Université de Tunis El Manar.
Ecole Nationale d'Ingénieurs de Tunis.
Département de Technologies d'Informations et
de Communications.

ù

Projet de fin d'année

Les techniques d'analyse de données dans les Data Lakes

Elaboré par :

Barhoumi Dhia Eddine
Hammali Wael

Encadré par :

Mr.Mohamed Ali Zoghlami

Classe : 1er année informatique

Année universitaire : 2024/2025

Remerciements

En cette année 2025, nous avons eu l'opportunité de réaliser notre projet de fin d'année. Ce projet, intitulé "Les techniques d'analyse de données dans les data lakes", a été un véritable challenge pour nous, mais aussi une source d'apprentissage. Nous réservons cette page pour mercier sincèrement toutes les personnes qui ont apporté leur assistance pour assurer le succès de notre travail. Nous voulons également adresser nos remerciements les plus sincères à notre encadreur Mr.Mohamed Ali Zoghlami, sa décision de nous choisir ce sujet de projet et son soutien continuont été essentiels pour notre réussite. Ses encouragements et ses conseils précieux ont été une source d'inspiration pour nous, nous poussant à donner le meilleur de nous-mêmes. Nous tenons également à exprimer nos sincères remerciements aux membres du jury pour leur intérêt envers notre projet en acceptant d'examiner notre travail et de le renforcer avec leurs propositions. Le succès de ce projet n'aurait pas été possible sans le soutien et la contribution de tous ceux que nous avons cités. Nous leur exprimons notre profonde gratitude et leur reconnaissons le mérite de leur implication.

Table des matières

Liste des Abréviations 6
0.1 Contexte : L'émergence des Data Lakes dans le domaine du Big Data 8
0.2 Problématique : Les défis liés à l'analyse des données non structurées et semi-structurées 8
0.3 Objectifs et méthodologie 9
Chapitre 1: Comprendre les Data Lakes et leur architecture **10**
1.1 Définition et caracteristiques des data lakes : 10
1.2 Différence entre data lake et data warehouse : 10
1.3 Les types de données dans un data lake : 12
1.3.1 Données brutes : 12
1.3.2 Données enrichies : 13
1.4 L'architecture des data lakes : 13
1.4.1 Composants clés : 13
1.4.2 Gouvernance des données dans les Data Lakes : 14
1.4.3 Exemples d'architectures 15
1.5 Enjeux de la gestion de données dans les data lakes : 17
1.5.1 Problématique de qualité de données : 17
1.5.2 Problémes de Sécurité des données : 17
1.6 Conclusion : 18
Chapitre 2: Techniques d'analyse des données dans les Data Lakes **19**
2.1 Types d'analyse des données 19
2.2 Les étapes de l'analyse de données 21
2.2.1 Prétraitement des données 21
2.2.2 Modélisation des données 23
2.2.3 Visualisation des données 23
2.2.4 Interprétation 24
2.2.5 Conclusion 24
Chapitre 3: Applications de l'analyse de données **25**
3.1 Secteur de la santé 25
3.2 Secteur E-business 25
3.3 Secteur des technologies 26
3.4 Secteur de la vente au détail 26

3.5 Conclusion : 27
Conclusion générale 28
Bibliographie 28

Table des figures

1.1 Différence entre data lake et data warehouse[4] 12
1.2 Multi-ponds architecture[15] 15
1.3 Multi-zones architecture[15] 16
1.4 Architecture of Amazon Web Services (AWS)[15] 17

Liste des abréviations

BI	Business Intelligence
API	Application Programming Interface
ETL	Extract,Transform,Load
ELT	Extract,Load,Transform
HDFS	Hadoop Distributed File System
JSON	JavaScript Object Notation
CSV	Comma-Separated Values
XML	eXtensible Markup Language
YAML	YAML Ain't Markup Language
ML	Machine Learning
RGPD	Règlement général sur la protection des données

Rèsumé

Avec l'essor du Big Data, les Data Lakes sont apparus comme une solution flexible face aux limites des entrepôts de données traditionnels. Ils permettent de stocker de grandes quantités de données structurées, semi-structurées et non structurées dans un même espace, sans schéma préalable. Cependant, cette flexibilité entraîne des défis, notamment pour l'analyse des données non structurées (ex. : vidéos, textes, logs), la gouvernance, la qualité des données et leur traitement en temps réel.

Ce rapport vise à explorer les méthodes d'analyse adaptées aux Data Lakes, en s'appuyant sur une étude des architectures, des outils (comme Spark, Hadoop, Tableau), et des approches analytiques.

Introduction :

0.1 Contexte : L'émergence des Data Lakes dans le domaine du Big Data

Avec l'émergence du Big Data, les entités doivent gérer une répartition massive de données en termes de volume, de diversité et de rapidité. Dans ce cadre, les systèmes conventionnels comme les data warehouses ont du mal à satisfaire la demande grandissante en matière de stockage et d'analyse. C'est dans ce contexte que les Data Lakes prennent toute leur importance, proposant une option de stockage de données à la fois flexible et évolutive. Les Data Lakes permettent de centraliser toutes les données, qu'elles soient structurées, semi-structurées ou non structurées, dans un même espace. Contrairement aux entrepôts de données qui nécessitent un schéma rigide avant l'intégration des données, les Data Lakes suivent une approche "schema-on-read", permettant d'explorer et de transformer les données à la demande. Cela le rend particulièrement adaptable aux environnements modernes où la diversité des sources de données est une constante. Par exemple, une entreprise de commerce électronique peut stocker dans un Data Lake des données de transactions structurées, des journaux d'activité semi-structurés et des avis clients non structurés sous forme de texte ou d'audio. Cette capacité à intégrer divers types de données sans transformation préalable est l'une des principales forces des Data Lakes.

0.2 Problématique : Les défis liés à l'analyse des données non structurées et semi-structurées

Cependant, l'utilisation des Data Lakes n'est pas sans défis. Parmi les problèmes majeurs se trouve la gestion des données non structurées et semi-structurées, qui représentent

souvent une grande partie des données collectées. Ces données, telles que les vidéos, les

images, les fichiers logs ou les tweets, manquent de structure définie, rendant leur organisation et leur analyse complexes.

En outre, l'absence de schéma rigide dans les Data Lakes, bien qu'étant un avantage, pose des problèmes pour la gouvernance des données. Les entreprises doivent s'assurer que les données sont accessibles, traçables et sécurisées. Les questions de qualité des données, telles que les doublons ou les valeurs manquantes, sont également amplifiées dans les Data Lakes où

les données brutes sont conservées.

