Breaking the Silo :
Optimize your Data Pipeline for Analytics and AI

Par Hettinga

IBM Enablement Leader – Unstructured Data

11th March 2019

Session Objectives

To show how IBM Software Defined Storage offerings address data management challenges in Analytics and AI use cases and help customers implement more efficient data pipelines

Content

- Data Management Challenges in Analytics and AI
- IBM Spectrum Storage for Analytics and AI
  - IBM Spectrum Scale
  - IBM Spectrum Discover
  - IBM Cloud Object Storage
- Data Unification using IBM Spectrum Scale
- Data Unification Case Studies
- Summary - IBM Spectrum Storage for AI

Data Management Challenges in Analytics and AI

Biggest Unstructured Data Challenges

Number of enterprises with **1,000 TB+** unstructured data stores grew **3X** from 2016 to 2017

39% of firms see sourcing, gathering, managing & **governing data** as their biggest **challenges** when using systems of insight

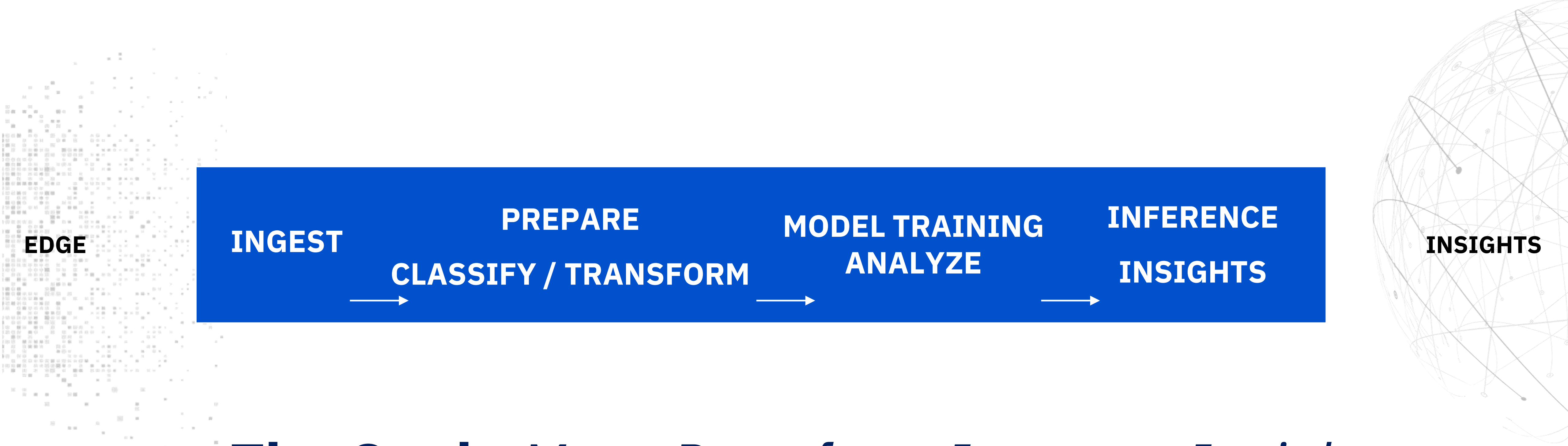
Source: Forrester Analytics, Global Business Technographics Data And Analytics Survey, 2017,
Global Business Technographics Data And Analytics Survey, 2016 (Enterprises with 1000+ employees)

Data Management Challenges in Analytics and AI

- Data ingest and preparation cycle are too time consuming
- Multi-source data aggregation
- Silos of infrastructure for various analytics use cases
- Multiple copies of same data without a single source of truth
- Analytics on stale data
- Need to securely manage and protect data for traceability
- Need for global accessibility and collaboration

