Lakehouse Analytics with Dremio

Dimitar D. Mitov
DXC Technology

DELIVERING EXCELLENCE FOR OUR CUSTOMERS AND COLLEAGUES

Introduction

Dimitar D. Mitov

Data Engineer

DXC Technology

https://github.com/ddmitov

https://www.linkedin.com/in/dimitar-mitov-12388982

Fortune 500

DXC Technology RANK 207

DXC Technology (NYSE: DXC) helps global companies run their mission-critical systems and operations while modernizing IT, optimizing data architectures, and ensuring security and scalability across public, private and hybrid clouds.

The world's largest companies and public sector organizations trust DXC to deploy services to drive new levels of performance, competitiveness, and customer experience across their IT estates.

DXC.com

Data Lakes, Data Warehouses and The Pyramid of Doom

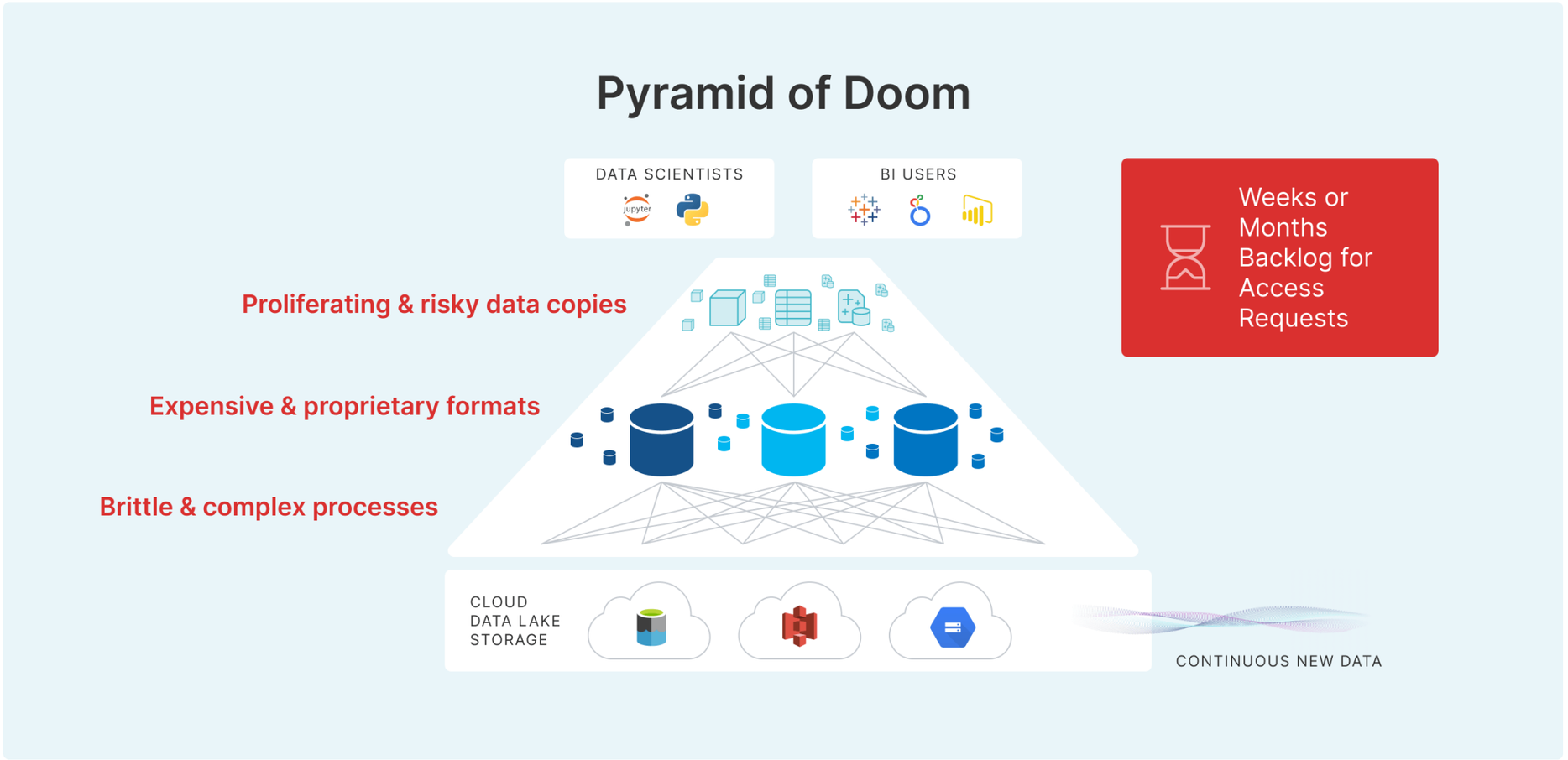
How complexity, cost and delays impair Data Analytics

What is a Data Lake?

AWS: *"A data lake is a **centralized repository** that allows you to store **all your structured and unstructured data at any scale**. You can **store your data as-is**, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions."*