Les défis techniques, comme le traitement en temps réel des flux de données ou l'indexation des données pour une récupération rapide, compliquent encore davantage leur utilisation. Par exemple, dans un cas d'utilisation tel que l'analyse des médias sociaux, l'extraction d'informations pertinentes à partir de millions de messages non structurés est une tâche ardue.

0.3 Objectifs et méthodologie

Face à ces défis, ce rapport a pour objectif d'explorer les techniques d'analyse de données adaptées aux Data Lakes. Nous chercherons à comprendre comment surmonter les obstacles liés à l'organisation et à l'analyse des données non structurées et semi-structurées. Pour cela, nous adopterons une méthodologie en trois volets :

— Une revue approfondie des concepts et architectures des Data Lakes, afin de comprendre leurs caractéristiques et leurs différences avec les entrepôts de données.

— Une analyse des techniques modernes pour le prétraitement, l'analyse et le traitement des données non structurées, en mettant l'accent sur des outils tels qu'Apache Spark, Hadoop, et Tableau.

Chapitre 1

Comprendre les Data Lakes et leur architecture

1.1 Définition et caracteristiques des data lakes :

Un data lake est une solution d'analyse de big data qui offre aux data scientists, analystes et professionnels de la BI la possibilité de :

- Collecter et stocker des données brutes de formats variés (structurées, semi-structurées, non structurées) en provenance de sources internes ou externes, tout en conservant leur format d'origine.
- Traiter ces données afin de les adapter aux besoins évolutifs de l'entreprise.
- Exploiter les données stockées pour réaliser des analyses statistiques, du reporting, des requêtes, de la visualisation ou encore du Machine Learning.
- Assurer la gestion des données en veillant à leur qualité, leur sécurité et leur cycle de vie[15]. ou

1.2 Différence entre data lake et data warehouse :

L'utilisation des data lakes et des data warehouses pour le stockage des données présente à la fois des avantages et des inconvénients. Chaque méthode adopte une approche différente pour gérer et analyser les données. Le data lake permet de traiter tous les types de données(structurées, semi-structurées et non structurées). Il stocke les données dans leur état brut, sans traitement ni transformation. Cette approche offre un avantage majeur permettant de conserver une quantité massive de données en un seul endroit, comprises celles issues des réseaux sociaux, des messages réseau, des fichiers journaux, des images et des vidéos. De plus, une grande partie des données des entreprises sont non structurées, ce qui les rend difficiles à gérer avec un data warehouse. L'architecture joue un rôle clé dans cette distinction. La conception plate du data lake facilite l'ajout et la suppression de données. En revanche, dans un data warehouse, ces opérations sont plus complexes, car les données sont regroupées dans des data marts

sous une forme prédéfinie, avec des métadonnées et une identification spécifiques. Grâce à des métadonnées riches et bien organisées, la gestion des données dans un data lake devient plus fluide et efficace. Cela offre une grande flexibilité et permet d'ajuster facilement l'utilisation des informations en fonction des besoins.

Il existe également une différence dans les méthodes de gestion et de traitement des données :

1. **Schema-on-Write (Schéma à l'écriture) :** l'architecture data warehouse utilise un processus ETL qui extrait et transforme les données avant de les charger dans un format structuré, c'est-à-dire que le traitement se fait avant le stockage.
   * **Avantages :** Requêtes rapides et optimisées.
   * **Inconvénient :** Difficile de gérer des données imprévues ou non structurées.
2. **Schema-on-read(Schéma à la lecture) :** l'architecture data lake utilise un processus ELT qui extrait et charge les données sous leur forme brute, puis les transforme au fur et à mesure des requêtes.
   * **Avantages :** Grande flexibilité.
   * **Inconvénient :** Difficile de gérer des données imprévues ou non structurées.

Le coût joue un rôle économique très important pour les entreprises. Bien que le data lake soit conçu pour résoudre les problèmes du data warehouse, son coût est généralement inférieur, car il élimine les dépenses liées au matériel (hardware)[9].

Phase de données	Métriques	Lac de données	Entrepôts de données
Ingestion	Sources de données	Tous	Externe / opérationnelle
	Types de données	structurées / traitées / non structurées / semi-structurées / brutes	Structurées /Traitées
	Métadonnées	Oui	Optionel
	Gouvernance de données	Nécessite une approche axée sur les métadonnées	sécurité des données facile à contrôler
	Vitesse de chargement	Rapide	Lent
	Transformation	Transformé lorsque les données sont prêtes à être utilisé.	Transformées avant le chargement des données
	Méthode de chargement	*Late binding*	*Early binding*
Stockage	Coûte de stokage	À faible coût	Coût modéré
	Système de stockage	Hadoop,SQL,NoSQL,Relationelle	SQL/Relationelle
	Modelisation	A la volée	Etoil ou flacon
	Volume de données	Approche Store-All pour d'énormes volumes de données	volumes de données modérés
	Securité	En développement	Pleinement développé
	Durée de stockage	Toutes les données qu'il pourrait utiliser à l'avenir pour toujours	Durée déterminée
Traitement	Type de traitement	*Temps réel / Batch*	*Batch*
	Agilité	Haut	Bas
	Schema	Schema-On-Read	Schema-On-Write
	Temps de traitement	Court	Complex
	Traitement de donnees	ELT	ETL

FIGURE 1.1 – Différence entre data lake et data warehouse[4]

1.3 Les types de données dans un data lake :

Il y a 2 catégories de données dans un Data Lake :

1.3.1 Données brutes :

Les données brutes sont organisées sous formes de quatre catégories :

— **Données structurées :** des données sous forme de bases de données relationnelles (SQL), fichiers CSV, JSON bien formé, par exemple : données clients, transactions ban- caires, inventaire.

— **Données non structurées :** : des données sous forme de texte brut, images, vidéos, fichiers audio, documents PDF, par exemple : emails, posts sur les réseaux sociaux, images

— **Données semi structurées :** des données sous forme de XML, JSON, YAML..., par exemple : logs d'applications, données issues d'API, fichiers de configuration.

— **Données en flux (Streaming Data) :** des données générées en temps réel, par exemple :

capteurs IoT, logs serveurs, données de marché financier.

1.3.2 Données enrichies :

Leur type initial est brut, mais elles sont nettoyées, normalisées ou traitées.

