Machine Learning "Soup to Nuts"

Jesse Gorzinski, Senior Business Architect, IBM
William Xiang, IBM: Software Developer, IBM
Adam Shedivy, IBM: AI and Open-Source, IBM

AI and IBM i use cases: three main categories

Data Analytics	Trend analysis Anomaly detection Natural language interrogation
Operations	Active monitoring / alerting Natural language administration Self-healing
Developer eXperience	CoPilot (help the developer write code) Code chatbots Conversion tools

Artificial Intelligence (AI)

Human intelligence exhibited by machines

AI can be defined as a technique that enables machines to mimic cognitive functions associated with human minds – cognitive functions include all aspects of learning, reasoning, perceiving, and problem solving.

Machine Learning (ML)

Systems that learn from historical data

ML-based systems are trained on historical data to uncover patterns. Users provide inputs to the ML system, which then applies these inputs to the discovered patterns and generates corresponding outputs.

Deep Learning (DL)

ML technique that mimics human brain function

DL is a subset of ML, using multiple layers of neural networks, which are interconnected nodes, which work together to process information. DL is well suited to complex applications, like image and speech recognition.

Foundation Model

Generative AI systems

AI model built using a specific kind of neural network architecture, called a transformer, which is designed to generate sequences of related data elements (for example, like a sentence).

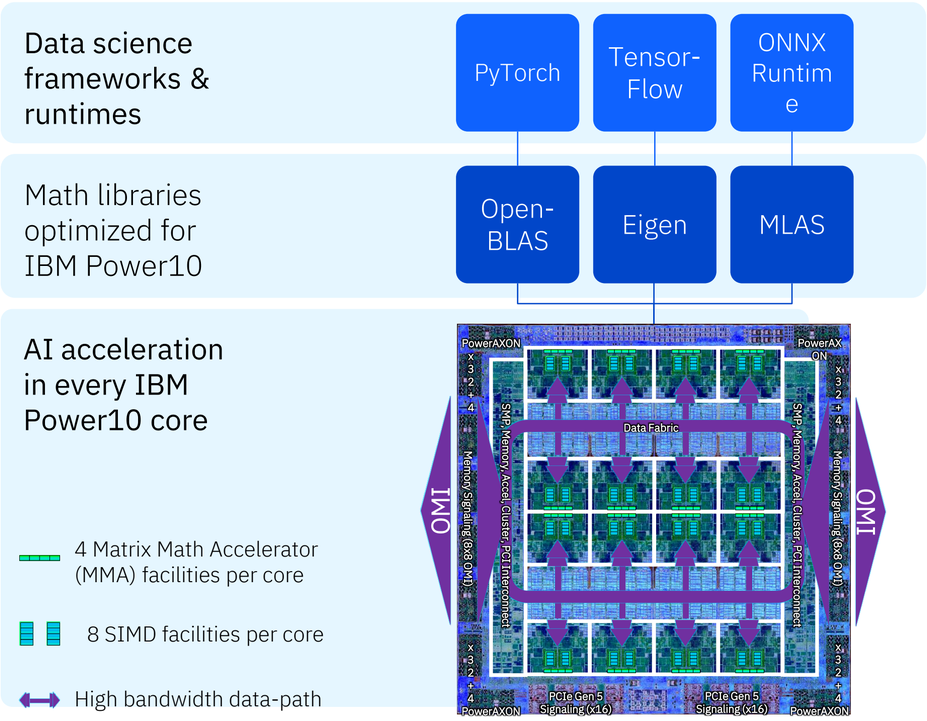
1950's

1980's

2010's

2020's

Accelerate AI Efficiently with AI Optimized Hardware



Each core has four MMA (Matrix Math Accelerator) facilities to accelerate matrix calculations that are used in many common AI workloads

Power 10 MMA Overview

Feature	AI Method	GPU	P10 with MMA
Training	Deep Learning	Best Fit (cost-perf)	Limited Benefit
	Machine Learning	Limited Benefit (cost-perf)	Best Fit (cost-perf)
	Foundation Model (like GenAI)	Best Fit (cost-perf)	Not Optimal
Re-training / Fine-tuning	Deep Learning	Best Fit (cost-perf)	Limited Benefit (cost-perf)
	Machine Learning	Not Applicable	Not Applicable
	Foundation Models (like GenAI)	Best Fit (cost-perf)	Limited Benefit (cost-perf)
Prompt Tuning (including RAG pattern)	Deep Learning	Not Applicable	Not Applicable
	Machine Learning	Not Applicable	Not Applicable
	Foundation Model (like GenAI)	Limited Benefit (cost-perf)	Best Fit (cost-perf)
Inference	Deep Learning	Limited Benefit (cost-perf)	Best Fit (cost-perf)
	Machine Learning	Limited Benefit (cost-perf)	Best Fit (cost-perf)
	Foundation Model (like GenAI)	Best Fit >3B depending on several factors (cost-perf)	Best Fit <3B depending on several factors (cost-perf)
SW Maintenance		Need to update GPU specific SW (CUDA, cuDNN, etc.)	Maintained by IBM / Partner

GPUs or Power 10 w/MMA*

*Please see speaker notes for details

Agenda

- Overview of AI stacks for IBM i
- High-level integration techniques
- Walk through IBM i end-to-end ML workflow with Rocket AI Hub on Power

AI Stacks

4 products, different values all on IBM Power 10

Watsonx

Enable GenAI workflows on IBM Power10 and benefit from:

- Secure end-to-end GenAI on-premises; no internet needed.
- 100% on-chip acceleration; no external accelerators needed.
- Domain adaptation via retrieval augmented generation (RAG).

IBM Cloud Pak for Data

Establish a holistic collaborative data & AI environment for heterogeneous stakeholders to work together (SMEs, data engineers, business analysts, ...) & converge AI with data by deploying AI models with mission-critical processes, data, and transactions while governing the end-to-end process.

Rocket AI Hub for IBM Power

Establish a development-centric data & AI environment (data scientists, developers, ops, scientists, ...) and get access to the most up-to-date frameworks & tools at no license costs on top of Red Hat OpenShift or vanilla Kubernetes. Start small and scale big over time.

RocketCE for IBM Power

Minimize the entry barrier for AI by natively using Python packages on Linux LPARS, without any container platform needed; allowing for minimal extensions of IBM Db2, SAP, and ORACLE landscapes. Benefit from over 200 packages optimized for IBM Power10.

Watsonx*

- Not available to run on Power without an accelerator yet, but clients can use Foundation Models and do inference on Power.

Note: You may opt to run Community Supported Open Source based AI solutions on Power

watsonx

Leverage foundation models to automate data search, discovery, and linking in watsonx.data

watsonx.ai

Train, validate, tune and deploy AI models

watsonx.data

Scale AI workloads, for all your data, anywhere

watsonx.governance

Enable responsible, transparent and explainable AI workloads

Leverage governed enterprise data in watsonx.data to seamlessly train or fine-tune foundation models

Direct, manage and monitor activities across the AI lifecycle, meeting risk and regulatory requirements with watsonx.governance

watsonx.ai: Data Science and MLOps

Build machine learning models automatically in the studio

Model training and development

Build experiments quickly and enhance training by optimizing pipelines and identifying the right combination of data

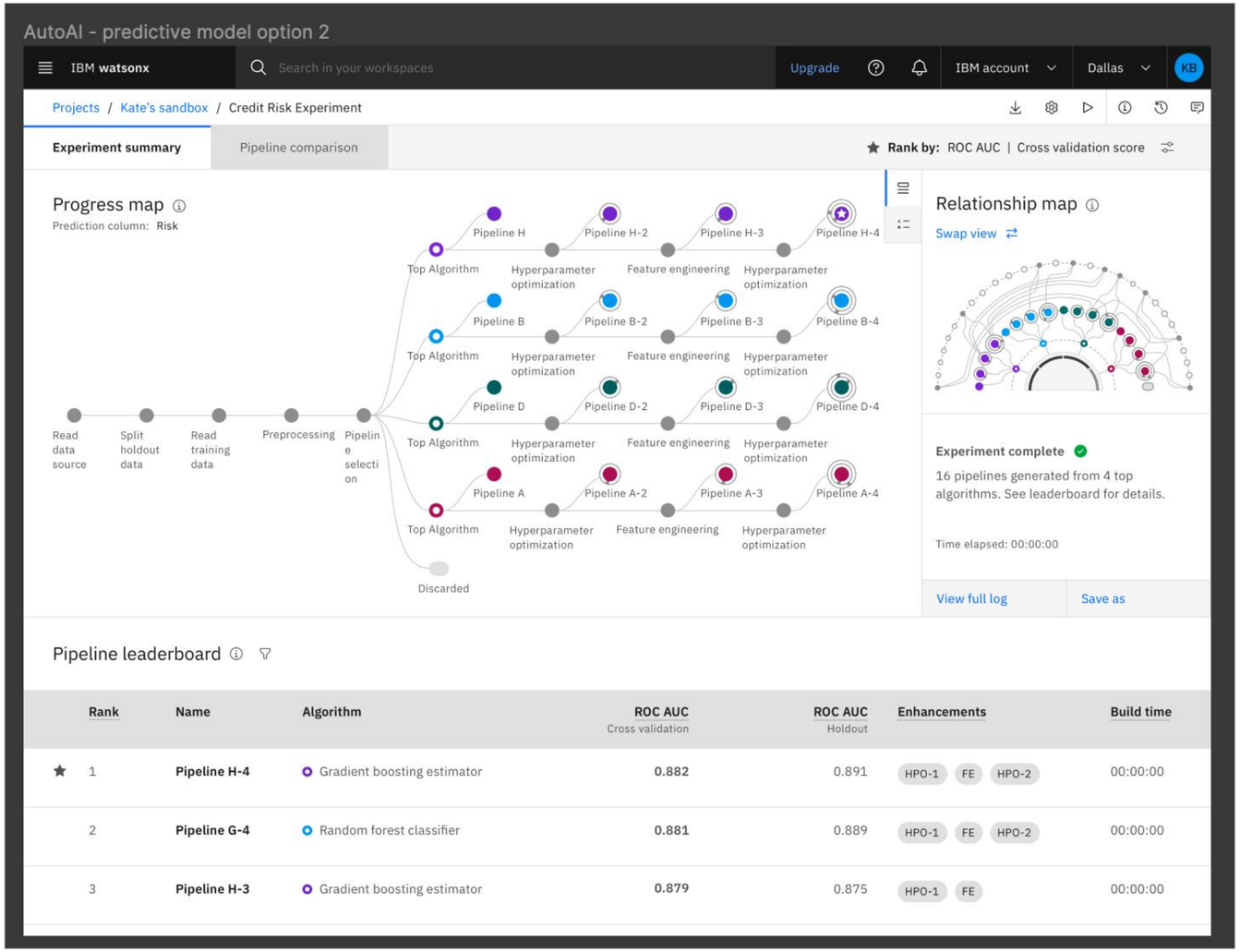
AutoAI, including preparing data for machine learning and generating and ranking candidate model pipelines

