

FORMATION INTRA — EDF

Data Management, Gouvernance & Intelligence Artificielle

Diffuser la culture Data et les standards du Groupe — de la stratégie à l'IA

2 jours · 14 heures

29-30 juin 2026 · EDF, 140 av. Viton, Marseille

Animé par GRO — Consultant expert Data Gouvernance, certifié DAMA DMBOK (CDMP)

Sparks Roca

Votre parcours sur 2 journées

JOUR 1 — Stratégie, gouvernance & organisation

- M1 · Introduction & enjeux
- M2 · Processus de valorisation
- M3 · Matrice de maturité
- M4 · Data Mesh & domaines
- M5 · Rôles & operating model
- M6 · Gouvernance Data & IA
- M7 · Modélisation & objets métier
- M8 · Données de référence & RGPD
- M9 · Catalogage des données
- **Atelier fil rouge + Quiz J1**

JOUR 2 — Qualité, architecture & IA

- M10 · Référentiels (Master/Reference)
- M11 · Qualité des données
- M12 · Outillage Data Management
- M13 · Architecture & plateforme
- M14 · IA & Data Science
- M15 · Outillage IA & MLOps
- M16 · Communautés Data/IA
- M17 · Organisation agile
- **Atelier Data-to-IA + Quiz J2**

GRO — 25 ans au service de la donnée

Certifications

DAMA DMBOK V2 (CDMP) · Collibra Data Steward. La formation s'appuie sur le référentiel international DAMA-DMBOK.

Double compétence

Business & IT : facilite la relation métiers / IT et la cohérence besoins ↔ solutions.

Outils maîtrisés

Collibra, Informatica, Atlan · MDM Semarchy · BI Power BI, Qlik, Spotfire, Snowflake.

Postures

Data Governance Officer, Data Quality lead, Product Owner agile, acculturation des métiers.

Quelques missions récentes

Safran Cabin — Data Office & gouvernance. **CMA CGM** — domaine Assets / IMO, Data Quality Excellence. **Airbus Helicopters** — data catalog & communauté BI. **Manutan** — stratégie data, 26 filiales.

Faisons connaissance

1 Qui êtes-vous ?

Prénom, direction ou métier, votre rôle au quotidien.

2 Votre rapport à la donnée

Quelles données manipulez-vous ? Pour quels usages ?

3 Un irritant

Un problème de données récurrent qui vous fait perdre du temps.

4 Une attente

Ce que vous aimeriez repartir avec à l'issue des 2 jours.

On y reviendra

Vos attentes sont notées au paperboard. Nous les reprendrons en synthèse, à la fin du Jour 2, pour vérifier qu'elles ont trouvé une réponse.

Ce que vous saurez faire à l'issue des 2 jours

1

Maîtriser les concepts

Donnée, information, connaissance ; Data Management vs gouvernance.

2

Cadrer la gouvernance

Rôles, instances et responsabilités (Owner, Steward, CDO...).

3

Évaluer & redresser la qualité

Mesurer un jeu de données et bâtir une démarche Data Quality.

4

Référentiels & métadonnées

Master Data, données de référence, catalogue et lignage.

5

Choisir les bons outils

Catalogue, MDM, DQ, plateformes data & IA.

6

Sécuriser & rester conforme

RGPD, éthique et sécurité ; ouvrir sur l'IA.

Comment nous allons travailler

Théorie ancrée

Chaque concept est illustré par un exemple concret et un retour d'expérience terrain.

Participatif

Questions, échanges et partage de vos situations : la formation se nourrit de vos cas.

Ateliers

Deux ateliers fil rouge (un par jour) en sous-groupes, sur des cas EDF.

Validation

Un quiz par jour pour ancrer les acquis, corrigé ensemble.

Le fil rouge

Deux idées guideront ces deux jours : « la valeur d'abord » (la gouvernance sert l'usage) et « pas d'IA fiable sans données gouvernées ». Gardez-les en tête.

JOUR 1

Stratégie, gouvernance & organisation Data

Le cadre stratégique, l'organisation et la maîtrise du patrimoine de données

MODULE 1 · Jour 1

Introduction & enjeux

Durée indicative : 15 min

Au programme de ce module :

- La donnée, un actif stratégique
- Donnée, information, connaissance
- Pourquoi la gouvernance

La donnée, un actif stratégique

« Data is an asset »

La donnée possède les attributs d'un actif : une valeur, un coût, un cycle de vie et un propriétaire. Elle se gère comme un capital de l'entreprise, pas comme un sous-produit de l'informatique.

- Valeur : elle éclaire des décisions et crée des services
- Coût : la produire, la stocker, la maintenir a un prix
- Risque : mal gérée, elle expose (erreurs, non-conformité)
- Propriété : quelqu'un doit en répondre

Exemple

Chez CMA CGM, traiter les données « Assets » (navires, ports) comme un actif a fiabilisé le reporting réglementaire IMO sur les émissions de CO₂.

Donnée, information, connaissance

1

Donnée

Fait brut, sans contexte. Ex : « 42 », « 78 ».

2

Information

Donnée mise en contexte. Ex : « 78 °C à 14h02 ».

3

Connaissance

Information interprétée. Ex : « seuil dépassé → risque ».

4

Décision / valeur

L'action éclairée qui crée de la valeur.

Exemple — le compteur communicant

« 78 » (donnée brute) devient « 78 kWh sur le site X à 14 h » (information), interprété en « consommation anormale » (connaissance), qui déclenche une vérification (décision). Une erreur sur un seul maillon fausse la décision finale.

Pourquoi la gouvernance des données ?

Sans gouvernance

- Données en silos, définitions divergentes
- Chiffres incohérents entre directions
- Temps perdu à recouper et corriger
- Risque de non-conformité (RGPD)
- Aucune base fiable pour l'IA

Avec gouvernance

- Données fiables, partagées, réutilisables
- Un langage commun entre métiers et IT
- Décisions fondées sur des faits
- Conformité et sécurité maîtrisées
- Un socle prêt pour l'analytique et l'IA

Processus de valorisation de la donnée

Durée indicative : 30 min

Au programme de ce module :

- Le cycle de vie de la donnée
- Focus collecte, stockage, usage
- La chaîne de valeur « value first »
- Le coût de la non-valorisation

Les 5 étapes du cycle de vie



Exemple — un relevé de consommation

Créer (le compteur émet un relevé) → Stocker (plateforme sécurisée) → Utiliser (facturation, pilotage réseau) → Maintenir (corriger un relevé aberrant) → Archiver puis purger (durée légale, RGPD). La gouvernance s'applique à chaque étape.

Étape 1 — Créer & collecter

La qualité naît à la source

C'est au moment de la création ou de la collecte que se joue l'essentiel de la qualité. Une donnée mal saisie ou mal captée se propage ensuite partout.

- Définir la source de référence (qui fait foi ?)
- Contrôler à la saisie (formats, valeurs autorisées)
- Documenter l'origine dès la collecte
- Éviter les ressaisies multiples, sources d'écarts

Exemple

Un capteur mal calibré ou un champ libre non contrôlé suffit à fausser tous les rapports en aval.

Stocker, utiliser, partager

Stocker (étape 2)

- Où ? base, datalake, cloud, on-premise
- Sécuriser : accès, chiffrement, sauvegarde
- Organiser pour retrouver (catalogue, M9)
- Maîtriser les coûts de stockage

Utiliser & partager (étape 3)

- Reporting, analyses, modèles d'IA
- Échange entre directions (interopérabilité)
- Exposer via data products, API
- Tracer les usages (qui consomme quoi)

Maintenir, archiver, détruire

4 Maintenir

Mettre à jour, corriger, contrôler la qualité dans le temps. Une donnée non maintenue se périmé.

5 Archiver

Conserver pour des raisons légales ou historiques, dans un stockage adapté et sécurisé.

Détruire

Supprimer quand la donnée n'est plus utile ni obligatoire : c'est une exigence RGPD (durée de conservation).

La rétention

Chaque type de donnée a une durée de conservation définie : ni trop courte (perte), ni trop longue (risque).

De la donnée à l'usage : « value first »

1

Capter le besoin

Partir d'un cas d'usage métier concret, pas de la donnée.

2

Sourcer & qualifier

Données nécessaires, sources, fiabilité.

3

Exposer & valoriser

Data products, dashboards, modèles IA.

4

Mesurer l'impact

Suivre l'usage et la valeur créée.

Principe directeur

On priorise les cas d'usage à forte valeur et faisabilité raisonnable. La gouvernance n'est pas une fin en soi : elle sert l'usage. C'est le fil conducteur de l'atelier de fin de journée.

Ce que coûte une donnée mal valorisée

≈ 30 %

du temps des équipes data passé à chercher, nettoyer et recouper des données peu fiables.

1 → 10 → 100

coût de la non-qualité : prévenir 1, corriger 10, subir l'erreur 100.

× IA

une IA n'amplifie de la valeur que si la donnée en amont est gouvernée et fiable.

Côté risque

- Décisions sur de mauvais chiffres
- Doublons, non-conformité, perte de confiance

Côté valeur

- Données réutilisables, time-to-market réduit
- Self-service, terrain prêt pour l'IA

Matrice de maturité Data

Durée indicative : 1 h

Au programme de ce module :

- Pourquoi mesurer la maturité
- Les 5 niveaux
- Les 6 axes d'évaluation
- Comment scorer
- Atelier d'auto-évaluation

Pourquoi mesurer sa maturité Data ?

1 Objectiver

Sortir des impressions : où en est-on vraiment, axe par axe ?

2 Langage commun

Métiers et IT partagent la même grille et les mêmes mots.

3 Révéler les écarts

Identifier les points faibles et les écarts entre directions.

4 Fixer une cible

Définir le niveau utile à atteindre, pas le niveau maximal partout.

Le piège à éviter

Viser le niveau 5 partout est coûteux et contre-productif. La maturité cible dépend des enjeux : un domaine critique mérite un niveau élevé, un domaine secondaire peut rester plus bas.

