IBM Spectrum Scale: Storage for AI

Simon Lorenz
Architect Big Data and Analytics

Storage for AI
Outline

- Customer Demands / Requirements
- Using IBM Spectrum Scale built-in Caching technology
- High Performance Tier (HPT)
- Data Accelerator for AI and Analytics (DA^3)
- Demo:
  AI Example Use Case: Detect Wildfires

Customer Demands / Requirements

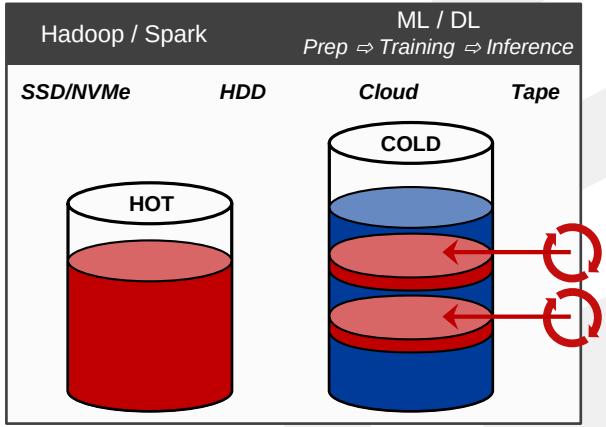
- Ease Data Scientist's and Storage Admins workflows:

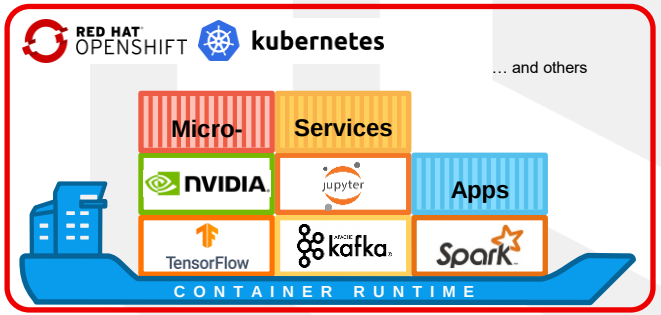
  „Dean, a data scientist, can find and prefetch the data he needs for his Machine Learning jobs without any knowledge about Spectrum Scale"

  - Lots of data! what do I need for my jobs, where is it placed?
  - How can I automate analytics on new ingested data?
  - How do I ensure my limited high performance storage does not run out of space?
  - How can I have multiple Data Scientist's work in parallel on my system?
  - How do I ensure my data in the high performance tier stays in sync with the capacity tier?
  - How can I easily / efficiently provision, test, deploy and scale my containerized workloads?

