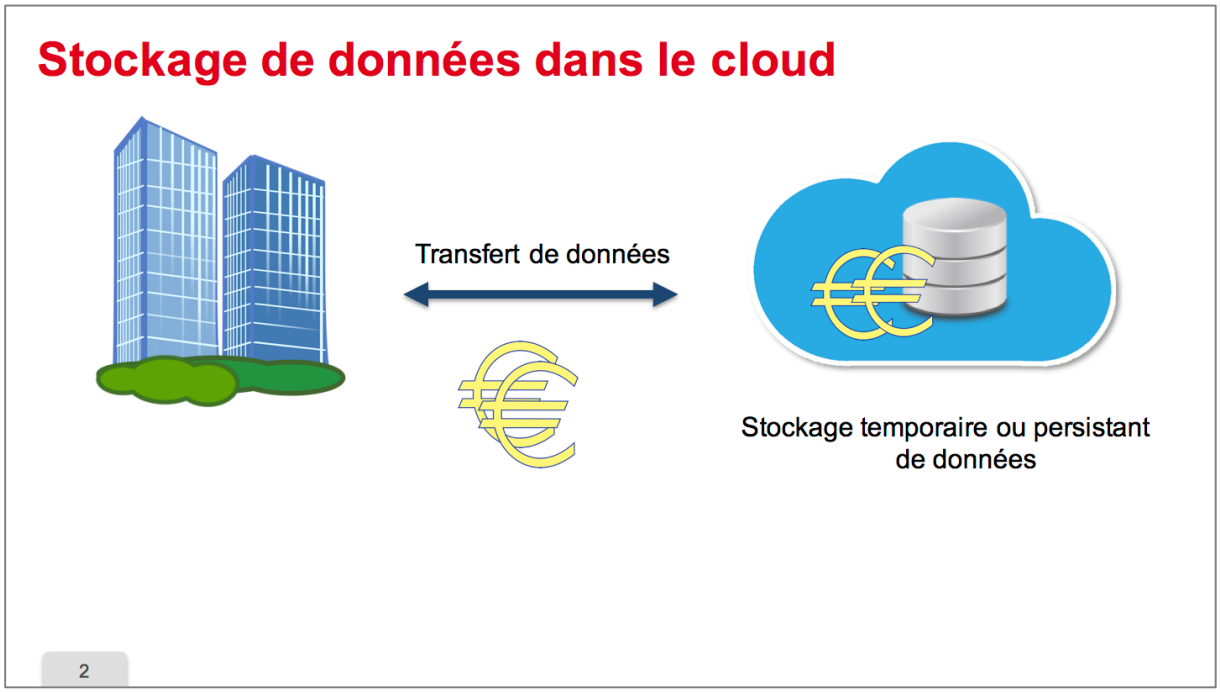
4. Gestion des données urbaines dans les nuages informatiques

- Brève histoire des nuages informatiques
- Modèles de service et de déploiement
- Technologie clé : la virtualisation
- IaaS : les points de vue utilisateur et fournisseur
- PaaS : programmation et déploiement des applications
- **Stockage de données**
- Traitement de données
- Traitement de flux de données

Les clouds offrent des services de stockage de données à la demande.

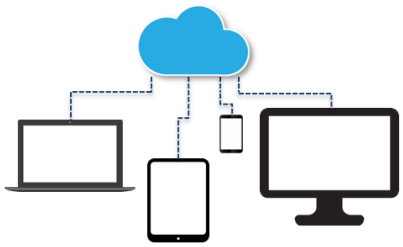


Les fournisseurs de services de stockage dans le cloud facturent non seulement le coût de l'espace de stockage dans le cloud mais également le transfert des données entre l'entreprise et le cloud dans les deux sens, ceci parce que les transferts de données monopolisent des ressources chez le fournisseur.
Plusieurs types de services de stockage peuvent être fournis.

Stockage en libre service (SaaS)

Utilisateur final

- Synchronisation de répertoires de données accédées depuis plusieurs machines
- Copie de sauvegarde de données (backup)
- Partage de données (travail collaboratif, public)



