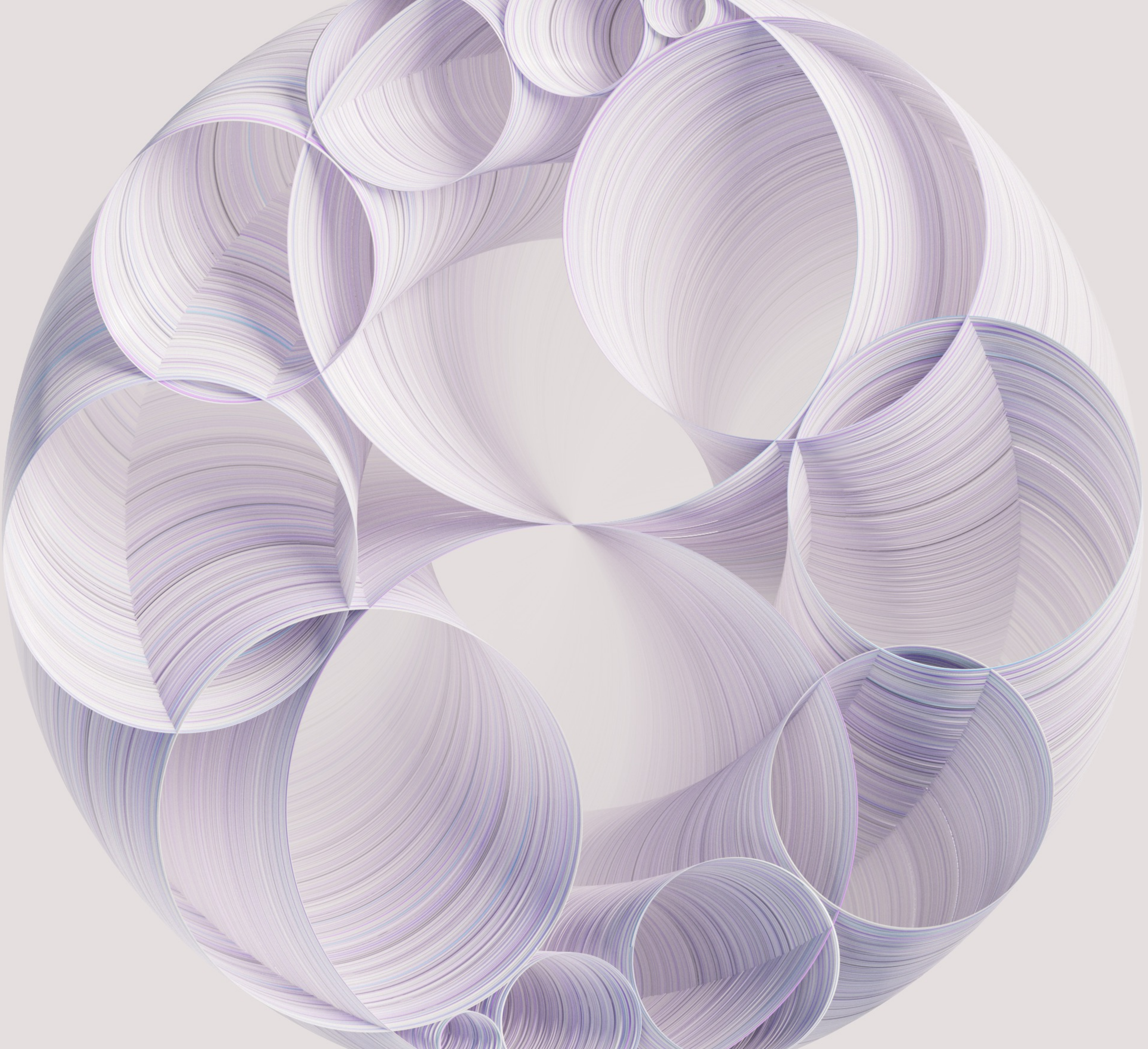
Bridging to the Lakehouse

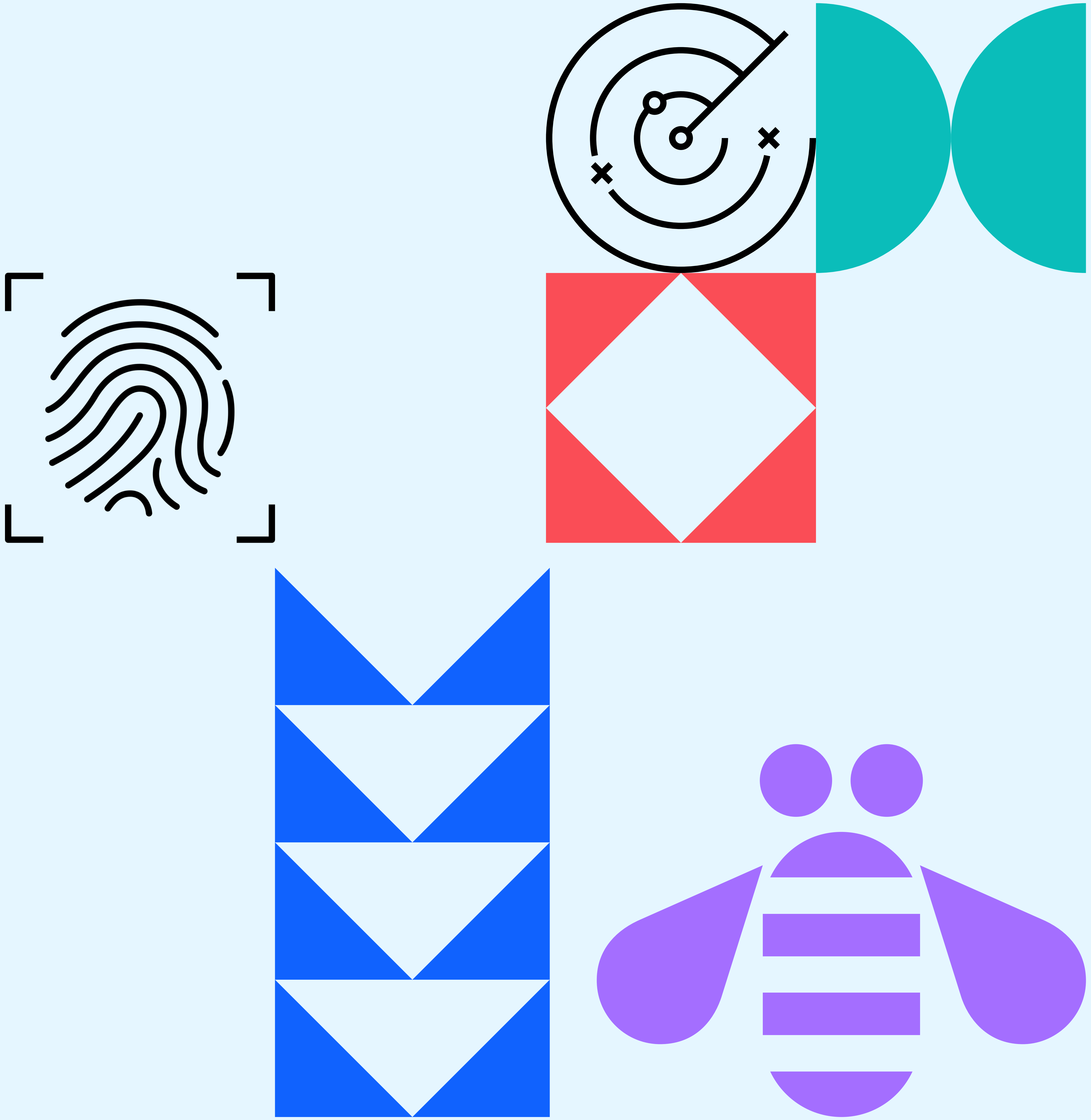
Connecting Db2 to watsonx.data



Francis Wong
Program Director – Db2 Development | IBM
fdewong@ca.ibm.com

Agenda

1. watsonx.data, an Introduction
2. watsonx.data Use Cases
3. Connecting watsonx.data to Db2



The speed, scope, and scale of generative AI impact is unprecedented

Sources: Statista; Reuters; Goldman Sachs; IBM Institute for Business Value; Gartner. Scale Zeitgeist: AI Readiness Report, a survey of more than 1,600 executives and ML practitioners

Massive early adoption

80%

of enterprises are working with or planning to leverage foundation models and adopt generative AI

Broad-reaching & deep impact

Generative AI could raise global GDP by

7%

within 10 years

Critical focus of AI activity & investment

Generative AI expected to represent

30%

of overall market by 2025

However, leaders are faced with unprecedented data challenges to scale AI

This environment leads to more cost and complexity for those who seek to govern data for AI.

