République Algérienne Démocratique et Populaire
Ministère de L'enseignement supérieur et de la Recherche Scientifique
Université de Tlemcen

Big Data, Cloud Computing

Autres dimensions de la technologie de l'information

Houcine MATALLAH

08/11/2022

Plan

1. Introduction
2. Big Data
3. Cloud Computing
4. Limites des systèmes relationnels dans les environnements distribués
5. Mouvance NoSQL
6. Contre Attaque NewSQL
7. Hadoop
8. Conclusion

Introduction

Changement d'échelle en volumétrie : Quelques chiffres

- PDG google : *«Tous les **2** jours nous nous créons autant de données que ce qui a été crée jusqu'au **2003**» s'appliquent aux données produites par le grand public*

- Université de Berkley (2003) :
  - **5 exaoctets** ($5x10^{18}$ oct) de données ont été crées en 2002
  - **92%** ont été stockées informatiquement
  - Croissance de la volumétrie est de **30% par an**

- IDC (2010) : Volume global des données numériques ≈ 35 zettaoctets **($35x10^{21}$ oct) en 2020**

Introduction

Changement d'échelle en volumétrie : Quelques chiffres

- Intel : **5 Milliards** d'objets connectés en **2013** – **50 Milliards** d'objets connectés en **2020**
- Nombre d'utilisateurs connectés à Internet depuis un **smartphone** a dépassé celui des utilisateurs connectés via un **PC** en 2013
- Nombre de pages web est environ 5 milliards pour plus d'un milliard de sites (2016)
- 3,81 Milliards d'internautes, soit 51% de la population en 2017
- 2,91 milliards d'inscrits sur les réseaux sociaux, soit 39% de la population en 2017

Introduction

Changement d'échelle en volumétrie : Quelques chiffres

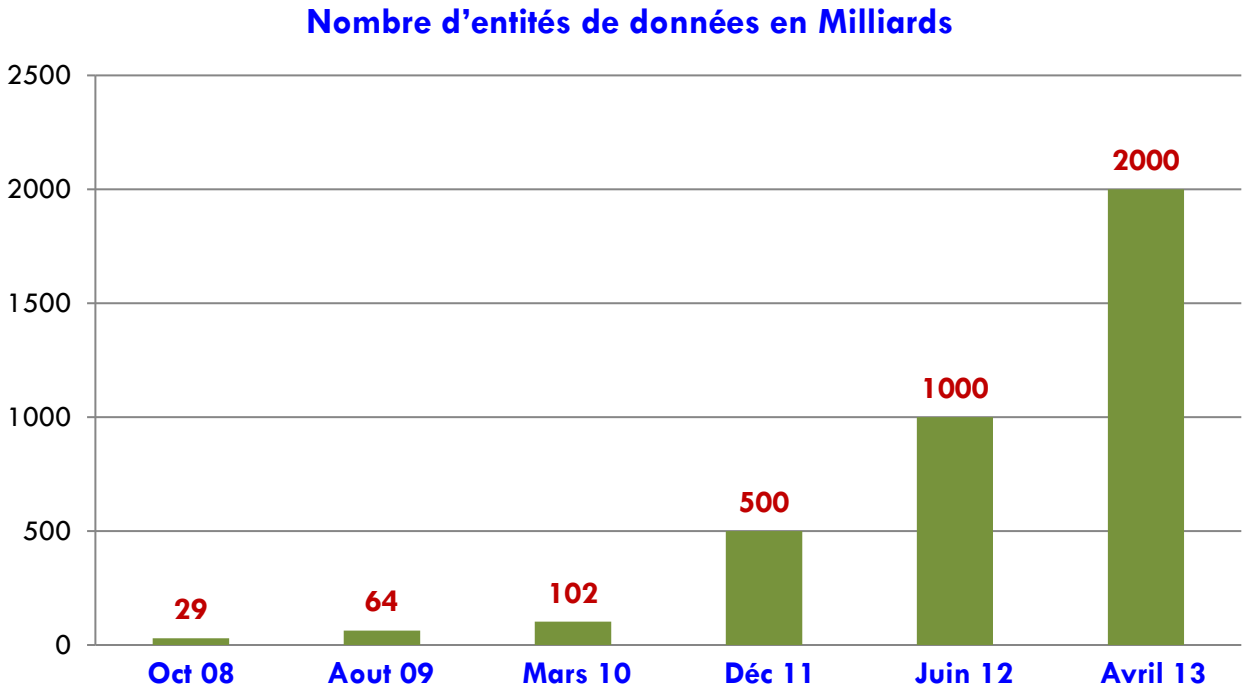
Chaque minute sur Internet (2019- 2020)

- 240 millions de photos aimées sur Facebook et 1 million de connexions
- 200 millions de mails envoyés
- 80 millions de mots traduits sur Google Translate
- 42 millions de messages envoyés sur WhatsApp
- 8 millions de snaps envoyés sur Snapchat
- 5 millions vidéos visionnés et 500 heures de vidéos téléchargées sur YouTube
- 4,2 millions de requêtes de recherches Google
- 3 millions de photos aimées sur Instagram
- 900 mille fichiers téléchargés sur Dropbox
- 200 mille appels sur Skype
- Amazon fait 260 mille $ de chiffre d'affaires en ligne (180 milliards $ en 2017
- 55 Milles d'utilisateurs connectés à Microsoft Teams

Introduction

Changement d'échelle en volumétrie : Quelques chiffres

- Ex : Croissance du nombre de fichiers enregistrés sur le site de stockage Cloud Amazon S3



Introduction

Changement d'échelle

Révolution technologique

Vulgarisation de l'informatique

- Explosion des données (Mega, Giga ⏩ Tera, Peta, Exaoctet,...)
- Changement d'échelle sur les 3 dimensions **Volumes, Types** et **Nombres**
- Accroissements massifs «déluge de données» ⏩ Difficultés rencontrées en matière de performance :
  - Modes d'accès traditionnels (SFL, SFD, SFP)
  - Systèmes de gestion de données classiques (SGBD Relationnels)

Introduction

Nouveaux concepts

- Changement d'échelle en volumétrie des données
- Les montées en charge
- Données hétérogènes et diversité des types (Structurées, Semi-structurées, Non structurées)
- Répartition géographique

Big Data | **Cloud Computing** | **NoSQL** | **NewSQL** | **Hadoop**

Big Data

Big Data

- ***Evolution des architectures des SGBD***

1. Architecture Centralisée
2. Architecture Client-Serveur
3. Bases de Données Distribuées
4. Bases de Données Parallèles
5. Entrepôts de Données (Data Warehouse)

Big Data

De nouveaux besoins : Big Data

- Explosion de données
- Ces masses de données **sans organisation** ni de **structure** sont crées puis accumulées
- **L'absence de classement** ou de **relation** évidente annonce la difficulté de recourir à des structures bien définies comme celles des BDR
- Le « ***déluge de données*** » dans le monde des sociétés commerciales de l'Internet, engendre des besoins d'analyse afin de tirer des **informations synthétiques** et **pertinentes**
- On crée environ 2,5 milliards Go tous les jours, émanant des différents domaines créés par les divers outils numériques : vidéos publiés, messages envoyés, signaux GPS, enregistrements transactionnels d'achats en ligne,.. **Ces volumes massifs de données sont baptisés Big Data**

Big Data

Définition

- Littéralement : **Données massives** ou **méga données**
- **Ensemble d'entités de données hétérogènes en extensibilité permanente** qui ne peuvent pas pris en charge par les systèmes de gestion de données classiques
- **Architecture distribuée et scalable** pour le traitement et le stockage de grands volumes de données
- L'émergence du Big Data est considérée comme une **nouvelle révolution industrielle** semblable à la découverte de la vapeur, de l'électricité, du téléphone et de l'informatique
- D'autres, qualifient ce phénomène comme étant le **dernier épisode de la troisième révolution industrielle,** dite celle de « l'information »

Big Data

Définition

- Concept et terminologie récente identifié par l'égalité des **3V** :

Volume : Big Data est associé à un **volume de données vertigineux,** se situant entre quelques dizaines de téraoctets et plusieurs pétaoctets en un seul jeu de données

Variété des types :

- Big Data, permettent de faire de la création, l'intégration, l'analyse, la reconnaissance, le classement des données de **différents types** comme des photos sur différents sites ou les messages échangés sur les réseaux sociaux, etc..
- La part importante et croissante des **données non structurées** est un des facteurs qui a conduit à cette appellation (*Etude d'IDC 2011: **le pourcentage des données non structurées est de 90%)***

Vélocité (ou Vitesse des échanges) : **Fréquence** à laquelle les informations sont générées, capturées, traitées, stockées et partagées

Big Data

Définition

- Un quatrième **V** pour **Valeur** et un cinquième pour la **Véracité** sont apparu ultérieurement (**+2V**):

Valeur Big Data désigne à la fois les grands volumes de données et la difficulté à extraire de cette masse de données **celles ayant suffisamment de valeur** pour justifier leur analyse

Véracité L'aptitude à juger la **crédibilité et la fiabilité** du nombre indéfini de données collectées qualifie la véracité du Big Data

Big Data

Définition

- Ex : création de données sur un réseau social comme Twitter illustre cette terminologie Big Data

Variété **Vélocité**

A l'occasion d'un événement sportif(classico) , **+8000** messages/s sont générées

Même si la taille de chaque message est faible, leur vélocité se doit d'être importante

Volume

A raison de 140 car/Tweet, la fréquence des échanges de Twitter produit plus de **9 To/jour**

Big Data

Architecture

- Couche matériel (infrastructure Layer)
- Couche stockage (Storage layer)
- Couche management et traitement
- Couche visualisation

Avantages de l'architecture

- Extensibilité (scalabilité)
- Performance
- Coût faible
- Disponibilité

Big Data

Sources et types de données

- **Sources de données structurées**
  - Générées par ordinateur ou par machine (Données de capteurs, Données Web, Données commerciales, Données financières, etc..)
  - Générées par l'homme (Données de saisie , Données de flux de clics, Données liées aux jeux, etc)

- **Sources de données non structurées**
  - Générées par ordinateur ou par machine (Images satellites, Données scientifiques, Photographies et vidéo, etc..)
  - Générées par l'homme (Textes et courrier interne d'une entreprise, Données des médias sociaux, Données mobiles, Contenu du site Web, etc..)