Développeur d'applications

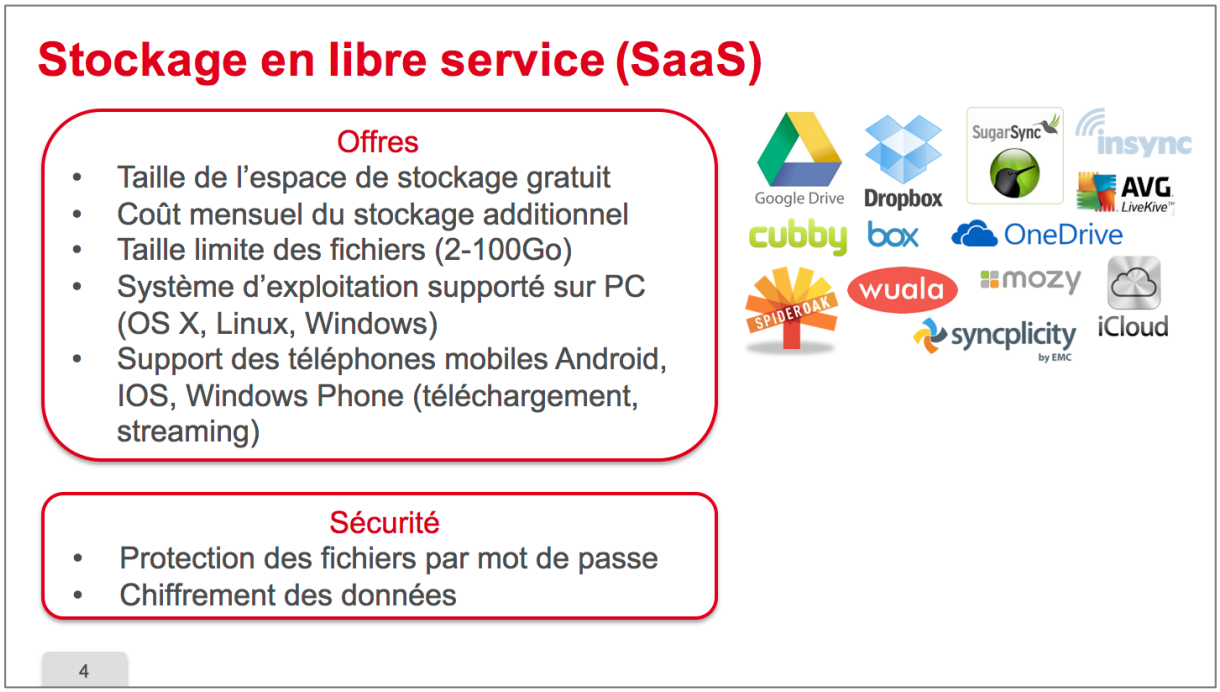
- API publique

3

Au niveau SaaS, des applications dédiées au stockage sont fournies.

En particulier, des applications comme dropbox qui permettent aux utilisateurs finaux de **synchroniser des répertoires de données** accédés depuis plusieurs machines. Par exemple, l'utilisateur peut accéder à ces données depuis son ordinateur personnel ou son smartphone.
On a également des applications de **sauvegarde de Back up de données** dans le cloud.
Il y a également des applications comme Google Docs qui permettent de **partager des données** entre plusieurs personnes dans le cadre d'un **travail collaboratif**. Ces applications permettent alors une édition coopérative d'un même document avec des modifications de chacun visibles en temps réel par tous les autres membres du groupe.
Enfin il y a également des applications qui permettent de rendre visibles des données comme par exemple des applications de partage de photos.

Les développeurs d'applications peuvent utiliser les applications de stockage que je viens de mentionner. En effet, toutes ces applications offrent une interface de programmation.



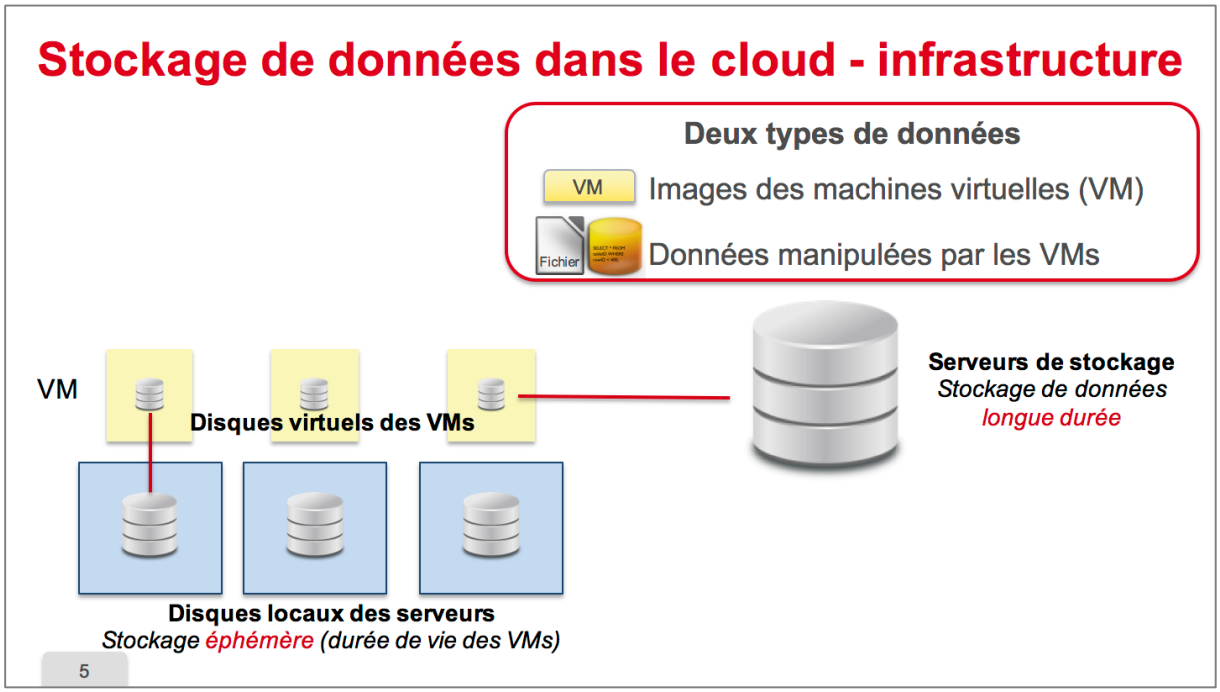
Donc de nombreux produits existent, comme en témoignent les logos qui illustrent ce transparent.

