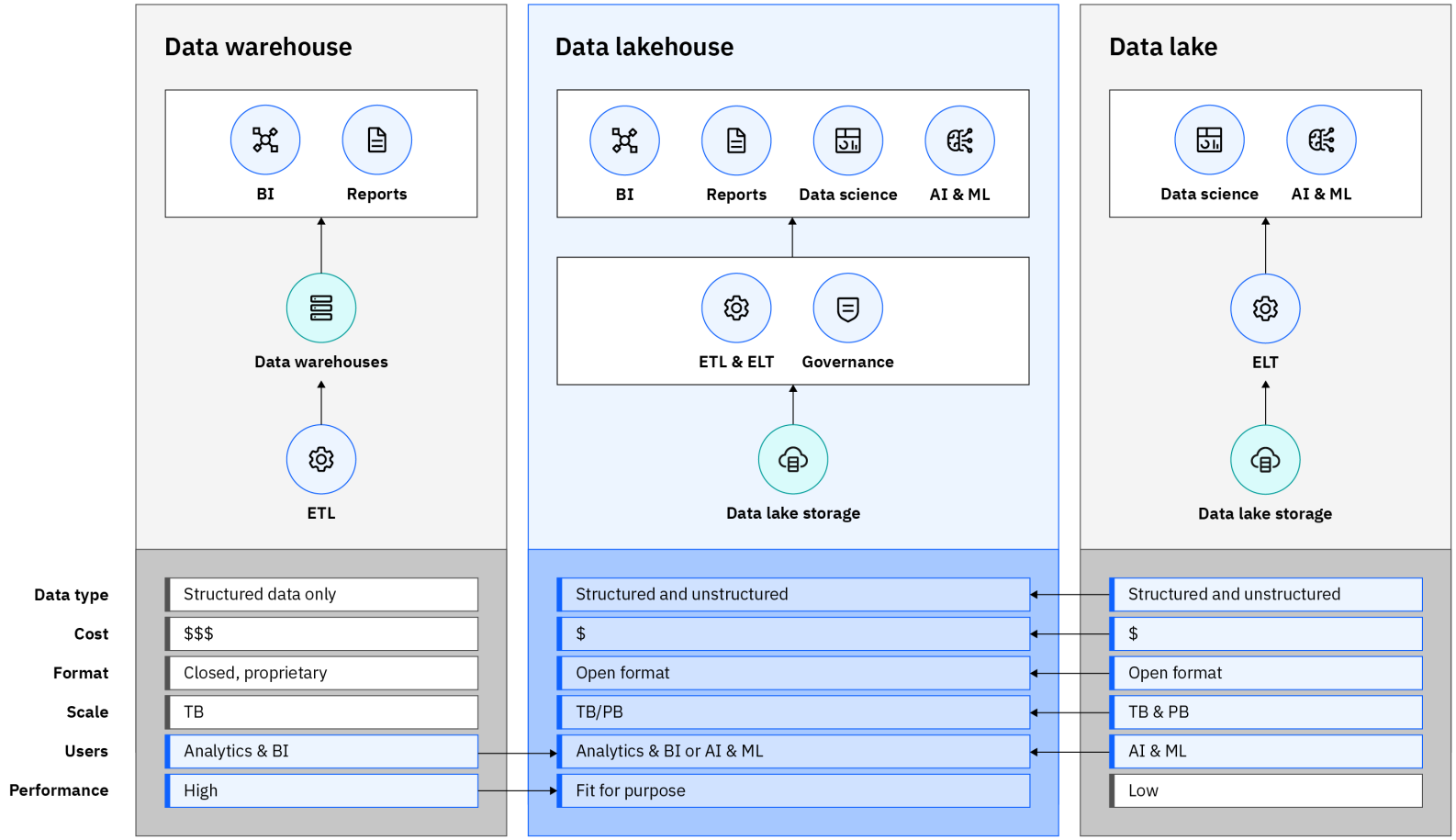
IBM® watsonx.data™

the only open, hybrid,
and governed data
store optimized for all data,
analytics and AI workloads.