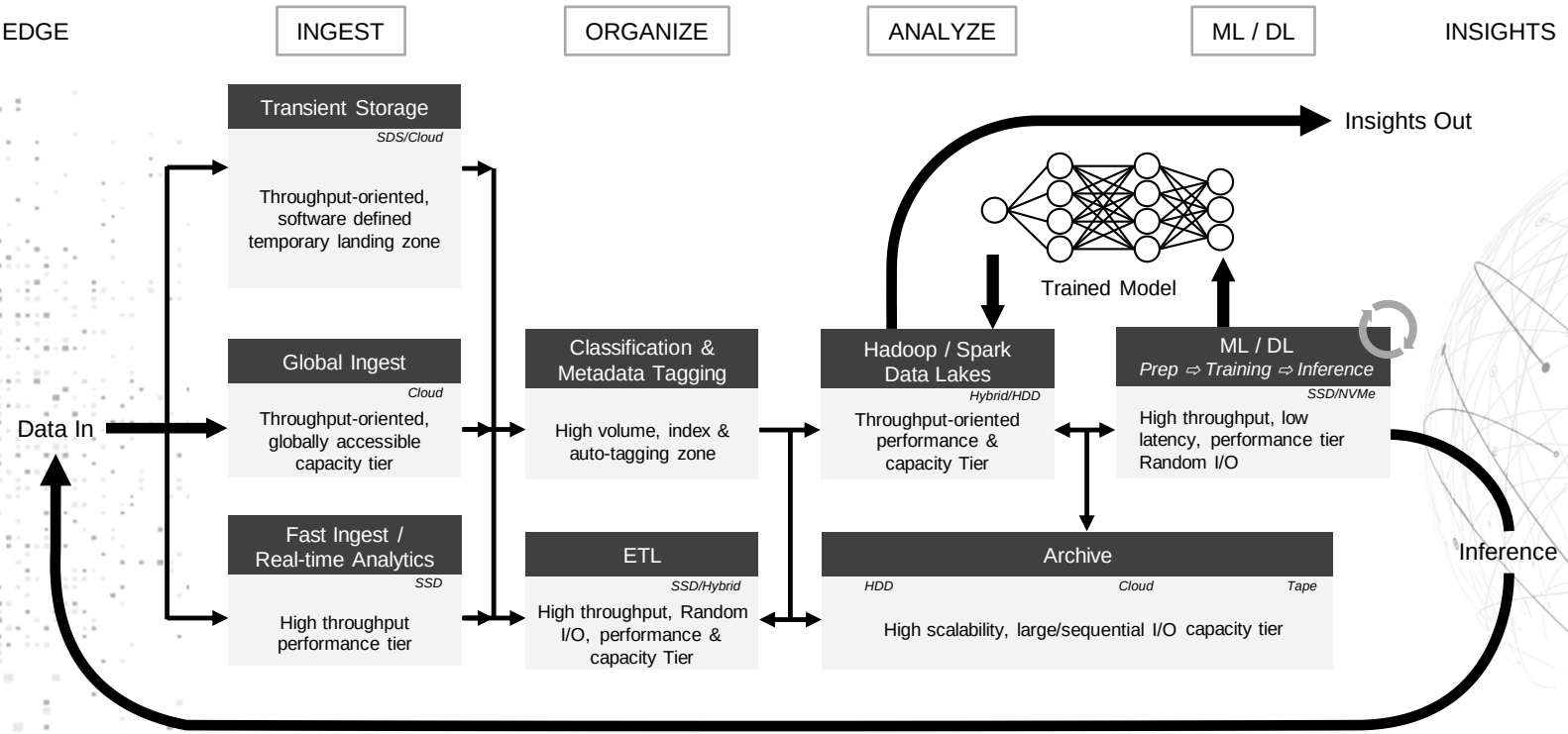
Customer Demands / Requirements

- Customers across all Industries are creating large PB to EB data stores.
- Vast majority of data is relatively cold, but still required for periodic trend analysis.
- Workloads more and more move into containers
- AI / Analytics requires high performance, low latency storage to keep expensive CPU / GPU / TPU / FPGA busy.





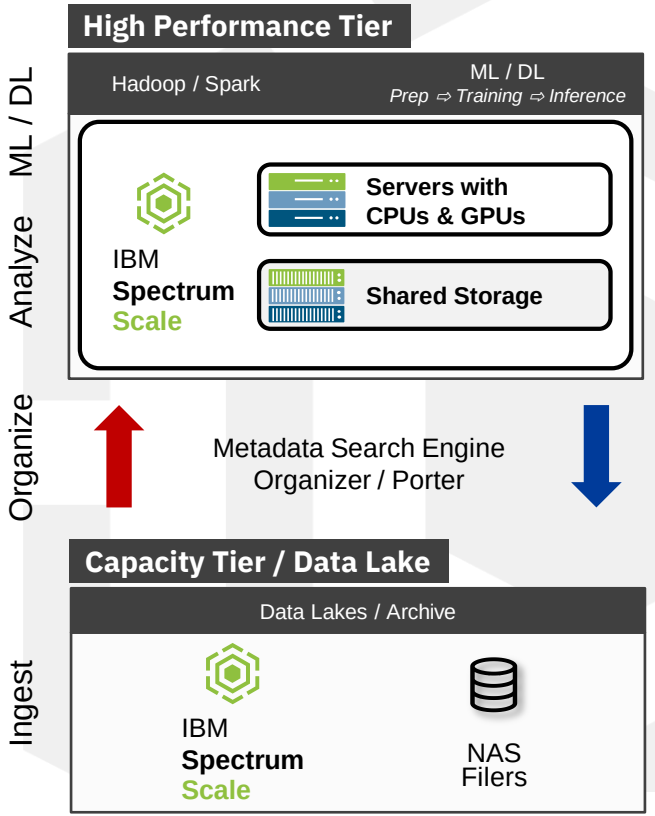
Enterprise Data Pipeline with IBM Spectrum Storage



