11. L'ingénierie des données

Contexte

L'un des défis de la science des données : mettre de grandes quantités de données dans des formats pouvant être **lus** par des algorithmes.

L'**ingénierie des données** est liée au traitement de ces données.

Après le traitement, les scientifiques des données développent des **preuves de concept** ; les ingénieurs IA/AA les traduisent en **modèles déployables**.

L'ingénierie des données existe depuis un certain ; avec l'essor du "**cloud computing**", l'expertise dans ce domaine devient aussi recherchée que celle en analyse de données (du moins, dans certains cercles).

Rôles et responsabilités (reprise)

Ingénieurs en données (ID)

- recevoir des données d'une source
- structurer, distribuer et stocker les données dans des lacs et des entrepôts de données
- créer des outils et des modèles de données que les SD utilisent

Ingénieurs AA

- déployer de modèles de données
- combler les écarts entre ID et SD
- faire passer des idées de validation de concept à grande échelle

Scientifiques des données

- recevoir des données procurées/fournies par l'ID
- extraire la valeur des données
- construire des modèles prédictifs de preuve de concept
- mesurer et améliorer les résultats
- construire des modèles analytiques

Rôles et responsabilités (reprise)

Dans les petites organisations, l'ingénierie et la science des données sont généralement **regroupées** dans sous un même toit.

Les grandes entreprises disposent d'ingénieurs de données **spécialisés**, qui construisent des **pipelines de données** et gèrent des **entrepôts de données** (en les alimentant en données et en créant des schémas de table pour assurer le suivi des données stockées).

En général, ID ≠ SD.

