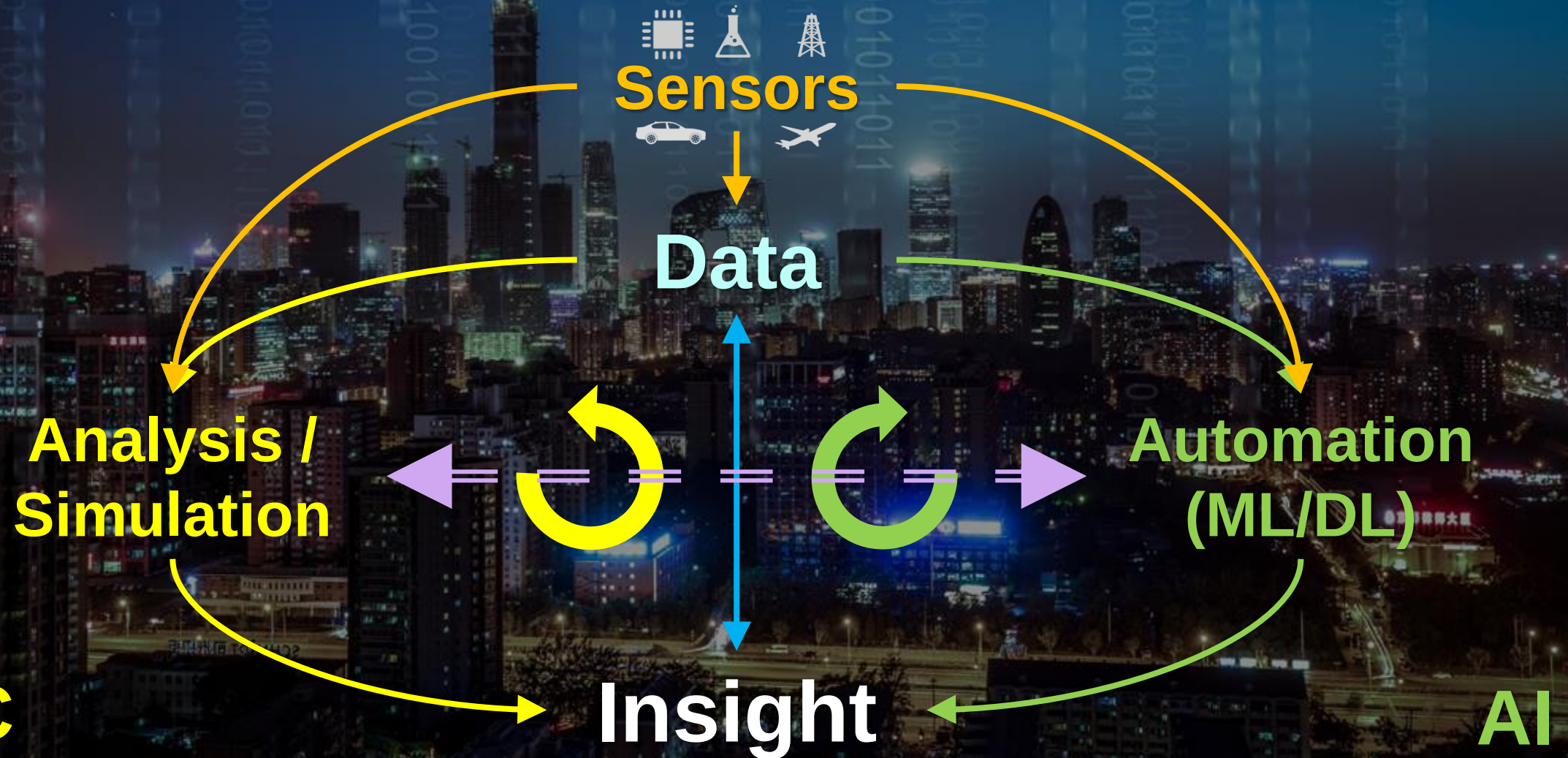


Spectrum Scale for AI

Ted Hoover
Program Director

IBM Spectrum Scale Development

The convergence of HPC, Big Data/AI and IoT enables better decisions, design, manufacturing and products.