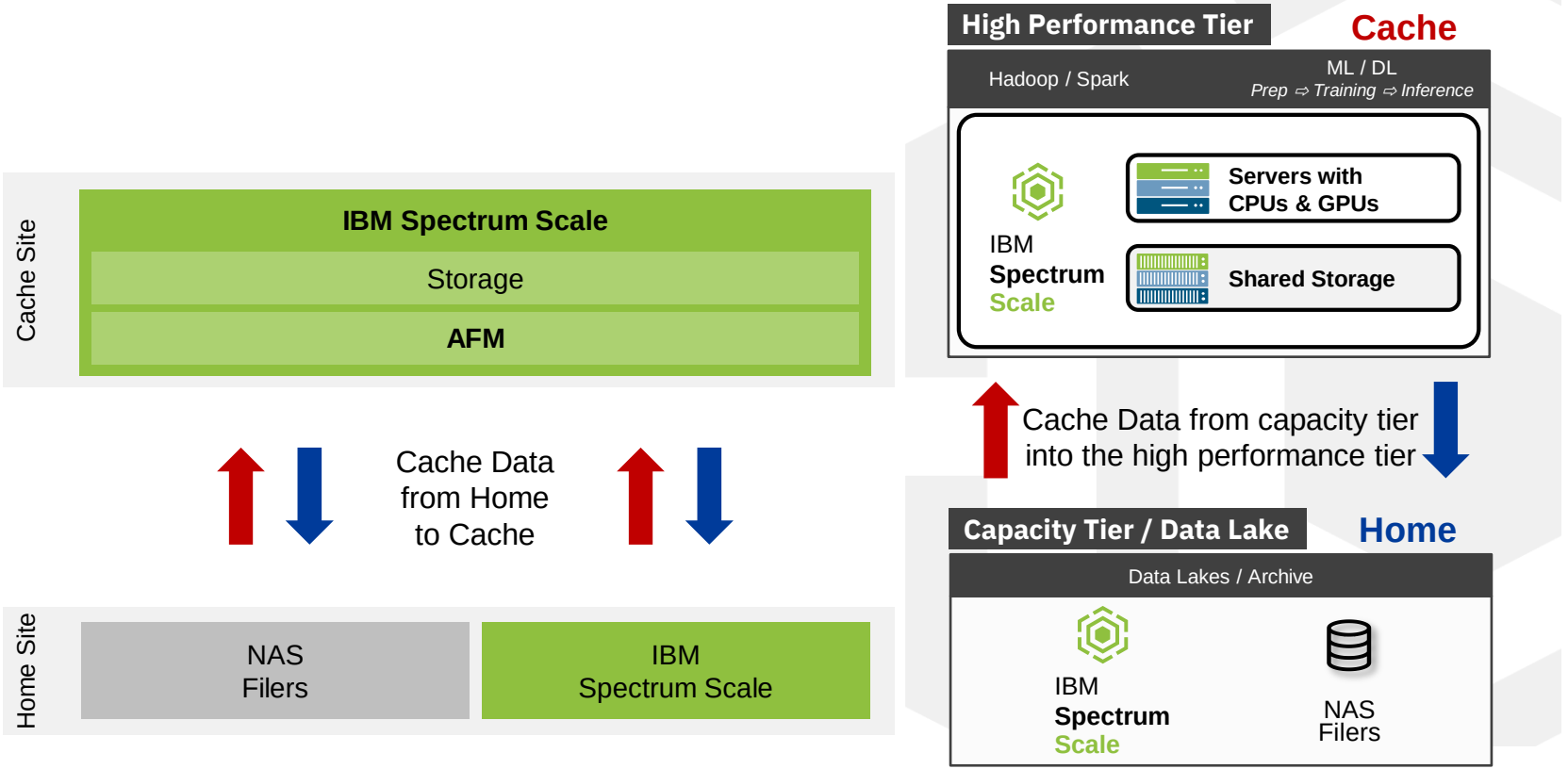
Performance vs. Capacity Storage

- **Performance Tier**
  - **AI and Analytics with high performing Accelerators**
  - **Maximize performance** of storage: **$/IOP & $/GB/s are key**
    - Low latency random I/O & High bandwidth sequential
    - Relatively small compared to Capacity tier (say 5-25%)
    - Can be Lower Durability, Lower Availability, Lower Reliability, if Architected properly
    - No Geo-distribution

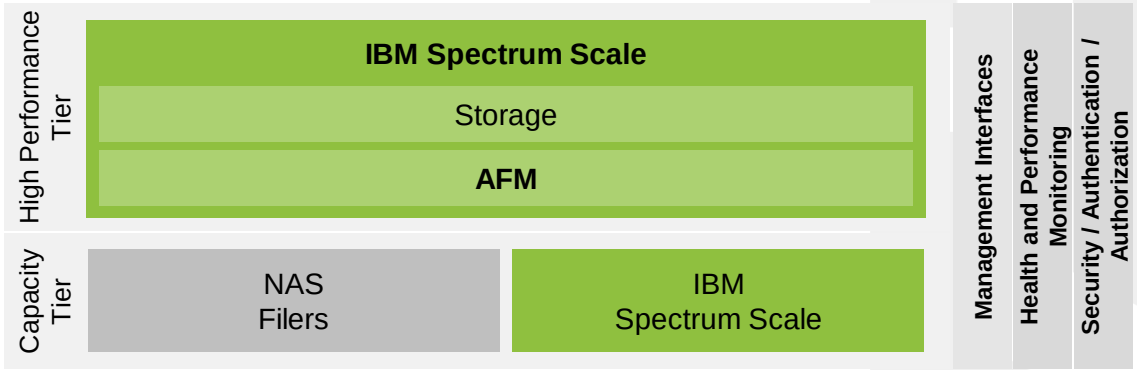
- **Capacity Tier** (aka “Data Lake”)
  - Data usually gets ingested here, large volumes
  - **Minimize** the cost of storage: **$/TB is key**
    - High Durability, Availability, Reliability, Geo-distribution



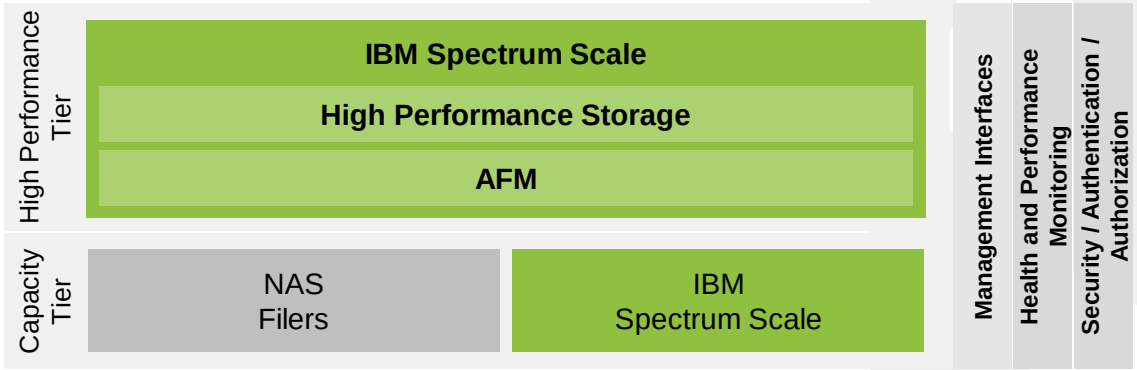
Using IBM Spectrum Scale built-in Caching technology



