#RedHatOSD

RED HAT FORUMS

Dremio: data virtualization in action

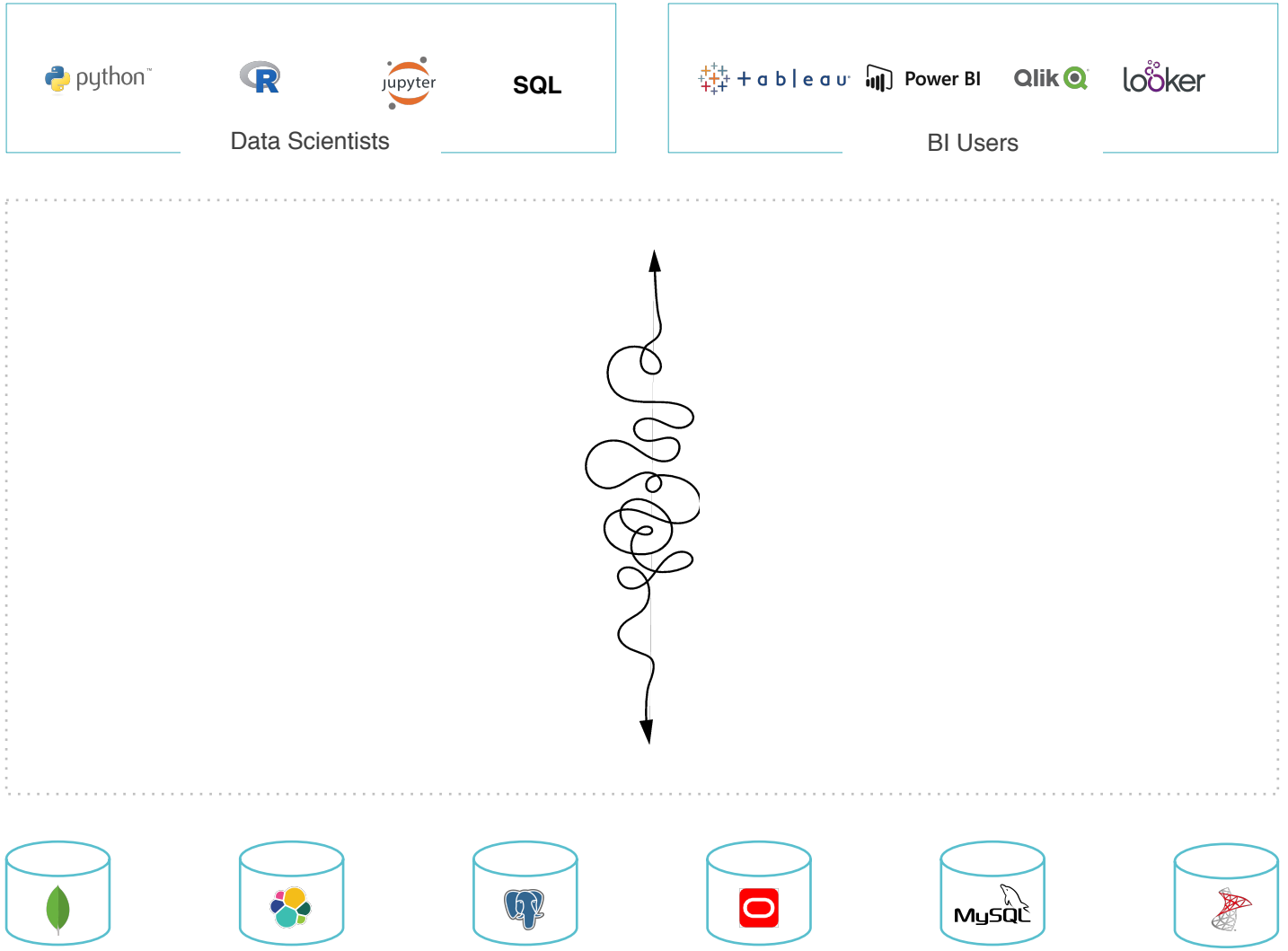
Stefano Pampaloni
Ceo - Seacom
3/12/19

Red Hat | RIOS RETE ITALIANA OPEN SOURCE SEACOM the leading open source architects

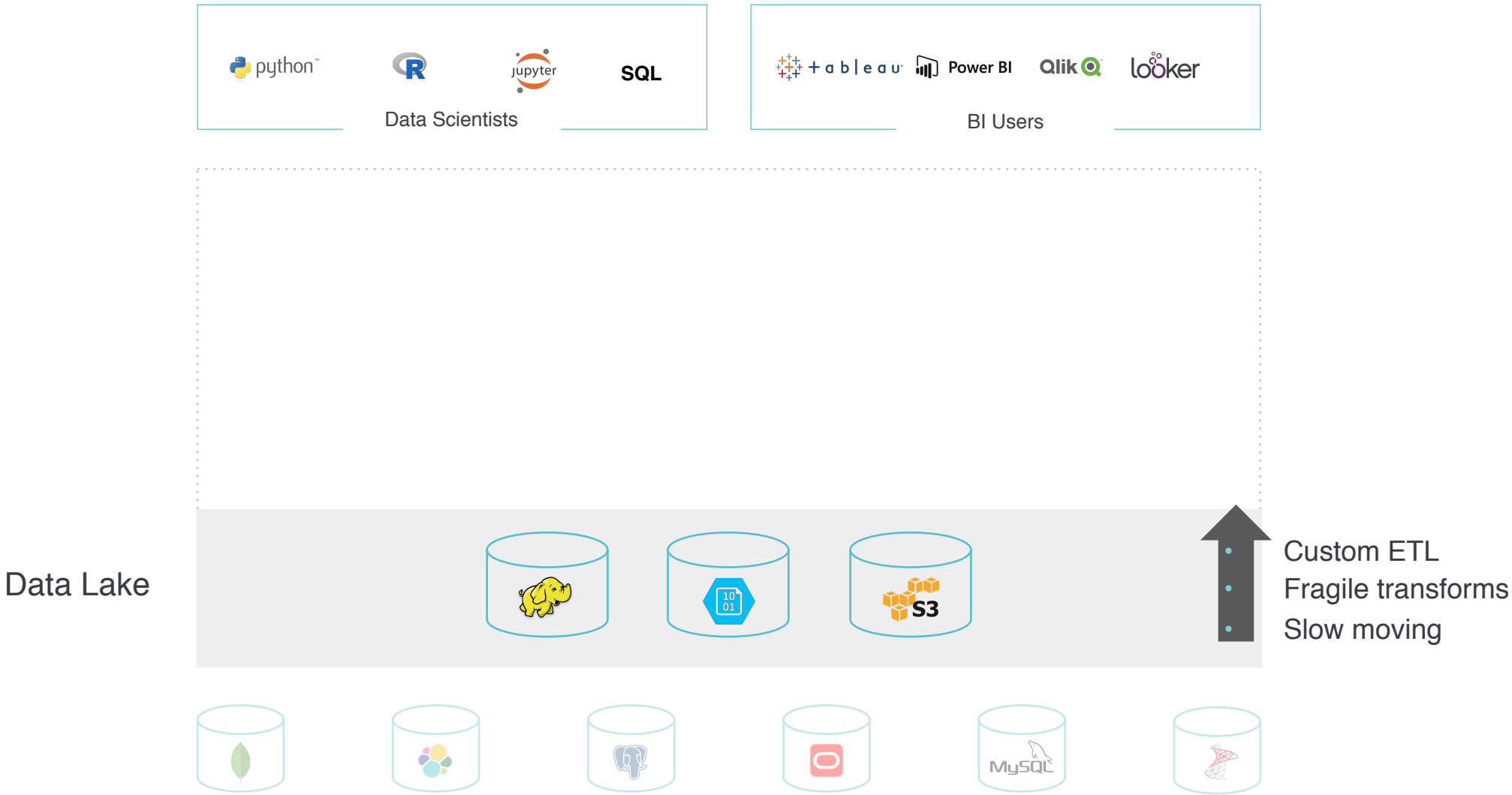
No moving data, no DWH, no cubes, no aggregation tables or extracts, No ETL and no nosql
Query SQL veloci direttamente sul Datalake, nessuna duplicazione dei dati, join fra sorgenti diverse, sicurezza e semantica centralizzata

