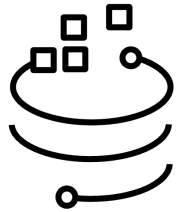
Data Lakehouse with watsonx.data and IBM Storage

IBM Storage Scale Days 2024

March 5-7, 2024 | Stuttgart Marriott Hotel Sindelfingen

Harald Seipp, IBM

Disclaimer



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. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

IBM Storage Software Strategy

Providing solutions around	01 IBM Storage for Hybrid Cloud	02 IBM Storage for Data and AI	03 IBM Storage for Data Resiliency
	Drive innovation and scale application modernization with container-enabled enterprise storage that deploys seamlessly across hybrid infrastructures with a simple and consistent user experience.	Accelerate business results and innovation and unlock the latent value of unstructured data across the data ecosystem by eliminating data silos, advancing data discovery and classification.	Reduce the threat exposure window from days to hours and proactively safeguard data with a multi-faceted and scalable data resiliency approach that defends against cybervulnerabilities from detection to recovery.
Designed for	Chief Technology Officer | IT Director | VP OpenShift Engineering | Director Open Infrastructure | Cloud Architect | Data Scientist	VP of Engineering | VP of Development | Chief Data Officer | Data Architect | HPC Specialist	IT Architect | Storage Architect | Chief Information Security Officer | IT Director
Leading with	IBM Storage Fusion	IBM Storage Scale IBM Storage Ceph	IBM Storage Defender
Delivered on	Edge – to – Core – to – Cloud		
That run on sustainable infrastructure	IBM Storage Fusion HCI System	IBM Storage Scale System	IBM Storage FlashSystem IBM Storage DS8K IBM Storage Tape

IBM’s Global Data Platform

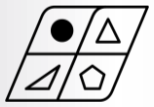
Data Sources + IBM’s Data Strategy = Business Value

Data Sources


Existing Data


Unstructured

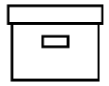

Semi-Structured


Mixed Media /Images /Genomics


New Generated IoT

AI/ML 

Analytics/HPC 

Backup/Archive 

Cloud Native 

Global Data Platform

Data Access Services
Multi-protocol & Performance

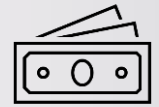
Data Caching or Core Services
Global Connectivity

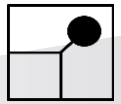
Data Management Services
Policy Automation

