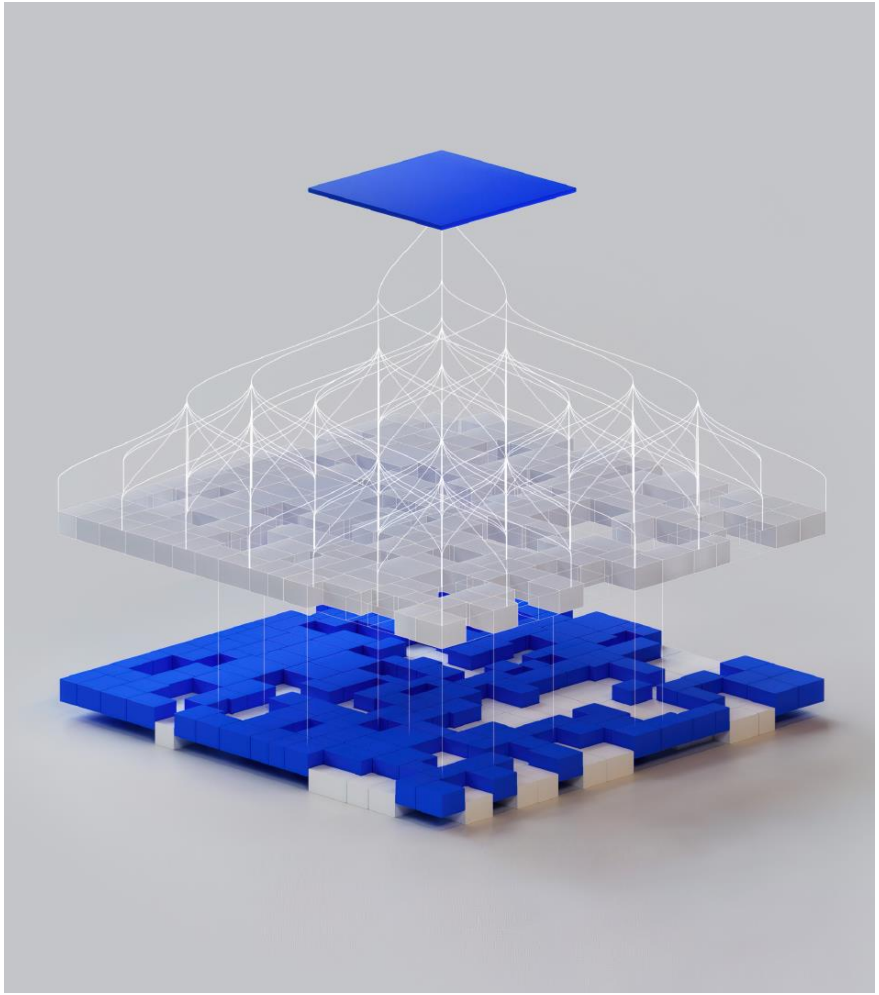
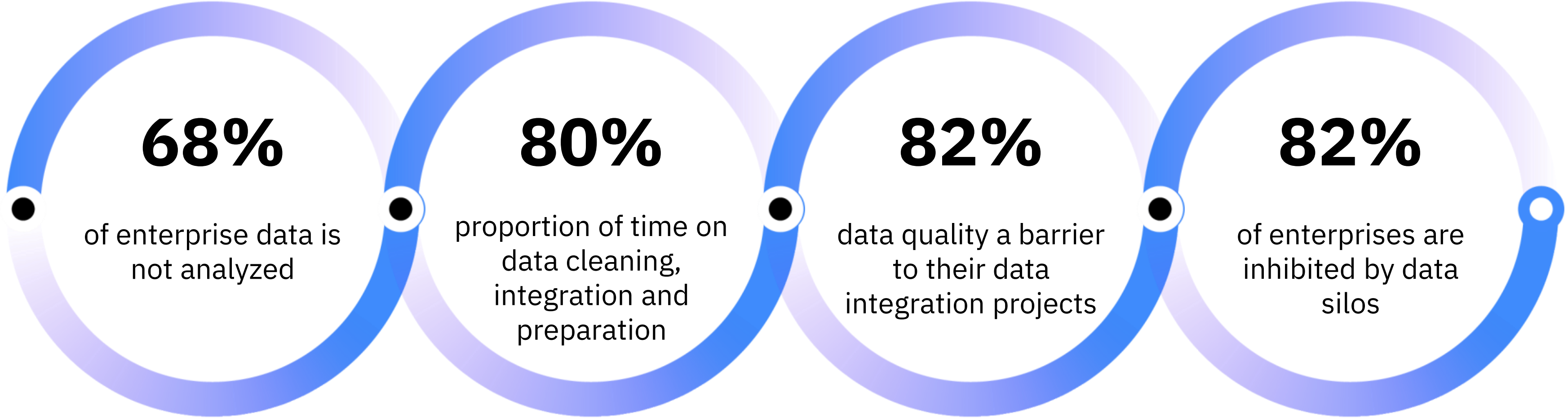
IBM Open Data Lakehouse



