

PowerAI (aka Watson ML Accelerator) update

Common Luxembourg

Double Tree Hotel, Luxembourg
February 28, 2019

Franz Bourlet – POWER Systems technical sales
IBM Belgium & Luxembourg
franz_Bourlet@be.ibm.com



Agenda

- PowerAI / WMLA update and editions
- Deep dive into WMLA Enterprise Edition
- H2O Driverless AI
- Intelligent Video Analytics
- Watson Studio Local
- PowerAI Vision

Agenda

- PowerAI / WMLA update and editions
- Deep dive into WMLA Enterprise Edition
- H2O Driverless AI
- Intelligent Video Analytics
- Watson Studio Local
- PowerAI Vision

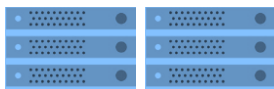
IBM PowerAI (aka Watson Machine Learning Accelerator)

Developer Ease-of-Use Tools

Open Source Frameworks:
Supported Distribution



Faster Training Times via
HW & SW Performance Optimizations



GPU-Accelerated
Power Servers



Storage

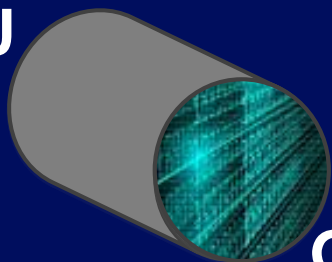
- Integrated & Supported AI Platform
- Ease of Use Tools for Data Scientists
- Red Hat and Ubuntu
- Easy installation (RPM/DEB)
- Cloud and on premise
- Bare metal and Docker-based
- 3-4x Speedup for AI Training

IBM adds value to curated, tested, and pre-compiled frameworks with WMLA

Large Model Support

Use system memory with GPUs to support more complex models and higher resolution data.

CPU



GPU

Distributed Deep Learning

Simplifies the process of training deep learning models across a cluster for faster time to results.

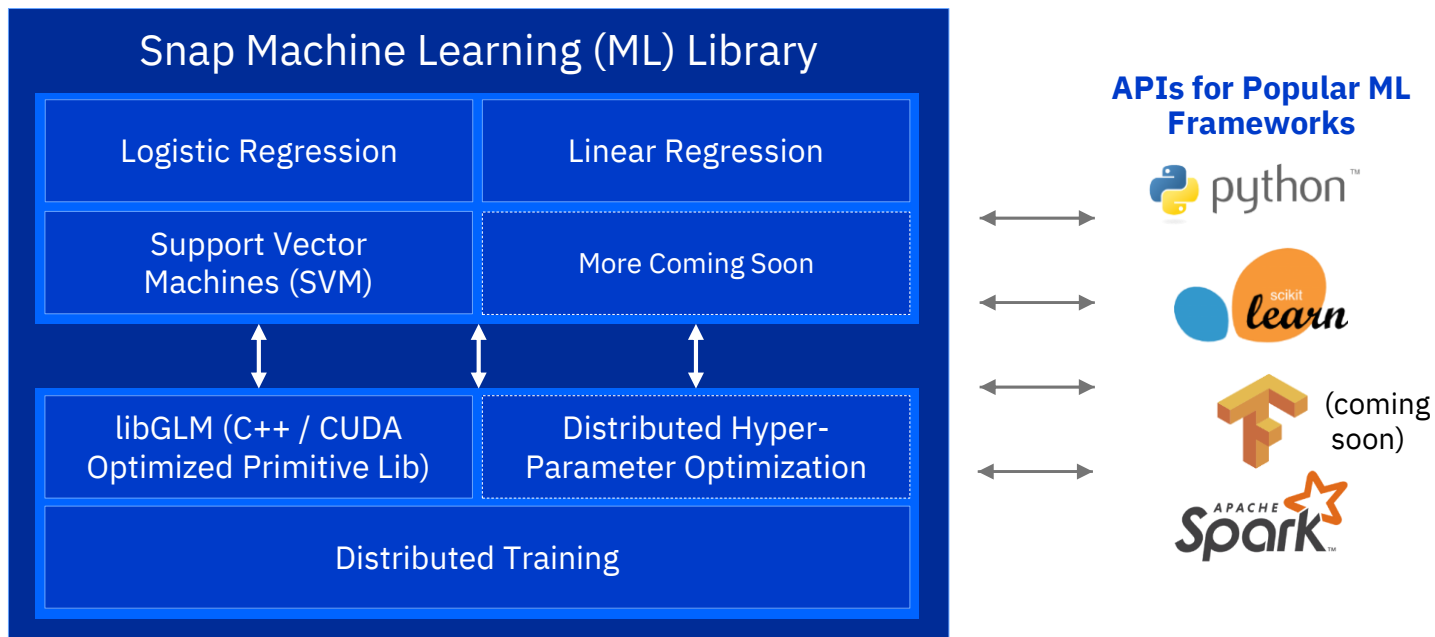


Fully Supported by IBM

PowerAI software and the accelerated Power servers it runs on are supported by IBM technical support.

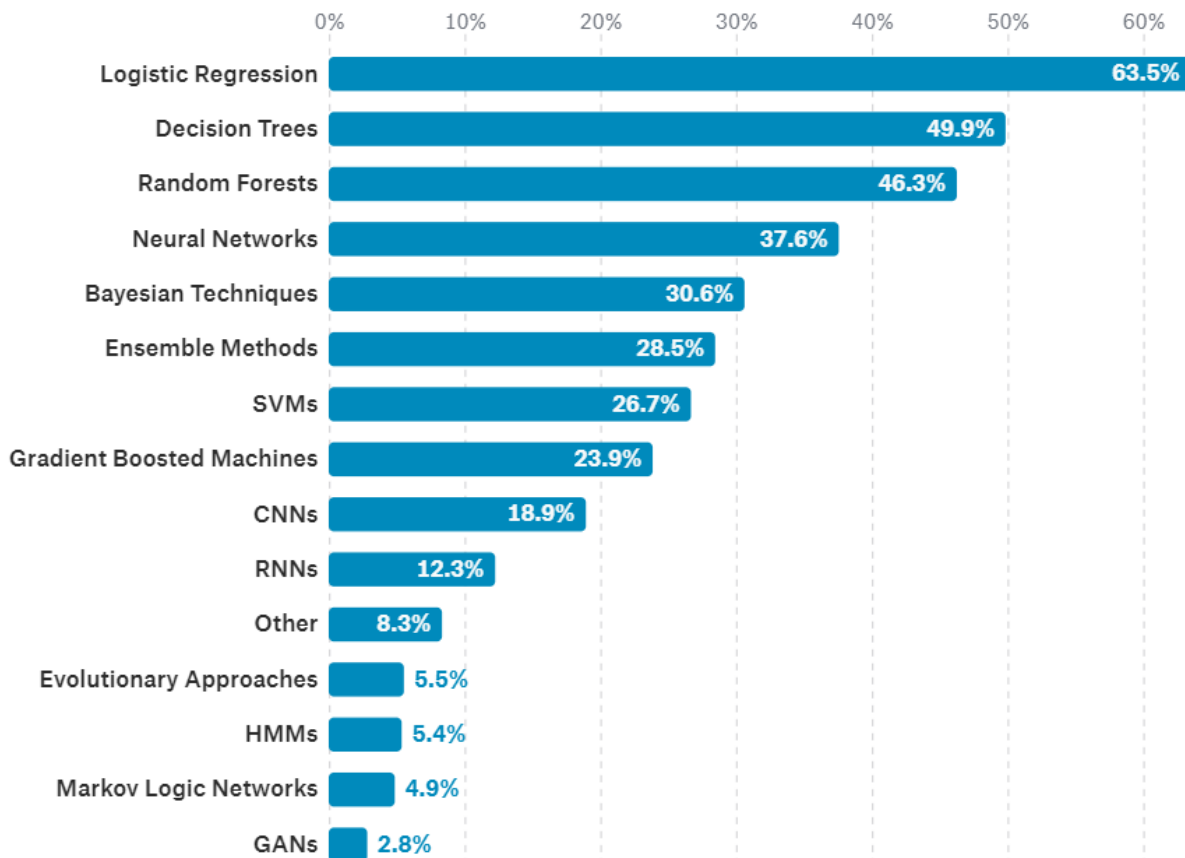


Snap ML – Accelerated Machine Learning Library for Distributed GPU



What data science methods are used at work?

Deep Learning is important and growing rapidly, but a Kaggle survey shows data scientists still rely primarily on machine learning algorithms



Enterprises rely on linear models for analytics

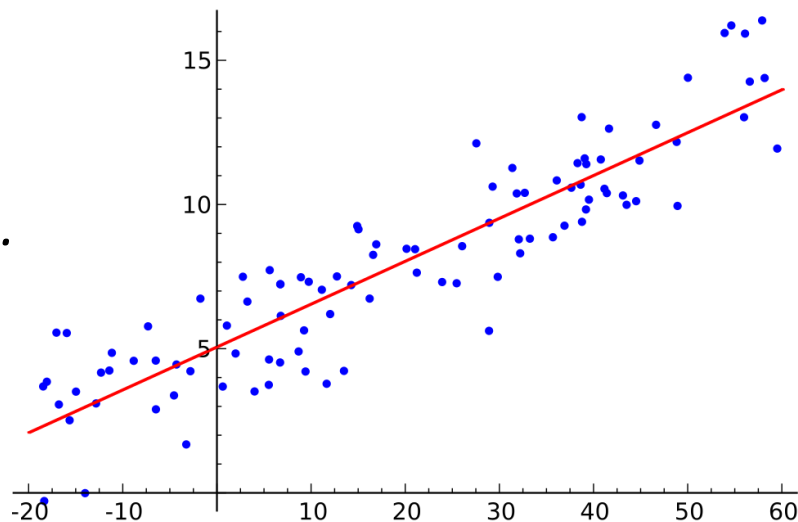
Regulated industries require explainable/interpretable predictions

Models are powered largely by text and structured data

Well understood tools; clients have deeper skill using Python, Spark, or R

Few of these linear models are cluster GPU accelerated, limiting both scale and speed...

