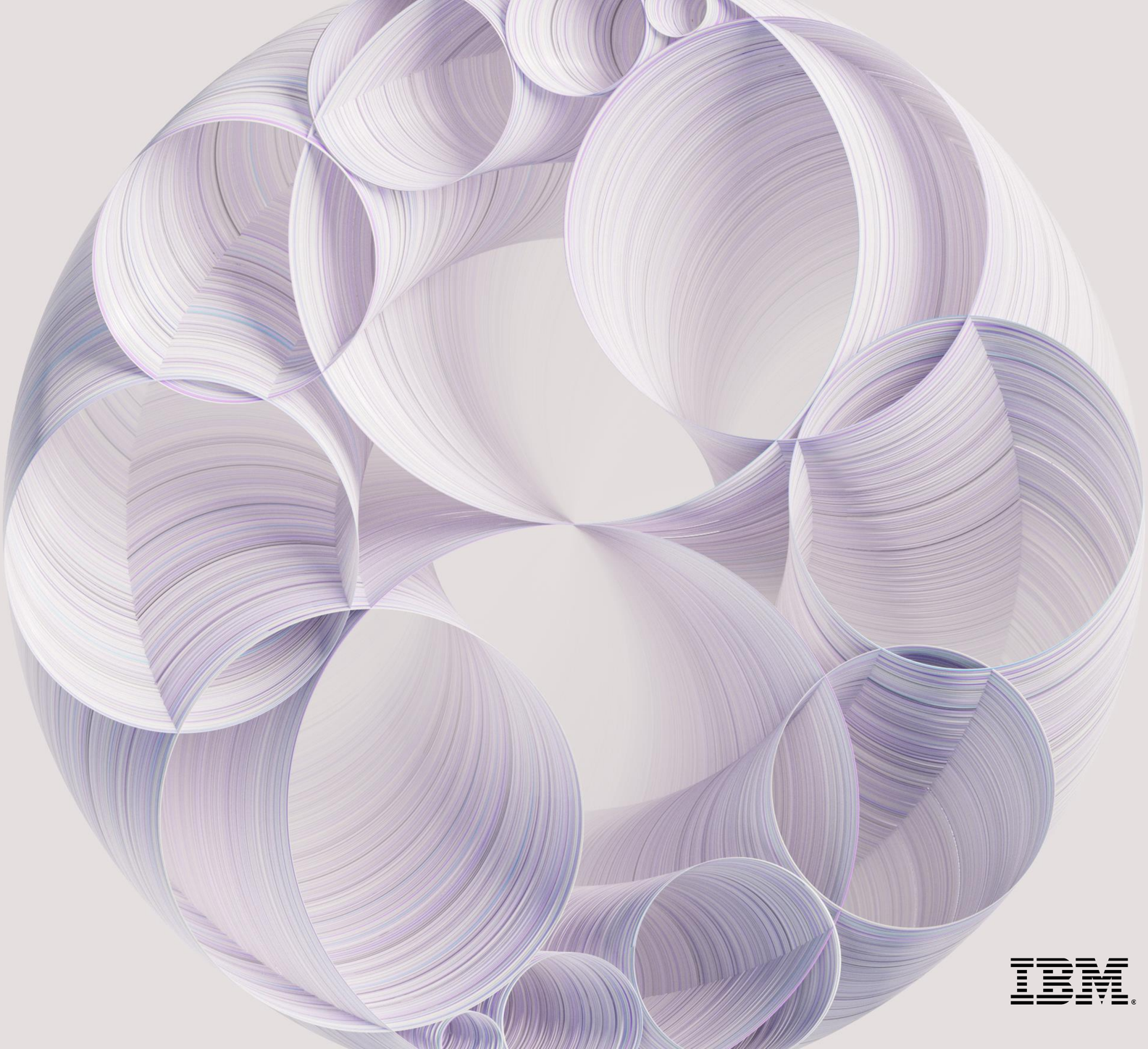
watsonx.data™

Scale AI workloads for all your data, anywhere

Data Server Day
Stockholm, Sweden
September 27, 2023

Kelly Schlamb
Data & AI, IBM
kschlamb@ca.ibm.com



IBM

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

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

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

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)

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

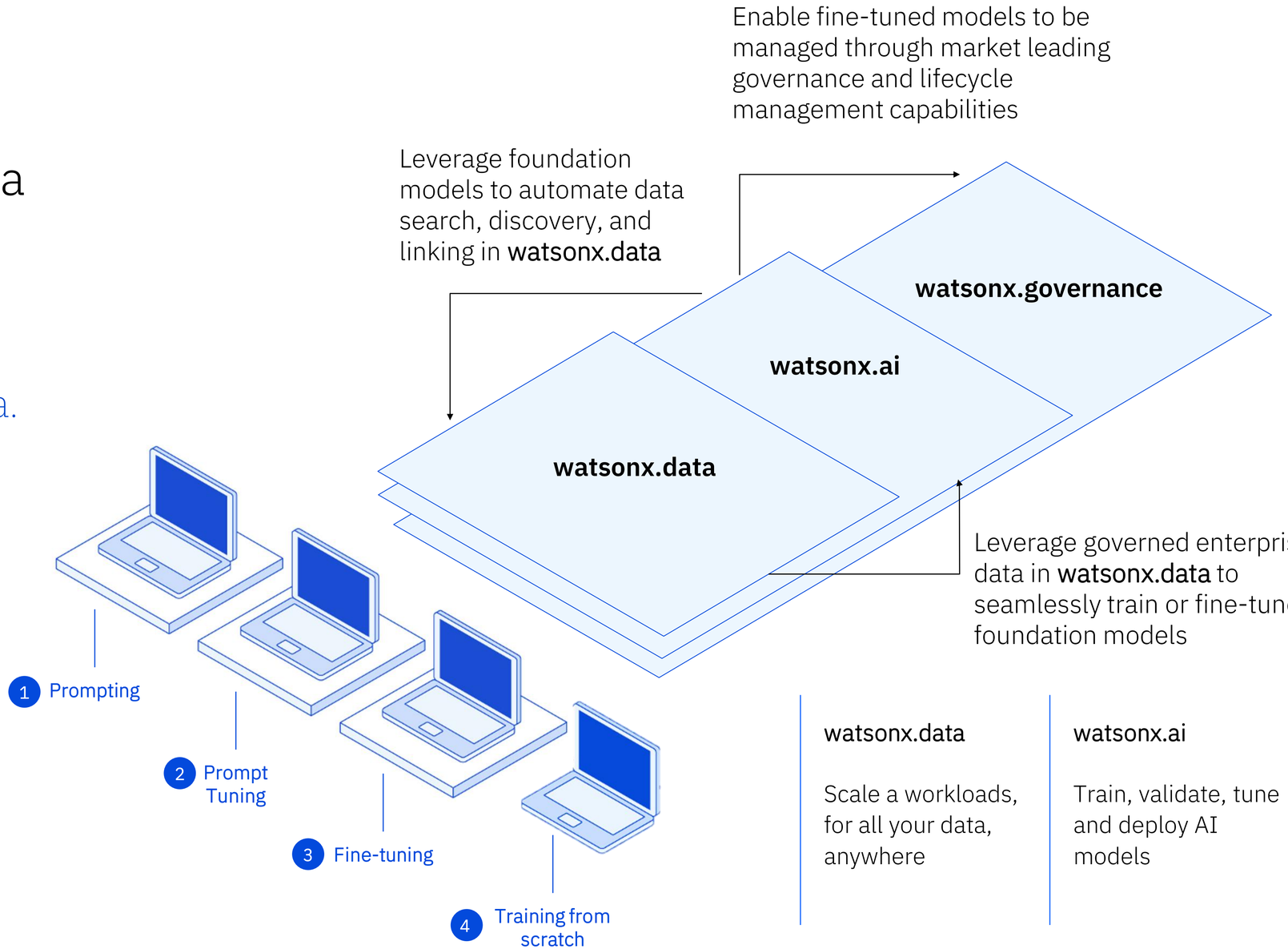
Introducing …

watsonx

watsonx

The platform for AI and data

Scale and accelerate the impact of AI with trusted data.



watsonx.data

Scale a workloads, for all your data, anywhere

watsonx.ai

Train, validate, tune and deploy AI models

watsonx.governance

Enable responsible, transparent and explainable AI workflows

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

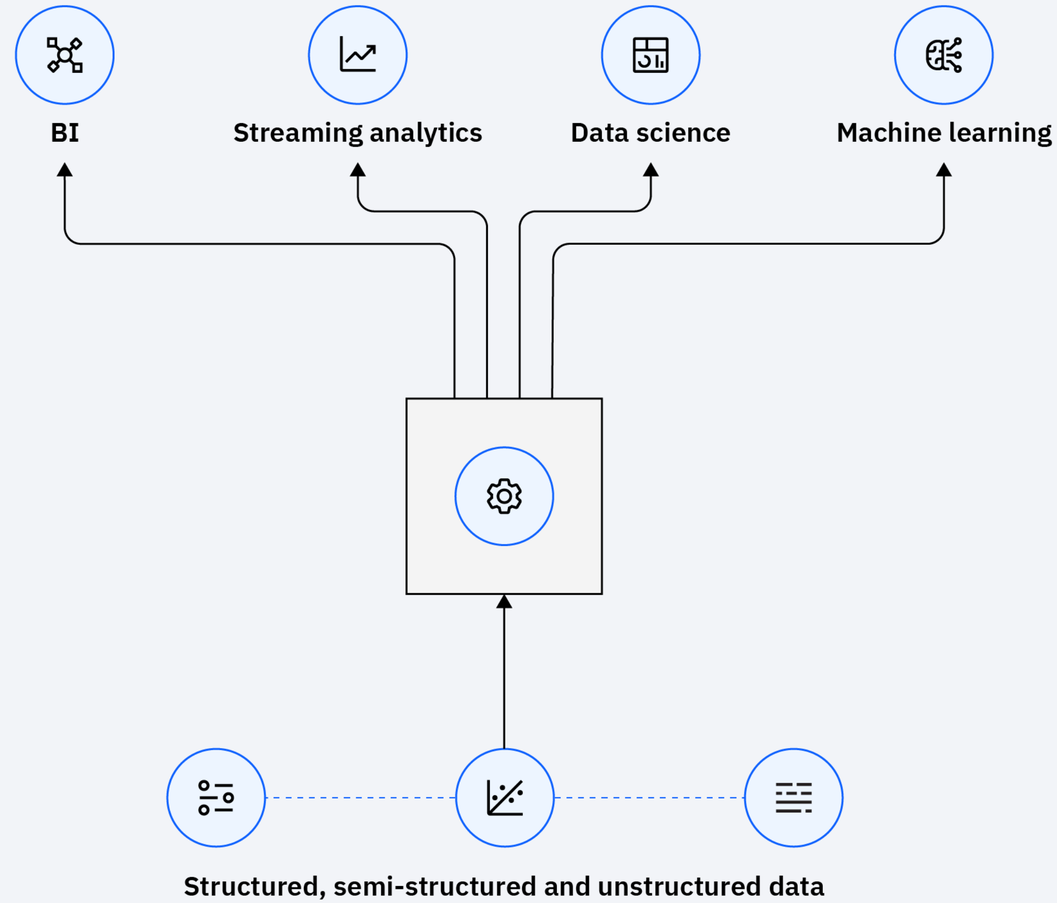
Accelerate responsible, transparent and explainable AI workflows