Our view of the market

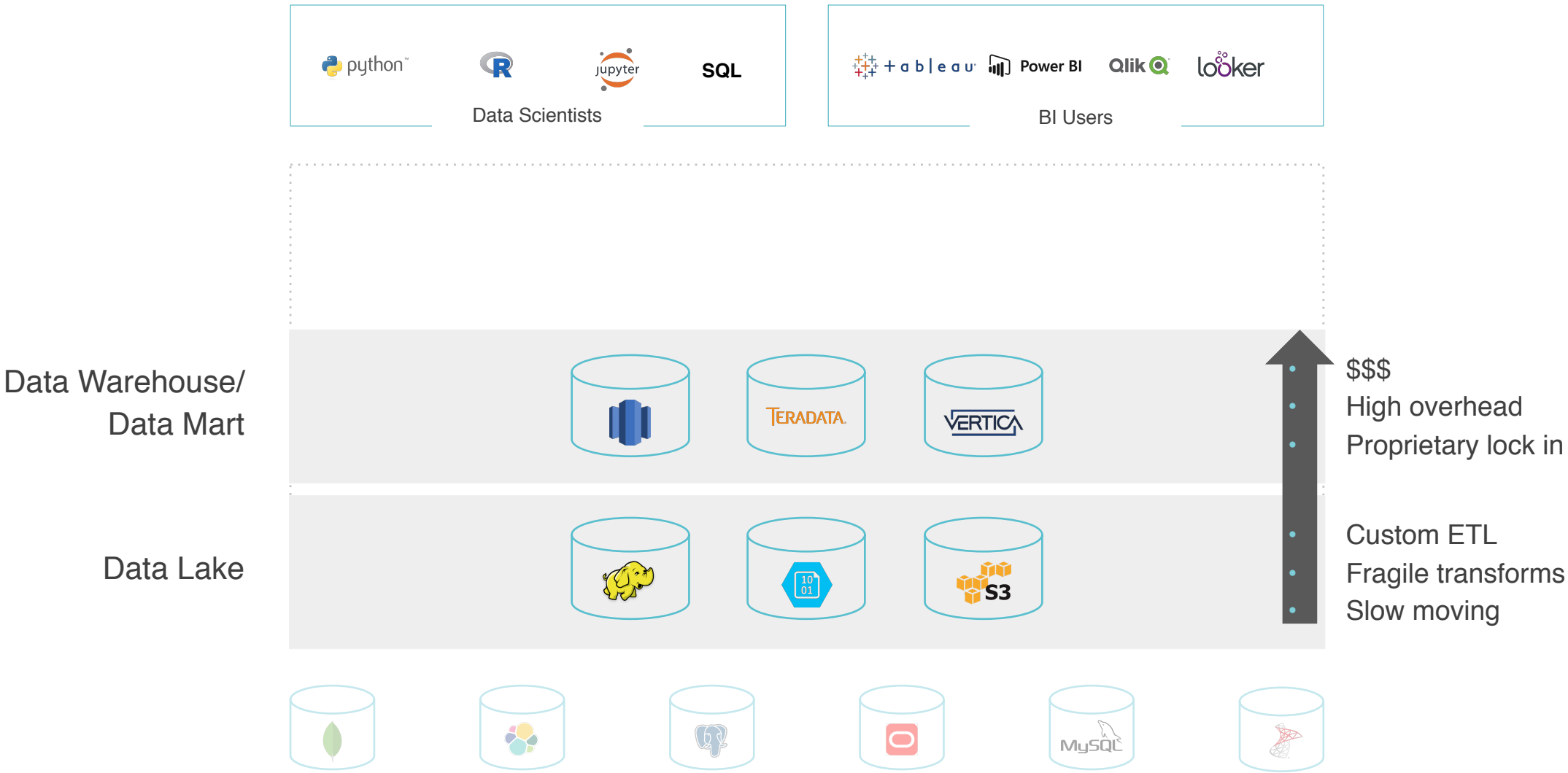
Your analysts and data scientists are hungry for data



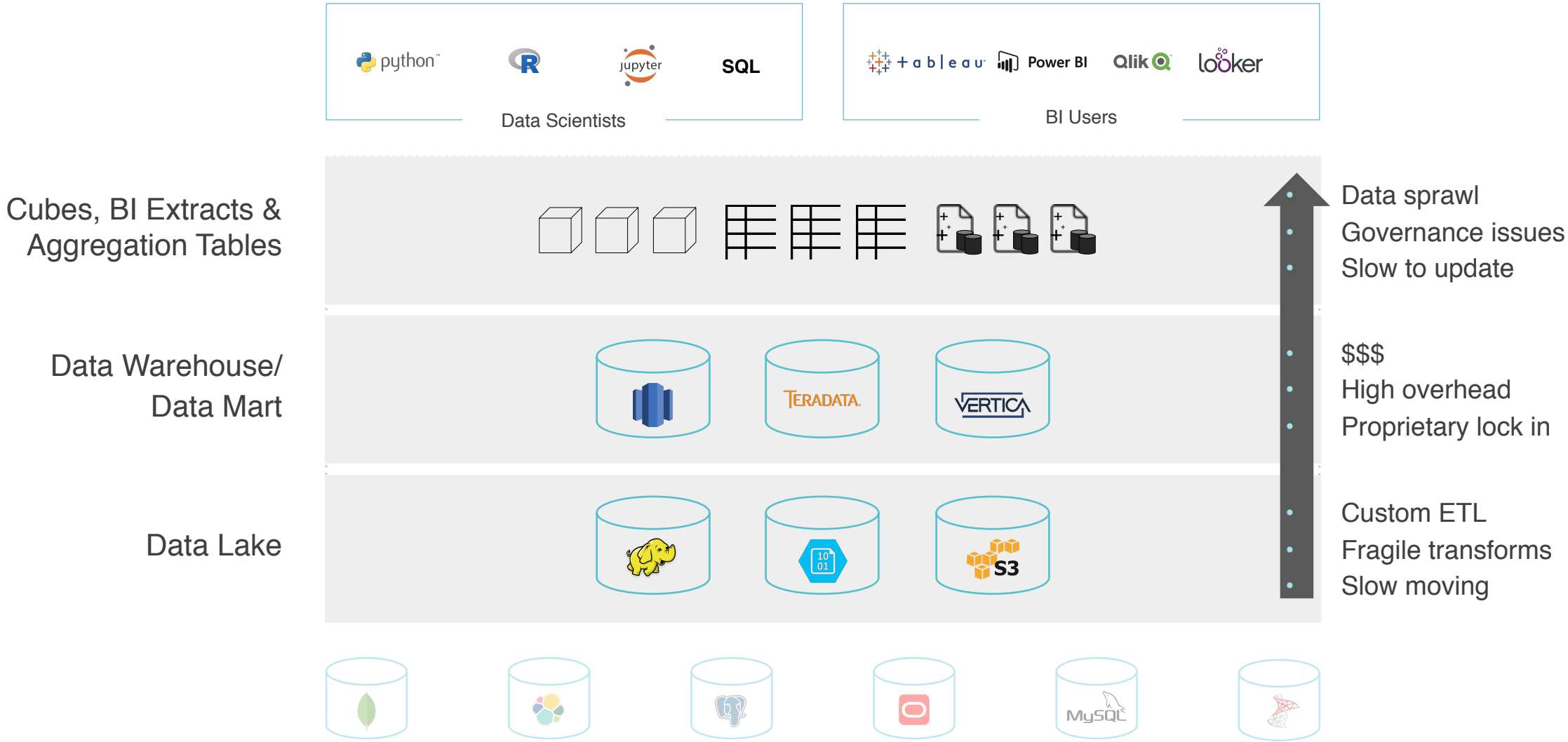
Data is a massive engineering project today