Nguyen Tuan Khang, khangnt@vn.ibm.com

Country Manager, IBM Software

Data complexity inhibits enterprises from becoming data-driven



68%

of enterprise data is not analyzed

80%

proportion of time on data cleaning, integration and preparation

82%

data quality a barrier to their data integration projects

82%

of enterprises are inhibited by data silos

There is no AI without IA

IA = Information Architecture

Sources: Seagate Technology Report, Corinium, Dun & Bradstreet

Challenges, old and new

Block Storage is a major cost of warehousing

Cost Performance Optimization

Balance Performance and cost on a per workload basis.

Data Silos cost businesses millions in replication and movement

Unified Virtual Warehouse

Query your globally distributed data intelligently.

Data is stored with multiple cloud vendors and on-prem

Built to be hybrid

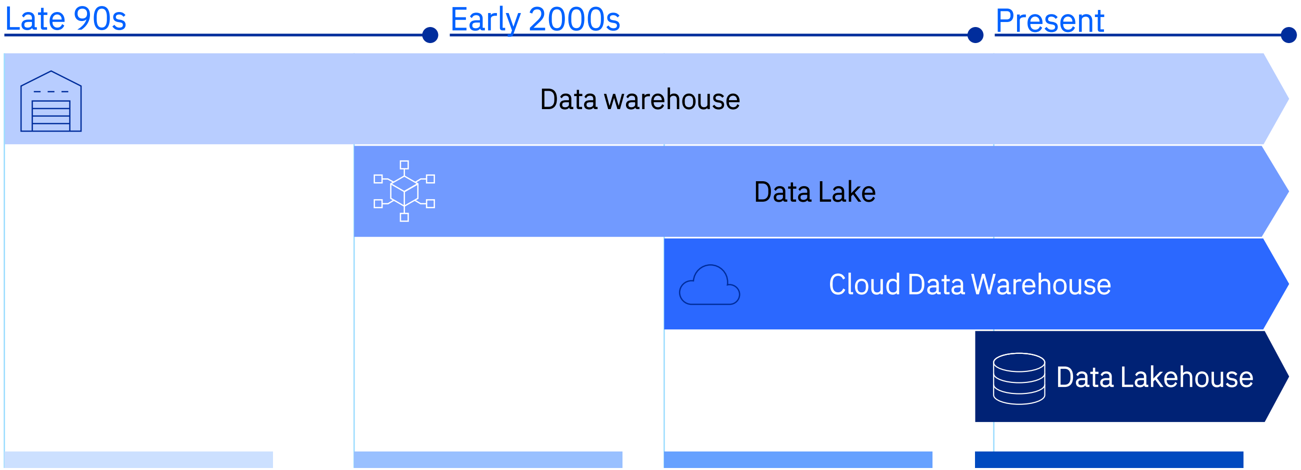
Designed to operate on data stored in hybrid environments