AI Examples in Every Industry



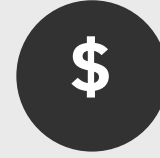
Autonomous
driving
Accident
avoidance



Location-based
advertising



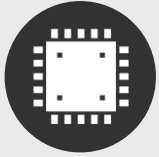
Sentiment analysis of
what's hot, problems



Market prediction
Fraud/Risk



Experiment sensor
analysis



Mfg. quality
Warranty
analysis



Clinical trials, drug
discovery,
Genomics



Captioning,
search, real time
translation



People & career
matching



Patient sensors,
medical image
interpretation



Drilling exploration
sensor analysis



Consumer
sentiment Analysis



Sensor analysis for
optimal traffic
flows

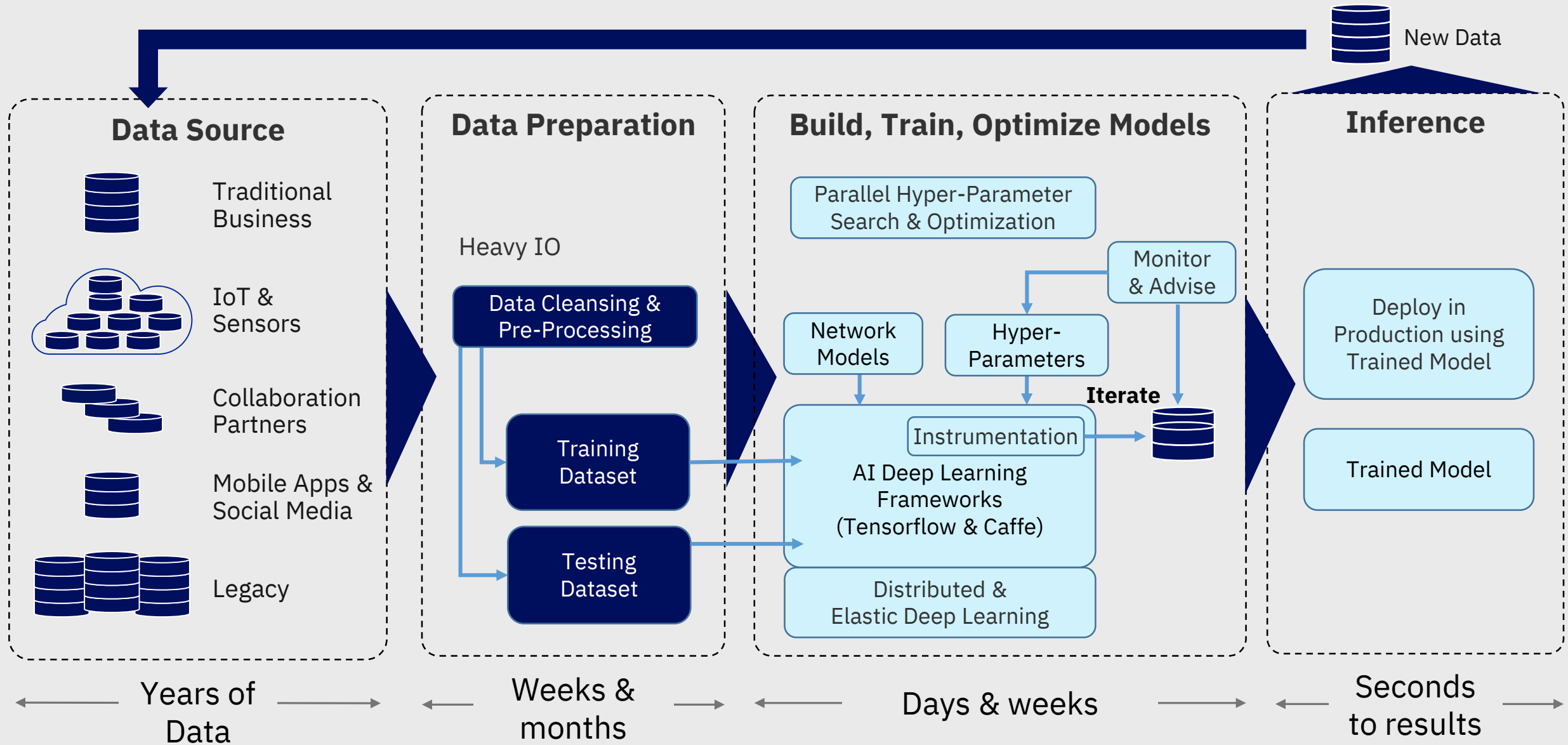


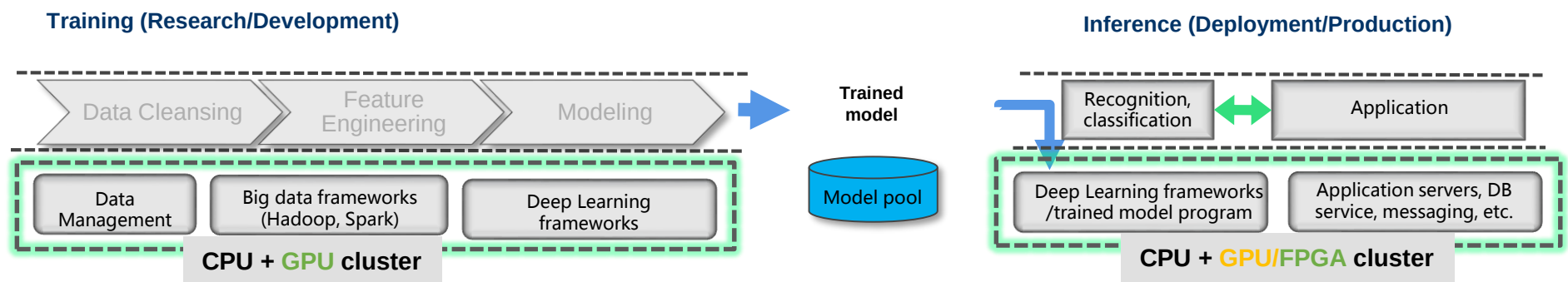
Smart Meter
analysis
for network
capacity,



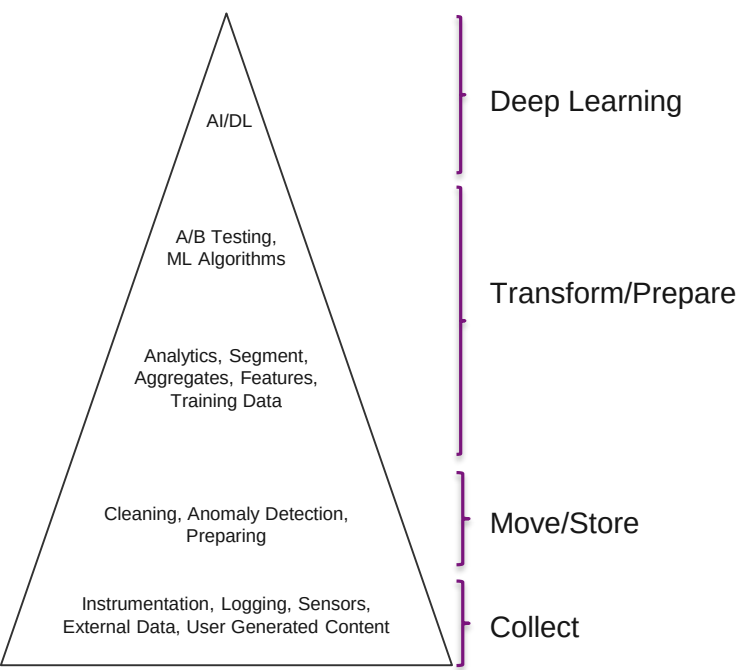
Threat analysis - social
media monitoring, video
Surveillance