Big Data

Big Data et Data Warehouse

Big Data	Data Warehouse
Les données sont conservées dans un **système de fichiers distribué** et scalable	Les données sont consolidées dans un **serveur central** de l'entrepôt de données
Les types de données manipulées sont **très variés**	Données manipulées **structurées**
Les données Big Data sont **analysées en ligne**	Les données dans la BI traditionnelle sont généralement **analysées en mode déconnecté**
Les données sont analysées en **temps réel** comme pour les moteurs de recherche ou e-commerce	les entrepôts sont alimentés **en intervalles discontinus,** hebdomadaires ou quotidiens
La technologie big data emploie un traitement **massivement parallèle**	Les DW s'articulent sur des architectures **Client/Serveur**

Big Data

Principaux acteurs

- Google (MapReduce, BigTable)
- Amazon (Dynamo, S3)
- Yahoo (Pnuts, S4)
- Facebook (Cassandra, Hive)
- Twitter (Storm, FlockDB)
- LinkedIn (Kafka, SenseiDB, Voldemort)
- LiveJournal (Memcached)
- Apache
- Etc…

Cloud Computing

Cloud Computing

De nouveaux besoins : Pourquoi le Cloud ?

Augmentation de la bande passante sur internet

Diminution des coûts des matériels informatiques

Grid Computing

Informatique distribuée

Virtualisation

Informatique de service

Cloud Computing

Définition

- « Modèle fournissant, à la demande et au travers d'un réseau, un ensemble partagé de ressources informatiques incluant des serveurs, des espaces de stockage, des applications, des traitements et des plates-formes de déploiement qui peuvent être rapidement mises en service avec un effort minimum de gestion et d'interaction avec le fournisseur de ce service » *National Institute of Standards and Technology*

- Le Cloud Computing permet de rendre une infrastructure matérielle et logicielle dynamique et flexible en exposant les capacités des Data-Centers comme étant un "réseau de services virtuels". Dans cette infrastructure, les utilisateurs peuvent accéder et déployer des applications à partir d'Internet suivant leurs demandes et la qualité de service exigée

- Le Cloud recouvre une fédération logique d'une multitude d'équipements répartis sur plusieurs sites sous forme de ressources informatiques reliées par le protocole Internet IP

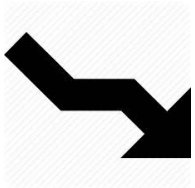
Cloud Computing

- *Pourquoi choisir le Cloud ?*

- **Speed**

- **Scale**

- **Economics**



Cloud Computing

Solution d'avenir

- Le Cloud représente l'avenir des infrastructures software et hardware. Cette tendance est justifiée par :
  - Coûts réduits de plus en plus (Service mesurable et facturable : Ordinateur loué dans le Cloud coûte moins de 5€/mois)
  - Offres attractives
  - Débits des connexions modernisés
- Les offres des services de type Cloud ne cessent de s'accroitre à tous les niveaux, à partir des applications spécifiques, embarqués, CRM, ERP, services web, librairies, jusqu'aux infrastructures complètes
- Un nombre très important d'entreprises ont déjà migré certaines de leurs services vers le Cloud ou envisagent de le faire dans le proche futur
- Le marché du Colud Computing devrait dépasser 241 Milliars $ en 2020 *(Salesforces.com)*

Cloud Computing

Secrets du succès du Cloud

- Aucun investissement initial (Sans infrastructure propriétaire)
- Facturer à l'usage (Pay As You Go) (Devis ou Proforma)
- Simplicité d'accès et d'utilisation de services complexes
- Réduction des coûts et du sous exploitation du matériel, licences, par consolidation et facturation réelle
- Libre-Service (self-service)
- Puissance de calcul élastique
- Transfert de risques

Cloud Computing

Modèles de services

- **SaaS**

  Software as a Service

  (Hard + SGBD ou SE + App)

- **PaaS**

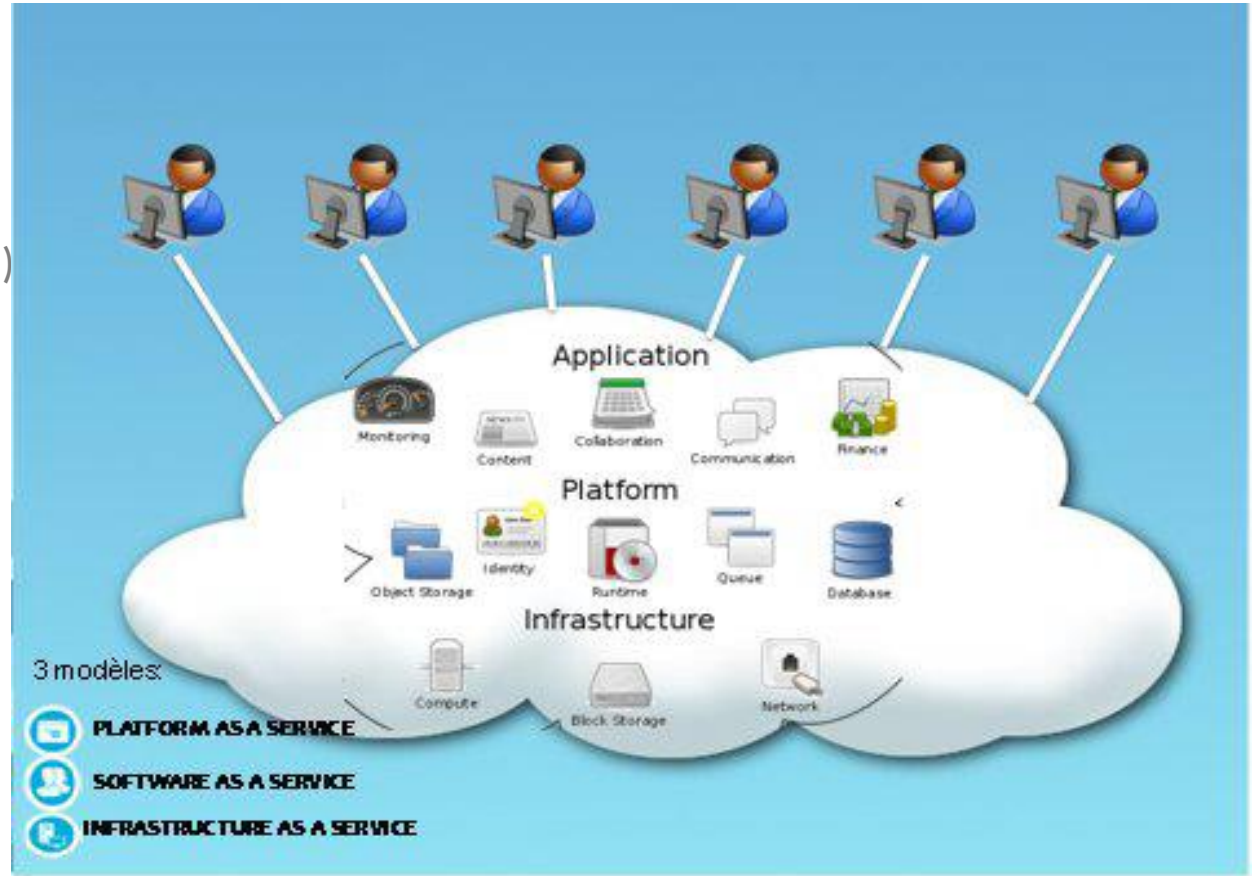
  Platform as a Service

  (Hard+ SGBD ou SE)

- **IaaS**

  Infrastructure as a Service

  (Hard)



Modèles de services

1. SaaS (Software as a Service)

- le logiciel est offert sous la forme d'un service
- Le fournisseur de Cloud gère entièrement sa plateforme matérielle et logicielle
- Les clients utilisent le logiciel fourni sans s'occuper de la pile en dessous (plateforme applicative, matériel…) ni l'installation du logiciel en question
- Ex : Google (Google Apps for Business), IBM (Lotus Life), Microsoft Azure (Office 365, Skype, Bing), Salesforce.com (Chatter), Messagerie électronique, logiciels de type CRM, etc.