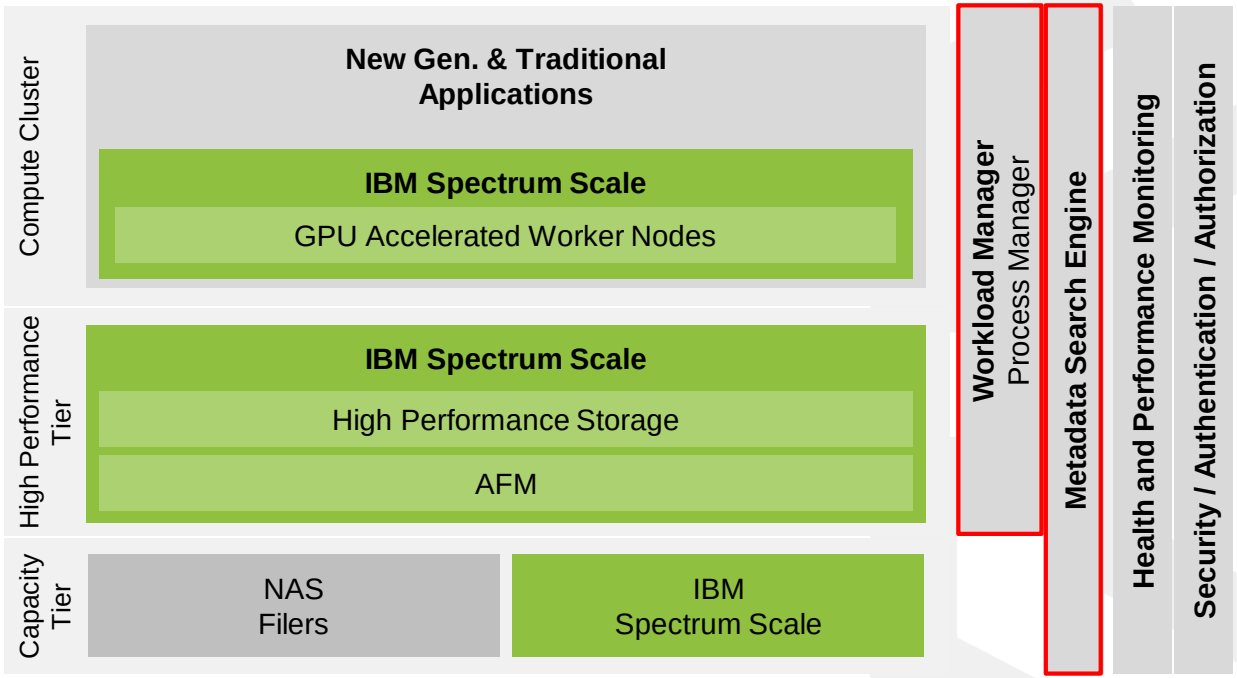
Using IBM Spectrum Scale built-in Caching technology