1.4 L'architecture des data lakes :

1.4.1 Composants clés :

1. **Stockage : gestion des données brutes**
   Le stockage est la pierre angulaire d'un Data Lake. Il permet de conserver des données brutes dans leur format d'origine, sans transformation préalable. Contrairement aux Data Warehouses, où les données sont structurées avant d'être stockées, un Data Lake accepte des données brutes, ce qui offre une grande flexibilité et pour cela il y a plusieurs technologies utulisées :
   — **HDFS (Hadoop Distributed File System) :** Un système de fichiers distribué conçu pour stocker de grands volumes de données sur des clusters.
   — **Amazon S3 :** Un service de stockage cloud qui offre une grande évolutivité et une durabilité élevée.
   — **Azure Data Lake Storage** : Une solution de stockage optimisée pour les charges de travail Big Data.
   ‘

2. **Ingestion : flux de données en temps réel et en batch**
   L'ingestion des données est le processus qui consiste à collecter et à importer des données dans le Data Lake. Elle peut se faire de deux manières principales :
   — **Ingestion en temps réel (streaming) :** Les données sont collectées et traitées au fur et à mesure qu'elles sont générées. Cela est essentiel pour les applications
   nécessitant une analyse immédiate, comme la surveillance des réseaux ou les transactions financières.
      — **Outils utilisés :**
      -Apache Kafka : C'est une plateforme distribuée de messagerie à haut débit, conçue pour la publication-souscription, le traitement de flux, la collecte de logs et la surveillance en temps réel[13].
      - Apache Flink : C'est une plateforme open-source puissante conçue pour le traitement distribué de données, capable de gérer efficacement à la fois les flux en continu et les traitements par lots et une solution flexible et performante pour les besoins d'analyse modernes[14].
   — **Ingestion par lots (batch) :** Les données sont collectées pendant une période donnée, puis traitées en une seule fois.

3. **Traitement :**
   Le traitement des données dans un Data Lake implique la transformation, le nettoyage et l'enrichissement des données brutes pour les rendre exploitables. Apache Spark est l'un des outils les plus populaires pour cette tâche, grâce à sa capacité à traiter de grands volumes de données en mémoire.
   - **Fonctionnalités clés de Spark :**
     - Traitement distribué pour une haute performance.
     - Prise en charge de plusieurs langages (Scala, Python, Java).
     - Intégration avec des sources de données variées (HDFS, S3, bases de données).
   - **Autres outils de traitement :**
     - **Apache Hive :** Pour exécuter des requêtes SQL sur des données stockées dans un Data Lake.
     - **Presto :** Un moteur de requêtes distribué pour des analyses interactives.

1.4.2 Gouvernance des données dans les Data Lakes :

1. **Gestion des métadonnées :**
   Les métadonnées sont des informations qui décrivent les données stockées dans le Data Lake. Elles jouent un rôle crucial pour la traçabilité et la facilité d'utilisation des données.
   - **Outils de gestion des métadonnées :**
     - **Apach Atlas :** c'est un outil de gestion des métadonnées pour Hadoop. Il permet de typifier et classer les métadonnées, de suivre l'origine des données (Data Lineage) pour la gestion des métadonnées et la traçabilité des données, de rechercher des sources de données et de sécuriser l'accès aux données en masquant certaines informations [7].
     - **AWS Glue Data Catalog :** Le AWS Glue Data Catalog est un système centralisé qui stocke toutes les métadonnées de vos données provenant de différentes sources. Il permet d'accéder facilement aux informations sur les formats de données, les schémas et les sources. Quand un travail AWS Glue ETL est lancé, il utilise ce catalogue pour obtenir des informations sur les données et s'assurer qu'elles sont bien transformées[11].

   - **Avantages :**
     - Facilite la recherche et l'accès aux données.
     - Améliore la compréhension des données pour les utilisateurs.
2. **Sécurité et contrôle d'accès :** La sécurité des données est un enjeu majeur dans les Data Lakes. Des mécanismes de contrôle d'accès et de chiffrement doivent être mis en place pour protéger les données sensibles[10].
   - **Mesures de sécurité :**
     - Chiffrement des données.

— Surveillance des accès et des activités.
— Mise en place de politiques de sécurité strictes.

1.4.3 Exemples d'architectures

Après avoir défini ce qu'est un lac de données, il est essentiel d'aborder son architecture. Cette section propose des exemples d'architectures de lac de données :

— **Multi-ponds architecture :** Cette architecture garantit l'ingestion de différents types de données structurelles.

La classification en trois types de données (analogiques, applicatives et textuelles) permet de retrouver les données plus rapidement et de faciliter leur analyse. Cependant, certains types de données ne sont pas pris en compte, comme les images, les vidéos et les fichiers audio.

De plus, dans ce lac de données, une fois que les données brutes sont traitées et envoyées vers d'autres réservoirs, elles ne restent plus dans le réservoir de données brutes[15].

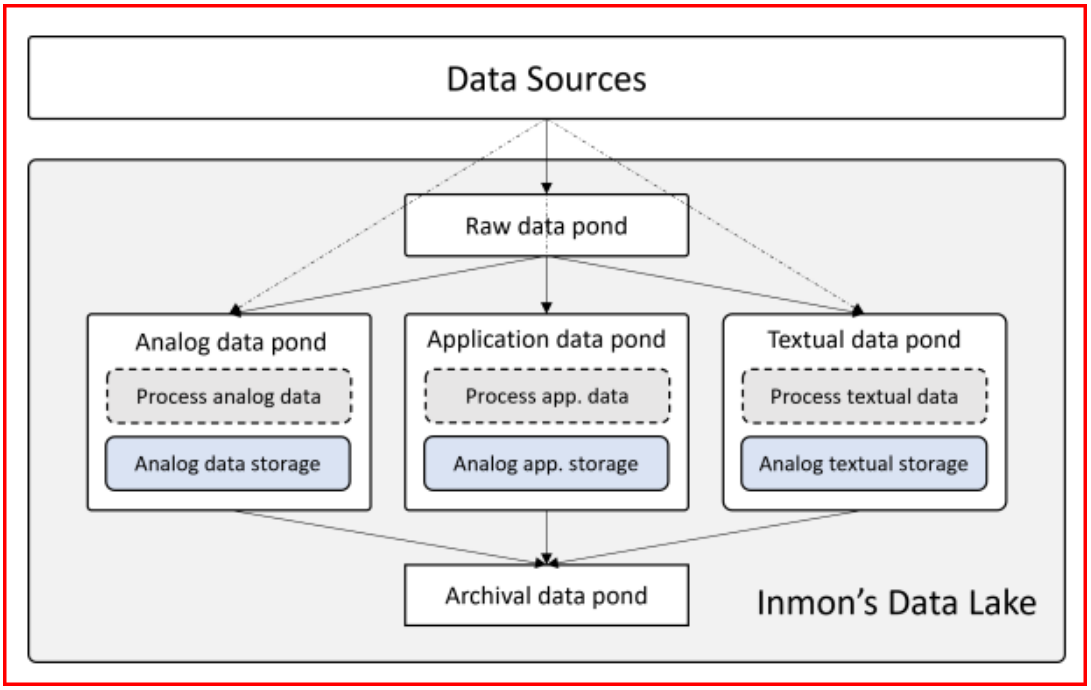


FIGURE 1.2 – Multi-ponds architecture[15]

— **Multi-zones architecture :** L'architecture indique les différentes fonctions d'un lac de données : ingestion, stockage, exploration, analyse et visualisation des données.

