Titre proposé

"Intégration du Data Mesh et de l'Apprentissage Fédéré dans une Plateforme de Gestion des Marchandises pour l'Amélioration de l'Efficacité Logistique et de la Confidentialité des Données"

Introduction

Au cours des dernières décennies, l'architecture des données a subi une transformation significative, passant d'un modèle centralisé à une approche plus distribuée et flexible. Cette évolution a été motivée par la prolifération des données, l'émergence de nouvelles technologies et la nécessité croissante pour les organisations de tirer parti de leurs données pour prendre des décisions plus éclairées. Dans ce contexte, l'émergence de la Data Mesh s'est révélée être une réponse innovante aux défis de gestion des données auxquels sont confrontées de nombreuses entreprises.

La Data Mesh représente une rupture par rapport aux approches traditionnelles en matière d'architecture des données en mettant l'accent sur la distribution de la propriété des données et la responsabilité de leur gouvernance au sein des différentes unités organisationnelles, appelées domaines de données. Cette approche vise à surmonter les limitations des architectures centralisées en favorisant une gestion plus agile et décentralisée des données, tout en facilitant la collaboration et la prise de décision à l'échelle de l'organisation.

Cependant, malgré les avantages potentiels offerts par la Data Mesh, des défis subsistent en ce qui concerne l'analyse des données dans un tel environnement distribué. En particulier, la nécessité de protéger la confidentialité des données tout en permettant une analyse efficace et collaborative demeure un enjeu majeur. C'est dans ce contexte qu'émerge la problématique de recherche de cette thèse : comment intégrer l'apprentissage fédéré dans le paradigme de la Data Mesh pour permettre une analyse de données décentralisée et respectueuse de la vie privée ?

L'objectif principal de cette thèse est de créer des meilleures pratiques pour la mise en œuvre du concept de data mesh au sein d'une organisation. Pour ce faire, plusieurs objectifs sont définis. Il s’agit d'explorer les possibilités d'intégration de l'apprentissage fédéré dans la Data Mesh, en concevant des méthodes et des techniques pour former des modèles d'apprentissage automatique de manière distribuée, tout en préservant la confidentialité des données. Ensuite, cette recherche vise à développer des stratégies pour sécuriser les produits de données intermédiaires générés dans le cadre de cette approche, afin de réduire les risques potentiels de fuites de données. Enfin, elle cherche à évaluer les avantages et les limites de cette approche hybride en termes d'efficacité, de sécurité et de respect de la vie privée. Cette évaluation se fera à travers des expérimentations pratiques et des études de cas réels pour une plateforme de gestion des marchandises englobant les domaines logistique, transport et assurance.

Motivation

La gestion des marchandises, qu'elles soient terrestres, maritimes ou aériennes, est confrontée à des défis croissants en termes de complexité logistique, de gestion des données, et de protection de la confidentialité des informations sensibles. Les approches traditionnelles de centralisation des données peuvent entraîner des problèmes de scalabilité, de gestion des silos de données, et de risques pour la confidentialité. L'intégration de concepts modernes tels que le data mesh et le federated learning peut offrir des solutions innovantes pour surmonter ces défis.

Etat de l’art

La revue de littérature sur les architectures des données, leurs limites et les principes clés de la Data Mesh met en lumière l'évolution du concept de Big Data depuis son introduction, ainsi que les défis rencontrés par les organisations dans la gestion et l'analyse de volumes de données toujours plus importants et complexes. Initialement caractérisé par le modèle des 3Vs (Volume, Variété, Vélocité), ce concept a évolué vers le modèle des 5Vs pour mieux représenter les défis actuels du traitement des données. En réponse à ces défis, des concepts tels que les entrepôts de données et les lacs de données ont émergé, fournissant des solutions pour stocker, gérer et analyser des quantités massives de données [1, 2]. Cependant, les approches centralisées de ces architectures posent des problèmes de scalability et de gouvernance, ce qui conduit à l'émergence de nouveaux paradigmes comme la Data Mesh [3].

