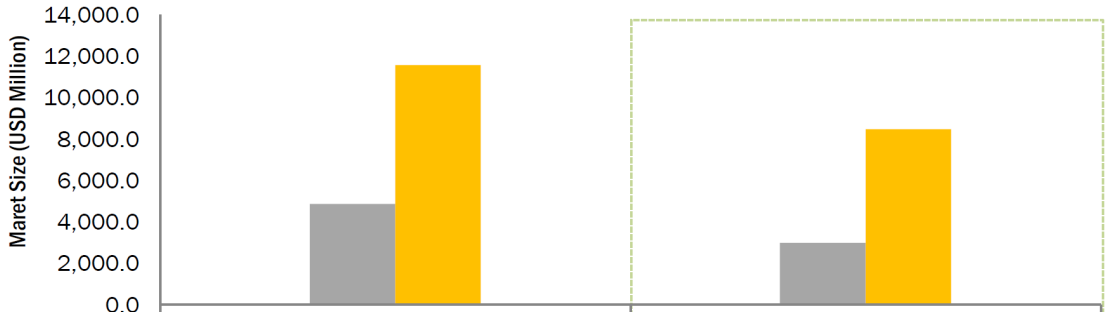
海量数据上云，跟我混

面向 Hadoop 和大数据分析的数据平台

周立旸 zhouly@cn.ibm.com
数据与 AI 存储及数据保护现代化
IBM 系统部

数据湖是企业数字化转型的重点领域

- 长期高速增长的IT领域
- AI 的普及进一步推动了企业数据湖的需求
- 客户在上市时间、性能和运营效率方面面临越来越大的挑战



	Solutions	Services
2019	4,866.9	2,983.1
2024	11,571.9	8,478.1
CAGR (2019-2024)	18.9%	23.2%

DATA LAKE MARKET SIZE, BY SOLUTION, 2017–2024 (USD MILLION)

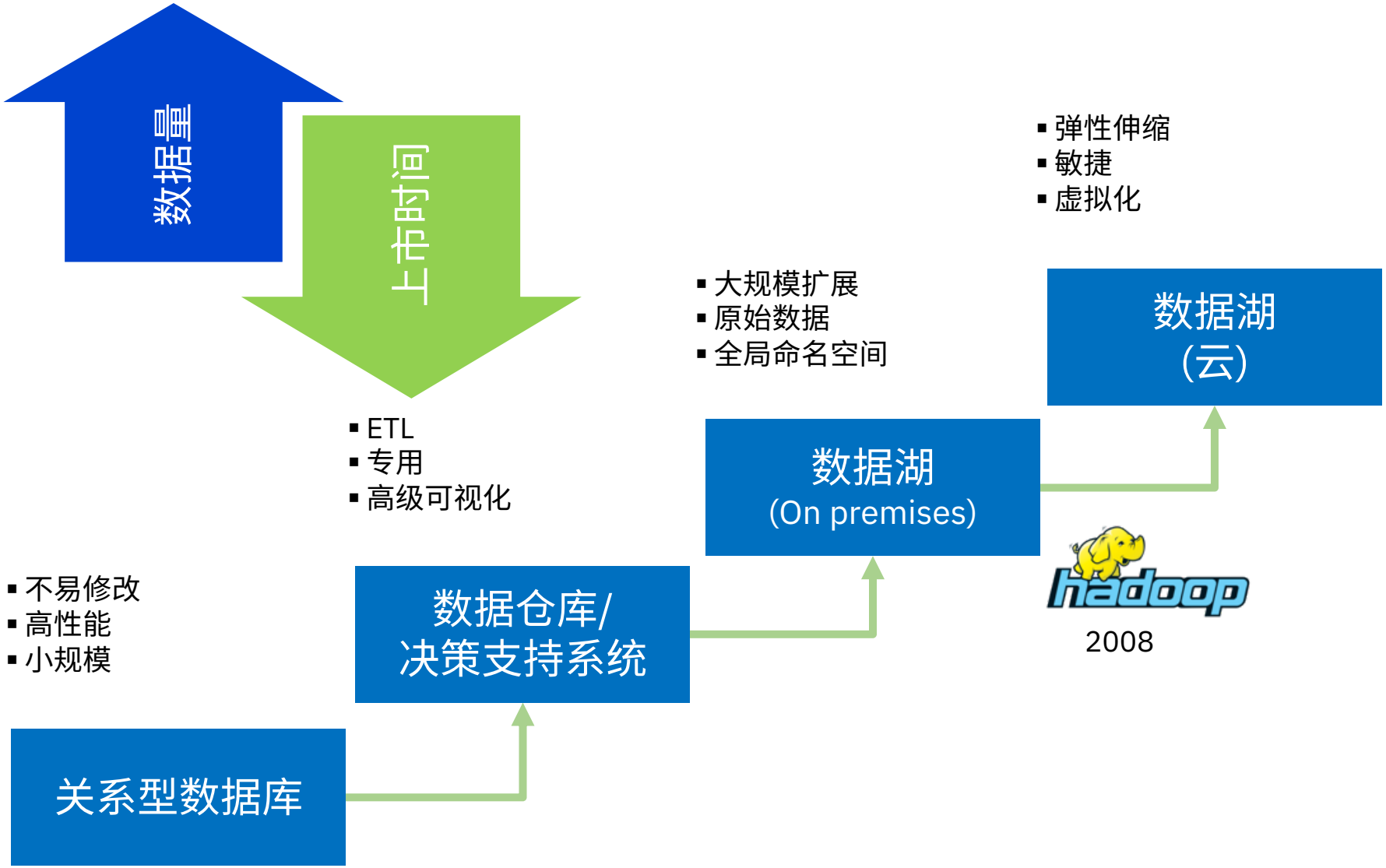
Solution	2017	2018	2019-e	2024-p	CAGR (2019–2024)
Data Discovery	491.5	587.7	704.9	1,519.8	16.6%
Data Integration and Management	618.5	747.8	907.6	2,095.5	18.2%
Data Lake Analytics	1,296.5	1,582.1	1,937.8	4,684.8	19.3%
Data Visualization	871.3	1,069.0	1,316.6	3,271.8	20.0%
Total	**3,277.8**	**3,986.6**	**4,866.9**	**11,571.9**	**18.9%**