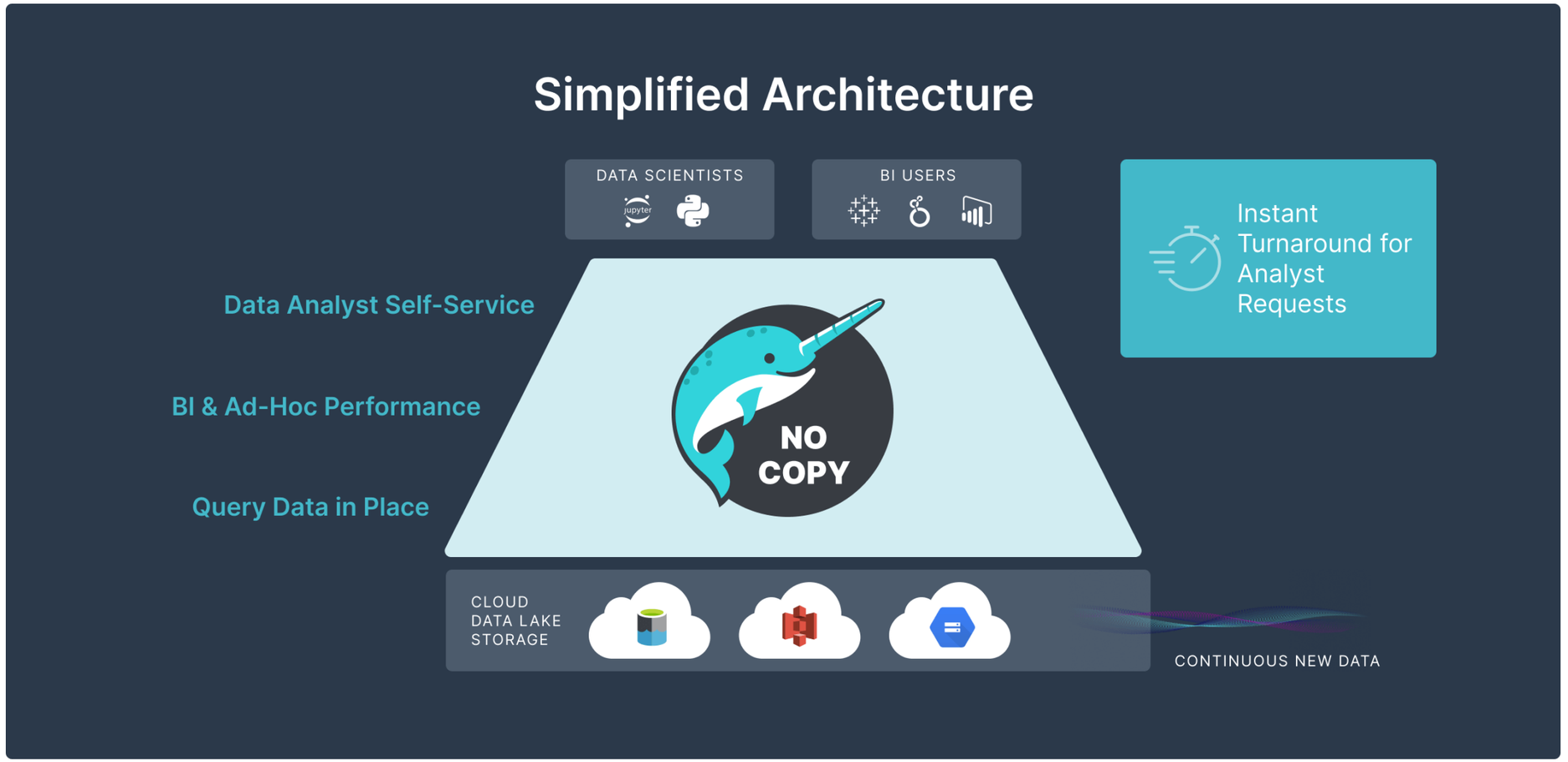
https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

The Pyramid of Doom



Waiting for a report or a dashboard or training data for ML?

Data Lakehouse using Dremio



What Is a Data Lakehouse?

Dremio: *"A lakehouse is a new type of data platform architecture that:*

- *Provides the **data management capabilities** of a data warehouse and takes advantage of the **scalability and agility** of data lakes*
- *Helps **reduce data duplication** by serving as the single platform for all types of workloads (e.g., BI, ML)*
- *Is **cost-efficient***
- *Prevents vendor lock-in and lock-out by leveraging **open standards**"*

https://www.dremio.com/blog/what-is-a-data-lakehouse/

What Is a Data Lakehouse?

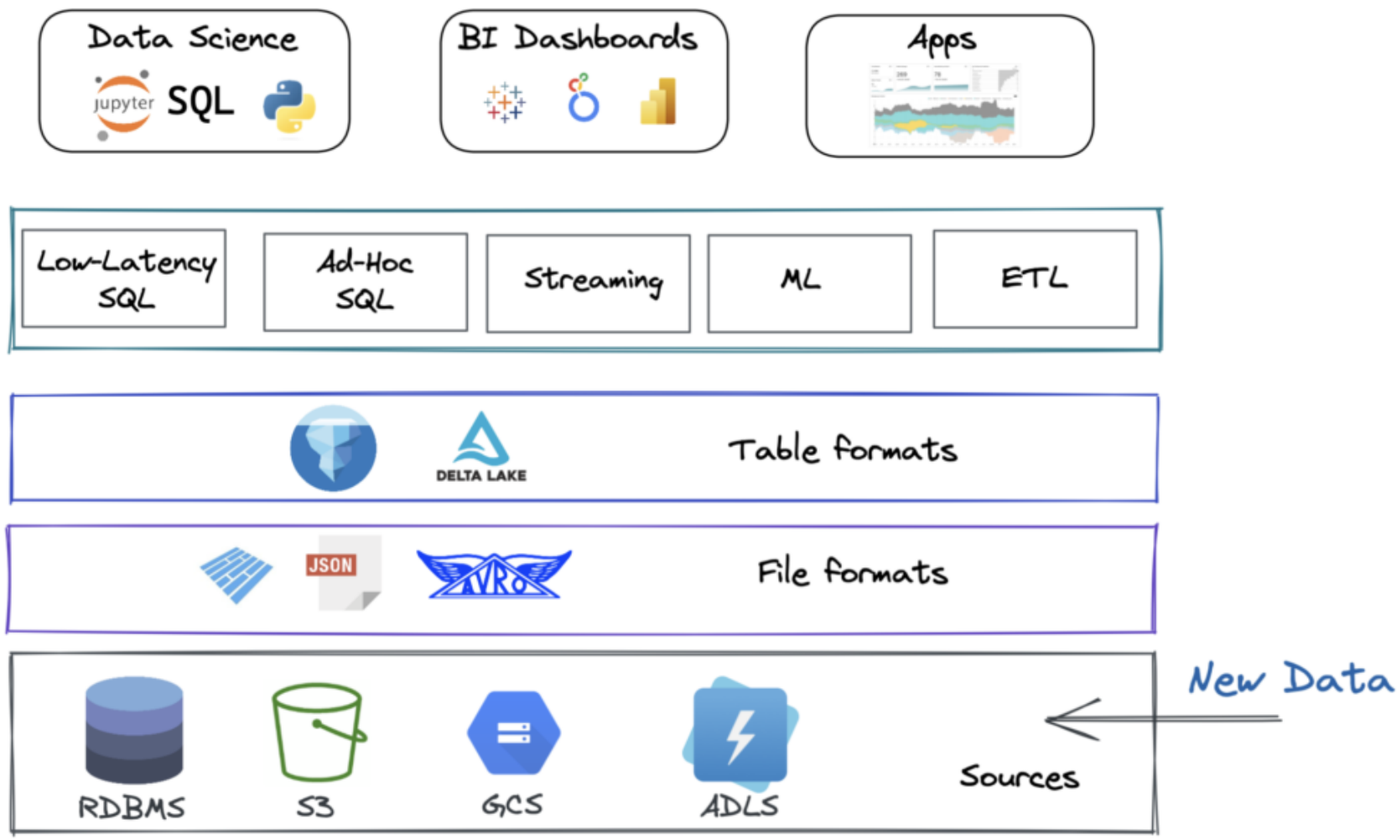
Databricks: *"A data lakehouse is a new, open data management architecture that combines the **flexibility**, **cost-efficiency**, and **scale** of data lakes with the **data management** and **ACID transactions** of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data."*

