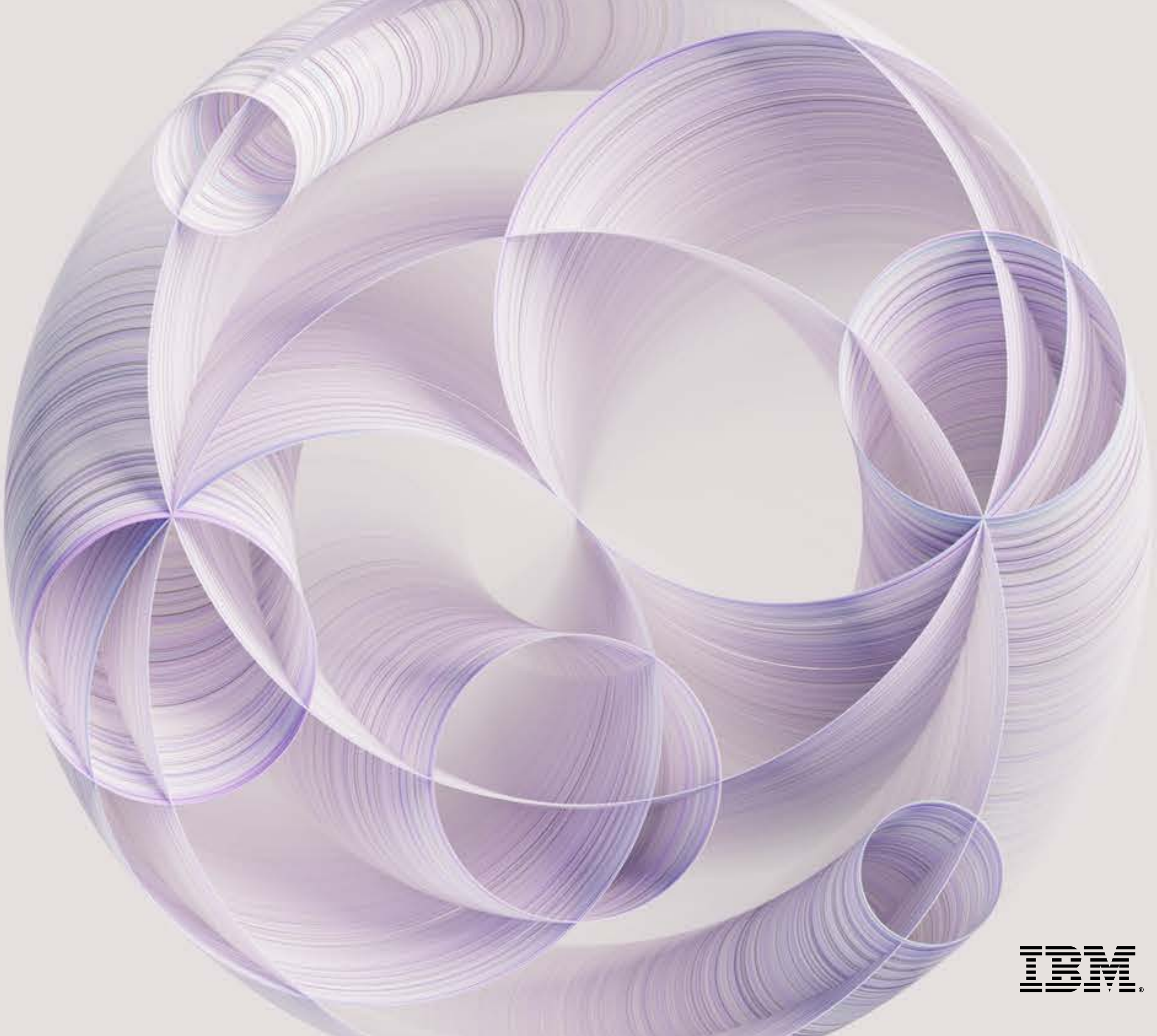
Watsonx.data

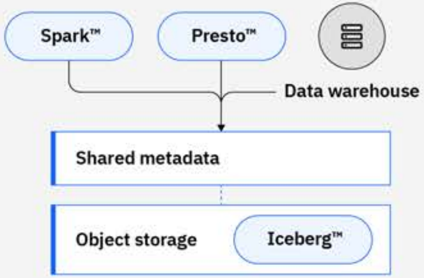
Day 1
Fundamentals



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

An open data store built for hybrid deployment of your analytics and AI workloads

