

# Sauvegarde

Clément Ghnassia

12 février 2012

# Chapitre 1

## Introduction

En Informatique, une sauvegarde consiste à faire des copies de données qui pourront être utilisées après un événement de pertes de données.

La sauvegarde a deux buts :

- La restauration de données après que celles ci aient été perdues ou endommagées. En effet, la perte de données est courant pour les utilisateurs. En effet 67% d'entre eux ont déjà eu ce type d'expérience.
- La restauration de données plus anciennes. Ici, la sauvegarde peut s'apparenter à de l'archivage.

Le fait d'avoir une bonne stratégie de sauvegarde n'est qu'un élément parmi tant d'autres dans un plan de restauration après un désastre. En effet les administrateurs ont tendance à penser que le fait de faire des sauvegarde régulière est suffisant. Il n'en est rien. Il faut penser plein d'autres situations comme la perte totale du système ou des applications, la destruction d'une base de données, ... pour envisager une reprise d'activité rapide.

Un système de sauvegarde doit prendre en compte la taille des données qu'il faut sauvegarder. En effet, même si cela dépend du type de données et de la compression plus ou moins efficace, il faut comprendre qu'une sauvegarde du système contiendra au moins une copie de toutes les données. Les capacités de stockage qui en découlent sont donc considérables, et un élément majeur de la choix d'une stratégie de sauvegarde. Il existe plusieurs modèles pour la sauvegarde de données, chacun étant adapté aux différentes situations, comme la quantité de données à sauvegarder, la capacité de stockage pour les sauvegardes, la redondance des données dans la sauvegarde, ou encore la rapidité de la restauration.

Avant que les données soient réellement sauvegardées, elles sont sélectionnées, extraites et manipulées. Lorsque les données sont importantes, cela peut vite devenir fastidieux. Des techniques ont donc été mises au point pour optimiser ces procédures. Cela peut aller de la résolution de problèmes lors de sauvegardes à chaud, de la compression, de l'encryption ou de la déduplication des données. Il ne faudra pas non plus négliger les facteurs humains dans les processus de sauvegardes et de restauration, qui tendent tout de même à être de plus en plus automatisés (pour les sauvegardes du moins).

# Chapitre 2

## Le stockage

### 2.1 Les modèles d'entrepôts de données

La première étape dans l'élaboration d'une stratégie de sauvegarde est de trouver le ou les modèles qui correspondront le mieux avec la situation, les contraintes et les objectifs recherchés. Le choix d'un modèle est primordial et doit être mûrement réfléchi. Il définira comment les données seront stockées et organisées. Cela aura donc un impact important en terme d'espace disque nécessaire, de sécurité, ou encore de vitesse et facilité lors d'une restauration.

#### 2.1.1 Non structurée

Même si cela semble la méthode la plus simple à mettre en oeuvre, elle n'est pas très efficace lors d'une perte de données et n'offre pas un très haut niveau en terme de restauration de données. De plus, elle nécessite une organisation méthodique de l'administrateur. Cette méthode consiste simplement à stocker des sauvegardes avec des informations sur la nature de ces sauvegardes et les dates où elles ont été réalisées.

#### 2.1.2 Images du système

Dans ce type de modèle, il s'agit de stocker une image complète du système à un instant donné. Une des applications courantes consiste à faire une image afin de copier à l'identique la configuration sur d'autres machines. Pas très performant, surtout en ce qui concerne l'optimisation de l'espace disque, il est de moins en moins utilisé pour la sauvegarde.

#### 2.1.3 Différentielle

Cette méthode consiste à réaliser une sauvegarde complète des données souhaitées, puis d'enregistrer à interval régulier toutes les différences depuis la dernière sauvegarde complète. Cela permet de ne pas dupliquer les données qui n'ont pas changées et offre donc un gain d'espace considérable, dépendant bien évidemment du nombre de données qui peuvent changer d'une sauvegarde à l'autre. Pour restaurer ces données, on aura besoin de la dernière sauvegarde complète et éventuellement la dernière sauvegarde avant la date que l'on souhaite pour la restauration.