Use predictions to optimize decisions, create and edit models in Python, in OPL or with natural language

Integrated visual modeling

Prepare data quickly and develop models visually to help visualize and analyze enterprise data to identify patterns and trends, explore opportunities, and make informed, insightful business decisions

- Uncover correlations
- Insight for hypotheses
- Find relationships and connections within the data



watsonx.ai: Synthetic Data Generator

Generate synthetic tabular data to address your data gaps

Create synthetic data at scale

Unlock your valuable insights by using synthetic data.

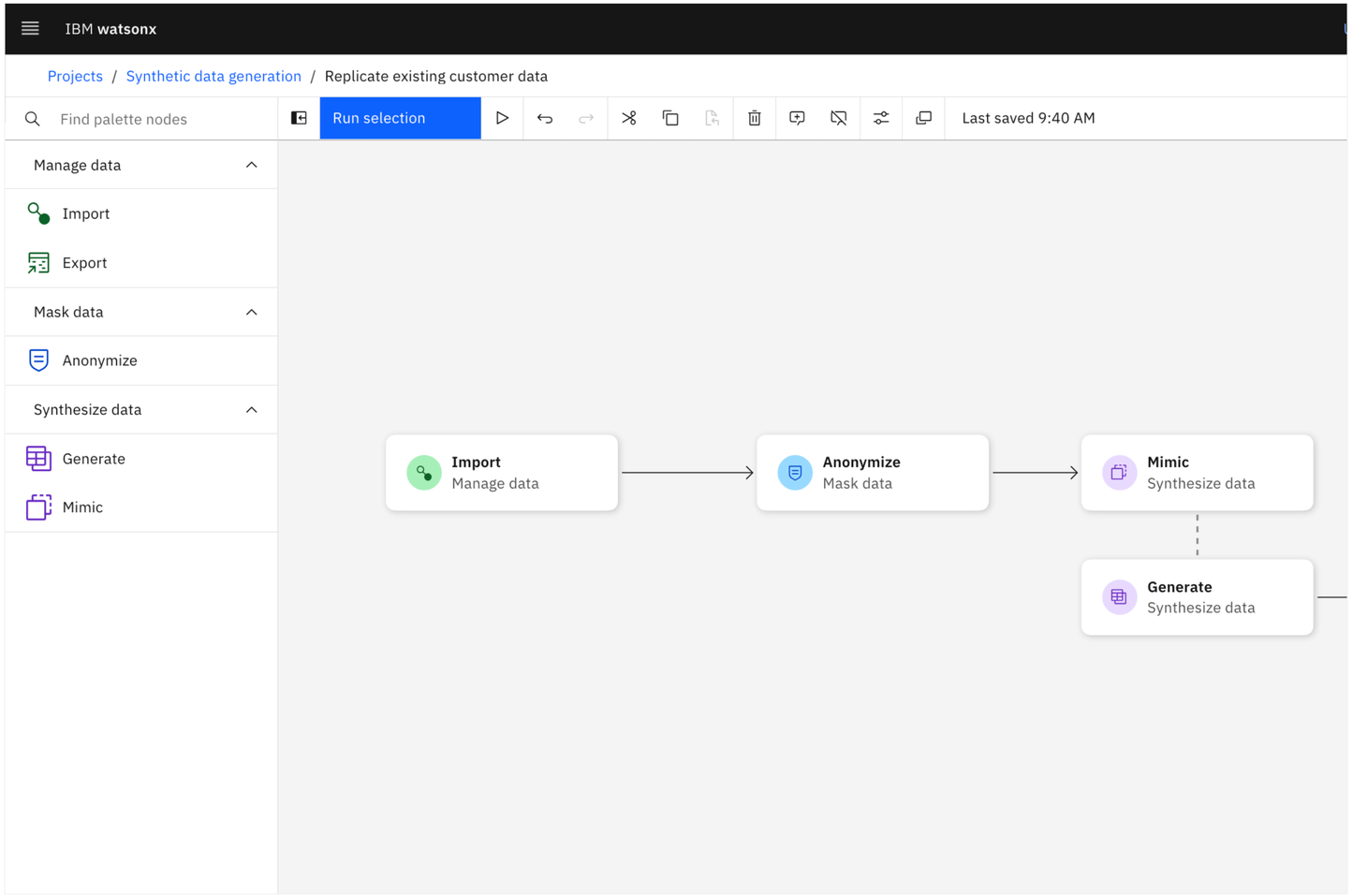
Create synthetic data using your existing data in a database or by uploading a file. If no data exists or can't be accessed, you can design your own data schema.

Address data gaps and create synthetic edge cases to expedite classical AI model training.

Select your model & privacy needs

Depending on your cost, fidelity, application, or data needs, you can select from multiple IBM models* to create your synthetic tabular data.

When using existing data, IBM models apply differential privacy to minimize your privacy risk and give you control over the level of privacy protection required for your organization.

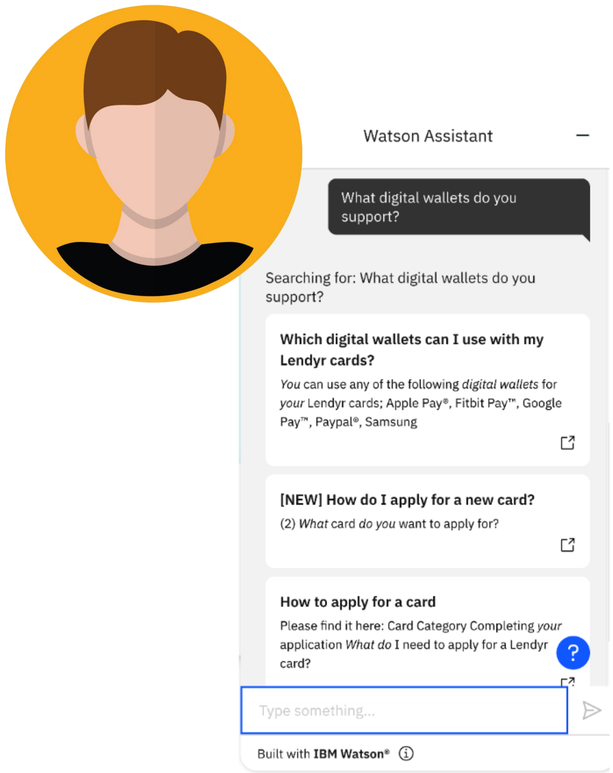


Evaluation metrics available in Q3 2024

IBM watsonx in action

Retrieval Augmented Generation (RAG) with Watsonx.ai.
RAG is available out of the box with Watsonx Assistant

Conversational search – Q&A for documents

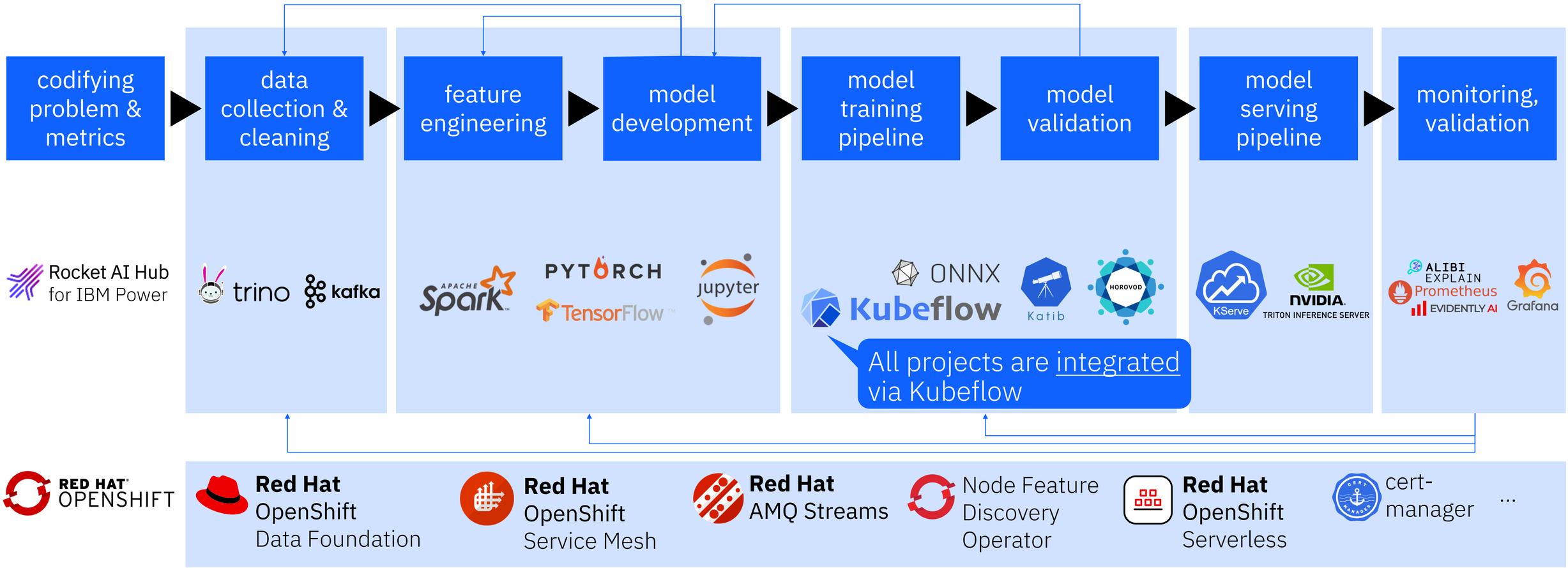


Augmented recovery generation (RAG):
Process of optimizing the result of a large language model. It uses a knowledge base external to reliable data sources used to train it before generating a response.

