Db2 and watsonx.data

The warehouse
meets the lakehouse

Data Server Day 2024
Stockholm, Sweden

Kelly Schlamb
WW Technology Sales Enablement, IBM
kschlamb@ca.ibm.com

Notices and disclaimers

© 2024 International Business Machines Corporation.
All rights reserved.

This document is distributed “as is” without any warranty, either express or implied. In no event shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.

Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM.

Not all offerings are available in every country in which IBM operates.

Any statements regarding IBM’s future direction, intent or product plans are subject to change or withdrawal without notice.

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at “Copyright and trademark information” at: www.ibm.com/legal/copytrade.shtml.

Certain comments made in this presentation may be characterized as forward looking under the Private Securities Litigation Reform Act of 1995.

Forward-looking statements are based on the company’s current assumptions regarding future business and financial performance. Those statements by their nature address matters that are uncertain to different degrees and involve a number of factors that could cause actual results to differ materially. Additional information concerning these factors is contained in the Company’s filings with the SEC.
Copies are available from the SEC, from the IBM website, or from IBM Investor Relations.

Any forward-looking statement made during this presentation speaks only as of the date on which it is made. The company assumes no obligation to update or revise any forward-looking statements except as required by law; these charts and the associated remarks and comments are integrally related and are intended to be presented and understood together.

Agenda

01 watsonx.data overview

02 Components

03 Data warehouse modernization & augmentation

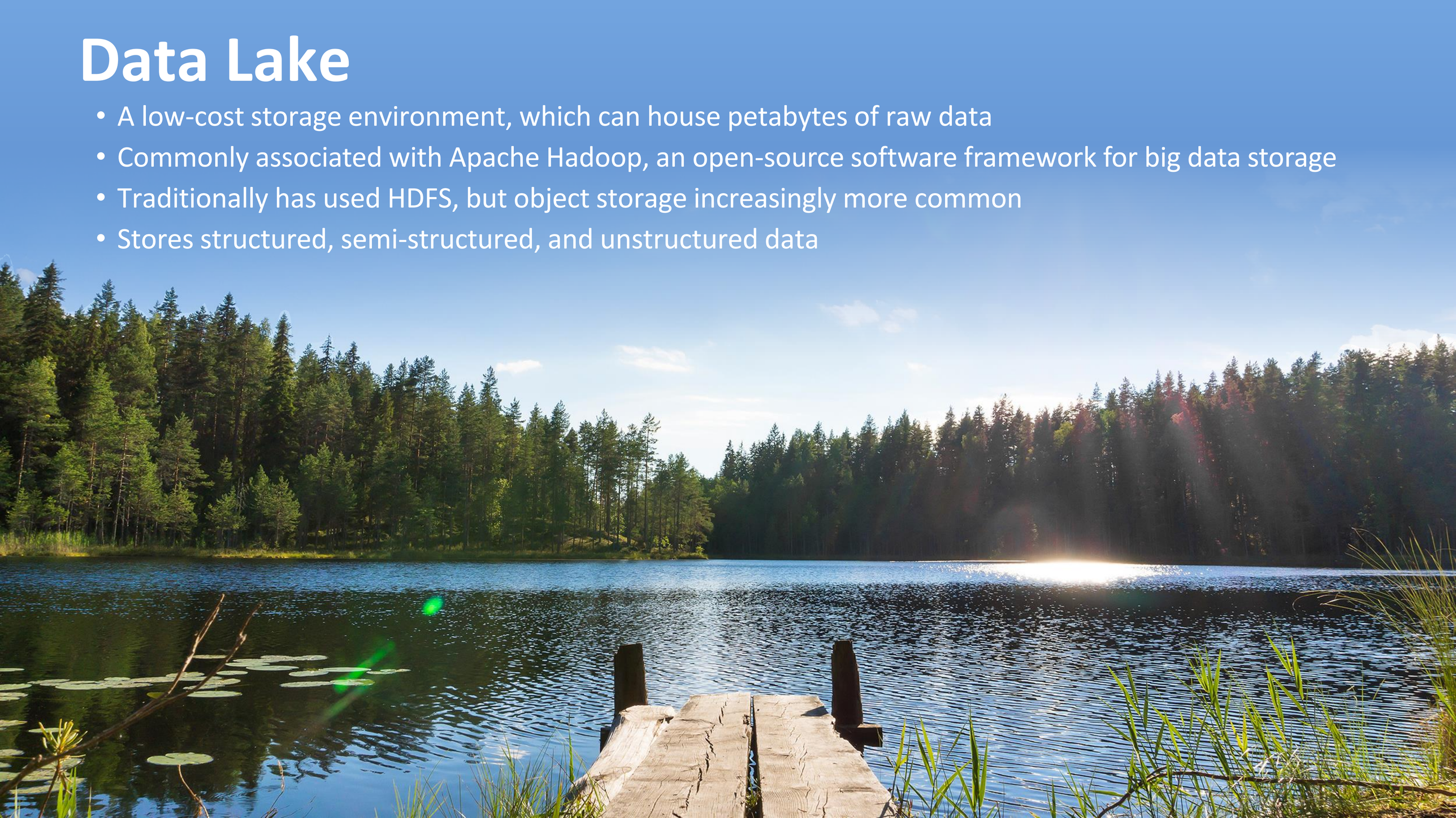
04 watsonx.data & Db2 WH data sharing

Data Warehouse

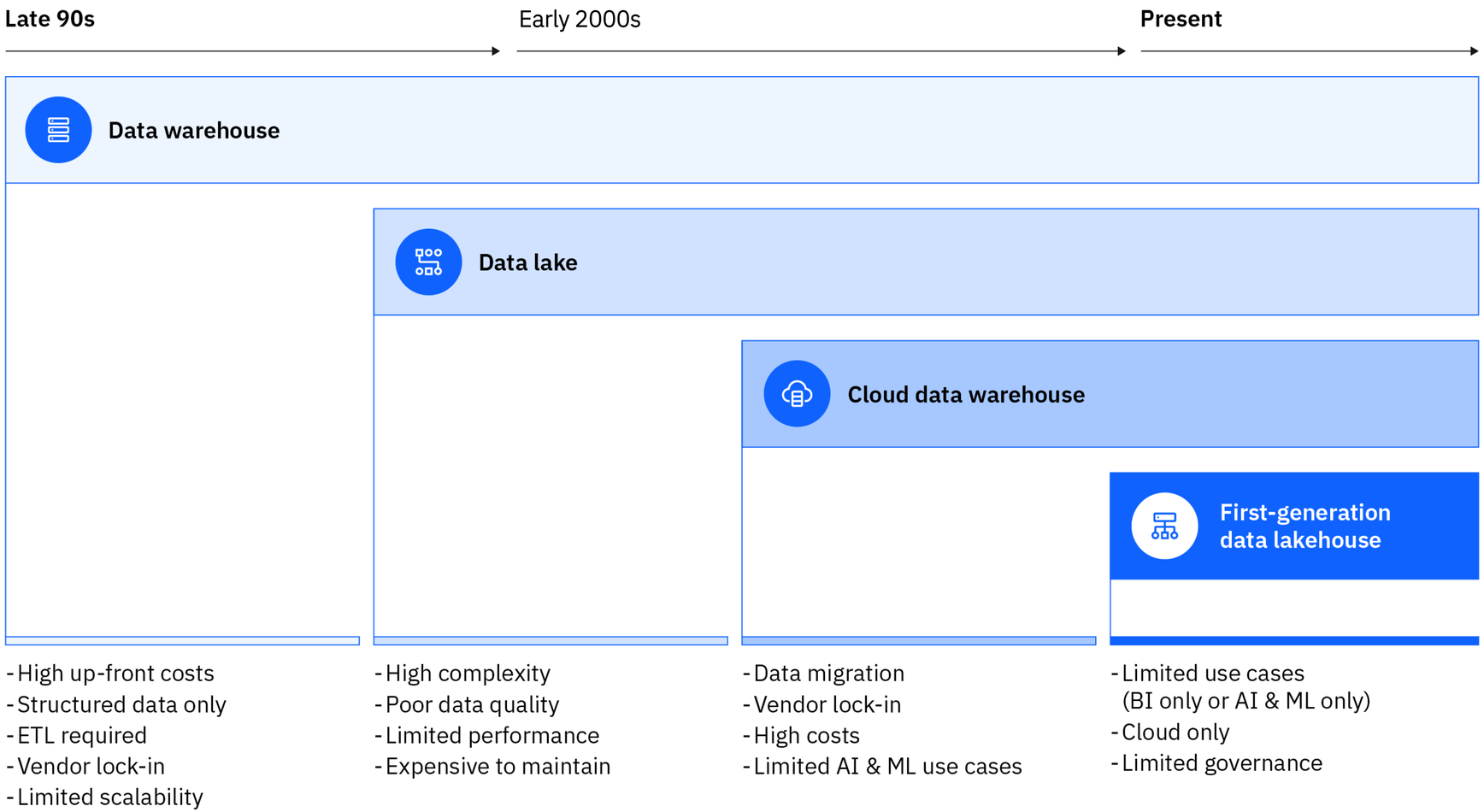
- Highly performant data management platform
- Data from multiple sources organized into a centralized, highly-structured relational database
- Primarily supports data analytics and business intelligence applications
- Data stored in proprietary formats on fast, expensive block-based storage devices

