels sont disponibles et avoir un comparatif permettrait de choisir la solution la plus adaptée.

Chapitre 2

Etat de l'art

2.1 Sauvegarde

Lors de l'élaboration d'une stratégie de sauvegarde, un très grand nombre de choix importants seront à faire, et cela à beaucoup de niveau. Un grand nombre de possibilités existent chacun offrant des avantages par rapport aux autres. Dans cette partie, nous allons explorer toutes les solutions existantes concernant les modèles d'entrepôts de données, les supports de sauvergarde et les gestions de dépots.

2.1.1 Modèles

Le modèle, c'est comment les données vont être organisées dans l'entrepôt de données, et donc de quelle manière les données seront sauvegardées. Il existe un nombre important de modèles, chacun offrant des avantages et des inconvénients.

Non structurée

Cette méthode consiste simplement à stocker des sauvegardes avec des informations sur la nature de ces sauvegardes et les dates auquelles elles ont été réalisées.

Images du système

Dans ce type de modèle, il s'agit de stocker une image complète du système à un instant donné

Différentielle

Une sauvegarde différentielle consiste à sauvegarder uniquement les données qui ont été modifiées depuis la dernière sauvegarde complète.

Incrémentale

Une sauvegarde incrémentale consiste à sauvegarder uniquement les données qui ont été modifiées depuis la dernière sauvegarde, quelle qu'elle soit.

Delta inversé

Dans ce type de modèle, on aura un mirroir des données à leur état au moment de leur dernière sauvegarde. On pourra revenir à un état antérieur grâce aux journaux de modifications sauvegardés.

Protection des données en continu (CDP)

Au lieu d'effectuer des sauvegardes périodiquement, les modifications sont envoyées instantanément vers l'espace de stockage dédié à la sauvegarde à travers le réseau.

2.1.2 Dispositifs de stockage

Le dispositif de stockage est le support matériel sur lequel les données sauvegardées seront stockées. Il doit être choisi avec reflexion, et est un élément central dans la stratégie de sauvegarde.

Bandes magnétiques

C'est le support le plus utilisé pour les sauvegardes en entreprise. C'est un bon compromis entre coûts, solidité et intégrité des données.

Disques Durs

De plus en plus utilisé, étant donné la baisse de prix de ce support. Le disque dur comme dispositif de stockage n'est pas aussi fiable en raison de leur solidité et du fait que l'on n'a aucune certitude sur la durée pendant laquelle l'intégrité des données est assurée.

Supports optiques

Les capacités de sotckages de ce support restent relativement faibles, et ne sont historiquement pas adaptés à une utilisation comme support de stockage, puisque la plupart sont de type WORM. Cette solution peut être envisageable pour l'archivage.

SSD

Bien que très interessants en raison de leur vitesse de lecture et d'écriture et leur solidité, ils ont deux inconvénients de taille : leurs coûts encore très élevés et le faible capacité de stockage.

Services de sauvegardes à distance

C'est le dispositif à la mode ces dernières années. En sauvegardant ces données via internet et en les confiant à une autre personne, cela permettra de se prémunir contre bien des scénarios. En revanche, cela implique de faire confiance à cette personne, même si un chiffrement des données permet de remédier à ce problème. De plus cela exige un débit montant très élevé, ce qui est encore rare, et qui limite la taille des données à sauvegarder.

2.1.3 Gestion du dépôt

Ici on définira comment les données seront accessibles. Il faudra choisir une méthode qui correspond aux besoins de sauvegardes en termes d'accessibilité, de sécurité et de coûts.

En ligne

Le support de stockage est en permanence connecté. Cela lui donne l'avantage d'être tout le temps accessible.

Intermédiaire

Moins chers, mais aussi moins accessibles que les systèmes en lignes, ces solutions dites intermédiaires semblent un bon compris encore accessibilité, sécurité des données et coûts.

Déconnecté

Sauf lorsque les sauvegardes et les restaurations sont réalisées, les sauvegardes sont complètement isolées du point de vue informatique et nécessite une intervention manuelle.

Externalisation

Les sauvegarde ou une partie des sauvegardes seront stockées hors-site pour se prémunir contre une catastrophe qui pourrait entraîner la destruction des sauvegardes si elles étaient stockées sur site.