There’s more data

Exploding data growth

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

In more locations

Multiple locations, clouds, applications and silos

82% of enterprises are inhibited by data silos.

In more formats

Documents, images, video

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

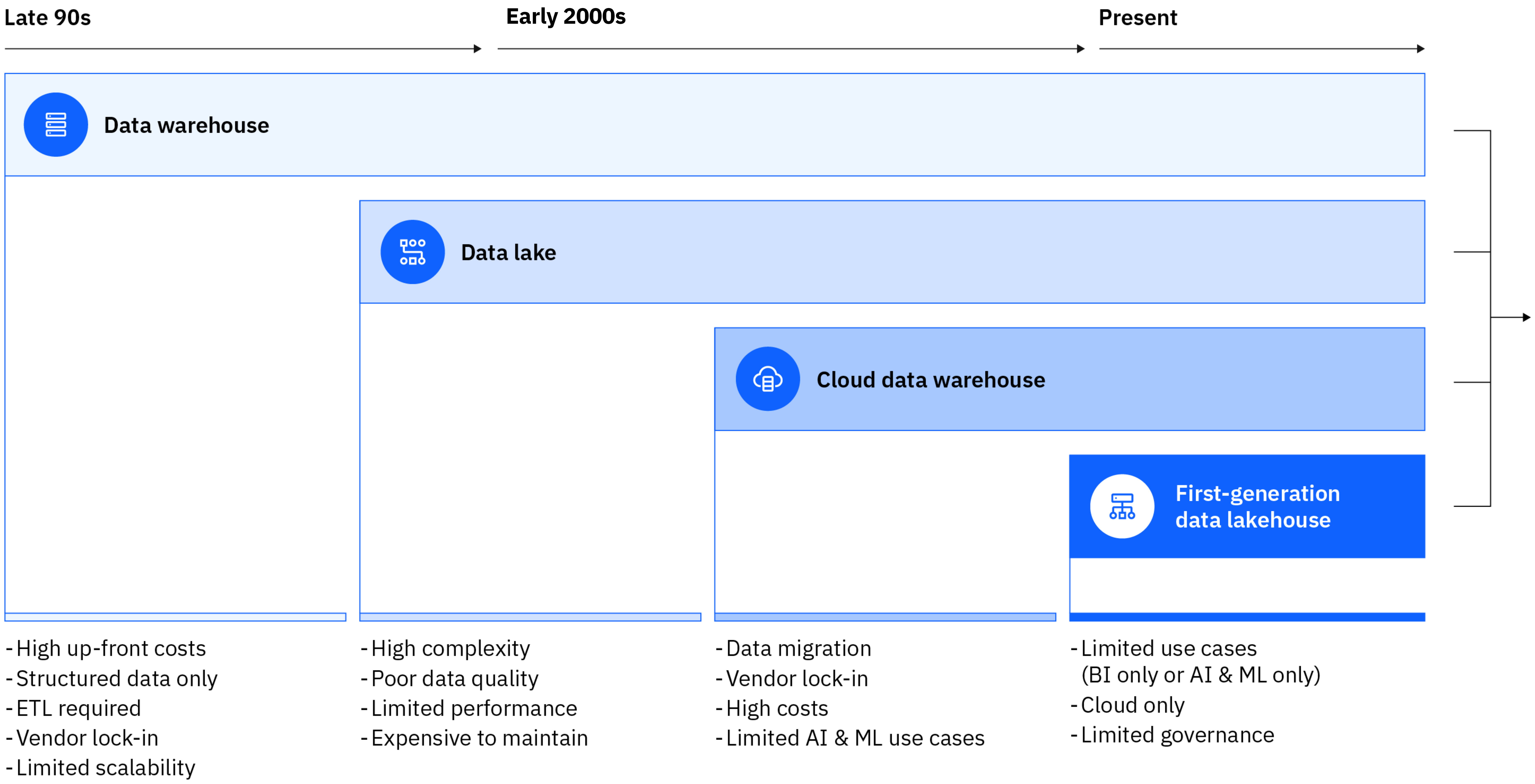
With less quality

Stale and inconsistent

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

Source: https://www.idc.com/getdoc.jsp?containerId=US49018922)

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



Late 90s

Early 2000s

Present

Data warehouse

Data lake

Cloud data warehouse

First-generation data lakehouse

- High up-front costs
- Structured data only
- ETL required
- Vendor lock-in
- Limited scalability

- High complexity
- Poor data quality
- Limited performance
- Expensive to maintain

- Data migration
- Vendor lock-in
- High costs
- Limited AI & ML use cases

- Limited use cases (BI only or AI & ML only)
- Cloud only
- Limited governance

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.

Enterprise leaders require a data architecture that can provide quick access to data, centralized governance and fit-for-purpose use.

1 Ability to scale AI while supporting compliance with lineage and reproducibility of data

2 Real-time analytics and BI that can connect to existing data in minutes without expensive duplicating or moving of data

3 Data sharing and self-service access for more users and more data while strengthening governance and security

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.