La Data Mesh est une architecture de données émergente qui prône la distribution des responsabilités et de la propriété des données au sein des différentes unités organisationnelles, appelées domaines de données. Elle permet de surmonter les limitations des architectures centralisées en favorisant une gestion plus agile et décentralisée des données, tout en facilitant la collaboration et la prise de décision à l'échelle de l'organisation. Les principes clés de la Data Mesh incluent la propriété des données par les domaines, la conceptualisation des données comme des produits, une plateforme de données en libre-service et une gouvernance computationnelle fédérée [4].

La littérature sur la data mesh fournit des principes généraux pour développer une architecture de data mesh, tels que la conception axée sur les domaines, la réflexion produit, la réflexion plateforme et la gouvernance computationnelle. Cependant, il n'existe pas d'outils et de méthodologies détaillés (validés empiriquement) pour appliquer ces principes afin de développer et de faire fonctionner une data mesh. Les organisations auraient également besoin de conseils et d'outils pour migrer systématiquement leurs architectures de données héritées vers une data mesh tout en évaluant les coûts et les avantages [5]. D’autres défis subsistent pour les scientifiques des données travaillant au sein des équipes de domaine. En particulier, l'application efficace des techniques d'apprentissage automatique à des données spécifiques au domaine pour produire des produits de données de haute qualité reste un défi majeur. Dans le contexte de l'architecture de données décentralisée de la Data Mesh, où chaque équipe de domaine possède ses propres données sans nécessité de les partager avec d'autres équipes, des méthodes comme l'apprentissage fédéré peuvent être explorées pour permettre la collaboration tout en préservant la confidentialité des données.

Les avancées récentes dans le domaine de l'apprentissage fédéré ont considérablement enrichi les possibilités offertes par cette approche pour l'analyse de données distribuée. L'apprentissage fédéré est une méthode d'apprentissage automatique qui encourage l'entraînement de modèles sur un large réseau de nœuds indépendants et décentralisés. Dans le contexte de la Data Mesh, ces nœuds correspondent à la diversité des domaines où les données résident naturellement. Cette méthodologie est étroitement alignée avec les principes fondamentaux de la Data Mesh, offrant ainsi une multitude d'avantages et en faisant un choix approprié pour les applications d'apprentissage automatique au sein de cette structure distribuée [6].

Un avantage clé de l'apprentissage fédéré est sa compatibilité avec la philosophie de la propriété décentralisée des données spécifique à chaque domaine, un aspect fondamental du modèle de la Data Mesh [7]. Contrairement aux problèmes associés à la copie des données dans l'apprentissage centralisé, l'apprentissage fédéré permet aux données de rester dans leur domaine d'origine tout au long de la phase d'apprentissage. Cette pratique réduit efficacement le besoin de duplication et de transfert des données, ce qui permet de résoudre les inefficacités associées et les risques potentiels

pour l'intégrité des données. De plus, l'apprentissage fédéré renforce le rôle des propriétaires de domaine dans le processus d'apprentissage automatique. En entraînant des modèles au sein de leurs domaines respectifs, les propriétaires de domaine peuvent exercer un contrôle et fournir des contributions au processus d'apprentissage. Cela améliore potentiellement la qualité et la pertinence des modèles, tout en respectant le principe de données en tant que produit, assurant ainsi que les données sont gérées et organisées dans leur contexte de domaine.

En outre, l'apprentissage fédéré adresse les préoccupations liées à la confidentialité et à la sécurité, qui sont souvent inhérentes à l'apprentissage centralisé. En maintenant les données dans leur domaine pendant l'entraînement du modèle, les données sensibles n'ont pas besoin d'être exposées à une autorité centrale, réduisant ainsi le risque de violations de données et de violations de la confidentialité. Ces avancées ont ouvert la voie à trois types distincts d'apprentissage fédéré : l'apprentissage fédéré horizontal, l'apprentissage fédéré vertical et l'apprentissage fractionné (Split Learning). Chacune de ces méthodes présente des caractéristiques uniques qui pourraient potentiellement être bénéfiques dans une architecture de données distribuée [8].

