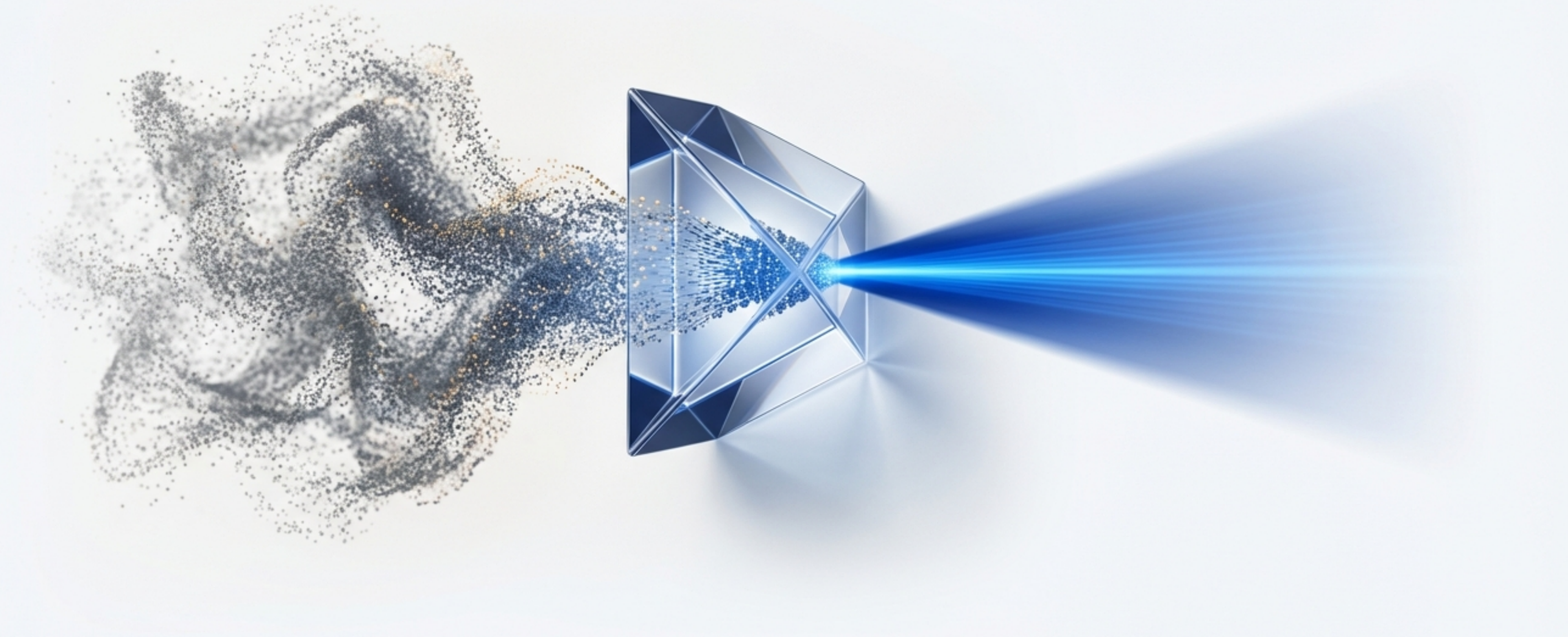


Le Voyage de la Donnée

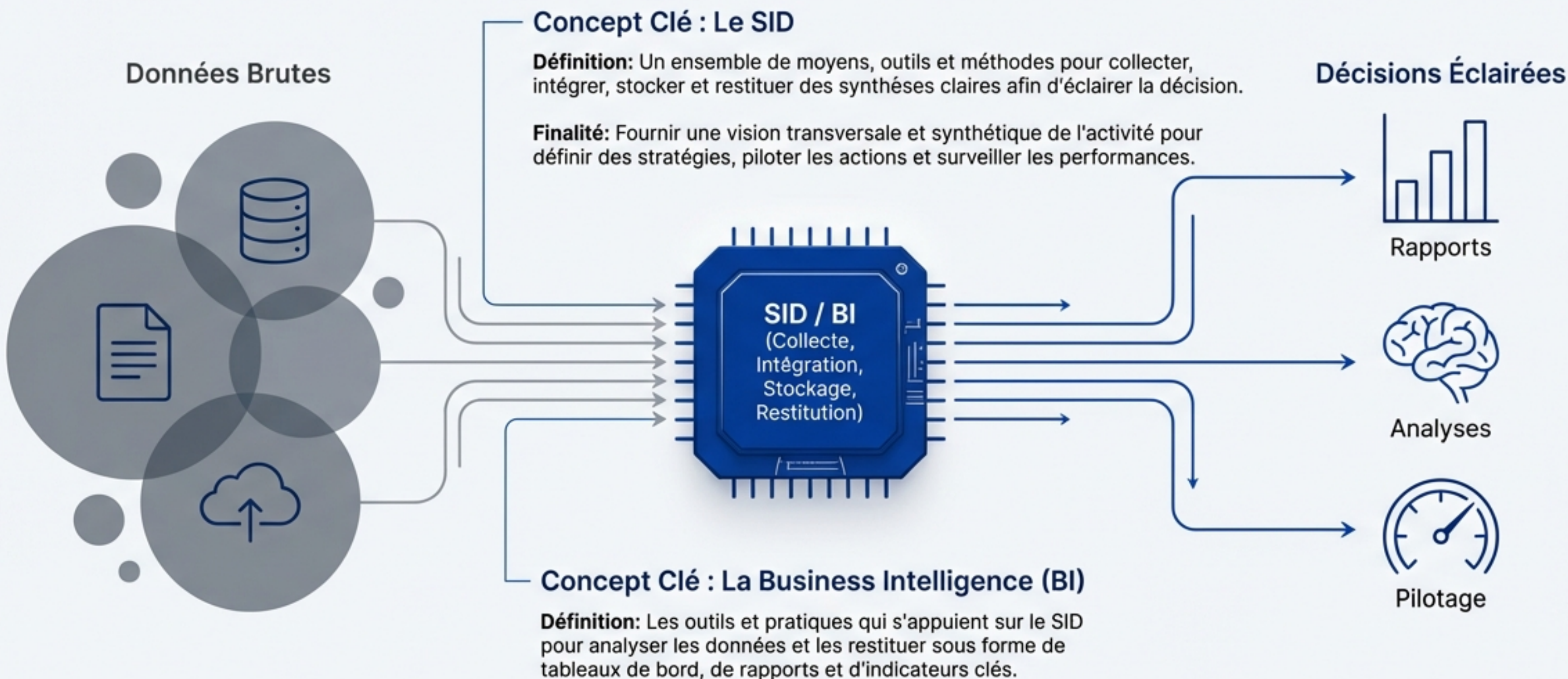
De la Donnée Brute à la Décision Stratégique



Un guide synthétique pour comprendre l'écosystème data moderne : technologies, métiers et applications.

La destination : Piloter l'entreprise par la donnée

L'objectif de tout écosystème data est de transformer les informations brutes en un avantage stratégique.
C'est le rôle du Système d'Information Décisionnel (SID).



Le paysage : Naviguer dans l'ère du Big Data

Les architectures modernes doivent gérer des données caractérisées par les '5 V'.