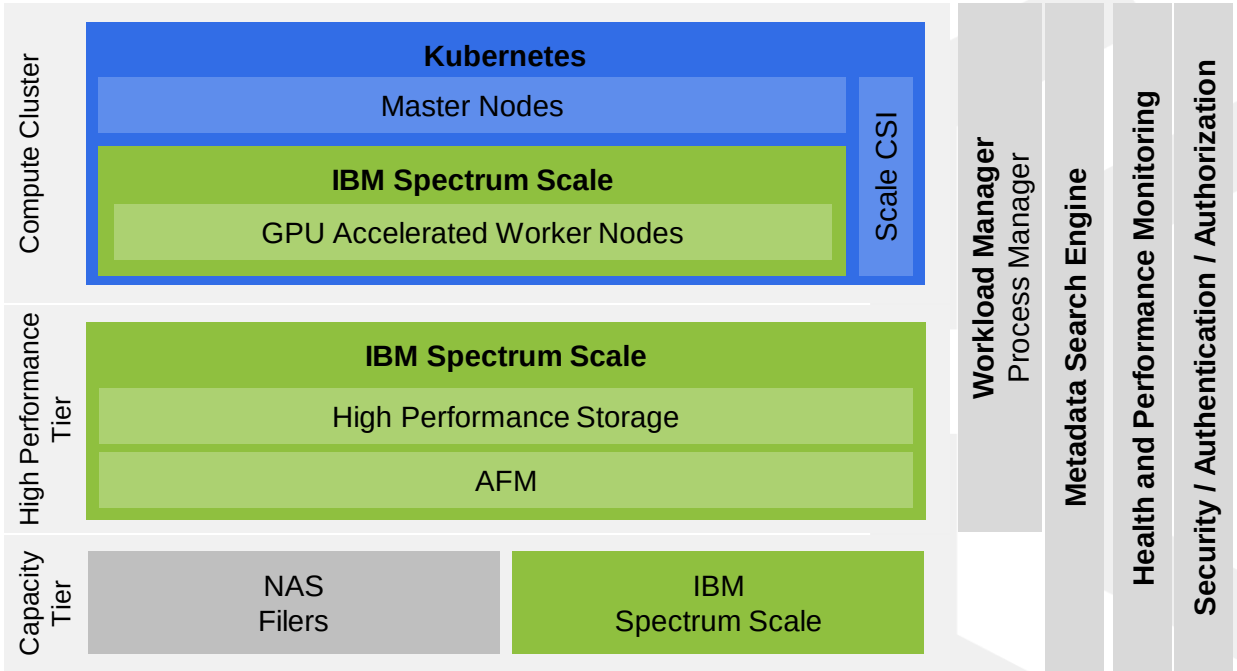
High Performance Tier (HPT) Technology



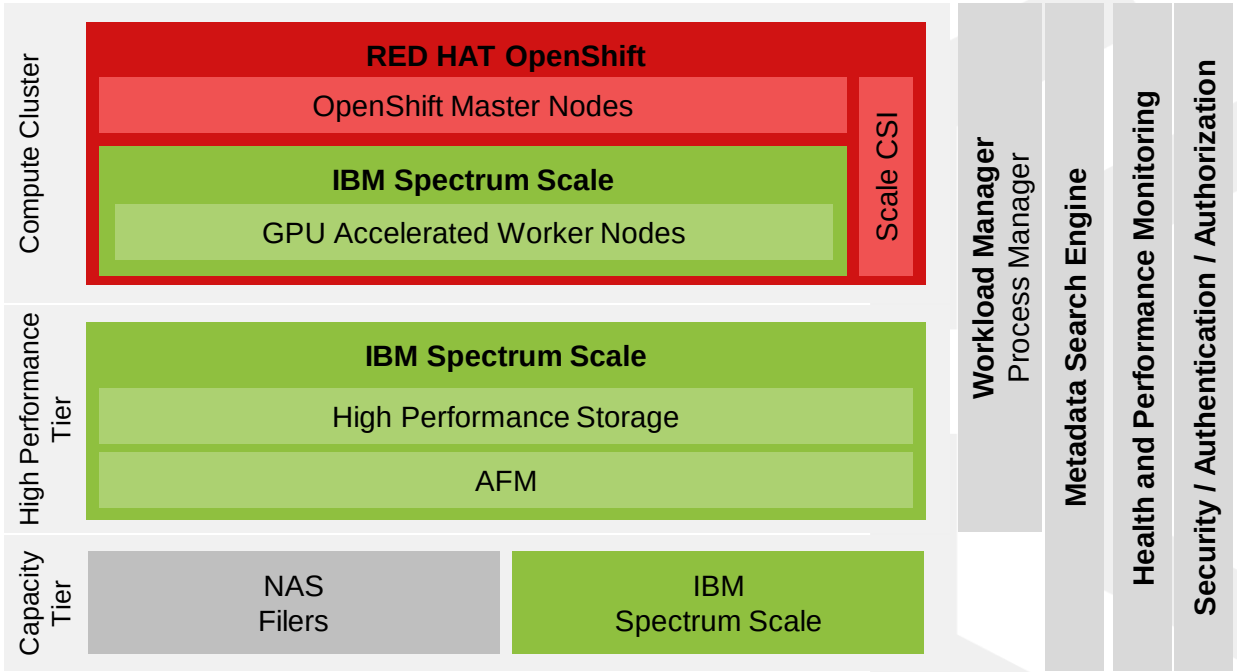
Data Accelerator for AI and Analytics (DA^3) Solution