Modèles de services

2. PaaS (Plateforme as a Service)

- le fournisseur offre une plateforme sous forme d'environnement complet sur laquelle des développeurs ou éditeurs de logiciels des clients peuvent déployer des applications

- La pile en dessous de cette plateforme à savoir le socle applicatif, le système d'exploitation, le matériel et le réseau, sont gérés par le fournisseur de service Notons que certaines offres PaaS exigent un langage de programmation spécifique

- Ex : Google AppEngine (PaaS où les développeurs peuvent réaliser leurs programmes en Python ou Java), Engine Yard (avec Ruby on Rails), Salesforce.com (langage propriétaire), Coghead SAP (langage propriétaire), Microsoft Azure.

Modèles de services

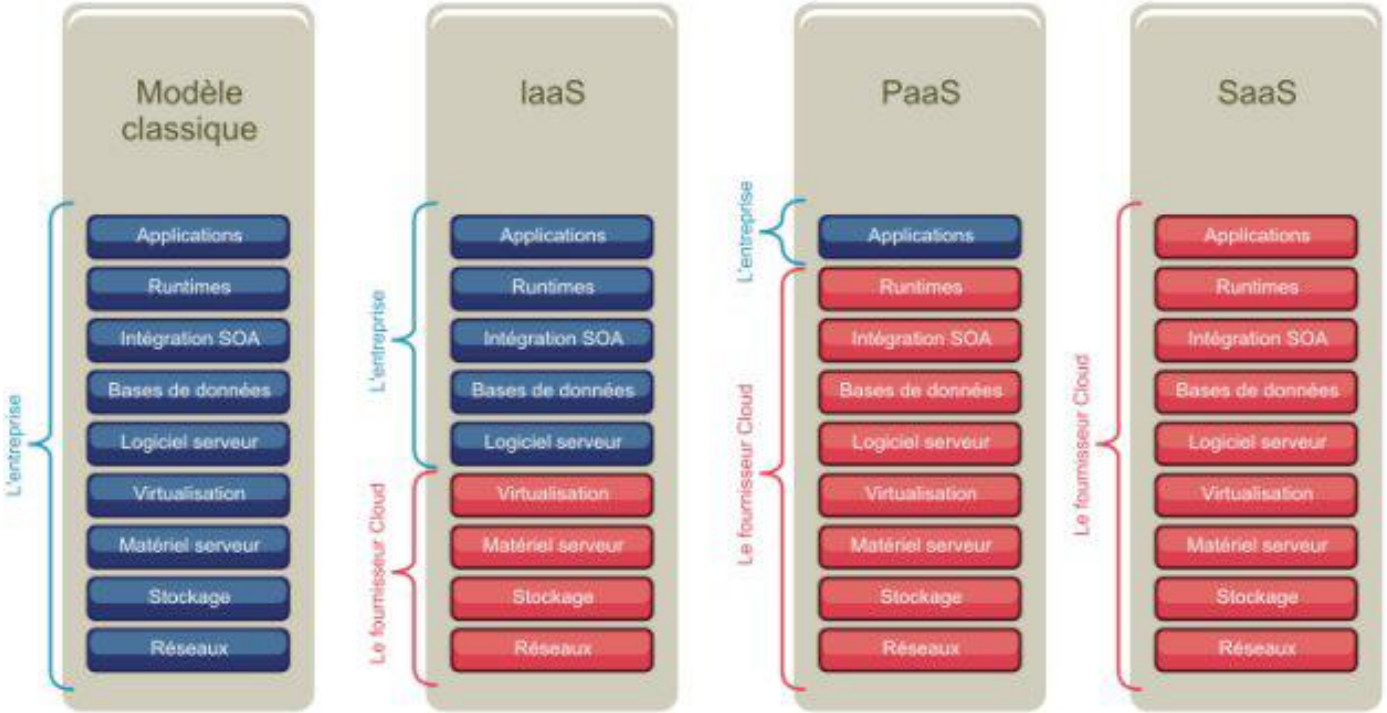
3. IaaS (Infrastructure as a Service)

- IaaS est la proposition d'un ensemble de services informatiques comprenant le matériel, le réseautage et le stockage
- Le fournisseur offre une plateforme sur laquelle les clients vont pouvoir déployer de ressources d'infrastructures dont la plus grande partie est localisée à distance dans des Data Centers
- Le client acquiert une ressource et est facturé pour cette ressource en fonction de la quantité utilisée et de la durée d'utilisation.
- Ex : Amazon EC2 (Web Services Elastic Compute Cloud) et Amazon S3 (Secure Storage Service), Microsoft Azure, HP (CloudSystem) CloudSigma, RackSpace.

Cloud Computing

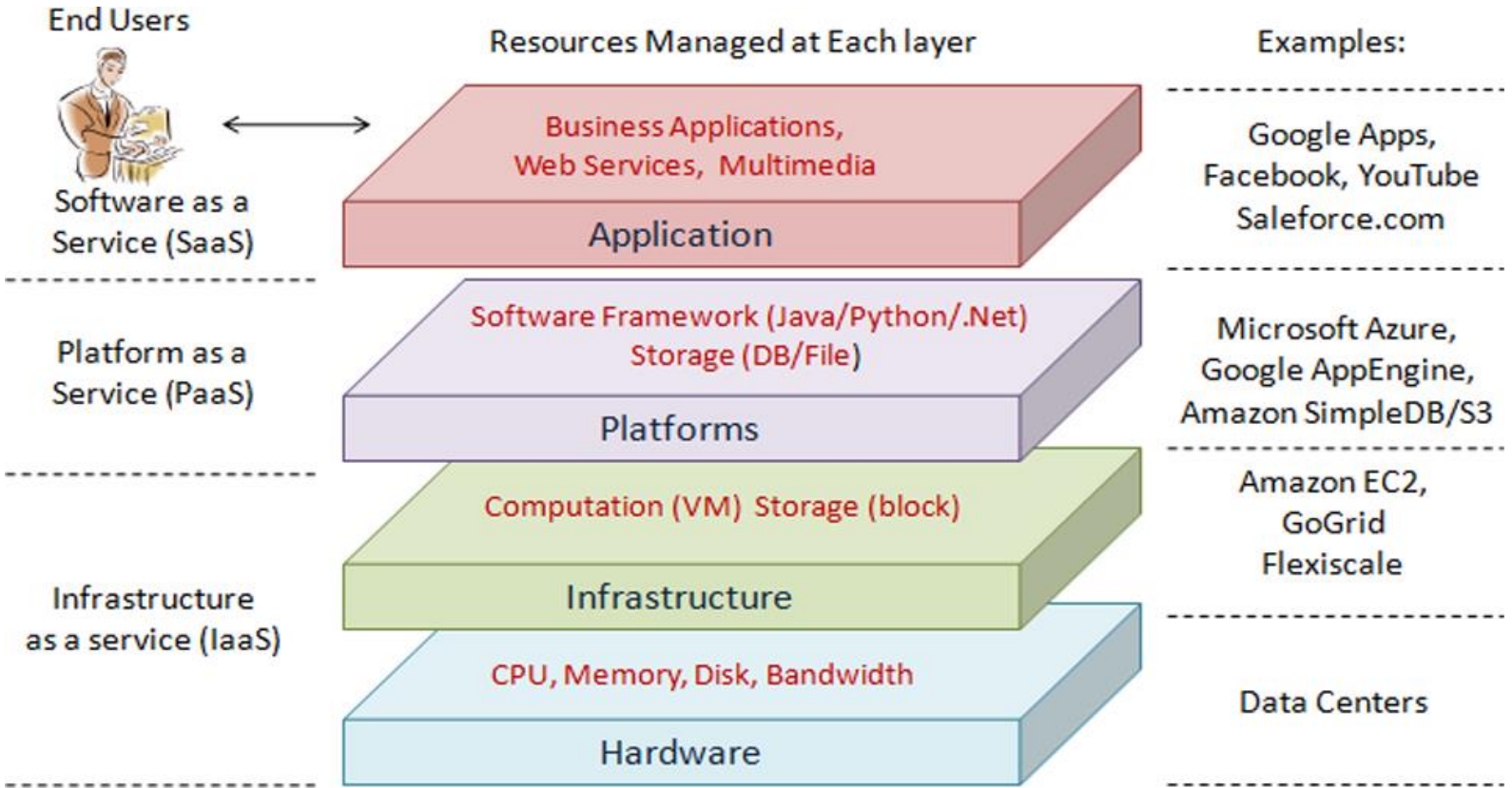
Modèles de services

- **Partage de responsabilité entre le client et le fournisseur dans les 3 modèles**



Cloud Computing

- ***Modèles de services en couches***



Cloud Computing

Modèles de déploiement

- 2 formes principales de déploiements du CC
  1. **Cloud Public**
  2. **Cloud Privé**

- 2 formes dérivées
  1. **Cloud Communautaire**
  2. **Cloud Hybride**

Modèles de déploiement

1. Cloud Public

- Le Cloud public est un ensemble de matériel, réseautage, stockage, services, applications et interfaces dont un tiers est propriétaire et exploité par d'autres sociétés et individus
- Ces fournisseurs externes possèdent des Data Center hautement extensibles masquant les détails de l'infrastructure au client
- Les services offerts sont accessibles via internet à tout le monde qui paye les services

Modèles de déploiement

1. Cloud Public

- Le principe est d'héberger des applications, en général des applications Web, sur un environnement partagé avec un nombre illimité d'utilisateurs
- Ce modèle offre un maximum de flexibilité pour le client et requiert de lourds investissements pour le fournisseur de services
- Les fournisseurs du Cloud public les plus connus sont Google, Microsoft et Amazon

