Exploring AI with IBM Z

Marcus Kraft
Program Manager Linux on Z Distributions & Ai
marcus.kraft1@ibm.com

zExpert Forum 98

Wislikofen, Switzerland

2024-04-09

Agenda

1 Use scenarios

2 Technical background

3 Where to start ?

Client use cases

FRAUD DETECTION

Business challenge
A large US bank was unable to score all transactions in real-time with an off-platform scoring engine due to network latency and inability to scale; as a result, 80% of transactions went unscored.

Business impact
100% real-time scoring
>20M$ in savings per year

Capabilities leveraged

GO python™

LightGBM

CLEARING & SETTLEMENT

Business challenge
A card processor wants to recognize high risk transactions and trades. AI analyzes risk patterns of a transaction that might fail and provide reason and remedies to fix the errors and missing values.

Business impact
83% reduction in failed trades
100s of saved employee hours

Capabilities leveraged

GO TensorFlow

RETAIL CRIME DETECTION

Business challenge
A large US retailer lacked the ability to fully model for risk as they only have data for their own business. Use of federated learning enabled them to share intelligence and models without exposing sensitive data.

Business impact
6x improved crime detection
72% fewer false positives

Capabilities leveraged

Watson Federated Learning

IBM Cloud Pak for Data

CHATBOT SERVICES

Business challenge
An IT services provider needs to automate vehicle registration process for European citizens. Chatbot enabled automation of repetitive task and offer personalized user experience to help solve their queries.

Business impact
26% reduced operation cost
65% boost in user satisfaction

Capabilities leveraged

OPENSHIFT spaCy

RASA TensorFlow

Client use cases

MEDICAL IMAGING

Business challenge
A large US health insurer needs to analyze large volume of medical records in near real time to validate process claims. Trained computer vision and deployed Deep Learning image analysis models.

Business impact
4x increased in fraud detection
83% increase in accuracy

Capabilities leveraged

PyTorch
ONNX

CLIMATE CHANGE IMPACT

Business challenge
A European environmental agency seeks to understand impact of climate change on coastal ecosystem by analyzing photos taken by field agents. Deployed DL image processing models for faster analytics.

Business impact
41x reduced energy consumed compared to x86 alternatives

Capabilities leveraged

TensorFlow
ONNX

Other explorations

AIRCRAFT ASSESSMENT

Business challenge
It is critical for airline industry to predict remaining cycles of an aircraft without any failure. But this requires analyzing extremely large volume of data at scale, to detect patterns and anomalies and predict risks.

Business impact
Proactively stop losses, lower operational, compliance costs.

Capabilities leveraged

TensorFlow
IBM Cloud Pak for Data
ONNX

SENTIMENT ANALYSIS

Business challenge
TripAdvisor study shows 81% of people frequently read reviews before booking hotel. Sentiment analysis and NLP helps hotel owners identify critical and appraising factors amongst customers.

Business impact
Proactively identify and fix gaps to improve customer exp.

Capabilities leveraged

OpenCV
IBM Cloud Pak for Data
ONNX

Major US health organization leveraging LinuxONE 4 for AI-based cancer research applications

LinuxONE 4 testing with Telum chip in January 2023 was 75% faster compared to GPU server

Client Challenge

- Client is the health solution and care delivery provider in US. They provide data, analytics, research, consulting and technology to hospitals, physicians and health plans. Client pharmacy services fills more than 1.5 billion prescriptions annually.
- Client datacenters were full and power usage at maximum. It needed to find solutions to scale its applications, while deferring or avoiding the expense of building a new datacenter.
- Client was encountering scalability issues on its x86 servers with GPUs, especially for new AI apps, such as pre-processing biopsy images for prostate cancer detection.
- The client was considering migrating workloads to Google Anthos cloud container platform.

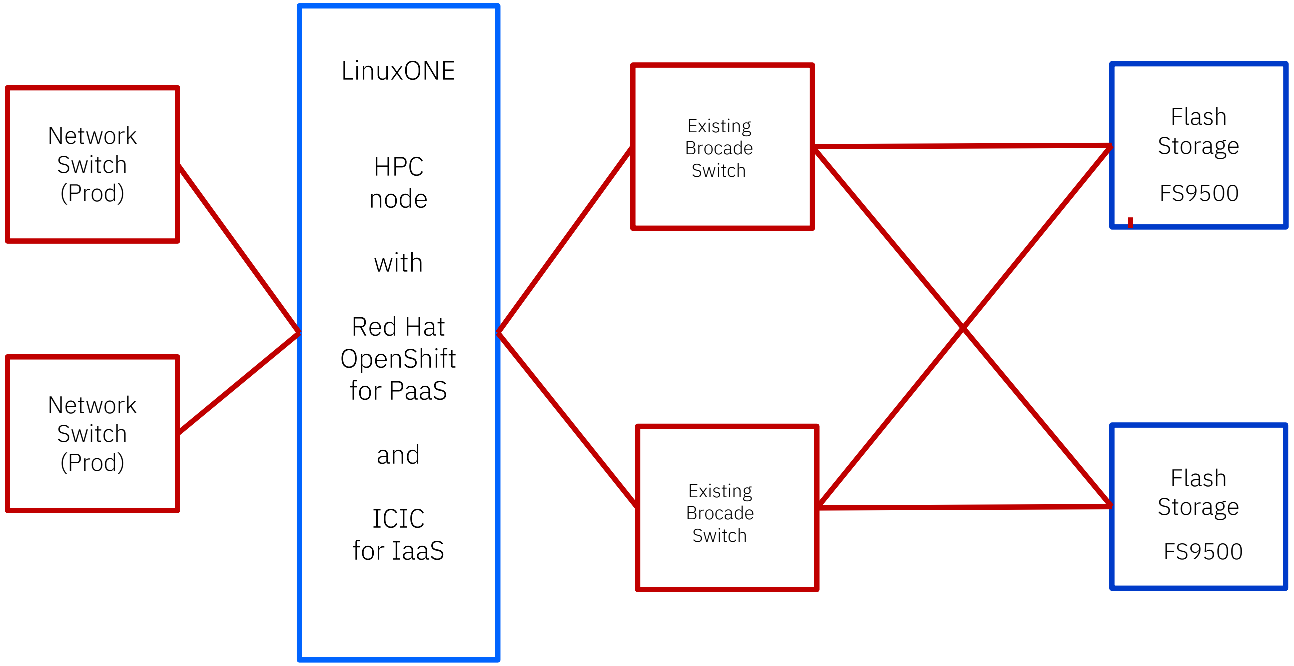
IBM Solution

- IBM proposed a solution based on LinuxONE with Red Hat OpenShift. The LinuxONE design consolidated workloads from x86/GPU servers, reducing its datacenter footprint and energy consumption.
- A hybrid cloud deployment plan was developed between IBM Expert Labs and the client using a combination of PaaS and IaaS models, offering AI application teams options for either self management or managed services.
- The Expert Labs team implemented two LinuxONE IV machines with RoCE, Open Systems Adapter (OSA), internal NVMe and FS9500 storage. Multiple virtual machines were configured on the LinuxONE servers to service different workload types, including Red Hat OpenShift as a foundation for for PaaS, IBM Cloud Infrastructure Center (ICIC) for IaaS.
- The complex implementation was a first of a kind with a promoting a high degree of collaboration with the development teams for both the LinuxONE and storage solutions

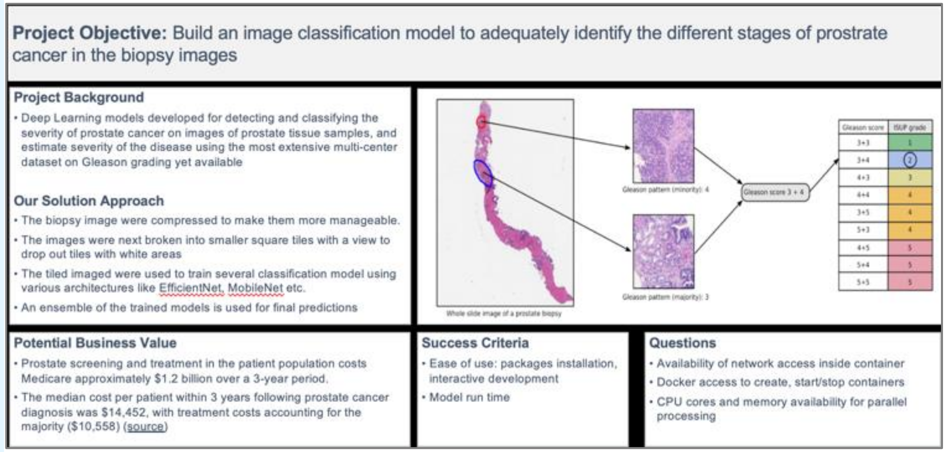
Client Benefits

- Client is now able to dynamically scale its AI application workloads on LinuxONE. Using the new Telum chip with client's prostate cancer detection app was 41% faster compared to z15 and 75% faster compared to its x86 servers with GPUs.
- LinuxONE was proven to be able to consolidate client's application workloads, while both reducing its datacenter footprint and consuming less energy than its x86 distributed servers.
- Its new sustainable hybrid cloud datacenter options – including PaaS and IaaS, give client's application teams broad flexibility and cost efficiency in service management models.
- Client has now developed a roadmap for further exploitation of LinuxONE for new applications that will enable sustainable growth.

Major US health organization leveraging LinuxONE 4 for AI-based cancer research applications



Prostate Cancer Detection Application – Test Result Details



Project Objective: Build an image classification model to adequately identify the different stages of prostrate cancer in the biopsy images

Project Background

- Deep Learning models developed for detecting and classifying the severity of prostate cancer on images of prostate tissue samples, and estimate severity of the disease using the most extensive multi-center dataset on Gleason grading yet available

Our Solution Approach

- The biopsy image were compressed to make them more manageable.
- The images were next broken into smaller square tiles with a view to drop out tiles with white areas
- The tiled imaged were used to train several classification model using various architectures like EfficientNet, MobileNet etc.
- An ensemble of the trained models is used for final predictions

Whole slide image of a prostate biopsy
Gleason pattern (minority): 4
Gleason pattern (majority): 3
Gleason score 3 + 4
Gleason score
ISUP grade
3+3
1
3+4
2
4+3
3
4+4
4
3+5
4
5+3
4
4+5
5
5+4
5
5+5
5

Potential Business Value

- Prostate screening and treatment in the patient population costs Medicare approximately $1.2 billion over a 3-year period.
- The median cost per patient within 3 years following prostate cancer diagnosis was $14,452, with treatment costs accounting for the majority ($10,558) (source)

Success Criteria

- Ease of use: packages installation, interactive development
- Model run time

Questions

- Availability of network access inside container
- Docker access to create, start/stop containers
- CPU cores and memory availability for parallel processing