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

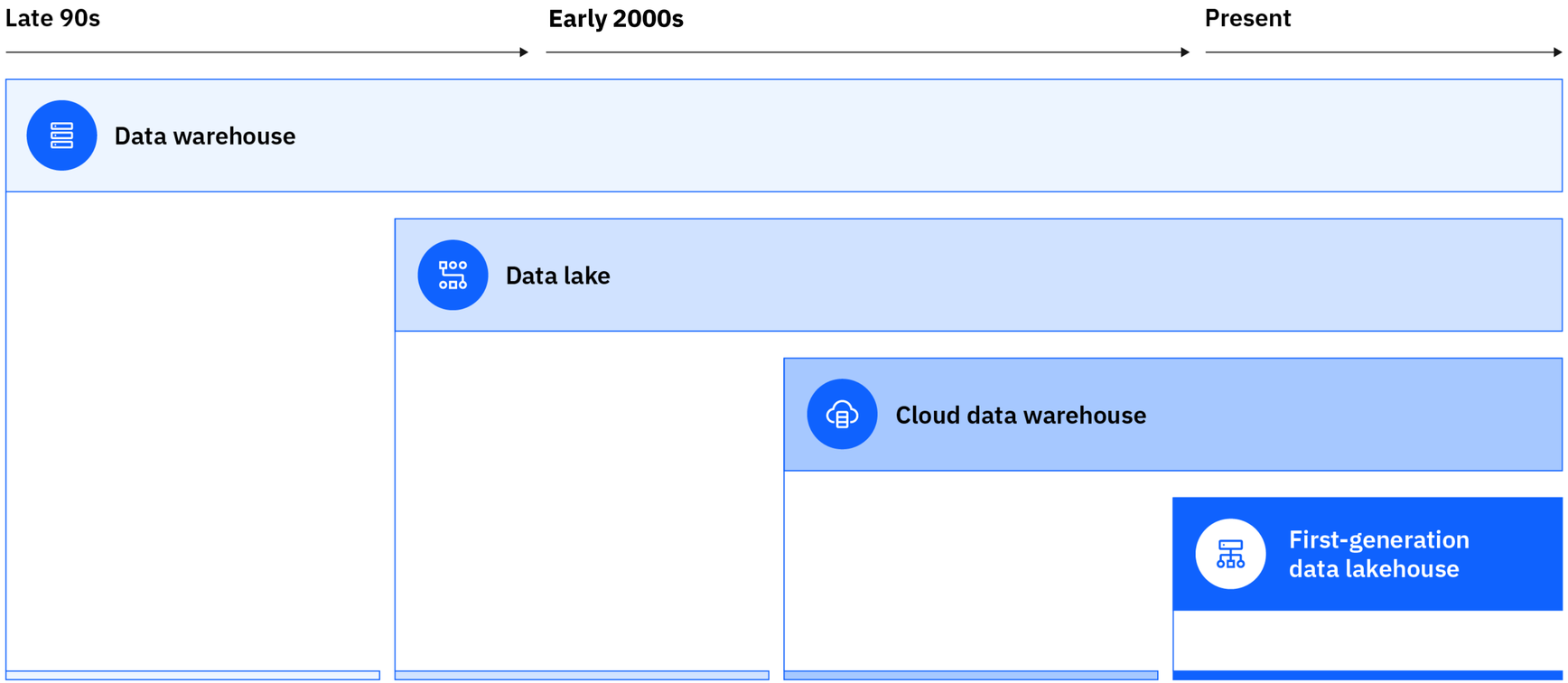
The data lakehouse

Data lakehouse (n):

A data lakehouse combines the high-performance characteristics of a data warehouse with the cost-efficiency, flexibility and scalability of a data lake to support highly complex data transformations and a wide variety of use cases



Traditional DM approaches have created more overall complexity and cost ... leading to the emergence of the data lakehouse

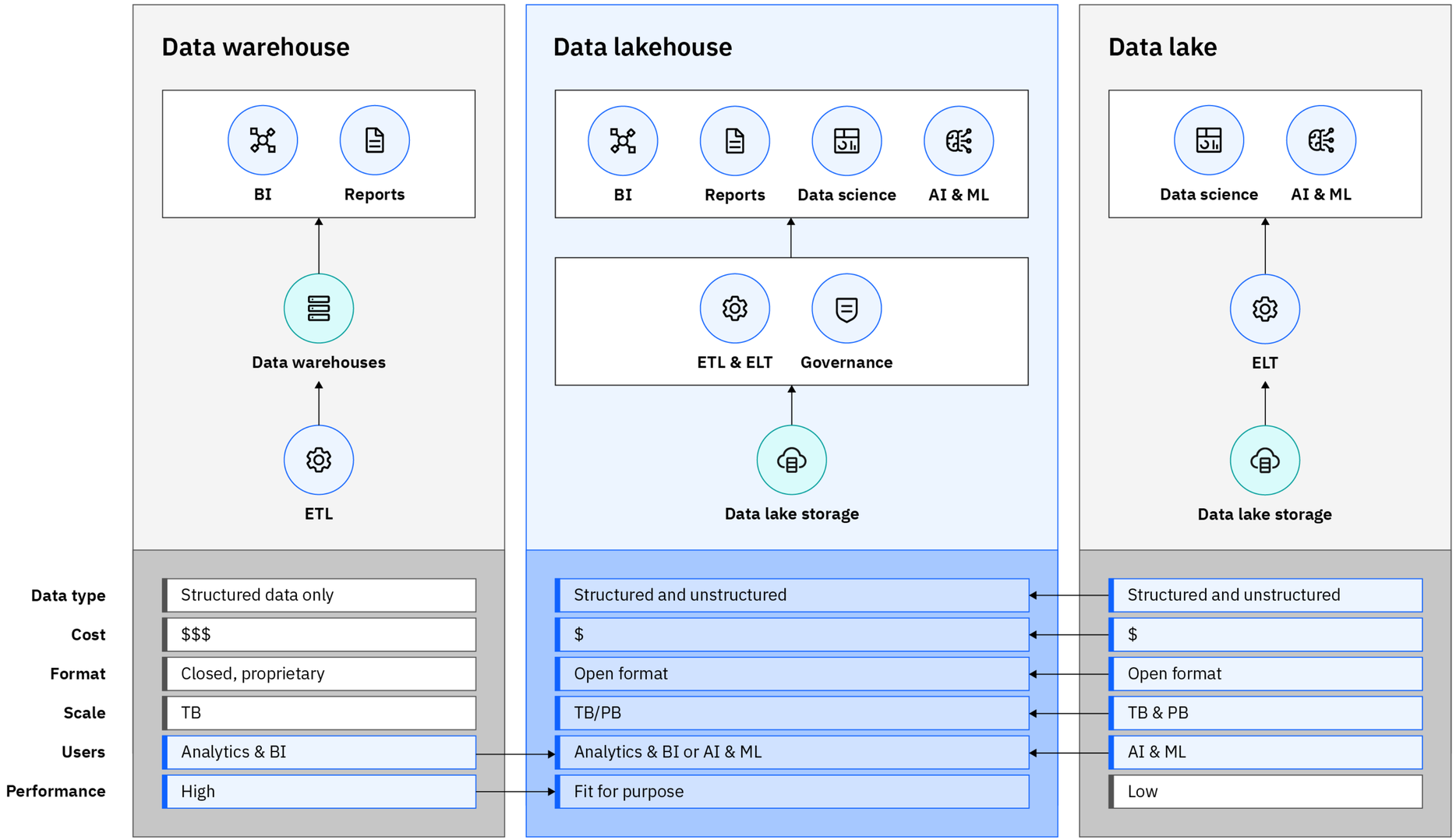


Today, leaders at most large enterprises manage their data and workloads using a mix of data repositories and data stores in hybrid environments.

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.

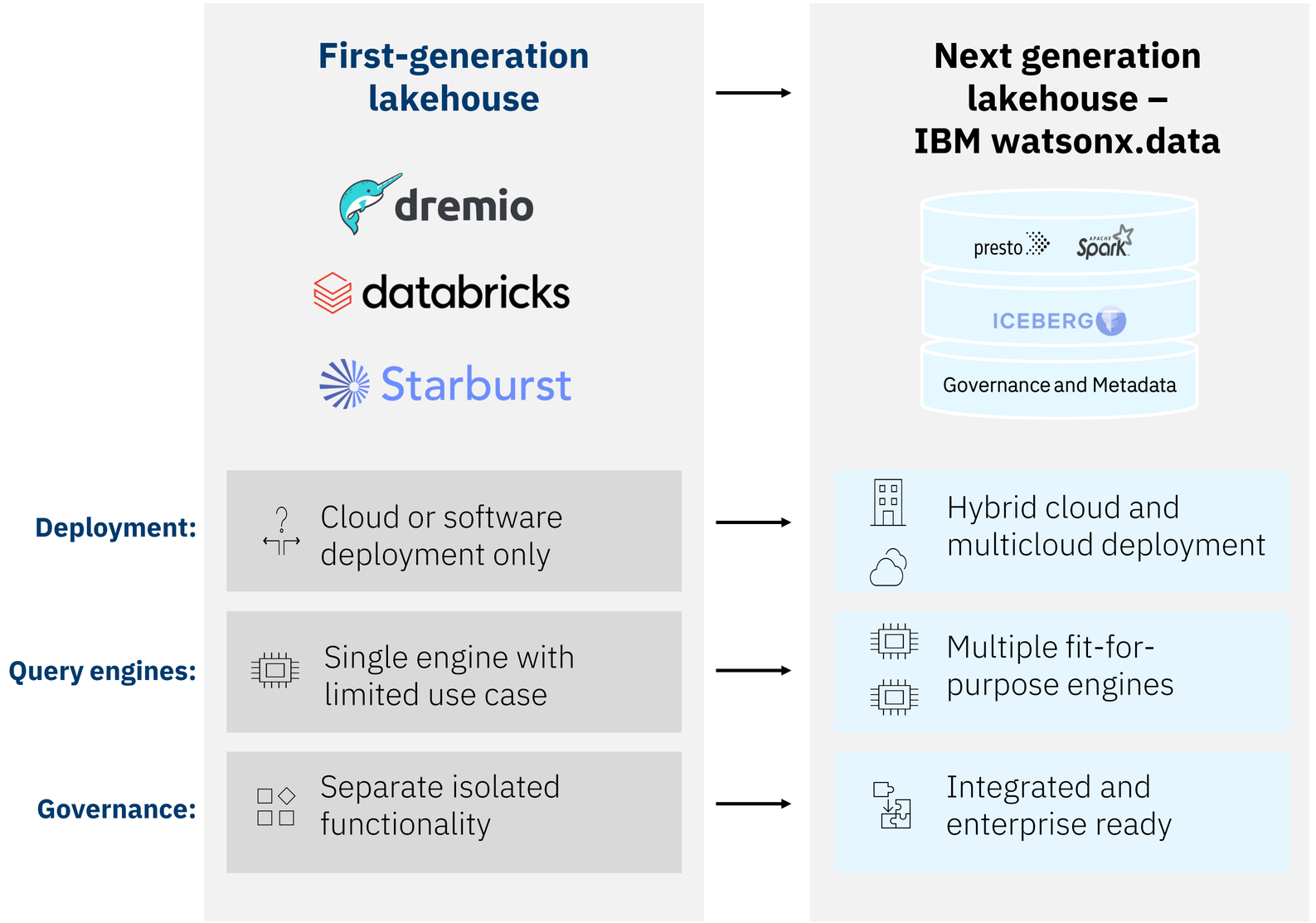
Lakehouses are meant to be a new class of data store that combines the best of data warehouses and data lakes