Modèles de déploiement

2. Cloud Privé

- Le Cloud privé est un ensemble de matériel, réseautage, stockage, services, applications et d'interfaces appartenant à une organisation et exploité par lui pour l'usage de ses employés, partenaires et clients
- Les ressources sont détenues et contrôlées par le département informatique de l'entreprise ou les services sont accessibles via le réseau privé
- Le Cloud privé est un environnement hautement contrôlé se trouvant derrière un pare-feu qui n'est pas ouvert à la consommation publique

Modèles de déploiement

2. Cloud Privé

- Si une organisation gère un projet Big Data qui exige un traitement de quantités massives de données, ce modèle pourrait être le meilleur choix en termes de latence et de sécurité
- Ce modèle correspond aujourd'hui à une évolution des Data Centers virtualisés et à l'émergence de l'IT as a Service : le système d'information et les équipements informatiques qui se transforment en centre de services pour le reste de l'entreprise
- Eucalyptus, OpenNebula et OpenStack sont des exemples de solutions proposées pour la mise en place d'un Cloud privé

Modèles de déploiement

3. Cloud Communautaire

- Extension du Cloud privé
- les ressources, services et la propriété sont partagés à l'échelle d'une communauté (ex. : à l'échelle d'un Etat, d'une ville, d'une académie, etc.).
- Les services sont accessibles via l'interconnexion de réseaux appartenant aux organisations de cette communauté

Modèles de déploiement

4. Cloud Hybride

- Une combinaison d'un Cloud privé combiné à l'utilisation de services de Cloud public
- Certains services sont accessibles via un réseau privé et d'autres via Internet
- Le futur devrait confirmer l'émergence du Cloud hybride qui consiste à la cohabitation et la communication entre un Cloud privé et un Cloud publique dans une organisation partageant des données et des applications

Cloud Computing

Inconvénients majeurs

- Problème de fiabilité de la bande passante vers le fournisseur
- Problème de sécurité et confidentialité des données externalisées

Cloud Computing

- ***Principales solutions propriétaires***
  - Amazon's Public Elastic Compute Cloud
  - Google big data services
  - Microsoft Azure
- ***Principales solutions libres***
  - Eucalyptus
  - OpenStack
  - OpenNubela

Cloud Computing

Solutions développées par des poids lourds du marché Cloud

1. Amazon's Public Elastic Compute Cloud (EC2)

- Un des prestataires de services IaaS les plus prestigieux
- Initialement, Amazon a construit une infrastructure massive pour fonctionner son propre commerce de détail, sauf que l'entreprise a découvert que ses ressources étaient sous-exploitées
- ils ont décidé en 2006, de tirer parti de ces ressources en louant son infrastructure à d'autres entreprises ou utilisateurs
- Amazon EC2 offre une scalabilité contrôlée par l'utilisateur, qui paye les ressources utilisées à l'heure (à partir de quelques centimes € / heure)

Solutions développées par des poids lourds du marché Cloud

2. Google big data services

- Google Compute Engine : fournit des machines virtuelles (VM) qui s'exécutent dans les centres de données innovants et sur le réseau de fibre optique mondial de Google
- Google Big Query: Permet d'exécuter des requêtes similaires à SQL à grande vitesse sur des gigantesques ensembles de données de plusieurs milliards de lignes
- Google Prediction API : Un Cloud basé sur les outils d'apprentissage pour de grandes quantités de données. La prédiction est capable d'identifier des modèles de données qui peuvent être analysés à des fins diverses, y compris la détection de fraude, l'analyse de désabonnement et l'impression du client
- Abonnement annuel par utilisateur 40 €

Cloud Computing

Solutions développées par des poids lourds du marché Cloud

3. Microsoft Azure

- Microsoft a produit un ensemble d'outils de développement, de supports de machines virtuelles et de services de périphériques mobiles dans une offre PaaS
- Pour les clients possédant une grande expertise en .Net, SQLServer et Windows, l'adoption de PaaS Azure n'est pas assez compliquée
- Pour répondre aux nouvelles exigences d'intégration de Big Data dans ses solutions, Microsoft a également ajouté Windows Azure HDInsight basé sur Hortonworks Data Platform (HDP), qui offre une compatibilité à 100% avec Apache Hadoop
- HDInsight prend en charge la connexion avec Microsoft Excel et d'autres outils d'informatique décisionnelle (BI)
- Fournisseurs de licences (Anc politique) → Fournisseurs de services (Nouv politique)
- Abonnement mensuel par utilisateur de 5,25 à 22 €

Solutions libres

1. Eucalyptus

- Eucalyptus est une plateforme open source de Cloud Computing apparu en 2007 du projet VGrADS de l'université de Californie, Santa Barbara. Il permet d'exécuter des machines virtuelles dans un IaaS virtualisé
- Eucalyptus organise l'IaaS de façon hiérarchique : les machines au niveau des feuilles, les clusters (groupe de machines) au niveau intermédiaire et le Cloud (ensemble de clusters) à la racine
- Eucalyptus est composé de cinq principaux éléments: Cloud Controller, Node Controler Cluster Controller, Walrus, Storage Controller

Solutions libres

2. Open Stack

- OpenStack est une offre IaaS 100% open source, créé en juillet 2010 par la NASA et l'hébergeur américain Rackspace. Ce projet est soutenu par plusieurs constructeurs tels que : AMD, Intel, Dell et Citrix
- OpenStack comprend le composant Open Stack Compute pour la création automatique et la gestion de grands groupes de serveurs privés virtuels
- OpenStack Stockage pour optimiser la gestion de stockage et répliquer le contenu sur différents serveurs et le mettre à disposition pour une utilisation massive de données
- OpenStack s'organise autour de trois composants et des API de communication : OpenStack Nova, OpenStack Swift, OpenStack Imaging Service

Solutions libres

3. Open Nebula

- OpenNebula a vu le jour en 2005 à l'université Complutense de Madrid dans le cadre du projet européen open source dont l'objectif est l'administration des IaaS virtualisés
- Il organise l'IaaS sous forme de clusters et de VLAN (réseaux virtuels)
- Un cluster contient un ensemble de machines physiques tandis qu'un VLAN est défini pour un ensemble de VM
- Lors de la création d'une VM, le client choisit la machine et le VLAN dans lequel il souhaite l'exécuter
- Les composants d'OpenNebula peuvent être répartis sur trois couches : Tools, Core, Drivers
- Une version commerciale d'OpenNebula (OpenNebulaPro)

Simulateurs Cloud

Plusieurs simulateurs de Cloud ont émergé pour palier au besoin de modélisation, simulation et expérimentation :

- Open Cirrus, solution payante développée par HP, Intel et Yahoo
- CloudSim, CloudAnalyst
- GreenCloud, NetworkCloudSim
- EMUSIM, MDC SIM
- CDOSim, iCanCloud
- VirtualCloud, TeachCloud
- Etc..

Modes de déploiement des bases de données dans le Cloud

3 modes de déploiement des bases de données dans les environnements Cloud :

- **Base de données sur l'instance d'une image de machine virtuelle (IaaS)** : Le client peut opter pour des bases de données dans le Cloud en payant un temps d'utilisation limité d'instances de machines virtuelles (Oracle, Amazon, Rackspace, IBM,..)

- **Base de données en tant que service (SaaS)** : Le client peut acheter l'accès à un service de base de données maintenu et fourni par un fournisseur Cloud de base de données (Cloud Database Provider)
  Amazon fournit trois services de BD : SimpleDB, DynamoDB, Amazon Aurora

- **Base de données hébergée et administrée dans le Cloud** : Le fournisseur n'offre pas une base de données en tant que service, mais il assure seulement l'hébergement et l'administration de cette base à la place de l'utilisateur (Rackspace)

Modèles de données utilisés

3 types d'architectures de BD déployées dans le Cloud :

- Les BD relationnelles traditionnelles dites **SQL**
  Oracle, SQL Server, IBM DB2, Ingres, PostgreSQL, MySQL,..
- Les nouvelles BD non-relationnelles dites **NoSQL**
  MongoDb, Hbase, Cassandra, CouchBase, Redis, OrientDB, Bigtable,..
- Les nouvelles BD s'inspirant sur le modèle relationnel dites **NewSQL**
  VoltDB, Spanner, ClustrixDB, NuoDB, SQLFire, IMDB, Drizzle, MemSQL,..