Site de sauvegarde

Il s'agit de réaliser un autre site opérationnel et autonome qui sera alimenté avec les sauvegardes du site principal. Si le site principal venait a être indisponible, c'est ce site qui prendrait le relais.

2.1.4 Données

Fichiers

La plupart des informations est organisée en fichiers. Il est logique que l'on se base sur les fichiers lors d'une sauvegarde. Toutefois, il peut être judicieux de se baser sur les blocs du fichier. Ainsi lors d'une sauvegarde incrémentale ou différentielle, uniquement les parties du fichier qui ont été modifiées seront sauvegardées.

Systèmes de fichiers

Il peut-être utile de sauvegarder des éléments relatifs au système de fichiers. Il s'agira de faire des sauvegardes complètes, qui consiste à sauvegarder le système dans son ensemble. On peut aussi penser aux identificateurs de changements qui permettront de savoir si des changements ont été opérés sur le fichier depuis la dernière sauvegarde. Le versionnage du système de fichier est aussi utile, car il permettra à un utilisateur de récupérer simplement d'anciennes versions automatiquement enregistrées localement.

2.1.5 Données en temps réel

Lorsque la sauvegarde est effectuée à chaud, c'est à dire que la sauvegarde a lieu alors que le système est utilisé normalement, il se peut que la sauvegarde de fichiers ouverts et en cours de modification pose des problèmes. En effet, comme c'est généralement le cas lors de sauvegardes de bases de données, les données ne reflètent pas leur état à un instant t puisqu'elles ont été modifiées entre le début et la fin de la sauvegarde. Si des données sont liées, elles peuvent ne plus être cohérentes entre elles, et donc pas exploitables.

2.1.6 Objectifs

On retiendra trois objectifs majeurs dans la sauvegarde qui nous permettront d'élaborer la stratégie de sauvegarde en conséquences.

Points de restauration

Cela correspondra au point dans le temps auquel le système sera restauré.

Temps de restauration

Lors d'un désastre, la durée nécessaire à la restauration sera un élément capital, car il définira le temps pendant lequel le système sera indisponible.

Protection et intégrité des données

La protection et l'intégrité des données est sont des éléments majeurs dans les stratégies de sauvegardes. Elles seront définies par les choix qui on été faits en matière de dispositifs de stockages et de gestions des dépôts, ainsi que le chiffrement des données.

2.1.7 Limites

Pour la création d'un schéma de sauvegarde, il faudra prendre en compte les limites de l'environnement.

2.1.8 Fenêtre de sauvegarde

On appelle fenetre de sauvegarde la durée pendant laquelle la sauvegarde est possible, en relation avec l'utilisation plus ou moins importante du système. Bien évidemment, il faudra prendre en compte cette durée dans la mise en place du schéma de sauvegarde.

2.1.9 Performances

Les sauvegardes sont souvent gourmandes en CPU, et ont donc un impact sur les performances du système.

2.1.10 Investissements

Les stratégies de sauvegardes necessitent plus ou moins d'espace disque, ce qui aura un certain coût au final.

2.1.11 Limites Réseau

Les limitations du réseau par lequel transiteront les données dans le cas de systèmes de sauvegarde ditribués sont autant de contraintes à ne pas négliger.

2.2 Optimisation

2.2.1 Déduplication

La déduplication de données est une technique qui permet de minimiser de l'espace de stockage. Elle consiste à ne pas répliquer les données déja existantes sur le disque. Un fichier est décomposé sous forme de blocs de données car des fichiers peuvent avoir des blocs en commum. Le mécanisme de déduplication crée une table avec les index de tous les blocs de données des fichiers présents sur le disque. La taille des blocs peut varier selon les mécanismes utilisés mais plus les blocs sont petits, plus il y aura de chance qu'un autre bloc soit identique et donc, plus la déduplication sera efficace. En général, cette taille ne dépasse pas les 128ko.

Quand un utilisateur dépose un fichier, le mécanisme crée ses index et regarde s'il n'y a pas des blocs déjà existants. Si des blocs sont similaires alors une simple référence aux blocs déjà existants sera crée. Le schéma ci-dessous montre comment la déduplication fonctionne. Les blocs étant de la même couleur sont considérés identiques.

Il existe deux types de dépuplication : la déduplication à la volée (à la source) et la déduplication hors ligne (à la destination). La déduplication à la volée analyse les fichiers avant de les stocker pour savoir s'ils n'existent pas déjà sur le disque. Cette technique utilise une forte consommation CPU et mémoire. L'autre technique consiste à copier dans un premier temps le fichier sur le disque avant de tester s'il existe déjà. Cela nécessite de prévoir un espace de stockage tampon plus important.