Data Resiliency Services
Cyber Secure Recovery

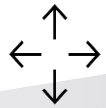
Business Value

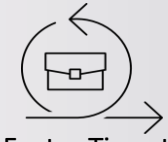

Resource Efficiencies


Cost Efficiency

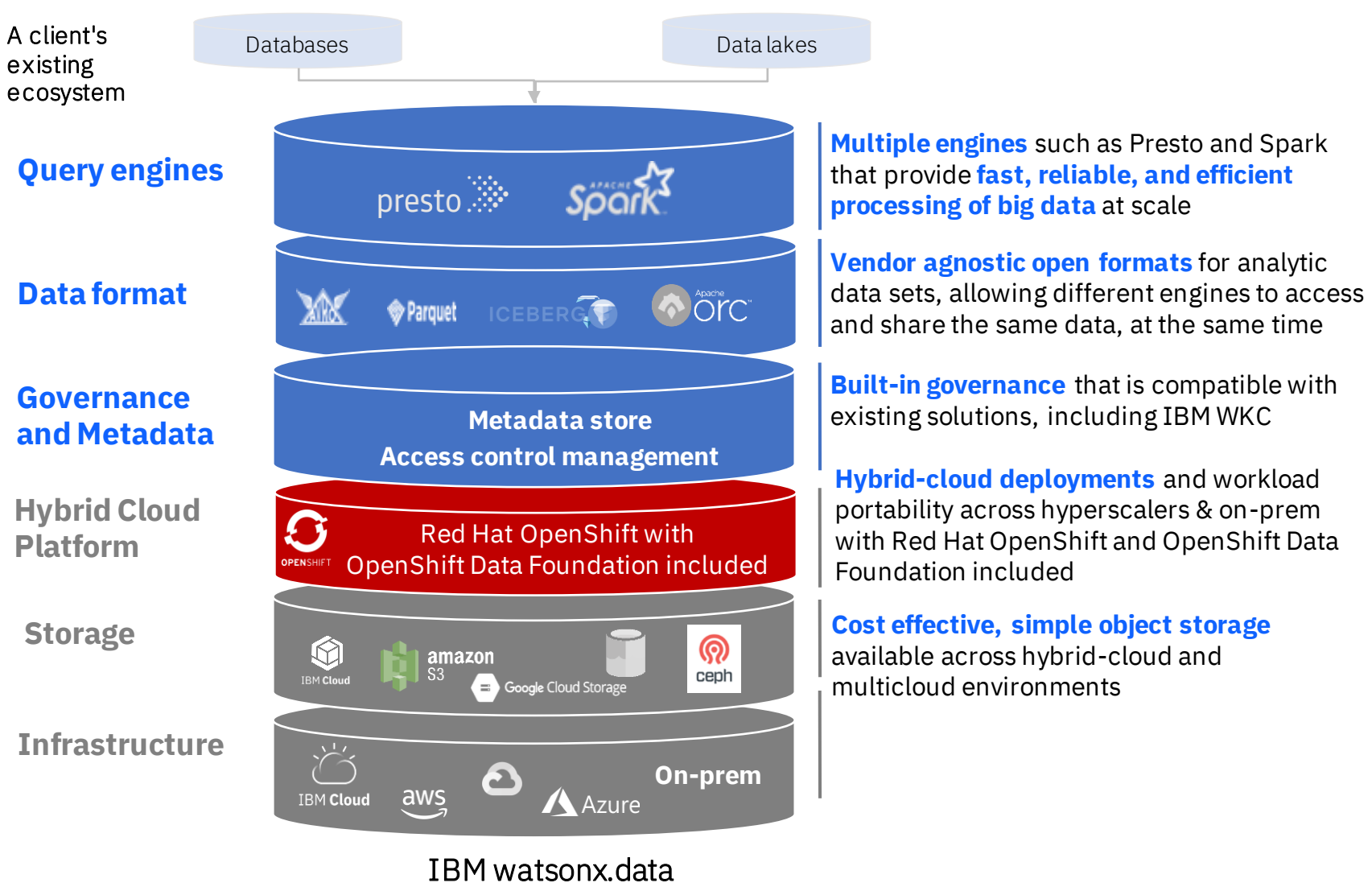

Security


Risk Reduction


Increase Agility


Faster Time to Market

Overview of the 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

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 compatible with existing solutions, including IBM WKC

Hybrid-cloud deployments and workload portability across hyperscalers & on-prem with Red Hat OpenShift and OpenShift Data Foundation included

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

Core lakehouse functionality

Ecosystem infrastructure

- **Optimize workload costs and performance using multi-engine functionality**
- **Access 100% of the data across databases and data lakes**
- **Ensure governance and reduce time to insight with centralized metadata and access management**
- **Reduce storage costs and facilitate data ingest**
- **Deploy on any infrastructure and optimize available resources**

Components for a modern Lakehouse solution

Engines

Engines enables querying, analytics and transformations for the lakehouse. The expectation is to enable the use of the **right engine for the right workloads & use cases.**

Metadata repository

A repository maintains schema and table metadata so that all engines and users can **consistently** locate and query against their data in a structured manner.

Data Storage

The data storage is where the data is physically stored. Customers would expect to use their own storage such as S3 or HDFS and locate them wherever they need. With **compute engines separated from storage**, concurrent accesses by multiple engines must be enabled in the Lakehouse