Méthodologie

L'approche méthodologique proposée dans cette thèse repose sur l'intégration de l'apprentissage fédéré dans le cadre de la Data Mesh, visant à permettre une analyse de données décentralisée et respectueuse de la vie privée. Cette approche vise à combiner les avantages de la Data Mesh en matière de distribution des responsabilités et de la propriété des données avec les capacités de l'apprentissage fédéré pour entraîner des modèles d'apprentissage automatique de manière décentralisée.

- Intégration de la Data Mesh :

La Data Mesh constitue le cadre architectural dans lequel cette recherche s'inscrit. L'approche de la Data Mesh repose sur la distribution de la propriété des données et des responsabilités de gouvernance aux différents domaines de données au sein de l'organisation. Dans le cadre de cette méthodologie, les principes de la Data Mesh seront intégrés pour organiser et structurer les données au niveau des domaines spécifiques. Chaque domaine de données sera considéré comme un nœud dans le réseau de l'apprentissage fédéré, où les données sont gérées localement et sont accessibles aux équipes de domaine pour l'entraînement des modèles. L'intégration de la Data Mesh, pour une plateforme de gestion des marchandises, permettra une gestion agile des données, tout en facilitant la collaboration et la gouvernance au niveau de chaque domaine.

- Utilisation de l'apprentissage fédéré :

L'apprentissage fédéré est une approche prometteuse pour l'analyse de données distribuée, permettant l'entraînement de modèles sur des données réparties sur un large réseau de nœuds indépendants. Dans le contexte de la Data Mesh, chaque nœud correspond à un domaine de données spécifique, où les données sont gérées localement par les équipes de domaine. L'utilisation de l'apprentissage fédéré permet aux équipes de domaine de former des modèles sur leurs propres données sans avoir besoin de les partager avec d'autres domaines ou de les centraliser, ce qui garantit la confidentialité et la sécurité des données. Cette approche favorise également la responsabilité et la gouvernance locales des données, en alignant étroitement les modèles d'apprentissage sur les besoins spécifiques de chaque domaine. Il s’agit donc de mettre en œuvre des algorithmes de federated learning pour améliorer les modèles de prédiction et d'optimisation logistique tout en préservant la confidentialité des données.

- Sécurisation des produits de données intermédiaires :

La sécurité des produits de données intermédiaires générés dans le cadre de cette approche est d'une importance capitale pour prévenir les risques potentiels de fuites de données et de violations de la vie privée. Pour ce faire, plusieurs méthodes et techniques peuvent être envisagées.

- Test et Evaluation

Il s'agit à la fin d'évaluer l'impact de l'intégration du data mesh et du federated learning sur l'efficacité logistique, la qualité des modèles, et la sécurité des données.

Résultats attendus

Les résultats anticipés de cette recherche sont multiples et visent à contribuer significativement au domaine de l'analyse de données dans un contexte décentralisé, notamment au sein de la Data Mesh. Parmi les résultats attendus figurent :

1. **Création de meilleures pratiques pour l'intégration de la Data Mesh dans une organisation :** Cette recherche vise également à créer des meilleures pratiques pour l'intégration de la Data Mesh dans une organisation et particulièrement dans une plateforme de gestion des marchandises. Ces pratiques comprendront des lignes directrices pour la conception et la mise en œuvre d'une architecture de données décentralisée, en mettant l'accent sur la distribution des responsabilités et de la propriété des données, ainsi que sur la gouvernance computationnelle. Ces meilleures pratiques aideront les organisations à migrer de manière systématique vers une architecture de données décentralisée tout en évaluant les coûts et les avantages de cette transition.
2. **Développement de méthodes efficaces pour l'analyse de données décentralisée :** L'intégration de l'apprentissage fédéré dans la Data Mesh devrait permettre le développement de méthodes novatrices pour l'analyse de données réparties sur un large réseau de domaines. Ces méthodes devraient offrir des solutions efficaces pour entraîner des modèles d'apprentissage automatique tout en préservant la confidentialité et la sécurité des données. De nouvelles méthodes d'apprentissage fédéré seront développées et évaluées pour améliorer la prédiction et l'optimisation logistique.
3. **Amélioration de la protection des produits :** En utilisant des techniques de sécurisation des données, nous nous attendons à renforcer la protection de la confidentialité dont les données sont utilisées dans le cadre de cette recherche. Les résultats devraient démontrer que les méthodes proposées permettent de garantir un niveau élevé de confidentialité tout en permettant une analyse efficace des données distribuées.
4. **Validation des avantages de la Data Mesh :** Les résultats obtenus devraient également valider les avantages de la Data Mesh en tant qu'architecture de données décentralisée. En montrant comment cette approche peut être combinée avec l'apprentissage fédéré pour une analyse collaborative et sécurisée des données, nous espérons démontrer sa pertinence et son potentiel dans un large éventail de scénarios applicatifs.