Proprietary formats prevent sharing and data proliferation

Open Data Formats

Share data among a variety of tools using Open Data Formats

Data Lakehouses are emerging technologies that solve for a new age of analytics



Late 90s

Early 2000s

Present

Data warehouse

Data Lake

Cloud Data Warehouse

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

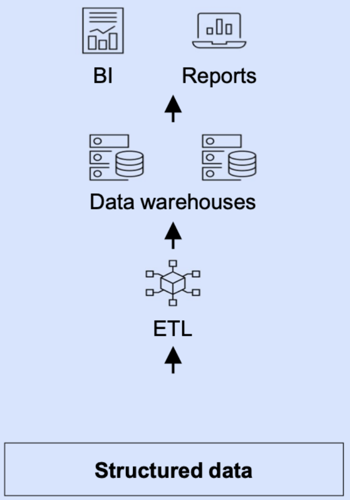
Data migration
Vendor lock-in
High costs
Limited AI/ML use cases

Still Early on Tech Curve
BI and AI in one place
Lower cost than warehouse
Open and Flexible

- Most enterprises today require two-tier architectures – both a data lake and multiple warehouses
- Data is moved and replicated for lake to warehouse, and the warehouse is still the access layer for key data
- Data Lakehouses combine the best of warehouses and data lakes
- **Data warehouse engine + Data Lake storage**

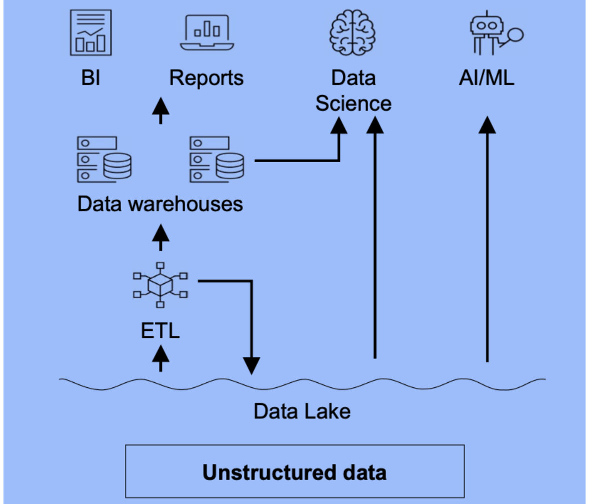
Where does a Lakehouse fit ?

Data Warehouse



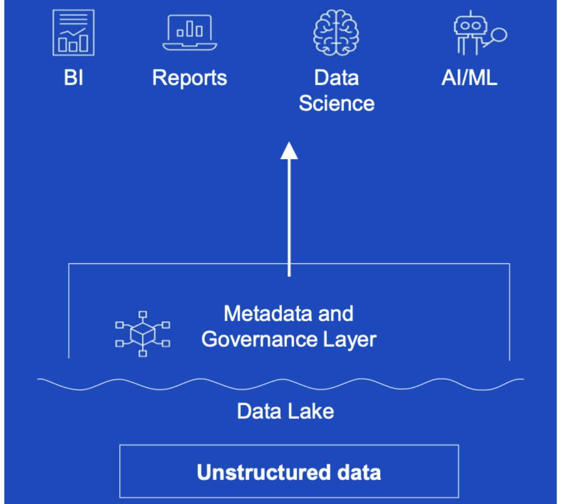
REFINED highly structured data
SMALL cleaned data sets
RELATIONAL

Data Lake



RAW data with minimal processing
LARGE data sets in many forms
UNDEFINED data for many use cases

Data Lakehouse



RAW data with minimal processing
LARGE data sets in many forms

Why Lakehouse

Current Challenges and Opportunity

Data Warehouse Challenges

- Proprietary data formats
- Vendor lock in
- Less flexible
- Elasticity scale limitations
- Expensive

Hadoop* Data Lake Challenges

- No ACID
- Poor in place Performance
- Failure to address real time requirements
- Narrow user focus mainly Data Science & ML
- Expensive to expand to generic BI and Analytical use cases
- High Skill to maintain and operate

Data Lakehouse

- Open Data File & Table formats
- ACID-compliancy – concurrent engine access
- Data Versioning
- Compute & Storage separation
- In-built Governance with policy enforcement
- Regulatory compliance (data lineage, origin, life cycle)

