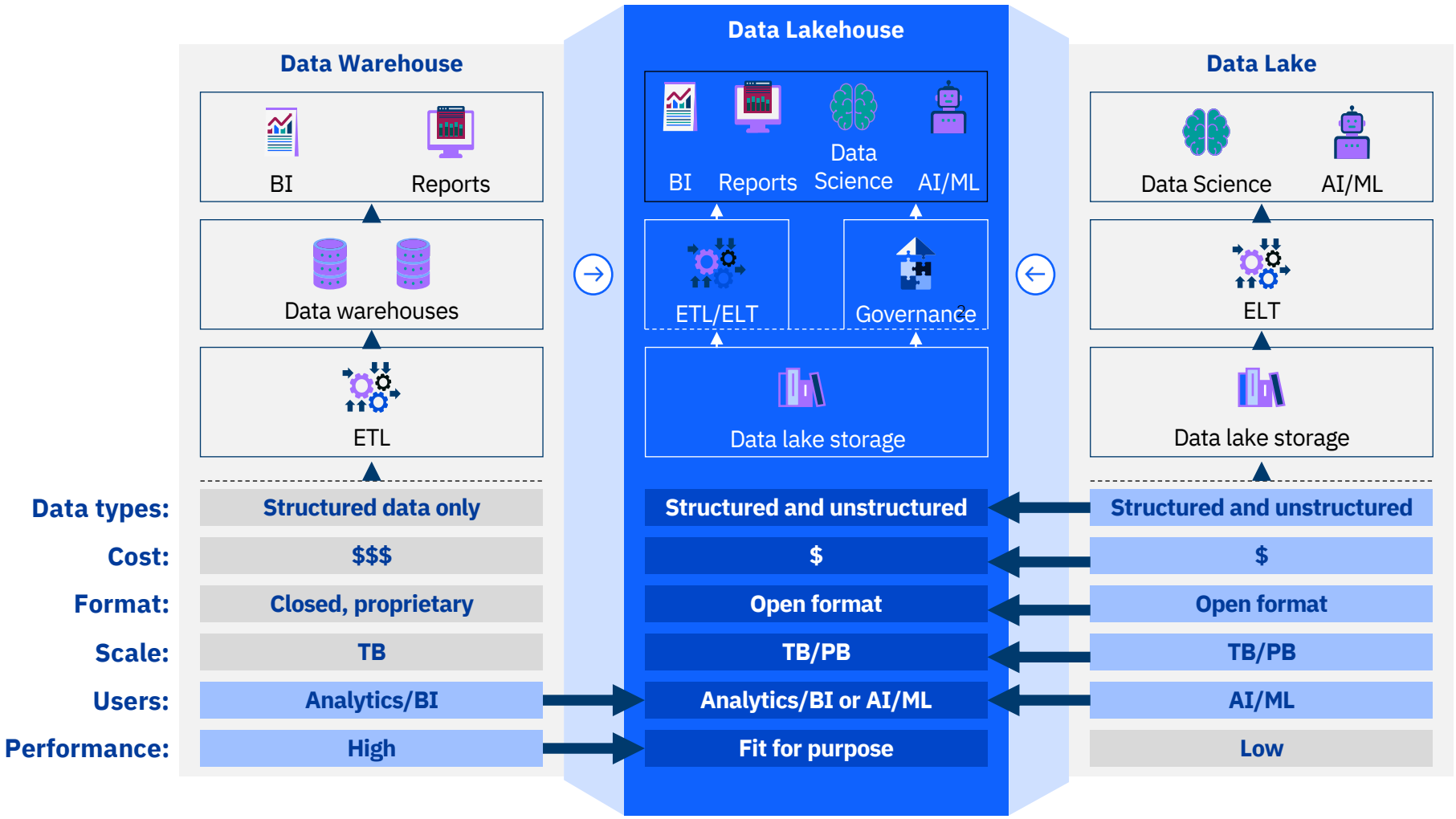
watsonx.data™

IBM Ecosystem Engineering – SI Lab

watsonx.data™

Lakehouses are meant to be a new class of data store that combines the best of data warehouses and data lakes

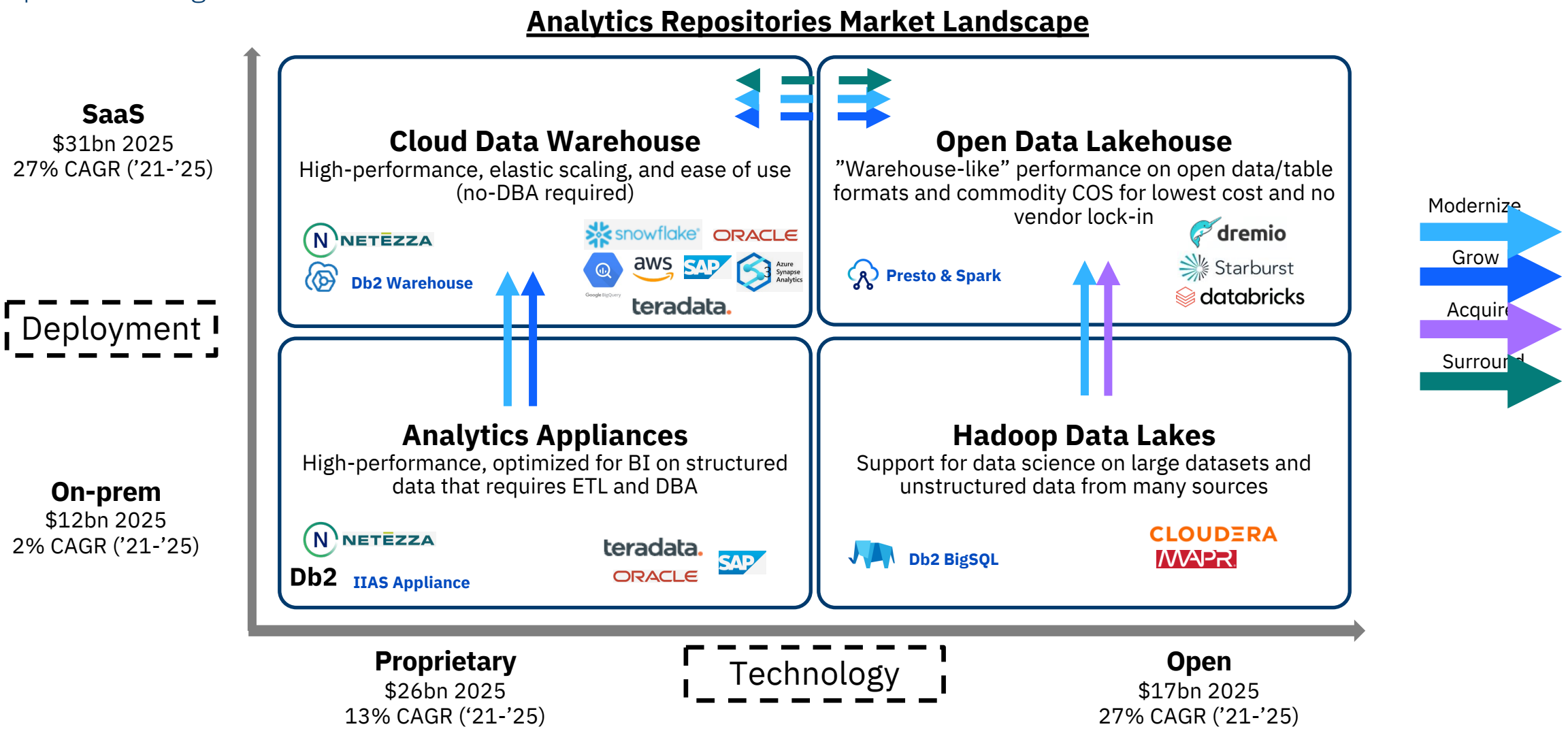


However, first generation lakehouses still have key constraints that limit their ability to address cost and complexity challenges:

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

Market dynamics

Major disruptions are driving the growth in the analytics repositories market from on-prem to SaaS and from proprietary to open technologies



The platform for AI and data

watsonx

Scale and accelerate the impact of AI with trusted data.

- watsonx.ai
- Train, validate, tune and deploy AI models

A next generation enterprise studio for AI builders to train, validate, tune, and deploy both traditional machine learning and new generative AI capabilities powered by foundation models. It enables clients to build AI applications in a fraction of the time with a fraction of the data.

watsonx.data

Scale AI workloads, for all your data, anywhere

Fit-for-purpose data store, built on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data.

watsonx.governance

Accelerate responsible, transparent and explainable AI workflows

End-to-end toolkit for AI governance across the entire model lifecycle to accelerate responsible, transparent, and explainable AI workflows.