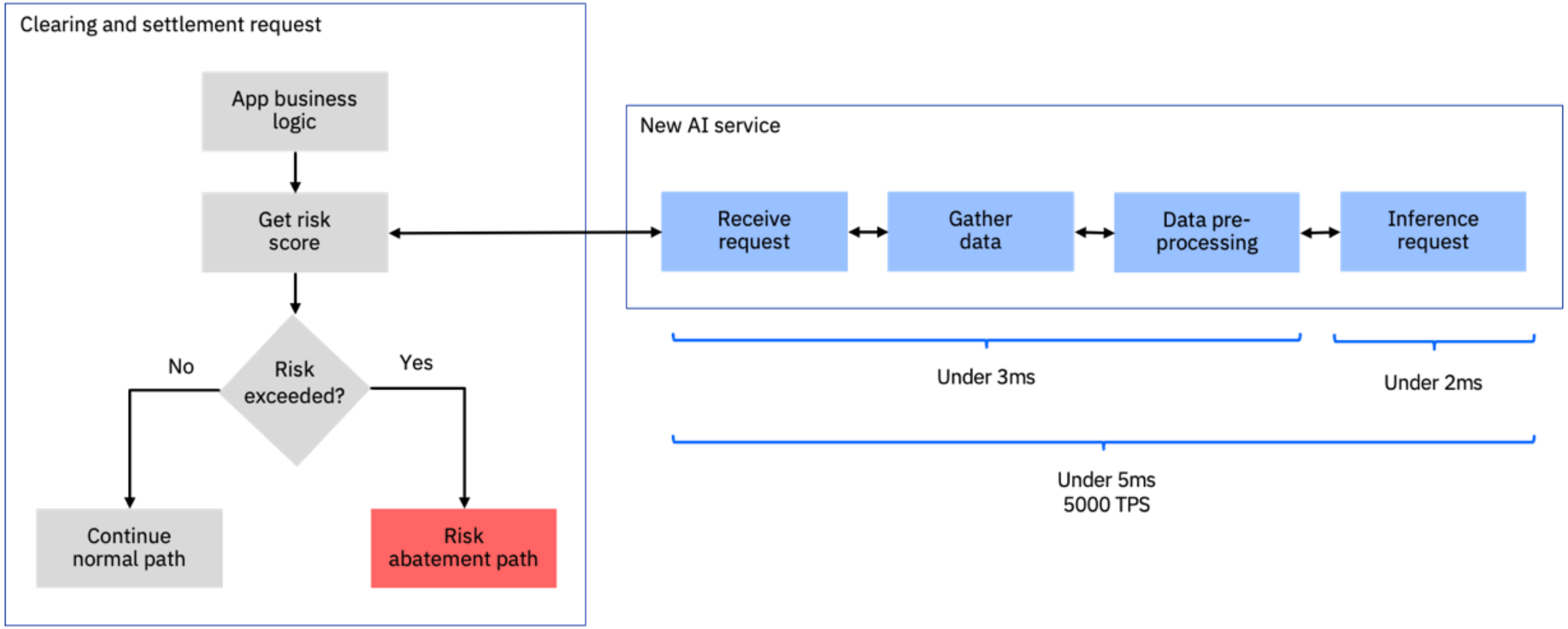
Specifications	z16 / LinuxONE4 (new)	z15/LinuxONE3 (old)	GPU Server
Server	**apsmd0158**	10.11.120.220	dbsld0127
OS	**Redhat**	Redhat	Redhat
Architecture	**s390x**	s390x	X86_64
Cores	**14**	62	64
Memory (RAM)	**126 GB**	500 GB	500 GB
GPU	**NA**	NA	16 GB

Metrics	z16 / LinuxONE4 (new)	z15/LinuxONE3 (old)	GPU Server
Total time	**4788 seconds**	6770 seconds	8416 seconds
Memory used (RAM)	**100 GB**	400 GB	400 GB
Cores used	**14**	62	64
Total Images	**512**	512	512
Subprocesses	**14**	48	48
Chunk size	**8**	16	16
Avg image size	**50 MB**	50 MB	50 MB

z16 & LinuxONE testing with Telum chip in January 2023 was **41% faster** compared to z15 and **75% faster** compared to GPU server

Major credit card processor utilizing AI for risk processing in clearing and settlement process

Comfortably achieved 5000 TPS under 5 millisecond latency with open-source AI framework of choice running on IBM Z



Clearing and settlement request

App business logic

Get risk score

Risk exceeded?

No

Yes

Continue normal path

Risk abatement path

New AI service

Receive request

Gather data

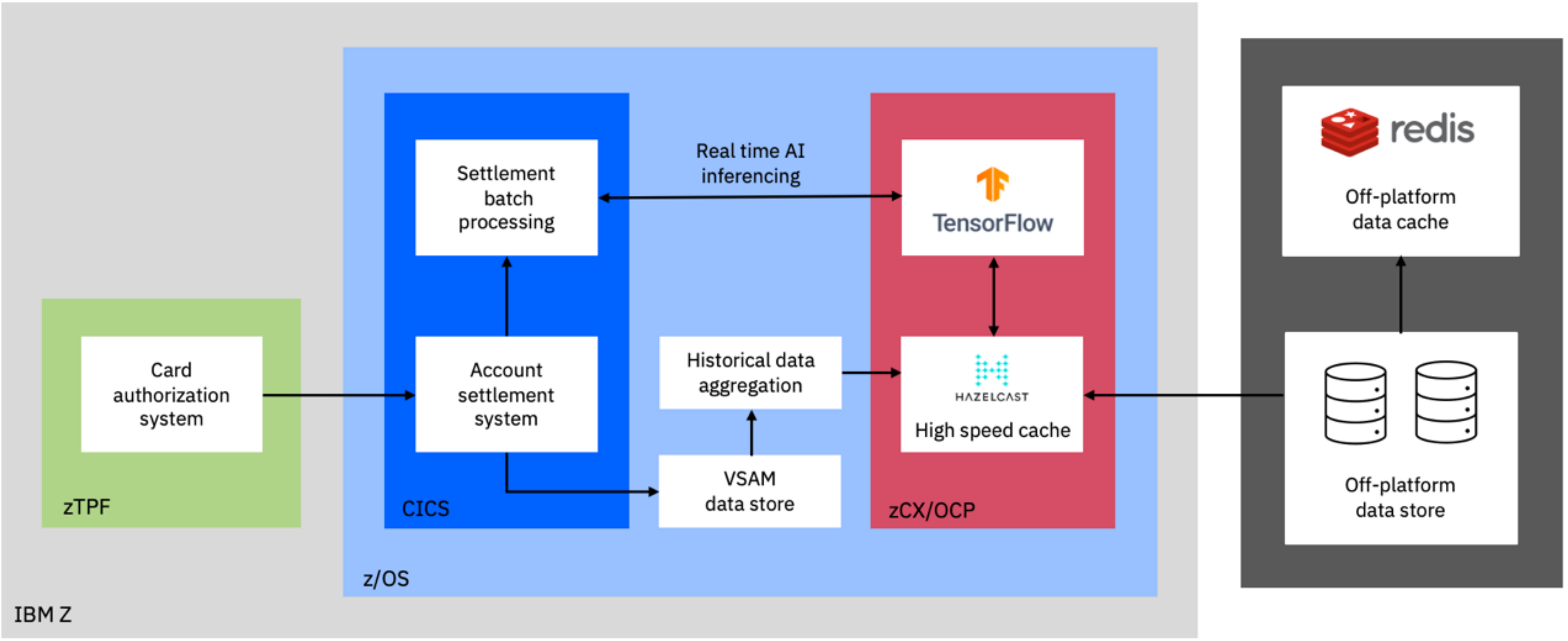
Data pre-processing

Inference request

Under 3ms

Under 2ms

Under 5ms
5000 TPS



IBM Z

z/OS

zTPF

Card authorization system

CICS

Account settlement system

Settlement batch processing

Real time AI inferencing

Historical data aggregation

VSAM data store

zCX/OCP

TensorFlow

HAZELCAST

High speed cache

redis

Off-platform data cache

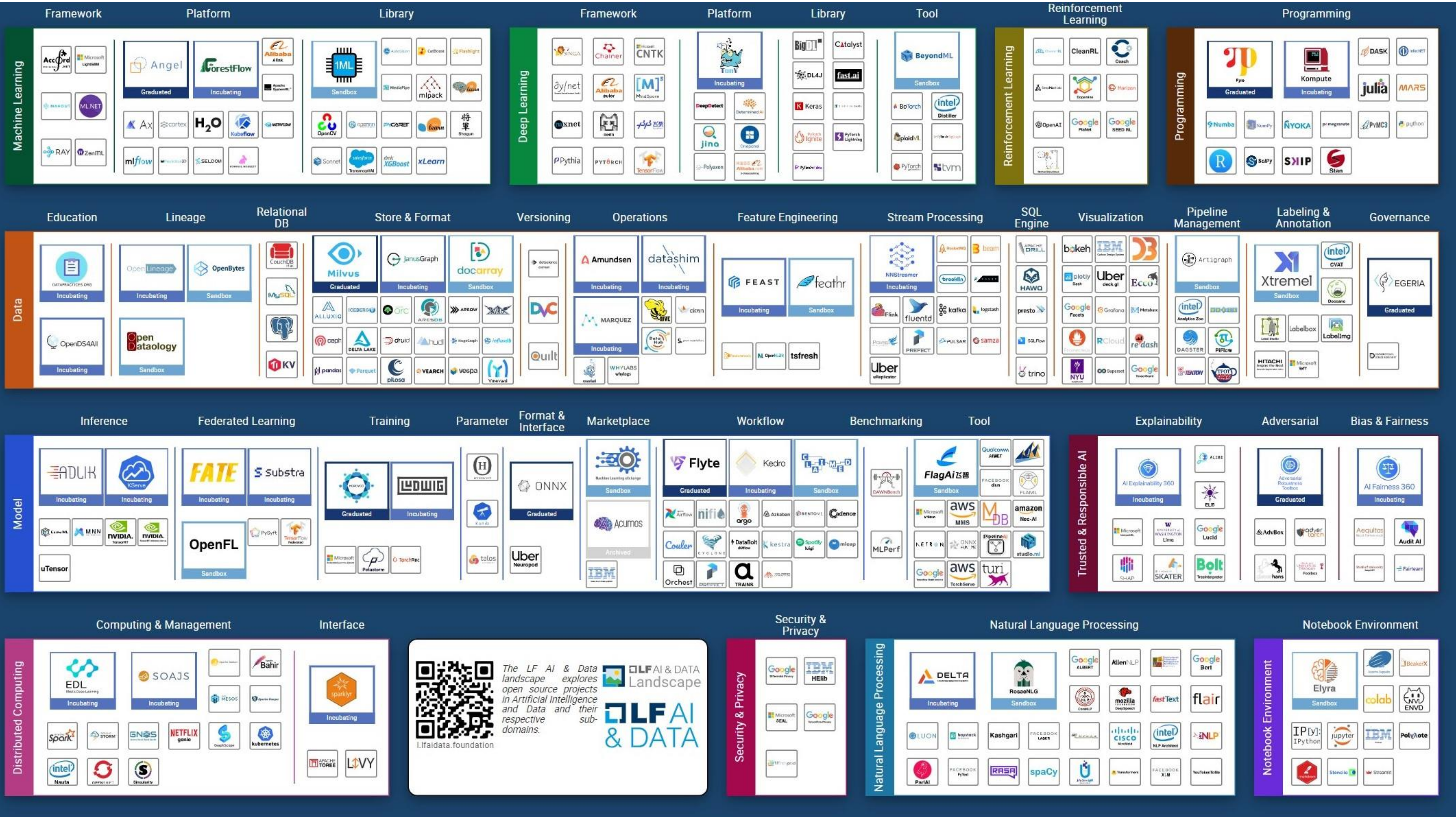
Off-platform data store

Agenda

1 Use scenarios

2 Technical foundations

3 Where to start ?



Machine Learning

Framework · Platform · Library

Deep Learning

Framework · Platform · Library · Tool

Reinforcement Learning

Reinforcement Learning

Programming

Programming

Data

Education · Lineage · Relational DB · Store & Format · Versioning · Operations · Feature Engineering · Stream Processing · SQL Engine · Visualization · Pipeline Management · Labeling & Annotation · Governance

Model

Inference · Federated Learning · Training · Parameter · Format & Interface · Marketplace · Workflow · Benchmarking · Tool

Trusted & Responsible AI

Explainability · Adversarial · Bias & Fairness

Distributed Computing

Computing & Management · Interface

The LF AI & Data landscape explores open source projects in Artificial Intelligence and Data and their respective sub-domains.