Les services se différencient par :

- la taille de l'espace de stockage gratuit fourni aux utilisateurs,
- le coût mensuel du stockage additionnel,
- la limite maximale sur la taille des fichiers qui peuvent être stockés,
- les systèmes d'exploitation supportés sur les PC,
- le fait que ces services supportent ou pas les smartphones et le type de support offert sur les smartphones.

La sécurité est bien entendue une préoccupation importante pour les propriétaires des données.

Certains services de stockage offrent une **protection des fichiers par mot de passe**, et/ou un **chiffrement des données**.



Voyons maintenant les services de stockage de données offerts par les clouds IaaS infrastructure.

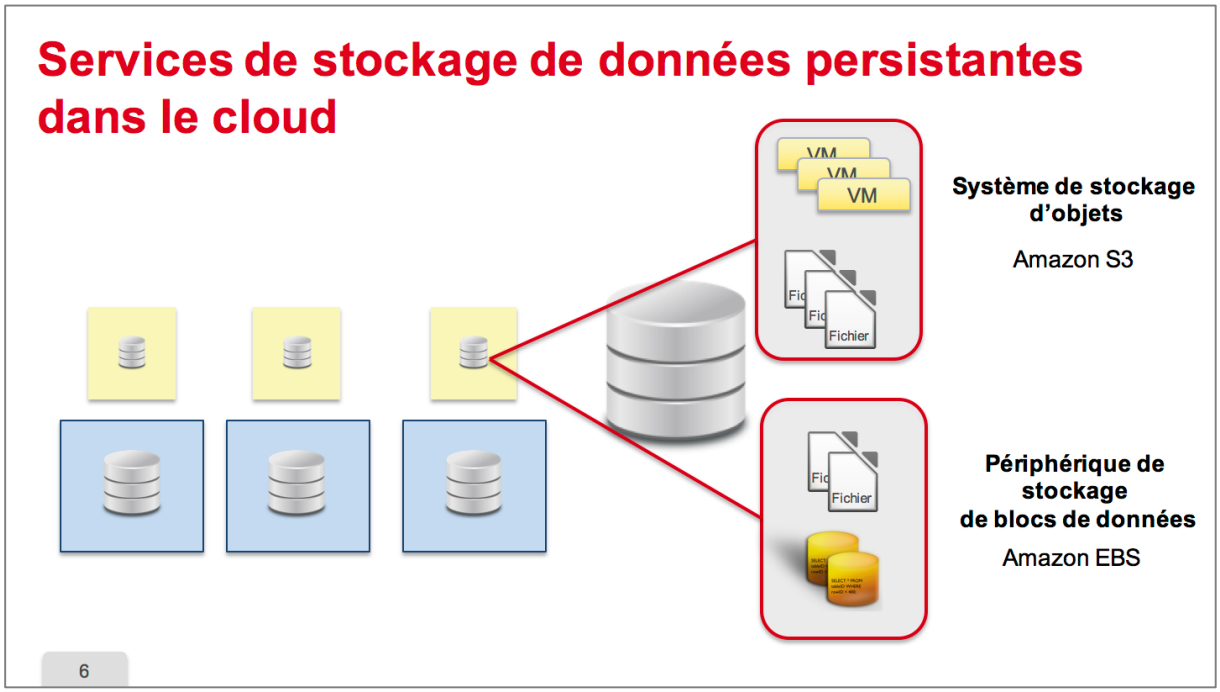
Au niveau infrastructure, deux types de données sont à considérer :

- d'une part les images de machine virtuelle,
- d'autre part, les données des utilisateurs qui sont manipulées par les machines virtuelles.

