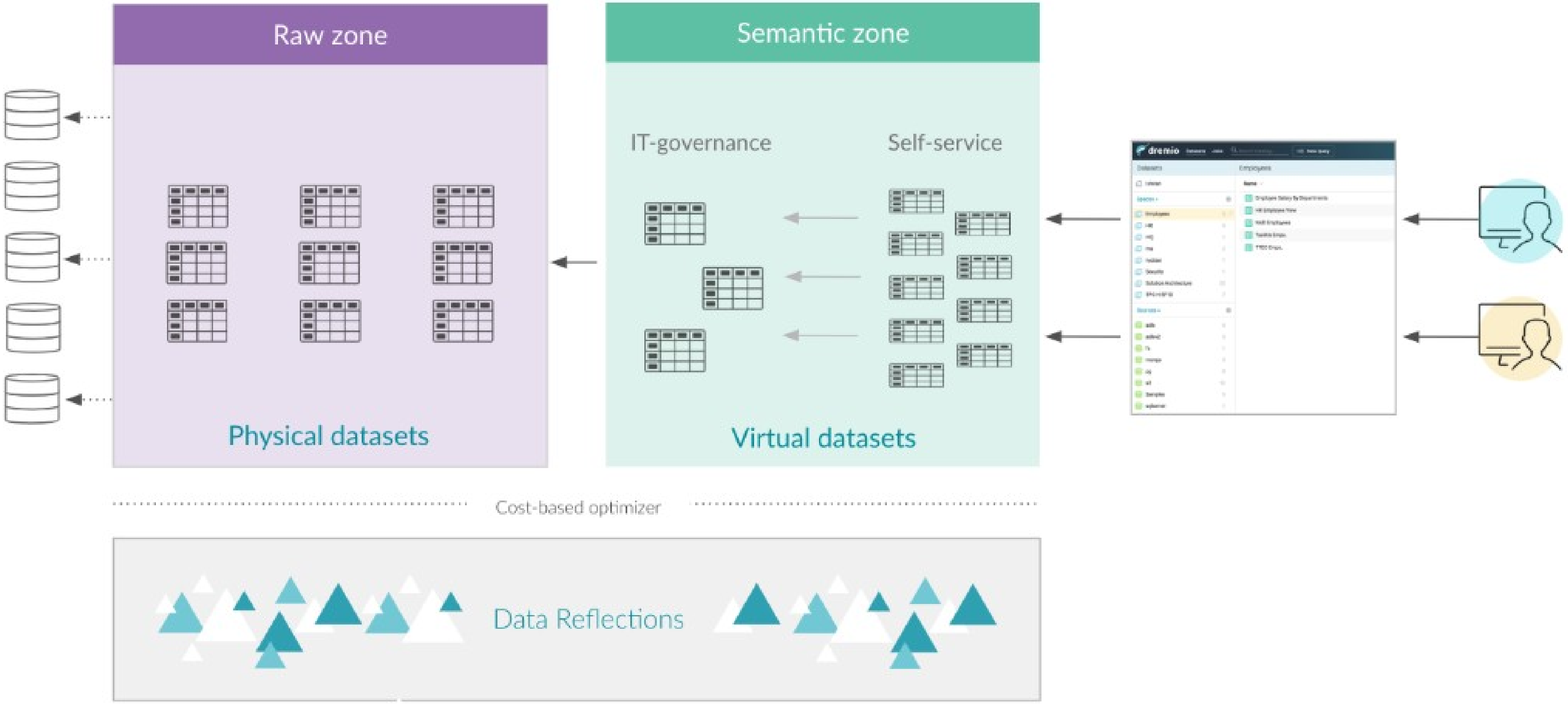
Dremio Introduction

What is Dremio?

- Dremio provides an almost “self-service” data platform that allows you to create virtual datasets from multiple sources
- Dremio acts as a read-only database while also providing simple configurations for modern data visualization tools, such as Power BI and Tableau
- “Dremio technologies like Data Reflections, Columnar Cloud Cache (C3) and Predictive Pipelining work alongside Apache Arrow to make queries on your data lake storage very, very fast”