IBM Spectrum Storage for Analytics and AI

Analytics and AI Data Pipeline



EDGE

INGEST

PREPARE

CLASSIFY / TRANSFORM

MODEL TRAINING

ANALYZE

INFERENCE

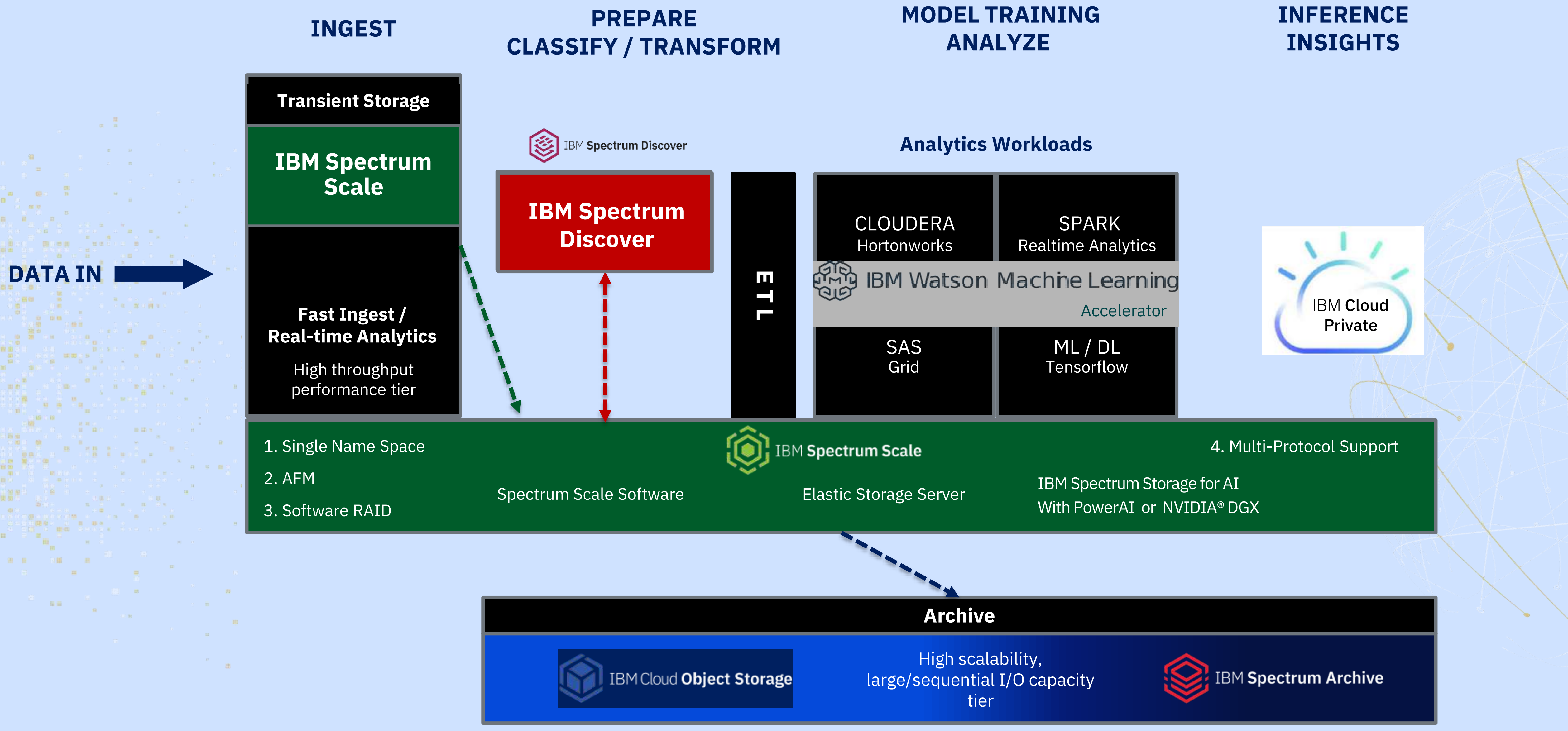
INSIGHTS

INSIGHTS

The Goal: *Move Data from Ingest to Insights*

Analytics and AI Data Pipeline with IBM Storage

The fastest path from ingest to insights



INGEST

PREPARE
CLASSIFY / TRANSFORM

MODEL TRAINING
ANALYZE

INFERENCE
INSIGHTS

DATA IN

Transient Storage

IBM Spectrum Scale

Fast Ingest /
Real-time Analytics

High throughput
performance tier

IBM Spectrum Discover

IBM Spectrum Discover

ETL

Analytics Workloads

CLOUDERA
Hortonworks

SPARK
Realtime Analytics

IBM Watson Machine Learning
Accelerator

SAS
Grid

ML / DL
Tensorflow

IBM Cloud
Private

1. Single Name Space
2. AFM
3. Software RAID

IBM Spectrum Scale

Spectrum Scale Software

Elastic Storage Server

4. Multi-Protocol Support

IBM Spectrum Storage for AI
With PowerAI or NVIDIA® DGX

Archive

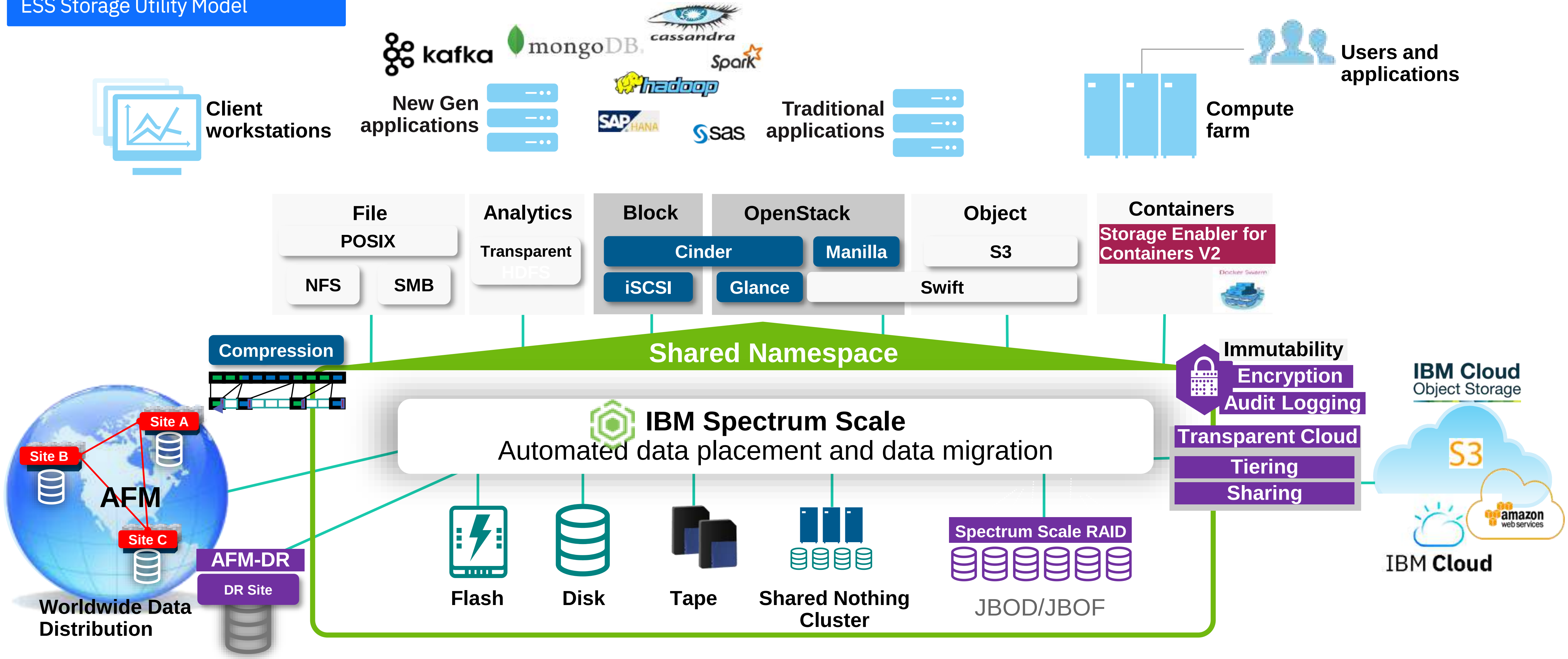
IBM Cloud Object Storage

High scalability,
large/sequential I/O capacity
tier

IBM Spectrum Archive

IBM Spectrum Scale - Unleash Storage Economics on a Global Scale

Licensed Editions
Data Access
Data Management
ESS Storage Utility Model



Consolidate all your unstructured data storage on spectrum scale with unlimited and painless scaling of capacity and performance. 4000+ clients using Spectrum Scale as data plane for Analytics and AI workloads