Data is a massive engineering project today



Data is a massive engineering project today



There's a better way,

- ✓ Works with any data source
- ✓ Works with any BI or Data Science tool
- ✓ 10x-1000x data acceleration
- ✓ Integrated Self-Service Platform
- ✓ Dynamic Security Controls
- ✓ Open source

Data Catalog

Data discovery interattiva, asset di livello enterprise

Data Curation

Combina, prepara e arricchisci i dati senza creare copie.

Data Acceleration

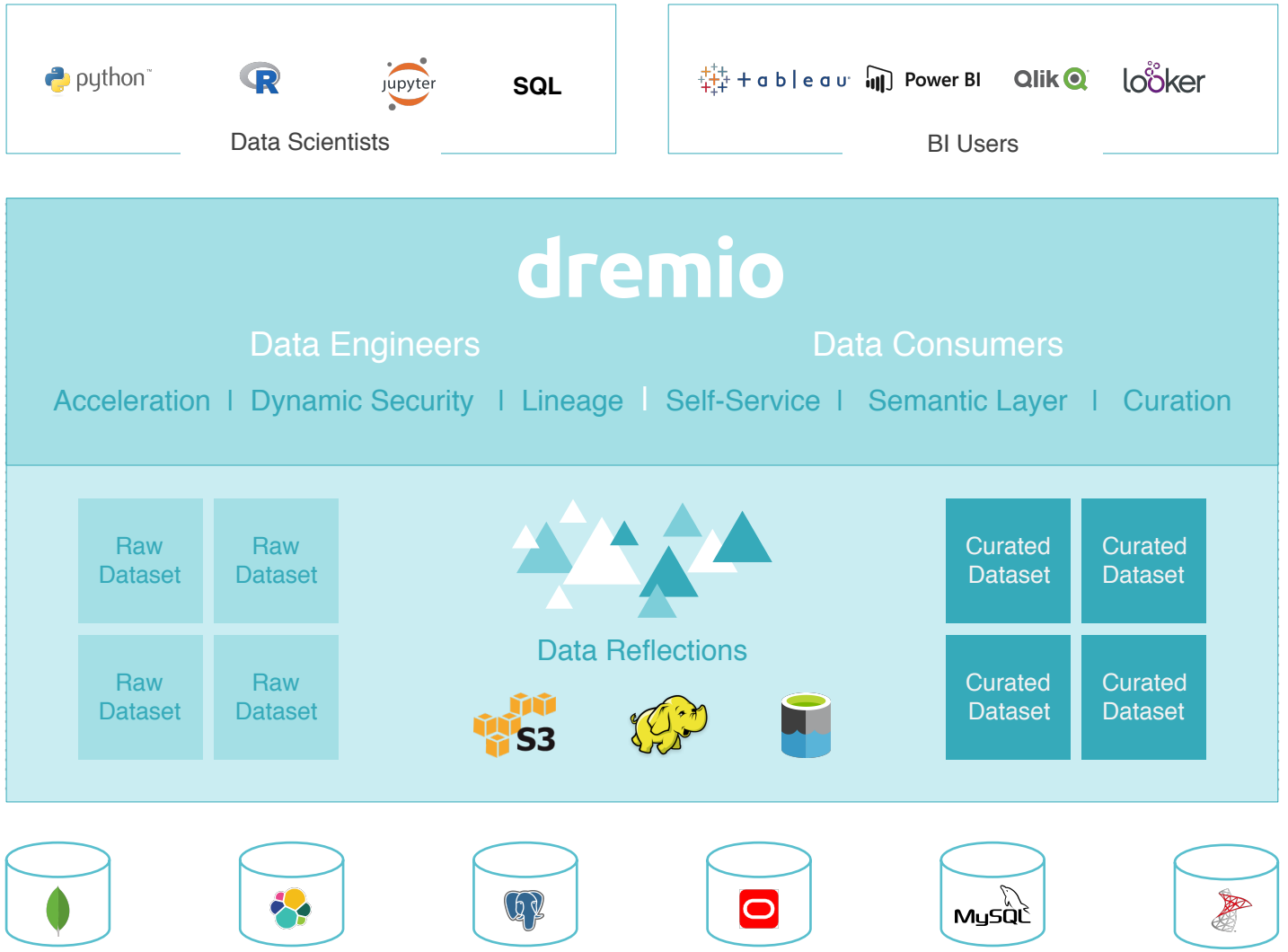
OLAP e query alla massima velocità

Senza utilizzare cubi o estrazioni di strumenti BI

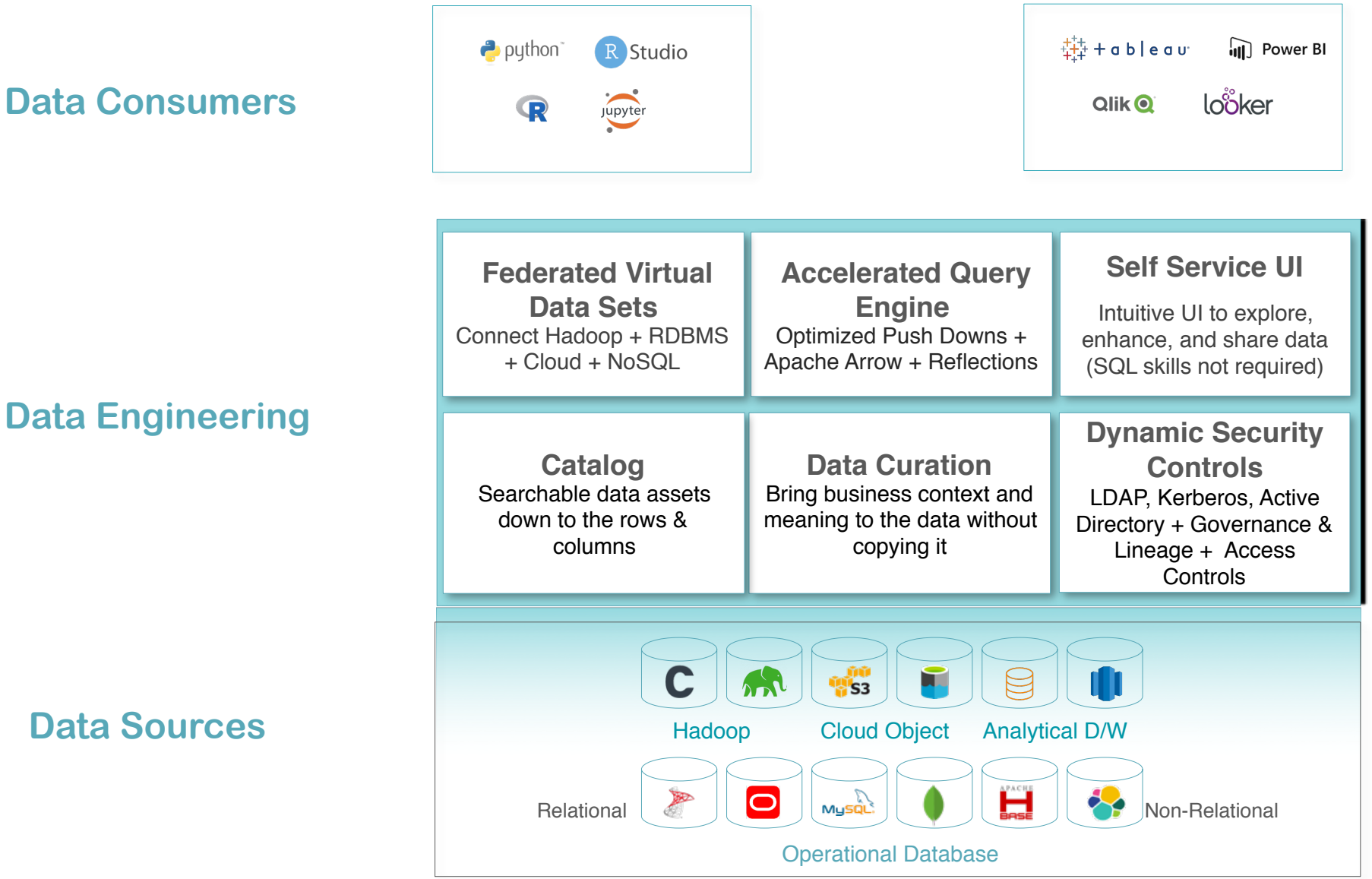
SQL Execution Engine

Motore basato su Apache Arrow, compatibile con ogni sorgente dati: RDBMS, NoSQL, S3, Hadoop