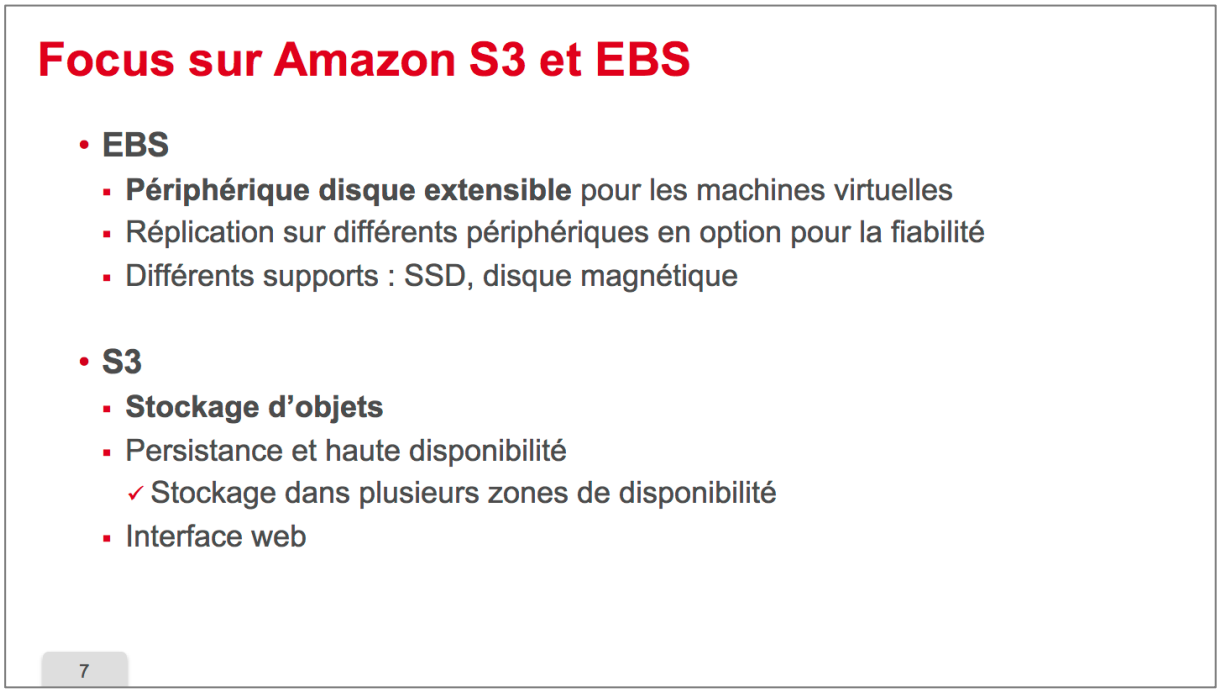
Du point de vue de l'espace de stockage, chaque serveur de calcul dans le Data Center dispose d'un disque local.

Le stockage sur les disques locaux des serveurs est un stockage éphémère par nature. Il est disponible le temps de l'exécution de la machine virtuelle.

En outre, le fournisseur de cloud d'infrastructures gère des serveurs qui sont dédiés au stockage des données pour cette fois-ci du stockage de longue durée. Les disques virtuels des machines virtuelles peuvent être stockés soit sur les disques locaux des serveurs de calcul, soit dans l'espace de stockage partagé.

