

05/01/2026

DATA REFINEMENT

Promo DIA2 - Hetic

Vincent Laurens

QUI SUIS-JE ?

VINCENT LAURENS

Consultant Data Cloud Engineer



Diplômé d'un MSC Data Management



Alternant Data Scientist/Engineer (1 an et demi)



DÉROULÉ DU MODULE

01

Partie théorique
(1 journée max)

02

Workshop :
Projet Data
Refinement (2 jours)

BASES ET CULTURE DATA



- COMPRENDRE CE QU'EST RÉELLEMENT UNE **DONNÉE**
- FAIRE LA **DIFFÉRENCE** ENTRE DONNÉE **BRUTE** ET DONNÉE **EXPLOITABLE**
- COMPRENDRE POURQUOI LA DATA EST AU CŒUR DES **DÉCISIONS**
- POSER LES **BASES**



QU'EST CE QU'UNE DONNÉE ?

UNE DONNÉE EST UNE **INFORMATION BRUTE**, NON INTERPRÉTÉE, DÉCRIVANT UN FAIT, UN ÉVÉNEMENT OU UNE MESURE

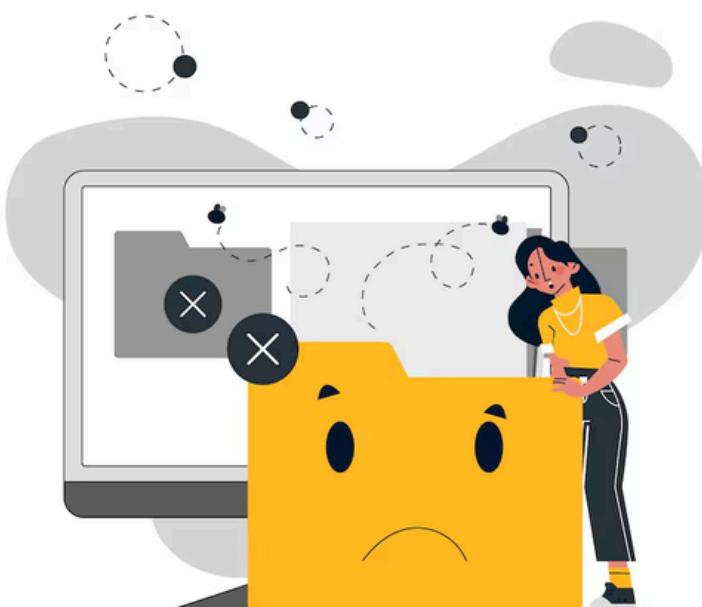
EXEMPLES :