e: estimated, p: projected

Source: “Data Lake Market: Global Forecast to 2024” , © Dec. 2019, MarketsandMarkets™

当数据湖遇见混合云

需要实现企业数据湖的现代化



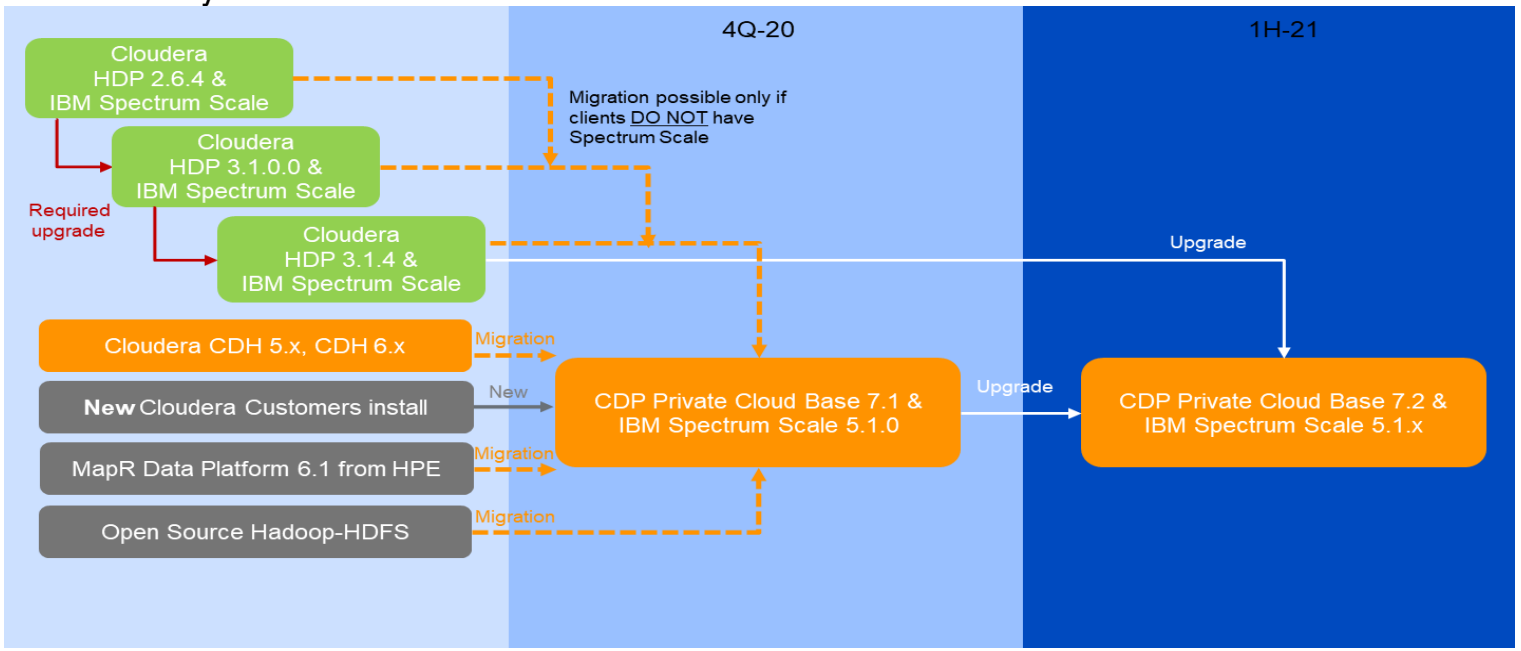
IBM 与 Cloudera 的战略合作，帮助客户实现更大价值

IBM + cloudera

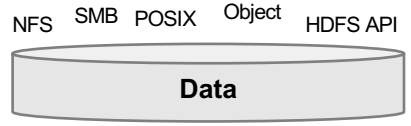
#1 open source Hadoop platform + IBM's leading value adds

- #1 SQL Engine for complex, analytical workloads
- Leader in Data Science (Source: Gartner)
- Leader in On-premise and Hybrid Cloud solutions
- Software defined storage with unmatched scalability

- #1 Open Source Hadoop Distribution
- 2500+ customers and 3000+ ecosystem partners
- Best-in-class 24x7 customer support
- Leading professional services and training