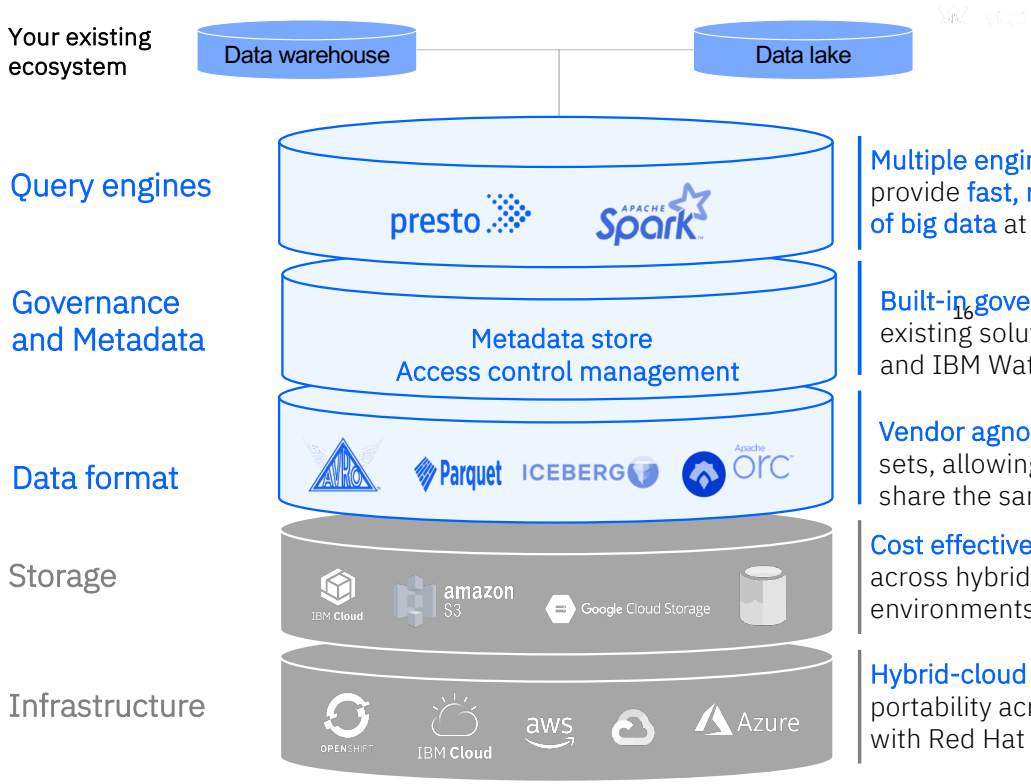
Lakehouses are a new approach meant to combine the advantages of data warehouses and data lakes, but first generation lakehouses still have key constraints



First generation lakehouses are still limited by their ability to address cost and complexity challenges:

- Single query engines set up to support limited workloads –typically just BI or ML
- Typically deployed over cloud only with no support for multi-/hybrid -cloud deployments
- Minimal governance and metadata capabilities to deploy across your entire ecosystem

Overview of the key components of the IBM watsonx.data: multiple query engines, open table formats and built-in enterprise governance



watsonx.data

Multiple engines such as Presto and Spark that provide **fast, reliable, and efficient processing of big data** at scale

Built-in governance that is compatible with existing solutions such as watsonx.governance and IBM Watson Knowledge Catalog

Vendor agnostic open formats for analytic data sets, allowing different engines to access and share the same data, at the same time

Cost effective, simple object storage available across hybrid-cloud and multi-cloud environments

Hybrid-cloud deployments and workload portability across hyperscalers and on-prem with Red Hat OpenShift

Core watsonx.data functionality

Ecosystem infrastructure

- Optimize workload costs and performance using multi-engine functionality
- Ensure governance and reduce time to insight with centralized metadata and access management
- Access all of your data across databases and data lakes
- Reduce storage costs and facilitate data ingest
- Deploy on any infrastructure and optimize available resources

Presto

- Presto is an open-source distributed SQL engine suitable for querying large amounts of data
- Supports both relational and non-relational sources
- Easy to use with data analytics and business intelligence tools
- Supports both interactive and batch workloads
- In watsonx.data, spin up one or more Presto compute engines of various sizes – cost effective, in that engines are ephemeral and can be spun up and shut down as needed

- Presto connectors allow access to data in-place, allowing for **no-copy data access and federated querying**
- Consumers are abstracted from the physical location of data
- A wide variety of data sources are supported, including:

What is a metastore?