Data Accelerator for AI and Analytics (DA^3) Solution

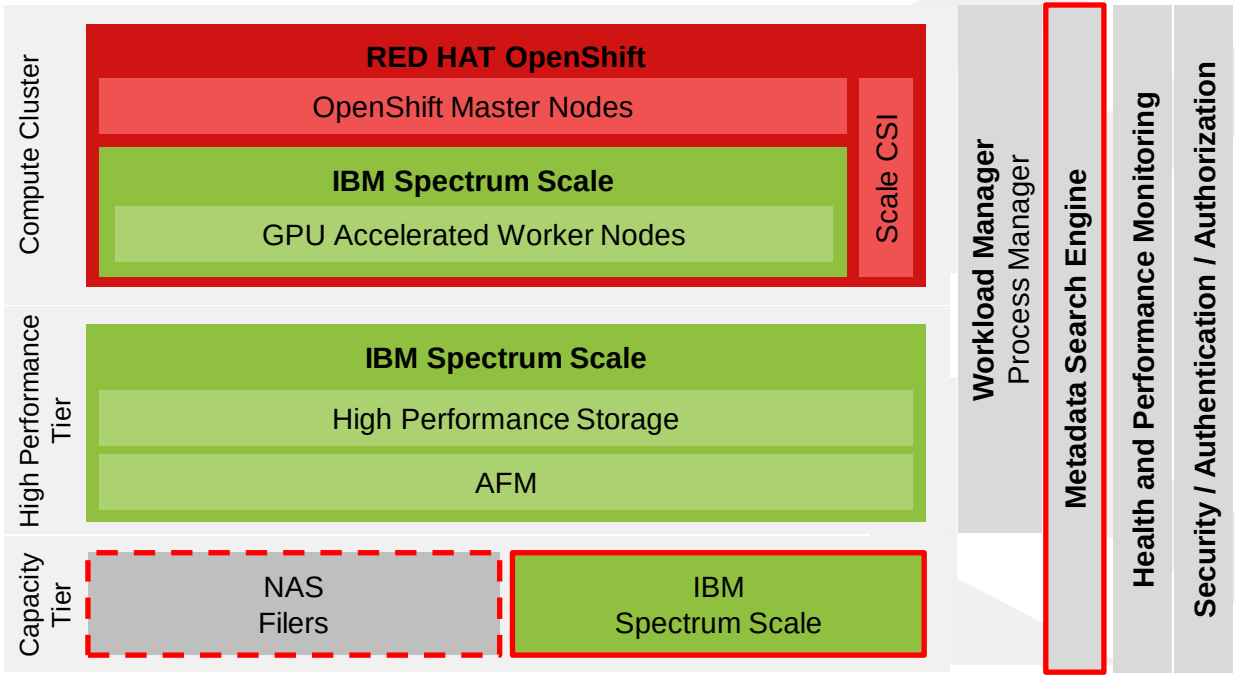


Data Accelerator for AI and Analytics (DA^3) Solution



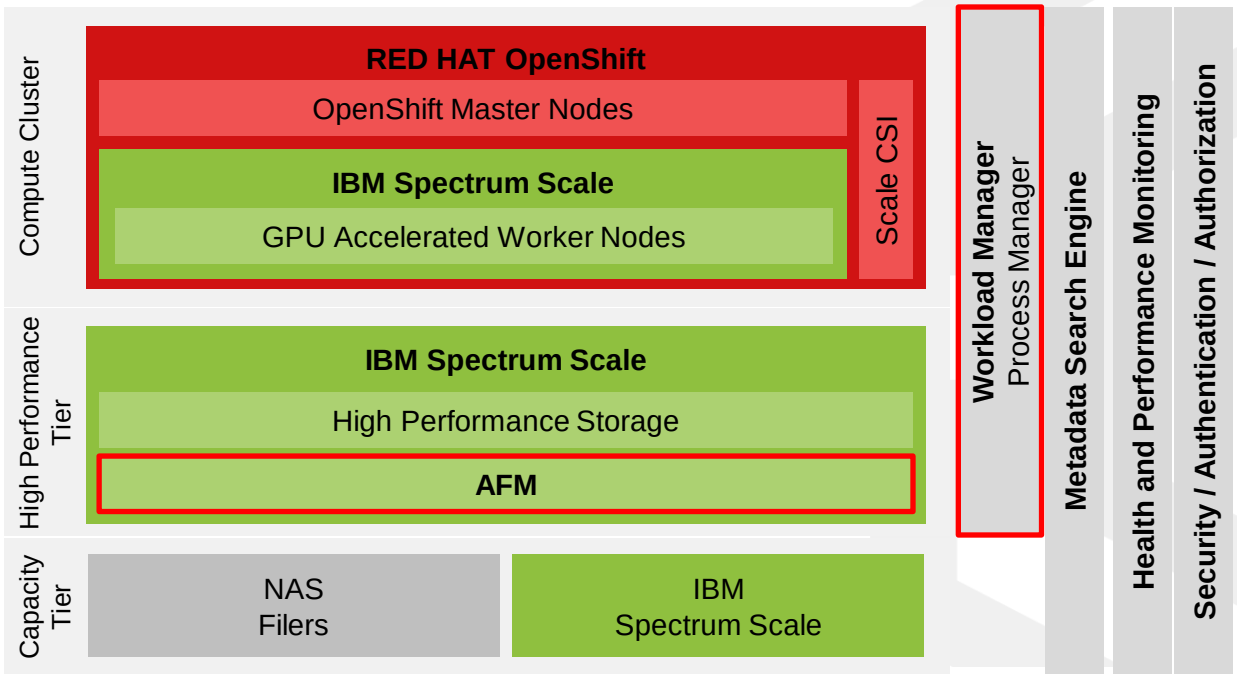
Data Accelerator for AI and Analytics (DA^3)

Lots of data! what do I need for my jobs, where is it placed?
How can I automate analytics on new ingested data?



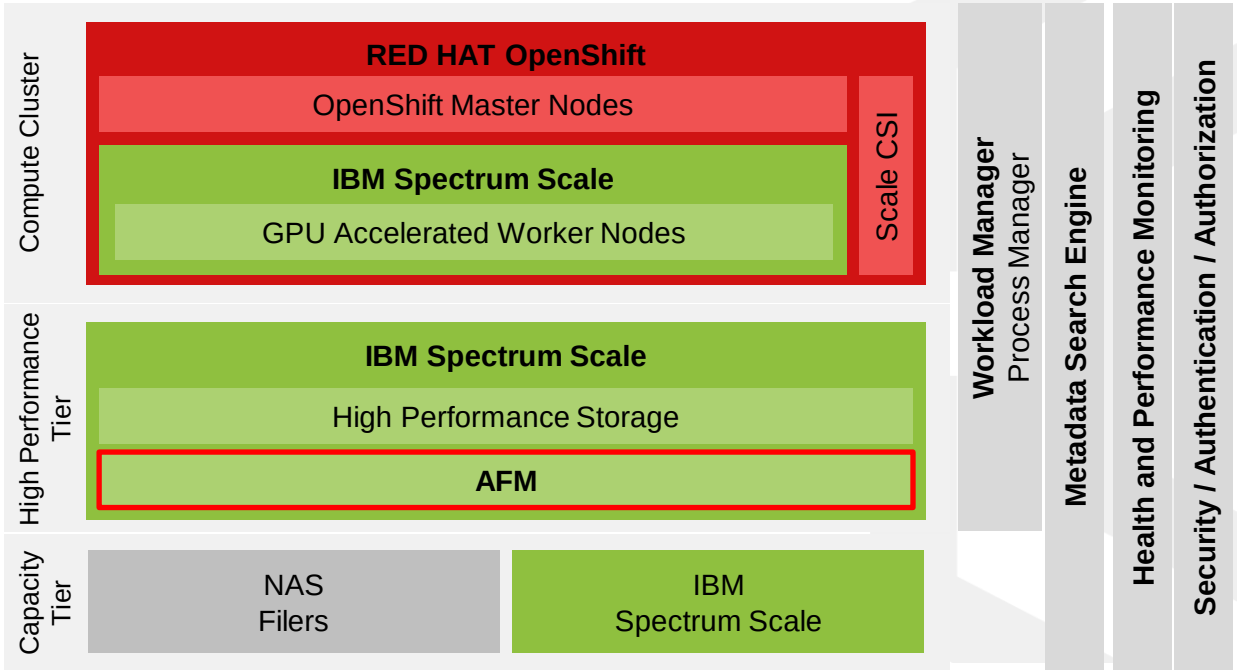
Data Accelerator for AI and Analytics (DA^3)

How do I ensure my limited high performance storage does not run out of space?
How can I have multiple Data Scientist's work in parallel on my system?