Volume

Quantités massives
(To à Po)



Vélocité

Données générées en
continu



Variété

Données structurées et
non structurées (JSON,
images, texte)



Véracité

Qualité, bruit, incertitudes



Valeur

Potentiel décisionnel

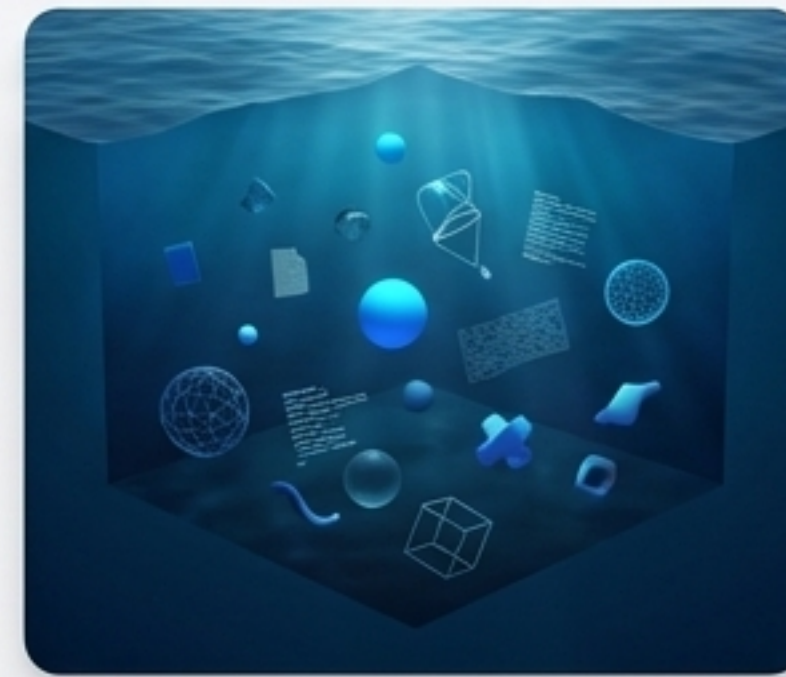
Deux approches de stockage stratégiques

Data Warehouse (Entrepôt de Données)



- Principe:** Stocke des données intégrées, historiques, nettoyées et organisées pour l'analyse.
- Structure:** Schémas en étoile, flocon, cubes multidimensionnels.
- Idéal pour:** Analyse structurée, reporting fiable et rapide.