https://www.databricks.com/glossary/data-lakehouse/

Oracle: *"… a data lakehouse takes the **flexible storage of unstructured data** from a data lake and the **management features and tools** from data warehouses …"*

https://www.oracle.com/data-lakehouse/what-is-data-lakehouse/

Data Lakehouse Concept



Dremio Architecture

Some basic concepts and elements

Dremio Deployment Models

I. Cloud Service Provider Environment:

1. AWS Edition, 2. Azure ARM

II. Hosted Kubernetes Environment:

1. Azure AKS, 2. Amazon EKS, 3. Google Cloud GKE

III. Shared Multi-Tenant Environment:

1. Hadoop using YARN, 2. MapR using YARN

IV. Standalone Cluster

Dremio Editions

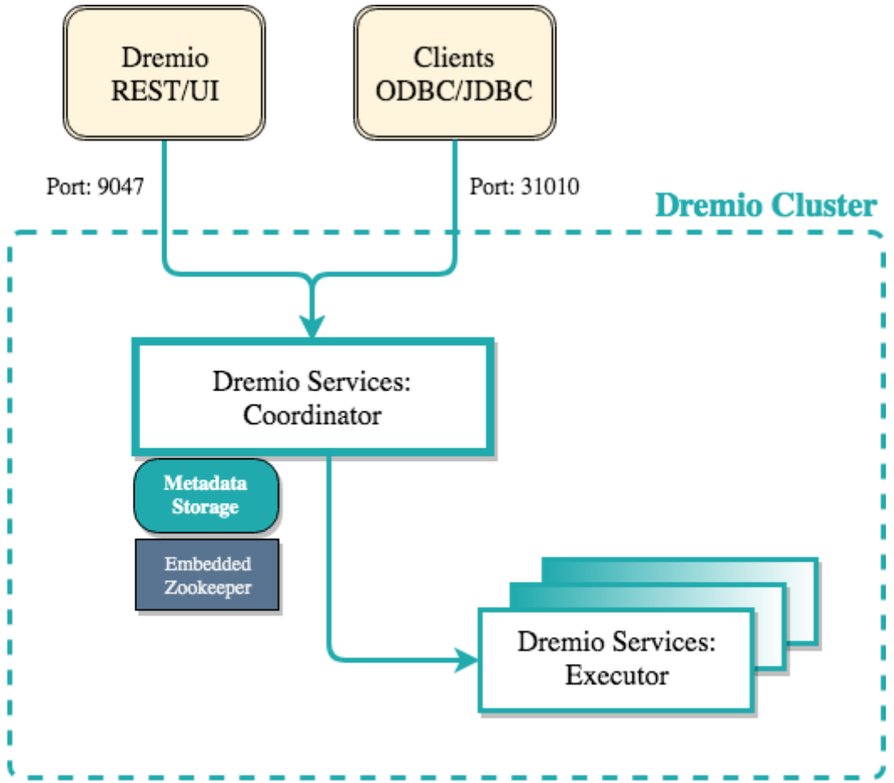
- **Standard (Software & Cloud):**

“A forever-free tier which provides everything you need to successfully build, automate, and query your data lake in production. This includes all of Dremio’s key features including reflections and self-service semantic layer, without any limits on query scale or concurrency.”

