Chapitre 5
Technologies de protection de données

1. Introduction

1.1 Un peu d'histoire

Il y a une vingtaine d'années, le titre de ce chapitre et même de cet ouvrage aurait probablement été simplement « Meilleures pratiques en matière de sauvegarde des données ». Un tel titre n'a plus de sens aujourd'hui en raison des évolutions successives des technologies de protection de données au fil des différentes ères informatiques.

La nécessité de conserver une copie des données s'est rapidement fait sentir dès l'avènement des premiers ordinateurs. Avant l'an 2000 qui marque le début de la bulle financière d'Internet et des premières start-up, aucune solution de sauvegarde à proprement parler et digne de ce nom existe alors.

La protection de la donnée ou de tout fichier se limite à copier celui-ci sur un autre support de stockage, par exemple la disquette informatique initialement utilisée. Une fois la copie des fichiers terminée, on l'extrayait de l'ordinateur, on la rangeait alors dans une boîte prévue à cet effet, pourvue d'une petite clé, disposée sur le bureau ou au mieux dans l'armoire adjacente.

Les risques afférents à la cybercriminalité n'étaient à l'époque pas d'actualité. L'administrateur ou l'opérateur en question devait se charger d'identifier manuellement au stylo le support de stockage et surtout se souvenir du lieu de stockage de la copie. Les informations portées sur le support se limitaient au strict minimum. Néanmoins, les indications manuscrites devaient être suffisamment précises pour retrouver rapidement la copie de la donnée au besoin. Les systèmes étaient fortement centralisés et chaque ordinateur disposait de son propre disque dur et de son propre lecteur de disquettes ou au mieux lecteur de bande magnétique.

L'avènement du réseau informatique permet par la suite d'interconnecter les ordinateurs entre eux. Il simplifie le transfert des données sans pour autant changer réellement la problématique de sauvegarde des données. Les copies sont orchestrées de manière quotidienne. La sauvegarde manuelle ou non se fait au mieux en fin de journée, sinon quand on y pense.

1.2 Une période sombre pour les données

Cette période de l'Âge sombre du numérique est néanmoins marquée par la perte de bon nombre de données, faute de les avoir sauvegardées dans les règles de l'art. On peut notamment citer le manque de viabilité et de fiabilité des premiers supports informatiques, très vulnérables à l'environnement extérieur (variation de températures, choc physique, champ électromagnétique…), l'absence de normes ou de standards au niveau des systèmes informatiques, l'utilisation de formats de données propriétaires ou devenus obsolètes...

Les conséquences de ces pratiques sont parfois catastrophiques. En 1975, la NASA lance le programme Viking, dont l'objectif est de faire atterrir sur Mars les premiers engins spatiaux américains. Au terme de la mission, il faudra quelque dix années avant de pouvoir procéder à l'analyse des bandes magnétiques contenant les enregistrements de la phase d'atterrissage des engins sur le sol martien.

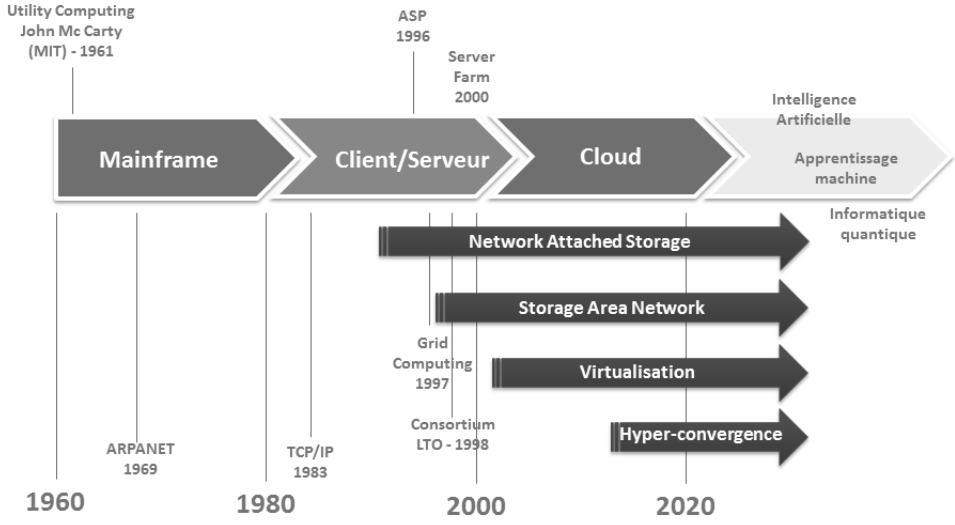
Un problème de taille se pose alors aux scientifiques de l'époque quant au bon déroulement de cette analyse. Le format de données était entièrement propriétaire et totalement inconnu des analystes de l'époque. Pire encore, les développeurs à l'origine du système d'enregistrement étaient décédés dans l'intervalle ou avaient quitté la NASA. Il n'existe alors aucune documentation technique sur le format employé.

