



M1151 · LECTURE ACADÉMIQUE

Initiation au management de la data

De la collecte à la décision : rendre les données fiables, gouvernées et utiles



Data Management
DATA MANAGEMENT

SÉANCE 1 / 3 Fondations, cycle de vie et sources

ENSEIGNANT

Pr. Youness BOUTYOUR

youness.boutyour@um5.ac.ma

PUBLIC CIBLE

Génie de la data

1ère année

INSTITUTION

Université Mohammed V de Rabat

Ecole Nationale Supérieure d'Informatique et d'Analyse des
Systèmes

L'école vous confie sa prochaine décision énergétique



- L'école veut réduire sa consommation sans dégrader les conditions d'enseignement.

QUESTION

?

Vous êtes l'équipe data : quel bâtiment faut-il examiner en premier ?

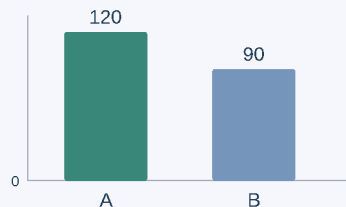
- **Vous commencez par lequel ?** Votez à main levée et justifiez votre choix.
- Quelle information vous manque pour défendre votre choix ?

BÂTIMENT	ÉNERGIE RELEVÉE
A — Enseignement	120 kWh
B — Laboratoires	90 kWh
C — Bibliothèque	70 kWh

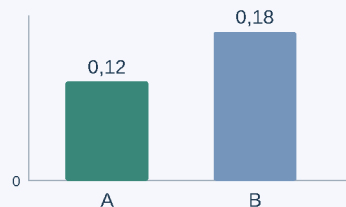
Une révélation qui change la lecture

- **A — Enseignement : 1 000 m², journée complète.** Comparaison possible.
- **B — Laboratoires : 500 m², journée complète.** Même période, mais surface différente.
- **C — Bibliothèque : deux heures de relevés manquantes.** Comparaison incomplète.

Énergie sur la journée (kWh)



Énergie / surface (kWh/m²)



Volume total $\neq$ intensité par mètre carré.

Il faut choisir l'indicateur adapté à la question.

Maintenez-vous votre choix ? Peut-on classer les trois bâtiments ?

- **A** consomme le plus en volume observé : 120 kWh .
- **B** consomme davantage par mètre carré : 0,18 kWh/m² contre 0,12 kWh/m² pour A.
- **C** ne peut pas être classé : des observations manquent.

Règle : avant de comparer, vérifier le périmètre, la période, l'unité et la complétude des données.

Le calcul est juste. La décision peut être mauvaise.

01 • Source et provenance

Qui a produit ces relevés, quand et selon quelle méthode ?

02 • Définition de la mesure

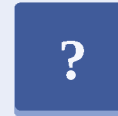
Que représentent exactement les `kWh`, pour quelle période et quel périmètre ?

03 • Complétude

Quelles observations manquent et cette absence change-t-elle la comparaison ?

04 • Responsabilité et usage

Qui contrôle l'indicateur et quelles limites faut-il signaler avant de décider ?



QUESTION DIRECTRICE

Avant de décider, que faut-il encore vérifier ?

Cette question guidera la suite de la séance.

À la fin de cette séance, vous saurez défendre votre décision

01 • Contextualiser

Relier la donnée à un domaine d'application, une source, une période, une unité et un périmètre.

02 • Retracer

Suivre le parcours d'une donnée et identifier les acteurs qui la transforment.

03 • Vérifier

Contrôler sa qualité et préciser ce qu'un indicateur permet — ou non — de conclure.

04 • Préparer la réutilisation

Vérifier l'accès, la licence, l'attribution, les restrictions et la réutilisation avec FAIR.

PRODUCTION ATTENDUE

« Je recommande le bâtiment X pour l'objectif Y, car l'indicateur Z apporte telle preuve. »

La recommandation précise le périmètre, la période, les données manquantes et les limites de l'interprétation.

Plan de la séance

SÉANCE 1 / 3 – FONDATIONS, CYCLE DE VIE ET SOURCES

01 Fondations du management des données

02 Cycle de vie des données

03 Identification et documentation des sources

04 Acteurs et responsabilités

05 Valeur et risques des données

06 Synthèse

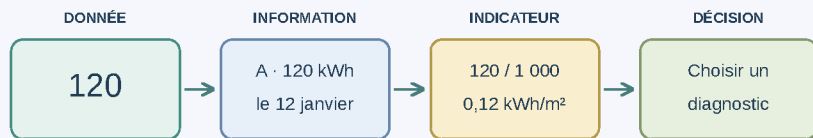
CHAPITRE

Fondations, cycle de vie et sources

01

Fondations du management des données

Une donnée n'est pas encore une décision



Même période · couverture vérifiée · usages comparables · responsable identifié

Donnée — qu'est-ce qui a été enregistré ?

120

Information — que signifie cette valeur ?

120 kWh pour le bâtiment A, le 12 janvier.

Indicateur — quelle comparaison permet-elle ?

0,12 kWh/m² pour comparer des surfaces différentes.

Décision — quelle action peut-on justifier ?

Prioriser — ou non — un diagnostic énergétique.

Définition opératoire

DÉFINITION OPÉRATOIRE

Le management des données est un cadre organisationnel et technique qui gouverne les données sur tout leur cycle de vie — de la planification à la suppression — afin de les rendre accessibles, compréhensibles, fiables, protégées et réutilisables pour un usage explicite.

Gouverner

attribuer les rôles, les responsabilités et les règles.