AGE: 24

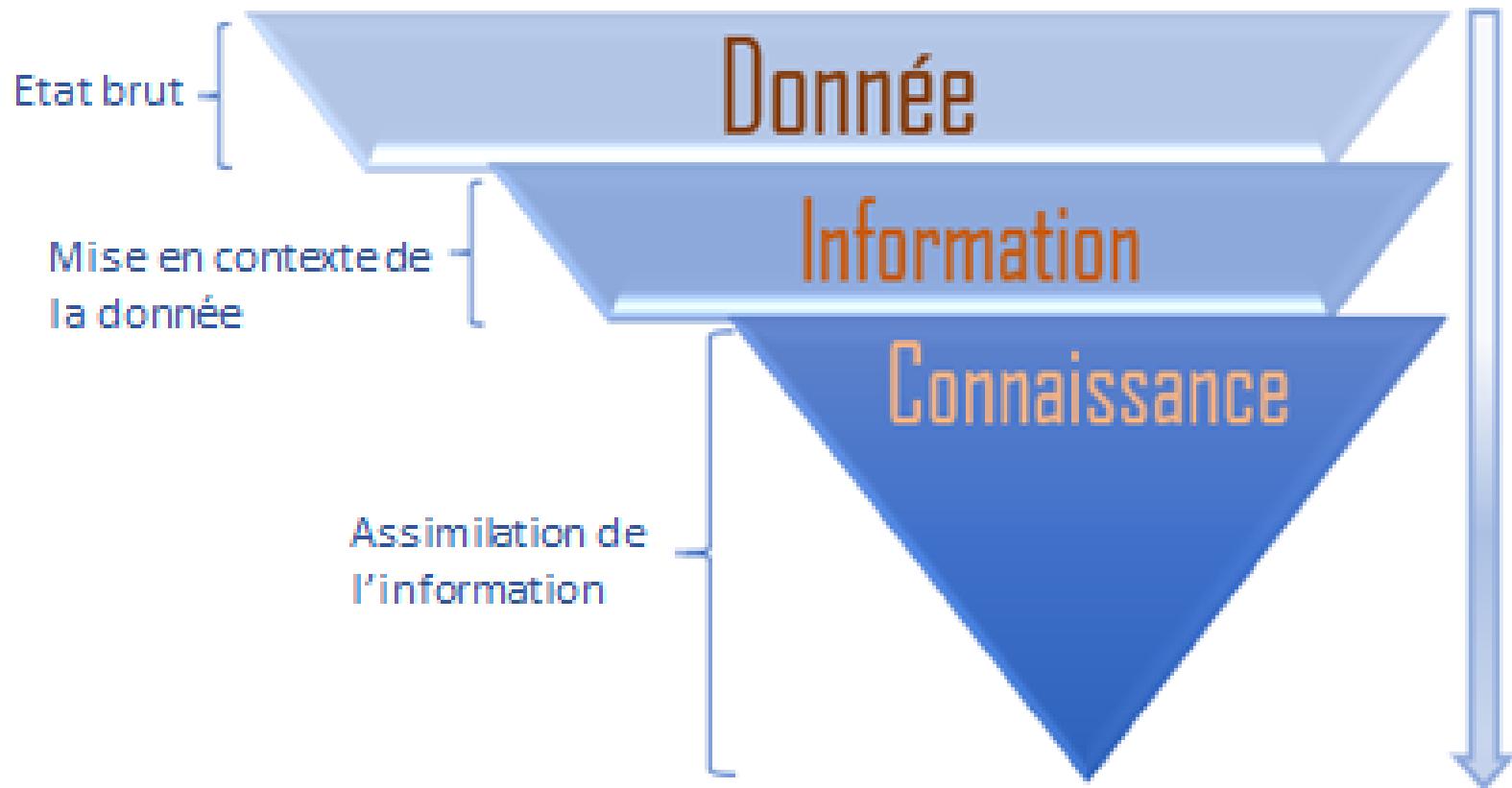
VILLE : PARIS

NOMBRE DE CLICS : 152

- UNE DONNÉE N'A **PAS DE SENS TOUTE SEULE**
- ELLE DEVIENT UTILE **QUAND ON LA CONTEXTUALISE**
- LA DONNÉE N'EST PAS **ENCORE UNE DÉCISION**



DONNÉE VS DÉCISION VS CONNAISSANCE



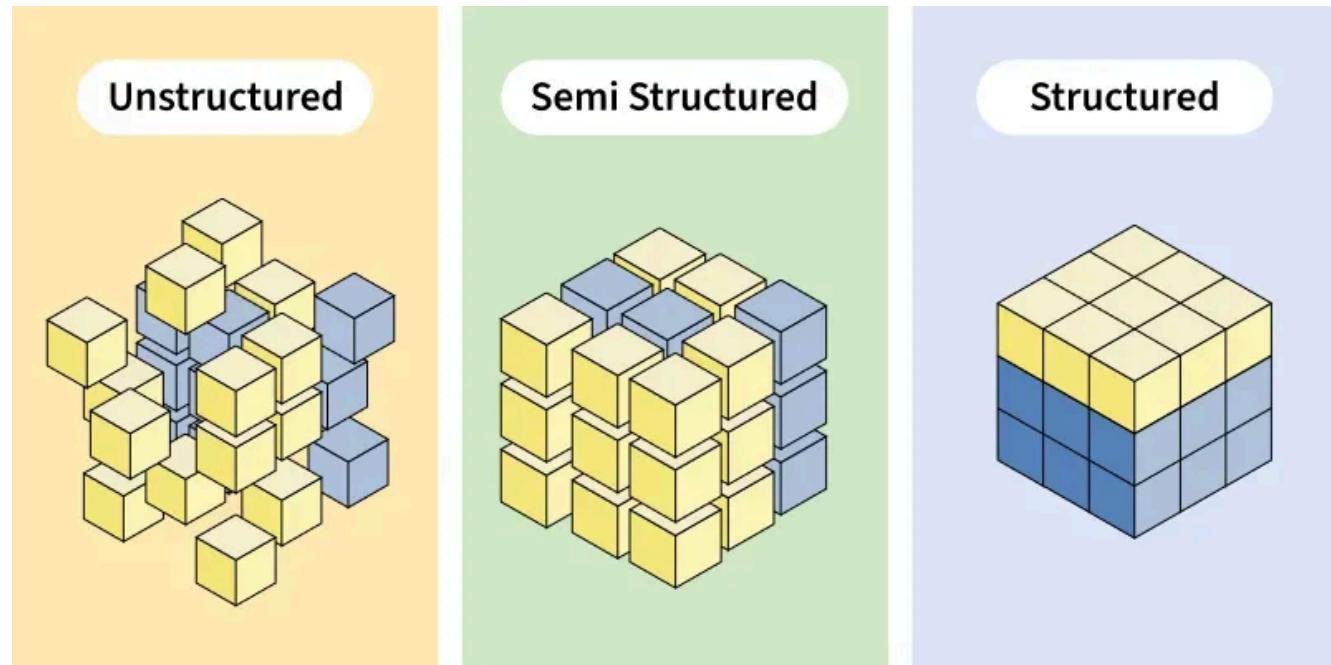
LA DATA SERT À PRENDRE DE MEILLEURES DÉCISIONS

DONNÉES BRUTE VS DONNÉE EXPLOITABLE

Critère	Donnée brute	Donnée exploitable
Définition	Donnée collectée telle quelle	Donnée prête à être utilisée
Qualité	Inconnue ou faible	Contrôlée et fiable
Valeurs manquantes	Souvent présentes	Gérées
Doublons	Possibles	Supprimés ou maîtrisés
Formats	Incohérents	Uniformes
Cohérence	Faible	Cohérente
Lisibilité	Difficile	Claire
Exploitabilité	Limitée	Directe

LES GRANDS TYPES DE DONNÉES

Type de données	Description	Exemples
Structurées	Données organisées en lignes et colonnes	Table Excel, base SQL
Semi-structurées	Données avec une structure partielle	JSON, XML, logs
Non structurées	Données sans structure fixe	Texte libre, images, vidéos



LES TYPES DE VARIABLES

Type de variable	Description	Exemples
Numérique	Valeurs mesurables ou calculables	Âge, prix, quantité
Catégorielle	Valeurs sous forme de catégories	Ville, pays, statut
Date / Temps	Informations temporelles	Date de commande, heure
Booléenne	Deux valeurs possibles	Oui / Non

IDENTIFIER LE TYPE DE VARIABLE EST INDISPENSABLE POUR BIEN NETTOYER ET TRANSFORMER LES DONNÉES

A QUOI SERT LA DATA ?

UTILISATIONS DE LA DATA :

- COMPRENDRE UN PHÉNOMÈNE
- MESURER UNE PERFORMANCE
- PRENDRE UNE DÉCISION
- PRÉDIRE UN COMPORTEMENT
- OPTIMISER UN PROCESSUS



EXEMPLES DE SECTEURS :

- MARKETING
- FINANCE
- LOGISTIQUE
- SANTÉ
- SPORT

POURQUOI LA QUALITÉ DES DONNÉES EST CRITIQUE ?