Snap ML changes this to accelerate machine learning



Snap ML: Training time goes from 1.1 hours to 90 seconds

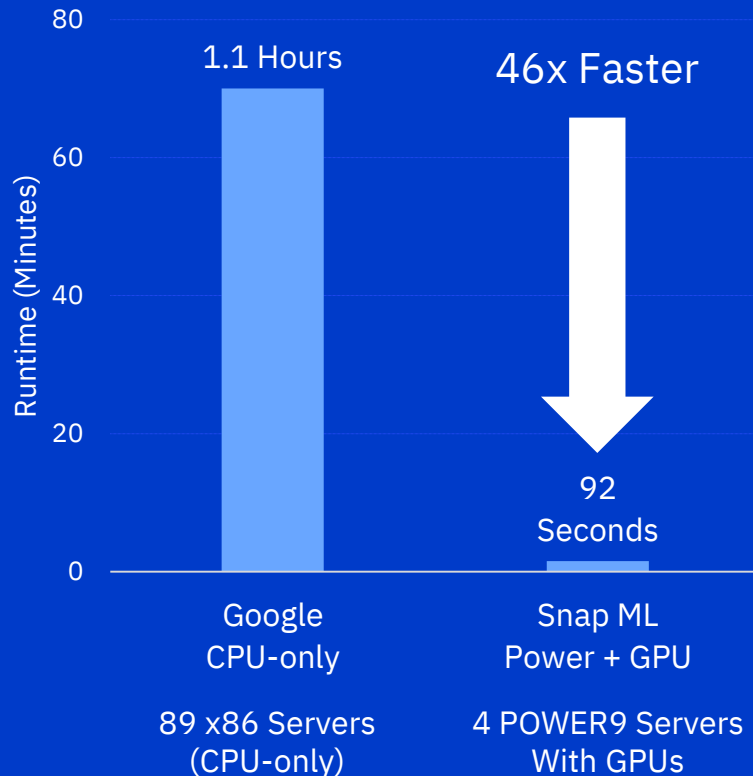
46x faster than previous record set by Google

Transparent to data scientists – plugs into existing machine learning codes

Generalized linear models: linear regression, logistic regression, support vector machines

FAST and EASY to cluster enable machine learning models for GPU and CPU

Logistic Regression in TensorFlow(CPU-only) vs Snap ML (with GPUs)



Simple to get started

Snap ML is available now as a no-charge download from IBM

Runs exclusively on the accelerated IBM Power AC922 or S822LC servers

Can accelerate and scale existing models, with little or no modification

Financial Services Usage Examples:

- Predict credit default **23x** faster than scikit-learn
- Speed up model training for credit card fraud detection: **32x** faster than TensorFlow, **12.5x** faster than scikit-learn
- Predict stock volatility from 10-k textual reports **35x** faster than Apache Spark

Introducing PowerAI / Watson ML Accelerator Enterprise Edition

Enterprise Edition

Deep Learning Impact

Data Management and ETL
Training visualization and monitoring
Hyper-parameter optimization

Spectrum Conductor

Multi-tenancy support & security
User reporting & charge back
Dynamic resource allocation
External data connectors

Distributed Deep Learning (>4 nodes)

Support Line L1-L3

Community Edition

Open Source Frameworks: Supported Distribution



Large Model Support

DDL (up to 4 nodes)

Agenda

- PowerAI / WMLA update and editions
- Deep dive into WMLA Enterprise Edition
- H2O Driverless AI
- Intelligent Video Analytics
- Watson Studio Local
- PowerAI Vision

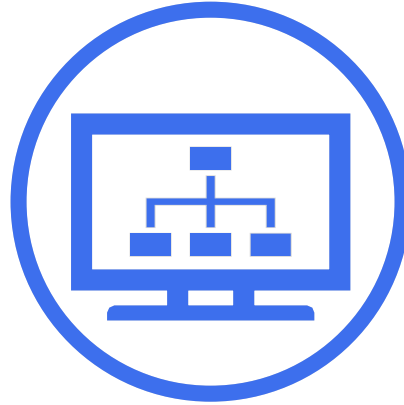
What enterprise customers struggle with in AI



