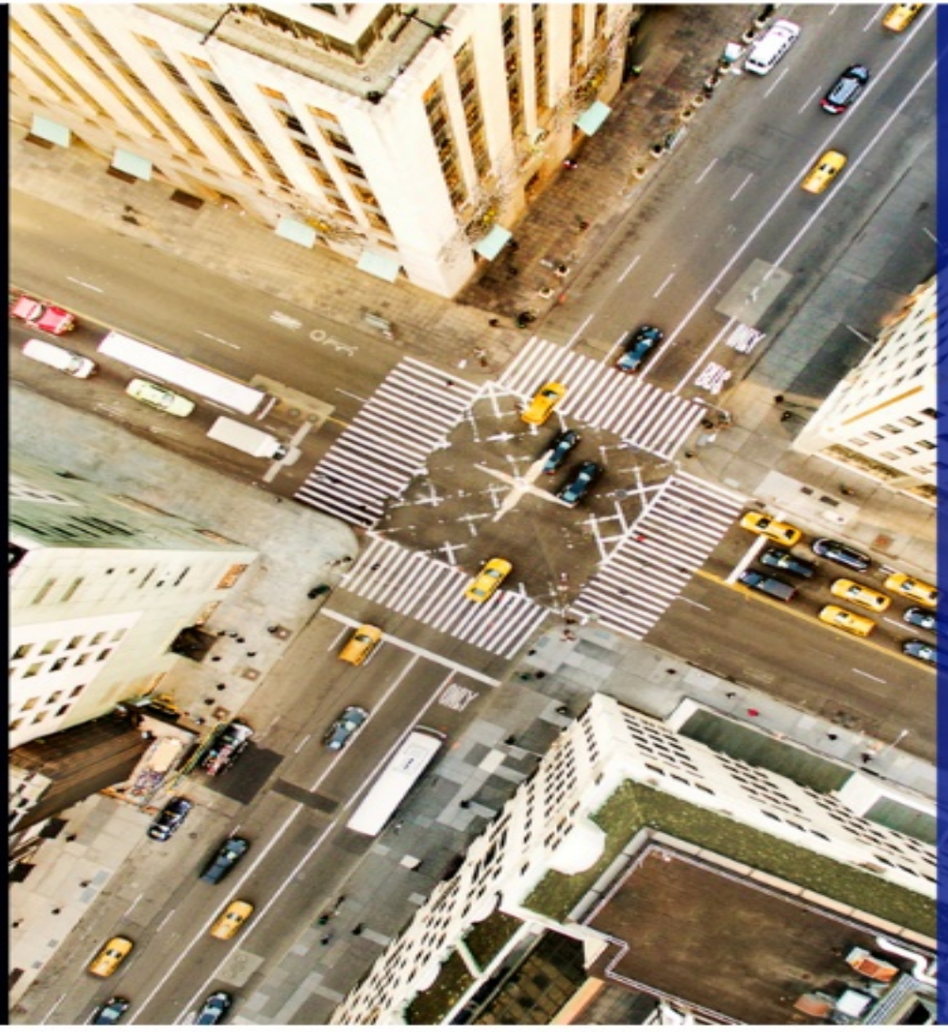
Accelerating AI Results in the Enterprise

Nick Werstiuk
Director Offering Management
werstiuk@ca.ibm.com
October 4, 2018



IBM

About this Presentation

IBM Systems is addressing the challenges organizations face in evolving their AI infrastructure from experimentation with PoCs, through growing multi-tenant, production systems with a goal towards expansion to enterprise scale, all while integrating into an organization's existing IT infrastructure.

With a set of easy to use, integrated software tools built on optimized, accelerated hardware, the IBM Systems AI architecture enables organizations to jump start AI and Deep Learning projects, speeds time to model accuracy and provides Enterprise-grade security, interoperability and support.