Documenter

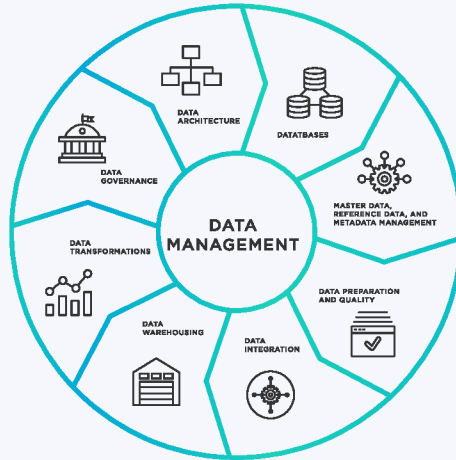
le sens, la qualité, la provenance et les transformations.

Contrôler

les accès, les risques, les preuves et les limites d'usage.

Le management des données ne décide pas à la place des métiers : il rend l'usage des données explicable et contrôlable.

Ce que la définition doit rendre possible



Pour l'utilisateur

Retrouver la bonne donnée, comprendre son sens et connaître ses limites avant de l'utiliser.

Pour l'organisation

Vérifier la qualité, protéger les usages, conserver les preuves et permettre la réutilisation.

Notions fondamentales — décrire ce qui est observé

DONNÉE Représentation codée d'une observation, d'une mesure ou d'un fait, enregistrée dans un format donné. Exemple : 120 dans une colonne de relevés.

INFORMATION Donnée interprétée dans un contexte explicite : objet, unité, période, lieu et signification sont précisés. Exemple : A · 120 kWh · 12 janvier · journée complète .

CONNAISSANCE Compréhension construite à partir d'informations reliées à des règles, des observations répétées ou un modèle explicatif. Exemple : une intensité supérieure de B observée sur plusieurs journées comparables.

TEST DE LECTURE Avant d'interpréter une valeur, demander : **que mesure-t-elle, sur quel périmètre, à quelle période et selon quelle méthode ?**

Notions fondamentales — mesurer pour décider

INDICATEUR Mesure calculée par une règle explicite sur un périmètre, une période et une population définis, avec une unité et une interprétation annoncées.

DÉCISION Choix d'une action ou d'une non-action, assumé par un responsable, à partir d'informations, de contraintes et d'un niveau d'incertitude explicites.

QUALITÉ DES DONNÉES Adéquation des données à un usage prévu, évaluée selon des dimensions telles que complétude, exactitude, cohérence, unicité, validité, fraîcheur et traçabilité.

LIMITE Un indicateur n'est pas une décision et une corrélation observée ne suffit pas à établir une cause.

Domaines d'application des données

Les données prennent leur valeur dans un domaine d'application : elles décrivent une activité, soutiennent une analyse et peuvent orienter une action.

01 • Industrie

Données : production, maintenance, qualité.

Usages : détecter les anomalies et optimiser les opérations.

02 • Services

Données : clients, transactions, tickets.

Usages : suivre l'activité et améliorer l'expérience utilisateur.

03 • Finance et assurance

Données : comptes, paiements, sinistres.

Usages : évaluer les risques et assurer la conformité.

04 • Recherche

Données : observations, mesures, résultats.

Usages : produire, vérifier et réutiliser des connaissances.

05 • Énergie

Données : compteurs, équipements, météo.

Usages : mesurer la performance et planifier les actions.

06 • Administration

Données : registres, démarches, territoires.

Usages : piloter les politiques publiques et les services.

Points communs à tous les domaines

- Quel que soit le domaine, les données doivent être **décrites et qualifiées** pour être comprises et jugées fiables.

Quatre dimensions pour une donnée maîtrisée



Une donnée maîtrisée est une donnée que l'on peut expliquer, prendre en charge, préserver et utiliser de façon responsable.

01 — Sémantique. Définit le sens, l'unité, le contexte et l'interprétation de la donnée.

02 — Organisationnelle. Précise les rôles, les responsabilités, les processus et les décisions associés.

03 — Technique. Garantit le stockage, la qualité, la sécurité, l'accès et la conservation.

04 — Juridique / éthique. Encadre les droits, les risques et les usages acceptables.

Trois formes de données, trois façons de les lire



Structurée

Schéma explicite, stable et contraint : attributs, types et relations sont définis avant l'enregistrement.

Formats : CSV, SQL, feuille tabulaire, Parquet.

Exemples : clients, transactions, catalogue, capteurs.

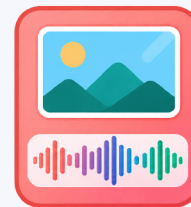


Semi-structurée

Clés, balises ou objets imbriqués : structure lisible par une machine, mais variable entre les enregistrements.

Formats : JSON, XML, YAML, NoSQL.

Exemples : API, configuration, document métier.



Non structurée

Contenu sans schéma tabulaire prédéfini ; son exploitation exige souvent extraction ou classification.

Formats : JPEG, PNG, PDF, MP3, WAV, vidéo.

Exemples : photo, audio, vidéo, courriel, rapport libre.

Une donnée peut avoir un format technique sans avoir une structure sémantique directement exploitable.

Une même observation, deux représentations

Un format décrit la manière d'encoder une donnée ; il ne modifie pas le fait observé.

CSV — représentation tabulaire

```
batiment,date,energie_kwh  
A,2026-01-12,120
```

- L'en-tête porte le schéma ; chaque ligne décrit une observation.
- Les colonnes gardent le même sens ; séparateur, types, unités et valeurs manquantes sont documentés.
- **Atout** : lecture directe avec un tableur, SQL ou Pandas

Invariant sémantique

Le bâtiment A a consommé 120 kWh le 12 janvier 2026.

