

Solution “Data Lake House Engine”

Dremio Cloud ou
Dremio On-premise



DREMIO-SYNALTIC • Janvier 2024



Agenda

1. Présentation Synaltic
2. Pourquoi Dremio
3. Capacités techniques Dremio



The background is a solid orange color with a pattern of thin, white, wavy lines that create a sense of movement and depth. The lines are more concentrated on the left side and fade out towards the right.

1

SYNALTIC

Experts en Data Management, passionnés d'Open Source et d'Innovation !



19

20 ans d'expérience dans le **Data Management** et plus de **250 projets réalisés** en France et dans les pays limitrophes.

30

30 collaborateurs certifiés intervenant en France et dans les pays limitrophes, **contribuant aux communautés** d'utilisateurs.

14

Acteurs innovants par sa R&D, depuis plus de 14 ans, nous pratiquons le CIR : nous accueillons un doctorant et participons à des projets collaboratifs européens.



Synaltic, expert Dremio depuis 2018 en France



Sur ces quinze dernières années, les solutions répondant aux enjeux du big data n'ont cessé de bouleverser le marché. C'est l'open source et le cloud qui auront le plus apporté d'innovation dans cette ère nouvelle où la donnée est devenue le nouvel eldorado. Dans ce paysage, Dremio a fait un pari : apporter simplicité, collaboration, convivialité sans perdre de vue la performance attendue par les utilisateurs dans le cadre d'une interrogation ad hoc.

Ce sont pour toutes ses raisons que Synaltic fait très vite le choix de Dremio, se rapproche rapidement de l'éditeur et de la communauté, et forme ses collaborateurs dès 2018.

Consultant Dremio Expert

- 20 ans d'expérience.
- Java J2EE, Data Warehouse, SQL, Talend, MariaDB ColumnStore, Hadoop, Apache Kafka, Apache Airflow
- Dremio : Architecture, Cloud, Pont Hadoop, Déploiement Yarn, Déploiement Standalone, Accélération, Couches Sémantique, Apache Iceberg.

Consultant Dremio Confirmé

- 5 ans d'expérience
- Python, Machine Learning, SQL, Data Warehouse, Apache Kafka, Oracle, Snowflake
- Dremio : Couches Sémantiques, Accélération, Accompagnement client découverte et prise en main de Dremio

Formations Synaltic - Dremio

- Système Avancé de gestion de base de données, SQL, Talend, Tableau, Data Warehouse, Gestion agiles de projets de données, Data Engineering, Data Visualisation, Gouvernance, Sécurité
- Dremio essential (Dremio utilisateurs métier, Data Curation), Dremio Avancée (Gouvernance, Accélération)
- "Compagnonnage" autour des projets Dremio

Synaltic partage son expertise Dremio

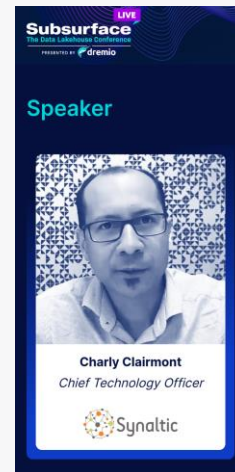
From Business Intelligence to Streaming Decisions

Subsurface - 2 Mars 2023

Retour sur 20 ans de data warehousing.

Mise en avant d'une architecture de livraison de données en continue vers un data lake S3 alimenté via Apache Kafka au format Apache Iceberg.

Dremio actualise en temps réel les jeux de données pour les livrer aux métiers.



OW2Con 2019, Paris Open Source Summit 2019, Big Data Paris 2020, Modern Data Stack France

The background is a solid orange color with a pattern of thin, white, wavy lines that create a sense of movement and depth. The lines are more concentrated on the left side and fade towards the right.

2

POURQUOI DREMIO

Au Sujet de Dremio

Le Data Lakehouse Facile et Ouvert

- **Fonctionnalités et performances type Data Warehouse** directement sur le Data Lake
- **Accélération des requêtes** pour éliminer les copies et les extraits de BI
- **Couche sémantique** pour permettre du libre-service gouverné
- **Multi-connecteurs** de base de données pour permettre des requêtes fédérées

Enterprise Adoption

- **+ de 1000 références** d'entreprises de tous les secteurs
- **5** des 10 entreprises du Fortune 10

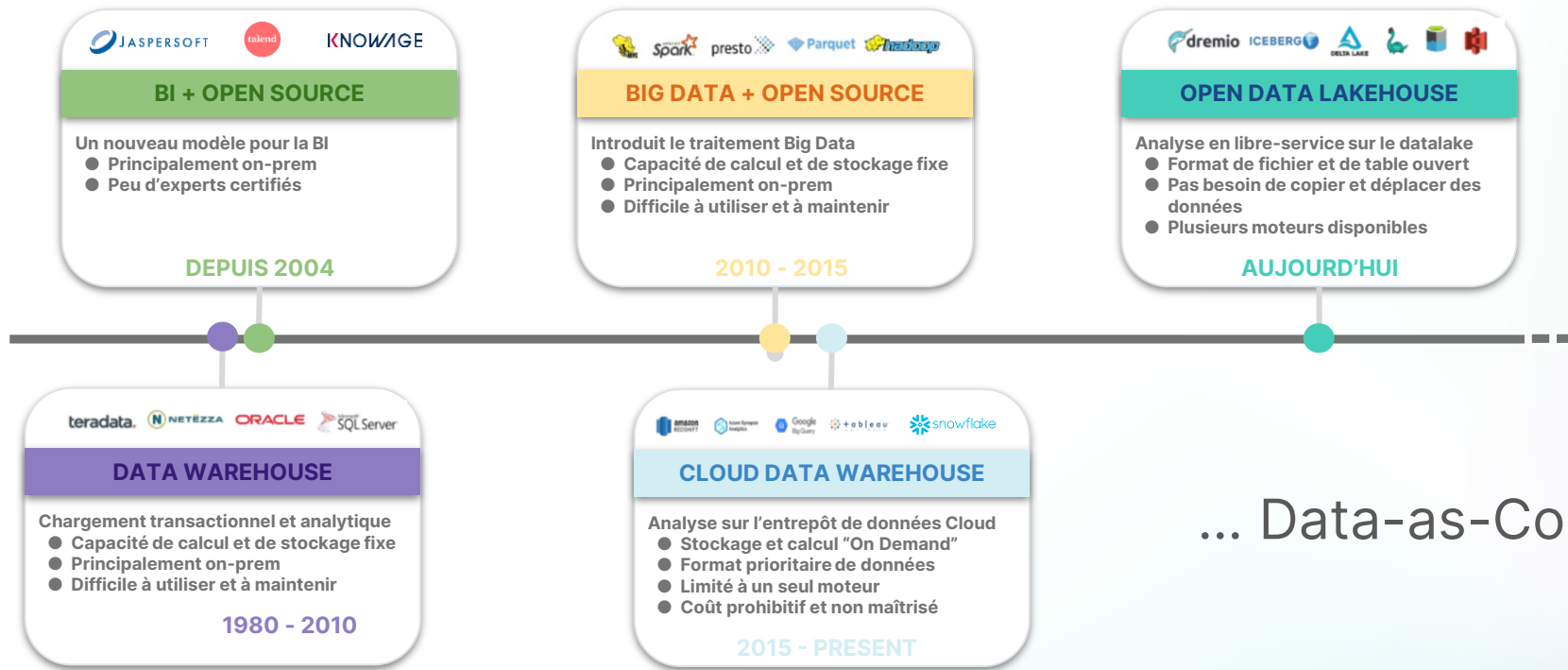
Open Source et Communauté

- Apache Arrow (**60M+** downloads/m), Apache Iceberg, Nessie
- Créateur et organisateur de la conférence **Subsurface** LIVE

