



COLLÈGE  
ROSEMONT

# Big Data : Systèmes de gestion de données

Par : Abderrazak Sahraoui

# Sommaire

- Base de données
- Entrepôt de données
- Lac de données
- Entrepôt lac de données

- Quand utiliser une base de donnée vs un entrepôt de données vs lac de données?
- Quel est le rôle de l'architecte dans la construction des systèmes de données?
- Quelle est la différence entre un schéma de base de données et schéma d'un entrepôt?
- Une donnée voyage-t-elle durant son cycle de vie entre BD, ED et LD ?
- Comment déterminer la fraîcheur d'une donnée ?
- Quelle est la source de données pour chaque structure ?

# Base de données



**What is a Database?**

RELATIONAL DATABASE

- Designed to capture and record data (OLTP)
- Live, real-time data
- Data Stored in tables with rows and columns
- Data is highly detailed
- Flexible Schema (how the data is organized)

The thumbnail also features a small video inset of a man with glasses and a beard, and three white database cylinder icons on a dark blue background.

[https://www.youtube.com/watch?v=-bSkREem8dM&ab\\_channel=AlexTheAnalyst](https://www.youtube.com/watch?v=-bSkREem8dM&ab_channel=AlexTheAnalyst)

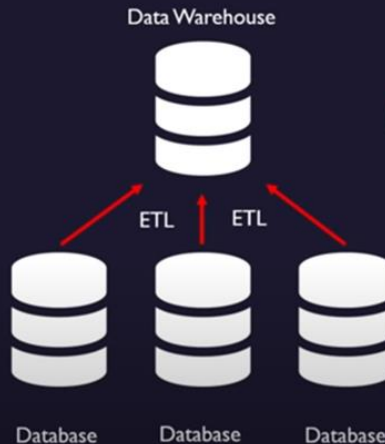
- Une base de données permet d'enregistrer des données provenant d'un processus OLTP (Online Transaction Process). Autrement dit, elle permet de capturer des données fraîchement créées par une application.
- Les données sont stockées dans les bases de données relationnelles sous forme de table de plusieurs colonnes et plusieurs lignes.
- Le schéma relationnel d'une table peut être aisément modifié comparé à d'autres structures. Il est possible d'ajouter de nouvelles colonnes à une table ou d'en supprimer ou modifier.

# Entrepôt de données

## What is a Data Warehouse?

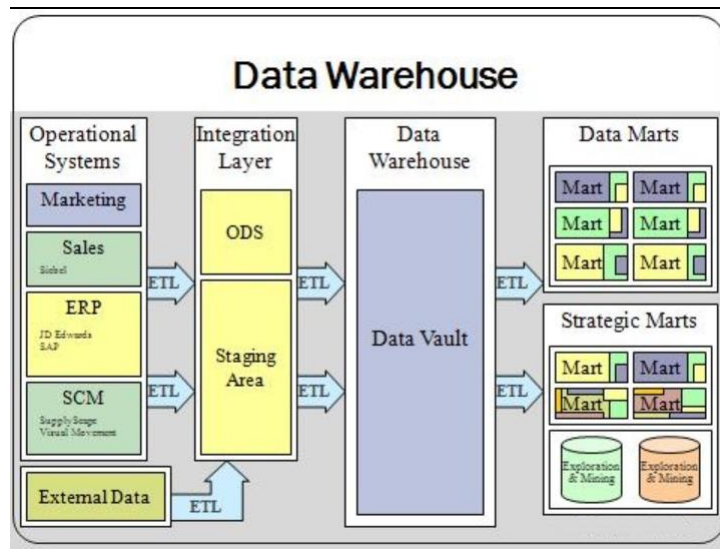
### RELATIONAL DATABASE

- Designed for analytical processing (OLAP)
- Data is refreshed from source systems – stores current and historical
- Data is summarized
- Rigid Schema (how the data is organized)



- Un entrepôt de données permet d'entreposer des données provenant de différentes sources de données. Lesquelles données sont destinées à être analysées par un processus OLAP (Online Analytics Process)
- Le but d'un entrepôt de données est de fournir une référence unique pour un ensemble de données pouvant servir dans la prise de décisions au sein de l'entreprise, et d'offrir les outils nécessaires aux processus analytiques BI (Business intelligence ou Informatique décisionnelle).