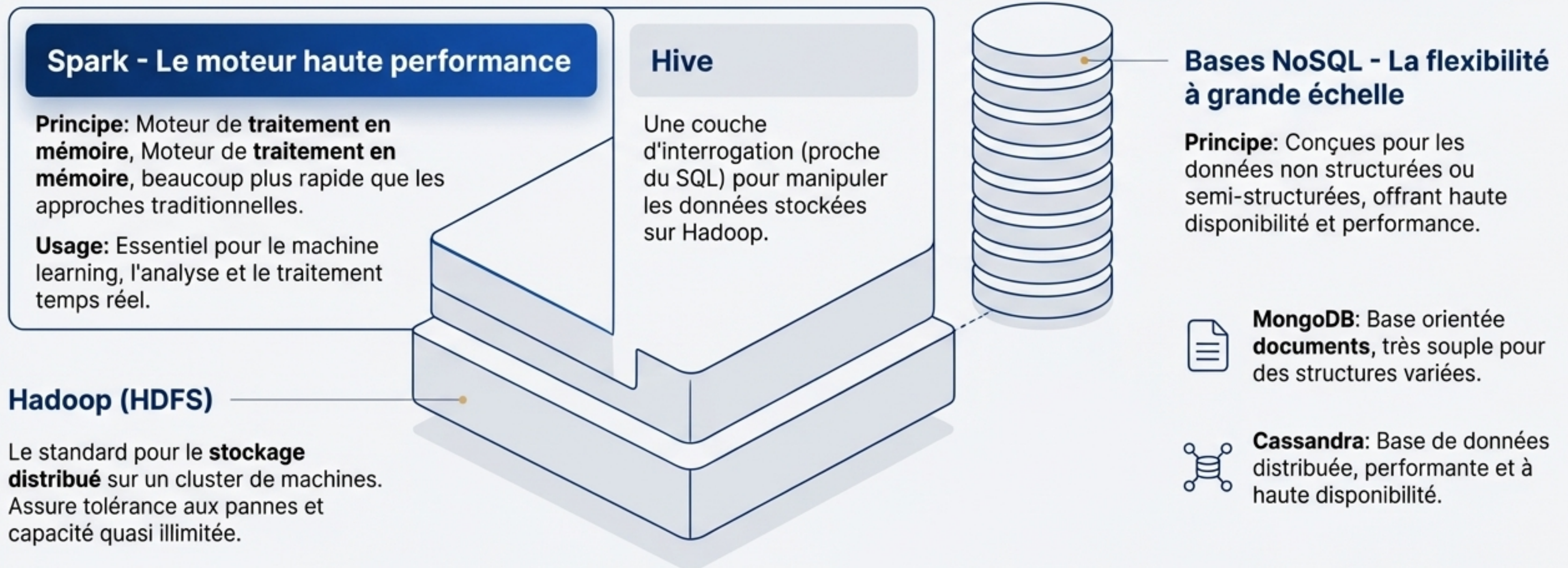
Data Lake



- Principe:** Stocke les données dans leur format brut, sans schéma préalable.
- Structure:** Flexibilité maximale.
- Idéal pour:** Données massives non structurées, travaux exploratoires, IA.

Les fondations technologiques du voyage

L'exploitation du Big Data repose sur un écosystème de technologies distribuées conçues pour le stockage et le traitement à grande échelle.



Rencontrez l'équipage : Les experts au cœur du voyage

La valorisation des données est une œuvre collective, orchestrée par des profils aux compétences complémentaires et spécialisées.

Les rôles clés dans la chaîne de valeur



L'Architecte de Données (Le Visionnaire)

Définit la stratégie globale, choisit les technologies et garantit la cohérence de l'écosystème data. Assure la **pérennité et la stabilité** des systèmes.



L'Ingénieur Big Data (Le Constructeur)

Intervient 'au début de la chaîne'. **Collecte, organise et rend exploitables** les volumes massifs de données. Construit et maintient les pipelines de données.



Le Data Analyst (L'Explorateur)

Analyse les données pour produire des rapports, des visualisations et des indicateurs.
Se concentre sur la **compréhension métier** et l'exploitation opérationnelle.



Le Data Scientist (Le Modélisateur)

Crée des **modèles statistiques et de machine learning** pour faire des prédictions, des classifications ou des recommandations personnalisées.

Data Analyst vs. Data Scientist : Deux approches, un même objectif



Data Analyst

Périmètre

Plus ciblé, orienté opérationnel, reporting, performance.

Question type

"Que s'est-il passé et pourquoi ?".

Missions principales

- Nettoyage et préparation des données.
- Mise en place d'indicateurs (KPI).
- Analyses statistiques descriptives.
- Production de rapports et visualisations.