MAUVAISE DONNÉE = MAUVAISE DÉCISION



EXEMPLES :

- MAUVAIS CIBLAGE CLIENT
- ERREURS FINANCIÈRES
- MAUVAIS REPORTING
- MODÈLES BIAISÉS

LES MÉTIERS DE LA DATA



QUELLES DIFFÉRENCES ENTRE LES PROFILS DATA ?

Data Engineer



ML Engineer



Data Scientist

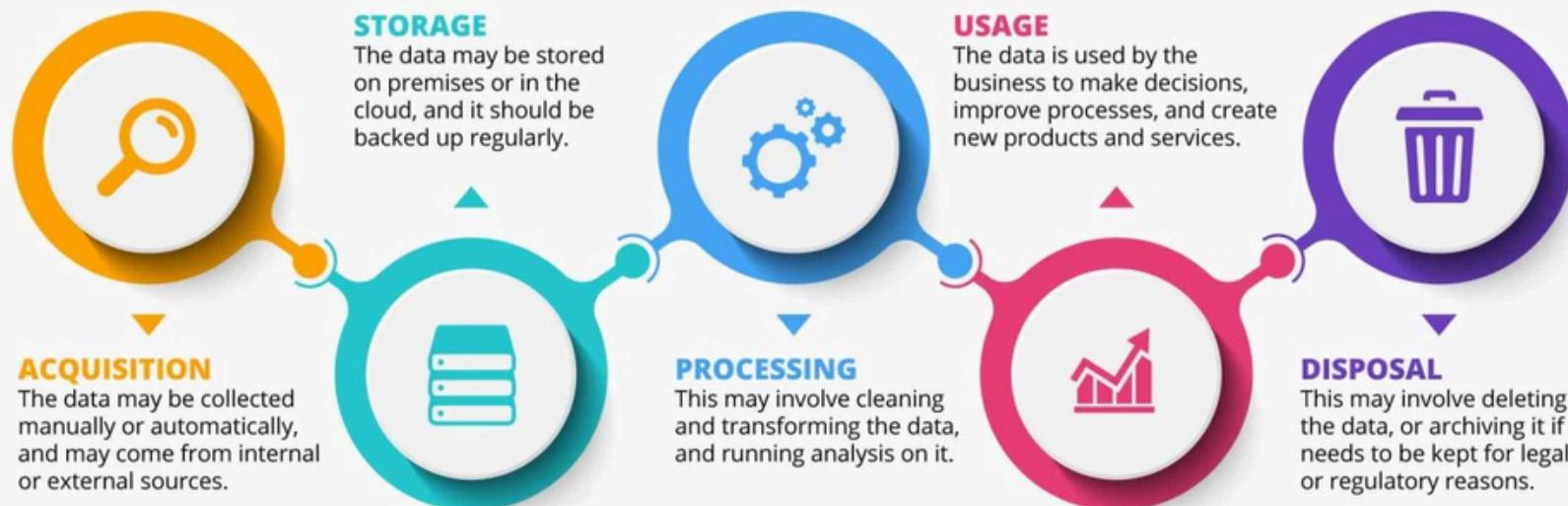


Data Analyst



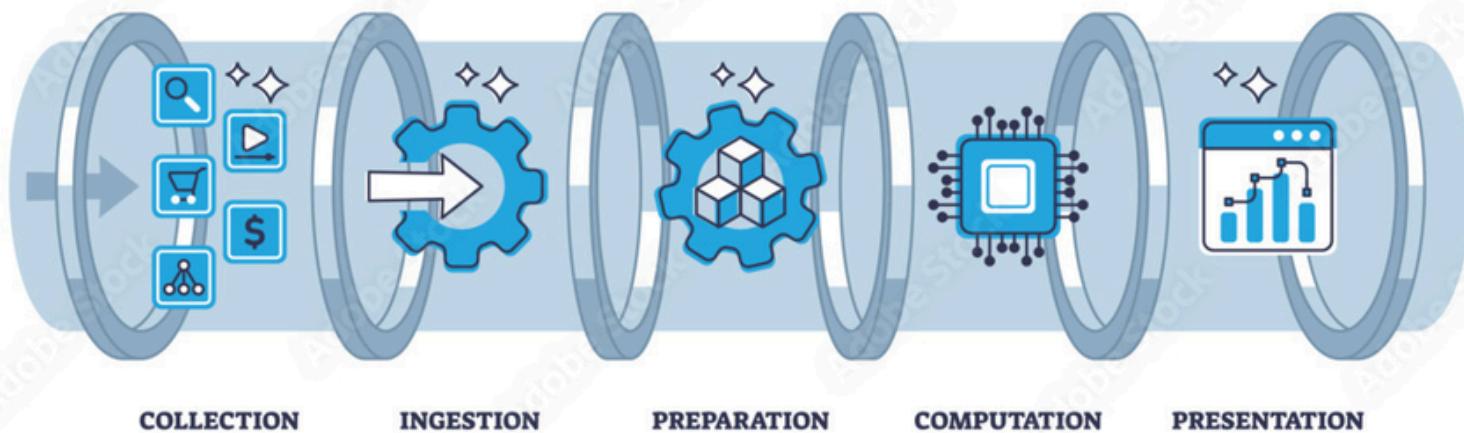
LE CYCLE DE VIE D'UNE DONNÉE

DATA LIFECYCLE MANAGEMENT



INGESTION ET TRANSFORMATION DES DONNÉES

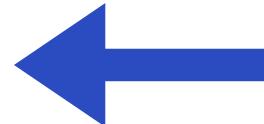
DATA PIPELINE



EXTRACT (E)

COLLECTER LES DONNÉES DEPUIS DIVERSES SOURCES

DATA SOURCES



OUTILS



LOAD (L) - DANS L'ETL

INSÉRER LES DONNÉES **TRANSFORMÉES** DANS UN SYSTÈME CIBLE.

