16 BIG DATA

1 RAISONS DE L'ESSOR DU BIG DATA

• Explosion des capacités de stockage et de calculs des ordinateurs

• Augmentation exponentielle du nombre de données générées et stockées

• Augmentation exponentielle du nombre de données

• Développement des technologies de virtualisation et du cloud computing

• Limites de la loi de Moore :

— la loi de Moore stipule que les ordinateurs deviennent au fil du temps plus petits, plus rapides, moins chers, à mesure que les transistors sur circuits intégrés deviennent plus efficaces

— il est de plus en plus difficile de réduire la taille des transistors et de les disposer toujours plus densément sur une puce

— cette limite physique a conduit à un arrêt de la croissance exponentielle de la puissance de calcul et a ainsi poussé à l'exploration de nouvelles méthodes pour traiter les données massives, comme celles utilisées dans le big data

• Génération sans cesse croissante de données, passant de 2 zettaoctets en 2010 (avec 80% de la donnée générée à usage humain) à 181 zettaoctets en 2020 (avec 18% de la donnée générée à usage humain)

• Augmentation de l'utilisation de données semi ou non-structurées

• Données de plus en plus connectées

• Limitations des systèmes classiques pour stocker, traiter et analyser des données massives, variées et changeantes

2 DÉFINITION DU BIG DATA

• Concept datant des années 70 avec les premiers datacenters

• **Big Data :** ensemble de technologies et de méthodes permettant de stocker, traiter et analyser des données massives, variées et changeantes à un coup financier, humain et temporel raisonnable

• **Tout est information :** principe que l'homme et la totalité du monde qui l'entoure peuvent être représentés comme des ensembles informationnels, dont la seule différence avec la machine est leur niveau de complexité, la vie deviendrait alors une suite de 0 et de 1, programmable et prédictible

• **Données structurées vs données non structurées :**

Données structurées	Données non structurées
Affichables sous forme de lignes, colonnes et dans des bases de données relationnelles	Non affichables sous forme de lignes, colonnes et dans des bases de données relationnelles
Nombres, dates et chaines de caractères	Images, fichiers audio, vidéo, traitement de texte, emails, feuilles de calcul
Estimées à 20% de données d'entreprise	Estimées à 80% de données d'entreprise
Nécessitent moins d'espace de stockage	Nécessitent plus d'espace de stockage
Plus faciles à gérer et à protéger avec des solutions héritées	Plus difficiles à gérer et à protéger avec des solutions héritées

• **Deux concepts clés du Big Data :**

— **le parallélisme et la distribution des algorithmes :** le traitement des données massives est réalisé par plusieurs algorithmes en même temps, chacun travaillant sur une partie différente des données afin de réduire le temps de traitement et d'optimiser la performance

— **la distribution du stockage des données :** les données massives sont stockées sur plusieurs machines pour garantir la disponibilité, la rapidité et la sécurité de l'accès et pour mieux gérer les pannes et minimiser les pertes de données

• **Domaines d'application du Big Data :** commerce, finance, santé, énergie, transport, agriculture, etc.

• Le Big Data est un domaine technique qui croise plusieurs spécialités de l'informatique :

— **informatique transactionnelle :** se concentre sur la gestion des transactions :

> une transaction est un ensemble d'opérations qui doivent être effectuées ensemble et qui doivent être toutes ou aucune effectuées

> les principes ACID (Atomicité, Cohérence, Isolation, Durabilité) sont importants dans ce domaine

— **informatique décisionnelle (ou BI) :** se concentre sur l'analyse des données pour aider les entreprises à prendre des décisions éclairées, cela inclut des outils pour la collecte, le stockage, l'analyse et la visualisation des données

— **informatique en temps réel :** se concentre sur le traitement de données qui doivent être traitées immédiatement (surveillance de systèmes, détection d'anomalies et sécurité informatique)

— **stockage et tri de données :** les données doivent être stockées de manière efficace et triées pour être accessibles rapidement (bases de données relationnelles et non relationnelles, systèmes de fichiers distribués, etc.)

— **traitement et analyse des données :** catégorisation, synthèse, prédiction et représentation des données

3 CARACTÉRISTIQUES DU BIG DATA : LES 9 V

• **Volume :** masse croissante de données qui nécessite des besoins spécifiques pour les stocker, les transporter et les analyser

• **Vitesse :** rapidité avec laquelle les données sont générées, traitées ou modifiées, répondant aux besoins des processus chronosensibles (bourses, stream, etc.) avec le risque pour l'Homme de perdre le contrôle sur les données

• **Variété :** diversité des types de données, dont 20% de données structurées et 80% de données semi ou non structurées

— **données structurées :** possèdent une structure prédéfinie (tableau, fichier, etc.) et observables en tableau ou base de données relationnelles traditionnelles

— **données non-structurées :** ne possèdent pas de structure prédéfinie, sont stockées dans leurs formats en mode natif (vidéo, audio, etc.), sont qualitatives et non quantitatives

— **données semi-structurées :** possèdent une structure, mais pas fixe ou prédéfinie, et des métadonnées typées (XML, JSON, etc.)