① 更快的分析，更少的占地空间

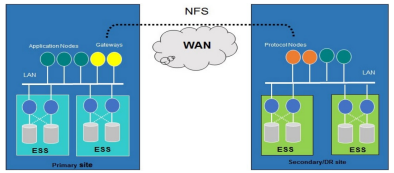


② 计算和存储分离，根据需求独立灵活扩展

③ 并行架构支持接近无限的扩展能力

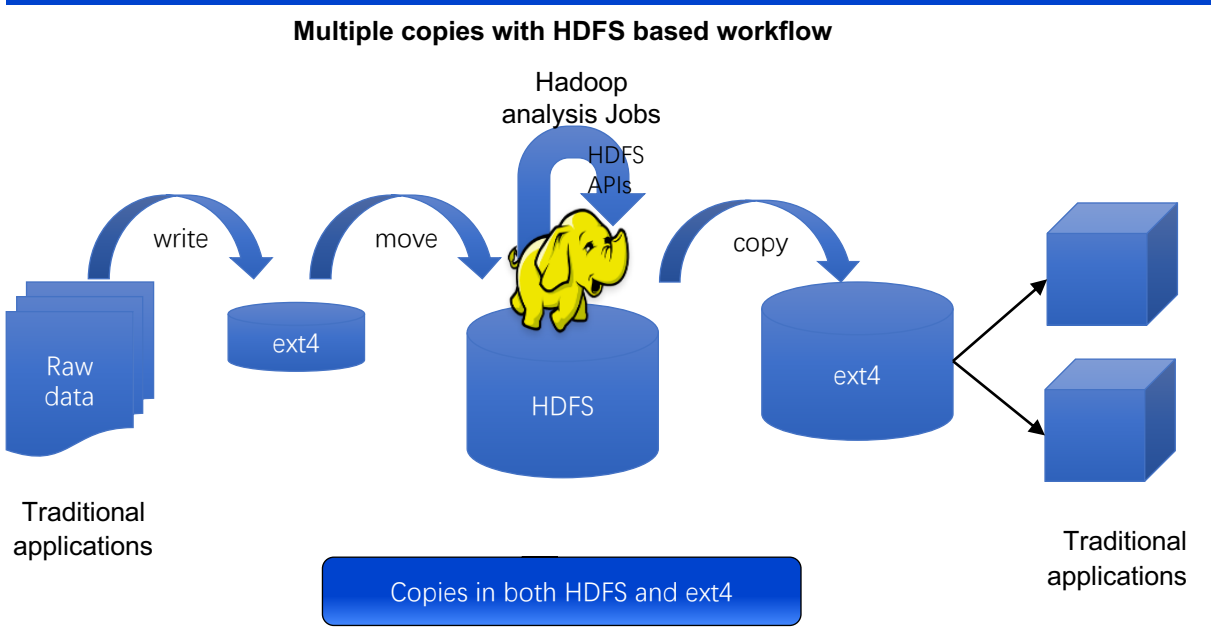
④ 全局命名空间实现数据整合

⑤ 容灾和其它企业级数据管理能力



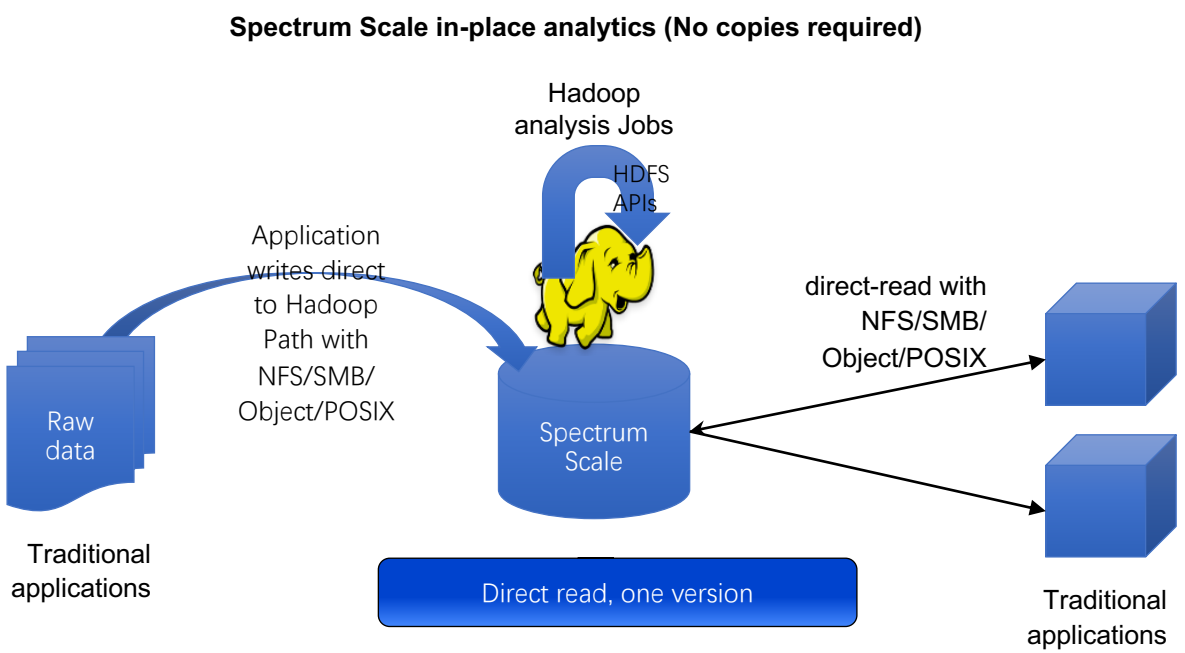
Spectrum Scale 为大数据分析提供优化的存储平台

数据科学家需要花费大量的时间将数据复制到 HDFS



Copy process can take hours/days & eventually results are based on stale data.