JSON — représentation par objet

```
{"batiment": "A", "date": "2026-01-12",  
  "energie": {"valeur": 120, "unite": "kwh"}}
```

- Les clés et les niveaux imbriqués regroupent des propriétés liées, comme la valeur et son unité.
- Les clés rendent le contexte explicite ; leur cohérence doit être contrôlée.
- **Atout** : échanger des objets riches ou des structures

Règle de conversion

Changer l'encodage, jamais le sens : préserver l'unité, la période, le périmètre et les valeurs.

De l'image au tableau exploitable



La photo conserve une organisation visuelle, pas encore des champs directement interrogeables.

Point de départ

Une photo n'est pas encore un tableau structuré : elle contient des pixels, une mise en page et des métadonnées éventuelles.

1. Extraire

- reconnaître les caractères et repérer les lignes et colonnes ;
- associer chaque valeur au bon champ.

2. Structurer et contrôler

- produire des champs nommés, des lignes et des types explicites ;
- vérifier dates, unités, valeurs manquantes et erreurs de lecture.

Règle de management des données

Une représentation structurée facilite le traitement, mais la qualité doit encore être démontrée.

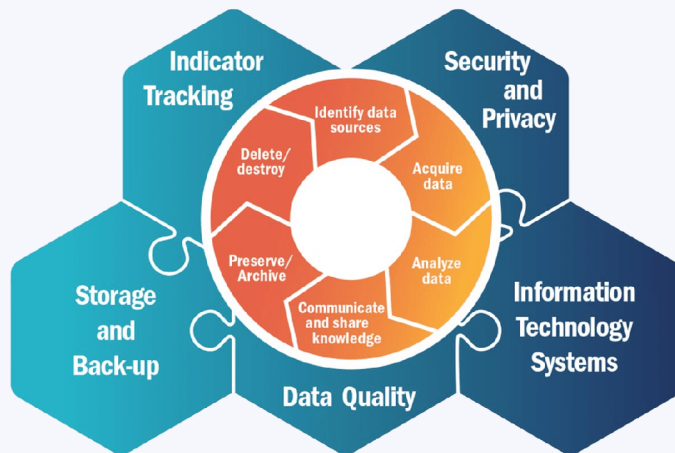
CHAPITRE

Fondations, cycle de vie et sources

02

Cycle de vie des données

Le cycle de vie des données



- Le cycle est révisable ; une anomalie, un nouvel usage ou une contrainte peut renvoyer à une étape antérieure.

DÉFINITION Le cycle de vie des données est l'ensemble des étapes qu'une donnée traverse : identifier sa source, l'acquérir, l'analyser, communiquer les résultats, la préserver ou l'archiver, puis la supprimer ou la détruire selon une règle explicite.

Garanties transversales

- documenter le sens, la provenance et les transformations ;
- contrôler la qualité et signaler les limites ;
- protéger les accès, les copies et les usages.

Garanties transversales : rendre chaque étape vérifiable

Une garantie transversale accompagne chaque étape du cycle : elle répond à une question concrète et laisse une preuve vérifiable.

Documenter

Que signifie la donnée et d'où vient-elle ?

Décrire sens, unité, période, source et producteur.

Preuve : fiche source et métadonnées.

Qualifier

Peut-on lui faire confiance pour cet usage ?

Contrôler couverture, complétude, cohérence, fraîcheur et anomalies.

Preuve : rapport de qualité et exceptions.

Protéger

Qui peut y accéder et dans quelles conditions ?

Limiter les accès, encadrer les usages et tracer les consultations sensibles.

Preuve : règle d'accès et journal des usages.

Préserver

Pourra-t-on retrouver la donnée et ses versions ?

Conserver le brut, les dérivés et les copies contrôlées pendant la durée définie.

Preuve : sauvegarde, version et responsable.

À retenir : ces garanties ne constituent pas une étape supplémentaire ; elles accompagnent l'identification, l'acquisition, l'analyse, le partage, la préservation et la suppression.

Identifier les sources

DÉFINITION Identifier les sources consiste à repérer l'origine, le périmètre et les responsables d'une donnée.

À examiner avant l'acquisition

- la question, la finalité, le périmètre et la période ;
- l'unité d'observation, la granularité et la fréquence ;
- le producteur, le responsable et les conditions d'accès ;
- le format et les conditions de réutilisation.

Enjeu général : une source identifiable permet d'évaluer la pertinence de la donnée avant de lui faire confiance.

Sortie : une source candidate est retenue, écartée ou conservée sous réserve, avec une justification minimale.

Sortie de l'étape : une source qualifiée

Question



Source candidate



Choix justifié

La fiche détaillée vient ensuite

Elle rassemble l'identité, le périmètre, l'accès, le format, la licence et les traces nécessaires à la réutilisation.

La section Sources approfondit cette qualification et la preuve d'acquisition.

Acquérir les données

DÉFINITION Acquérir consiste à obtenir une représentation de données depuis une source, en conservant la trace de l'opération.

Acquisition — entrée brute et traçabilité.

Obtenir une représentation exploitable

- conserver l'entrée brute et sa date d'acquisition ;
- enregistrer les paramètres et la version de la source ;
- séparer le brut des copies transformées ;
- protéger l'accès à la source et aux copies.

Enjeu général : l'acquisition doit être répétable, vérifiable et réversible autant que possible.

Règle : un tiers doit retrouver la source, le moment et les paramètres utilisés.

Analyser les données

DÉFINITION Analyser consiste à produire une information ou un indicateur à partir de données qualifiées, selon une règle explicite.

Qualifier avant d'interpréter

- l'unité, la fréquence et la période ;
- la granularité et la couverture attendue ;
- les valeurs manquantes et les anomalies ;
- les règles de transformation et d'agrégation.

Enjeu général : une analyse transforme des observations en indicateurs, sans effacer les anomalies ni les limites.

Deux fichiers peuvent parler d'un même phénomène tout en étant incompatibles si leur granularité diffère.