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

IBM watsonx.data is the next evolution of current first-generation lakehouses



IBM watsonx.data is the **only** lakehouse with **multiple open-source query engines** allowing clients to **optimize costs and performance by pairing the right workload with the right engine**

Run all workloads from **a single point of entry**

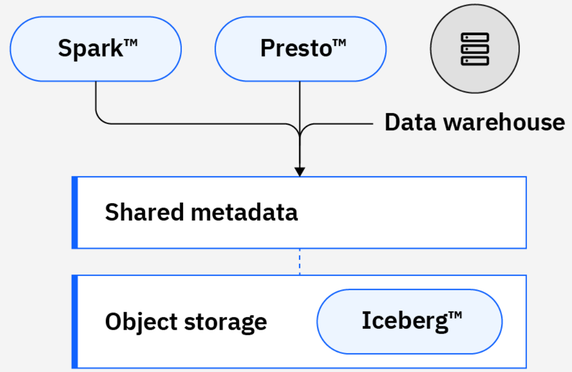
Deploy anywhere **with full support for hybrid cloud and multicloud** environments

Shared metadata across multiple engines eliminates the need to re-catalog, accelerating time to value while ensuring governance and eliminating costly implementation efforts

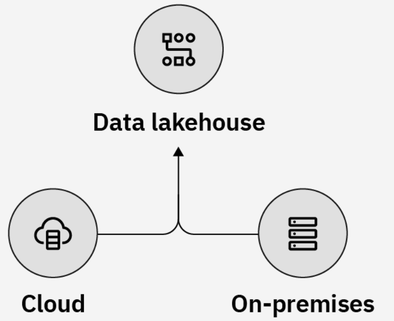
Access all your data across hybrid cloud through a single point of entry

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

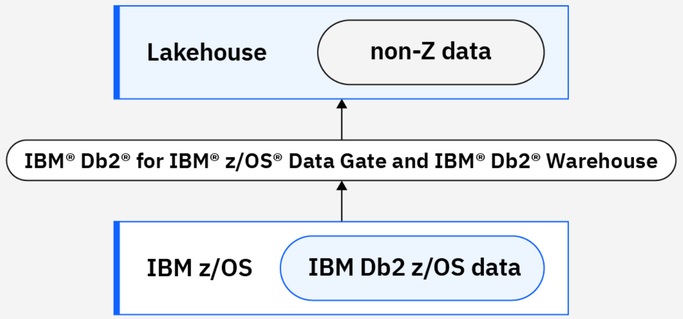
1 Share a single copy of data with tools that can read open data formats to minimize data duplication



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



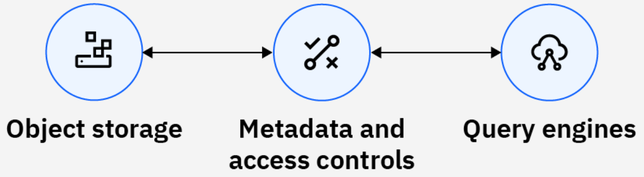
3 Synchronize and incorporate Db2 for z/OS data for lakehouse analytics



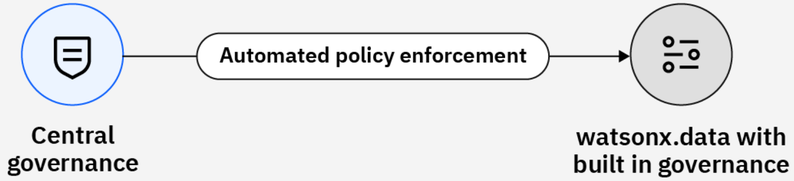
Get started in minutes with built-in governance, security and automation

Accelerate time to trusted analytics and AI

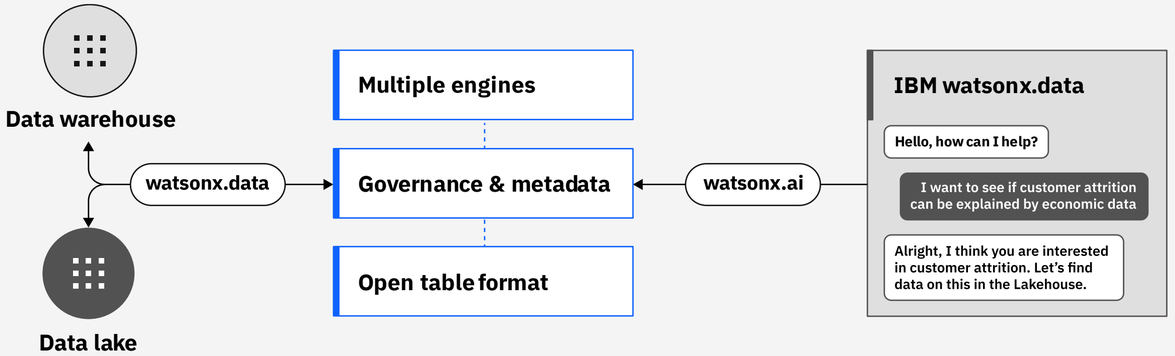
Connect to your existing analytics data and deploy fit-for-purpose query engines in minutes



Address enterprise compliance and security using built-in centralized governance across your data ecosystem



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

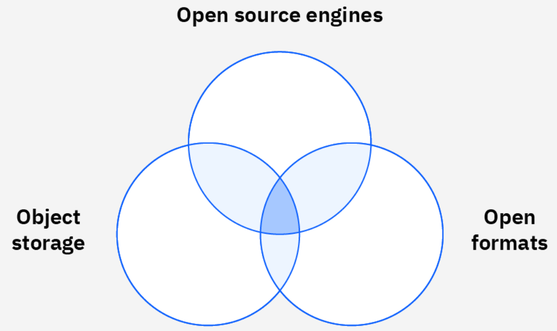


Reduce your data warehouse costs by up to 50%* by optimizing workloads

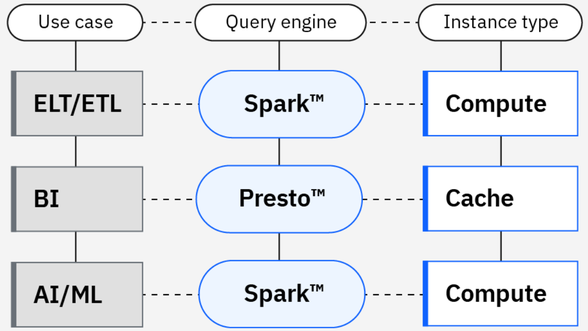
Optimize workloads from your data warehouse when you take advantage of low-cost object storage and fit-for-purpose query engines

*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.

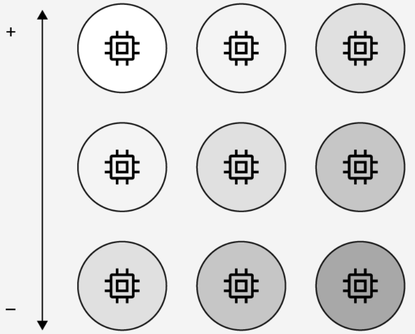
1 Share data between multiple analytics engines



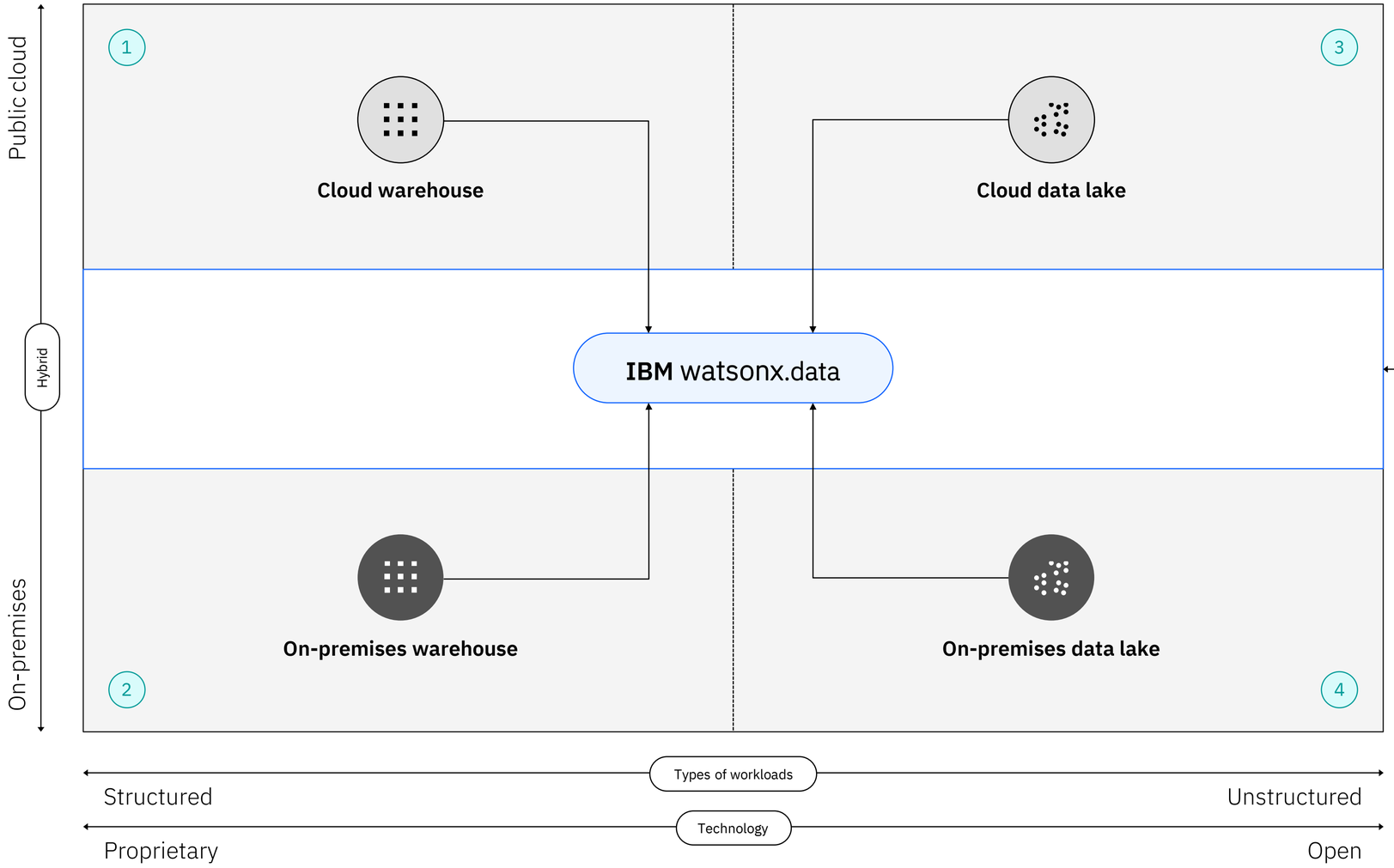
2 Use fit-for-purpose compute and cache-optimized instances



3 Scale up and scale down automatically



Access all your data quickly and optimize your data architecture with multi-engine support and hybrid deployment of analytics and AI workloads



1

Optimize costly cloud warehouses
Make the most of fit-for-purpose query engines and compute resources

2

Optimize & access on-premises warehouses
Use low-cost object storage and fit-for-purpose engines