dremio
The Data Lake Engine

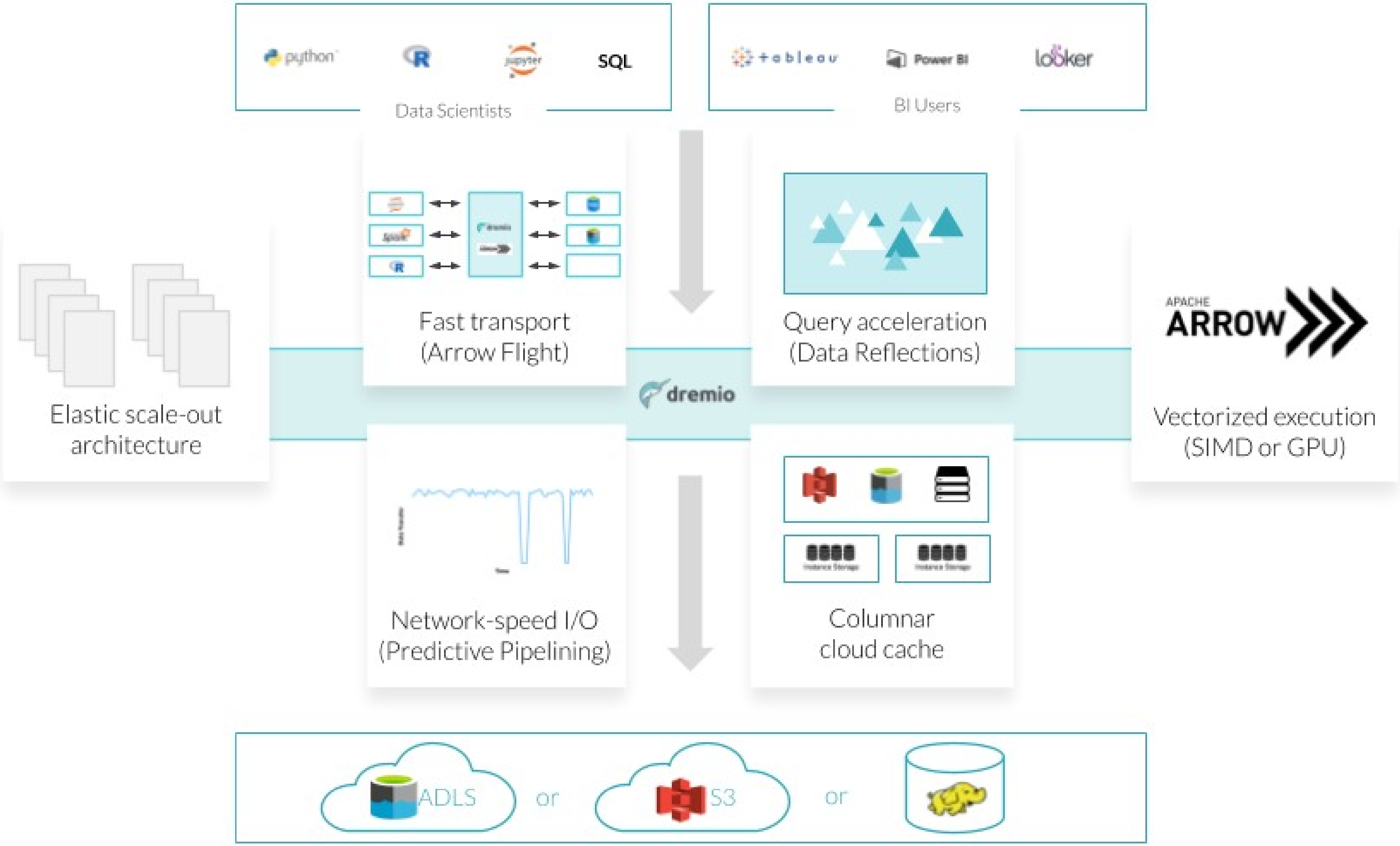
Dremio Logical Structure



Dremio Architecture

- Dremio appears just like a relational database,
- Dremio exposes ODBC, JDBC, REST and Arrow Flight interfaces.
- Easily connect any BI or data science tool e.g. Jupyter Notebooks, Power BI or Tableau.

Dremio Architecture



Data Scientists

BI Users

Fast transport (Arrow Flight)

Query acceleration (Data Reflections)

Elastic scale-out architecture

Vectorized execution (SIMD or GPU)

Network-speed I/O (Predictive Pipelining)

Columnar cloud cache

ADLS or S3 or

Deployments

In production, Dremio will be deployed as a cluster of nodes, each fulfilling a role:

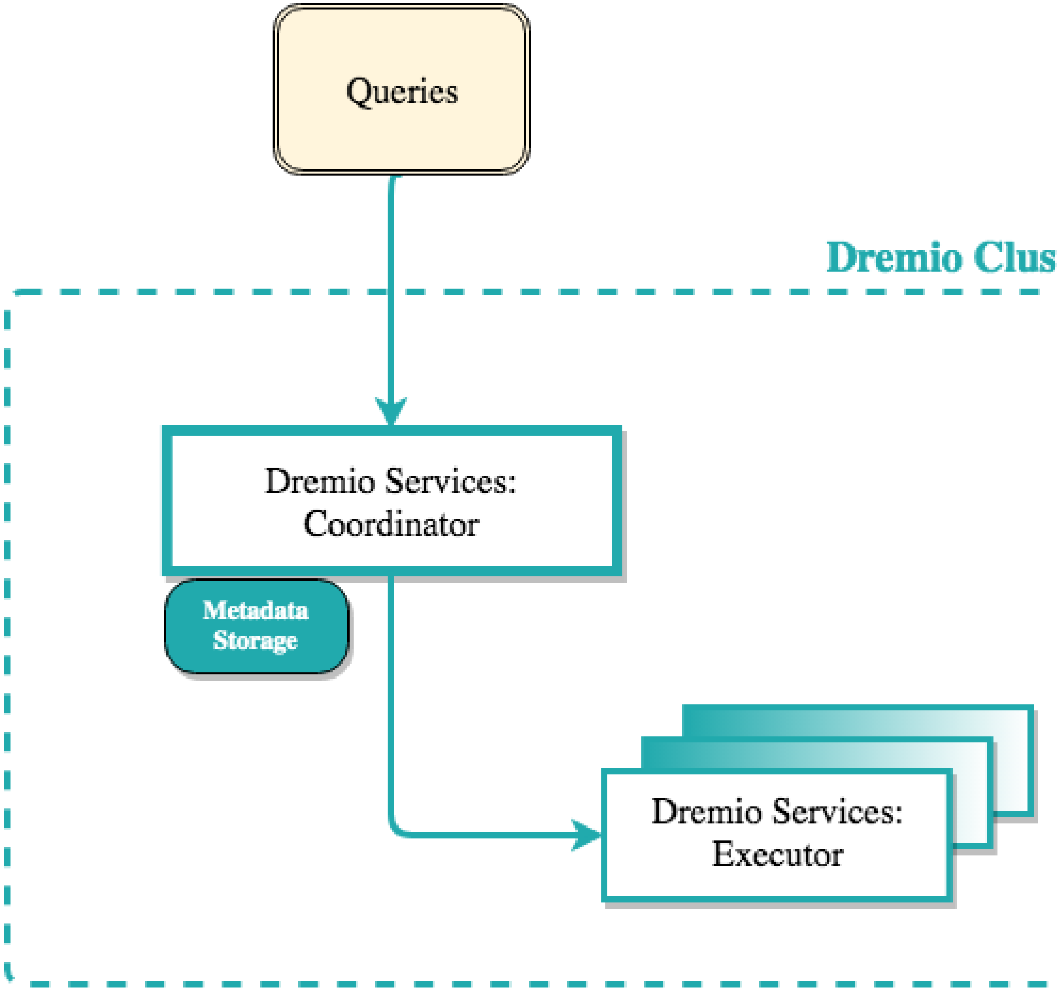
- Master Coordinator Role
  - Manages metadata, also responsible for:
    - Query planning
    - Serving Dremio's UI
    - Handling client connections, including the REST API
- Secondary Coordinator Role
  - Improve concurrency and distribute query planning for ODBC and JDBC client requests.
- Executor Role
  - Executor nodes execute queries.

Single Node Vs Cluster

- In single node deployments, both execution and coordination happens on the same node.

- In cluster deployments, a given node may only have a single role: either a coordinator or an executor. Multiple roles per node are not supported in cluster deployments.

Cluster Architecture



Queries

Dremio Cluster

Dremio Services:
Coordinator

Metadata
Storage

Dremio Services:
Executor

Physical Vs Virtual Datasets

- Dremio refers to the original (raw) source data as a “physical” dataset.
- Physical datasets cannot be modified by Dremio
- Virtual Datasets are derived from Physical datasets or other virtual datasets.
- Virtual datasets are defined by the steps needed for their creation, e.g. transformations, filters, joins etc.
- Virtual datasets are not copies of the physical dataset and so they use very little space.
- Virtual datasets will always reflect the current state of the physical datasets they are derived from.

Reflections

- Dremio accelerates data operations using “reflections”
- A reflection maintains one or more physically optimized representations of a dataset.
- Data Reflections are transparent to end users, so they can be added and revised without changing the SQL of client applications.
- The query optimizer can accelerate a query by utilizing one or more Data Reflections to partially or entirely satisfy that query, rather than processing the raw data in the underlying data source.

Types of Reflections

There are various types of Data Reflections:

- Raw reflections
  - These include one or more fields from the anchor dataset, sorted, partitioned and distributed by specific fields.
- Aggregation reflections
  - These include one or more dimension and measure fields from the anchor dataset, sorted, partitioned and distributed by specified fields.
- External reflections
  - An un-managed reflection, which allows users to leverage existing datasets and summary tables built in external system as reflections in Dremio.

Questions?

?

?

?

?

?

?