A new tier in data analytics: **Data-as-a-Service Platform**

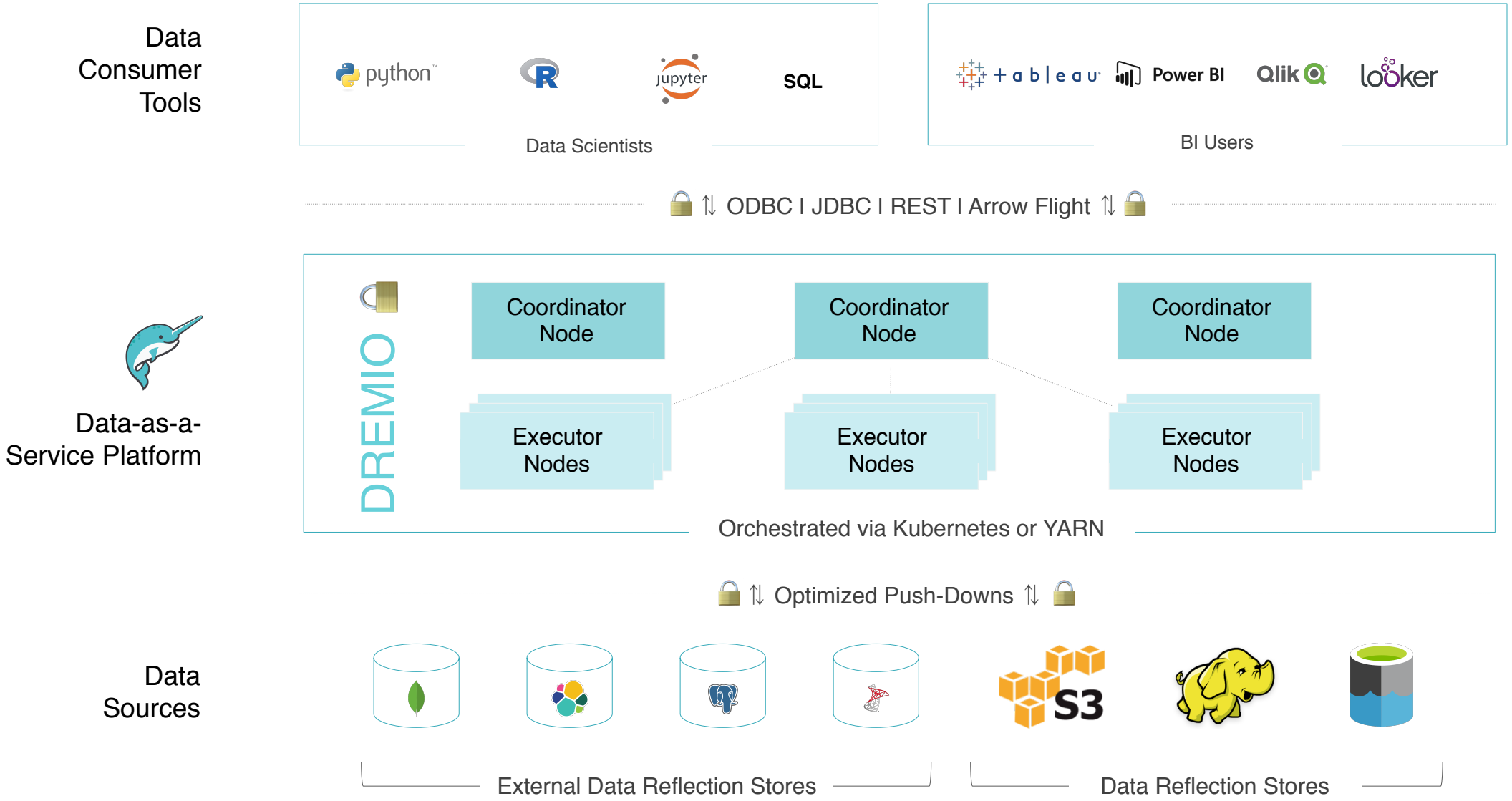


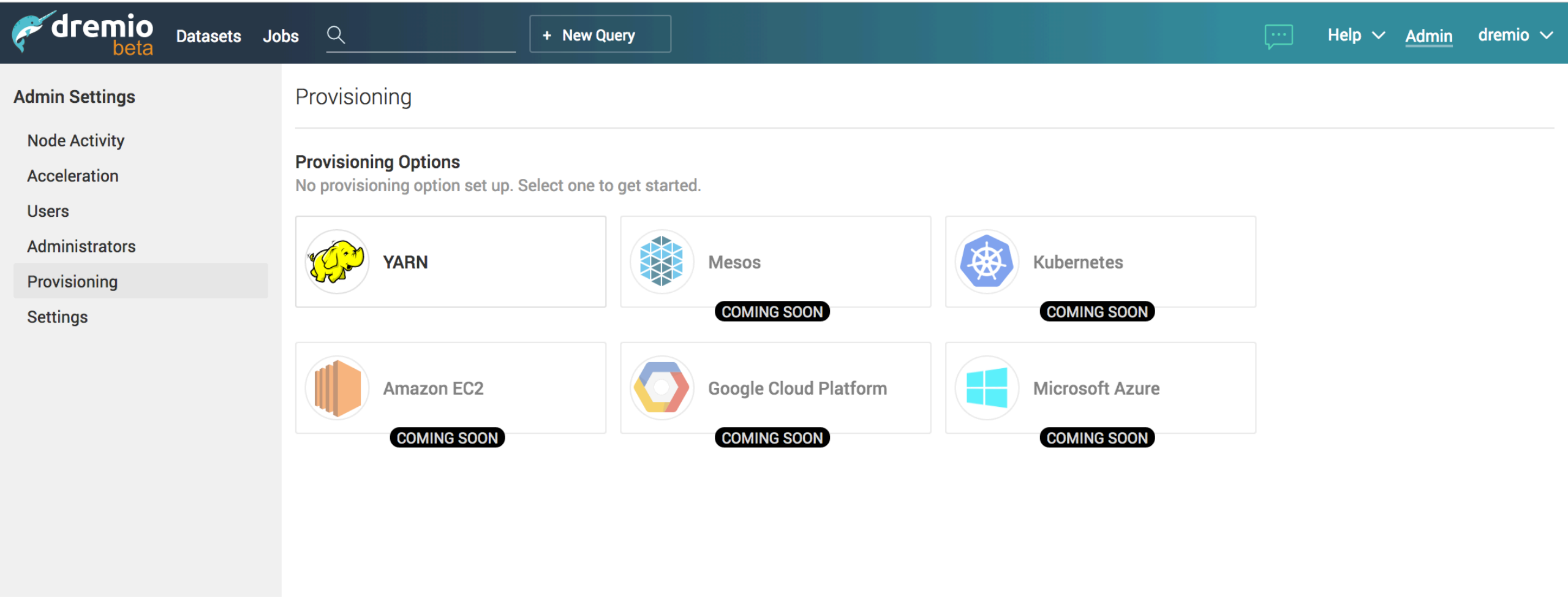
Dremio's Data-as-a-Service Platform



Architecture & Deployment

Deployment Architecture




dremio beta
Datasets
Jobs
+ New Query
Help
Admin
dremio
Admin Settings
Node Activity
Acceleration
Users
Administrators
Provisioning
Settings
Provisioning
Provisioning Options
No provisioning option set up. Select one to get started.
YARN
Mesos
COMING SOON
Kubernetes
COMING SOON
Amazon EC2
COMING SOON
Google Cloud Platform
COMING SOON
Microsoft Azure
COMING SOON

Why Dremio Wins

There's a better way,

- ✓ Vertically Integrated Query Engine
- ✓ Data Reflections
- ✓ Integrated Self-Service Platform
- ✓ Dynamic Security Controls
- ✓ Open Source

Vertically Integrated Query Engine

Optimized Push-Downs

Apache Arrow Execution

Data Reflections™

Data Reflections™

Data Reflections improve productivity & reduce data copies

TODAY

User picks best optimization

User-specific Cubes, BI Extracts, Aggregation Tables

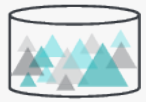
Domain-specific Data Marts