**Faster Time
to Results &
Accuracy**



**Increased
Resource
Utilization**



**Simplified
Management**



**Enterprise
Solution**

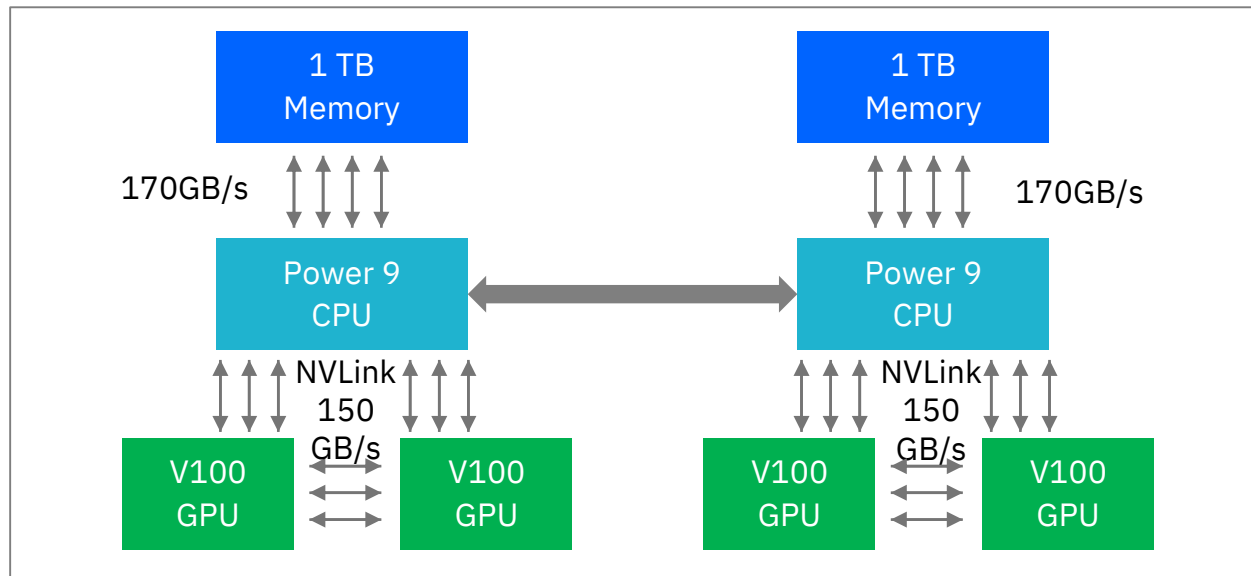
End-to-end NVLink 2.0 connectors

End-to-end NVLink connections

No PCIe bottleneck

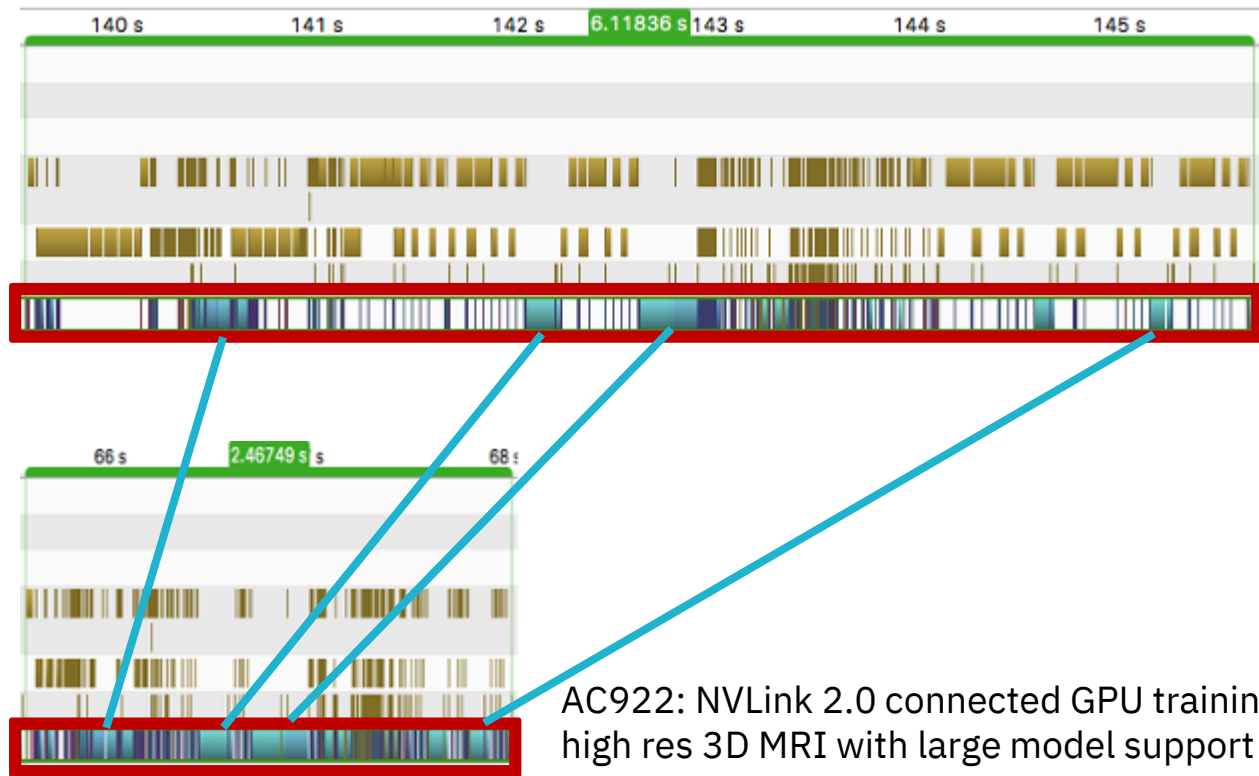


Data location transparent to GPUs



IBM AC922 Power System

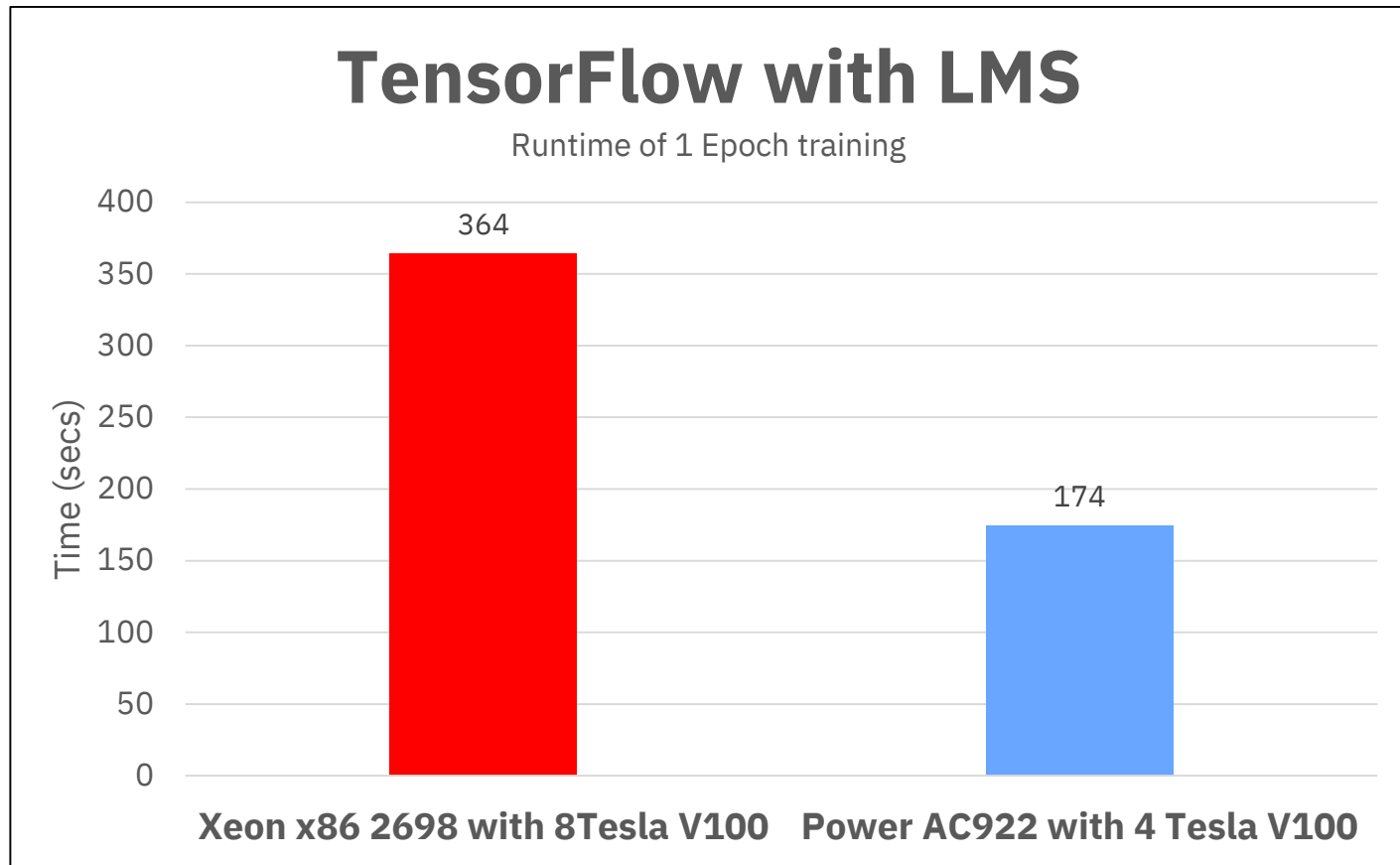
Combining NVLink and LMS



DGXx: PCIe connected GPU training one high res 3D MRI with large model support

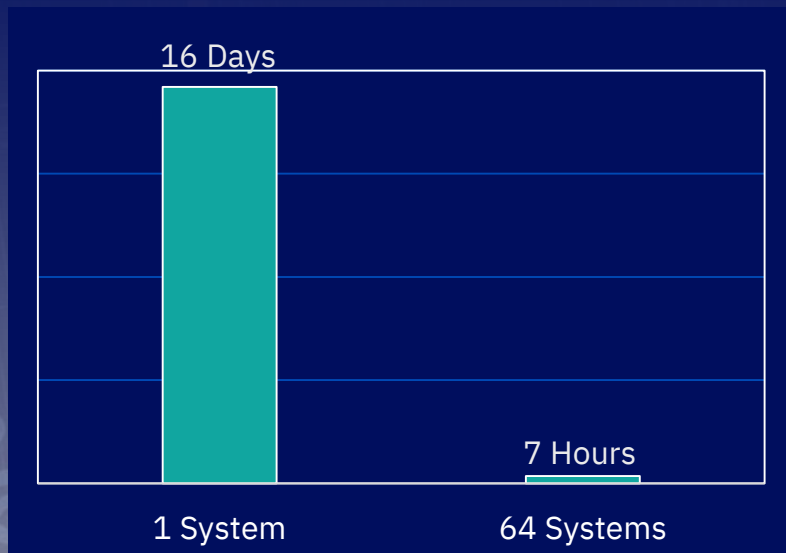
AC922: NVLink 2.0 connected GPU training one high res 3D MRI with large model support

POWER9 with 4 GPUs is 2.1x faster than x86 with 8 GPUs

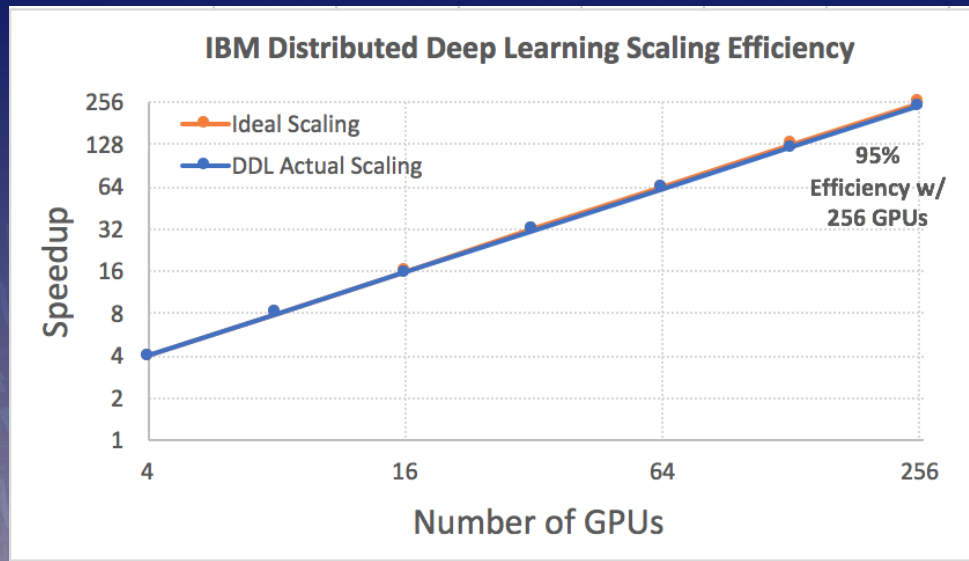


Distributed Deep Learning (DDL)

*16 Days Down to 7 Hours:
58x Faster*



Near Ideal Scaling to 256 GPUs and Beyond



ResNet-101, ImageNet-22K, Caffe with PowerAI DDL, Running on Minsky (S822Lc) Power System

Choose your scale out distributed training model

Distribution Model	Benefit
Bring Your Own Framework & Native Distribution Engines	Frameworks included in the IBM PowerAI distribution (e.g. Tensorflow) and frameworks with their own native distribution engine (e.g. CaffeOnSpark, etc.)
Distributed Deep Learning (DDL) Very Large Scale-out Single Model	Single user, very large distribution for high-performance training
Elastic Distributed Training Resource Sharing & Multitenancy	Concurrent, dynamic and fault tolerant sharing of resources across many tenants and jobs

Hyperparameter search & optimization

- Find the best hyperparameters using cognitive algorithms running in parallel, refining the values as the search progresses
- Supported Algorithms
 - Random Search
 - Bayesian
 - TPE
 - Hyperband
 - More to come...