Error: Could not find a suitable point for the given distance

De l'observation au résultat — contrôler chaque transformation.

Lecture : l'analyse qualifie les données, applique une règle explicite puis produit une représentation adaptée à la question.

Limite : un résultat n'est interprétable que si la qualité, la méthode et le contexte sont documentés.

Communiquer et partager

DÉFINITION Communiquer et partager consiste à rendre un résultat compréhensible et accessible dans des conditions d'usage explicites.

Partage — préparer, documenter, autoriser puis réutiliser.

Rendre le résultat réutilisable

- décrire le contenu, le périmètre et la période ;
- indiquer la provenance, la version et les transformations ;
- documenter l'unité, les valeurs manquantes et les limites.

Enjeu général : partager ne consiste pas seulement à transmettre un fichier ; il faut transmettre son contexte.

Point de vigilance : adapter le niveau de détail aux utilisateurs, aux risques et aux conditions d'usage.

Préserver et archiver

DÉFINITION Préserver et archiver consiste à maintenir une donnée retrouvable pendant la durée justifiée par son usage ou sa valeur.

Conservation — distinguer l'usage actif, la préservation et l'archivage.

Conserver selon le besoin

- **Préserver** : garder la donnée disponible pour un usage actif ;
- **Archiver** : la conserver pour une consultation ultérieure ;
- maintenir les versions, les preuves et les copies contrôlées.

Enjeu général : distinguer la disponibilité immédiate de la conservation à long terme.

Point de vigilance : définir durée, responsable, support et traitement des copies.

Supprimer ou détruire

DÉFINITION Supprimer ou détruire consiste à mettre fin de manière autorisée et traçable à la conservation d'une donnée.

Suppression — décision, exécution et preuve.

Mettre fin à la conservation

- vérifier l'échéance et les obligations applicables ;
- identifier les copies, dérivés et sauvegardes concernés ;
- enregistrer la décision, le responsable et la date d'exécution ;
- conserver la preuve de suppression ou de destruction.

Enjeu général : la suppression doit porter sur le bon périmètre et respecter les obligations de conservation.

Règle : supprimer n'est pas effacer silencieusement ; c'est une opération autorisée, traçable et vérifiable.

03

Identification et documentation des sources

Une source ne se résume pas à une URL

DÉFINITION Une source est un dispositif qui produit, publie ou met à disposition une représentation de données. Pour la réutiliser, il faut pouvoir en comprendre l'origine, le contenu, le mode d'accès et les conditions d'usage.

- **Qui ?** producteur, responsable et contact.
- **Quoi ?** objet, périmètre, unité et période.
- **Comment ?** accès, paramètres, version et instant de collecte.



Une adresse localise une ressource ; la documentation permet d'identifier l'observation réellement utilisée.

Droits d'utilisation des données disponibles sur le Web

PRINCIPE Une donnée accessible techniquement n'est pas nécessairement librement réutilisable. L'usage dépend de la licence, de la finalité autorisée et des conditions fixées par le producteur ou le site.

Avant l'acquisition

- identifier le producteur et la licence ;
- distinguer le droit d'auteur, les droits sur une base de données et les conditions contractuelles du site ;
- vérifier les règles applicables aux données personnelles, l'attribution, les restrictions et la finalité ;
- prévoir la mention de la version et de la date d'accès.

Accès

La ressource peut-elle être obtenue ?

Licence / droit d'auteur

Quel droit encadre le contenu ?

Base de données

L'extraction et la réutilisation sont-elles permises ?

Données personnelles

Quelles obligations de protection s'appliquent ?

Conditions contractuelles

Quelles règles le site impose-t-il ?

Attribution / diffusion

Qui citer et quelles limites respecter ?

Règle de prudence

En cas de licence absente ou ambiguë, limiter ou suspendre la réutilisation et demander une clarification. Ne jamais contourner une authentification, une restriction technique ou une protection du site.

Trace : conserver l'URL, la licence, l'attribution, la version et la date de consultation.

Types de sources : qualifier avant de collecter



DÉFINITION Le type de source décrit la manière dont une donnée est produite, obtenue ou maintenue. Il ne doit pas être confondu avec son format, son mode d'accès ou son niveau de qualité.

Source primaire ou originale

Produit ou enregistre directement l'observation étudiée.

Source secondaire ou dérivée

Réutilise, agrège, transforme ou republie des données déjà produites.

Source de référence

Est officiellement désignée comme référence maintenue pour un domaine ou un attribut.

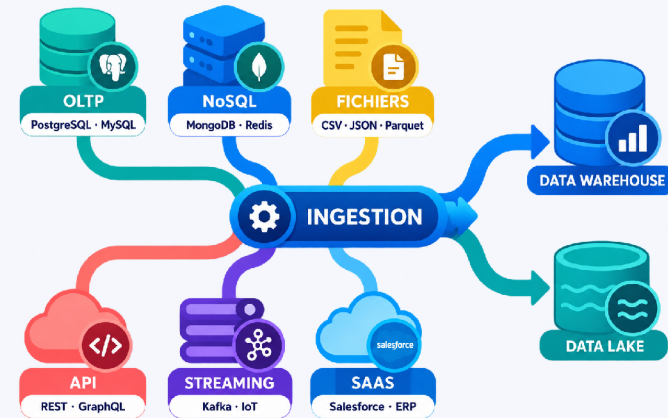
À retenir : une source de référence peut être primaire ou secondaire ; « interne/externe » décrit un autre axe de qualification.

Sources des données en data engineering

DÉFINITION Une **source de données** est un point d'origine qui fournit des données brutes ou quasi brutes avant leur ingestion et leur transformation.