Les pipelines de données

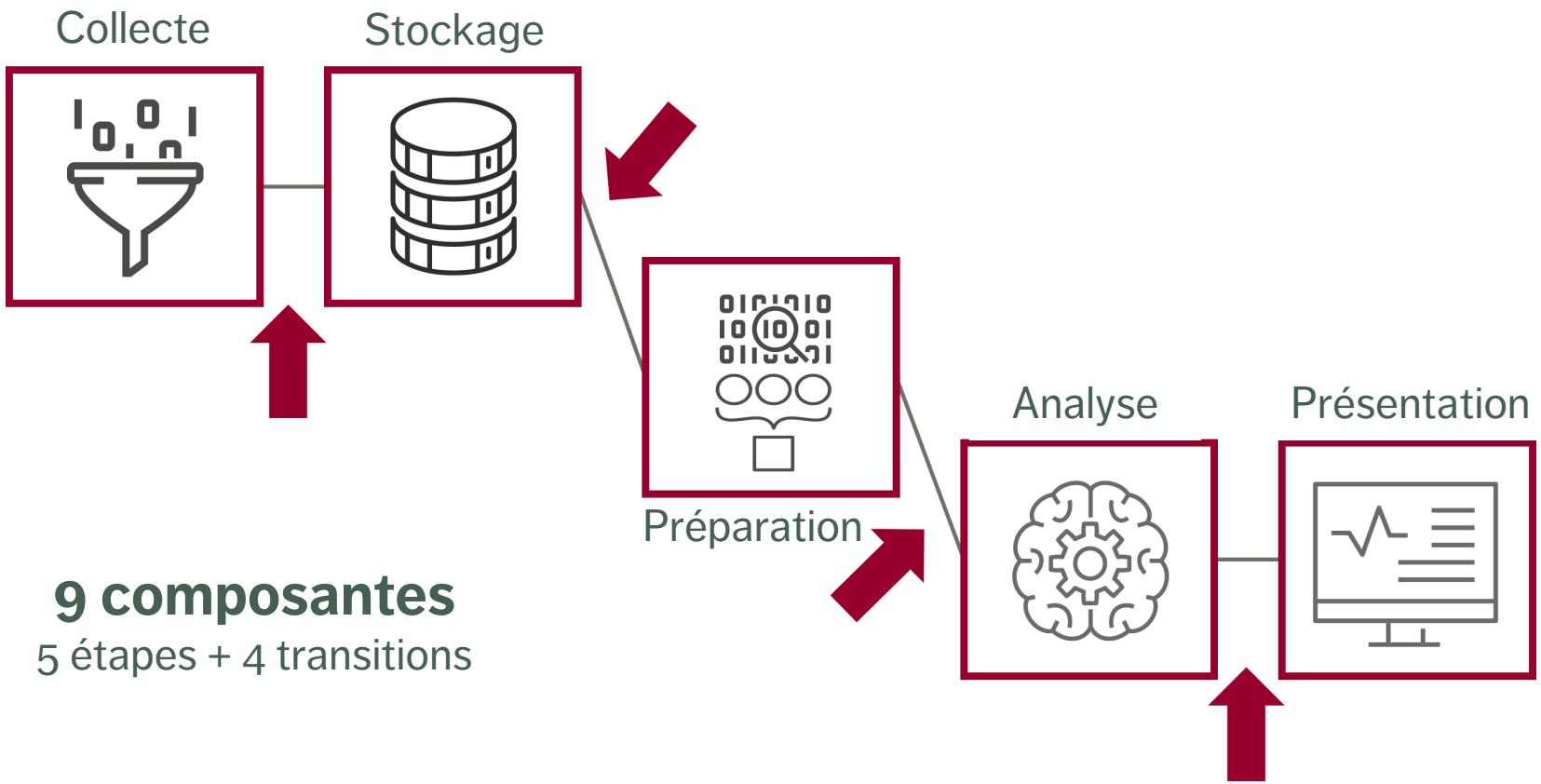
Ingénierie des **données**

- les opérations qui créent des **interfaces** et des **mécanismes** pour le flux/l'accès à l'information
- mise en place d'une **infrastructure de données**, préparation des données pour une analyse plus poussée par des SD

Les données peuvent provenir de nombreuses **sources** (et types de sources), et dans une variété de formats et de tailles.

Transformer tout cela en un processus que les SD peuvent utiliser et dont ils peuvent tirer du sens est connu sous le nom de **construction d'un pipeline de données**.

Les pipelines de données



9 composantes

5 étapes + 4 transitions

Les pipelines de données

Principal défi en matière d'ingénierie des données :

- construire un pipeline qui **s'exécute en temps réel** (ou presque) **à chaque fois qu'il est sollicité**
- afin que les utilisateurs obtiennent des **informations actualisées** avec des **délais minimaux**

Les pipelines conceptuels sont transmis aux ingénieurs AA pour le **déploiement** et la **production**. Certains des travaux entourant cette tâche comprennent :