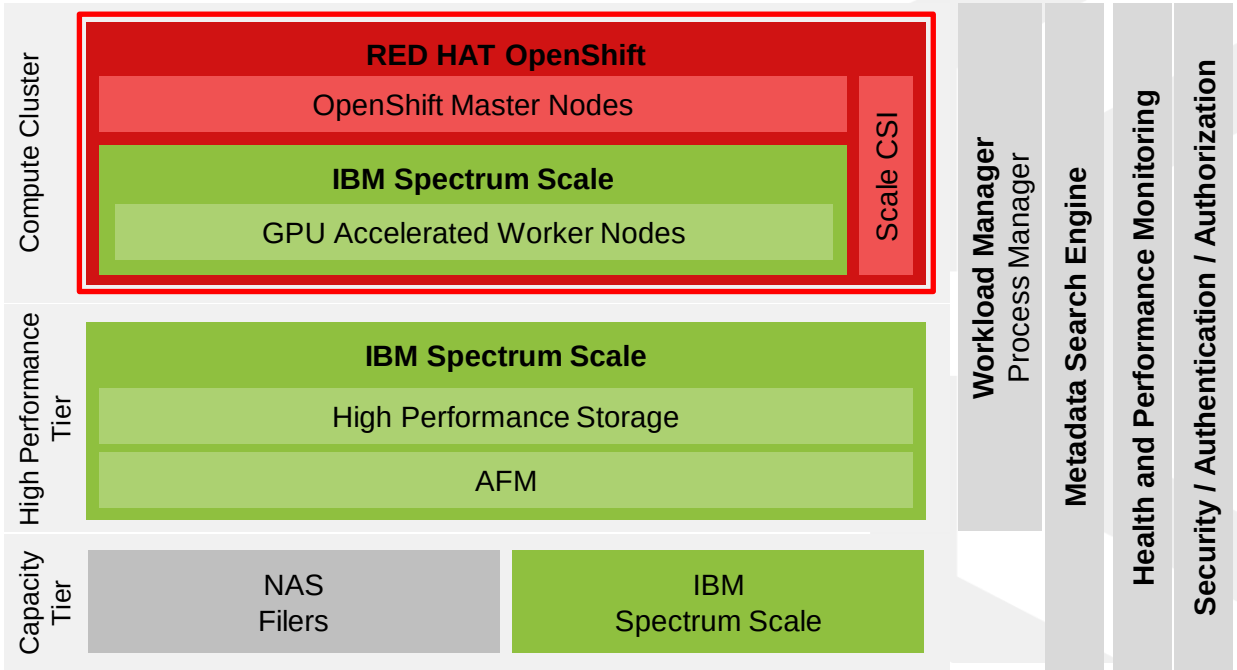
Data Accelerator for AI and Analytics (DA^3)

How do I ensure my data in the high performance tier stays in sync with the capacity tier?



Data Accelerator for AI and Analytics (DA^3)

How can I easily / efficiently provision, test, deploy and scale my containerized workloads?



Summary: Data Accelerator for AI and Analytics (DA^3)

- Brings the right data closer to AI & analytics applications at the right time
  - Faster model training, inference or real-time analytics
  - Without another “persistent” copy of the data, keeping a single source of truth
- Can be deployed on IBM Spectrum Scale NVMe appliance, ESS3000
- Integrates with
  - Workload Managers (start with IBM Spectrum LSF)

    Handle job scheduling and balancing load on compute nodes
  - Metadata Search Engines (start with IBM Spectrum Discover)

    Enable meta-data management functions such as tagging, classification to automate selection of right datasets for prefetching into the cache
- End to end security and monitoring
- Eases Data Scientist’s and System Admins workflows

Just slideware?