- Manages metadata for the tables in the lakehouse, including:
  - Schema information (column names, types)
  - Location and type of data files
- Similar in principle to the system catalogs of a relational database
- Shared metastore ensures query engines see schema and data consistently
- May be a built-in component of a larger integration/governance solution

HMS used by watsonx.data

- Hive metastore (HMS) is a component of Hive, but can run standalone
- Open-source
- Manage tables on HDFS and cloud object storage
- Pervasive use in industry

AWS Glue Data Catalog

- Component of AWS Glue integration service
- Inventories data assets of AWS data sources
- Includes location, schema, and runtime metrics

Microsoft Purview Data Catalog

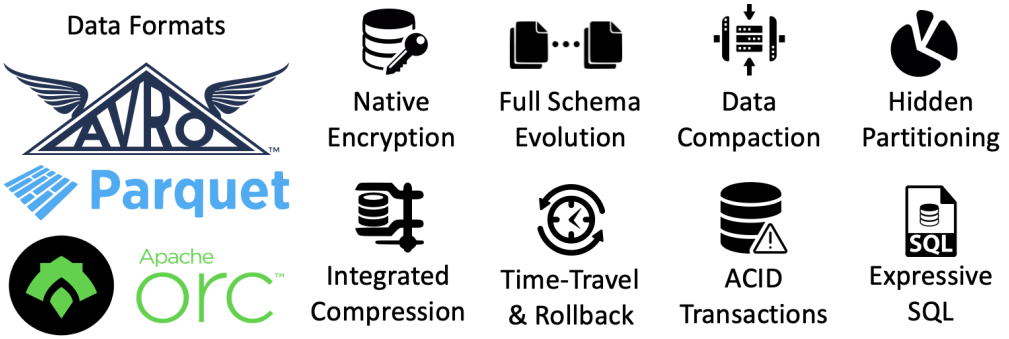
- Component of Microsoft Purview data governance solution
- Helps manage on-premises, multicloud, and SaaS data
- Offers discovery, classification, and lineage

Databricks Unity Catalog

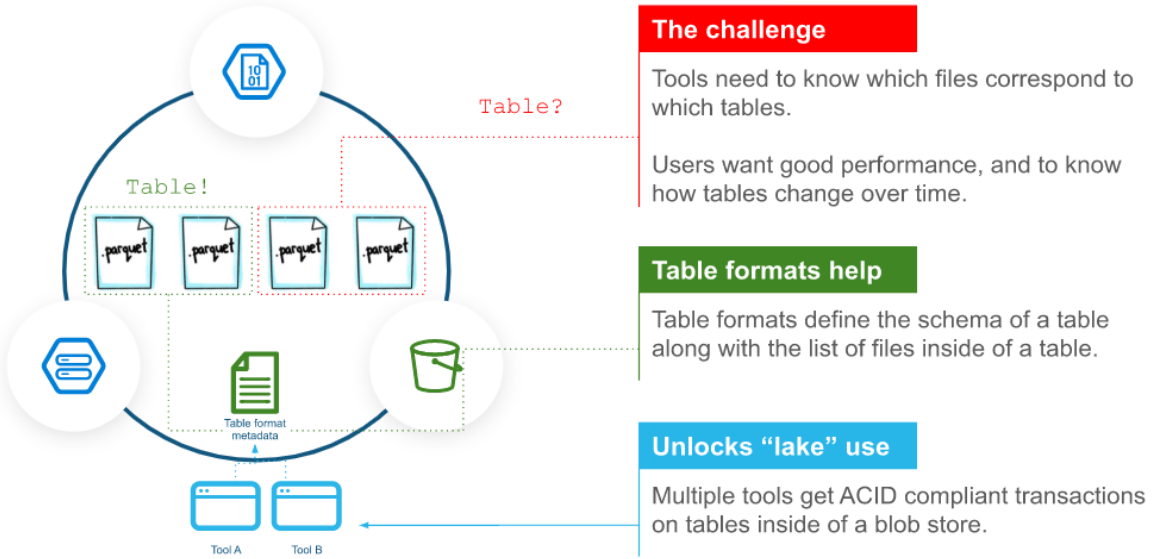
- Provides centralized access control, auditing, lineage, and data discovery across a Databricks lakehouse
- Contains data and AI assets including files, tables, machine learning models, and dashboards

What is Apache Iceberg?