Data Lake

- A low-cost storage environment, which can house petabytes of raw data
- Commonly associated with Apache Hadoop, an open-source software framework for big data storage
- Traditionally has used HDFS, but object storage increasingly more common
- Stores structured, semi-structured, and unstructured data



Traditional approaches to addressing data management challenges have created more overall complexity and cost, which has led to the emergence of data lakehouse architectures



Today, leaders at most large enterprises manage their data and workloads using a mix of data repositories and data stores in hybrid environments.

The overall cost across all these repositories remains high.

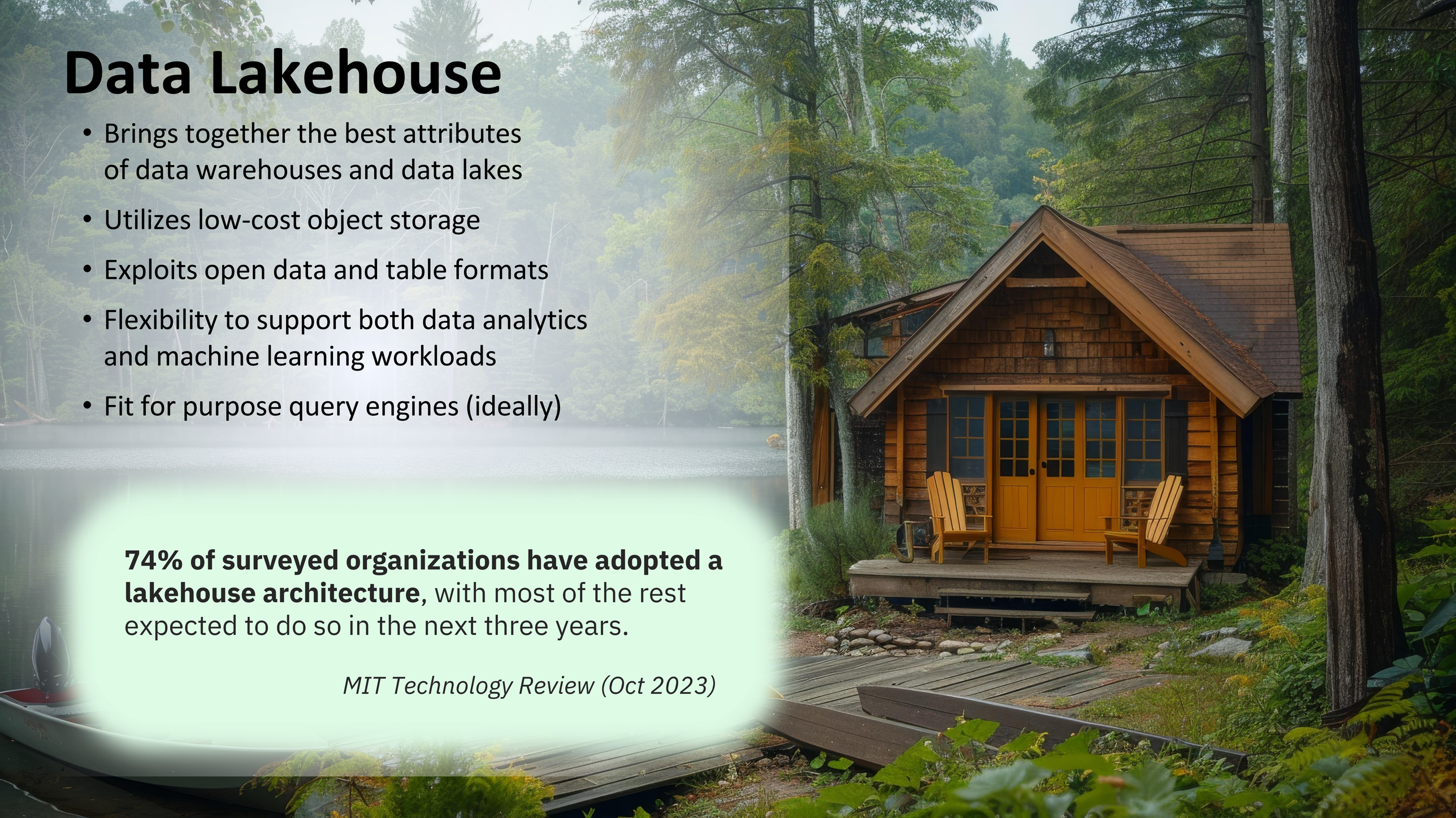
It’s difficult for leaders to effectively leverage and govern the data across multiple environments and use enterprise data for analytics and AI.

