Les fondements du Big Data

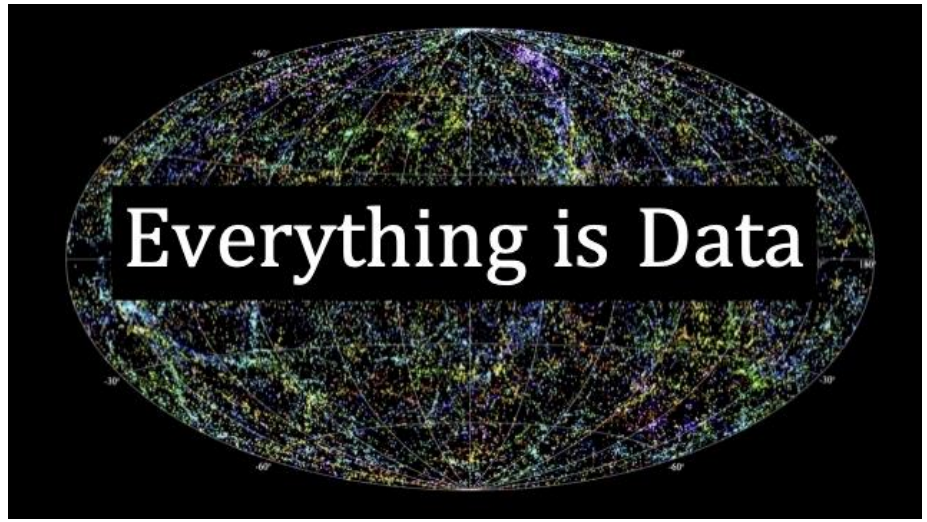
Mohamed El Marouani

TDIA 2

1. Qu'est ce que les données?
2. Types des données
3. Impact des données
4. Caractéristiques des données (les V)
5. Data Journey
6. Qu'est ce que Big Data?
7. Big Data: cas d'utilisation
8. Evolution du Big Data
9. Paysage du Big Data

Qu’est ce que les données ?

- Les **données (Data)** sont toutes les informations que vous collectez et qui ont été organisées et structurées de manière à pouvoir être analysées.
  « *Data are a collection of discrete or continuous values that convey information, describing the quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols that may be further interpreted formally* “ - Wikipedia
- Les données sont collectées à chaque fois que vous effectuez un achat, que vous naviguez sur un site web, que vous voyagez, que vous passez un appel téléphonique ou que vous publiez un message sur un site de médias sociaux.
- Les données peuvent provenir de nombreuses sources, notamment de capteurs, d'enquêtes, d'expériences, d'observations ou d'enregistrements existants (données historiques), comme les transactions financières.



Qu’est ce que les données ?

- Les organisations modernes considèrent les données comme **leur actif le plus précieux**, car elles fournissent des informations sur le comportement des clients, les tendances du marché, les performances des produits, etc. qui aident à prendre des décisions éclairées sur l'affectation des ressources.
- **La théorie de l'information** a poussé le concept de données beaucoup plus loin (Shannon, 1948). La théorie de l'information est un domaine d'étude qui cherche à comprendre la nature et l'origine de l'information et, selon cette étude, tout peut être considéré comme des données. **Cela inclut les objets physiques et les concepts abstraits tels que les idées ou les émotions**. En outre, les données sont définies comme tout ensemble de symboles qui transmettent un sens lorsqu'ils sont interprétés par un récepteur. Par conséquent, tout ce qui a une forme de **représentation symbolique** (par exemple, **des séquences d'ADN**, **des mots**, **des nombres**) peut être classé comme données dans ce contexte.