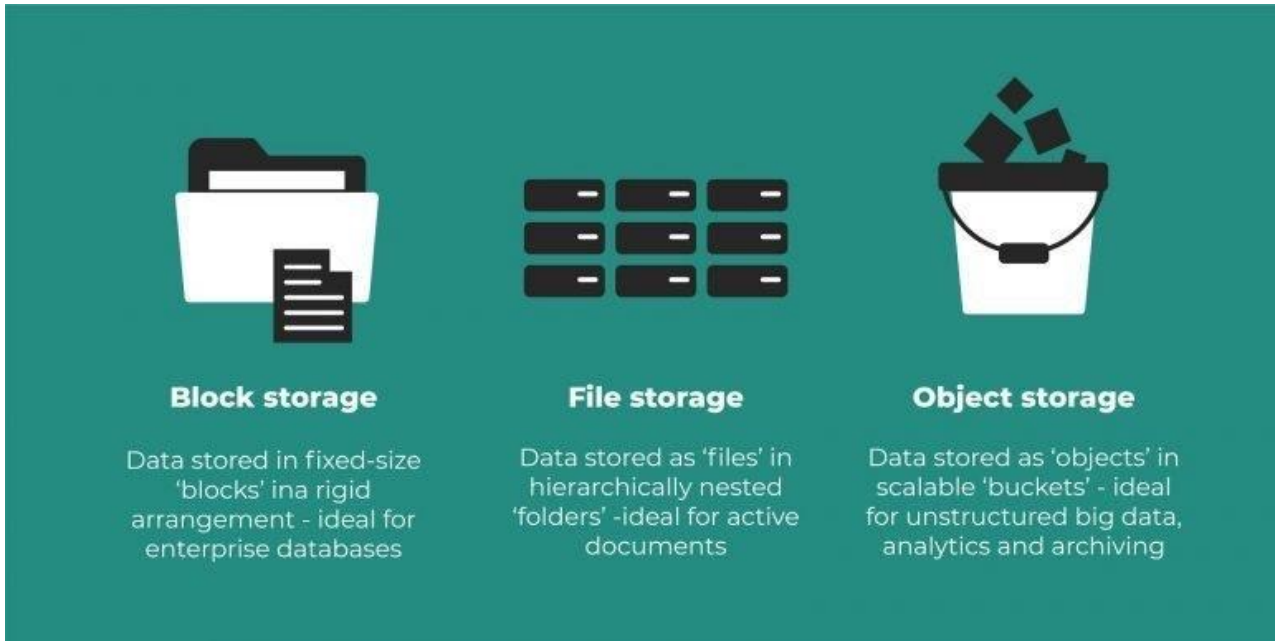
ICEBERG



- High-performance format for huge analytic tables
- Brings the simplicity of SQL to big data and data lakes
- Fully open source and accessible
- Rapidly becoming the industry standard



What is object storage?



Object storage:

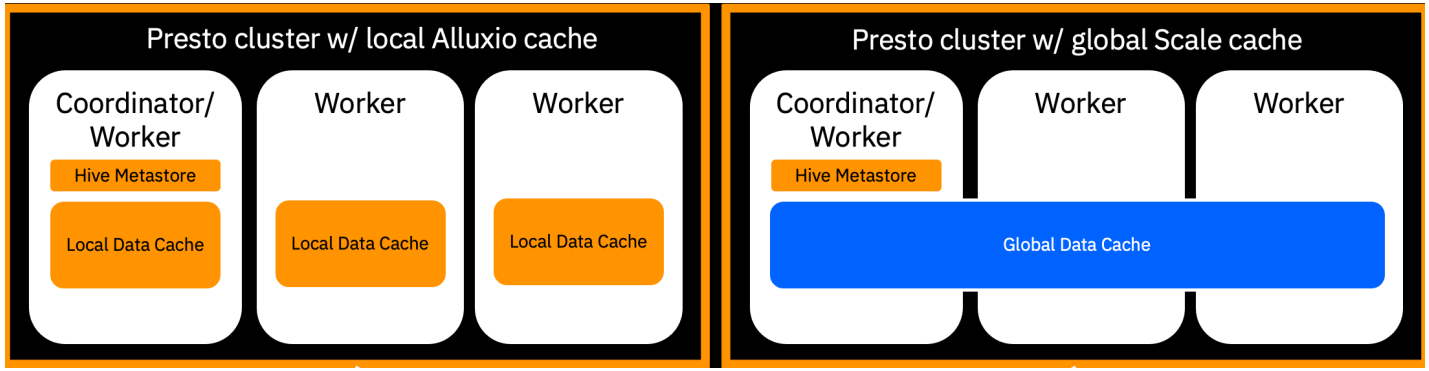
- Low cost
- Near unlimited scalability
- Extreme durability & reliability (99.999999999%)
- High throughput
- High latency (but can be compensated for)
- Basic units are *objects*, which are organized in *buckets*