Data Lakehouse

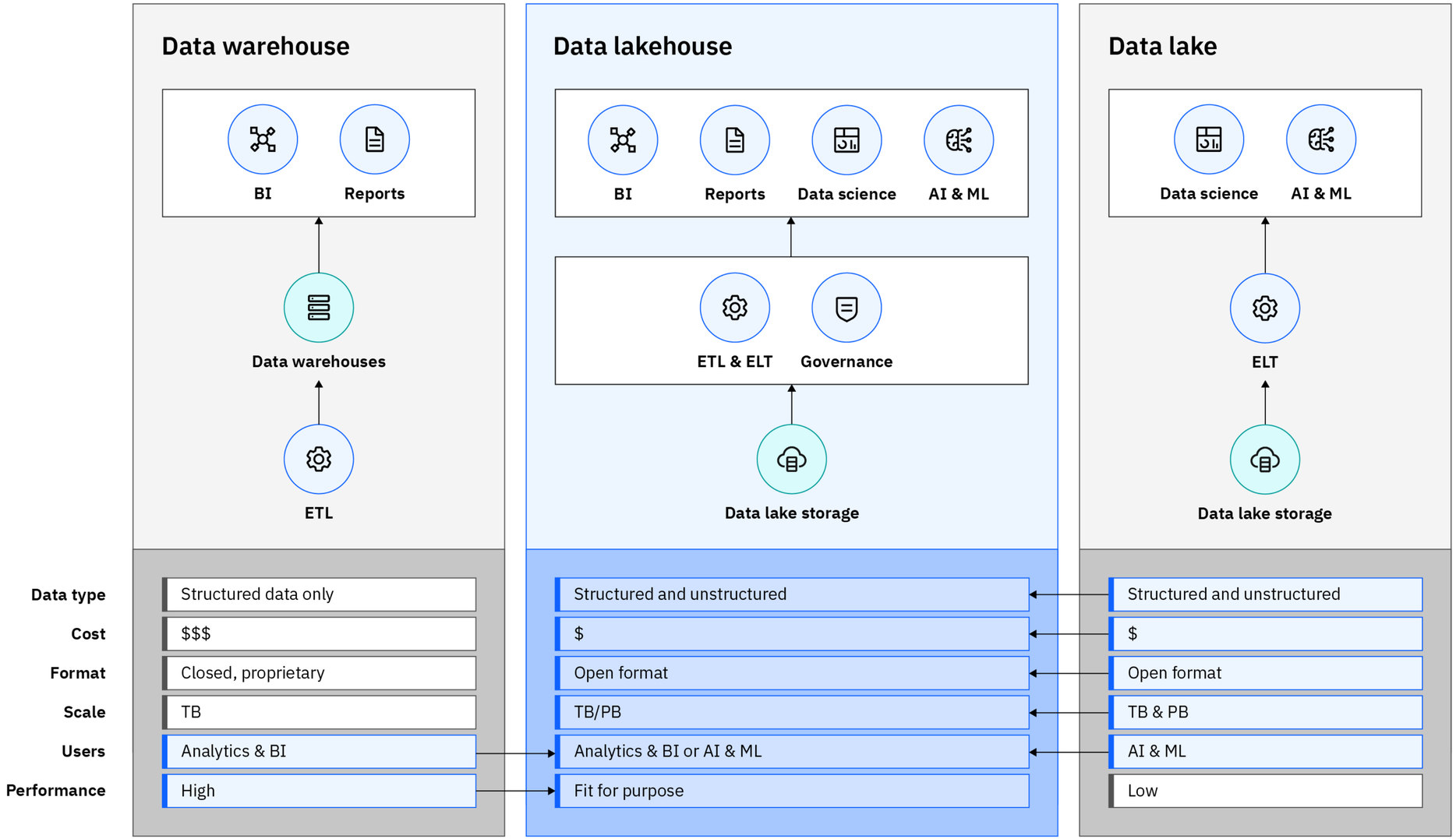
- Brings together the best attributes of data warehouses and data lakes
- Utilizes low-cost object storage
- Exploits open data and table formats
- Flexibility to support both data analytics and machine learning workloads
- Fit for purpose query engines (ideally)

74% of surveyed organizations have adopted a lakehouse architecture, with most of the rest expected to do so in the next three years.

MIT Technology Review (Oct 2023)



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



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 on cloud only with no support for multi-/hybrid-cloud deployments
- Minimal governance and metadata capabilities to deploy across the entire ecosystem

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 you 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

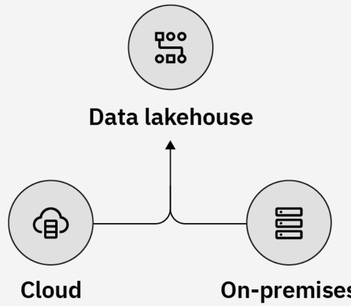
watsonx.data

Scale AI workloads, for all your data, anywhere

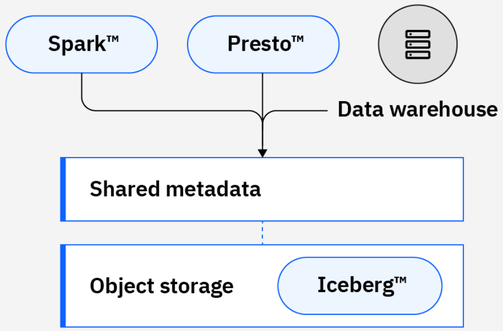
A hybrid, open data lakehouse to power AI and analytics with all your data, anywhere – supported by querying, governance, and open data formats to access and share data.

Seamlessly deploy across any cloud or on-premises environment in minutes with workload portability through Red Hat® OpenShift®.

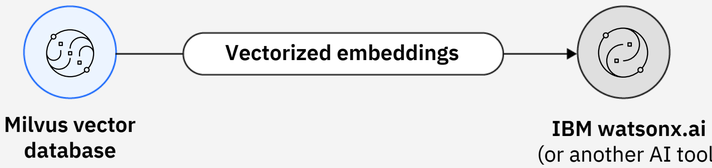
Access all your data through a single point of entry across all clouds and on-premises environments.



Reduce the cost of your data warehouse by up to 50%* through workload optimization across multiple query engines and storage tiers.



Unify, curate, and prepare data for AI



*When comparing published 2023 list prices normalized for VPC hours of IBM watsonx.data to several major cloud data warehouse vendors. Savings may vary depending on configurations, workloads and vendors.

watsonx.data

Use Cases

Data warehouse optimization

Optimize workloads from your data warehouse by choosing the right engine for the right workload, at the right cost. Replace ETL jobs and reduce costs of your data warehouse by up to 50% through workload optimization.

Data lake modernization

Augment Hadoop data lakes using watsonx.data and access better performance, security, and governance, without migration or ETL

Mainframe data for AI

Unleash the power of mainframe data for AI and analytics in watsonx.data with integration to IBM Data Gate for watsonx and Data Virtualization Manager for z/OS. Readily virtualize or replicate data to Iceberg for analytics and AI.

Datastore for Generative AI

Unify, curate, and prepare data efficiently for AI models and applications. Integrated vectorized embedding capabilities enable RAG use cases at scale across large sets of your trusted, governed data.