AI Examples in Every Industry

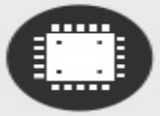
Autonomous driving
Accident avoidance

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

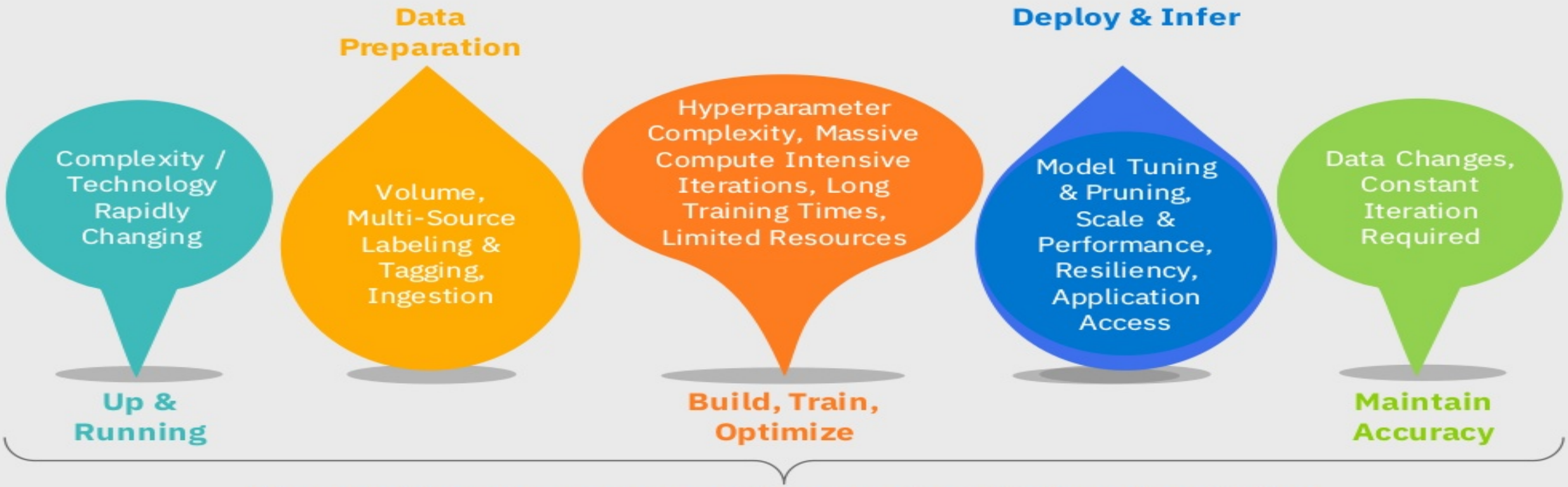
Consumer sentiment Analysis

Sensor analysis for optimal traffic flows

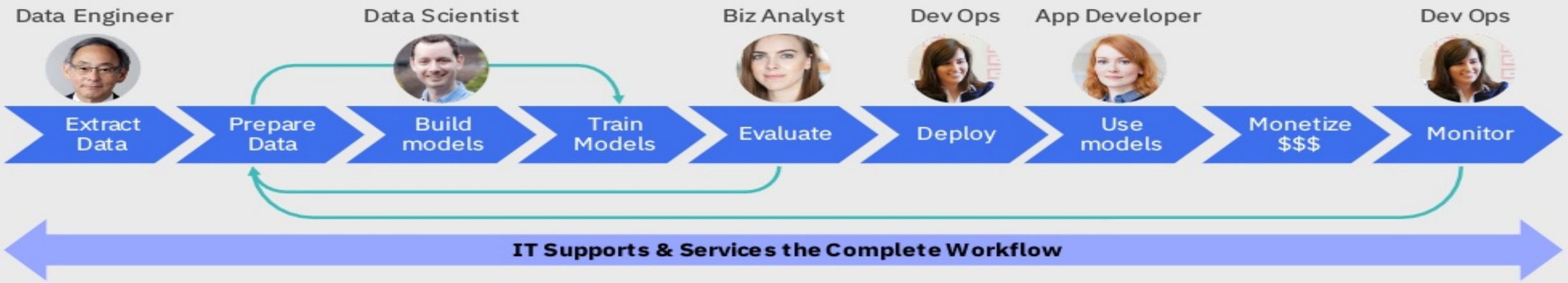
Smart Meter analysis for network capacity