Spectrum Scale 支持多种数据访问接口，无需复制



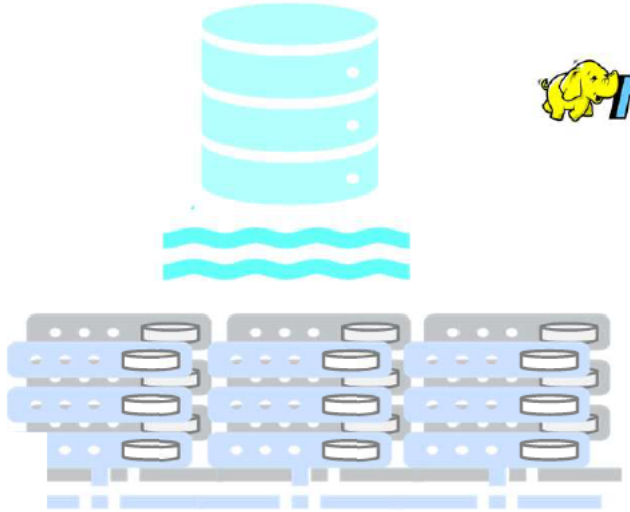
数据保护成本巨大 – HDFS 缺省三副本保护

存放 5PB 数据，HDFS 需要 15PB 裸容量的存储

Spectrum Scale 纠删码仅需 30% 数据保护开销

采用 Spectrum Scale ECE，仅需要 6.5PB 裸容量

整合数据湖



hadoop

"The Hadoop 3.0 spec acknowledges that, at some point, too much cheap becomes expensive" -ZDNet

