Le Data Lakehouse

LIENARD, Thomas, thomas.lienard@thelio.fr THELIO, Bègles (Orateur)

ANGLADE, Christophe, christophe.anglade@databricks.com , DATABRICKS, Paris (Orateur)

Thématique :

Thème 2 : Structuration des données : ETL, Indexation, Données Structurées, Data Management, Algorithme d'Apprentissage, Classification par Machine Learning..

Résumé : ***Datalake, Datawarehouse, Datalakehouse, vers une convergence des solutions ?***

Mots clés : *Datalake, Datawarehouse, Datalakehouse, Lakehouse*

1. Introduction

Les Datawarehouses ont été développés à la fin des années 1980 et constituent un endroit idéal pour stocker des "données structurées". Ce sont des bases de données relationnelles conçues pour les requêtes et les analyses, et contiennent normalement des données historiques qui ont été extraites de données transactionnelles. Les Datalakes, en revanche, sont des espaces de stockage non relationnels, centralisés et consolidés pour les données brutes, telles que les données structurées, semi-structurées et non structurées.

L'utilisation originale du terme "data lakehouse" est attribuée à une entreprise appelée Jellyvision (un client de Snowflake). Snowflake a repris le nom, et l'a promu en 2017, en décrivant leurs efforts pour combiner le traitement des données structurées avec un système sans schéma. AWS a ensuite commencé à utiliser le terme pour décrire ses services de données et d'analyse "lake house architecture". L'un des principaux atouts de la data lakehouse est appelé une couche transactionnelle structurée, qui a été développée par Databricks en 2019.

2. Méthodologie

Dans un premier temps nous nous intéresserons à l'évolution des données depuis 20 ans, évolution qui justifie l'apparition de nouveaux paradigmes.

En analysant les différents outils du marché, nous verrons :

- A quelles problématiques répondent ces outils ?
- Quelles approches ont choisi les acteurs du secteur pour développer leurs solutions ?
- Comment donner du sens à ces modèles concurrents ?

- Pourquoi y a-t-il une telle disparité entre les approches ?
- Que choisir entre les bases et les référentiels open-source (spark/delta) et propriétaires (snowflake/relational) ?

3. Originalité / perspective

Au-delà de la comparaison des outils, toujours intéressante à faire, nous mettrons en avant les causes profondes de ces évolutions. Ceci permettra aux auditeurs d'avoir les clés pour mieux décrypter les discours marketing et se faire un avis éclairé sur les dernières évolutions.

PRESENTATION DATABRICKS :

Issue du milieu universitaire et de la communauté open source, Databricks a été fondée en 2013 par les créateurs d'Apache Spark™, Delta Lake et MLflow. La plateforme Lakehouse de Databricks combine le meilleur des data lakes et des data warehouses pour offrir la fiabilité, la gouvernance renforcée et les performances des data warehouses avec l'ouverture, la flexibilité et la prise en charge du machine learning des data lakes.

PRESENTATION THELIO :

THELIO, cabinet de conseil et d'intégration data, est né pour relever un défi : démystifier le monde de la data pour la rendre accessible à toutes les entreprises. Né à Bordeaux et continuant sa croissance à Lyon, THELIO mobilise ses consultants experts pour accompagner ses clients sur l'ensemble de la chaîne de valeur de la donnée, du cadrage de leur stratégie data à sa mise en œuvre.

https://www.thelio.fr/

ORATEURS :

Christophe Anglade,

Solution Architect chez Databricks, j'apporte mon expertise dans le design, l'architecture, l'implémentation et la mise à l'échelle de solutions Data.

J'ai travaillé pendant plus de 6 ans en tant que consultant dans différent domaines Data (BI, Data-Engineering, Architecte Data) et dans la construction de plateformes Cloud.

Aujourd'hui j'accompagne principalement de grands comptes internationaux dans l'élaboration de leur stratégie Data autour d'architectures Lakehouse.

https://www.linkedin.com/in/christopheanglade/

Thomas Lienard,

Data-Architect chez Thélio, j’accompagne nos clients dans la définition et la réalisation d’architectures Data dans un cadre agnostique.

Spécialisé sur le cloud-engineering de modern data plateformes, J’interviens aussi bien sur des projets de migration vers le cloud que sur des projets d’industrialisation et de structuration de la donnée.

linkedin Thomas LIENARD