Fondée en 2014, Version 1.0 en 2017

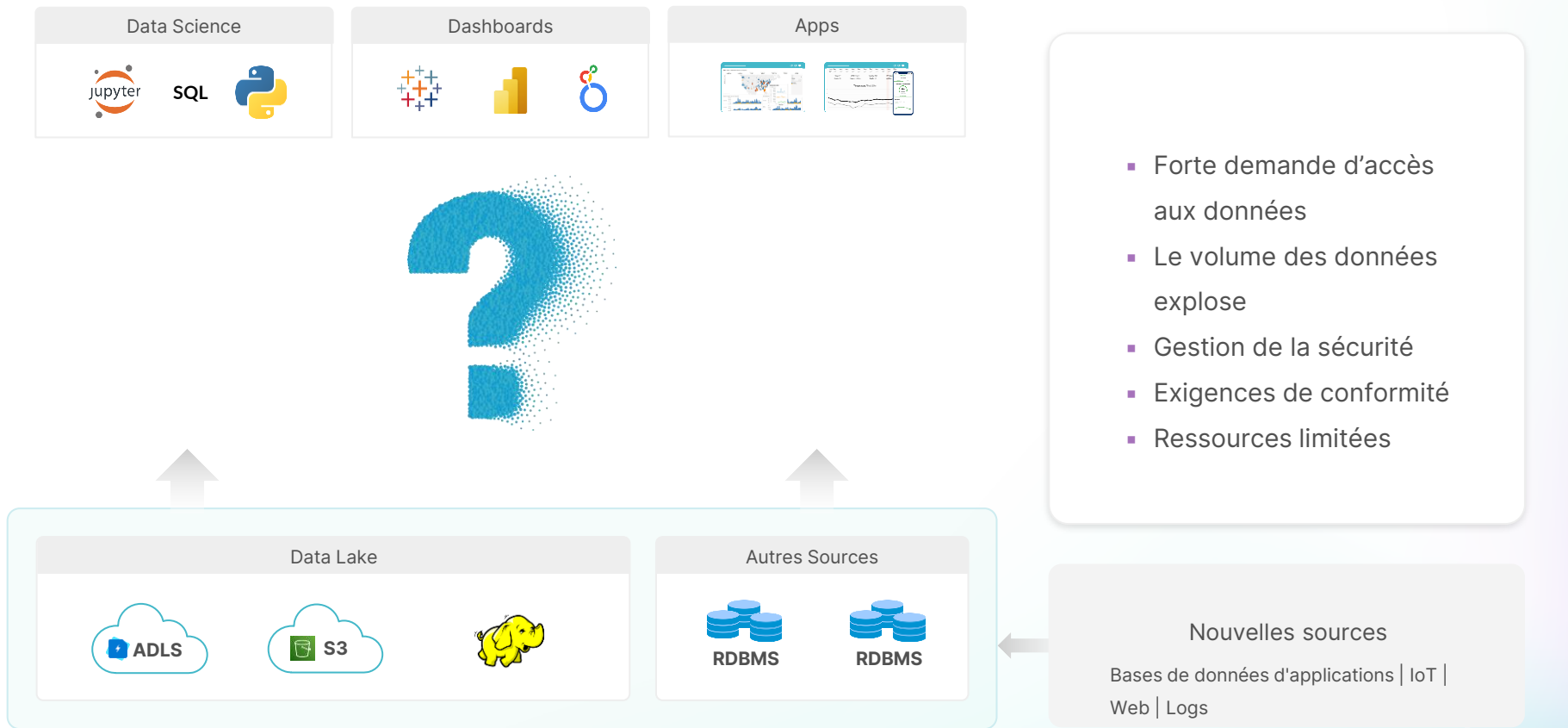


Historique du Data Warehouse à l'Open Data Lakehouse

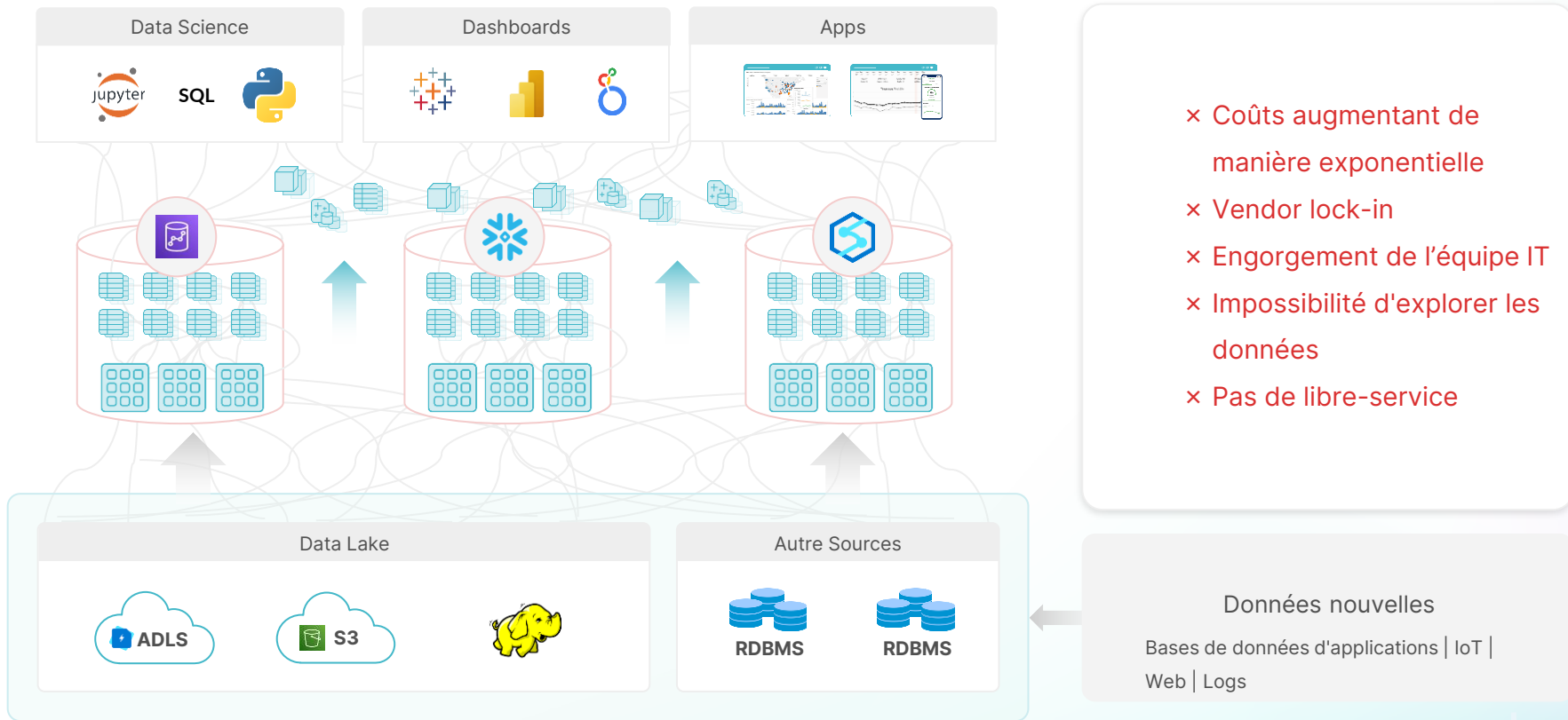


... Data-as-Code ?

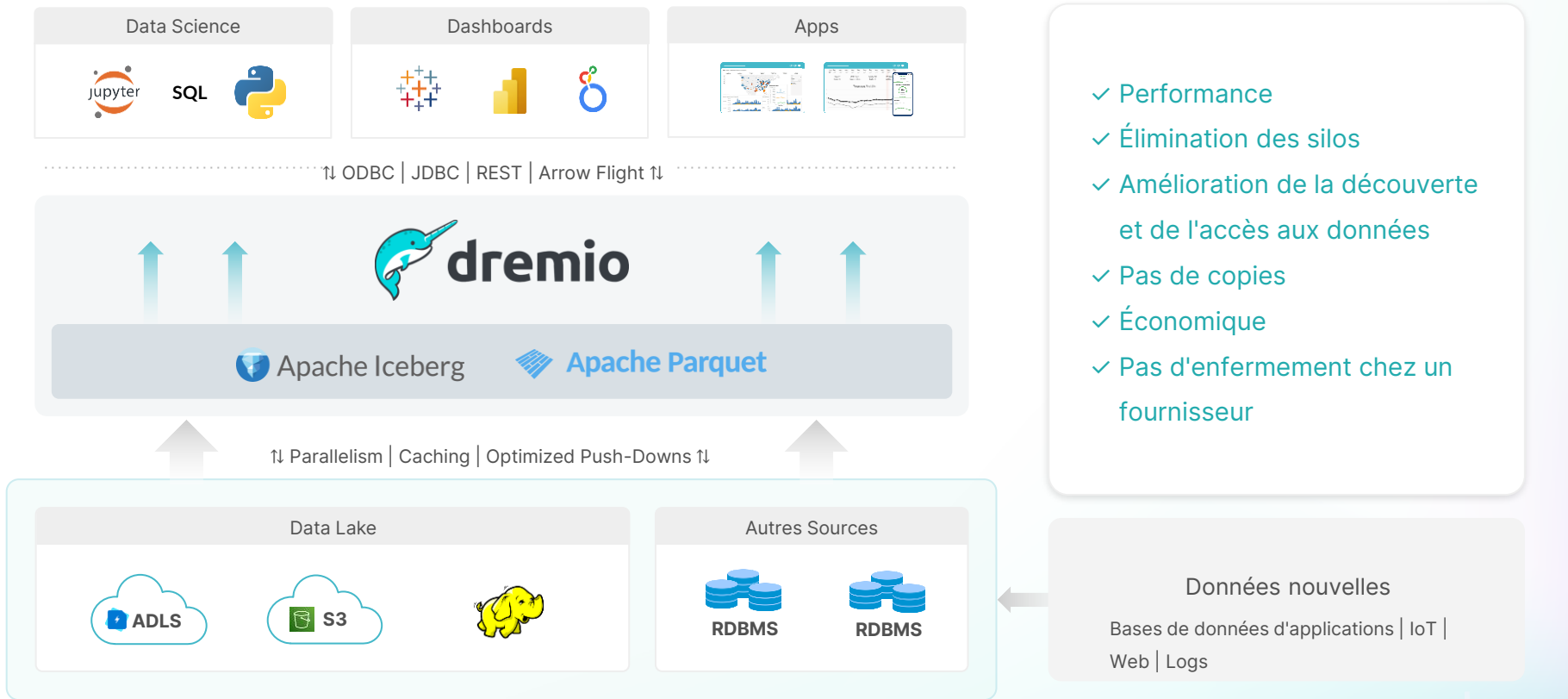
Les entreprises veulent démocratiser les données... comment ?