Introducing Project Saagar: the IBM Data Lakehouse

Data *Lake*
+
Ware*house*

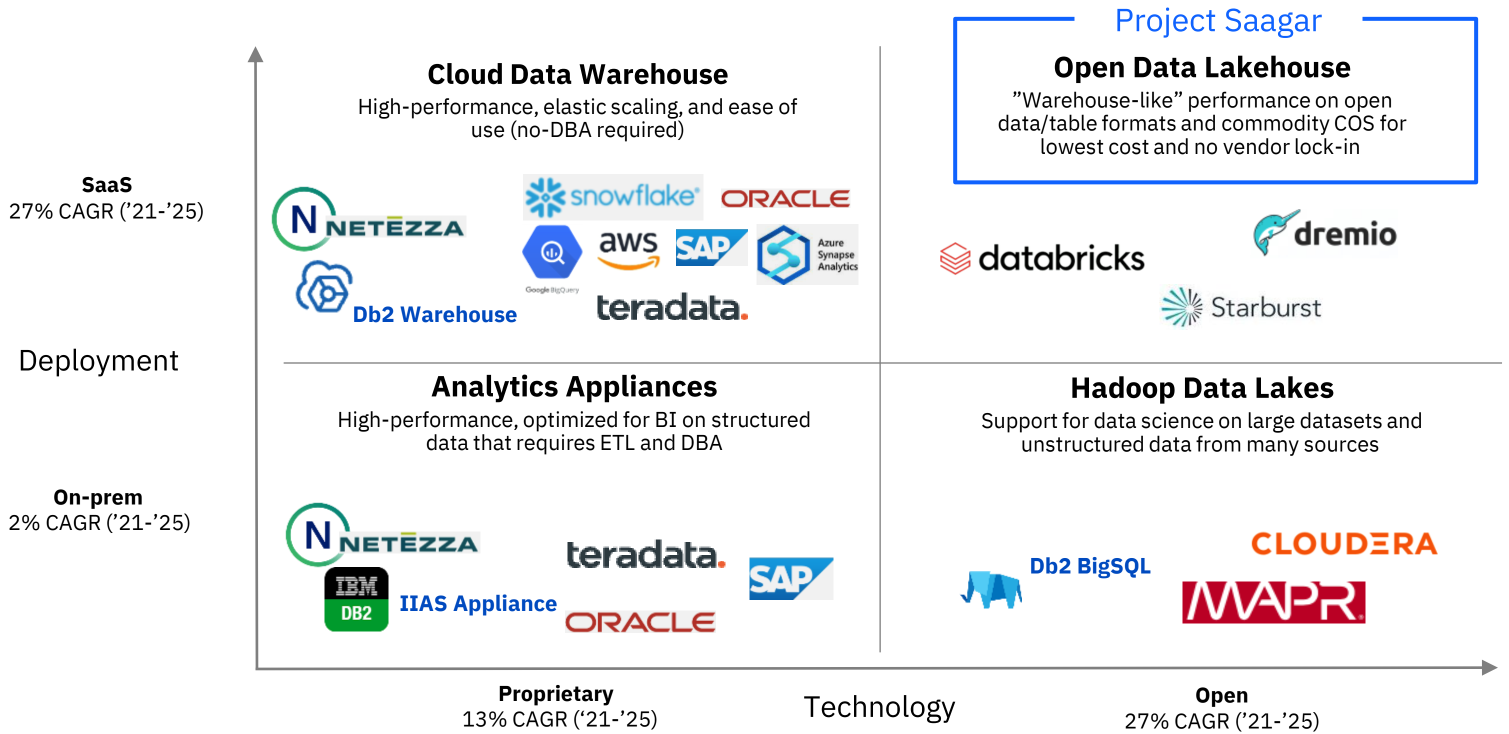
Our vision - An open, flexible, cost-effective, governed data management service

The IBM data lakehouse brings together the advantages of data warehouses and data lakes within a new architectural approach while leveraging open-source technologies such as the **Presto** SQL engine and the **Iceberg** table format.

It enables organizations to store data on low-cost object storage while ensuring data is **open**, available and **governed** for the various business needs and the different analytical engines organizations are using today.