Source Data

DREMIO

Dremio picks best optimization

Shared Experience Model



Data Reflections

Source Data

Data Reflections get you there faster by going a shorter distance

Like an Index

- ✓ Users don't change their behavior, dashboards, etc
- ✓ Dramatically improves query efficiency
- ✓ Many options for different optimizations
- ✓ Queries can be resolved without push-down
- ✓ System-managed maintenance, security

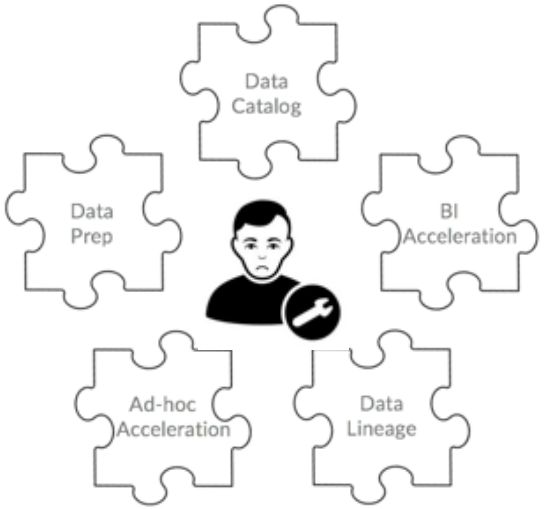
Like a materialized view

- ✓ Full power of SQL – join, filter, transform, etc
- ✓ Row-level fidelity, or summarized data
- ✓ Pre-computed aggregations, joins
- ✓ Pre-sort, pre-partition
- ✓ Columnarized, compressed on file system

#RedHatOSD

Integrated Self-Service Platform

Integrated Self-Service Platform



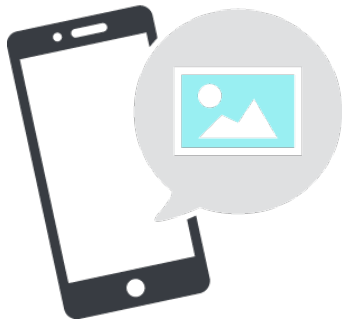
Niche Products

- x Limited to “weakest link”
- x IT-operated, no self-service
- x Complex, manual integrations
- x Proprietary technologies
- x High TCO

Dremio

- ✓ Greater than sum of its parts
- ✓ Self-service, faster time to insight
- ✓ Low operational overhead
- ✓ No vendor lock in
- ✓ Low TCO