Work flow and data flow is complex





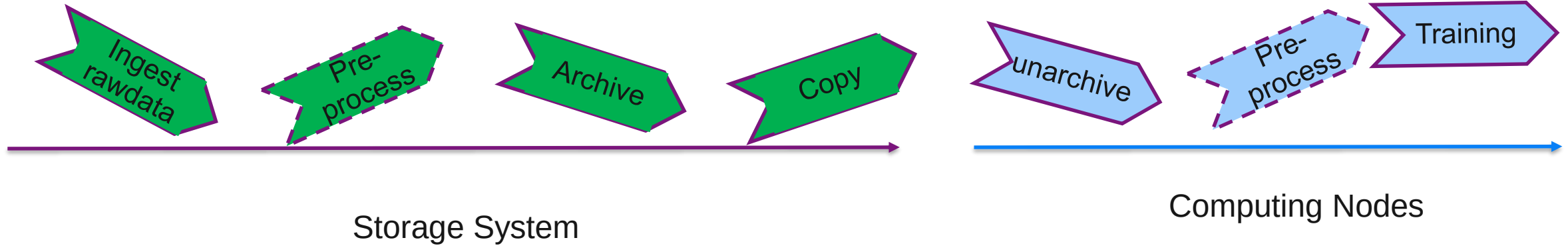
Data Hierarchy in Deep Learning



Phase	Tools and Applications
Deep Learning	<i>Frameworks:</i> TensorFlow(Apache), Caffe(BSD), Torch(BSD), Theano(free), CNTK(free), Neon(Apache) <i>IDE:</i> *IBM Power AI Enterprise/CwS with deep learning, Nvidia Digits, *Watson <i>Spectrum Scale:</i> POSIX for Power AI;
Transform/Prepare	<i>Machine Learning:</i> IBM SPSS, * IBM DSX , SAS <i>Spectrum Scale:</i> SPSS works over GPFS POSIX/HDFS; SAS works over GPFS POSIX
Move/Store	<i>Platform:</i> Hadoop, Spark, POSIX/NFS/SMB <i>Spectrum Scale:</i> POSIX/HDFS interface;
Collect	Variable data ingestion end devices <i>Spectrum Scale:</i> Supports Swift/S3 object interface for data ingest

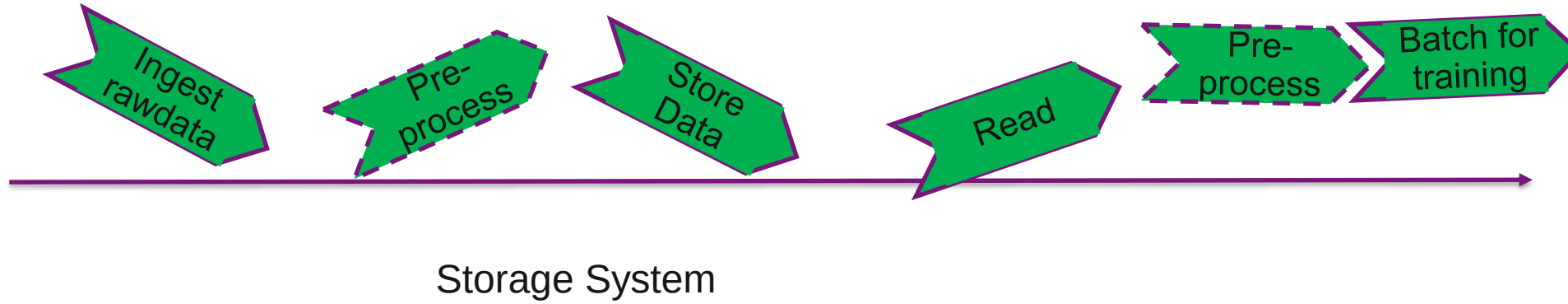
Ref: Monica Rogati – AI Hierarchy of Needs

High Throughput



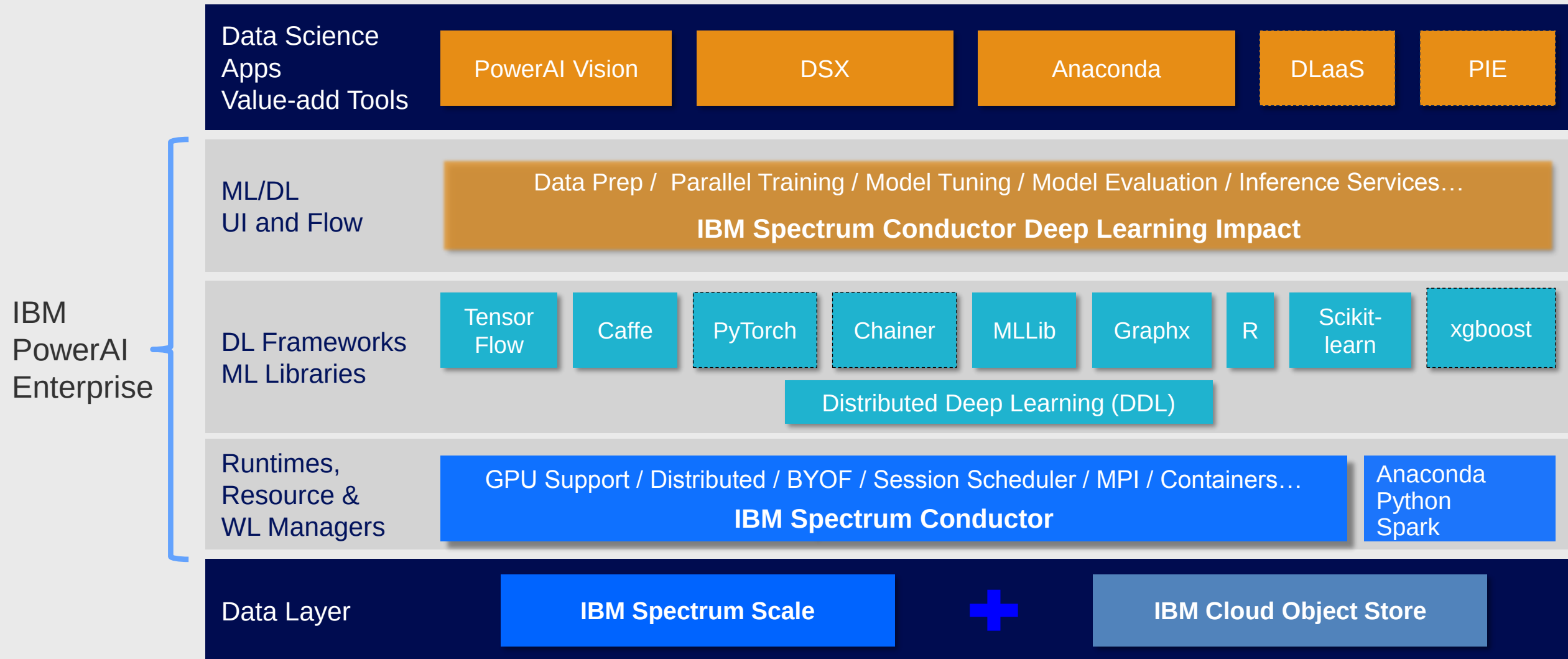
- “pre-process” could be done in Storage side or in Computing Nodes side(it’s flexible)
- “Storage” could be distributed file system or NAS. If so, “copy” could be “read archived data from distributed file system or NAS on Computing Nodes”. If “Storage” is local file system, “copy” will be “scp the archived data”
- “Archive” is to archive a lot of small images into a larger archived file
- Usually, “Unarchive” could be done in memory(it could be in local fast disks)
- High throughputs are required for Storage System