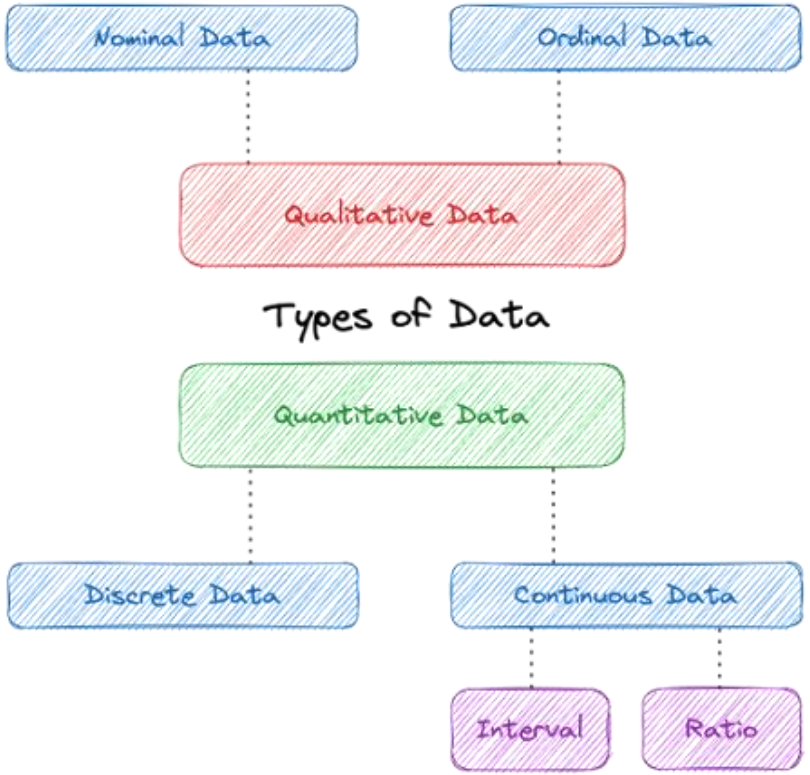
Types des données

D'un point de vue purement statistique, les données peuvent être classées en deux grandes catégories en fonction de leur valeur :

- **Les données quantitatives (numériques)** : il s'agit de toute information qui peut être exprimée, mesurée et comparée à l'aide de valeurs numériques, telles que des nombres entiers ou des nombres réels. Parmi les exemples de données quantitatives, on peut citer la taille, le poids, la longueur, les relevés de température, la taille d'une population ou des éléments dénombrables tels que le nombre d'élèves dans une salle de classe.
  Ce type de données peut être subdivisé en données **discrètes** (nombres entiers) ou **continues** (décimales).
  - **Données continues** : données quantitatives qui peuvent être divisées de manière significative en niveaux plus fins. Elles peuvent être mesurées sur une échelle ou un continuum. Elles peuvent avoir presque n'importe quelle valeur numérique : n'importe quelle valeur dans un intervalle fini ou infini (intervalle) ou une valeur qui compare deux nombres ou plus (rapport). Les exemples incluent la taille, le poids, la température, la vitesse, l'IMC et le temps.
  - **Données discrètes** : consistent en des valeurs finies, numériques et dénombrables. Les valeurs discrètes ne peuvent pas être divisées en parties. Les variables discrètes comprennent les dénombrements (par exemple, le nombre d'enfants dans un ménage), le nombre total de produits ou les indicateurs binaires (oui/non, vrai/faux).