Dans un contexte de serveur de messagerie et de fichiers centralisés, la déduplication de données peut très rapidement économiser de nombreux gigaoctets d'espace disque ainsi que la diminution de la bande passante qui aurait été utilisée pour la sauvegarde. En effet, dans le cas où un même mail de 1Mo est envoyé à cinquante destinataires alors l'économie du disque sera de 50-1 megaoctets (stockage d'un seul mail). La déduplication est faite pour des fichiers tels que des documents bureautiques ou des machines virtuelles qui ont souvent de nombreux blocs en commun.
Le terme inverse de la déduplication est la réhydratation. Elle fait appel à la table des index afin de

renvoyer tous les blocs de données référencés pour un fichier demandé.

Certain outils comme LessFS mise en relation avec un système de fichiers ZFS permettent de dédupliquer et de compresser les blocs de données. Cela permet de gagner encore plus d'octets sur le disque mais nécessite une consommation mémoire et CPU plus importante.

2.2.2 ZFS

Introduction

Le système de fichier ZFS (Zettabyte File System) a été conçu par Sun en 2005 et est sous licence CDDL. Il n'était disponible que sous Solaris mais est devenu récemment disponible sous linux. Il est l'un des systèmes de fichiers les plus intéressants du marché. En effet, ZFS intègre de nombreux avantages que d'autres n'ont pas. Voici une liste de ses principaux avantages :

- Pas de limites pratiques (taille des disques, fichiers, ...)
- Garantir la sécurité des données (intégrité, disponibilité)
- Administration simplifiée
- Gestionnaire de volume intégré
- Compression
- Snapshot
- Duplication
- Quotas et réservation d'espace
- Performances élevées
- Indépendant de l'architecture matérielle

ZFS est un système de fichier 128 bits contrairement aux autres systèmes qui sont de 64 bits. Ainsi ses limites sont de 16 milliards de milliards fois plus autant dire qu'il n'a quasi pas de limite. Afin d'optimiser ses performances, ZFS utilise tout l'espace disponible de la RAM pour créer un énorme cache. Ce procédé s'appelle ARC (Adaptive replacement cache). Il peut poser problème aux autres processus qui testent la mémoire inutilisée avant de ce lancer mais cette mémoire est souvent inutilisée. Il peut être partagé via le réseau avec d'autres systèmes de fichiers comme nfs ou samba. Ainsi même depuis des systèmes qui ne le supporte pas, il sera accessible.

Stockage

ZFS fonctionne avec un pool. C'est un ensemble de périphériques qui fournissent de l'espace pour le stockage et la duplication des données comme le raid logiciel.
Voici les différentes unités de base de stockage de données :

- Disques : entiers ou juste une partition
- Fichiers dans un autre système de fichiers
- Miroirs : 2 (ou plus) disques, partitions ou fichiers
- Raid-z : plusieurs disques, variante de RAID-5

ZMirror est un miroir classique. Il utilise les mécanismes de checksum pour valider les lectures sur un composant et bascule sur le second s'il détecte une erreur puis corrige le composant défaillant (si possible). Le système Raid-z est similaire au procédé Raid 5. Il utilise les checksums (SHA-256 + fletcher) et repose sur le copy on write : supprime le "write-hole".

2.2.3 Compression

Tout comme la déduplication, la compression est une technique qui permet d'économiser de l'espace de stockage. Chaque fichier est constitué d'une succession de millions de bits 0 ou 1. La compression

permet de diminuer le nombre de bits que constitue un fichier en changeant la succession de bits de départ. Suivant l'algorithme de codage utilisé, le taux de compression peut différer. Les algorithmes d'encodage sont plus ou moins efficaces selon le type de fichier compressé.
Il existe deux types de compression : la compression avec perte et sans perte. La compression sans perte signifie qu'après la décompression, le fichier sera identique au fichier compressé. C'est le plus souvent utilisé sur des documents, des fichiers exécutables ou des archives. Ces données étant principalement des caractères texte, ils ne peuvent pas être modifiés. Les formats de documentation tels que txt, doc ou pdf sont donc compressés sans perte. Tant qu'à la compression avec perte, les fichiers décompressés ne seront pas exactement identiques au fichier original mais les informations seront sensiblement les mêmes. Les types de fichiers utilisés par cette compression sont les images, les sons et les vidéos. Cett technique se repose sur la limitation des sens de l'homme comme la vision et l'audition. L'homme ne pourra donc pas identifier les différences entre le fichier original et le fichier après décompressage. Les formats de fichiers jpeg, avi ou mp3 sont donc compressés avec pertes.
Pour chaque technique de compression, il existe plusieurs algorithmes de codage.