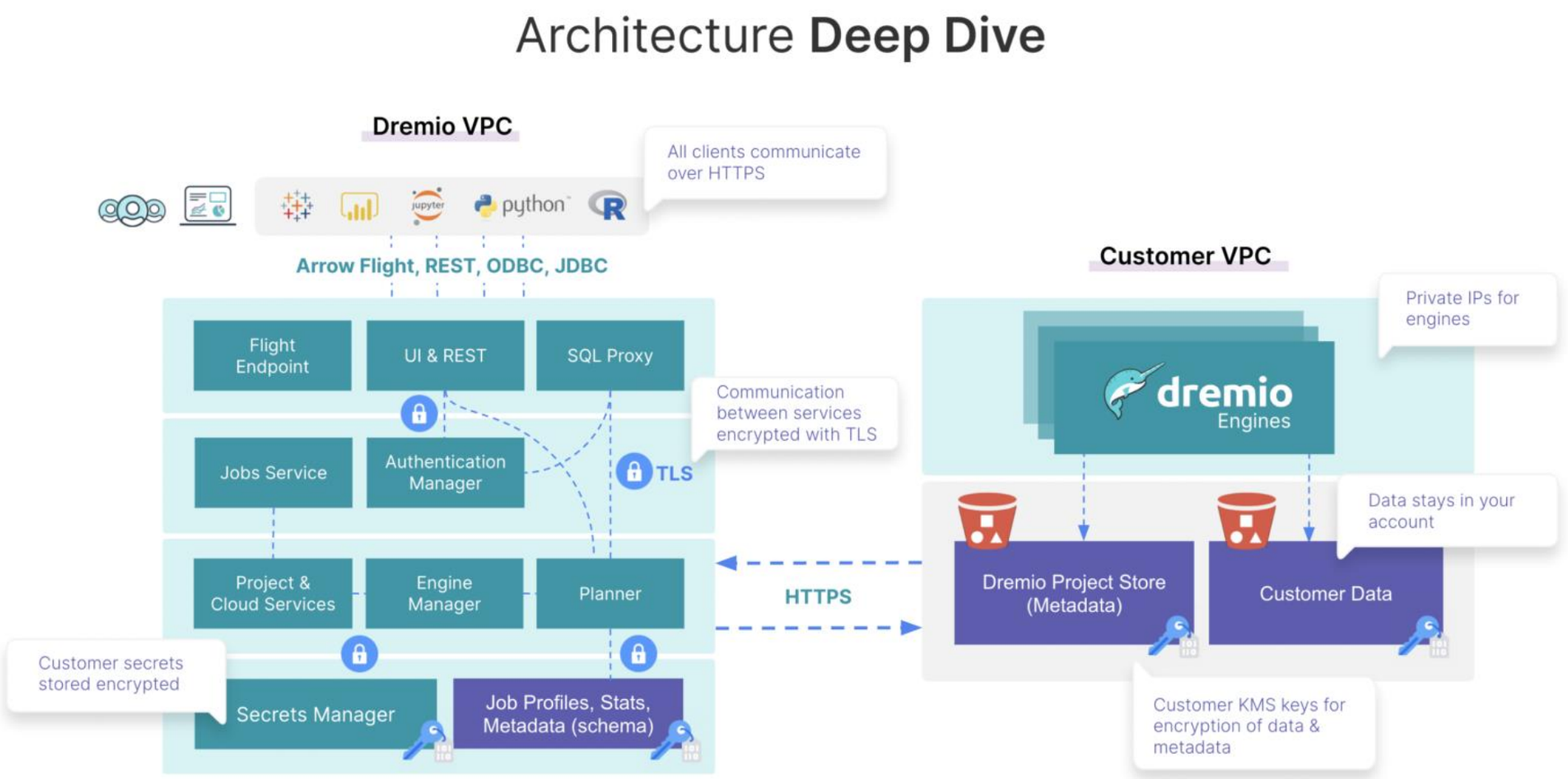
- **Enterprise (Cloud):**

“Includes everything in Standard edition, plus advanced security features (such as custom roles, enterprise IdPs, SCIM, and custom encryption keys) and enterprise support.”