Cependant, cette architecture permet aux utilisateurs de personnaliser la collecte des données avant la phase d'ingestion et en dehors du lac de données, ce qui rend difficile la récupération de la source des données.

De plus, l'absence de gouvernance des données dans le lac de données entraîne l'impossibilité d'assurer la sécurité et le contrôle de la qualité des données[15].

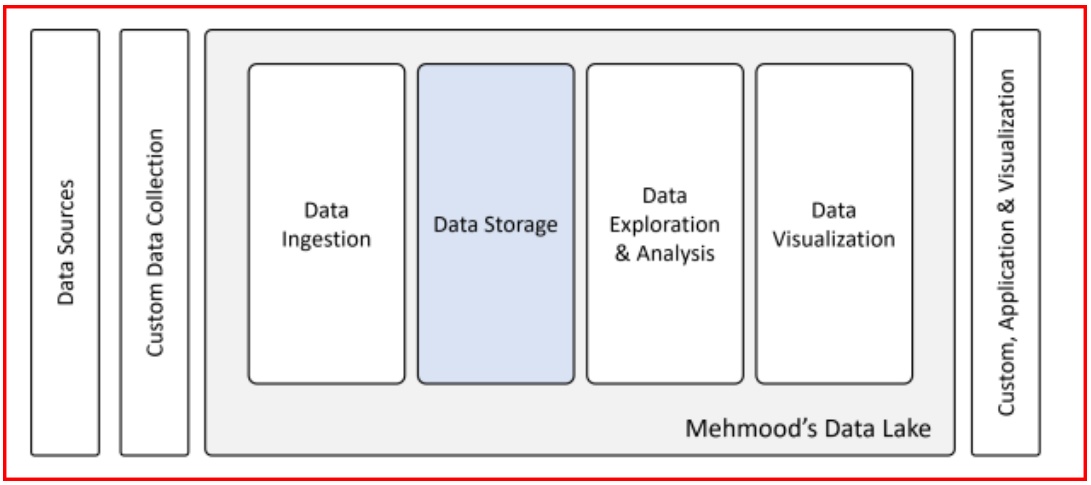


FIGURE 1.3 – Multi-zones architecture[15]

— **Architecture of Amazon Web Services (AWS) :** Cette architecture présente clairement le flux de données dans un lac de données. Toutes les données sont ingérées et centralisées en vue de leur analyse.

Cependant, le flux de données est unidirectionnel dans ce schéma. Les données brutes sont ingérées et stockées dans le stockage central, mais aucune donnée issue des processus de transformation n'est créée ni stockée dans ce même stockage central.

Néanmoins, les données nettoyées et transformées peuvent être directement réutilisées par les utilisateurs afin d'améliorer l'efficacité du traitement des données[15].

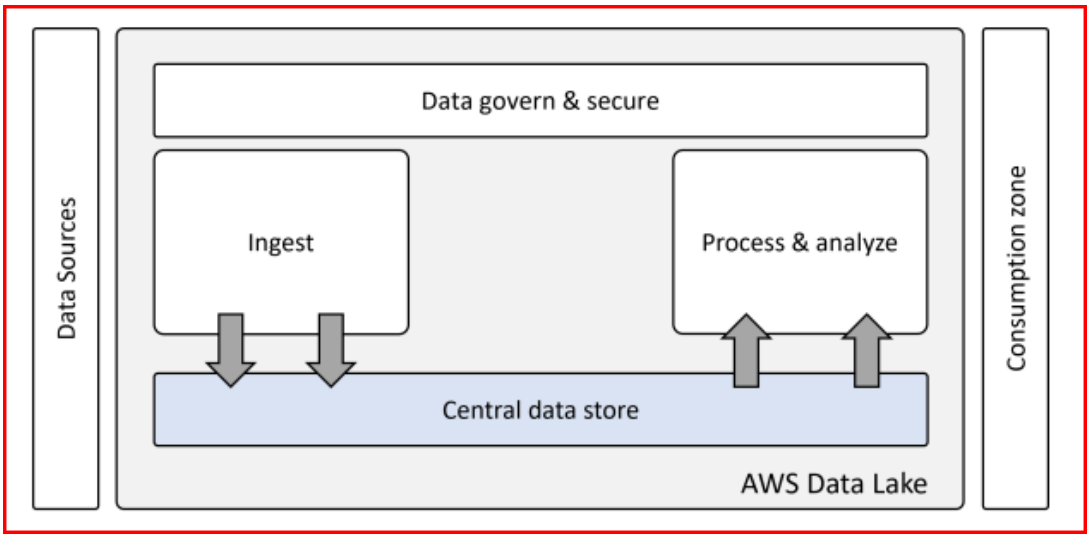


FIGURE 1.4 – Architecture of Amazon Web Services (AWS)[15]

1.5 Enjeux de la gestion de données dans les data lakes :

1.5.1 Problématique de qualité de données :

Les problèmes de qualité des données en Data Lake sont amplifiés par la volumétrie et l'hétérogénéité, rendant les méthodes traditionnelles inadaptées. Trois formes d'hétérogénéité existent : technologique (différents outils et infrastructures), des schémas (modèles et représentations variés) et au niveau des instances (valeurs contradictoires entre sources).

Les données s'organisent en quatre niveaux : sources de données, relations multiples, relations simples et attributs. L'intégration des données doit surmonter ces hétérogénéités pour garantir une qualité optimale[12].