Santander

60%

finanz informatik

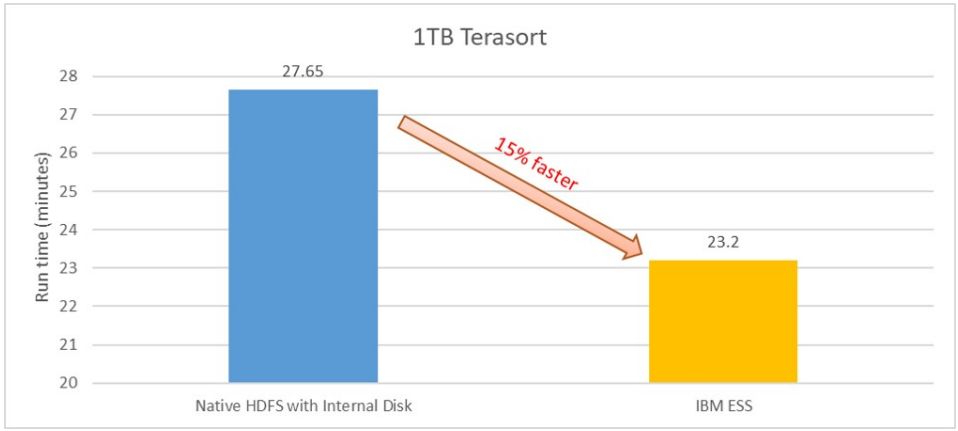
GM

UPMC
University of Pittsburgh Medical Center

ESS

Ford

FNB
First National Bank



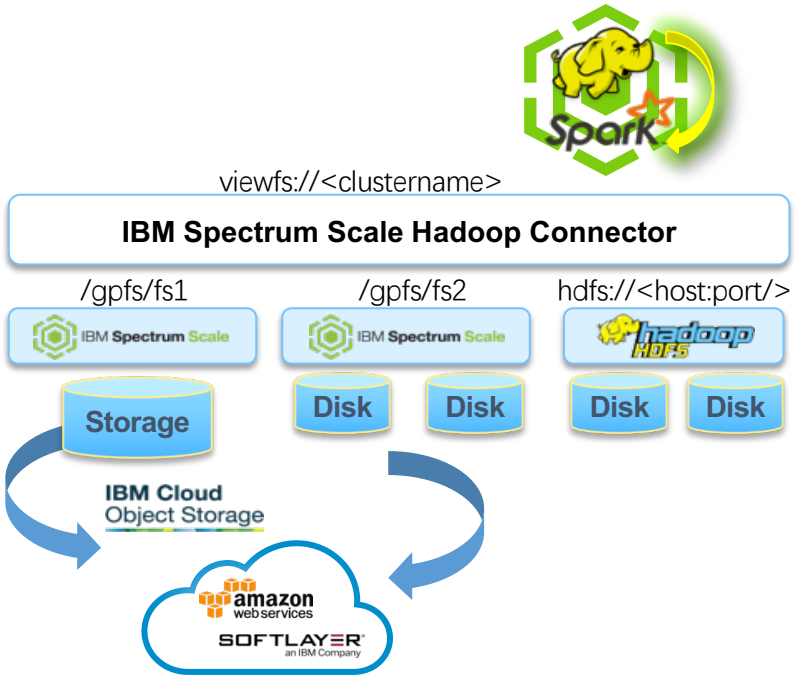
Blog: Hadoop Performance for disaggregated compute and storage configurations based on IBM Spectrum Scale Storage