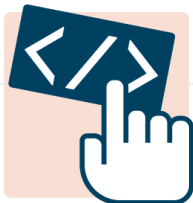
Data Warehouses : Coûteux, Propriétaires, Complexes



Dremio Data Lakehouse: Facile, Ouvert, 1/10ème du coût



L'avantage Dremio



Analytique en libre-service

Interface utilisateur moderne et intuitive

Vue unifiée des données
(en local, hybride et Cloud)

Requêtes fédérées



Données ouvertes, pas de lock-in

Basé sur des normes impulsées par
la communauté, notamment :

- Apache Parquet
- Apache Iceberg
- Apache Arrow



Performance à 1/10ème du coût

Requêtes performantes

Haute concurrence

Pas de copies de données coûteuses
à gérer

3

CAPACITÉS TECHNIQUES

Data as a Product

- Qualité
- Réutilisation
- Documentation

Self-Service (Libre Service)

- Accès
- Catalogue
- Partage

Data Mesh

Orienté Métier

- Structure
- Propriété
- Responsabilité
- Producteurs / Consommateurs

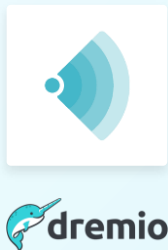
Gouvernance Fédérée

- Sécurité
- Conformité
- Interopérabilité
- Données Personnelles

Architecture Dremio

BI tools, data science notebooks, SQL editors

↕ ODBC | JDBC | REST | Arrow Flight ↕



Intuitive U/I

SQL Runner, No-code UI for self-service data curation/sharing

Semantic Layer

Unified view of data with view/table hierarchy, built-in data catalog and lineage

Query Engine

- Apache Arrow-based, vectorized execution
- Query acceleration - Data Reflections, Federated queries
- Elastic engines, Multi-engine architecture, workload management, query routing

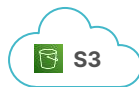
Governance & Security

RBAC, fine-grained access control, authentication, auditing and query history

↕ Iceberg, Delta Lake | Parquet, ORC, JSON, CSV Readers | ARP Connectors ↕

VM | K8S | YARN

Cloud Object Storage



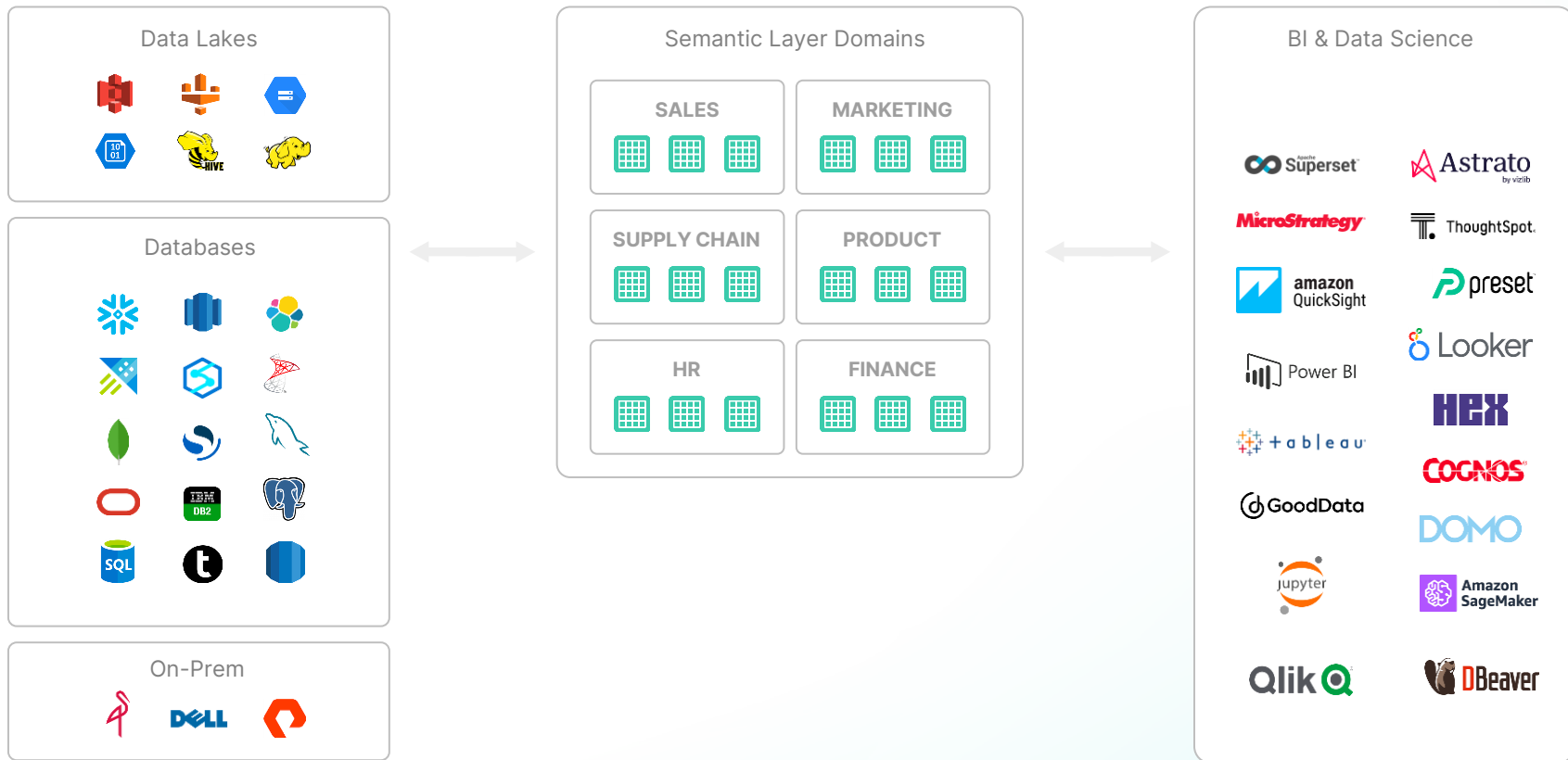
On-Prem



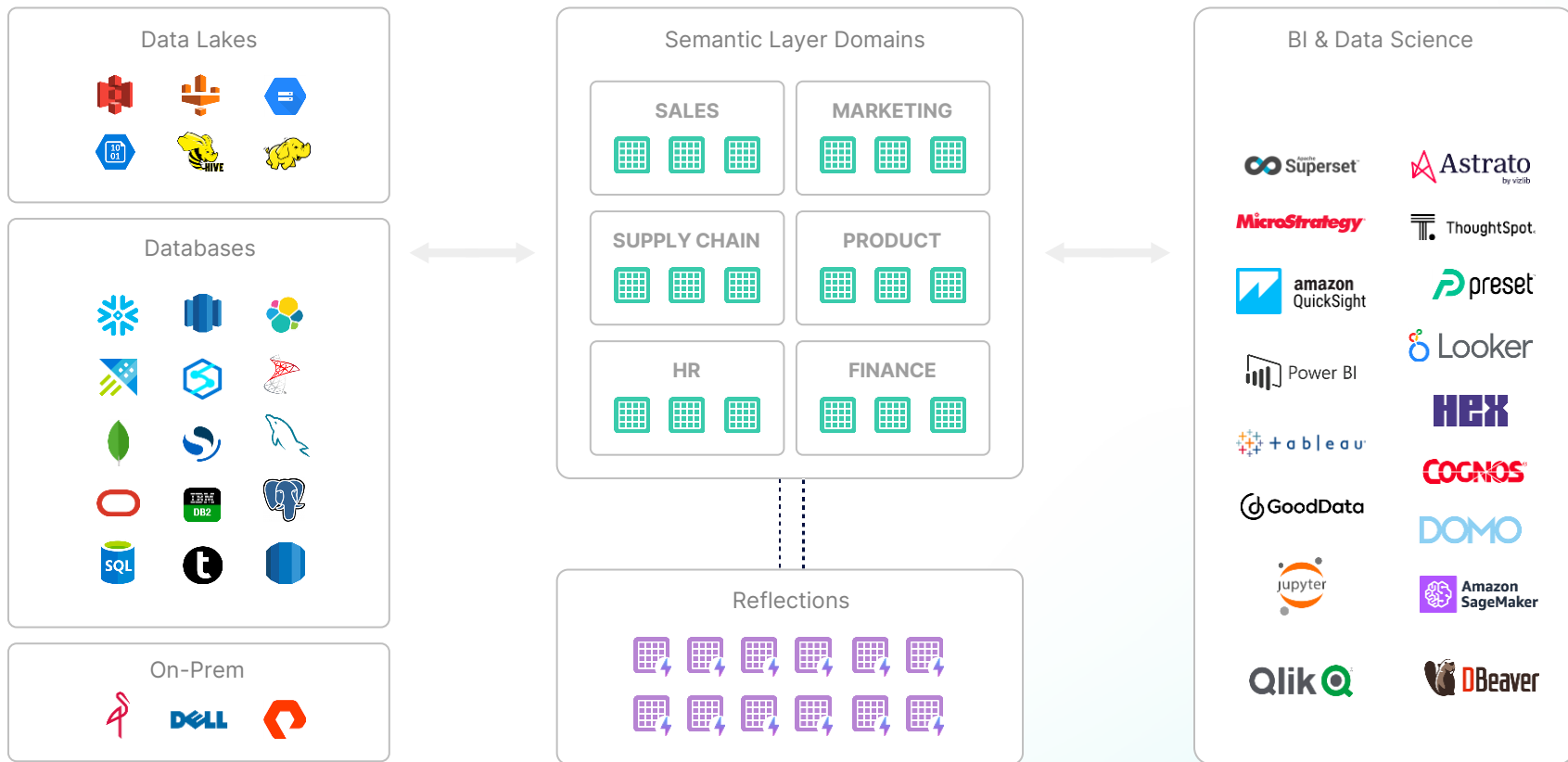
A collage of logos for various data science and business intelligence tools, arranged in a grid-like fashion. The logos include:

- Superset**: A logo with two interlocking circles and the text "Superset".
- Astrato**: A logo with a stylized red 'A' and the text "Astrato by vizib".
- MicroStrategy**: A logo with the text "MicroStrategy" in red.
- ThoughtSpot**: A logo with a stylized 'T' icon and the text "ThoughtSpot".
- amazon QuickSight**: A logo with the Amazon logo and the text "amazon QuickSight".
- preset**: A logo with a green 'P' icon and the text "preset".
- Looker**: A logo with a colorful 'L' icon and the text "Looker".
- HEX**: A logo with the text "HEX" in blue.
- Tableau**: A logo with a colorful 'T' icon and the text "tableau".
- Cognos**: A logo with the text "COGNOS" in red.
- GoodData**: A logo with a 'G' icon and the text "GoodData".
- DOMO**: A logo with the text "DOMO" in blue.
- jupyter**: A logo with a colorful 'J' icon and the text "jupyter".
- Amazon SageMaker**: A logo with a purple 'S' icon and the text "Amazon SageMaker".
- Qlik**: A logo with the text "Qlik" in green.
- DBBeaver**: A logo with a beaver icon and the text "DBBeaver".

Couche Sémantique Orientée Domaine - Data layer

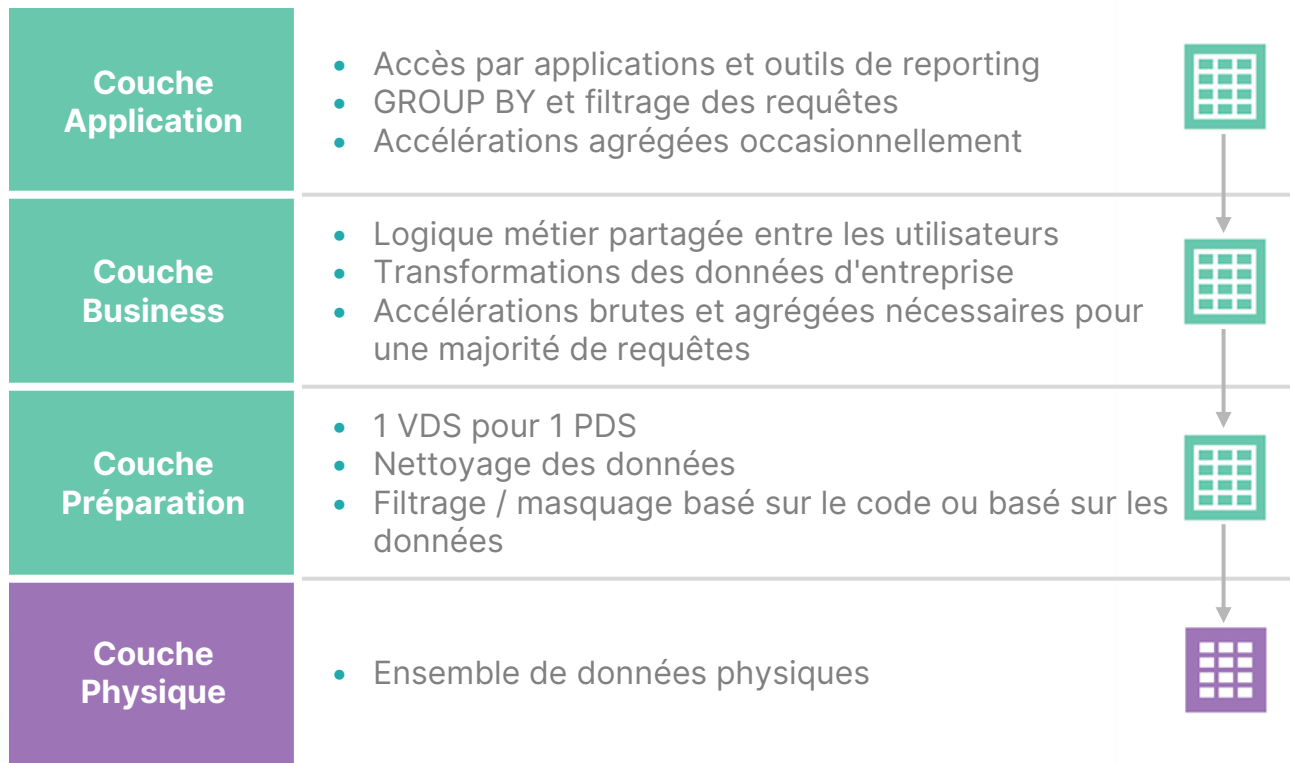


Couche Sémantique Orientée Domaine - Accélération

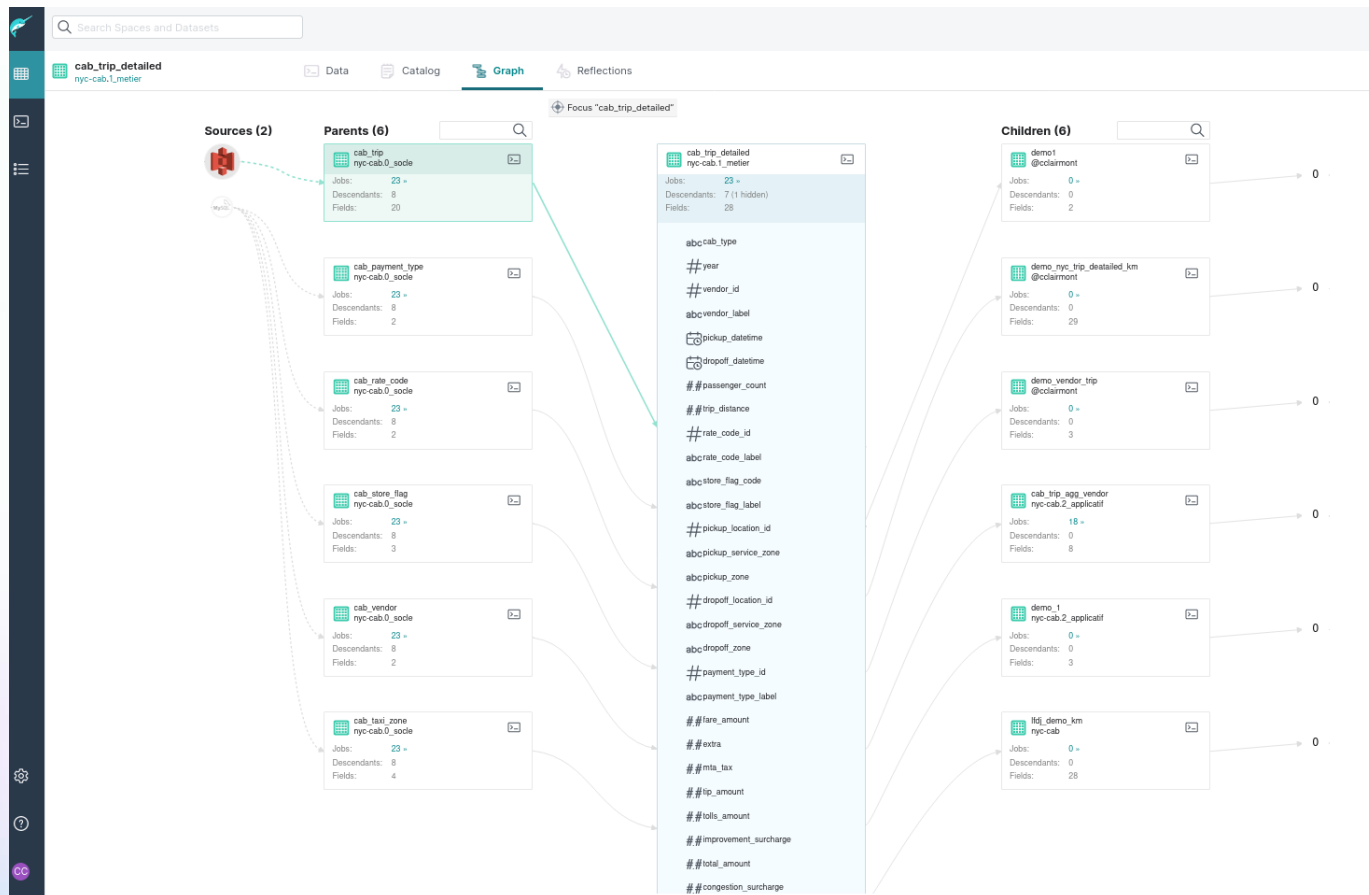


Formalisation de la Couche Sémantique

- Analogue aux techniques de modélisation standard des data warehouse
- Favorise une solution plus performante et nécessitant peu de maintenance
- Réutilisation des vues logiques (VDS)
- Agilité des équipes de développement et les utilisateurs métier
- Meilleur contrôle sur les données