Plusieurs mois d'analyse et d'ingénierie inverse (*reverse engineering*) seront nécessaires à l'extraction des précieuses données de ce qu'il convient d'appeler une véritable boîte noire. Si une telle situation n'est plus souhaitable à l'heure actuelle, il convient de garder à l'esprit l'aspect pérennité des données sauvegardées, mais également le facteur humain, car les acteurs d'aujourd'hui ne seront pas forcément ceux qui procéderont à la restauration demain.

1.3 Voguent les données

Les évolutions technologiques successives en matière d'informatique ont des répercussions directes sur le volume de données produit et échangé quotidiennement au niveau mondial. Elles influent également sur la manière dont tout un chacun y a accès et enfin sur les moyens à mettre en œuvre pour en garantir la disponibilité et la résilience.

Au début du XXI^{e} siècle, la donnée devient le sujet de toutes les convoitises et quelques entreprises américaines comprennent rapidement l'importance de la collecter à grande échelle afin de la monétiser, et ce, très souvent à l'insu de son propriétaire. La donnée est désormais « l'or Invisible » de ce siècle et les GAFAM (Google Amazon Facebook Apple Microsoft) l'ont parfaitement saisi. Les technologies de protection de données actuelles sont nées des concepts majeurs ayant traversés les différentes ères informatiques.



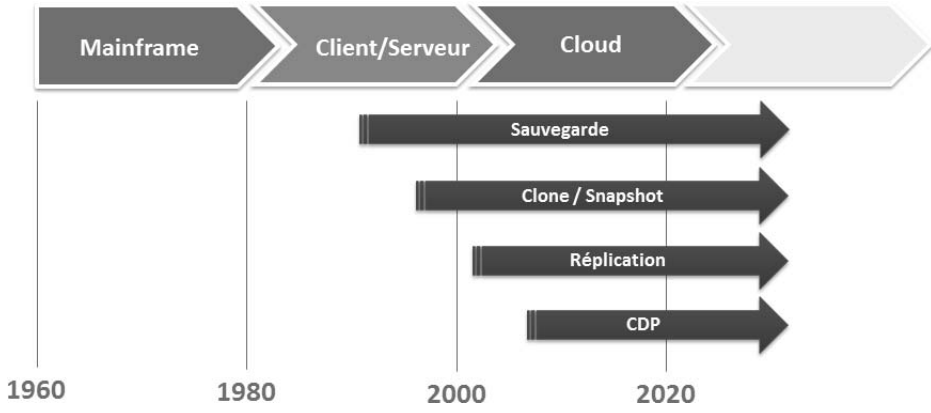
1.4 Des nouvelles exigences et nouvelles technologies

Le 1er janvier 1983, le réseau ARPANET adopte le protocole TCP/IP, la base d'Internet. La volatilité des supports de stockage, associée à une plus grande mobilité des données susceptibles désormais d'être échangées, et au temps de traitement nécessaire pour obtenir un résultat issu d'un calcul, se double de l'impératif de préserver les informations.

Les premières solutions de sauvegarde mettant en application trois principes majeurs apparaissent vers la fin des années 1980. Les autres technologies de protection de données, à savoir la réplication, le cliché instantané (*snapshot*) puis la protection continue des données (CDP - *Continuous Data Protection*) viennent épauler les solutions de sauvegarde traditionnelles ; ces dernières ne suffisant pas toujours à répondre favorablement aux nouvelles exigences en matière de résilience et de disponibilité des données.

Remarque

La protection des données n'est pas une simple affaire de sauvegarde et de restauration, mais un ensemble de processus, méthodes et technologies à mettre en œuvre, constituant une fonction régalienne de toute entreprise publique ou privée.