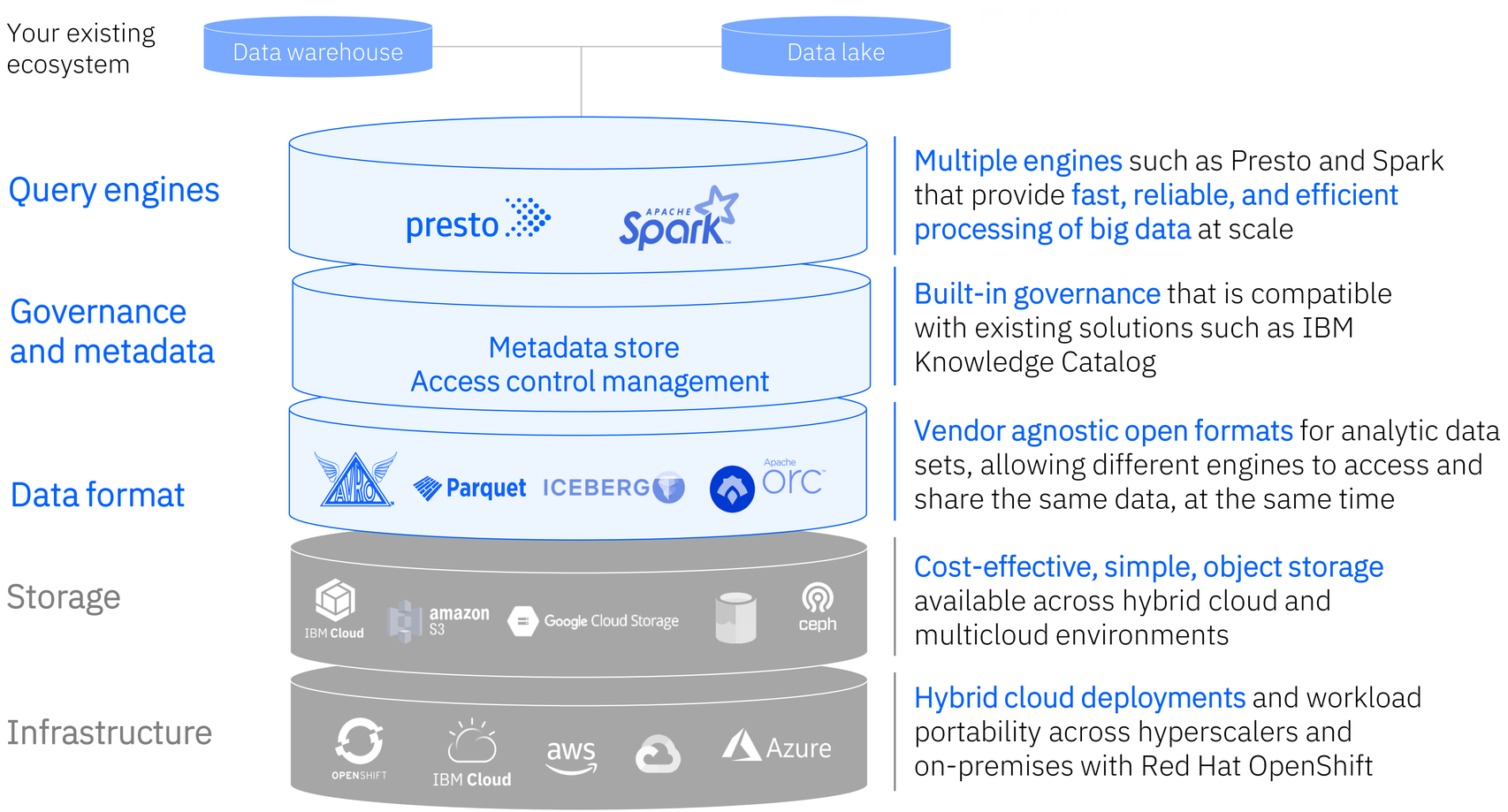
3 4

Modernize data lakes
Run existing reporting and enable new AI workloads without the cost and complexity of Hadoop

1 2 3 4

Deploy across hybrid cloud and multicloud
Seamlessly deploy to both the public cloud and to your existing on-premises investment

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



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 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

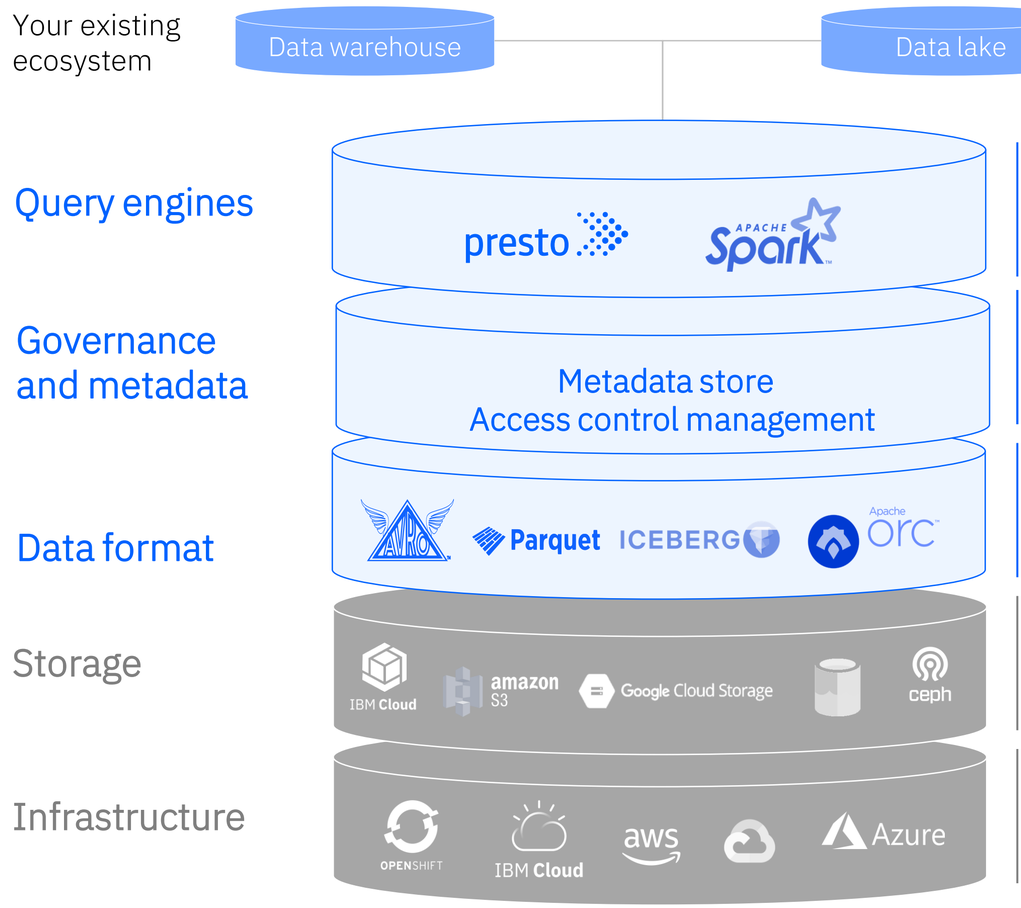
Hybrid cloud deployments and workload portability across hyperscalers and on-premises with Red Hat OpenShift

Core watsonx.data functionality

Ecosystem infrastructure

- Optimize workload costs and performance using multi-engine functionality
- Strengthen 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

Key components of 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 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-premises with Red Hat OpenShift

Completely open.
No lock-in!

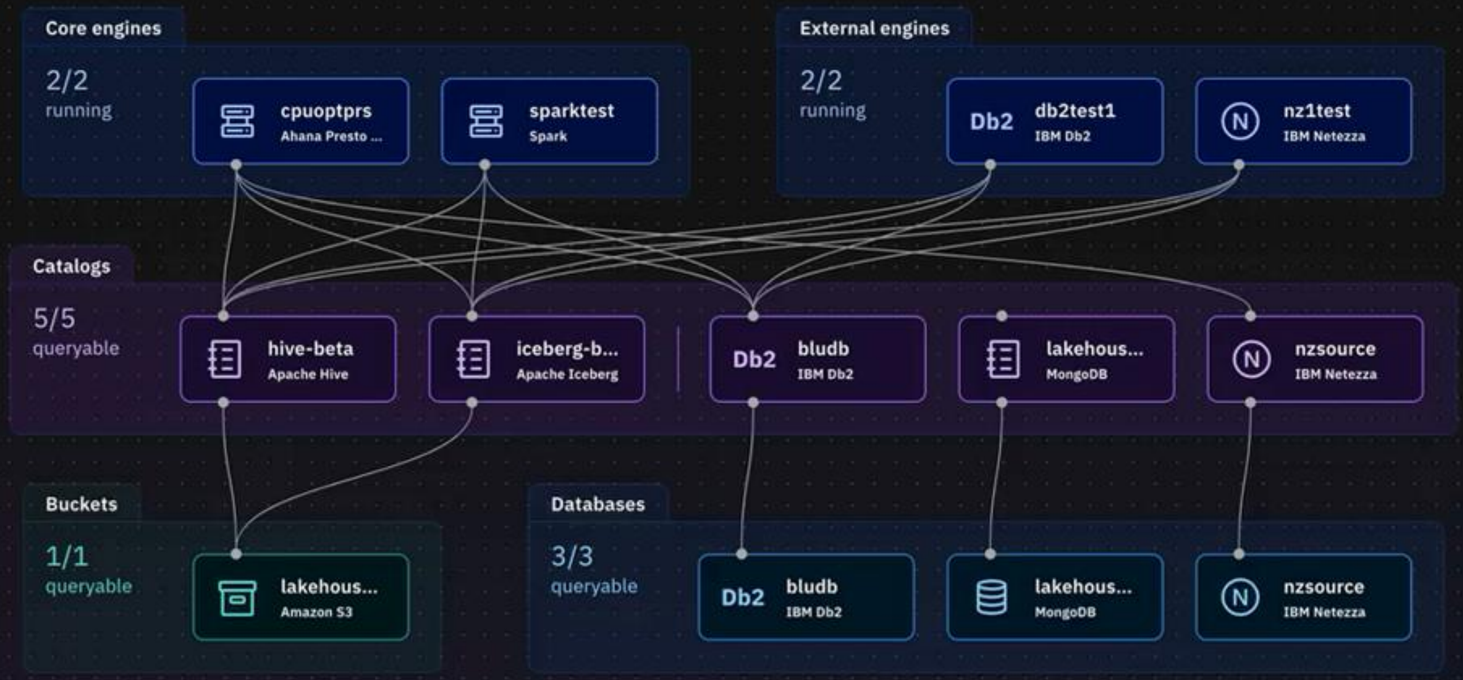
Built on a foundation of industry-embraced open-source technologies.

Infrastructure manager

Define and associate your infrastructure components.

Search your system

Add component



Data explorer

Browse your schemas and tables by engine.

Engine
cpuoptprs

Search for loaded tables

Create

Catalogs associated

- bludb
  - "dp_target" 4
  - audit
  - db2ins1
  - db2inst1
  - defect_test
  - dp_bvt_trgt
  - dp_pen_source
  - dp_pen_target
  - dp_source
  - dp_target
  - gosales
  - gosalesct
  - gosalesdw
  - gosaleshr
  - gosalesmr
  - gosalesrt
  - i60uajfj
  - ibm_rtmon
  - ibmconsole
  - information_schema
  - nullid
  - nv_target

bludb ▸ "dp_target" ▸

"customer"

Table schema | Data sample

Search for columns

Add column +

Name	Data type	Comment
store_id	integer	
active	integer	
first_name	varchar(22)	
email	varchar(80)	
address_id	integer	
last_name	varchar(24)	
create_date	timestamp	
customer_id	integer	
activebool	smallint	
last_update	timestamp	

Items per page: 25 | 1–10 of 10 items | 1 of 1 page

Query workspace

Build and run queries against your data.