Limites des systèmes relationnels

dans les environnements distribués

Limites des systèmes relationnels
dans les environnements distribués

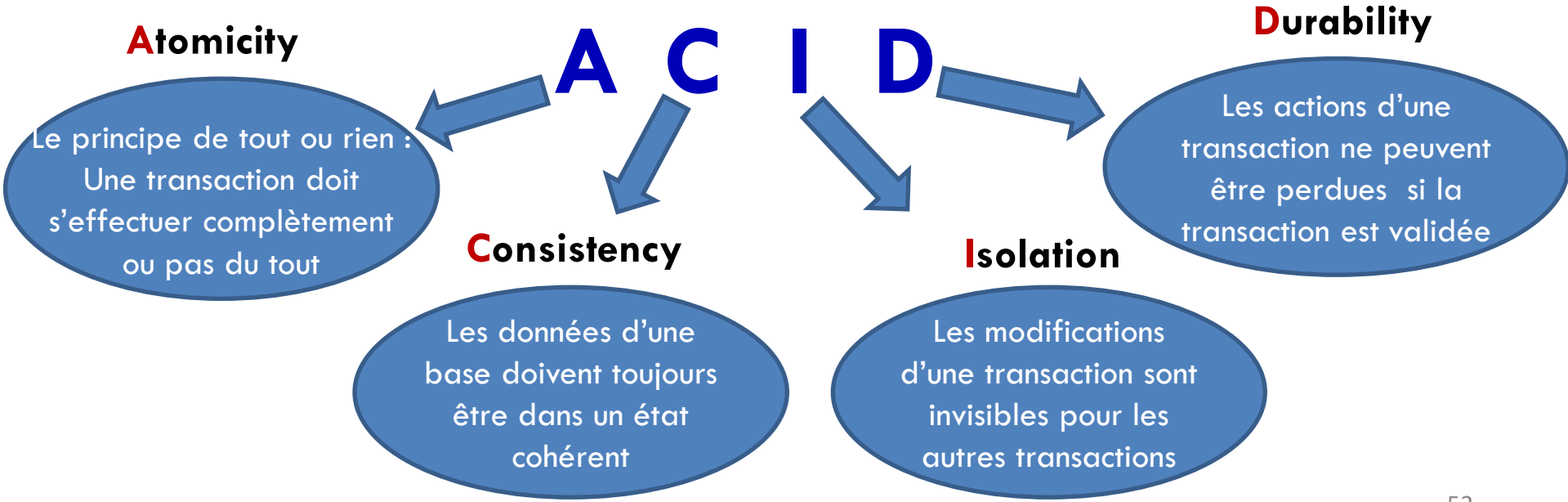
Modèle relationnel : Rappel

- Modèle introduit par **Edgar Codd** (**IBM 1970**), Fondé sur des concepts mathématiques
- les données sont stockées dans une **table,** l'ensemble de ces tables constituent une **base de données relationnelle** ayant un **schéma**
- Le schéma est une **représentation structurelle** du contenu de la base de données, définissant les **tables,** les **champs** des tables et les **relations entre les deux**
- Les données de chaque table peuvent être lues, supprimées et mises à jour en utilisant le **langage standard** pour les bases de données relationnelles **SQL**
- La gestion, le stockage, la mise à jour, le partage, la cohérence et la sécurité des données sont assurés par un **SGBD Relationnel**

Limites des systèmes relationnels
dans les environnements distribués

- ***Propriétés ACID : Rappel***

Les SGBDR répondent bien au besoin transactionnel en respectant les contraintes **ACID**, qui permettent de garantir la fiabilité d'une transaction



Limites des systèmes relationnels
dans les environnements distribués

Caractéristiques des systèmes de gestion des données

- **Performance**
  - ***Latence*** : le temps d'attente avant l'arrivée du premier octet
  - ***Débit*** : La quantité d'octets transférés

- **Extensibilité** : Capacité de prise en charge des accroissements de toute nature, volumétrie, nombre de clients, etc.. avec des temps de réponse +/- constant
  - ***Verticale*** (*scale-up*) : garder le même nombre de serveurs, en haussant leurs puissances
  - ***Horizontale*** (*scale-out*) : ajouter d'autres serveurs dans le réseau

- **Disponibilité** : Capacité de fonctionnement en continu sans défaillance

Limites des systèmes relationnels
dans les environnements distribués

- ***Disponibilité d'une base de données***

Disponibilité en %	Arrêt annuel
90% (un neuf)	36.5 jours
99% (deux neufs)	3.65 jours
99.9% (trois neufs)	8.76 heures
99.99% (quatre neufs)	52.56 minutes
99.999% (cinq neufs)	5.26 minutes
99.9999% (six neufs)	31.5 secondes
99.99999% (sept neufs)	3.15 secondes

Limites des systèmes relationnels
dans les environnements distribués

Contraintes principales

Les systèmes relationnels ne peuvent pas être déployés dans un environnement Cloud à grande échelle, en conservant les mêmes performances :

- **Application des propriétés ACID en milieu distribué**
- **Scalabilité limitée**
- **Requête de jointure non optimale**
- **Gestion des objets hétérogènes**
- **Types de données limités**
- **Langage de manipulation**
- **Pauvreté sémantique**

Mouvance NoSQL

NoSQL

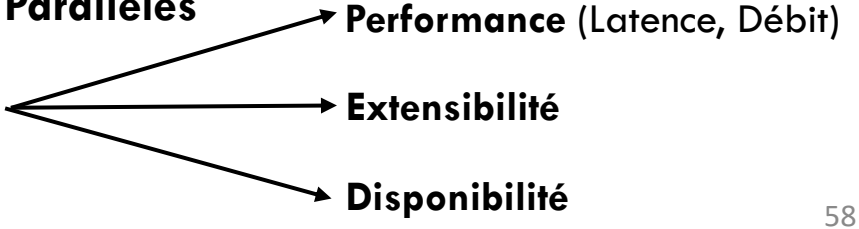
Définition

- 2009 : Terme **NoSQL** choisi pour intituler tous les SGBD de type non-relationnel

NoSQL

Concepts de base

- Solutionner les difficultés rencontrées par les SGBDR dans la gestion des données classées « Big Data »
- Proposer des **alternatives** aux bases de données relationnelles
- Abondan de la représentation matricielle de l'information et le langage SQL
- S'adapter aux nouvelles tendances et architectures du moment
- Manipuler des **grandes masses de données et distribués**
- S'appuyer sur les **Systèmes de Fichiers Parallèles**
- Ensemble de concepts qui insistent sur



NoSQL

Caractéristiques

- Plus d'enregistrements dans des tables
- Pas de jointures
- Pas de schéma imposé
- Acidité relative (Pas obligatoirement ACID)
- Travaille sur plusieurs processeurs
- Utilise la notion "ne rien partager" entre noeuds (shared-nothing)
- Supporte les modèles de traitements parallèles
- Supporte des requêtes distribuées

Notons aussi que le NoSQL n'est pas :

- Un langage opposé au SQL
- Toujours open source
- Toujours lié au Big Data ou Cloud Computing

NoSQL

Contexte dans lequel les grands systèmes ont été construits

- **Amazon** : (DynamoDB) e-commerce
- **Facebook** : (Cassandra puis HBase) fonction de recherche dans la boîte de réception
- **Google** : (Bigtable) Service de base de données NoSQL Big Data de Google. Cette base de données est utilisée par beaucoup de services Google, tels que la recherche, Analytics, Maps et Gmail.
- **Yahoo** : (Pnuts) construit pour stocker les données des utilisateurs qui peuvent être lues ou écrites sur toutes les page Web, stocker les listes de données pour les pages de shopping de Yahoo, stocker les données pour servir ses applications de réseau social
- **LinkedIn** : (Voldemort) gérer les mises à jour en ligne à partir de diverses caractéristiques d'écritures intensives sur le site internet

NoSQL

Relationnel vs NoSQL

Relationnel	NoSQL
Schéma défini au départ	Schémaless
Structure rigide	Structure dynamique
Gestion des valeurs Null	Pas de valeurs Null
Transactions ACID	Transactions BASE
Scalabilité Verticale	Scalabilité Horizontale sans limites
Insiste sur la Cohérence des données	Insiste sur la Disponibilité et la Performance
Mauvaise gestion de gros volumes de données	Conçu pour les gros volumes de données
Langage SQL	Différents Langages selon le modèle utilisé

NoSQL

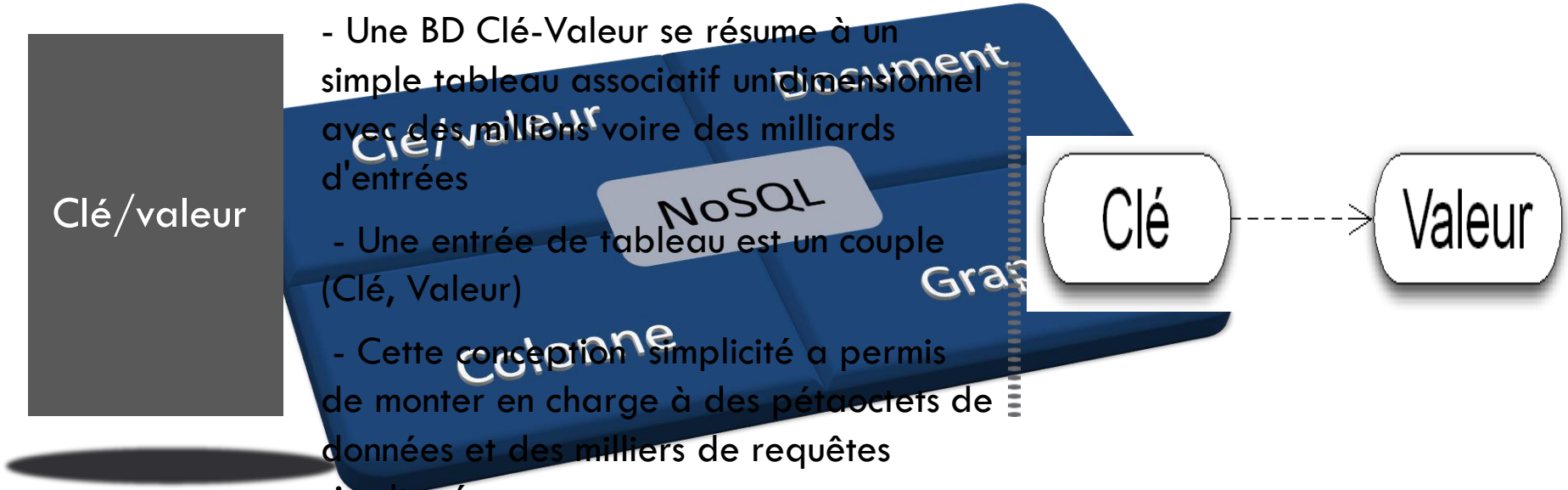
Architectures

Clé/valeur

- Une BD Clé-Valeur se résume à un simple tableau associatif unidimensionnel avec des millions voire des milliards d'entrées