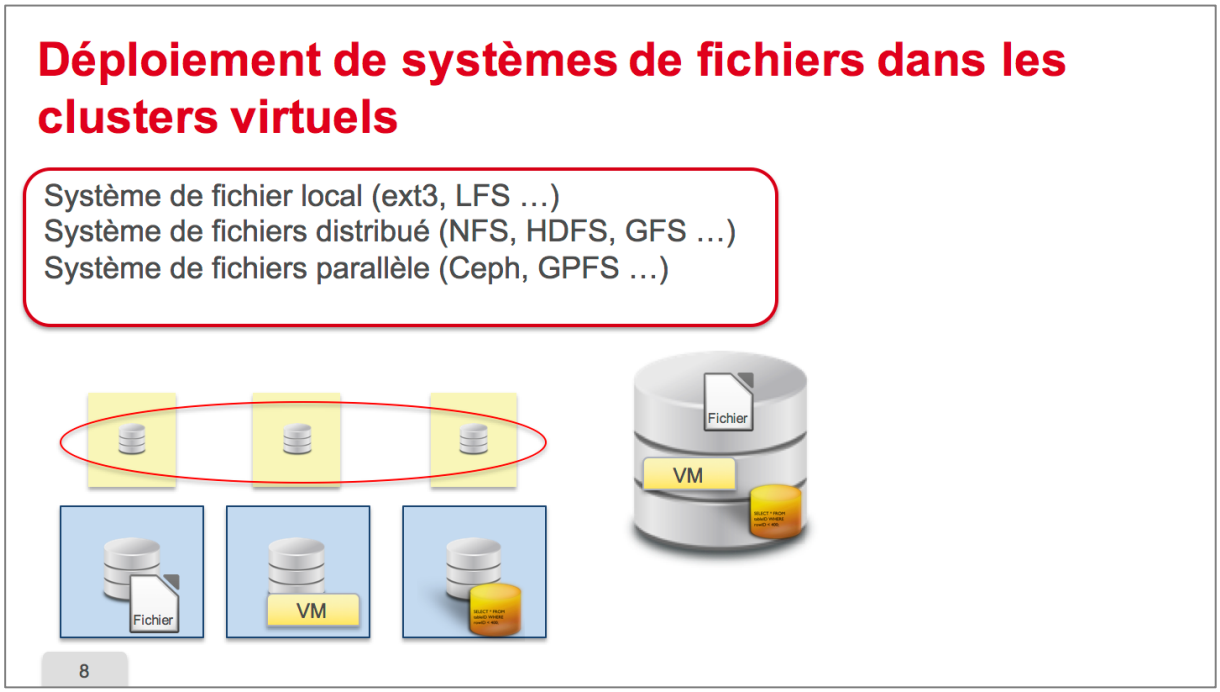


Plusieurs types de stockage de données persistantes existent, les systèmes de stockage d'objets et les périphériques de stockage de blocs de données.
Les systèmes de stockage d'objets sont par exemple le service S3 offert par Amazon. Ce type de systèmes de stockage est utilisé pour stocker les images de machine virtuelle et les fichiers des machines virtuelles.
Un périphérique de type blocs de données peut être utilisé par les machines virtuelles comme un disque sur lequel elles peuvent installer le système de fichiers de leur choix. Le système EBS d'Amazon est un exemple de ce type-là. En général, les machines virtuelles utilisent ce système de stockage pour ranger les fichiers ou des bases de données.
Un petit focus donc sur les deux services de stockage persistants offerts par Amazon, à savoir S3 et EBS.

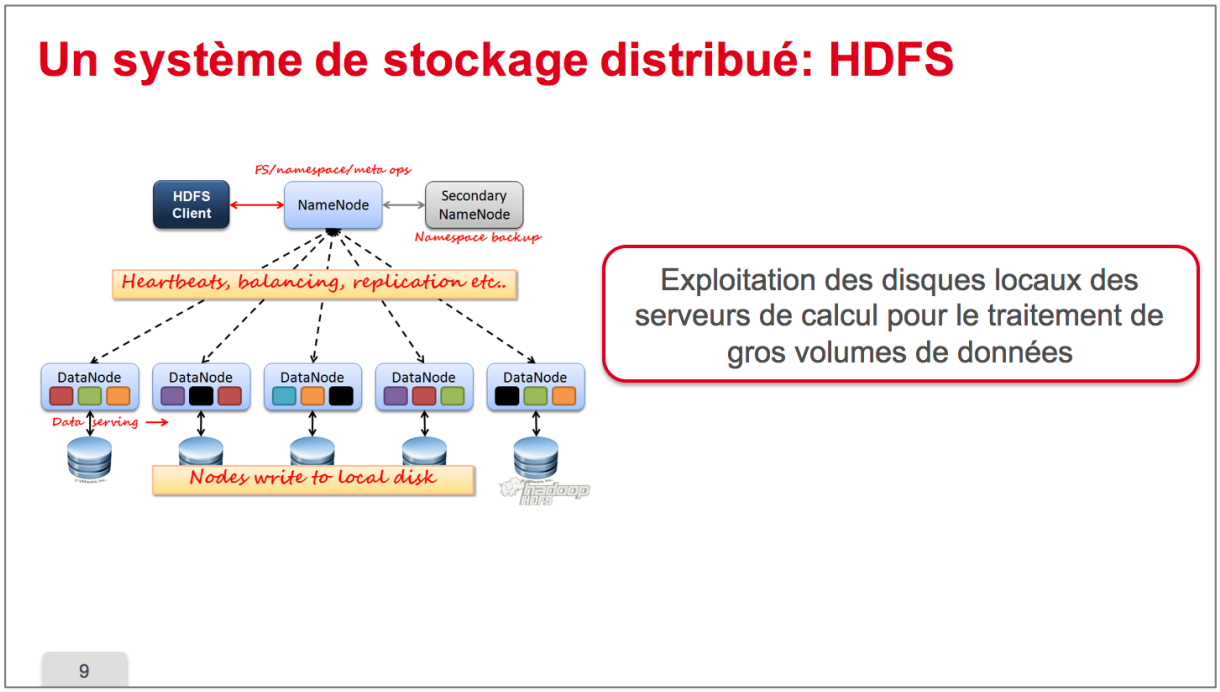


Donc EBS est un **périphérique disque extensible**. Les données sont répliquées sur plusieurs périphériques de manière optionnelle pour garantir de la fiabilité. Et EPS peut utiliser différents supports, soit des disques magnétiques, soit sur SSD.

Le service S3 quant à lui permet de **stocker des objets**, il offre de la **persistance** et de la **haute disponibilité**. Amazon stocke les données de S3 dans plusieurs zones dites de disponibilité, c'est à dire dans plusieurs Data Center géographiquement distants. Ce service est **accessible à travers une interface Web**.



