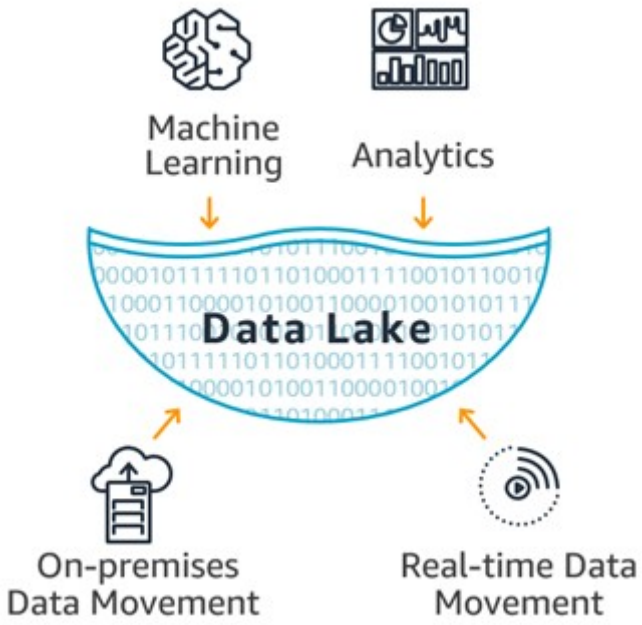
Data Lake

Lucile Sautot
2020

Le concept général



Source de l’image : Amazon

Définition

Un lac de données est une collection de données telle que :

- **Les données n’ont pas de modèle fixe**
- **Tous les formats de données sont possibles**
- **Les données n’ont pas été transformées**

Définition

Un lac de données est une collection de données telle que :

- **Les données sont conceptuellement présentes à un seul endroit … mais peuvent être physiquement distribuées**

Définition

Un lac de données est une collection de données telle que :

- **Les données sont utilisées par un ou plusieurs experts en sciences des données**

Définition

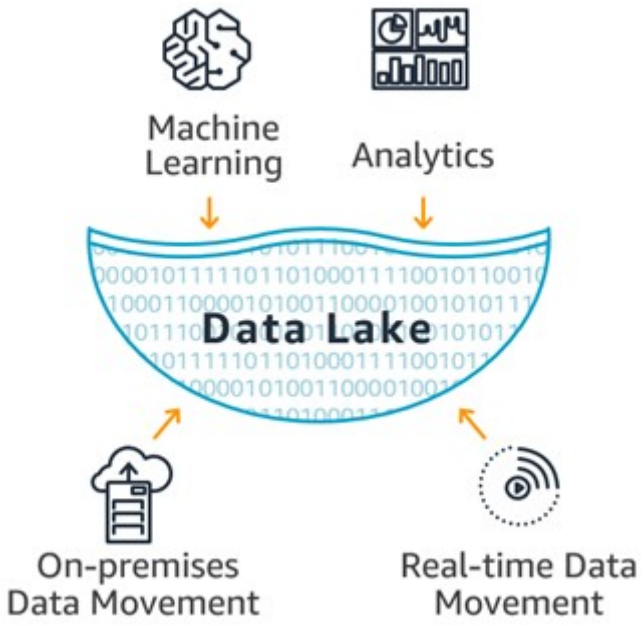
Un lac de données est une collection de données telle que :

- **Les données sont associées à un catalogue de métadonnées**
- **Les données sont associées avec des règles et des méthodes pour leur gouvernance**

En résumé (très grossier)