TYPES DE FORMATS



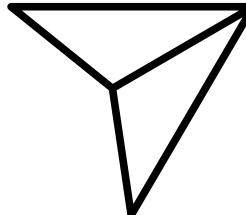
FORMATS BRUTS



FORMATS COLONNES



DONNÉES
STRUCTURÉES

- 
- MODES DE CHARGEMENTS :**
- **FULL** LOAD (ÉCRASEMENT COMPLET)
 - **INCREMENTAL** LOAD (DELTA/APPEND)

CIBLES COURANTES



DATABASES (SGBD, DATAWAREHOUSE)



DATALAKE (GCS/S3)



SERVEUR INTERNE (FTP/SFTP)

LOAD (L) - DANS L'ELT

INSÉRER LES DONNÉES **EXTRAITES** DANS UN SYSTÈME CIBLE.

TYPES DE FORMATS



FORMATS BRUTS

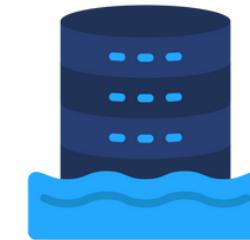


FORMATS COLONNES

CIBLES COURANTES



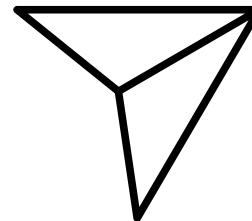
DATABASES (SGBD, DATAWAREHOUSE)



DATALAKE (GCS/S3)



SERVEUR INTERNE (FTP/SFTP)



MODES DE CHARGEMENTS :

- **FULL** LOAD (ÉCRASEMENT COMPLET)
- **INCREMENTAL** LOAD (DELTA/APPEND)

TRANSFORM (T)

NETTOYER, ENRICHIR ET STRUCTURER LES DONNÉES

TYPES DE TRANSFORMATIONS



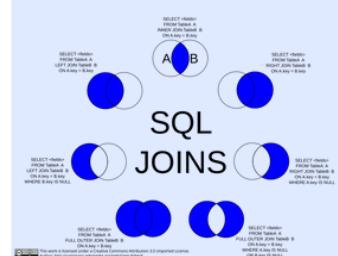
NULLS, TYPES,
DOUBLONS



MAPPING

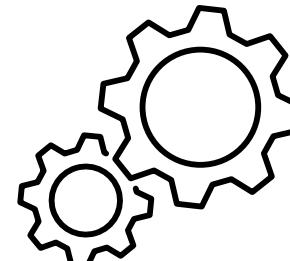


AGRÉGER LA
DONNÉE



JOINTURES

OUTILS



ANALYSE DE DONNÉES

COMPRENDRE LA DONNÉE

Process of Data Analysis

Visualization and Communication



Define Objectives

01

Statistical Analysis



Data Collection

02

Exploratory Data Analysis



Data Cleaning and Data Processing

VISUALISATION DES DONNÉES

RENDRE LES RÉSULTATS COMPRÉHENSIBLES

TYPES DE VISUALISATIONS



GRAPHIQUES

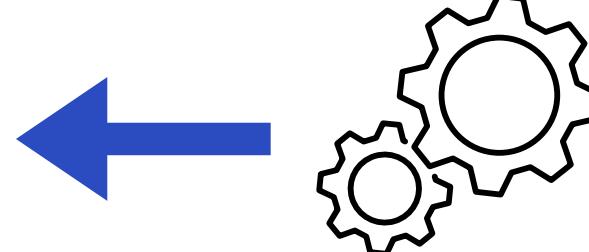


TABLEAUX DE
BORD



KPI

OUTILS



ADAPTÉ AU PUBLIC CIBLE
(MANAGER, CLIENT...)

+ a b | e a u®

DÉCISION

OBJECTIF : AGIR

Data-Driven Decision Making Examples

Inventory Management



A retail store uses sales data to predict stock levels and reorder products just in time to meet customer demand.

Marketing Campaigns



A company analyzes customer purchase history to personalize email marketing campaigns, increasing engagement and sales.

Customer Service



A telecom company uses call center data to identify common issues and improve customer service processes, reducing wait times and increasing satisfaction.

Product Development



A tech firm uses user feedback data to prioritize features in the next software update, ensuring alignment with customer needs.

Pricing Strategy



An e-commerce business leverages competitive pricing data to adjust prices in real-time, optimizing profit margins and staying competitive.



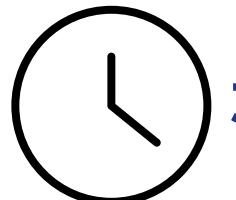
New Horizons[®]
An Edgewood Group Brand

EXERCICE

DÉcrire les étapes du cycle de vie des données
d'une application de livraison de repas

MÉTHODOLOGIE :

- 1.COLLECTE DES DONNÉES
- 2.STOCKAGE DES DONNÉES
- 3.NETTOYAGE DES DONNÉES
- 4.TRANSFORMATION DES DONNÉES
- 5.ANALYSE DES DONNÉES
- 6.VISUALISATION DES DONNÉES
- 7.DÉCISION



30 MIN

DATA QUALITY & REFINEMENT



- COMPRENDRE CE QU'EST LA QUALITÉ DES DONNÉES
- IDENTIFIER LES ERREURS COURANTES
- COMPRENDRE LE RÔLE DU DATA REFINEMENT



POURQUOI LA QUALITÉ DES DONNÉES EST UN ENJEU MAJEUR ?