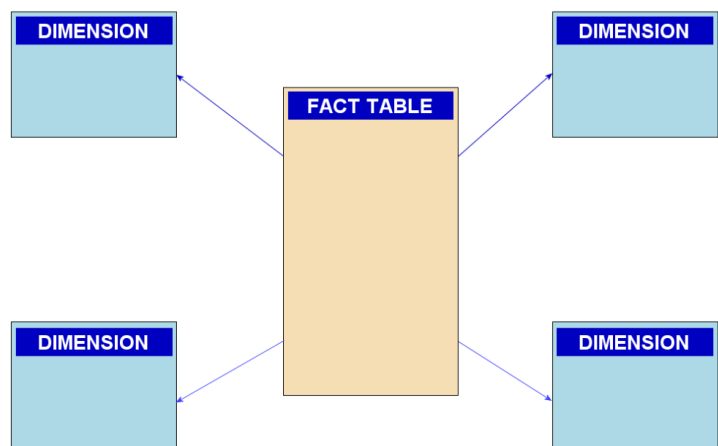
# Entrepôt de données



- En amont, Les données arrivent à l'entrepôt par le biais d'un processus ETL (Extract, Transforme et Load). Les données sont extraites de sources localisées dans des systèmes transactionnels en production. Les données sont épurées ou transformées par filtrage, codage et certification.
- Les données de l'entrepôt peuvent être conservées sous deux formes :
  - sous forme élémentaire et détaillée.
  - sous forme agrégée selon des axes ou des dimensions d'analyse prévues. il est impossible de retrouver le détail et la profondeur des indicateurs une fois ceux-ci agrégés. (par exemple, si l'on a agrégé les résultats par mois, il ne sera plus possible de faire une analyse par journée).
- En aval, les données peuvent être restituées aux usagers par des outils OLAP de :
  - requêtes ou reporting,
  - cubes ou hypercubes,
  - fouille de données.

[https://fr.wikipedia.org/wiki/Entrep%C3%B4t\\_de\\_donn%C3%A9es](https://fr.wikipedia.org/wiki/Entrep%C3%B4t_de_donn%C3%A9es)

# Entrepôt de données



- Les entrepôts utilisent le modèle dit en étoile où les tables sont réparties en deux catégories : tables de faits et tables de dimension.
- Chaque modèle en étoile est constitué d'une table centrale de faits contenant les mesures comme montant, quantité, etc. et de plusieurs tables de dimension comme le temps (jour, mois, année) nomenclature (famille de produit, sous-famille, ...) segmentation clientèle (sexe, tranche âge,...)
- La jointure dans un modèle en étoile entre table de faits et tables de dimension est facilitée (optimisée) par la présence d'une clé calculée à partir des clés des tables de dimension ce qui facilite l'analytique.
- Le modèle dit en flocon est une variante du modèle en étoile où les tables de dimensions sont normalisées évitant ainsi la redondance et permettant un gain d'espace de l'ordre de 5 à 10%.