Dans un cluster virtuel, il est possible de déployer différents types de systèmes de fichiers selon les besoins des applications.
Cela peut être un système de fichiers local comme ext3 sur Linux ou LFS qui sera accessible dans la VM où il est installé.
Cela peut être aussi un système de fichier distribué comme NFS, HDFS, GFS qui va permettre de partager des fichiers entre plusieurs machines virtuelles.
Ou cela peut être aussi un système de fichiers parallèles comme Ceph, GPFS, pour permettre des entrées sorties efficaces.
Dans un cloud d'infrastructures, les options pour le stockage des données des applications qui s'exécutent dans les machines virtuelles sont donc multiples. Donc le choix et la configuration du stockage est une tâche ardue pour les utilisateurs de clusters virtuels.



Un système de fichier distribué, communément déployé dans le cloud, est le système HDFS, il est notamment utilisé pour l'exécution d'application MAP reduce, comme on le verra dans la séquence suivante.

La particularité du système de fichiers HDFS est d'exploiter les disques locaux présents sur les différents nœuds de calcul pour effectuer un traitement en parallèle sur de gros volumes de données.

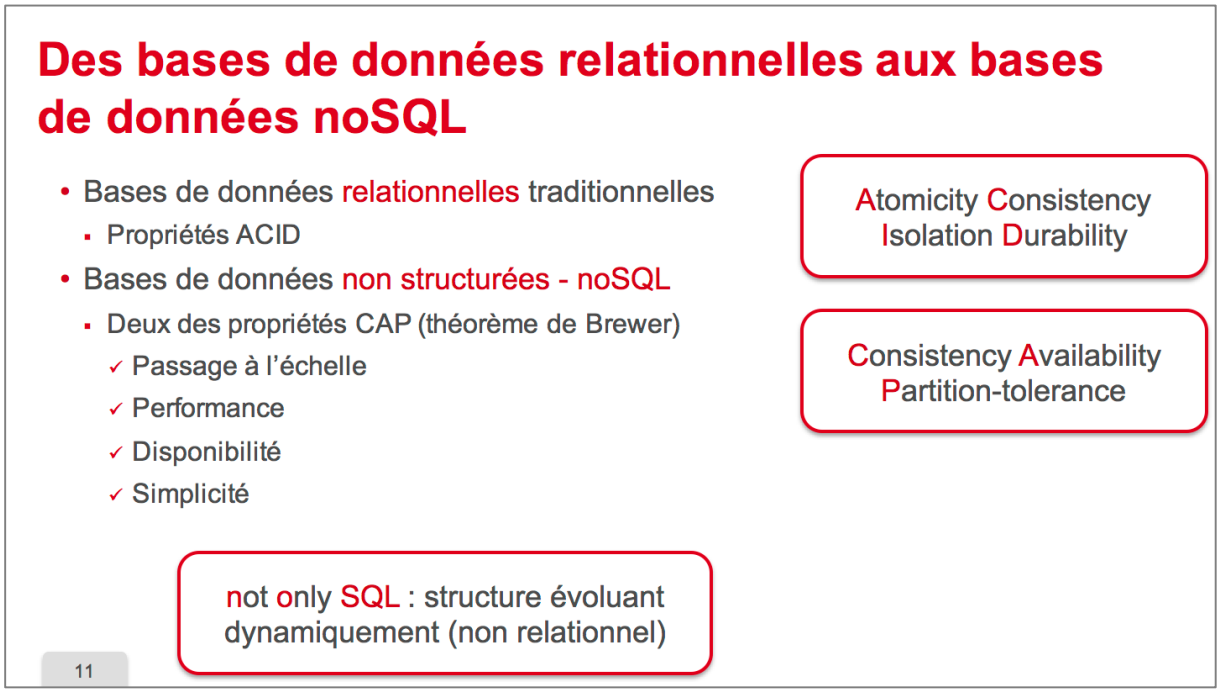
Ce système de fichiers gère automatiquement la réplication des données sur plusieurs disques.

Impact de l'évolution des données sur la conception des bases de données dans le monde du cloud

- Volumes de données croissants
- Variété des données
- Vélocité
- Données semi- ou non structurées
- Données liées entre elles

10

À l'heure du big Data, des **volumes de données croissants** sont produits, les **données sont extrêmement variées** et le **rythme de production** peut être **très rapide**.
Il peut s'agir de **données semi ou non structurées** et bien **souvent les données sont liées entre elles**.