可用容量提高一倍，性能比HDFS**快15%**

自动分层到磁带或云，节约70%+成本

提供企业级的数据平台

- 同步/异步的容灾
- 备份/归档性能
- 基于策略的数据自动分层
- 不同类型的配额管理
- 审计日志和目录监控



Capability		IBM Spectrum Scale (with HDFS Transparency)	HDFS
In-place analytics for file and object		Yes. All in place with support for POSIX, NFS, SMB, HDFS, and Object with concurrent access. Enables centralized, enterprise-wide data lakes.	Limited support with NFS gateway. No support for SMB, Object, or POSIX.
Performance		Comparable or better performance than HDFS in equivalent hardware configurations.	Same as IBM Spectrum Scale HDFS transparency.
Scalability (maximum number of nodes, files, and data)		IBM Spectrum Scale includes parallel file system architecture that differs from scale-out architecture of HDFS. No single metadata server is in the architecture as a bottleneck. Metadata serving function is distributed across the cluster. Test limit for number of files per file system is 9 billion. IBM Spectrum Scale production deployments are available beyond this test limit.	HDFS can scale up to 350 million files with a single name node because of scale-out architecture limitation. Supports only single or a pair of high availability NameNode, which becomes a bottleneck. Users must use federation functions to overcome this limitation.
If centralized storage is supported, what are the advantages?		Yes. Supports storage area network (SAN)-based shared storage and IBM Elastic Storage Server.	Not supported.
Supports storage-rich server		Yes	Yes
Supports tiering to tape and cloud Object Storage		Yes	No
Data reliability by using replication and erasure coding		Erasure codes from IBM Spectrum Scale RAID in IBM Elastic Storage Server, or data replication from IBM Spectrum Scale.	Support data replication for workload and erasure code for cold data.
Supports enterprise data backup		Yes, with IBM Spectrum Protect™ and Veritas NetBackup.	Does not support IBM Spectrum Protect or Veritas NetBackup.
Supports disaster recovery		Yes, Sync or ASync mode.	Only available for Hbase or Hive.
Supports Remote Direct Memory Access (RDMA)		Yes, when hardware is available.	Not supported.
Improve I/O performance through native client in compute node (Short Circuit Read/Write)		Yes, supports. Can use IBM Spectrum Scale Native Client and high-performance network, such as RDMA over InfiniBand to improve I/O bandwidth and latency and reduce CPU resource.	No native client on compute node.
Security	Secure data at rest	Yes, supports IBM ISKLM and Vormetric key manager and is FIPS-complaint,	Yes
	Secure data in motion	Yes	Yes
	Immutability	Yes	No
	Authentication	Yes	Yes
	Authorization	Yes	Yes
	Auditing	Yes	Yes

Spectrum Scale 大数据平台的应用场景

应用场景	HDFS + 服务器内置盘	IBM Spectrum Scale/ESS 的优势
Spectrum Scale 用作 HDFS 存储	三副本方式增加了用户总体存储成本 需要通过数据复制支持多种数据源和应用，效率低下	✓ 更简单，更便宜 ✓ 减少不必要的成本（30% vs 200% 的数据冗余，减少不必要的数据复制（多协议支持） ✓ 易于管理
HDFS 存储分层/整合	需要更高性能满足 AI 和高性能分析等应用需求 需要控制成本并支持更高灵活性	✓ 易于扩展 ✓ 易于管理 ✓ 降低成本
HDFS 备份	现有架构难以满足合规性或网络安全弹性的需求 数据量大，传统方式不能满足备份窗口的时间需求	✓ 快速备份/恢复数据 ✓ 低成本的数据备份解决方案 ✓ 支持混合多云环境
数据摄取层	难以满足流应用等高速数据摄取的需求 性能和可扩展性均收到限制	✓ 软件定义，灵活扩展性能和容量 ✓ 多协议支持满足不同应用的数据摄取需求
新一代工作负载	难以支持混合多云环境下新一代应用的需求，如容器化的应用	✓ 支持 AI、多云、5G、物联网等新型应用 ✓ 实现网络安全弹性
数据容灾	无法满足企业级应用的容灾需求	✓ 支持各种双活、主备容灾方式 ✓ 经过数十年验证