- Most notable provider for object storage is Amazon S3 (Simple Storage Service)
  - Other vendors offer S3-compatible object storage

Global Data Caching for multi-engine Lakehouse

Caching is an important concept in ***Lakehouse***, which allows for improved performance over object storage, by temporarily storing critical data on local disks

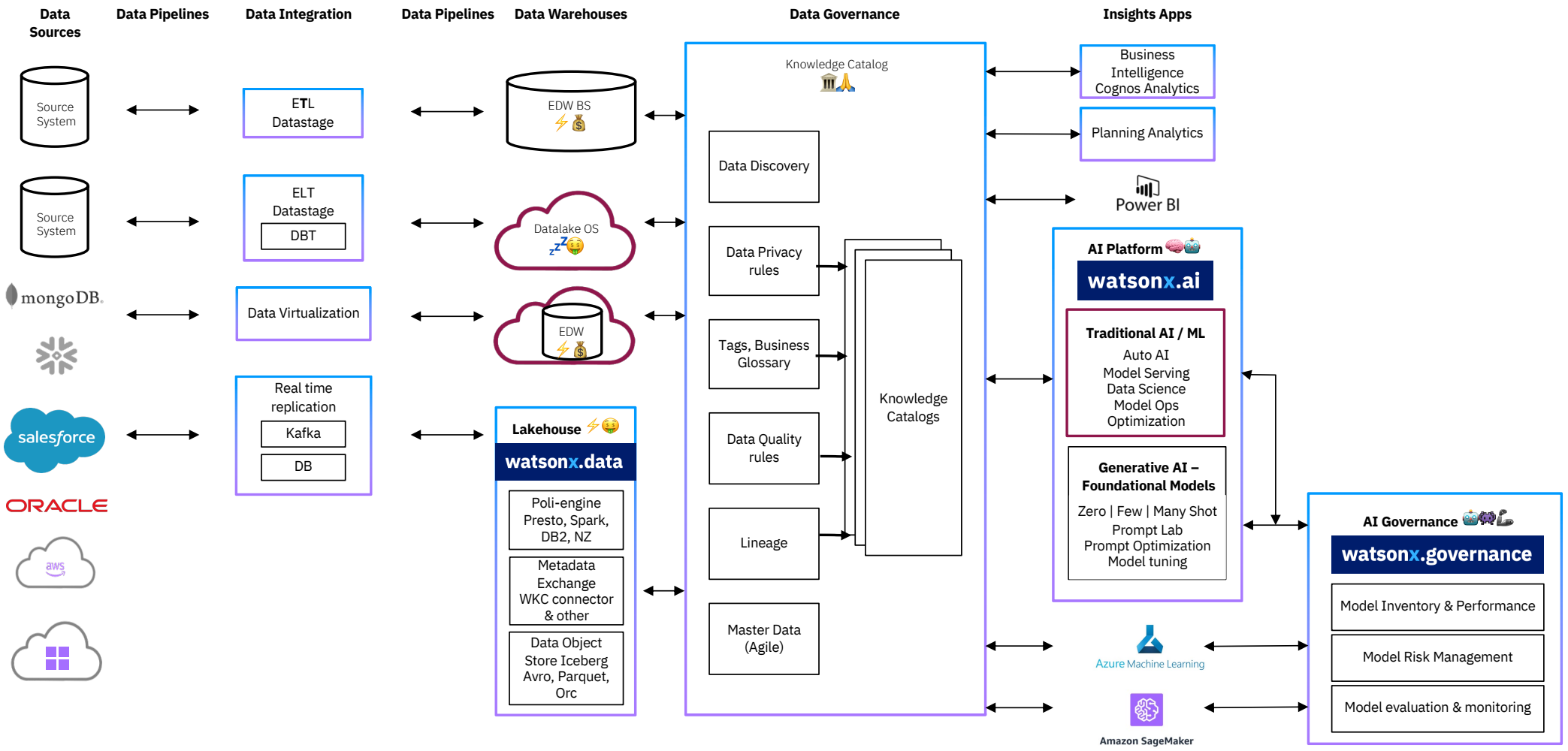
Spectrum Scale provides ***Global Data Caching*** capabilities that **differentiate** from how any other vendor approaches caching from object storage