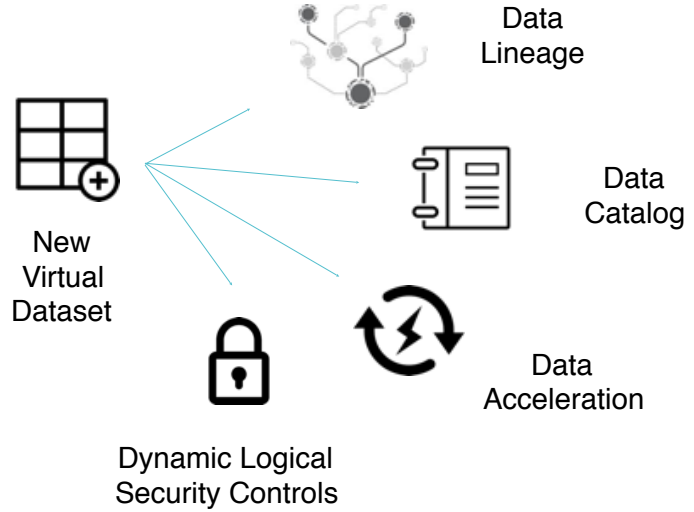
More than the sum of its parts



iPhone

When you take a great photo …

- ✓ Location tagged
- ✓ One click share with family on Facebook
- ✓ Automatic archive to Google Photos
- ✓ Search on people, place, thing in photo



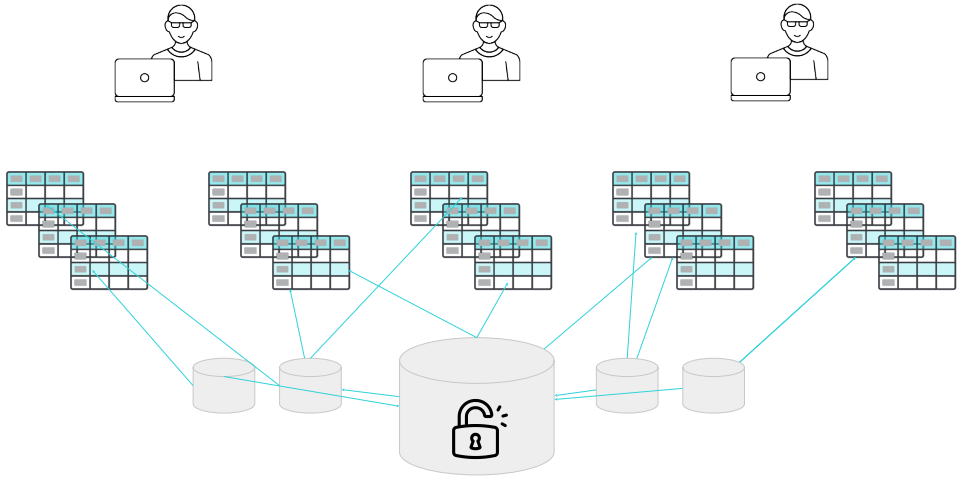
Dremio

When you create a Virtual Dataset …

- ✓ Data Provenance & Lineage automatically tracked
- ✓ Searchable in Data Catalog
- ✓ Accelerated by ancestor Data Reflections
- ✓ Dynamic security controls applied at runtime

Dynamic Security Controls

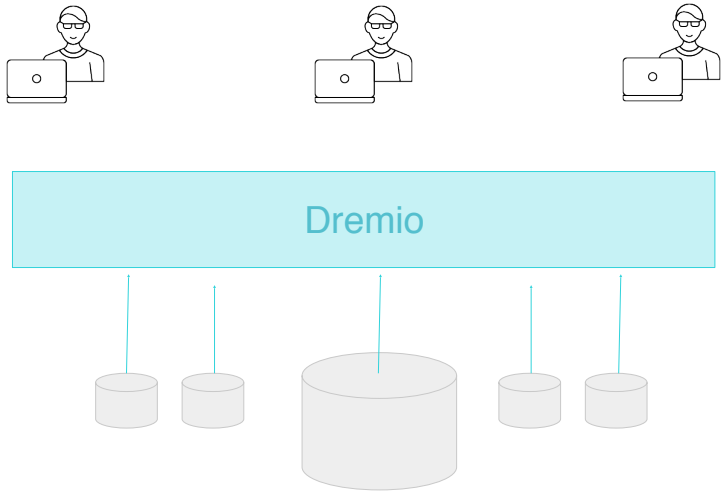
Control your risk around Shadow Analytics



Physical security controls

Users are taking matters into their own hands:

- x Copying data into spreadsheets
- x Downloading into 3rd party storage services
- x Taking outside your enterprise access controls

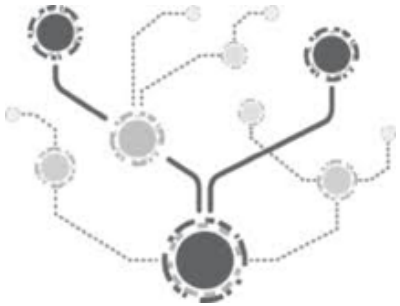


Attract more bees with honey

Keep users coming back to a governed, secure environment

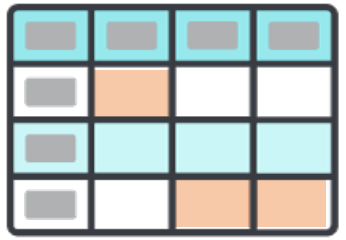
- ✓ Self-service: removes reasons to “go-around”
- ✓ Simple, intuitive access for all approved datasets
- ✓ Great experience with any tool, high performance

Secure and govern analytics with confidence



Automatic Data Provenance & Lineage

- ✓ Across all enterprise sources, virtual datasets
- ✓ Easily analyze patterns of use across users, tools
- ✓ Evaluate “what if” scenarios for downstream analysis
- ✓ Query time determination of access control behavior