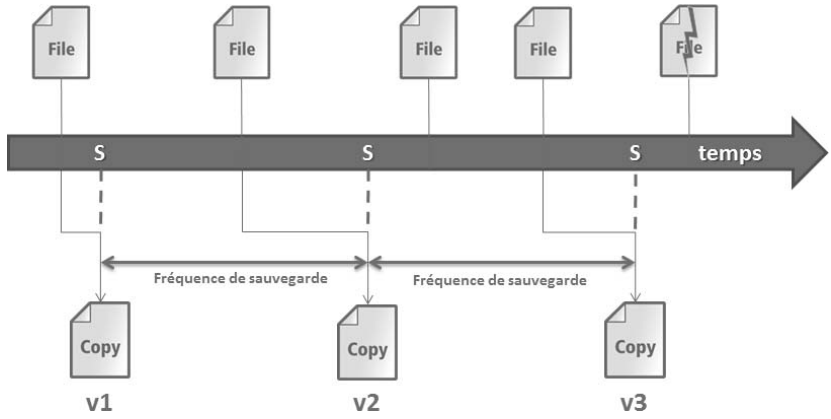


1.5 Les familles SPit et APiT

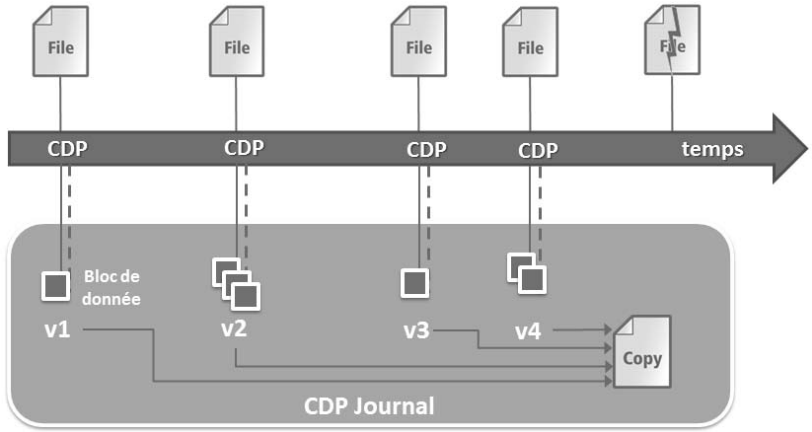
Les technologies de protection des données énoncées précédemment peuvent être classées en deux grandes familles distinctes.

1.5.1 APiT (Any Point-in-Time)

Il convient de traduire l'acronyme APiT (*Any Point-in-Time*) par « à tout instant dans le temps », en l'occurrence dans le passé. Les solutions de sauvegarde traditionnelles appartiennent naturellement à cette famille. Le principe de mémorisation décrit plus loin dans ce chapitre à la section Les trois principes élémentaires est mis à contribution par la sauvegarde afin de pouvoir restaurer une donnée à un instant précis dans le temps. Toute technologie appartenant à cette famille est par conséquent à même de conserver plusieurs versions d'une même donnée et de les restaurer, à la demande, depuis un historique plus ou moins important selon la technologie utilisée et la durée de conservation paramétrée.



Il convient de pondérer le terme « à tout instant dans le temps ». La restauration d'un fichier traditionnel se matérialise par une version disponible parmi celles mémorisées lors des différentes sauvegardes. En revanche, la réalisation de sauvegarde « à chaud » de bases de données relationnelles permet de restaurer celles-ci à un instant très précis, tel qu'une transaction, en exploitant les mécanismes de journalisation. L'exemple ci-dessus illustre, ici, la restauration de trois versions de sauvegarde d'un même fichier alors que celui-ci a fait l'objet de quatre modifications. La fréquence de sauvegarde détaillée ultérieurement est ici supérieure à celle de la modification du fichier. Fort de ce constat, la technologie de protection continue des données (*Continuous Data Protection* - CDP), appartenant à cette même famille, résout ce problème.



Grâce à cette technologie, tout bloc de données modifié au niveau d'un fichier par exemple est immédiatement enregistré comme une nouvelle version du fichier dans un fichier de journalisation : le CDP Journal. Le temps séparant l'écriture de la donnée sur le stockage primaire et l'enregistrement du ou des blocs modifiés dans le journal CDP va être déterminant pour définir si la technologie est dite « *Near CDP* » ou « *True CDP* ». Certaines solutions de protection de données mettent en œuvre la technologie CDP au moyen d'un ordonnanceur. Celui-ci déclenche à un intervalle de temps très rapproché (tous les quarts d'heure ou moins, par exemple) un cliché instantané afin de procéder à l'écriture du ou des blocs de données modifiés dans le journal CDP. La technologie est alors « *Near CDP* ». Dans le cas contraire, ces deux opérations sont quasi synchrones.

1.5.2 SPiT (Single Point-in-Time)

Il convient de traduire cet acronyme dans la langue de Molière par « un seul instant dans le temps ». En d'autres termes et contrairement aux technologies appartenant la famille APiT, une seule version de la donnée est conservée. Cette dernière version est la plus à jour possible. On parle alors de « fraîcheur » de la donnée. La réplication des données appartient à cette famille. Cette technologie de protection des données est largement répandue afin de disposer des données les plus à jour sur un site de secours. La technologie de réplication est mise en œuvre à différents niveaux de l'infrastructure, qu'il s'agisse de répliquer des fichiers, des fichiers journaux d'une base de données, des machines virtuelles, des volumes disques (LUN) entiers ou encore une copie de données précédemment sauvegardée sur disque. La réplication s'exécute selon des modes synchrone ou asynchrone détaillés ultérieurement.