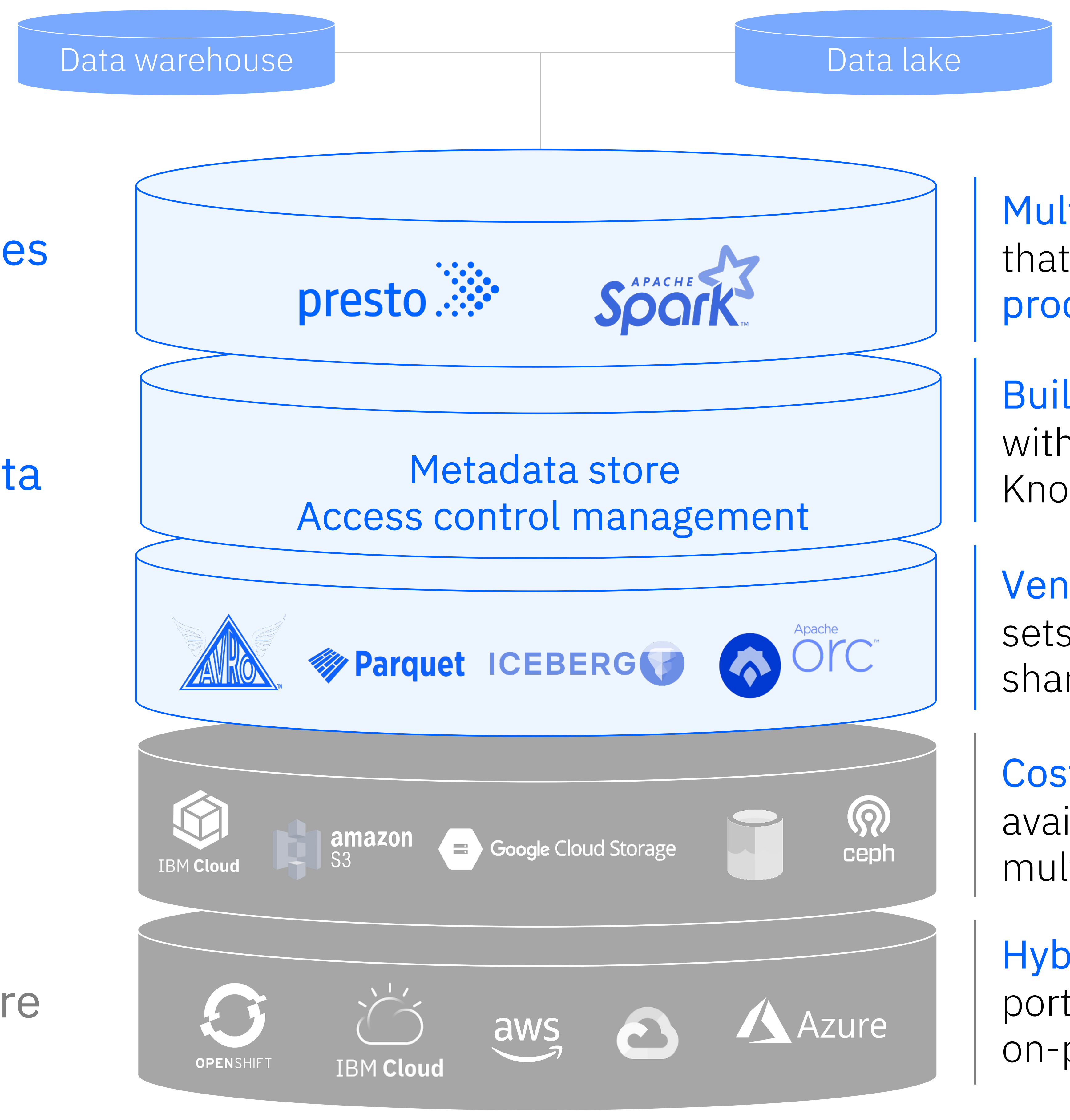
Optional

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

Core watsonx.data functionality

Ecosystem infrastructure

Your existing ecosystem



watsonx.data

Query engines

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

Optimize workload costs and performance using multi-engine functionality

Governance and metadata

Built-in governance that is compatible with existing solutions such as IBM Knowledge Catalog

Strengthen governance and reduce time to insight with centralized metadata and access management

Data format

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

Access all of your data across databases and data lakes

Storage

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

Reduce storage costs and facilitate data ingest

Infrastructure

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

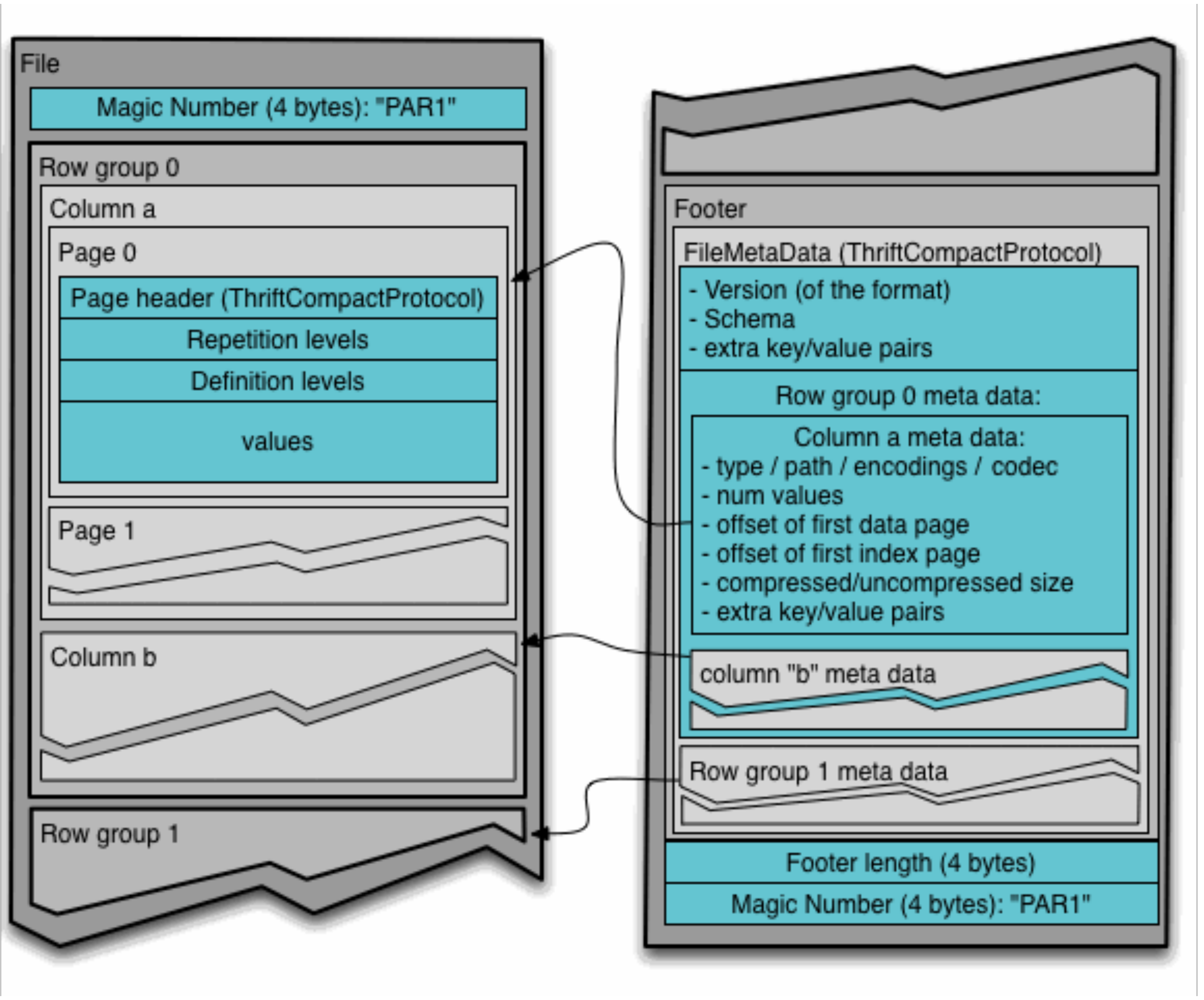
Deploy on any infrastructure and optimize available resources

Benefits of Open Data Formats (Parquet)

Open

Open Source. Reference implementation / format specifications publicly available

Support available for multiple tools and multiple programming languages. No vendor lock in.