1.5.2 Problémes de Sécurité des données :

La sécurité des données est un enjeu critique, en particulier pour les données sensibles ou réglementées. Les Data Lakes doivent être protégés contre les accès non autorisés, les fuites de données et les cyberattaques lors de chaque phase du cycle de vie Big Data[12]. Dans cette

partie, nous expliquons comment la sécurité des données est assurée à chaque étape de leur cycle de vie dans un lac de données :

A. **Collection de données :** Cette phase consiste à récupérer des informations provenant de sources diverses. Il est essentiel de s'assurer que les données proviennent de sources fiables et que leur transfert est sécurisé. Les attaques comme le phishing, le spoofing, et le spamming sont des techniques utilisées par les pirates pour collecter des données confidentielles. Des pratiques de sécurité et de sensibilisation sont nécessaires pour contrer ces menaces.

B. **Stockage de données :** Une fois collectées, les données doivent être stockées en toute sécurité. Les systèmes de stockage Big Data actuels répondent aux exigences de performance, mais ne garantissent pas toujours la sécurité. Des menaces comme le vol de disques, l'accès non autorisé et les attaques basées sur l'exploration de données peuvent affecter la sécurité des données stockées, notamment sur le Cloud, où l'emplacement physique des données et les accès ne sont pas toujours garantis.

C. **Analyse de données :** L'analyse des données permet d'extraire des informations importantes. Le chiffrement est une méthode efficace de sécurisation, mais l'énorme volume de données dans le Big Data rend cette tâche plus complexe. Les menaces incluent l'exploration non autorisée des données, la ré-identification des individus à partir des ensembles de données et les erreurs résultant de mauvaises corrélations entre les données.

D. **Création de connaissances :** Les informations extraites peuvent influencer les décisions des entreprises. Ces données sensibles sont susceptibles de faire l'objet d'attaques, comme le phishing ou l'usurpation d'identité. De plus, la vie privée des individus peut être menacée, notamment lors de l'analyse de données personnelles. Les réglementations comme le RGPD exigent la protection et parfois la destruction des données sensibles après leur utilisation, mais ce défi reste souvent.

1.6 Conclusion :

L'architecture des data lakes repose sur trois principaux composants : le stockage, l'ingestion et le traitement des données. Le stockage permet de conserver des données brutes, offrant une grande flexibilité avec des technologies comme HDFS, Amazon S3, et Azure Data Lake Storage. L'ingestion des données peut se faire en temps réel (streaming) ou par lots (batch), utilisant des outils tels qu'Apache Kafka et Apache Flink. Le traitement des données est réalisé à l'aide d'outils comme Apache Spark, permettant de transformer et analyser efficacement de grands volumes de données.

Chapitre 2

Techniques d'analyse des données dans les Data Lakes

2.1 Types d'analyse des données

L'analyse des données se décline en quatre types principaux : descriptive, explicative, prédictive et prescriptive, chacun apportant une approche spécifique pour exploiter les données.

1. **Analyse descriptive :**
   L'analyse descriptive constitue la première étape de l'exploration des données. Elle vise à résumer et à caractériser les données dans leur forme brute. Cette approche repose sur le calcul de statistiques fondamentales, telles que la moyenne, la médiane, l'écart-type, ainsi que l'analyse de la distribution des valeurs. Son objectif principal est de :

   — Fournir une vue d'ensemble précise et cohérente des données afin de mieux comprendre leur structure et leurs caractéristiques principales.
   — Identifier les tendances générales, détecter les valeurs extrêmes et repérer les anomalies.
   — Mettre en évidence les relations potentielles entre les différentes variables.

   Bien que l'analyse descriptive soit indispensable pour appréhender les données, elle ne permet pas d'expliquer les causes profondes des phénomènes observés. Par ailleurs, elle ne peut être utilisée isolément pour prédire des événements futurs ou pour élaborer des recommandations stratégiques. De ce fait, elle est souvent accompagnée d'analyses prédictives et prescriptives pour offrir une compréhension plus approfondie et orientée vers l'action.

2. **Analyse prédictive :**
   L'analyse prédictive est une approche d'analyse de données qui repose sur l'utilisation d'algorithmes pour anticiper des événements ou comportements futurs. Son objectif principal est de fournir des informations stratégiques aux entreprises et organisations, afin

de les aider à prendre des décisions éclairées. Les modèles prédictifs sont développés à l'aide d'algorithmes de machine learning tels que la régression linéaire, les arbres de décision ou les réseaux neuronaux. Parmi les objectifs visés, on retrouve notamment :

— La prévision des ventes à venir.
— La détection des comportements frauduleux.
— La prédiction des attitudes des consommateurs.
— La prévision des défaillances d'équipement.

Malgré sa puissance, l'analyse prédictive présente certaines limites. Étant donné que ces modèles se fondent sur des données historiques, leurs prévisions peuvent devenir moins précises si les conditions futures diffèrent de celles observées dans le passé. De plus, des biais dans les données ou dans les algorithmes employés peuvent fausser les résultats et affecter la fiabilité des prédictions.

3. **Analyse prescriptive :**
   L'analyse prescriptive est une approche avancée qui cherche à fournir des recommandations pratiques permettant d'optimiser les résultats dans un contexte spécifique. Elle utilise des modèles algorithmiques sophistiqués pour analyser différentes options et suggérer les actions les plus adaptées pour atteindre les objectifs fixés.
   L'objectif principal de l'analyse prescriptive est d'améliorer les résultats, de faciliter la prise de décisions automatisées et d'anticiper les effets des diverses actions. Parmi ses contraintes, on retrouve la nécessité de disposer de données de haute qualité, la complexité des modèles à mettre en œuvre et la difficulté de considérer l'intégralité des facteurs pertinents pouvant influencer les décisions.

4. **Analyse exploratoire ou explicative :**
   L'analyse explicative est un type d'analyse de données qui s'efforce de comprendre les liens de causalité entre différentes variables. Contrairement à l'analyse descriptive, qui se limite à décrire les données, l'analyse explicative cherche à clarifier les raisons pour lesquelles certains phénomènes se produisent de la manière observée.
   L'objectif principal de l'analyse explicative est d'identifier les relations causales entre les variables et de comprendre comment ces dernières influencent une variable cible. Cela inclut la capacité à prédire les valeurs futures d'une variable en se basant sur les données passées et présentes d'autres variables.
   Bien que l'analyse explicative soit un outil précieux, elle présente certaines limitations. Elle peut ne pas tenir compte de tous les facteurs pertinents et peut ne pas saisir toute la complexité du phénomène étudié. De plus, des biais dans les données peuvent altérer les conclusions tirées de l'analyse. Enfin, les modèles employés pour réaliser cette analyse ne sont pas toujours généralisables à tous les contextes et peuvent nécessiter des ajustements pour mieux s'adapter aux spécificités de chaque situation.