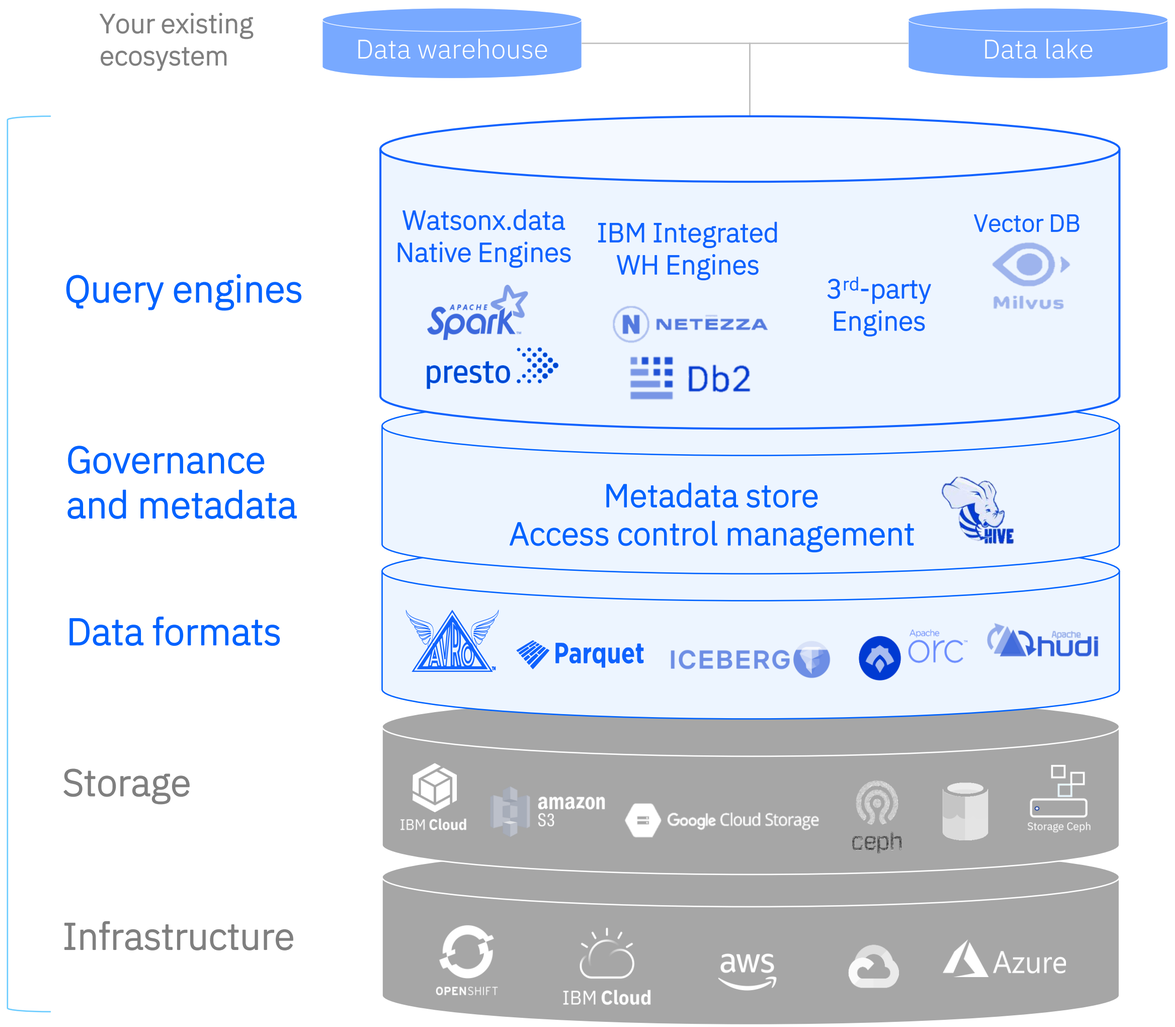
Generative AI powered data insights

Leverage Gen-AI infused in watsonx.data to find and understand data and unlock new data insights through semantic search — no SQL required. Unleash cryptic structured data using auto-generated semantic metadata in natural language for easy self-service access to data.

IBM **watsonx.data** – the next generation data lakehouse

Completely open.
No lock-in!

Built on a foundation of industry-embraced open-source technologies.



watsonx.data

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

Milvus for **semantic searching and RAG** uses cases

Built-in governance that is compatible with existing solutions such as watsonx.governance and IBM 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 multicloud environments

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

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

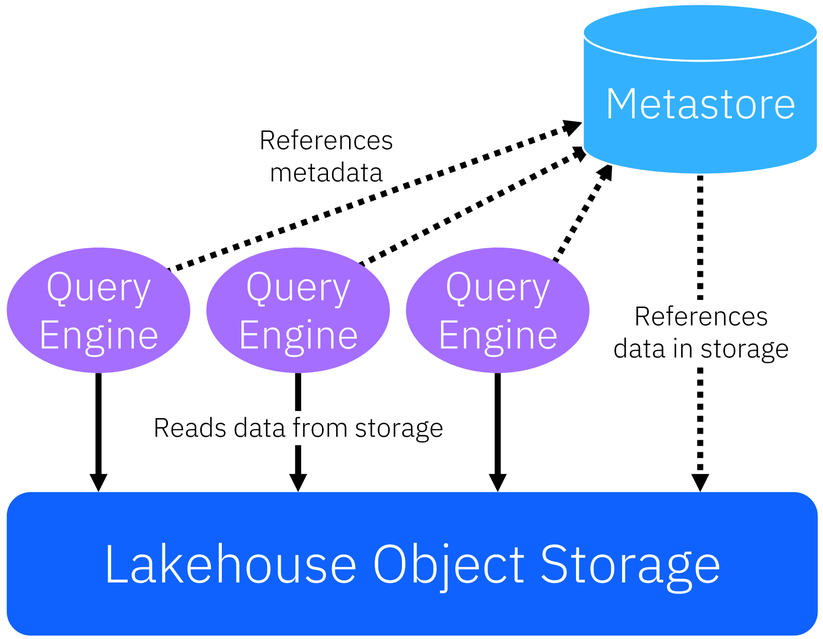
- Component of Microsoft Purview data governance solution
- Helps manage on-premises, multicloud, and SaaS data
- Offers discovery, classification, and lineage

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

Hive Metastore (HMS)

- Open-source **Apache Hive** was built to provide an SQL-like query interface for data stored in Hadoop
- **Hive Metastore (HMS)** is a component of Hive that stores metadata for tables, including schema and location
- HMS can be deployed standalone, without the rest of Hive (often needed for lakehouses, like watsonx.data)
- Query engines use the metadata in HMS to optimize query execution plans
- The metadata is stored in a traditional relational database (PostgreSQL in the case of watsonx.data)
- In watsonx.data, IBM Knowledge Catalog integrates with HMS to provide policy-based access and governance



presto

Make sense of all your data, any size, anywhere

Get the insights you need with Presto, a fast and flexible open-source SQL query engine

Scalable architecture

- Designed for analytic queries
- Uses open source query engines

Pluggable Connectors

- Allows access to external data sources without moving data
- Wide variety of connector for cloud and on- premises data sources