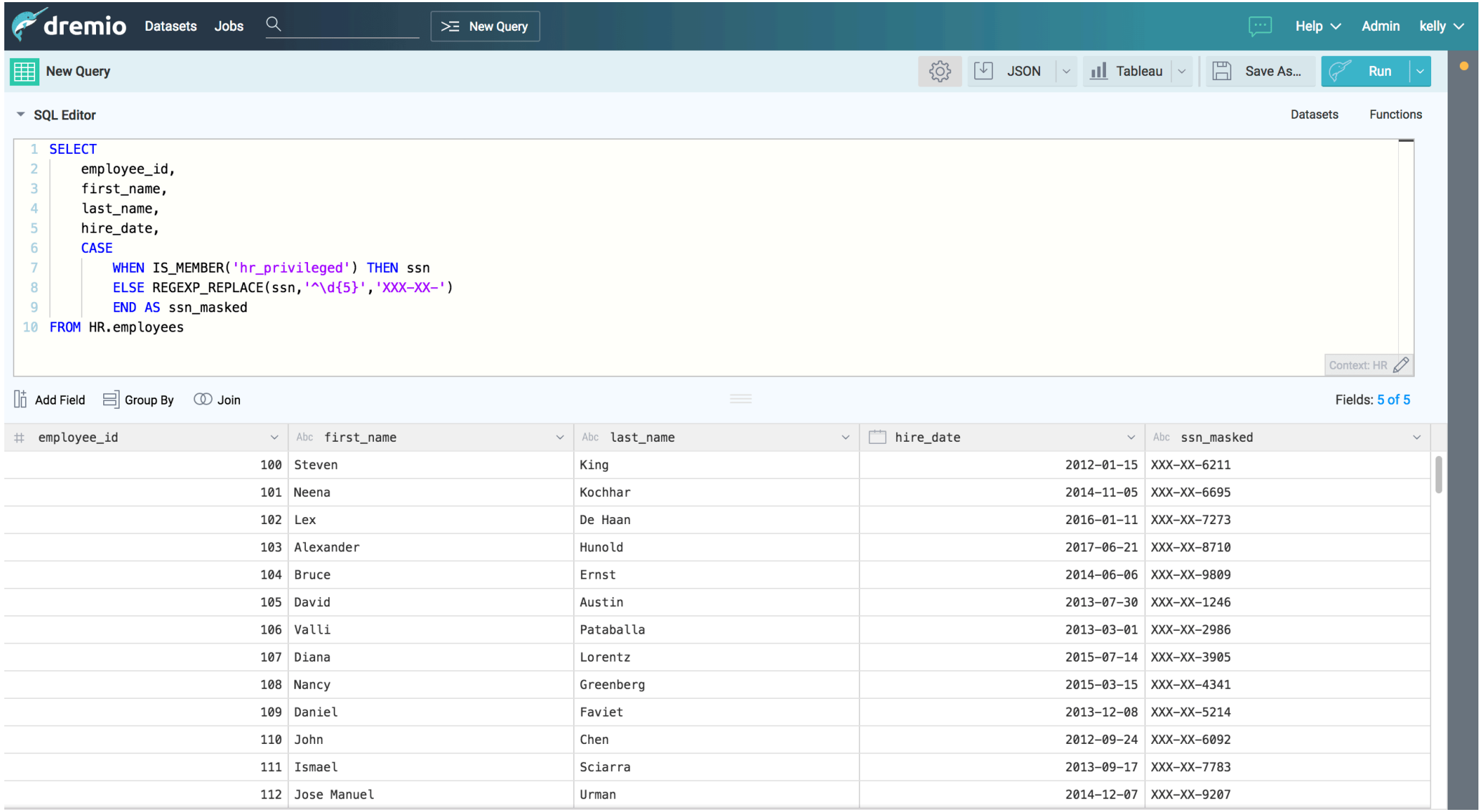
Row & Column-level Access Control

- ✓ Query time determination via RBAC
- ✓ Row & column level controls for all sources
- ✓ Dynamic masking based on LDAP/AD membership
- ✓ Extend the capabilities of all your sources

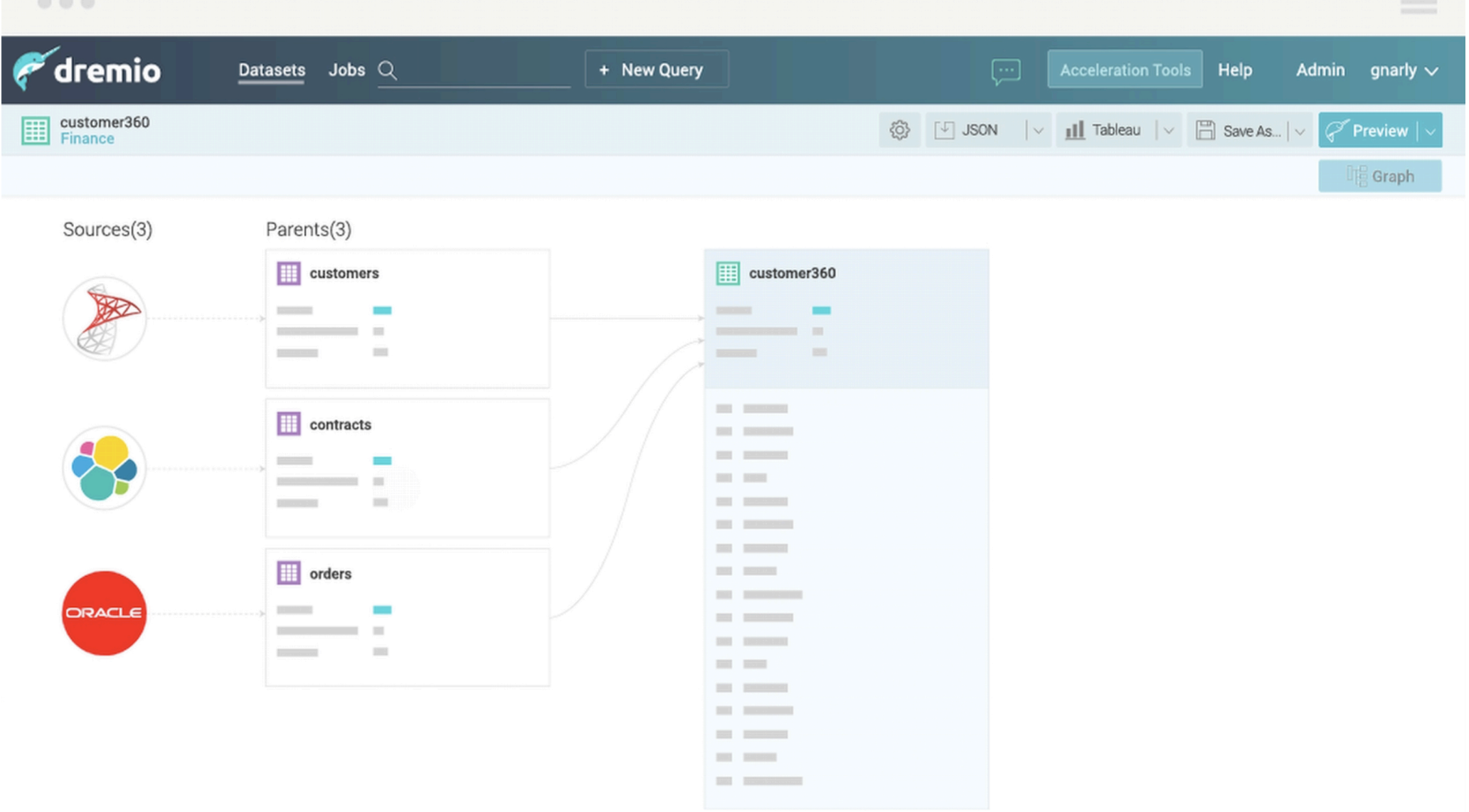
Masking & row/ column-level access controls

```
1  SELECT
2      employee_id,
3      first_name,
4      last_name,
5      hire_date,
6      CASE
7          WHEN IS_MEMBER('hr_privileged') THEN ssn
8          ELSE REGEXP_REPLACE(ssn,'^\d{5}','XXX-XX-')
9          END AS ssn_masked
10 FROM HR.employees
```

- ✓ Any source: S3, ADLS, HDFS, RDBMS, etc.
- ✓ IS_MEMBER() integrated with LDAP
- ✓ Dynamic evaluation
- ✓ Masking, transformation, joins
- ✓ SQL-based