Types des données

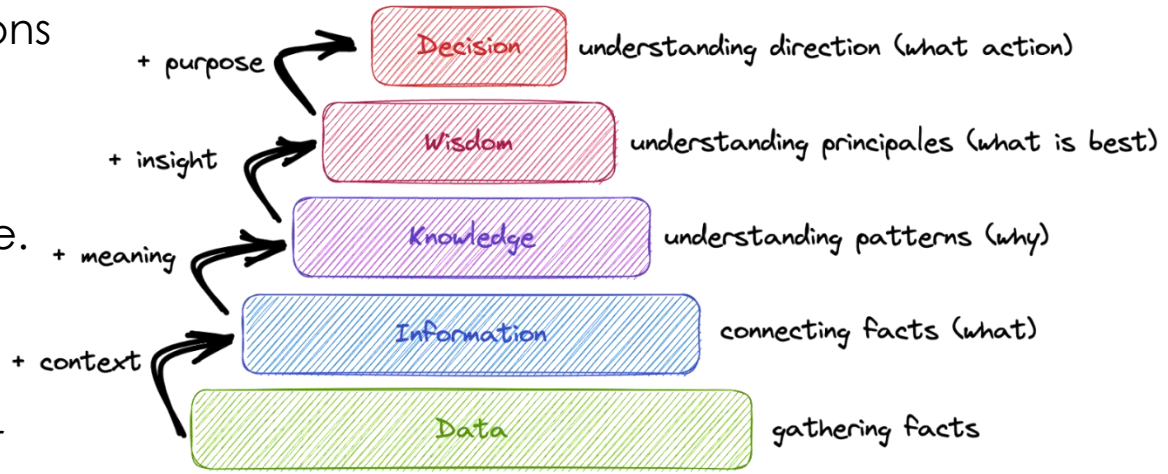
- **Données qualitatives (catégorielles)** : il s'agit d'informations non numériques telles que les opinions, les sentiments, les perceptions et les attitudes. Ces données peuvent répondre à des questions telles que : « Comment cela s'est-il produit ? » ou "Pourquoi cela s'est-il produit ?". Parmi les exemples de données qualitatives, on peut citer le sexe, les classements, les dénombrements, etc. Ce type de données peut être divisé en deux catégories : **les données nominales** et **les données ordinales**.
    - **Données nominales** : un type de données catégoriques qui n'a pas de valeur numérique ou d'ordre. Il s'agit de noms, d'étiquettes ou de catégories qui classent et organisent les informations en groupes distincts. Les exemples incluent le sexe (homme/femme), la nationalité (marocain/français) et les couleurs (vert/bleu).
    - **Données ordinales** : ce type de données est associé à un ordre ou à un classement. Les exemples incluent les classements tels que 1er, 2ème et 3ème ; les notes telles que A+, B- et C/D ; et les notes élevées, moyennes et basses.



Impact des données

Le **modèle DIKW** décrit la relation entre les données, l'information, la connaissance et la sagesse.