2.2 Les étapes de l'analyse de données

2.2.1 Prétraitement des données

Le prétraitement des données est une étape clé dans l'analyse des données. Il consiste à transformer les données brutes en un format structuré et exploitable pour l'analyse. Ce processus garantit que les données sont correctement préparées et prêtes pour les étapes suivantes, telles que l'exploration, la modélisation et l'interprétation[3]. Les étapes de prétraitement sont :

1. **Découvrir les données :** L'analyse exploratoire des données (AED) est une étape clé visant à révéler les informations cachées dans les données. Développée par John Tukey dans les années 70, elle est souvent comparée à un travail de détective, où l'analyste cherche à dégager une histoire plausible à partir des données. L'AED utilise des méthodes quantitatives (moyenne, médiane, variance, écart-type) et graphiques (histogrammes, boîtes à moustaches) pour mieux comprendre la distribution, la tendance centrale et la dispersion des variables. Cette approche permet d'identifier des tendances, des anomalies et de poser des hypothèses en vue d'analyses plus approfondies.
2. **Structurer les données :** La structuration des données est une étape essentielle de leur nettoyage et de leur préparation, visant à les organiser pour une analyse statistique spécifique. Cette phase permet de corriger les éventuelles irrégularités ou anomalies susceptibles d'affecter la fiabilité des modèles d'analyse. Il est donc crucial d'examiner en détail le processus de nettoyage des données afin d'assurer des résultats plus précis et pertinents.
3. **Le nettoyage des données :** Le nettoyage des données est essentiel pour assurer la qualité de l'analyse. Les neuf conseils suivants abordent, à l'aide d'exemples pratiques, une série de problèmes couramment rencontrés lors du nettoyage des données[6] :
   (a) **La vérification de l'orthographe**
   (b) **Suprimer les doublants**
   (c) **Trouver et remplacer :** Grâce à des remplacements ciblés, il est possible de nettoyer les données de manière efficace et de les structurer de façon pertinente, en mettant en évidence certaines tendances ou répétitions.
   (d) **Majuscules et minuscules**
   (e) **Espaces et caractères qui ne s'impriment pas**
   (f) **Chiffres et signes**
   (g) **Date et Heure**
   (h) **Fusionner et Fractionner les colones**
   (i) **Données d'un sous-ensemble** Conserver seulement les colones/rangées et observations d'intérêt.
4. **Enrichir les données :** Un jeu de données peut parfois comporter des valeurs manquantes, ce qui peut compromettre la qualité des réponses apportées à la question de recherche. Il

peut alors être nécessaire de rechercher d'autres sources de données afin de les fusionner pour compléter l'information.[6].

5. **Valider les données :** La validation des données implique de vérifier les types des variables et leur format, détecter les valeurs aberrantes via des fonctions d'agrégation, et s'assurer de la cohérence des données, notamment géographiques, pour éviter toute erreur d'analyse.

Parmi les outils utilisés lors du prétraitement des données, on peut citer :

— **Apache Hadoop :** c'est une plateforme open source développée par la fondation Apache, utilisant Java pour la gestion et le traitement de grandes quantités de données, qu'elles soient structurées ou non. Conçu pour fonctionner au sein d'un environnement distribué, Hadoop permet de stocker ces données sur des clusters, offrant ainsi une solution adaptée à la gestion de volumes massifs et variés. Elle résout les défis associés à la volumétrie, la diversité et la complexité des données en fournissant un espace de stockage évolutif et une capacité de traitement élevée. Hadoop est également capable de gérer simultanément un grand nombre de tâches, qu'elles soient analytiques ou de traitement. Son efficacité et sa flexibilité en font un choix privilégié pour des géants du web tels que Twitter, LinkedIn, eBay et Amazon, qui l'utilisent pour exploiter leurs énormes ensembles de données et optimiser leurs opérations[14].

— **Apache Spark :** Apache Spark est une plateforme open source de traitement distribué, spécialement conçue pour les charges de travail liées au big data. Grâce à la mise en cache en mémoire et à l'optimisation des requêtes, il permet d'effectuer des analyses rapides sur des données de toutes tailles. Spark propose des API pour le développement en Java, Scala, Python et R, et permet la réutilisation du code pour différentes tâches, telles que le traitement par lots, les requêtes interactives, l'analyse en temps réel, l'apprentissage automatique et le traitement des graphes. Ce système est utilisé par des entreprises de divers secteurs, telles que FINRA, Yelp, Zillow, DataXu, Urban Institute et CrowdStrike[1].

— **Databricks :** Databricks est une plateforme cloud dédiée à la gestion et à l'analyse de vastes ensembles de données grâce au moteur open source Apache Spark. Elle propose un espace de travail unifié permettant aux data scientists, ingénieurs et analystes métier de collaborer, concevoir et déployer des applications basées sur les données. Conçue pour simplifier et optimiser le traitement des big data, Databricks offre des outils et services pour la préparation des données, l'analyse en temps réel et l'apprentissage automatique. Parmi ses fonctionnalités clés figurent la prise en charge de divers formats de données, l'intégration avec les principales bibliothèques et frameworks de data science, ainsi qu'une scalabilité flexible adaptée aux besoins des utilisateurs[8].

* **Remarques :**

(a) La base de Databricks repose sur Apache Spark, orienté vers les architectures modernes de Data Lake, et ajoute des services qui facilitent le travail avec les big data pour les utilisateurs.

(b) Le choix de l'un de ces outils dépend de deux critères : le traitement en temps réel ou le stockage et le traitement de grandes quantités de données. Pour le traitement en temps réel, on utilise Databricks ou Apache Spark, en fonction de l'architecture, en choisissant l'un ou l'autre selon l'architecture. Pour le stockage et le traitement de grandes quantités de données, le choix se porte sur Apache Hadoop.

2.2.2 Modélisation des données

Une fois les données nettoyées et préparées, elles sont soumises à un processus de modélisation visant à mettre en lumière des tendances et des liens significatifs. Pour cela, on mobilise diverses méthodes statistiques telles que la régression, la classification ou encore le clustering. Ces analyses permettent ensuite de tirer des enseignements et de faire émerger des relations pertinentes à partir des modèles construits.

— **La régression :**
Cette méthode est utilisée pour comprendre la relation entre une variable dépendante et une ou plusieurs variables indépendantes, en vue de créer un modèle permettant de prédire la variable dépendante à partir des variables indépendantes.