Tune Hyperparameters for model:Caffe-vgg19-flower

* Hyperparameter search type: Random Search

Tuning Parameter Settings

Input the parameters that will be tuned

* Optimizer (select at least 1):
☒ SGD
☐ AdaDelta
☐ AdaGrad
☐ Adam
☐ Nesterov
☐ RMSProp

* Learning rate range:

* Weight decay range:

Overview					
Hyperparameter Tuning					
Training					
Validation Results					
Framework:	TensorFlow (Distributed training with IBM Fabric and auto-scaling)		Spark instance group:	d1m	
Model files:	/shared/dli/models/TensorFlow/inceptionv3-dong-tuning-20180424083651		Batch size:	32	
Dataset:	flowers-incept				
Hyperparameter					
Learning rate policy:	fixed	Base learning rate:	0.020041713	Learning rate decay:	0
Staircase:	True	Solver type:	Momentum	Momentum:	0.013016915
Decay:	0.1	Epsilon:	1	Maximum iterations:	5000

Inference using REST API

Single Inference is available via REST API in IBM PowerAI Enterprise

New Inference: Eric-elastic-demo-1-20180327103709-20180327182422-Inference

Inference name: Eric-elastic-demo-1-20180327103709-20180327182422-Inference-20180411212505

Threshold: 0.1

* Files for inference: No files selected.

Inferences : Manage deep learning model inferences. [Show/Hide](#) | [List Operations](#) | [Expand Operations](#)

GET	/platform/rest/deeplearning/v1/inferences	Get all inference instances for a model
POST	/platform/rest/deeplearning/v1/inferences	Create a new inference from the model training.
POST	/platform/rest/deeplearning/v1/inferences/startpredict	Start predicting an inference model
GET	/platform/rest/deeplearning/v1/inferences/weightfile	Retrieves the file name of the latest weight file
DELETE	/platform/rest/deeplearning/v1/inferences/{predictName}	Deletes a prediction
GET	/platform/rest/deeplearning/v1/inferences/{predictName}	Get the inference instance details
GET	/platform/rest/deeplearning/v1/inferences/{predictName}/predicts	Get the prediction results for an inference
PUT	/platform/rest/deeplearning/v1/inferences/{predictName}/stop	Stops a prediction

Choose your inference architecture for visual models

Data Center: Train model & Compile to Edge

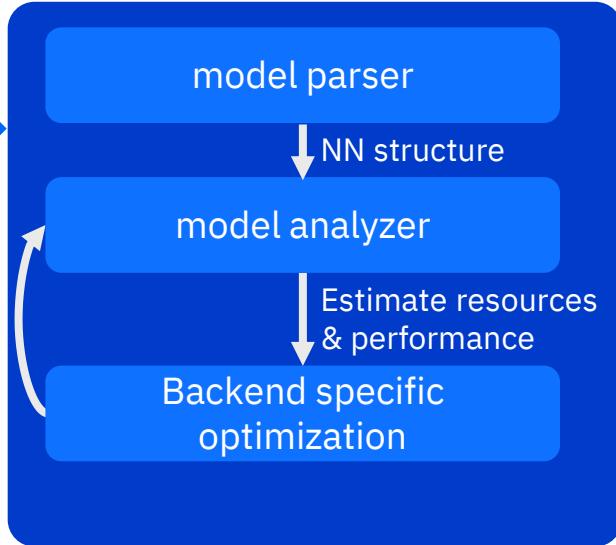
IBM PowerAI Inference Engine



Trained model



CPU + GPU



Map to Different Platforms



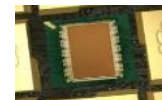
Cloud or Edge



Embedded FPGA



Embedded GPU



Neural network processor

CPUs, GPUs

Data preparation for deep learning

Import from different formats

New Dataset

Create a dataset from:

LMDBs

TensorFlow Records

Images for Object Classification

Images for Object Detection

Images for Vector Output

CSV Files

Other

Cancel

Transform, split and shuffle

New Dataset

Create a dataset from images for object detection.

* Dataset name:

Create in Spark instance group:

dli-sig

* Training folder:

i

The training folder must contain an Object.

* Portion of training images for validation:

%

* Portion of training images for testing:

%

* Split algorithm:

hold-out

☐ Double the number of images in the dataset by creating a resized copy of each existing image

Data preparation for deep learning

Preview Results

voc-partial-data

Overview

State: **Finished** Run duration: **0.0 minutes**

This dataset is generated from Image, CSV or Object detection, run as Spark application.

Dataset details

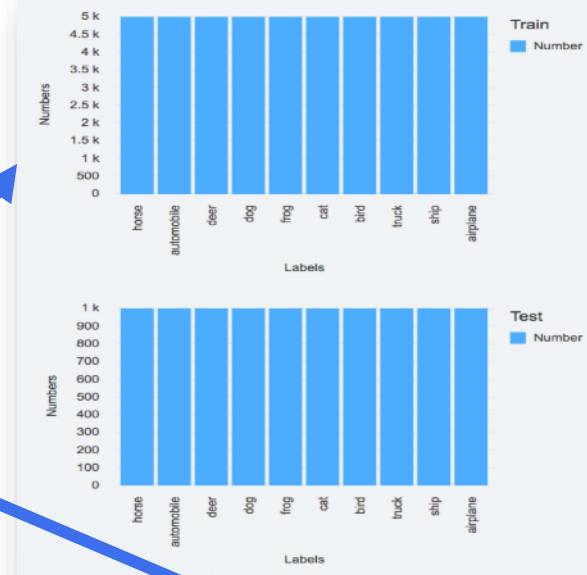
DBBackend: ObjectDetection
Submitted: 6/14/2017, 10:05:26 PM
Training directory: /gpfs/difs1/i004/datasets/voc-partial-data/ImageSets/Main/train.txt
Test directory: /gpfs/difs1/i004/datasets/voc-partial-data/ImageSets/Main/test.txt
Validation directory: /gpfs/difs1/i004/datasets/voc-partial-data/ImageSets/Main/val.txt

Image details

Image type:
Width*Height: 0*0
Resize transformation:
Split algorithm: hold-out

Image Review

[Train Images Review](#) [Test Images Review](#) [Validation Images Review](#)



Training images

Showing 1 to 10 of 434 entries

← 1 2 3 4 5 ... 44 →

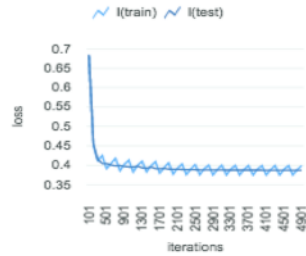
Deep Learning Impact Insight - Training visualization

monitor, analyze, optimize

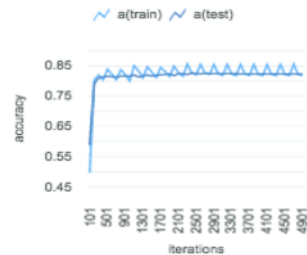
Algorithm's view and Quality Optimization



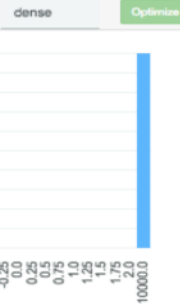
Learning Curves



Learning Curves



Weight histogram



Algorithm's view and Quality Optimization

Time

Optimize

Time Estimation



0

Iteration Eclipse



Current Iteration:4901

Suggestion

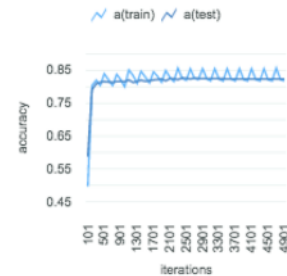
Back

learning_rate=0.15

**Recommended
Learning rate to re-
train the model.**

Learning Curves

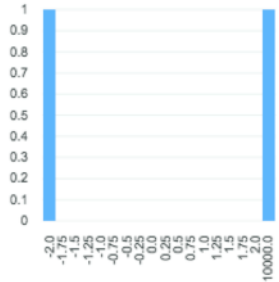
Optimize



Weight histogram

dense

Optimize



Dynamic, policy based resource plan

- Sharing while preserving ownership
- Change the plan 'on the fly' while workload is running
- Allocations flex during runtime to reflect business priorities – Dynamic Allocation
- Enables application level SLA management

The screenshot shows the IBM Platform Symphony Advanced Edition interface. The main window displays the 'Resource Plan' for 'ComputeHosts'. It includes a table with columns for Consumer, Model type: Ownership, Lend | Limit, Borrow | Limit, and Model type: Share. The table lists various consumers like Symphdemo, SymTesting, SampleApplications, SymExec, and MapReduceConsumer, each with its own Lend and Borrow limits. Two blue arrows point from the 'Lend | Limit' and 'Borrow | Limit' columns to the 'Lend Details' and 'Borrow Details' panels on the right.