Purpose optimized engines

Watson Query (Orchestrator)
VISION/Not-yet-reality

Apache Spark/Flink/Other

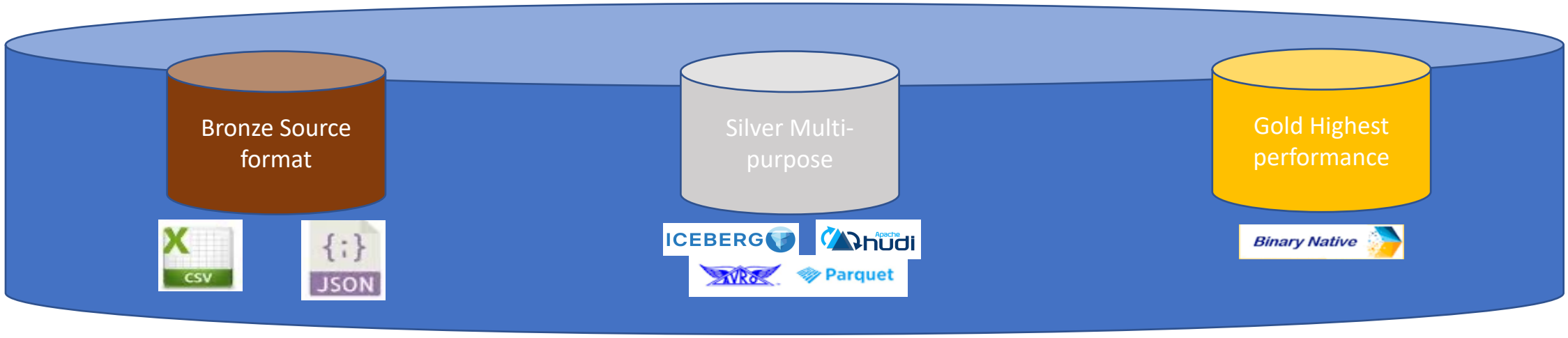
- Large scale batch analytics.
- Exabyte scale.
- Data engineering+transformation.
- Low volume batch transactions.

Presto

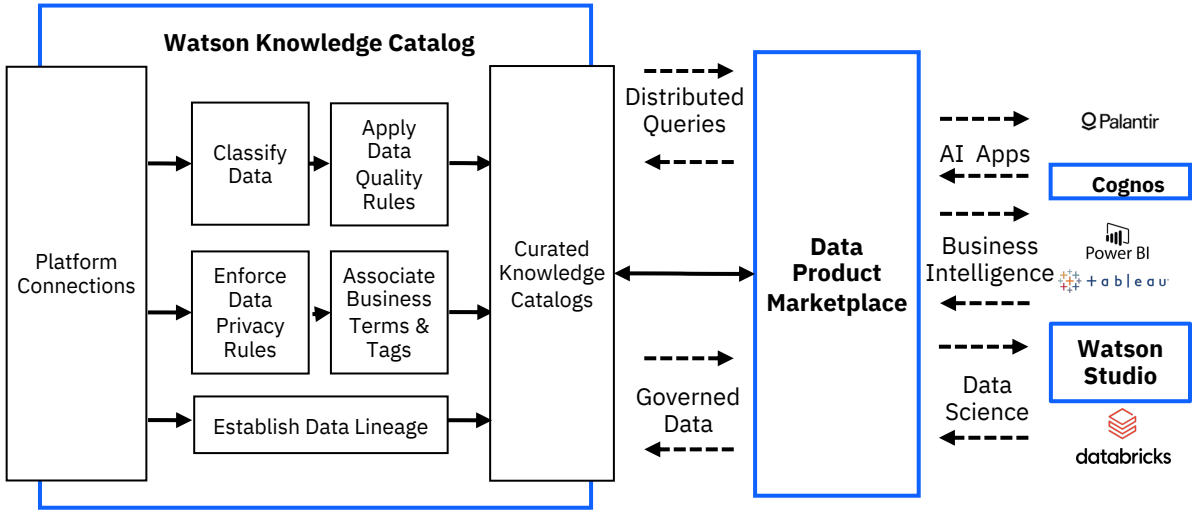
- Interactive queries+Adhoc analytics.
- Lightweight scalable engines.
- Low volume batch transactions.

Data Warehouse(s)

- High performance BI+Analytics.
- High Concurrency.
- High volume transactions.