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 market is valued at around **$29B** (warehouse and data lake) and it is projected to **17%** CAGR ('21-26')



Project Saagar

Open Data Lakehouse

"Warehouse-like" performance on open data/table formats and commodity COS for lowest cost and no vendor lock-in

Cloud Data Warehouse

High-performance, elastic scaling, and ease of use (no-DBA required)

Db2 Warehouse

Analytics Appliances

High-performance, optimized for BI on structured data that requires ETL and DBA

IIAS Appliance

Hadoop Data Lakes

Support for data science on large datasets and unstructured data from many sources

Db2 BigSQL

SaaS
27% CAGR ('21-'25)

Deployment

On-prem
2% CAGR ('21-'25)

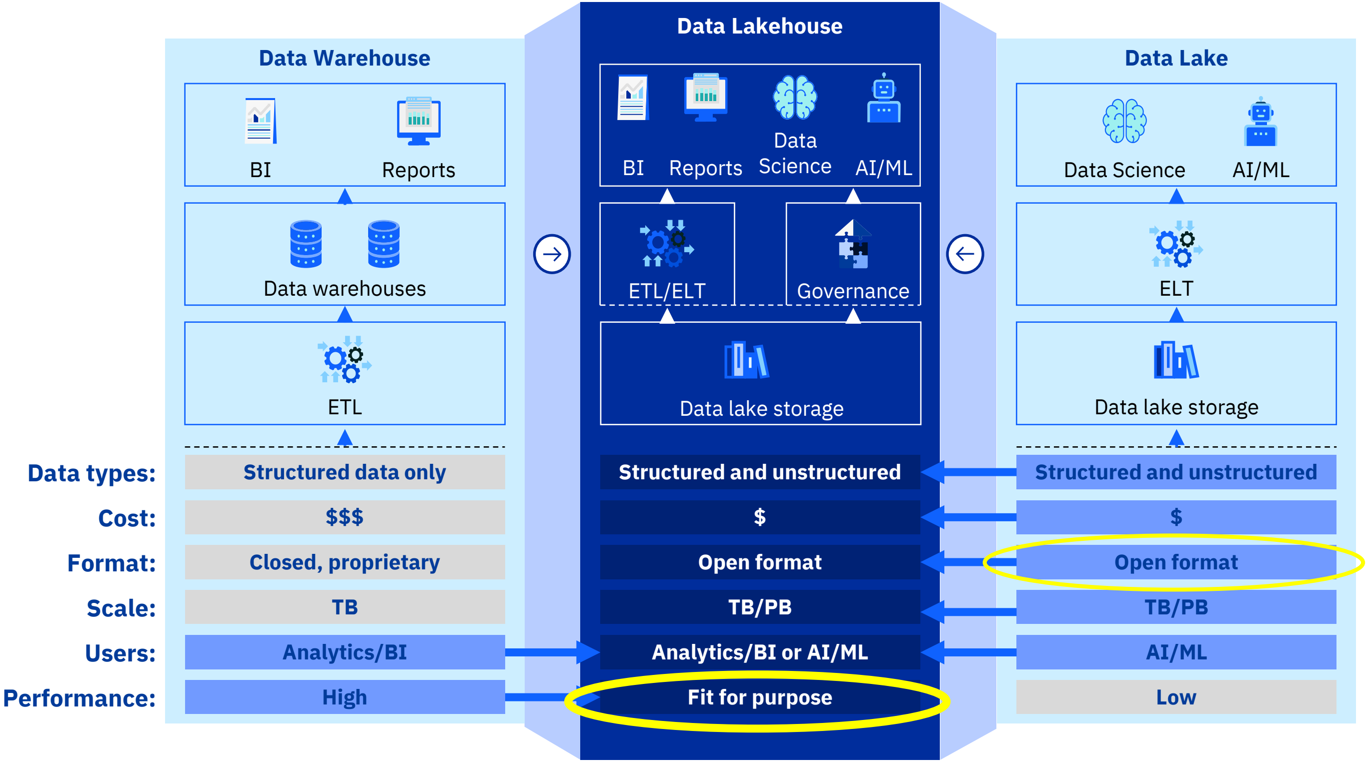
Proprietary
13% CAGR ('21-'25)

Technology

Open
27% CAGR ('21-'25)

Sources: IDC Data Management Forecast (November 2021), IDC BDA Forecast (June 2021), MI modeling