IBM Spectrum Scale – Parallel Architecture for Performance Scaling

OAK RIDGE
National Laboratory

Summit System

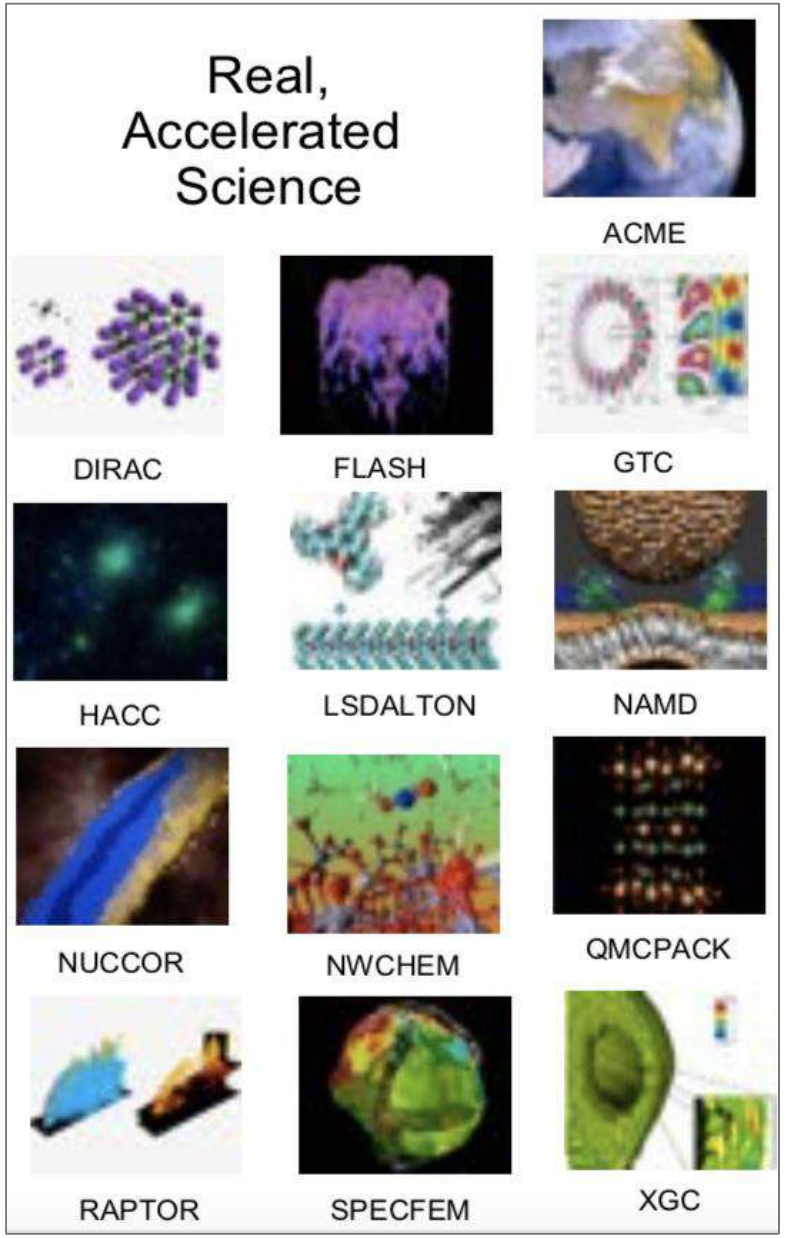
- 4608 nodes, each with:
  - 2 IBM Power9 processors
  - 6 Nvidia Tesla V100 GPUs
  - 608 GB of fast memory
  - 1.6 TB of NVMe memory
- 200 petaflops peak performance for modeling and simulation
- 3.3 ExaOps peak performance for data analytics and AI

2.5 TB/sec

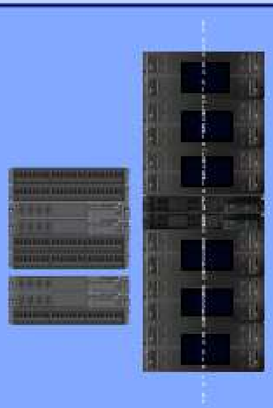
Throughput to storage architecture

250 PB

HDD storage capacity



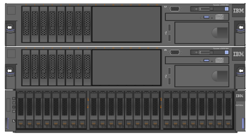
IBM Spectrum Scale offers Deployment Choice

	Type	Software licenses	Hardware	Details
IBM Spectrum Scale	Software *Per TiB license*	▪ Data Access Edition, ▪ Data Management Edition	Bring your own servers, storage, network.	Combine with IBM ESS, or other IBM and other vendor storage/server hardware.
IBM ESS 	Bundled H/W + S/W *Per-drive license, or per-TiB license. Storage Utility Model*	▪ Data Access Edition or ▪ Data Management Edition ▪ Storage Utility Model available for ESS: billed for monthly capacity usage	Bundled servers and storage. Includes IBM SSR software for advanced RAID/erasure coding.	Storage building block. Spectrum Scale based. Add ESS or Spectrum Scale + IBM or other vendor Storage//Server
IBM Cloud amazon web services	Cloud	IBM Cloud AWS: Spectrum Scale Bring-your-own-license on AWS Marketplace	Provided by Cloud vendor	AWS install via catalogue

IBM Spectrum Scale as an Integrated Solution

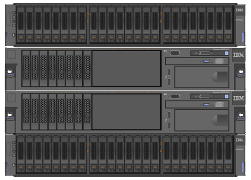
- *Speed*

- *Capacity*

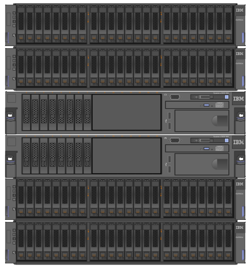


- **Model GS1S**
- 24 SSD

- **14 GB/s**



- **Model GS2S**
- 48 SSD



- **Model GS4S**
- 96 SSD

- **40 GB/s**

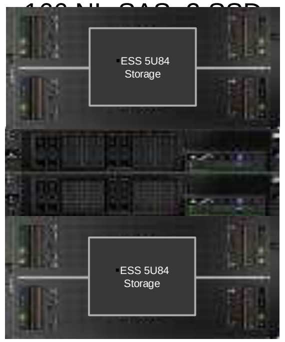
- **Model GH14S:**
- 1 2U24 Enclosure SSD
- 4 5U84 Enclosure HDD
- 334 NL-SAS, 24 SSD