Les 5 niveaux de maturité



Lecture

On progresse rarement de façon homogène : on peut être « Piloté » sur la qualité et « Initial » sur le catalogue. La matrice photographie ce profil pour prioriser.

Du niveau 1 au niveau 5 : ce qui change

Niveaux 1-2 (réactif)

- On découvre les problèmes quand ils explosent
- Chacun gère « ses » données dans son coin
- La qualité dépend des personnes, pas des processus
- Peu ou pas d'outils, beaucoup d'Excel

Niveaux 4-5 (pilote)

- On anticipe : indicateurs, alertes, prévention
- Rôles, instances et règles partagés
- La qualité est mesurée et pilotée
- Outils en place : catalogue, MDM, dashboards

Les 6 axes d'évaluation

Stratégie & sponsoring

Vision data, soutien de la direction, feuille de route et moyens.

Organisation & rôles

Rôles tenus (Owner, Steward...), instances de gouvernance.

Qualité des données

Règles, mesure, remédiation et contrôle dans le temps.

Métadonnées & catalogue

Documentation, glossaire, catalogue, lignage.

Architecture & outils

Socle technique, intégration, outils de gouvernance.

Culture & usages

Acculturation, adoption, communautés, partage.

Comment attribuer une note

Noter chaque axe de 1 à 5

Pour chaque axe, on choisit le niveau qui décrit le mieux la situation réelle, avec une justification courte. On note ce qui est, pas ce qu'on voudrait.

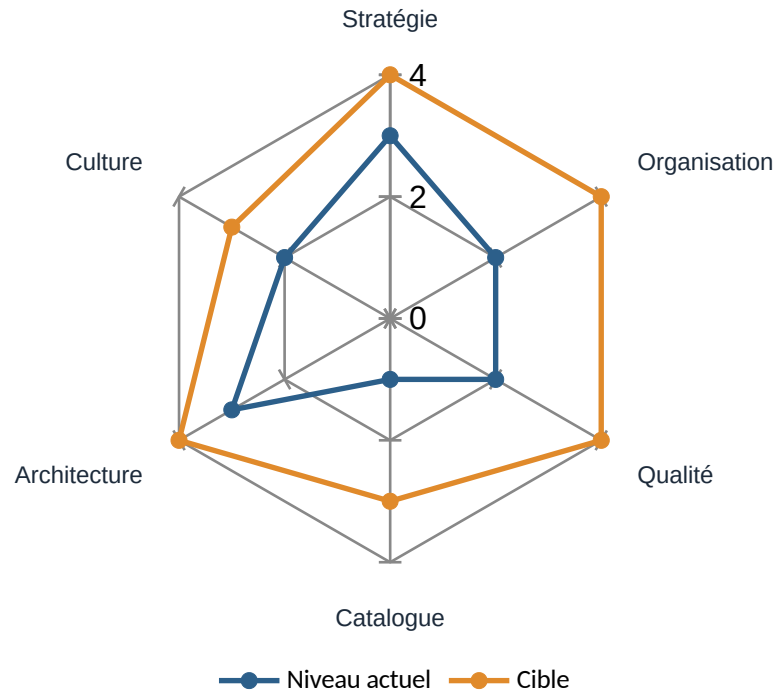
- Une note par axe, pas une note globale
- Une justification factuelle en une ligne
- On reporte les 6 notes sur un radar
- On compare les perceptions entre participants

Exemple

« Qualité : 2/5 — des contrôles existent sur le CRM mais rien n'est mesuré ni suivi ailleurs. »

Visualiser le profil de maturité

Le radar de maturité



Lire le radar

- L'aire bleue = la situation actuelle
- L'aire orange = la cible visée à 12-18 mois
- L'écart entre les deux = le plan de progrès
- Les axes les plus « creusés » sont prioritaires
- Exemple ci-contre : le catalogue est le point faible

Atelier : auto-évaluez votre maturité

Étape 1 — Noter (20 min)

Chacun note son périmètre de 1 à 5 sur les 6 axes, avec justification.

Étape 2 — Visualiser

On reporte les notes sur le radar : le profil apparaît.

Étape 3 — Comparer (15 min)

Restitution croisée : les écarts de perception sont riches d'enseignements.

Étape 4 — Agir

2-3 actions prioritaires par axe critique, vers le plan d'action.



Retour d'expérience — Démarche utilisée chez Manutan : workshops d'état des lieux filiale par filiale, avant de bâtir la stratégie pour le COMEX.

Data Mesh & domaines de données

Durée indicative : 30 min

Au programme de ce module :

- Le modèle centralisé et ses limites
- Les 4 principes du Data Mesh
- Le domaine de données
- Le data product
- La gouvernance fédérée

Les limites du modèle centralisé

Le goulot d'étranglement

Dans un modèle centralisé, une seule équipe data traite toutes les demandes de toutes les directions. Ce modèle atteint vite ses limites quand l'entreprise grandit.

- L'équipe centrale devient un goulot d'étranglement
- Elle est éloignée du sens métier de la donnée
- Les délais s'allongent, la frustration monte
- Le modèle ne passe pas à l'échelle

Exemple

Plus l'entreprise a de domaines et de données, plus une équipe centrale unique sature.

Centralisé vs organisation par domaines

Approche centralisée

- Une équipe centrale traite tout
- Cohérence forte, mais goulot d'étranglement
- Éloignée du métier
- Difficile à passer à l'échelle

Organisation par domaines (Data Mesh)

- La donnée appartient au domaine qui la connaît
- Donnée traitée comme un produit
- Plateforme & règles communes en self-service
- Gouvernance fédérée

Les 4 principes du Data Mesh

1 Propriété par domaine

Chaque domaine métier est responsable de ses données, de bout en bout.

2 Donnée comme produit

Conçue pour être consommée : documentée, fiable, avec un niveau de service.

3 Plateforme self-service

Un socle commun pour produire et exposer sans tout réinventer.

4 Gouvernance fédérée

Standards communs au Groupe, appliqués localement par chaque domaine.

Qu'est-ce qu'un domaine de données ?

Un périmètre métier responsable

Un domaine regroupe les données liées à une même activité métier, sous la responsabilité d'une équipe qui les connaît et en répond.

- Ex : Clients, Production, Maintenance, RH, Finance
- Un responsable identifié pour le domaine
- Porte définitions, qualité, exposition, accès
- Frontières claires avec les autres domaines

Exemple

Chez CMA CGM, j'étais responsable de la gouvernance d'un domaine (Lignes commerciales, puis Assets) : un découpage très concret.

La donnée « comme un produit »

Le data product

On ne livre plus une table brute, mais un « produit » de données pensé pour ses consommateurs : fiable, documenté, avec un niveau de service annoncé.

- Donnée + métadonnées + qualité + règles d'accès
- Documenté et facile à trouver (catalogue)
- Réutilisable par les autres domaines
- Un niveau de service : fraîcheur, disponibilité

Exemple

Un « produit » Clients certifié, que toutes les directions consomment au lieu de recréer leur propre version.

Tenir ensemble cohérence et autonomie

Ce qui est commun (fédéré)

- Standards de qualité et de sécurité
- Règles de nommage et définitions partagées
- Plateforme et outils communs
- Conformité (RGPD) au niveau Groupe

Ce qui est local (domaine)

- La connaissance métier de la donnée
- La production et la qualité au quotidien
- Les décisions d'accès à ses données
- L'exposition des data products

Rôles & fonctions — l'operating model

Durée indicative : 1 h

Au programme de ce module :

- Panorama des rôles
- CDO, Owner, Steward, Custodian
- Rôles dérivés
- Le stewardship au quotidien
- Le RACI métier ↔ IT
- Les 3 modèles d'operating

Qui fait quoi dans la gouvernance ?

CDO porte la stratégie data et sponsorise la filière au niveau Groupe.

Data Owner responsable métier d'un périmètre : définitions, règles, qualité, accès.

Data Steward pilote opérationnel : documente, contrôle, traite la qualité au quotidien.

Data Custodian garant technique (IT) : stockage, sécurité, sauvegarde, mise en œuvre.

Data User consommateur de la donnée, qui respecte les règles d'usage.

Exemple — le domaine Maintenance

L'Owner (responsable Maintenance) définit « équipement » et les accès ; le Steward documente et contrôle les fiches ; le Custodian (IT) héberge la GMAO ; le CDO arbitre avec la Production. À ne pas confondre : Owner (décide) ≠ Custodian (exécute).

Le Chief Data Officer

CDO — sponsor & stratège

Le CDO porte la vision data au plus haut niveau. Il fait de la donnée un sujet de direction, arbitre les priorités et garantit la valeur et la conformité à l'échelle du Groupe.

- Définit la stratégie et la feuille de route data
- Sponsorise la filière, débloque les moyens
- Arbitre les conflits de priorités
- Rend compte de la valeur créée à la direction

Exemple

Sur mes missions Safran et Manutan, l'accompagnement du CDO et la formalisation de la stratégie pour le COMEX étaient au cœur du dispositif.

Le Data Owner (métier)

Owner — il décide

Le Data Owner est un responsable métier. Il définit le sens et les règles de « ses » données et en répond. Ce n'est pas un rôle informatique.

- Définit les données et leurs règles métier
- Fixe le niveau de qualité attendu
- Décide qui peut accéder à ses données
- Est redevable de la valeur et des risques

Exemple

Le responsable du domaine « Production » est Owner des données de production : il dit ce qu'est une « unité produite » et qui peut la voir.

Le Data Steward

Steward — il opère

Le Data Steward est le bras opérationnel de la gouvernance au quotidien. Il met en musique les règles définies par l'Owner et veille à la qualité.

- Documente la donnée (glossaire, catalogue)
- Contrôle la qualité et traite les anomalies
- Répond aux questions sur « sa » donnée
- Anime la communauté de son périmètre

Exemple

C'est le rôle que j'ai le plus accompagné en mission : acculturer et outiller les Stewards fait toute la différence.