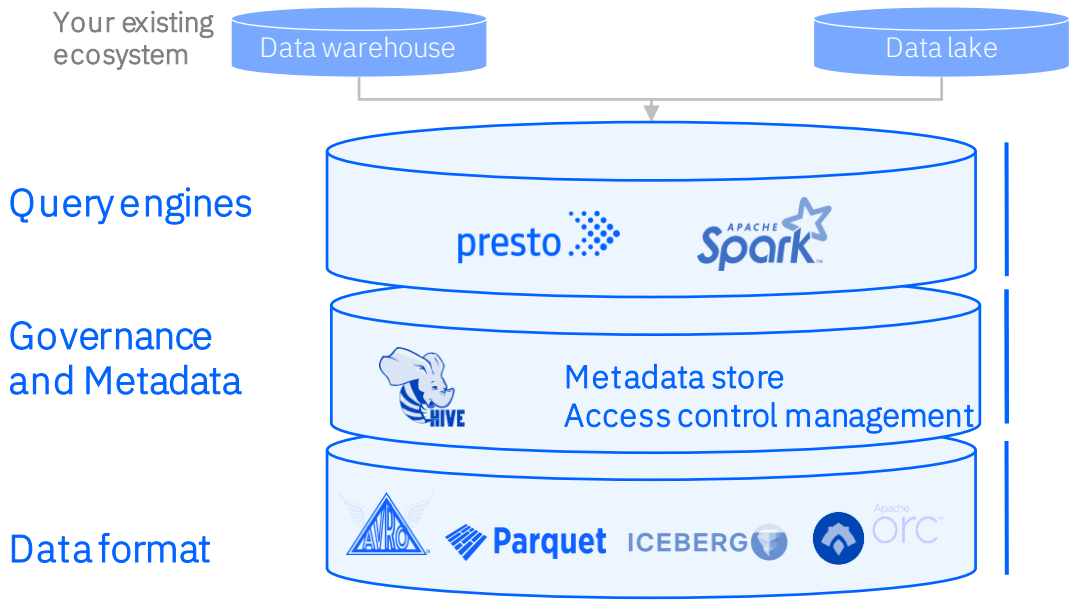
Data Governance

A lakehouse needs to be able to **enforce Data policies** for access, privacy and safe harbor or sovereignty regulations

Cloud & on-premise Service

Lakehouses need to be deployable or burstable **anywhere** and even span clouds in a hybrid fashion. Applications and end users would need to access lake houses engines from anywhere

IBM Storage Software in watsonx.data (on-premises)



IBM Storage Fusion
OpenShift Appliance (HCI)
or
Software Defined Storage
(Data Foundation)

IBM Storage Scale
(File and object)

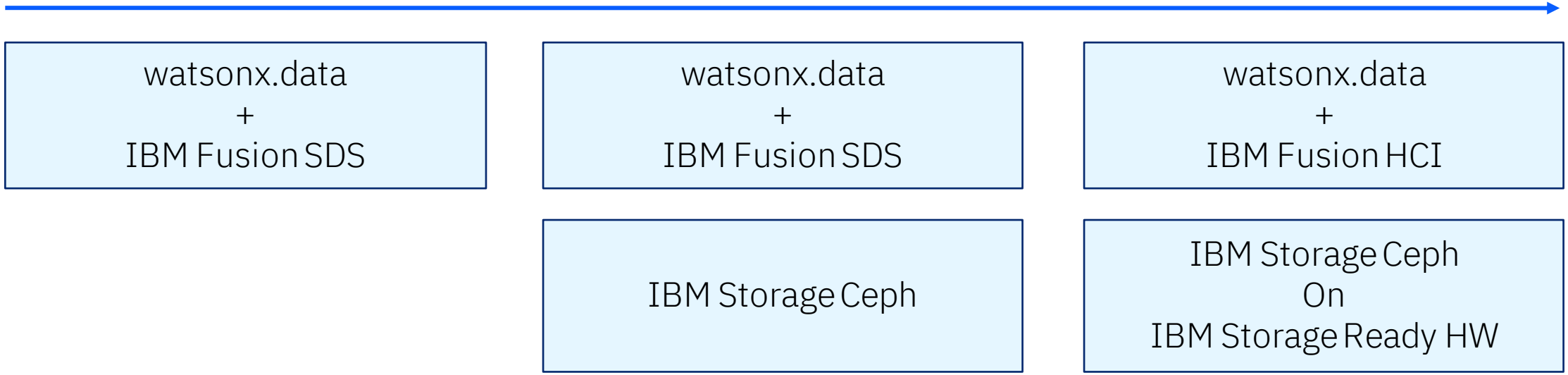
IBM Storage Ceph
(File, Object, Block)