- **Model GH24S:**
- 2 2U24 Enclosure SSD
- 4 5U84 Enclosure HDD
- 334 NL-SAS, 48 SSD

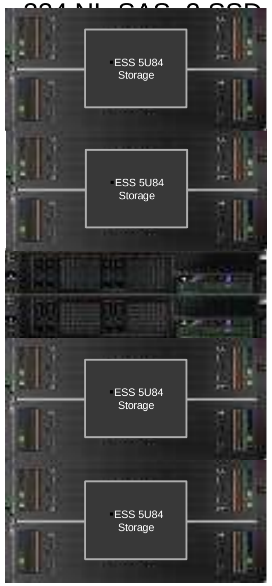
- **Model GL1Sz:**
- 1 Enclosures, 9U

- **6 GB/s**

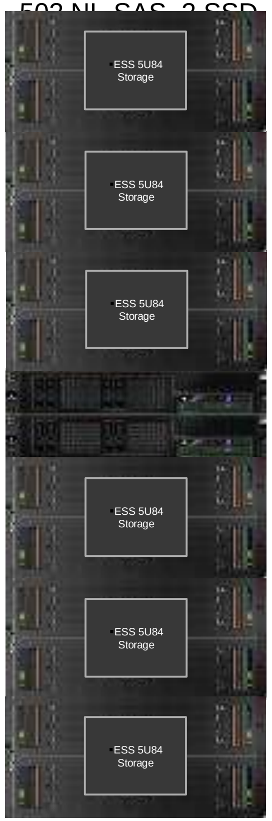
- **Model GL2S:**
- 2 Enclosures, 12U
- 166 NL-SAS, 2 SSD



- **Model GL4S:**
- 4 Enclosures, 20U
- 334 NL-SAS, 2 SSD



- **Model GL6S:**
- 6 Enclosures, 28U
- 502 NL-SAS, 2 SSD



Why IBM Spectrum Scale for Analytics/AI workloads ?

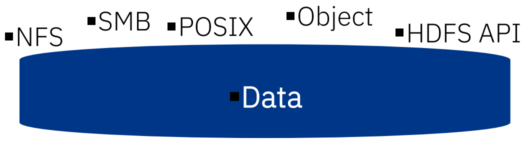
Unmatched Scalability and Performance with the most optimized storage footprint

Performance leadership in AI benchmarks

40GB/s and 300TB in 2U*, Linear scaling of 120GB/s in 6U*

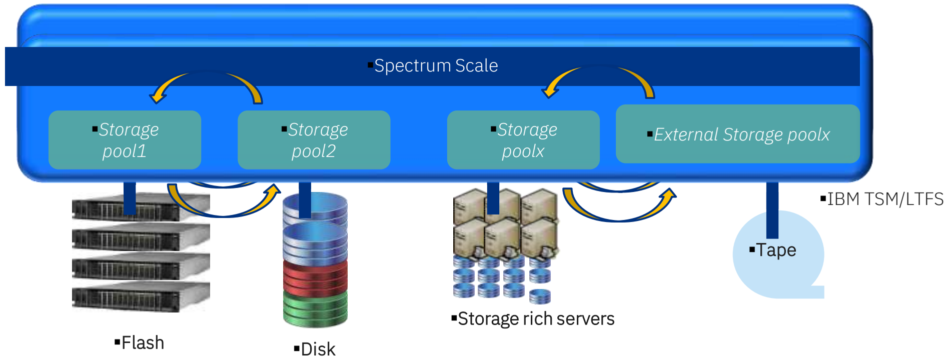
- * With Spectrum Scale NVMe appliance – PDF document

Reduce datacenter with in-place analytics



- Access to the data using any of the industry standard protocols.
- No need to maintain separate copies for different applications.

Full Data Life Cycle Management



- Policy based auto tiering between storage pools

Extreme scalability with parallel file system

Data + Metadata Node	Data + Metadata Node	Data + Metadata Node	Data + Metadata Node	..

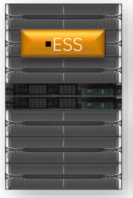
- Scale to billions of files. No centralized metadata node bottleneck.

Flexible storage architectures

IBM Spectrum Scale

- Install SW in hyperconverged mode

- OR



- in Shared storage mode

- Support for flexible and hybrid architectures under common namespace. Enabled for running containerized workloads.

Global namespace that spans geographies

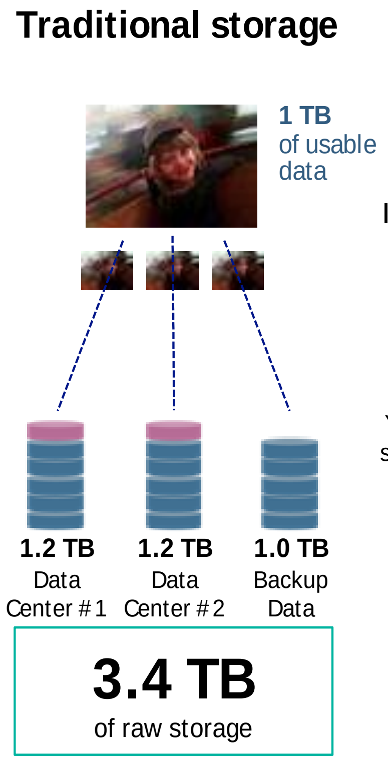
Stretch clusters and Active – Active replicas of data for real time global collaboration

IBM Cloud Object Storage – #1 Object Store by IDC 2018

- **Flexible for any app**
  - Use On Premise, Managed Cloud or Hybrid Cloud
  - Use as a Service - Dedicated or Public
  - Deploy to both traditional and native Cloud applications
  - Provides Active Archive and Cold tier
  - Global ingest capability
- **Client proven enterprise scale**
  - Shared nothing architecture, with strong consistency
  - Scalable namespace mapping with no centralized metadata
  - Highly reliable and available with replication
  - Distributed rebuilder to maintain consistency
  - Distributed collection and storage of statistics needed for management
  - APIs for integration with external management applications
  - Automated network installation
- **Simplicity delivers big advantage**
  - Manages all storage from a single pane of glass with zero down time – on-premises, in the cloud or both
  - Uses fewer administrative resources than traditional storage
  - Requires no extra management for storage high availability, backup or disaster recovery