- **On ne modélise pas les données …**
- **Mais la gestion des données.**

Architecture d’un data lake



Source de l’image : Amazon

Architecture d’un data lake

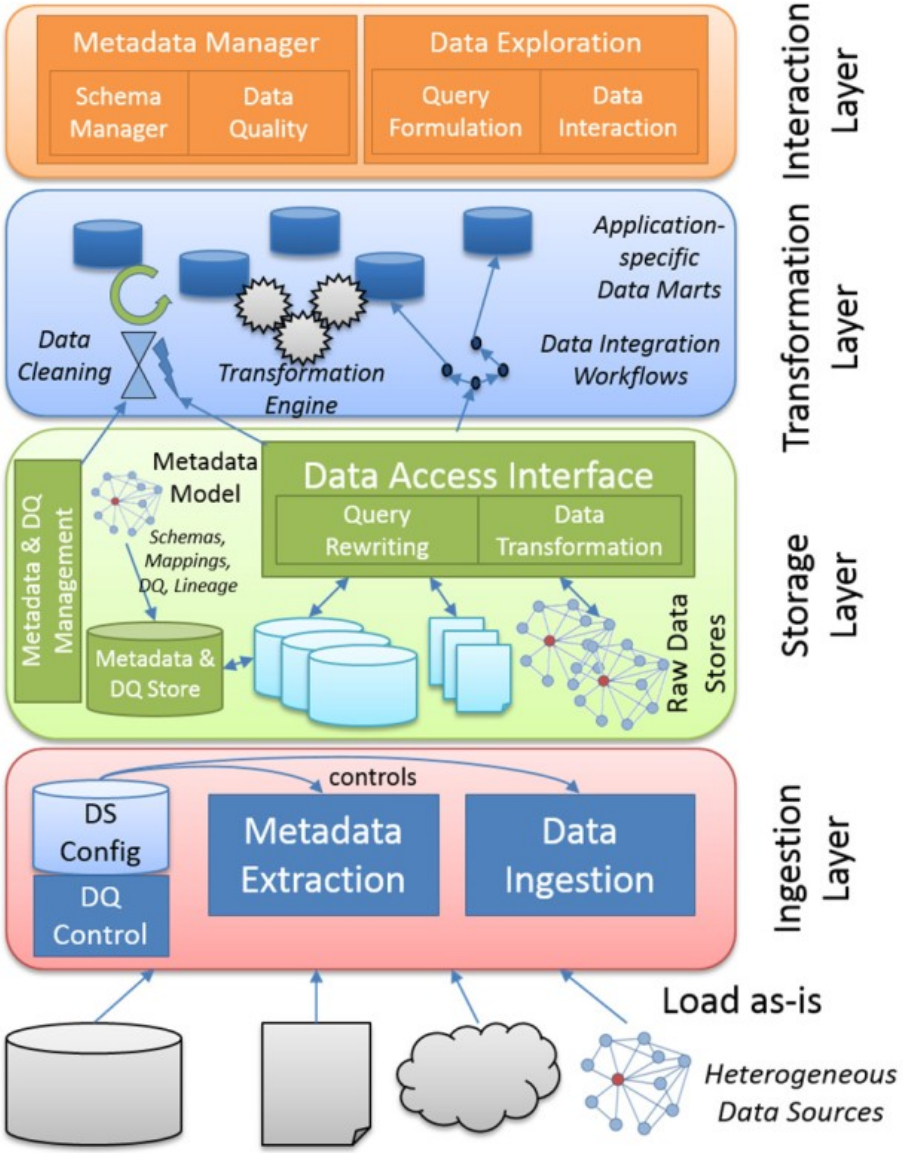


Figure 2.1. *Architecture of a data lake*

Architecture d’un data lake

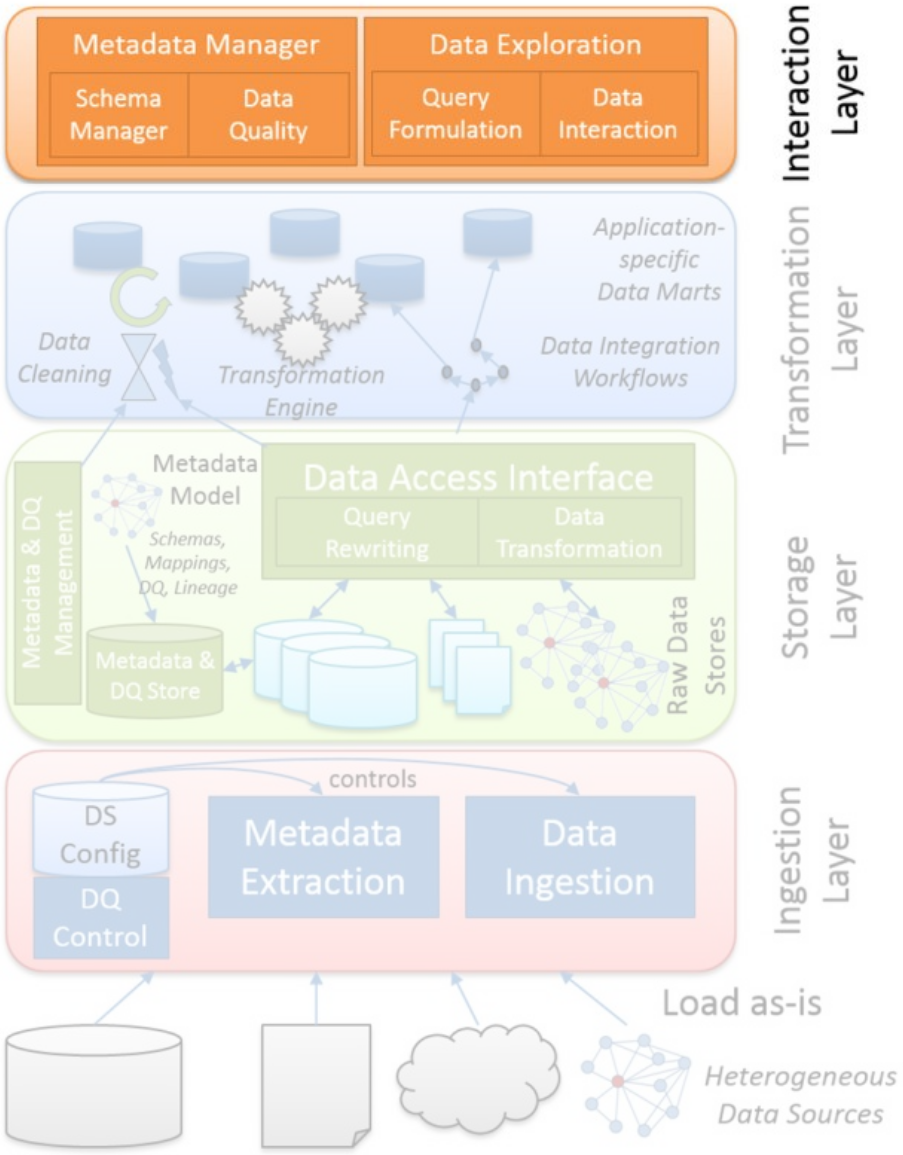


Figure 2.1. *Architecture of a data lake*

Architecture d’un data lake

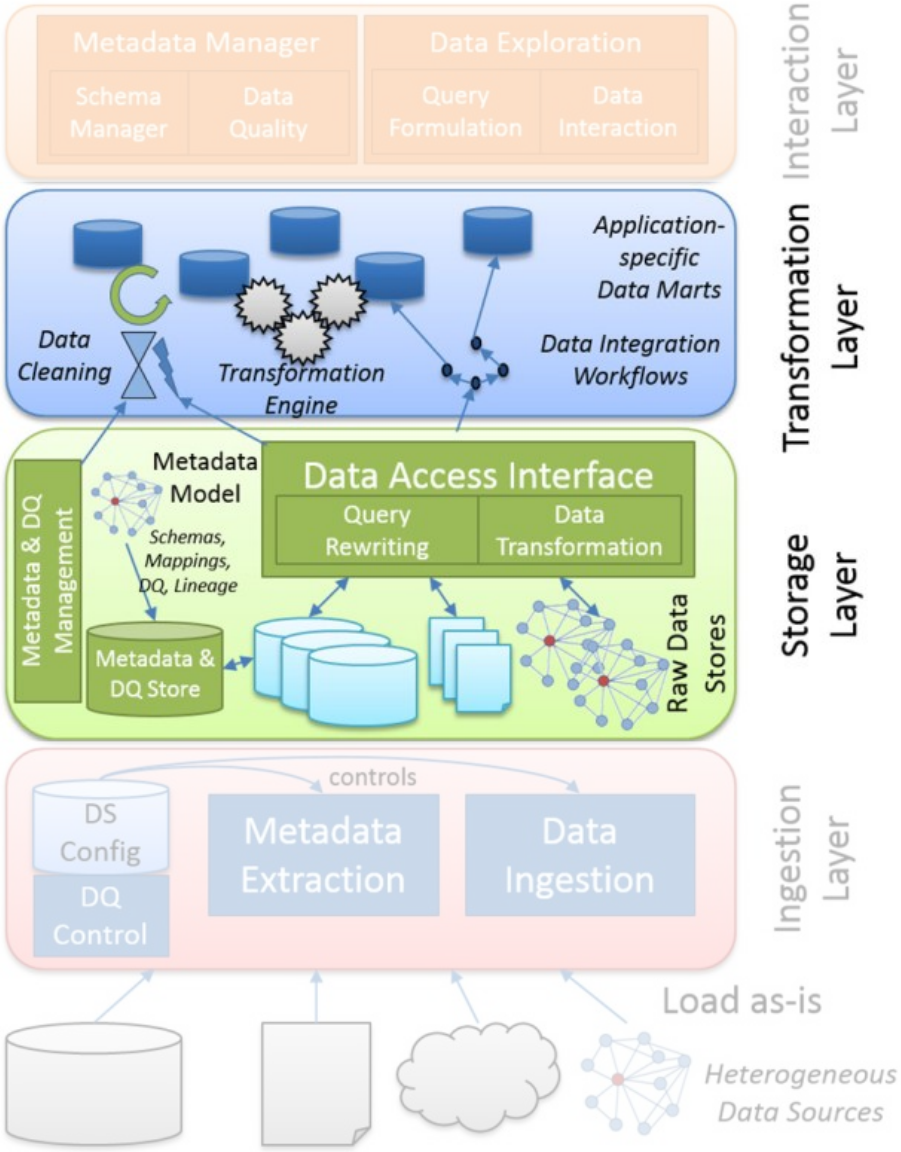