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



Spark™

Presto™

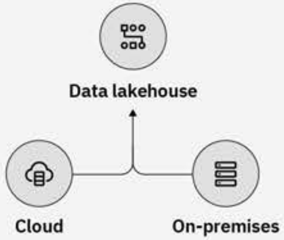
Data warehouse

Shared metadata

Object storage

Iceberg™

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

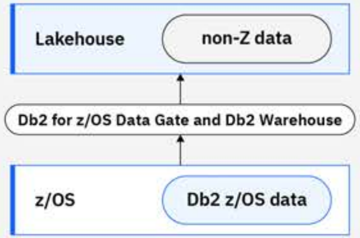


Data lakehouse

Cloud

On-premises

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



Lakehouse

non-Z data

Db2 for z/OS Data Gate and Db2 Warehouse

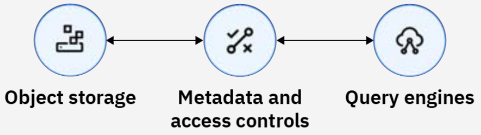
z/OS

Db2 z/OS data

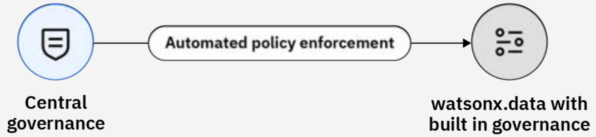
Get started in minutes with built-in governance, security and automation.

Accelerate time to trusted analytics and AI

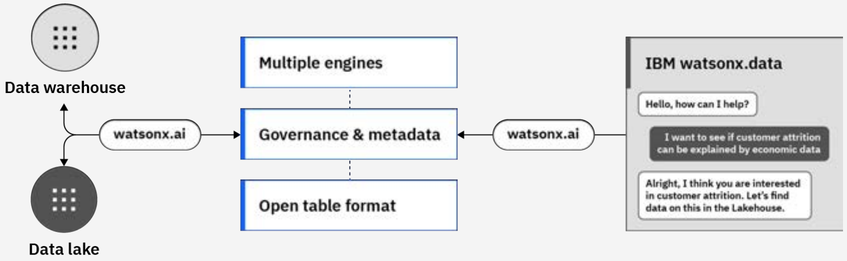
Connect to your existing analytics data and deploy fit-for-purpose engines in minutes



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


Central governance
Automated policy enforcement
watsonx.data with built in governance

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



Hello, how can I help?

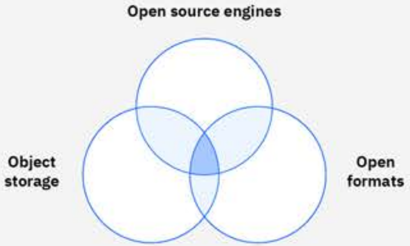
I want to see if customer attrition can be explained by economic data

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

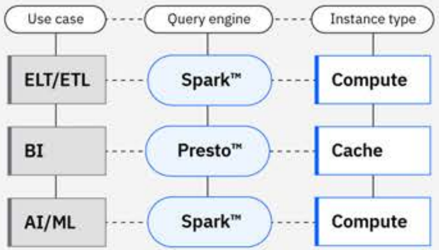
Reduce your data warehouse costs by up to 50%* by optimizing workloads

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

1 Share data between multiple analytics engines

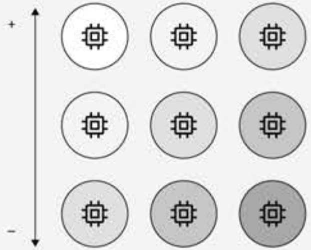


2 Use fit-for-purpose compute and cache-optimized instances



Use case	Query engine	Instance type
ELT/ETL	Spark™	Compute
BI	Presto™	Cache
AI/ML	Spark™	Compute

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.