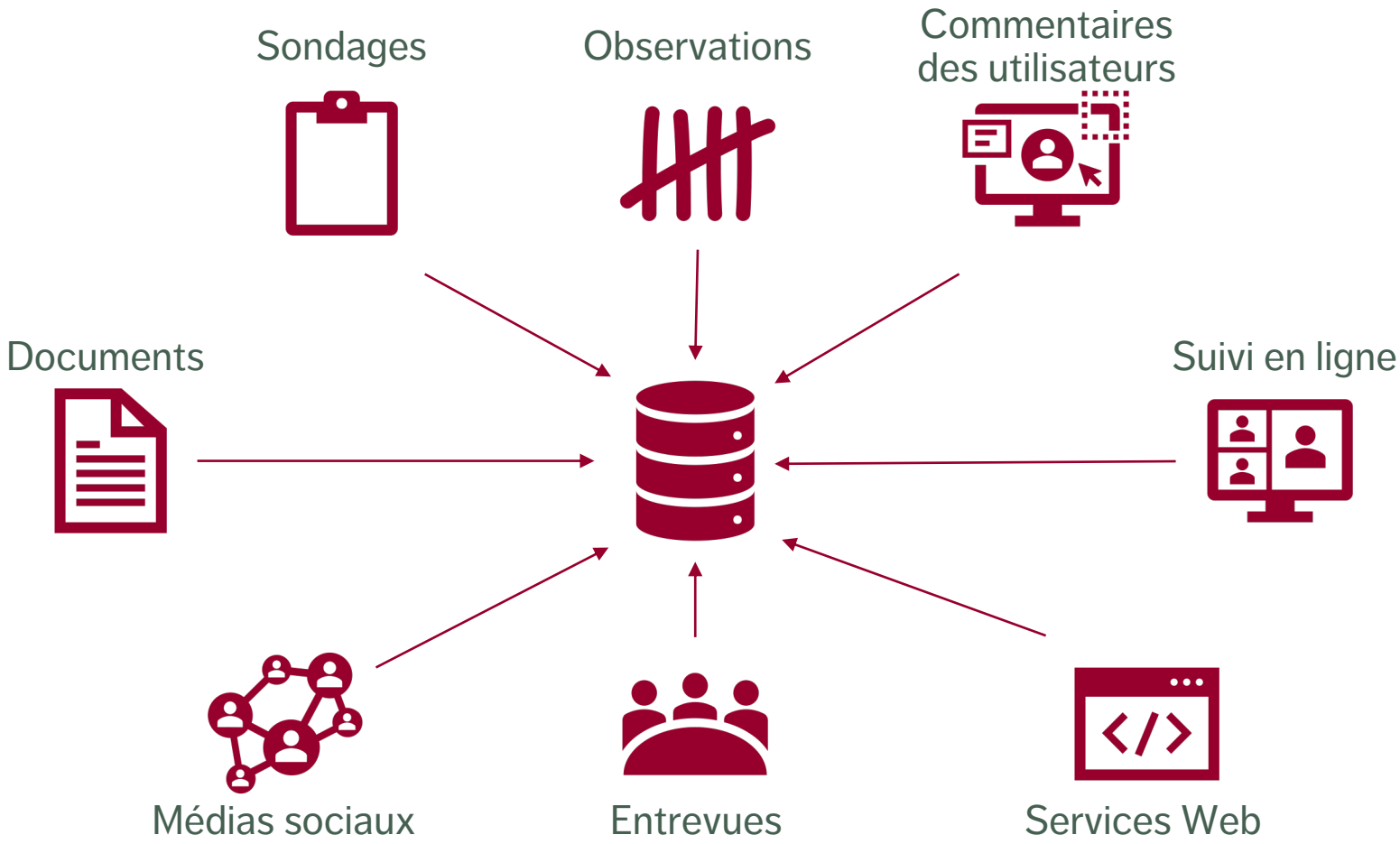
- contrôles de la qualité des données
- optimisation de la performance des requêtes
- la création d'un écosystème d'intégration/livraison continue pour les changements de modèles
- ingéstion des données provenant de diverses sources dans le modèle de données
- transfert des techniques d'AA et de SD aux systèmes distribués

Les pipelines de données

Thèmes communs (opérations/framework/tâches/sources) pour les étapes du pipeline :

- **collecte de données :** applications, applications mobiles, microservices, dispositifs de l'Internet des objets (IoT), sites web, instrumentation, journalisation, capteurs, données externes, contenu généré par l'utilisateur, etc.
- **le stockage des données :** Gestion des données de référence (MDM), entrepôt, lac de données, etc.
- **intégration/préparation des données :** ETL, intégration de données en flux, etc.
- **analyse des données :** apprentissage automatique, analyse prédictive, tests A/B, expériences, intelligence artificielle (IA), apprentissage profond, etc.
- **livraison et présentations :** tableaux de bord, rapports, microservices, notifications push, email, SMS, etc.