Focalisation

Analyse directe, compréhension métier.



Data Scientist

Périmètre

Vision plus globale, utilisant des données variées.

Question type

"Que va-t-il se passer et que pouvons-nous faire ?".

Missions principales

- Création de modèles statistiques avancés.
- Développement d'algorithmes de machine learning.
- Production de recommandations personnalisées.
- Analyses prédictives et prescriptives.

Focalisation

Modélisation avancée, algorithmes complexes.

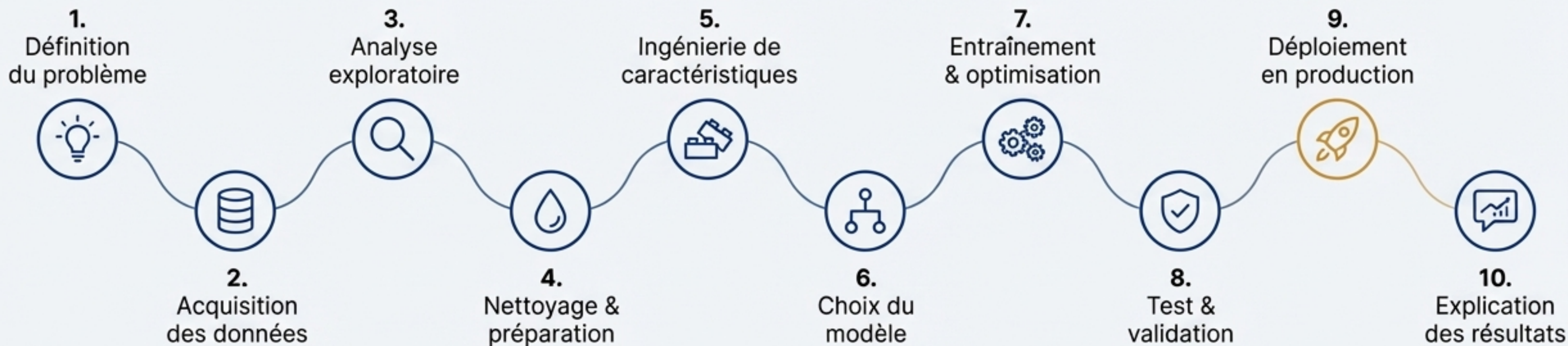
Le moteur du voyage : Le Machine Learning

Un champ de l'IA fondé sur des méthodes statistiques et mathématiques permettant à un ordinateur **d'apprendre à partir de données** et **d'améliorer ses performances** sans être explicitement programmé pour chaque tâche.