• **Volatilité :** durée de vie d'une donnée, c'est-à-dire le temps pendant lequel cette donnée est pertinente et utile

Remarque : Pour gérer la volatilité des données, il est important d'estimer la durée de vie d'une donnée afin de déterminer quand elle sera obsolète et de prévoir son traitement et sa prise en charge en conséquence.

• **Valeur :** profit que l'on peut tirer d'une donnée

Remarque : Pour extraire de la valeur d'une donnée, il est souvent nécessaire de réaliser des opérations de traitement, comme le regroupement, les filtres, la classification ou la hiérarchisation.

• **Vulnérabilité :** sécurité des données : nécessité de mettre en place une structure sécurisée pour protéger les données contre les pirates informatiques

• **Véracité :** fiabilité et confiance des données, c'est-à-dire la mesure dans laquelle les données sont vraies et précises, nécessitant des outils de vérification et de validation comme le recoupement et l'enrichissement des données

• **Validité :** conformité et précision des données, c'est-à-dire la mesure dans laquelle les données sont correctes, pertinentes et représentatives du phénomène qu'elles sont censées décrire

Remarque : Environ 60% du temps d'un scientifique est consacré au nettoyage des données avant analyse pour s'assurer de leur validité.

• **Visibilité :** capacité à avoir une vision précise et claire des données

— Cette capacité diminue à mesure que le volume de données augmente.

— Des outils spécialisés sont nécessaires pour améliorer la visibilité des données.

— La manière dont les données sont représentées (couleur, forme, etc.) peut avoir un impact sur la façon dont l'analyse de la donnée est perçue.

4 TECHNOLOGIES DU BIG DATA

• Pour répondre aux besoins du Big Data, des systèmes permettant de dépasser les limites des systèmes traditionnels ont été développés.

4.1 Architectures scalables

• **Scalabilité :** adaptation de la taille et/ou de la puissance d'un système informatique pour répondre aux changements de la charge de travail

• **Scalabilité verticale :** augmentation des ressources internes : augmentation de la puissance (processeur, RAM, stockage) d'un système

— solution la plus simple et la plus rapide à mettre en œuvre

— fréquemment utilisée dans les systèmes traditionnels

— augmentation exponentielle du cout du matériel

— faible adaptabilité aux changements de la charge de travail

— problèmes en cas de panne (single point of failure (élément d'un système informatique ou d'une infrastructure qui, s'il échoue, peut entrainer une panne complète ou significative du système.))

• **Scalabilité horizontale :** augmentation des ressources externes : augmentation du nombre de machines de faible puissance pour augmenter la puissance globale

— solution économique la plus adaptée

— pseudo-linéarité des performances

— découpages et réplications des données

— augmentation exponentielle des échanges

— architectures de réseaux complexes (cluster)

— synchronisation des données

4.2 Solutions de stockage pour les données non structurées

• L'avènement du Big Data a nécessité l'adaptation des systèmes de stockage pour s'adapter à la quantité et aux types de données.

• **Théorème de CAP (triangle de CAP) :** tout système ne peut garantir que 2 des 3 propriétés suivantes :

— **Cohérence :** chaque lecture de données renverra la valeur la plus récente et que chaque écriture de données sera propagée à toutes les copies

— **Disponibilité :** le système est toujours accessible pour la lecture et l'écriture de données

— **Tolérance aux pannes :** le système continue à fonctionner même si certaines parties du réseau (ou des nœuds) ne peuvent pas communiquer entre elles en raison d'une panne du réseau ou d'autres problèmes

• **Limitations des bases de données relationnelles :**

— manque d'adaptabilité des schémas de données

— limitées aux données structurées (< 20%)

— dépenses massives en temps et ressources en cas de modifications de la structure des données

• Le NoSQL a été développé pour pallier les limitations des bases de données relationnelles.

• **Avantages du NoSQL :**

— prise en charge des données structurées, semi-structurées et structurées

— adaptabilité des schémas de données

• **Principes BASE :** utilisés à la place des principes ACID en NoSQL :

— **Basically available :** disponibilité des données à tout moment (réplication des données)

> Le système doit pouvoir répondre aux requêtes de tout utilisateur même en cas de pannes.