- DONNÉES ISSUES DE **MULTIPLES SOURCES**
- **ERREURS HUMAINES** FRÉQUENTES
- DONNÉES **INCOMPLÈTES** OU **INCOHÉRENTES**
- **VOLUME** IMPORTANT

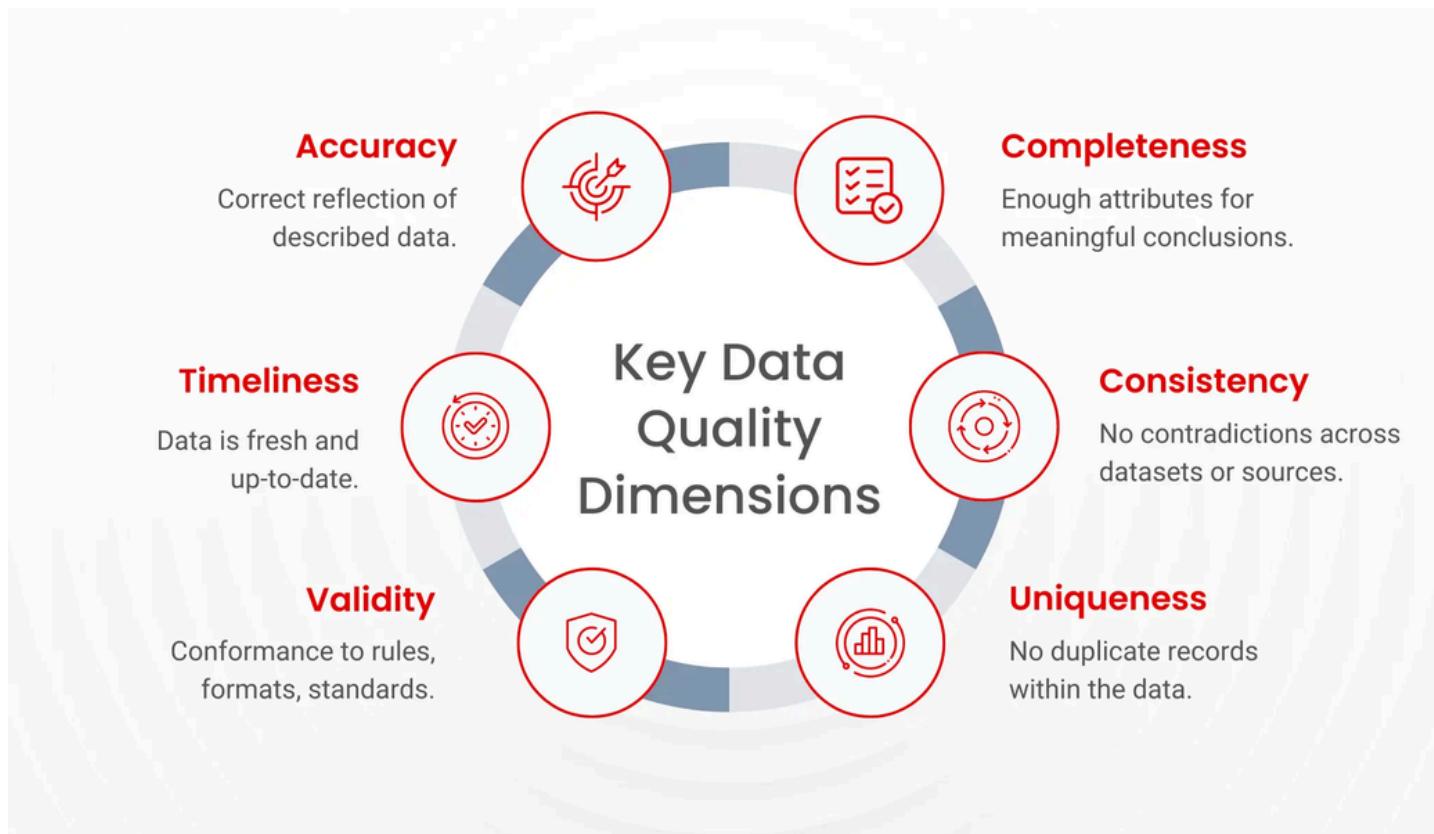


EXEMPLES DE PROBLÈMES COURANTS:

- CHAMPS VIDES
- VALEURS INCOHÉRENTES
- FORMATS DIFFÉRENTS
- DOUBLONS
- ERREURS DE SAISIE

QU'EST CE QUE LA DATA QUALITY ?

LA DATA QUALITY MESURE LA CAPACITÉ DES DONNÉES À ÊTRE UTILISÉES SANS RISQUE D'ERREUR POUR UN OBJECTIF DONNÉ.



QOS ET QOD

QOS (QUALITÉ DE SERVICE)