#### 2.1.4 Incrémentale

Cette méthode repose sur le même principe qu'une sauvegarde de type différentiel, à la différence que lors d'une sauvegarde incrémentale, seuls les données modifiées depuis la dernière sauvegarde, complète ou non, seront enregistrées. L'avantage par rapport à une sauvegarde de type différentielle est un gain en espace disque, car il y'aura moins de redondance. En revanche, la restauration s'avèrera plus longue et plus fastidieuse étant donné qu'on aura besoin de rejouer toutes les sauvegardes depuis la dernière sauvegarde complète et éventuellement jusqu'à la dernière sauvegarde à la date désirée.

### 2.1.5 Delta inversé

Ici, le fonctionnement est inversé. En effet, plutôt que de faire une sauvegarde complète et d'enregistrer les modifications par la suite comme c'est le cas avec les sauvegarde de type différentiel et incrémental, On aura toujours une sauvegarde complète (un miroir) correspondant à l'état des données à l'instant de la dernière sauvegarde effectuée. Le fonctionnement est trivial : après avoir effectué une sauvegarde complète, chaque sauvegarde inversée aura pour effet une synchronisation du miroir avec l'état réel des données. En même temps, les modifications qui permettront de revenir à un état antérieur seront enregistrées. Ainsi, pour restaurer les données au dernier état, il suffira juste de restaurer le miroir, et une restauration à un instant antérieur impliquera de rejouer les sauvegardes de manière décroissante dans le temps depuis le miroir jusqu'au moment souhaité.

### 2.1.6 Protection des données en continu (CDP)

Aussi appelé sauvegarde en temps réel, c'est la méthode la plus sûre en terme de protection des données. Au lieu d'effectuer des sauvegarde périodiquement, les modifications sont envoyées instantanément vers l'espace de stockage dédié à la sauvegarde à travers le réseau. Bien que faisant penser à de la réplication (pour de la haute disponibilité par exemple), cela est complètement différent car ce n'est pas une copie parfaite des données qui est réalisée, mais un système de logs et de journaux qui enregistrent les modifications et qui permettront de revenir à un état antérieur pour une éventuelle restauration.

## 2.2 Dispositifs de stockage

Aujourd'hui, les périphériques pour le stockage de données sont nombreuses et variées. Il s'agira ici d'analyser chacun de ces dispositifs et d'en comprendre les avantages et les inconvénients, en particulier dans le cas d'une utilisation pour la sauvegarde.

### 2.2.1 Bandes Magnétiques

Aujourd'hui encore, c'est sans doute le support le plus utilisé pour le stockage de données, et plus particulièrement pour la sauvegarde et l'archivage. La raison est tout d'abord financière. En effet, même si la tendance est à une réduction des différences de prix entre les bandes magnétiques et les disques durs, c'est encore actuellement la solution la moins chère en terme de capacité de stockage. Il faut noter que les bandes magnétiques correspondent bien aux contraintes de sauvegarde car elles fournissent un accès séquentiel, ce qui a un impact bénéfique sur les vitesses de lecture et d'écritures. De plus, on sait que les bandes magnétiques, si elles sont correctement conservées, assureront l'intégrité de leur contenu sur plusieurs décennies. Enfin, les bandes magnétiques sont utilisées depuis un certain temps, et leurs caractéristiques sont donc totalement maîtrisées. En revanche, le problème survient au niveau des formats. En effet, la plupart du temps les formats de bandes magnétiques sont nombreux et souvent propriétaires. Utiliser des bandes magnétiques engendrera donc une dépendance forte vis à vis des constructeurs, et ajoute des contraintes en cas de changement de solution. De plus, pour les grosses structures, l'investissement sera d'autant plus important que l'achat d'un automate sera nécessaire.

### 2.2.2 Disques durs

Ce moyen est de plus en plus utilisé pour la sauvegarde et l'archivage, mais dans bien des cas, les solutions à base de bandes magnétiques lui sont toujours préférées, et ce à juste titre. A première vue, l'utilisation de disques durs pour la sauvegarde et l'archivage de données semblerait la meilleure solution. En effet, le prix des disques durs se rapproche des bandes magnétiques à capacité équivalente. De plus, les disques durs ont des avantages certains par rapport aux bandes magnétiques en terme de temps d'accès, de disponibilité, de capacité, et de simplicité d'utilisation. Ils ne nécessitent pas d'équipement spécifique, et permettent une utilisation dans des contextes différents, s'adaptant plus aux besoins, comme par exemple leur emplacement, aussi bien au niveau local que sur le réseau. On pourra aussi ajouter des procédés permettant de réduire les tailles des sauvegardes, comme des