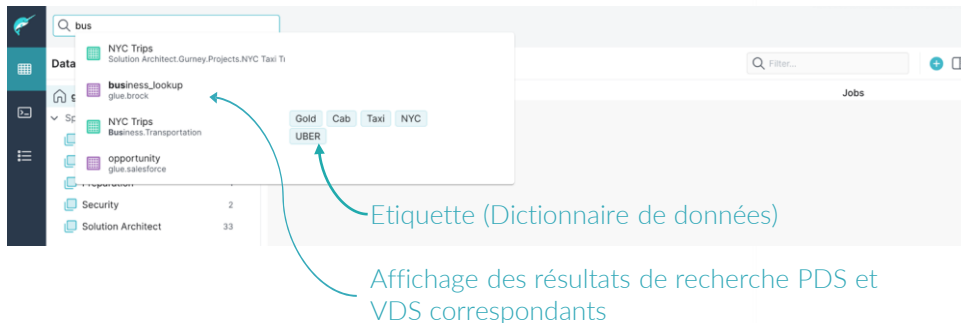
Data Lineage



Interface Intuitive: Recherchez dans votre catalogue

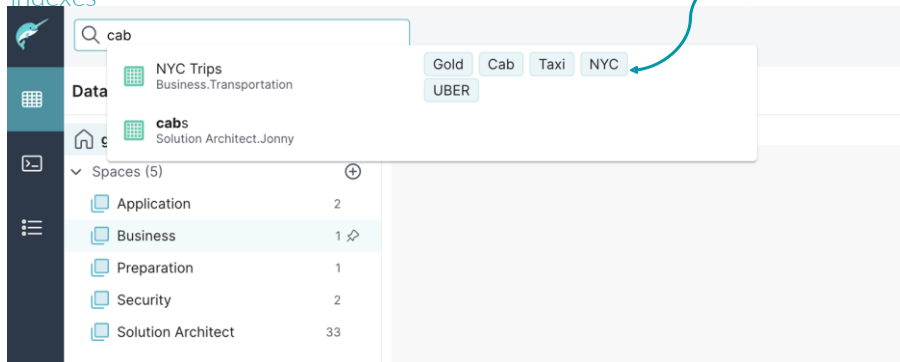
- Recherché par:
 - Nom du jeu de données physique et virtuel
 - Nom du répertoire et du schéma
 - Nom des colonnes
 - Mots clés
- La correspondance de sous-chaîne est prise en charge
- La fonctionnalité des étiquettes s'apparente à un dictionnaire de données.

Recherche dans le catalogue global



La recherche par balise affiche les résultats indexés

Navigation par Mots clé



Interface Intuitive: Wiki & Mots Clés

- Lors de l'affichage d'un PDS/VDS, cliquez sur le bouton "Wiki" vous amène au wiki et au contenu des mots clés pour l'ensemble de données
- Vous pouvez l'utiliser pour tirer partie des connaissances organisationnelles, prendre des décisions plus éclairées sur l'utilisation de l'ensemble de données, voir le contenu connexe, etc.
- Le contenu du wiki est composé de texte, de liens et d'images

The screenshot shows the Dremio Wiki interface for the 'NYC Taxi Trips' dataset. The interface includes a search bar at the top, a navigation sidebar on the left, and a main content area. The main content area displays the 'NYC Taxi Trips dataset' with a large yellow 'TAXI' sign image. Below the image, there is a 'Dimensions' section listing various attributes like pickup_date, pickup_datetime, dropoff_date, dropoff_datetime, rate_code, and store_and_fwd_flag. On the right side, there is a 'Tags' section with buttons for 'Gold', 'Cab', 'Taxi', 'NYC', and 'UBER', and a 'Columns' section listing 20 columns including vendor_id, pickup_date, pickup_datetime, dropoff_date, dropoff_datetime, passenger_count, trip_distance_mi, pickup_longitude, pickup_latitude, and rate_code.

Search Spaces and Datasets

NYC Trips
Business, Transportation

Data Catalog Graph Reflections

Wiki

NYC Taxi Trips dataset



Dimensions

- pickup_date - Date when the customer is picked up
- pickup_datetime - Timestamp when the customer is picked up
- dropoff_date - Date when the customer was dropped off
- dropoff_datetime - Timestamp when the customer was dropped off
- rate_code - The final rate code in effect at the end of the trip
- store_and_fwd_flag - Indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server

Tags

Gold Cab Taxi NYC UBER Add tag

Columns 20

Abc vendor_id

pickup_date

pickup_datetime

dropoff_date

dropoff_datetime

passenger_count

trip_distance_mi

pickup_longitude

pickup_latitude

Abc rate_code

Export de Données

Cette fonctionnalité peut être activée ou non

Télécharger les données résultats de vos requêtes

a) Dremio vous permet de télécharger le jeu de résultats sous forme de fichier JSON, CSV ou Parquet.

a) Vous pouvez aussi copier le jeu de résultats dans le presse-papiers. Si le jeu de résultats est trop volumineux, téléchargez le contenu du tableau pour obtenir les résultats complets.



pickup_datetime	# passenger_count	## trip_distance_mi	## fare_amount	## tip_a
2013-02-10 21:59:00.000	2	1.61	7.0	
2013-02-10 22:00:00.000	6	0.98	6.0	
2013-02-10 21:57:00.000	1	1.09	6.0	

API d'accès aux Données

Obtenir les données via l'API

L'API REST Dremio est organisée par type de ressources telles que les sources. Elle est conçue autour des principes RESTful. HTTP **GET** est utilisé pour recupérer les ressources existantes, **POST** en créé de nouvelles, **PUT** les met à jour et **DELETE** les supprime.

Les objets manipulés par l'API sont les suivants :

- ☐ Catalog API
- ☐ Grant API
- ☐ User API
- ☐ Roles API
- ☐ Reflection API
- ☐ Job API
- ☐ Workload Management API
- ☐ Node Collections API
- ☐ Source API
- ☐ SQL API
- ☐ Accounts API
- ☐ Vote API
- ☐ KVStore API

Pour exporter des données on utilisera l'API SQL

```
curl -X POST \
  http://localhost:9047/api/v3/sql \
  -H 'Authorization: dremio1cmq10aoro8o0el8rj0fjh6n0' \
  -H 'Content-Type: application/json' \
  -d '{
    "sql": "SELECT * FROM SFIncidents.weapons_laws WHERE PdDistrict = '\\\\'CENTRAL\\\\'\\' LIMIT 3"
  }'
```

Vérifier le statut du job

```
GET localhost:9047/api/v3/job/23a4bf43-52aa-5d29-c473-0083c44fe700
```

Obtenir les résultats de la requête

```
GET localhost:9047/api/v3/job/23a4bf43-52aa-5d29-c473-0083c44fe700/results
```