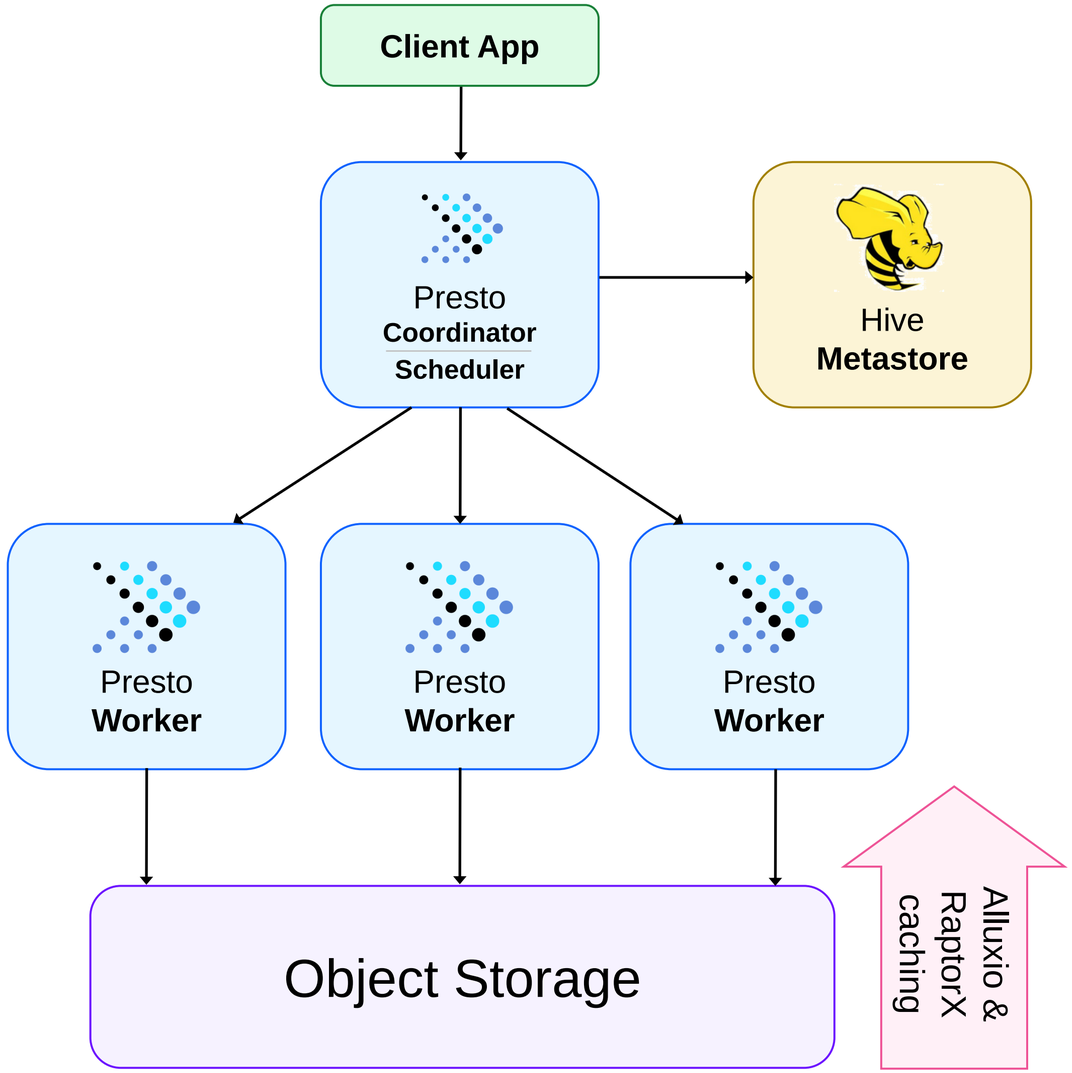
Performance

- MPP architecture for processing large data sets
- Can scale worker nodes as needed

Presto architecture

The structure of Presto is similar to that of classical MPP database management systems.

- **Client:** Issues user query and receives final result.
- **Coordinator:** Parses statement, plans query execution, and manages worker nodes. Gets results from workers and returns final result to client.
- **Workers (Java/C++):** Execute tasks and process data.
- **Connectors:** Integrate Presto with external data sources like object stores, relational databases, or Hive.
- **Caching:** Accelerated query execution through metadata and data caching (provided by Alluxio and RaptorX).



Powered by

presto

Adobe

Digital advertising platform

Over 2000 daily reports and 100s of pipelines on a 7 PB data lake with over 400 billion records

Bolt

Ride-hailing, micromobility rentals, and food delivery in Europe and Africa

Up to 100,000 daily queries (over 1.5 million queries per month) with over 2000 active internal users on 2 PB data lake

Meta

Social media

30,000 queries per day with 1000 daily active users on a 300 PB data lake

Uber

Ride-hailing, food delivery

Over 100 million queries per day with 7000 weekly active users on a 50 PB data lake

ByteDance

Internet technology

Over 2 million queries per day for business intelligence and one-off use cases

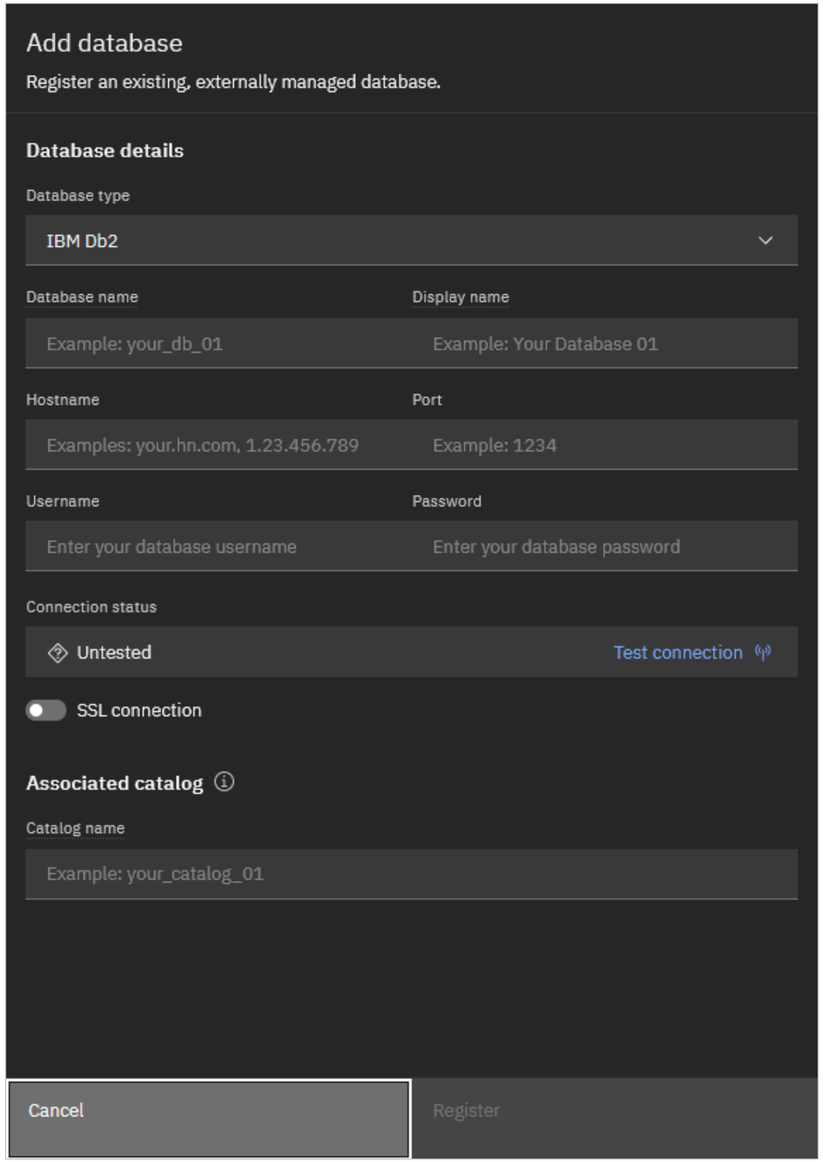
twilio

Communications API technology

Over 2700 active internal users running 1 million queries scanning 40 PB of data per month

Presto connectors in watsonx.data for federated data access

- IBM Db2
- IBM Netezza
- IBM Data Virtualization Manager for z/OS
- IBM Informix
- Apache Druid
- Apache Kafka
- Apache Pinot
- Amazon Redshift
- BigQuery
- Cassandra
- ClickHouse
- Elasticsearch
- MongoDB
- MySQL
- Oracle
- PostgreSQL
- Prometheus
- Redis
- SAP HANA
- SingleStore
- Snowflake
- SQL Server
- Teradata
- *... with more to come*



Data warehouse modernization

Workload cost optimization and overall cost reduction

The benefits:

- Offload workloads with lower performance requirements (e.g. historical data that must remain queryable, and workloads not suited for a data warehouse query engine)
- Workloads can be executed with the query engine that is ”fit for purpose” versus a single, expensive data warehouse engine
- Multiple data sources across different locations can reside in a single lakehouse versus multiple data warehouses

The components of watsonx.data used by this use case

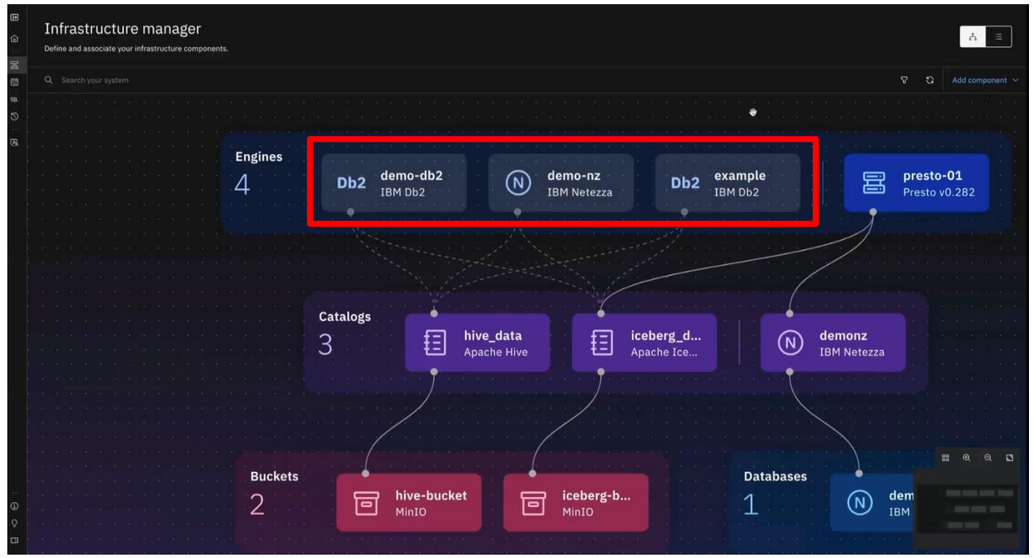
- Iceberg tables
- Apache Spark
- Presto
- Presto connectors (federated data access)