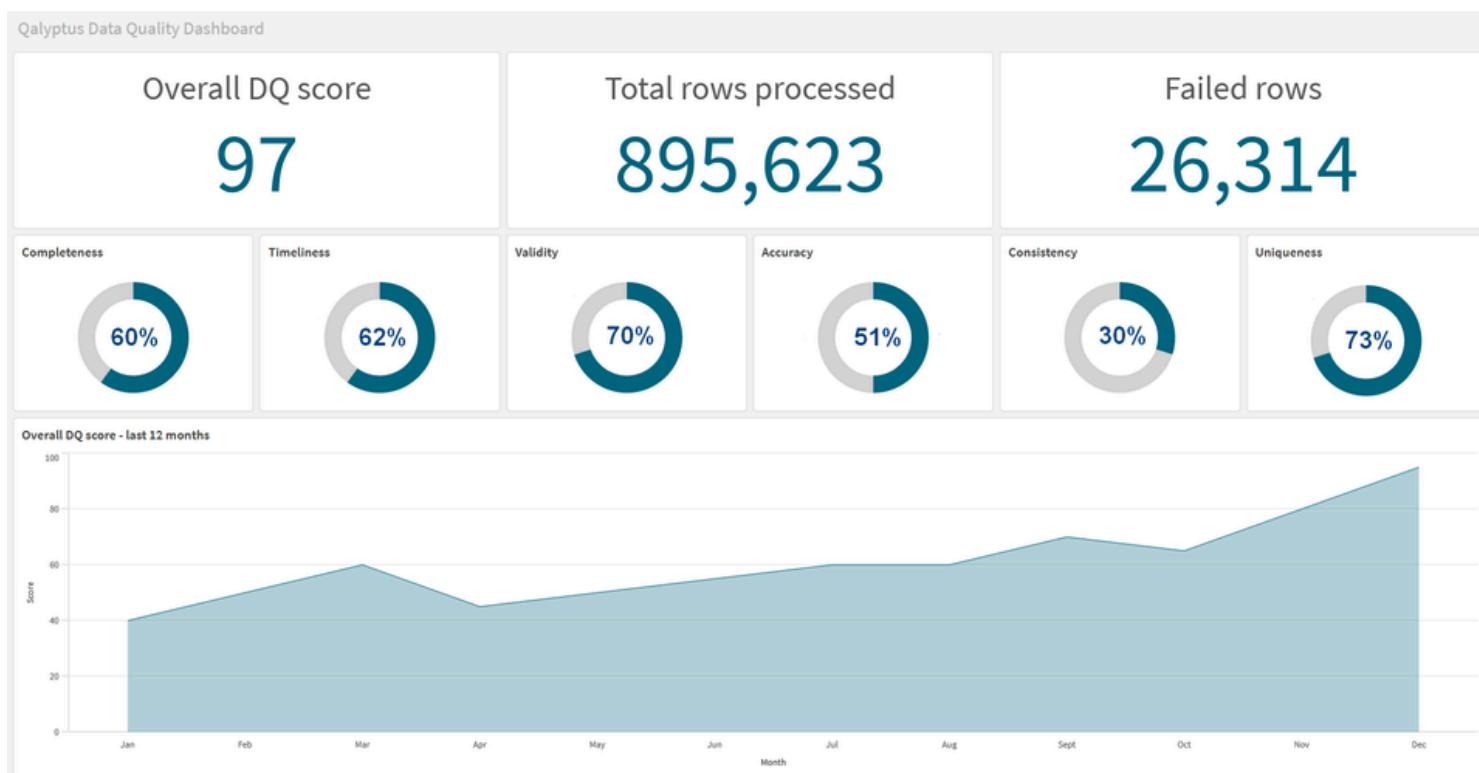
- **DISPONIBILITÉ**: LES DONNÉES SONT ACCESSIBLES QUAND ON EN A BESOIN
- **FIABILITÉ**: PAS DE PANNES, PAS DE PERTES DE DONNÉES
- **PERFORMANCE** : LES REQUÊTES ET TRAITEMENTS SONT RAPIDES
- **SCALABILITÉ** : SUPPORTE LES VOLUMES CROISSANTS
- **SÉCURITÉ** : ACCÈS CONTRÔLÉ, DONNÉES PROTÉGÉES
- **TRAÇABILITÉ** : CHAQUE TRANSFORMATION EST IDENTIFIABLE

QOD (QUALITÉ DE DONNÉES)

- **EXACTITUDE** : LES DONNÉES SONT JUSTES
- **COMPLÉTUDE** : TOUTES LES DONNÉES NÉCESSAIRES SONT PRÉSENTES
- **COHÉRENCE** : PAS DE CONTRADICTION ENTRE LES SOURCES
- **UNICITÉ** : PAS DE DOUBLONS
- **FRAÎCHEUR** : DONNÉES À JOUR
- **CONFORMITÉ** : RESPECTE LES FORMATS, RÈGLES MÉTIER, ETC.

IDENTIFIER LES PROBLÈMES DE QUALITÉ

- Explorer les données
- Statistiques
- Visualisation
- Règles métiers



GOUVERNANCE ET DATA QUALITY

LA DATA GOUVERNANCE EST L'ENSEMBLE DES **RÈGLES, PROCESSUS ET OUTILS** QUI PERMETTENT DE **GÉRER LES DONNÉES D'UNE ENTREPRISE** DE MANIÈRE FIABLE, SÉCURISÉE ET CONFORME.



DATA CATALOG

DOCUMENTER TOUTES LES DONNÉES DISPONIBLES

PERMET DE CONNAITRE :