Compression sans perte

Parmi les algorithmes sans perte, il y a les algorithmes tels que Lempei-Ziv ou le codage RLE (Run-Length Encoding) qui consistent à remplacer des suites de bits utilisées plusieurs fois dans un même fichier. D'autres algorithmes comme l'algorithme de codage Huffman détermine les suites de bits et plus une suite est utilisée souvent, plus la suite qui la replacera sera courte. Voici les différents types de compression sans perte :

- L'algorithme Lempel-Ziv
  - LZ77
  - LZ78
  - LZO
- L'algorithme RLE
- Codage par modélisation de contexte
  - Prédiction par reconnaissance partielle (PPM)
  - Pondération de contextes (CM)
- L'algorithme de codage Huffman

Compression avec pertes

La compression avec pertes s'utilisent donc sur des données perceptibles par l'homme comme les sons, les images ou les vidéos. Elles suppriment les données que l'homme ne perçoit pas ou quasiment pas. Ainsi pour le format JPEG 2000, la compression est de 1 bit/pixels au lieu de 24 bits/pixels. La compression avec pertes est une technique irréversible c'est à dire qu'il ne sera pas possible de retrouver le fichier original. Il existe trois grandes familles de compression avec pertes : la compression par prédiction, par transformation et la compression basée sur les récurrences fractales de motif.
Voici les différents types de compression avec pertes :

- Compression par prédiction
- Compression par transformation
- La norme JPEG
- Compression par ondelette
- Compression basée sur les récurrences fractales de motif

Chapitre 3

Conclusion

Après avoir effectuer un nombre considérable de recherches bibliographiques sur la sauvegarde en général, les techniques d'optimisation de ces programmes, et les logiciels permettant la mise en oeuvre de sauvegardes, il s'agira maintenant de tester les différents logiciels de sauvegarde. Pour cela, il faudra mettre en place une infrastructure qui nous permettra de réaliser ces tests sur une échelle réelle.

La plateforme de tests devra nous permettre d'effectuer différents tests, et ce dans différentes situations d'utilisation. Les différents scénarios d'utilisation qui sont envisagés seront une base de données, un serveur mail, un serveur de fichiers, et des machines virtuelles.

Différentes méthodes de sauvegardes seront aussi envisagées, tel que des sauvegardes complètes, des sauvegardes continues, des snapshots, et des sauvegardes hors-site.

Nous réfléchirons aussi à l'impact de ces sauvegardes dans le cas de différents incidents. En effet, il faudra analyser et comprendre l'efficacité de ces sauvegardes dans le cas d'un fichier malencontreusement effacé ou corrompu, d'un système hors service et dont les données sont irrécupérables, ou encore une catastrophe majeure qui impacterait l'ensemble du site tel qu'un feu ou une innondation.

Il serait interessant de se pencher sur la délégation de tâches pour les utilisateurs, aussi bien au niveau de la sauvegarde que de la restauration. En effet, les utilisateurs ont tendance à vouloir effectuer eux-même ces opérations, surtout quand il s'agit de mauvaises manipulations de leur part.

Comme one l'a vu, la sauvegarde en elle-même n'est pas une solution complète mais fait partie d'un mécanisme permettant d'assurer l'intégrité des données. Aussi, on s'interressera aux concepts qui font partie de ce mécanisme, tel que les plans de reprise d'activité. L'élaboration de ces plans est tout aussi importante que la création d'une stratégie de sauvegarde, car elle indiquera clairement la marche à suivre en fonction des différents évènements, et permettront de gagner un temps précieux.

On s'interressera aussi plus longuement sur les types de sauvegardes à adopter dans un contexte de machines virtuelles. En effet, avec l'augmentation croissante des machines virtuelles, réaliser la sauvegarde de celles-ci alors qu'elles sont en cours d'utilisation peut devenir un vrai casse-tête.