File

Magic Number (4 bytes): "PAR1"

Row group 0

Column a

Page 0

Page header (ThriftCompactProtocol)

Repetition levels

Definition levels

values

Page 1

Column b

Row group 1

Footer

FileMetaData (ThriftCompactProtocol)

- Version (of the format)
- Schema
- extra key/value pairs

Row group 0 meta data:

Column a meta data:

- type / path / encodings / codec
- num values
- offset of first data page
- offset of first index page
- compressed/uncompressed size
- extra key/value pairs

column "b" meta data

Row group 1 meta data

Footer length (4 bytes)

Magic Number (4 bytes): "PAR1"

Optimized

Column organized for analytics use case fast reads & compression optimization

Self describing with file footer & pages carrying statistics enabling data skipping / predicate pushdown

A New Class of Open Data Formats
Apache Iceberg

ICEBERG

Full open-source, Open Data Table format, quickly becoming an industry standard

Relies on Open Data File formats for storage, but provides an additional layer of metadata for enhanced capabilities

Native Encryption

Full Schema Evolution

Data Compaction

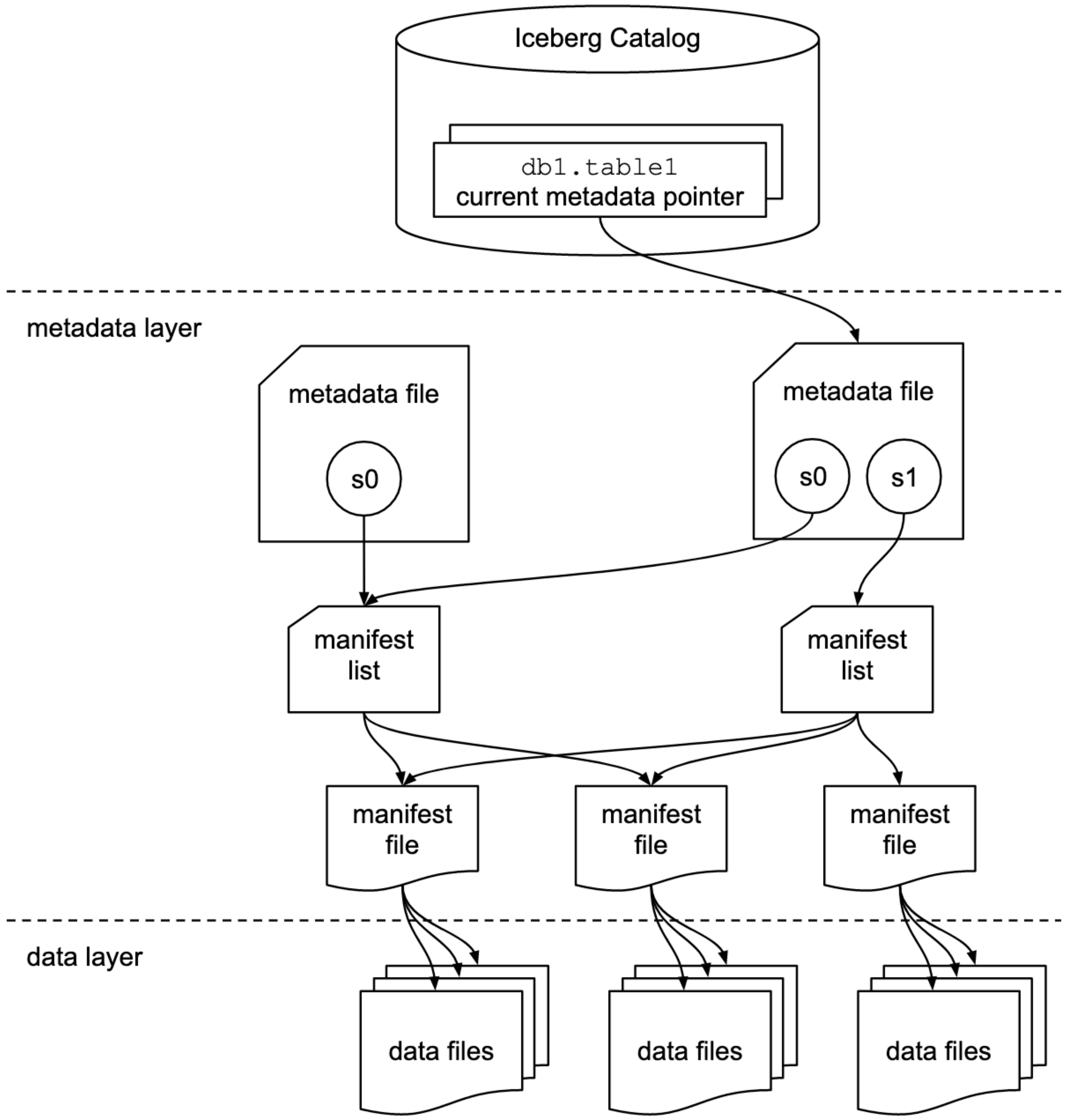
Hidden Partitioning

Integrated Compression

Time-Travel & Rollback

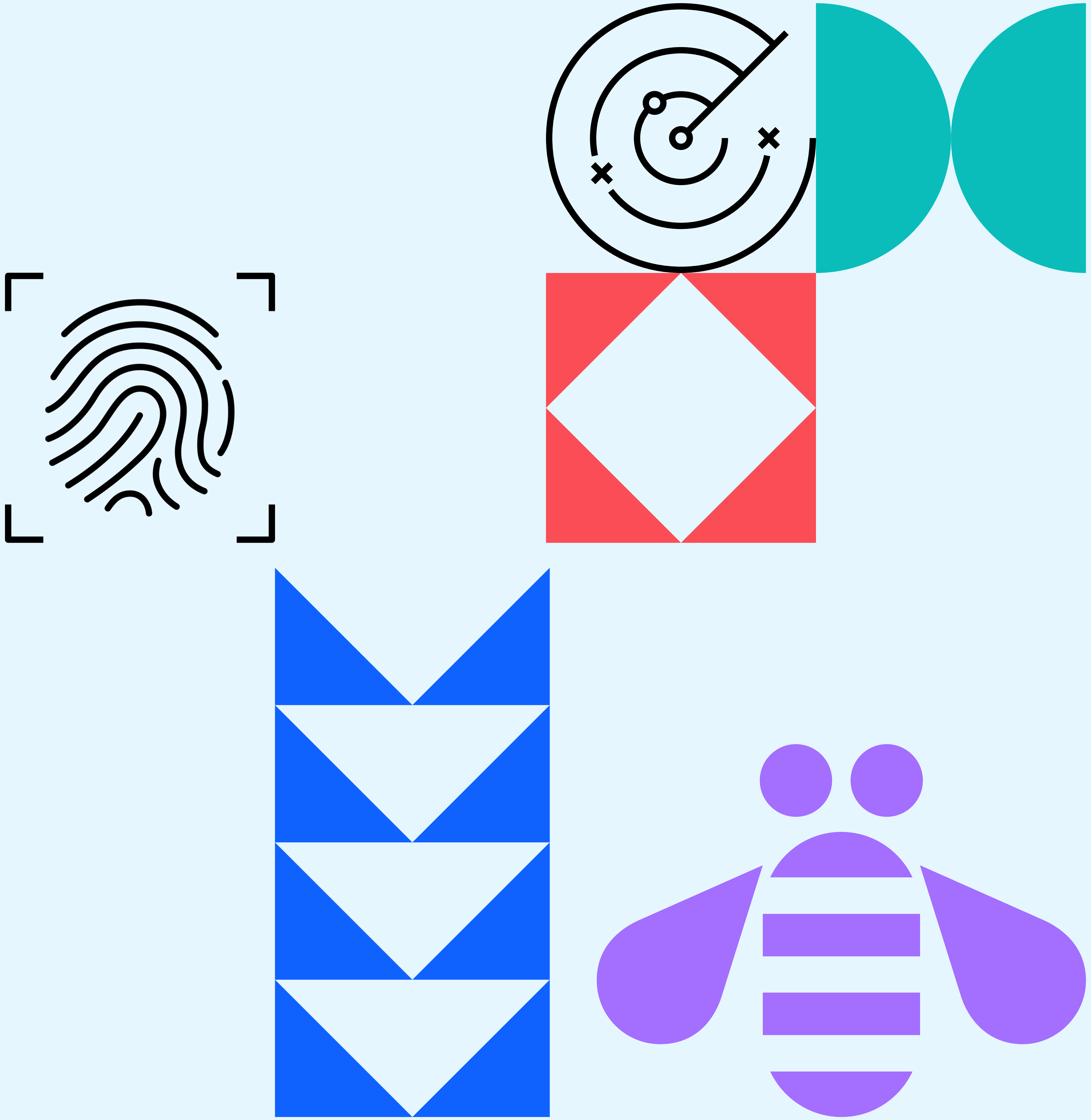
ACID Transactions

Expressive SQL



Agenda

1. watsonx.data, an Introduction
2. watsonx.data Use Cases
3. Connecting watsonx.data to Db2



Use Cases

Share data through an open format

Eliminate data silos by sharing Db2 tables with data lakes and lakehouse engines.