- **Bases et fichiers** : OLTP, NoSQL, fichiers et stockage objet.
- **Services et applications** : API, services web, SaaS, CRM et ERP.
- **Événements** : streaming, messages applicatifs et capteurs IoT.

Point de méthode : le format et le protocole d'accès se documentent séparément du type de source.



Data warehouse, data lake et data lakehouse

Trois environnements cibles, trois logiques d'usage :

01 — Data warehouse

Entrepôt intégré et structuré : données nettoyées pour la business intelligence, le reporting et les indicateurs.

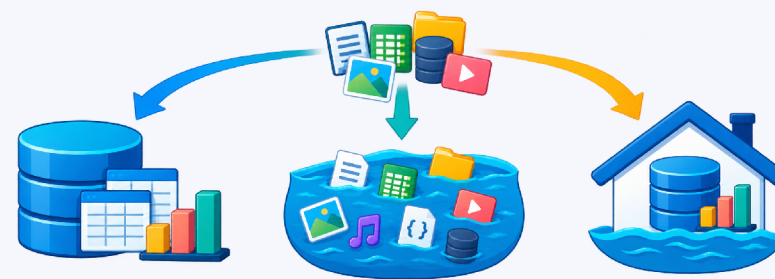
02 — Data lake

Lac de données flexible : données brutes ou transformées, structurées, semi-structurées ou non structurées.

03 — Data lakehouse

Architecture unifiée : souplesse de stockage du lake et performances analytiques du warehouse.

À retenir : ce sont des environnements de stockage et d'exploitation ; une architecture peut en combiner plusieurs.



Même flux de données, environnements adaptés à des usages différents.

Structurer
BI et reporting

Conserver
données variées

Unifier
analytics et ML

Évaluer une source : la grille 5P

Cette grille proposée dans le cours aide à vérifier cinq conditions avant de réutiliser une source.



Les 5P transforment une intuition sur la source en décision documentée.

Qui ?
Provenance

Quoi ?
Périmètre

Quand ?
Périodicité

Sous quelles conditions ?
Permission

Quelles traces ?
Preuve

Règle : une réponse inconnue devient une limite à documenter.

Une source se décrit sur trois axes

Trois niveaux à distinguer : *qui* produit ou publie la donnée ?
comment l'obtenir ? *sous quelle forme* est-elle transmise ?

01 — Origine

Producteur, responsable, ressource et contexte de création.

À consigner : organisation, rôle, intitulé et date de publication.

02 — Mode d'accès

Canal ou opération utilisé pour obtenir la donnée et retracer l'acquisition.

À consigner : URL ou endpoint, protocole, authentification et date de collecte.

03 — Format

Représentation technique qui détermine comment la donnée se lit et se convertit.

À consigner : type de fichier, encodage, schéma, compression et version.

Exemple filé

Un même jeu de données peut être décrit ainsi :

Origine — catalogue institutionnel
producteur : organisme public

Accès — API REST
requête documentée et datée

Format — JSON
clés et structure à interpréter

À retenir : changer de format modifie la représentation, pas l'origine de la donnée. La conversion doit toutefois être documentée et attribuée.

Choisir un mode d'accès et conserver la preuve

DÉFINITION **Acquérir** consiste à obtenir une représentation de données depuis une source, selon un accès documenté et en conservant la preuve de l'opération.

Critère de choix : retenir le mode qui répond au besoin, respecte les conditions d'accès et peut être rejoué ou vérifié.

Fichier publié — télécharger une version précise.
Preuve : URL, date, version ou hash, fichier original.

API — envoyer une requête paramétrée.
Preuve : endpoint, paramètres, horodatage et réponse brute ; exclure les secrets.

Page HTML — extraire des éléments du document.
Preuve : URL, instant, sélecteurs, méthode et copie de référence.

Vérifier avant d'acquérir

- la licence, les conditions d'utilisation et les limites de diffusion ;
- la stabilité de la page, de l'API ou de la version publiée ;
- la possibilité de conserver le brut et de rejouer l'opération.

Règle de traçabilité

Le mode d'accès indique *comment* obtenir la donnée ; il ne définit pas *qui* en est le producteur.

TP 1 — 4 h : pratiquer les trois modes avec Python, puis diagnostiquer et nettoyer le jeu en conservant le brut, la requête et la méthode d'extraction.

La fiche source : un contrat minimal de réutilisation

Elle rassemble les éléments nécessaires pour qu'une autre personne retrouve, comprenne et discute la ressource utilisée.

Identité

titre, identifiant, producteur, contact et finalité.

Acquisition

URL ou chemin, méthode, paramètres, instant et version.

Périmètre

objet, population, période, lieu, unité et définition.

Réutilisation

format, licence, qualité connue, limites et conditions d'usage.

CHAMP	EXEMPLE À CONSERVER
Version et brut	v1.2 ; fichier original non modifié
Accès et instant	endpoint ou chemin ; date et fuseau
Définition et qualité	unités, couverture, valeurs manquantes
Permission et limite	licence, restrictions et usages exclus

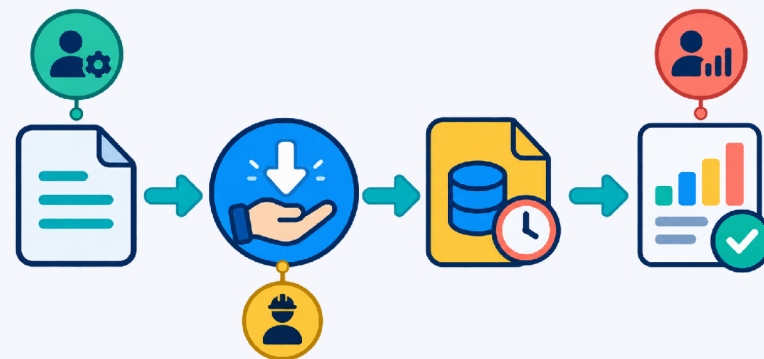
Règle : une permission non vérifiée reste une question ouverte, pas une autorisation implicite.