IBM Cloud Object Storage information dispersal

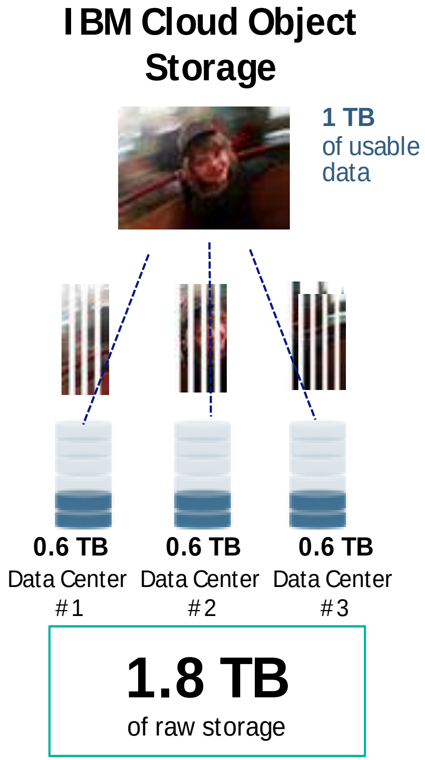
Redefining availability and economics of data storage



IBM Cloud Object Storage requires less than half the storage and 70% lower TCO*.

You can lose a disk, a server or even a whole site due to failure or disaster, and still quickly recover 100% of your data.

Slices are distributed geographically for durability and availability.



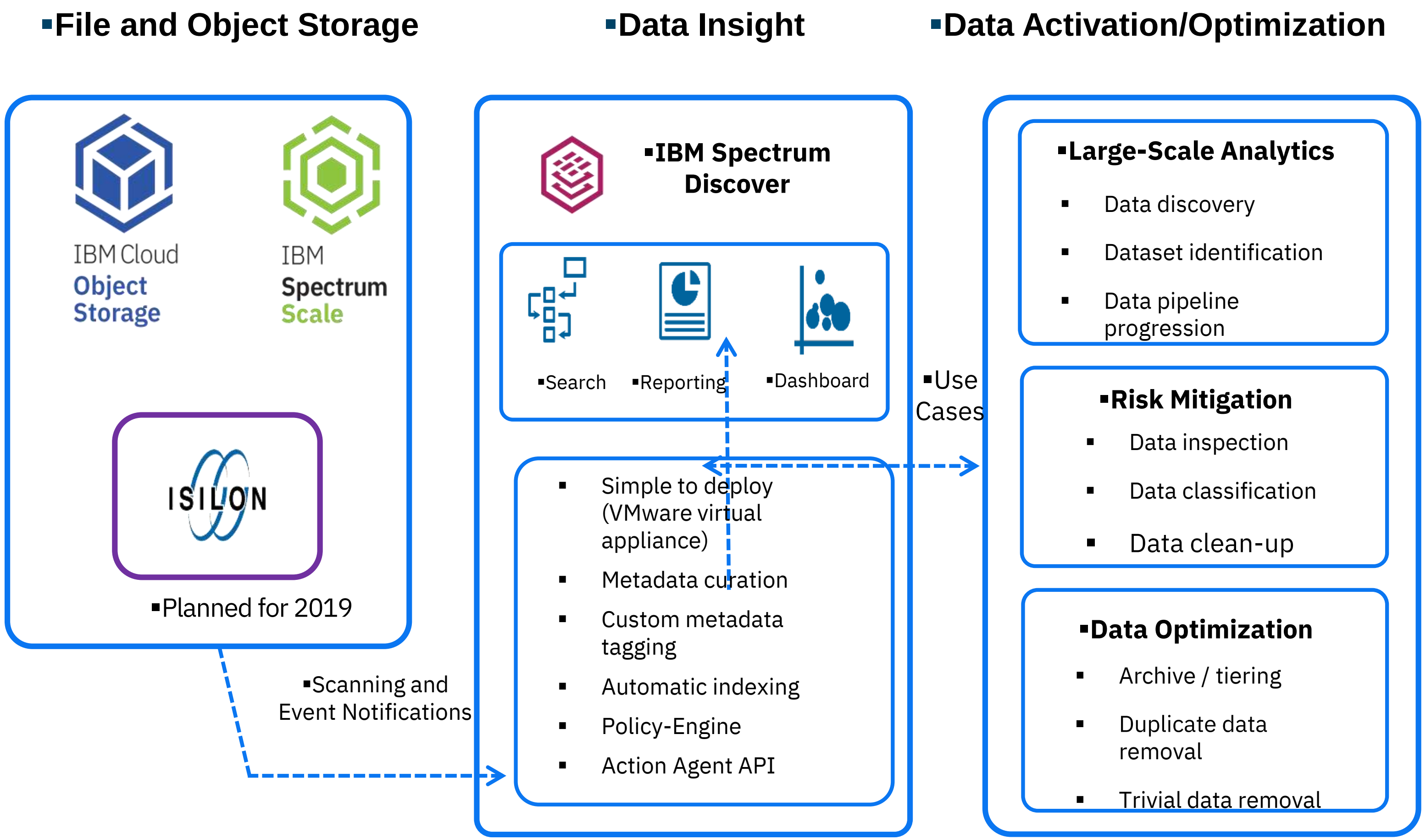
Traditional storage requires 3.4 TBs raw storage capacity for 1 TB of usable storage.

Our object storage requires only 1.8 TBs raw storage capacity for 1 TB of usable storage.