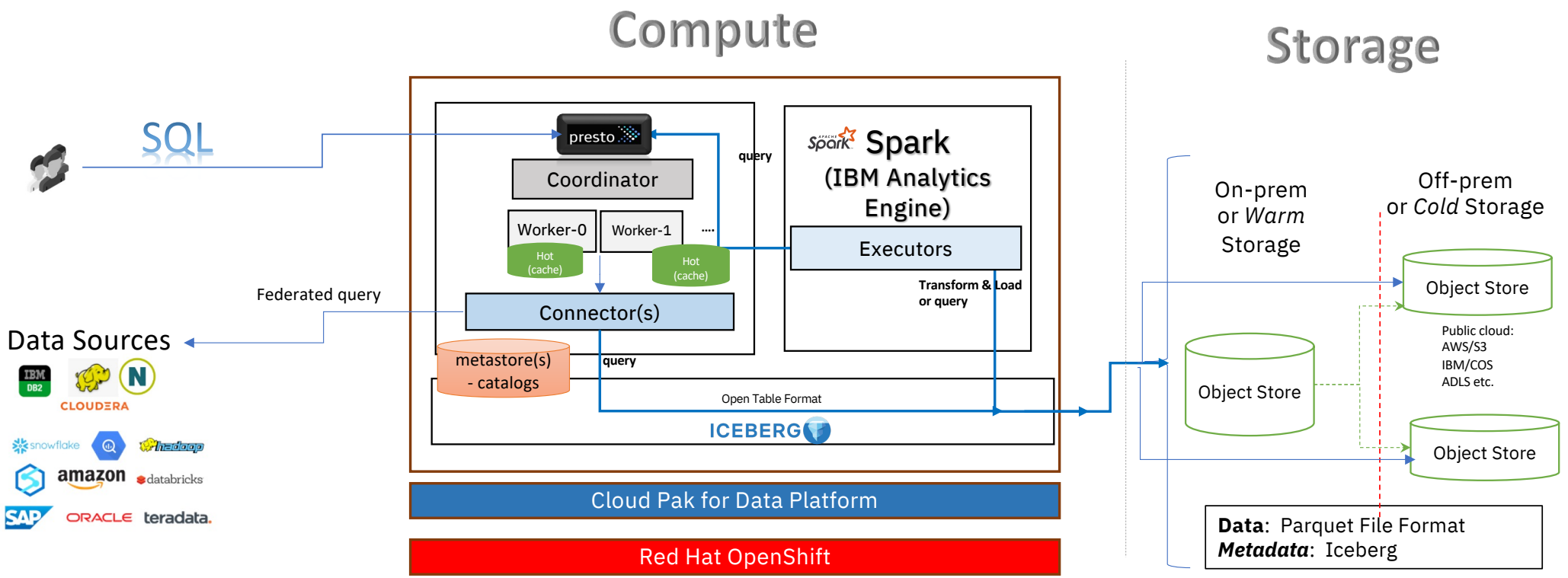
- Every compute engine, and every worker of that engine **must spend time "caching" data that** isn't local
- Data can be duplicated multiple times across local caches and coordinators aren't aware of who has what
- Cache is lost when engine is turned off – ie. Presto instance gets shut down

- Cached only ONCE = faster performance
- Cache is **shareable**, every engine within Lakehouse can access the same cache – Db2W, NZ, Presto, Spark
- Data is only cached once across participating nodes
- Cache persists as long as the cache service persists even if individual engines are destroyed

How **watsonx** fits in your architecture



watsonx.data on Cloud Pak for Data



- Compute & storage separation, ***independently scalable***
  - Support for ***multiple engines*** *(initially Presto & Spark)*
  - and ***multiple Storage buckets*** – example: on-prem & off-prem for *tiered* Warm/Cold
- **in-place** access to data sources to *avoid data copies*
  - ***Federated Queries***
- Easy ***data*** **copy** to target Object Stores (or archive from warm -> cold)
- **Open** formats & standards: complement with other technologies & custom apps