Le Data Custodian (IT)

Custodian — il met en œuvre

Le Data Custodian est le garant technique. Il héberge, sécurise et rend disponible la donnée, en appliquant les règles décidées par le métier.

- Stockage, performance, sauvegarde
- Sécurité technique : accès, chiffrement
- Met en œuvre les règles définies par l'Owner
- N'décide pas du sens métier de la donnée

Exemple

Quand l'Owner dit « cette donnée est confidentielle », le Custodian implémente les habilitations et le chiffrement.

Les autres rôles utiles

Data Architect

Conçoit la structure et les flux de données à l'échelle de l'entreprise (voir M13).

Data Quality Analyst

Spécialiste de la mesure et de la remédiation de la qualité (voir M11).

Data Protection Officer (DPO)

Garant de la conformité RGPD et du respect des droits des personnes (voir M8).

Data Scientist / Analyst

Exploite la donnée pour produire analyses et modèles (voir M14).

Le Data Steward au quotidien

1

Documenter

Glossaire, catalogue : définitions, règles, propriétaires.

2

Contrôler

Suivre les indicateurs qualité, détecter les anomalies.

3

Arbitrer

Traiter les demandes d'accès, répondre aux questions.

4

Animer

Lier producteurs et consommateurs, diffuser les pratiques.

Le bon profil (DAMA)

Un Steward efficace est avant tout un expert reconnu de son domaine ET un bon communicant. La maîtrise des outils vient ensuite : ce n'est pas un rôle « informatique ».

Faire travailler métiers et IT ensemble

1

Définir

Le métier (Owner) définit sens, règles, qualité attendue.

2

Produire & exposer

L'IT (Custodian) met en œuvre, stocke, sécurise, expose.

3

Contrôler

Le Steward mesure, alerte, suit la remédiation.

4

Arbitrer

Les instances tranchent priorités et conflits (M6).

Clé de réussite

La frontière n'est pas « métier OU IT » mais une collaboration outillée. Un RACI clair par type de donnée supprime les zones grises où « personne n'est responsable ».

Les 3 modèles d'organisation

Centralisé

Une équipe pilote tout. Cohérence maximale, mais moins agile et éloigné du terrain.

Décentralisé

Chaque entité gère ses données. Proche du métier, mais hétérogénéité et doublons.

Fédéré (recommandé)

Standards communs + exécution locale. Le meilleur compromis, vers lequel s'oriente EDF.

Pourquoi le fédéré

Il combine la cohérence du centralisé et l'agilité du décentralisé. C'est le modèle naturel du Data Mesh : standards Groupe + autonomie des domaines.

Gouvernance de la filière Data & IA

Durée indicative : 30 min

Au programme de ce module :

- Qu'est-ce que gouverner
- Le Data Office
- Le comité & les communautés
- Les circuits de décision
- Politiques & conformité by design

Gouverner, c'est quoi exactement ?

Gouverner la donnée

La gouvernance définit qui décide quoi, comment et à quel niveau, à propos des données. C'est le système de pilotage qui rend la donnée fiable et utilisable par tous.

- Des rôles et des responsabilités clairs
- Des instances qui décident et arbitrent
- Des règles (politiques) connues de tous
- Des indicateurs pour piloter

Exemple

Sans gouvernance, chaque projet réinvente ses règles : la cohérence d'ensemble devient impossible.

Le Data Office : la cheville ouvrière

Le Data Office anime

Le Data Office est l'équipe qui fait vivre la filière data. Il ne décide pas seul : il outille, accompagne et coordonne les domaines.

- Définit méthodes, standards et outils
- Accompagne les domaines et les Stewards
- Maintient le catalogue et la documentation
- Rend compte au comité et au COMEX

Exemple

Au sein des Data Offices Safran et Manutan, j'ai contribué à structurer méthodes, standards et accompagnement des métiers.

Comité de gouvernance & communautés

Le comité de gouvernance

- Instance de décision, sponsorisée par le CDO
- Fixe priorités, politiques, moyens
- Arbitre les conflits entre domaines
- Se réunit à intervalle régulier

Les communautés / groupes de travail

- Stewards et experts par domaine ou sujet
- Partage des pratiques et entraide
- Résolution collective de problèmes
- Diffusion de la culture data

Gouverner aussi l'IA

Le comité IA / éthique

À côté de la gouvernance des données, la gouvernance de l'IA encadre les usages des modèles : conformité, éthique, risques et validation des cas sensibles.

- Valider les cas d'usage IA sensibles
- Vérifier conformité (RGPD) et éthique
- Évaluer les risques (biais, sécurité)
- S'appuie sur des données déjà gouvernées

Exemple

L'IA générative impose des règles claires : que peut-on envoyer à un modèle ? comment vérifier ses réponses ?

Les circuits de décision

1

Remonter

Un besoin, un problème qualité ou un conflit est remonté par un domaine.

2

Instruire

Le Data Office analyse, documente, propose des options.

3

Décider

Le comité arbitre : priorité, règle, moyens.

4

Décliner

La décision devient politique / standard, appliquée localement.

Exemple — un arbitrage en comité

Deux directions calculent différemment le « taux de disponibilité » d'une installation. Le Data Office instruit et compare ; le comité tranche la définition de référence ; elle devient un standard inscrit au glossaire. Fini les deux chiffres contradictoires au COMEX.

Politiques & conformité « by design »

Politique de gouvernance

Document de référence : principes, rôles, règles d'accès, qualité, sécurité, conformité.

Politiques par domaine

Qualité, métadonnées, sécurité, archivage... chaque sujet a ses règles déclinées.

« By design »

Conformité et qualité intégrées dès la conception, pas ajoutées après coup.

Mesurable

Des indicateurs et un suivi : une politique non mesurée n'est pas appliquée.

Modélisation des données & objets métier

Durée indicative : 1 h

Au programme de ce module :

- Le problème du langage
- L'objet métier & le glossaire
- Les 3 niveaux de modèle
- Focus conceptuel / logique / physique
- Relationnel vs dimensionnel
- La démarche de modélisation

Quand chacun parle sa langue

Le même mot, des sens différents

« Client », « Site », « Contrat » n'ont pas la même définition d'une direction à l'autre. Résultat : des chiffres incohérents et des réunions à se mettre d'accord sur les mots.

- Le « nombre de clients » diffère selon qui compte
- Les échanges de données deviennent risqués
- On perd du temps à réconcilier
- La confiance dans les chiffres s'érode

Exemple

Un « client actif » : depuis 3 mois ? 12 mois ? avec une commande ? Chaque service a sa réponse.

L'objet métier : une définition partagée

L'objet métier

Un objet métier est une définition partagée d'un concept de l'entreprise, avec ses attributs et ses règles. C'est la brique d'un langage commun.

- Un concept : Client, Compteur, Site, Contrat...
- Une définition unique qui fait foi
- Des attributs clés et des règles de gestion
- Un propriétaire (Data Owner)

Exemple

Chez Manutan, l'accompagnement portait sur un data modèle commun aux 26 filiales dans la suite Informatica.

Le glossaire métier

Le glossaire fait foi

Le glossaire métier est le dictionnaire partagé des concepts de l'entreprise. Il fige les définitions et sert de référence à tous, métiers comme IT.

- Une entrée par concept, validée par l'Owner
- Le « contrat de sens » entre directions
- La base du catalogue de données (M9)
- Vivant : maintenu par les Stewards

Exemple

Une fois « client actif » défini au glossaire, tous les rapports utilisent la même règle.

Conceptuel, logique, physique

1

Modèle conceptuel

Les concepts métier et leurs relations, sans technique. Langage des métiers.

2

Modèle logique

Entités, attributs, clés, cardinalités. Indépendant de la techno.

3

Modèle physique

Tables, colonnes, index, types. Dépend de l'outil et de la base.

Sens de lecture

On va du sens métier (conceptuel) vers la technique (physique). Les trois niveaux doivent rester cohérents entre eux.

Le modèle conceptuel

Le « quoi » métier

Le modèle conceptuel décrit les grands concepts et leurs relations, en langage métier, sans aucune considération technique. C'est le niveau qui se discute avec les opérationnels.

- Entités : Client, Commande, Produit...
- Relations : « un Client passe des Commandes »
- Compréhensible par un non-informaticien
- Sert à valider la compréhension commune

Exemple

Un schéma simple où l'on voit Client → Commande → Produit, sans aucune table ni colonne.

Du logique au physique

Modèle logique

- Entités détaillées avec attributs et clés
- Cardinalités (1-N, N-N)
- Normalisation pour éviter la redondance
- Indépendant de la base de données

Modèle physique

- Tables et colonnes réelles
- Types de données, index, contraintes
- Optimisé pour une base précise
- Réalisé et maintenu par l'IT

Relationnel vs dimensionnel

Modèle relationnel

- Tables liées par des clés, normalisées
- Évite la redondance, garantit la cohérence
- La norme pour les applications & le transactionnel
- Ex. base d'un ERP, d'un CRM

Modèle dimensionnel

- Tables de « faits » + « dimensions » (étoile)
- Optimisé pour l'analyse et le reporting
- La norme pour la BI et le Data Warehouse (M13)
- Ex. ventes par produit / région / temps

À retenir

On ne choisit pas l'un « contre » l'autre : le relationnel sert la production, le dimensionnel sert l'analyse. Les deux coexistent dans le SI.

Comment se déroule une modélisation



Exemple — l'objet métier « Site »

« Site » = une adresse de livraison aux Achats, une unité de production à la Production, un lieu de travail aux RH. La solution : un objet métier « Site » défini, arbitré et porté par un Owner, inscrit au glossaire commun.

Données de référence & données contraintes

Durée indicative : 1 h

Au programme de ce module :

- Donnée personnelle
- Les 6 principes RGPD
- Les droits des personnes
- Sécurité : le triangle CIA
- Vulnérabilité, menace, risque
- Protéger : chiffrement, anonymisation
- Classification & données contraintes

Qu'est-ce qu'une donnée personnelle ?