Threat analysis, social media monitoring, video Surveillance

Pain Points – Deep Learning Pipeline

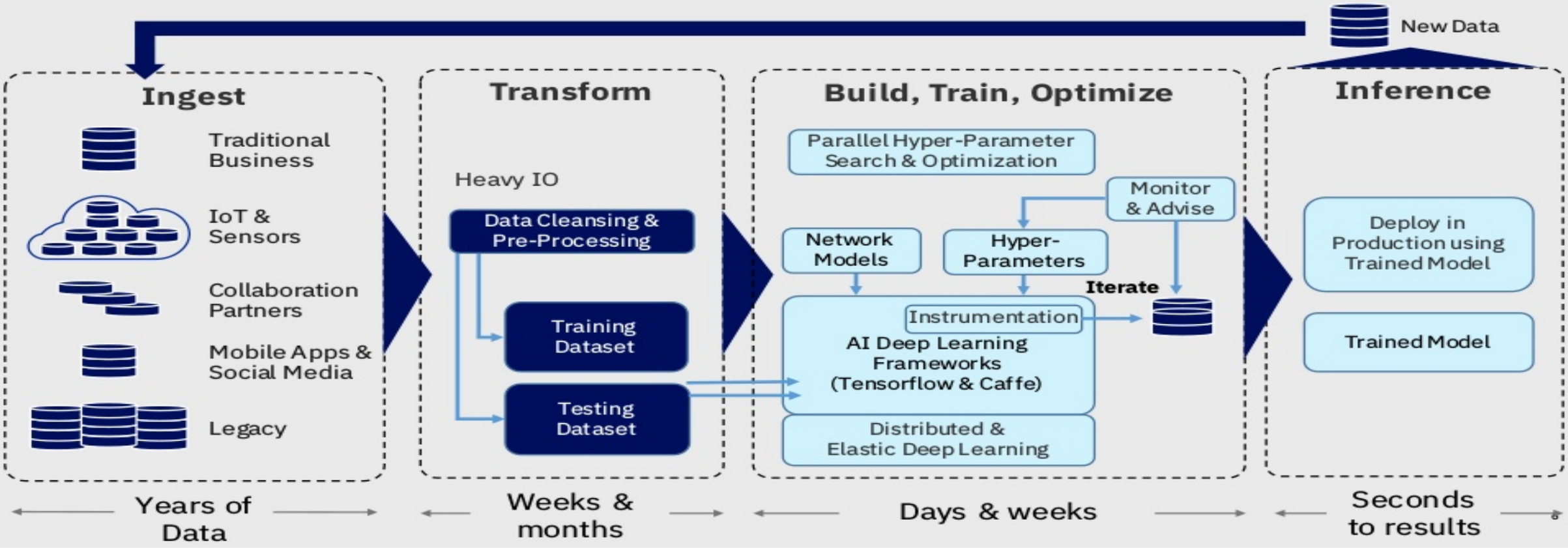


Data Science for AI/Deep Learning is a Team Sport



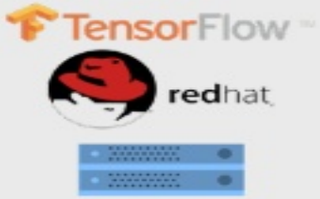
Building cognitive apps using deep learning **requires** multiple skillsets
Connected infrastructure for data, development and iteration.
A common data platform and workflow is crucial for enterprise success.

Deep Learning Work flow and data flow is complex



Key Capabilities in PowerAI Enterprise

Simplicity: Integrated Platform that Just Works

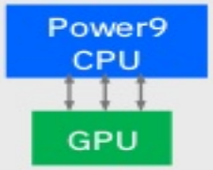


Curate, Test, and Support Fast Moving Open Source

Provide Enterprise Distribution on RedHat

Easy to deploy Enterprise Platform across the AI workflow

Ease of Use, Unique Capabilities



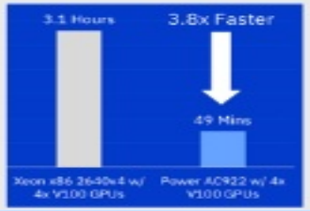
Large data & model support due to NVLink

Acceleration of Spark, Analytics & ML

AutoML

Elastic Training: Scale GPUs as Required

Faster Model Training Time



Faster Training Times in Single Server

Scalability to 100s of Servers (Cluster level Integration)