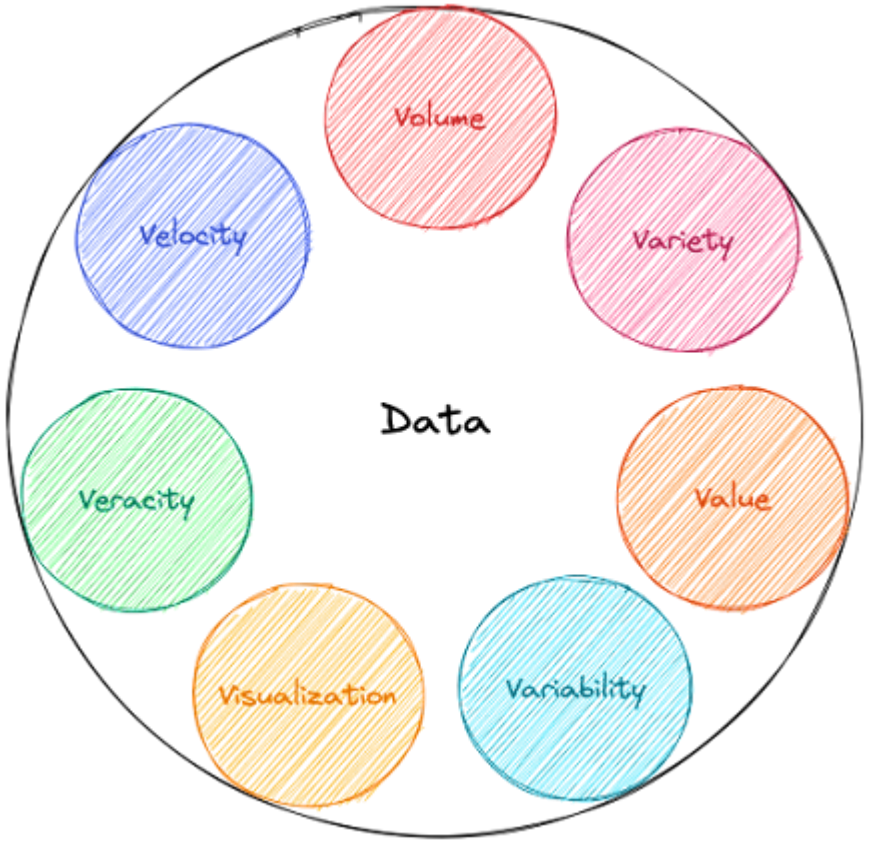
- **Données / Data :**
  - Considérées comme la matière première d'une prise de décision avisée.
  - Fournissent une base objective pour tirer des conclusions ou prendre des décisions.
- **Information / Information :**
  - Issue de l'analyse des données à l'aide de méthodes comme les statistiques ou l'apprentissage automatique.
  - Permet de découvrir des schémas auparavant non évidents.
- **Connaissance / Knowledge :**
  - Résulte de la transformation des informations en savoir structuré.
  - Sert de base aux processus de prise de décision.
- **Sagesse / Wisdom :**
  - Implique l'application des connaissances avec expérience et jugement.
  - Permet de prendre des décisions éclairées sur les stratégies et actions futures.



Caractéristiques des données (les V)

Les cinq caractéristiques principales et innées des données sont :

- **Volume :** La quantité de données qu'une organisation génère et stocke.
- **Vélocité :** la vitesse à laquelle les données sont générées et la vitesse à laquelle elles se déplacent et peuvent être traitées pour en tirer des informations exploitables.
- **Variété :** la diversité des données. Les organisations peuvent collecter des données à partir de sources multiples, dont le format peut varier. Les données collectées peuvent être structurées, semi-structurées ou non structurées.
- **Véracité :** fait référence au niveau de confiance et de fiabilité des données collectées. En d'autres termes, il s'agit de la qualité et de l'exactitude des données. Les données collectées peuvent comporter des éléments manquants, être inexactes ou ne pas être en mesure de fournir une valeur réelle.
- **Valeur :** se réfère à la valeur que les données peuvent apporter sur ce que les organisations peuvent en faire. Cette caractéristique est directement liée à la signification et au contexte qu'une organisation peut donner aux données collectées.