techniques de de-duplication. Mais l'utilisation de ce type de support dans un contexte de sauvegarde, et surtout d'archivage possède un inconvénient de taille. Tout d'abord la fragilité du support n'est pas adapté à ce type d'utilisation et n'assure donc pas l'intégrité des données stockées. En effet là où la manipulation et l'externalisation de ces données est courante, la fragilité des disques durs est donc critique. De plus, il n'a pas été établi sur combien de temps les données pourraient être conservées sur le long terme sans risque d'altération de celles-ci. Si cette contrainte n'est pas insurmontable pour des sauvegardes où la conservation de données en vue d'une restauration n'auraient pas beaucoup de sens, elle l'est en revanche beaucoup plus pour l'archivage, qui nécessite souvent une conservation des données à long terme.

### **2.2.3 Supports optiques**

Si ces supports sont courants et très utilisés par les particuliers pour stocker des données, de par son prix et sa simplicité d'utilisation, il ne semble pas que ce soit un support adapté dans le cas de sauvegarde et d'archivage de quantités importantes de données, comme c'est souvent le cas dans une entreprise. En effet ces supports sont généralement d'une faible capacité de stockage et sont souvent de type WORM (écriture mais pas effacement). Toutefois, il existe des solutions qui peuvent correspondre à des petites structures ne nécessitant pas énormément de capacité de stockage. Des systèmes de chargeurs permettent une automatisation, ce qui aura pour effet d'éviter le contact des disques avec l'être humain, souvent facteur négatif dans l'intégrité des données.

### **2.2.4 Disquette**

Les disquettes, très à la mode dans les années 1980, n'est évoqué ici que pour des raisons historiques. Très limité en terme de capacité de stockage, ce support est aujourd'hui totalement obsolète.

### **2.2.5 SSD**

Ce type de support, aussi appelé support à base de mémoire flash ont pour particularité d'être extrêmement rapide, aussi bien en lecture, qu'en écriture. Ces éléments ne possèdent pas le défaut de fragilité des disques durs. En revanche le ratio prix /capacité est extrêmement élevé, et les capacités sont plus limitées que les disques durs classiques, ce qui rend leur utilisation comme support pour la sauvegarde ou l'archivage impossible.

### **2.2.6 Services de sauvegardes à distance**

Avec la popularisation des accès internet à haut débit, c'est une méthode en vogue et qui se développe de plus en plus. Ces services permettent d'effectuer des sauvegardes en ligne, ce qui semble être la solution la plus fiable en terme d'intégrité des données, puisque ces sauvegardes peuvent être stockées n'importe où dans le monde. En revanche, cela oblige de faire confiance à un tiers puisqu'il aura accès à toutes les données. Même en chiffrant ces données avant de les envoyer, il faudra que ce service soit sûr. En effet les conséquences peuvent être dramatiques si un problème survient chez le tiers alors qu'on doit effectuer une restauration. De plus, les particuliers ont en général un faible débit montant, ce qui est contraignant dans le cas de sauvegarde sur une grande quantité de données.

## **2.3 Gestion du dépôt**

Il existe aussi plusieurs types de gestion de dépôts. En effet il faudra choisir une méthode qui corresponde aux besoins de sauvegardes en termes d'accessibilité, de sécurité et de coût. Ces méthodes ne sont pas exclusives et pourront être combinées pour arriver à une gestion des sauvegardes et de leurs dépôts plus fine et plus élaborée.

### **2.3.1 En ligne**

Ce type de stockage est en permanence connecté. Cela lui donne l'avantage d'être le plus accessible et permet d'effectuer une restauration avec le moins de manipulation au préalable. Ce système de stockage peut se mettre en place à l'aide de disques internes, ou comme c'est le cas dans les grandes structures, à l'aide de baies SAN. En plus d'être couteuse, la gestion des dépôts en ligne induit des problèmes de sécurité. En effet, les données étant accessibles en permanence, elles peuvent facilement être altérées, de manière intentionnelle, ou suite à une erreur de manipulation.