- Cost-effective (no model retraining, quick to implement),
- Accurate (sourced, up-to-date response),
- Flexible/Scalable.

https://research.ibm.com/blog/retrieval-augmented-generation-RAG

Machine Learning Operations (MLOps) Platform



Rocket AI Products

	Rocket CE	Rocket AI Hub
Built on open source?	Yes	Yes
Evaluate at no-cost?	Yes	Yes
Runs on Power?	Yes	Yes
Requires container platform?	No	Yes
Function	Core AI libraries	Core AI libraries AND: orchestration tools, graphical pipeline tools, and container-based workload management

High Level Integration Techniques

New HTTP Functions for SQL (QSYS2)

HTTP_GET, HTTP_POST, HTTP_PUT, HTTP_DELETE

Two part blog series:

– https://techchannel.com/SMB/09/2021/new-sql-http-functions-part-1
  - Overview of the new services
– https://techchannel.com/Trends/09/2021/sql-http-part-2
  - Integrating with OSS

A very different approach: Apache Camel

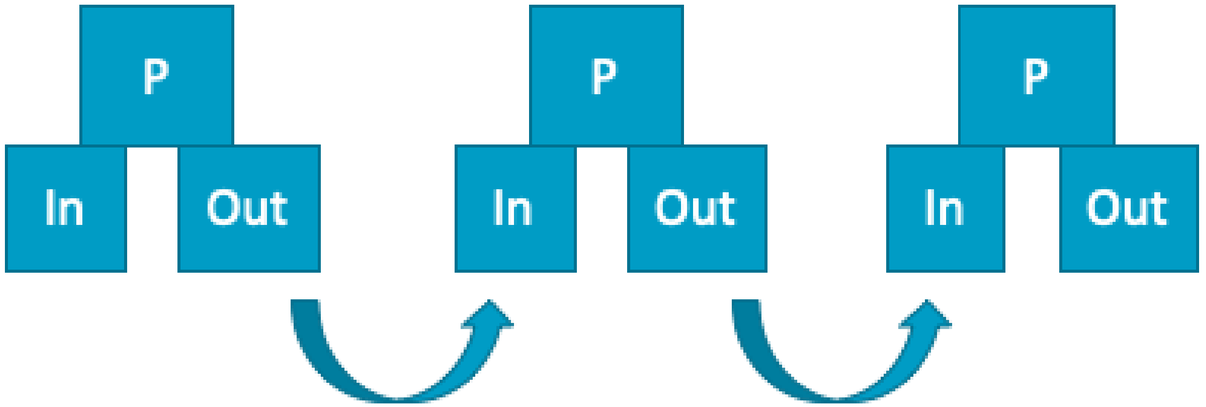
A Java-based integration framework

As Jesse says, "it can be used to connect anything to anything"

Information about Camel with IBM i:
– https://ibmi-oss-docs.readthedocs.io/en/latest/camel/README.html

How Does it Work?

Exchanges can be chained together – like piping commands through *NIX – to form a Camel Route
The "Out" message of a previous Exchange becomes the "in" message of a new Exchange
This defines the route



```
CREATE OR REPLACE FUNCTION COOLSTUFF.FRENCH (
            MESSAGE_DATA CLOB(64512)
    )
    RETURNS  CLOB(64512) CCSID 1208  LANGUAGE SQL  SPECIFIC COOLSTUFF.FRENCH NOT DETERMIN
    SET OPTION ALWBLK = *ALLREAD, ALWCPYDTA = *OPTIMIZE, COMMIT = *NONE, DECRESULT = (31,
    BEGIN
        DECLARE UNIQUIFIER VARCHAR(100);
        DECLARE LOCAL_MESSAGE_DATA_UTF8 CLOB(64512) CCSID 1208;
        DECLARE LOCAL_KEY_DATA VARCHAR(1000);
        SET UNIQUIFIER = QSYS2.JOB_NAME CONCAT QSYS2.THREAD_ID CONCAT CURRENT_TIMESTAMP;
        SET LOCAL_KEY_DATA = RPAD(UNIQUIFIER, 100, 'J');
        CALL QSYS2.SEND_DATA_QUEUE_UTF8(
            DATA_QUEUE_LIBRARY => 'COOLSTUFF',
            DATA_QUEUE => 'FRENCHQ',
            MESSAGE_DATA => MESSAGE_DATA,
            KEY_DATA => LOCAL_KEY_DATA
        );
        SELECT MESSAGE_DATA_UTF8
            INTO LOCAL_MESSAGE_DATA_UTF8
            FROM
                TABLE (
                    QSYS2.RECEIVE_DATA_QUEUE(
                        DATA_QUEUE_LIBRARY => 'COOLSTUFF',
                        DATA_QUEUE => 'FRENCHQ2', KEY_DATA => LOCAL_KEY_DATA, KEY_ORDER =>
                        WAIT_TIME => 10.050)
                );
        RETURN LOCAL_MESSAGE_DATA_UTF8;
    END;
```

Now we have AI in a UDF!

*demo.sql

```
1
2 VALUES( COOLSTUFF.FRENCH('The dog is loud'));
3
```

```
00001
Le chien est bruyant
```

WatsonX client SDK for Db2
https://github.com/IBM/WatsonX-SDK-Db2-IBMi

Run SQL Scripts - oss74dev.rch.stglabs.ibm.com(R1078935)

File Edit Search View Connection Run Explain Monitor Editor Tools Help

*Untitled 1

```
41
42
43
44 VALUES COOLSTUFF.WATSONXAI('why are armadillos so cute?')
45
```

```
00001
Armadillos are cute for a variety of reasons. Here are some
```

Done: 1 rows retrieved.

Messages Environment VALUES COOLSTUFF.WATSONXAIHTTP('wh

Connected to relational database R1078935 on oss74dev.rch.stglabs.ibm.com as JGORZINS - 220472/QUSER/QZDASOINIT using JDBC configu

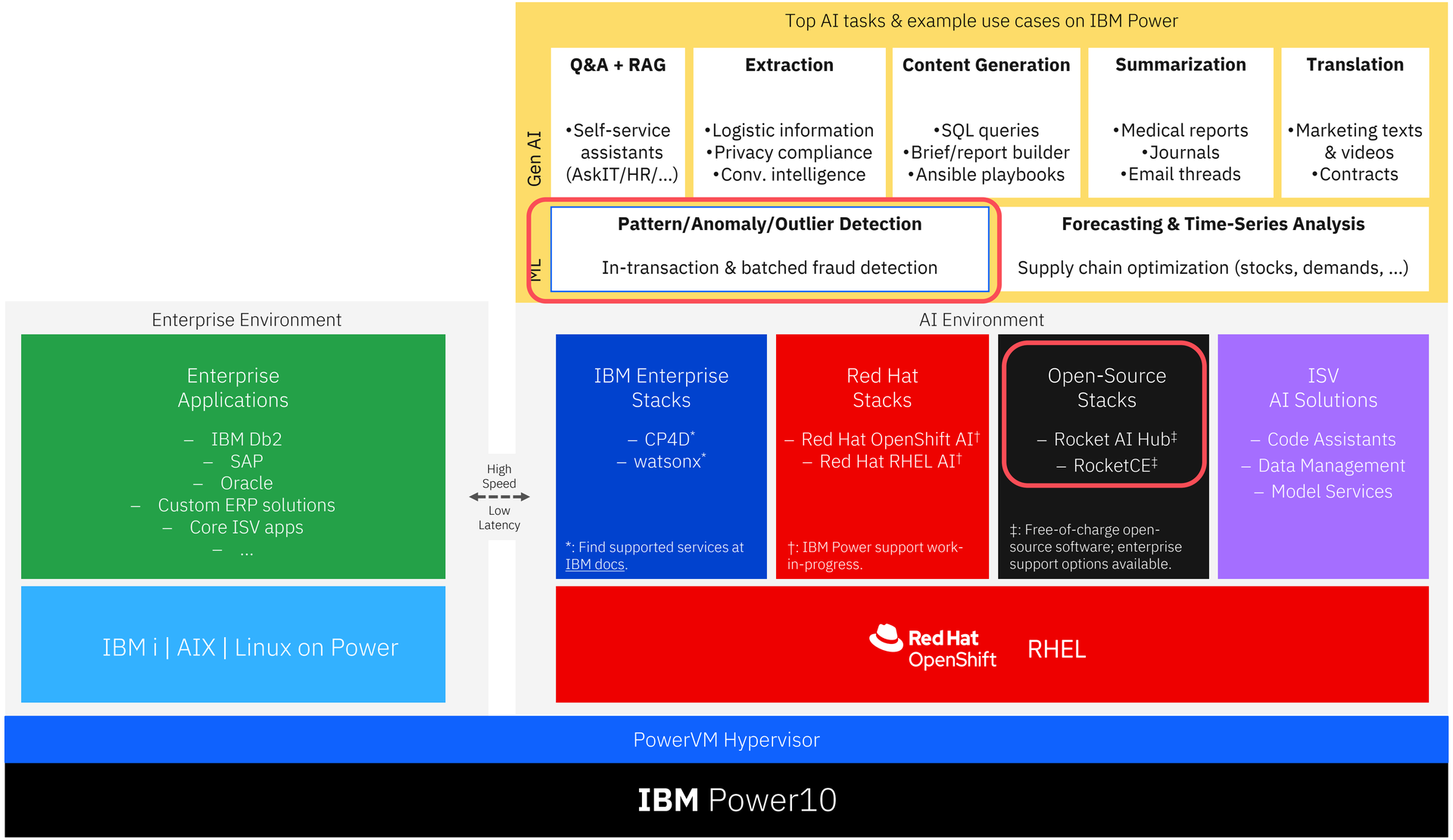
Recently changed:

- Rebuilt authentication layer
- Access WX from different geos
- Inference API to leverage custom-tuned models
- Multi-model support

Workflow with Rocket AI Hub

AI for business with **IBM** Power:

Reference architecture



SaaS

Empower individuals to work without expert knowledge across a variety of business processes & applications.

PaaS (Data & AI)

Leverage generative AI & machine learning — tuned with your data.

PaaS (OS)

Build on top of Red Hat OpenShift or start natively with RHEL.

IaaS

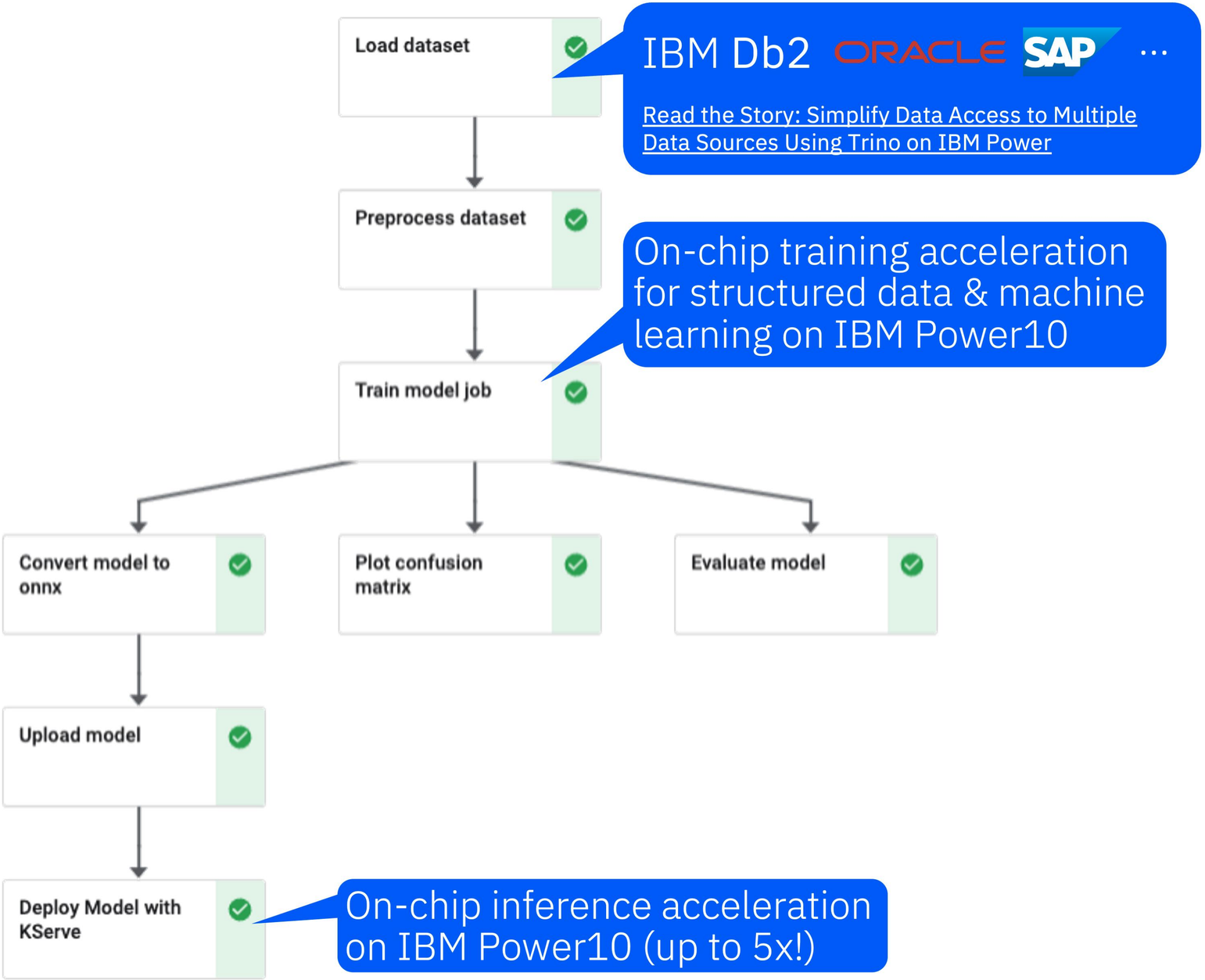
Accelerate, converge & safeguard AI efficiently with your data & workflows. On- and off-premises.

Pipelines

...for end-to-end orchestration

Data Scientist

- Reusable
- Versionable
- Repeatable
- Automatable
- Auditable



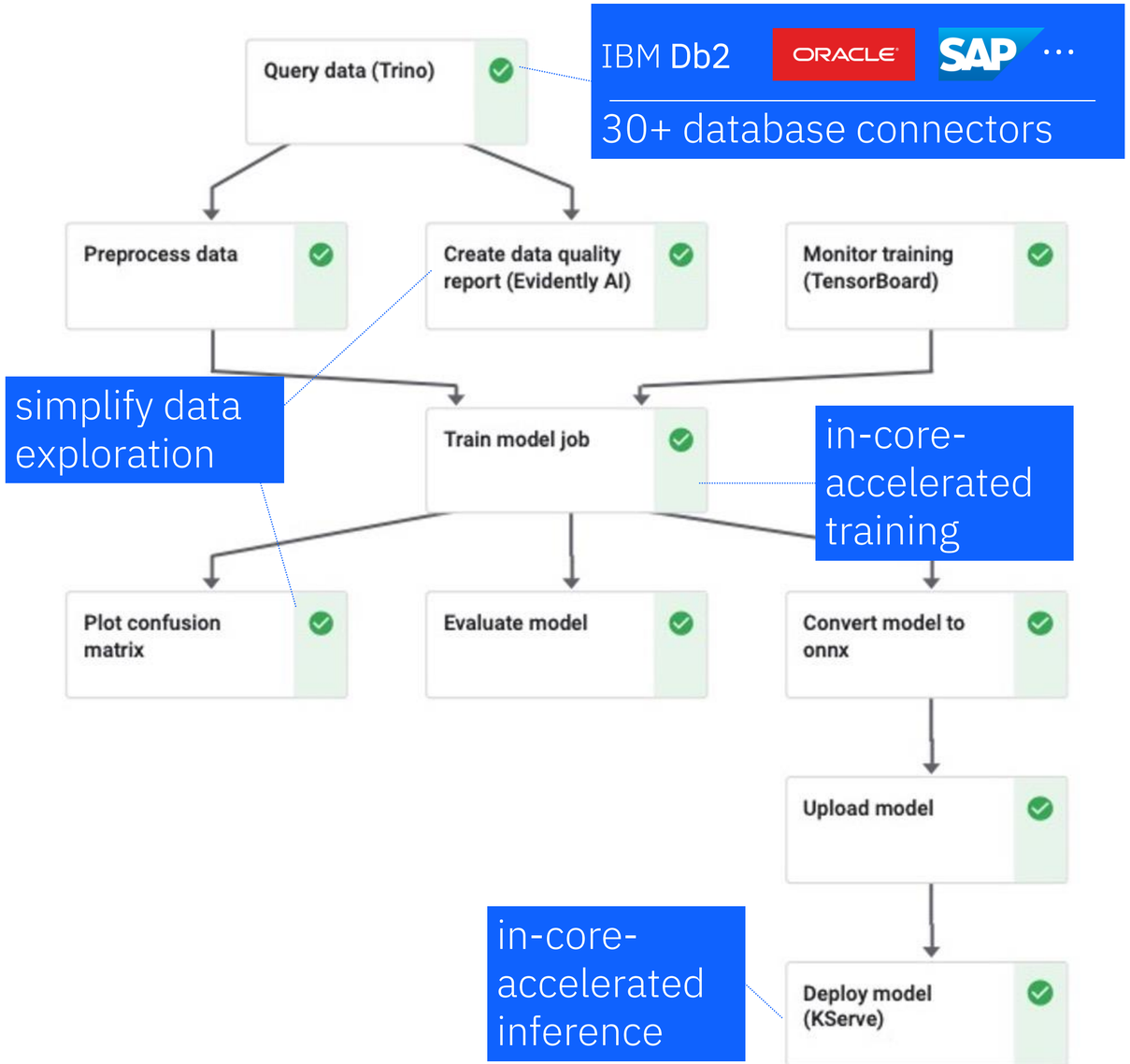
Rocket AI Hub for IBM Power

integrates the Open-Source AI portfolio.