High IOPS



- Raw data or preprocessed data are small pictures(e.g. 20KB per picture)
- High IOPS is required for Storage System

2018 Reference Architecture for AI Infrastructure: Software



IBM AI Architecture from Experimentation to Expansion



Experimentation

Single Tenant



Scientist's workstations



IBM Power Systems AC922



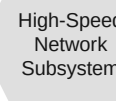
Internal SAS drives & NVM's

Stabilization & Production

Secure Multitenant



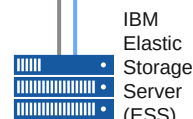
Existing Organization Infrastructure



High-Speed Network Subsystem



Training Cluster
IBM Power Systems AC922



IBM Elastic Storage Server (ESS)

Expansion

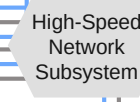
Enterprise Scale / Multiple Lines of Business



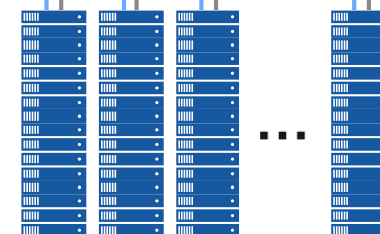
Existing Organization Infrastructure

Master & Failover Master Nodes
IBM Power Systems LC921 & LC922

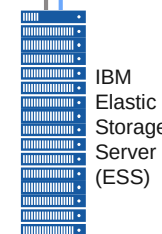
Login Nodes
IBM Power Systems LC921 & LC922



High-Speed Network Subsystem



Training & Inference Cluster
IBM Power Systems AC922, LC921 & LC922



IBM Elastic Storage Server (ESS)

Services & Support

IBM PowerAI Enterprise

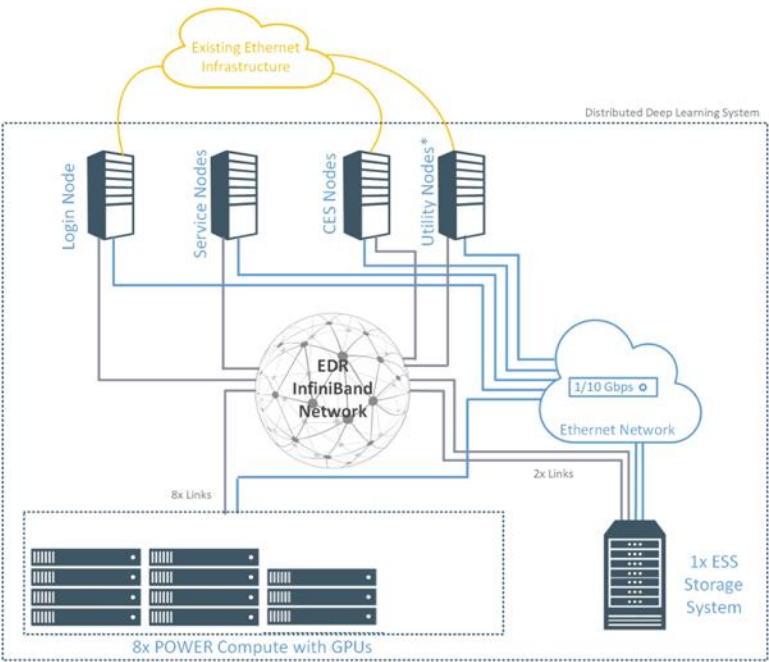
Red Hat Enterprise Linux (RHEL)