- Une entrée de tableau est un couple (Clé, Valeur)

- Cette conception simplicité a permis de monter en charge à des pétaoctets de données et des milliers de requêtes simultanées

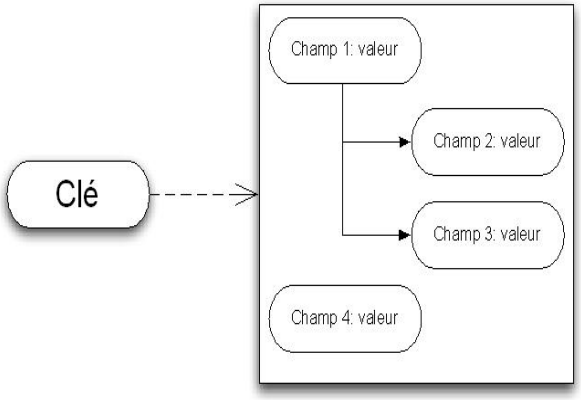


- **Key-value store :** Redis, Memcached, Amazon DynamoDB, Riak, Ehcache Hazelcast (utilisé dans le site LinkedIn), OrientDB, Oracle NoSQL, Berkeley DB, etc.

Architectures

Orientée Document

- Les clés ne sont plus associées à des valeurs sous forme de bloc binaire mais à un document dont le format n'est pas imposé.

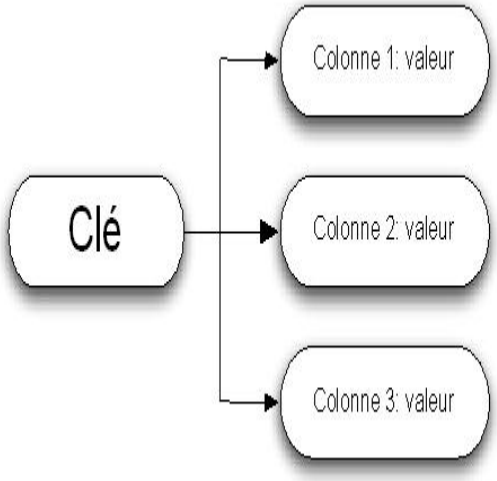


- **Document store :** MongoDB, CouchDB, CoucHBase, AmazonDynamoDB, MarkLogic, RavenDB, OrientDB, Cloudant, GemFire, RethinkDB, Datameer, Microsoft Azure DocumentDB, ArangoDB, PouchDB, etc.

Architectures

Orientée Colonne

- Les bases de données orientées colonne vont stocker les données de façon à ce que toutes les données d'une même colonne soient stockées ensemble

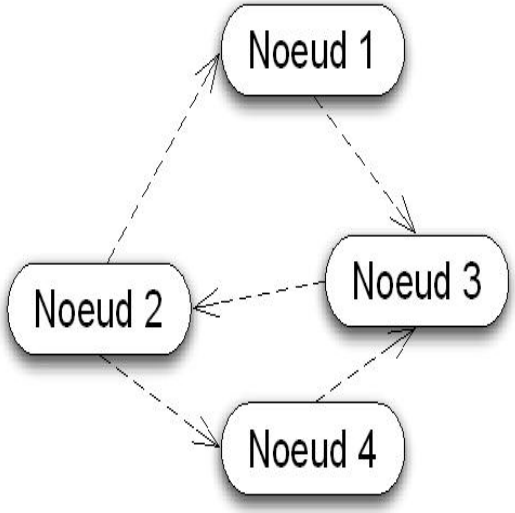


- **Column family :** Cassandra (Apache), HBase (Apache), Bigtable (Google), Accumulo (Apache), Hypertable, etc.

Architectures

Orientée Graphe

- Palier à des problèmes impossibles à résoudre avec des BDD relationnelles
- Cas d'utilisation typique : Les réseaux sociaux où l'aspect graphe prend tout son sens



- **Graph store :** Neo4j, OrientDB, Titan, ArangoDB, Giraph, InfiniteGraph, Sqrrl, Sparksee, InfoGrid, HyperGraphDB, FlockDB, VelocityGraph, GlobalsDB, GraphDB, etc.

NoSQL

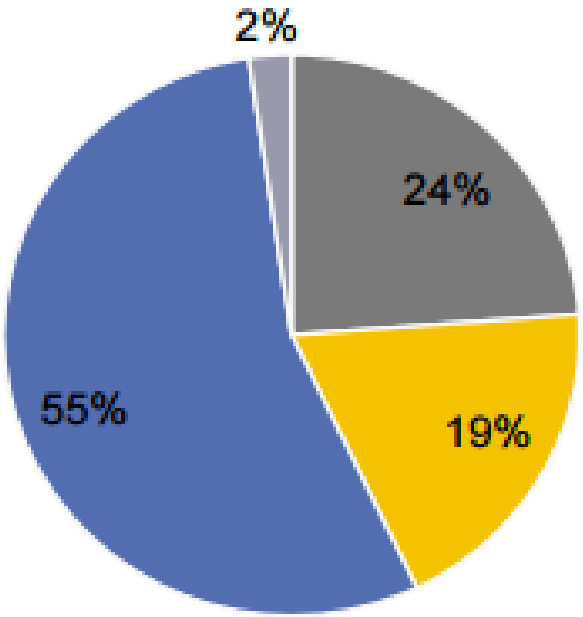
Architectures

- Solutions hybrides qui ont été conçues pour répondre à des besoins spécifiques

- **Multi-modèles :** Amazon DynamoDB, Microsoft Azure Cosmos DB, Datastax Enterprise, Ignite, ArangoDB, Apache Drill, Virtuoso, etc.

NoSQL

- ***Types BD majoritairement utilisées***



- Orientée colonnes (Cassandra, HBase...)
- Orientée clé-valeur (Redis, Riak...)
- Orientée documents (MongoDB, Couchbase, ElasticSearch...)
- Orientée graphes (Neo4J, Titan...)

NoSQL

Classement de popularité SQL et NoSQL

Top 30 (Janvier 2021)

Mesurée à base :

- Nombre de citations dans les sites Web
- Intérêt général au système
- Fréquence des discussions techniques sur le système
- Nombre d'offres d'emploi dans lesquelles le système est mentionné
- Nombre de profils dans les réseaux professionnels, dans lesquels le système est mentionné
- Pertinence dans les réseaux sociaux

- **Document store (3)**
- **Column family (2)**
- **Key-value store (2)**
- **Graph store (1)**
- **Multi-modèles (2)**

361 systèmes dans le classement, janvier 2021

Rang jan 2021	Rang dec 2020	Rang Janv 2020	SGBD	Modèle de base de données	But jan 2021	But dec 2020	But Janv 2020
1.	1.	1.	Oracle	Relationnel , multimodèle	1322,93	-2,66	-23,75
2.	2.	2.	MySQL	Relationnel , multimodèle	1252,06	-3,40	-22,60
3.	3.	3.	Microsoft SQL Server	Relationnel , multimodèle	1031,23	-6,85	-67,31
4.	4.	4.	PostgreSQL	Relationnel , multimodèle	552,23	+4,65	+45,03
5.	5.	5.	MongoDB	Document , multimodèle	457,22	-0,51	+30,26
6.	6.	6.	IBM Db2	Relationnel , multimodèle	157,17	-3,26	-11,53
sept.	sept.	↑ 8.	Redis	Valeur-clé , multi-modèle	155,01	+1,38	+6,26
8.	8.	↓ sept.	Elasticsearch	Moteur de recherche , multimodèle	151,25	-1,24	-0,19
9.	9.	↑ dix.	SQLite	Relationnel	121,89	+0,21	-0,25
dix.	dix.	↑ 11.	Cassandra	Colonne large	118,08	-0,76	-2,59
11.	11.	↓ 9.	Microsoft Access	Relationnel	115,33	-1,41	-13,24
12.	12.	↑ 13.	MariaDB	Relationnel , multimodèle	93,79	+0,18	+6,34
13.	13.	↓ 12.	Splunk	Moteur de recherche	87,66	+0,66	-1,01
14.	14.	↑ 15.	Teradata	Relationnel , multimodèle	72,59	-1,24	-5,70
15.	↑ 16.	↑ 25.	Base de données Microsoft Azure SQL	Relationnel , multimodèle	71,36	+1,87	+43,16
16.	↓ 15.	↓ 14.	Ruche	Relationnel	70,43	+0,16	-13,81
17.	17.	↓ 16.	Amazon DynamoDB	Multi-modèle	69,14	+0,01	+7,11
18.	18.	↑ 20.	SAP Adaptive Server	Relationnel	54,61	-0,27	+0,02
19.	19.	↑ 22.	Neo4j	Graphique	53,79	-0,85	+2,13
20.	↑ 21.	↓ 17.	Solr	Moteur de recherche	52,48	+1,24	-4,08
21.	↓ 20.	↓ 19.	SAP HANA	Relationnel , multimodèle	50,87	-1,63	-3,82
22.	22.	↓ 18.	FileMaker	Relationnel	47,39	-0,31	-7,72
23.	23.	↓ 21.	HBase	Colonne large	46,28	-0,64	-7,06
24.	24.	↑ 26.	Google BigQuery	Relationnel	36,00	+0,24	+9,24
25.	25.	↓ 24.	Microsoft Azure Cosmos DB	Multi-modèle	32,97	-0,57	+1,46
26.	26.	↓ 23.	Couchbase	Document , multimodèle	31,63	-0,19	-0,41
27.	27.	↑ 32.	InfluxDB	Des séries chronologiques	26,32	+0,17	+5,18
28.	28.	28.	Memcached	Valeur clé	25,97	+0,08	+0,86
29.	↑ 31.	↑ 31.	Amazon Redshift	Relationnel	22,92	+0,50	+1,33
30.	↓ 29.	30.	Oiseau de feu	Relationnel	22,75	-0,08	+0,20
31.	↓ 30.	↓ 27.	Informix	Relationnel , multimodèle	22,45	-0,19	-2,68
32.	32.	↓ 29.	Vertica	Relationnel , multimodèle	21,21	-1,00	-1,44
33.	↑ 34.	33.	Netezza	Relationnel	19,00	+0,67	-0,62
34.	↓ 33.	↑ 35.	Spark SQL	Relationnel	18,74	-0,55	+1,61
35.	↑ 36.	↓ 34.	CouchDB	Document	16,35	-0,51	-2,02

