

L'Architecture des Données : la pyramide DAMA

Architecture d'Entreprise

L'Architecture des Données : la pyramide DAMA

Architecture d'Entreprise

III — LE CARACTÈRE

« Data is an asset with unique properties — it is not depleted by use, but can degrade in quality if poorly governed. » — DAMA, Data Management Body of Knowledge (DMBOK 2.0)

« Une donnée sans maître est une donnée qui pourrit. »

Vidéo d'introduction

 Télécharger cette leçon en PDF

La mine de charbon qu'on n'arrive pas à exploiter

En 2019, Marc dirige la Direction Data de GreenTech. Il hérite d'une situation que toute organisation moderne redoute : une **mine de données inexploitée**. GreenTech a quinze ans d'historique de production d'énergie, dix ans de données clients, trois ans de données IoT de capteurs. Au total : 47 téraoctets de données, 217 tables dans 12 bases différentes, des millions de lignes par jour.

Marc a un budget, une équipe, et une demande de la direction : « *exploite ces données pour faire des économies et améliorer la satisfaction client* ». Trois projets pilotes sont lancés en parallèle :

- Un projet de maintenance prédictive des éoliennes (avec les données IoT)
- Un projet de scoring du risque client (avec les données contractuelles)
- Un projet de détection des fraudes (avec les données de paiement)

Les trois projets piétinent. Après quatre mois, aucun n'a livré de valeur en production. L'équipe data passe son temps à comprendre où sont les données, à réconcilier des formats incohérents, à traiter des valeurs manquantes ou aberrantes. La direction commence à douter : « *finalement, l'IA, c'est de la poudre aux yeux* ».

Marc fait alors ce qu'il aurait dû faire dès le début : **regarder les données elles-mêmes**, pas les modèles qu'on veut leur appliquer. Il découvre une réalité brutale.

- Les données de capteurs IoT ont 15 % de valeurs manquantes sur les éoliennes les plus anciennes (capteurs défectueux jamais remplacés).
- Les données clients ont trois identifiants différents selon les applications — un même client peut avoir trois identifiants.
- Les données de paiement sont stockées dans un format propriétaire d'un éditeur, dont l'éditeur a cessé le support en 2018.
- Aucune des tables n'a de **propriétaire** identifié. Personne n'est responsable de la qualité, de la complétude, de la documentation.

Marc pose alors la question qui aurait dû être posée en premier : « **quelle est l'architecture de nos données ?** ». Cette question déclenche un chantier de six mois qui n'a rien à voir avec l'IA, mais qui rend l'IA possible.

C'est ce que cette leçon cherche à vous transmettre : **comment l'architecture des données est le prérequis à toute initiative d'exploitation, et comment la pyramide DAMA donne le cadre de ce travail.**

Voici la pyramide DAMA, qui sert de cadre à toute évaluation de maturité data :

Premier principe — Les données sont un actif, pas un sous-produit

La première erreur conceptuelle consiste à considérer les données comme un sous-produit des applications. Cette erreur rend toute initiative data impossible à coordonner, et condamne les projets à dépendre de la bonne volonté des détenteurs applicatifs.

Le DAMA DMBOK 2.0 est sans ambiguïté sur ce point : « *Data is an asset with unique properties* ». Les données sont un actif de l'organisation, au même titre que les bâtiments, les machines, les brevets. Elles ont trois propriétés uniques :

- **Non-déplétion** — l'usage d'une donnée ne la consomme pas (contrairement à une matière première). Une même donnée peut servir mille usages.
- **Dégradation possible** — la qualité d'une donnée peut se dégrader (obsolescence, valeurs manquantes, formats incompatibles). Contrairement aux matières premières, les données peuvent se dégrader même quand elles sont stockées.
- **Externalités** — l'usage d'une donnée par un service peut affecter un autre service (un changement de format cassant les consommateurs en aval).

Ces trois propriétés justifient que les données soient gérées comme un actif, pas comme un sous-produit. Cela signifie : un propriétaire identifié pour chaque donnée, une politique de qualité documentée, un cycle de revue, un budget dédié.