IBM Power System & x86 Servers

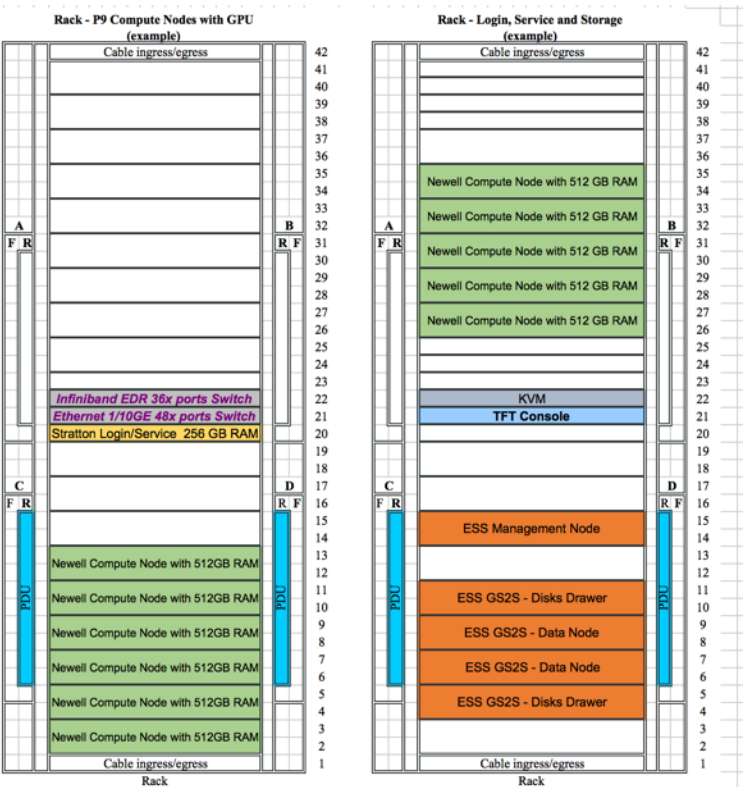
IBM Spectrum Scale / IBM Elastic Storage Server (ESS)