Donnée à caractère personnel

Toute information se rapportant à une personne physique identifiée ou identifiable, directement ou indirectement. La notion est large.

- Directe : nom, prénom, e-mail, photo
- Indirecte : identifiant, IP, n° de badge
- Sensible : santé, opinions, biométrie (régime renforcé)
- Même croisées, des données anodines peuvent ré-identifier

Exemple

Un identifiant technicien + un site + un horaire peut suffire à identifier une personne : c'est de la donnée personnelle.

Les 6 principes du RGPD

Licéité & loyauté

Une base légale (dont le consentement) et un traitement transparent.

Finalité

Une finalité déterminée, explicite et légitime ; pas de détournement.

Minimisation

Collecter le strict nécessaire à la finalité, rien de plus.

Exactitude

Des données exactes et tenues à jour.

Limitation de conservation

Conserver le temps utile, puis archiver ou supprimer.

Intégrité & confidentialité

Protéger par des mesures de sécurité adaptées.

Le consentement valable

Un consentement affirmatif

Lorsque le traitement repose sur le consentement, celui-ci doit résulter d'une action positive claire. Le silence ou une case pré-cochée ne valent pas consentement.

- Libre : sans contrainte ni conséquence négative
- Spécifique : pour une finalité précise
- Éclairé : la personne sait à quoi elle consent
- Univoque : un acte positif et révocable

Exemple

Une case à cocher non pré-remplie, accompagnée d'une explication claire de l'usage.

Les droits des personnes

Accès

Savoir quelles données sont détenues et y accéder.

Rectification

Faire corriger des données inexactes.

Effacement

Demander la suppression (« droit à l'oubli »), sous conditions.

Portabilité

Récupérer ses données dans un format réutilisable.

Opposition

S'opposer à un traitement (ex. prospection).

Limitation

Geler temporairement un traitement contesté.

Impact opérationnel

Répondre à ces droits dans les délais suppose de savoir où sont les données : le catalogue et le lignage (M9) rendent le RGPD applicable.

Le triangle de la sécurité : CIA

1 Confidentialité

Seules les personnes habilitées accèdent à la donnée (habilitations, chiffrement).

2 Intégrité

La donnée n'est ni altérée ni corrompue (contrôles, journaux).

3 Disponibilité

La donnée est accessible quand il le faut (sauvegardes, résilience).

Sécurité & gouvernance

La sécurité met en œuvre techniquement ce que la gouvernance décide : qui a le droit d'accéder à quoi. Les deux sont indissociables.

Vulnérabilité, menace, risque

1

Vulnérabilité

Une faille : mot de passe faible, accès trop large, absence de chiffrement.

2

Menace

Ce qui peut exploiter la faille : attaque, erreur, vol, fuite.

3

Risque

La probabilité × l'impact si la menace exploite la vulnérabilité.

Exemple — classer les données d'un projet

Données techniciens (personnelles) → base légale, minimisation, accès restreint. Disponibilité des installations (sensible) → niveau confidentiel/restreint. L'Owner classe, le Custodian applique habilitations et chiffrement. La protection suit la sensibilité.

Les techniques de protection

Habilitations donner accès au strict nécessaire (principe du moindre privilège).

Chiffrement rendre la donnée illisible sans la clé, au repos et en transit.

Pseudonymisation remplacer les identifiants par des codes réversibles avec une clé séparée.

Anonymisation rendre la ré-identification impossible (irréversible) — sort du champ RGPD.

Masquage cacher une partie de la donnée (ex. n° de carte ****1234).

Traçabilité journaliser les accès pour détecter et auditer.

Classer pour protéger

La donnée contrainte

Une donnée soumise à des règles spécifiques : RGPD, secret des affaires, réglementations sectorielles, sensibilité pour l'activité. Elle impose des précautions renforcées.

- On attribue un niveau : public, interne, confidentiel, restreint
- Le niveau détermine accès, conservation, traçabilité
- L'Owner décide de la classification
- Le Custodian la met en œuvre techniquement

Exemple

Sur les applications IMO (CMA CGM), classer et qualifier les données conditionnait directement l'usage autorisé.

Catalogage des données

Durée indicative : 45 min

Au programme de ce module :

- La métadonnée
- Métier vs technique
- Le catalogue de données
- Le self-service
- Glossaire & dictionnaire
- Le lignage (lineage)

La métadonnée : donnée sur la donnée

La métadonnée

Une métadonnée décrit une donnée : son sens, son origine, son format, sa qualité, sa sensibilité. C'est ce qui transforme un chiffre en information exploitable et gouvernable.

- Répond à : c'est quoi ? d'où ça vient ? qui en répond ?
- Quelle qualité ? quel niveau de confidentialité ?
- Sans elle, une donnée est un chiffre sans contexte
- Elle est le carburant du catalogue

Exemple

« 78 » seul ne dit rien. Avec ses métadonnées (température, équipement X, capteur Y, °C) il devient exploitable.

Deux familles de métadonnées

Métadonnées métier

- Sens de la donnée — ex. « conso = kWh facturés »
- Règle de gestion / qualité — ex. valeur ≥ 0
- Propriétaire & domaine — ex. domaine Clients, Owner Facturation
- Sensibilité & usage — ex. donnée personnelle, usage facturation

Métadonnées techniques

- Table, colonne, type — ex. FACT_CONSO.conso_kwh (DECIMAL)
- Source & fréquence — ex. plateforme compteurs, MAJ chaque nuit
- Volumétrie & localisation — ex. 30 M lignes/mois, datalake Cloud
- Transformations — ex. relevés horaires → agrégat mensuel

Exemple — la donnée « consommation mensuelle d'un client »

Côté métier : « kWh facturés sur la période », propriétaire = responsable Facturation, donnée personnelle à usage de facturation. Côté technique : colonne FACT_CONSO.conso_kwh (DECIMAL), issue de la plateforme compteurs, rafraîchie chaque nuit, agrégée depuis les relevés horaires.

Le catalogue de données

Le « Google » de vos données

Le catalogue recense le patrimoine de données et le rend trouvable, compréhensible et fiable. C'est le point d'entrée pour qui cherche une donnée.

- Rechercher par mot-clé, domaine, usage
- Comprendre : définition, propriétaire, règles
- Faire confiance : source, fraîcheur, qualité
- Le socle de l'autonomie des utilisateurs

Exemple

Migrer le catalogue de Collibra vers Atlan (CMA CGM) et animer le catalogage des cas d'usage (Safran) : deux missions très concrètes.

Le self-service de la donnée

1 Trouver seul

L'utilisateur cherche dans le catalogue sans solliciter l'IT à chaque fois.

2 Comprendre seul

Il accède à la définition, aux règles et au niveau de qualité.

3 Demander l'accès

Un circuit clair de demande d'accès, tracé et arbitré.

4 Réutiliser

Il consomme un data product existant plutôt que de recréer la donnée.

Le bénéfice

Le self-service libère du temps IT et accélère les métiers — à condition que le catalogue soit alimenté et tenu à jour. Un catalogue vide ne sert à personne.

Glossaire métier & dictionnaire de données

Glossaire métier

- Les définitions des concepts (le « quoi »)
- Langage commun, validé par les Owners
- Fait foi dans toute l'entreprise
- Relié aux objets métier (M7)

Dictionnaire de données

- La description technique (le « comment »)
- Tables, champs, formats, types
- Maintenu avec l'IT
- Relié au glossaire : sens ↔ technique

Le lignage (data lineage)

1

Source

Système d'origine où la donnée est créée.

2

Transformations

Calculs, jointures, règles appliquées au fil des traitements.

3

Cible / usage

Rapport, dashboard ou modèle d'IA qui consomme la donnée.

À quoi ça sert

Le lignage retrace la donnée de la source à l'usage. Il permet l'analyse d'impact d'un changement, le diagnostic d'une erreur et la preuve de conformité (« d'où vient ce chiffre ? »).

Le lignage en pratique

« Ce KPI est faux ! »

Le lignage transforme une enquête à l'aveugle en diagnostic guidé. Sans lui, retrouver l'origine d'une erreur dans un rapport peut prendre des jours.

- On remonte : rapport → calcul → table → source
- On identifie l'étape qui introduit l'erreur
- On mesure l'impact : quels autres rapports touchés ?
- On corrige à la cause, pas au symptôme

Exemple

Un chiffre d'affaires erroné dans un dashboard : le lignage montre qu'une règle de conversion de devise a changé en amont.

ATELIER FIL ROUGE — JOUR 1

Gouverner les données d'un projet

Mise en situation (45 min) — par groupes de 3 à 4

À partir d'un cas EDF : identifier les données critiques, désigner Owner / Steward / Custodian, repérer les données personnelles ou contraintes, et proposer 3 règles de gouvernance prioritaires.

Restitution : 5 min par groupe — consignes détaillées dans le cahier d'exercices.

QUIZ — JOUR 1

Validez vos acquis du Jour 1

13 questions • QCM • 15 minutes • auto-corrigé

Accessible en ligne (lien / QR code) ou sur papier. Correction commentée en groupe juste après.

Récap du Jour 1 : actif data • cycle de vie • maturité • domaines • rôles • gouvernance • objets métier • RGPD/sécurité • catalogue.

JOUR 2

Qualité, architecture & Intelligence Artificielle

De l'usine à données fiable jusqu'aux cas d'usage IA du Groupe

Référentiels de données

Durée indicative : 1 h

Au programme de ce module :

- Master vs Reference vs Transactionnel
- Exemples concrets
- Le golden record
- Les défis du MDM
- Le MDM comme capacité
- Les 3 architectures MDM
- Gouvernance des référentiels

Master, Reference, Transactionnel

Master Data les sujets cœur, partagés et durables : Client, Fournisseur, Produit, Site.

Données de référence les listes qui classent les autres données : pays, devises, codes, statuts.