Optimize Workloads

Use the most appropriate tool for the task at hand without having to move or copy the data

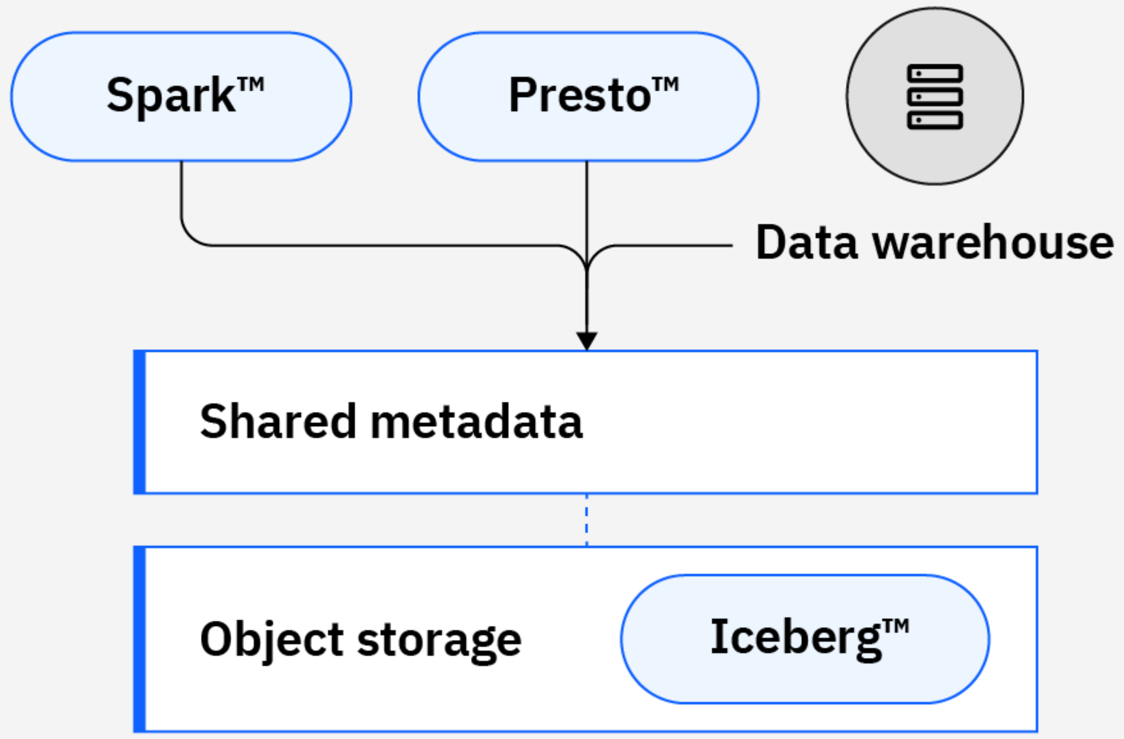
Warehouse Augmentation

Gain new insights from your warehouse data by combining Db2 Warehouse and data lakes platform data through open formats engine.

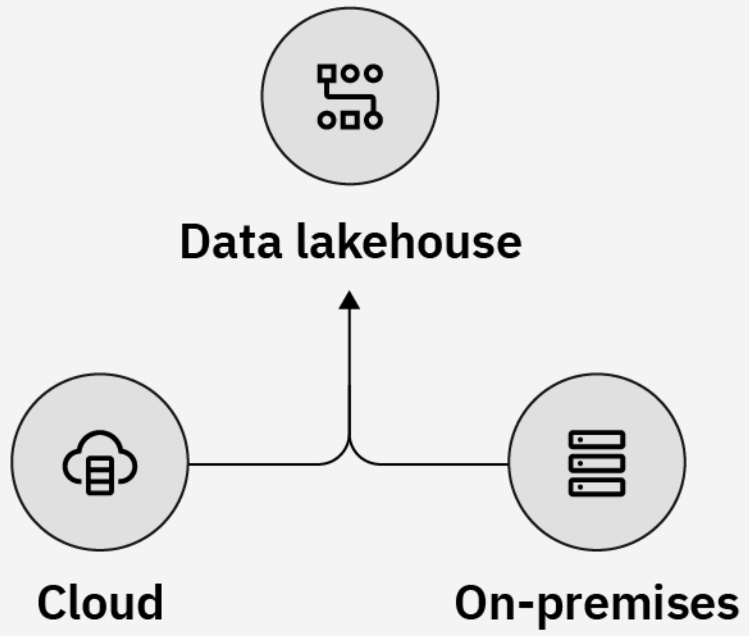
Share Data Through an Open Format

An open data store, based on an open lakehouse architecture built for hybrid deployment of your data, analytics, and AI workloads

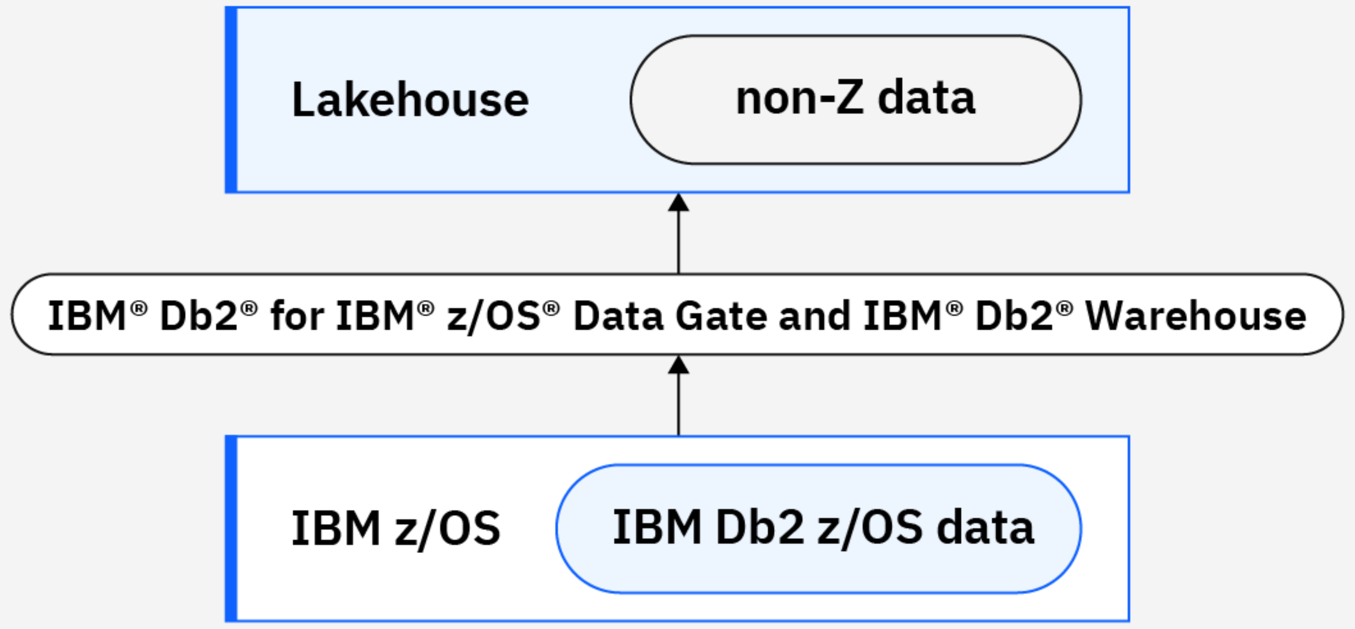
1. Share a single copy of data with tools that can read open data formats to minimize data duplication