### **2.3.2 Intermédiaire**

Moins chers, mais aussi moins accessibles que les systèmes en lignes, ces solutions dites intermédiaires semblent un bon compromis encore accessibilité, sécurité des données et coûts. Cela permet aussi de rester assez réactif en cas de restauration. Un bon exemple serait l'utilisation d'une bibliothèque de bandes magnétiques automatisée, qui tout en offrant une solution déconnectée, permet d'effectuer une restauration rapidement.

### **2.3.3 Déconnecté**

Sauf lorsque les sauvegardes et les restaurations sont réalisées, les sauvegardes sont complètement isolé du point de vu informatique et nécessite une intervention manuelle de la part du responsable. Dans le cas de bandes magnétiques, il s'agira d'une insertion dans un lecteur. Dans le cas d'un disque dur, il s'agira de le connecter à un ordinateur par exemple. Cette solution est donc orientée pour la sécurité des données et l'accessibilité dépendra de l'emplacement de stockage (typiquement sur site ou hors site)

### **2.3.4 Externalisation**

Le 11 Septembre l'a montré, le fait d'avoir des sauvegardes n'empêche pas la perte irrémédiable de données. Pour se prémunir contre des désastres de cette ampleur tel que des feux ou des tremblements de terre, il faudra opter pour un stockage d'une partie ou de toutes les sauvegardes hors site. Ces endroits où seront stockées ces données peuvent être plus ou moins évolués, d'un simple bureau chez l'administrateur, jusqu'aux chambres fortes. Dans ce cas, la mise en place d'une restauration peut s'avérer plus longue, et pose donc des problèmes d'accessibilités.

### **2.3.5 Site de sauvegarde**

Si lors d'un évènement particulièrement désastreux, les équipements informatiques sont hors d'état de fonctionner, le fait d'avoir des sauvegardes à jour ne permettra pas une restauration, et l'indisponibilité du système sera alors inévitable. Dans le cas de cette éventualité, une solution particulièrement couteuse consiste de fournir à un autre site des équipements informatiques nécessaires pour prendre le relais. Ainsi, en alimentant le second site avec des sauvegardes à jour provenant du premier site, celui-ci pourra fonctionner indépendamment et de manière autonome, jusqu'à un rétablissement.

# Chapitre 3

## Les données

Lors de la conception de stratégies de sauvegardes, une étape importante consiste à établir quelles seront les données qui sont critiques et qui auront besoin d'être sauvegardées. Il faudra bien évidemment considérer les fichiers importants pour l'utilisateur, mais aussi ceux qui sont nécessaires au bon fonctionnement de l'ordinateur. Cette étape lors de la conception de stratégies de sauvegardes est vitale. En effet, choisir de sauvegarder un nombre trop important de fichiers aurait pour conséquences un remplissage de l'espace de stockage et empêcherait l'application de la stratégie de sauvegardes à court terme. Au contraire, ne pas sauvegarder assez de données aurait, comme on peut s'en douter, des conséquences désastreuses lors d'une perte de données.

### 3.1 Fichiers

Aujourd'hui, toute donnée informatique est structurée en fichiers.

#### 3.1.1 Sauvegarde de fichiers

C'est évidemment ce qui sera copié dans le cas de sauvegardes classiques.

#### 3.1.2 Sauvegarde de parties de fichiers

Lorsqu'on modifie un fichier, une sauvegarde de type différentielle ou incrémentale aura pour effet d'inclure toute le fichier dans la nouvelle sauvegarde, même si celui-ci n'a été que peu modifié. Afin de résoudre ce problème de redondance de données, une solution consiste à sauvegarder uniquement les blocs qui ont été modifiés plutôt que tout le fichier. Les procédés de compression et de dé-duplication sont une solution à ce problème.

### 3.2 Systèmes de fichiers

Il peut être judicieux dans certains cas de sauvegarder des éléments relatifs au système de fichiers.

#### 3.2.1 Sauvegarde complète

Plutôt que de copier l'ensemble des fichiers contenus dans le système, une solution consiste à sauvegarder le système de fichiers dans son ensemble. Cette technique s'apparente à la réalisation d'images disque. En copiant tout le système de fichiers à l'aide d'outils de bas niveau (comme dd), ce type de sauvegardes est souvent plus rapide, surtout quand les données sont composées de petits fichiers fragmentés. En revanche, la réalisation de ce type de sauvegardes ne peut pas généralement se faire à chaud, c'est à dire que le système de fichier devra être démonté durant toute la durée de la sauvegarde. De plus, les avantages de ce type de sauvegarde sur un système de fichier peut remplir seront vite effacés car un outil de bas niveau copiera tous les blocs du système de fichiers, sans faire la distinction entre des blocs contenant des données et des blocs non indexés.