Données transactionnelles les opérations : une commande, un relevé, un mouvement. Elles s'appuient sur les deux autres.

Comment les distinguer

Master = « de qui / de quoi parle-t-on ? » (sujets). Reference = « dans quelle catégorie ? » (listes). Transactionnel = « que s'est-il passé ? » (événements).

Les données maîtres (Master Data)

Les sujets de l'entreprise

Les données maîtres décrivent les objets cœur de l'activité, partagés par de nombreux processus et durables dans le temps. Elles ont beaucoup d'attributs.

- Client, Fournisseur, Produit, Site, Équipement...
- Beaucoup d'attributs, propres à l'entreprise
- Donnent le contexte des transactions
- Enjeu : une version unique et fiable

Exemple

Le fournisseur « Schneider Electric » : un même sujet utilisé par les achats, la compta, la maintenance.

Les données de référence

Les listes de classification

Les données de référence sont des listes de valeurs qui servent à catégoriser les autres données. Souvent normalisées et partagées au-delà de l'entreprise.

- Pays, devises, codes ISO, unités, statuts
- Peu d'attributs, très stables
- Souvent issues de normes externes (ISO)
- Enjeu : une référence unique et à jour

Exemple

Le code devise « EUR », le code pays « FR », le statut « EN_SERVICE ».

Le « golden record »

La version qui fait foi

Le golden record est la version unique, complète et fiable d'une donnée maître, consolidée à partir de toutes les sources qui la détiennent.

- Une seule « vérité » pour le client, le produit...
- Construite par rapprochement (matching)
- Selon des règles de survivance (quelle valeur gagne ?)
- Distribuée à toutes les applications

Exemple

Le même client dans le CRM, l'ERP et un Excel, avec des écarts → un seul golden record consolidé.

Les défis d'un projet référentiel

Les doublons

Le même client existe sous plusieurs formes : il faut détecter et rapprocher (matching).

Les écarts entre sources

Quelle valeur retenir quand les sources divergent ? Règles de survivance à définir.

La mise à jour

Quand une donnée change, toutes les applications doivent rester cohérentes.

La gouvernance

Qui est Owner de la donnée maître ? Qui valide les règles ? Sans réponse, le référentiel se dégrade.



Retour d'expérience — C'est le cœur d'un projet MDM : transformer des données dispersées et contradictoires en un référentiel unique digne de confiance.

Le MDM comme capacité

1

Identifier

Repérer les données maîtres critiques et leurs sources.

2

Rapprocher

Détecter les doublons, faire correspondre les enregistrements.

3

Consolider

Construire le golden record selon des règles définies.

4

Distribuer

Mettre la donnée de référence à disposition des applications.

Exemple — le golden record « Fournisseur »

« Schneider Electric » existe en 4 variantes (CRM, ERP, GMAO, Excel) avec des écarts. Le MDM les rapproche, construit une fiche unique selon des règles de survivance, puis la distribue à toutes les applications.

Les 3 architectures MDM

1

Registre (Registry)

Un index pointe vers les données restées dans les sources.
Simple, mais requêtes complexes.

2

Hub transactionnel

Les applications passent par le hub pour lire/écrire. Gouvernance forte, mais coût de transformation.

3

Approche consolidée

Hybride : sources conservées + référentiel partagé. Bon compromis, complexité d'intégration.

RX

Retour d'expérience — MDM outillé avec Semarchy : le choix d'architecture dépend toujours du niveau d'imbrication métier et de la maturité des systèmes sources.

Quelle architecture pour quel contexte ?

Plutôt Registre

- Quand les sources doivent rester maîtresses
- Pour démarrer vite, à moindre coût
- Quand on veut surtout « voir » la donnée consolidée

Plutôt Hub / Consolidé

- Quand on veut piloter et écrire la donnée maître
- Pour une gouvernance forte et durable
- Quand la qualité du référentiel est critique

Gouverner les référentiels

Un Owner par référentiel

Le référentiel Client a un propriétaire métier qui en répond.

Des règles explicites

Règles de rapprochement, de survivance, de mise à jour, validées et documentées.

Une qualité mesurée

Taux de doublons, complétude, fraîcheur du golden record suivis dans le temps.

Un cycle de vie

Création, validation, mise à jour, retrait des données maîtres : un processus, pas du cas par cas.

Qualité des données

Durée indicative : 1 h

Au programme de ce module :

- Enjeux & coût de la non-qualité
- « Fit for use »
- Les 8 dimensions
- Traduire en règles
- Le cycle d'amélioration
- Corriger à la source
- Outils & dashboards
- Atelier

Pourquoi la qualité est critique

Conséquences

Décisions erronées, perte de confiance, temps perdu à recouper, coûts cachés, non-conformité, IA biaisée.

Tout le monde concerné

La qualité se construit à la source (saisie) et se maintient tout au long du cycle de vie.

Un investissement

La qualité a un coût, mais la non-qualité en a un bien plus élevé (caché).

Préalable à l'IA

Une IA fiable suppose des données fiables. La DQ prépare le terrain (M14).



Retour d'expérience — « Data Quality Excellence » chez CMA CGM : dashboards de pilotage de la qualité des données sensibles, sous Snowflake et Qlik.

Le coût de la non-qualité

1

coût de PRÉVENIR : contrôler à la saisie, à la source.

10

coût de CORRIGER : nettoyer la donnée une fois l'erreur entrée.

100

coût de SUBIR : décision erronée, client perdu, sanction.

La leçon du « 1-10-100 »

Plus on laisse une erreur avancer dans la chaîne, plus elle coûte cher. Investir à la source est toujours le meilleur rapport coût/bénéfice.

Une qualité « adaptée à l'usage »

« Fit for use »

Une donnée n'est pas bonne ou mauvaise dans l'absolu : elle est bonne si elle répond à l'usage attendu. La qualité se définit toujours par rapport à un besoin métier.

- Le niveau exigé dépend de l'usage
- Une donnée « assez bonne » pour A peut être insuffisante pour B
- On ne vise pas la perfection, mais l'aptitude à l'usage
- Les Owners définissent le niveau attendu

Exemple

Une adresse approximative suffit pour des statistiques régionales, mais pas pour une livraison.

Identifier les données critiques (stratégiques)

VITALE

Elle alimente un processus critique (facturation, production) ou un rapport de direction.

DÉCISIVE

Si elle est fausse, une mauvaise décision business est prise. Sa fiabilité impacte les résultats.

GÉRÉE par un Owner

Un expert métier (Data Owner) peut en garantir la justesse, la définition et l'origine.

Comment la repérer ? 3 questions

Impact business : si cette donnée est fausse pendant 24 h, qui appelle ? Quel impact financier ou client ?

Impact décisionnel : alimente-t-elle un KPI ou un rapport clé pour le management / le CODIR ?

Impact opérationnel : combien de processus ou d'applications critiques s'arrêtent (ou se trompent) si elle est manquante ?

Exemples de données critiques

Périmètre	Donnée critique	Pourquoi vitale ? (impact si fausse)
Production	« Statut d'un équipement »	Maintenance mal planifiée, arrêts non anticipés, indisponibilité des installations.
Clients	« Statut du compte » (actif/résilié)	Facturation erronée, mauvais ciblage, prévisions faussées, insatisfaction client.
Finance	« Centre de coût »	Imputations budgétaires fausses, reporting erroné, décisions d'investissement biaisées.
RH	« Code emploi »	Erreurs de paie, masse salariale faussée, organigrammes incorrects.
Énergie	« Coefficient de conversion »	Volumes et bilans énergétiques faux, conformité réglementaire compromise.

À retenir

Quelques dizaines de données critiques suffisent souvent à couvrir l'essentiel des risques. C'est là qu'on concentre la mesure et la remédiation — pas sur les milliers de champs du SI.

Les 8 dimensions de la qualité

Complétude · Completeness

Toutes les données requises sont présentes.

Exactitude · Accuracy

La donnée reflète la réalité.

Cohérence · Consistency

Mêmes valeurs d'un système à l'autre.

Unicité · Uniqueness

Pas de doublon : une entité une seule fois.

Validité · Validity

Valeurs conformes au format / domaine attendu.

Fraîcheur · Timeliness

À jour et disponible au bon moment.

Intégrité · Integrity

Liens entre données préservés (référentielle).

Vraisemblance · Reasonability

Donnée plausible (pas de naissance en 1850).

En pratique : on cible les dimensions clés

Ces dimensions offrent un langage commun (« la complétude des e-mails est à 82 % »). Mais on ne les mesure pas toutes, partout : pour chaque donnée critique, on retient les 2 à 3 dimensions qui comptent vraiment pour son usage. Mesurer les 8 systématiquement serait coûteux et illisible.

Traduire les dimensions en règles

Complétude règle : taux de valeurs nulles = 0 sur les champs obligatoires.

Unicité règle : aucune clé (nom + e-mail + naissance) en double.

Validité règle : l'e-mail respecte un format ; la date est calendaire.

Cohérence règle : le site est dans la liste de référence normalisée.

Vraisemblance règle : l'âge est compris entre 18 et 67 ans.

Mesurer = règle + score

Chaque règle s'exécute automatiquement et produit un score, suivi dans le temps. C'est ainsi que la qualité devient mesurable et pilotable.