LF AI & DATA Landscape

LF AI & DATA

l.lfaidata.foundation

Security & Privacy

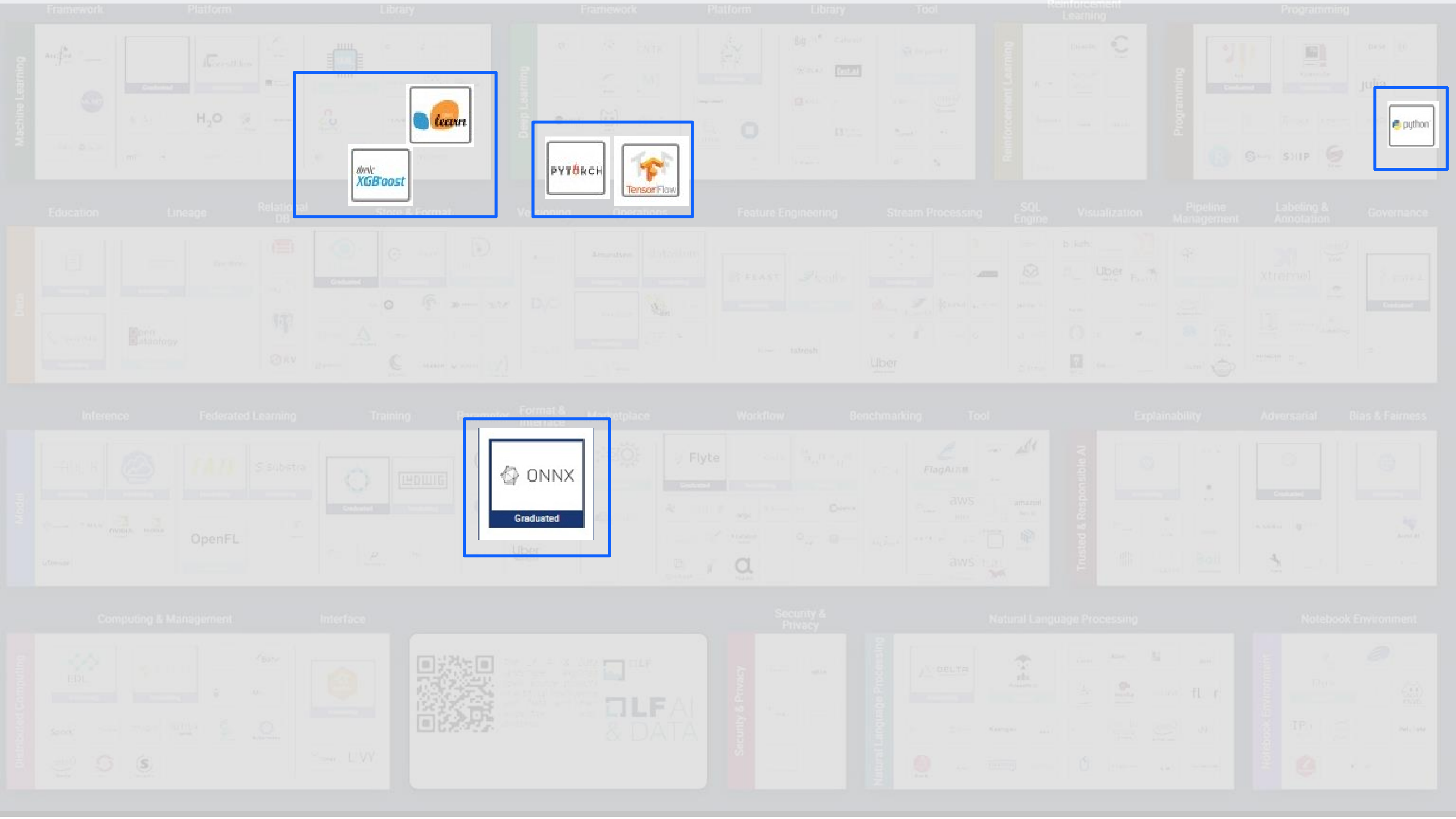
Security & Privacy

Natural Language Processing

Natural Language Processing

Notebook Environment

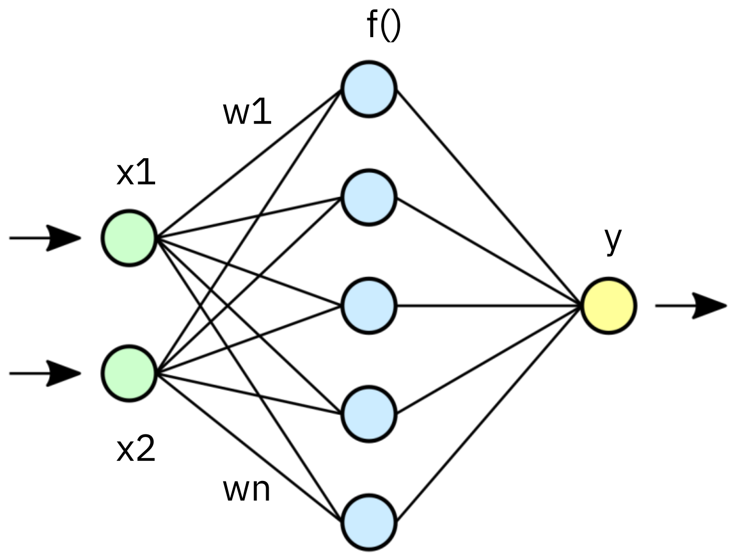
Notebook Environment


Machine Learning
Framework
Platform
Library
learn
XGBoost
Deep Learning
Framework
Platform
Library
Tool
PYTORCH
TensorFlow
Reinforcement Learning
Programming
python
Data
Education
Lineage
Relational DB
Store & Format
Versioning
Operations
Feature Engineering
Stream Processing
SQL Engine
Visualization
Pipeline Management
Labeling & Annotation
Governance
Model
Inference
Federated Learning
Training
Parameter
Format & Interface
Marketplace
Workflow
Benchmarking
Tool
ONNX
Graduated
Trusted & Responsible AI
Explainability
Adversarial
Bias & Fairness
Distributed Computing
Computing & Management
Interface
Security & Privacy
Natural Language Processing
Notebook Environment
LF AI & DATA

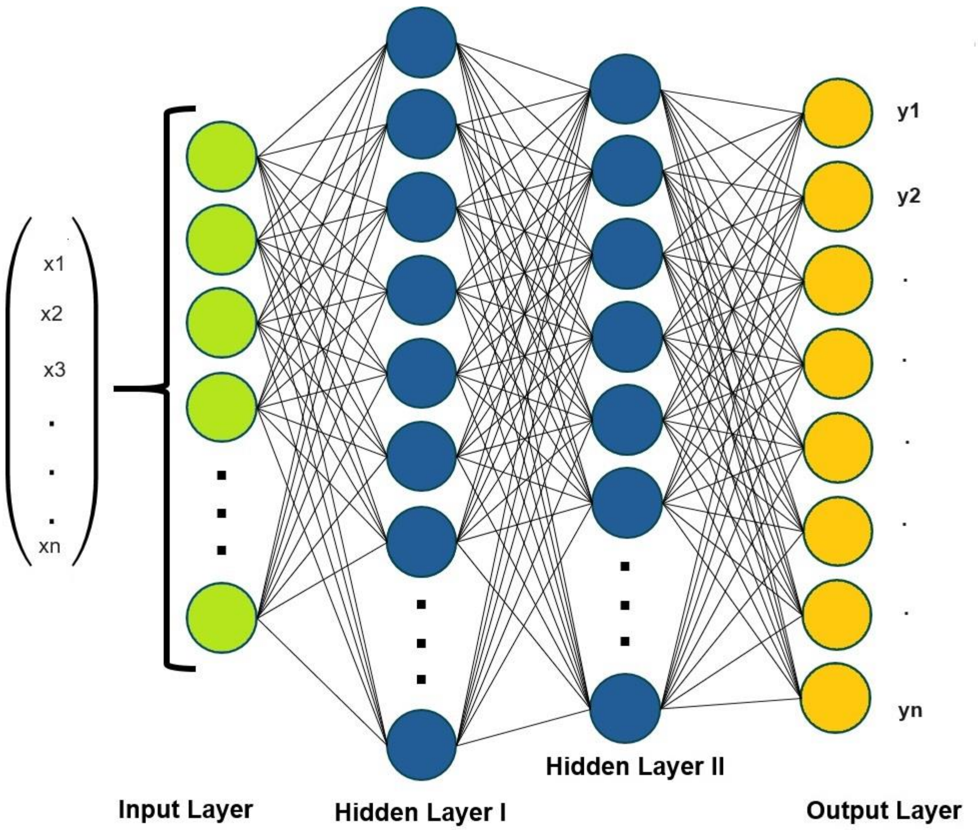
Neural Networks NN

Examples

Simple neural network



Deep neural network



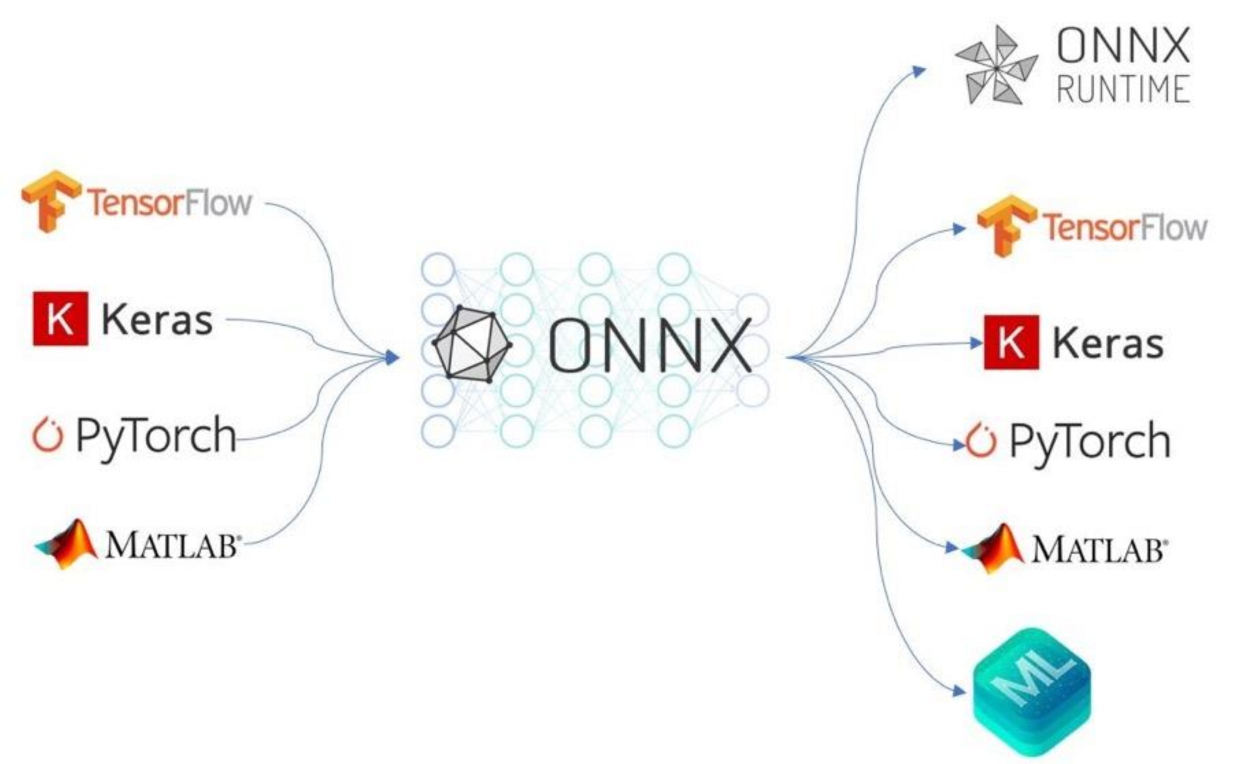
NN Model: inputs, weights, activation functions, output and topology ...