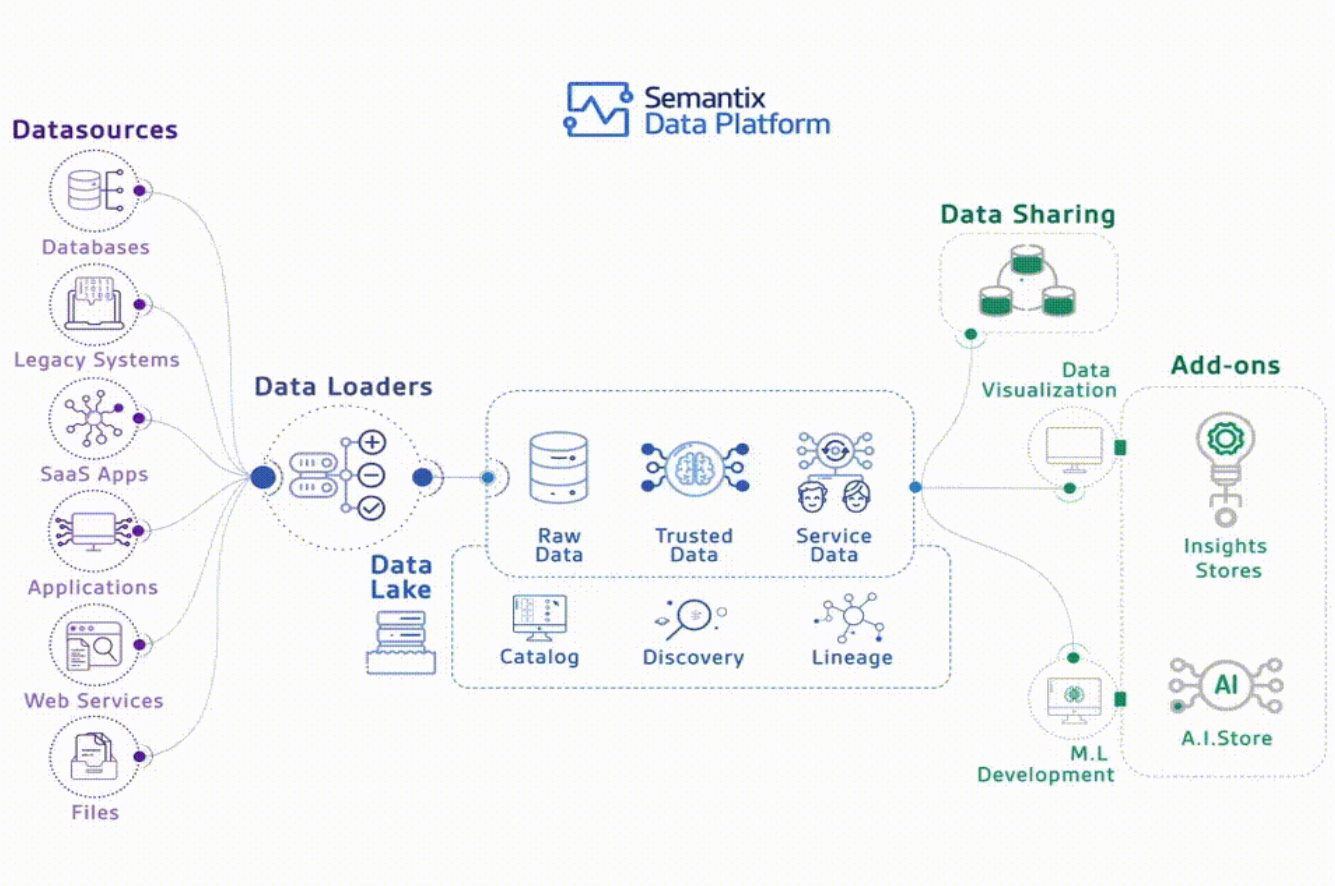


Caractéristiques des données (les V)

Dans le domaine du marketing, les experts en la matière ont commencé à utiliser deux caractéristiques supplémentaires qui ne sont pas innées aux données mais qui peuvent avoir un impact significatif sur les informations générées à partir de celles-ci. Ces deux caractéristiques sont les suivantes

- **Variabilité :** une mesure de la variation des valeurs dans chaque variante de données. Ce concept est lié au contexte des données et à la signification qui leur est donnée. Dans une organisation, la signification peut changer constamment, ce qui a un impact significatif sur l'homogénéisation des données. Ce concept diffère de celui de variété : Imaginez un café qui propose six mélanges de café différents (c'est la variété), mais si vous prenez le même mélange tous les jours. Il a un goût différent chaque jour ; c'est la variabilité.
- **Visualisation :** La visualisation est essentielle dans le monde d'aujourd'hui. L'utilisation de tableaux et de graphiques pour visualiser de grandes quantités de données complexes est beaucoup plus efficace pour transmettre du sens que des données brutes dans des feuilles de calcul remplies de chiffres et de formules.