[https://fr.wikipedia.org/wiki/%C3%89toile\\_\(mod%C3%A8le\\_de\\_donn%C3%A9es\)](https://fr.wikipedia.org/wiki/%C3%89toile_(mod%C3%A8le_de_donn%C3%A9es))

# Comparatif

## Key Differences

- Databases are designed for Transactions, Data Warehouses are designed for analytics and reporting
- Databases data is fresh and detailed, Data Warehouses data is refreshed periodically and is summarized
- Databases work slowly for querying large amounts of data and can slow down transactional processes, Data Warehouses don't interfere with any processes and are generally faster



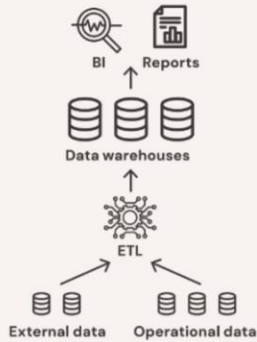


# Comparatif

Caractéristique	Base de données de production	Data warehouses	Datamarts
Opération	gestion courante, production	référentiel, analyse ponctuelle	analyse récurrente, outil de pilotage, support à la décision
Modèle de données	entité/relation	3NF, étoile, flocon	étoile, flocon
Normalisation	fréquente	maximum	rare (redondance d'information)
Données	actuelles, brutes, détaillées	historisées, détaillées	historisées, agrégées
Mise à jour	immédiate, temps réel	souvent différée, périodique	souvent différée, périodique
Niveau de consolidation	faible	faible	élevé
Perception	verticale	transverse	horizontale
Opérations	lectures, insertions, mises à jour, suppressions	lectures, insertions, mises à jour	lectures, insertions, mises à jour, suppressions
Taille	en gigaoctets	en téraoctets	en gigaoctets

# Data Warehouse

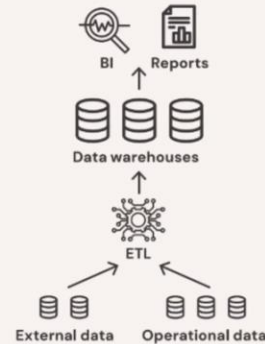
## Data warehouse



### Pros:

- Business intelligence (BI)
- Analytics
- Structured & clean data
- Predefined schemas

## Data warehouse



### Cons:

- No support for semi or unstructured data
- Inflexible schemas
- Struggled with volume and velocity upticks
- Long processing time

# Lac de données

## What is a Data Lake?

### RELATIONAL DATABASE

- Designed to capture raw data (structured, semi-structured, unstructured)
- Made for large amounts of data
- Used for ML and AI in its current state or for Analytics with processing
- Can organize and put into Databases or Data Warehouses

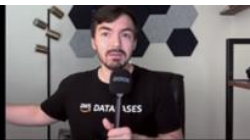


[https://fr.wikipedia.org/wiki/Lac\\_de\\_donn%C3%A9es](https://fr.wikipedia.org/wiki/Lac_de_donn%C3%A9es)

- Un lac de donnée permet le stockage rapide de données massives hétérogènes dans leur format original ou avec peu de transformation.
  - Données structurées issues de bd relationnelles.
  - Données issues de bases NoSQL
  - Données semi-structurées (fichiers CSV, journaux, xml, json,...)
  - Données non structurées ( emails, documents, pdf
  - Fichiers blob (images, audio, vidéo)
- Les lacs sont utilisés par des ingénieurs de données et des scientifiques de données pour des applications en apprentissage machine et intelligence artificielle.
- Lorsqu'une donnée arrive au lac, elle se verra dotée d'un identifiant et de balises de métadonnées. Lorsqu'un besoin se présente, le Data Lake est parcouru pour y rechercher des informations pertinentes. L'analyse de ces données permet alors d'apporter de la valeur et de répondre à ce besoin.
- Le stockage se fait en utilisant l'architecture d'un cluster Hadoop.
- Les données peuvent être conservées dans le lac pour un usage ultérieur non prédéterminé.