Storage Fusion Integrated (Storage Scale or Data Foundation)

IBM Storage Discover
(Metadata catalog and governance)

IBM Storage Defender
(Data resilience: backup and cyber resilience)

Initial approach for IBM Storage in watsonx.data

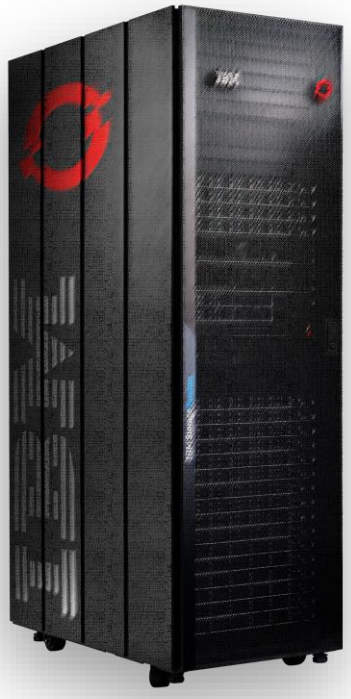


- Converged approach for POCs and other small environments.
- Internal Ceph Object (Data Foundation) for persistence.
- 100TB *usable* Fusion SDS included.

- Decoupled storage and compute.
- External Ceph object for persistence.
- 500TB *usable* IBM Storage Ceph included.

- Complete solution including hardware from IBM.
- 500TB *usable* IBM Storage Ceph included.

IBM Storage Fusion HCI System for AI/ML



"Lack of proper storage infrastructure is one of the key reasons many AI projects fail..."

IDC Strategic Imperatives for AI Storage Infrastructure June 2022

NVIDIA A100 GPUs

Optional Fusion HCI GPU servers to support up to 6 NVIDIA A100 GPUs

Scale-out parallel file system

High performance access to Petabytes of unstructured data through POSIX and CSI interfaces

Data cataloging

Scan and label unstructured data on NAS filers, S3 object stores, and SMB shares to build meta-data index catalogs

Data Mobility and Global Data Platform

Use Active File Management services to transparently ingest data from existing NAS filers and object stores into the Fusion scale-out parallel file system