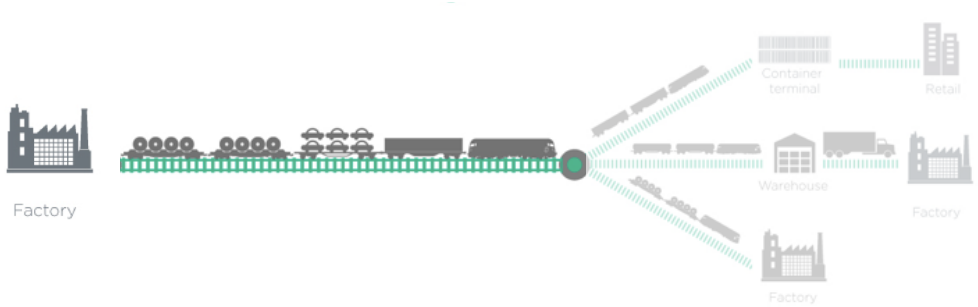


employee_id	first_name	last_name	hire_date	ssn_masked
100	Steven	King	2012-01-15	XXX-XX-6211
101	Neena	Kochhar	2014-11-05	XXX-XX-6695
102	Lex	De Haan	2016-01-11	XXX-XX-7273
103	Alexander	Hunold	2017-06-21	XXX-XX-8710
104	Bruce	Ernst	2014-06-06	XXX-XX-9809
105	David	Austin	2013-07-30	XXX-XX-1246
106	Valli	Pataballa	2013-03-01	XXX-XX-2986
107	Diana	Lorentz	2015-07-14	XXX-XX-3905
108	Nancy	Greenberg	2015-03-15	XXX-XX-4341
109	Daniel	Faviet	2013-12-08	XXX-XX-5214
110	John	Chen	2012-09-24	XXX-XX-6092
111	Ismael	Sciarra	2013-09-17	XXX-XX-7783
112	Jose Manuel	Urman	2014-12-07	XXX-XX-9207


dremio
Datasets
Jobs
+ New Query
Acceleration Tools
Help
Admin
gnarly
customer360
Finance
JSON
Tableau
Save As...
Preview
Graph
Sources(3)
Parents(3)
customers
contracts
orders
ORACLE
customer360

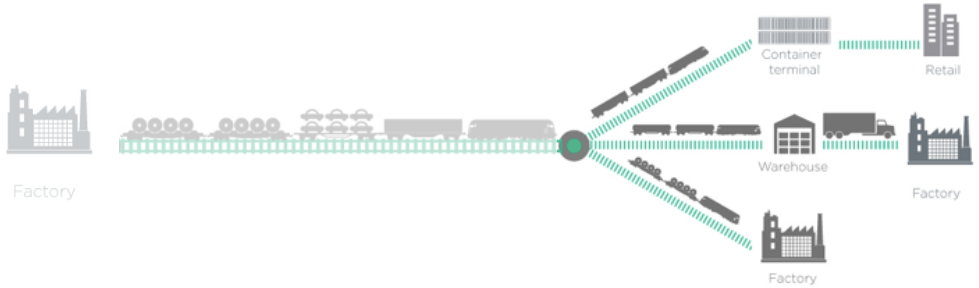
Dremio and ETL

ETL falls into two workload patterns



Long-Haul ETL

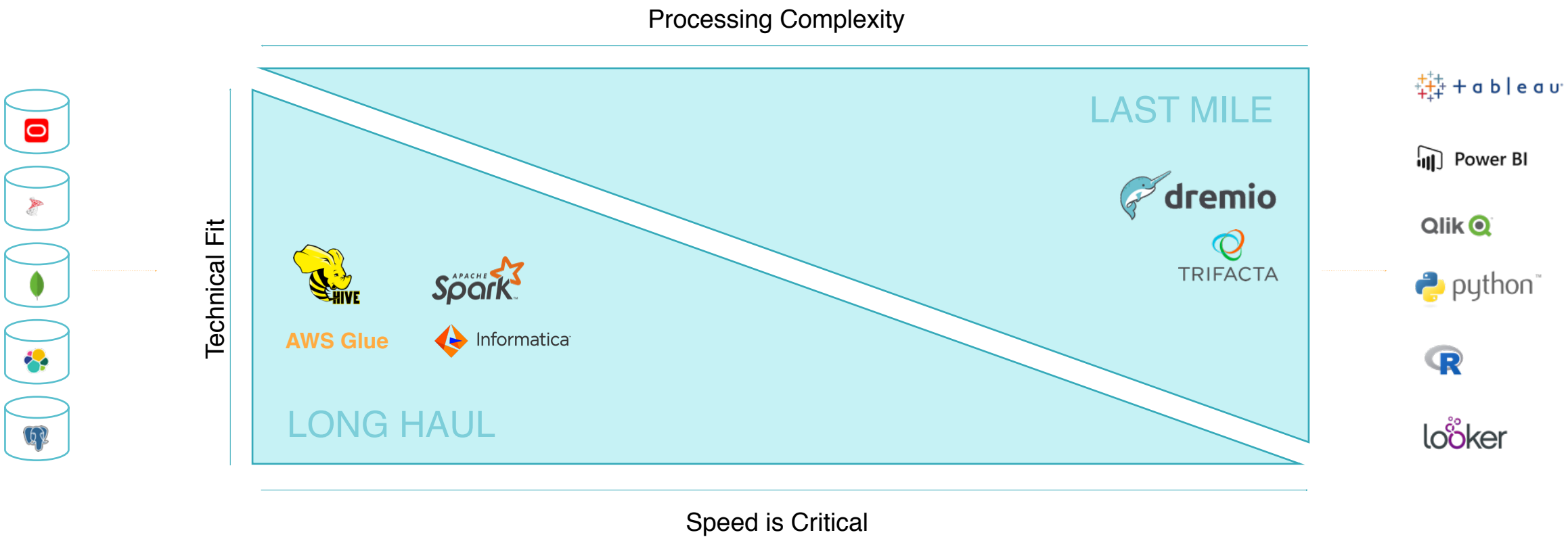
- ✓ Bulk operations, multi-hour runtime
- ✓ Branching/forking logic
- ✓ Human steps in the process
- ✓ 100s of tables Snowflake schema
- ✓ Sessionization



Last-Mile ETL

- ✓ Filtered subsets of data
- ✓ Aggregations
- ✓ Calculated fields
- ✓ Data type conversions
- ✓ SQL-based transformations

ETL: Long-Haul vs. Last-Mile



Open Source

Dremio is open source

Apache licensed, built on Apache projects

- ✓ No vendor lock-in
- ✓ Faster development cycles
- ✓ Widely used, more hardened code
- ✓ More resources to learn, operate
- ✓ Not dependent on single vendor for viability

Editions for all levels of use

	Community Edition	Enterprise Edition
Powerful & Intuitive UX	✓	✓
Scales from 1-1000+ Nodes	✓	✓
Data Connectors	✓	✓
Data Acceleration	✓	✓
Business Continuity	✓	✓
Enterprise Security		✓
Multi-tenant Workload Management		✓
End-to-end Encryption over TLS		✓
Data Provenance & Lineage		✓
Enterprise Data Connectors		✓
Crowd-Sourced Acceleration Priority		✓
Support 24x365		✓
License	Apache 2.0	Commercial

#RedHatOSD

RED HAT FORUMS

Per visualizzare video e altri contenuti vai sul sito RIOS tramite questo link: www.reteitalianaopensource.net/form-redhat-2019

THANK YOU

linkedin.com/seacomsrl

Facebook.com/seacomsrl

youtube.com/riosacademy

twitter.com/seacomsrl

Red Hat | RIOS RETE ITALIANA OPEN SOURCE SEACOM the leading open source architects