Data objects | Saved queries

Engine
cpuoptprs

Search for loaded tables

Catalogs associated

- bludb
  - "dp_target" 4
  - audit
  - db2ins1
  - db2inst1
  - defect_test
  - dp_bvt_trgt
  - dp_pen_source
  - dp_pen_target
  - dp_source
  - dp_target
  - gosales
  - gosalesct
  - gosalesdw
  - gosaleshr
  - gosalesmr
  - gosalesrt
  - i60uajfj
  - ibm_rtmon
  - ibmconsole
  - information_schema

Untitled 1 •

Explain | Run on cpuoptprs

```
1 SELECT *
2 FROM "iceberg-beta"."default"."order_detail" AS details
3 LEFT JOIN "bludb"."gosales"."order_header" AS header
4 ON details.order_number=header.order_number
5 LIMIT 10;
```

No results yet.

Build and run a query in the text editor above to view its results here.

IBM watsonx.data v0.9.0/nightly

Hey Daniel 👋😊, how can I help you?

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

Alright, I think you're interested in customer churn. Let's see if we find data on this topic in the lakehouse

Show me the details of table CSTINSIGHT.INDIVIDUAL_CST_CHURN_FACT.

Alright, here you can take a look at data in this table.

Enter text

HELP

Send

< Back to Semantic Search

ID
CSTINSIGHT.INDIVIDUAL_CST_CHURN_FACT

Name
Individual Cst Churn Fact

Columns
19

Rows
1,175

Package
Isw pack master package

Schema
Cstinsight

Entity
Individual customer churn fact

Timestamp
04/04/2023 18:44:25

Source
Native

Open Semantic Expansion

Send to IBM Cognos Analytics

Add to collection

Individual customer churn fact

Show Semantic Enrichment

Description
The number of customers that churned from one service provider to another.

Primary Key: CDR_DT_ID, IDV_CST_PRFL_ID, CST_RSDNC_AREA_ID, AGRM_ID

Tags: cost of acquisition, customer retention, customer acquisition

Schema | Data | Quality

Search...

Column	Description	Concept	Concept description
PPN_TMS	The time the data was collected	Population Timestamp	The date and time at which the population was measured.
AGRM_ID	Agreement id	Arrangement Identifier	Identifies the numbers or codes by which an Arrangement may be uniquely identified. These identifiers may be industry recognized identifiers or internally generated identifiers. Each Involved Party to the Arrangement may have its own Arrangement Identification; for example, a Clearing Arrangement between two financial institutions may be assigned an ID of #3456 by Bank A and #6543 by Bank B.
PD_ID	Product id	Product Identifier	Specifies the unique identification that is assigned to the Product. The identifier may be an industry recognized identifier or an internally generated identifier.
NUM_OF_LOST_CST	The number of customers that were lost during the period of time	Number Of Lost Customers	A measure that identifies the number of lost Involved Parties or parties that terminated arrangements within a given period of time with whom the Financial Institution had Product Arrangements.
NUM_OF_NEW_CST	The number of new customers	Number Of New Customers	A measure that identifies the number of new Involved Parties with whom the Financial Institution has Product Arrangements within a given period of time.
NUM_OF_CST_RTNG	The number of customers that returned to the service provider	Number Of Returning Customers	A measure that identifies the number of returning customers having the same arrangement as before, with the Financial Institution.

Let’s take a closer look at some of these open-source concepts/technologies...

What is object storage?

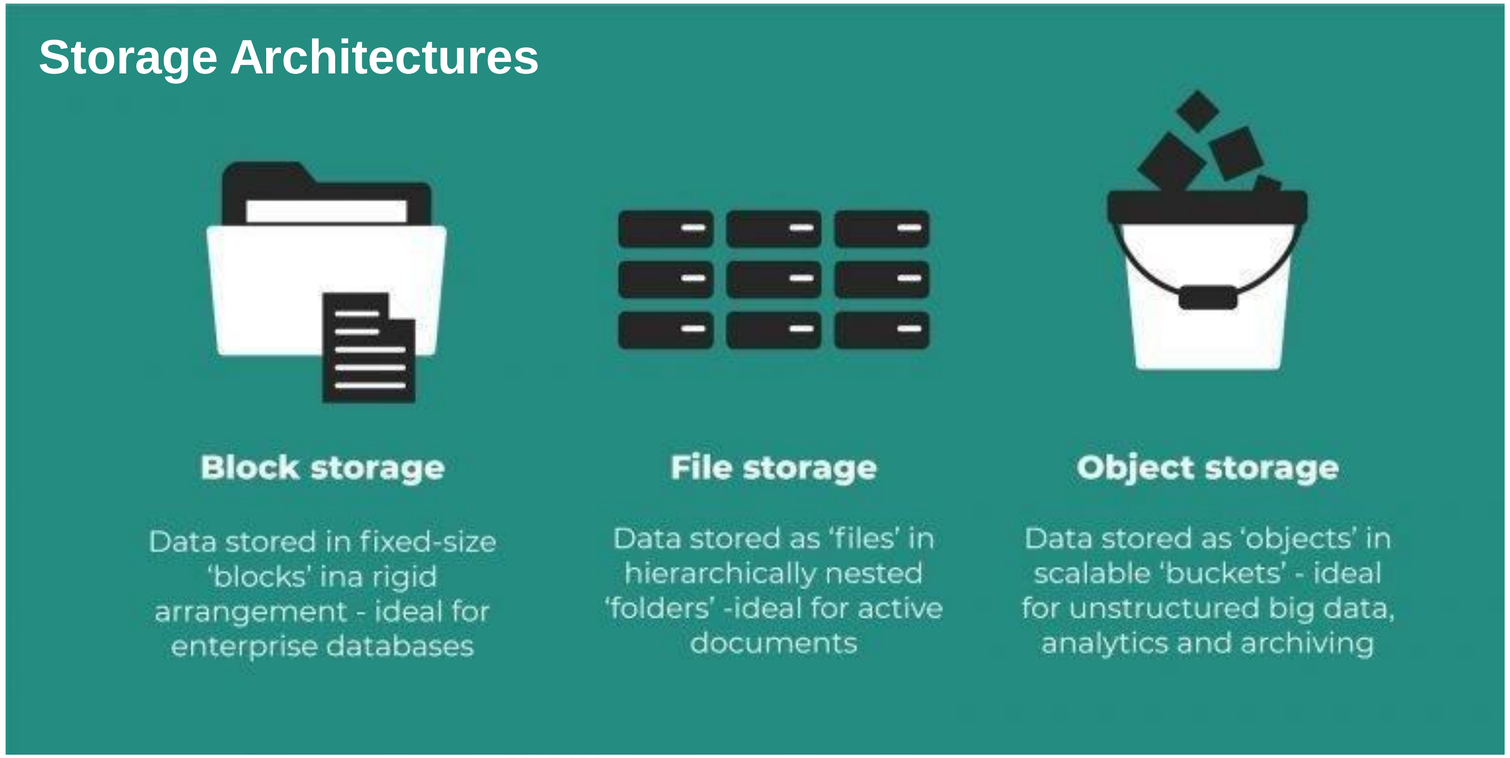


Image source: https://www.openpr.com/news/2367430/global-object-storage-market-market-revenue-market-growth

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

amazon S3 · IBM Spectrum Fusion · ceph · DELL EMC ISILON · RedHat OpenShift · IBM Cloud Object Storage · NetApp

Common open data file formats

Commonly used in data lakes and lakehouses

CSV

- Human-readable text
- Each row corresponds to a single data record
- Each record consists of one or more fields, delimited by commas

{ JSON }

- Human-readable text
- Open file and data interchange format
- Consists of attribute-value pairs and arrays
- JSON = JavaScript Object Notation

Parquet

- Open-source
- Binary columnar storage
- Designed for efficient data storage and fast retrieval
- Highly compressible
- Self-describing

Apache orc

- Open-source
- Binary columnar storage
- Designed and optimized for Hive data
- Self-describing
- Similar in concept to Parquet

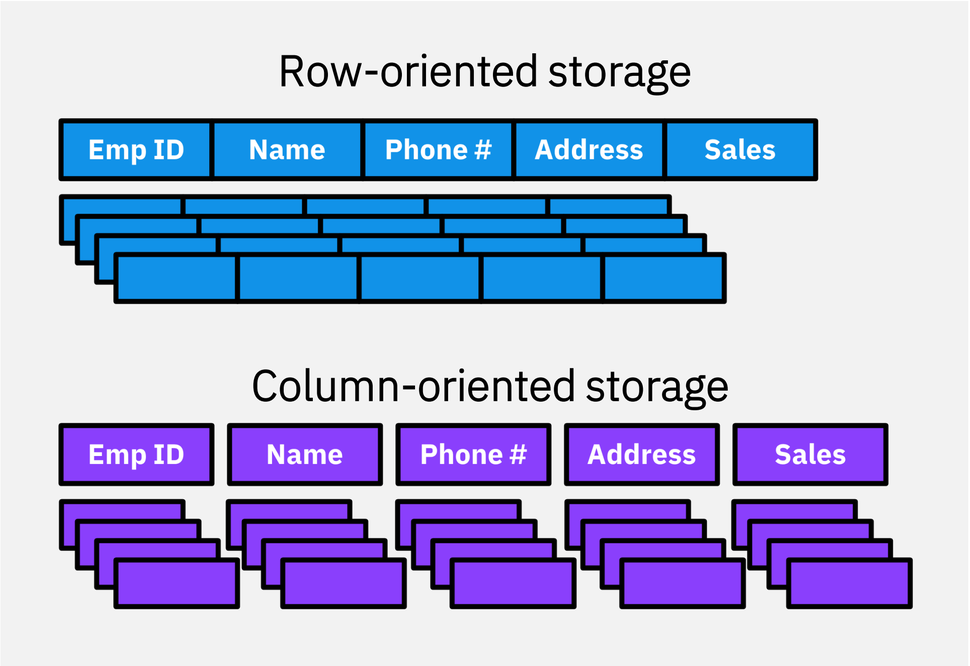
AVRO

- Open-source
- Row-oriented data format and serialization framework
- Robust support for schema evolution
- Mix of text/binary

Apache Parquet

Parquet is an open file format designed to support fast data processing for complex data

- Open-source
- Columnar storage
- Highly compressible
- Self-describing
- Schema evolution



Why do these things matter in a lakehouse?

- Performance of queries directly impacted by size and amount of file(s) being read
- Ability to read/write data to an open format from multiple runtime engines enables collaboration
- Size of data stored, amount of data scanned, and amount of data transported affect the charges incurred in using a lakehouse (depending on the pricing model)

Table management and formats

Sits “above” the data file layer

Organizes and manages table metadata and data

Typically supports multiple underlying disk file formats (Parquet, Avro, ORC, etc.)