La contrefaçon classique consiste à dire « nos données sont dans les bases de nos applications, donc elles sont gérées ». Cette formulation cache une réalité : les données ne sont gérées que par accident, au gré des évolutions applicatives, sans vue d'ensemble.

***Et toi ?** Combien de jeux de données critiques de ton organisation ont un **propriétaire identifié** ? Combien ont une politique de qualité documentée ? Combien ont fait l'objet d'une revue annuelle ?*

Deuxième principe — La pyramide DAMA distingue cinq niveaux

Le DAMA DMBOK organise les activités de gestion des données en une pyramide à cinq niveaux. Chaque niveau répond à une question différente. Sauter un niveau rend les niveaux supérieurs intenables.

La pyramide DAMA se présente ainsi, du plus opérationnel au plus stratégique :

1. **Niveau 1 — Données opérationnelles** — les données brutes, telles qu'elles sont créées et consommées par les applications transactionnelles.
2. **Niveau 2 — Données de référence (Master Data)** — les données partagées par plusieurs applications, qui doivent être cohérentes entre elles (clients, produits, fournisseurs).
3. **Niveau 3 — Données historisées (Data Warehouse)** — les données consolidées dans un entrepôt, pour l'analyse historique.

4. **Niveau 4 — Données analytiques (BI / Reporting)** — les indicateurs, dashboards, rapports construits à partir du Data Warehouse.
5. **Niveau 5 — Données décisionnelles (Advanced Analytics / IA)** — les modèles prédictifs, les algorithmes d'optimisation, les recommandations.

La règle d'or est **l'ordre des niveaux** : on ne peut pas faire du niveau 5 (IA) sans le niveau 4 (BI), ni du niveau 4 sans le niveau 3 (historisation), etc. Cette règle est souvent violée dans les organisations qui veulent « faire de l'IA » sans avoir stabilisé les niveaux inférieurs.

Pour GreenTech, Marc a compris le problème après quatre mois de tâtonnement : il essayait de faire du niveau 5 (maintenance prédictive) sans avoir stabilisé les niveaux 1 et 2 (capteurs IoT, données contractuelles). Le chantier de remédiation a commencé par les niveaux les plus bas.

***Et toi ?** Sur quelle strate de la pyramide DAMA ton organisation est-elle mature ? Combien de niveaux sont vraiment outillés et maintenus ? Quel est le niveau le plus faible ?*

Troisième principe — Le Master Data Management résout le problème des données éclatées

Le MDM (Master Data Management) est la pratique qui consiste à désigner un propriétaire et une source de vérité pour chaque donnée partagée. Sans MDM, l'organisation vit dans le chaos des données dupliquées.

Le problème classique du MDM est celui des **données client**. Dans une organisation sans MDM, le même client peut apparaître sous trois identifiants différents selon qu'il a été créé dans le CRM, dans l'ERP, ou dans le système de facturation. Trois identifiants, trois adresses, trois historiques de paiement. Toute analyse client est impossible.

Le MDM résout ce problème en trois étapes :

1. **Identifier les données partagées** — clients, produits, fournisseurs, employés, sites géographiques. Liste courte, 5 à 10 entités en général.
2. **Désigner un propriétaire** — un rôle identifié (pas un service), qui a la responsabilité de la qualité de la donnée.
3. **Établir une source de vérité** — une application ou un référentiel qui contient la version canonique de la donnée. Les autres applications consomment cette version (en lecture) ou la mettent à jour (avec validation).

Pour GreenTech, Marc a commencé le MDM par les **clients** (3 mois) puis les **sites de production** (2 mois). Le résultat : un identifiant unique client, trois applications qui consomment ce même identifiant. Le scoring du risque client, bloqué depuis six mois, est devenu un projet de six semaines.

***Et toi ?** Combien d'applications dans ton organisation contiennent une version de la donnée « client » ? Combien ont des identifiants différents ? Combien ont un propriétaire identifié ?*

Quatrième principe — Le Data Warehouse consolide, mais ne déduplie pas

Le Data Warehouse (DW) est le lieu où les données sont historisées pour l'analyse. Sa fonction n'est pas de dédupliquer, mais de fédérer des sources multiples en respectant leur diversité.

