

À propos d'Apache Iceberg

Vous méritez mieux que des CSV par FTP !

L'Évolution des Échanges de Données

Le CSV est depuis longtemps un standard en data engineering, tout comme Excel l'est pour les métiers. Couplé au FTP, il permet l'échange de fichiers de données entre différents systèmes. Cependant, cette méthode présente des limitations significatives qui impactent la qualité et l'efficacité des échanges.

Cas d'Usage

Prenons l'exemple d'une place de marché collectant des données sur le matériel agricole disponible à l'échange entre exploitants. Chaque heure, tous les partenaires transmettent leur inventaire à une plateforme centralisée.

Défis Actuels :

- **Absence de Standardisation** : Les adhésions ont eu lieu au fil du temps sans imposer un format commun.
- **Manque d'Analyse** : La plateforme n'est utilisée qu'à des fins d'inventaire instantané, sans analyse approfondie pour détecter des tendances ou des pannes.
- **Intégrité des Données** : Aucun moyen de garantir que les fichiers CSV parviennent intacts via FTP.

La Réalité Terrain

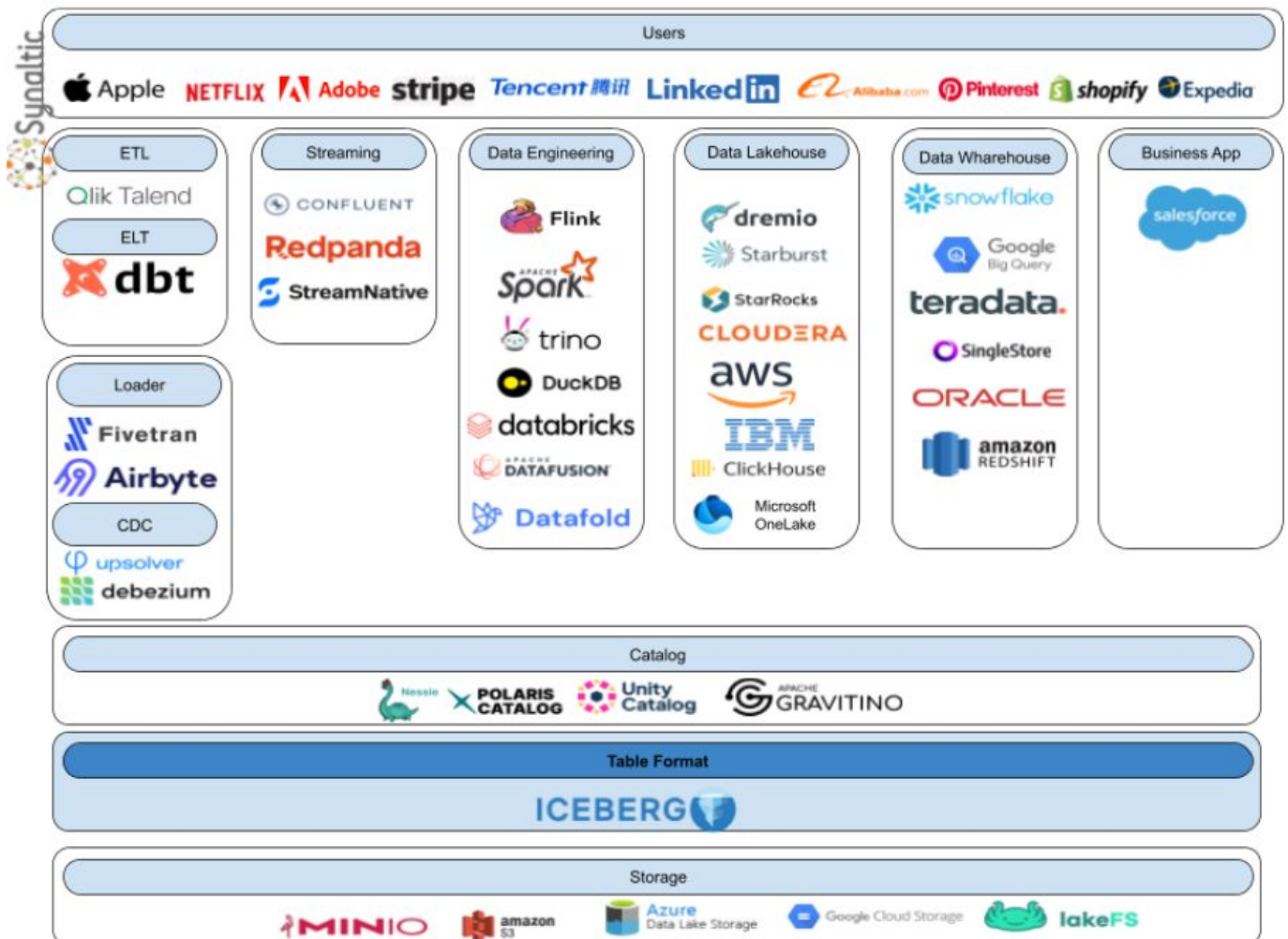
Cette situation est courante dans de nombreux secteurs, comme l'assurance. Les plateformes d'échange reposant sur des fichiers CSV sont légion, souvent pleines de fichiers inutilisables.