Les deux phases de l'apprentissage

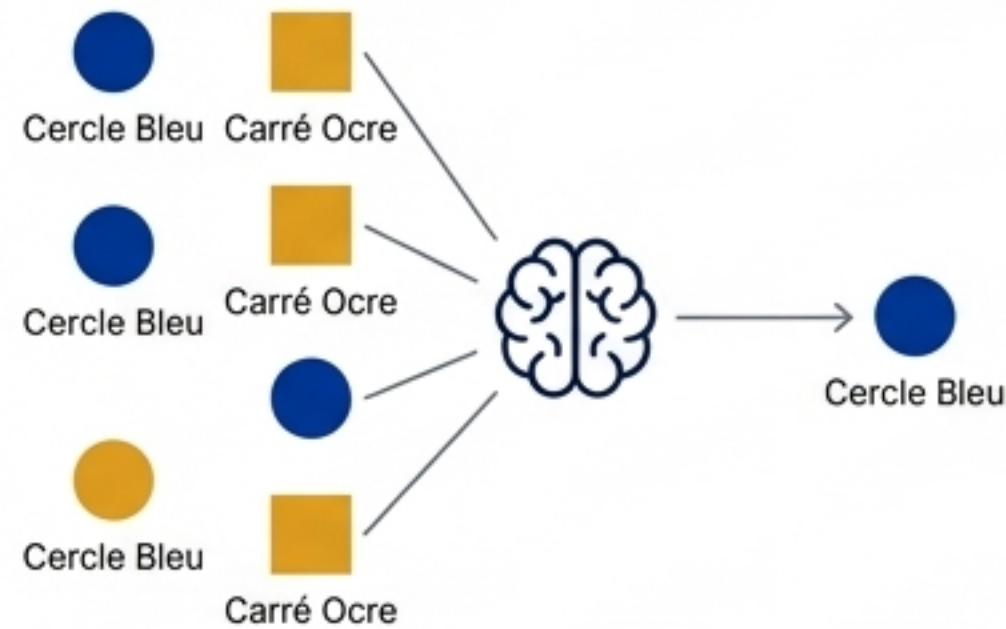


Les grandes étapes d'un projet de Machine Learning



Les grandes familles d'apprentissage automatique

Apprentissage Supervisé

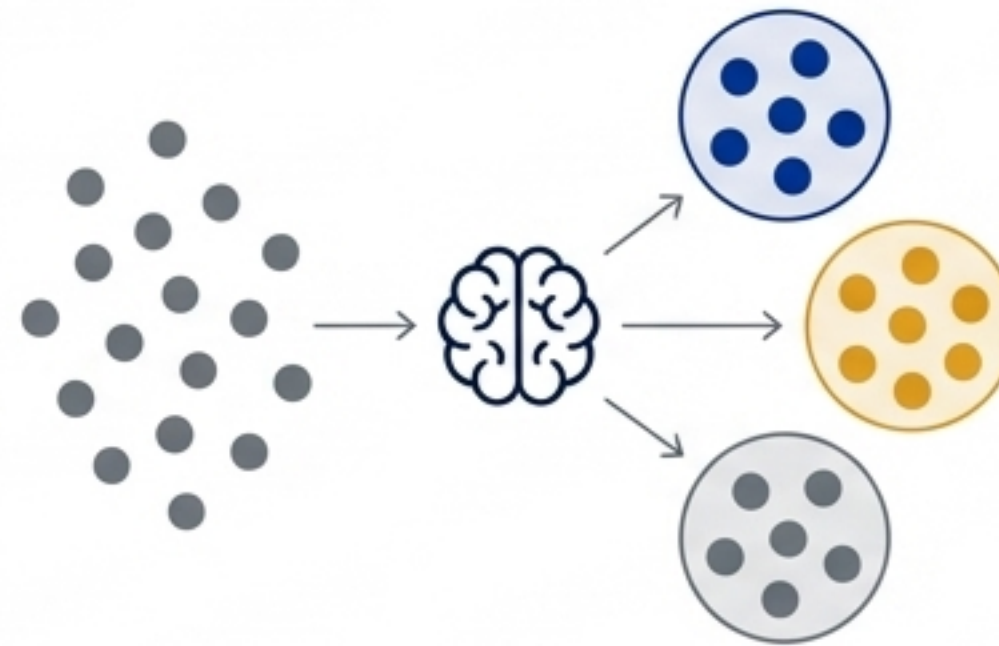


Principe : L'algorithme apprend à partir de données **étiquetées** (la bonne réponse est connue).

Objectif : Apprendre une fonction $Y = f(X)$.

Exemples d'usage : Prédiction météo, filtrage de spam, détection de fraude.

Apprentissage Non Supervisé

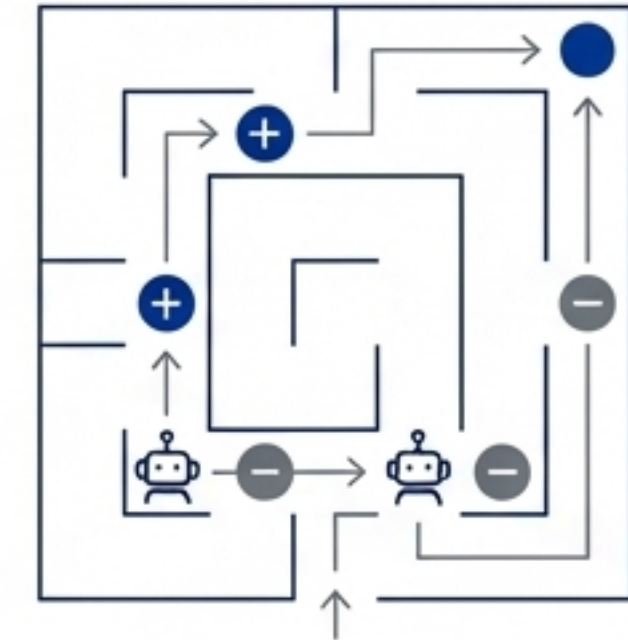


Principe : L'algorithme explore des données **non étiquetées** pour découvrir des **structures cachées**.

Objectif : Trouver des motifs et des groupes naturels.

Exemples d'usage : Segmentation de clientèle, détection d'anomalies.

Apprentissage par Renforcement



Principe : Le modèle apprend **en interagissant avec un environnement** via un système de **récompenses** et de **pénalités**.

Objectif : Apprendre une stratégie optimale par essais/erreurs.