Leads to Faster Insights and Better Economics

Open AI Platform w/ Ecosystem Partners

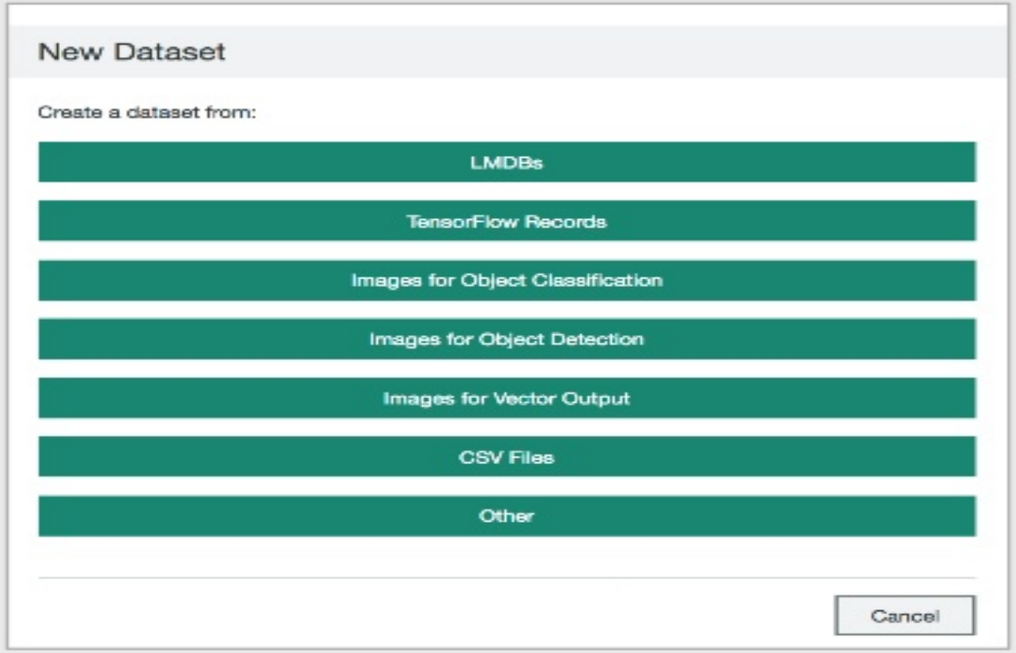
Platform that Partners can build on

Software Partners: H2O, IBM, Anaconda

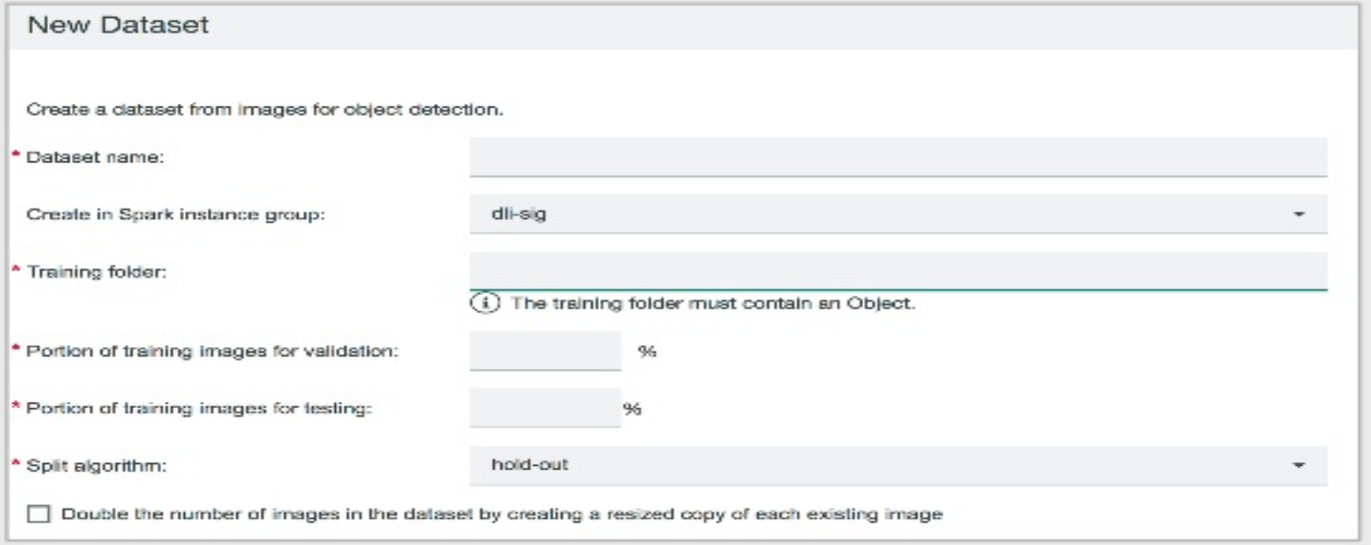
SIs, Solution Vendors & Accelerator Partners