Rocket AI Hub for IBM Power uses Kubeflow's MLOps pipelines to automate the data science workflow end-to-end. We provide a catalog of reusable pipeline components that integrate best-of-breed open-source AI tools and are optimized for IBM Power10. This makes data science simple & efficient on your trusted platform.

Our open-source AI blogs:

- MLOps with Kubeflow on IBM Power (Link)
- Simplify data access to multiple data sources using Trino on IBM Power (Link)
- True Hybrid Cloud for ML – Or: How I can burst my training to x86 & deploy back to Power (Link)
- Training on Steroids – Leveraging Distributed Training in Kubeflow (Link)



MLOps: Automating end-to-end AI workflows

Example: Real-time fraud detection close to IBM i data

Rocket AI Hub for IBM Power

Live data from:
- ATM machines
- Online banking
- …

Trusted data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	User	Card	Year	Month	Day	Time	Amount	Use Chip	Merchant Name	Merchant City	Merchant State	Zip	MCC	Is Fraud?	Errors?
2	0	0	2019	4	1	4:08	$39.58	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
3	0	0	2019	4	2	4:12	$36.98	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
4	0	0	2019	4	4	4:20	$40.36	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
5	0	0	2019	4	4	10:54	$7.98	Chip Transaction	Walmart	Brandon	FL	33510	5311	No	
6	0	0	2019	4	5	4:46	$49.87	Swipe Transaction	Nissan Service	Brandon	FL	33510	7538	No	
7	0	0	2019	4	5	10:59	$11.27	Chip Transaction	Applebees	Brandon	FL	33511	5812	No	

Credit Card Transactions

IBM Db2

AI Model

1 Integrate & ingest mission critical data

trino

2 Train & fine-tune AI models

EVIDENTLY AI Explore

jupyter TensorFlow Experiment & Train

TensorBoard TensorFlow Monitor & Evaluate

ONNX Export

3 Inference close to mission critical data

IBM i | PASE for i

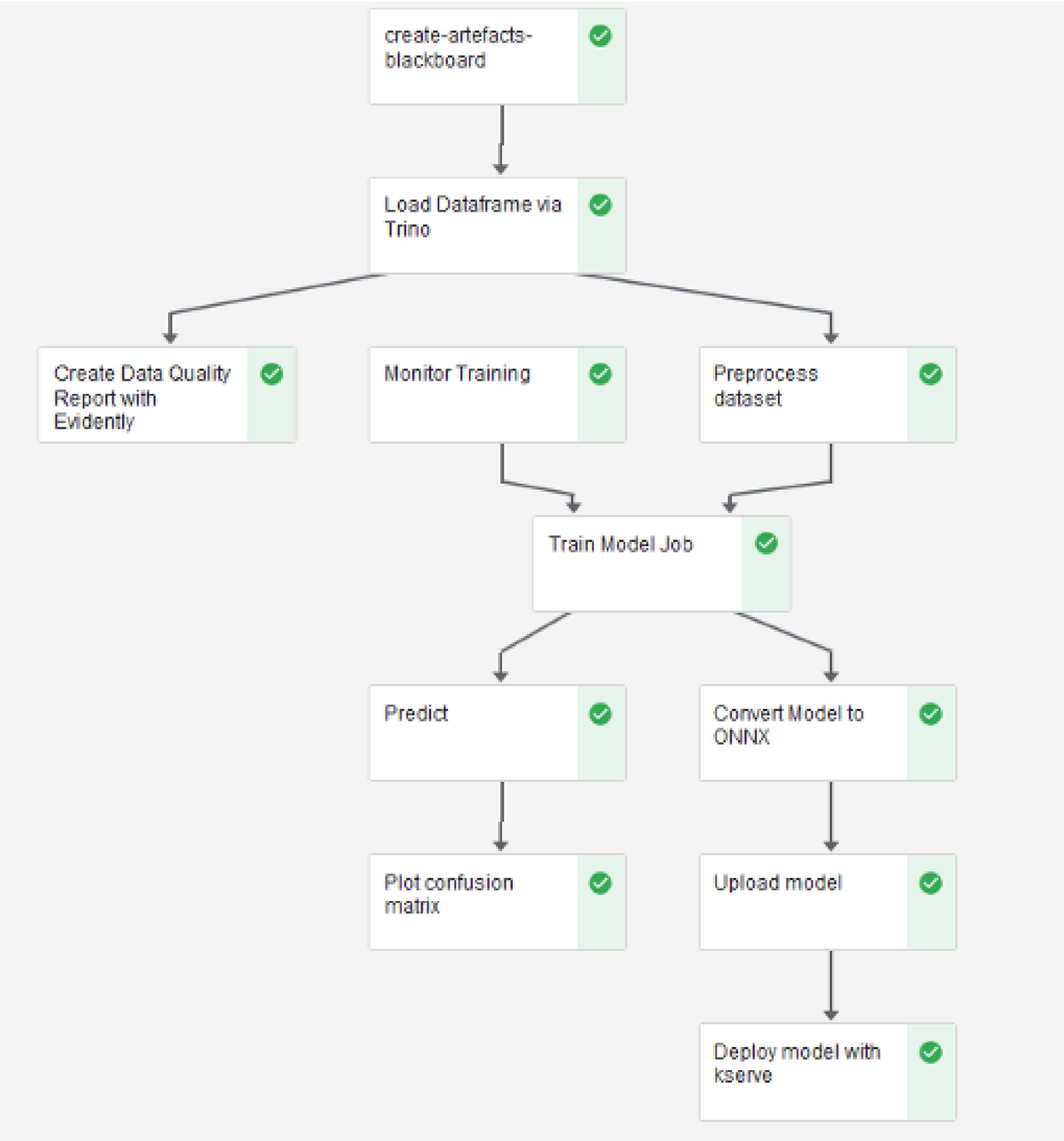
Linux on Power

Kubeflow End-to-End Orchestration

Red Hat OpenShift

IBM Power10

The pipeline automatically loads data from IBM i, trains a model & deploys the newly trained AI model to both KServe and IBM i*.



* On system inference forthcoming

MLOps: Automating end-to-end AI workflows

Example: Real-time fraud detection close to IBM i data

Rocket AI Hub for IBM Power

Live data from:
- ATM machines
- Online banking
- …

Trusted data

User	Card	Year	Month	Day	Time	Amount	Use Chip	Merchant Name	Merchant City	Merchant State	Zip	MCC	Is Fraud?	Errors?
0	0	2019	4	1	4:08	$39.58	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
0	0	2019	4	2	4:12	$36.98	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
0	0	2019	4	4	4:20	$40.36	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
0	0	2019	4	4	10:54	$7.98	Chip Transaction	Walmart	Brandon	FL	33510	5311	No	
0	0	2019	4	5	4:46	$49.87	Swipe Transaction	Nissan Service	Brandon	FL	33510	7538	No	
0	0	2019	4	5	10:59	$11.27	Chip Transaction	Applebees	Brandon	FL	33511	5812	No	

Credit Card Transactions

IBM Db2

AI Model

1 Integrate & ingest mission critical data

trino

2 Train & fine-tune AI models

EVIDENTLY AI Explore

jupyter TensorFlow Experiment & Train

TensorBoard TensorFlow Monitor & Evaluate

ONNX Export

3 Inference close to mission critical data

IBM i | PASE for i

Linux on Power

Kubeflow End-to-End Orchestration

Red Hat OpenShift

IBM Power10

Example: 24M credit card transactions of a fictional bank

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	User	Card	Year	Month	Day	Time	Amount	Use Chip	Merchant Name	Merchant City	Merchant State	Zip	MCC	Is Fraud?	Errors?
2	0	0	2019	4	1	4:08	$39.58	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
3	0	0	2019	4	2	4:12	$36.98	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
4	0	0	2019	4	4	4:20	$40.36	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
5	0	0	2019	4	4	10:54	$7.98	Chip Transaction	Walmart	Brandon	FL	33510	5311	No	
6	0	0	2019	4	5	4:46	$49.87	Swipe Transaction	Nissan Service	Brandon	FL	33510	7538	No	
7	0	0	2019	4	5	10:59	$11.27	Chip Transaction	Applebees	Brandon	FL	33511	5812	Yes	

Most transactions are legitimate – but also frauds exist.