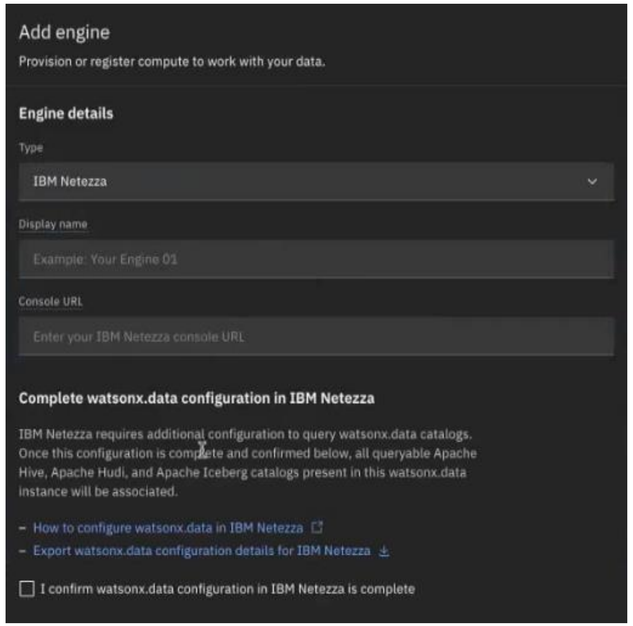
- Advantages versus competitors:
  - True hybrid-cloud – deploy anywhere
  - Includes open-source Spark and Presto query engines; most other lakehouses provide only a single query engine
  - Integrates with Netezza Performance Server and Db2 to allow data warehouse engines for queries that require optimum performance
  - Apache Iceberg is widely adopted which allows more data to be easily queried using watsonx.data

Db2 Warehouse and Netezza integration with watsonx.data

- New functionality in Db2 WH and Netezza (NPSaaS):
  - Ability to work with open data format tables in object storage (e.g. Db2's DATALAKE tables)
  - Integration with watsonx.data's metastore (w/ syncing of metadata for tables in object storage)

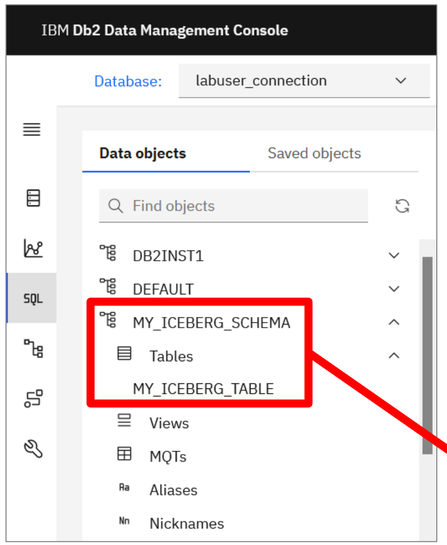
- Db2 WH and Netezza can be registered as "External Engines" in the watsonx.data console



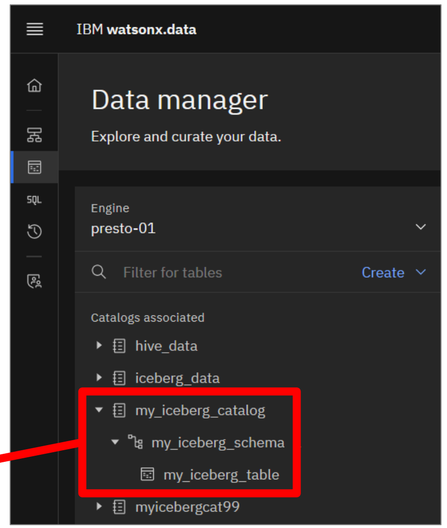


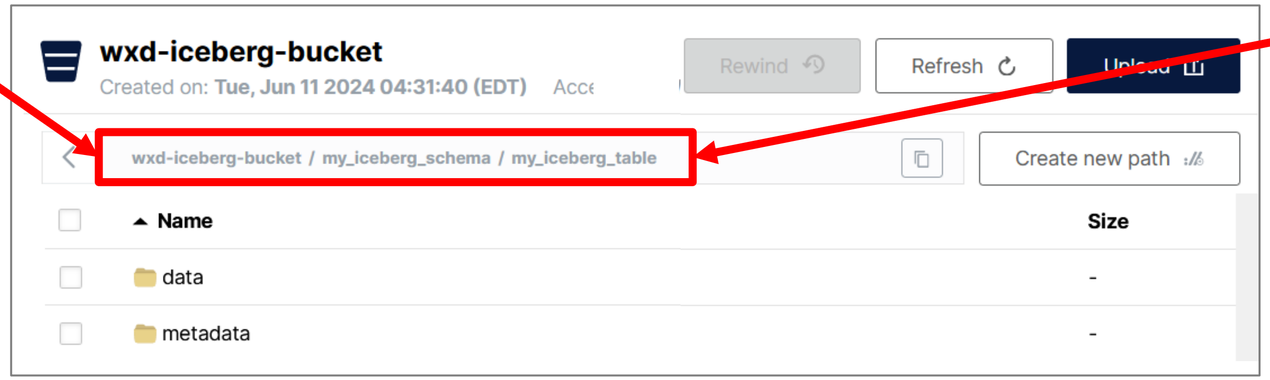
Db2 Warehouse

watsonx.data

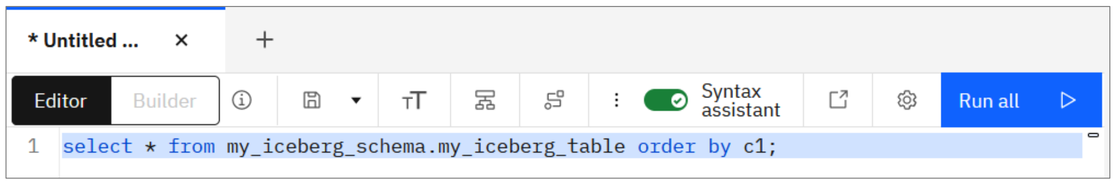


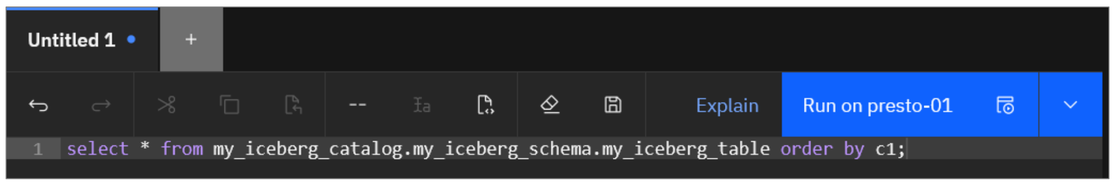
now also...