> Les requêtes peuvent être obsolètes : les données auxquelles les requêtes accèdent peuvent ne pas être à jour ou représenter l'état le plus récent des données.

— **Soft-state :** notion selon laquelle les données sont constamment dans un flux d'utilisation

> Les données peuvent être utilisées par plusieurs utilisateurs en même temps.

> La cohérence n'est pas garantie.

— **Eventual consistency :** notion selon laquelle, à un moment donné, le système parviendra à une cohérence des données

> La synchronisation des données est faite en arrière-plan.

> Les données peuvent être obsolètes : pendant le processus de synchronisation en arrière-plan, il peut y avoir un délai pendant lequel certaines copies des données peuvent ne pas être à jour par rapport à d'autres copies.

• **4 modèles de stockage NoSQL :**

— **Modèle clé/valeur :** les données sont organisées sous forme de paires clé/valeur

> Chaque donnée est associée à une clé unique qui permet de l'identifier, et elle peut être de n'importe quel type de données.

> Ex : Redis (StackOverFlow), Riak (GitHub), Memcached (Wikipédia), Voldemort (LinkedIn)

— **Modèle colonne :** les données sont stockées sous forme de table dénormalisée

> Ce modèle est similaire aux bases de données relationnelles, mais il est optimisé pour le stockage de grandes quantités de données et la lecture de colonnes spécifiques plutôt que de lignes entières.

> Les données sont organisées en familles de colonnes regroupées, où chaque famille peut contenir un ensemble de colonnes.

> Ex : Cassandra (Nasa), HBase (Facebook, Xiaomi), BigTable (GCP)

— **Modèle document :** les données sont stockées sous forme de documents au format JSON, avec un identifiant unique et des propriétés clé/valeur

> Les valeurs peuvent également être d'autres documents, ce qui permet de représenter des structures de données complexe.

> Ex : MongoDB (SEGA, ThermoFisher Scientific), CouchDB (CERN)

— **Modèle graphe :** les données sous forme de graphes, où les nœuds représentent des entités et les arcs représentent les relations entre ces entités

> Les nœuds et les arcs peuvent avoir des propriétés clé/valeur associée.

> Ex : Neo4j (Orange, Airbus), OrientDB, Titan

4.3 Architectures distribuables et massivement parallèles

• **Technologies de distribution :** les technologies du Big Data utilisent des architectures distribuées pour la répartition des stockages et des traitements

• **Distribution des stockages :** les données sont réparties et dupliquées sur plusieurs machines pour un stockage distribué, améliorant l'accessibilité des données, permettant une scalabilité horizontale et assurant la tolérance aux pannes

• **Stockage distribué :** différents niveaux de stockage distribué sont utilisés :

— la réplication du stockage

— la répartition des données

— le sharding : technique utilisée dans le domaine du stockage de données distribuées pour diviser les données en fragments plus petits appelés « shards » et les répartir sur plusieurs machines ou serveurs, chaque shard contient une partie des données totales, ce qui permet de répartir la charge de travail et d'optimiser les opérations de lecture et d'écriture

• **Réplication du stockage :** la réplication des données sur plusieurs machines améliore l'accessibilité des données, offre une scalabilité horizontale et assure la tolérance aux pannes

• **Trois types d'approches réplicatives :**

— le système maitre-esclave (postgreSQL, MongoDB) : une instance principale (maitre ou master) est responsable de la gestion des opérations d'écriture (insertions, mises à jour, suppressions) tandis que les instances de réplication (esclaves ou slaves) reproduisent ces opérations à partir du maitre, les requêtes de lecture peuvent être traitées à la fois par le maitre et les esclaves

— le système multi-maitre (CouchDB) :

> plusieurs nœuds agissent en tant que maitres, ce qui signifie qu'ils peuvent tous accepter les opérations d'écriture

> chaque nœud maitre se réplique vers les autres nœuds maitres, garantissant ainsi la cohérence des données

> les requêtes de lecture peuvent être traitées par n'importe quel nœud maitre

— le système sans maitre ou décentralisé (Cassandra) : il n'y a pas de distinction stricte entre maitres et esclaves : chaque nœud dans le cluster est capable de traiter les opérations d'écriture et de lecture, et les données sont réparties de manière équilibrée sur l'ensemble du système

4.4 Algorithmes distribués et parallèles

• **Architectures distribuables et massivement parallèles :**

— conçues pour répartir le traitement des données sur plusieurs nœuds ou serveurs

— permettent une exécution parallèle des opérations, accélérant ainsi le traitement

— particulièrement adaptées au traitement de grands volumes de données

— Les architectures massivement parallèles utilisent des milliers de nœuds de calcul pour une puissance de traitement encore plus importante.