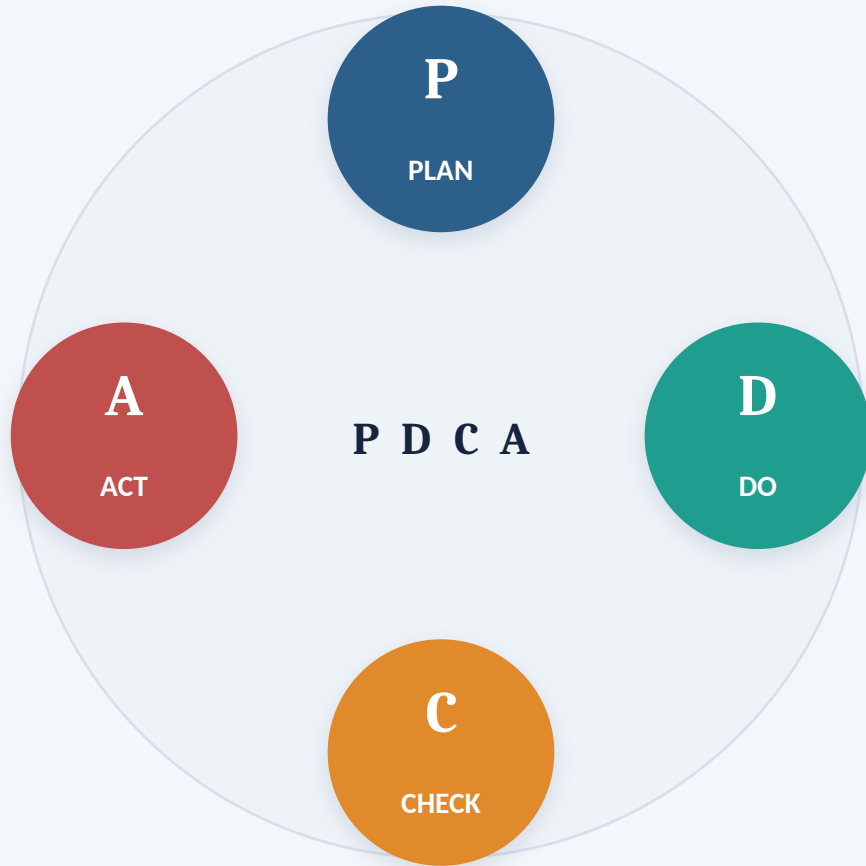
Le cycle d'amélioration continue



Exemple — un KPI de production faux

Le tableau de bord COMEX affiche un volume incohérent. Le lignage révèle que deux sites utilisent une unité de mesure différente (défaut de cohérence). On corrige la règle à la source, pas le chiffre dans le rapport.

La roue de l'amélioration continue (PDCA)



P **Plan** — Diagnostic qualité + plan d'action documenté.

D **Do** — Actions correctives menées et tracées (≈ 4 semaines).

C **Check** — Mesure du score après action + comparaison au point de départ.

A **Act** — Pérennisation : standards, fiches données, registre mis à jour.

Corriger à la source, pas en aval

Nettoyer en aval (à éviter)

- On corrige dans le rapport, à la main
- L'erreur revient au prochain chargement
- Effort répété, jamais résolu
- Chacun « patche » dans son coin

Corriger à la source (à viser)

- On traite la cause racine (saisie, règle)
- L'erreur ne se reproduit plus
- Effort unique, bénéfice durable
- Tout le monde profite de la correction

Outiller la qualité

Profilage

Analyser un jeu de données pour révéler aberrations, NULL, doublons, formats.

Règles & contrôles

Exécuter des règles en continu, avec seuils d'alerte.

Dashboards qualité

Suivre les scores par domaine et dimension (Owners, Data Office).

Workflow de remédiation

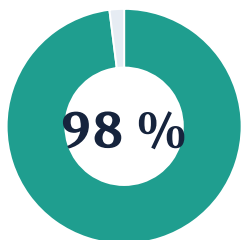
Affecter les anomalies aux Stewards, suivre jusqu'à résolution.



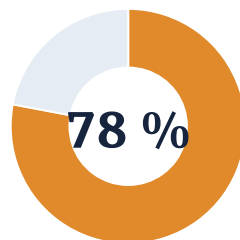
Retour d'expérience — Indicateurs sur les données essentielles, rapports de suivi et dashboards de pilotage (CMA CGM, Airbus) : la mesure rend la qualité pilotable.

Le tableau de bord qualité : ce qu'on suit vraiment

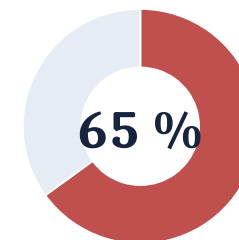
Un score de qualité (0-100) par donnée critique — lecture simple, du rouge au vert



« Statut du compte »



« Centre de coût »



« Téléphone client »

Ce qu'on mesure

% de complétude, exactitude, fraîcheur ; nombre d'anomalies ouvertes / résolues (M-1 → M).

Ce qu'on en fait

Mesurer, publier et améliorer ensemble. Le Data Owner pilote ; le tableau de bord est la boussole partagée.

Atelier : évaluez un jeu de données

1 Repérer

Dans un extrait « Techniciens », identifier les anomalies (doublet, e-mail vide, date impossible...).

2 Nommer

Associer chaque anomalie à sa dimension de qualité.

3 Régler

Proposer une règle de contrôle qui l'aurait détectée.

4 Restituer

Comparer entre groupes : a-t-on vu les mêmes anomalies ?

Consignes

L'atelier complet (extrait de données + grille à remplir) est dans votre cahier d'exercices. 20 min en sous-groupes.

Outillage : les outils du Data Management

Durée indicative : 1 h

Au programme de ce module :

- Pourquoi des outils
- Catalogue & métadonnées
- MDM & Data Quality
- Intégration & analytics
- L'offre du marché
- Bien choisir
- Les pièges

L'outil au service de la démarche

L'outil n'est pas la stratégie

Les outils industrialisent et passent à l'échelle ce que la gouvernance a défini. Mais aucun outil ne remplace les rôles, les règles et les cas d'usage.

- L'outil automatise et fait gagner du temps
- Il rend la gouvernance visible et partagée
- Il ne décide ni du sens ni des règles
- Mal cadré, il devient un « cimetière de données »

Exemple

Un catalogue déployé sans Stewards pour l'alimenter reste vide : l'outil seul ne crée pas la valeur.

Outils de catalogue & métadonnées

Catalogue & gouvernance

Ces plateformes recensent les données, gèrent le glossaire, le lignage et les rôles. C'est la colonne vertébrale d'une gouvernance outillée.

- Recenser, documenter, rendre trouvable
- Glossaire, dictionnaire, lignage intégrés
- Gestion des rôles et workflows de validation
- Exemples : Collibra, Informatica, Atlan, DataGalaxy

Exemple

J'ai pratiqué Collibra, Informatica et Atlan en mission, dont une migration Collibra → Atlan.

Outils MDM & Data Quality

Outils MDM

- Gèrent le référentiel des données maîtres
- Rapprochement, golden record, distribution
- Exemples : Semarchy, Informatica MDM, Stibo

Outils Data Quality

- Profilage, règles, monitoring, remédiation
- Souvent intégrés aux suites de gouvernance
- Ou via SQL / Snowflake pour les contrôles

Intégration & restitution

Intégration / pipelines

- ETL / ELT pour déplacer et transformer
- API et iPaaS pour les échanges
- Orchestration des flux entre systèmes

Stockage & analytics

- Snowflake, datalake / lakehouse (M13)
- BI : Power BI, Qlik, Spotfire
- Exposer, analyser, restituer la donnée

Comprendre l'offre du marché

Suites intégrées

Un éditeur couvre plusieurs briques. Cohérence, mais dépendance à un fournisseur.

Best of breed

Le meilleur outil par besoin, à intégrer. Souplesse, mais complexité.

Cloud vs on-premise

SaaS élastique ou interne, selon souveraineté et contraintes.

Open source vs éditeur

Coût et liberté vs support, maintenance et garanties.

Tendance

Le marché évolue vite, avec une montée des plateformes intégrées « data + IA » dans le cloud. L'important reste l'adéquation à vos besoins, pas l'outil à la mode.

Bien choisir et déployer

1

Partir du besoin

Quel problème de gouvernance ? Quels usages, quels utilisateurs ?

2

Vérifier l'intégration

Connecteurs et interopérabilité avec l'existant.

3

Soigner l'adoption

Ergonomie, accompagnement, formation.

4

Gouverner l'outil

Rôles, droits, maintien du contenu dans le temps.

Critère décisif

Un outil adopté et alimenté à 60 % de ses capacités vaut mieux qu'un outil riche que personne n'utilise. L'adoption prime sur les fonctionnalités.

Les pièges classiques

Acheter avant de cadrer

Choisir l'outil avant d'avoir défini rôles, règles et cas d'usage.

Le « big bang »

Vouloir tout déployer d'un coup au lieu de démarrer sur un périmètre pilote.

Le catalogue fantôme

Un outil déployé mais non alimenté, qui se vide de sens.

Oublier l'adoption

Négliger la formation et l'accompagnement : l'outil reste inutilisé.

Architecture de données & plateforme

Durée indicative : 1 h 30

Au programme de ce module :

- Le rôle de l'architecture
- Modèle d'entreprise & flux
- DWH, Lake, Lakehouse
- Le data swamp
- Intégration : ETL, API, ESB
- Cloud, on-premise, hybride
- Exposition & self-service
- Gouvernance by design

À quoi sert l'architecture de données

Le plan d'urbanisme

L'architecture de données définit comment les données sont structurées, stockées, déplacées et exposées, en cohérence avec la stratégie. C'est le plan d'urbanisme du patrimoine.

- Sans plan, chaque projet construit dans son coin
- Avec plan, les données circulent et se réutilisent
- Elle aligne le technique sur les usages métier
- Elle prépare le passage à l'échelle (et à l'IA)

Exemple

Une ville sans plan d'urbanisme devient ingérable : il en va de même pour les données.

Modèle de données & flux

Modèle de données d'entreprise

- Vue d'ensemble des grands objets métier
- Partagée entre toutes les directions
- Relié aux objets métier (M7) et au catalogue
- Évite les définitions divergentes

Cartographie des flux

- D'où vient la donnée, où va-t-elle
- Par quels traitements elle passe
- Relié au lignage (M9)
- Révèle les dépendances et les risques

Le Data Warehouse

L'entrepôt fiabilisé

Le Data Warehouse stocke des données structurées, modélisées et fiabilisées, destinées au reporting et à la BI. Le schéma est défini en amont (« schema on write »).