- *LE CONTENU D'UNE TABLE*
- *SES UTILISATEURS*
- *SON BUT*

Capabilities of a Modern Data Catalog

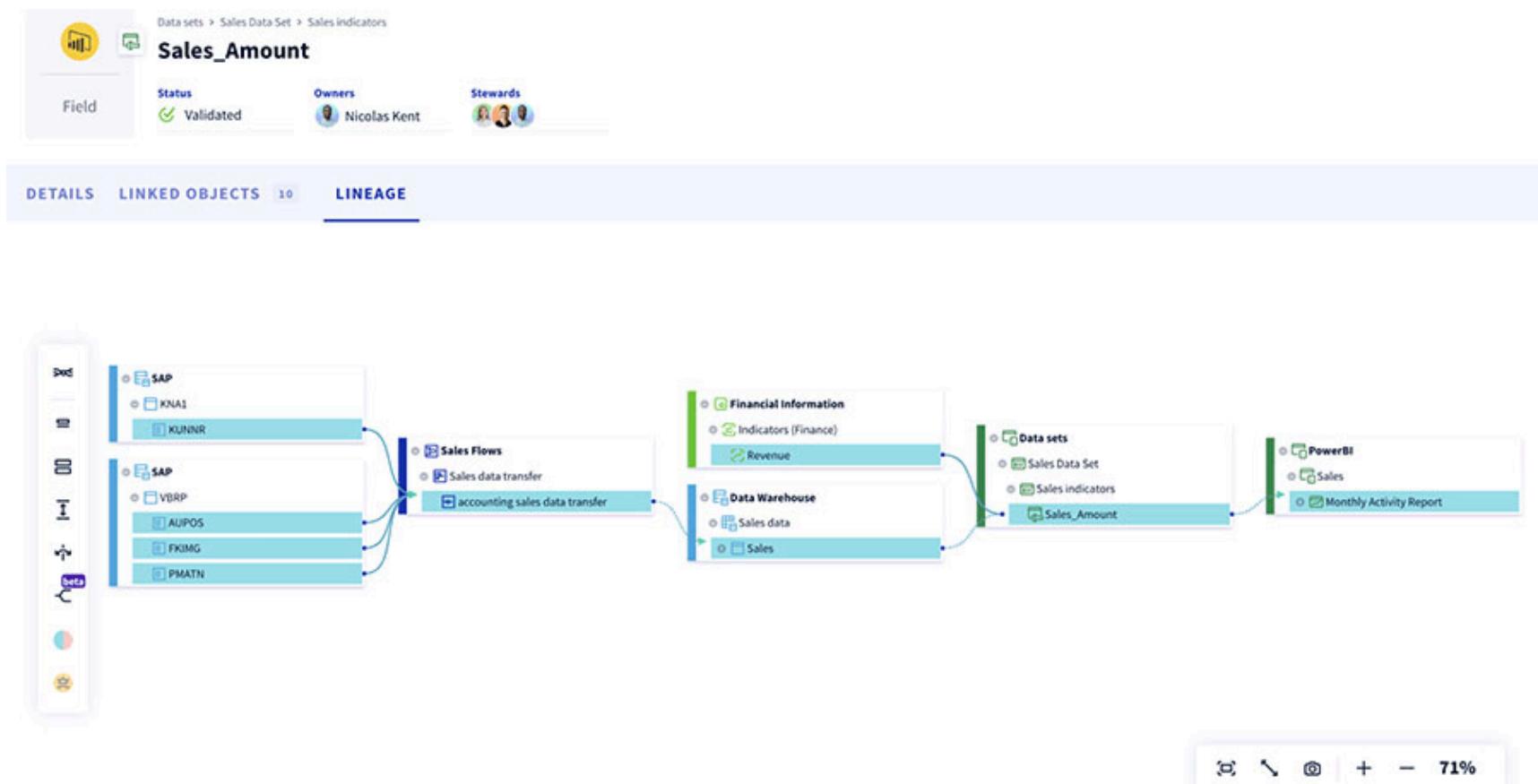
 CENTRAL DATA REPOSITORY Metadata context plane for your organization's data estate	 METADATA MANAGEMENT Bi-directional flow of metadata for active metadata management
 DATA DISCOVERY Google for your data to effortlessly locate the right asset for your project	 DATA LINEAGE Complete visibility from data source to BI, enriched with metadata context
 COLLABORATION A unified workspace for diverse teams to interact and collaborate on data	 DATA GOVERNANCE Federated governance to balance democratization, quality and compliance
 INTEGRATION CAPABILITIES Connect seamlessly with best-in-class tools across the data stack	 GENERATIVE AI CAPABILITIES Automated documentation, natural language queries, and more

atlan

DATA LINEAGE

SUIVRE LE PARCOURS D'UNE DONNÉE

- INTÉGRÉ À UNE SOLUTION DATACATALOG OU DATAWAREHOUSE
- UTILE POUR COMPRENDRE UN RÉSULTAT PAR EXEMPLE



SÉCURITÉ ET CONFORMITÉ

- RGPD (DONNÉES PERSONNELLES)
- ANONYMISATION (MASQUER)
- GÉRER LES ACCÈS (IAM)

Types Of Data Security



CONCLUSION - PARTIE 1



LA DONNÉE BRUTE N'EST PAS EXPOITABLE

- CONTIENT DES ERREURS, INCOHÉRENCES ET DES MANQUES
- DOIT ÊTRE TRAVAILLÉE AVANT ANALYSE OU DÉCISION

LA QUALITÉ DES DONNÉES EST ESSENTIELLE

- MAUVAISE QUALITÉ = MAUVAISES DÉCISIONS
- TOUJOURS EVALUER LA DATA QUALITY

LE DATA REFINEMENT REND LA DONNÉE UTILISABLE

- REGROUPE :
 - NETTOYAGE
 - TRANSFORMATION
 - VALIDATION
- PERMET DE PASSER DE DONNÉE BRUTE À DONNÉE EXPLOITABLE

LA DONNÉE SUIT UN CYCLE DE VIE

- DE LA COLLECTE À LA DÉCISION
- DATA REFINEMENT AU COEUR DU CYCLE

LA LOGIQUE MÉTIER EST INDISPENSABLE

- LES CHOIX DOIVENT ÊTRE COHÉRENTS AVEC LE CONTEXTE
- TOUJOURS JUSTIFIER SES DÉCISIONS

CONCLUSION - PARTIE 2



L'OUTIL EST SECONDAIRE

- MÉTHODE ET REFLEXION > TECHNOLOGIE

LA DONNÉE DOIT ÊTRE DOCUMENTÉE

- EXPLIQUER LES CHOIX RÉALISÉS
- FACILITER LA COMPRÉHENSION ET LA RÉUTILISATION

Savoir rendre une donnée fiable et exploitable est une compétence data fondamentale

PROJET FINAL

RÉALISER UN REFINEMENT COMPLET D'UN DATASET BRUT

L'OBJECTIF EST DE :

- APPLIQUER LES NOTIONS DE DATA QUALITY ET DATA REFINEMENT
- NETTOYER ET TRANSFORMER UN DATASET BRUT
- PRODUIRE UNE DONNÉE EXPLOITABLE ET COHÉRENTE
- JUSTIFIER LES CHOIX EFFECTUÉS

SOURCES DE DONNÉES :

- CHOIX DU DATASET SUR LA SLIDE SUIVANTE
- FORMAT : CSV
- DONNÉES STRUCTURÉES
- CONTIENT VOLONTAIREMENT :
 - VALEURS MANQUANTES
 - DOUBLONS
 - ERREURS DE FORMAT
 - INCOHÉRENCES

TÂCHES ATTENDUES :