Resource Plan

Resource Group: ComputeHosts | Time Intervals and Settings

Slot allocation policy

Consumer	Model type: Ownership	Consumer Rank	Lend Limit	Borrow Limit	Model type: Share
▼ symphdemo	Owned Slots: 108				
▼ SymTesting					
■ Symping61	0	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1
Total	0	-	-	-	-
Balance	0	-	-	-	-
▼ SampleApplications					
■ SOASamples	0	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1
■ EclipseSamples	0	50	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1
Total	0	-	-	-	-
Balance	0	-	-	-	-
▼ SymExec					
■ SymExec61	0	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1
Total	0	-	-	-	-
Balance	0	-	-	-	-
▼ MapReduceConsumer					
■ MapReduce61	40	0	<input checked="" type="checkbox"/> Details	<input checked="" type="checkbox"/> Details	<input checked="" type="checkbox"/> 1
■ MapReduceHighPriority	10	0	<input checked="" type="checkbox"/> Details	<input checked="" type="checkbox"/> Details	<input checked="" type="checkbox"/> 1
■ MapReduceDefault	30	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1
Total	40	-	-	-	-
Balance	0	-	-	-	-
■ SampleAppCPP	0	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1
■ GpuTestApp	0	0	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/> 1
Total	40	-	-	-	-
Balance	68	-	-	-	-

Lend Details:

MapReduce61 (ComputeHosts, 00:00-24:00)

Total lend limit:

Lend to these consumers

Consumers to lend to	Lend / Limit
▼ symphdemo	
▼ SymTesting	
■ Symping61	<input type="checkbox"/>
▼ SampleApplications	
■ SOASamples	<input type="checkbox"/>
■ EclipseSamples	<input type="checkbox"/>
▼ SymExec	
■ SymExec61	<input type="checkbox"/>
▼ MapReduceConsumer	
■ MapReduce61	<input checked="" type="checkbox"/>
■ MapReduceHighPriority	<input checked="" type="checkbox"/>
■ MapReduceDefault	<input checked="" type="checkbox"/>
■ SampleAppCPP	<input type="checkbox"/>
■ GpuTestApp	<input type="checkbox"/>

Expand All | Collapse All

Apply | Revert | Close

Borrow Details:

MapReduceDefault (ComputeHosts, 00:00-24:00)

Total borrow limit:

Borrow from consumers

Consumers to borrow from	Borrow / Order
▼ symphdemo	
▼ SymTesting	
■ Symping61	<input type="checkbox"/>
▼ SampleApplications	
■ SOASamples	<input type="checkbox"/>
■ EclipseSamples	<input type="checkbox"/>
▼ SymExec	
■ SymExec61	<input type="checkbox"/>
▼ MapReduceConsumer	
■ MapReduce61	<input checked="" type="checkbox"/> 1
■ MapReduceHighPriority	<input checked="" type="checkbox"/> 2
■ MapReduceDefault	<input type="checkbox"/>
■ SampleAppCPP	<input type="checkbox"/>
■ GpuTestApp	<input type="checkbox"/>

Expand All | Collapse All

Apply | Revert | Close

Elastic Distributed Training – Example

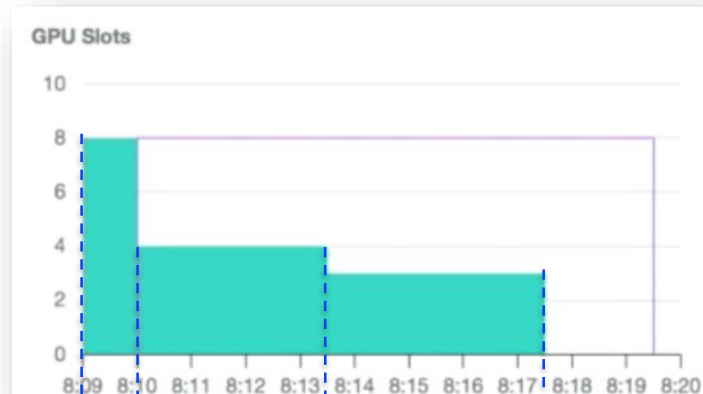
Environment :

- Two (2) POWER9 servers with four (4) GPUs
- Eight (8) GPUs total
- Policies
 - Fairshare
 - Preemption
 - Priority

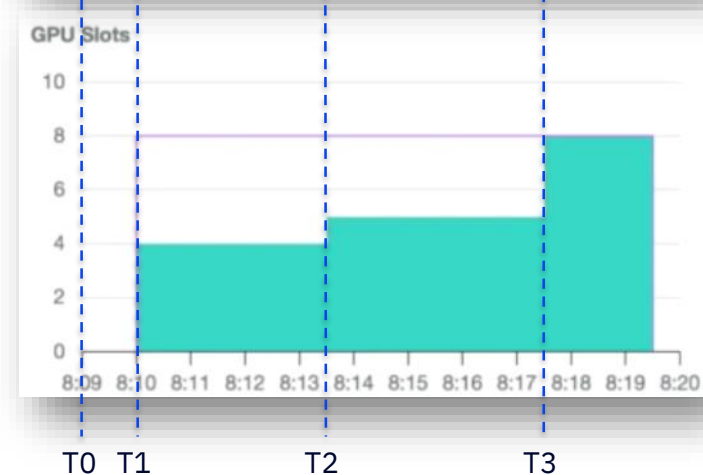
Timeline :

- **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

Job #1

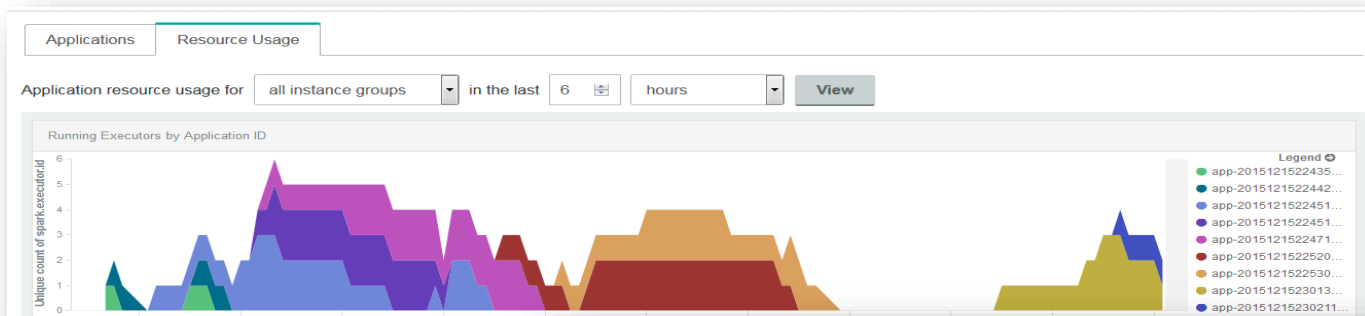


Job #2

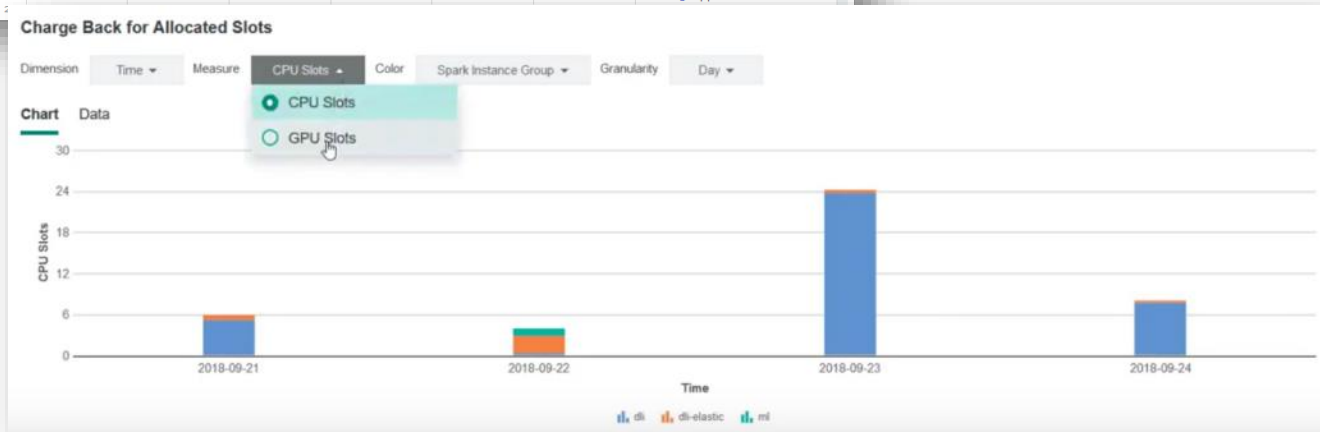


Monitor and chargeback accounting

PowerAI Enterprise includes detailed monitoring



PowerAI resource
chargeback reporting



Production proven with enterprise security



Authentication

Production proven support for Kerberos, Siteminder, AD/LDAP and OS authentication



Impersonation

Allow different lines of business to define production execution user



Authorization

Fine grained access control

Role based control (RBAC)

Spark binary life cycle, notebook updates, deployments, resource plan, reporting, monitoring, log retrieval, and execution



Encryption

SSL & daemon authentication

Storage encryption with IBM Spectrum Scale

Agenda

- PowerAI / WMLA update and editions
- Deep dive into WMLA Enterprise Edition
- **H20 Driverless AI**
- Intelligent Video Analytics
- Watson Studio Local
- PowerAI Vision

H2O.ai Company

Company	Founded in Silicon Valley in 2012 Series C Investors: Wells Fargo, NVIDIA, Nexus Ventures, Paxion Ventures
Products	<ul style="list-style-type: none">• H2O Open Source Machine Learning (14,000 organizations)• H2O Driverless AI – Automatic Machine Learning
Leadership	Market Leader recognized by Gartner, Forrester, InfoWorld, Constellation Research
Team	130+ AI experts (Kaggle Grandmasters, Distributed Computing and Visualization experts)
Global	Mountain View, London, Prague, Chennai