Dremio Basic Cluster Architecture



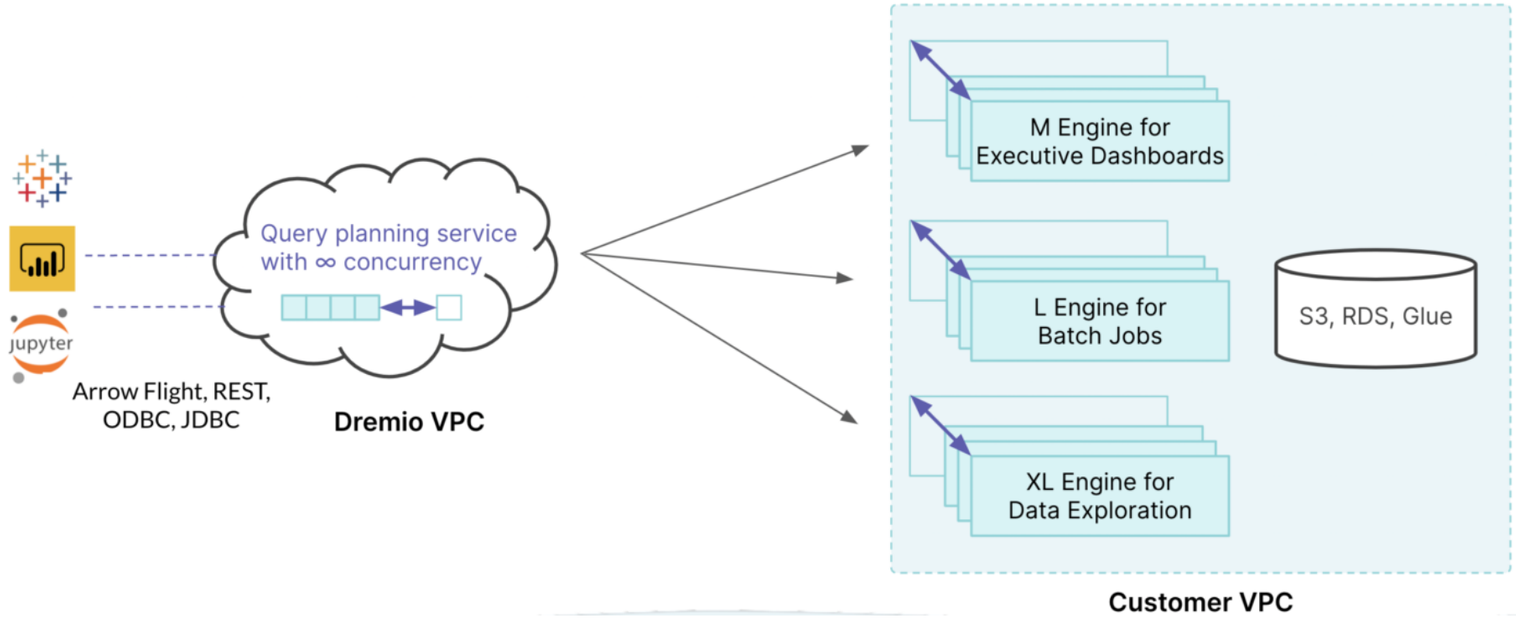
Dremio Cloud Architecture



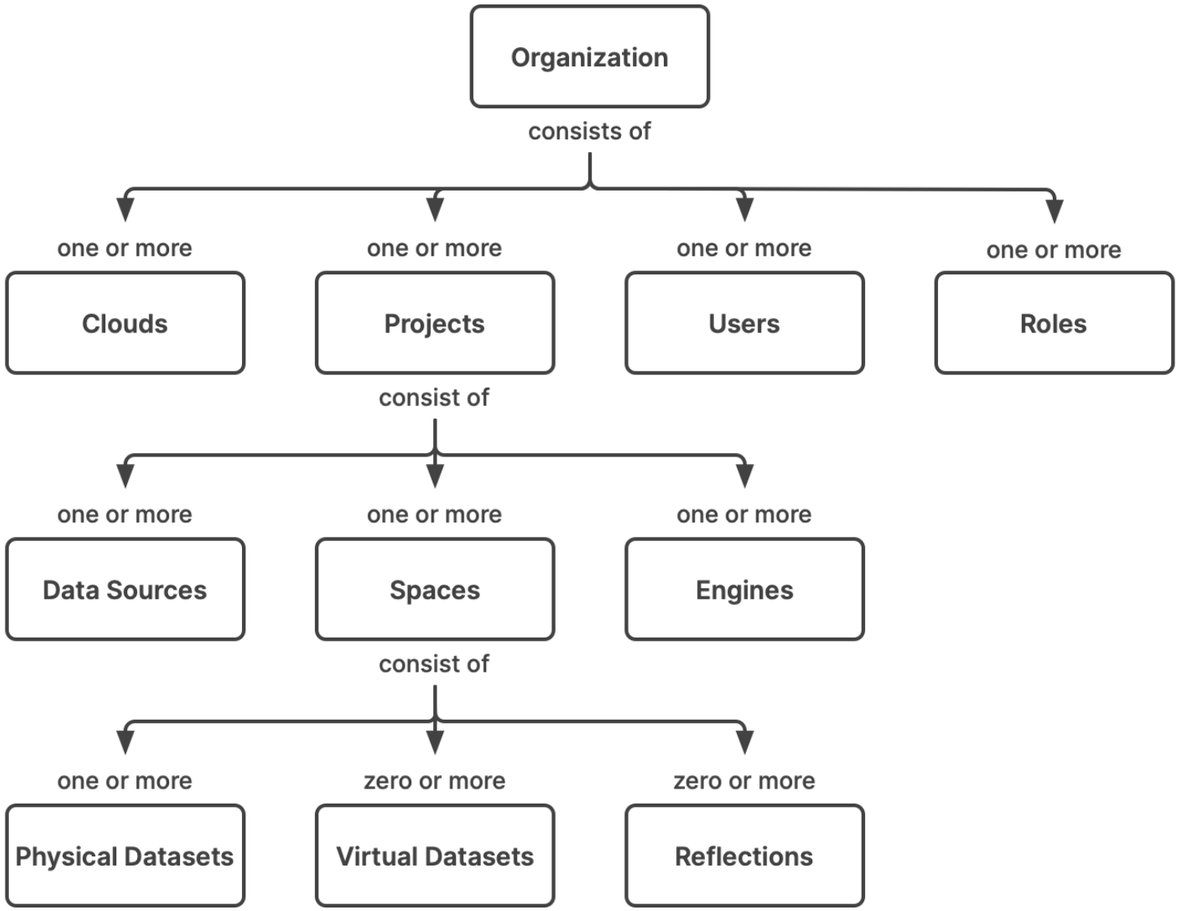
Dremio Cloud Principles

Frictionless, Infinite Scale, Enterprise-grade

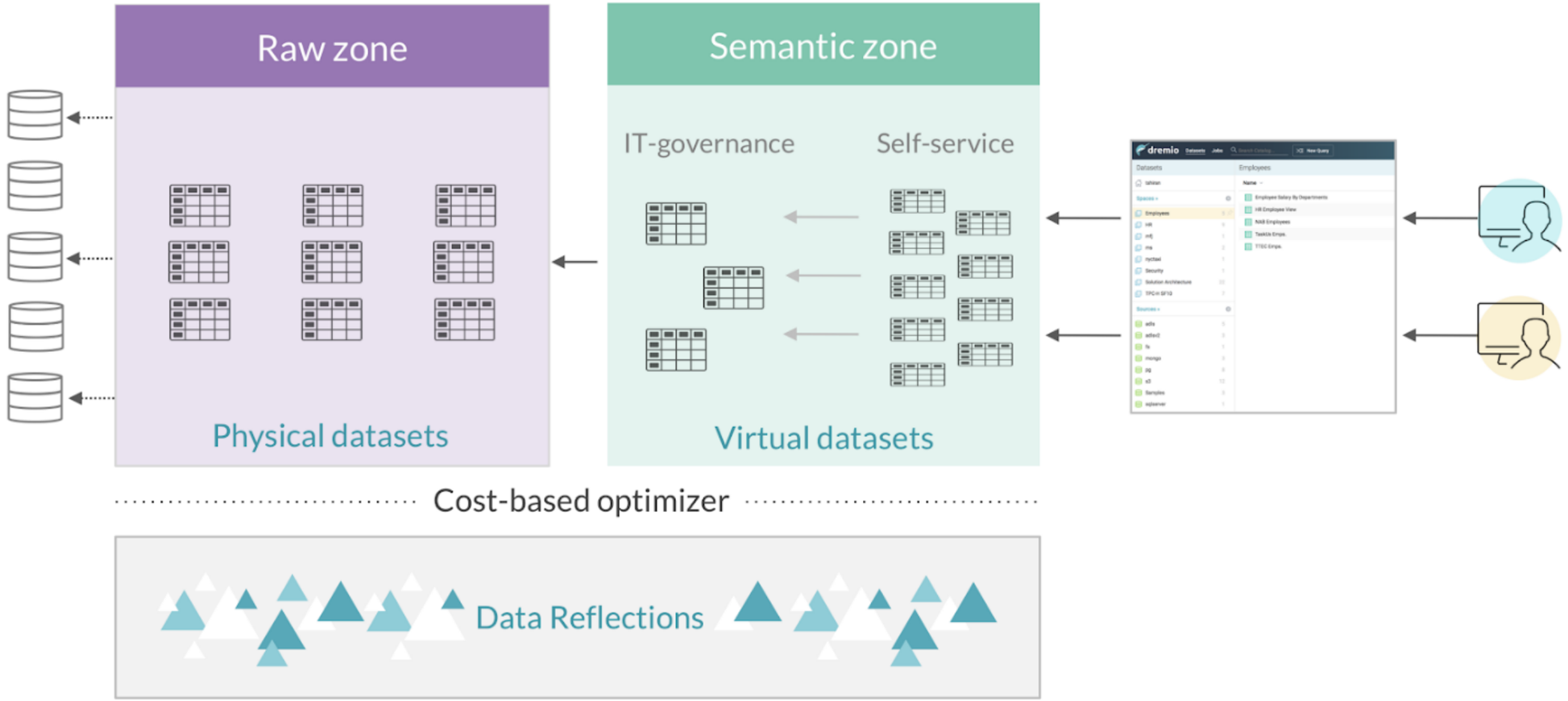
1. Passwordless, bi-directional BI integrations
2. Global control plane and query planner
3. Automatic rule-based query routing
4. Isolated auto-scaling execution engines
5. Data stays in company's cloud storage buckets