Data Preparation for Deep Learning

Import from different formats

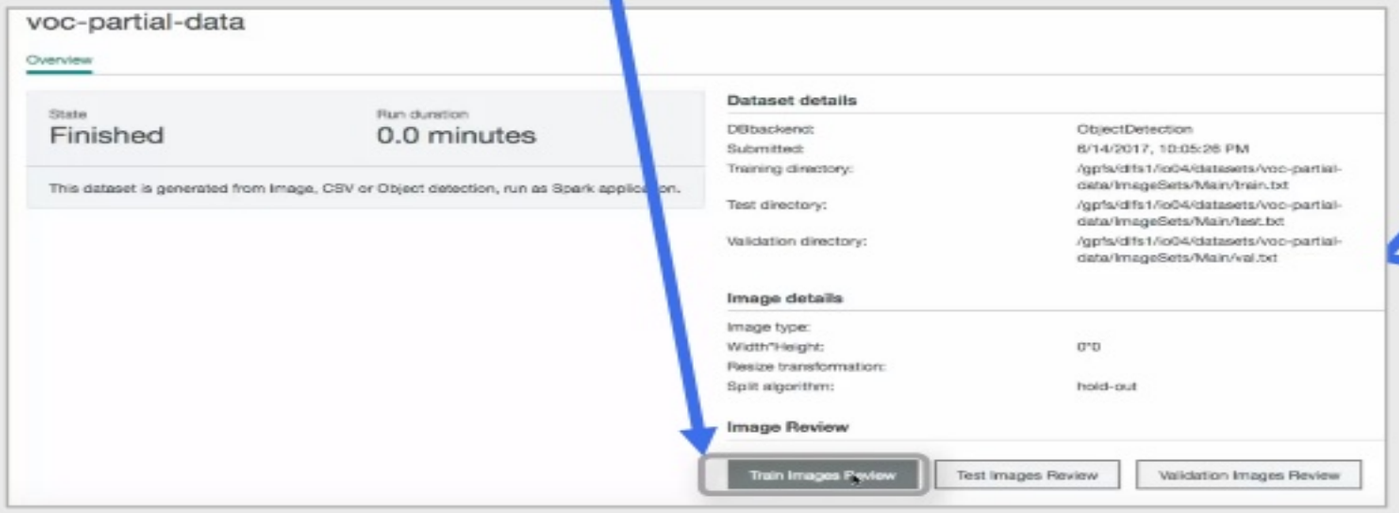


Transform, split and shuffle



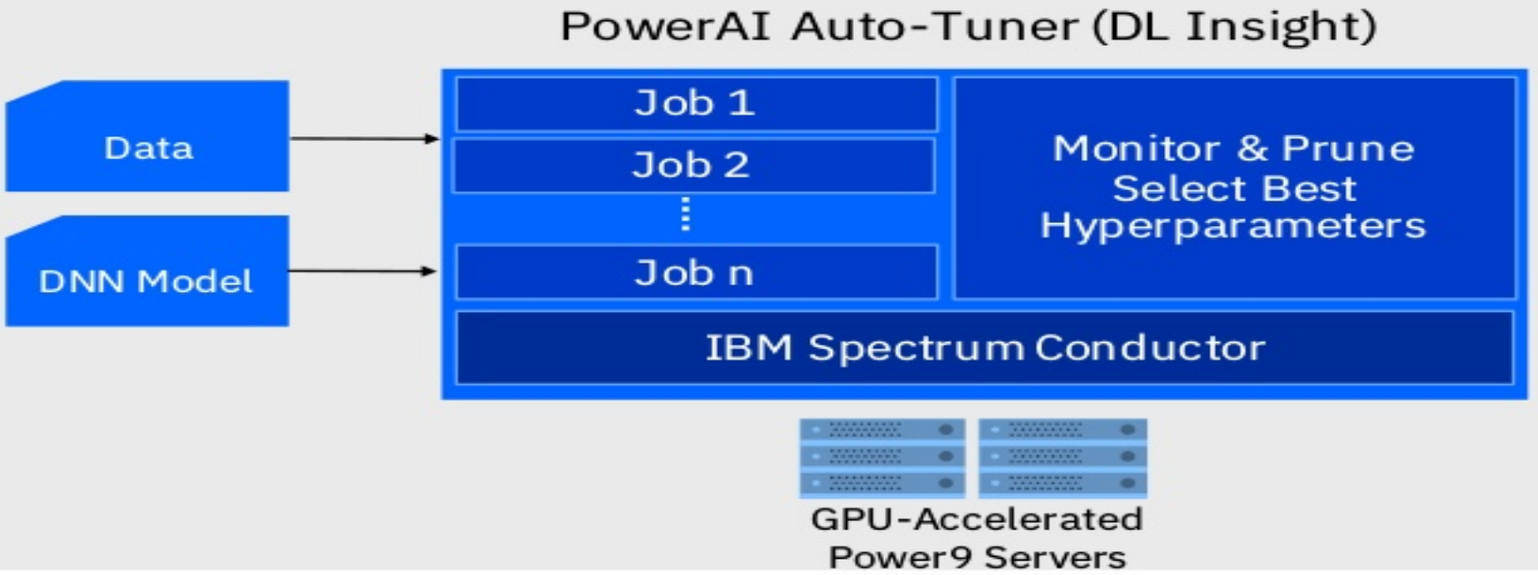
Data Preparation for Deep Learning

Preview Results

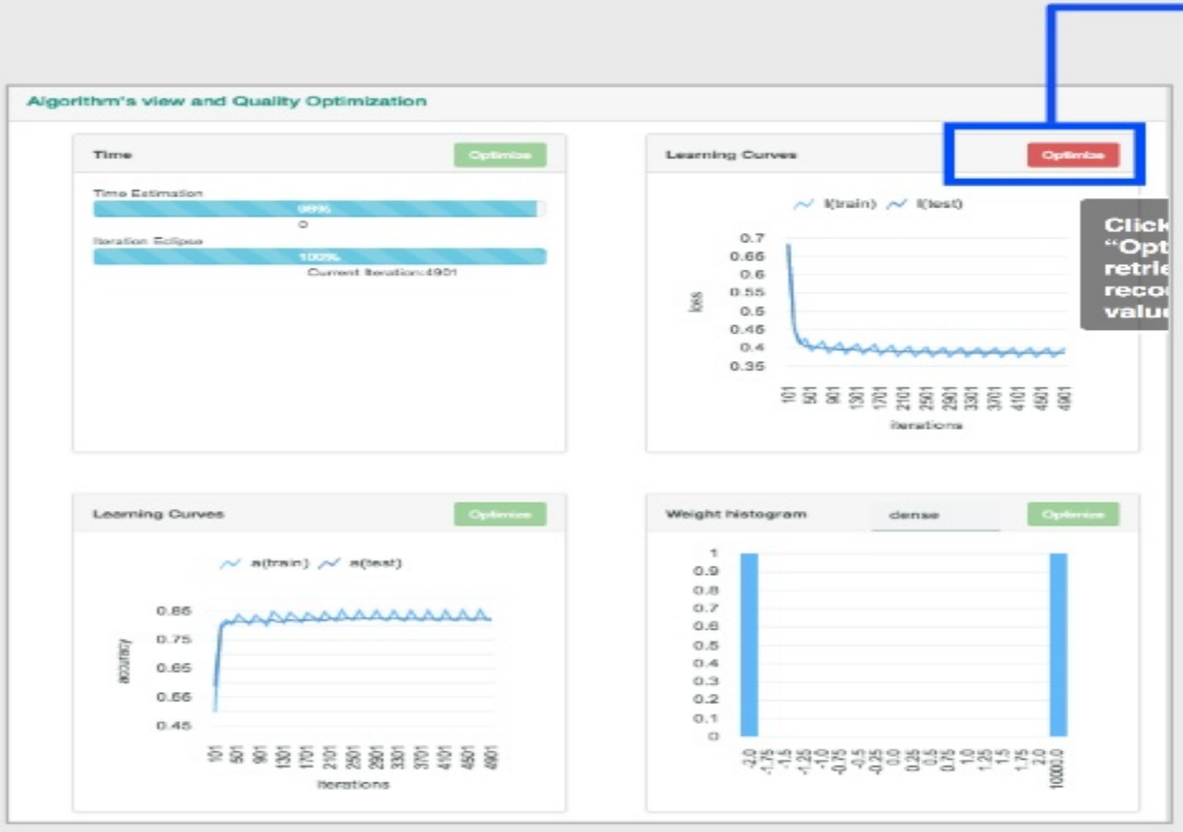


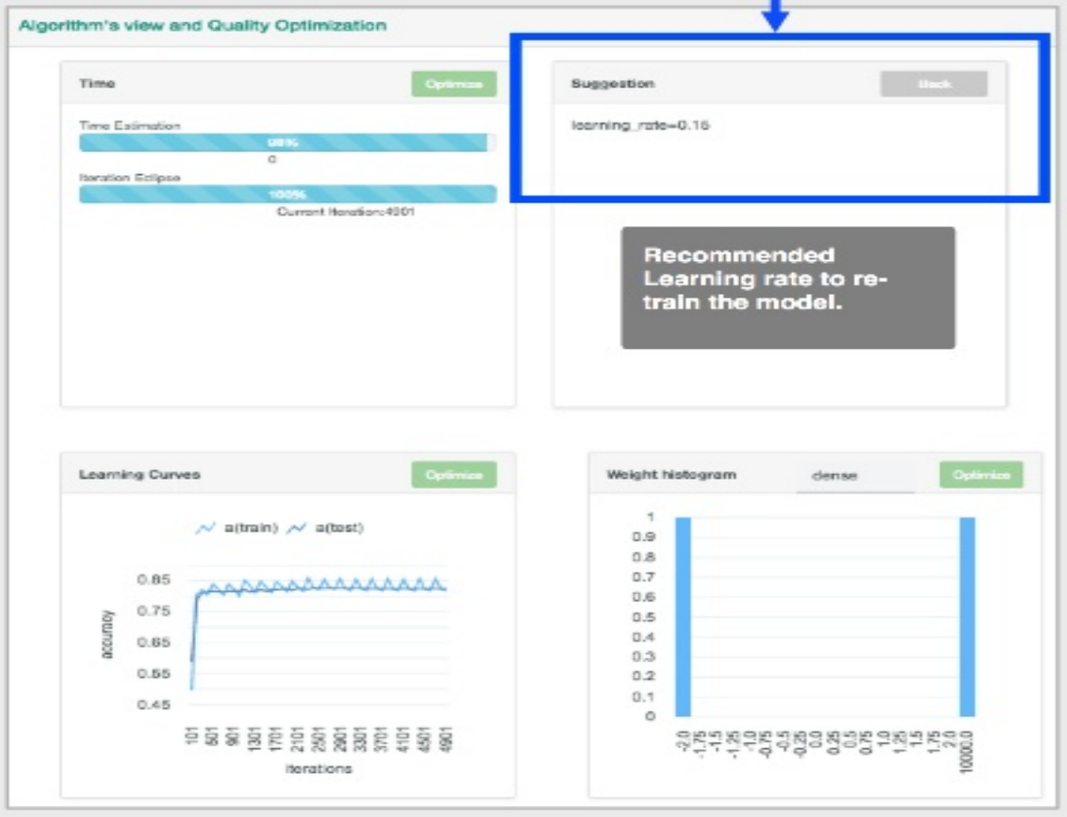
Auto Hyper-Parameter Tuning

- Data scientists run 100s of jobs with different Hyper-parameters