- Données propres, certifiées, historisées
- Modèle dimensionnel (faits / dimensions, M7)
- Idéal pour le reporting fiable et le COMEX
- Moins souple pour les données brutes ou non structurées

Exemple

Les tableaux de bord financiers réglementaires reposent typiquement sur un Data Warehouse.

Le Data Lake

Le lac de données brutes

Le Data Lake stocke tous types de données (structurées ou non) à très grande échelle et à bas coût. Le schéma est appliqué à la lecture (« schema on read »).

- Accueille brut, semi-structuré, non structuré
- Stockage massif et économique
- Idéal pour l'exploration et la data science
- Demande de la gouvernance pour rester exploitable

Exemple

Des téraoctets de logs de capteurs IoT stockés bruts, pour entraîner des modèles de maintenance prédictive.

Le Lakehouse

La convergence

Le Lakehouse combine la souplesse et le coût du Data Lake avec la fiabilité et la gouvernance du Data Warehouse. C'est la tendance actuelle des architectures.

- Un seul socle pour brut et données fiabilisées
- Souplesse du lake + qualité du warehouse
- Évite de dupliquer les données entre deux mondes
- Facilite le continuum BI → data science → IA

Exemple

Une plateforme unique où cohabitent reporting certifié et bacs à sable de data science.

Warehouse, Lake, Lakehouse

Data Warehouse structuré, certifié, schema on write — pour la BI et le reporting fiable.

Data Lake brut, massif, bas coût, schema on read — pour l'exploration et la data science.

Lakehouse convergence des deux — un socle unique, gouverné, du reporting à l'IA.

Comment choisir

Souvent, les trois coexistent. Le choix dépend de l'usage : fiabilité certifiée (warehouse), exploration (lake) ou unification (lakehouse).

Le piège du « data swamp »

Du lac au marécage

Un Data Lake sans gouvernance ni catalogue se transforme en « marécage » : on y déverse des données qu'on ne retrouve plus, qu'on ne comprend plus et auxquelles on ne peut plus se fier.

- Données non documentées = inexploitable
- Personne ne sait ce qui est fiable
- Le coût de stockage augmente sans valeur
- La gouvernance (catalogue, qualité) est la digue

Exemple

Un lake de 100 To où plus personne ne sait quelles tables sont à jour ni qui en est responsable.

Déplacer et transformer la donnée

ETL / ELT extraire, transformer, charger — par lots ou en flux, pour alimenter les entrepôts.

API & services exposer la donnée à la demande, en temps réel, de façon contrôlée.

ESB / iPaaS bus et plateformes cloud pour orchestrer les échanges entre applications.

Modèle pivot (canonique) un format commun qui évite de relier chaque application à toutes les autres.



Retour d'expérience — Ma spécialisation historique (IBM) portait sur l'échange de données entre applications : l'interopérabilité reste le nerf de la guerre.

Pourquoi un modèle pivot ?

Sans modèle pivot

- Chaque appli connectée à toutes les autres
- Nombre de connexions qui explose (N^2)
- Un changement casse plusieurs liens
- Maintenance coûteuse et fragile

Avec modèle pivot

- Chaque appli parle un format commun
- Connexions linéaires, pas exponentielles
- Un changement isolé et maîtrisé
- Échanges fiables et tracés

Cloud, on-premise, hybride

Cloud

Élasticité, services managés, paiement à l'usage. Rapide à déployer.

On-premise

Hébergé en interne. Contrôle total, contraintes de souveraineté satisfaites.

Hybride

Le meilleur des deux : composer avec l'existant et les exigences de souveraineté.

Le critère data

Où sont les données sensibles ? La localisation et la conformité guident le choix.

Exposer la donnée en self-service

Exposer, pas enfermer

Une plateforme moderne expose la donnée de façon contrôlée pour que les métiers se servent en autonomie, sans solliciter l'IT à chaque demande.

- Via API, data products, vues partagées
- Self-service encadré par le catalogue (M9)
- Accès tracés et soumis à habilitation
- Responsabilise les domaines (Data Mesh, M4)

Exemple

Un domaine publie un data product « Sites » que les autres consomment via une API documentée.

Gouvernance & sécurité « by design »

Sécurité intégrée

Habilitations, chiffrement, traçabilité branchés nativement sur la plateforme.

Catalogue branché

On ne consomme que ce qui est documenté et gouverné.

Qualité contrôlée

Les contrôles qualité s'exécutent dans la plateforme, en continu.

Transparent pour l'usage

Bien conçue, la plateforme rend la gouvernance presque invisible pour l'utilisateur.

Exemple — un service de prévision de consommation

Data Lake (météo + capteurs bruts) → Data Warehouse (reporting certifié) → data product « prévision » exposé via API → catalogue, qualité et sécurité branchés sur la plateforme.

IA & Data Science

Durée indicative : 1 h

Au programme de ce module :

- IA, ML, Data Science
- BI vs Data Science
- Les types d'IA
- L'IA générative
- Le processus Data Science
- La donnée, carburant de l'IA
- Cas d'usage EDF

IA, Machine Learning, Data Science

Intelligence artificielle techniques permettant à une machine de réaliser des tâches « intelligentes » (analyser, prédire, générer).

Machine Learning l'algorithme apprend des règles à partir de données, au lieu d'être programmé explicitement.

Deep Learning un ML à base de réseaux de neurones, très performant sur image, texte, son.

Data Science la discipline qui combine statistiques, programmation et métier pour créer de la valeur.

L'emboîtement

Data Science englobe le ML qui englobe le Deep Learning. L'IA est le terme parapluie. Le point commun : tout repose sur des données.

Business Intelligence vs Data Science

Business Intelligence

- Décrit le passé : « que s'est-il passé ? »
- Tableaux de bord, indicateurs, reporting
- S'appuie sur des données structurées et fiables
- Réponses connues, agrégées

Data Science

- Prédit et prescrit : « que va-t-il se passer ? »
- Modèles statistiques et d'apprentissage
- Exploite aussi des données brutes et massives
- Réponses probabilistes, à valider

Les grands types d'IA

1 Analyse prédictive

Anticiper : panne d'équipement, demande, consommation.

2 Classification / scoring

Catégoriser ou noter : risque, segment, priorité.

3 Optimisation

Meilleure décision sous contraintes : planning, énergie.

4 IA générative

Produire du contenu : texte, code, image, synthèse.

Le bon type pour le bon besoin

On ne « fait pas de l'IA » en général : on choisit le type adapté à la question métier. La plupart des cas d'usage industriels relèvent du prédictif et de l'optimisation.

L'IA générative

Générer du contenu

L'IA générative produit du texte, du code, des images à partir d'instructions. Elle ouvre de nouveaux usages mais apporte des enjeux spécifiques de confiance.

- Assistants, synthèse, rédaction, support
- Atout : productivité et accès à l'information
- Risque : « hallucinations » (réponses fausses mais crédibles)
- Exige vérification, confidentialité, éthique

Exemple

Un assistant qui résume la documentation interne — utile, à condition de vérifier ses réponses et de maîtriser ce qu'on lui envoie.

Le processus Data Science

1

Comprendre le besoin

Question métier, valeur attendue, critères de succès.

2

Préparer les données

Collecte, nettoyage, qualité — souvent 70 % du travail.

3

Modéliser

Entraîner, comparer, évaluer la performance.

4

Déployer

Mettre en production, intégrer au métier.

5

Surveiller

Suivre la dérive, ré-entraîner, gouverner.

RX

Retour d'expérience — Cas type (Airbus) : exploiter les données de maintenance pour optimiser les plans d'entretien — l'exemple parfait d'un projet « data-to-IA ».

La donnée, carburant de l'IA

70 %

du temps d'un projet IA consacré à préparer et fiabiliser les données.

GIGO

« Garbage in, garbage out » : des données fausses produisent un modèle faux.

Biais

des données non représentatives créent des modèles biaisés.

Ce que la gouvernance apporte à l'IA

- Données fiables, documentées, conformes
- Traçabilité : d'où viennent les données ?

Les enjeux propres à l'IA générative

- Confidentialité de ce qu'on envoie au modèle
- Vérification des réponses, éthique

Des cas d'usage parlants

Maintenance prédictive

Anticiper les pannes d'installations à partir des données capteurs et d'historique.

Prévision de consommation

Prévoir la demande pour optimiser production et achats d'énergie.

Optimisation

Planning, logistique, pilotage énergétique sous contraintes.

Assistants documentaires

IA générative encadrée pour accéder vite à l'information interne.

Exemple — la maintenance prédictive

Besoin : réduire les arrêts subis. Données : capteurs + historique fiabilisés (M11). Modèle : estime le risque de panne par équipement. Production : alerte les équipes, surveillé et ré-entraîné (MLOps, M15).

Outillage IA & MLOps

Durée indicative : 45 min

Au programme de ce module :

- Environnements de développement
- Feature store & catalogue de modèles
- Le MLOps
- Surveiller un modèle
- L'IA générative en entreprise
- Sécurité & garde-fous

Où l'on développe l'IA

Le bac à sable du data scientist

Les environnements de développement permettent d'expérimenter (langages, frameworks ML, notebooks) sur des données gouvernées, avant toute mise en production.

- Notebooks et librairies (Python, frameworks ML)
- Accès à des données cataloguées et de qualité
- Isolés de la production
- Reproductibles et traçables

Exemple

Un data scientist teste un modèle de prévision sur un jeu de données documenté, pas sur une extraction sauvage.

Feature store & catalogue de modèles

Feature store

- Référentiel de variables prêtes à l'emploi
- Documentées, fiables, partagées
- Évite de recalculer les mêmes variables
- Accélère et fiabilise les projets

Catalogue de modèles

- Recense modèles, versions, performances
- Trace les usages autorisés
- Facilite l'audit et la conformité
- Un modèle = un actif gouverné

Le MLOps : industrialiser l'IA

1

Versionner

Données, code et modèles tracés et reproductibles.