One software stack from experimentation to expansion

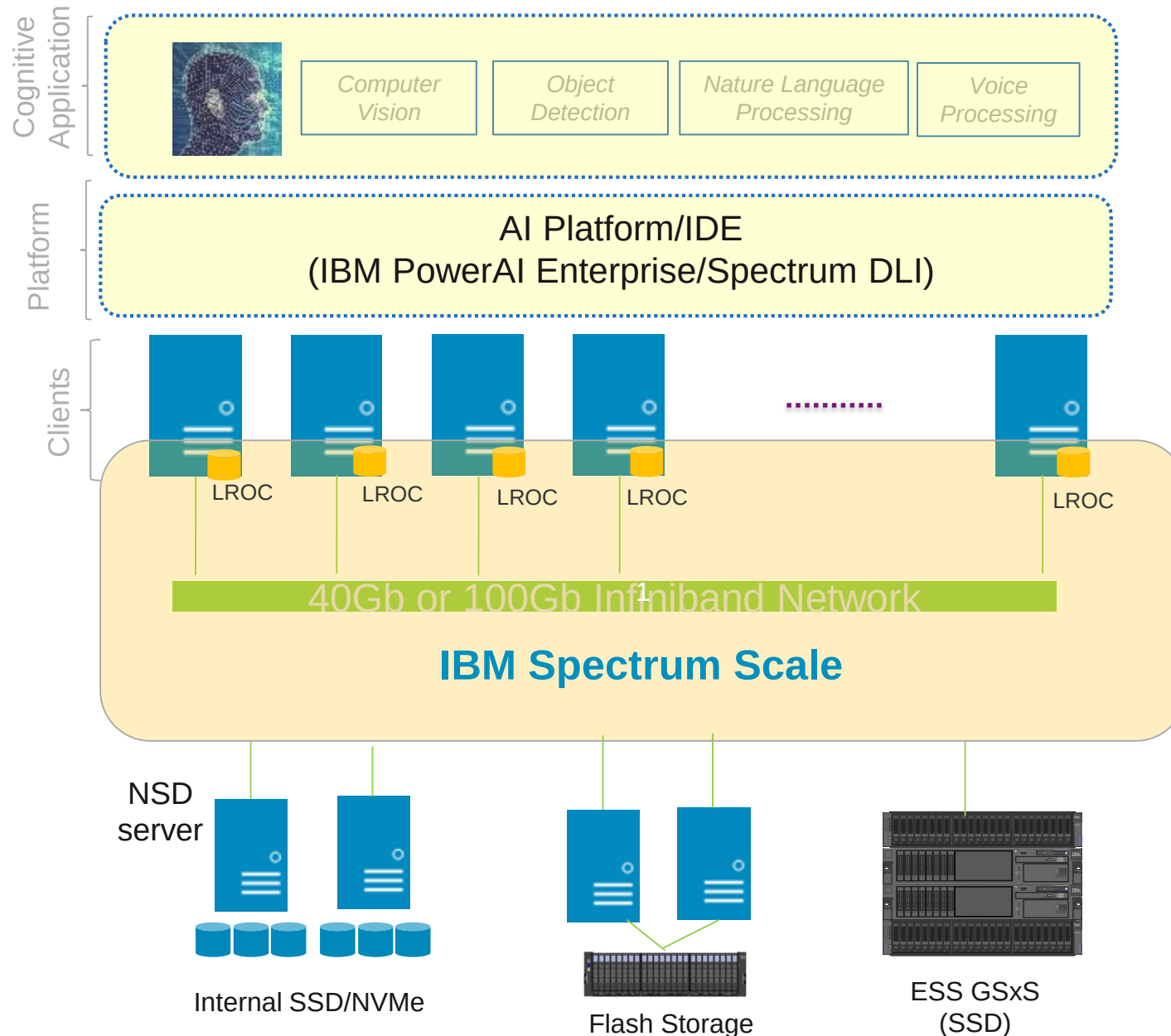
A System that Scales



QTY	Part
1	1 Gb RackSwitch
1	10 Gb RackSwitch
1	100 Gb InfiniBand
1	IBM Power System LC921 (9006-12P)
11	IBM Power System AC922 (8335-GTC)
1	ESS Management Server (5148-21L)
2	Elastic Storage Server Data Server (5148-22L)