  - Learning rate, Decay rate, Batch size, Optimizers (GradientDecedent, Adadelta, Momentum, RMSProp, ..)
- Auto-Tuner searches for good hyper-parameters by launching 10s of jobs & selecting the best ones
  - 3 search approaches: Random, Tree-based Parzen Estimator (TPE), Bayesian



Runtime Training Visualization
Monitor, Analyze,& Optimize





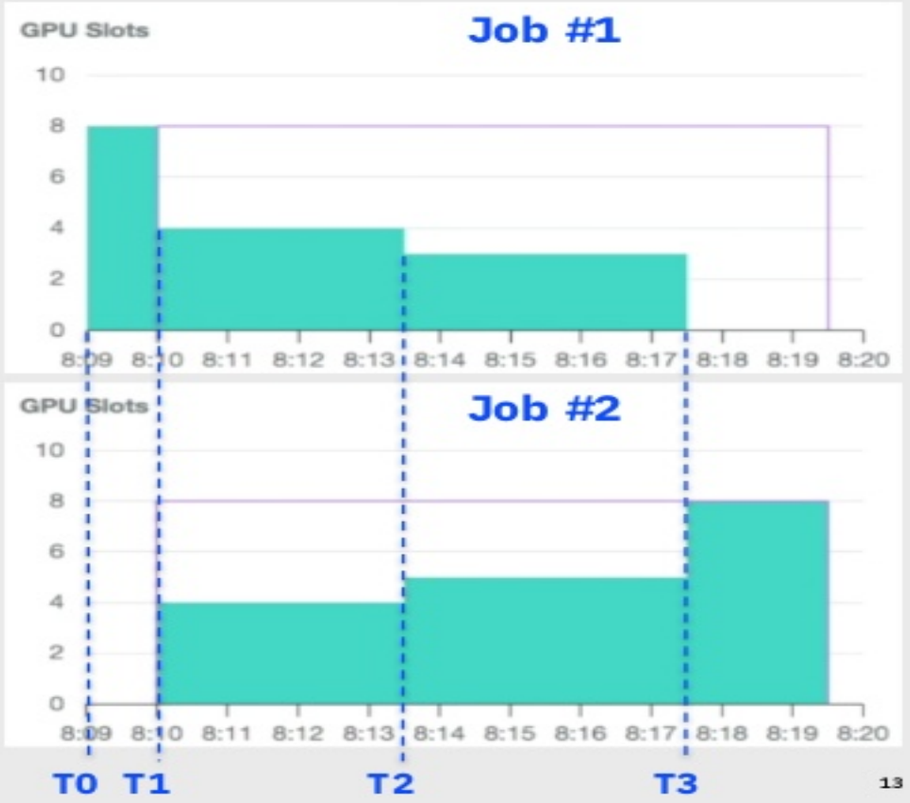
Choose your Distributed Training Approach