La provenance relie la source au résultat

DÉFINITION La provenance est l'histoire d'une donnée : son origine, les activités qui l'ont transformée et les agents qui y ont contribué.

- **Entité** : fichier, réponse ou table considérée comme un objet.
- **Activité** : collecte, conversion, nettoyage ou analyse.
- **Agent** : producteur, opérateur ou responsable de validation.

TEST DE TRAÇABILITÉ D'où vient la donnée ? Qu'a-t-on fait ? Qui peut expliquer l'écart entre l'entrée et le résultat ?



Une transformation devient explicable lorsqu'elle relie l'entrée, l'opération, la version et le responsable.

Vérifier avant de réutiliser

Une donnée documentée n'est pas automatiquement une donnée adaptée. La dernière vérification porte sur l'usage envisagé.

Retrouver

La ressource, sa version, son instant et son fichier brut.

Comprendre

Le sens, le périmètre, l'unité et la méthode.

Autoriser

La licence, les restrictions et les personnes habilitées.

Déclarer

Les limites, les anomalies et les usages non justifiés.

TRANSITION

Documenter une source suppose aussi de savoir qui la produit, qui la contrôle et qui autorise son usage. La prochaine section répartit ces responsabilités entre les acteurs.

04

Acteurs et responsabilités

Les acteurs et les responsabilités

DÉFINITION Un **acteur** est une personne, une équipe ou un système auquel une activité, une décision ou une obligation de preuve est attribuée dans le cycle de vie des données.

Producteur

Crée, saisit ou enregistre une observation dans un dispositif.

Data steward

Maintient les définitions, métadonnées, règles de qualité et anomalies.

Data analyst

Produit une analyse en documentant méthode, périmètre et limites.

Propriétaire métier

Porte le sens, la finalité et les conditions d'usage d'un domaine.

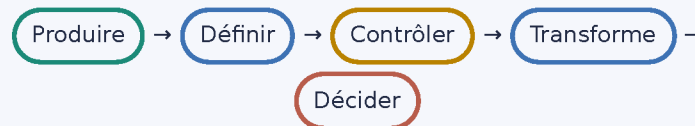
Data engineer

Conçoit les flux, stockages et transformations qui rendent les données disponibles.

Décideur

Interprète le résultat et assume la décision et ses conséquences.

Un même parcours, plusieurs responsabilités



Exemple

Le dispositif, le métier, le steward, l'équipe data et le responsable habilité contribuent chacun à une étape distincte.

Déclinaison campus

Le service patrimoine, le référent énergie et l'équipe data peuvent porter ces fonctions. Ces intitulés sont illustratifs : une même personne peut cumuler plusieurs rôles, mais les responsabilités doivent rester explicites.

Une décision de donnée distribue quatre responsabilités

GOUVERNANCE DES DONNÉES

Organisation des droits de décision, des responsabilités, des contrôles et des règles d'usage.

Décider — qui autorise l'usage et en assume la finalité ?

Exécuter — qui collecte, transforme ou diffuse la donnée ?

Contrôler — qui vérifie la qualité, les accès et les conditions ?

Expliquer — qui peut justifier le sens, la méthode et les limites ?

Exemple : publier un indicateur pour piloter une action

Décider : le responsable métier habilité autorise la finalité et l'usage.

Exécuter : l'ingénieur ou l'analyste collecte, transforme et diffuse le résultat.

Contrôler : le steward et le responsable d'accès vérifient qualité, droits et conditions.

Expliquer : l'analyste rend la méthode, la provenance et les limites compréhensibles.

Déclinaison campus : avant de classer les bâtiments, le responsable doit savoir si la couverture, l'unité et le périmètre autorisent la comparaison.

À retenir : une responsabilité se traduit par une décision, une action et une preuve.

Une responsabilité se prouve par une trace

RÈGLE Une responsabilité est traçable lorsqu'une trace permet de relier un acteur habilité, une décision ou une action, sa date et la preuve produite.

La question n'est pas seulement « qui intervient ? », mais « qui répond de quoi, à quel moment et avec quelle preuve ? »

01 — Qualité

Qui qualifie une anomalie et valide la correction ?

Exemples : valeur incohérente ou observation absente.

Campus : confirmer les deux heures manquantes pour C.

Preuve : fiche d'anomalie, correction et validation.

02 — Accès

Qui autorise l'usage, pour quelle finalité et avec quelles limites ?

Exemples : partager un indicateur avec un nouveau destinataire.

Campus : comparer les bâtiments sans exposer de données inutiles.

Preuve : règle, autorisation et journal des accès.

03 — Conservation

Qui fixe la durée et décide d'archiver ou de supprimer ?

Exemples : appliquer une échéance à une série de données.

Campus : conserver le brut, les dérivés et leurs versions.

Preuve : calendrier, décision et journal d'exécution.

Une gouvernance devient vérifiable lorsque l'on peut reconstituer : acteur → action → date → preuve.

Trois capacités transversales de gouvernance

DÉFINITION Une **capacité de gouvernance** désigne un ensemble organisé de pratiques, de responsabilités et de contrôles permettant de maîtriser les données.

Curate — structurer et qualifier

Organiser, décrire et évaluer les données pour qu'elles soient retrouvables et interprétables.

Exemples :

- données clients ou produits : catalogue, propriétaires et règles de qualité.

Protect — protéger et encadrer

Maîtriser les accès, les risques, les droits et les conditions d'utilisation.