# Lac de données

**DATA LAKES**



- 01** Schema-on-read
- 02** Can store structured, semi-structured or non-structured data
- 03** Built to process large amounts of data at a lower cost than data warehouses
- 04** Very flexible and easy to make changes to
- 05** Tracks historical information but it isn't always as well structured as the data warehouse
- 06** Used by data scientists/data engineers and other highly technical users (unless you put another layer on top like Hive)

[https://www.youtube.com/watch?v=ExpRL0m9BcA&ab\\_channel=SeattleDataGuy](https://www.youtube.com/watch?v=ExpRL0m9BcA&ab_channel=SeattleDataGuy)

- Les avantages des lacs de données sont :
  - la rationalisation du stockage des données,
  - la réduction des coûts de stockage,
  - et la facilitation de l'accès pour l'analyse et la prise de décisions d'une façon globale.
- Les inconvénients sont :
  - la difficulté à conserver un lac de données propre et organisé,
  - la difficulté à organiser et maintenir une gouvernance des données efficace,
  - le temps nécessaire à traiter et analyser les données stockées à l'état brut.
  - L'expertise requise pour rechercher, analyser et traiter les données de manière pertinente et créatrice de valeur, souvent confiées aux Data Scientists
  - la sécurité, la confidentialité et les problématiques liées aux données personnelles et au respect des réglementations.
- Plusieurs environnements fournissent des services complets permettant la gestion d'un lac de données. La plupart d'entre eux sont basés sur la technologie Hadoop et fournissent des installations en local (MapR, Cloudera, Hortonworks) ou dans le Cloud (Microsoft Azure, Google Cloud Platform, Amazon S3)

# Lac de données

## Data warehouse vs. data lake

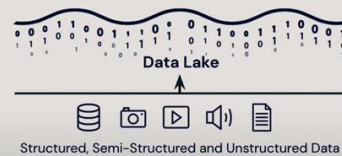
	DATA WAREHOUSE	DATA LAKE
DATA TYPES	Structured, processed data from operational databases, applications and transactional systems	Structured, semistructured and unstructured data from sensors, apps, websites, etc.
PURPOSE	Predefined purpose for business intelligence, batch reporting and data visualization	May not have a predefined purpose; typically used for machine learning, deep analysis and discovery
USERS	Data engineers, business analysts, data analysts	Data engineers, data scientists
SCHEMA POSITION	Schema-on-write	Schema-on-read
BENEFITS	Categorized historical data stored in a single repository with ease of access for the end user	Data stored in its native format, allowing flexibility for data scientists to analyze and develop models from diverse data sources

# Lac de données

## Data Lakes

### Pros:

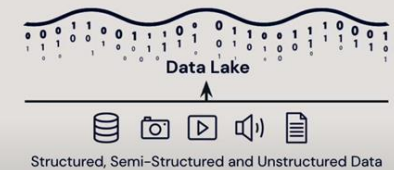
- Flexible data storage
- Streaming support
- Cost efficient in the cloud
- Support for AI and Machine Learning



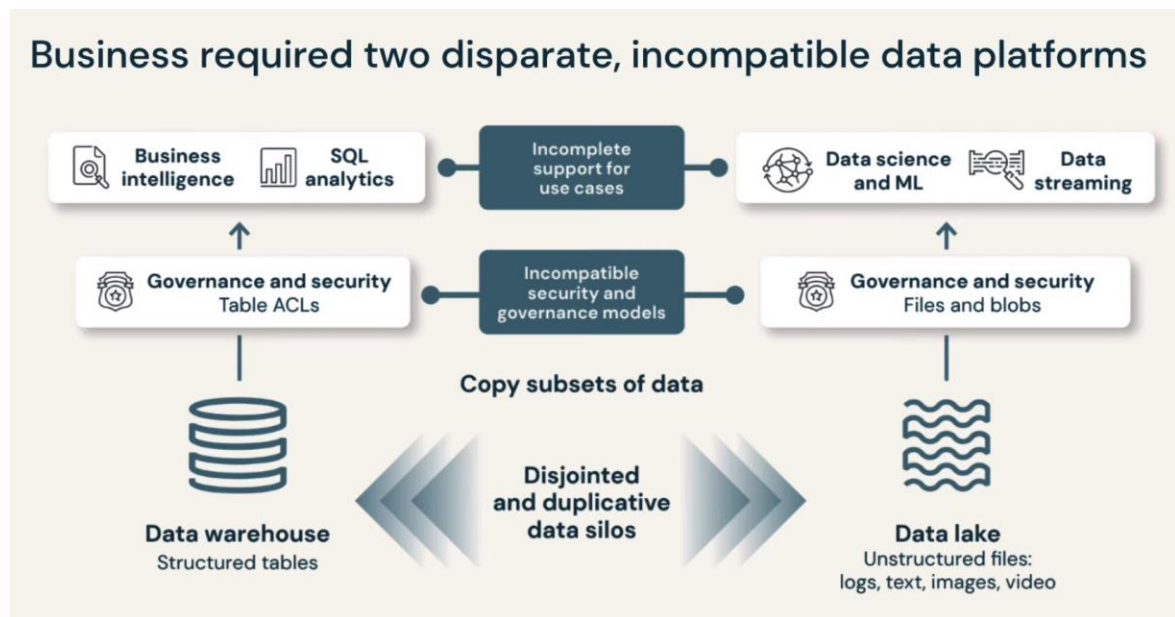
### Cons:

- No transactional support
- Poor data reliability
- Slow analysis performance
- Data governance concerns
- Data warehouses still needed

## Data Lakes



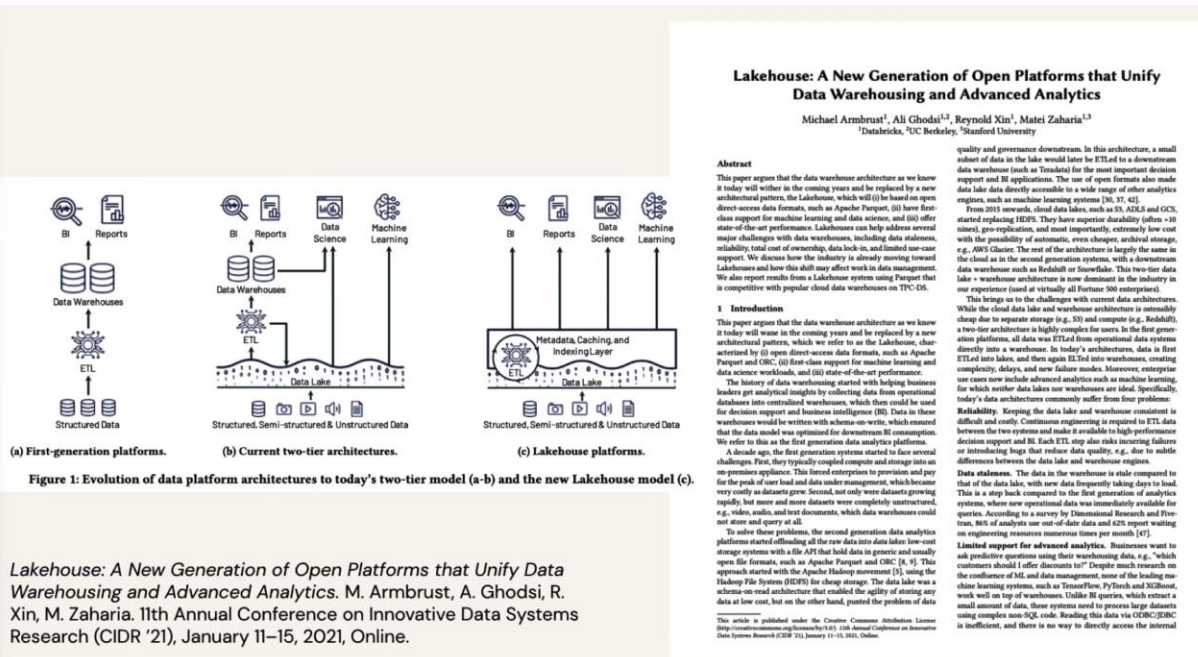
# Entrepôt et lac de données



- Besoin d'une plateforme de données qui combine les avantages des deux solutions et élimine leurs inconvénients.