Data Journey



Data Journey (parcours des données) se fait en plusieurs étapes. Les principales sont **l'ingestion**, **le stockage**, **le traitement** et **la distribution**. Chaque étape comporte son propre ensemble d'activités et de considérations. Le parcours des données comporte également une notion d'activités « sous-jacentes », c'est-à-dire des activités critiques tout au long du cycle de vie. Il s'agit notamment de la sécurité, de la gestion des données, du DataOps, de l'orchestration et de l'ingénierie logicielle.

Data Journey

1 - L'ingestion des données

L'ingestion des données est la première étape du cycle de vie des données. C'est à ce stade que les données sont **collectées** à partir de diverses **sources internes** telles que les bases de données, les systèmes de gestion de la relation client (CRM), les systèmes d'information de gestion (ERP), les systèmes existants, les **sources externes** telles que les enquêtes et les fournisseurs tiers. Il est important de s'assurer que les données acquises sont **exactes** et à jour afin de pouvoir les utiliser efficacement dans les étapes suivantes du cycle.

À ce stade, **les données brutes sont extraites** d'une ou de plusieurs sources de données, répliquées, puis intégrées dans un support de stockage d'atterrissage. Ensuite, vous devez prendre en compte les caractéristiques des données que vous souhaitez acquérir pour vous assurer que l'étape d'ingestion des données dispose de la technologie et des processus adéquats pour atteindre ses objectifs.

Data Journey

2 - Stockage des données

Le stockage des données désigne la manière dont les informations sont **conservées** après leur acquisition. Il repose sur des **plateformes sécurisées et fiables**, intégrant des mécanismes de sauvegarde essentiels à la reprise après sinistre. Par ailleurs, des contrôles **d'accès stricts** doivent être mis en place afin de protéger les données sensibles contre toute tentative d'accès non autorisé ou malveillant.
Le choix d'une **solution de stockage** est une étape déterminante du cycle de vie des données, bien qu'il s'agisse d'un processus complexe influencé par plusieurs facteurs. Parmi les principales **caractéristiques du stockage**, on distingue :

- **Le cycle de vie des données** : la manière dont elles évoluent au fil du temps.
- **Les options de stockage** : les différentes méthodes permettant d'optimiser leur conservation.
- **Les couches de stockage** : la structuration des données en fonction de leur importance et de leur accessibilité.
- **Les formats de stockage** : le mode d'organisation des données selon leur fréquence d'accès.
- **Les technologies de stockage** : les infrastructures sur lesquelles reposent les données.

Bien que le stockage constitue une phase distincte du parcours des données, il s'intègre étroitement aux autres étapes clés, telles que l'ingestion, la transformation et la mise à disposition des données. Il intervient à différents points du pipeline de traitement, se connectant aux systèmes sources et influençant la manière dont les données sont exploitées à chaque phase. Ainsi, la stratégie de stockage adoptée a un impact direct sur l'efficacité globale du cycle de vie des données.

Data Journey

3 - Traitement des données

Une fois les données saisies et stockées, elles doivent être exploitées pour devenir véritablement utiles. L'étape suivante du cycle de vie des données est la **transformation**, qui consiste à convertir les données brutes en informations exploitables pour les différents cas d'utilisation en aval.
Le traitement des données repose sur une série de **transformations de base**, essentielles pour garantir la cohérence et l'exactitude des informations. Ces transformations incluent :

- **La conversion des types de données** : transformation des chaînes de caractères (dates, valeurs numériques) en types de données adaptés.
- **La normalisation des enregistrements** : harmonisation des formats et structuration des données.
- **L'élimination des erreurs** : suppression des entrées incorrectes ou incohérentes.

À mesure que le pipeline de traitement progresse, des **transformations plus avancées** peuvent être nécessaires, telles que :

- **L'adaptation ou la normalisation du schéma des données**, pour assurer une compatibilité avec les systèmes en aval.
- **L'agrégation de données à grande échelle**, notamment pour les besoins de reporting.
- **La transformation des données en vecteurs de caractéristiques (embeddings)**, pour les intégrer à des modèles d'apprentissage automatique.