Implications potentielles

Les résultats de cette recherche ont des implications potentielles importantes pour la pratique et la recherche future dans le domaine de l'analyse de données et de la sécurité des données. En mettant en avant les avantages en termes de sécurité des données et de protection de la vie privée, cette recherche pourrait contribuer à orienter les pratiques en matière de gestion et d'analyse des données

dans un contexte décentralisé. En effet, les résultats attendus pourraient transformer la manière dont les données sont gérées et utilisées dans l'industrie logistique, offrant des avantages tangibles en termes de performance et de protection des données. De plus, les méthodes et techniques développées pourraient servir de base à de futures recherches visant à explorer d'autres aspects de la Data Mesh et de l'apprentissage fédéré, ou à étendre leur application à d'autres domaines d'étude à savoir l'intégration de la blockchain. Enfin, les résultats de cette recherche pourraient également inspirer de nouvelles approches et initiatives visant à renforcer la confidentialité et la sécurité des données dans un monde de plus en plus axé sur les données.

Ce sujet de thèse offre une perspective interdisciplinaire qui peut intéresser à la fois les domaines de la science des données, de l'ingénierie des systèmes d'information, et de la logistique, en répondant à des défis concrets et actuels dans le secteur des marchandises.

Plan de Travail

Le plan de travail de cette recherche comprend les étapes suivantes :

1. Conception de l'Approche Méthodologique : Dans cette phase initiale, une revue approfondie de la littérature existante sur la Data Mesh, l'apprentissage fédéré et les meilleures pratiques pour migrer vers une architecture de données distribuée sera effectuée. Une approche méthodologique intégrant ces concepts de manière cohérente, en mettant l'accent sur l'intégration du Data Mesh dans l'organisation, la migration vers cette architecture et l'implémentation de l'apprentissage fédéré, sera définie particulièrement pour une plateforme de gestion des marchandises.

2. Collecte des Données : Une fois l'approche méthodologique définie, les données nécessaires à la recherche seront collectées. Cela pourrait inclure des données sur les infrastructures existantes, les processus organisationnels et les besoins métier, afin de comprendre les défis et les opportunités liés à l'intégration du Data Mesh dans l'organisation et à l'utilisation de l'apprentissage fédéré. Il y aura dans cette étape l'identification des domaines de données pertinents (ex : données de transport terrestre, données de transport maritime, données clients, etc.).

3. Implémentation des Méthodes : Les techniques d'intégration du Data Mesh dans l'organisation seront mises en œuvre, en développant des outils et des processus pour faciliter la migration vers cette architecture. Parallèlement, les infrastructures nécessaires à l'apprentissage fédéré seront mises en place, en concevant des systèmes permettant l'entraînement de modèles sur des données réparties de manière sécurisée et respectueuse de la vie privée. Spécifiquement, des algorithmes de federated learning adaptés à la plateforme seront implémentés. Des modèles seront entrainés localement dans chaque domaine et agrégation des modèles de manière sécurisée.

4. Évaluation des Résultats : Une fois un prototype de la plateforme intégrant le data mesh et le federated learning développé, les résultats obtenus seront évalués en termes d'efficacité, d'impact sur la sécurité et la confidentialité des données et aussi d'impact organisationnel et de satisfaction des parties prenantes. Les résultats seront analysés surtout par rapport aux gains en efficacité logistique. Les avantages et les défis de l'approche de migration vers une architecture de données distribuée seront analysés, en mettant en évidence les leçons apprises et les meilleures pratiques identifiées pour l'utilisation de l'apprentissage fédéré.