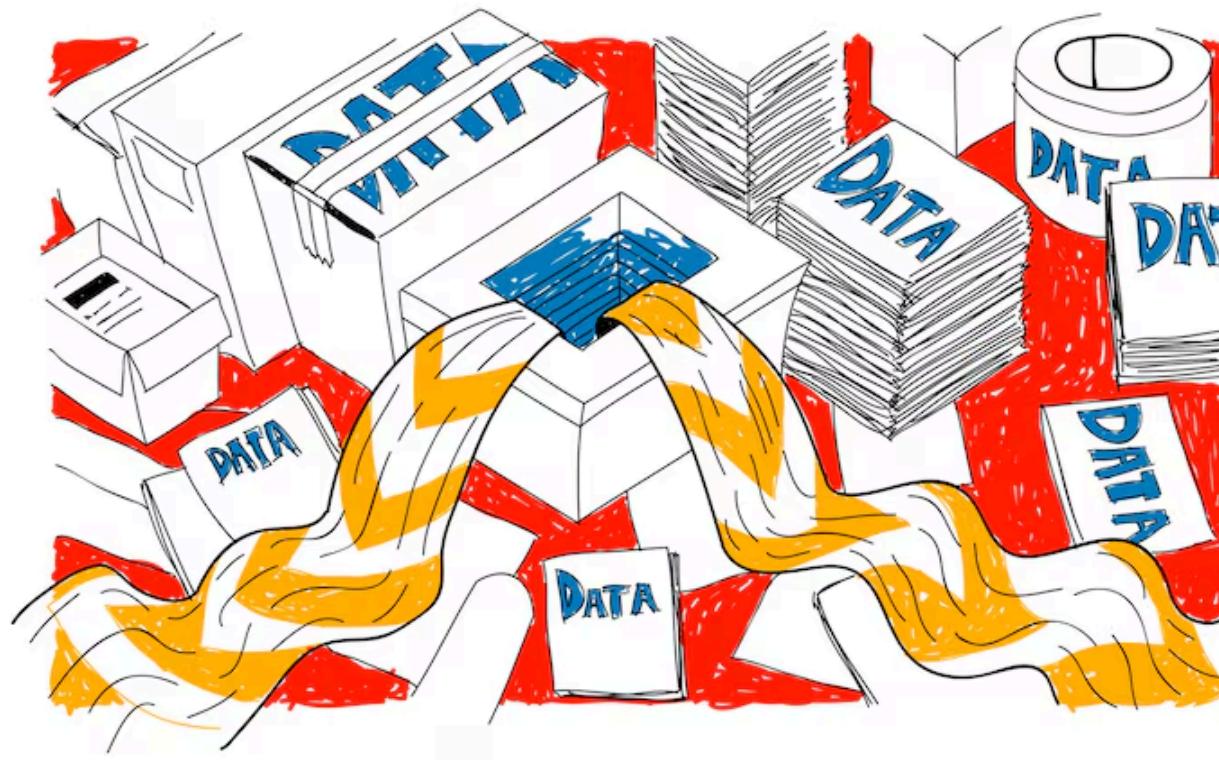
- 1.COMPRENDRE LE CONTEXTE ET LES DONNÉES
- 2.IDENTIFIER LES PROBLÈMES DE QUALITÉ
- 3.NETTOYER LES DONNÉES
- 4.TRANSFORMER LES DONNÉES
- 5.VÉRIFIER LA QUALITÉ FINALE

PROJET FINAL

RÉALISER UN REFINEMENT COMPLET D'UN DATASET BRUT

LISTE DES DATASETS :

- [HTTPS://WWW.KAGGLE.COM/DATASETS/AHMEDMOHAMED2003/CAFE-SALES-DIRTY-DATA-FOR-CLEANING-TRAINING](https://www.kaggle.com/datasets/ahmedmohamed2003/cafe-sales-dirty-data-for-cleaning-training)
- [HTTPS://WWW.KAGGLE.COM/DATASETS/RASHIKRAHMANPRITOM/DATA-SCIENCE-JOB-POSTING-ON-GLASSDOOR](https://www.kaggle.com/datasets/rashikrahmanpritom/data-science-job-posting-on-glassdoor)
- [HTTPS://GITHUB.COM/EYOWHITE/MESSY-DATASET](https://github.com/EYOWHITE/messy-dataset)



PROJET FINAL

RÉALISER UN REFINEMENT COMPLET D'UN DATASET BRUT

DATA-REFINEMENT-PROJECT/

```
|-  
|   |- DATA/  
|   |   |- RAW/  
|   |   |- PROCESSED/  
  
|-  
|   |- NOTEBOOKS/  
|   |   |- 01_EXPLORATION  
|   |   |- 02_CLEANING  
|   |   |- 03_TRANSFORMATION  
  
|-  
|   |- REPORTS/  
|   |   |- RAPPORT.PDF  
  
|-  
|   |- README.MD  
|   |- REQUIREMENTS.TXT
```

PROJET FINAL

RÉALISER UN REFINEMENT COMPLET D'UN DATASET BRUT

ÉLÉMENTS À RENDRE:

- **CODE SOURCE** SUIVANT L'ORGANISATION DÉTAILLÉE DANS LA SLIDE PRÉCÉDENTE VIA LIEN GITHUB À REMPLIR DANS LE CSV
- **DATASET FINAL NETTOYÉ** (CSV)
- **RAPPORT** (1-2 PAGES)

IMPORTANT :

- LES **CHOIX** DOIVENT ÊTRE **JUSTIFIÉS**
- LE **RAISONNEMENT** ET LA **LOGIQUE** DOIVENT ÊTRE MIS EN AVANT
- LA **LOGIQUE** EST PLUS IMPORTANTE QUE LA **PERFECTION TECHNIQUE**



<https://shorturl.at/fSmOt>

DATE/HEURE MAX DE RENDU : MERCREDI 08/01/2026 À 23H59