Distribution Model	Benefit
Bring Your Own Framework & **Native Distribution Engines**	Frameworks not included in the IBM PowerAI distribution and frameworks with their own native distribution capabilities (e.g., Distributed TensorFlow, Horovod, CaffeOnSpark, etc.)
Distributed Deep Learning (DDL) Very Large Scale-out Single Model	Single user, very large distribution and high-performance training
Elastic Distributed Training Resource Sharing & Multitenancy	Concurrent, dynamic and fault tolerant sharing of resources across many tenants and jobs. Transparent integration with popular frameworks like Pytorch

Elastic Distributed Training – Quality of Service
Transparent, Elastic & Resilient

Environment
- Two (2) POWER8 servers with four (4) GPUs
- Eight (8) GPUs total
- Policies
  - Fairshare
  - Preemption
  - Priority

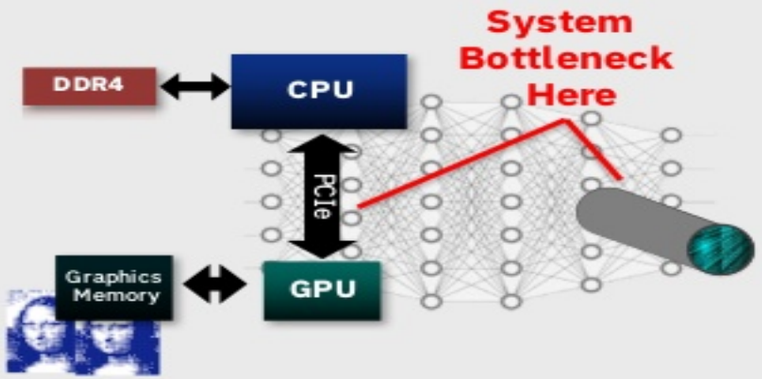
Timeline
- **T0** - Job 1 starts, uses all available GPUs
- **T1** - Job 2 starts, Job 1 gives up four GPUs
- **T2** - Job 2 priority change, Job 1 gives up GPUs
- **T3** - Job 1 finishes, Job 2 uses all GPUs