Lakehouse = data warehouses + data lakes



	Data Warehouse	Data Lakehouse	Data Lake
Data types:	**Structured data only**	**Structured and unstructured**	**Structured and unstructured**
Cost:	**$$$**	**$**	**$**
Format:	**Closed, proprietary**	**Open format**	**Open format**
Scale:	**TB**	**TB/PB**	**TB/PB**
Users:	**Analytics/BI**	**Analytics/BI or AI/ML**	**AI/ML**
Performance:	**High**	**Fit for purpose**	**Low**

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 and Analytics
2. Typically deployed over cloud only with no support for multi-/hybrid -cloud deployments
3. Minimal governance and metadata capabilities to deploy across your entire ecosystem

IBM Lakehouse’s key components

key components multi-engine, open format and built-in enterprise governance

Core Lakehouse functionality

Ecosystem architecture

Your existing ecosystem

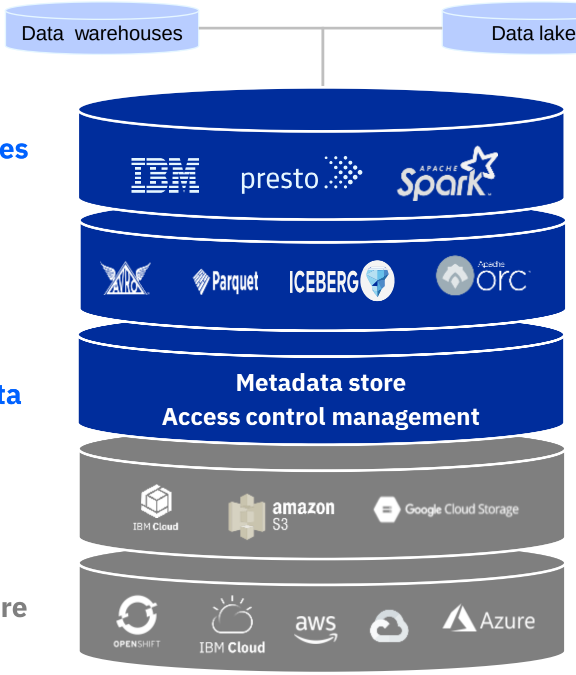
Query engines

Data format

Governance and Metadata

Storage

Infrastructure


Data warehouses

Data lakes

Metadata store

Access control management

The IBM Open Lakehouse

Multiple engines. Presto and Spark are **open-source**, distributed engines that provide **fast, reliable, and efficient processing** at scale