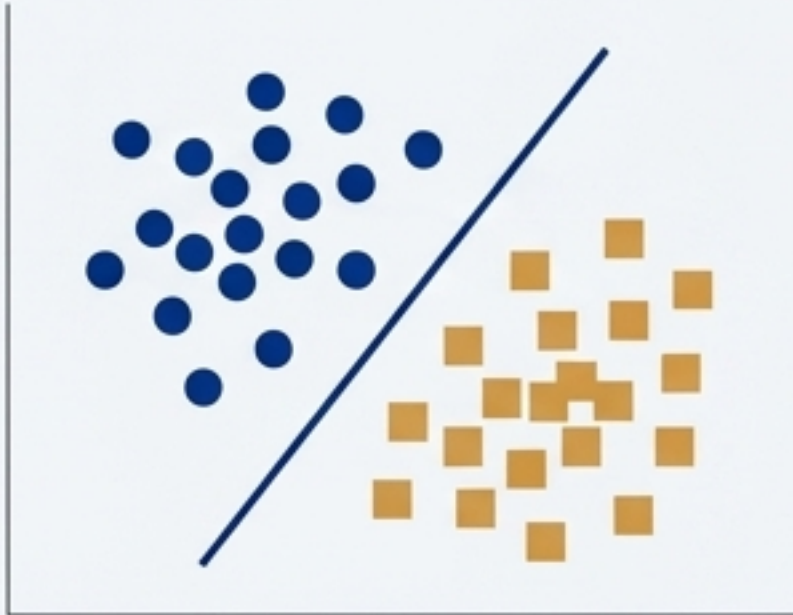
Exemples d'usage : Jeux (AlphaGo), robotique, voiture autonome.

Note : D'autres approches hybrides existent, comme l'apprentissage **semi-supervisé** (peu de données étiquetées, beaucoup de non étiquetées) et **auto-supervisé** (le système génère ses propres étiquettes).

Plongée dans l'apprentissage supervisé : Prédire et Classifier

Le supervisé se divise en deux tâches fondamentales, selon la nature de la sortie souhaitée.

La Classification



Objectif : Prédire une **catégorie** discrète.

Exemples : Oui/Non, Chat/Chien, Client à risque/Non à risque.

Algorithmes courants :

- Régression logistique
- Arbres de décision
- Random Forest (Forêts d'arbres décisionnels)
- SVM (Machines à vecteurs de support)

La Régression



Objectif : Prédire une **valeur continue**.

Exemples : Un prix, une température, un chiffre d'affaires.

Algorithmes courants :

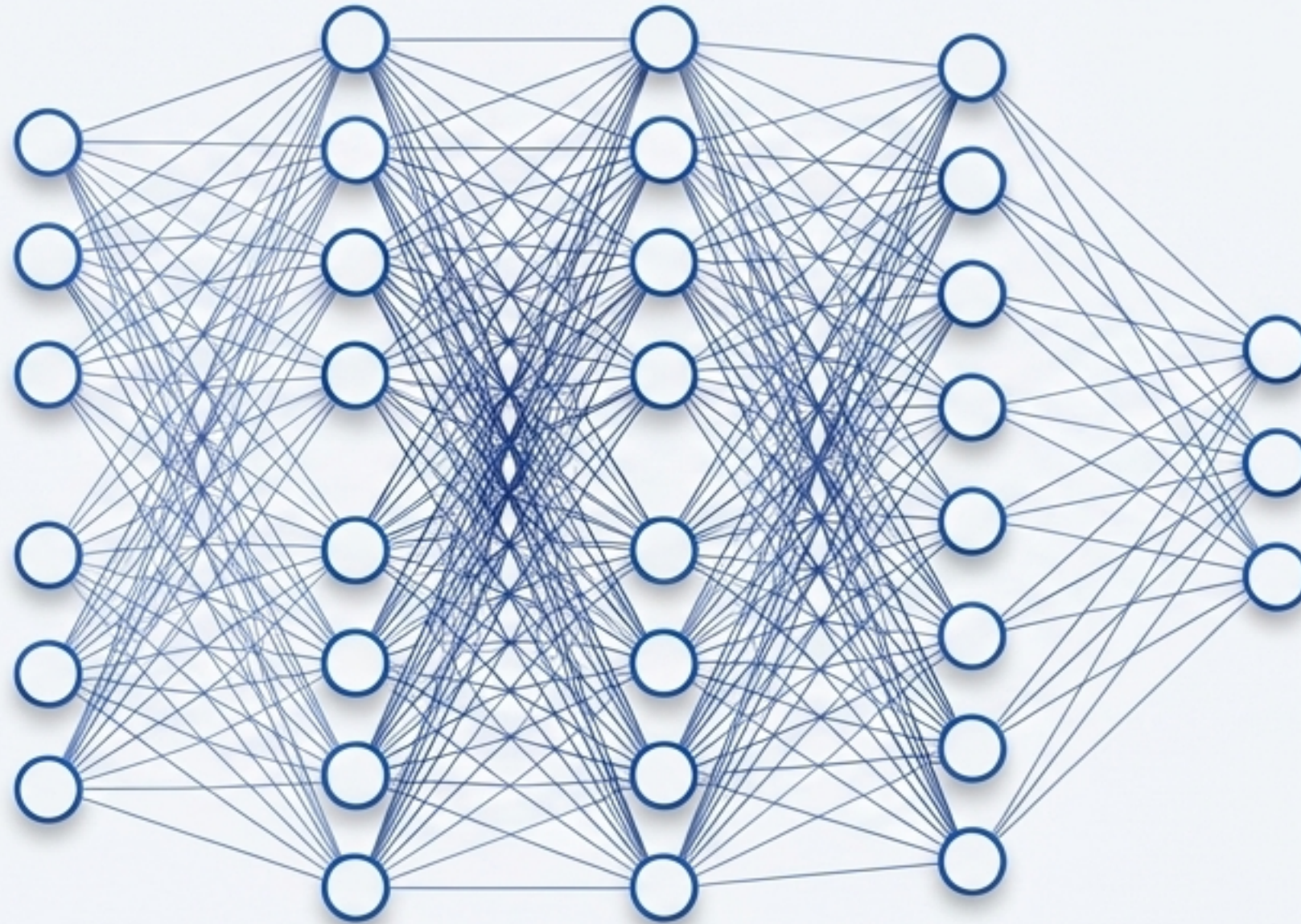
- Régression linéaire
- Réseaux de neurones (peuvent aussi faire de la classification)