Common data file formats

Computer systems and applications store data in files

Data can be stored in binary or text format

File formats can be open or closed (proprietary/lock-in)

Open formats (Parquet, ORC, and Avro) are 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

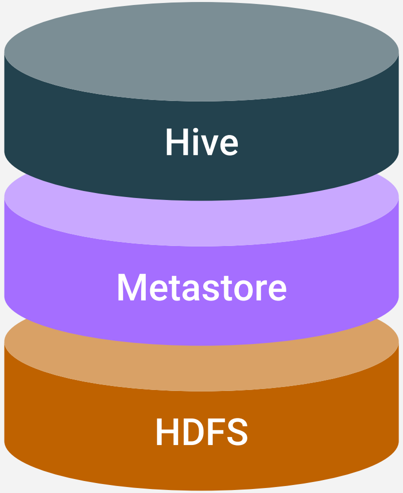
- Open-source
- Row-oriented data format and serialization framework
- Robust support for schema evolution
- Mix of text/binary

THE OPEN DATA LAKEHOUSE DIFFERENTIATOR

“IBM / Cloudera Shares a Joint Vision”

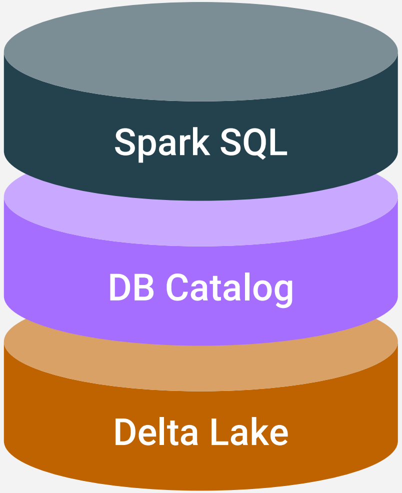
The “Old” Way

Hive SQL over data in HDFS



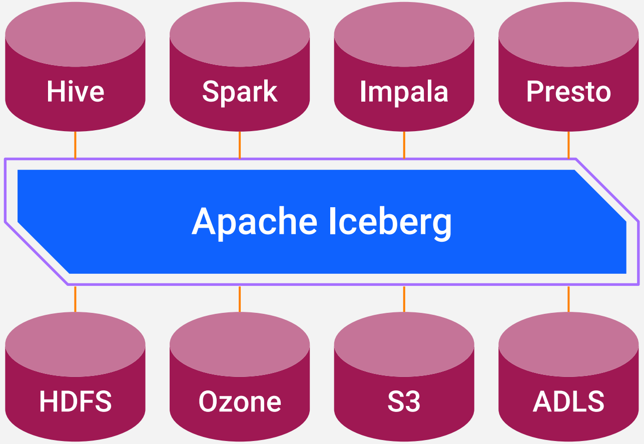