5. Rédaction de la Thèse : Enfin, la thèse sera rédigée en intégrant les résultats de la recherche, les discussions théoriques et les implications pratiques. Cette phase comprendra également la rédaction des conclusions et des recommandations pour les organisations souhaitant migrer vers une architecture de données distribuée basée sur le Data Mesh et l'utilisation de l'apprentissage fédéré pour une analyse de données sécurisée et respectueuse de la vie privée.

Conclusion :

La nécessité impérieuse pour les organisations d'adopter des approches novatrices telles que la Data Mesh réside dans les défis sans cesse croissants auxquels elles sont confrontées. La prolifération exponentielle des données, combinée à des impératifs de confidentialité et de sécurité toujours plus stricts, met en évidence la nécessité d'une révision fondamentale des méthodes de gestion et d'analyse des données existante. Les architectures traditionnelles peinent à répondre à ces nouveaux défis, souvent confrontées à des problèmes d'évolutivité, de gouvernance et de sécurité.

Dans ce contexte, la Data Mesh émerge comme une réponse pertinente et innovante. En distribuant la propriété des données et la responsabilité de leur gouvernance à travers des domaines de données autonomes, elle offre une approche agile et décentralisée, permettant une gestion plus efficace des données tout en préservant leur confidentialité et leur sécurité. Cependant, malgré son potentiel, la littérature existante sur la Data Mesh reste limitée, laissant un vide que cette thèse s'efforce de combler.

En mettant en lumière les défis auxquels les organisations sont confrontées et en soulignant l'insuffisance des approches actuelles, cette recherche démontre l'importance cruciale de repenser nos modèles de gestion des données. En proposant une intégration de l'apprentissage fédéré dans le paradigme de la Data Mesh, cette thèse offre une voie prometteuse pour relever ces défis, ouvrant ainsi de nouvelles perspectives pour une analyse de données distribuée, sécurisée et respectueuse de la vie privée.

Références

[1] Jaroslav Pokorný. Database Architectures: Current Trends and their Relationships to Requirements of Practice. Advances in Information Systems Development. 2007

[2] Athira Nambiar and Divyansh Mundra. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. Big Data and Cognitive Computing 6(4):132. 2022. DOI: 10.3390/bdcc6040132

[3] AWS. Qu'est-ce qu'un Data Mesh ? Disponible sur https://aws.amazon.com/fr/what-is/data-mesh/

[4] Anton Dolhopolov, Arnaud Castelltort and Anne Laurent. Implementing Federated Governance in Data Mesh Architecture. Future Internet 2024, 16(4), 115; https://doi.org/10.3390/fi16040115

[5] Nemania Borovits, Indika Kumara, Damian A. Tamburri and Willem-Jan Van Den Heuvel. Privacy Engineering in the Data Mesh: Towards a Decentralized Data Privacy Governance Framework. International Conference on Service-Oriented Computing. 2023

[6] Haoyuan Li and Salman Toor. Empowering Data Mesh with Federated Learning. ACM Knowledge Discovery and Data Mining; 25th - 29th August, 2024; Barcelona, Spain. isbn: 978-1-4503-XXXX-X/18/06. Available on : https://arxiv.org/html/2403.17878v2

[7] Ehsan Hallaji, Roozbeh Razavi-Far, Mehrdad Saif, Boyu Wang, Qiang Yang. Decentralized Federated Learning: A Survey on Security and Privacy. IEEE Transactions on Big Data. Apr. 2024, pp. 194-213, vol. 10. DOI: 10.1109/TBDATA.2024.3362191

[8] Jie Wen, Zhixia Zhang, Yang Lan, Zhihua Cui, Jianghui Cai & Wensheng Zhang. A survey on federated learning: challenges and applications. International Journal of Machine Learning and Cybernetics. Volume 14, pages 513–535, (2023)