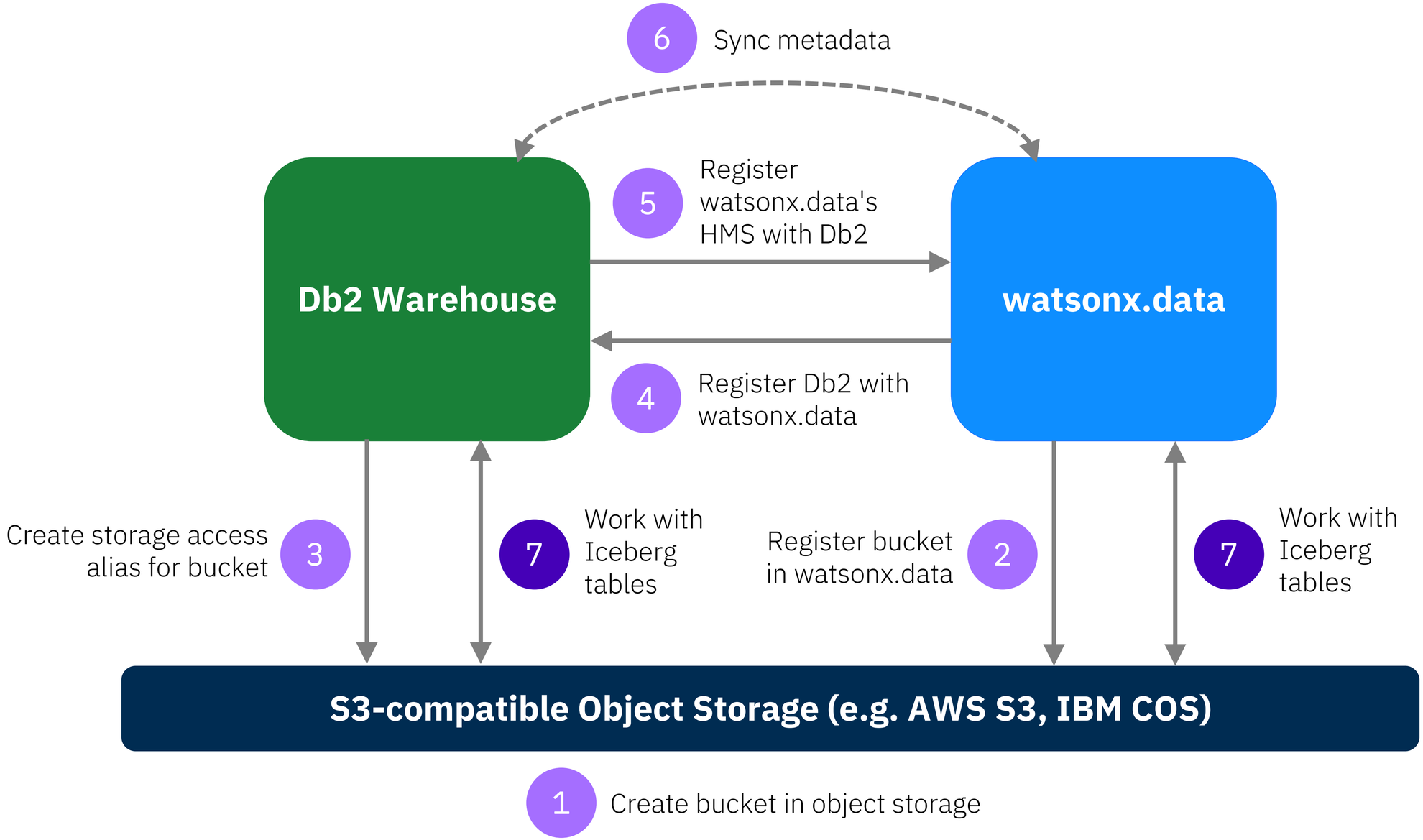


Concurrent access
to the same data





Configuring watsonx.data and Db2 Warehouse



Example steps for metastore integration in Db2WH

- Create storage access alias referencing the S3 bucket:

```
CALL SYSIBMADM.STORAGE_ACCESS_ALIAS.CATALOG('wxdalias', 'S3', '<S3-endpoint>',
  '<S3-accessKey>', '<S3-secretAccessKey', 'wxdbucket', '', 'I', '')
```

- Register watsonx.data's metastore:

```
CALL REGISTER_EXT_METASTORE('wxd', 'type=watsonx-data,uri=thrift://<hostname>:<port>', ?, ?)
CALL SET_EXT_METASTORE_PROPERTY('wxd', 'use.SSL', 'true', ?, ?)
CALL SET_EXT_METASTORE_PROPERTY('wxd', 'ssl.cert', '/etc/ssl/certs/lh-ssl-ts.crt', ?, ?)
CALL SET_EXT_METASTORE_PROPERTY('wxd', 'auth.mode', 'PLAIN', ?, ?)
CALL SET_EXT_METASTORE_PROPERTY('wxd', 'auth.plain.credentials', '<HMS-user>:<HMS-password>', ?, ?)
```

- From Db2WH, sync with watsonx.data's metastore and query an existing Iceberg table in the datalake:

```
CALL EXT_METASTORE_SYNC('wxd', wxdschema1', '.*', 'SKIP', 'CONTINUE', NULL);
SELECT * FROM WXDSCHEMA1.TABLE1;    (No CREATE DATALAKE TABLE statement needed here in Db2!)
```

Examples of creating tables in watsonx.data from Db2 WH

- SQL can be run from any Db2 WH connected application (e.g. CLP, DMC, 3rd-party app)
- Hive table example:

```
CREATE DATALAKE TABLE HSCHEMA.EMP_TABLE (EMP_ID INT, EMP_NAME CHAR(20)) STORED AS PARQUET
  LOCATION 'DB2REMOTE://wxdalias//hschema/emptable'
  TBLPROPERTIES('bigsql.external.catalog'='wxd');

INSERT INTO HSCHEMA.EMP_TABLE VALUES (1, 'Row1'),(2, 'Row2'),(3, 'Row3'),(4, 'Row4');

INSERT INTO HSCHEMA.EMP_TABLE VALUES (5, 'Row5'),(6, 'Row6'),(7, 'Row7'),(8, 'Row8');

SELECT * FROM HSCHEMA.EMP_TABLE ORDER BY EMP_NAME;
```

- Iceberg table example:

```
CREATE DATALAKE TABLE ISCHEMA.SALES_HISTORY STORED AS PARQUET STORED BY ICEBERG
  LOCATION 'db2remote://wxdalias//db2tables/ischema/sales_history'
  TBLPROPERTIES('iceberg.catalog'='wxd'
  AS SELECT * FROM SALES_SCHEMA.SALES_DATA;

SELECT SELLER_ID, SUM(SALE_TOTAL) AS SALE_TOTAL_SUM, SUM(GROSS_PROFIT) AS GROSS_PROFIT_SUM
  FROM ISCHEMA.SALES_HISTORY
  GROUP BY SELLER_ID ORDER BY SELLER_ID LIMIT 20;
```