TensorFlow on IBM Z

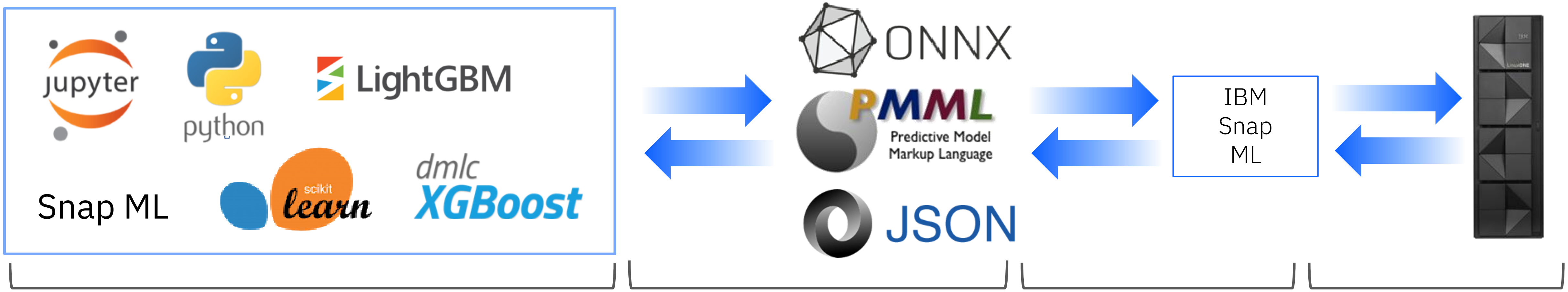
- TensorFlow on IBM Linux on Z and LinuxONE provides a leading open-source deep learning tools and framework for data scientists.
- Data Scientists can now leverage one of most popular ML/DL library. TensorFlow can be used for several use cases including NLP/text-based applications, image recognition, voice search and many more. It is used in popular apps like Facebook's DeepFace image recognition system, Apple's Siri for voice recognition.
- TensorFlow on Linux for Z enables infusion of DL & ML into mission critical workloads at scale leveraging Integrated Accelerator for AI
- TensorFlow can be accessed via:
  - Linux distribution on zCX (z/OS Container Extensions)
  - IBM Cloud Pak for Data v4.6 and later
  - IBM LinuxONE container image repository

ONNX

- **O**pen **N**eural **N**etwork e**X**change is a standard format for representing machine learning and deep learning models
- ONNX makes it easy to move and port AI models between different frameworks and hardware
- ONNX Models are generated by supported wide range of DL and ML frameworks or converted from other formats using conversion tools
- ONNX Models are designed to support multiple frameworks and runtimes
- They are executed/accelerated across hardware and execution environments
- Benefits to Z clients:
    - Build and train AI models using familiar frameworks
    - Make AI assets portable onto Z ecosystem
    - Optimize to seamlessly leverage software and hardware accelerators on IBM Z and LinuxONE 4



Machine Learning on IBM Z



Build and train ML models in any popular framework on any platform of your choice

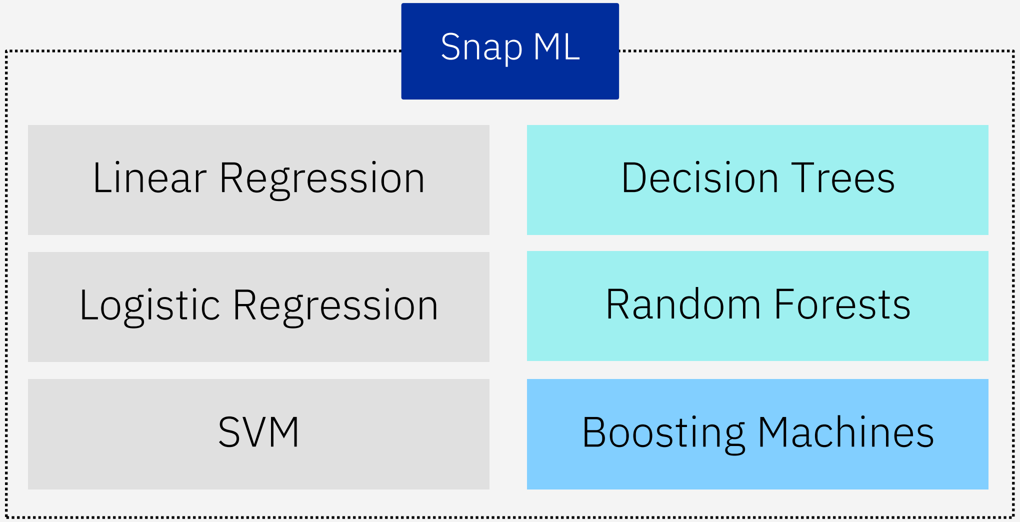
Convert ML models to ONNX, PMML or JSON interchange format and run on zIIPs engine

IBM Snap ML generates a program optimized for performance by executing on AIU

Deploy inference service on Z and infuse AI into applications

Snap ML

- Snap ML is an open-source library being developed and maintained by IBM Research that supports acceleration of training and inference of popular machine learning models
- Provides high-performance implementation for:
  - Generalized linear models
  - Tree-based models
  - Gradient boosting models
- Supports IBM Z & LinuxONE 4 Integrated Accelerator for AI
- Packaged along with IBM Cloud Pak for Data for Z, IBM Watson Machine Learning for Z and available as standalone container images via PyPi
- Documentation available here: https://snapml.readthedocs.io/en/latest/
- Examples of usage available here: https://github.com/IBM/snapml-examples



Snap ML

Linear Regression

Decision Trees

Logistic Regression

Random Forests

SVM

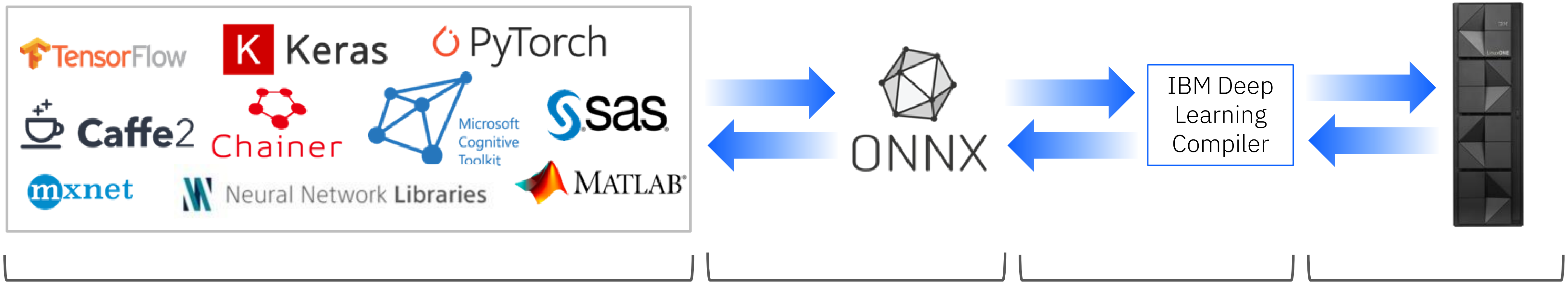
Boosting Machines

Fast	**Efficient**	**Accurate**	**At scale**
Train and re-train on new data online	Optimized to use fewer hardware resources	Make accurate decisions and predictions	Learn from all available data
Large parameters, model searches	Increased ROI and TCO	Easily create enterprise models for use cases like fraud detection, risk analysis, etc.	More data, better models, higher accuracy
Make fast decisions			Handle large volume of data

IBM Z Deep Learning Compiler (zDLC)

- **zDLC** enables clients to build and train deep learning models using their choice of open-source frameworks like TensorFlow, PyTorch, Keras, etc. and deploy and inference on LinuxOne 4 leveraging AIU
- Compile deep learning models optimised to leverage Integrated Accelerator for AI to deploy and run on IBM LinuxONE 4 and IBM Z16
- zDLC is powered by open-source project **ONNX**-**M**ulti-**L**evel-**I**ntermediate-**R**epresentation (ONNX-MLIR)
- **ONNX-MLIR** optimizes ONNX model graphs compilation into machine native code.
- The compiled shared library files can be embedded in applications independent of the original framework's dependencies and packages
- zDLC supports C++, Java, Python APIs
- Packaged along with IBM Watson Machine Language for Z, IBM Cloud Pak for Data for Z and standalone container images

Deep Learning on IBM Z



IBM Deep Learning Compiler

Build and train DL models in any popular framework on any platform of your choice

Convert DL models to ONNX interchange format and run on zIIPs engine

IBM DLC generates a program optimized for performance by executing on AIU

Deploy inference service on IBM Z and LinuxONE 4 and infuse AI into applications

IBM Telum Processor

Centralized on-chip accelerator shared by all cores

New Neural Network Processing Assist instruction

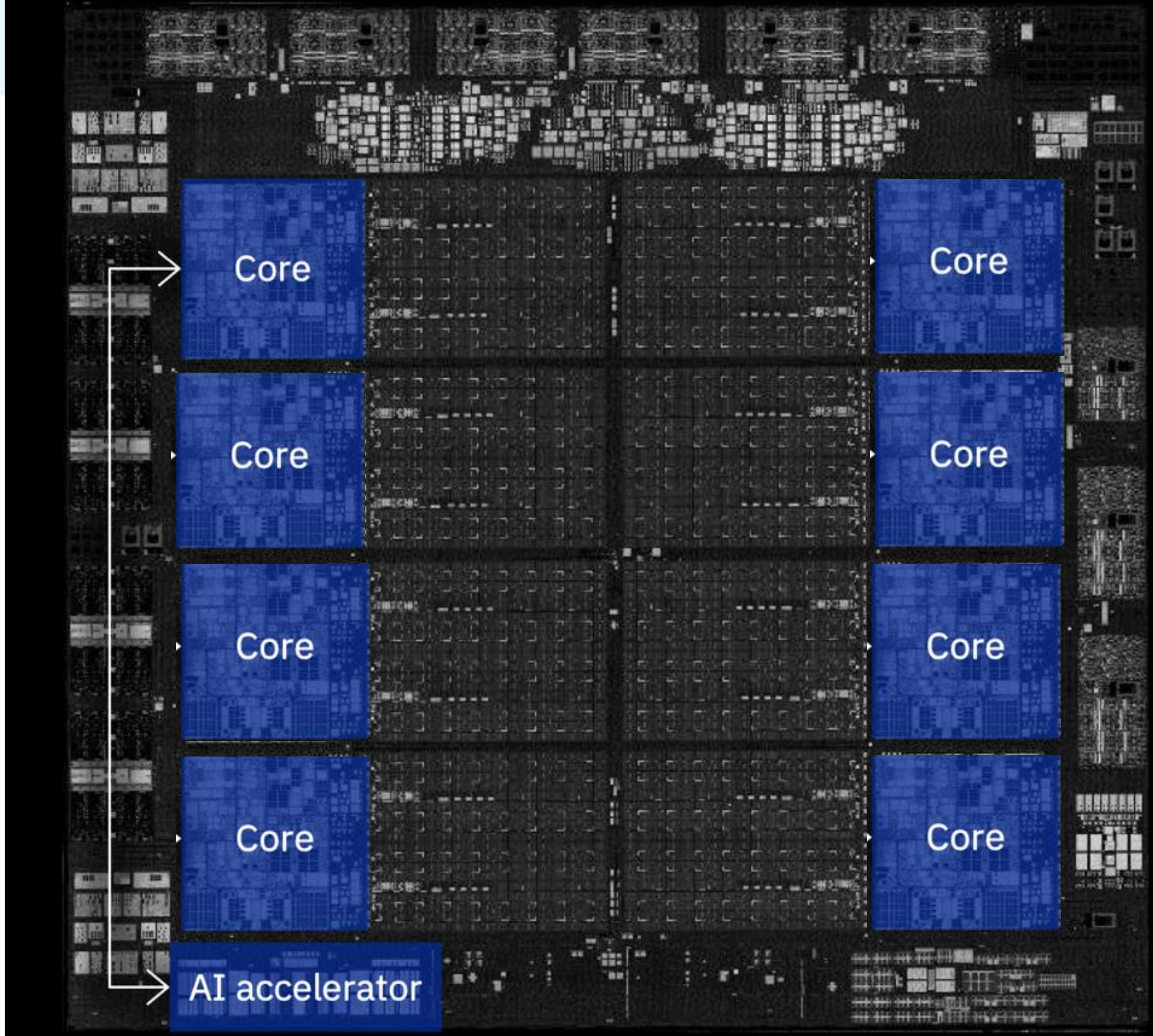
- Memory-to-memory CISC instruction
- Operates directly on tensor data in user space
- Matrix Multiplication, Convolution, Pooling, Activation Functions