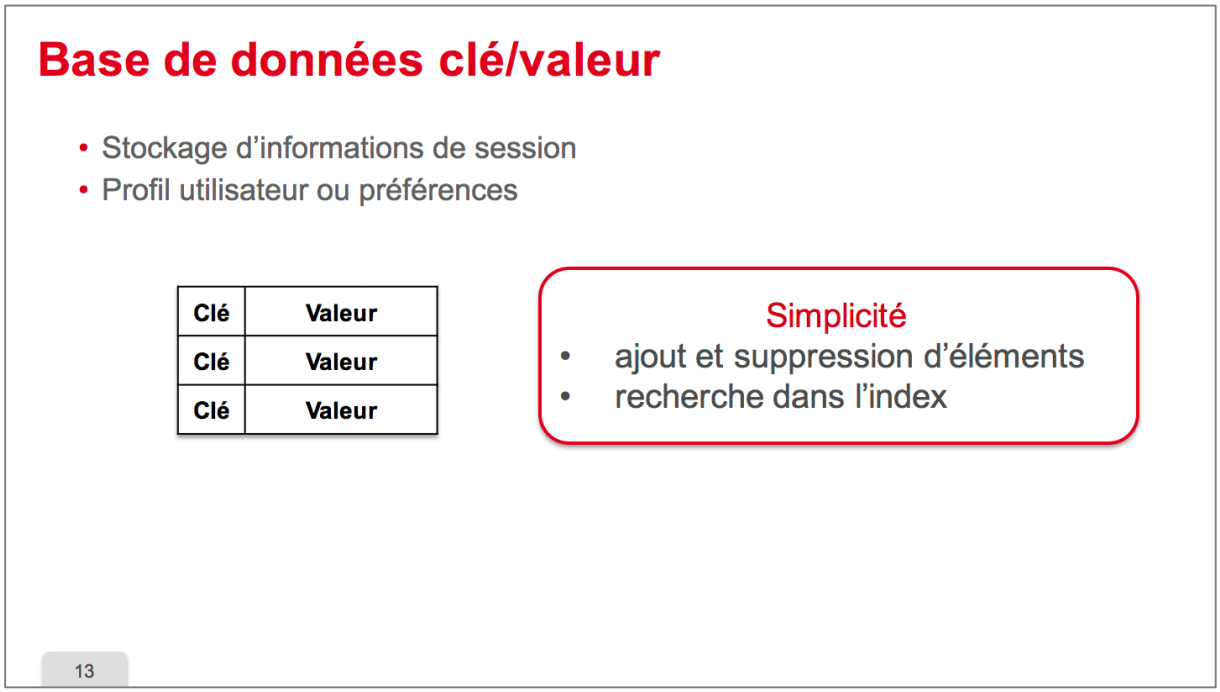
Donc ces évolutions ont donné naissance à de nouveaux types de base de données que nous allons aborder dans la deuxième partie de cette séquence.

Les bases de données traditionnelles sont fondées sur des schémas de données qui sont fixes et définis dès la conception de ces bases de données.
Ces systèmes de base de données traditionnels garantissent les propriétés ACID (Atomicity Consistency Isolation Durability) :

- **Atomicité** des mises à jour tout ou rien,
- **Cohérence** des données en présence d'écritures concurrentes,
- **Intégrité** et **Persistance** des données en dépit de défaillances ou d'attaque.

Il faut savoir que la mise en œuvre des propriétés ACID entraîne des surcoûts importants du fait des besoins de synchronisation, ce qui les rend inadaptées pour bon nombre d'applications nouvelles.

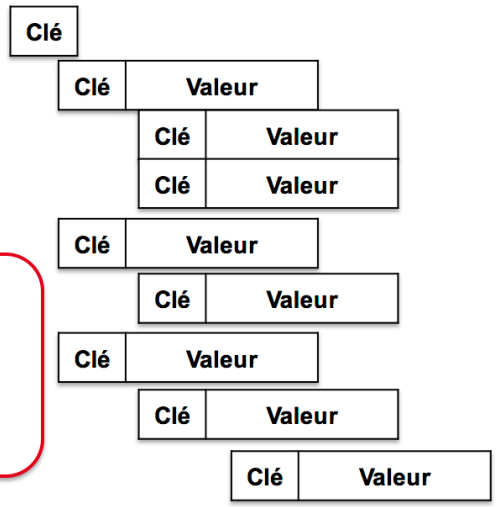
Dans les années 2000 des bases de données de nouvelle génération ont vu le jour, les bases de données non structurées ou not only SQL : structure évoluant dynamiquement (non relationnel).
Dans ce nouveau type de base de données, le schéma des données peut évoluer au fil du temps, par ajout de nouveaux attributs.
Les différentes bases de données et noSQL font divers compromis entre les trois propriétés mentionnées. Bien souvent, elles affaiblissent la cohérence au profit des deux autres propriétés.



Ce sont les plus simples, elles gèrent des index clés/valeur avec seulement trois types d'opérations : la recherche dans l'index, l'ajout ou la suppression d'éléments dans l'index. Elles sont utilisées par exemple pour la gestion de profils utilisateurs ou de préférences dans les applications.