• **Algorithmes distribués et parallèles :**

— conçus pour fonctionner sur des architectures distribuables et massivement parallèles.

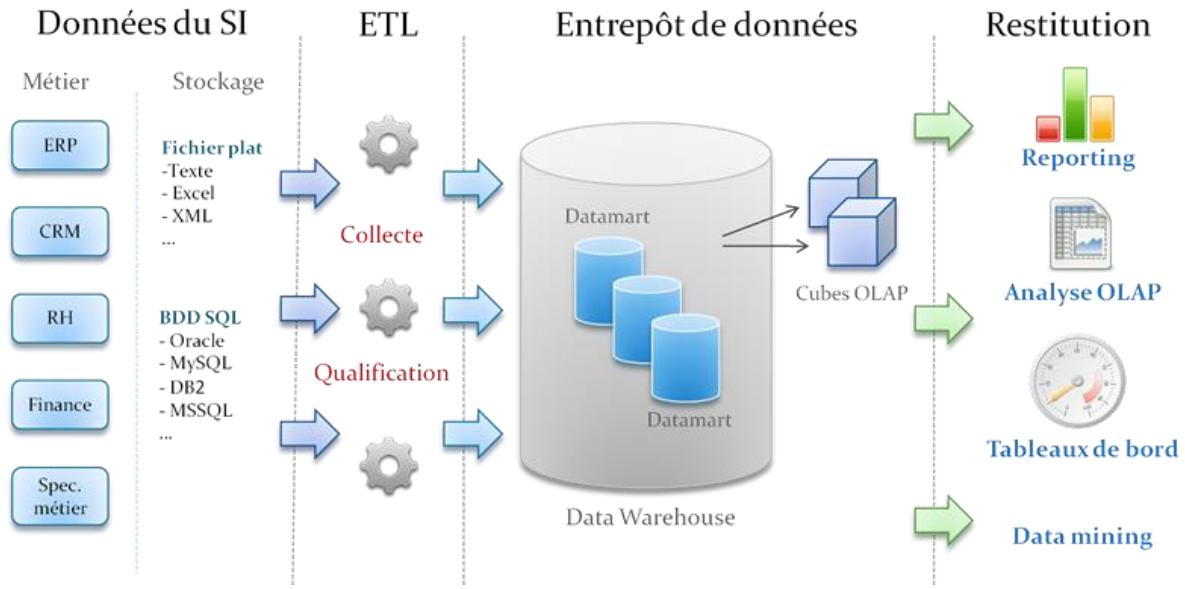
— divisent les tâches en sous-tâches pouvant être exécutées en parallèle sur différents nœuds de calcul

— accélèrent le traitement global en exploitant le parallélisme

— utilisés pour des opérations telles que le tri, la recherche, l'agrégation ou l'apprentissage automatique sur de grands ensembles de données

— Exemple de cette approche : algorithme MapReduce, développé par Google en 2005 et popularisé par le projet open-source Hadoop, permettant de réaliser des opérations massives sur des données complexes

5 ARCHITECTURE DE DONNÉES



• Quelques termes importants :

— **Datacenter :** installation physique qui abrite des serveurs, des équipements de stockage, des infrastructures réseau et d'autres composants nécessaires pour stocker, gérer et traiter les données d'une organisation

— Datahub : utilisé pour désigner différents concepts :

> emplacement centralisé où les données sont collectées, organisées et partagées au sein d'une organisation

> plateforme ou à une infrastructure logicielle qui facilite la gestion et l'accès aux données, permettant aux utilisateurs de rechercher, explorer et utiliser les données de manière efficace

— **Datalake :** architecture de stockage de données flexible et évolutive qui permet de stocker de grandes quantités de données brutes et non structurées provenant de diverses sources, afin de permettre des analyses et des traitements ultérieurs.

— **Datamart :** sous-partie d'un entrepôt de données qui est spécifiquement conçue pour répondre aux besoins analytiques d'un domaine ou d'un département spécifique au sein d'une organisation. Il contient des données agrégées, prétraitées et structurées pour faciliter les analyses et les rapports ciblés.

— **Datawarehouse (entrepôt de données) :** base de données centralisée qui consolide et intègre des données (généralement structurées, nettoyées et organisées en vue de l'analyse) provenant de différentes sources au sein d'une organisation, optimisée pour les opérations de requêtes analytiques et de génération de rapports

— **ETL (Extract, Transform, Load) :** processus d'extraction, de transformation et de chargement des données depuis différentes sources vers un entrepôt de données ou un datalake