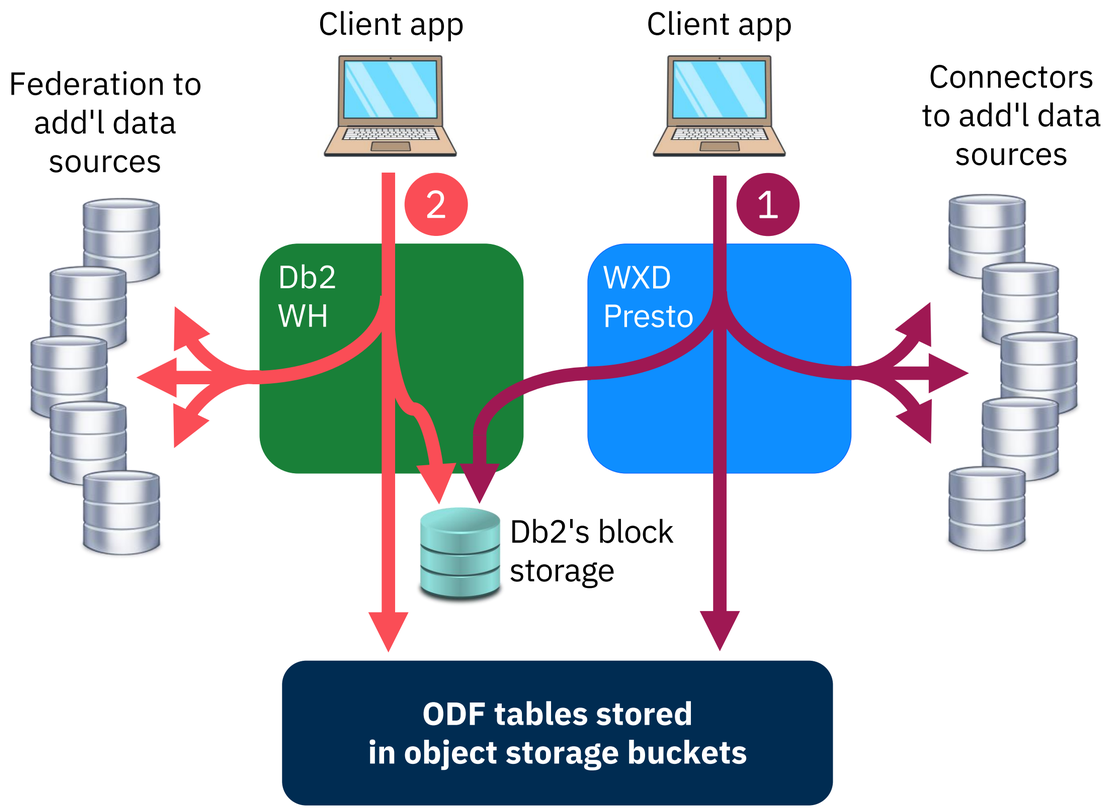
Flexible data access options

1. **Through Presto in watsonx.data**
   (offloaded workloads from the warehouse)
   - Tables in object storage (including Db2's lakehouse tables)
   - Federated access to tables in various data sources (inc. Db2 and Netezza)

2. **Through Db2 Warehouse**
   (leveraging Db2 WH query engine & compute)
   - Db2 tables stored locally (or natively in object storage in Db2WoC Gen 3)
   - Tables stored in open data formats on object storage
   - Federated access to various IBM and non-IBM data sources

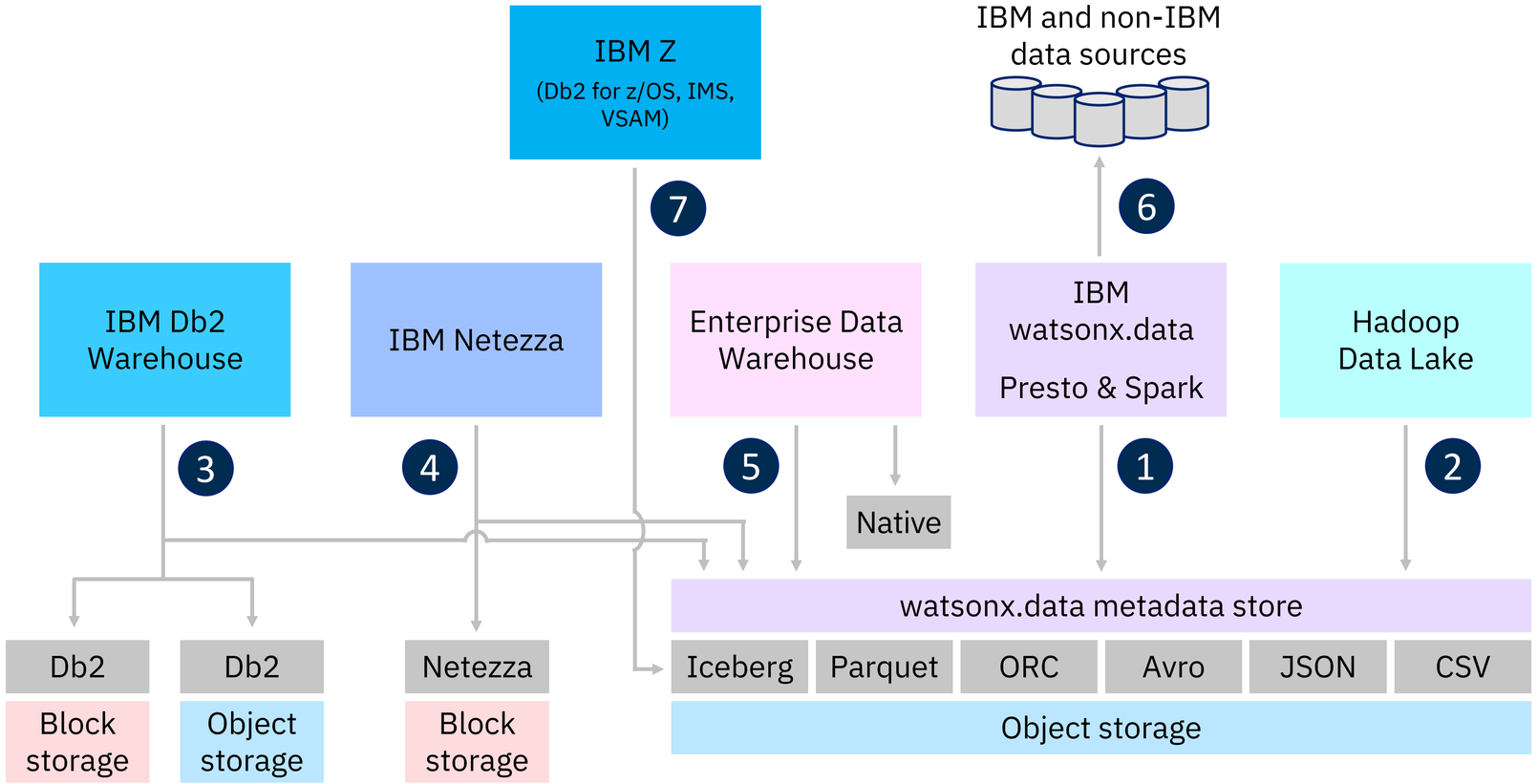
3. **Through Netezza** *(not shown)*
   (leveraging Netezza query engine & compute)
   - Tables stored locally
   - Tables stored in open data formats on object storage

4. **Through other query engines (e.g. Spark)** *(not shown)*
   (use fit-for-purpose engines depending on workload)



Db2 Warehouse with watsonx.data

IBM watsonx.data and your data ecosystem



1. Watsonx.data's native **Presto** and **Spark** engines work with open data and table formats.
2. Sync metadata with watsonx.data. Convert legacy file storage structures to Iceberg (over time).
3. **Natively store Db2 WH data in open data formats.** Offload/promote between Db2 WH and watsonx.data.
4. Access lakehouse data **natively through Netezza**.
5. For DWs that "speak" Iceberg, offload data/workloads to lakehouse.
6. Presto's connectors allow for **federated data access** to many different data sources, without having to move or copy data.
7. With Data Gate for watsonx, **replicate mainframe transactional data to Iceberg**, where it can be used for analytics and AI workloads.

Additional resources

My blogs:
- How to create datalake tables in Db2WH
- How to integrate Db2WH with watsonx.data

IBM watsonx.data documentation:
- SaaS (IBM Cloud, AWS)
- Software

IBM watsonx.data & Db2 WH integration:
- Accessing watsonx.data steps
- Remote storage connectivity steps
- Registering NPS/Db2WH engines in watsonx.data
- Using DATALAKE tables in Db2 Warehouse

Going to TechXchange in Las Vegas in October?
- Attend my hands-on lab on Db2 datalake tables and watsonx.data integration (session #1848)