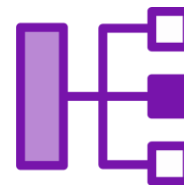
Performance



Architecture Distribuée

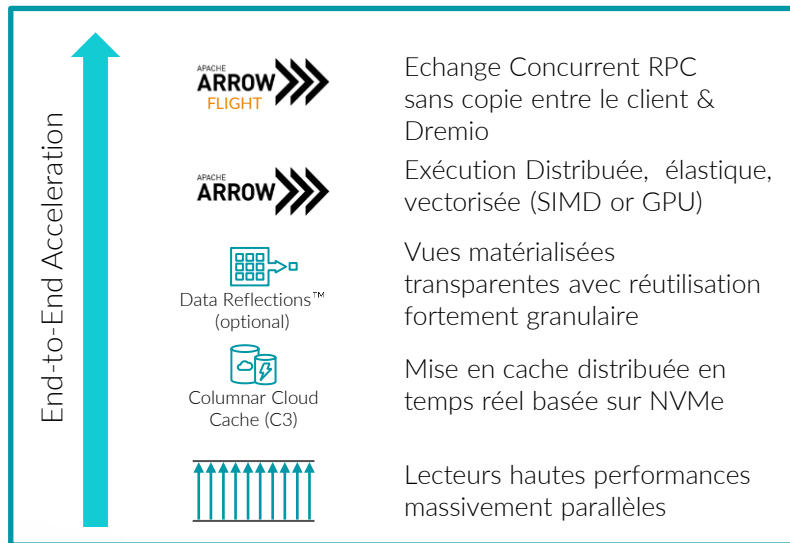


Accélération de Requêtes

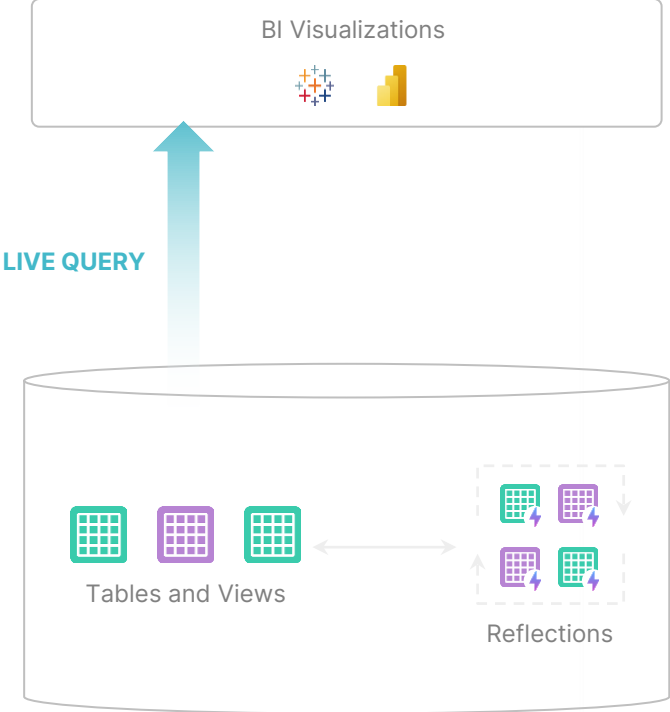
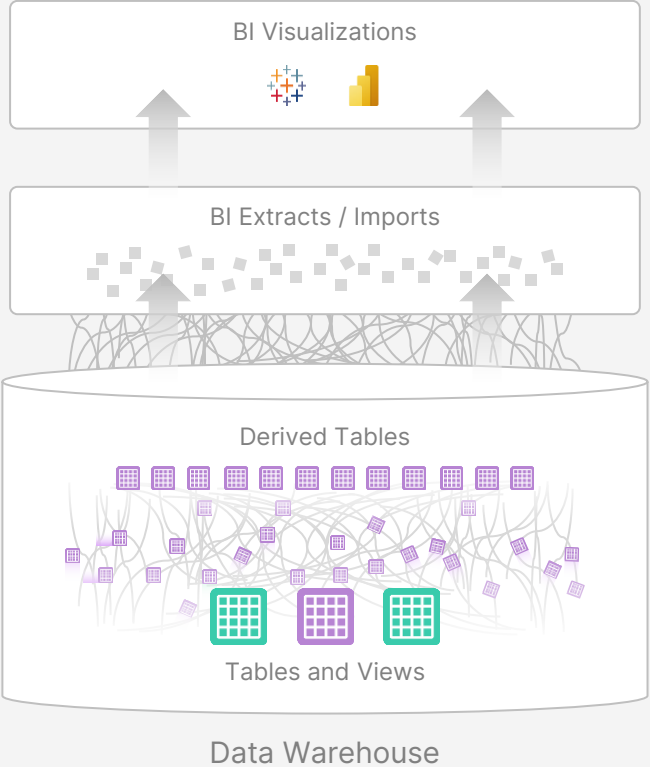


Gestion de la montée en charge

Performance à tous les niveaux



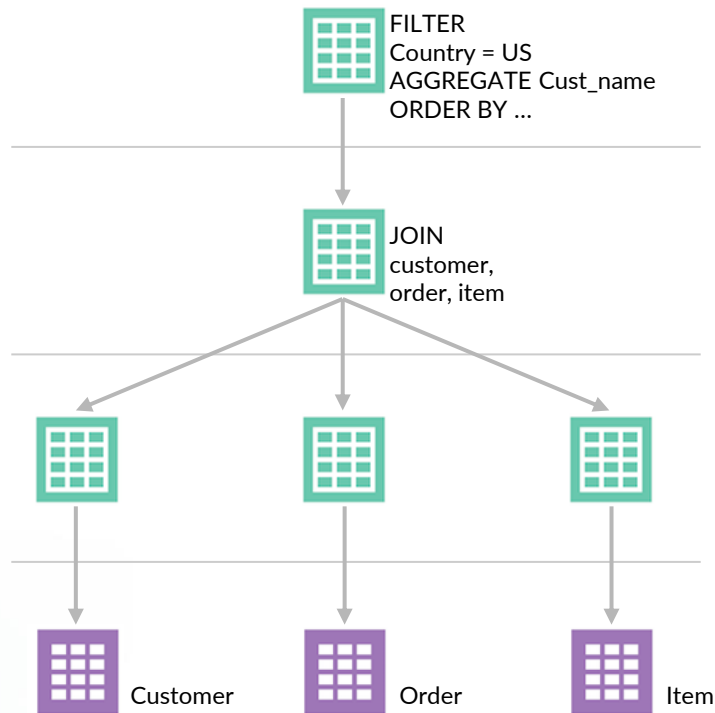
Les réflexions éliminent le besoin d'extractions/importations BI



Les accélérations sont stockées et optimisées

- Représentations physiquement optimisées des données source, pour satisfaire partiellement ou entièrement une requête à la place des données source
- Chaque réflexion est stockée avec une expression SQL algébrique associée
- Le planificateur évalue les expressions algébriques pour les substitutions au moment de l'exécution
- Les réflexions peuvent inclure des données provenant de plusieurs sources de données.