### **3.2.2 Identificateurs de changement**

Certains systèmes de fichiers évolués permettent à chaque fichier d'avoir un bit d'archive qui permettra de savoir si des changements ont été opérés sur le fichier depuis la dernière sauvegarde.

### **3.2.3 Versionnage du système de fichiers**

Lorsque elle est mise en place sur un système de fichiers, une version de chaque fichier modifié est automatiquement conservé. Ainsi, l'utilisateur qui le souhaite peut restaurer un fichier à un instant donné, sans avoir à effectuer aucune manipulation au préalable.

## **3.3 Données en temps réel**

Lorsque la sauvegarde est effectuée à chaud, c'est à dire 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. On a donc un problème d'intégrité.

### **3.3.1 Sauvegarde par snapshots**

Disponible sur certains systèmes de stockage, cette fonction permet de faire une copie du système de fichier à un instant donné. Lors de la prise d'un snapshot, les activités du système sont très brièvement gelées, et on a donc plus de risque d'incohérence des données.

### **3.3.2 Sauvegarde de fichier ouvert**

Certains logiciels permettent la sauvegarde de fichiers ouverts simplement en vérifiant si les fichiers sont utilisés, ou encore à l'aide de verrous.

### **3.3.3 Sauvegarde à froid**

Ce terme est généralement utilisé pour les bases de données. Lors de ce type de sauvegardes, la base de données devient indisponible le temps de la sauvegarde, et ceci pour éviter toute modification de fichier qui risquerait d'entraîner des incohérences.

### **3.3.4 Sauvegarde à chaud**

Cette fois, ce type de sauvegarde, généralement sur des bases de données, permet d'effectuer des sauvegardes sans pour autant rendre la base indisponible. Un image des données est d'abord réalisée à chaud, pendant que les modifications sont stockées dans des journaux. A la fin de la réalisation de l'image initiale, les journaux sont rejoués pour avoir une copie des données conforme à la date de restauration.

## **3.4 Méta-données**

Au-delà des fichiers, il faudra aussi penser à sauvegarder des informations relatives au systèmes

### **3.4.1 Secteur de boot**

Indispensable pour le démarrage du système, il peut être utile dans certains cas de le sauvegarder.

### **3.4.2 Schémas de partitions**

La aussi, il peut être utile sauvegarder ces informations qui pourraient être utiles lors d'une restauration rapide du système.



### **3.4.3 Méta-données des fichiers**

Ces méta-données stockent, entre autre, les permission, les ACL et les propriétaires de chaque fichier.

### **3.4.4 Méta-données du système**

Ce sont des informations nécessaires pour le système.

## **3.5 Optimisation**

Beaucoup de méthodes d'optimisation pour la sauvegarde et l'archivage sont utilisées, et ont pour rôle d'améliorer les performances de ce processus, comme par exemple les vitesses de sauvegardes, les vitesses de restauration, ou encore l'intégrité et la sécurité des données

### **3.5.1 Compression**

Ce mécanisme permet d'influer sur l'espace qu'occupera les sauvegardes sur l'espace de stockage. En effet, il existe de nombreux algorithmes et de procédés plus ou moins efficaces et dépendant du type de données ciblées qui permettront d'alléger les données en taille et donc d'optimiser l'espace disque.

### **3.5.2 Dé-duplication**

Tandis que la compression n'est ciblé que sur un fichier, voir un groupe de fichiers, la dé-duplication est elle gérée par le support de stockage lui-même. En analysant les blocs déjà présents sur le support, il optimisera l'espace de stockage en préférant indexer ces blocs plutôt que de les réécrire, ce qui aura pour conséquences une redondance nulle et une optimisation de l'espace disque maximale.

### **3.5.3 Duplication**

Les stratégies de sauvegardes ont parfois recourt à de la duplication afin de manipuler et d'optimiser ces données dans le cas d'une éventuelle restauration, ou tout simplement pour stocker les données dans deux endroits différents.

### **3.5.4 Encryption**

## Résumé