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

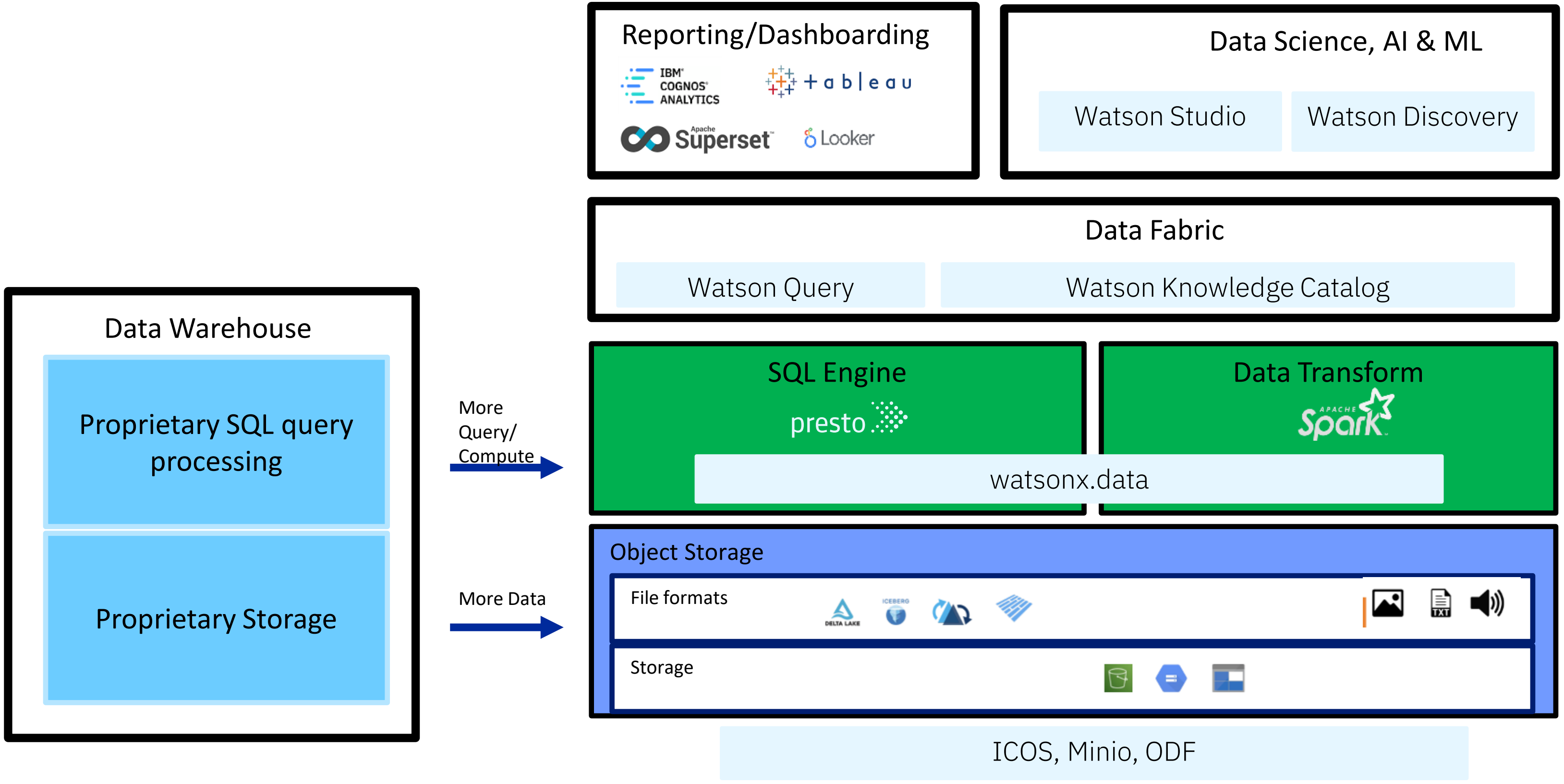
Built in governance that is developed on **open-source solutions**. Also compatible with existing solutions, including IBM WKC

Cost effective, simple object storage for your Lakehouse. Available across hybrid-cloud and multi-cloud environments

Multi-cloud integrations with IBMs Kubernetes platform, OpenShift (or other 3rd party Kubernetes)

- **Optimize workload costs and performance using multi-engine functionality**
- **Access 100% of your data across lakes, warehouses and other lakehouses**
- **Ensure governance and reduce time to insight with single shared metadata and access management**
- **Reduce storage costs compared to costly relational warehouse storage**
- **Deploy on any infrastructure with everything you need built in**

Data Warehouse to AI Open Lakehouse


Data Warehouse

Proprietary SQL query processing

Proprietary Storage

More Query/ Compute

More Data

Reporting/Dashboarding

Data Science, AI & ML

Watson Studio

Watson Discovery

Data Fabric

Watson Query

Watson Knowledge Catalog

SQL Engine

Data Transform

watsonx.data

Object Storage

File formats

Storage

ICOS, Minio, ODF

IBM Data and AI product portfolio

Designing exceptional customer and employee experiences by making data ready for AI — and everyone

IBM Cloud Pak for Data
Out-of-the-box cloud data platform to deliver your services on the cloud of your choice

Data Insights and Applications
Accelerate data analytics and AI

Business Analytics
- Business Analytics Enterprise
- Cognos Analytics
- Planning Analytics

Business Apps
- Watson Assistant
- Watson Discovery
- Watson Orders

Data Fabric & AI Lifecycles
Establish a data architecture to simplify data access and automate data and AI lifecycles

Data Science and MLOps
- Watson Studio
- Watson Machine Learning

Data Governance
- Watson Knowledge Catalog
- Match 360
- Manta

AI Governance
- OpenPages
- Watson Studio
- Watson OpenScale
- Watson Knowledge Catalog