Search

Untitled-1

```
1  select * from demo.fraud
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS IBM I COMMENTS

USER	CARD	YEAR	MONTH	DAY	TIME	AMOUNT	USE CHIP	MERCHANT NAME
0	0	2002	9	1	06:21	$134.09	Swipe Transaction	3527213246127877000
0	0	2002	9	1	06:42	$38.48	Swipe Transaction	-727612092139916000
0	0	2002	9	2	06:22	$120.34	Swipe Transaction	-727612092139916000
0	0	2002	9	2	17:45	$128.95	Swipe Transaction	3414527459579107000
0	0	2002	9	3	06:23	$104.71	Swipe Transaction	5817218446178736000
0	0	2002	9	3	13:53	$86.19	Swipe Transaction	-714667074812520100
0	0	2002	9	4	05:51	$93.84	Swipe Transaction	-727612092139916000
0	0	2002	9	4	06:09	$123.50	Swipe Transaction	-727612092139916000
0	0	2002	9	5	06:14	$61.72	Swipe Transaction	-727612092139916000

Loaded 100. More available. 060071/QUSER/QZDASOINIT

DB2 FOR I

SCHEMA BROWSER
- DEMO
  - Tables
    - card_transactions
    - fraud card transactions
      - user NUMERIC, 0 def.
      - card NUMERIC, 0 def.
      - year NUMERIC, 0 def.
      - month NUMERIC, 0 def.
      - day NUMERIC, 0 def.
      - time CHAR(5), ' ' def.
      - amount CHAR(9), ' ' def.
      - "USE CHIP" CHAR(18), ' ' def.
      - "MERCHANT NAME" NUMERIC, 0 def.
      - "MERCHANT CITY" CHAR(26), ' ' def.
      - "MERCHANT STATE" CHAR(32), ' ' def.
      - zip NUMERIC, 0 def.
      - mcc NUMERIC, 0 def.
      - errors? CHAR(52), ' ' def.
      - "IS FRAUD?" CHAR(3), ' ' def.
    - orders
  - Triggers

STATEMENT HISTORY

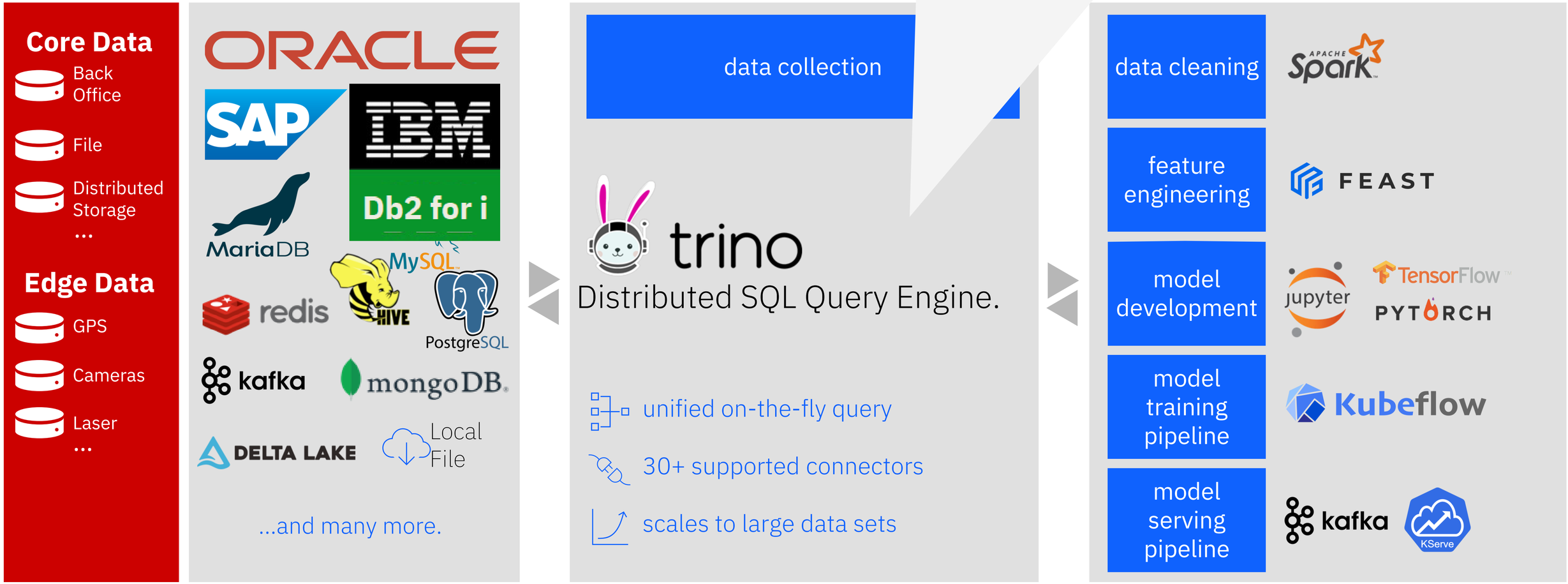
SQL JOB MANAGER
- New job (1) 060071/QUSER/QZDASOINIT
- Saved Configuration
  - demo
  - demo2

EXAMPLES

Launchpad 0 0 0 p114ibmi01.pbm.ihost.com Live Share New job (1) Go Live Spell Reload Prettier

Unify Data Collection with Trino

```
SELECT * FROM mongodb.weather.ny w JOIN
postgresql.public.stockhistory s ON w._id = s.Date
JOIN db2fori.transactions t ON s.Date = t.transDate
WHERE s.Date < date \'2022-08-05\' ORDER BY date
ASC LIMIT 42;
```



[See our blog: https://community.ibm.com/community/user/powerdeveloper/blogs/natalie-jann/2022/11/07/simplify-data-access-using-trino-on-ibm-power]

Access Db2 for i Data via Trino Python API

In [3]:

```python
import json
import os
import requests

from tensorflow import keras
import pandas as pd

%load_ext lab_black
```

In [43]:

```python
def get_data_table(rows: int):
    import pandas as pd
    from trino.dbapi import Connection

    with Connection(
        host="trino.trino",
        port="8080",
        user="anybody",
        catalog="jtopen",
        schema="demo",
    ) as conn:
        link = conn.cursor()
        link.execute(f"SELECT * FROM fraud LIMIT {rows}")
        return pd.DataFrame(link.fetchall(), columns=[i.name for i in link.description])


rdf = get_data_table(100000)
print(f"Retrieved {len(rdf)} rows")
rdf.head()
```

```
Retrieved 100000 rows
```

Out[43]:

	user	card	year	month	day	time	amount	use chip	merchant name	merchant city	merchant state	zip	mcc	errors?
0	0	0	2002	9	1	06:21	$134.09	Swipe Transaction	3527213246127876953	La Verne	CA	91750	5300	...
1	0	0	2002	9	1	06:42	$38.48	Swipe Transaction	-727612092139916043	Monterey Park	CA	91754	5411	...
2	0	0	2002	9	2	06:22	$120.34	Swipe Transaction	-727612092139916043	Monterey Park	CA	91754	5411	...
3	0	0	2002	9	2	17:45	$128.95	Swipe Transaction	3414527459579106770	Monterey Park	CA	91754	5651	...
4	0	0	2002	9	3	06:23	$104.71	Swipe Transaction	5817218446178736267	La Verne	CA	91750	5912	...

Kubeflow Central Dashboard

https://kubeflow.apps.b2s001.pbm.ihost.com/_/pipeline/?ns=user-example-com#/runs/details/a864491c-e9cb-4d64-84d9-cfe19ea037b7

Kubeflow

user-example-com (Owner)

- Home
- Notebooks
- Tensorboards
- Volumes
- Models
- Experiments (AutoML)
- Experiments (KFP)
- Pipelines
- Runs
- Recurring Runs
- Artifacts
- Executions
- Object Store
- Monitoring

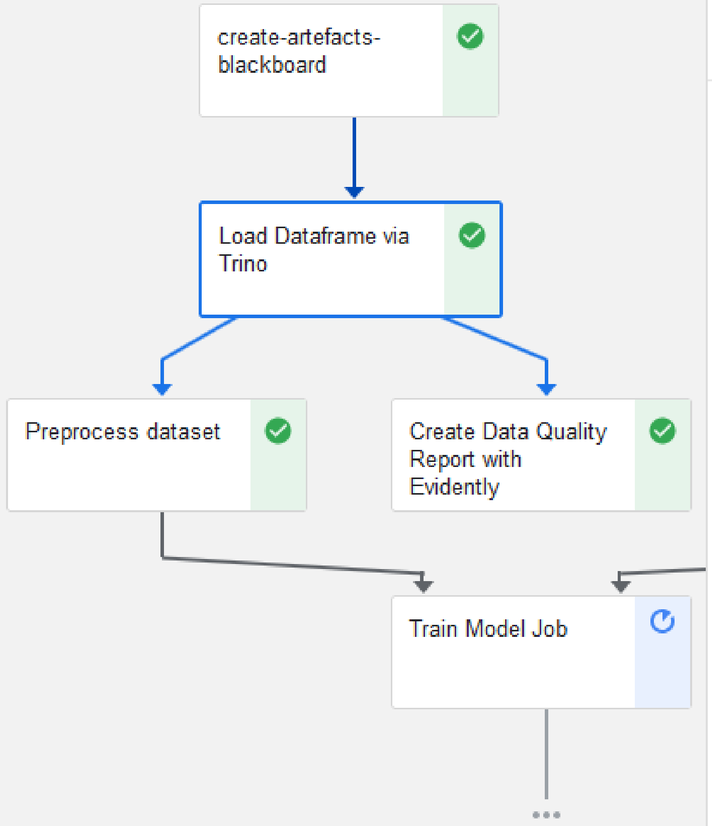
Experiments › ibmi-fraud

Retry | Clone run | Terminate | Archive

Clone of ibmi-fraud

Graph | Run output | Config

Simplify Graph



fraud-detection-qt8np-3281257605

Input/Output | Visualizations | Details | Volumes | Logs | Pod | Events | ML

Load Dataframe via Trino

fraud-detection-qt8np

run_id
fraud-detection-qt8np

Custom Properties

input:columns_query
SHOW COLUMNS FROM jtopen.demo.fraud

input:host
trino.trino

input:port
8080

input:query
SELECT * FROM jtopen.demo.fraud limit 1000000

input:user
anybody

kfp_pod_name
fraud-detection-qt8np-3281257605

Outputs

Kubeflow

user-example-com (Owner)

Home
Notebooks
Tensorboards
Volumes
Models
Experiments (AutoML)
Experiments (KFP)
Pipelines
Runs
Recurring Runs
Artifacts
Executions
Object Store
Monitoring

Manage Contributors

GitHub
Documentation

Privacy • Usage Reporting
build version dev_local

Experiments > ibmi-fraud

Retry Clone run Terminate Archive

ibmi-fraud

Graph Run output Config

Simplify Graph

create-artefacts-blackboard

Load Dataframe via Trino

Create Data Quality Report with Evidently

Monitor Training

Train Model Job

Predict

Plot confusion matrix

Runtime execution graph. Only steps that are currently running or have already completed are sho

fraud-detection-bwtnn-1512622821

Input/Output Visualizations Details Volumes Logs Pod Events ML Metadata