2. Connect to and access data remotely across hybrid cloud with the ability to cache remote sources



3. Synchronize and incorporate Db2 for z/OS data for lakehouse analytics

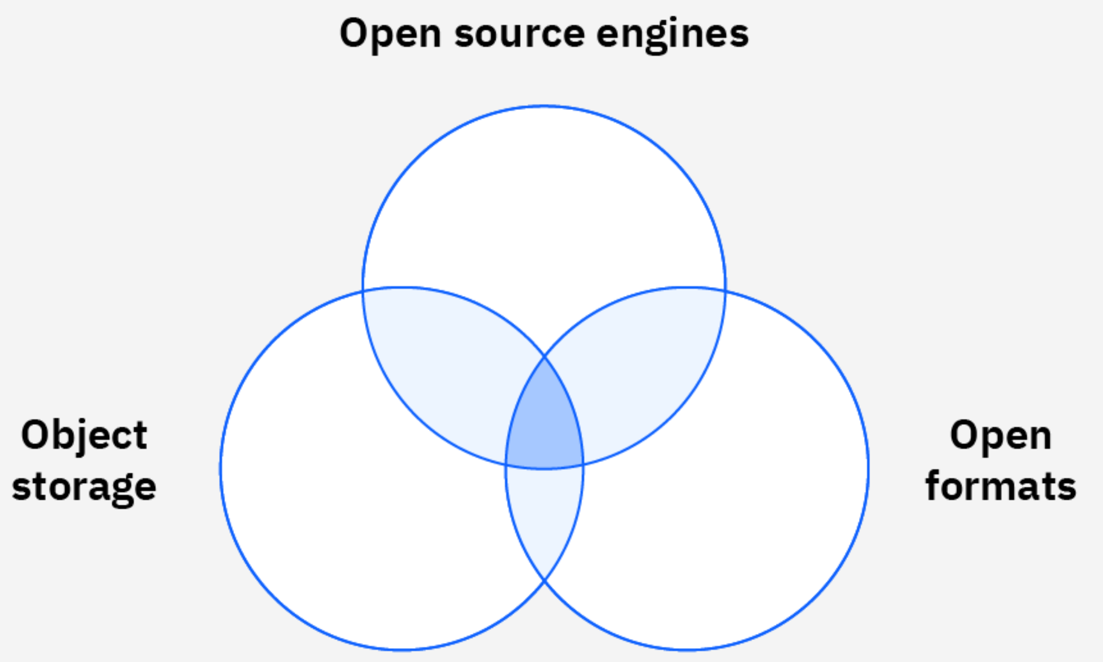


Optimize Workloads

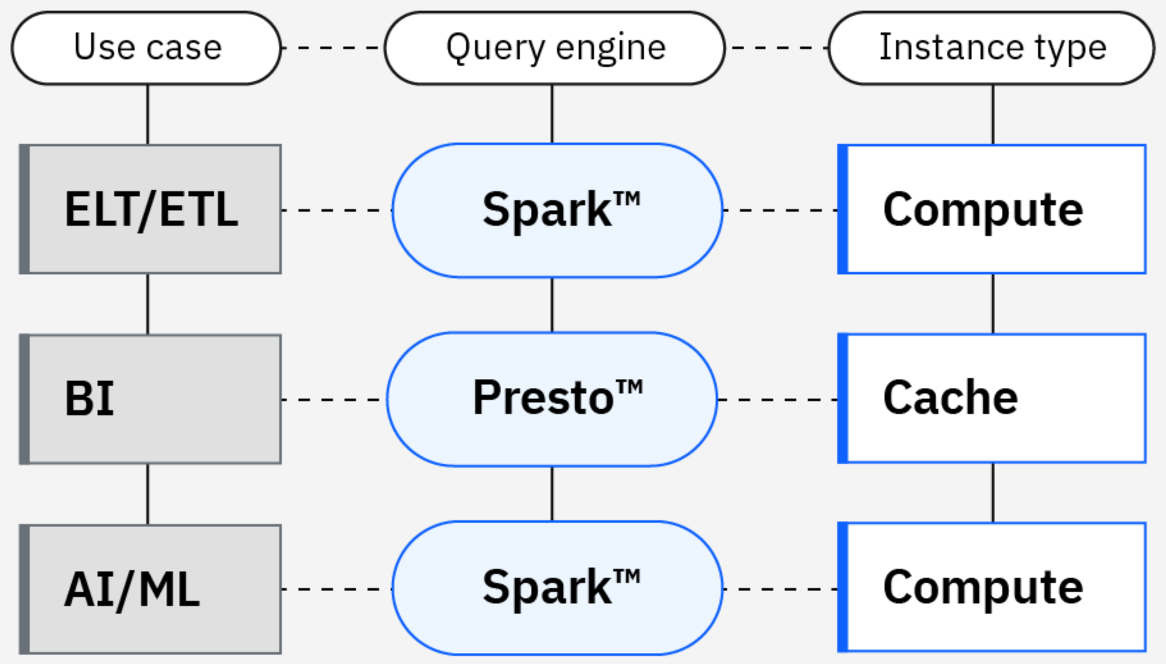
Optimize workloads from your data warehouse when you take advantage of low-cost object storage and fit-for-purpose query engines

Reduce data warehouse costs by up to 50%* by optimizing workloads

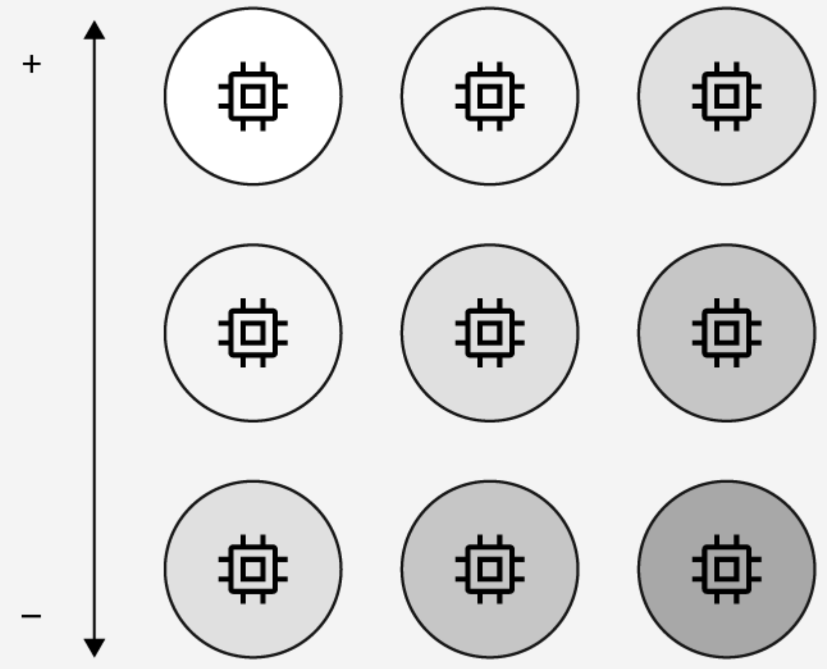
1 Share data between multiple analytics engines



