

# Glossary

## Concepts d'architecture

- **Data Warehouse** : entrepôt de données centralisé, structuré et optimisé pour l'analyse (BI, reporting). Données nettoyées et modélisées.
- **Data Lake** : stockage massif de données brutes (structurées ou non) dans leur format d'origine, à faible coût.
- **Data Lakehouse** : hybride combinant la flexibilité du data lake et les garanties transactionnelles/performances du warehouse (ex : Databricks, Delta Lake).
- **Data Mesh** : approche décentralisée où chaque domaine métier possède et expose ses données comme des "produits", avec une gouvernance fédérée.
- **Data Fabric** : couche d'intégration unifiée et automatisée (souvent via métadonnées) qui connecte des sources hétérogènes sans les déplacer.
- **Architecture Médaille** : organisation des données en couches Bronze (brut), Silver (nettoyé), Gold (agrégé, prêt pour le métier). Popularisée par Databricks.
- **Data Mart** : sous-ensemble d'un warehouse dédié à un domaine métier précis (finance, marketing...).
- **Architecture Lambda** : combine un traitement batch et un traitement temps réel en parallèle.
- **Architecture Kappa** : tout en streaming, un seul pipeline (simplification de Lambda).
- **ODS (Operational Data Store)** : zone intermédiaire contenant des données opérationnelles quasi temps réel, avant le warehouse.

## Modélisation

- **Modèle en étoile (Star Schema)** : table de faits centrale entourée de tables de dimensions.
- **Modèle en flocon (Snowflake Schema)** : variante avec dimensions normalisées.
- **Data Vault** : méthode de modélisation (hubs, links, satellites) conçue pour l'historisation et l'auditabilité à grande échelle.
- **SCD (Slowly Changing Dimensions)** : techniques pour gérer l'évolution des dimensions dans le temps (Type 1 : écrasement, Type 2 : historisation...).
- **OLTP vs OLAP** : transactionnel (écritures rapides, ex : base d'une app) vs analytique (requêtes lourdes en lecture).

## Gouvernance et gestion

- **Data Catalogue** : inventaire des actifs data avec métadonnées, permettant la découverte et la documentation (ex : DataHub, Collibra, Alation).
- **Data Lineage** : traçabilité du parcours d'une donnée, de sa source à sa consommation.

- **Data Governance** : ensemble des règles, rôles et processus garantissant qualité, sécurité et conformité des données.
- **MDM (Master Data Management)** : gestion des données de référence (clients, produits) pour avoir une version unique de la vérité.
- **Data Quality** : mesure et amélioration de la fiabilité des données (complétude, cohérence, fraîcheur).
- **Data Contract** : accord formel entre producteur et consommateur de données définissant schéma, SLA et qualité attendue.

## Traitement et pipelines

- **ETL / ELT** : Extract-Transform-Load vs Extract-Load-Transform (la transformation se fait dans le warehouse, approche moderne).
- **dbt (data build tool)** : outil de transformation SQL dans le warehouse (le "T" de ELT), avec versioning, tests et documentation.
- **Apache Spark** : moteur de calcul distribué en mémoire pour le traitement de gros volumes (batch et streaming).
- **Apache Kafka** : plateforme de streaming distribuée pour l'ingestion et la diffusion d'événements en temps réel.
- **Apache Airflow** : orchestrateur de pipelines (DAGs) pour planifier et superviser les workflows data.
- **CDC (Change Data Capture)** : capture des modifications d'une base source pour les propager en aval sans tout recharger.
- **Streaming vs Batch** : traitement continu au fil de l'eau vs traitement par lots planifiés.
- **Reverse ETL** : renvoi des données du warehouse vers les outils opérationnels (CRM, marketing).

## Moteurs de requête et stockage

- **Impala** : moteur SQL massivement parallèle (MPP) pour requêter directement les données Hadoop/HDFS à faible latence.
- **Presto / Trino** : moteur SQL distribué capable de requêter plusieurs sources hétérogènes (fédération).
- **Hive** : couche SQL historique au-dessus de Hadoop, orientée batch.
- **HDFS** : système de fichiers distribué de Hadoop.
- **Parquet / ORC** : formats de fichiers colonnaires optimisés pour l'analytique.
- **Delta Lake / Iceberg / Hudi** : formats de tables transactionnels (ACID) au-dessus du data lake.
- **MPP (Massively Parallel Processing)** : architecture où les requêtes sont réparties sur plusieurs nœuds (Snowflake, BigQuery, Redshift).

## Divers

- **Data Product** : jeu de données packagé, documenté et consommable comme un produit (concept central du data mesh).
- **Semantic Layer** : couche d'abstraction qui traduit les données techniques en métriques métier cohérentes (ex : "chiffre d'affaires" défini une seule fois).

- **Sandbox / Staging** : environnements d'expérimentation ou de préparation avant production.
  - **Data Observability** : monitoring de la santé des pipelines et des données (fraîcheur, volume, schéma).
- 

Revision #1

Created 20 July 2026 16:23:15 by EMB

Updated 20 July 2026 16:23:30 by EMB