```
time="2024-04-23T18:16:41.625Z" level=info msg="capturing logs" argo=true
INFO 2024-04-23 18:16:42,091: Establishing Trino connection...
INFO 2024-04-23 18:16:42,093: Querying data...
INFO 2024-04-23 18:17:06,881: Retrieved 1000000 rows.
INFO 2024-04-23 18:17:06,881: Querying column names...
INFO 2024-04-23 18:17:07,309: Using columns: ['user', 'card', 'year', 'month', 'day', 'time', 'amount', 'use chip',
INFO 2024-04-23 18:17:28,353: Finished.
time="2024-04-23T18:17:29.192Z" level=info msg="/tmp/outputs/dataframe/data -> /var/run/argo/outputs/artifacts/tmp/ou
time="2024-04-23T18:17:29.192Z" level=info msg="Taring /tmp/outputs/dataframe/data"
```

MLOps: Automating end-to-end AI workflows

Example: Real-time fraud detection close to IBM i data

Rocket AI Hub for IBM Power

Live data from:
- ATM machines
- Online banking
- …

Trusted data

User	Card	Year	Month	Day	Time	Amount	Use Chip	Merchant Name	Merchant City	Merchant State	Zip	MCC	Is Fraud?	Errors?
0	0	2019	4	1	4:08	$39.58	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
0	0	2019	4	2	4:12	$36.98	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
0	0	2019	4	4	4:20	$40.36	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
0	0	2019	4	4	10:54	$7.98	Chip Transaction	Walmart	Brandon	FL	33510	5311	No	
0	0	2019	4	5	4:46	$49.87	Swipe Transaction	Nissan Service	Brandon	FL	33510	7538	No	
0	0	2019	4	5	10:59	$11.27	Chip Transaction	Applebees	Brandon	FL	33511	5812	No	

Credit Card Transactions

IBM Db2

AI Model

1 Integrate & ingest mission critical data

trino

2 Train & fine-tune AI models

EVIDENTLY AI Explore

jupyter TensorFlow Experiment & Train

TensorBoard TensorFlow Monitor & Evaluate

ONNX Export

3 Inference close to mission critical data

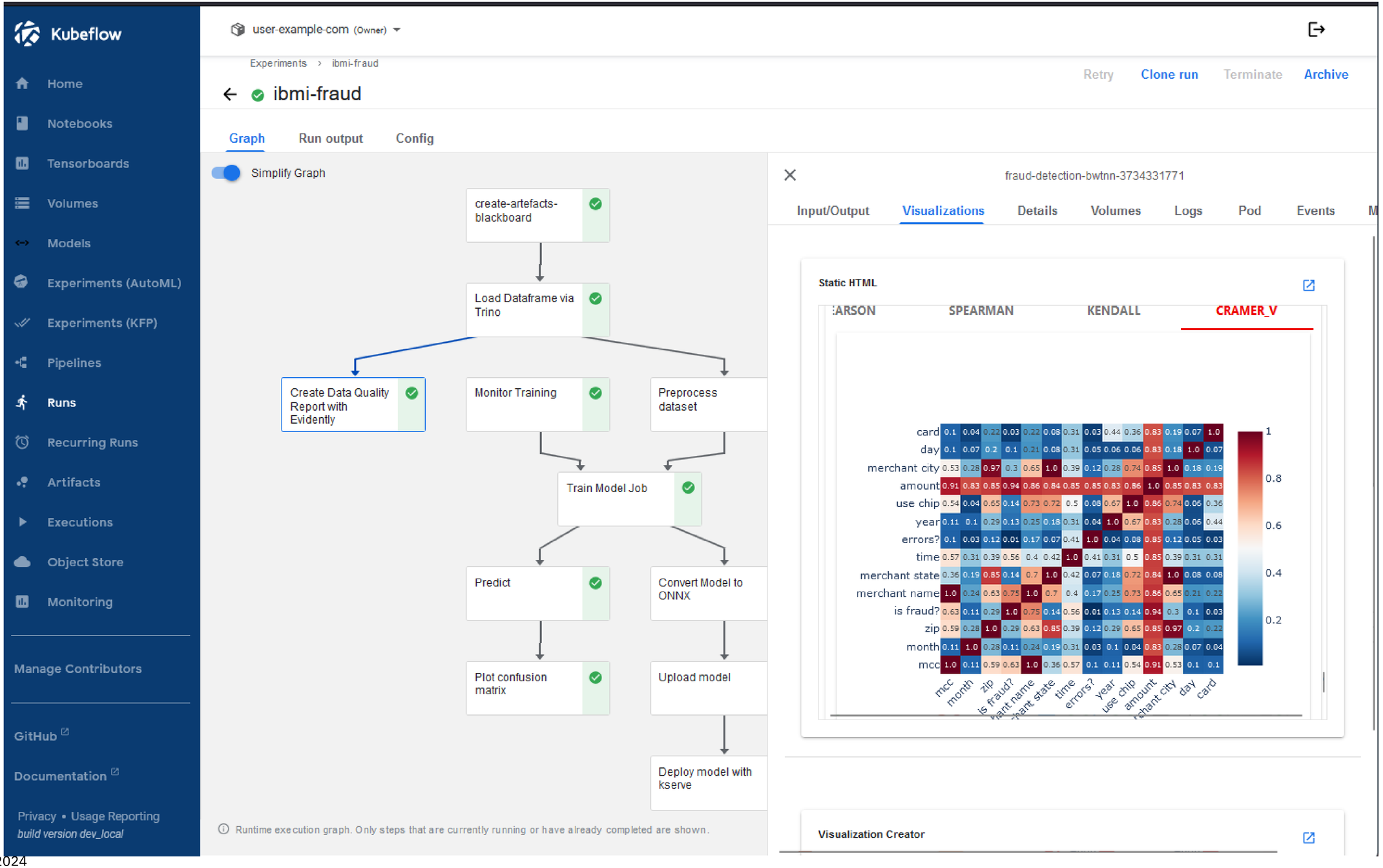
IBM i | PASE for i

Linux on Power

Kubeflow End-to-End Orchestration

Red Hat OpenShift

IBM Power10



Burst training to where it belongs best.

Train model job

- ✓ Only reserve when & what is needed
  → Save costs
- ✓ No need to operate own GPU hardware
  → Reduce complexity
- ✓ Hybrid: IBM Power / x86 / Q
  → Avoid vendor lock-in

Read the Story: True Hybrid Cloud for ML – Or: How I can burst my training to x86 & deploy back to Power

Structured data | classical machine learning | fine-tuning

RED HAT OPENSHIFT, IBM Z

Unstructured data | deep learning

RED HAT OPENSHIFT, NVIDIA, intel, kubernetes, Open Telekom Cloud, IBM Cloud

Big data jobs (HPC, foundation models, ...)

RAY, APACHE Spark, IBM Spectrum LSF, IBM Z, IBM Cloud

Quantum simulation & jobs

RED HAT OPENSHIFT, Qiskit, IBM Z, Q

MLOps: Automating end-to-end AI workflows

Example: Real-time fraud detection close to IBM i data

Rocket AI Hub for IBM Power

Live data from:
- ATM machines
- Online banking
- …

Trusted data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	User	Card	Year	Month	Day	Time	Amount	Use Chip	Merchant Name	Merchant City	Merchant State	Zip	MCC	Is Fraud?	Errors?
2	0	0	2019	4	1	4:08	$39.58	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
3	0	0	2019	4	2	4:12	$36.98	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
4	0	0	2019	4	4	4:20	$40.36	Chip Transaction	Kelly Auto Repair	Brandon	FL	33510	7538	No	
5	0	0	2019	4	4	10:54	$7.98	Chip Transaction	Walmart	Brandon	FL	33510	5311	No	
6	0	0	2019	4	5	4:46	$49.87	Swipe Transaction	Nissan Service	Brandon	FL	33510	7538	No	
7	0	0	2019	4	5	10:59	$11.27	Chip Transaction	Applebees	Brandon	FL	33511	5812	No	

Credit Card Transactions

IBM Db2

AI Model

1 Integrate & ingest mission critical data

trino

2 Train & fine-tune AI models

EVIDENTLY AI Explore

jupyter TensorFlow Experiment & Train

TensorBoard TensorFlow Monitor & Evaluate

ONNX Export

3 Inference close to mission critical data

IBM i | PASE for i

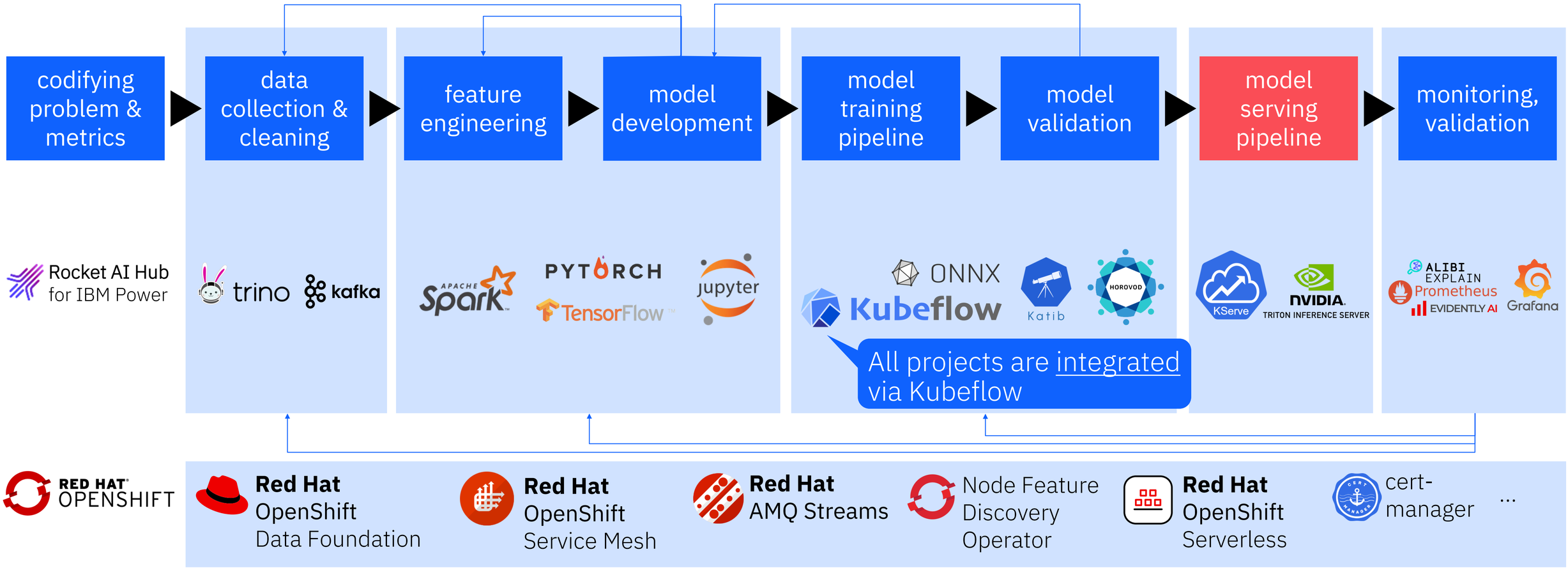
Linux on Power

Kubeflow End-to-End Orchestration

Red Hat OpenShift

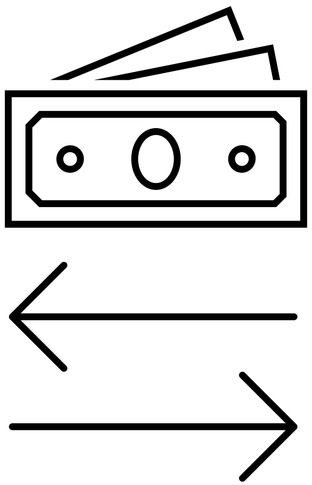
IBM Power10

Machine Learning Operations (MLOps) Platform



Model Serving Pipeline
Optimized for IBM Power 10

Inference Data



KServe/Triton Inference Server

ONNX

PyTorch

TensorFlow

→ Up to 5x acceleration on IBM Power 10*
→ No GPU hardware needed!

Inference Result

That's a **Fraud!**

* 5x improvement in per socket inferencing throughput for large size 32b floating point inferencing models from Power9 E980 (12-core modules) to Power10 E1080 (15-core modules).
Based on IBM testing using PyTorch, OpenBLAS on the same BERT Large with SQuAD v1.1 data set