“Some” other way

Spark SQL over Delta tables



The Iceberg way

Multi-function analytics over all your data

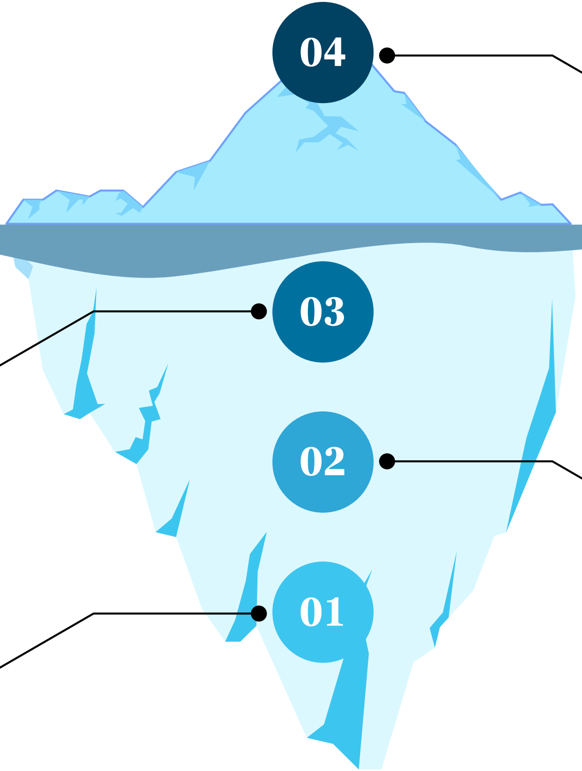


Powerful Engines

Open-source Engines for each use case

presto

APACHE Spark



HIVE

Catalog

IBM use Hive Metastore to catalog Iceberg Table Formats

ICEBERG

Table Format

Organize the data where it lives into tables using Iceberg

Parquet

File Format

Store Data in Apache Parquet Files (Open Columnar Format)

Storage

amazon S3

IBM Cloud Object Storage

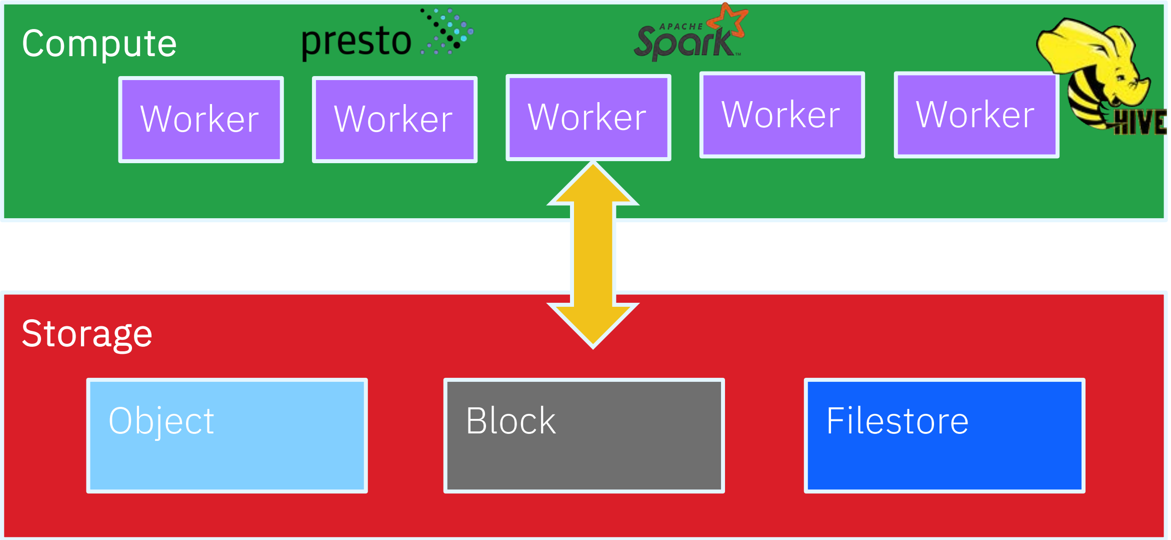
Hyperscaler Provider (AWS, Azure or IBM)

Which cluster type?

Depending on your use-case!