L'un des principaux défis de cette phase réside dans la **précision et l'efficacité du traitement**, qui nécessite une puissance de calcul importante. Sans stratégies d'optimisation adéquates, ce processus peut s'avérer coûteux à long terme. Ainsi, une gestion efficace des ressources et des performances est essentielle pour garantir un traitement rapide et fiable des données.

Data Journey

4 - Servir les données

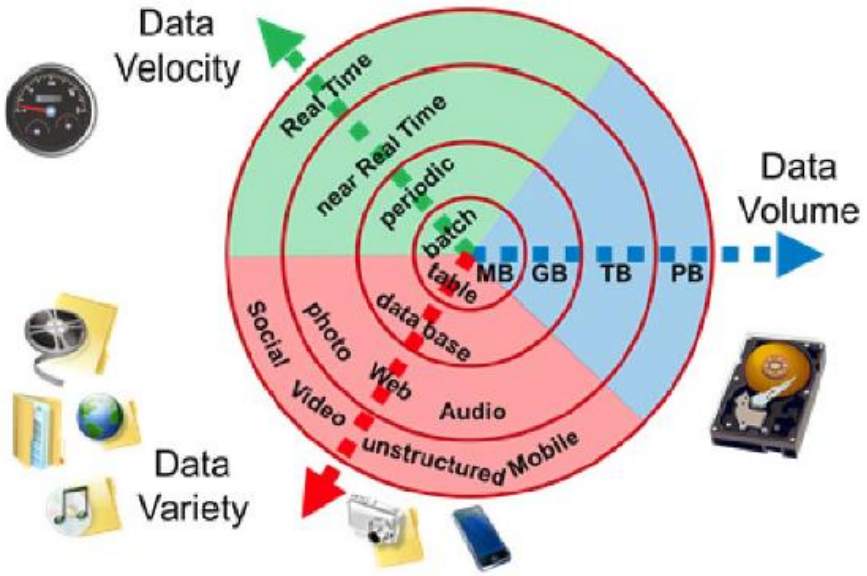
Vous avez atteint la dernière étape du parcours des données. Après avoir été ingérées, stockées et transformées en structures cohérentes et exploitables, il est temps d'en extraire toute la **valeur**.
Le **service de données** est l'étape où les informations prennent tout leur sens. C'est ici que les **ingénieurs BI**, les **ingénieurs en machine learning** et les **data scientists** appliquent des techniques avancées pour générer des insights pertinents. Parmi les approches les plus courantes, on retrouve :

- **L'analyse des données** : interprétation et exploration des informations stockées afin d'orienter la prise de décision et d'anticiper les tendances futures.
- **La visualisation des données** : utilisation d'outils spécialisés pour représenter graphiquement les résultats et faciliter leur compréhension.

Cette phase constitue l'aboutissement du cycle de vie des données, permettant de transformer des ensembles bruts en ressources stratégiques et exploitables pour l'entreprise.

Qu’est ce que Big Data?

Le **Big Data**, ou **données massives**, désigne des ensembles de données si **volumineux**, **variés** et **générés à une telle vitesse** qu'ils dépassent les capacités des outils traditionnels de gestion et d'analyse de données.



Ces caractéristiques sont souvent résumées par les "3V" :

- **Volume** : Quantité massive de données générées.
- **Variété** : Diversité des types de données (structurées, non structurées, semi-structurées).
- **Vélocité** : Vitesse à laquelle ces données sont produites et doivent être traitées.

Certaines définitions ajoutent deux autres "V" :

- **Véracité** : Fiabilité et qualité des données.
- **Valeur** : Potentiel des données à générer des informations utiles.

Big Data: cas d'utilisations

Secteur financier (Détection de la fraude)

- **Descriptif :** Les institutions financières analysent les comportements de transactions pour détecter et prévenir les fraudes en temps réel.