Virtual and physical datasets



Reflection



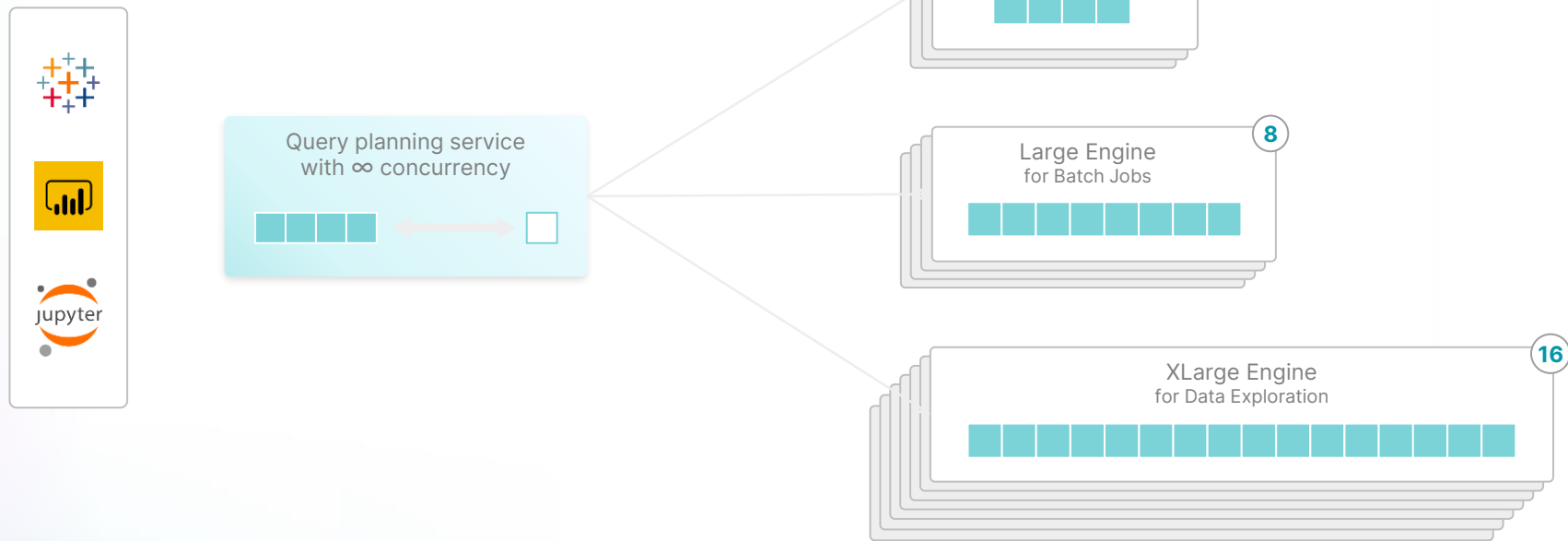
JOIN
customer, order,
item

FILTER
country = US

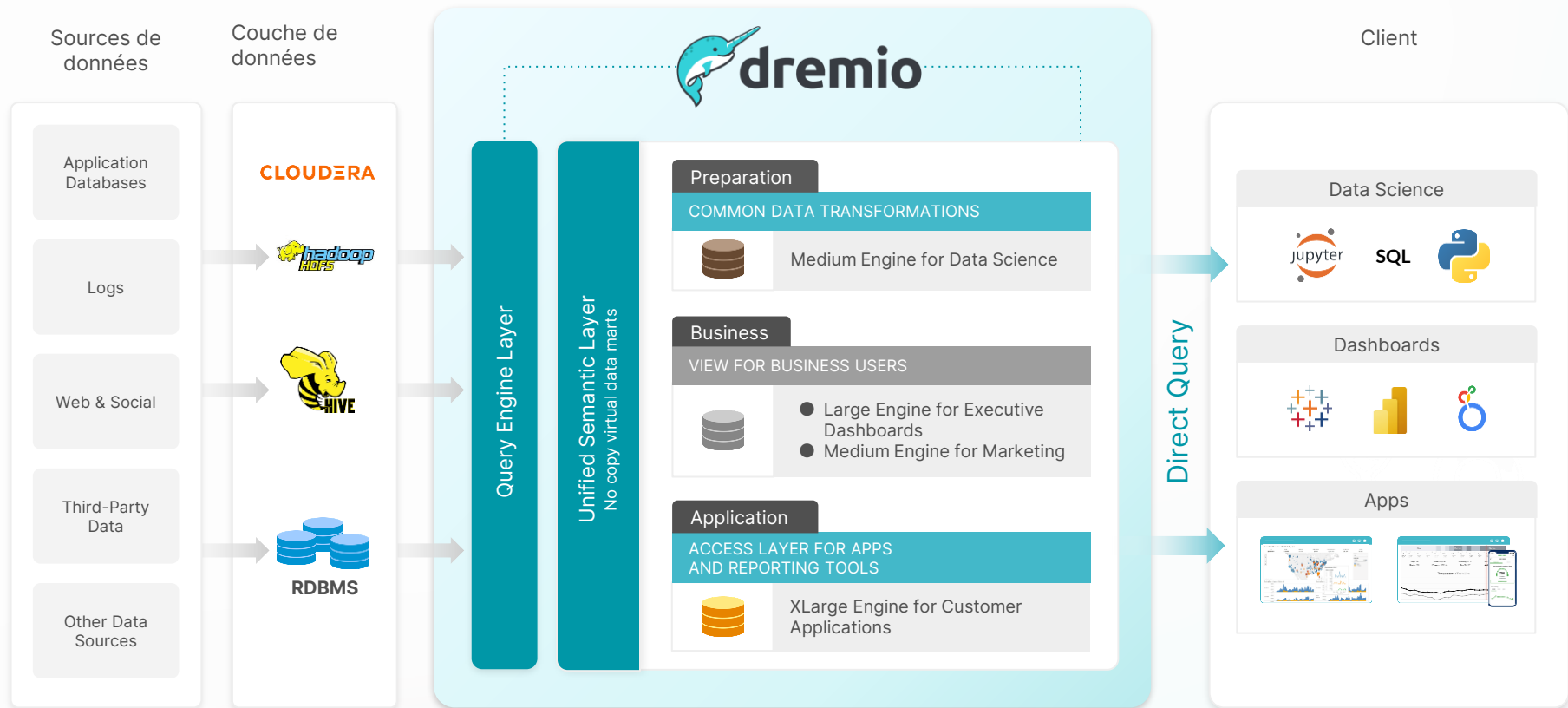
AGGREGATE
cust_name

ORDER BY ...

Architecture concurrence illimitée



Exemple d'architecture "Open Data Lakehouse"



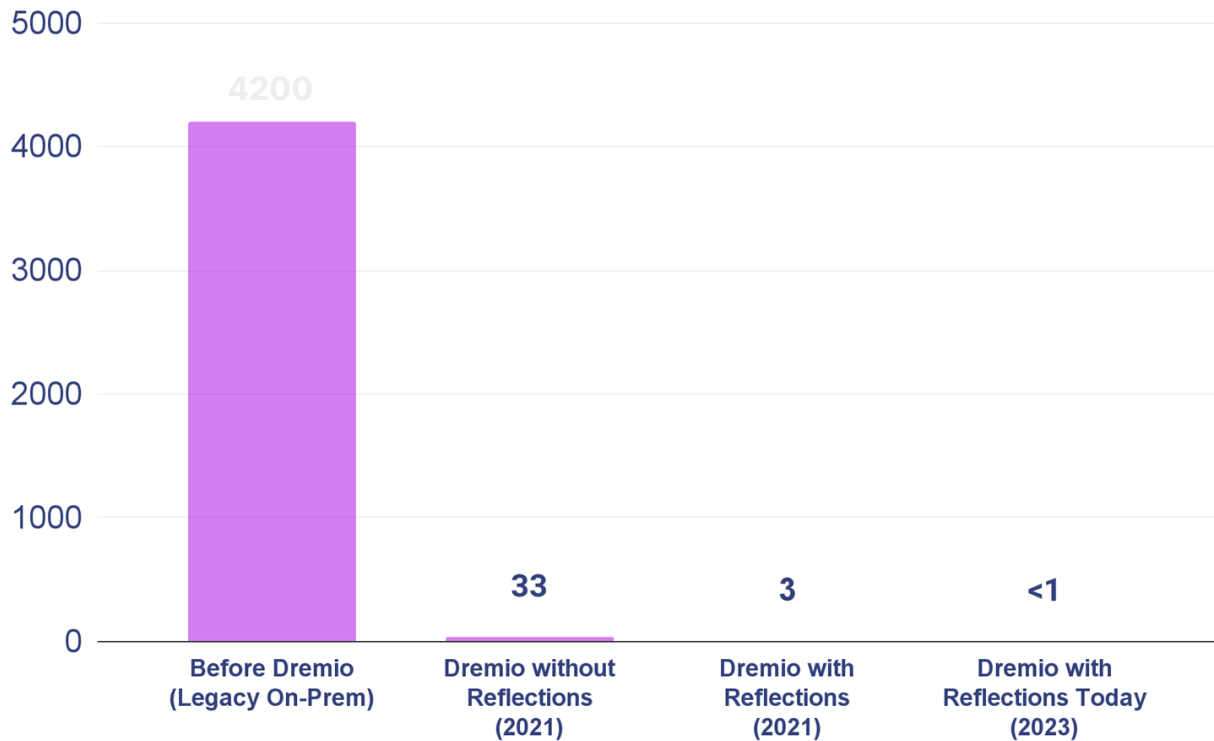
Des Performances en amélioration constante

Benchmark mené avec

RenaissanceRe

Premier fournisseur mondial de
réassurance et d'assurance

Query Performance in Seconds



L'accélération des requêtes permet une BI en quasi temps-réel

Observabilité

Observabilité en temps réel pour les réflexions, y compris les informations d'actualisation et d'utilisation

Recommandations

Recommandations automatiques de réflexion basées sur la charge de travail des requêtes

Mise à jour Incrémentale & Automatique

Mises à jour automatiques incrémentielles et basées sur les partitions lorsque les tables sont mises à jour

Reflection Matching

Amélioration de la correspondance de réflexion, des conseils et de l'explicabilité (pourquoi cela ne correspondait-il pas ?)

Accélération Autonome / Implicite



Dremio crée/supprime automatiquement des réflexions pour accélérer vos charges de travail

Name	Type	Dataset	Current Footprint	Last Refresh Duration	Accelerated Count	Refresh Job History
dataset960_agg960	Aggregation	dataset960	9.05 KB	00:00:46	5	History
dataset960_raw960	Raw	dataset960	799.60 KB	00:00:44	15	History
dataset960_2_agg960	Aggregation	dataset960.2	8.19 KB	00:00:34	0	History
dataset960_2_raw960	Raw	dataset960.2	8.23 KB	00:00:44	0	History

Open Table Formats Réduire la maintenance et Améliorer DML



DML à faible latence

UPDATE/DELETE à haut débit : fusion à la lecture, suppressions d'égalité

Statistiques tables Apache Iceberg (puffin)

Index Z-Order

Partitions adaptatives



Ingestion

Ingestion partielle avec la clause SQL CONTINUE

Nouveaux formats et emplacements de fichiers

Conversions natives : Hive -> Iceberg, Object Storage -> Iceberg

Conversions natives : Delta Lake -> Iceberg

Ingestion optimisée des séries temporelles

Ingestion depuis Kafka



Améliorations des performances, de la fiabilité et des coûts



- Optimisation Plan d'exécution
- Mise en cache de l'ensemble de résultats
- Amélioration filtres "pushdown" et élimination des partitions
- Gestion et détections des Asymétries (Skew)
- Nouveau lecteur Parquet et prise en charge de RLE_DICTIONARY
- Amélioration des performances d'Apache Iceberg



Performance Plan d'exécution



Performance Exécution



Performance I/O



Fiabilité

- Annulation instantanée de requête
- "Spilling" pour tous les opérateurs
- Allocation de mémoire dynamique en temps réel



Résilience complète des requêtes (no OOM) et une plus grande simultanéité



Réduction des Coûts

- Compression améliorée pour Parquet avec zstd
- Support processeurs ARM



Réduction de 40 % des coûts d'infrastructure

La Fédération de requêtes réduit l'utilisation de l'ETL

Requête au travers de votre Data Lakehouse et vos Autres Databases :

Recently Shipped

Snowflake connecteur
DB2 connecteur
Dremio-to-Dremio connecteur

Q1

New: OpenSearch
Enhanced: MongoDB

Q2

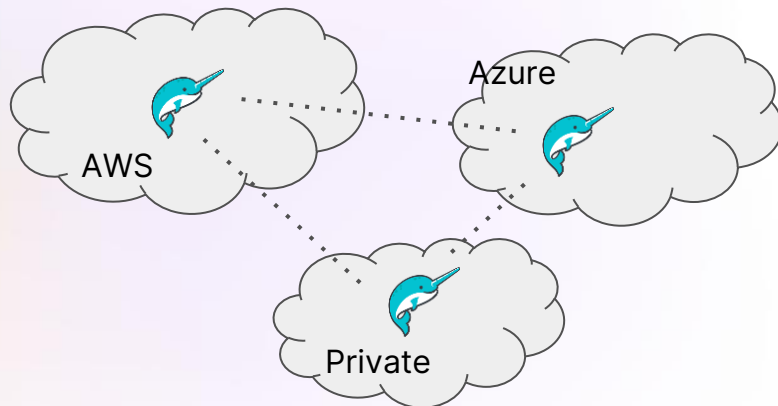
Enhanced: Oracle
Enhanced: SQL Server
Enhanced: Snowflake