Data Integration
- DataStage
- Databand
- Watson Query
- Watson Pipelines
- Data Replication

Data Sources
Store, manage and unlock data

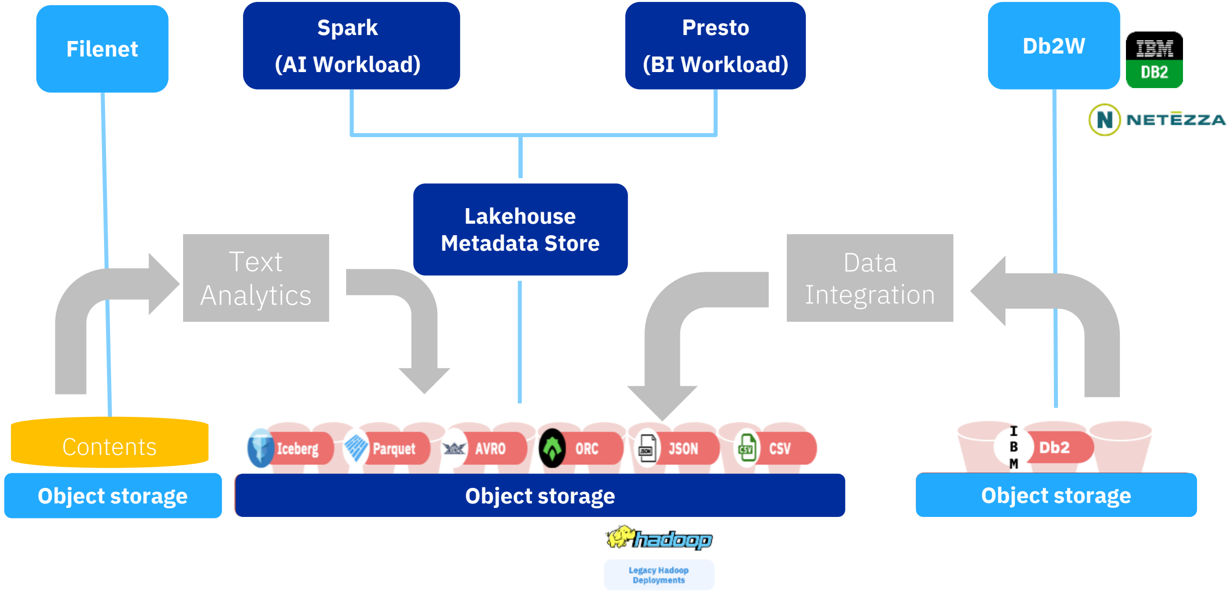
Transactional
- IBM Db2
- IBM Informix

Warehouse
- IBM Db2
- Netezza

Lakehouse
- Project Saagar

Strategic Partners
- Cloudera
- DataStax
- EDB
- MongoDB
- SingleStore

IBM Lakehouse ecosystem for maximum workload coverage



Modules

- Data Warehouse
- ECM
- Lakehouse
  - Addon: Data Virtualization
  - Addon: Data Integration to transfer from DW to Lakehouse
  - Addon: Text Analytics for extract unstructured contents to Lakehouse

Next Roadmap

- Data Governance
- MLOps

The IBM Open Data Lakehouse

The IBM data lakehouse brings together the advantages of data warehouses and data lakes within a new managed cloud service and self-managed on any Cloud or on-premise.

A Low-Cost and extensible Query Engine
Presto is an **open-source, fast** and reliable **SQL engine** for Data Analytics and data lake houses.

A proven and reliable metadata repository
The hive **metastore** is the de facto standard in open-source data lake metadata management

Stores data in Object Store buckets in the Iceberg open data format to **facilitate data access and sharing** across applications

Open, Flexible, and Modular
Lakehouse is designed to enable customers to standardize their data formats and metadata with unified data governance

Evolve your big data platform
The simplest path to upgrade from traditional big data platforms, either as a side car or moving data to cloud object storage

Limitless scalability and elasticity
Explore, shape, and analyze data at any scale by separating storage and compute

Integrates readily with Db2 Warehouse and Netezza to support the **right data engine for the right workloads at the right cost**

IBM