- **Solution :**
  - Analyse des modèles de transactions avec des algorithmes de détection d'anomalies.
  - Utilisation du Machine Learning pour reconnaître des comportements suspects.
  - Mise en place de systèmes de scoring en temps réel (ex. systèmes de scoring de carte bancaire).

- **Challenges :**
  - Faux positifs qui peuvent impacter l'expérience client.
  - Besoin de traitements en temps réel pour bloquer rapidement les fraudes.
  - Complexité liée aux volumes de données générés par les transactions globales

Big Data: cas d'utilisations

Smart Cities (Gestion intelligente du trafic urbain)

- **Descriptif :** Les villes intelligentes utilisent le Big Data pour optimiser la circulation et réduire les embouteillages.

- **Solution :**
  - Analyse des flux de circulation en temps réel à partir des capteurs et caméras.
  - Modélisation des itinéraires optimaux en fonction des conditions actuelles.
  - Intégration avec des applications mobiles (ex. Google Maps).

- **Challenges :**
  - Traitement et synchronisation de données en temps réel provenant de différentes sources.
  - Sécurité et protection contre les cyberattaques.
  - Acceptation par les citoyens et respect de leur vie privée.

Big Data: cas d'utilisations

Secteur de la santé (Prédiction des maladies et personnalisation des traitements)

- **Descriptif :** L'analyse de grandes quantités de données médicales (dossiers patients, imageries médicales, données génétiques) permet de détecter des tendances et de proposer des traitements personnalisés.

- **Solution :**
  - Utilisation d'algorithmes d'apprentissage automatique pour identifier des modèles dans les données de santé.
  - Intégration de données issues de capteurs portables (montres connectées, bracelets de santé).
  - Plateformes de stockage et de traitement cloud (ex. AWS, Google Cloud Healthcare).

- **Challenges :**
  - Protection des données sensibles et conformité avec les réglementations (RGPD, HIPAA).
  - Interopérabilité des différents systèmes hospitaliers.
  - Fiabilité des algorithmes et des recommandations médicales.

Evolution du Big Data

Bien que le concept de Big Data soit relativement nouveau, la nécessité de gérer des jeux de données volumineux remonte aux années 1960 et 70, avec les premiers data centers et le développement des bases de données relationnelles.

- **Passé:** En 2005, on assista à une prise de conscience de la quantité de données que les utilisateurs généraient sur Facebook, YouTube et autres services en ligne. **Apache Hadoop**, une infrastructure open source créée spécifiquement pour stocker et analyser de grands jeux de données, fut développé cette même année. **NoSQL** commença également à être de plus en plus utilisé à cette époque.

- **Présent:** Le développement d'infrastructures open source, telles qu'Apache Hadoop et, plus récemment, **Apache Spark**, a été primordial pour la croissance du Big Data, car celles-ci facilitent l'utilisation du Big Data et réduisent les coûts de stockage. Depuis, le volume du Big Data a explosé. Les utilisateurs génèrent toujours d'énormes quantités de données, mais ce ne sont pas seulement les humains qui les utilisent.
  Avec l'avènement de **l'Internet of Things (IoT)**, de plus en plus d'objets et de terminaux sont connectés à Internet, collectant des données sur les habitudes d'utilisation des clients et les performances des produits. L'émergence du Machine Learning a produit encore plus de données.

- **Futur:** Alors que le Big Data a fait des progrès, sa valeur continue de croître à mesure que l'**IA générative** et l'utilisation du **Cloud Computing** se développent dans les entreprises. Le cloud offre une évolutivité considérable, les développeurs peuvent simplement faire fonctionner rapidement des clusters dédiés pour tester un sous-ensemble de données. En outre, les **bases de données graphiques** deviennent de plus en plus importantes, avec leur capacité à afficher d'énormes quantités de données de manière à rendre les analyses rapides et complètes.

Le paysage Big Data