Firmware running on core and AI Accelerator

- Address translation and access check for tensor data
- Prefetching of tensor data into L2 cache
- Coordination of data staging and compute

Enterprise class availability & security

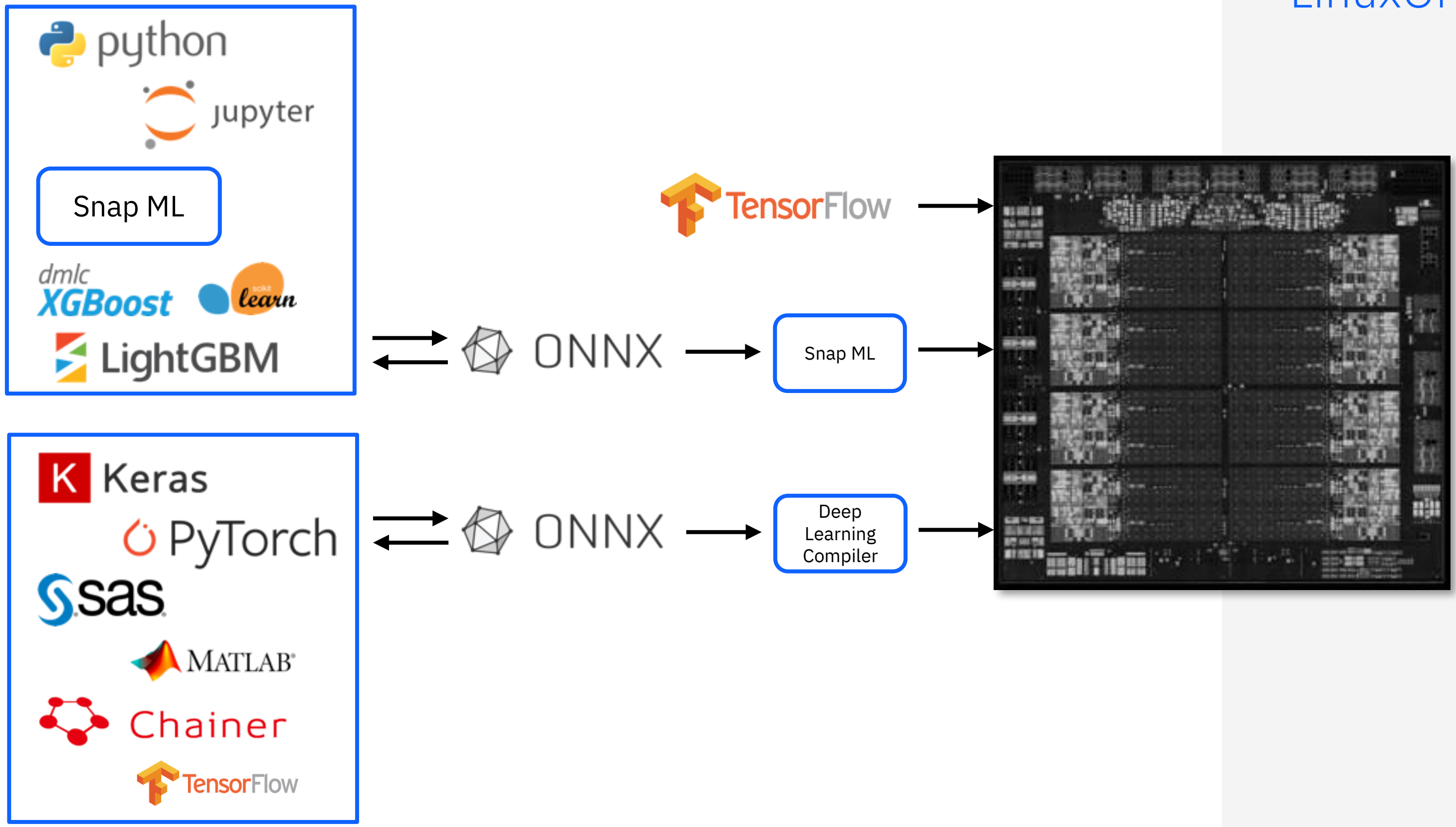
- Virtualization
- Direct memory access with all protection mechanisms
- Error checking and recovery



A holistic approach to AI

Seamlessly integrate AI into existing enterprise workload stack

Build & Train anywhere



Deploy AI at scale on LinuxONE 4

Applications

Banking
Finance
Insurance

Retail
Hospitality
Transport

Healthcare
Government
…

Offerings

IBM Cloud Pak for Data

IBM Watson

IT Operations

INSTANA an IBM Company

turbonomic

Cloud Pak for AIOps

Frameworks

PyTorch, TensorFlow, Keras, Apache Spark, scikit learn, ANACONDA, ONNX

Libraries, Compilers

python, Java, GO, GCC, C++, OpenBLAS, Deep Learning Compiler

OS, Virtualization

Linux, Ubuntu, SUSE, OPENSHIFT

Agenda

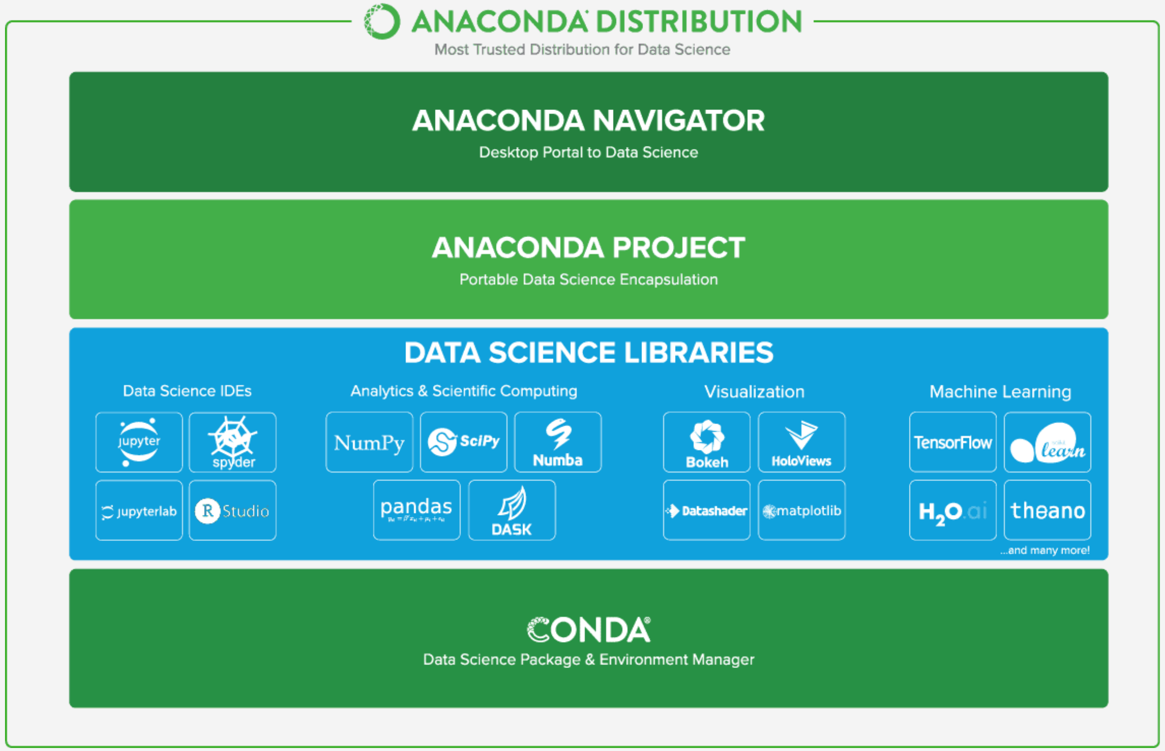
1 Use scenarios

2 Technical background

3 Where to start ?

Anaconda on IBM LinuxONE

- Anaconda on IBM Linux on Z and LinuxONE provides a leading open-source data science tools and framework for data scientists.
- Data scientists can leverage industry standard frameworks to discover, visualize, build, and train their models in tools and frameworks they're used to, such as scikit-learn, PyTorch, Matplotlib and more.
- Enabling infusion of AI & ML into mission critical workloads and processes leveraging Integrated Accelerator for AI
- Available Editions - Miniconda, Anaconda Individual Edition, Anaconda Commercial Edition
- Quick links:
  - Installation Instructions
  - White Paper
  - Guided Example



IBM Z
Container Image Registry

A registry for open source container images*

- Building blocks for creating workloads
- IBM Z versions of popular images
- Foundational distros, languages, databases, web serving, CI/CD infrastructure
  https://ibm.github.io/ibm-z-oss-hub/main/main.html

Hosted at the IBM Container Registry

- Images are built from scratch by IBM
- IBM controls the contents of the channel
- Images are scanned by IBM Vulnerability Manager
- Reports available to review
- Image digest hashes published to enable secure pull
- Images can be deployed in Linux on IBM Z and IBM z/OS zCX

Accommodate common IBM Z security requirements by working with a source you can trust to deliver container images built using best practices

* This program has important terms and conditions for use of the images of this registry. Please see the program agreement for a full details of these terms.

scikit learn
alpine Linux
Ubuntu
ONNX MLIR
NGINX
node JS
jupyter
Ruby
GitHub
mongoDB
Jenkins
GO
Postgres
APACHE ZooKeeper
Apache Tomcat
TensorFlow

Any many more!

IBM Cloud Pak for Data on Z

Simplify delivery of Data and AI services to the business

Collect and ingest data

DataStage, Apache Spark

Store and access data

Db2, Db2 warehouse, IDV, Big SQL, Data Gate

Prepare and govern data

DataStage, Watson Studio

Analytics and data science

Apache Spark, Watson Studio

Build and train models

Watson Studio, Watson Machine Learning

Deploy and service models

Watson Studio, Watson Machine Learning

Automate model lifecycle

Watson Studio, Watson Machine Learning, Auto ML

Monitor deployed models

Watson Studio, Open Scale

Opensource capabilities

Snap ML, Anaconda, TensorFlow, PyTorch Python libraries, etc.