NoSQL

- ***Classement de popularité Top 10*** ***(Janvier 2021)***

Système	Classement NoSQL	Classement général
MongoDB (D)	1	5
Redis (V)	2	7
Cassandra (C)	3	10
AmazonDB (M)	4	17
Neo4j (G)	5	19
HBase (C)	6	23
Microsoft Azure Cosmos DB (M)	7	25
Couchbase (D)	8	26
Memcached (V)	9	28
CouchDB (D)	10	35

NoSQL

Faiblesses ou insuffisances

- Non-existence de langages de requête normalisés tels que SQL
- Manque d'architectures et d'interfaces normalisées
- Cohérence de données abandonnée relativement au profit de la haute disponibilité
- Sécurité des données : Les systèmes NoSQL se base sur l'application pour la protection des données
- Maturité et stabilité : Les bases relationnelles ont une longueur d'avance sur ce point. Les utilisateurs sont familiers avec leur fonctionnement et ont confiance en elles
- Moins de documentation et d'outils disponibles

Contre Attaque NewSQL

NewSQL

Motivation

- Cherche à regrouper les avantages du NoSQL et du SQL
- Exploiter les atouts du modèle relationnel (Fondements, SQL, Acidité)
- Architecture qui reprend les avantages du NoSQL et comble son principal désavantage par une éventuelle cohérence des données
- Fournir au modèle relationnel les avantages d'extensibilité horizontale et la tolérance aux pannes assurées par les solutions NoSQL
- Gérer les énormes volumes de données

NewSQL

Définition

- NewSQL est une nouvelle architecture logicielle qui propose de repenser le stockage des données pour prendre en charge les masses d'informations
- Catégorie de bases de données modernes, qui a émergé du monde NoSQL mais reste différent sur plusieurs aspects
- Modèle de stockage distribué potentiellement en mémoire qui peut être requêté classiquement par une interface SQL
- Elle profite des architectures distribuées et des évolutions sur le plan du matériel pour coller aux nouvelles tendances
- Certains auteurs : « **Scalable RDBMS** »

NewSQL

Architecture

- Cette nouvelle architecture est née du croisement de 3 types d'architectures, relationnelle, non-relationnelle et grille de données appelée cache distribuée
- S'appuyer sur un stockage distribué issu des architectures NoSQL, pour supporter des accès transactionnels à fort débit
- Utiliser la distribution et la réplication des données pour assurer la scalabilité et la disponibilité des données
- La plupart des solutions NewSQL proposent un stockage en mémoire distribué sur plusieurs machines sous forme de grille de données (Base de données en mémoire)
- Une architecture distribuée fournissant de meilleures performances par rapport aux solutions classiques de type SGBD Relationnel

NewSQL

Caractéristiques

- Conserver le modèle relationnel au cœur de son système
- Se base sur le SQL comme langage commun de requêtage
- Un schéma relationnel avec des limitations pour faciliter la distribution des données et des traitements
- Supporte les contraintes ACID
- Adopte un mécanisme qui évite d'imposer de verrous lors d'opérations concurrentes de lecture avec les opérations d'écritures
- Architecture sans maître et capable de tourner sur un nombre important de nœuds sans affecter de goulot d'étranglement

NewSQL

Avantages

- Scalabilité horizontale
- Lecture en temps réel plus facile
- Temps de réponse rapide
- Meilleures performances par rapport aux SGBDR
- Gestion de gros volumes de données hétérogènes
- Conserve les concepts du modèle relationnel et les propriétés ACID
- Solution économique

NewSQL

Intérêt

- BD NewSQL destinées aux entreprises qui gèrent des données sensibles et stratégiques

- BD qui requièrent une certaine évolutivité, mais aussi une cohérence supérieure à ce qu'apportent les bases NoSQL

NewSQL

- *Catégories*

NewSQL

- ***Leaders de la technologie***
  - VoltDB
  - Spanner
  - ClustrixDB
  - NuoDB
  - TransLattice Elastic Database
  - SQLFire
  - GridGain IMDB
  - Drizzle
  - MemSQL

NewSQL

Inconvénients

- Non encore normalisés autant que les systèmes SQL traditionnels
- Les architectures en mémoire peuvent être inappropriées pour des volumes dépassant quelques téraoctets et pour des machines moins puissantes
- Offre un accès partiel aux outils riches des SGBDR

Hadoop

Hadoop

Présentation

- **Implantation de référence** des Big Data et du Cloud Computing
- Hadoop n'est pas un sigle : ***Doug Cutting*** s'inspira de la pelluche de son fils de 3 ans, un éléphant jaune, pour le logo ainsi que pour le nom de ce nouveau framework

- Mise en œuvre du modèle de programmation **MapReduce**
- Framework Java libre conçu pour la création d'applications distribuées et extensibles
- Les serveurs du futur disposeront d'une couche supplémentaire de logiciels de base Hadoop

Hadoop

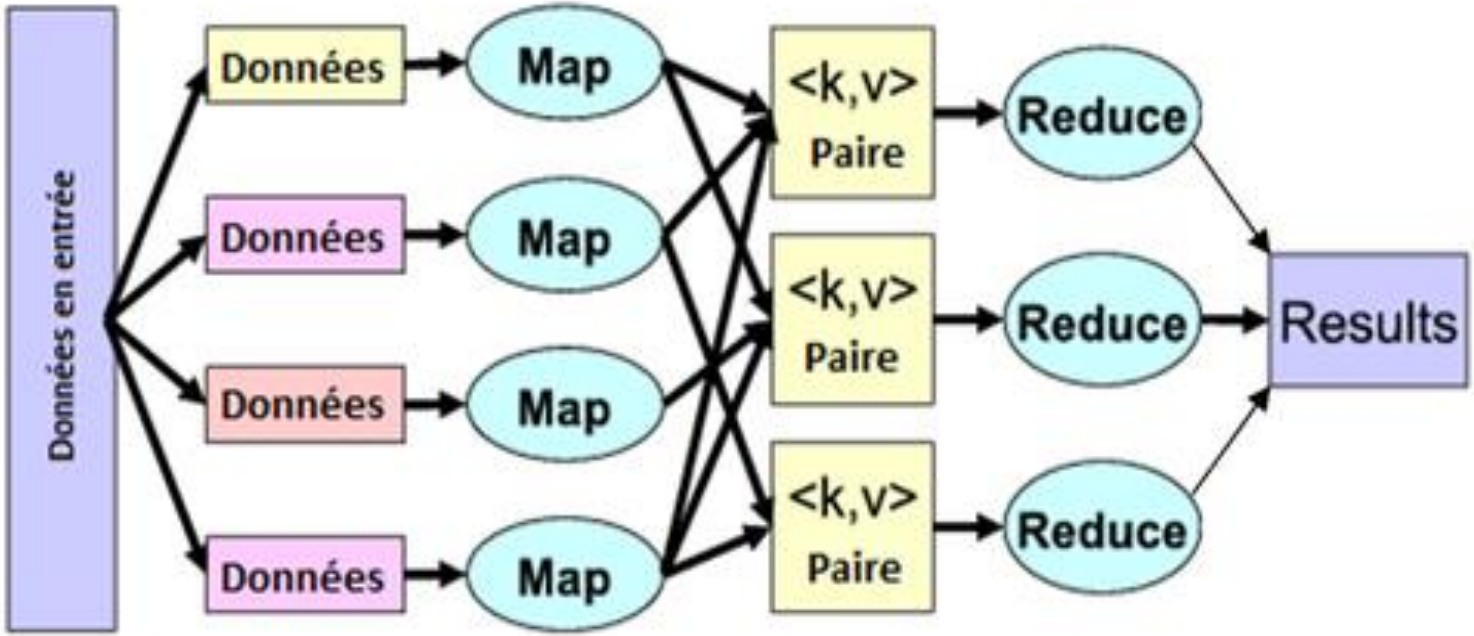
Architecture

- Mode d'implantation en partie niveau OS et en partie à l'extérieur
- La partie la plus proche du SE est le système de gestion de fichiers HDFS
- Plusieurs distributions : Cloudera, MapR Technologies, Hortonworks
- Architecture : HBase, ZooKeeper, Hive, Pig, Sqoop, **MapReduce**, **HDFS**