2

Automatiser

Pipelines d'entraînement et de déploiement (CI/CD pour l'IA).

3

Déployer

Mettre le modèle en production de façon maîtrisée.

4

Surveiller

Détecter la dérive, mesurer la performance, ré-entraîner.

MLOps = DevOps appliqué à l'IA

C'est ce qui fait passer un modèle « du notebook à la production » de façon fiable et répétable. Sans MLOps, un modèle reste un prototype.

Un modèle en production se gouverne

La dérive (drift)

Un modèle se dégrade avec le temps : le monde change, les données d'entrée évoluent, la performance baisse. Il faut surveiller et ré-entraîner.

- Suivre la performance dans le temps
- Détecter la dérive des données d'entrée
- Surveiller les biais et les écarts
- Ré-entraîner quand nécessaire

Exemple

Un modèle de prévision calibré avant une crise énergétique devient faux si on ne le ré-entraîne pas.

L'IA générative en entreprise

Assistants encadrés

Copilotes internes pour rédaction, synthèse, support — avec garde-fous.

RAG sur données internes

Brancher l'IA sur la documentation gouvernée pour des réponses fiables et sourcées.

Bénéfices

Accès rapide à l'information, productivité, montée en compétence.

Conditions

Données de qualité, encadrement des usages, vérification humaine.

Sécurité & garde-fous de l'IA

Maîtriser les entrées

Ne pas envoyer de données sensibles à un modèle non maîtrisé.

Vérifier les sorties

Contrôler les réponses (hallucinations) ; garder l'humain dans la boucle.

Tracer

Journaliser les usages pour l'audit et la conformité.

Encadrer

Des règles d'usage claires et un comité IA qui valide les cas sensibles (M6).

Communautés Data/IA & ressources

Durée indicative : 30 min

Au programme de ce module :

- Le Data Office, point d'entrée
- Les communautés d'experts
- Le catalogue, ressource vivante
- Monter en compétence (DAMA, CDMP)
- Passer à l'action

Le Data Office, votre premier réflexe

À qui s'adresser ?

Pour toute question data — une définition, un accès, une règle, un outil — le Data Office est le point d'entrée. Il oriente, accompagne et met en relation.

- Méthodes, standards, catalogue, outils
- Accompagnement des projets et des Stewards
- Met en relation avec le bon expert
- Le réflexe avant de « bricoler » dans son coin

Exemple

Plutôt que recréer un énième fichier Clients, on demande au Data Office s'il existe déjà un référentiel.

Les communautés d'experts

Communauté des Stewards

Partage des pratiques de catalogage et de qualité entre domaines.

Communauté BI / analytics

Échange autour des outils de restitution et des bonnes pratiques.

Communauté data science / IA

Réutilisation de modèles, retours d'expérience, veille.

Pourquoi s'impliquer

On apprend plus vite, on évite les erreurs déjà commises, on diffuse la culture.



Retour d'expérience — L'animation de communautés (BI Spotfire chez Airbus, communautés data en mission) fait partie du métier : la culture data se diffuse par le réseau.

Le catalogue, ressource du quotidien

La ressource vivante

Le catalogue de données (M9) n'est pas qu'un outil de gouvernance : c'est la ressource où chacun trouve, comprend et fait confiance à la donnée disponible.

- Le premier endroit où chercher une donnée
- Définitions, propriétaires, qualité, lignage
- Vivant : alimenté par les Stewards
- Le socle du self-service et de l'autonomie

Exemple

Avant de lancer une analyse, on consulte le catalogue pour voir si la donnée existe et si elle est fiable.

Aller plus loin

Le référentiel DAMA-DMBOK

Le corpus international (11 domaines) qui structure cette formation.

La certification CDMP

Certified Data Management Professional (DAMA), dont votre formateur est titulaire.

Ressources internes

Formations, guides, documentation et standards du Data Office.

La veille

Le domaine évolue vite (IA, plateformes) : rester curieux et connecté.

Passer à l'action

Votre premier pas

La meilleure façon d'ancrer cette formation est de l'appliquer tout de suite, sur un cas concret de votre périmètre, en petit.

- Choisir UN cas d'usage de votre quotidien
- Identifier les données, leur Owner, leur qualité
- Appliquer la démarche vue ensemble
- S'appuyer sur le Data Office et les communautés

Exemple

Documenter dans le catalogue les 5 données les plus critiques de votre service : un premier pas concret et utile.

Organisation agile

Durée indicative : 30 min

Au programme de ce module :

- Pourquoi l'agile pour la data
- Itérations & feedback
- Time to market
- Le Product Owner & le backlog
- Les cérémonies
- Agile + gouvernance

Pourquoi l'agile pour la data ?

Avancer par itérations

Les besoins data évoluent et la donnée se découvre en avançant. Livrer par petits incréments réduit le risque et délivre de la valeur plus tôt qu'un projet « tunnel ».

- On apprend en faisant, on ajuste vite
- On livre de la valeur tôt, pas à la fin
- On réduit le risque de se tromper longtemps
- Le métier voit et oriente régulièrement

Exemple

Plutôt qu'un data product « parfait » dans 12 mois, une première version utile dans 6 semaines, puis on améliore.

Le cycle itératif

1

Prioriser

Choisir l'incrément à plus forte valeur dans le backlog.

2

Réaliser

Développer sur un sprint court (1 à 3 semaines).

3

Démontrer

Montrer le résultat au métier, recueillir le feedback.

4

Ajuster

Réordonner le backlog selon les retours et apprentissages.

Le moteur

Le feedback régulier du métier est le carburant de l'agilité : il évite de construire pendant des mois quelque chose qui rate la cible.

Accélérer le « time to market »

Approche « tunnel »

- Tout spécifier, tout construire, livrer à la fin
- Valeur livrée très tard
- Risque élevé de rater la cible
- Difficile de corriger en cours de route

Approche agile

- Livrer vite un premier incrément utile
- Valeur dès les premières semaines
- On corrige et on améliore en continu
- Le risque est réparti, pas concentré à la fin

Le Product Owner & le backlog

Le porteur de la valeur

Le Product Owner porte la vision métier du produit data. Il construit et priorise le backlog pour maximiser la valeur livrée à chaque incrément.

- Représente le métier auprès de l'équipe
- Priorise le backlog (la liste des choses à faire)
- Décide de ce qui est livré, et dans quel ordre
- Fait le lien avec les parties prenantes

Exemple

Product Owner en train SAFe (Airbus) : prioriser le backlog au plus près de la valeur, sprint après sprint.

Les rituels agiles

Planification de sprint

Choisir ce que l'équipe s'engage à livrer sur le sprint.

Point quotidien

Un court échange chaque jour : avancement et obstacles.

Revue (démonstration)

Montrer l'incrément au métier et recueillir le feedback.

Rétrospective

Améliorer la façon de travailler en équipe, en continu.

Concilier agilité et rigueur

Un mythe à dépasser

« Agile = pas de règles » est faux. L'agilité organise la rigueur autrement.

Qualité dans la Definition of Done

Un incrément n'est « fait » que s'il respecte qualité et sécurité.

Gouvernance dans le flux

Catalogage, classification et contrôles sont des tâches du sprint.

Le bon équilibre

Assez de cadre pour la confiance ; assez de souplesse pour la vitesse.

ATELIER PRATIQUE — JOUR 2

Simulation d'un projet « Data-to-IA »

Mise en situation (1 h) — par groupes de 3 à 4

À partir d'un besoin métier EDF : qualifier le cas d'usage, identifier les données et leur qualité, définir la gouvernance (rôles, sécurité), esquisser l'architecture, et décrire le passage à l'IA jusqu'à la mise en production.

De la détection du besoin à la mise en production — trame dans le cahier d'exercices.

QUIZ — JOUR 2

Validez vos acquis du Jour 2

13 questions · QCM · 15 minutes · auto-corrigé

Accessible en ligne (lien / QR code) ou sur papier. Correction commentée en groupe juste après.

Récap du Jour 2 : référentiels & MDM · qualité · outils · architecture · IA & data science · MLOps · communautés · agilité.

Les 6 clés à retenir

1

La donnée est un actif

On la gère avec une stratégie, des rôles et de la valeur en ligne de mire.

2

Pas de valeur sans gouvernance

Rôles clairs, instances, politiques : qui décide, qui contrôle, qui protège.

3

La qualité se mesure et se pilote

Dimensions, règles, scores et remédiation à la source.

4

Référentiels & catalogue

Master Data fiable + catalogue/métadonnées = confiance et réutilisation.

5

Sécurité & conformité by design

RGPD, sécurité, éthique intégrés dès la conception.

6

L'IA repose sur la donnée

Une IA fiable suppose des données gouvernées et de qualité.

Votre feuille de route personnelle

1

Choisir

Un cas d'usage concret de votre périmètre.

2

Cadrer

Données, Owner, qualité, sensibilité.

3

Outils

Documenter dans le catalogue, fixer des règles.

4

Mesurer

Un premier indicateur de qualité ou d'usage.

5

Partager

S'appuyer sur le Data Office et les communautés.

3 actions à 90 jours

Dans votre cahier, formalisez 3 actions concrètes et réalistes à mener au retour. C'est le meilleur moyen de transformer la formation en résultats.

Ressources & certification

Le référentiel DAMA-DMBOK

Le corpus international de référence du Data Management qui structure cette formation.

La certification CDMP

Certified Data Management Professional (DAMA) : valider et faire reconnaître ses compétences.

Votre Data Office & communautés

Le meilleur point d'appui au quotidien : standards, catalogue, entraide.

Mettre en pratique

Appliquer la démarche sur un cas réel, étape par étape.

Merci !

Des questions, un projet data à challenger ? Continuons l'échange.

GRO — Consultant expert Data Gouvernance (CDMP)

guillaumeroca@grsc.fr

Formation Sparks Roca — merci de compléter le questionnaire de satisfaction