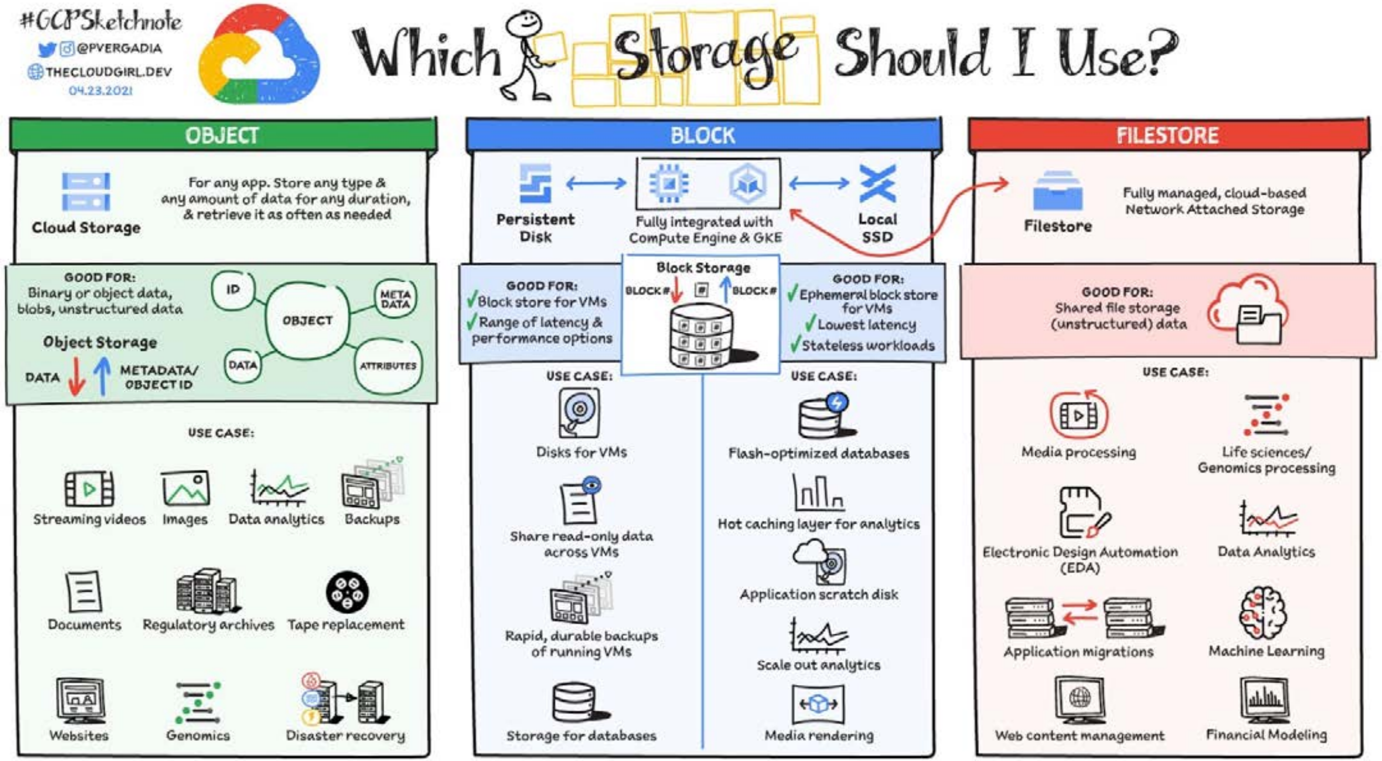
There are three models you should consider for your cluster:

- Compute
- Balanced
- Storage



Which storage?

Recommendations for Storage Decision



#GCPSketchnote
@PVERGADIA
THECLOUDGIRL.DEV
04.23.2021

Which Storage Should I Use?

OBJECT

Cloud Storage

For any app. Store any type & any amount of data for any duration, & retrieve it as often as needed

GOOD FOR:
Binary or object data, blobs, unstructured data

Object Storage

USE CASE:

- Streaming videos
- Images
- Data analytics
- Backups
- Documents
- Regulatory archives
- Tape replacement
- Websites
- Genomics
- Disaster recovery

BLOCK

Persistent Disk ⟷ Fully integrated with Compute Engine & GKE ⟷ Local SSD

Block Storage

GOOD FOR:
✓ Block store for VMs
✓ Range of latency & performance options

GOOD FOR:
✓ Ephemeral block store for VMs
✓ Lowest latency
✓ Stateless workloads

USE CASE:

- Disks for VMs
- Share read-only data across VMs
- Rapid, durable backups of running VMs
- Storage for databases

USE CASE:

- Flash-optimized databases
- Hot caching layer for analytics
- Application scratch disk
- Scale out analytics
- Media rendering

FILESTORE

Filestore

Fully managed, cloud-based Network Attached Storage

GOOD FOR:
Shared file storage (unstructured) data

USE CASE:

- Media processing
- Life sciences/ Genomics processing
- Electronic Design Automation (EDA)
- Data Analytics
- Application migrations
- Machine Learning
- Web content management
- Financial Modeling