Train Larger More Complex Models

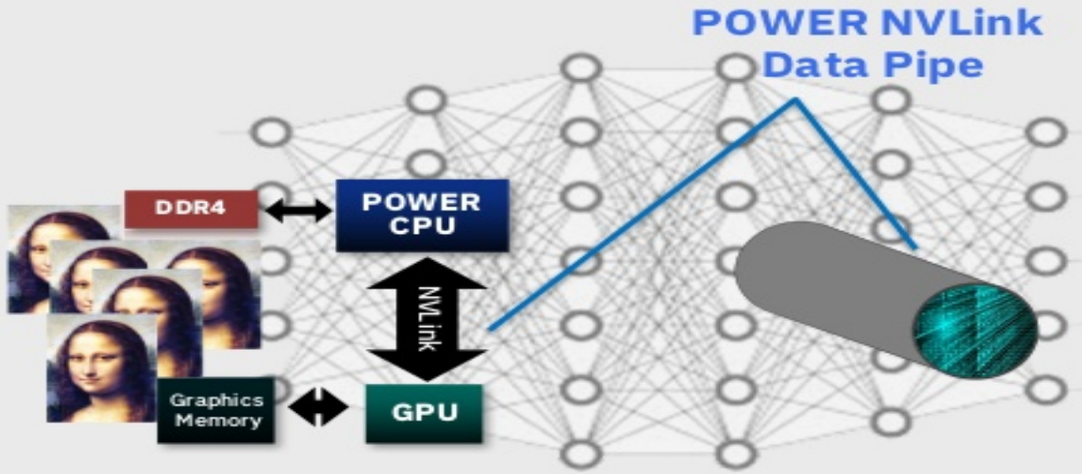
Traditional Model Support

Limited memory on GPU forces tradeoff in model size / data resolution



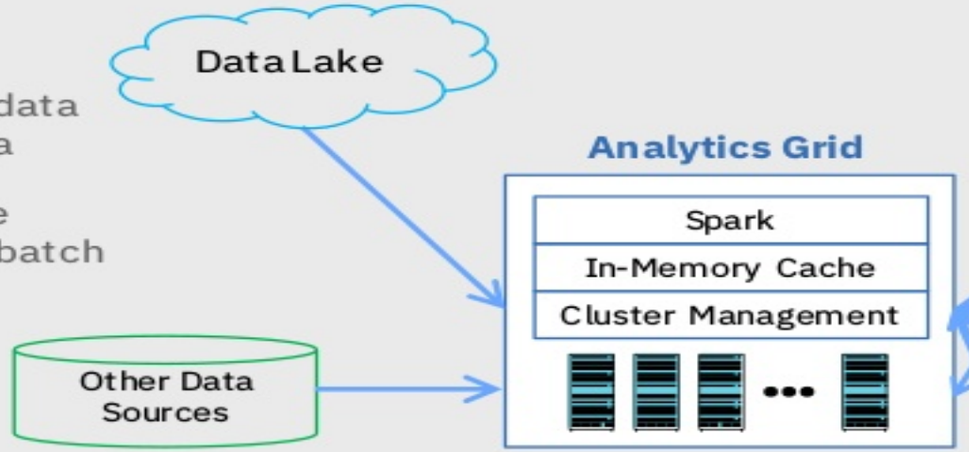
Large Model Support

Use system memory and GPU to support more complex and higher resolution data



High Performance Data Science as Service

Top 5 Bank in USA
- Spark centric environment
- Load data from Hortonworks data lake and other enterprise data sources
- Hundreds of users, interactive notebooks, ad-hoc queries & batch reports



Data Scientists
- Via node book, analyze data and develop new apps

Data Engineering
- Consume ETL as service - develop new ETL apps

Fraud Detection Teams
- Run third party ML application
- Run risk reporting

Risk & Quant Teams
- Develop risk models, train DNN/CNN models

Credit Scoring with DL
- Using Deep Learning for Credit Application Scoring

- High performance **POWER9** environment for compute & memory intensive workloads
- **Spark in memory** analytics for real time/ad-hoc query/analysis & batch reporting
- **Dis-aggregated compute/storage infrastructure**; Scale compute & storage independently
- **True multi-tenant**: Many LOBs, users, resource plans & application SLAs
- **Multiple Spark versions**, multiple notebook versions
- **GPU acceleration**, Shared RDD optimization & **new innovations around ML/DL**
- Support Spark, batch, micro-service application frameworks in the same environment

AI Adoption Cycle

Experimentation

- Single node
- Single user/tenant
- Small scale data
- Algorithm prototyping, hyperparameter optimization

Production

- Expanding use cases
- Multi-node
- Cluster
- Medium scale data
- Security

Expansion

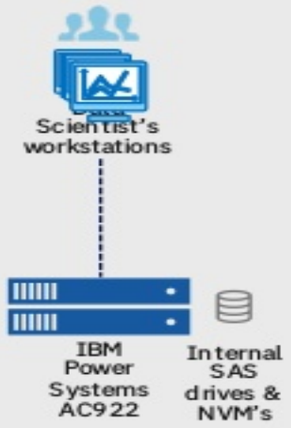
- Data Science Shared Service
- Multitenant
- Upstream data pipeline
- Model iteration
- Scalable Inference

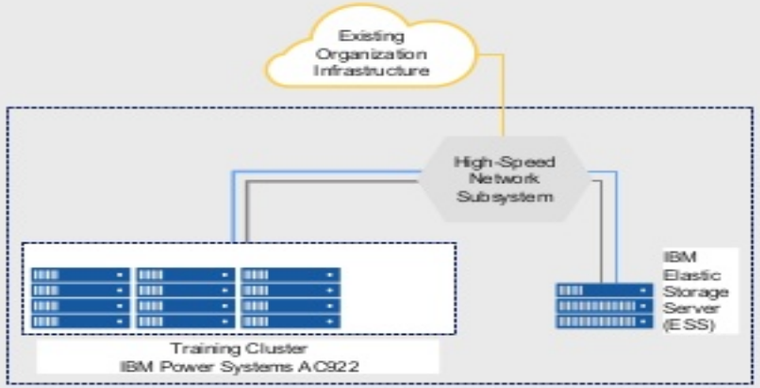
IBM AI Architecture from Experimentation to Expansion

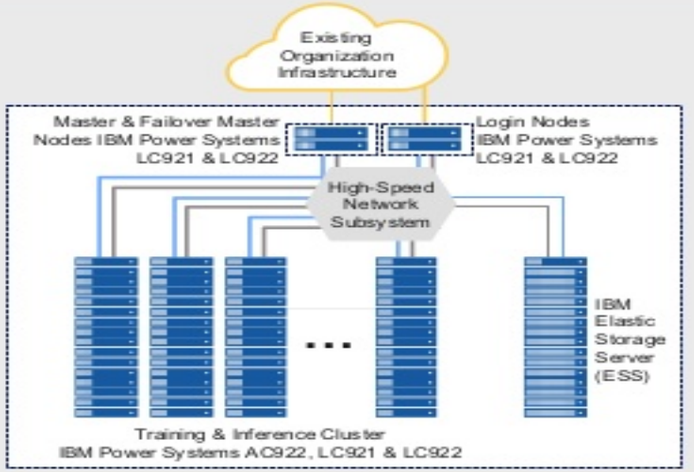
Experimentation
Single Tenant

Stabilization & Production
Secure Multitenant

Expansion
Enterprise Scale / Multiple Lines of Business







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

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 mange 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 supercomptuters in the world, and purpose built for AI workloads.

Thank You !

IBM has many ways to help you get started

Get the detailed reference architecture materials	AI Discovery Workshops to evaluate where you want to start and where you want to go.	Deployment of POC Projects with PowerAI, PowerAI Enterprise and the reference architecture	Expansion and scaling into initial production or enterprise wide deployment