2 Use fit-for-purpose compute and cache-optimized instances



3 Scale up and scale down automatically

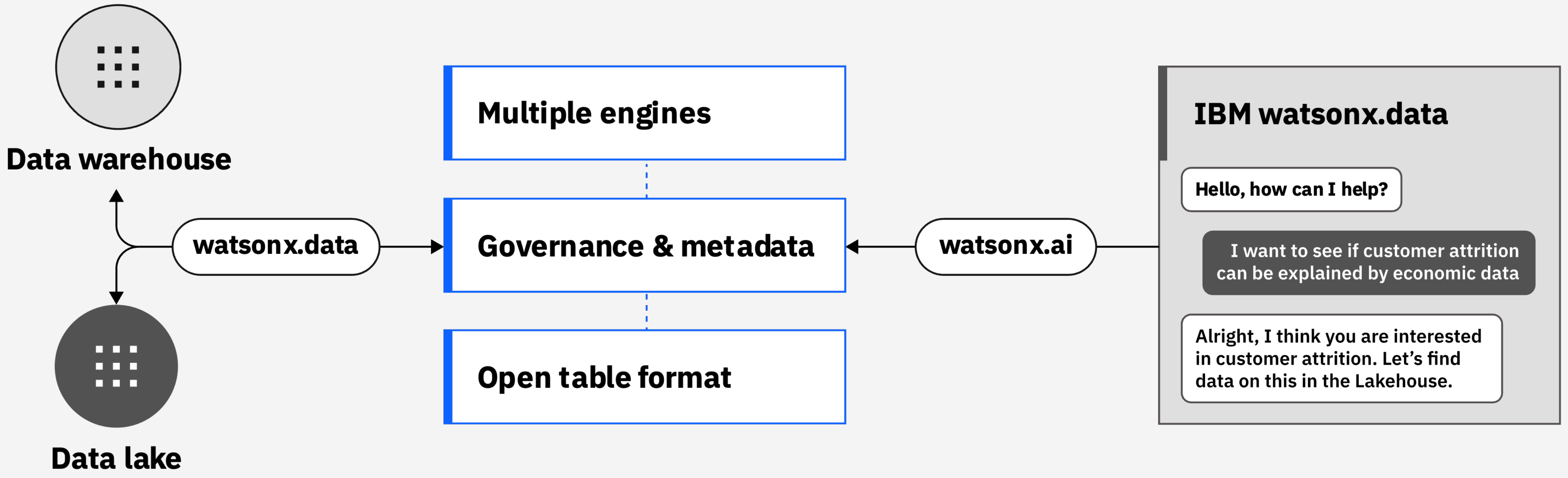


*When comparing published 2023 list prices normalized for VPC hours of IBM watsonx.data to several major cloud data warehouse vendors. Savings may vary depending on configurations, workloads and vendors.

Warehouse Augmentation

Accelerate time to trusted analytics and AI

Use foundation models to discover, augment, refine and visualize watsonx.data data and metadata



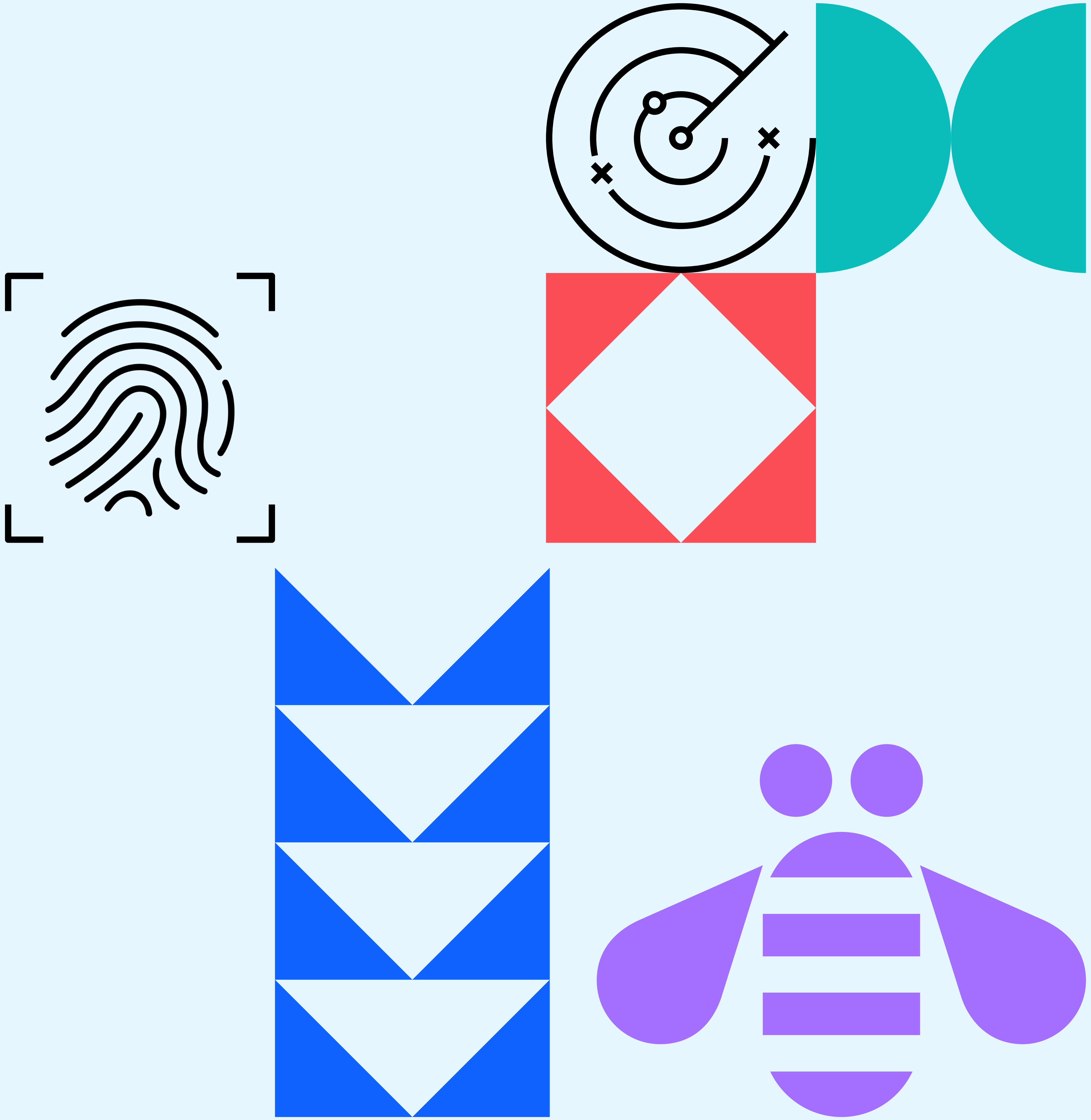
Hello, how can I help?

I want to see if customer attrition can be explained by economic data

Alright, I think you are interested in customer attrition. Let's find data on this in the Lakehouse.