Universal Data Governance for Hybrid/Multi-Cloud

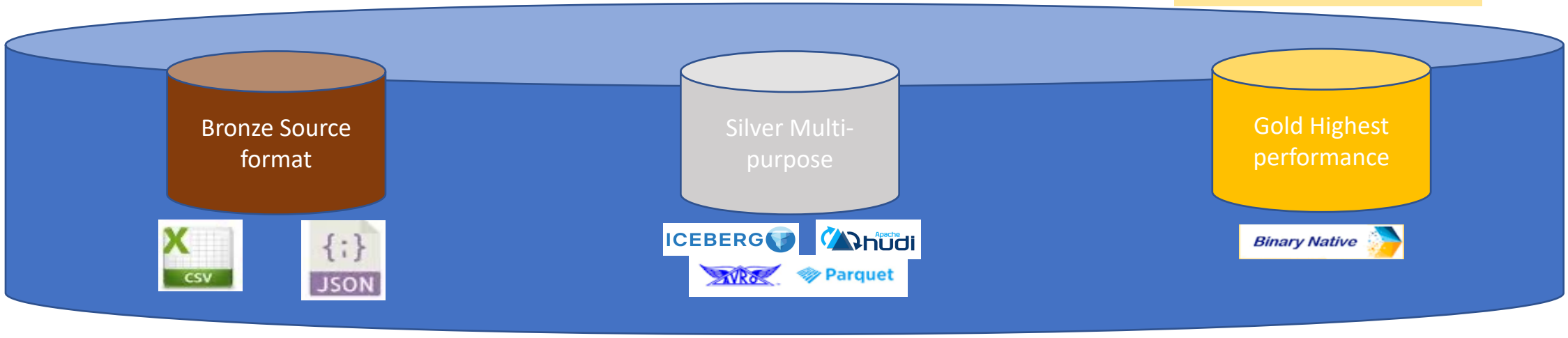


Statistical Analysis, Machine learning and Artificial Intelligence

Files

Tables

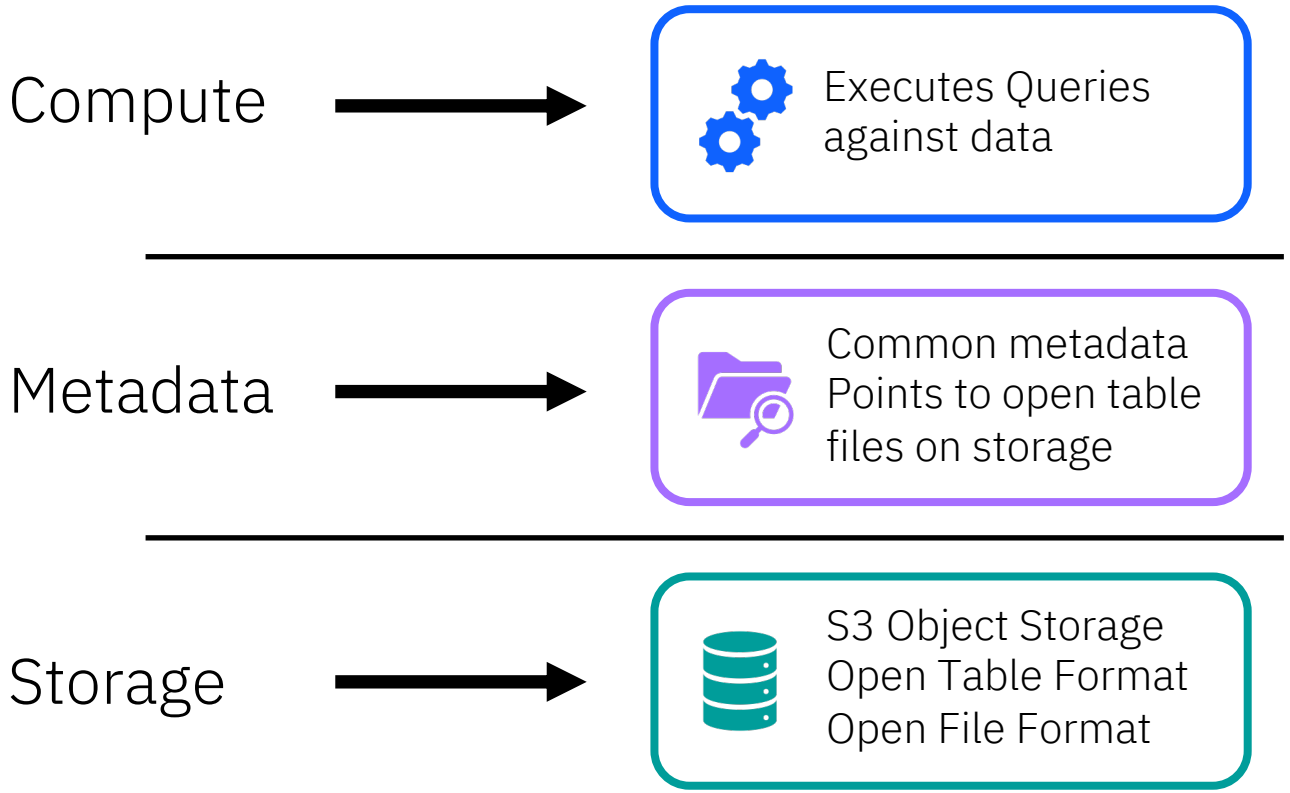
Reports/Dashboard



watsonx.data

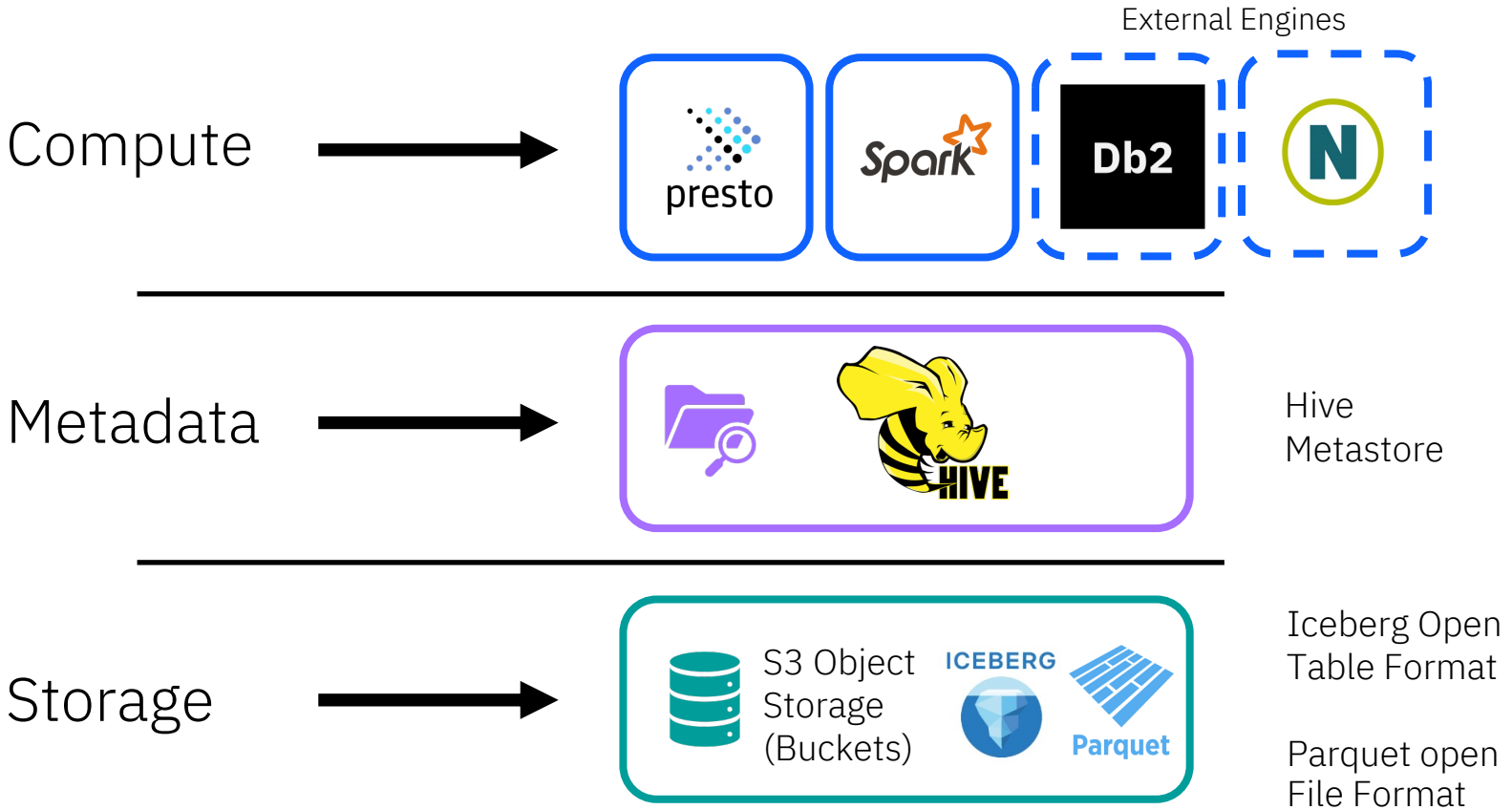
What is a lakehouse architecture?

General Lakehouse Architecture



What is the watsonx.data architecture?

watsonx.data Architecture

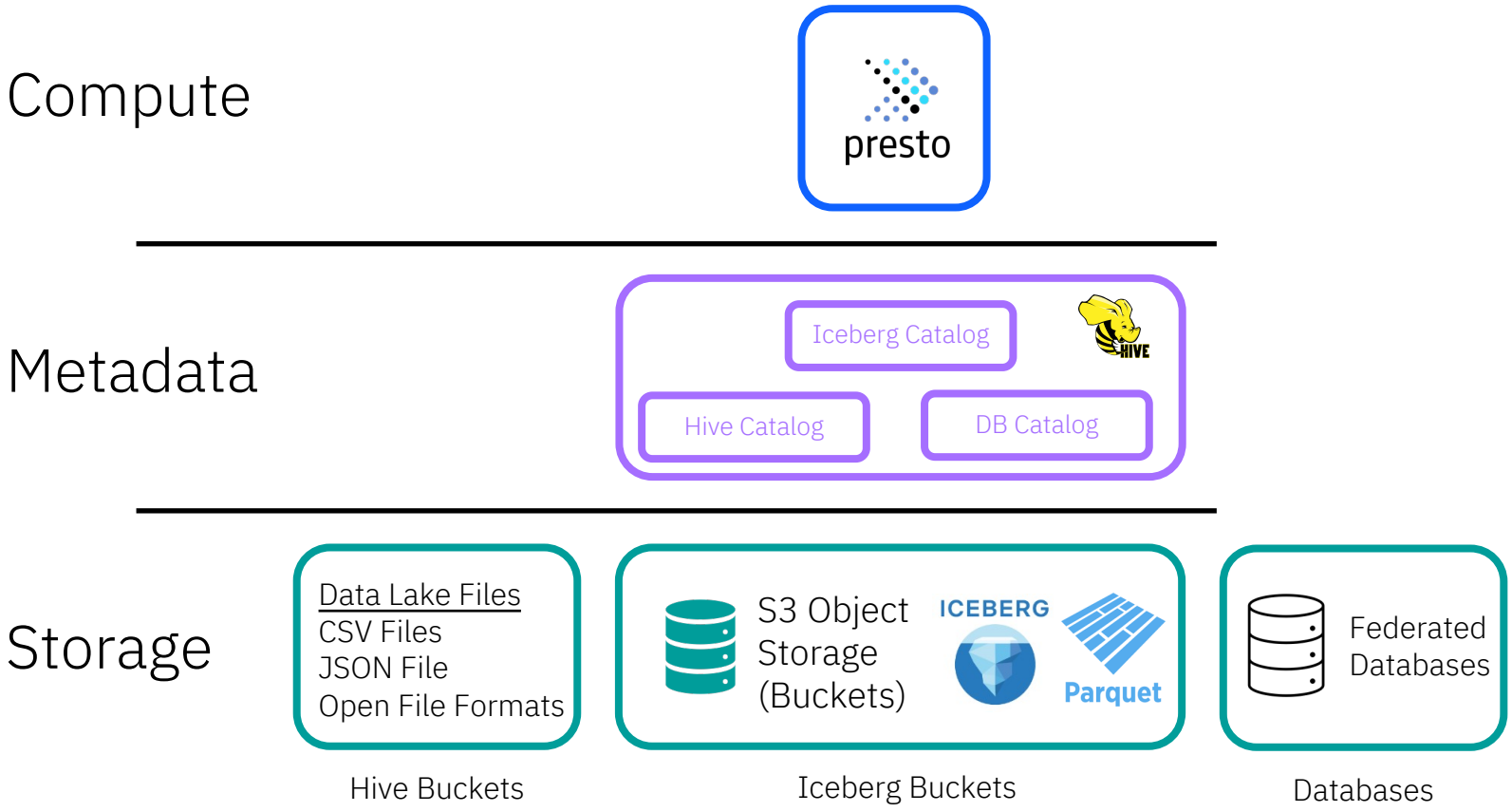


The Presto Engine adds functionality:

Connectors to non-iceberg data files

Connectors to databases (some allowing CRUD operations)

watsonx.data Architecture



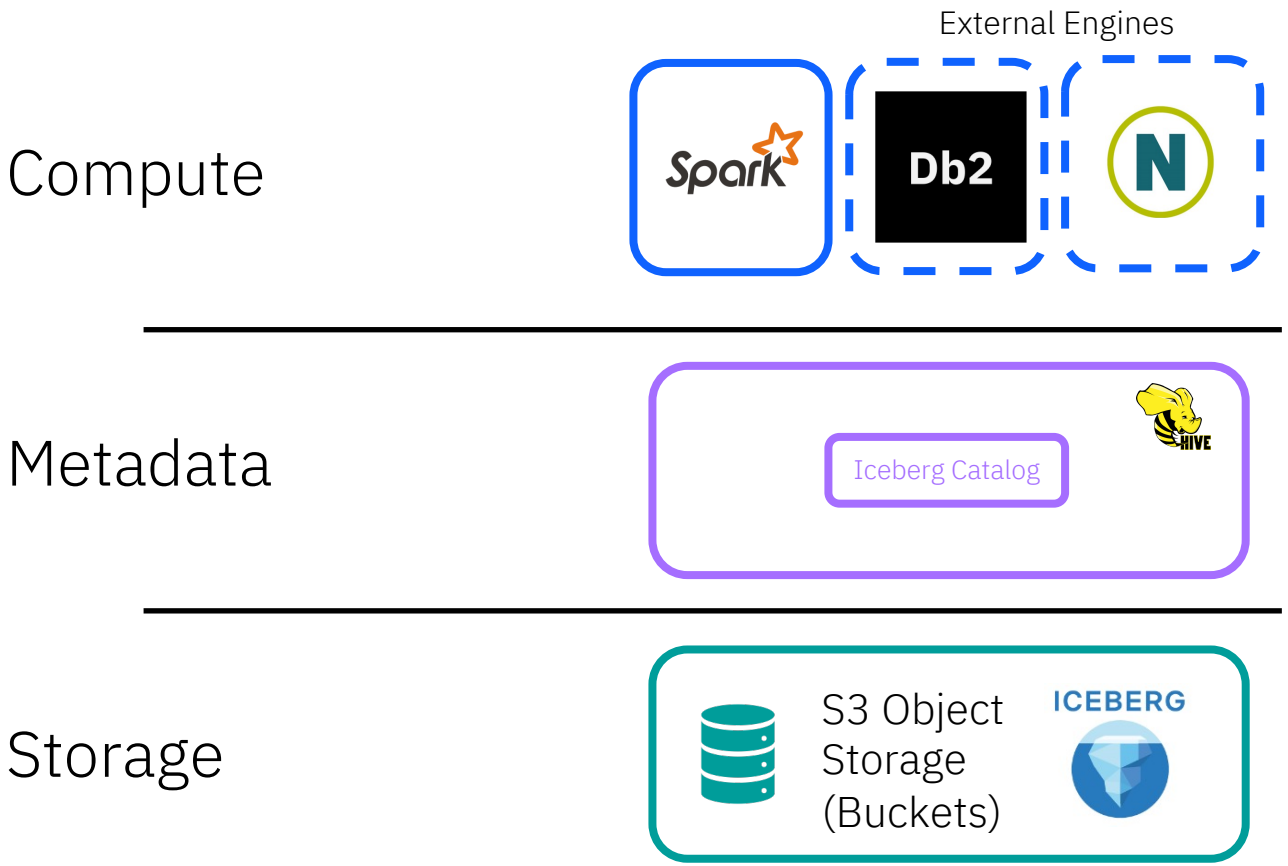
Spark, Db2, Netezza Engines

Can connect to Iceberg tables

Can access their own data objects/table within their environments

Cannot access Presto connectors

watsonx.data Architecture

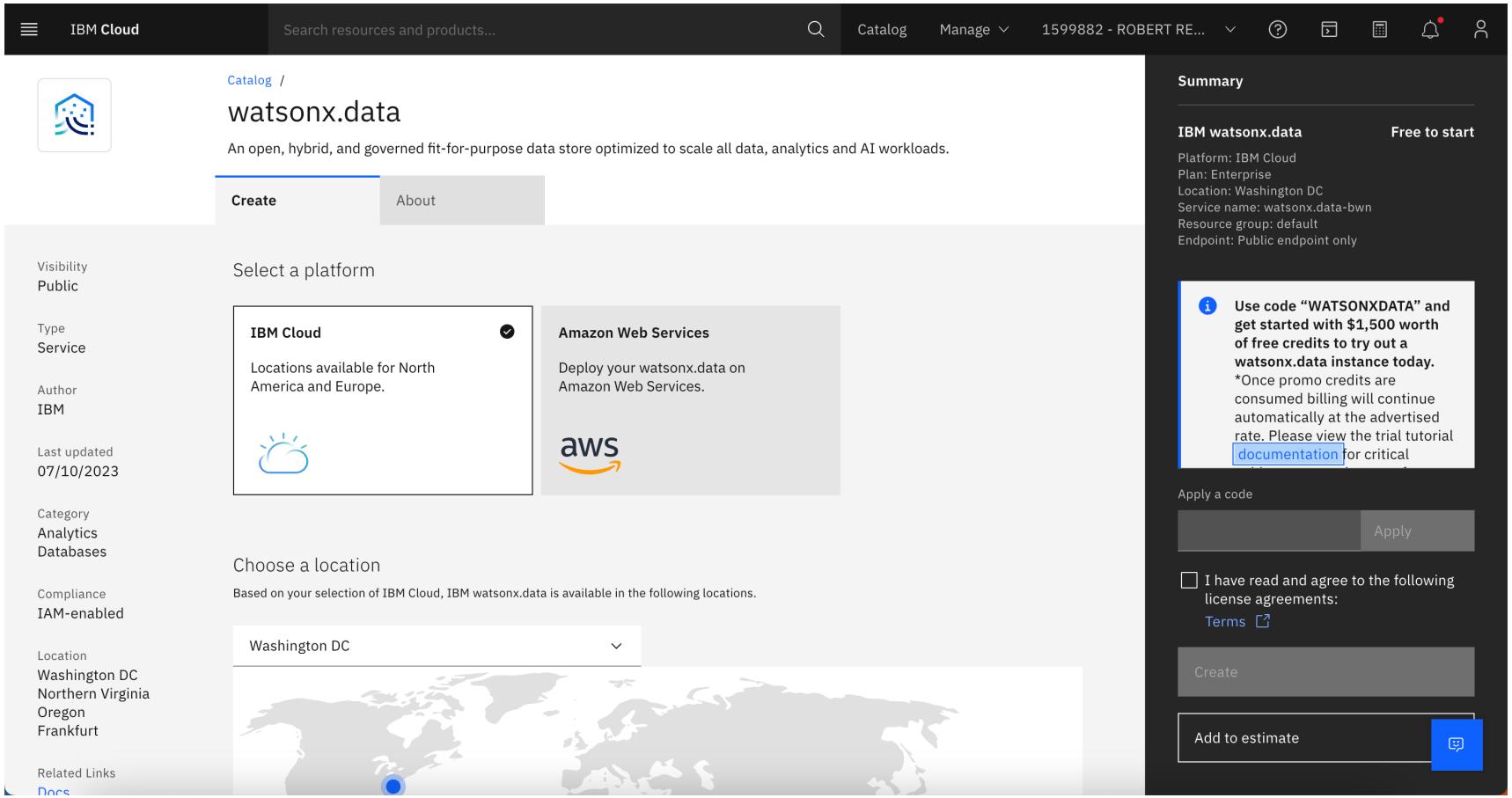


watsonx.data

Deployment options

https://cloud.ibm.com/lakehouse

SaaS → IBM Cloud
AWS



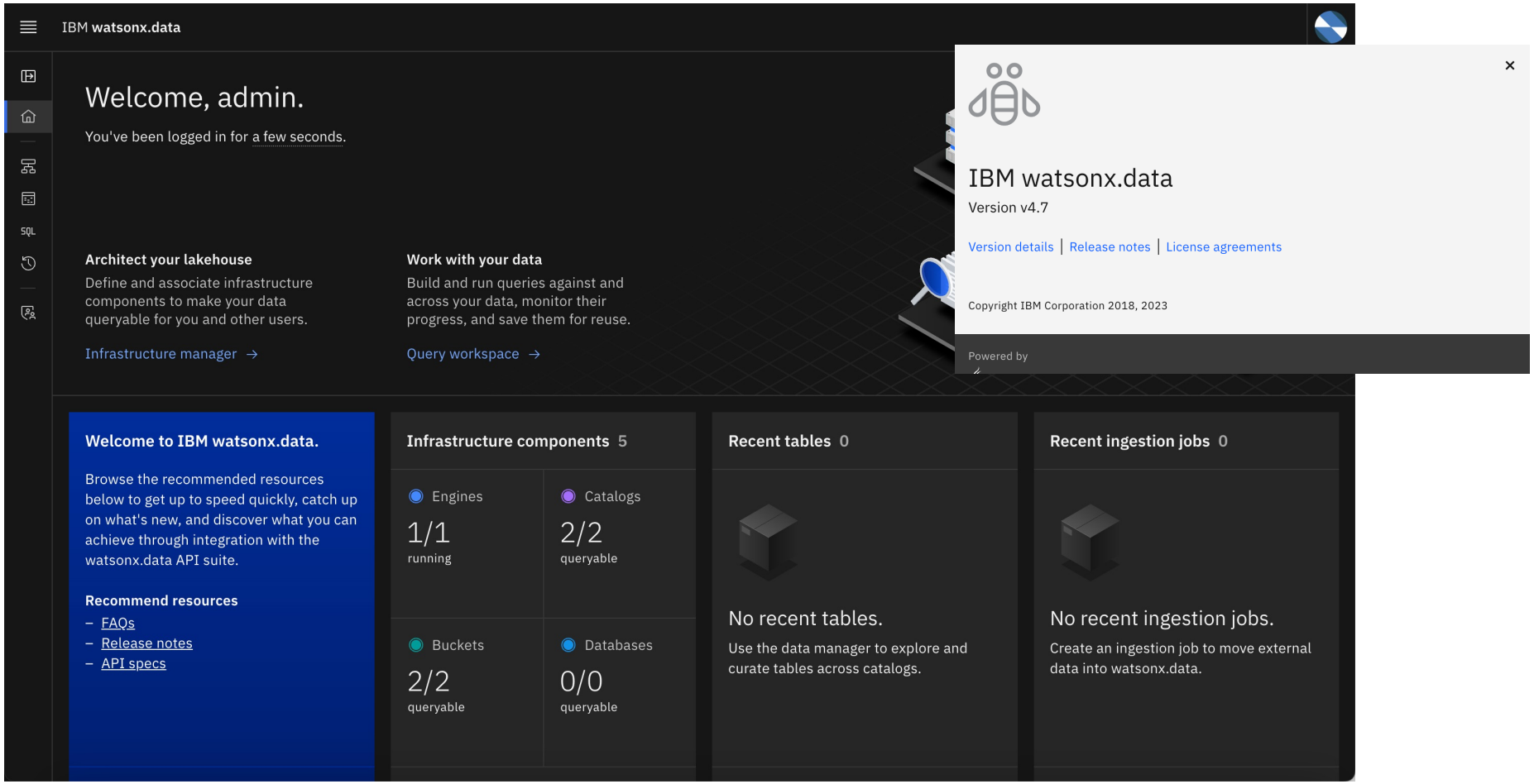
watsonx.data

- Engine Scaling
- Multi-Engine Support
- Caching Support
- For Enterprise Workloads