À la pointe : Le Deep Learning et les réseaux de neurones

Le Deep Learning (apprentissage profond) est une sous-branche du Machine Learning utilisant des **réseaux de neurones profonds** (avec de nombreuses couches cachées, ≥ 3).

Pourquoi est-ce une révolution ?

- **Principe** : Capacité à apprendre des représentations hiérarchiques et complexes directement à partir des données brutes.
- **Accélérateur** : Le calcul parallèle massif des **GPU** (vs CPU) a permis des gains de performance de $\times 10$ à $\times 100$, rendant l'entraînement de ces modèles possible.



Domaines de domination du Deep Learning

-  Images (Vision par ordinateur)
-  Sons (Reconnaissance vocale)
-  Textes (Traitement du Langage Naturel - NLP)

Domaines où le Machine Learning classique reste meilleur

-  Données tabulaires (< 10 000 lignes)
-  Problèmes nécessitant une forte **explicabilité**

Traitements sur des données irrégulières ou non continues

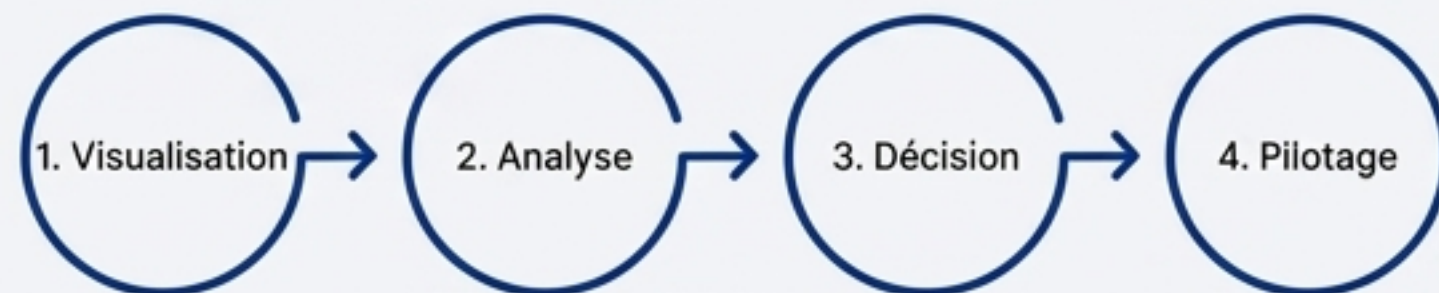
L'arrivée : Transformer l'analyse en pilotage stratégique

La finalité du voyage est de mettre à disposition des décideurs une information claire, consolidée et pertinente via les outils de la Business Intelligence.

Les outils de restitution et de pilotage

-  **Le Reporting:** Fournit des états réguliers et des indicateurs clés pour suivre l'activité.
-  **Les Tableaux de Bord (Dashboards):** Regroupent les indicateurs de performance (KPI) pour une visualisation immédiate et une prise de décision rapide.
-  **Les Analyses OLAP (On-Line Analytical Processing):** Permettent une exploration en profondeur des données sur plusieurs axes (temps, lieu, produit) pour des analyses complexes.

Le cycle de la décision éclairée



Les règles du voyage : Gouvernance et défis éthiques

La puissance de la data s'accompagne de responsabilités et de défis techniques et éthiques majeurs.

La Gouvernance et la Conformité



Le DPO (Data Protection Officer) : Le 'chef d'orchestre de la conformité'. Son rôle est de garantir l'application du **RGPD**, de conseiller l'entreprise et d'assurer la liaison avec la **CNIL**.