Base de données orientée document

- Application avec différents types d'objets et des recherches sur différents attributs
  - Analyse de données web en temps-réel
    - Nombre de pages vues
    - Nombre de visiteurs



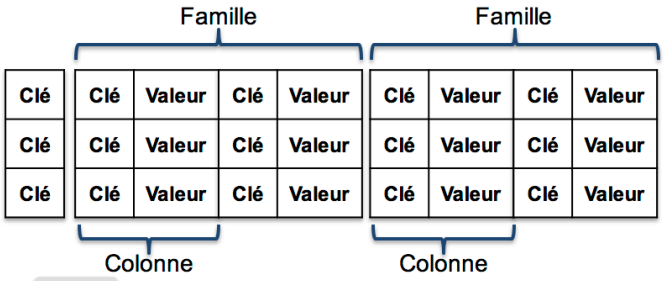
Document : tout type d'objet sans pointeur
- Documents imbriqués, listes
- Index secondaires
- Lignes de taille variable
- Ajout d'attributs dynamiquement

14

Elles sont un peu plus sophistiquées que les précédentes. On appelle document tout type d'objet sans pointeur.
Ces bases de données gèrent des documents imbriqués des listes des index secondaires. Les lignes d'un enregistrement sont de taille variable et il est possible d'ajouter des attributs dynamiquement.
Ce type de base de données est utilisé pour des applications dans lesquelles plusieurs types de données sont gérés et pour lesquelles les données n'ont pas besoin d'être mises à jour de manière immédiate. Cela peut être le cas d'applications d'analyse de données web en temps réel qui calculent par exemple le nombre de pages vues, le nombre de visiteurs donc il manipule les notions de pages et de visiteurs.

Base de données orientée colonnes

- Même type d'utilisation que les bases de données orientées document
  - Débit plus élevé
  - Garanties de cohérence plus fortes
- Exemple
  - Site de commerce électronique



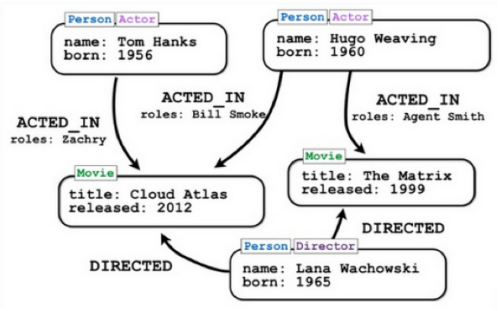
- Gros volumes de données scalaires
- Structuré en famille de colonnes
- Ajout de colonnes dynamiquement
- Partitionnement vertical et horizontal automatique

Elles ont le même type d'usage que les bases de données orientées documents. Cependant, elles sont conçues pour des débits de données plus importants et des garanties de cohérence plus fortes.
Elles sont utilisées par exemple pour les données de sites de commerce électronique. Elles gèrent de gros volumes de données scalaires et sont structurées en famille de colonnes. Donc il est possible d'ajouter des colonnes dynamiquement.
Les données de ce type de base de données sont automatiquement partitionnées horizontalement par groupes de lignes stockées sur différents sites et également verticalement par famille de colonnes à des fins de passage à l'échelle et d'efficacité. Le partitionnement permet d'équilibrer la charge sur plusieurs sites et de mettre en place efficacement de la réplication de données.

Base de données orientée graphe

- Réseaux sociaux
- Services de géolocalisation
- Moteurs de recommandation

3 notions :
- ✓ Nœuds
- ✓ Relations
- ✓ Propriétés



16

Elles sont utilisées dans des applications de type réseaux sociaux, des services de géolocalisation et des moteurs de recommandation.
Dans ce type de base de données, trois notions sont gérées : les **nœuds**, les **relations** entre les nœuds et les **propriétés** qui vont s'appliquer aux nœuds et aux liens entre les nœuds.
Sur l'exemple, le graphe représente des acteurs des films et des réalisateurs, les acteurs sont caractérisés par des propriétés comme leur nom, leur date de naissance.
Les arcs sont également étiquetés avec des mentions telles que "a joué dans" ou "a dirigé".

Déploiement de systèmes de fichiers et bases de données dans le cloud

- IaaS – Mise à disposition d'images de machines virtuelles préconfigurées
- PaaS – Gestion automatique du déploiement et de l'élasticité des systèmes de stockage

Système de fichiers : XtreemFS
Base de données relationnelle : mySQL
Base de données clé/valeur : Scalarix

17

Que ce soit les bases de données ou les systèmes de fichiers, il existe des outils qui permettent de les déployer facilement dans le cloud que ce soit au niveau IaaS ou au niveau plateforme.
Dans le cas des clouds de type IaaS, les systèmes de fichiers ou les bases de données sont installés et préconfigurés dans des images de machine virtuelle qui sont mises à disposition des utilisateurs dans les places de marché.
Certains services de plateforme permettent de déployer et gérer l'élasticité des systèmes de fichiers et des bases de données.
Pour reprendre l'exemple de la séquence précédente de ConPaas, cette plateforme permet de déployer un système de fichiers distribués élastiques XtreemFS et permet également de déployer la base de données relationnelle MySQL, ainsi qu'une base de données noSQL orientée clés valeurs Scalarix.
Tous les systèmes de stockage que nous venons d'évoquer trouvent des applications dans les villes intelligentes qui génèrent des volumes de données considérables qui sont exploitées par une multitude d'applications au service des citoyens.

Illustrations & photos : crédits

p. 2 : by netalloy, domain public, http://www.freestockphotos.biz/stockphoto/15362

p. 9 : © Hadoop HDFS

p. 10 : Domaine public