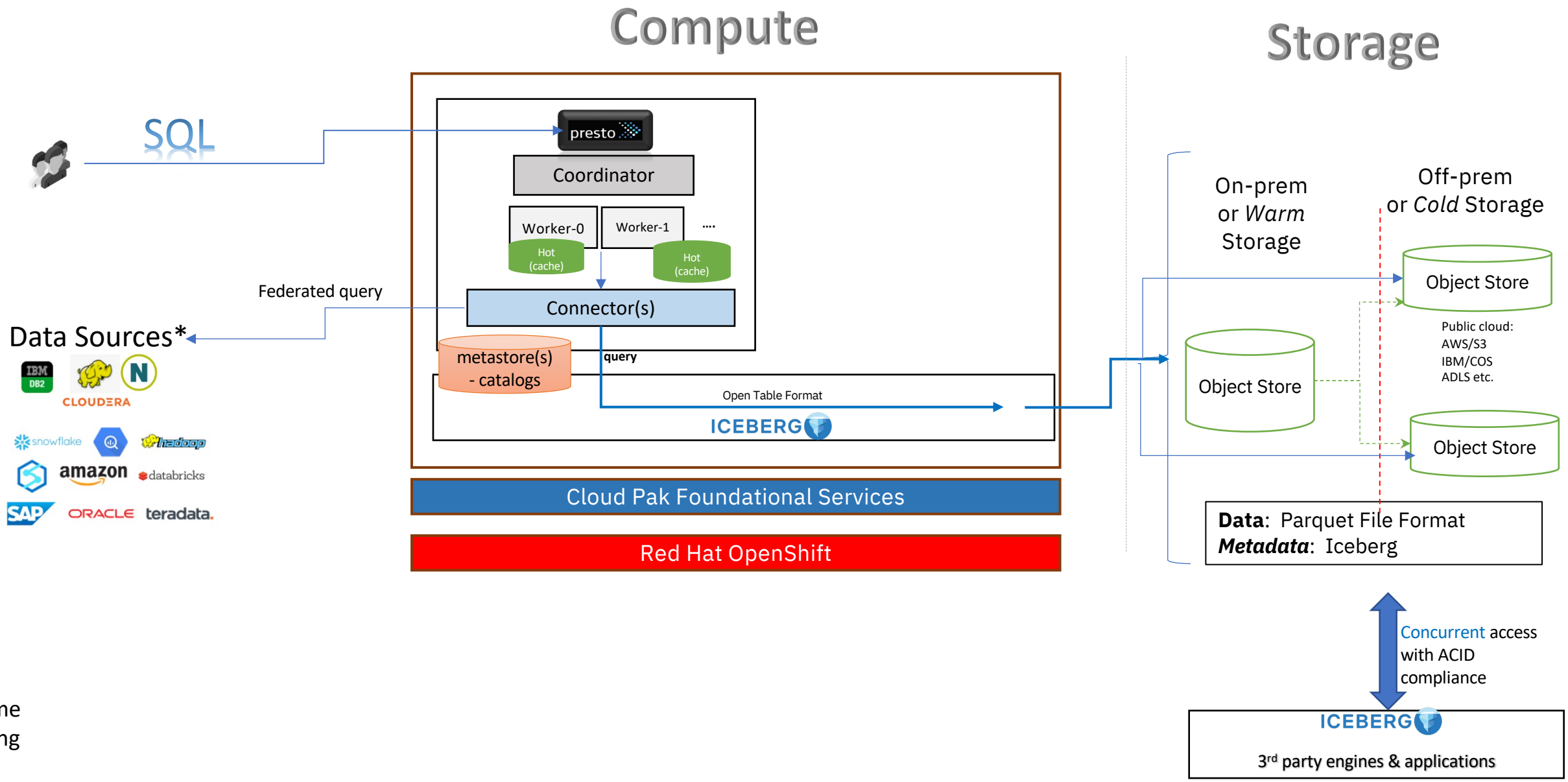
- Runs on Red Hat Openshift

Deployment options

Software (on-premises) → Stand alone or CP4D Cartridge

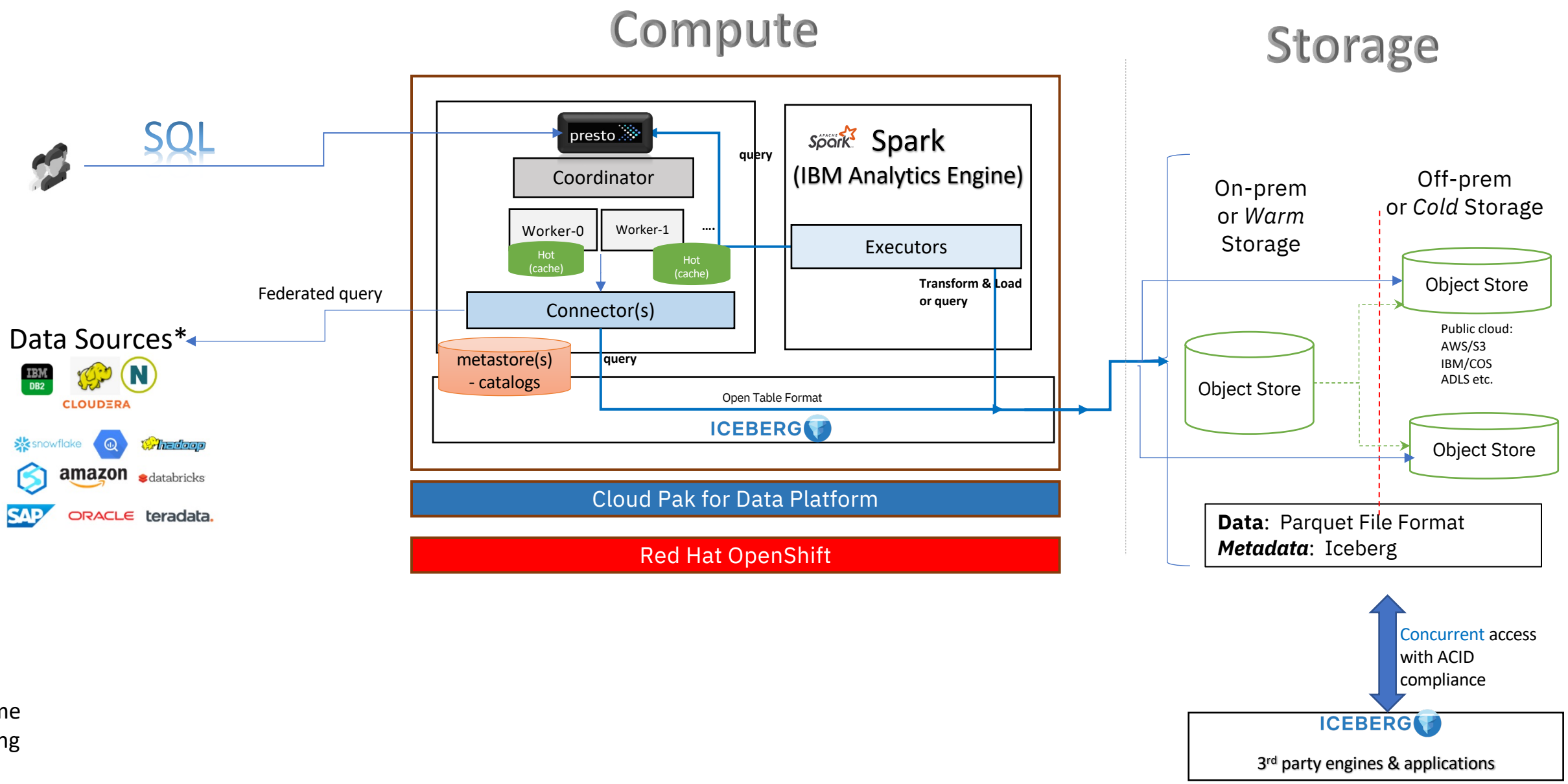


watsonx.data on Openshift – stand alone



* Some coming soon

watsonx.data on Cloud Pak for Data



* Some coming soon

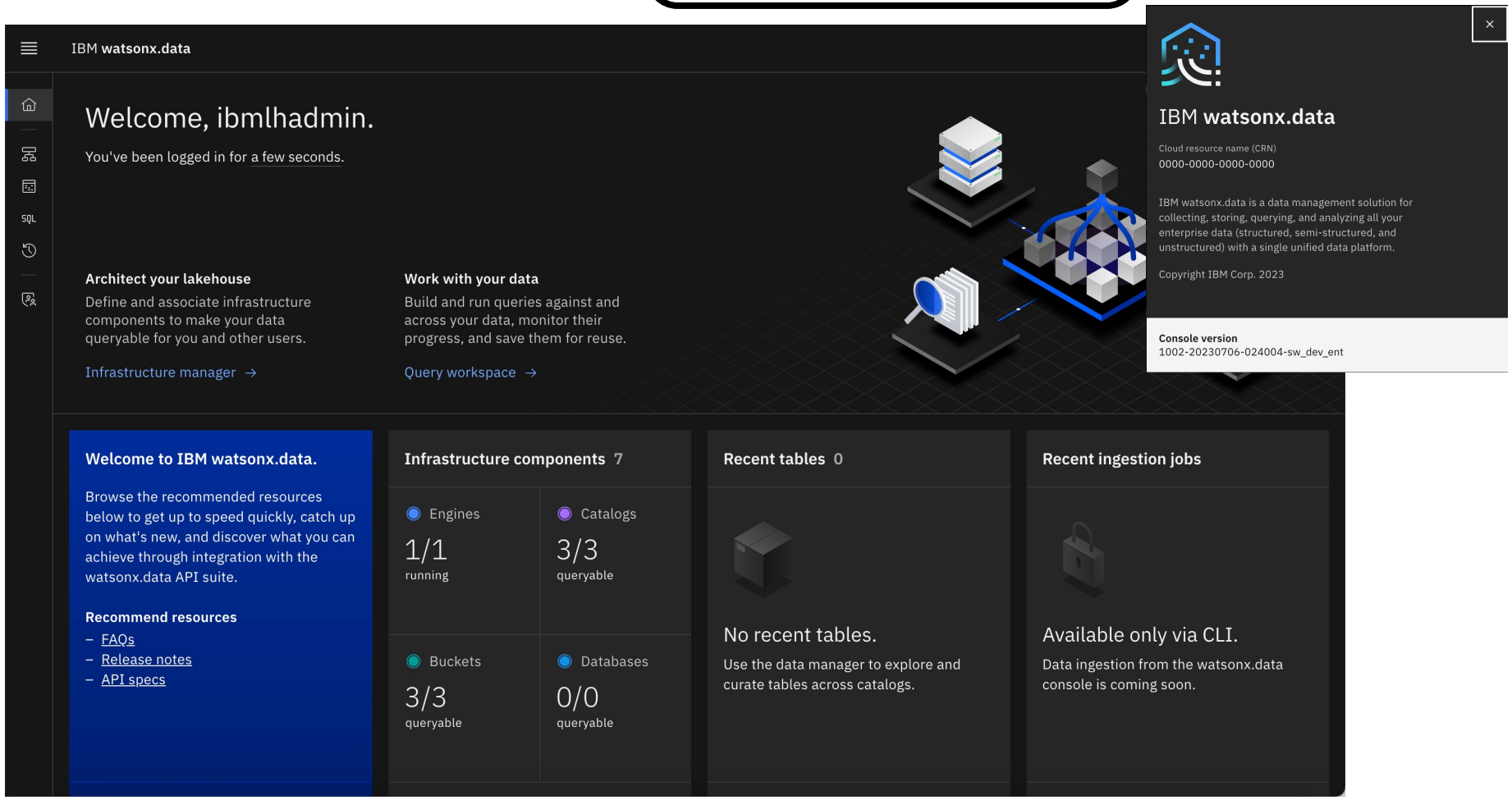
watsonx.data

Deployment options

- Single Presto Engine
- Cannot Scale
- Single User
- No Caching
- Used to test/play
- Not for Production

- Install on Linux (Recommended)
- Runs on Docker/Podman

Developer Edition → Runs on Podman or Docker



File Formats

- Row based
- Indexed
- Less compression than Parquet and ORC but faster write speeds
- Up to 3x faster read times
- Good schema evolution support
- Can be used as a landing area prior to additional data transformations

- Columnar
- Originally also designed for Hadoop
- Works very well with Spark
- Good for traditional OLAP queries because of its columnar format
- Carries schema, is self describing
- Data stored in pages
- Good for complex nested data structures

- Also columnar
- Data stored in stripes
- Indexed rows, compatible with HDFS
- KPIs for compression and runtime are very similar to parquet
- Parquet better for Spark, ORC better for Hive/Hadoop

Open Source Table Formats

- Separation of compute, data, and storage
  1. Leverage low-cost, infinitely scalable object storage
  2. Standardized
     - open file formats (Parquet, ORC, DWRF, JSON, …)
     - table formats (Apache Iceberg, LF Delta, Apache Hudi)
  3. Accessed by scalable compute engines of choice (Presto, Spark, etc.)

Sample Discovery Questions

- What type of database are they using today
- Is cost an issue or starting to become an issue
- What are the use cases and how do they serve the business
- What are the required SLAs? Who are the users accessing? How fast do they need answers?
- Are there ETL workloads within the EDW, what are the time windows
- Is there historical data in the EDW that is not frequently accessed but still required to be there?
- Overall data size and instance size

IBM Modernize + Cross Sell

Customer Goal: Customer has an IBM on-premises solution and want to move to next generation of self-managed and SaaS offerings

Customer motivation:

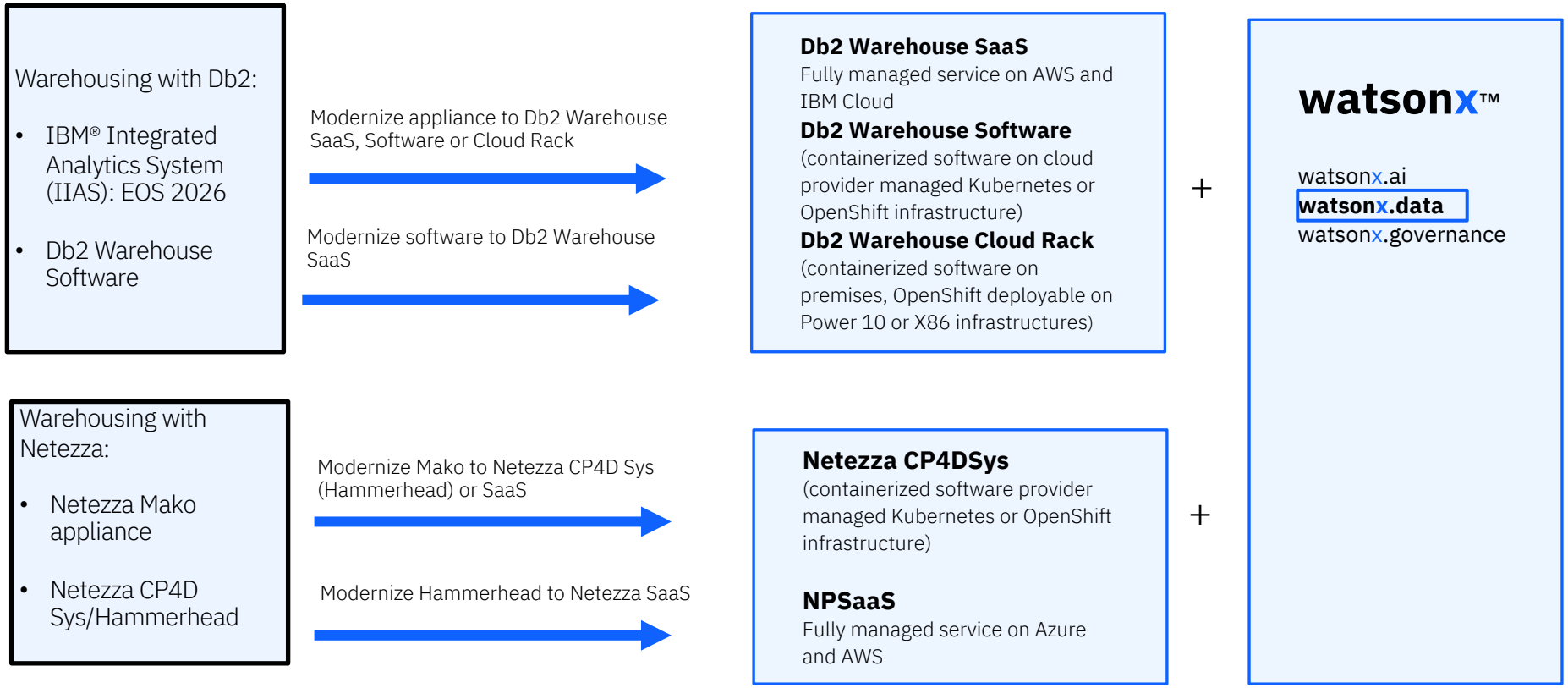
- Looking for next generation features
- Limited compute or data storage resulting in offloading or expanding data infrastructure
- Difficulty scaling up and down for workloads
- New use cases: AI/ML, Real-time analytics, data engineering, data sharing

Challenges to Address:

- System migration compatibility
- Sizing management
- Cost controls on the cloud
- Access to all data across hybrid cloud without duplication/movement

Key Messages:

1. Modernize your database appliance with like for like compatibility and support new use cases
2. Optimize workloads for AI with fit for purpose engines and reduce warehouse costs by 50%, access all governed data across hybrid cloud



IBM Db2 for z/OS Cross-sell

Customer Goal: Customer has IBM Db2 for z/OS and wants to access transactional data from mainframe for AI use cases

Customer motivation:

- Brand new use cases with mainframe transactional data: AI/ML, Real-time analytics, data engineering, data sharing

Challenges to Address:

- System migration compatibility

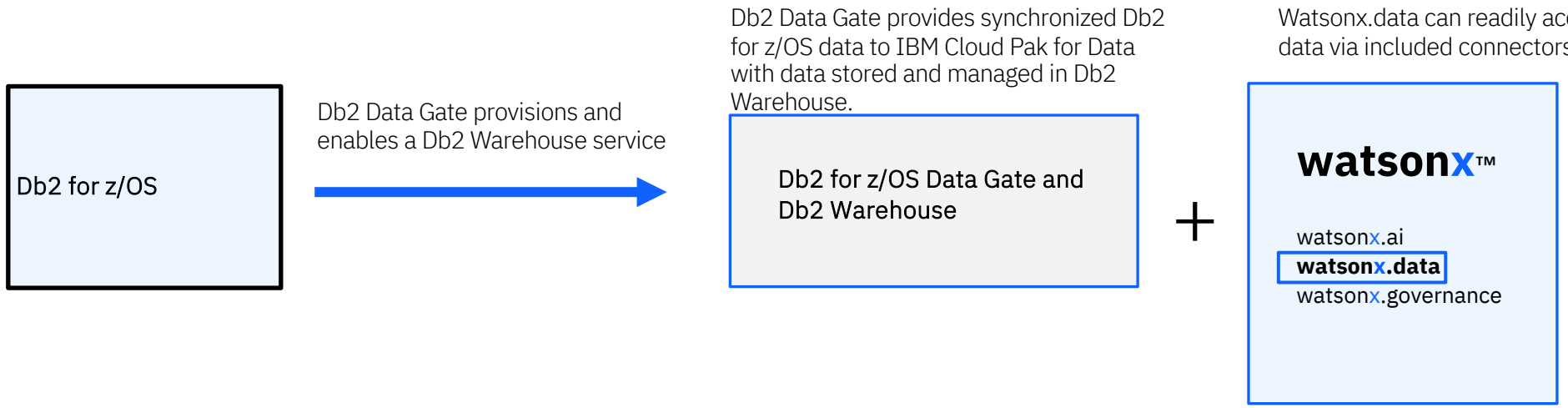
Key Messages z/OS and Db2 Warehouse:

- An integrated, optimized synchronization feature maintains currency between source Db2 for z/OS data on IBM Z and Db2 Warehouse targets
- Db2 Warehouse is a relational databases that delivers advanced data management and analytics capabilities

Key Messages z/OS and .data:

- Access and analyze data from mainframe systems in near-real time
- Consume less processor capacity with built-in synchronization
- Use the most up-to-date data from mainframe for machine learning models

---



Competitors to IBM data warehouse & watsonx.data

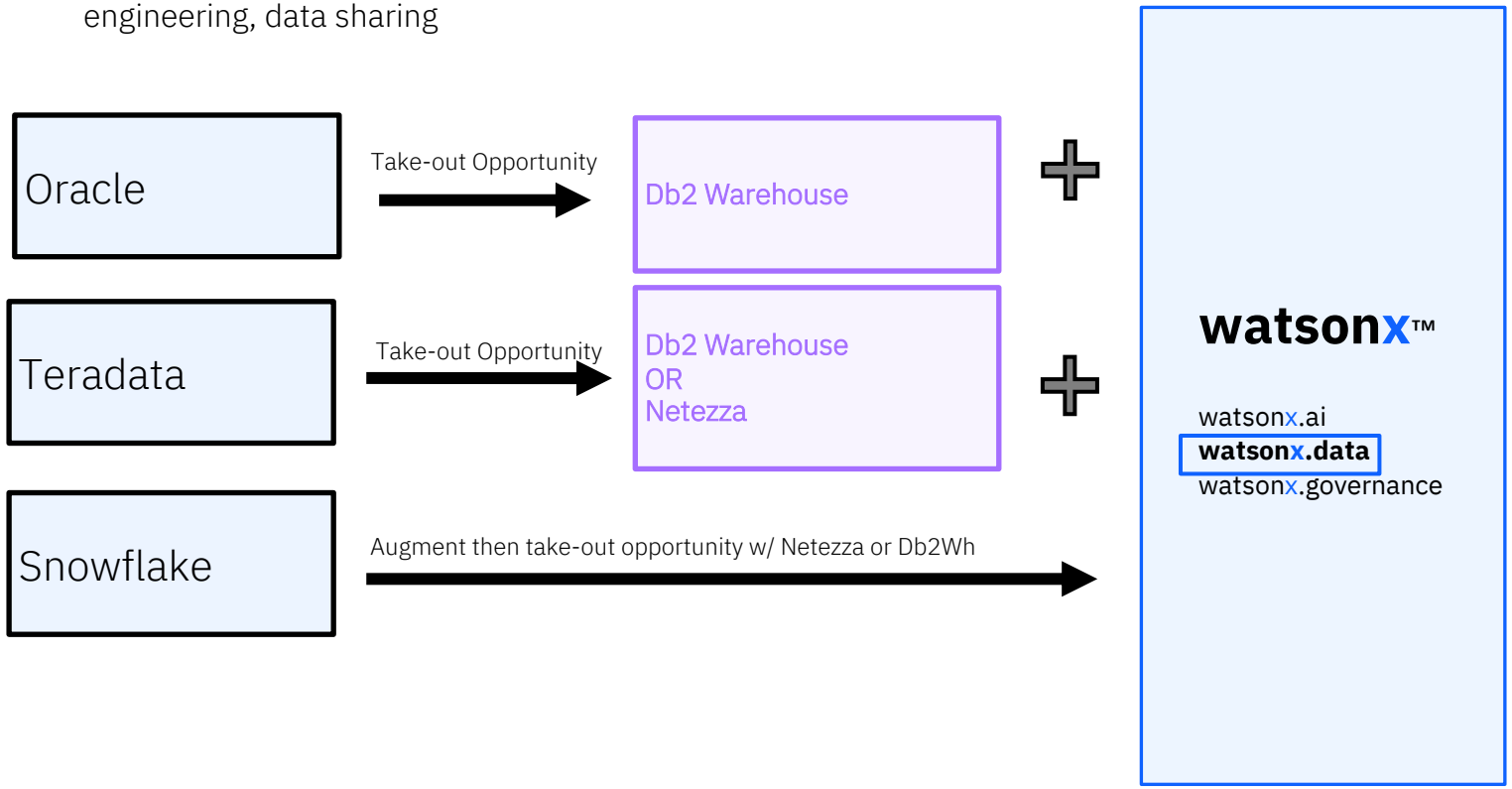
Customer Goal: Customer has competitor solution and is dissatisfied with expense of current solution or desires an IBM solution