IBM Spectrum Discover Overview

Data Insight for Analytics, Governance, & Optimization

- Automate cataloging of unstructured data by capturing metadata as it is created
- Enable comprehensive insight by combining system metadata with custom tags to increase storage admin & data consumer productivity
- Leverage extensibility using the API, custom tags, and policy-based workflows to orchestrate content inspection & activate data in AI, ML, & analytics workflows



- File and Object Storage

IBM Cloud Object Storage

IBM Spectrum Scale

ISILON

- Planned for 2019

- Scanning and Event Notifications

- Data Insight

- IBM Spectrum Discover

- Search
- Reporting
- Dashboard

- Simple to deploy (VMware virtual appliance)
- Metadata curation
- Custom metadata tagging
- Automatic indexing
- Policy-Engine
- Action Agent API

- Use Cases

- Data Activation/Optimization

- Large-Scale Analytics
  - Data discovery
  - Dataset identification
  - Data pipeline progression
- Risk Mitigation
  - Data inspection
  - Data classification
  - Data clean-up
- Data Optimization
  - Archive / tiering
  - Duplicate data removal
  - Trivial data removal

Data Unification with IBM Spectrum Scale

Common data layer that can be accessed by multiple applications

Build more efficient workflow / pipeline

Improve data governance

Reduce storage footprint

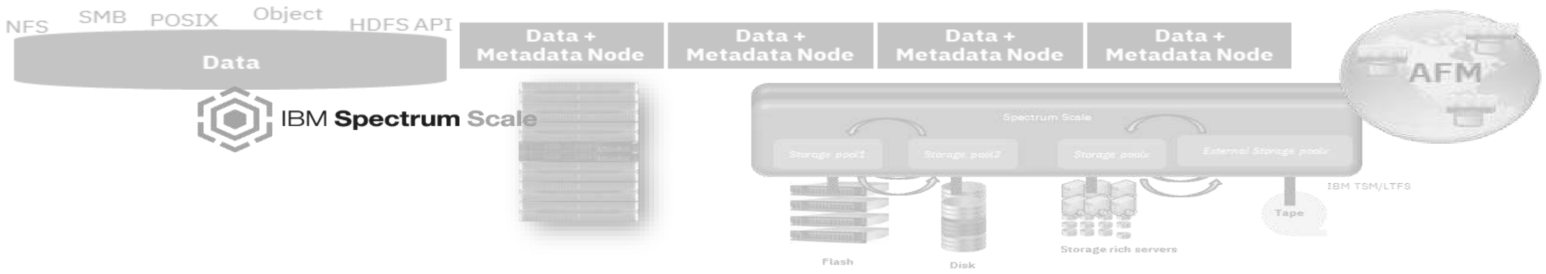
Data Unification with IBM Spectrum Scale

EDGE

INGEST → PREPARE CLASSIFY / TRANSFORM → MODEL TRAINING ANALYZE → INFERENCE INSIGHTS

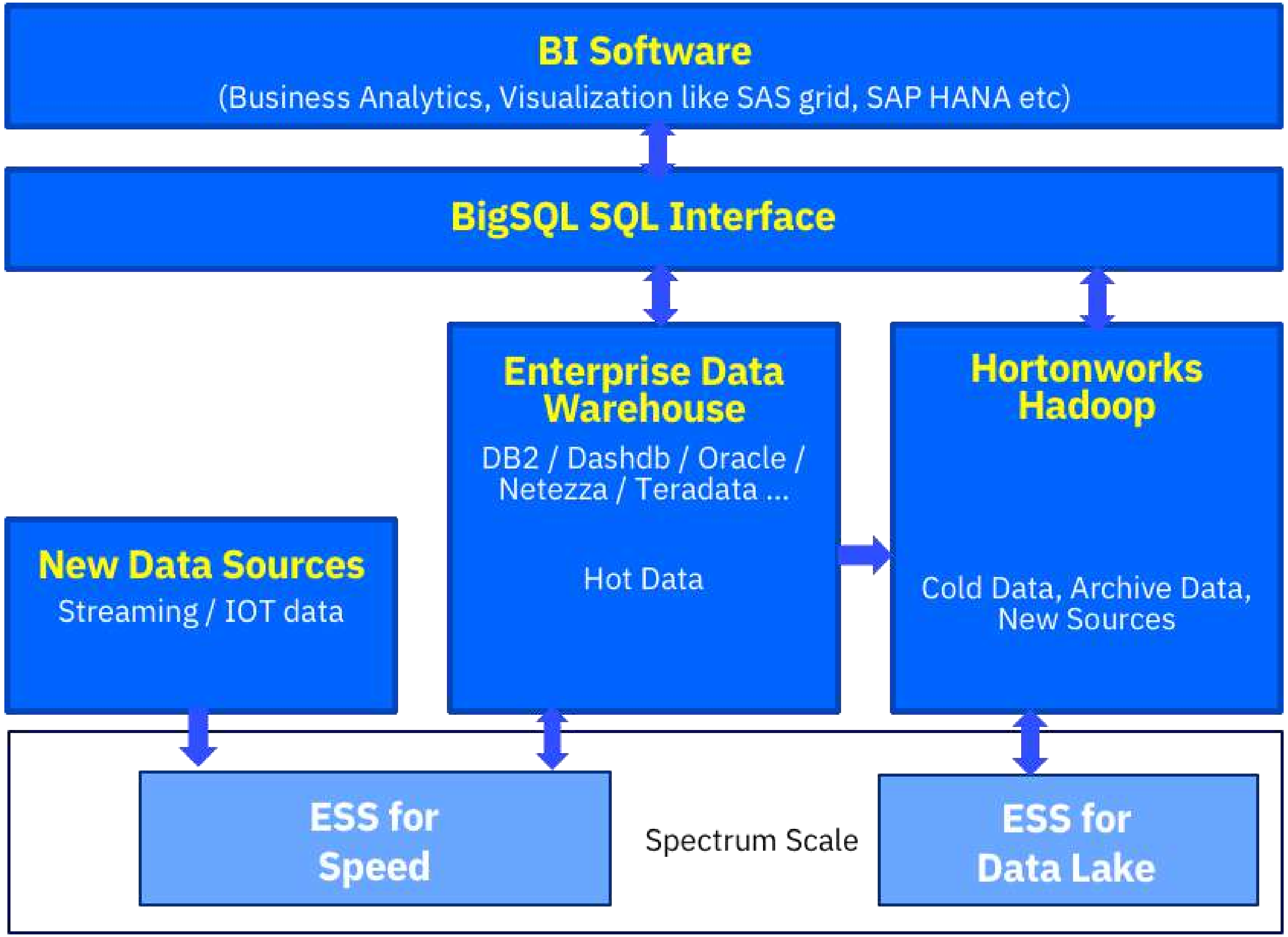
INSIGHTS

← **IBM Spectrum Scale Namespace** →



Data Unification Case Studies

EDW Optimization
Simplify data management using common storage between EDW and Hadoop