Growing Worldwide Open Source Community



14,000 Companies
using H2O



155,000
data scientists

222 OF THE **500** FORTUNE
 **H₂O**

8 OF TOP 10
BANKS

7 OF TOP 10
INSURANCE COMPANIES

4 OF TOP 10
HEALTHCARE COMPANIES



H2O World
NYC, London, SF



130K Meet-up Members



H2O.ai Product Suite

H₂O-3

In-memory, distributed
machine learning algorithms
with H2O Flow GUI

Spark + H₂O
SPARKLING
WATER

H2O AI open source engine
integration with Spark

H₂O4GPU

GPU-accelerated
machine learning package

DRIVERLESSAI

Automatic feature
engineering, machine
learning and interpretability

Open Source

- 100% open source – Apache V2 licensed
- Built for data scientists – interface using R, Python on H2O Flow (interactive notebook interface)
- Enterprise Support subscriptions

- Built for domain users, analysts and data scientists – GUI based interface for end-to-end data science
- Fully automated machine learning from ingest to deployment
- Licensed on a per seat basis (annual subscription)

H2O.ai is a Recognized Leader in AI and ML

2018 Gartner Magic Quadrant for Data Science and Machine Learning Platforms

Figure 1: Magic Quadrant for Data Science and Machine-Learning Platforms



“Technology leadership ... with a distinguished vision”

“the quasi-industry standard”

Forrester Wave: Notebook-Based Predictive Analytics And Machine Learning Solutions, Q3 2018



“H2O.ai’s future is automated machine learning”

“its bright future is in Driverless AI”

Top 3 Artificial Intelligence (AI) and Machine Learning (ML) Software Solution

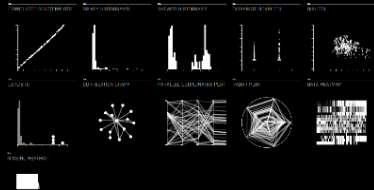


“its vision of creating an AI and ML tool that ultimately aims to allow almost everyone within the business to create their own predictive models”

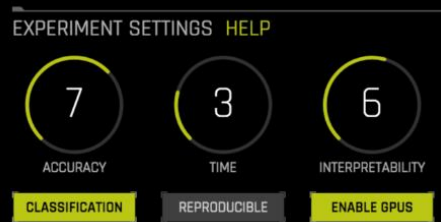
H2O Driverless AI – Simple, Fast, Accurate, Interpretable

H₂O.ai

Automatic Data Visualization



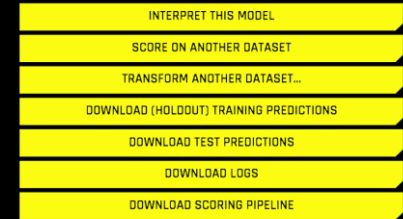
Fast and Accurate Results



Industry Leading Interpretability



Easy Deployment for Low Latency Models



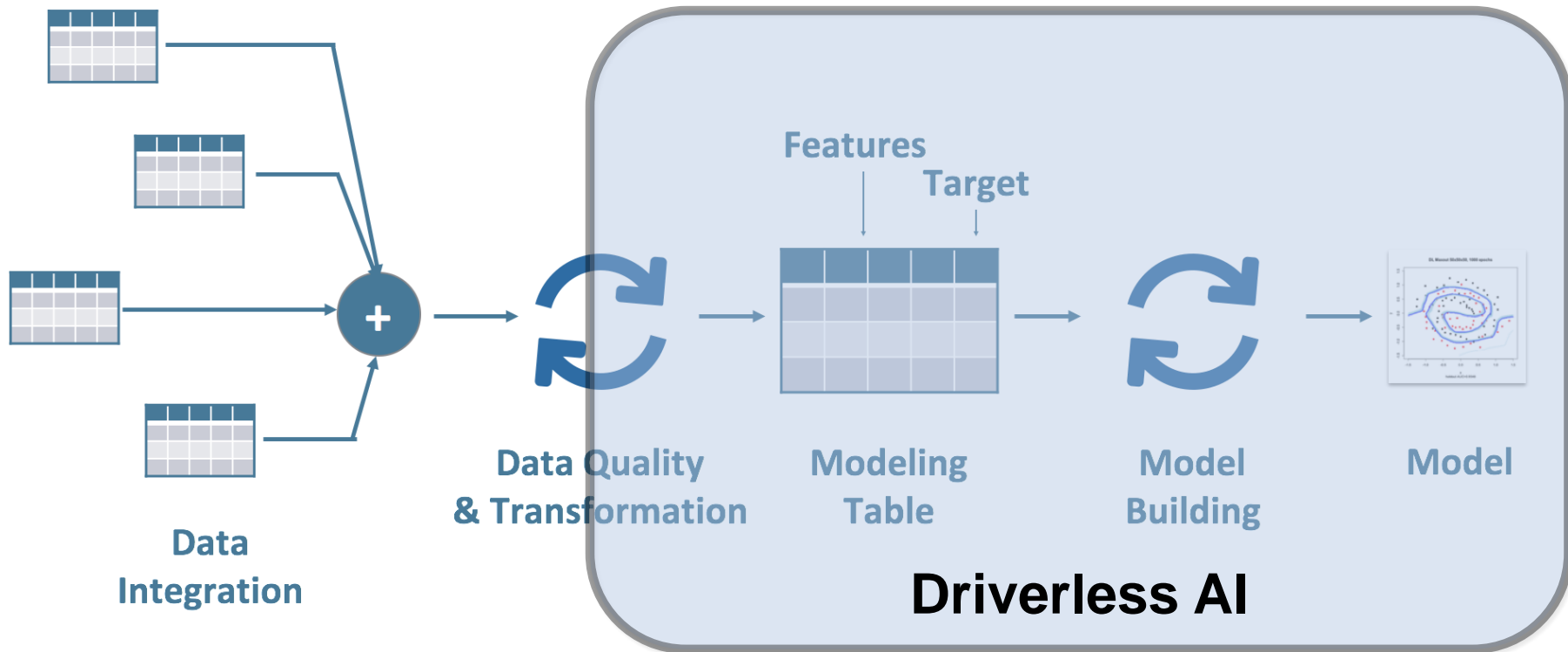
- Automatic generation of visualizations and graphs to explore your data before the model-building process
- Most relevant graphs shown for the given data set
- Identify outliers and missing values

- “Data Scientist in a Box”
- Simple interface
- Automatic feature engineering to increase accuracy
- Automatic recipes for solving wide variety of use-cases
- Automatic tuning to find and tune the right ensemble of models

- Trusted results with explainability and transparency
- Interpretability for debugging, not just for regulators
- Get reason codes and model interpretability in plain English
- K-Lime, LOCO, partial dependence and more

- Production-ready, stand-alone scoring pipelines that are easy for IT to deploy and manage
- Python and Java
- Streamlined scoring code to deploy on any device: on the edge, mobile, ...
- Very fast (milliseconds) to satisfy today’s real-time apps

Driverless AI: Automates Data Science and Machine Learning Workflows



The Driverless AI Experience

< H2O.ai Experiment **desusupe**

DRIVERLESS AI 1.3.0 - AI TO DO AI

Licensed to IBM (SN26193 - For evaluation only, not for production use)

TRAINING DATA

DATASET

creditcard.csv

ROWS

24K

COLUMNS

25

DROPPED COLS

--

VALIDATION DATASET

--

TEST DATASET

--

TARGET COLUMN

default payment next

FOLD COLUMN

--

WEIGHT COLUMN

--

TIME COLUMN

[OFF]

TYPE

bool

COUNT

23999

UNIQUE

2

TARGET FREQ

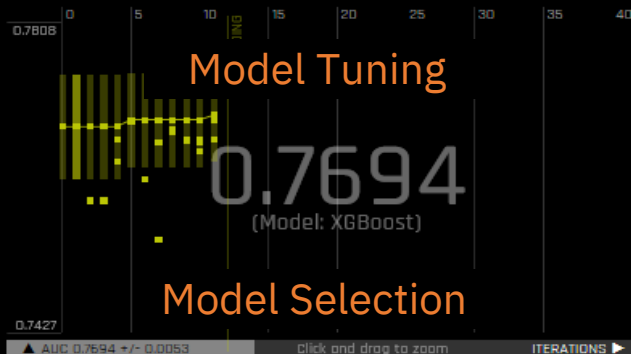
5369

ASSISTANT

TRAINED 4/4 ENSEMBLE BASE LEARNERS
[XGBOOST]



ITERATION DATA - VALIDATION



VARIABLE IMPORTANCE

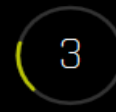
43_NumToCotTE:MARRIAGE:PAY_0:PAY_2:0	1.00
42_TruncSVD:MARRIAGE:PAY_0:PAY_2:PAY_3:0	0.71
47_TruncSVD:PAY_0:PAY_3:PAY_AMT5:1	0.57
48_NumToCotWoE:PAY_0:PAY_5:PAY_AMT2:0	0.35
10_PAY_0	0.30
27_ClusterDist6:PAY_0:5	0.16
27_L	0.15
32_L	0.13
50_NumToCotTE:PAY_2:PAY_5:0	0.11
59_NumToCotTE:LIMIT_BAL:PAY_4:0	0.10
57_ClusterTE:ClusterID76:BILL_AMT3:PAY_0:0	0.09
31_ClusterDist6:BILL_AMT1:4	0.08
63_NumToCotTE:BILL_AMT2:PAY_0:PAY_2:PAY_3:PAY_5:PA...	0.07
26_NumToCotWoE:LIMIT_BAL:PAY_4:0	0.07

[DATASETS](#) [EXPERIMENTS](#) [MLI](#) [HELP](#) [PY_CLIENT](#) [MOJO2-RUNTIME](#) [MESSAGES\[0\]](#) [LOGOUT](#) [KSchlamb](#)