Exemples :

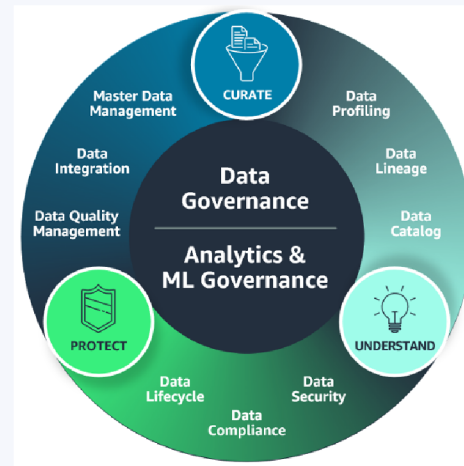
- données RH ou financières : accès selon les rôles et journaux de consultation.

Understand — comprendre et expliquer

Rendre le sens, la provenance, les transformations et les limites compréhensibles.

Exemples :

- indicateur d'activité : source ERP, transformations et tableau de bord.



À retenir : une donnée maîtrisée est à la fois organisée, protégée et explicable.

Source : AWS — Data Governance.

05

Valeur et risques des données

Valeur et risques : décider de l'usage

VALEUR

Bénéfice concret attendu d'un usage de données : mieux diagnostiquer, coordonner une activité, réduire un coût ou éclairer une décision.

RISQUE

Éventualité d'une conséquence défavorable liée à une donnée, à un traitement ou à un usage : erreur d'interprétation, décision injuste, exposition inutile ou résultat impossible à justifier.

La valeur et le risque dépendent du contexte d'usage



Finalité

Quelle décision ou quel service l'usage doit-il soutenir ?

Public

Qui peut consulter, interpréter ou réutiliser le résultat ?

Preuves

Quelles traces permettent de vérifier le résultat ?

Limites

Quel risque reste présent malgré les contrôles ?

Exemples

Professionnel : un indicateur de maintenance peut prioriser une intervention, mais une mesure incomplète peut déplacer le risque vers un autre équipement.

Campus : les relevés peuvent cibler un diagnostic énergétique, mais un classement fondé sur C incomplet peut conduire à une mauvaise priorité.

Un indicateur répond à une question précise

Mesure absolue — V

Question : quelle quantité totale a été observée ?

Usage : estimer une charge, un volume ou un coût global.

Limite : dépend de la taille, de l'exposition ou du volume d'activité.

Contrôle : vérifier période, unité et couverture.

Mesure normalisée — R

Question : quelle quantité rapportée à un dénominateur faut-il comparer ?

Formule : $R = V / D$, avec un dénominateur défini et comparable.

Limite : dépend de la qualité du dénominateur et du contexte.

Contrôle : vérifier que D est mesuré sur le même périmètre.

Exemples

- industrie : énergie totale d'une ligne / énergie par unité produite ;
- services : nombre d'incidents / incidents pour 1 000 utilisateurs ;
- campus : kWh observés / kWh par m².

Principe

Le dénominateur transforme la question ; il ne rend pas automatiquement les entités comparables. Sa définition, sa période et sa qualité doivent être documentées.

Lecture rapide : V répond à « combien ? » ; $R = V / D$ répond à « combien par rapport à quelle base ? ».

Le classement dépend de la question

Un classement est défendable seulement si les observations sont comparables sur les dimensions qui déterminent la question.

Objet

Les unités comparées mesurent-elles le même phénomène ?

Période

Les observations couvrent-elles le même intervalle temporel ?

Dénominateur

La base de normalisation est-elle définie, stable et pertinente ?

Couverture

Les valeurs manquantes ou exclusions ont-elles le même traitement ?

Exemples : comparer des lignes de production, des services pour 1 000 utilisateurs ou des bâtiments selon leur consommation totale et leur surface.

À retenir : un indicateur décrit une relation mesurée ; il n'explique pas à lui seul sa cause et ne constitue pas une preuve de performance.

Conséquence : si l'objet, la période, le dénominateur ou la couverture diffère, le classement doit être qualifié ou suspendu.

Quatre critères avant tout usage

Avant de diffuser ou d'automatiser un résultat, il faut vérifier ses conditions d'emploi. Les quatre critères forment une porte de contrôle, pas une nouvelle étape technique.

01 — Pertinence

La donnée répond-elle à la question et au périmètre étudié ?

Preuve : finalité, périmètre et indicateur définis.

Exemple : un KPI de fidélisation doit mesurer la population réellement concernée.

03 — Légitimité

Finalité, droits, proportionnalité et conditions d'accès sont-ils établis ?

Preuve : règle d'accès, licence ou validation d'usage.

Exemple : une donnée RH n'est accessible qu'aux fonctions autorisées.

02 — Qualité

Erreurs, manques, biais et incertitudes sont-ils connus et mesurables ?

Preuve : contrôles, couverture et rapport d'anomalies.

Exemple : une série de capteurs doit signaler ses intervalles manquants.

04 — Responsabilité

Un acteur habilité peut-il contrôler, expliquer et assumer le résultat ?

Preuve : rôle nommé, validation et journal de décision.

Exemple : un responsable métier valide l'usage d'un indicateur réglementaire.

Arbitrer un usage : publier, différer ou limiter

CAS GÉNÉRAL

Une équipe souhaite diffuser un résultat alors qu'une partie des preuves reste incomplète. Le niveau d'affirmation doit rester proportionné à ce qui a été vérifié.

Décision à prendre : quel niveau de diffusion est compatible avec la pertinence, la qualité, la légitimité et la responsabilité établies ?

Publier

Le périmètre, la qualité et les conditions d'usage couvrent le résultat annoncé.

Trace : version et validation.

Différer

Une donnée manquante, un contrôle ou une validation peut modifier l'interprétation.

Trace : anomalie et action attendue.

Limiter

Un constat partiel reste utile si sa couverture, son incertitude et ses limites sont explicites.