Pas de données, pas d'IA ! Sans qualité de données, que des hallucinations !

Sans qualité de données, on obtient des "hallucinations" plutôt que des informations fiables. À l'ère de l'IA générative, la piètre qualité des données entraîne des modèles peu performants.

Apache Iceberg : Une Révolution dans les Échanges de Données

Apache Iceberg s'impose comme un format universel, adoptant une standardisation qui améliore significativement la qualité des échanges de données. En passant des fichiers CSV aux échanges de données structurées, Apache Iceberg assure non seulement une meilleure standardisation mais aussi une intégrité et une gestion des données inégalées.



Échanger des Données : la Base de Notre Économie Moderne

L'échange de données est crucial dans notre économie actuelle. Aucune organisation n'est totalement autonome et dépend de données provenant de multiples sources pour fonctionner efficacement.



Charly Clairmont

Chez Synaltic, nous mettons un point d'honneur à utiliser des formats ouverts pour donner le contrôle total de leurs données à nos clients.

Apache Iceberg et le Catalogue de Données

Le format ouvert d'Apache Iceberg a démontré la nécessité de cataloguer les tables enrichies de métadonnées telles que les types et les statistiques. Ces métadonnées permettent aux moteurs de calcul de parcourir efficacement les fichiers, rendant le catalogue indispensable.

La gestion des données est en pleine révolution avec Apache Iceberg, offrant un format standard et un protocole d'accès universel compatible avec tout moteur de calcul.



Pourquoi les Catalogues sont Essentiels

Les catalogues tels que Nessie et Polaris renforcent l'importance des fonctionnalités suivantes :

Portabilité des Données et des Tables

Les catalogues enregistrent chaque table et l'emplacement de ses métadonnées, permettant une référence unique et évitant les duplications inutiles. Cela facilite la création, la référence et le partage sécurisé des données.

Accès Unique et Concurrent

Les catalogues assurent un contrôle de la concurrence, vérifiant les métadonnées avant et après chaque transaction pour maintenir les garanties ACID. Cela permet aux organisations d'utiliser le moteur de requête le plus efficace pour chaque cas d'usage.

Gouvernance Portable, Centralisée, Partagée

Les catalogues évoluent pour intégrer des règles d'accès cohérentes sur différents moteurs de requête. Nessie et Polaris permettent de définir ces règles une seule fois, simplifiant ainsi la gestion et renforçant la sécurité.

Versionnage du Catalogue

Nessie de Dremio introduit le concept de "versionnage du catalogue", capturant des instantanés de toutes les tables et leurs références, et supportant les commits comme Git. Cela facilite les transactions multi-tables, les restaurations et la création d'environnements sans duplication de données, contribuant à une gestion efficace et écologique des données.

Gestion de Lakehouse

Déployer un catalogue Apache Iceberg transforme les plateformes de données en véritables « gestionnaires de Lakehouse ». Le catalogue intégré dans Dremio offre des fonctionnalités avancées de gouvernance, d'optimisation des tables et de gestion des versions, simplifiant l'adoption et la migration vers Iceberg.

Flexibilité et Connectivité

Dremio fonctionne aussi bien dans des environnements cloud que sur site, se connectant de manière transparente aux lacs de données, aux entrepôts de données et aux bases de données. Cette flexibilité permet une offre d'analyse puissante et unifiée.

Avec Dremio, les entreprises peuvent tirer parti de leurs infrastructures existantes tout en intégrant de nouvelles technologies sans friction. La compatibilité de Dremio avec divers systèmes permet une intégration fluide, garantissant que les données restent accessibles et utilisables, quelles que soient les évolutions technologiques futures.

Mise en Œuvre Efficace

Déployer Dremio avec un catalogue Apache Iceberg permet de sécuriser l'accès aux données et de garder le contrôle total. Utilisez Apache Spark pour l'ingestion, Snowflake pour le calcul, et rendez les données accessibles à des outils comme Power BI grâce à Dremio.

Cette mise en œuvre simplifiée permet aux organisations de se concentrer sur l'analyse et l'exploitation des données sans se soucier des complications techniques. Les capacités de Dremio et d'Apache Iceberg assurent que les données sont traitées avec la plus grande précision, offrant ainsi des insights précieux et exploitables pour les décideurs.

Avantages Clés :

- **Scalabilité** : La solution s'adapte facilement à l'évolution des besoins en données.
- **Interopérabilité** : Fonctionne harmonieusement avec une vaste gamme de systèmes et de plateformes.
- **Sécurité** : Des contrôles d'accès robustes garantissent que les données sensibles sont protégées à chaque étape.

En optant pour Dremio et Apache Iceberg, les entreprises peuvent non seulement améliorer leur gestion des données mais aussi obtenir des avantages compétitifs significatifs grâce à une analyse plus rapide et plus précise.