AI offerings for IBM Z

AI for Business

IBM Offerings

IBM Cloud Pak for Data for Z

Open-source

- ONNX
- zDLC
- Apache Spark
- scikit learn
- ANACONDA
- TensorFlow
- NVIDIA TRITON INFERENCE SERVER

AI for IT

IT Operations

- INSTANA an IBM Company
- turbonomic
- IBM Cloud Pak for Watson AI Ops

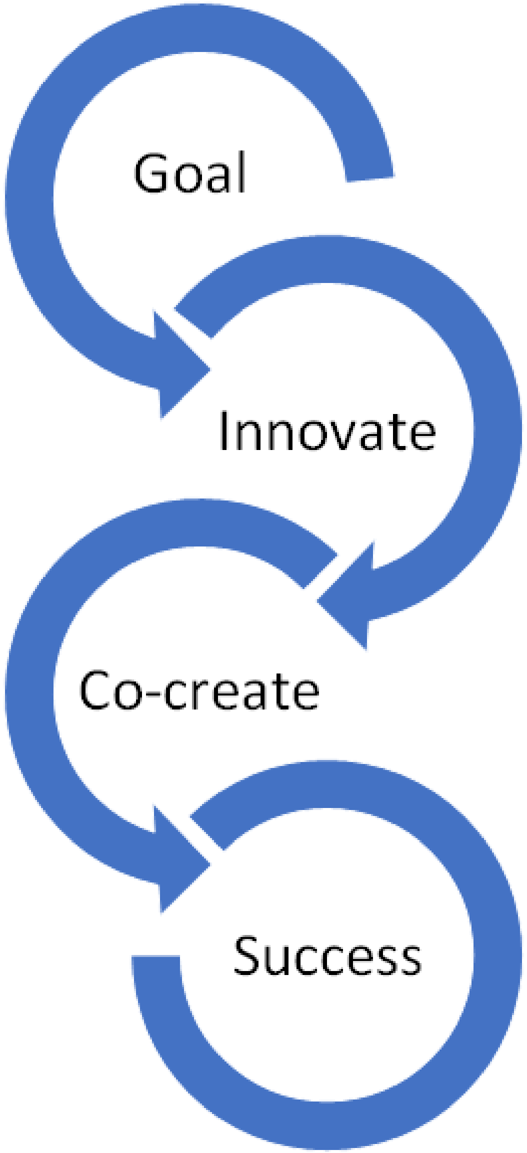
AI on IBM Z and LinuxONE Discovery Workshop

Objectives & Outcomes

- Understand the technology required for an AI on IBM Z solution
- Identify, define, prioritize and assess appropriate AI use cases. Choose a use case that has high value and is feasible.
- Scope an MVP, deliver a reference architecture and partner to deliver a POC for the selected use case.

Requested Personas/Roles

- Application Engineer
- SRE/Infrastructure Engineer
- Line of Business
- Chief Data Officer (CDO)
- Data Scientist



AI on IBM Z 4 Recap

Modernize applications by embracing AI

- Build cloud-native solutions with existing infrastructure and skills
- Work with best-in class IBM offerings that support every stage of AI model life cycle
- Leverage popular opensource ML/DL libraries and frameworks optimized using IBM

Embed real time AI insights at scale

- Develop high performance, low latency AI solutions that can infuse insights without impacting SLAs
- Score 100% workload with in-transaction inferencing
- Implement critical business solutions like fraud detection, AML, credit risk assessment, and more

Build resilient, sustainable AI solutions

- Industry-first quantum safe allows encryption and protection of all transactions and data on the platform
- Offer high availability up to 99.999999% (8x 9s); which is up to 0.3 seconds of downtime annually
- Reduce energy consumption, space and carbon footprint

Engage with us:

- aionz@us.ibm.com
- AI on IBM Z and LinuxONE Community
- https://ibm.github.io/ai-on-z-101/
- Contact us directly

Sites
Journey to AI on IBM Z Content Solution link
IBM Z and Cloud Mod Center AI Page link
Real-Time analytics and AI on the IBM mainframe link

Blogs
TensorFlow blog: link
ONNX blog: link

Demos
Watson Machine Learning Demo link
Anti-Money Laundering with AI on Z link
Fraud Detection Demo link

Redbooks
Optimized Inferencing and Integration with AI on IBM Z Introduction, Methodology, and Use Cases: link
Demystifying Data with AI on IBM Z –POV: link
Art of the Possible with AI on IBM Z link

Paper
IDC: The business value of the transformative mainframe link
Operationalizing Fraud Prevention on IBM z16: Reducing Losses in Banking, Cards, and Payments link

Open Source
IBM Z and LinuxONE container Image Registry: link
TensorFlow on IBM Z and LinuxONE container Image Registry: link
Anaconda Partnership link

IBM

Engage with us

Sites
Journey to AI on IBM Z Content Solution link
IBM Z and Cloud Mod Center AI Page link
Real-Time analytics and AI on the IBM mainframe link

Blogs
TensorFlow blog: link
ONNX blog: link

Demos
Watson Machine Learning Demo link
Anti-Money Laundering with AI on Z link
Fraud Detection Demo link

Redbooks
IBM Cloud Pak for Data on IBM Z link
Optimized Inferencing and Integration with AI on IBM Z Introduction, Methodology, and Use Cases: link
Demystifying Data with AI on IBM Z –POV: link
Art of the Possible with AI on IBM Z link

Paper
IDC: The business value of the transformative mainframe link
Operationalizing Fraud Prevention on IBM z16: Reducing Losses in Banking, Cards, and Payment link

Open Source
IBM Z and LinuxONE container Image Registry: link
TensorFlow on IBM Z and LinuxONE container Image Registry: link
Anaconda Partnership link

Contact Us
aionz@us.ibm.com
AI on IBM Z and LinuxONE Community
https://ibm.github.io/ai-on-z-101/

AI is disrupting every industry

Enterprises want to turn their data into a competitive advantage

Banking

Modernize loans with smart lending, reduce risk with predictive analysis and fraud detection

Insurance

Provide best-in-class claim experiences, meet compliance requirements, and identify fraud

Healthcare

Offer personalized treatment, detect disease faster and automate drug discovery and diagnosis

Telecom

Optimize infrastructure, offer self diagnosis & troubleshooting, dynamic ads and offers

Supply chain

Predict demand and supply trends, improve ops visibility, remove bottleneck/constraints

Technology

Increase productivity with automation, improve operations and service management

Government

Improve public services, insight-driven policy making, identify and prevent fraud

Automotive

Enhance driving experience, offer predictive maintenance services, optimize processes

Retail

Create personalized experiences, improve in-store operations, manage risks and increase profits

Major credit card processor utilizing AI for risk processing in clearing and settlement process

Comfortably achieved 5000 TPS under 5 millisecond latency with open-source AI framework of choice running on IBM Z

Client Challenge

- Client is a multinational financial services corporation headquartered in San Francisco. They provide digital payments globally, most commonly through branded credit cards, debit cards and prepaid cards.
- As part of their credit card clearing and settlement process, they want to determine which trades have a high-risk exposure before settlement.
- Existing app is a z/OS batch workload that uses rules-based approach today with a stringent of 5000 TPS and <5ms SLA requirements.
- Need to augment their existing rules-based approach with ML and DL model-based predictions without impacting the SLA.
- Client's data science org primarily develops, trains, deploy models onto home-grown x86 platform using open-source TensorFlow and TensorFlow Serving platform.
- Client was facing scaling issues with large latency delays when inferencing on x86 platform, far exceeding their SLA requirements.

IBM Solution

- IBM proposed a solution based on Linux on Z with Red Hat OpenShift running on z/OS Container Extension. Co-locating AI workloads onto z/OS helped client drastically reduce latency and achieve superior performance and scaling.
- AI for Z Solutions team and AI for Z Development team worked closely with the client team to deliver TensorFlow and TensorFlow Serving images supported on IBM Z. They helped client migrate existing TensorFlow models onto Z and further fine tune it to improve efficiency to meet client's stringent SLAs.
- Additionally, they delivered XGBoost models using Snap ML and NVidia Triton Inferencing Server (TIS) on Z.
- Snap ML models offer improved accuracy for certain operations as compared to TensorFlow DL models.
- NVidia TIS is a high-performance data pre-processing and ML/DL model inferencing server. AI for Z Solutions team compiled TIS on C++ and was able to deliver more than 27,000 TPS with under 1.4ms response time including data pre-processing.

Client Benefits

- Client is now able to infuse AI-based risk predictions into their clearing and settlement transactions in real time while comfortably maintaining their existing SLA requirements.
- Adopting RHOCP offers additional operational agility to the client team to deliver solutions onto their hybrid cloud platform using the same toolchain and delivery pipeline.
- Client's data science org can continue to innovate by leveraging familiar opensource platform of choice and deliver AI solutions onto IBM Z without the need to skill up on Z.

Next steps

- Clearing and Settlement application embedded with AI-based risk predictions going live in 1H 2024.
- Augment Snap ML models and TIS into their solution post going live.
- Start working on future projects including fraud detection and infusing business insights in real time on Z
- In addition to opensource explore IBM offerings like WMLz for future projects.

IBM Cloud Pak for Data on Z

Simplify delivery of Data and AI services to the business

Analyze data and infuse AI

- 50+ analytics services, AI apps and industry solutions.
- Manage end to end lifecycle of AI models – build, train, test, deploy, monitor, manage.
- Work with familiar opensource capabilities alongside IBM's market leading offerings

Organize data

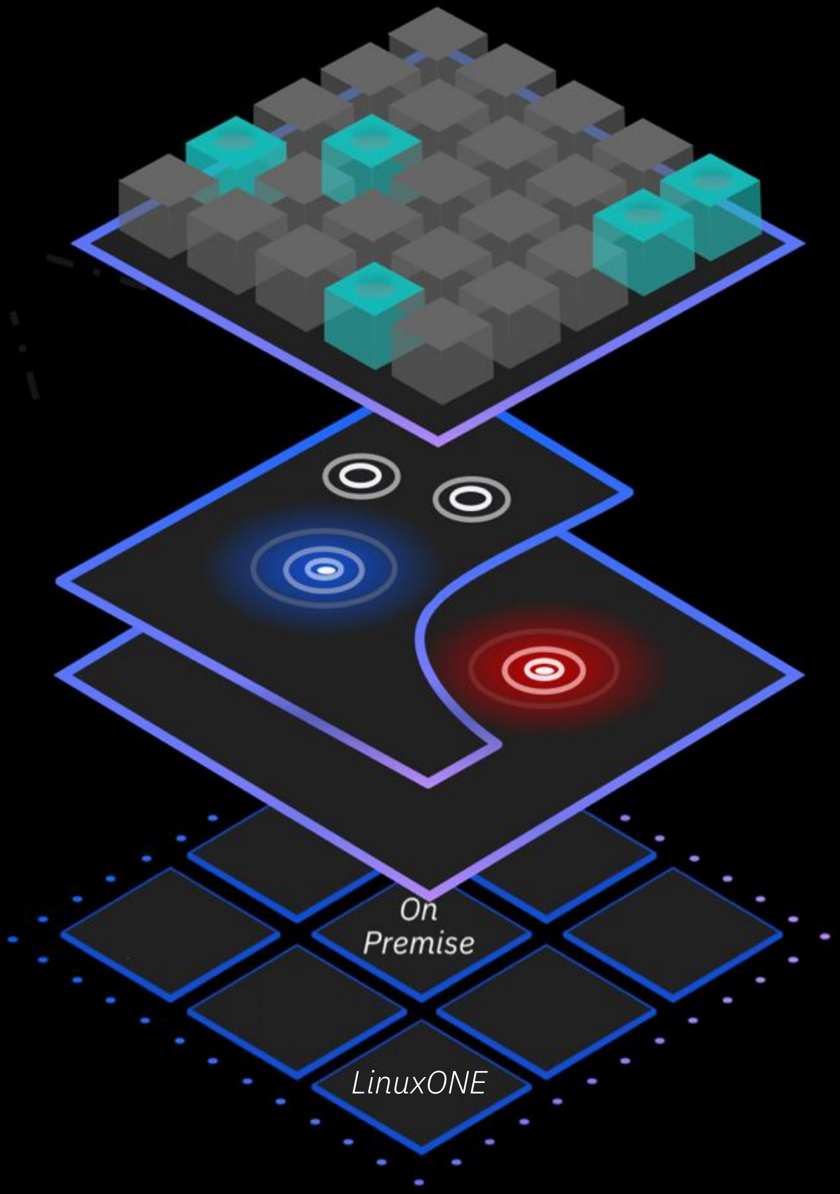
- Catalog and govern all enterprise data, models, rules, and insights through a common experience