BI Software
(Business Analytics, Visualization like SAS grid, SAP HANA etc)
BigSQL SQL Interface
Enterprise Data Warehouse
DB2 / Dashdb / Oracle / Netezza / Teradata ...
Hot Data
Hortonworks Hadoop
Cold Data, Archive Data, New Sources
New Data Sources
Streaming / IOT data
ESS for Speed
Spectrum Scale
ESS for Data Lake

Archive Data away from EDW

- Move cold or rarely used data to Hadoop as active archive
- Store more of data longer

Offload costly ETL process

- Free your EDW to perform high-value functions like analytics & operations, not ETL
- Use Hadoop for advanced ETL

Optimize the value of your EDW

- Use Hadoop to refine new data sources, such as web and machine data for new analytical context

Control cluster sprawl

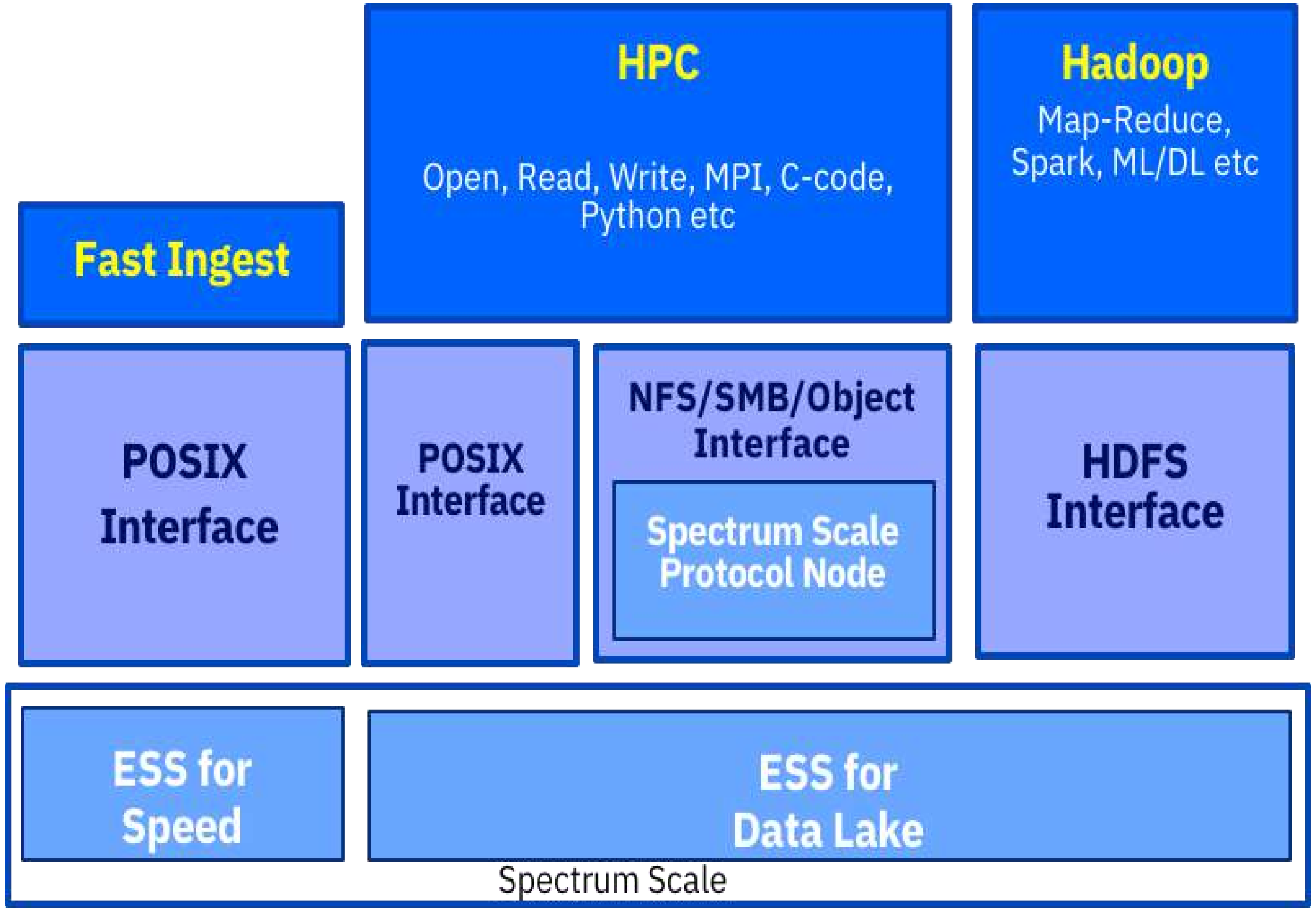
- Grow storage independent of compute with ESS
- POWER servers deliver 1.7x throughput compared to Hortonworks on x86
- Up-to 60% less storage footprint

Reduce migration effort & skillset gap

- Use existing investment in Oracle/DB2/Netezza skills
- BigSQL allows you to migrate applications without major code rewrites and additional SQL development

A Financial Services company in Europe is optimizing their DB2 warehouse using Hortonworks Hadoop; and is using ESS as the common storage behind DB2 and Hadoop.

Integrated HPC and Hadoop
Efficiently transform data into insights with single data lake for HPC & Hadoop



Extend HPC to add modern analytics capabilities

- Efficient movement of data between modern and traditional applications with common namespace
- Spectrum Scale in-place analytics capabilities enable accessing the same data using NFS/SMB/Object/POSIX/HDFS without requiring any modifications to the data
- Improve data reliability and governance with single data lake