Hadoop

MapReduce

- Mc
  et
- Po
- Pe
  (pl
- Co
  - M
  - Re
- Poi



arallèles

e nœuds

s nœuds

licités

Hadoop

Hadoop Distributed File System (HDFS)

- La partie gestion des données de cette implantation est ***Hadoop Distributed File System***
- HDFS est un système de fichiers distribué, extensible et portable développé par Hadoop à partir du GFS (Google File System)
- Écrit en Java, conçu pour stocker de très gros volumes de données sur un grand nombre de machines
- Permet l'abstraction de l'architecture physique de stockage, afin de manipuler un système de fichiers distribué comme s'il s'agissait d'un disque dur unique

Architecture HDFS

- **Composants d'Hadoop DFS**
  - ***Job tracker*** : un aiguilleur de tâches centralisé
  - ***Task trackers*** : des gestionnaires de traitements locaux aux serveurs
  - ***Data node*** : les nœuds de stockage
  - ***Name node*** : un serveur central de métadonnées
  - ***Secondary Name node*** : un serveur de secours

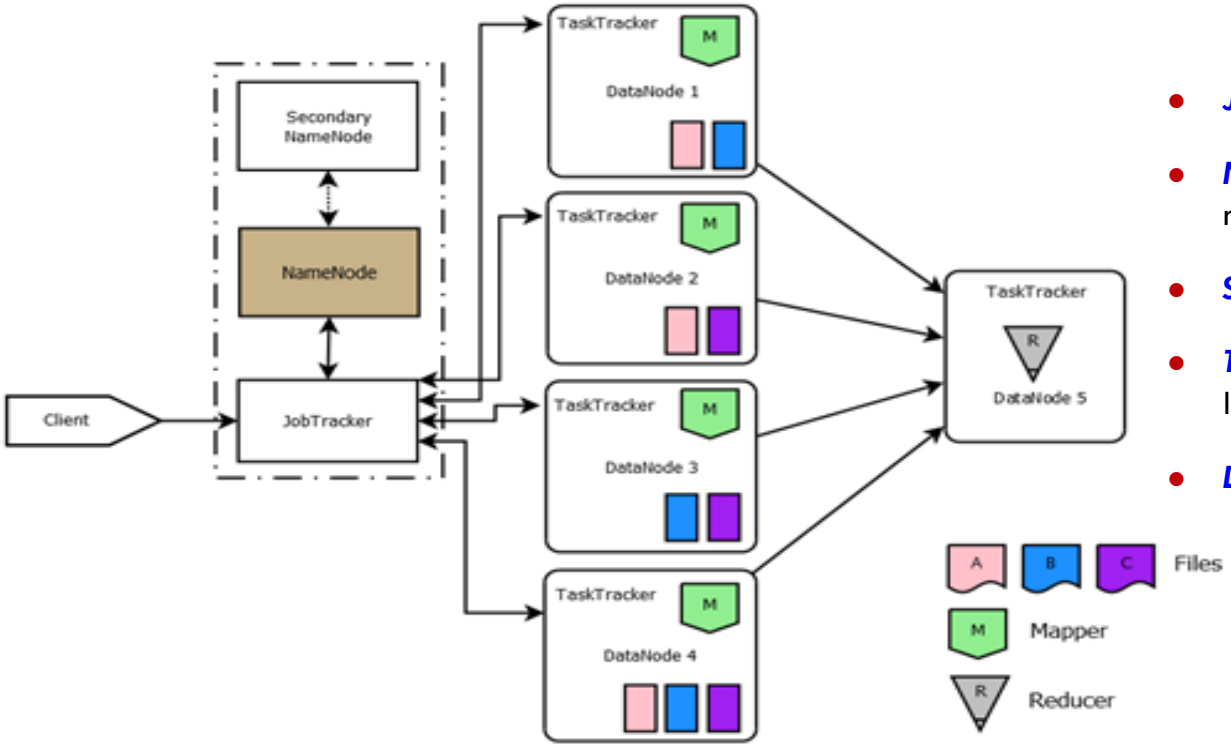
Traitement des métadonnées dans HDFS

- **Service de métadonnées centralisé**
  - HDFS adopte un service de métadonnées séparé avec isolation et centralisation du service
  - La centralisation est motivée par la garantie de l'intégrité des métadonnées et la fiabilité du service

- **Séparation des chemins des données et des métadonnées**
  - Les données circulent parallèlement entre clients et *Data nodes*
  - Le Job Tracker récupère auprès Name Node le layout des blocs composant le fichier accédé (lise des Data nodes)
  - Goulot d'étranglement d'accès aux données est dégagé par un parallélisme massif
  - Le *Name Node* reste à l'écart du circuit des données

Hadoop

Architecture HDFS



Typical Components of a Hadoop Cluster (Menon, 2013)

- ***Job tracker*** : Aiguilleur de tâches centralisé
- ***Name node*** : Serveur central de métadonnées
- ***Secondary Name node*** : Serveur de secours
- ***Task trackers*** : Gestionnaires de traitements locaux
- ***Data node*** : Nœuds de stockage

Hadoop

Approche Hadoop vs Approche BD Rel

	BD Rel	Hadoop
Données	Conçues pour données structurées	Conçues pour données non structurées
Schéma des données	Statique	Inexistant ou dynamique
Volume des données	Ordre de plusieurs Go	Ordre de plusieurs Po
Mise à jour des données	Fréquente	Ecriture unique (WORM)
Interface de requêtage	SQL	Pg Mapreduce en Java, SQL+,C++,HiveSql,..
Type d'accès	Interactif et traitement par lots	Traitement par lots
Granularité de travail	Requête entière	Selon la taille de données
Force principale	Grande performance pour requêtage sur les données struct.	Extensibilité en volumétrie et en nombre de nœuds de traitement
Schéma d'infrastructure matérielle	Centralisée	Distribuée
Extensibilité	Verticale (Scale-up)	Horizontale (Scale-out)

Conclusion

- ***Big Data, Cloud Computing : Recherches dans le domaine***

  - Sujets à la mode et d'actualité
  - Domaine très dynamique et prometteur
  - Tous les travaux dans le contexte seront très appréciés
  - Beaucoup de journaux et conférences sont focalisés sur le sujet
  - Opportunités très intéressantes de publications

Big Data, Cloud Computing
Projets de Recherches Réalisés

- ***PFE Master***
  - **Etudes comparatives NoSQL**
    - 2014 – 2015 : MongoDB vs Hbase
    - 2015 – 2016 : MongoDB, Couchbase, Cassandra, HBase, Redis, OrientDB
  - **Etudes comparatives NoSQL vs SQL**
    - 2016 – 2017 : MongoDB vs MySQL
  - **Etudes comparatives NewSQL vs SQL**
    - 2016 – 2017 : VoltDB vs MySQL

PFE Master

- **Migration : MySQL vers MongoDB**
  - 2014 – 2015 : BD d'un réseau social
- **Migration : Oracle vers MongoDB**
  - 2015 – 2016 : BD des passeports biométriques

Big Data, Cloud Computing

Projets de Recherches Réalisés

- ***PFE Master (Co-tutelle Univ Tlemcen - Univ Poitiers France)***

- 2017-2018 : Optimisation des requêtes dans des environnements parallèles : Application au Big Data
- 2018-2019 : **AUTOPILOT**-AUTOmatisation de l'identification et de la recommandation des paramètres responsables de l'amélioration du départ des PILOTes élites de la BMX Race
- 2019-2020 : Contribution au développement d'une nouvelle approche générique de distribution de big data
- 2020-2021 : Automatisation de la recommandation de compétences pour les offres d'emplois

Big Data, Cloud Computing

Projets de Recherche en Cours

- ***Doctorat (Co-tutelle Univ Tlemcen - Univ Poitiers France)***

  - 2017-2018 : Gestion personnalisée des Big Data : vers un cadre applicatif générique
  - 2018-2019 : Traitement des Big Data : vers une gestion parallèle, optimale et déclarative
  - 2018-2019 : Traitement des Big Data : vers une intégration de la qualité lors de l'évaluation des requêtes
  - 2020-2021 : Vers un système modulaire pour la gestion des BIG DATA

Big Data, Cloud Computing
Projets de Recherches Réalisés

Publications

- **« Towards a New Model of Storage and Access to Data in Big Data and Cloud Computing »**
  - *International Journal of Ambient Computing and Intelligence (**IJACI**) – Décembre 2017*
- **« Experimental Comparative Study of NoSQL Databases : HBase versus Mongodb by YCSB »**
  - *International Journal on Computer Systems Science & Engineering (**IJCSSE**) – Juillet 2017*
- **« Evaluation of NoSQL Databases: MongoDB, Cassandra, HBase, Redis, Couchbase, OrientDB »**
  - *International Journal of Software Science and Computational Intelligence **(IJSSCI)**– Oct, Dec 2020*
- **« Comparative Study Between the MySQL Relational Database and the MongoDB NoSQL Database »**
  - *International Journal of Software Science and Computational Intelligence **(IJSSCI)**– Juin 2021*