Figure 2.1. *Architecture of a data lake*

Architecture d’un data lake

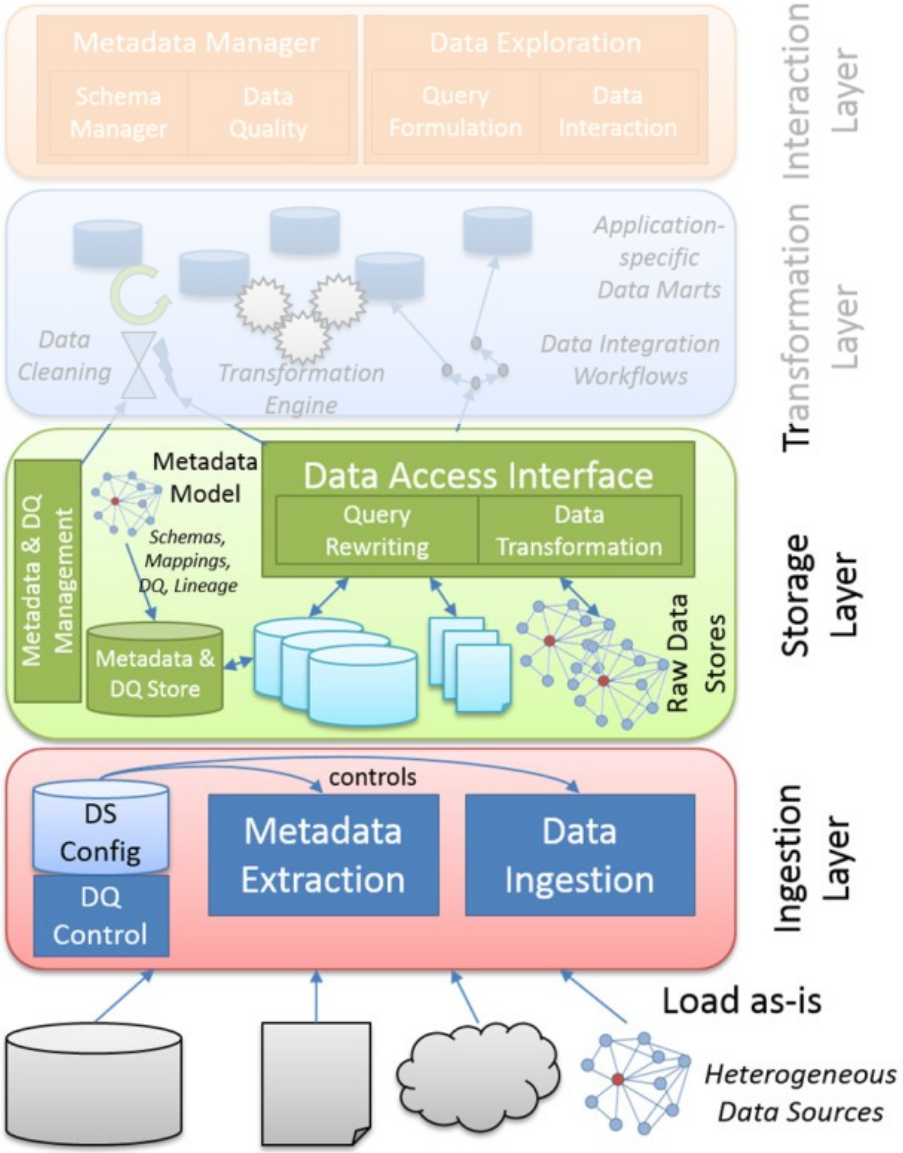
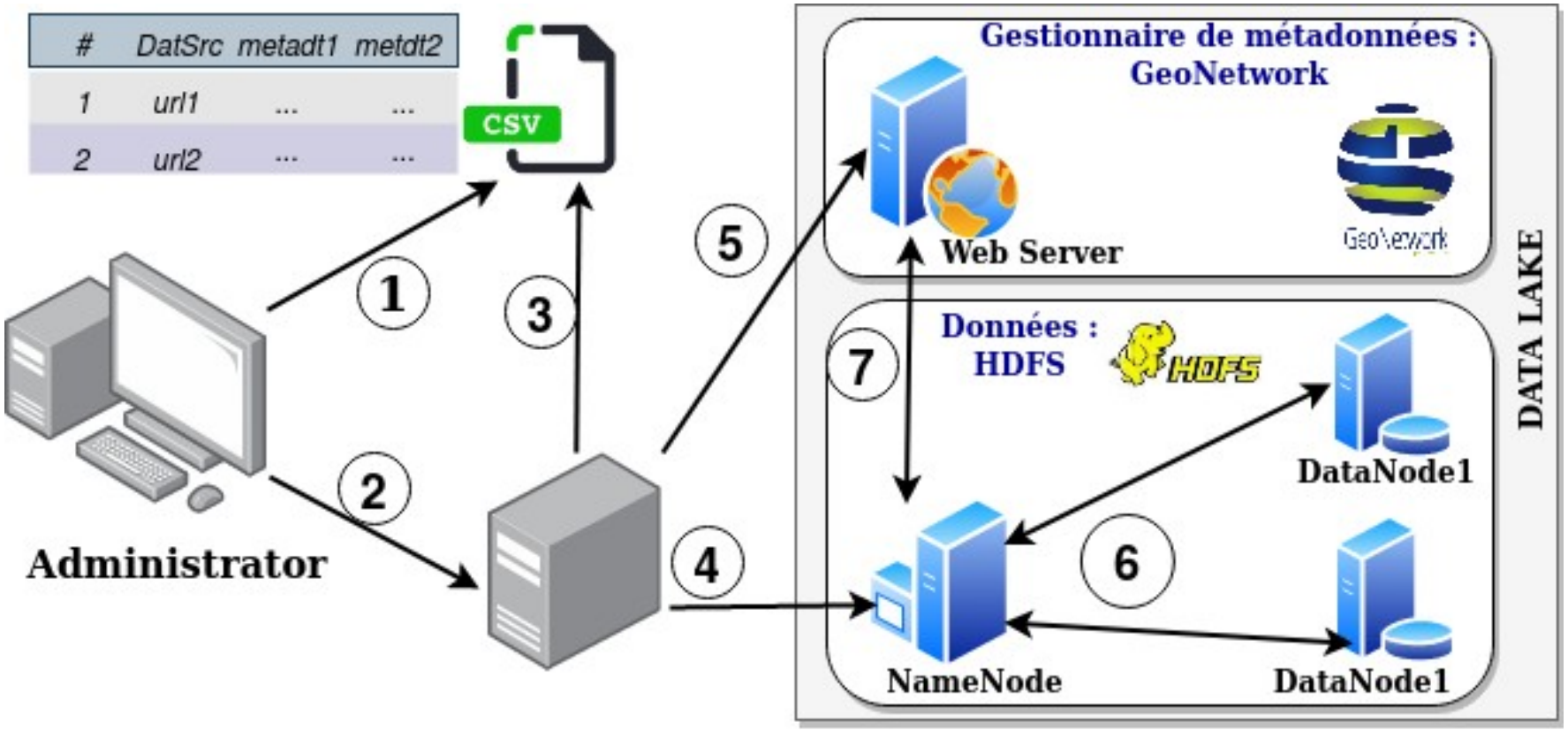


Figure 2.1. *Architecture of a data lake*

Un exemple de début de mise en place



1 - Liste des jeux de données à télécharger dans un fichier .csv

2 - Exécuter le programme pour parcourir et collecte sur le web

3 - Lecture du fichier .csv

4 - Insertion des jeux de données dans HDFS

5 - Générer la fiche de méta-données et d'index

6 - Distribution des données dans les datanodes

7 - Indexation de données dans HDFS à partir de GeoNetwork

Data base, data warehouse ou data lake ?

- **Base de données relationnelles**
- **Base de données NoSQL**
  - Orientée documents
  - Orientée graphes
- **Entrepôt de données**
- **Lacs de données**

Un peu de lecture

- **Anne Laurent, Dominique Laurent, and Cédrine Madera, eds. Data Lakes. John Wiley & Sons, 2020.**
- **Sawadogo, Pegdwendé, and Jérôme Darmont. "On data lake architectures and metadata management." Journal of Intelligent Information Systems (2020)**
- **Rodrique Kafando, Rémy Decoupes, Lucile Sautot, Maguelonne Teisseire. Spatial Data Lake for Smart Cities: From Design to Implementation. AGILE: GIScience Series, 2020, 1**

Merci à tous !