L'école Inmon (Bill Inmon, *Building the Data Warehouse*, 1990) prescrit une approche **top-down** : on conçoit d'abord le

modèle de données global de l'organisation (le « grand sujet »), puis on alimente. L'école Kimball (Ralph Kimball, *The Data Warehouse Toolkit*, 1996) prescrit une approche **bottom-up** : on modélise par sujet d'analyse (« data mart »), et le DW émerge de l'union des data marts.

Les deux écoles ont leurs partisans. Le consensus actuel penche pour une **approche hybride** : un modèle de données global pour la gouvernance (Inmon), des data marts pour l'analyse opérationnelle (Kimball).

La contrefaçon classique consiste à utiliser le DW comme un **dénormaliseur** — on prend les données des applications, on les dénormalise, on stocke, et on espère que ça suffise. Cette approche produit un DW qui ressemble aux données sources, sans valeur ajoutée.

Pour GreenTech, Marc a adopté l'approche hybride : un modèle de données global documenté (3 mois), puis trois data marts pour les trois projets pilotes (maintenance, scoring, fraude). Le DW a émergé progressivement, sans céder à la tentation du « grand DW » qui résoudrait tout.

Et toi ? Ton organisation a-t-elle un DW ? Quelle est son architecture — Inmon, Kimball, ou hybride ? Combien de data marts sont en exploitation active ?

Cinquième principe — La gouvernance des données n'est pas un département, c'est une pratique

La gouvernance des données ne se décrète pas. Elle se pratique au quotidien par les propriétaires de données, soutenus par un petit noyau de coordination.

Le DAMA DMBOK identifie les **rôles clés** de la gouvernance :

- **Data Owner** — la personne responsable de la qualité d'une donnée. C'est un rôle métier, pas un rôle IT. Le propriétaire du « client » est le directeur commercial, pas le DSI.
- **Data Steward** — l'opérationnel qui fait le travail quotidien de qualité des données (dédoublonnage, validation, documentation). C'est un rôle technique ou métier, soutenu par le Data Owner.
- **Data Custodian** — la personne qui gère l'infrastructure technique des données (bases, sauvegardes, sécurité). C'est un rôle IT.
- **Chief Data Officer (CDO)** — le dirigeant qui porte la stratégie data. C'est un rôle exécutif, avec budget et mandat.

La contrefaçon classique consiste à créer un « département Data » qui centralise tous les rôles. Cette organisation rend les Data Owners irresponsables (« c'est le département data qui gère ») et le département data infréquentable. La parade est de **garder la gouvernance distribuée** : les Data Owners restent dans les métiers, soutenus par un petit noyau de CDO + Data Stewards.

Pour GreenTech, Marc a obtenu trois Data Owners nommés (directeur commercial pour les clients, directeur technique pour les sites, directrice financière pour les contrats), un Data Steward dédié, et un mandat CDO rattaché au COMEX. Aucun département data créé. La gouvernance est restée distribuée.

Et toi ? Ton organisation a-t-elle des Data Owners identifiés ? Un CDO ? Combien de Data Stewards à temps plein sur les données critiques ?

Application — Mon plan de remédiation data

À rédiger séance tenante. Cinq sections précises.

1. **Diagnostic** — Sur quelle strate de la pyramide DAMA ton organisation est-elle mature ? Quel est le niveau le plus faible ? Quelle est la donnée partagée la plus éclatée (clients, produits, fournisseurs) ?
2. **Cible** — À 12 mois, quelles strates doivent être opérationnelles ? Quel est le périmètre MDM prioritaire ?
3. **Quick wins** — Trois actions à impact rapide (en dessous de 30 jours) : dédoublonnage d'une table, documentation d'un glossaire, nomination d'un Data Owner.
4. **Programme** — Trois chantiers à 6 mois : MDM client, DW unifié, gouvernance formalisée.
5. **Indicateurs** — Trois KPIs pour mesurer le progrès : taux de dédoublonnage client, complétude des données critiques, nombre de Data Owners nommés.

Ce plan est ton premier livrable d'architecture des données. Il est court, mais il est actionnable.