Offers database-like capabilities, like ACID transactions

- Open-source
- Designed for large, petabyte-scale tables
- ACID-compliant
- Capabilities include schema evolution, partition evolution, and table version rollback – all without re-writing data
- Advanced data filtering
- Time-travel queries

- Open-source
- Manages the storage of large datasets on HDFS and cloud object storage
- Includes ACID transactions, upserts/ deletes, advanced indexes, streaming ingestion services, concurrency, data clustering, and asynchronous compaction
- Multiple query options: snapshot, incremental, and read-optimized

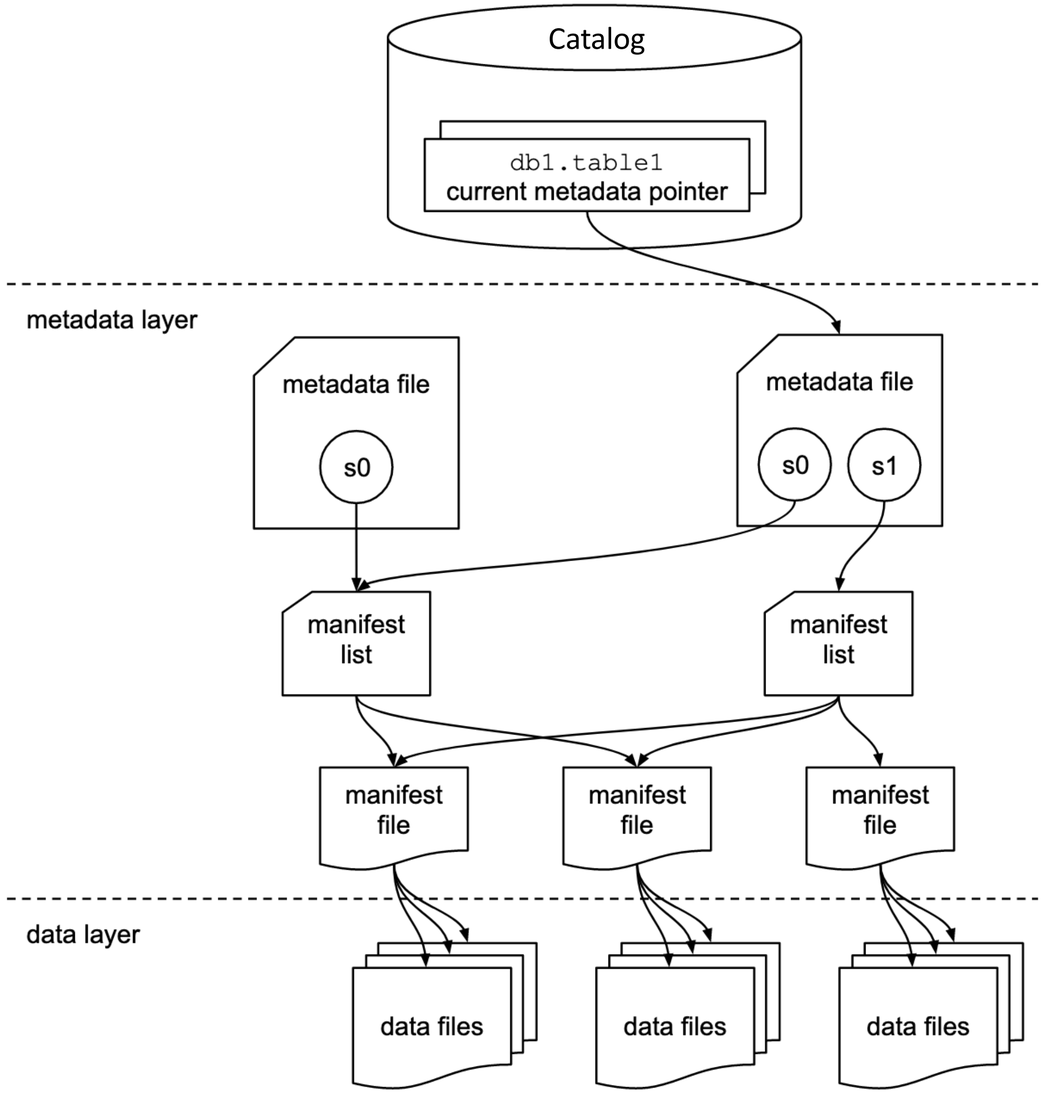
- Open-source, but Databricks is primary contributor and controls all commits to the project – so effectively “closed”
- Foundation for storing data in the Databricks Lakehouse Platform
- Capabilities include indexing, data skipping, compression, caching, and time-travel queries
- Designed to handle batch as well as streaming data

Why Apache Iceberg for data lakehouses?

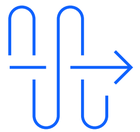
Open-source data table format that helps simplify data processing on large dataset stored in data lakes

People love it because it has:

- SQL — Use it to build the data lake and perform most operations without learning a new language
- Data Consistency — ACID compliance (not just append data operations to tables)
- Schema Evolution — Add/remove columns without distributing underlying table structure
- Data Versioning — Time travel support that lets you analyze data changes between update and deletes
- Cross Platform Support — Supports variety of storage systems and query engines (Spark, Presto, Hive, +++)



ACID transactions



ACID refers to a set of properties of database transactions intended to **guarantee data validity** despite errors, power failures, and other mishaps

Atomicity

Guarantees that each transaction is a single event that either succeeds or fails completely; there is no half-way state.

Consistency

Ensures that data is in a consistent state when a transaction starts and when it ends, guaranteeing that data is accurate and reliable.

Isolation

Allows multiple transactions to occur at the same time without interfering with each other, ensuring that each transaction executes independently.

Durability

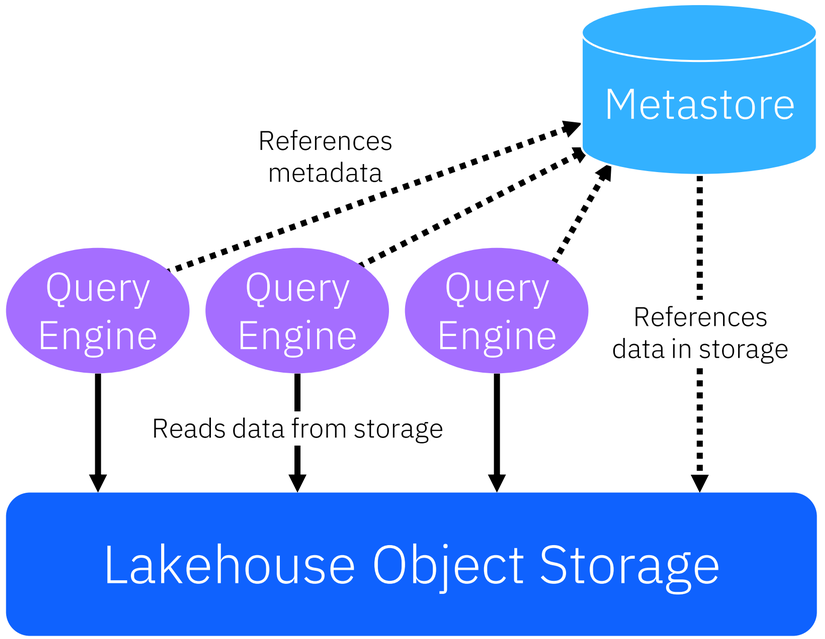
Means that data is not lost or corrupted once a transaction is submitted. Data can be recovered in the event of a system failure, such as a power outage.

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, data, and metadata consistently

Hive Metastore (HMS)

- Component of Hive (but can run standalone)
- In watsonx.data, IBM KC integrates with HMS for policy-based access and governance
- Query engines use the metadata in HMS to optimize query execution plans



Presto

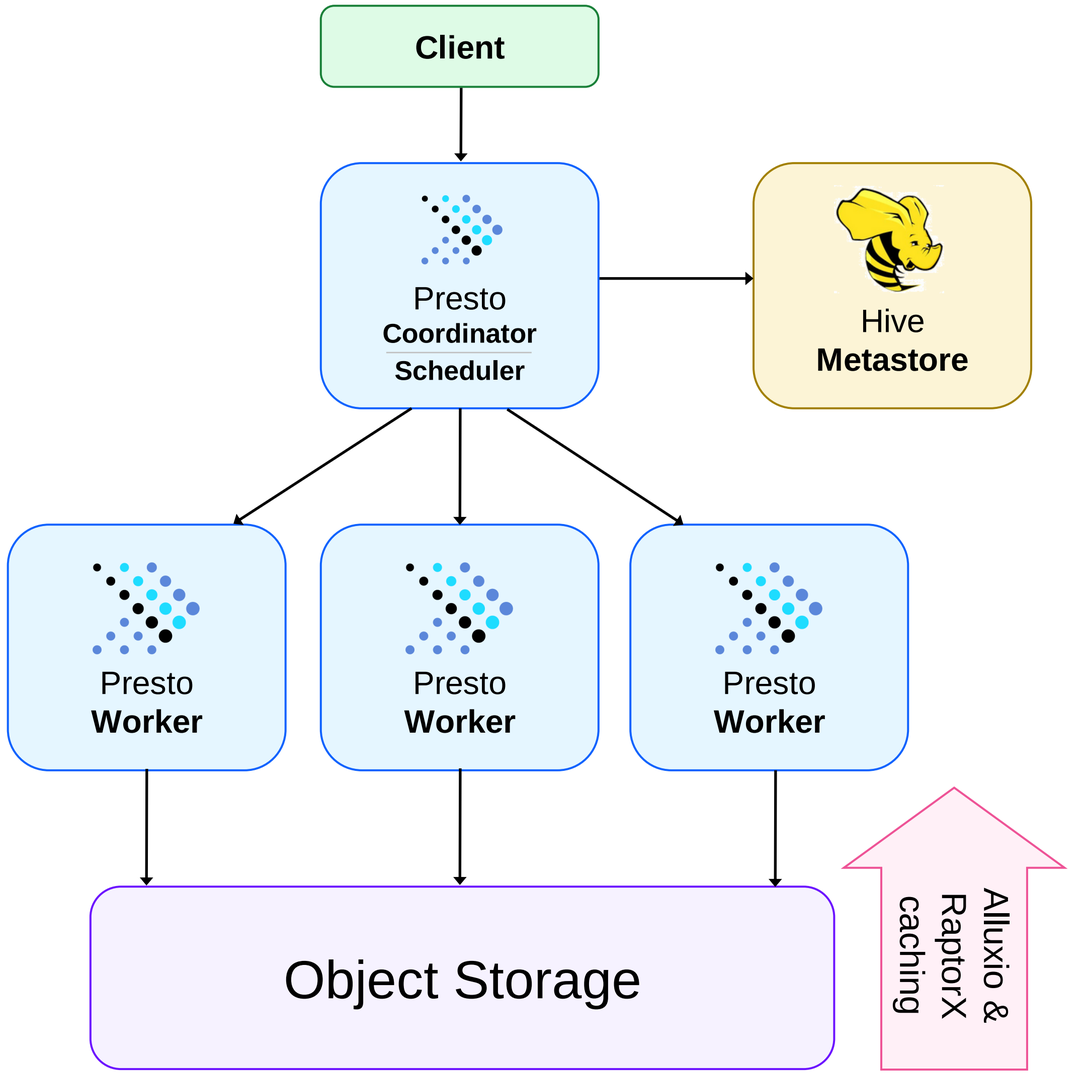
- Presto is an open-source distributed SQL engine suitable for querying large amounts of data
- Supports both relational and non-relational sources
- Easy to use with data analytics and business intelligence tools
- Supports both interactive and batch workloads
- In watsonx.data, spin up one or more 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:

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:** 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).