Agenda

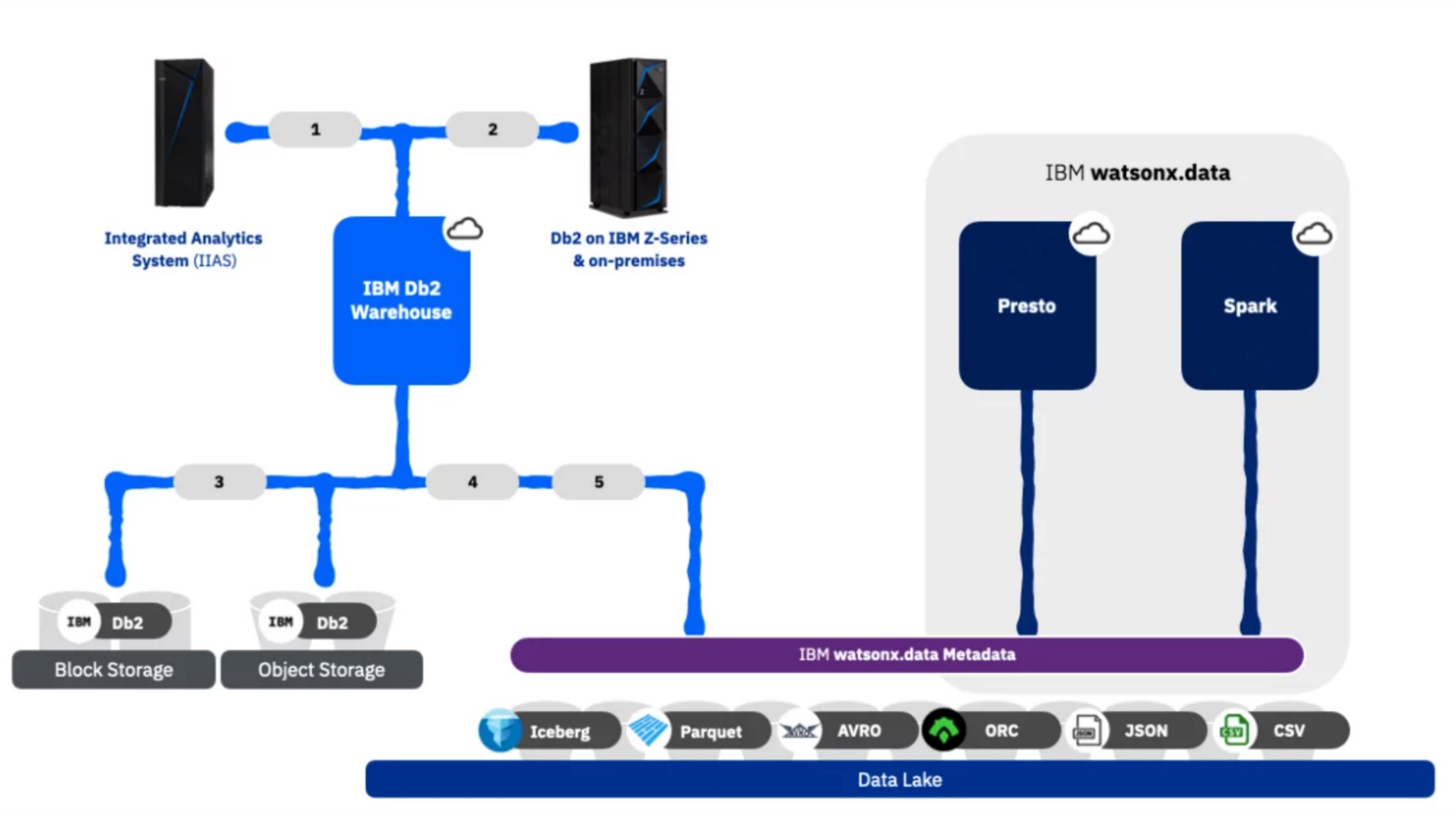
1. watsonx.data, an Introduction
2. watsonx.data Use Cases
3. Connecting watsonx.data to Db2



Watsonx.data and Db2

Sharing data & tables across the 2 systems.

Using the best tool for the workload at hand.



Connecting Db2 with watsonx.data

1. Set up a STORAGE ACCESS ALIAS to connect to the Object Storage service

```
 CALL SYSIBMADM.STORAGE_ACCESS_ALIAS.CATALOG(‘myalias’, S3',
‘s3.eu-south-2.amazonaws.com’, ‘****’, ‘****’, ‘mybucket’, ‘some/path',
‘R', ‘datalake-user-role')
```

2. Register the Watsonx.data metastore

```
 CALL REGISTER_EXT_METASTORE('watsonxdata',
'type=watsonx.data,uri=thrift://hmsauth1.fyre.ibm.com:9083', ?, ?)

 CALL SET_EXT_METASTORE_PROPERTY('watsonxdata', 'use.SSL', 'true',
?, ?)
```

3. You can now share tables between Db2 & watsonx.data
(See next slides)

Source: https://www.ibm.com/docs/en/db2-warehouse?topic=watsonxdata-configuring-db2-integration

Importing a Table from Watsonx.data

```
CALL EXTERNAL_CATALOG_SYNC('metastore-name', 'schema-name', 'table-name', 'exist-action', 'error-action', ‘options')
```

- Brings the table definition into the Db2 catalog. The data is shared between the 2 systems. Need to re-synch if the schema of the table changes.

- Multiple tables & schemas can be specified using regular expression.

- The *metastore-name* is the name used to register the metastore when setting up the connection.

- If a table is REPLACEd, it is dropped and re-created.
  - Working on improving that.

Exporting a Table to watsonx.data

– Regular tables

```
 CREATE DATALAKE TABLE hiveschema.db2exported(id int, name varchar(32))
STORED AS PARQUET LOCATION 'DB2REMOTE://hive-
bucket//hiveschema/db2exported' TBLPROPERTIES('bigsql.external.catalog’ =
'watsonxdata')
```

– Iceberg tables

```
 CREATE DATALAKE TABLE iceberg.db2exported(id INT, name VARCHAR(32))
STORED AS PARQUET STORED BY ICEBERG LOCATION 'DB2REMOTE://iceberg-
bucket//iceberg/db2exported' TBLPROPERTIES('iceberg.catalog' = 'watsonxdata')
```

– The table is created in both the Db2 & watsonx.data catalog and data is shared.

– The value of the property is the name used to register the metastore when setting up the connection.

A few gotchas

1. Db2 has a 20 mins (by default) data cache for DATALAKE tables.

   Force its refresh with the HCAT_CACHE_SYNC stored procedure when you insert data into a shared from watsonx.data.

2. Some INSERT statement may implicitly create new partitions. For shared tables, they will not be registered in the other system metastore.

   In Db2, run MSCK REPAIR TABLE on the table.
   In watsonx.data run system.sync_partition_metadata on the table.

3. Schema evolution for shared table is disabled in Db2 and must be done from the watsonx.data side.

watsonx.data Console

IBM **watsonx.data**

Infrastructure manager

Define and manage your Lakehouse components.

Search input text

Add component +

FS
SQL
Engines 4
engine_01
Presto
Devspark
Spark
Nzexam
IBM Netezza
Db2_01
IBM Db2
Catelogs 4
Hive_beta
Apache Hive
Iceberg
Apache Ice...
Db2_data
IBM Db2
Nzsource
IBM Netezza
Buckets 2
Objectibm
IBM Cloud...
Storages
Amazon S3
Databases 2
Db2_data
IBM Db2
Nzsource
IBM Netezza

A few links

Introducing the next generation of Db2 Warehouse on ibm.com

Better together: IBM watsonx.data and IBM Db2 on ibm.com

Accessing watsonx.data on IBM Db2 Warehouse Docs

Accelerating your Datalake tables with a Cache of Db2 Warehouse MQTs idug.org

Thank you

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.

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

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

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.