Cas pratique — Refonte du SI Décisionnel de Mutuelle Plus

Contexte sectoriel : assurance santé, dixième acteur français, 1 000 000 de personnes protégées, 800 collaborateurs. Mutuelle Plus fait face à une crise de son modèle historique : population assurée vieillissante (dépenses de santé en hausse), perte de vitesse face aux concurrents digitaux (résiliations en hausse, 72 heures de délai pour répondre à un adhérent là où le marché est à l'instantanéité).

Diagnostic de l'existant. Le SI est verrouillé par le progiciel **Gest+**. Le Data Warehouse atteint **15 To** de données. Les **batches nocturnes** de calcul de prestations débordent sur les horaires d'ouverture, empêchant les agences d'accéder aux dossiers à jour dès 9h. L'absence de centralisation force les conseillers à utiliser des **tableaux locaux**, brisant la continuité de service lors des mobilités géographiques des adhérents. Impact financier : **150 000 €/an** de pertes directes liées aux incidents de paiement, et hausse continue du taux de résiliation faute de réactivité.

Parties prenantes (matrice de Mendelow) :

Partie prenante	Quadrant	Attente clé
Directeur des Risques	Gérer activement	Conformité RGPD, pilotage solvabilité
DSI	Gérer activement	Désendettement technique Gest+, urbanisation
Directeur Financier	Satisfaire	Retour à l'équilibre des comptes, ROI des projets
Data Analysts	Tenir informé	Accès à une donnée propre pour modéliser
Directeurs d'agences	Tenir informé	Vue 360° du client, fin de la double saisie

Trajectoire de transformation TOGAF ADM :

- **Phase préliminaire + Phase A** : principes d'architecture (partage de l'information, indépendance vis-à-vis de l'éditeur).
- **Transition 1 (Hybridation)** : mise en place d'un bus de données pour extraire les flux de Gest+ en temps réel, soulageant le Data Warehouse et réduisant les erreurs de remboursement.
- **Transition 2 (Convergence)** : déploiement du CRM centralisé et décommissionnement progressif des silos « tableaux » en agence.
- **Architecture cible** : SI orienté services (SOA) facilitant l'intégration de garanties spécifiques pour les agents territoriaux.

Livrables attendus de l'apprenant :

1. Une note de Vision (Phase A) positionnant la cible data à 24 mois.
2. Un schéma logique de données cible (Phase C) avec MDM client.
3. Une roadmap de migration (Phases E/F) structurée en trois vagues.
4. Un Business Case incluant la réduction du churn estimée (3 % de churn évité = ~2,4 M€ de CA stabilisé).

Pièges classiques à éviter :

- Sous-estimer la latence lors du transfert des 15 To vers le nouveau référentiel.
- Oublier la conformité RGPD (projet prioritaire identifié dès le départ).
- Négliger la moyenne d'âge des équipes (43 ans) qui nécessite un plan d'accompagnement au changement IT.

Critères d'évaluation : la note est validée si l'apprenant (1) chiffre l'économie attendue, (2) propose une trajectoire incrémentale, (3) identifie les trois risques majeurs et leurs parades.

Encadré Au lecteur / À l'animateur

Question d'ouverture : « Combien de jeux de données critiques ont un propriétaire identifié aujourd'hui dans ton organisation ? »

Profils à observer :

- **Le data scientist qui veut faire de l'IA sans avoir stabilisé les niveaux 1-2.** Vole au-dessus de la pyramide DAMA. *Question* : « Quel est ton taux de valeurs manquantes sur la donnée critique ? »
- **Le MDM-maniaque qui veut tout consolider.** Vise le « grand MDM unifié » qui résoudra tout. *Question* : « Combien de données partagées sont dans ton périmètre ? »
- **Le data Owner par défaut.** Désigne le DSI comme propriétaire de toutes les données, ce qui rend personne responsable. *Question* : « Qui est nommé responsable de la donnée client ? »
- **L'isolé qui documente seul.** Produit un glossaire de 300 termes que personne ne consulte. *Question* : « Combien de fois ton glossaire a-t-il été consulté pour une décision ? »