Trace : réserve et périmètre.

RÈGLE DE DÉCISION

Choisir l'usage le moins affirmatif compatible avec les preuves disponibles. Une donnée incomplète peut soutenir un diagnostic exploratoire, mais pas nécessairement un classement ou une décision définitive.

FAIR : vérifier la réutilisabilité

DÉFINITION **FAIR** est un cadre de réutilisabilité : une ressource doit être retrouvable, accessible, interopérable et réutilisable pour un usage explicite.

Findable — trouvable

Identifiée par un nom, une version, un identifiant ou un emplacement.

Accessible — accessible

Obtenue par un accès documenté, avec responsable, droits et restrictions.

Interoperable — interopérable

Interprétable par d'autres outils grâce au schéma, aux unités et aux vocabulaires.

Reusable — réutilisable

Reprise correcte grâce à la provenance, la qualité, les limites, la licence et le contexte.

Preuves typiques : identifiant, schéma, dictionnaire, licence, journal de qualité et historique des versions.

Exemples

- professionnel : un indicateur mensuel de ventes est catalogué, accessible au service autorisé, exprimé dans un schéma stable et accompagné de sa source ERP ;
- campus : une série énergétique indique son emplacement, son mode d'accès, ses unités, sa période, sa couverture et ses conditions de réutilisation.

Attention : FAIR ne signifie pas « donnée parfaite ». Le cadre rend les conditions de réutilisation explicites ; il ne remplace ni les contrôles de qualité ni la gouvernance.

Réutilisateur
retrouve, comprend et cite la ressource

Producteur
décrit, maintient et signale les limites

Décider sous contrainte : valeur, risque et preuve



Questionner

Définir la finalité, le périmètre et l'indicateur adapté.

Qualifier

Vérifier qualité, légitimité, provenance, accès et réutilisabilité.

Assumer

Nommer le responsable, déclarer les limites et conserver les preuves.

Dossier minimal de décision : finalité, indicateur, contrôles effectués, responsable nommé et limites déclarées.

Une décision de données est défendable lorsqu'elle relie une question, un indicateur, des preuves et un responsable.

06 Synthèse

Les acquis essentiels : de la donnée à la décision

01 — Décrire

Préciser le sens, l'unité, la période, le périmètre et le format.

02 — Retracer

Relier la source, l'acquisition, les transformations, les versions et les preuves.

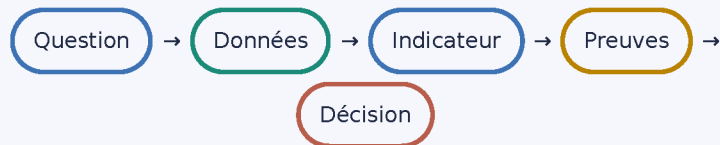
03 — Vérifier

Contrôler qualité, comparabilité, accès, légitimité et réutilisabilité.

04 — Assumer

Nommer les responsables, déclarer les limites et justifier l'usage retenu.

Une donnée devient exploitable lorsqu'elle peut être expliquée, vérifiée et réutilisée pour une finalité précise.



Fil directeur : le management des données organise les conditions qui rendent cette chaîne compréhensible, contrôlable et responsable.

Critère de réussite : un tiers doit pouvoir retrouver la source, comprendre l'indicateur, vérifier les contrôles et identifier le responsable.

La règle de synthèse : décider avec des preuves

Une décision de données est défendable lorsqu'un tiers peut en reconstituer le raisonnement et en discuter les limites.

Question

Finalité, périmètre et population concernés.

Indicateur

Formule, unité, période et dénominateur.

Preuves

Source, contrôles, versions et anomalies.

Responsable

Validation, accès, limites et décision.

Test de défendabilité : une autre personne peut-elle retrouver la donnée, comprendre le calcul, vérifier les contrôles et identifier qui répond du résultat ?

Publier

si les preuves couvrent l'usage annoncé

Différer

si une incertitude peut changer la décision

Limitier

si un constat partiel reste utile et explicable

Formulation attendue : « Nous recommandons *usage* sur *périmètre*, car *preuves* ; sous réserve de *limites*, avec validation de *responsable*. »

Prochaine séance : acquérir avec Python

TP 1 — 4 h : lire un CSV avec Pandas, interroger une API avec Requests, extraire un tableau d'une page statique avec BeautifulSoup, puis produire un diagnostic qualité et un jeu nettoyé.

source → acquisition → brut conservé → exploration → observations documentées

Livrables : notebook étudiant de la version choisie complété, fichiers bruts, fiche de source (URL, instant, méthode) et synthèse : lignes, colonnes, types, manques, doublons et limites des modes d'acquisition.

Le pipeline et les indicateurs relèvent du TP 2. Le scraping dynamique et le contournement d'accès sont hors périmètre.

Références pour aller plus loin

- USGS — Data Lifecycle : étapes et garanties transversales ; rapport de Faundeen et al., 2013, figure 1.
 - Microsoft Learn — Identify data formats : tables et représentations.
 - Microsoft Learn — Explore analytical data processing : figure Bronze / Silver / Gold.
 - AWS — Data Governance : organiser, protéger, comprendre.
 - Google Cloud — Data governance : responsabilités et cycle de conservation.
 - data.gouv.fr — Licence Ouverte 2.0 : réutilisation, attribution et conditions d'usage.
 - UK Data Service — Exercise: Data documentation : audit des informations nécessaires à la réutilisation d'un jeu de données, adapté dans le TD.
 - UK Data Service — Exercise: FAIRness of a dataset : grille Findable, Accessible, Interoperable, Reusable, adaptée dans le TD.
- Les références ne remplacent pas l'analyse : chaque source doit être remplacée dans son contexte, son périmètre et sa licence.