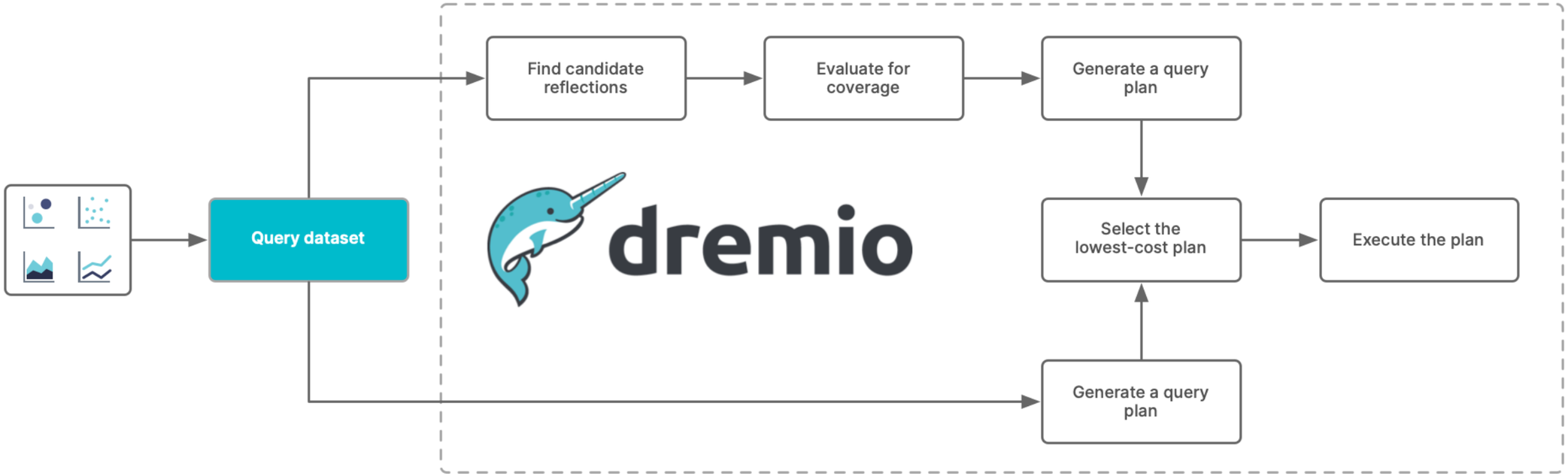
Objects in Dremio Cloud



Dremio Datasets and Reflections



Dremio Query Acceleration with Reflections



Dremio Client Interfaces – Web User Interface

- **Software:**

http://coordinator-hostname:9047

- **Cloud:**

https://app.dremio.cloud/

https://app.eu.dremio.cloud/

Dremio Client Interfaces – REST API

- **Software:**

http://coordinator-hostname/apiv2/login

http://coordinator-hostname/api/v3

- **Cloud:**

https://api.dremio.cloud/v0/

https://api.eu.dremio.cloud/v0/

Dremio Client Interfaces – JDBC

- **Software:**

jdbc:dremio:direct=coordinator-hostname:31010

- **Cloud:**

jdbc:dremio:direct=sql.dremio.cloud:443

jdbc:dremio:direct= sql.eu.dremio.cloud:443

Dremio Client Interfaces – ODBC Driver for Arrow Flight SQL

- **Software:**

coordinator-hostname:32010

- **Cloud:**

data.dremio.cloud:443

data.eu.dremio.cloud:443

Dremio Client Interfaces – Apache Arrow Flight SQL

- **Software:**

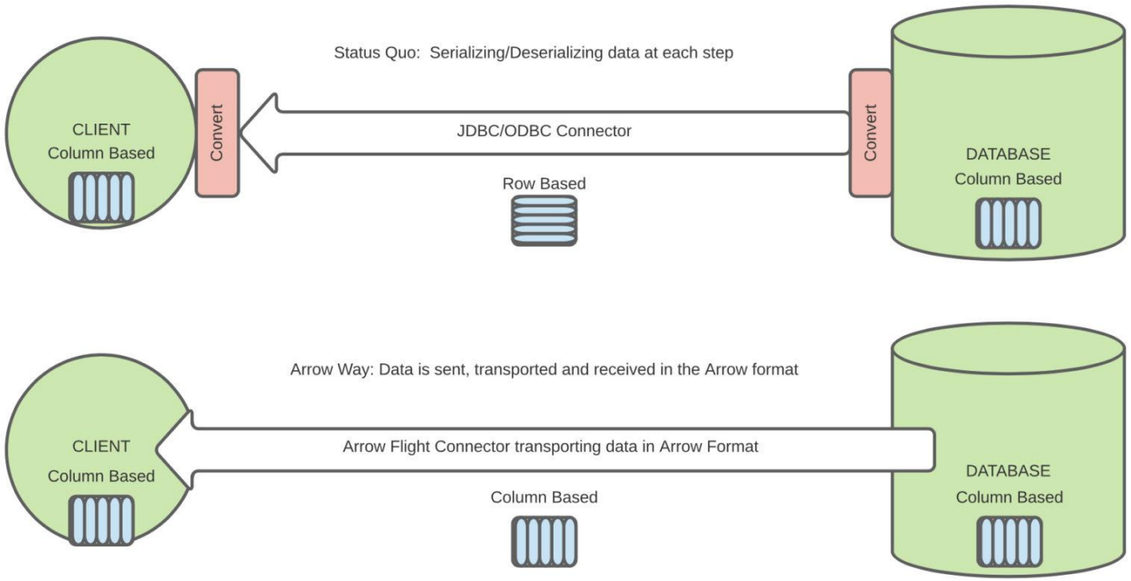
coordinator-hostname:32010

- **Cloud:**