La collecte de données



ETC – Extraire

Extraire | Transformer | Charger

ETC – Transformer

Structure

Types de données

Agrégation

Nettoyage

Rejoindre

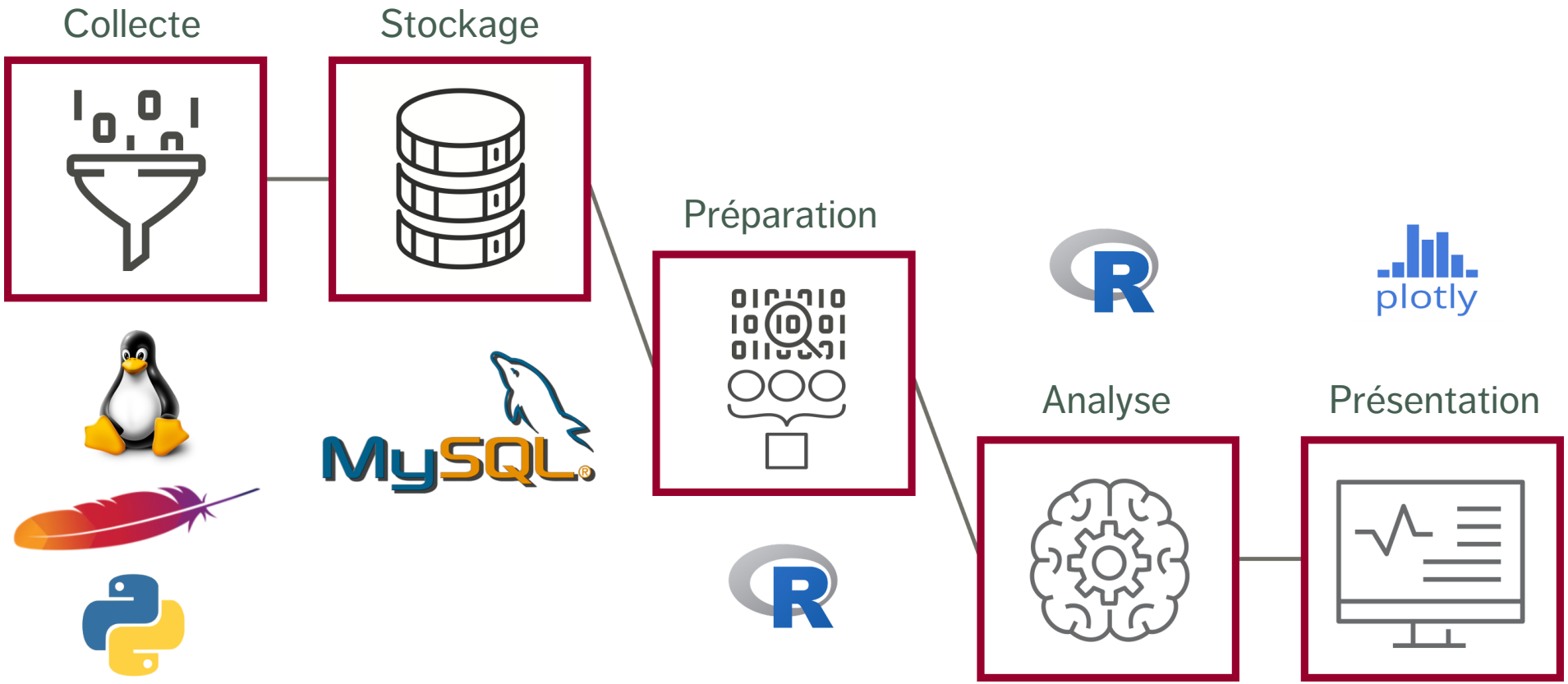
Regroupement

Extraire | **Transformer** | Charger

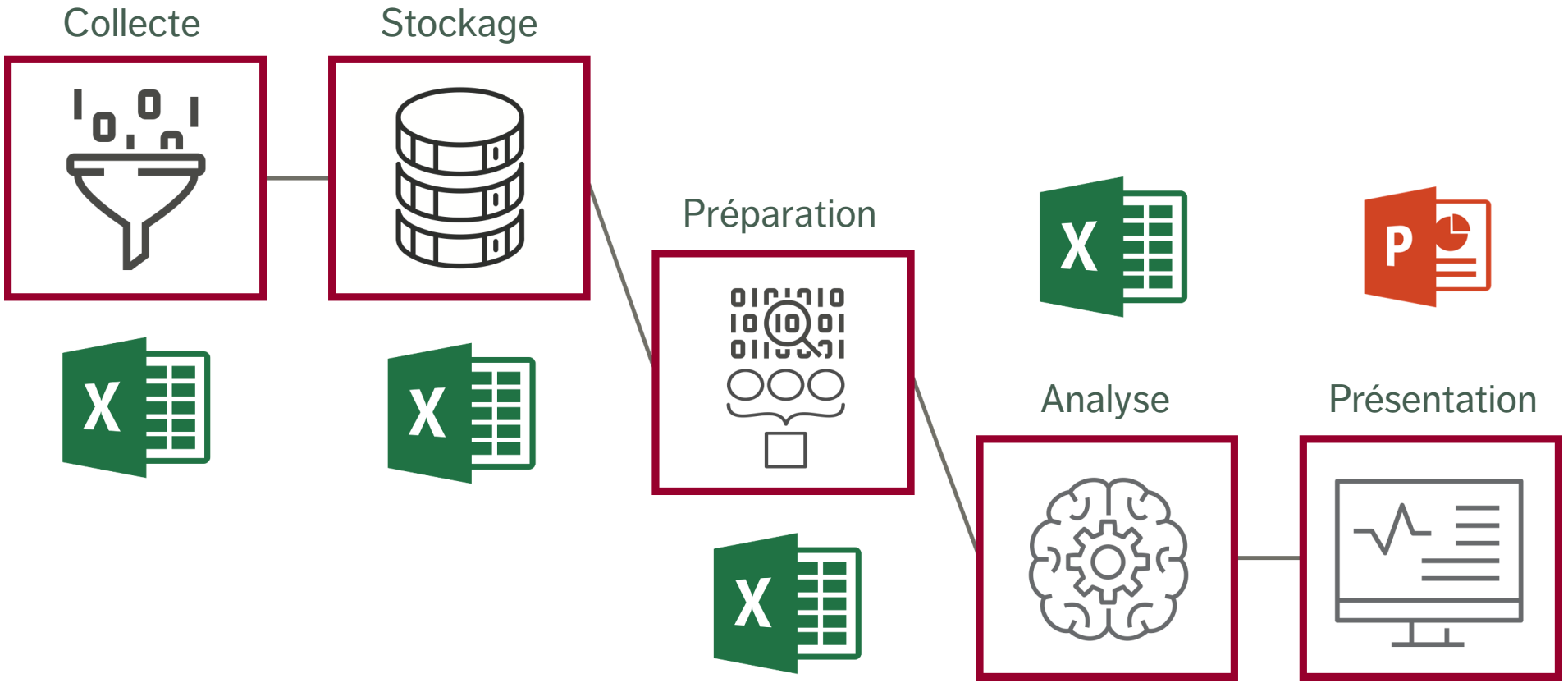
ETC – Transformer

Extraire | Transformer | **Charger**

Un pipeline de données “Open Source”



Un pipeline de données GdC (?)



Les outils de pipelines de données

Les pipelines permettent aux utilisateurs de diviser les tâches importantes en une série de petites étapes séquentielles, ce qui peut aider à **optimiser** chaque étape.

E.g., si vous utilisez TensorFlow pour la composante d'analyse d'un pipeline DL qui consiste en un seul grand script, **tout**, de la collecte des données à la présentation, doit utiliser TensorFlow, ce qui peut ne pas être optimal.

Les outils de pipeline de données sélectionnent le meilleur cadre/ langage pour chaque composante/ tâche du pipeline :

- Luigi (Spotify)
- Airflow (AirBnB)
- scikit-learn
- pandas/tidyverse
- etc.

Les outils d'ingénierie des données

Il est peu probable qu'un ID puisse maîtriser tous les outils d'ingénierie de données possibles, mais les équipes ID ont une plus grande **couverture :**

- **bases de données analytiques** (Big Query, Redshift, Synapse, etc.)
- **ETC** (Spark, Databricks, DataFlow, DataPrep, etc.)
- **moteurs de calcul évolutifs** (GKE, AKS, EC2, DataProc, etc.)
- **orchestration de processus** (AirFlow/Cloud Composer, Bat, Azure Data Factory, etc.)
- **déploiement et mise à l'échelle de plateforme** (Terraform, outils personnalisés, etc.)
- **outils de visualisation** (Power BI, Tableau, Google Data Studio, D3.js, ggplot2, etc.)
- **programmation** (tidyverse, numpy, pandas, matplotlib, scikit-learn, scipy, Spark, Scala, Java, SQL, T-SQL, H-SQL, PL/SQL, etc.)

La gouvernance des données

La gouvernance des données englobe :

- les **personnes** ;
- les **processus**, et
- les **technologie de l'information**