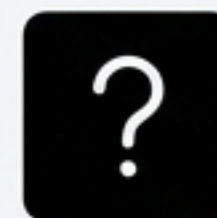
Enjeux :

- Sécurité des données sensibles, contrôle des accès, respect de la vie privée.

Les limites et problèmes du Machine Learning



Biais des algorithmes : Un modèle peut apprendre et amplifier des biais présents dans les données, menant à des discriminations involontaires (ex: recrutement, scoring financier).



Le problème de la 'boîte noire' : La complexité de certains modèles (ex: réseaux de neurones) rend leurs décisions difficiles à expliquer. Des outils d'**explicabilité** (SHAP, Feature Importance) sont nécessaires pour y remédier.



Corrélation n'est pas causalité : Les modèles apprennent des **corrélations**, pas des relations de cause à effet.



Qualité des données : Les modèles sont sensibles aux valeurs manquantes, au bruit et aux données déséquilibrées.

L'impact : Quand la donnée transforme le monde réel

L'écosystème data, de l'infrastructure aux algorithmes, alimente des innovations dans tous les secteurs.



Systemes de
recommandation



Diagnostic médical
assisté



Détection de fraude



Vision par ordinateur



Traitement du
langage naturel (NLP)



Optimisation
industrielle

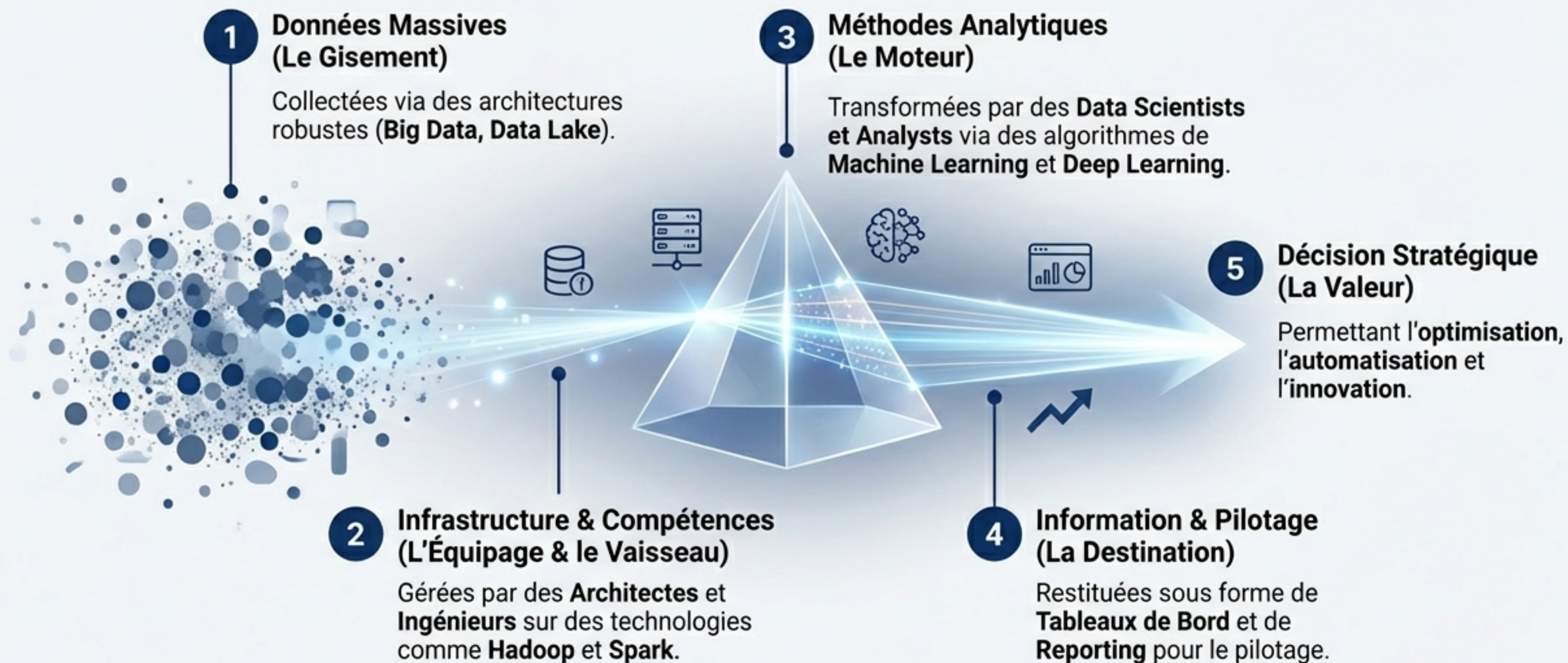


Trading
algorithmique



Robotique

Synthèse : La chaîne de valeur de la donnée



La maîtrise de cette chaîne est aujourd'hui un levier fondamental de performance et de compétitivité.