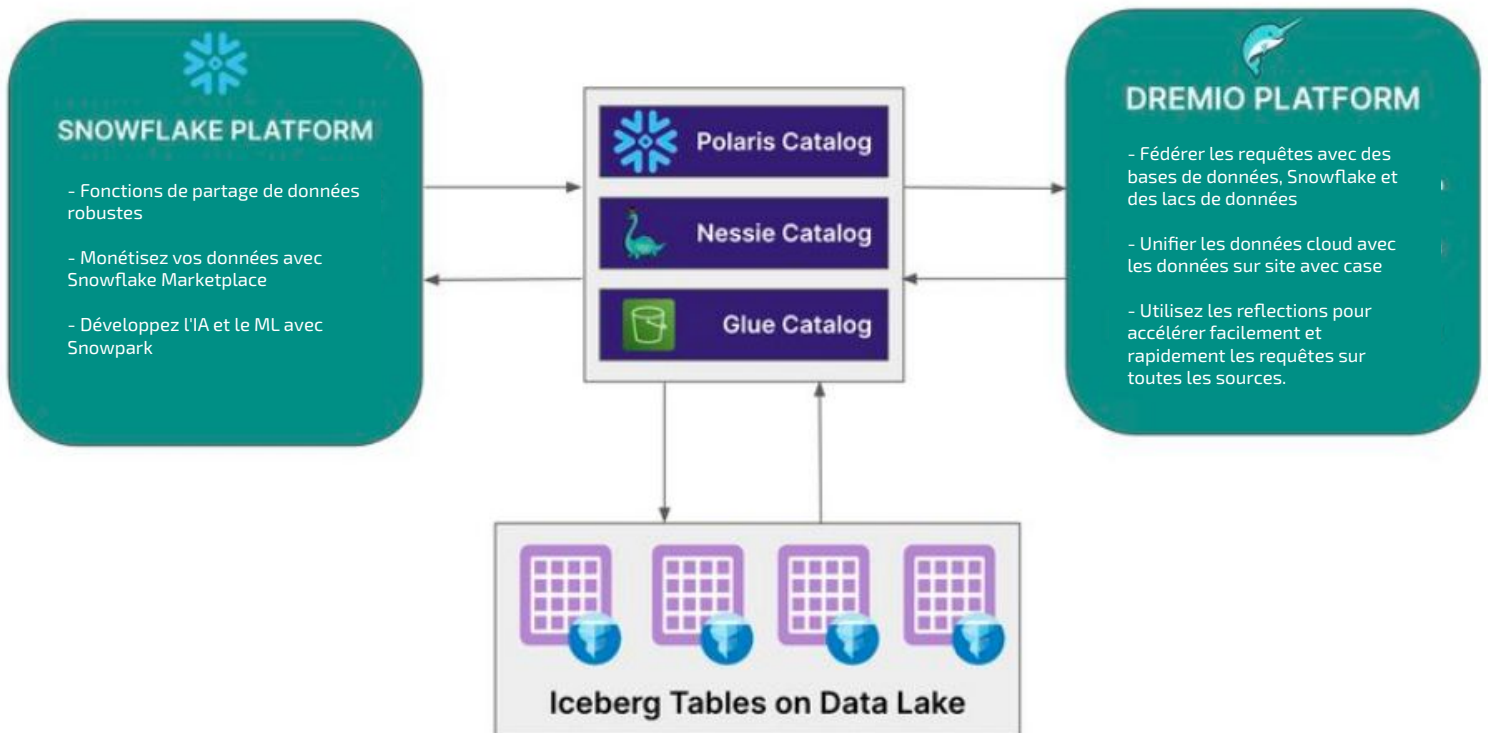
Et Notre Cas d'Usage dans Tout Ça ?

Transformer les organisations face aux changements actuels n'est pas simple. Chez Synaltic, nous nous positionnons comme des vulgarisateurs avant-gardistes, aidant les organisations à garder le contrôle de leurs données.

Pour une organisation collectant des données sur le matériel agricole :

- **Simplification et Sécurisation de l'Accès** : Fini l'échange de fichiers, place aux échanges de données structurées via un "Data Contract".
- **Évolution sans Altération** : Le schéma de la table peut évoluer sans altérer les données existantes.
- **Historisation et Analyse** : Les processus ETL bénéficient de l'historisation, permettant des analyses et des programmes de machine learning directement à partir des tables Apache Iceberg.

Ce scénario est réalisable dès aujourd'hui avec les technologies actuelles, offrant des solutions pratiques et modernes pour la gestion des données.



PARTENAIRE DREMIO DEPUIS 2021

Depuis 2004, Synaltic est un des spécialistes en Data Management et assure la promotion des alternatives innovantes, et particulièrement celles qui sont Open Source.

Après avoir signés les premiers clients français pour TALEND (2006) et TABLEAU (2012), Synaltic déploie, dès 2018, l'**Open Data Lakehouse Dremio** du journal *Le Point*.

En 2021, Synaltic devient le premier partenaire DREMIO en France et référence la solution au catalogue UGAP en 2024.

ACHAT INSTALLATION

Achat de la licence en EURO
+
Installation par les experts francophones de Synaltic

"ONBOARDING" FORMATION

3 étapes en 25 jours
+
3 modules de formation
+
20 heures d'expertise

SUPPORT EXPERTISE

Gestionnaire de compte francophone
+
SLA
+
Expertise "On Demand"