Red Hat OpenShift | IBM Storage Fusion HCI System | NVIDIA

watsonx.data and Storage

Fusion HCI Rack

watsonx.data
DB2, Netezza, Presto, Spark

watsonx.ai
Inferencing/Fine tuning

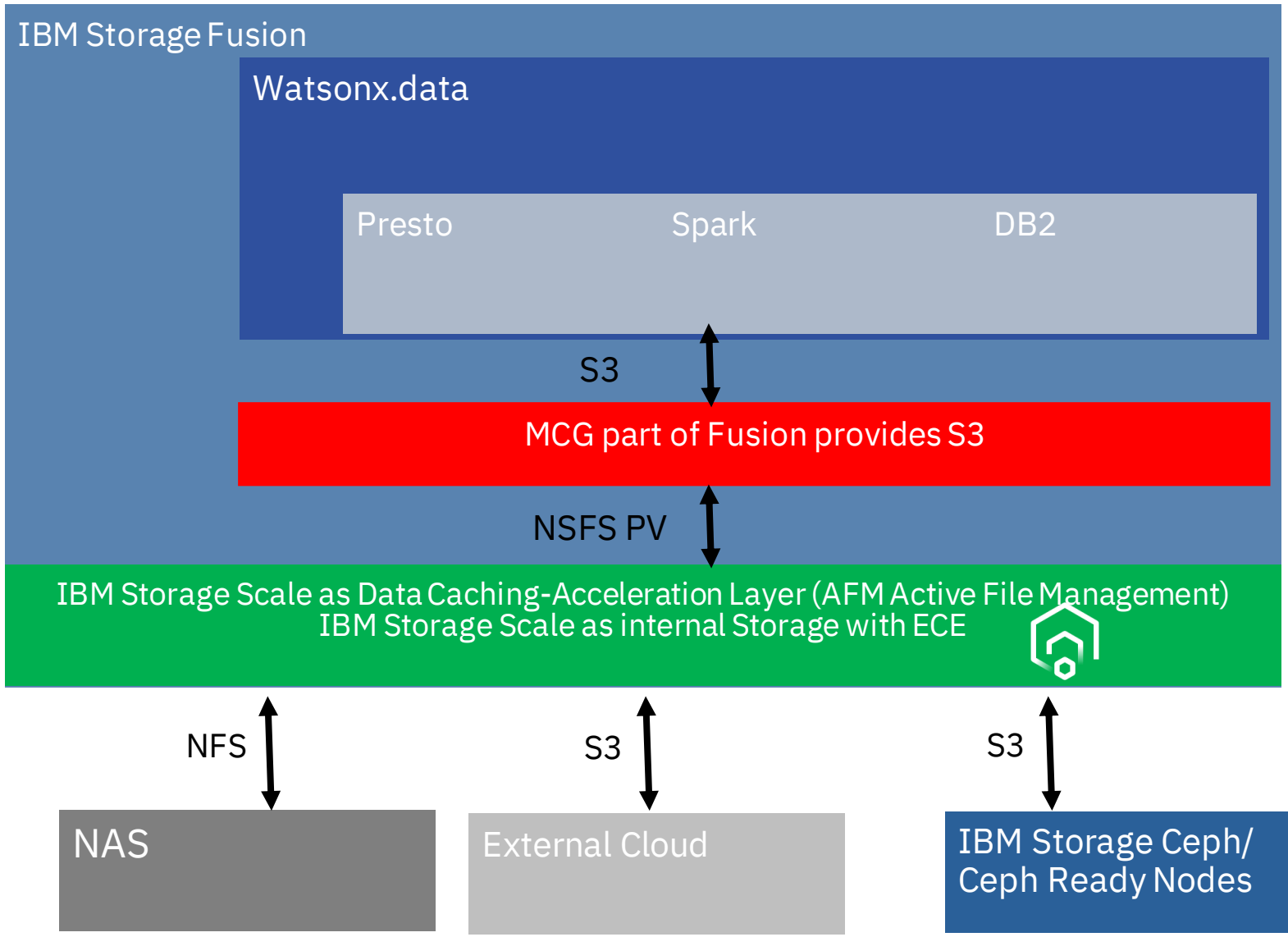
watsonx.governance

1 IBM Fusion HCI: watsonx compute and data acceleration appliance

IBM Storage Ceph:
watsonx persistent storage

2 Ceph Ready Nodes

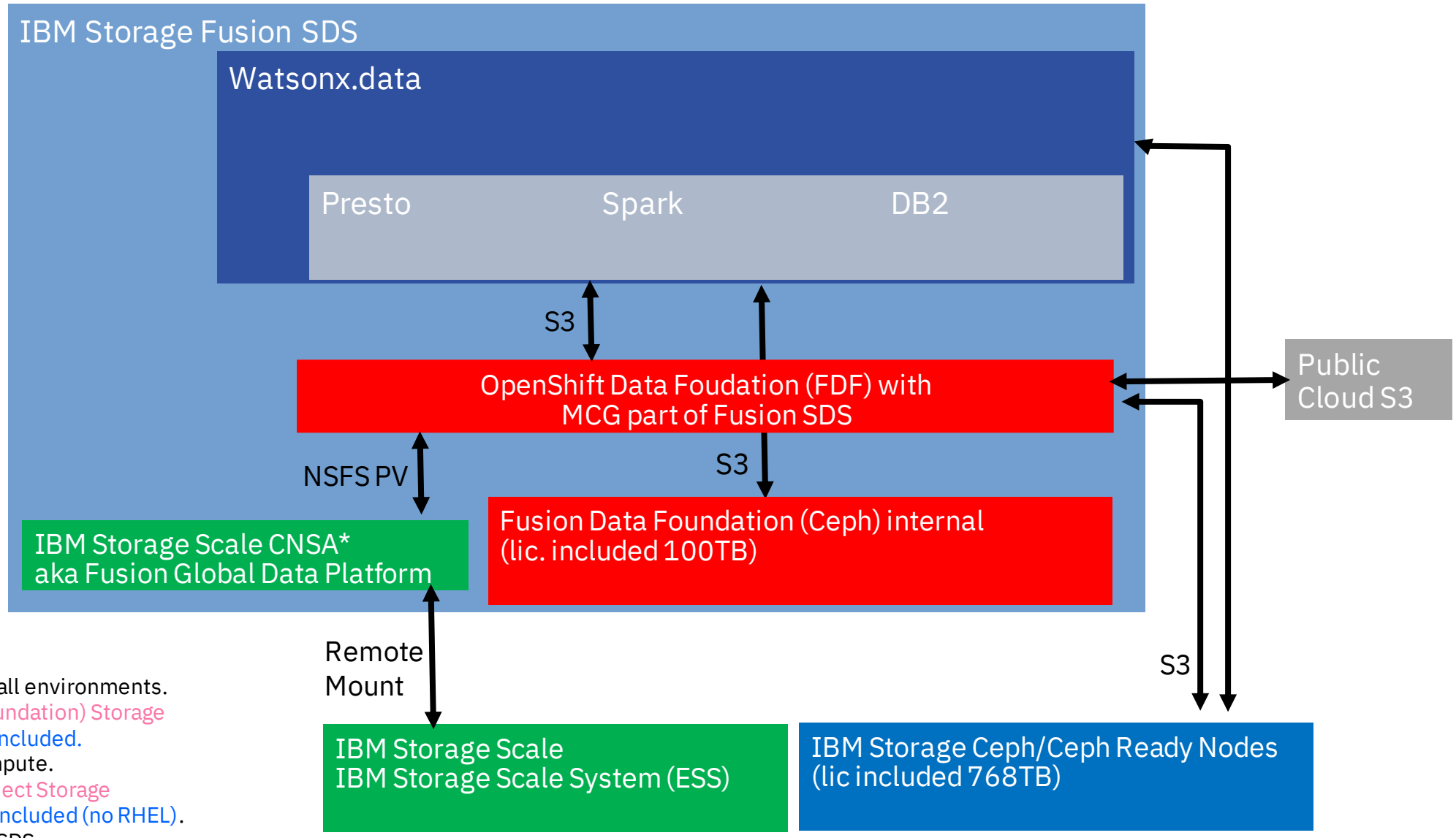
Fusion HCI: Storage options with watsonx.data



Watsonx.data

- 768TB raw IBM Storage Ceph included (no RHEL)
- 256 VPC Fusion

Fusion SDS: Storage options with watsonx.data



Fusion SDS

Watsonx.data:
SDS Converged- for POCs and small environments.

- Internal Ceph (Fusion Data Foundation) Storage
- 100TB useable in Fusion SDS included.

SDS Decoupled- Storage and Compute.

- External IBM Storage Ceph Object Storage
- 768TB raw IBM Storage Ceph included (no RHEL).
- 100 TB useable FDF in Fusion SDS

* Container Native Storage Access

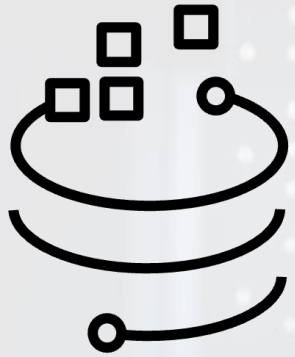
Benefits of using Storage Scale as Accelerator

Early benchmark testing has proven that one can gain

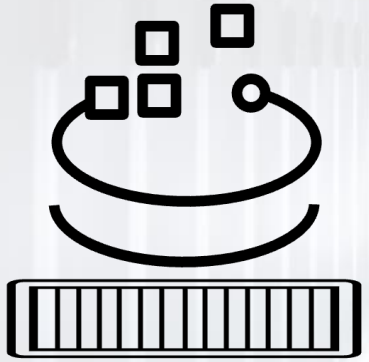
5-15x

TPC-DS query performance improvement with Storage Acceleration on

Thank you for using



Storage Scale



Storage Scale System