Collect data

- Virtually connect, manage and query data assets no matter where they live
- Provision databases, data virtualization in seconds

Cloud-native deployment

- Run on OpenShift K8s deployment
- Provide private cloud, hybrid cloud platform leveraging your existing infrastructure and skills
- Build modern, interactive apps infused with intelligence



Robust AI ecosystem

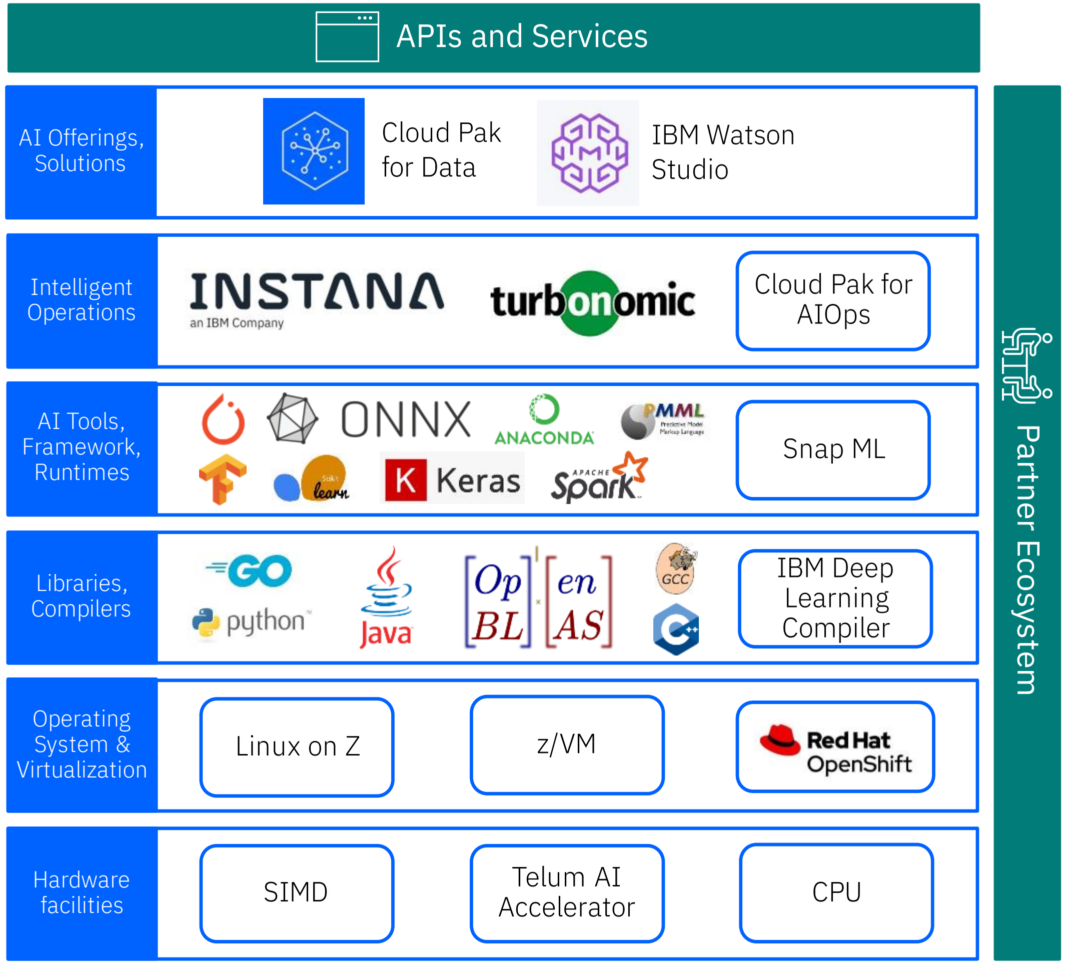
Modernize data infra and securely store and access enterprise data

Develop, train, deploy and manage complete AI lifecycle with ease

Utilize familiar opensource frameworks with optimized performance

Provide hybrid cloud platform while leveraging existing infra and skills

Build modern, interactive apps with embedded intelligence



APIs and Services

AI Offerings, Solutions: Cloud Pak for Data; IBM Watson Studio

Intelligent Operations: INSTANA an IBM Company; turbonomic; Cloud Pak for AIOps

AI Tools, Framework, Runtimes: ONNX; ANACONDA; PMML; scikit learn; Keras; Apache Spark; Snap ML

Libraries, Compilers: GO; python; Java; OpenBLAS; GCC; IBM Deep Learning Compiler

Operating System & Virtualization: Linux on Z; z/VM; Red Hat OpenShift

Hardware facilities: SIMD; Telum AI Accelerator; CPU

Partner Ecosystem

Enterprise-fit AI infrastructure

- Industry-first on-chip AI accelerator for Enterprise workloads
- Develop, train, deploy and manage complete AI lifecycle with ease
- IBM and opensource AI tools optimized to leverage to leverage AI accelerator
- Robust AI ecosystem optimized to leverage AI accelerator
- Embed Real-time insights for high volume workload without impacting SLAs

IBM Telum Processor

More than just a matrix multiplier!!!

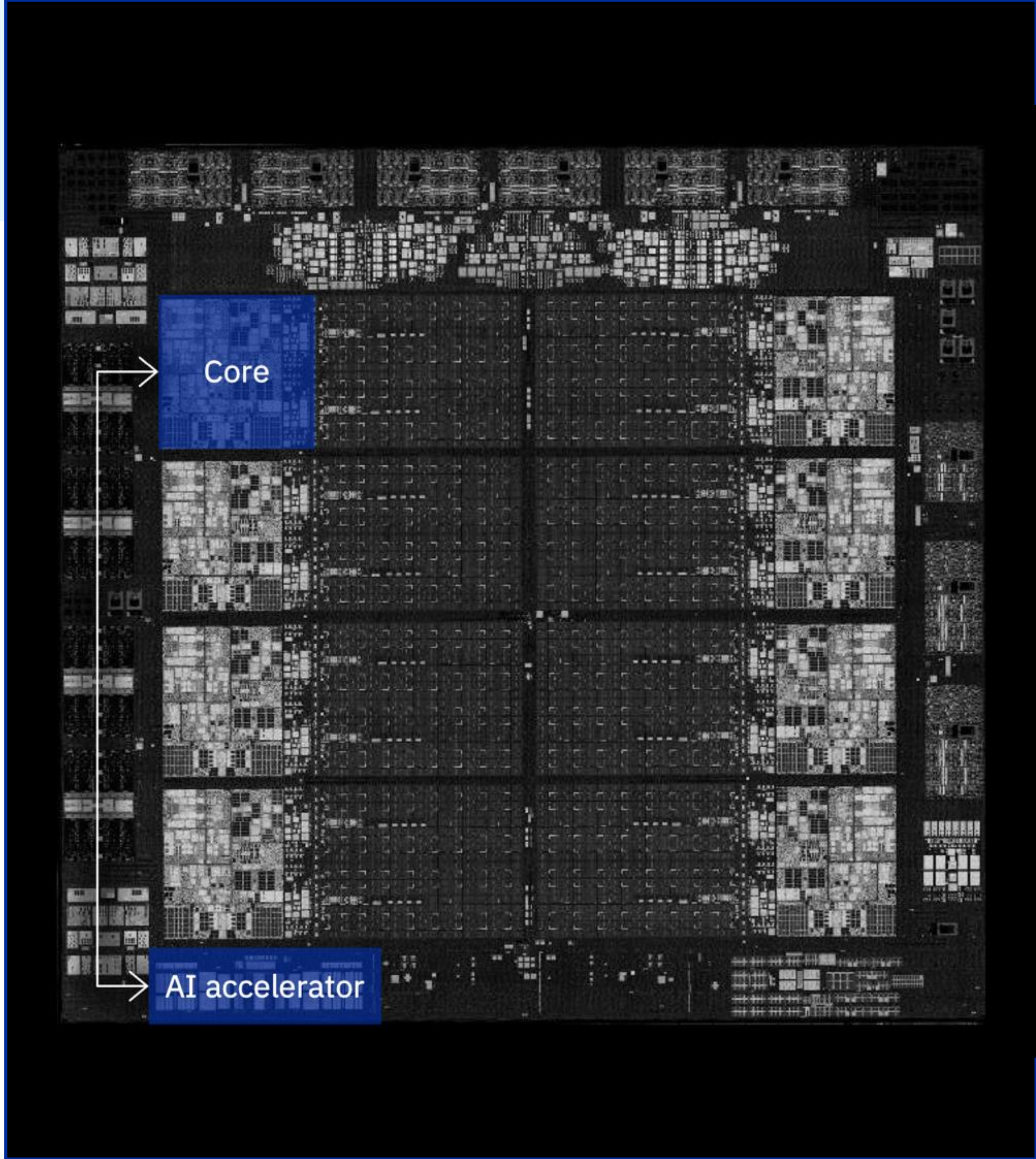
Operations supported on the accelerator

- LSTM Activation
- GRU Activation
- Fused Matrix Multiply, Bias op
- Fused Matrix Multiply (w/ broadcast)
- Batch Normalization
- Fused Convolution, Bias Add, Relu
- Max Pool 2D
- Average Pool 2D
- Softmax
- Relu
- Tanh
- Sigmoid
- Add
- Subtract
- Multiply
- Divide
- Min
- Max
- Log

AI functions and macros extracted via NPPA instructions

- Elementwise, Activation
- Normalisation, Pooling
- Matrix multiplication
- Convolution
- Conv + Scale + Activation
- MatMul + Scale / Activation
- RNN Activation

Accelerator is extensible such that new functions can be added with every firmware updates