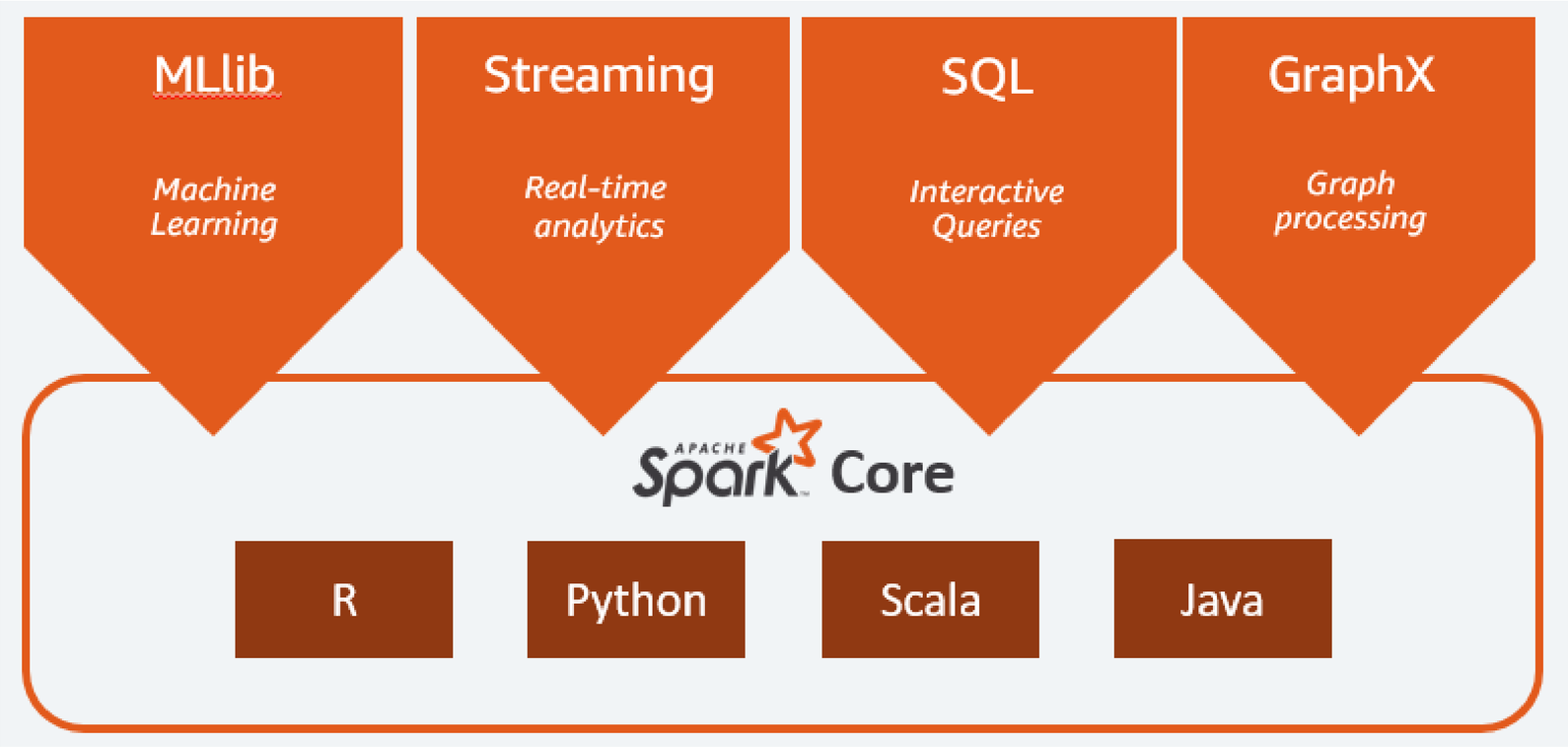


Apache Spark

Apache Spark is an open-source data-processing engine for large data sets. It is designed to deliver the computational speed, scalability, and programmability required for *big data*, specifically for streaming data, graph data, ML, and AI applications.

Spark has libraries that extend the capabilities to ML, AI, and stream processing.

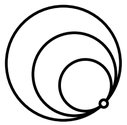
- Apache Spark MLlib
- Spark Streaming
- Spark SQL
- Spark GraphX



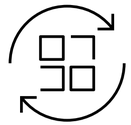
The IBM approach to a data lakehouse architecture combines the best of IBM with the best of open source

Best-in-class cost and performance optimizations for compute and storage

Built-in integrations with IBM data repositories and data fabric

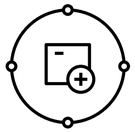


Deep expertise and capabilities in data and storage



Open and vendor-agnostic across architectural tiers

Enables hybrid, multicloud deployments with the Red Hat OpenShift platform



The best of open source

presto

Apache Spark

AVRO

HIVE

ICEBERG

Parquet

Apache orc

Use cases

Deploy AI/ML at scale

Build, train, tune, deploy, and monitor trusted AI and ML models for mission-critical workloads with data in IBM watsonx.data; strengthen compliance with lineage and reproducibility of data used for AI.

Apply real-time analytics and BI

Combine data from existing sources with new data in watsonx.data to unlock new, faster insights without the cost and complexity of duplicating and moving data across different environments.

Streamline data engineering

Reduce data pipelines, simplify data transformation, and enrich data for consumption using SQL, Python or an AI-infused conversational interface.

Share data responsibly

Enable self-service access for more users to more data while you strengthen security and compliance with centralized governance and local automated policy enforcement.

Powered by

presto

Adobe

Digital advertising platform

2K+ daily reports and
100s of pipelines on a
7 PB data lake with
400B+ records

Bolt

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

100K daily queries with
2K active internal users on a
2 PB data lake

Meta

Social media

30K queries per day with
1K daily active users on a
300 PB data lake

Uber

Ride-hailing, food delivery

100M+ queries per day with
7K weekly active users on a
50 PB data lake

ByteDance

Internet technology

2M+ queries per day for
business intelligence and
one-off use cases

twilio

Communications
API technology

2700 active internal users,
running 1M queries,
scanning 40 PB data per month

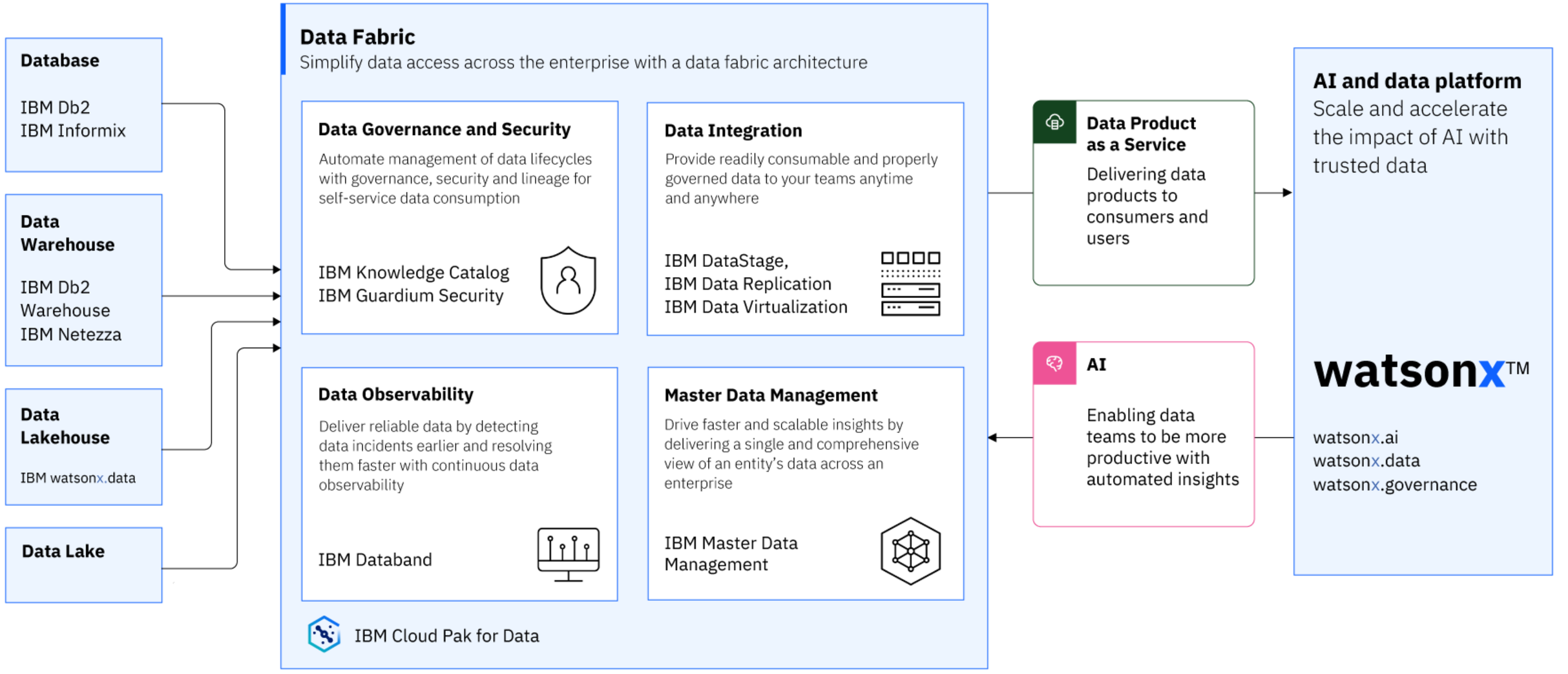
watsonx.data

True hybrid cloud

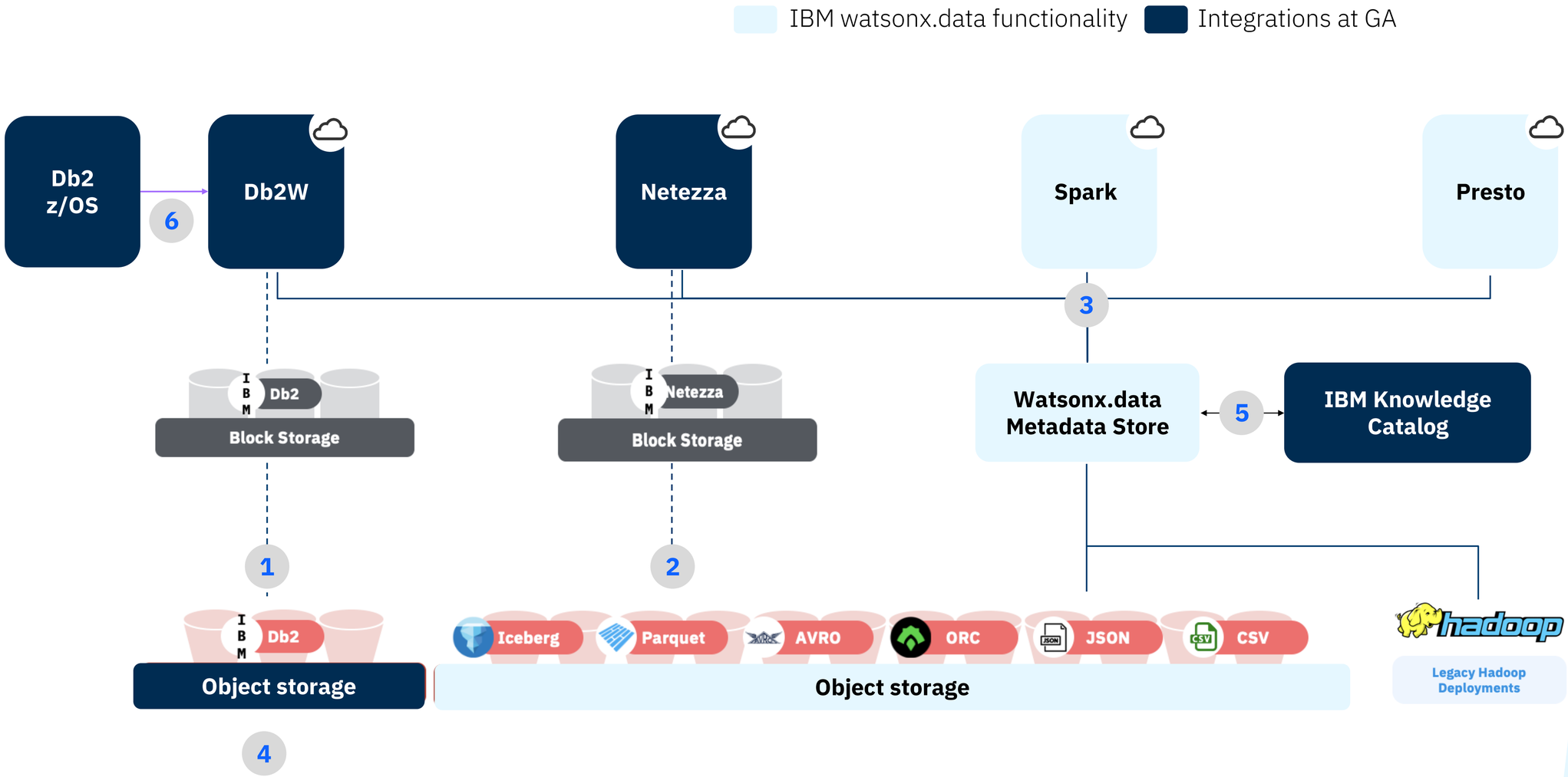
Deployment options:

- Software-as-a-Service (IBM Cloud, Amazon Web Services)
- Cloud Pak for Data cartridge
- Standalone on Red Hat OpenShift Container Platform

Investments in a trusted data foundation will accelerate and scale AI

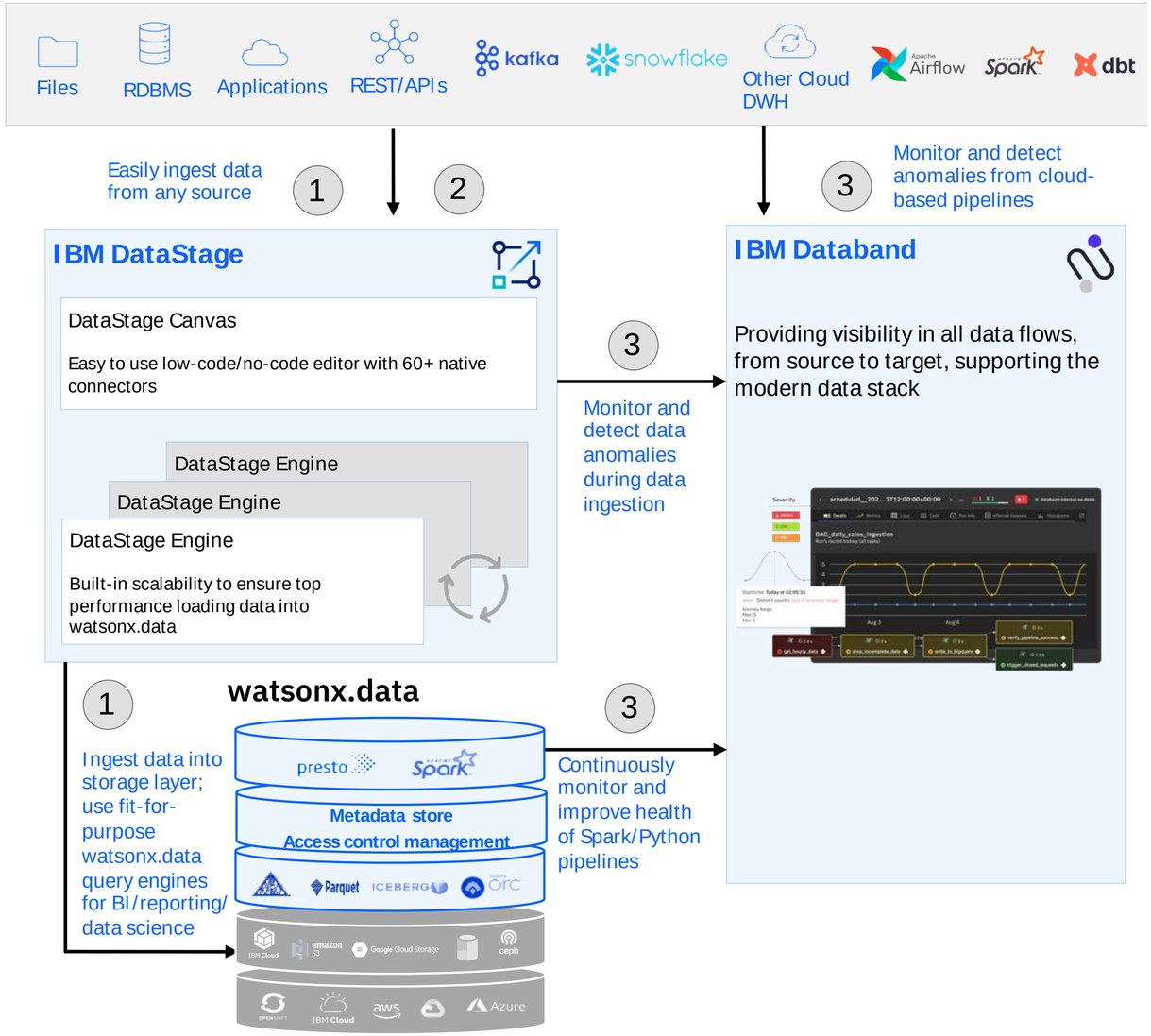


The integrated IBM watsonx.data ecosystem for maximum workload coverage and optimal price-performance



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. **KC policies enforced by the lakehouse** via metadata service
6. Analyze Z data **easily and securely** with Db2 for z/OS Data Gate

Effortlessly populate with trusted data leveraging best-in-class data ingestion and observability



watsonx.data + IBM DataStage

Easily build EL(T) pipelines with an intuitive visual design

1 **Ingest data from any source**
Leverage 60+ native connectors to ingest data into watsonx.data from any type of source, ensuring top performance with built-in engine scalability

2 **Reduce cost by offloading data from cloud data warehouses**
Offload data from cloud data warehouses to enable shifting workloads like BI, reporting, or data science to fit-for-purpose query engines

watsonx.data + IBM Databand

Continuously detect and resolve data quality incidents

3 **Monitor, detect, and resolve data quality incidents**
Monitor and improve the health of DataStage, Spark, or Python pipeline workloads running on watsonx.data; detect data anomalies and accelerate issue resolution

What IBM offers

Why IBM?

Open	IBM’s AI is based on the best open technologies available
Trusted	IBM’s AI is transparent, responsible, and governed
Targeted	IBM’s AI is designed for enterprise and targeted at business domains
Empowering	IBM’s AI is for value creators, not just users

watsonx.data

Scale AI workloads, for all your data, anywhere

A fit-for-purpose data store, based on an open lakehouse architecture, supported by querying, governance and open data formats to access and share data

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

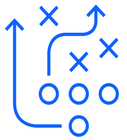
Get started in minutes with built-in governance, security and automation.

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

*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.

Three ways to **get started with watsonx.data today**

IBM’s investment in partnering with clients



Free trial

Experience watsonx.data and test out core capabilities with a free trial.

[Try our free trial]

Client briefing

Discussion and custom demonstration of IBM’s generative AI watsonx point-of-view and capabilities. Understand how watsonx.data can be leveraged in any businesses AI strategy.

2-4 hours

Pilot program

Watsonx pilot developed with IBM AI engineers. Prove watsonx.data value for the selected use case(s) with a plan for adoption.

1-4 weeks

IBM watsonx.data free trial

Get started for free with

IBM watsonx.data

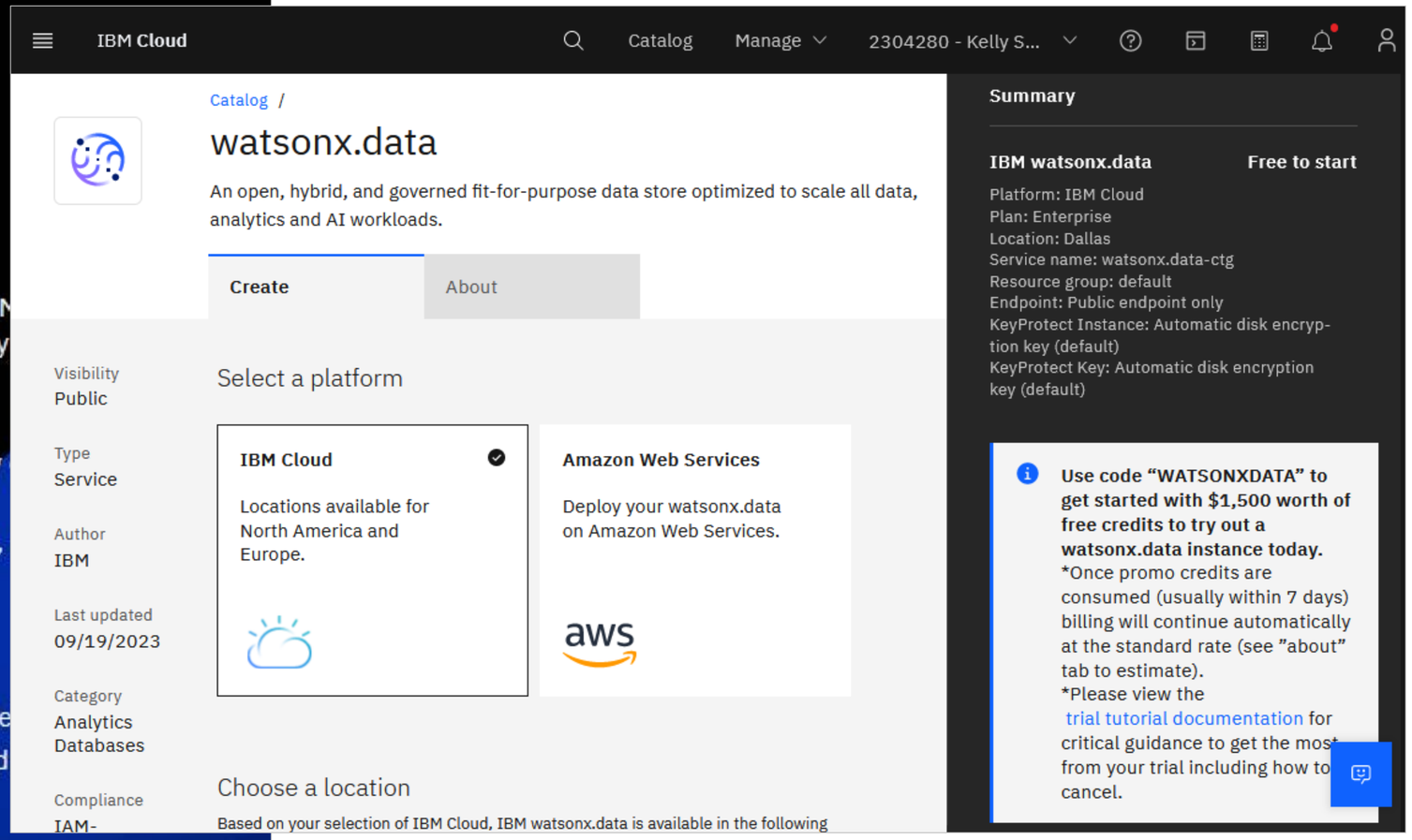
Get $1,500 to test drive our features

In our free to start trial, you'll receive $1,500 to test drive an IBM open data lakehouse instance on IBM Cloud. You'll be able to try capabilities, including:

- Elastic scaling and pause of multiple query engines
- Support for open data and table formats to share a single copy engines
- A simple, integrated console with built in governance, security, automation to get started in minutes.

To start your free trial:

1. Log in or create an IBM Cloud account using the form on the le
2. Go to the watsonx.data lakehouse catalog page (https: //cloud /lakehouse) choose IBM infrastructure.
3. Apply the promo code WATSONXDATA.
4. Click Create.



[Free trial]