```
curl  -s -k -X POST https://demo-application-onnx-user-example-com.apps.b2s001.pbm.ihost.com/api/model/
predict -H "Content-Type: application/json" -d '{
    "index": 1,
    "user": 2,
    "card": 4,
    "year": 2022,
    "month": 9,
    "day": 2,
    "time": "14:09",
    "amount": "$149345.84",
    "use chip": "Online Transaction",
    "merchant name": 3452760747765970571,
    "merchant city": "ONLINE",
    "merchant state": "",
    "zip": 0,
    "mcc": 3174,
    "errors?": "",
    "is fraud?": "Yes"
}'
```

```
curl  -s -k -X POST https://demo-application-onnx-user-example-com.apps.b2s001.pbm.ihost.com/api/model/
predict -H "Content-Type: application/json" -d '{
    "index": 1,
    "user": 2,
    "card": 4,
    "year": 2022,
    "month": 9,
    "day": 2,
    "time": "14:09",
    "amount": "$149345.84",
    "use chip": "Online Transaction",
    "merchant name": 3452760747765970571,
    "merchant city": "ONLINE",
    "merchant state": "",
    "zip": 0,
    "mcc": 3174,
    "errors?": "",
    "is fraud?": "Yes"
}'
```

```
{
  "result": 0.24492061138153076,
  "time": 51.978
}
```

1 Integrate & ingest mission critical data

trino

2 Train & fine-tune AI models

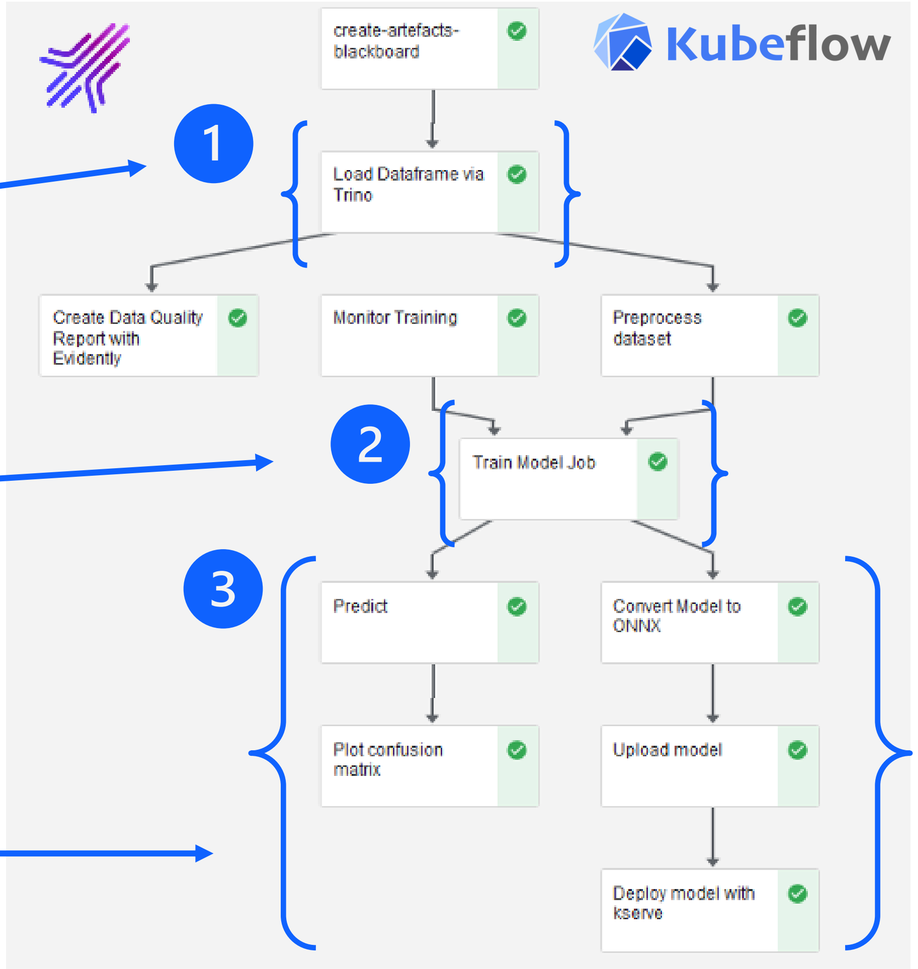
EVIDENTLY AI Explore

jupyter TensorFlow Experiment & Train

TensorBoard TensorFlow Monitor & Evaluate

ONNX Export

3 Inference close to mission critical data



Linux on Power

Kubeflow End-to-End Orchestration

Red Hat OpenShift

IBM Power10

Any Questions?

A Personal Note


SEPTEMBER 2019 IBMSYSTEMSMAG.COM
Power Systems
IBM Systems
THE OPEN-SOURCE REVOLUTION
PAGE 16
Node.js
PAGE 22
IBM Q
PAGE 34
JESSE GORZINSKI
BUSINESS ARCHITECT,
OPEN SOURCE ON IBM i

IBM i continuous innovation continuous integration

IBM i + AI

For More Information

IBM i
continuous innovation
continuous integration

Links You Need	Twitter	#Hashtags
IBM i Home Page: https://www.ibm.com/it-infrastructure/power/os/ibm-i (find link to Forrester Study and updated IBM i Strategy Whitepaper) IBM Strategy Whitepaper: https://www.ibm.com/it-infrastructure/us-en/resources/power/i-strategy-roadmap/ IBM Client Success: https://www.ibm.com/it-infrastructure/us-en/resources/power/ibm-i-customer-stories/ Support Life Cycle: https://www.ibm.com/support/lifecycle/ License Topics: https://www-01.ibm.com/support/docview.wss?uid=nas8N1022087 Fortra IBM i Marketplace Survey https://www.fortra.com/resources/guides/ibm-i-marketplace-survey-results	@IBMSystems @COMMONug @IBMChampions @IBMSystemsISVs @IBMiMag @ITJungleNews @SAPonIBMi @SiDforIBMi	#PowerSystems #IBMi #IBMAIX #POWER9 #LinuxonPower #OpenPOWER #HANAonPower #ITinfrastructure #OpenSource #HybridCloud #BigData

IBM

Thank you!!