Customer motivation:

- Costs too high with existing software or cloud solution
- Difficult relationship with existing software provider
- Need flexibility to scale new use cases and all workloads running on same instance: AI/ML, Real-time analytics, data engineering, data sharing

Challenges to Address:

- System migration compatibility
- Cost controls on the cloud - this is the # one complaint we hear from Snowflake customers!
- Sizing management
- Access to all data across hybrid cloud without duplication/movement



Key Messages:

Why watsonx.data?

- Access all data for AI across hybrid cloud
- Optimize workloads for AI and reduce warehouse costs by 50%
- Built-in governance
- Leverage watsonx platform for trusted AI and generative AI capabilities

Why Db2 Warehouse?

- Simplify migration with Db2 compatibility mode (Oracle, Sybase, PostgreSQL)
- Fully managed SaaS offerings, or containerized deployments for self-managed software
- Seamless integration with .data and IBM ecosystem
- Use object storage to save on storage costs

Why Netezza?

- Leverage similar PostgreSQL construct to reduce migration complexity
- Fully managed SaaS offerings, or containerized deployments for self-managed software
- AI-infused elastic scaling for cost predictability
- Seamless integration with .data and IBM ecosystem

Concepts and market entry points

Three key concepts for IBM watsonx.data

1. Presto is a next-generation open-source SQL engine designed to run efficiently over data lakes.

2. Warehouses and first-generation lakehouses are monolithic, and not optimized to work on all workloads. Only IBM watsonx.data's multi-engine architecture allows for true workload optimization.

3. Iceberg is an open-table format that allows multiple engines to access the same data – this means, Snowflake, Netezza, and IBM watsonx.data can all access data in Iceberg at the same time.

Market entry points + use cases

1. Warehouse Optimization narrative

- Competitive and IBM (Netezza + Db2)
- Talk track – Client is concerned with the spend on traditional warehouse today – looking to optimize for both performance and cost
- Value prop: Cost optimization and openness through the shared meta layer and fit-for-purpose engines
- Use case – Snowflake write-intensive workloads moving to Spark and/or Presto, thus reducing cost of Snowflake virtual data warehouses

2. Modernizing data lake narrative

- Modernizing storage architecture to facilitate shared metadata and fit-for-purpose engines
- Talk track – Converting legacy file storage structures into open-file structures and assigning those into the shared meta layer, thus facilitating fit-for-purpose engines
- Value prop: Modernize with warehouse-like performance for querying data in open formats, built-in governance
- Use case – Move from HDFS to open-source iceberg within a consolidated shared meta layer

Business Drivers

250%

The aggregate volume of data stored is set to grow over 250% in the next 5 years.

82%

82% of enterprises are inhibited by data silos.

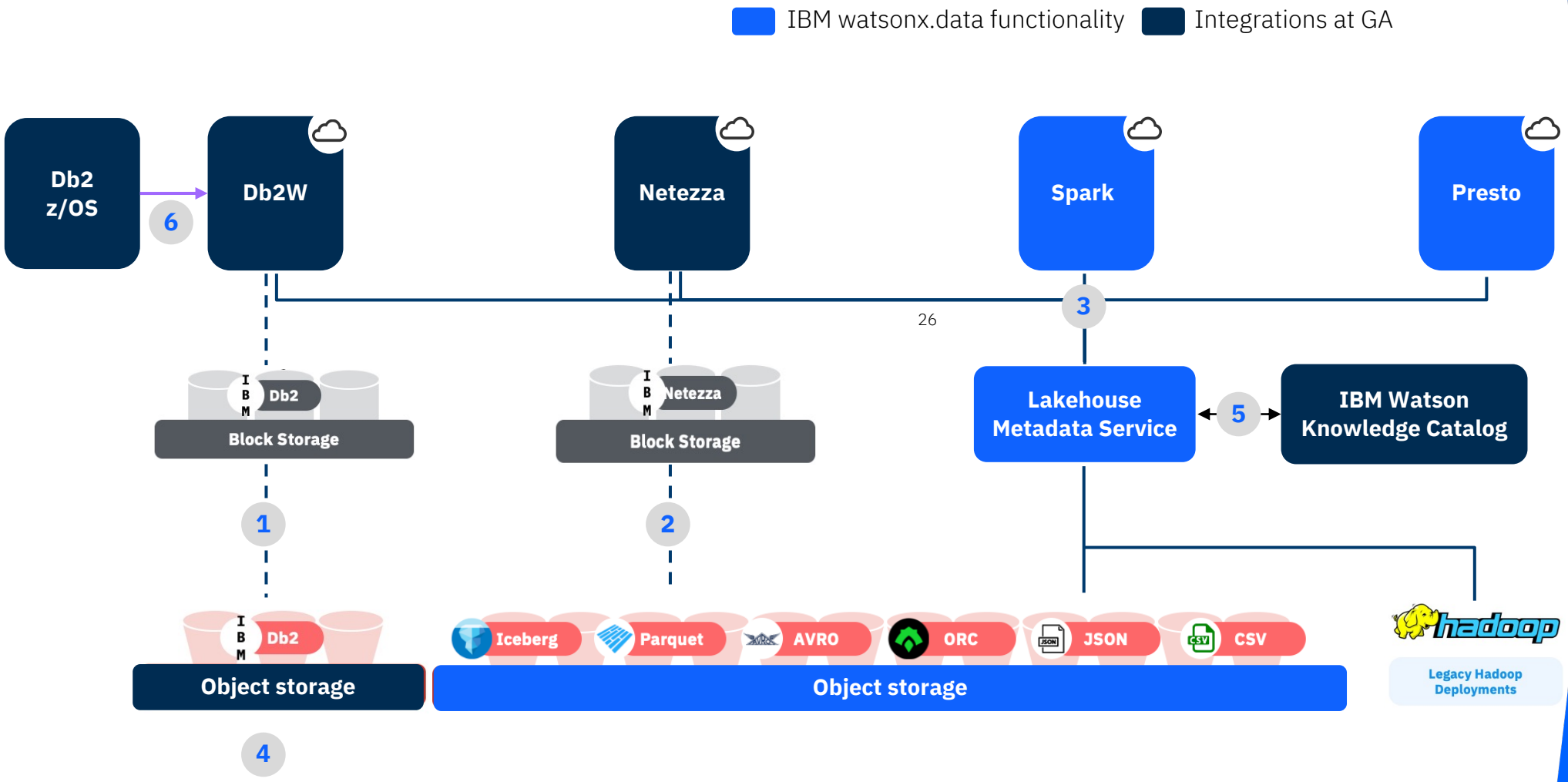
80%

80% of time is spent on data cleaning, integration and preparation.

82%

82% of enterprises say data quality is a barrier on their data integration projects.

The integrated IBM watsonx.data ecosystem for maximum workload coverage and optimal price-performance



1. Warehouses can access data in the lakehouse
2. Easily Promote data between the warehouse and lakehouse
3. Query routing service, multiple engines can access same data lake data
4. The lakehouse can access data residing in Db2/Netezza
5. WKC policies enforced by the lakehouse via metadata service
6. Analyze z data easily and securely with Db2 for z/OS Data Gate

Gartner Magic Quadrant for Data Quality Solutions

Forrester wave: Data Management for Analytics

watsonx.data Differentiators

- No other data lakehouse offering has **integrated data warehouse engines** in addition to the Apache Spark and open-source query engines
- The cloud hyperscalers (AWS, Microsoft Azure, and GCP) along with Databricks provide no **hybrid cloud deployment capability**
- **Deployment flexibility in other clouds** – no other data lakehouse offering can be deployed as easily across different cloud platforms
- Other data lakehouse competitors do NOT have the level of experience with mission critical applications, and **level of research in query optimization and query processing**, as IBM
- Watsonx.data also can be coupled with **Watson Knowledge Catalog** – a leader in the Gartner Magic Quadrant
- Watsonx.data and its selection of Apache Iceberg and Presto delivers **an open solution versus a single contributor open-source lock-in**