IBM Telum Processor

Combining compute and data movers

Aggregate of >6 TFLOPS / chip

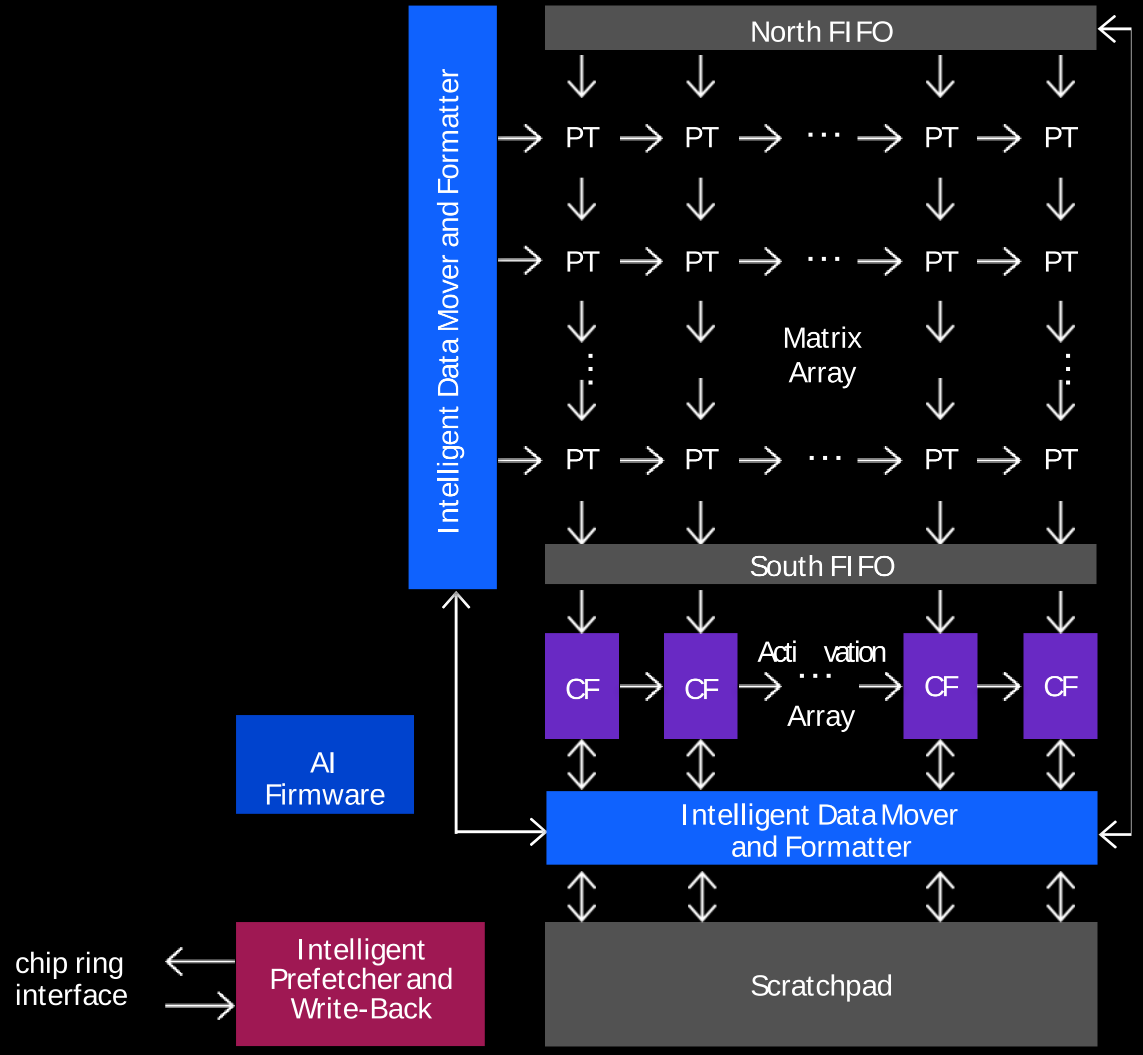
- Over 200 TFLOPS on 32-chip system

Compute Arrays

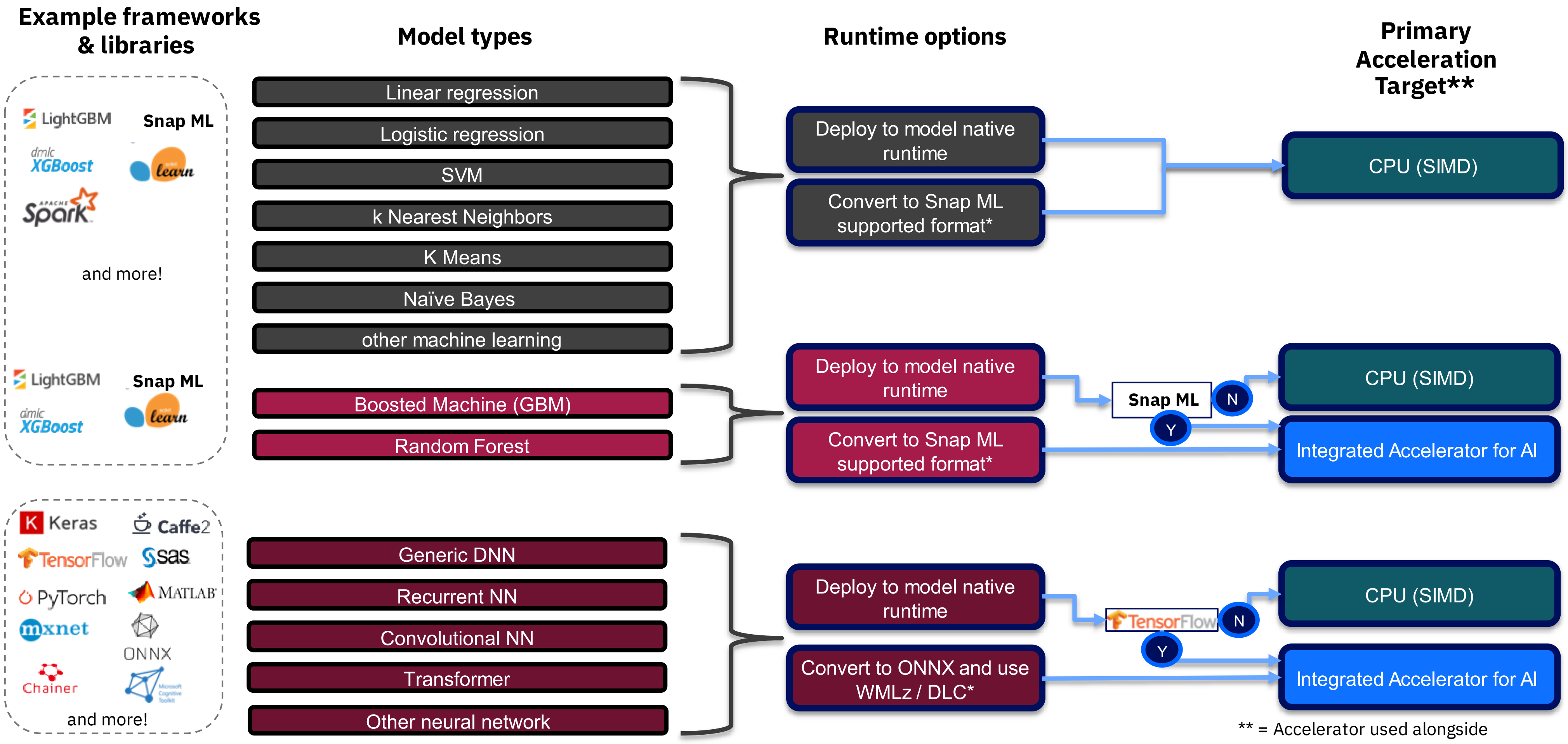
- 128 processor tiles with 8-way FP-16 FMA SIMD
  - Optimized for matrix multiplication and convolution
- 32 processor tiles with 8-way FP-16/FP-32 SIMD
  - Optimized for activation functions & complex operations

Intelligent Prefetcher and Data Movers

- 200+ GB/s read/store bandwidth from/to cache
- 600+ GB/s bandwidth between engines
- Multi-zone scratchpad for concurrent load, execution and store



Choosing optimal deployment path



Example frameworks & libraries

Model types

Runtime options

Primary Acceleration Target**

Linear regression
Logistic regression
SVM
k Nearest Neighbors
K Means
Naïve Bayes
other machine learning

Deploy to model native runtime
Convert to Snap ML supported format*
CPU (SIMD)

Boosted Machine (GBM)
Random Forest

Deploy to model native runtime
Convert to Snap ML supported format*
Snap ML
N
Y
CPU (SIMD)
Integrated Accelerator for AI

Generic DNN
Recurrent NN
Convolutional NN
Transformer
Other neural network

Deploy to model native runtime
Convert to ONNX and use WMLz / DLC*
TensorFlow
N
Y
CPU (SIMD)
Integrated Accelerator for AI

* = where applicable

** = Accelerator used alongside SIMD/CPU operations.

AI on IBM Z Discovery Workshop

IBM **Client Engineering for Systems**

IBM Z and LinuxONE
Quick Reference Guide

AI on IBM Z Discovery Workshop

Overview

The AI landscape has gone through a tremendous transformation over the past number of years. Clients are starting to recognize their need to proactively perform analytical scoring and inferencing on many of their transactional workloads in real time. IBM has made great strides in improving ease of deploying an analytical model directly on z/OS. By keeping the analytics on IBM Z, we can perform our scoring/inferencing with "live" data already on IBM z Systems because the latency of making an analytical query on platform is much lower than going off platform. The new zAIU on-chip accelerator on the IBM z16 will decrease the latency even more as well as decrease AI scoring consumption of resources on GPs. This on-platform approach also ensures that AI workloads benefit from the traditional IBM Z qualities of services.

This workshop will demonstrate the business value of performing analytical scoring and inferencing directly on IBM Z and will bring the user through a journey of taking a model trained in many industry standard analytical frameworks off platform and easily deploying that model onto IBM Z. It will also show how then we can easily take existing applications and quickly modify them to add inferencing and/or scoring using this deployed model.

IBM **Client Engineering**

Target Audience

- IBM Z clients who are interested in leveraging AI and analytic capabilities to gain new insights from their IBM Z based workloads.
- Clients interested in understanding how IBM Z can enable them to score every transaction, so no opportunity is missed due to latency.
- Those interested in seeing how IBM Z enables scoring (or inference) using industry standard open-source frameworks directly on platform which allows for minimal impact to transactional workload SLAs while still providing for all of the qualities of service of IBM Z.

Why Use This Service?

- Are your analytical insights sometimes too late to do much good?
- Are you interested in running many of the popular analytic frameworks directly on IBM Z easily?
- Would you like some hands-on experience deploying models on IBM Z and having your applications leverage them?
- Would you like to learn how AI on IBM Z can save you on operational costs and improve security of your data?

Benefits

This workshop will ensure that you have the architecture, practices, and hands-on experience to successfully begin your AI on IBM Z journey.

Service Provided

- Discover some of the AI on IBM Z technologies and their value
  - Understanding key technologies and their place in your AI journey, including the new on-chip accelerator on the IBM z16
  - Leveraging TensorFlow and TensorFlow serving to deploy models for online inference using zCX
  - Utilizing the Watson Machine Learning for z/OS Online Scoring Community Edition to service Open Neural Network Exchange (ONNX) models
- Streamlining model deployment and development and management with WMLz
- Advise on planning steps required to deploy your model onto IBM Z
- Illustrate the AI on IBM Z technology through real world use case demonstrations and showcases
- Apply workshop discoveries to real life through interactive hands-on experience

Optional

- Define client-specific AI use cases
- Confirm use case functional requirements and outcomes
- Design client solution functional architecture and configuration
- Develop practical implementation plan

Deliverables

- Advice on best practices using AI on IBM Z technologies
- Skills enablement for AI on IBM Z
- Identify possible use cases for client environment
- Recommendations for initial AI on IBM Z adoption

Contacts

Contact us at ce4s@ibm.com or your local IBM Client Engineering for Systems team

Also consider the IBM Client Engineering for Systems **Data and AI Client Workshop** offering to gain an understanding of more of the specifics of AI technologies using DVM, Db2 Analytics Accelerator, Db2 AI for, Db2 for z/OS Data Gate as well as the IBM Cloud Pak for Data..

Notices and disclaimers

-
- **U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.**
- Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.
- IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."
- **Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.**

- Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those.
- Customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.
- References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.
- Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.
- It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers

- Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**
- The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml

IBM