另可选用一体化的高性能 IBM Elastic Storage System

最大化可用性

- 内置世界领先的Decluster RAID 技术，大幅缩减数据重建时间，故障不影响性能
- 平均无故障时间远远超过硬件生命周期

最大化性能

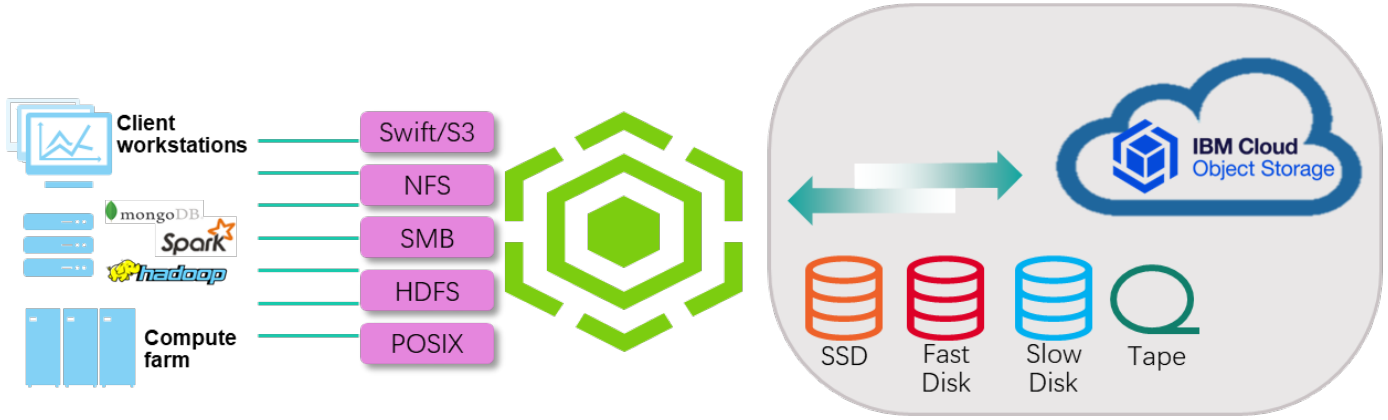
- 提供基于全闪存和磁盘的模块，最大化数据访问性能

领先的数据管理功能

- Spectrum Scale 实现多元化的数据管理
- 按需扩展，可整合其它存储环境

经济性

- 在增强容错性的同时提供更高的可用容量
- 节约空间、能耗，降低总体成本



速度 ← → 容量

IBM Elastic Storage System 3000

2U24 Enclosure
12 or 24 NVMe drives

3rd generation ESS Flash

Up to 40 GB/s*
100% read, InfiniBand

IBM Elastic Storage System 5000

SLx SCx

3rd generation ESS HDD

Up to 55 GB/s*
100% read, InfiniBand

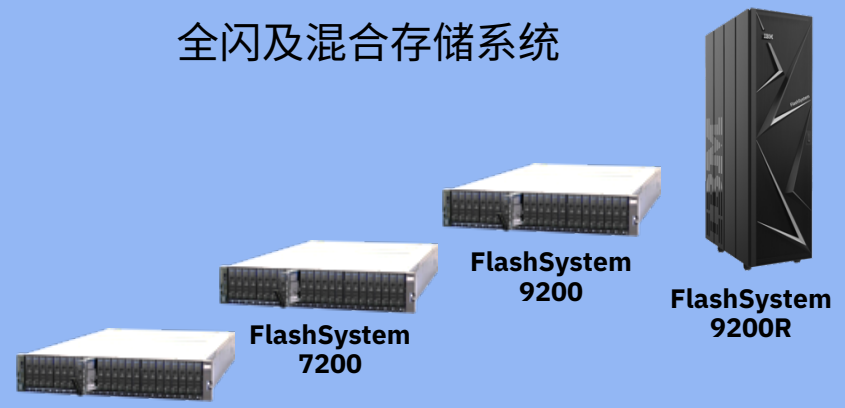
期望有机会为你提供 IBM 全球领先的存储产品和服务

混合多云存储
Storage for Hybrid Multicloud

IBM **Storage Insights**

IBM **Spectrum Virtualize**

全闪及混合存储系统



SVC

Storage for Z

AI 和大数据存储
Storage for AI & Big Data

IBM **Spectrum Discover**

IBM **Spectrum Scale**

IBM Cloud **Object Storage**

Elastic Storage System

Cloud Object Storage

网络

IBM **Spectrum Protect** Suite

数据保护现代化与网络安全
Cyber Resiliency & Modern Data Protection

混合云　快照　磁带　虚拟机　容器

感谢聆听!

您可以访问 IBM 在线实验室 https://csc.cn.ibm.com/labs/portal
(一杯咖啡的时间，体验最新主机,存储真机操作)

或致电热线: 400-669-2039

IBM