- AI Example Use Case: Detect Wildfires

  Workflow replicated in our Germany Kelsterbach Lab

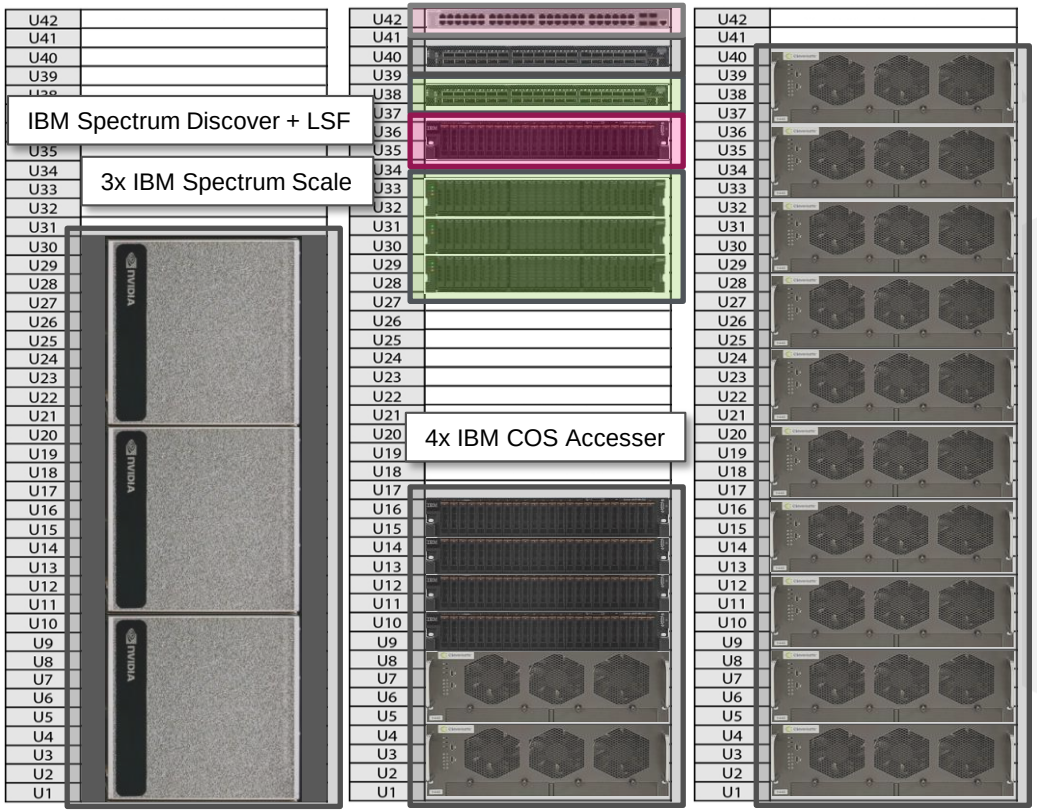
Wildfires

Objective:

- The purpose of the study is to apply AI techniques to detect smoke as a better way to fight fires
  - Detect Wildfires automatically
  - Deliver real time alerts about new burns to active fire fighters
- AI/DL neural networks and inferencing utilized to analyze images and video files to monitor smoke and wildfires

Pictures and Video:
Smoke detection study NVIDIA, Microsoft, IBM

Configuration 2019



Configuration 2019


Smoke
Smoke
Smoke
Smoke
Smoke
Smoke
Smoke
Smoke
Smoke
Smoke

Blueprint and Interview

Interview

Blueprint

Designing and Building End-to-End Data Pipelines Using IBM Storage for AI and Big Data with NVIDIA DGX Systems

High-Performance IBM Storage for AI and Big Data with NVIDIA DGX-2 POD Systems for Enterprise Data Pipelines

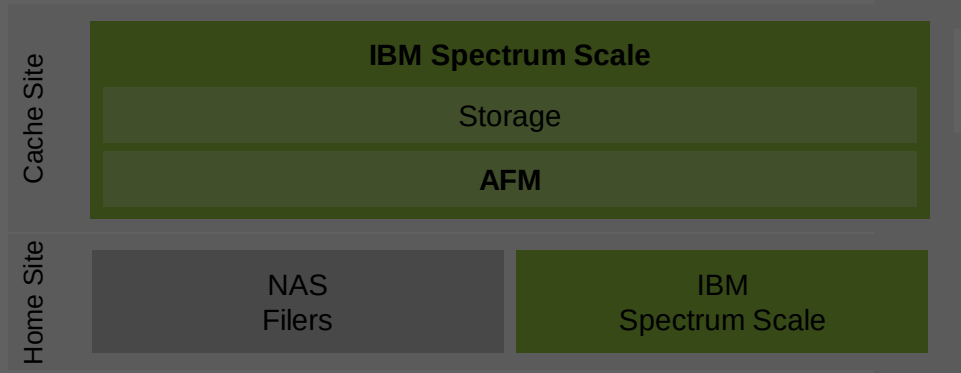
IBM Cloud Object Storage

IBM Spectrum Scale

IBM Spectrum Discover

Live DA^3 Insides / Demo

> 20min. Demo and additional Insights: Not available online.



Live DA^3 Insides / Demo

- Build the Demo in relatively short time, without much knowledge of the other products. Following colleagues were more or less involved and helped hands on or by pointing to the right knowledge center pages:

Thanks to:

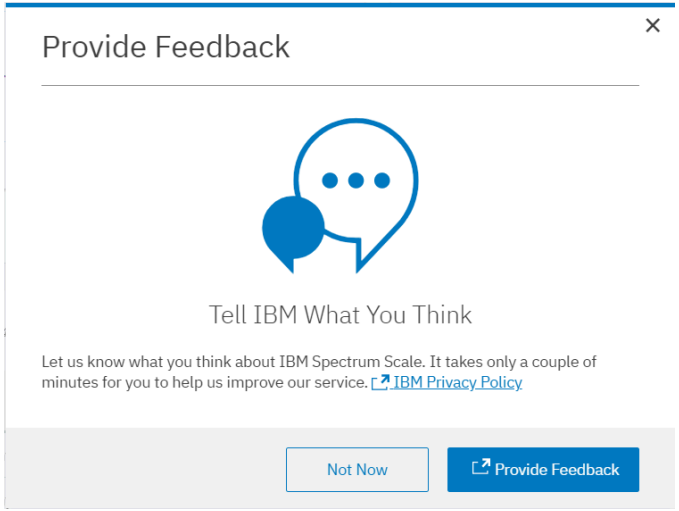
- Discover
  - Joseph Dain
  - Raul Saba
- LSF
  - Christof Westhues
  - Bill McMillan
  - Mark Black
- COS
  - David Wohlford
  - Derek Gascon
  - Shibhani Rai
- Scale
  - Constantine Arnold
  - Gero Schmidt
  - Harald Seipp
  - Tomer Perry
  - Marc Eshel
  - Wei Gong
  - Smita Raut
  - Przemyslaw Podfigurny
  - Piyush Chaudhary, for the “I’m sure, You will make it!”

Thank you!

Data Accelerator for AI and Analytics
Online Survey

Please help us to improve Spectrum Scale with your feedback

- If you get a survey in email or a popup from the GUI, please respond
- We read every single reply



Or contact: Simon Lorenz <simon.lorenz@de.ibm.com>