Spectrum Scale for Deep Learning



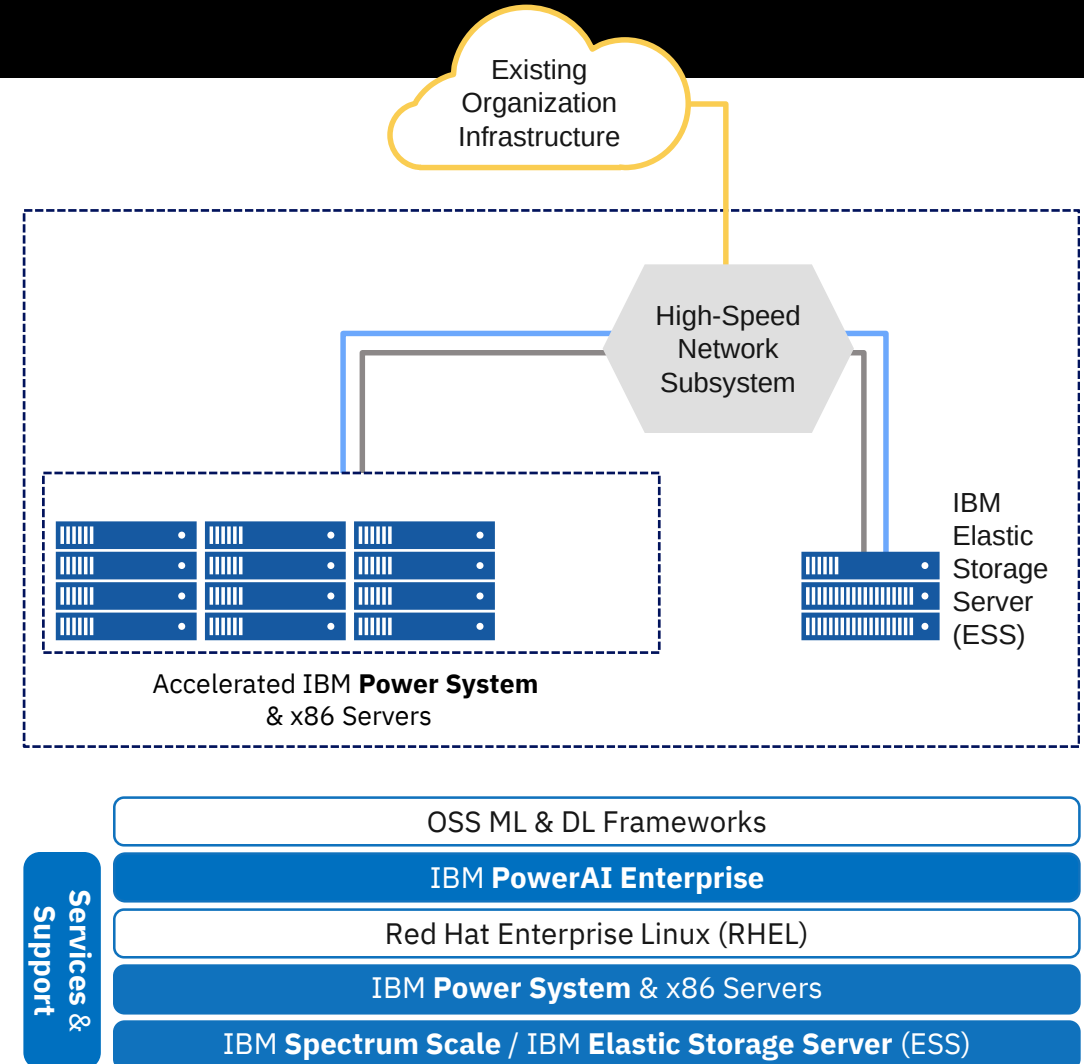
- **Recommended Hardware Configuration**
 - ✓ Spectrum Scale server/client configuration
 - ✓ High IOPS SSD or NVMe as storage or ESS GSxS/SSD or Flash
 - ✓ 40Gb/100Gb Infiniband or RoCE ethernet for RDMA
- **Spectrum Scale advantages in DL**
 - ✓ SMB/NFS/POSIX/HDFS/object interface supports for different data ingestion for training and inference
 - ✓ Parallel read for high IOPS/throughput requirements
 - ✓ Full data lifecycle management
- ✓ **What we are doing**
 - ✓ Best practice guide(in progress)
 - ✓ Mmap() performance enhancements
 - ✓ Further IO optimization for write-once-but-read-many-times:
 - ✓ Prefetch/cache all data in LROC locally(in progress)
 - ✓ Tuning profile for cognitive workloads

A Reference Architecture for AI Infrastructure

IBM Systems

Supports your AI journey from PoC/experimentation into production & then scale-out to support your organization/enterprise. Based on the extensive work done with our clients building their AI environments

- Reduce complexity, time & risk of building & running
- Improves data science efficiency & productivity
- Building block approach
- Speeds time to accuracy
- End-to-end AI workflow support - from data ingestion & preparation through building, training & optimizing models, & into production & inference
- Enterprise resiliency
- Integrated HW & SW solution based on opens source & IBM solutions
- IBM services & support



<https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=87016787USEN>

Thank You.
IBM Storage & SDI

A series of thick, blue diagonal stripes of varying lengths and orientations, creating a dynamic, abstract pattern in the bottom right corner of the slide.

Based on real world experience

Different workloads, but built with many of the same building blocks

Wells Fargo: Financial Risk Modeling

Using AI to enhance financial risk models and provide validation to meet regulatory requirements and business goals.

Automotive Sensor IoT: Transforming data from the edge to useful insights

From global data to insight, they manage large data as objects, extracted to run AI.

Top 5 Global Bank: Building a better client profile using Spark and AI

Managing multi-platform data ingest with distributed computing and ML/DL to normalize, clean and tag data to build client behavior profiles .

IBM Global Chief Data Office: One Common Enterprise Data Backbone

The backbone at the core of every business process for a single version of the truth, providing data, computing, analytics & AI.

CORAL: National Lab Supercomputers built for AI

The most powerful and smartest supercomputers in the world, and purpose built for AI workloads.