Pièges d'animation :

- Ne pas laisser la discussion dériver vers les outils (Data Lake, Snowflake, BigQuery). Rappeler que l'architecture des données est d'abord une question de gouvernance.
 - Ne pas confondre Data Lake et Data Warehouse. Le Data Lake stocke des données brutes ; le Data Warehouse les historise avec un modèle.
-

Ce que la leçon 8 vous demande

Ouvrez votre dossier d'architecte à la page suivante. Préparez :

- **Un diagnostic** de la maturité data de ton organisation selon la pyramide DAMA.
 - **Un périmètre MDM prioritaire** : 1 à 3 données partagées, avec Data Owner nommés.
 - **Trois quick wins data** à lancer dans les 30 jours.
-

Sources de cette leçon : DAMA International, *Data Management Body of Knowledge (DMBOK 2.0)* · Bill Inmon, *Building the Data Warehouse* (1990) · Ralph Kimball, *The Data Warehouse Toolkit* (1996) · Cas GreenTech (anonymisé)

Leçon suivante : **L8 — L'Architecture Applicative : la cartographie des systèmes.**

Quiz — 18 questions pour valider votre compréhension

Trois series de six questions. Prenez un crayon, écrivez vos réponses dans l'espace prévu. Score indicatif : 14/18 minimum pour valider la leçon.

A — COMPREHENSION

A1 Quelles sont les trois propriétés uniques des données selon le DAMA DMBOK ? Pourquoi justifient-elles qu'on les traite comme un actif ?

A2 Quels sont les cinq niveaux de la pyramide DAMA ? Pourquoi doit-on respecter l'ordre des niveaux ?

A3 Qu'est-ce que le Master Data Management (MDM) ? Quelles sont les trois étapes pour le mettre en place ?

A4 Quelle est la différence entre l'approche Inmon et l'approche Kimball pour le Data Warehouse ? Pourquoi l'approche hybride est-elle recommandée ?

A5 Quels sont les quatre rôles clés de la gouvernance des données (Data Owner, Steward, Custodian, CDO) ? Quelle est la contrefaçon à éviter ?

A6 Pourquoi un projet d'IA (niveau 5) sans Data Warehouse (niveau 3) ni MDM (niveau 2) est-il condamné à piétiner ?

B — AUTO-EVALUATION

B1 Combien de jeux de données critiques de ton organisation ont un propriétaire identifié ? Une politique de qualité documentée ? Une revue annuelle ?

B2 Sur quelle strate de la pyramide DAMA ton organisation est-elle mature ? Quel est le niveau le plus faible ?

B3 Combien d'applications dans ton organisation contiennent une version de la donnée « client » ? Combien ont des identifiants différents ? Combien ont un propriétaire identifié ?

B4 Ton organisation a-t-elle un Data Warehouse ? Quelle est son architecture — Inmon, Kimball, ou hybride ? Combien de data marts sont en exploitation ?

B5 Ton organisation a-t-elle des Data Owners identifiés ? Un CDO ? Combien de Data Stewards à temps plein ?

B6 Combien de jeux de données critiques ont un propriétaire identifié aujourd'hui dans ton organisation ?

C — ENGAGEMENTS A 30/45/60 JOURS

C1 Diagnostiquer la maturité data de mon organisation selon la pyramide DAMA (5 niveaux, score 1-5). (Échéance : J+7.)

C2 Définir le périmètre MDM prioritaire : 1 à 3 données partagées avec Data Owners nommés. (Échéance : J+14.)

C3 Lancer trois quick wins data : dédoublonnage d'une table, documentation d'un glossaire, nomination d'un Data Owner. (Échéance : J+30.)

C4 Lancer un chantier MDM client sur 3 mois : identifier le propriétaire, établir la source de vérité, réconcilier les applications. (Échéance : J+90.)

C5 Définir trois KPIs data : taux de dédoublonnage client, complétude des données critiques, nombre de Data Owners nommés. (Échéance : J+45.)

C6 Documenter la gouvernance data : matrice RACI des données critiques, revue trimestrielle. (Échéance : J+60.)