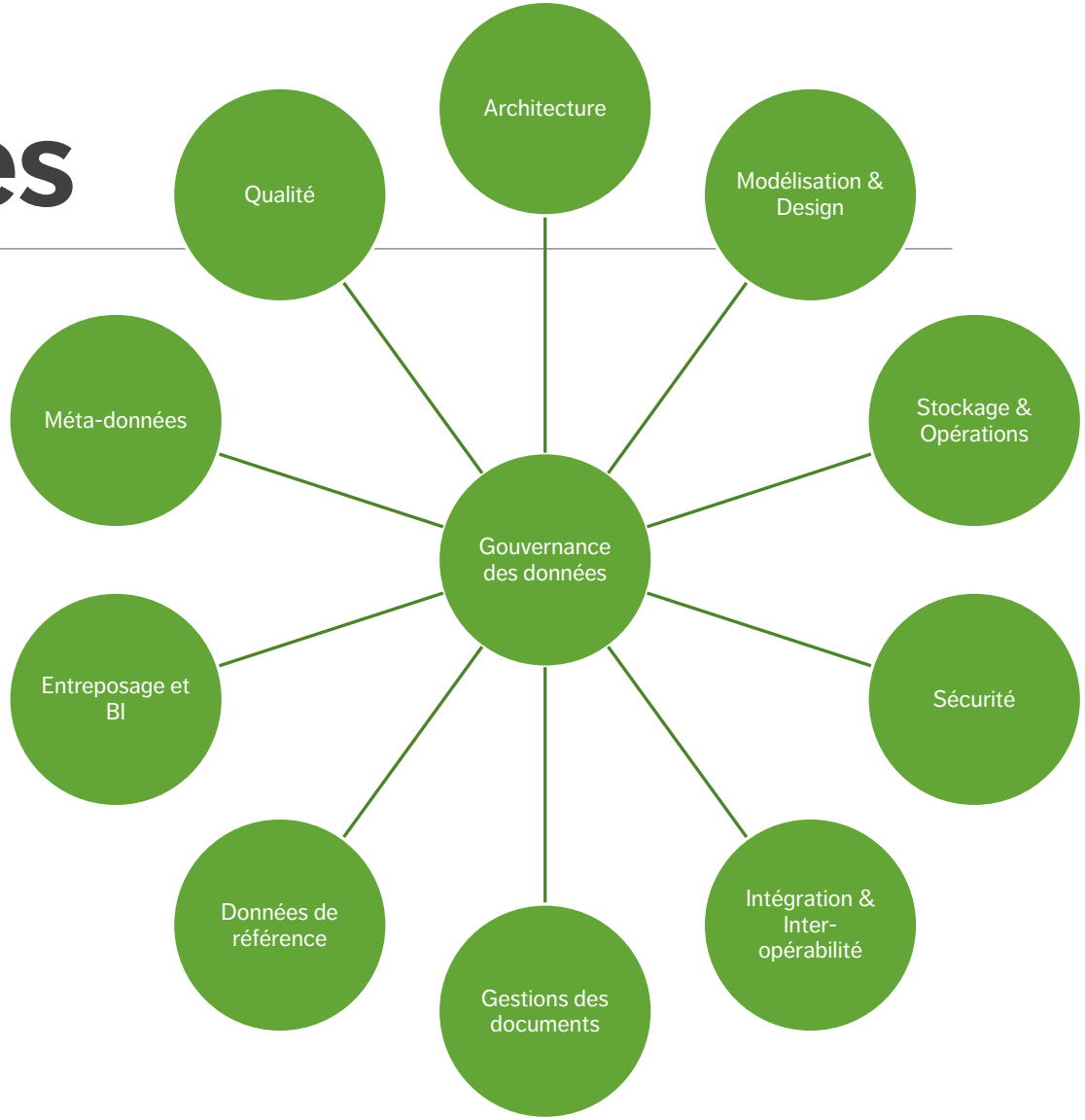
Onl'utilise pour créer un traitement **cohérent/approprié** des données d'une organisation à travers l'entreprise.

Elle fournit la base, la stratégie, et la structure pour garantir que les données sont gérées comme un **actif** et transformées en informations **significatives**.

La gouv. des données

Objectifs :

- création d'une culture de données libre service
- établir des règles internes pour leur utilisation
- mettre en œuvre les exigences de conformité
- améliorer les communications
- augmenter la valeur des données
- réduire les coûts associés aux données
- gérer continuellement les risques
- assurer une existence continue



Lectures suggérées

L'ingénierie des données

Data Understanding, Data Analysis, Data Science
Data Engineering and Management

Background and Context

Data Engineering

- Data Pipelines
- Automatic Deployment and Operations
- Scheduled Pipelines and Workflows
- Data Engineering Tools

Exercices

L’ingénierie des données

1. À quoi ressemble votre pipeline de science des données (ou celui de votre organisation) ? Pourrait-il être amélioré ?
2. Identifiez des cas où vous avez rencontré des problèmes liés à la disponibilité, la facilité d’utilisation, la cohérence, l'intégrité, la qualité, la sécurité ou la fiabilité des données.
3. Complétez tous les exercices précédents que vous n'avez pas eu l'occasion de terminer.