Ingest fast and improve time to insight

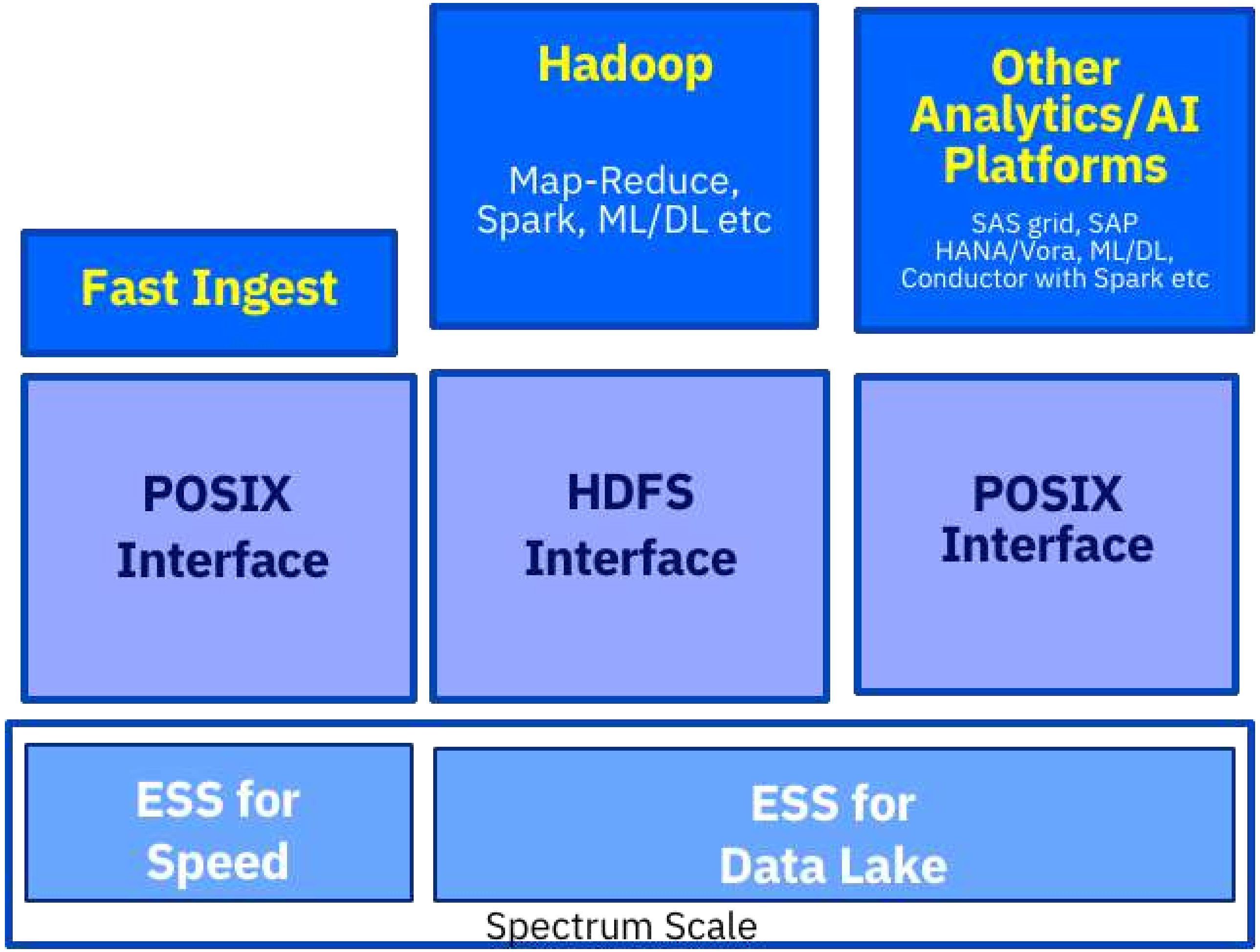
- POSIX interface combined with ESS Flash storage gives super fast ingest ability
- Common namespace enables running some edge analytics at the ingest layer as well

Control cluster sprawl

- Grow storage independent of compute with ESS
- Up-to 60% less storage footprint
- POWER servers deliver 1.7x throughput compar

NASA and a Healthcare company from middle east are using common Spectrum Scale data lake to efficiently get insights using traditional HPC and Hadoop analytics.

Unified Analytics/AI Workflows
Single data lake for Hadoop and non-Hadoop workloads


Fast Ingest
Hadoop
Map-Reduce, Spark, ML/DL etc
Other Analytics/AI Platforms
SAS grid, SAP HANA/Vora, ML/DL, Conductor with Spark etc
POSIX Interface
HDFS Interface
POSIX Interface
ESS for Speed
ESS for Data Lake
Spectrum Scale

All analytics workflows on common storage

- Improve data reliability and governance with single data lake for Hadoop and non-Hadoop analytics setups
- Build ML/DL workflows that use multiple analytics platforms
- Share data across analytics workflows as appropriate

Ingest fast and improve time to insight

- POSIX interface combined with ESS Flash storage gives super fast ingest ability

Control cluster sprawl

- Grow storage independent of compute with ESS
- Up-to 60% less storage footprint
- POWER servers deliver 1.7x throughput compared to Hortonworks on x86

A bank in South Africa is implementing HDP and SAS grid software on a common ESS based infrastructure.

Summary – IBM Spectrum Storage for AI

IBM Spectrum Storage for AI supercharges your AI data pipeline with **storage solutions optimized for the unique demands of AI**.

Integrating industry-leading servers, ISV / open source software and IBM software-defined storage, IBM Spectrum Storage for AI delivers simplified deployment, groundbreaking performance, and extended data management to drive developer productivity with the fastest path to insights.

Power Systems

NVIDIA

Hortonworks

IBM Spectrum Storage for AI – Available Solutions

https://www.ibm.com/it-infrastructure/storage/ai-infrastructure

- IBM Spectrum Storage for Hadoop/Spark workloads
  - IBM Spectrum Scale and Hortonworks/Cloudera Integration
  - IBM Spectrum Scale and IBM Spectrum Conductor for Spark Integration
- IBM Spectrum Storage for AI with NVIDIA DGX
  - IBM Spectrum Scale and NVIDIA DGX Reference Architecture
- IBM Spectrum Storage for AI with Power Systems
  - IBM Spectrum Scale and Power AC922 Reference Architecture
- IBM Spectrum Connect – Storage Enabler for Containers
- IBM Spectrum Storage for AI in Autonomous Driving

Contacts

Pallavi Galgali
IBM Offering Manager – Storage Solutions for Analytics / AI
pgalgali@us.ibm.com
+1-914-433-9882

Par Hettinga
IBM Enablement Leader – Unstructured Data
par@nl.ibm.com
+31-(0)6-53359940

Christopher Maestas
IBM Senior Architect Spectrum Scale
cdmaestas@us.ibm.com
+1-505-321-8636