EXPERIMENT SETTINGS



ACCURACY



TIME



INTERPRETABILITY

CLASSIFICATION

REPRODUCIBLE

ENABLE GPUS

SCORER

GINI

MCC

F05

F1

F2

ACCURACY

LOGLOSS

AUC

AUCPR

Quickly Start Experiment

[Options](#) [Log](#) [Trace](#)

CPU

MEM

ROC

PREC-RECALL

LIFT

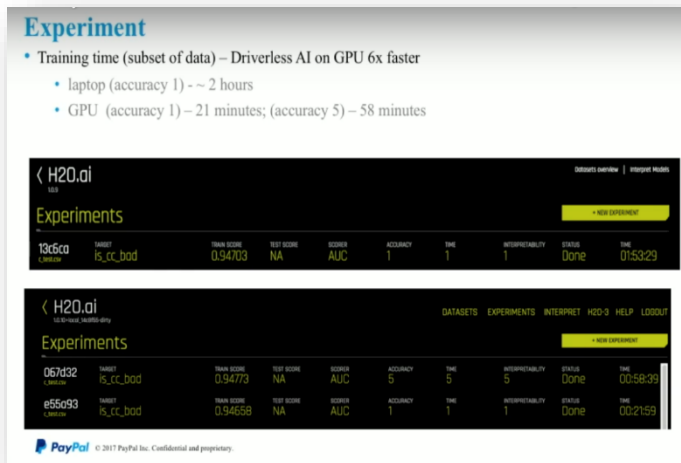
GAINS

GPU USAGE

GPU1

GPU2

- Driverless AI matched **10 years** of expert feature engineering
- Increased accuracy from **0.89 to 0.947 (6%)** in detecting fraudulent activity
- **6X** speed up when running on an IBM Power GPU-based server



Experiment

- Training time (subset of data) – Driverless AI on GPU 6x faster
 - laptop (accuracy 1) – ~ 2 hours
 - GPU (accuracy 1) – 21 minutes; (accuracy 5) – 58 minutes

Model	Train Score	Test Score	Score	Accuracy	Time	Interpretability	Status	Time
13c5c0	0.94703	NA	AUC	1	1	1	Done	01:53:29
067d32	0.94773	NA	AUC	5	5	5	Done	00:58:39
e55e93	0.94658	NA	AUC	1	1	1	Done	00:21:59

PayPal © 2017 PayPal Inc. Confidential and proprietary.

“Driverless AI is giving amazing results in terms of feature and model performance”

Venkatesh Ramanathan
Senior Data Scientist, PayPal

- Improved **years** of expert feature engineering
- **Increased accuracy** of existing credit risk scoring in **less time**
- **2x propensity to buy** for new bank products
- Accelerated by IBM Power Systems AC922



“We were also able to double the propensity for our banking customers to accept an offer of credit products, such as credit cards... We plan to use the platform for more use cases in the future.”

Ruben Diaz
Data Scientist,
Visión Banco

Integration with WMLA EE

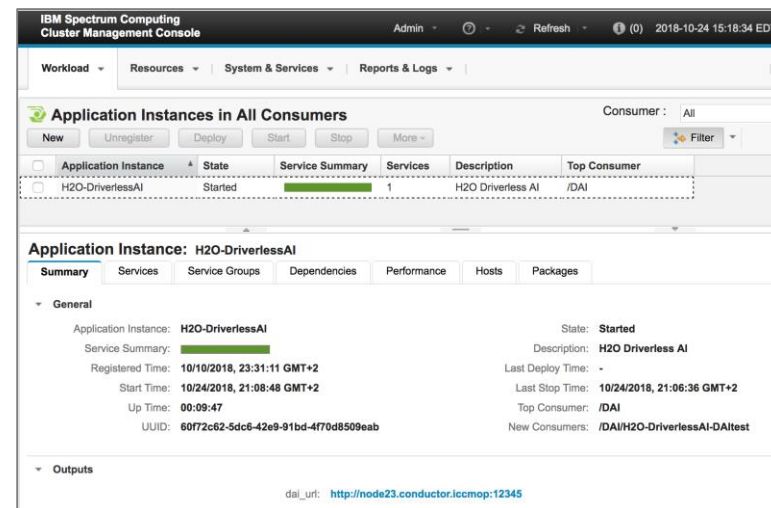
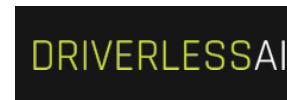
Deploy and manage instances of H2O Driverless AI with PowerAI Enterprise (through IBM Spectrum Conductor)

What's available today (as of Feb. 2019)

- Multiple instances of Driverless AI deployed across the cluster (up to one instance per host)
- Automatic failover: if Driverless AI goes down, it is restarted on another host
- Easy start/stop of Driverless AI through the web interface
- User permissions management
- Shared file system for data and logs

Roadmap

- Running Driverless AI across multiple machines
- Running multiple instances on a single host
- Integration of Driverless AI authentication with PowerAI Enterprise authentication
- Log retrieval from IBM Spectrum Conductor web interface
- ...



Agenda

- PowerAI / WMLA update and editions
- Deep dive into WMLA Enterprise Edition
- H2O Driverless AI
- **Intelligent Video Analytics**
- Watson Studio Local
- PowerAI Vision

IBM Intelligent Video Analytics



SAFETY & SECURITY

Identify and monitor people and objects to improve public safety.

Key Capabilities:

- Facial recognition
- Object missing / left behind
- Trip wires / Intrusion detection
- Detect potential weapons
- People count and movement
- Post Event video analysis (forensics)



WORKPLACE SAFETY

Spot safety risks and ensure compliance to workplace regulations.

Key Capabilities:

- Safety equipment compliance – hats, vests, suits, gloves etc.
- Heat Maps
- Detection of workplace risks (spills, fallen objects, animals, etc.)
- People count
- Crowd formation



COMMERCIAL/RETAIL ANALYTICS

Analyze demographics and patterns of movement to gain insights.

Key Capabilities:

- Heat map
- Track summary
- Demographics
- Anomalous direction
- Trip wires, complex alerting
- Inventory management

Agenda

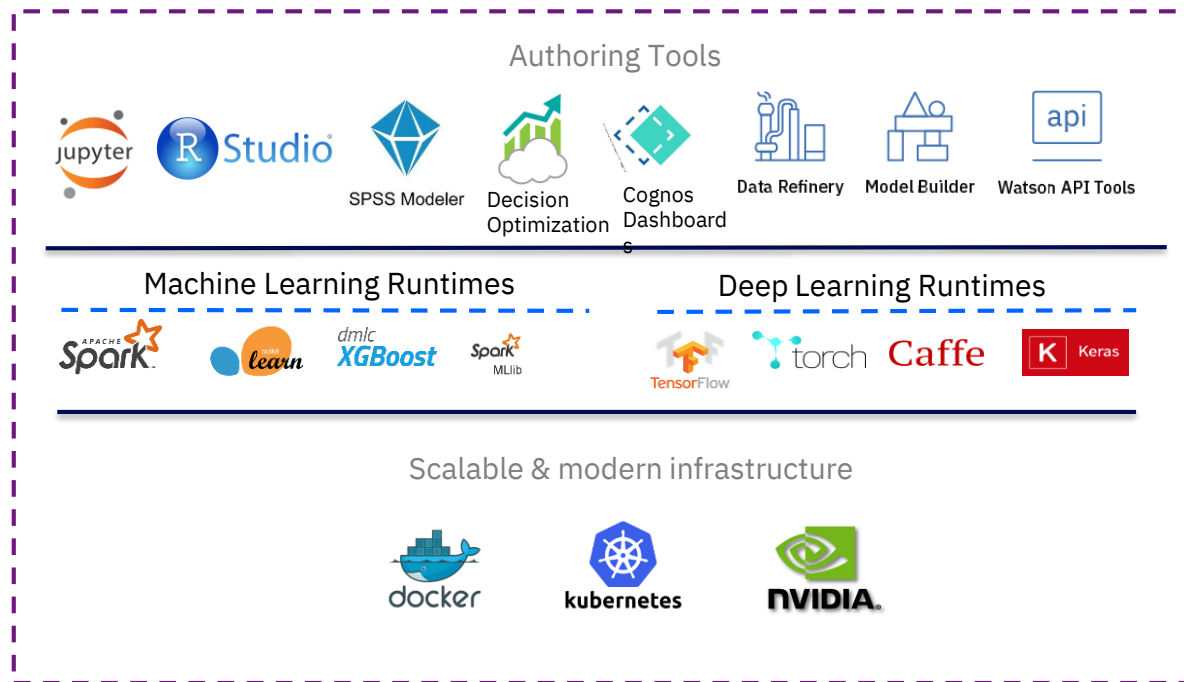
- PowerAI / WMLA update and editions
- Deep dive into WMLA Enterprise Edition
- H2O Driverless AI
- Intelligent Video Analytics
- **Watson Studio Local**
- PowerAI Vision

Watson Studio : Tools for supporting data scientists

- Create, train and collaborate
- Best of breed open source & IBM tools
- Code (R, Python or Scala) and no-code/visual modeling tools

- Most popular open source frameworks
- IBM best-in-class frameworks

- Fully managed or customer managed service
- Container-based resource management
- Elastic pay as you go cpu/gpu power



Deployment options:



Desktop - Windows & macOS


Fully Managed on **IBM Cloud**

Local – Deploy it anywhere





Private Cloud - Embedded in ICP4D

Watson Studio user interface

 IBM Data Science Experience Local


▼ 58 Trial Days Left 

Projects > DSX_Local_Workshop_KG > All




All Notebooks RStudio Models SPSS Modeler Streams Scripts Data Sets Other Files Published Assets