— **L'analyse de clusters :**
Cette méthode est utilisée pour explorer des ensembles de données de même type ou présentant des caractéristiques similaires, en regroupant les données corrélées au sein de clusters.

— **La classification :**
Cette méthode permet de regrouper les objets en différentes catégories en fonction de leurs caractéristiques, et d'utiliser le modèle pour estimer la valeur de la variable dépendante à partir des variables indépendantes[2].

2.2.3 Visualisation des données

Il s'agit d'une étape qui permet à l'utilisateur de comprendre plus facilement les résultats, et de diagnostiquer les tendances ainsi que les relations à partir de nouvelles représentations des données, telles que des graphiques ou des tableaux. Chaque type de visualisation possède ses propres avantages et s'adapte à des besoins d'analyse spécifiques :

— **Les graphiques à barres :** permettent de comparer efficacement différents clusters de données.

— **Les graphiques en ligne :** facilitent l'exploration des tendances dans les données au cours du temps.

— **Les diagrammes circulaires :** offrent une visualisation claire des statistiques pour chaque catégorie[2].

2.2.4 Interprétation

Après la visualisation et la compréhension des données, l'étape suivante consiste à prendre des décisions éclairées, soit dans le but d'optimiser les gains, soit pour vérifier la qualité et la justesse des résultats obtenus lors des étapes précédentes, en les comparant à d'autres données de référence[2].

2.2.5 Conclusion

Une fois les données visualisées et analysées, il devient nécessaire de prendre des décisions afin d'optimiser les résultats ou de valider la qualité et l'exactitude des conclusions tirées lors des étapes précédentes, en les confrontant à des données de référence supplémentaires[2].

Chapitre 3

Applications de l'analyse de données

3.1 Secteur de la santé

L'utilisation des données dans le domaine médical transforme progressivement la manière dont les soins sont prodigués. En s'appuyant sur des informations précises issues de dossiers patients, les établissements de santé, qu'ils soient des hôpitaux, des cliniques ou des acteurs de l'industrie pharmaceutique, peuvent affiner leurs diagnostics, adapter les traitements et améliorer globalement la qualité des soins. Cette approche permet également de mieux comprendre les mécanismes des maladies, leurs évolutions et les réactions des patients face à différentes thérapies, ouvrant ainsi la voie à la mise au point de médicaments innovants et à l'optimisation des traitements déjà existants.

L'analyse des données joue aussi un rôle fondamental dans la détection de tendances médicales et l'identification de schémas récurrents liés aux maladies. En croisant un large éventail d'informations provenant de dossiers médicaux électroniques, d'enquêtes de santé ou encore d'essais cliniques, les professionnels de santé peuvent mieux cerner les origines des pathologies, identifier les facteurs de risque et anticiper leur propagation. Les technologies d'intelligence artificielle, en particulier les algorithmes de machine learning, renforcent ces capacités prédictives en s'appuyant sur les données passées pour estimer les risques futurs pour un patient donné.

La crise du COVID-19 a mis en exergue l'importance stratégique de cette approche. Face à l'urgence sanitaire mondiale, les gouvernements et les institutions médicales ont intensifié la collecte et l'analyse de données afin de suivre la dynamique de l'épidémie, comprendre les profils les plus vulnérables et ajuster les pratiques cliniques en fonction des résultats observés. Ces analyses ont contribué à éclairer les choix politiques, à orienter les plans d'action sanitaires et à renforcer les mesures de protection des populations[2].

3.2 Secteur E-business

Les entreprises s'appuient de plus en plus sur l'analyse de données pour améliorer leur prise de décision, explorer de nouveaux marchés et enrichir l'expérience client. Dans un environne-

ment commercial très concurrentiel, se démarquer devient essentiel pour rester en tête.

Les mégadonnées jouent un rôle clé à chaque étape du cycle de vente : qu'il s'agisse d'anticiper les tendances produits, d'estimer la demande ou encore d'optimiser la gestion des points de vente. Grâce à ces volumes massifs d'informations, les entreprises sont en mesure de repenser leurs stratégies et de trouver des approches innovantes pour répondre aux attentes du marché[5].

3.3 Secteur des technologies

Dans le secteur des technologies, l'analyse de données est devenue un levier essentiel pour améliorer les produits et les services proposés aux utilisateurs. En recueillant des informations sur les comportements et les préférences des usagers, les entreprises technologiques peuvent affiner leurs offres, concevoir de nouvelles solutions et perfectionner celles déjà existantes.

Par exemple, les entreprises spécialisées dans la mobilité exploitent les données pour optimiser les trajets de livraison, améliorer la fluidité des transports et proposer des services plus efficaces. Dans le domaine des technologies de l'information, l'analyse des données permet de mieux cerner les menaces en cybersécurité et de renforcer la protection des systèmes informatiques. Quant à l'industrie du jeu vidéo, elle s'appuie sur les données des joueurs pour ajuster les mécaniques de jeu et offrir une expérience plus immersive et personnalisée.

Les géants du numérique, comme Google, Apple, Facebook, Amazon et Microsoft regroupés sous l'acronyme GAFAM ont très tôt misé sur la puissance des données. Ils utilisent l'analyse de données pour perfectionner leurs services, qu'il s'agisse d'algorithmes de recommandation, de publicité ciblée ou d'interfaces utilisateur intelligentes. Google, par exemple, adapte son moteur de recherche en fonction des requêtes les plus fréquentes, tandis qu'Amazon personnalise les suggestions d'achat en fonction des habitudes de navigation et d'achat de ses clients.

Plus récemment, des entreprises comme OpenAI exploitent l'analyse de données à grande échelle pour entraîner des modèles avancés en traitement du langage naturel, comme ChatGPT. Ces technologies trouvent des applications variées : traduction automatique, reconnaissance vocale ou encore assistance intelligente.

En somme, l'analyse de données s'impose comme un pilier central dans le développement de produits technologiques innovants, permettant aux entreprises de rester à la pointe et de répondre au plus près des attentes des utilisateurs[2].

3.4 Secteur de la vente au détail

Dans le secteur du commerce, l'exploitation des données permet aux entreprises de mieux comprendre les comportements d'achat, d'évaluer la performance de leurs produits et d'affiner leurs stratégies marketing. En analysant les informations issues des achats effectués en magasin

ou en ligne, les enseignes peuvent identifier les tendances de consommation et ajuster leur offre en fonction des attentes du marché.