Q3

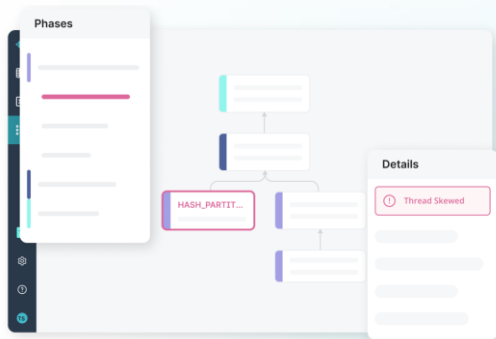
New: Druid
New: Vertica
New: SAP HANA

Requête au travers des **multiples environnements**

- Multi-region
- Multi-cloud
- Hybrid cloud



Expérience Améliorée pour les Analystes et les Data Scientists



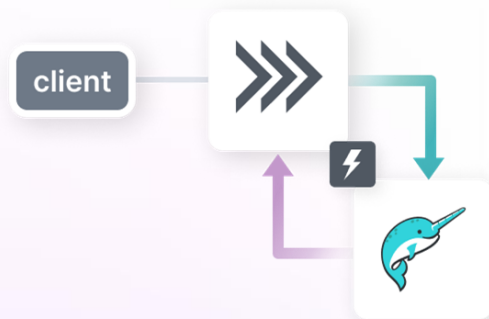
Dremio SQL Runner & Profiler

- Enhanced auto-complete (Q1)
- In-product SQL ref and tutorials (Q1-Q2)
- Dataset cards (Q1)
- Query profiler recommendations (Q3)
- Persistent query tabs (Q3)



Amélioration des connecteurs Clients

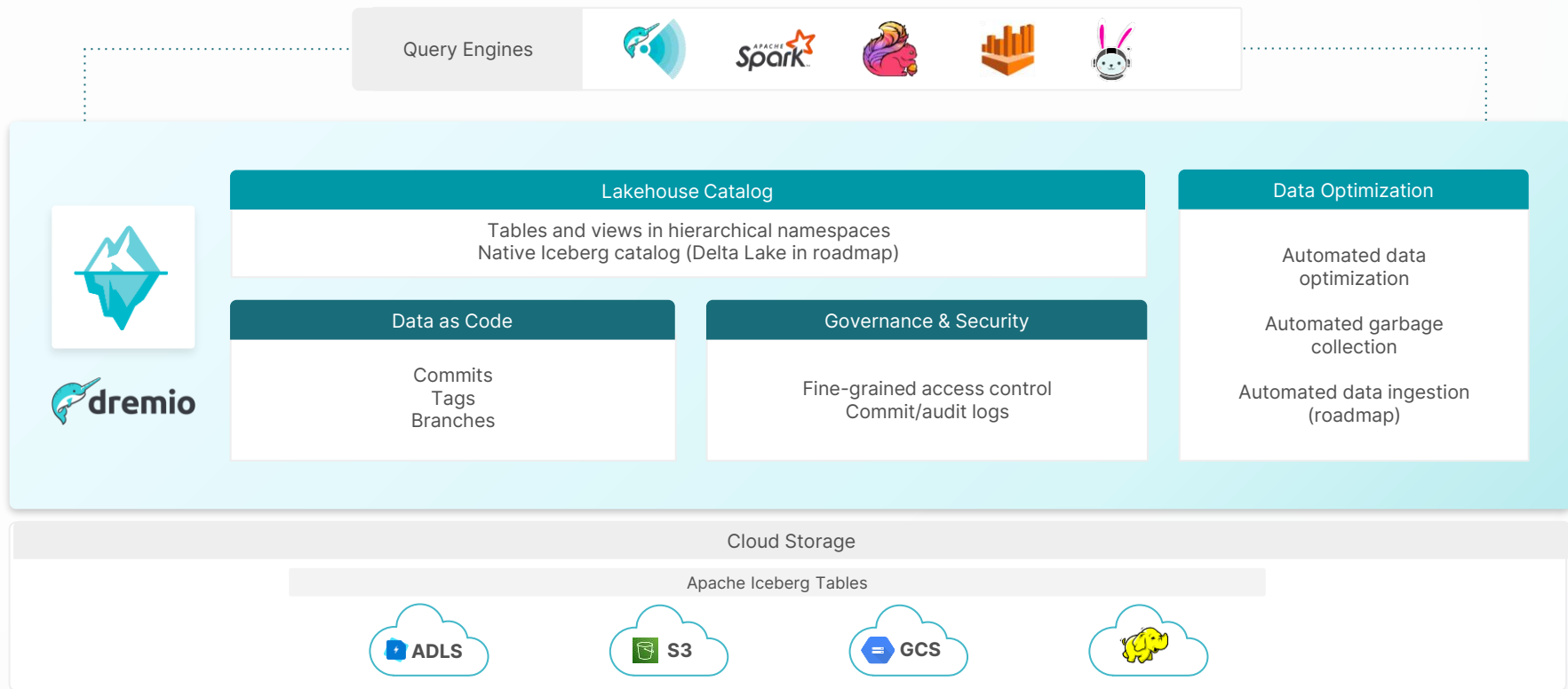
- Power BI & Tableau SSO (Recently Shipped)
- AWS QuickSight (Q2)
- Superset SSO (Q3)



Arrow Flight SQL

- New Python integrations: pyarrow, SQLAlchemy (Q1)
- Arrow Flight SQL for PBI & Tableau (Q2-Q3)
- Parallel Arrow Flight SQL endpoint (Q4)

Dremio Arctic est un **Service de Gestion de Data Lakehouse**



The background is a solid orange color with a pattern of thin, white, wavy lines that create a sense of movement and depth. The lines are more concentrated on the left side and fade out towards the right.

4

CHIFFRES CLEFS

Des performances accrues pour un coût réduit !

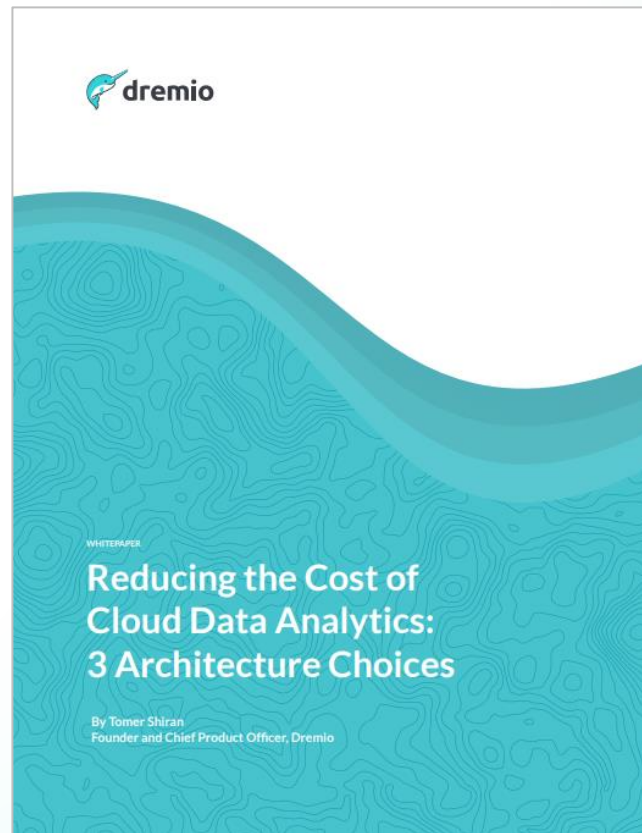
4 à 100 fois plus rapide

L'accélération des requêtes (Apache Arrow, C3 et Data Reflections) augmente les performances de 4 à 100 fois par rapport aux moteurs de requête simples.

Pour 90% moins cher

L'accélération des requêtes se traduit par une réduction des coûts de 75 % (4x) ou plus. Les moteurs élastiques offrent une réduction supplémentaire de 60 % (2,5x) ou plus. Dremio est donc capable de réduire vos dépenses d'infrastructure de 90% ($4 * 2,5 = 10$).

Source : [Dremio, des performances accrues pour un coût réduit - Synaltic](#)





Notre téléphone

01 55 28 89 40



Nos bureaux

24 Rue de l'Église, 94300 Vincennes



Notre site web

www.synaltic.fr



Marc SALLIERES

Synaltic
Président
msallieres@synaltic.fr

Charly CLAIRMONT

Synlatic
Directeur Technique
cclairmont@synaltic.fr

Guillermo KIRKPATRICK

Dremio
guillermo.kirkpatrick@dremio.com