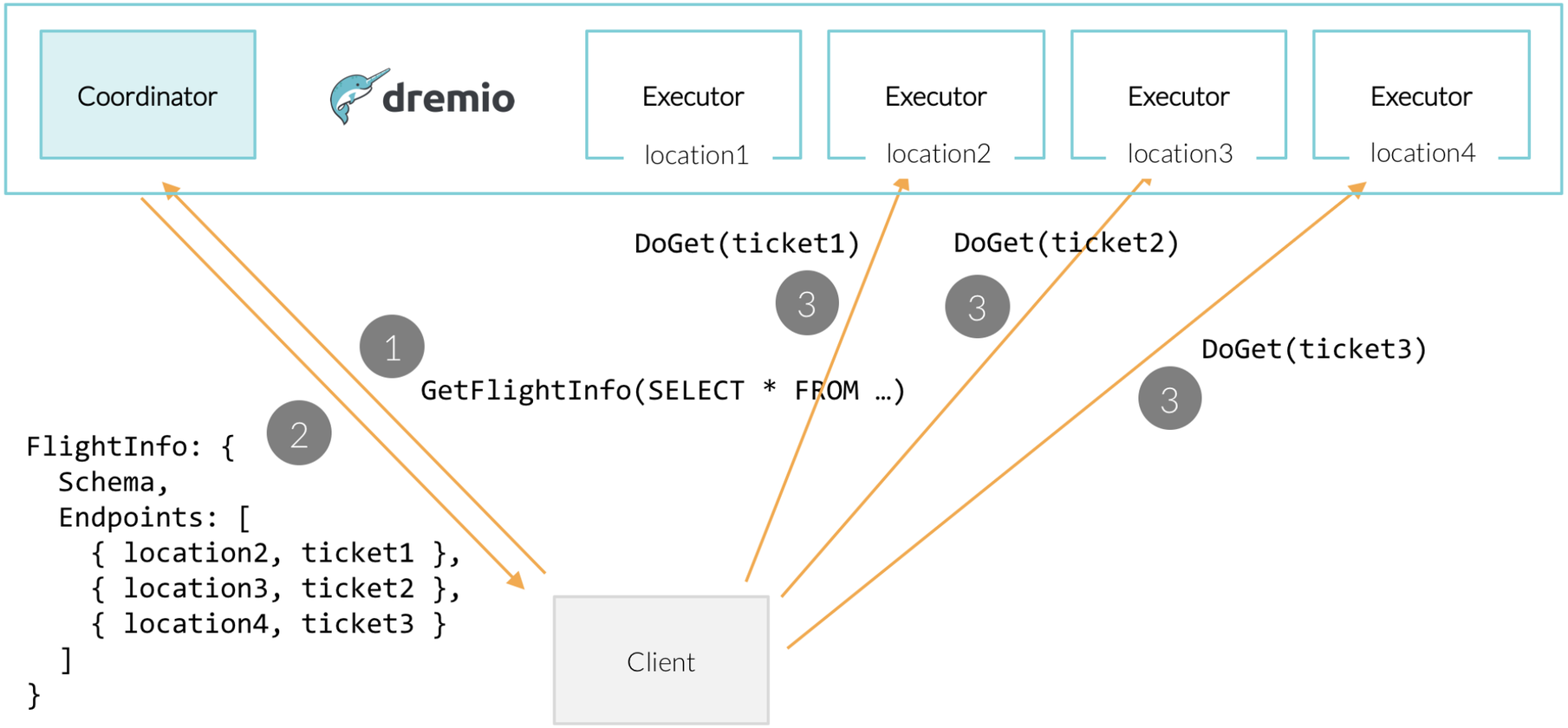
data.dremio.cloud:443

data.eu.dremio.cloud:443

ODBC/JDBC vs Apache Arrow Flight



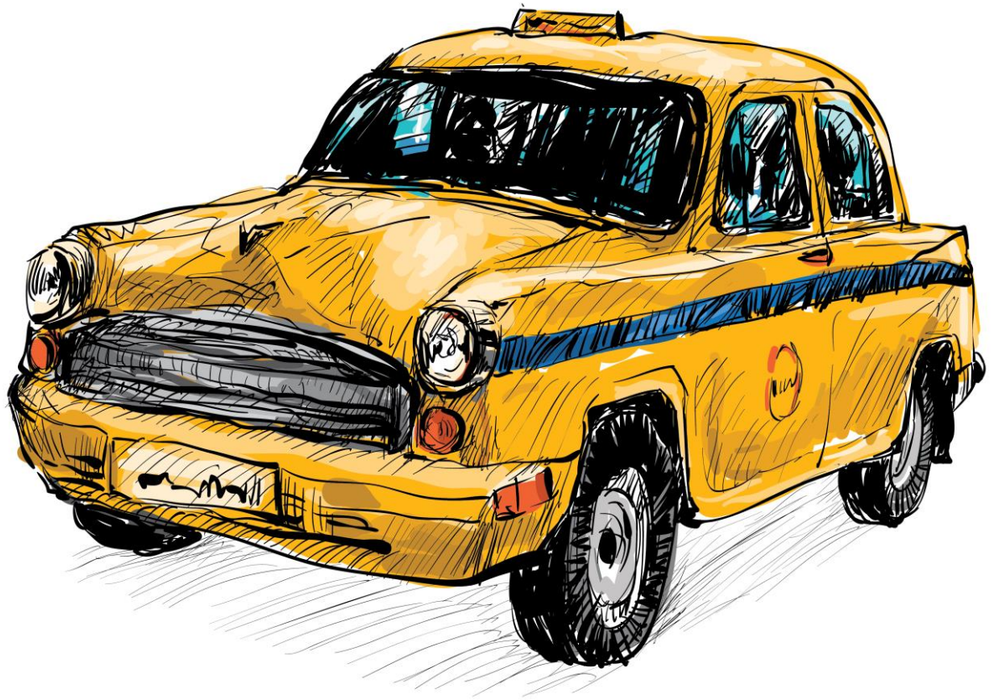
Apache Arrow Flight Workflow



Demo Time

How Dremio, Databricks and Superset can play together

NYC TLC Yellow Taxi Trip Records



Credits

Credits

- https://docs.dremio.com/cloud/api/
- https://docs.dremio.com/cloud/arrow-flight/
- https://docs.dremio.com/cloud/client-applications/jdbc/
- https://docs.dremio.com/cloud/getting-started/editions/
- https://docs.dremio.com/software/deployment/deployment-models/
- https://docs.dremio.com/software/deployment/standalone/standalone-cluster/
- https://docs.dremio.com/software/drivers/dremio-jdbc-driver/
- https://docs.dremio.com/software/rest-api/overview/
- https://www.dremio.com/blog/dremio-cloud-under-the-hood/
- https://www.dremio.com/blog/enabling-open-data-lakes-with-dremio-and-delta-sharing/

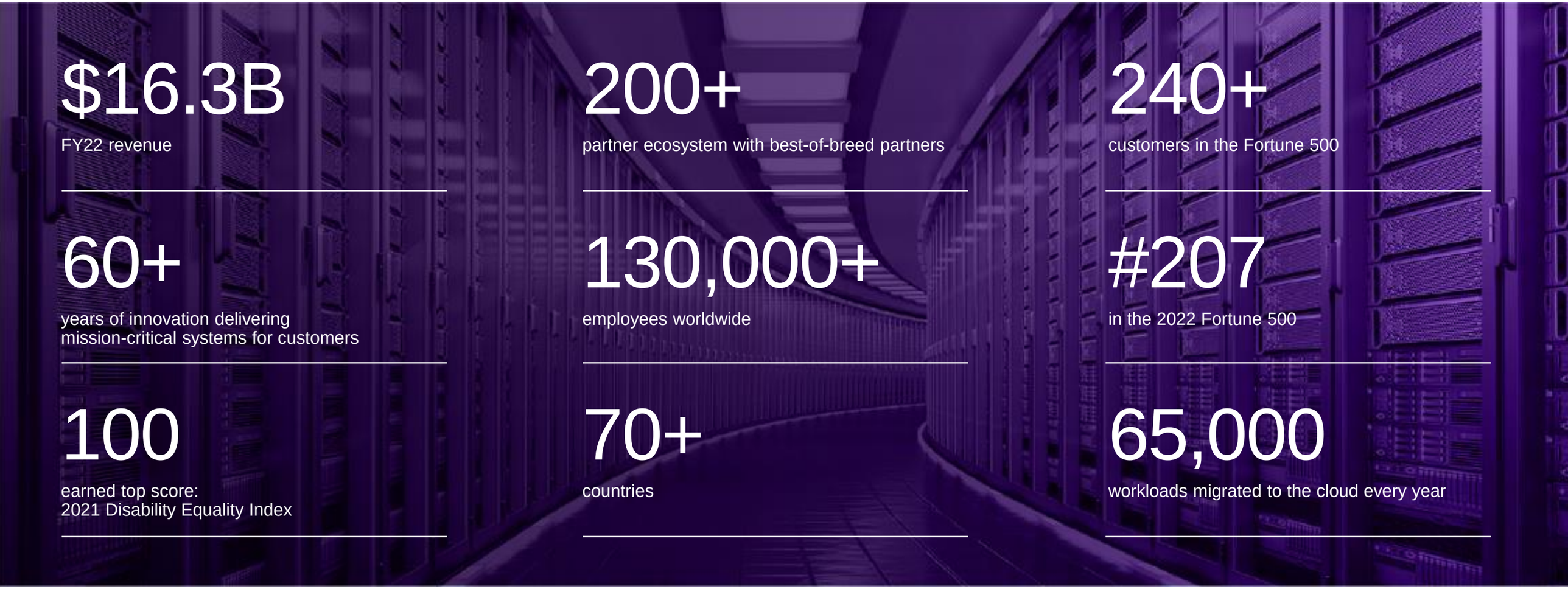
Credits

- https://www.dremio.com/blog/is-time-to-replace-odbc-jdbc/
- https://www.dremio.com/blog/what-is-a-data-lakehouse/
- https://docs.dremio.com/cloud/getting-started/editions/
- https://www.dremio.com/platform/sonar/
- https://www.dremio.com/resources/guides/apache-iceberg-an-architectural-look-under-the-covers/
- https://www.dremio.com/subsurface/an-introduction-to-apache-arrow-flight-sql/
- https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/
- https://www.databricks.com/glossary/data-lakehouse
- https://www.freepik.com/free-photo/busy-man-waiting-something-important_13133922.htm
- https://www.oracle.com/data-lakehouse/what-is-data-lakehouse/
- https://www.vecteezy.com/vector-art/1312450-color-sketch-of-an-old-taxi

Questions and answers

DXC at a glance

Fortune 500
DXC Technology RANK 207



$16.3B
FY22 revenue

200+
partner ecosystem with best-of-breed partners

240+
customers in the Fortune 500

60+
years of innovation delivering mission-critical systems for customers

130,000+
employees worldwide

#207
in the 2022 Fortune 500

100
earned top score:
2021 Disability Equality Index

70+
countries

65,000
workloads migrated to the cloud every year