L'étude des habitudes d'achat joue un rôle clé dans cette démarche. En observant des données telles que les produits les plus souvent achetés, la fréquence d'achat ou encore les canaux utilisés (en ligne ou physique), les entreprises obtiennent une vision plus précise des préférences des consommateurs. Ces informations leur permettent d'orienter leurs décisions en matière de conception de produits, de mise en rayon ou encore de ciblage publicitaire.

Par ailleurs, l'analyse des ventes et des retours clients offre un aperçu précieux sur la performance des produits. Elle aide à identifier les articles qui rencontrent le plus de succès ainsi que ceux qui peinent à séduire les clients, facilitant ainsi les ajustements en production, en distribution ou même en communication.

En ce qui concerne le marketing, les données collectées permettent de mesurer l'efficacité des campagnes publicitaires et de mieux allouer les ressources. Les entreprises peuvent repérer les canaux les plus performants, ajuster leurs messages en fonction du public cible et saisir de nouvelles opportunités de croissance. L'objectif est clair : concevoir des actions marketing plus pertinentes, plus rentables et plus personnalisées[2].

3.5 Conclusion :

Avec l'augmentation exponentielle des données, leur analyse devient essentielle pour les entreprises afin d'améliorer la prise de décisions, anticiper les tendances et identifier de nouvelles opportunités. Ce rapport a exploré comment la collecte et l'exploitation des données permettent aux entreprises de mieux comprendre leurs marchés et leurs clients, grâce à des outils comme le machine learning et les systèmes de recommandation.

L'avenir de l'analyse des données réside dans l'intégration de diverses sources et dans l'accès à distance aux services, pour une prise de décision plus précise et plus rapide.

Conclusion générale

Ce rapport a permis de mettre en lumière les principaux enjeux liés à l'utilisation des Data Lakes dans le cadre du Big Data. Tout en offrant une flexibilité sans précédent dans le stockage de données structurées, semi-structurées et non structurées, les Data Lakes posent des défis considérables, notamment en matière de gestion de données non structurées, de gouvernance, de qualité et de traitement en temps réel.

L'analyse des architectures et des outils spécifiques comme Apache Spark, Hadoop et Tableau a démontré l'importance d'adopter des approches adaptées pour tirer parti de la richesse des données stockées dans un Data Lake. Ces outils, combinés à des techniques de prétraitement et d'analyse modernes, permettent de transformer des données brutes en informations utiles, tout en surmontant les obstacles liés à leur organisation et à leur analyse.

Enfin, bien que les Data Lakes apportent des solutions innovantes et évolutives, leur gestion et leur exploitation efficace nécessitent des stratégies rigoureuses pour assurer la qualité et la sécurité des données, tout en garantissant leur accessibilité. Les entreprises qui adoptent cette technologie devront donc investir dans des systèmes robustes de gouvernance des données et dans des outils d'analyse adaptés pour en maximiser les avantages.

Les perspectives d'avenir pour les Data Lakes, notamment avec les progrès en matière de traitement en temps réel et de gestion des données non structurées, ouvrent de nouvelles avenues pour les organisations cherchant à exploiter pleinement le potentiel du Big Data.

Bibliographie

[1] Amazon Web Services. Amazon web services (aws) - cloud computing services, 2025. URL https://aws.amazon.com/. Accessed : 2-April-2025.

[2] Analytics. L'analyse des données : définition et types, 2025. URL https://analytics.fr/analyse-de-donnees/definition-analyse-donnees/#:~:text=A.-,L'analyse%20descriptive,type%2C%20la%20distribution%2C%20etc. Consulté le 3 avril 2025.

[3] Astera. Data preprocessing, 2025. URL https://www.astera.com/fr/type/blog/data-preprocessing/. Accessed : 2025-04-02.

[4] M. Derakhshannia. *Gestion et optimisation de l'architecture logistique de lacs de données*. PhD thesis, Université de Montpellier, 2022. URL https://theses.hal.science/tel-03851629/. Consulté le 4 avril 2025.

[5] M. B. Diop, P. O. Dieye, and J. Barthelemy. Fouilles de données massives. 2024. URL https://www.researchgate.net/profile/Mamadou-Diop-21/publication/380347633_Fouilles_de_donnees_massives/links/663704393524304153635bca/Fouilles-de-donnees-massives.pdf.

[6] eCampusOntario. Nettoyage de données dans le processus de gestion des données de recherche, 2025. URL https://ecampusontario.pressbooks.pub/gdrcanada/chapter/nettoyage-de-donnees-dans-le-processus-de-gestion-des-donnees-de-recherche/. Consulté le 10 avril 2025.

[7] A. S. Foundation. Apache atlas, 2025. URL https://fr.wikipedia.org/wiki/Apache_Atlas. Accessed : 2025-04-02.

[8] GeeksforGeeks. Introduction to databricks, 2025. URL https://www.geeksforgeeks.org/introduction-to-databricks/. Accessed : 2-April-2025.

[9] I. Hassan. Storage structures in the era of big data : from data warehouse to lakehouse. *Journal of Theoretical and Applied Information Technology*, 102(6), 2024. URL https://www.jatit.org/volumes/Vol102No6/16Vol102No6.pdf.

[10] C. Madera. *L'évolution des systèmes et architectures d'information sous l'influence des données massives : les lacs de données*. PhD thesis, Université Montpellier, 2018. URL https://theses.hal.science/tel-02138983/document.

[11] A. W. Services. Aws glue data catalog, 2025. URL https://docs.aws.amazon.com/prescriptive-guidance/latest/serverless-etl-aws-glue/aws-glue-data-catalog.html. Accessed : 2025-04-02.

[12] M. Talha. *Big Data entre Qualité & Sécurité de données*. PhD thesis, Université Cadi Ayyad (Marrakech, Maroc), 2022. URL https://theses.hal.science/tel-04011007/document.

[13] D. Vohra and D. Vohra. Using apache kafka. *Kubernetes Microservices with Docker*, pages 347–371, 2016. URL https://link.springer.com/chapter/10.1007/978-1-4842-1907-2_12.

[14] S. Yessad, S. Boussoufa-Lahlah, W. Ferroudj, K. Djellaoui, and L. Bouallouche-Medjkoune. Architecture d'analyse de big data basée sur l'apprentissage profond pour le stationnement intelligent : Étude et comparaison des outils, 2023. URL https://hal.science/hal-04219657/document. Consulté le 4 avril 2025.

[15] Y. Zhao. *Metadata management for data lake governance*. PhD thesis, Université Toulouse 1 Capitole, 2021. URL https://publications.ut-capitole.fr/44574/1/ZhaoYan2021.pdf. Consulté le 3 avril 2025.