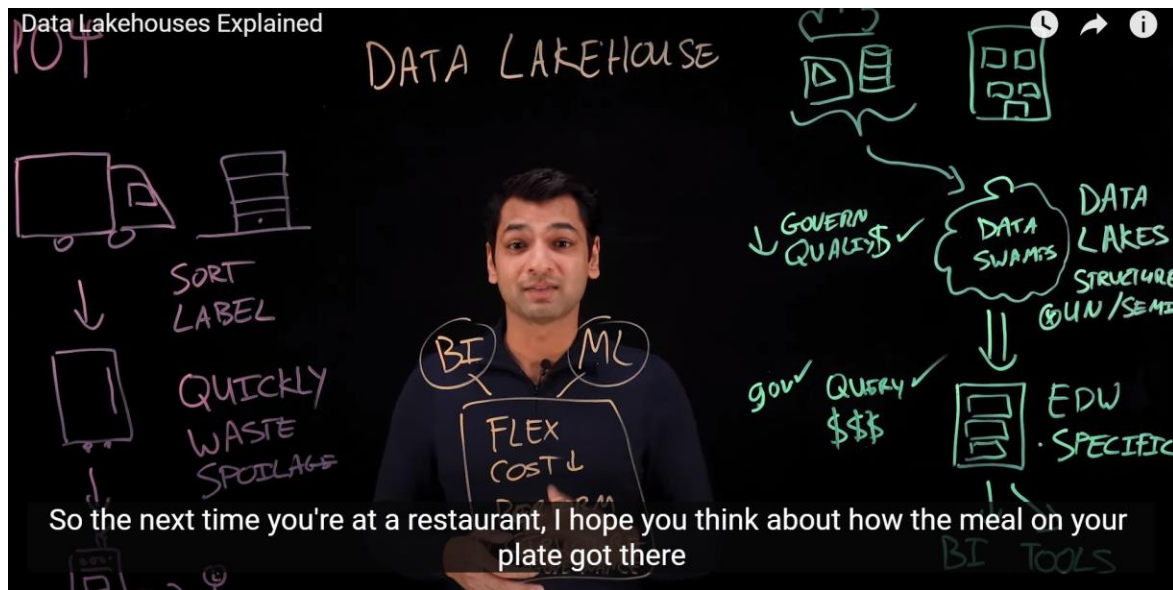
# Databricks Lakehouse



<https://www.databricks.com/learn/training/lakehouse-fundamentals-accreditation>

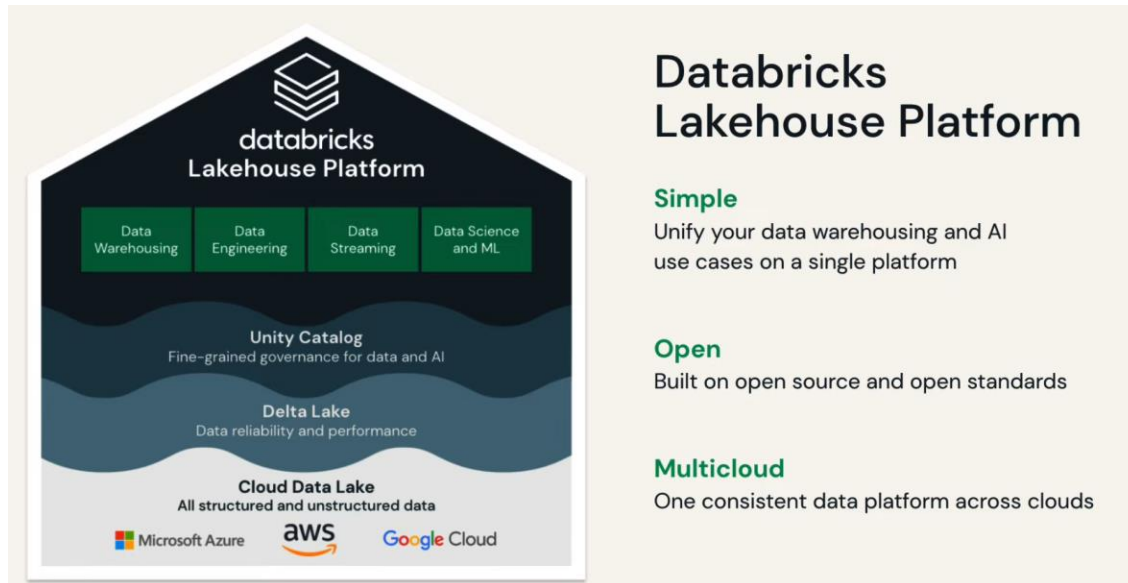


# Lakehouse : Entrepôt lac de données



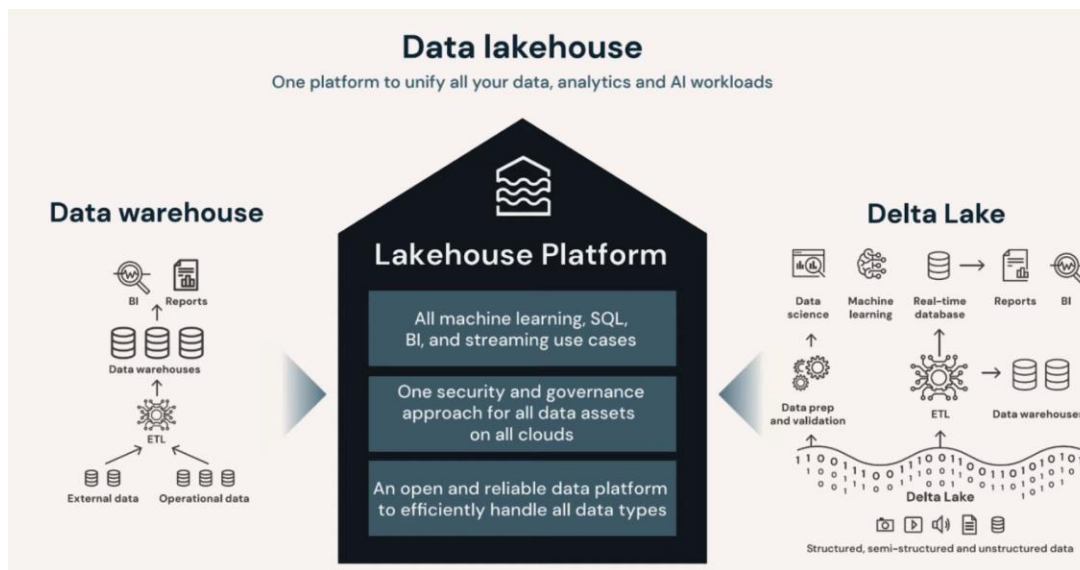
[https://www.youtube.com/watch?v=Enu-EH7RHMH&ab\\_channel=IBMTTechnology](https://www.youtube.com/watch?v=Enu-EH7RHMH&ab_channel=IBMTTechnology)

# Databricks Lakehouse



[https://www.youtube.com/watch?v=Enu-EH7RHMM&ab\\_channel=IBMTechology](https://www.youtube.com/watch?v=Enu-EH7RHMM&ab_channel=IBMTechology)

# Entrepôt lac de données



- Databricks offre une architecture hybride unifiant une plateforme d'entrepôt de données et une plateforme de lac de données.

# Entrepôt lac de données

## Key features of a data lakehouse.

- Transaction support
- Schema enforcement and governance
- Data governance
- BI Support
- Decoupled storage from compute
- Open storage formats
- Support for diverse data types
- Support for diverse workloads
- End-to-end streaming