Notebooks [view all \(10\)](#)

 [add notebook](#)

NAME	STATUS	ENVIRONMENT	TOOL	LAST MODIFIED	
 PMMLTestClient		Jupyter with Python 2.7, Scala 2.11, R 3.4.1	JUPYTER	03-20-2018	
 TelcoChurn_Deploy		Jupyter with Python 2.7, Scala 2.11, R 3.4.1	JUPYTER	03-19-2018	
 TelcoChurnEvalScript		Jupyter with Python 2.7, Scala 2.11, R 3.4.1	JUPYTER	03-19-2018	
 TelcoChurn_SparkML		Jupyter with Python 2.7, Scala 2.11, R 3.4.1	JUPYTER	03-19-2018	
 TelcoChurn_SparkML			JUPYTER	03-19-2018	

RStudio [view all \(0\)](#)

 [open RStudio](#)

NAME	TYPE	LAST MODIFIED
------	------	---------------

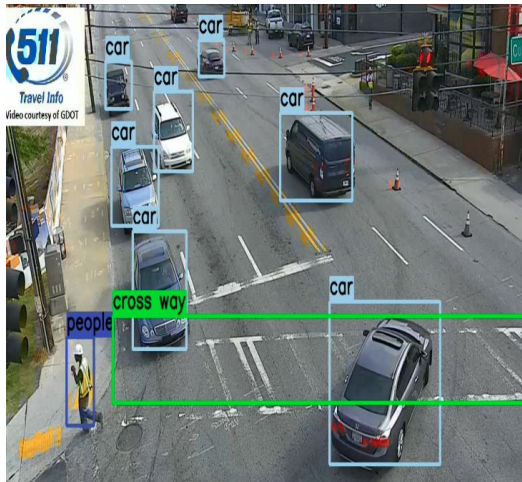
you have no rstudio files

Agenda

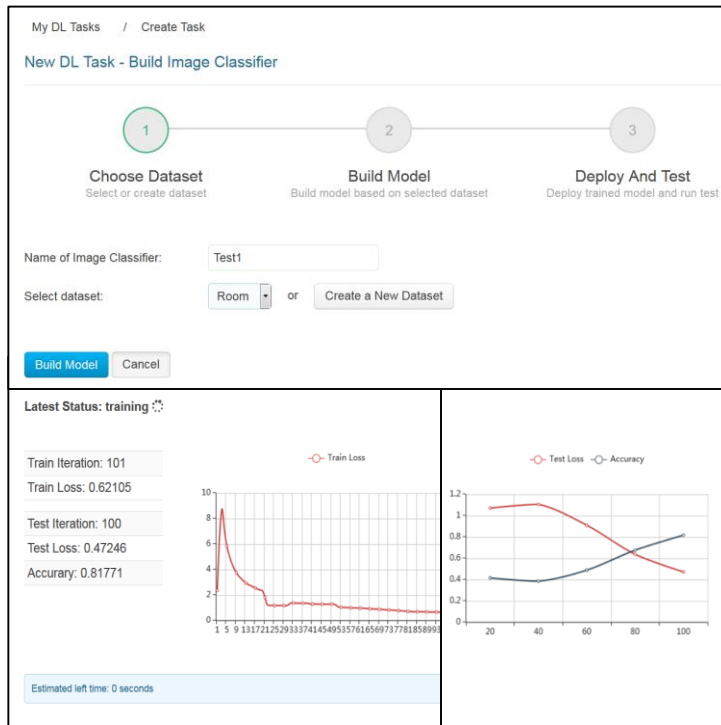
- PowerAI / WMLA update and editions
- Deep dive into WMLA Enterprise Edition
- H2O Driverless AI
- Intelligent Video Analytics
- Watson Studio Local
- **PowerAI Vision**

PowerAI Vision : deep learning for the rest of us

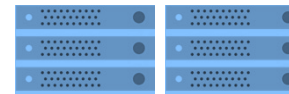
Label Image or Video Data



Auto-Train AI Model



Package & Deploy AI Model



AI SW portfolio on POWER Systems : the full picture

AI for
Data Scientists and
non-Data Scientists

**Watson
Studio
Local**

Data scientist
toolkit

**Intelligent
Video
Analytics**

Video streams
insights

**PowerAI
Vision**

Auto-DL for
Images & Video

**H2O
Driverless AI**

Auto-ML for Text
& Numeric Data,
NLP

Watson ML
Accelerator CE

PowerAI: Open Source ML Frameworks



TensorFlow™

Caffe



PyTorch

Large Model Support (LMS)

Distributed Deep Learning
(up to 4 nodes)

Watson ML
Accelerator EE

Distributed Deep Learning
(DDL – 1000s of nodes)

Auto Hyper-parameter
Tuning

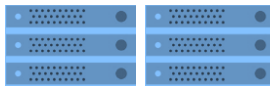
IBM Spectrum Conductor with Spark

Cluster Virtualization,
Dynamic Resource Orchestration,
Multiple Frameworks, Distributed Execution Engine

**Deep Learning Impact
(DLI) Module**

Data & Model
Management, ETL,
Visualize, Advise

Accelerated
Infrastructure



Accelerated Servers



Storage

Comparing AI offerings on Power Systems

		Deep Learning			ML and DL	Machine Learning
		Power AI Base	Power AI Enterprise	AI Vision	Watson Studio Local	H2O Driverless AI
Offering	Description	Deep Learning	Deep Learning for the Enterprise	Deep Learning with Video tools	Notebook oriented development environment for ML and DL	Automated Machine learning
	Pricing Model	Free download	Commercial	Commercial	Commercial	Commercial
	Support	Available from IBM	IBM L 1-3 Included	IBM L1-3 Included	Available from IBM	H2O L 1-3
Applications	Text & Numeric	Yes	Yes	No	Yes	Yes
	Images	Yes	Yes	Yes	Yes	No
	Video	-	Optional add-on	Yes		No
	Primary Persona	Data Scientist	Data Scientist	Line of Business	Data Scientist	Data Scientist
	Second persona	IT	IT	IT	IT	Line of Business
	User Skill Level	High	Medium to high	Low	Medium to high	Low to Medium
	Strengths	Rapid deployment, high performance, scale	enterprise grade, High performance, rapid Deployment	Rapid deployment, simple GUI high performance	Notebook based development environment, strong collaboration, model management	Simplified deployment, intuitive user interface, automatic pipelines, "explainability" for models, end to end automation
Platform	Distributed DL (DDL)	1-4 nodes	1-thousands of nodes	Coming	Coming	-
	Large Model Support	Yes	Yes	Coming	Coming	-
	Server(s)	S822LC or AC922	S822LC or AC922	S822LC or AC922	S822LC or AC922, LC922	S822LC, AC922, LC921/922
IBM Products						
	Spectrum MPI (DDL)	Limited to 4 nodes	Included	?		Optional add-on
	Spectrum Conductor DLI	Optional add-on	Included	?	Optional Add On	Optional add-on
	IBM DSX Local	Optional add-on	Optional add-on	No		Optional add-on
Cloud	IBM Cloud Public	Yes	No	Trial only	Watson Studio	?
	IBM Cloud Private	Yes	Coming	Yes	Yes	Coming

Great reading

Draft Document for Review February 15, 2019 12:17 pm: 50204-5469-00



IBM PowerAI: Deep Learning Unleashed on IBM Power Systems

Dino Quintero
Bing He
Bruno C. Faria
Alfonso Jara
Chris Parsons
Shota Tsukamoto
Richard Wale



Analytics

Power Systems

IBM is a registered trademark of International Business Machines Corporation. All other trademarks are the property of their respective owners.



AI and Big Data on IBM Power Systems Servers

Harish R. Rajagopal
Aravind Subramanian
Robert Fendley de Lima
Srinivas Murthy Reddy
Srinivas Venkatesh
Srinivas Venkatesh
Srinivas Venkatesh



Analytics

Power Systems

IBM

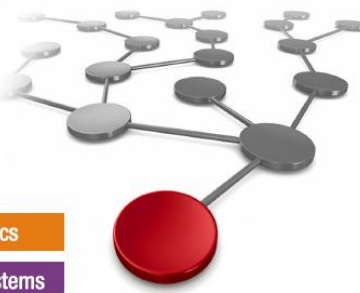
Redbooks

Draft Document for Review February 15, 2019 12:00 pm: REDP-5472-00



IBM Power System AC922 Introduction and Technical Overview

Alexandre Bicas Caldeira



Analytics

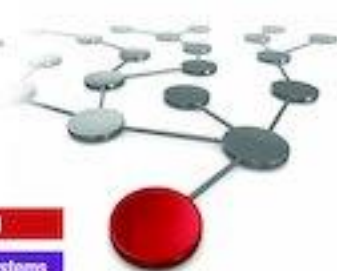
Power Systems

IBM is a registered trademark of International Business Machines Corporation. All other trademarks are the property of their respective owners.



Networking Design for HPC and AI on IBM Power Systems

Youssef Foad
Waqar Foad
Hassan Foad



Cloud

Power Systems

IBM

Redpaper

Free download on <http://www.redbooks.ibm.com/>

Great links

- [Watson ML Accelerator homepage](#)
- [Watson ML Accelerator for developers](#)
- [Try Watson ML Accelerator](#)
- [PowerAI Vision homepage](#)
- [Download PowerAI Vision](#)
- [Watson ML Accelerator & PowerAI Vision FAQ](#)
- [AC922 cognitive system](#)

