ORACLE

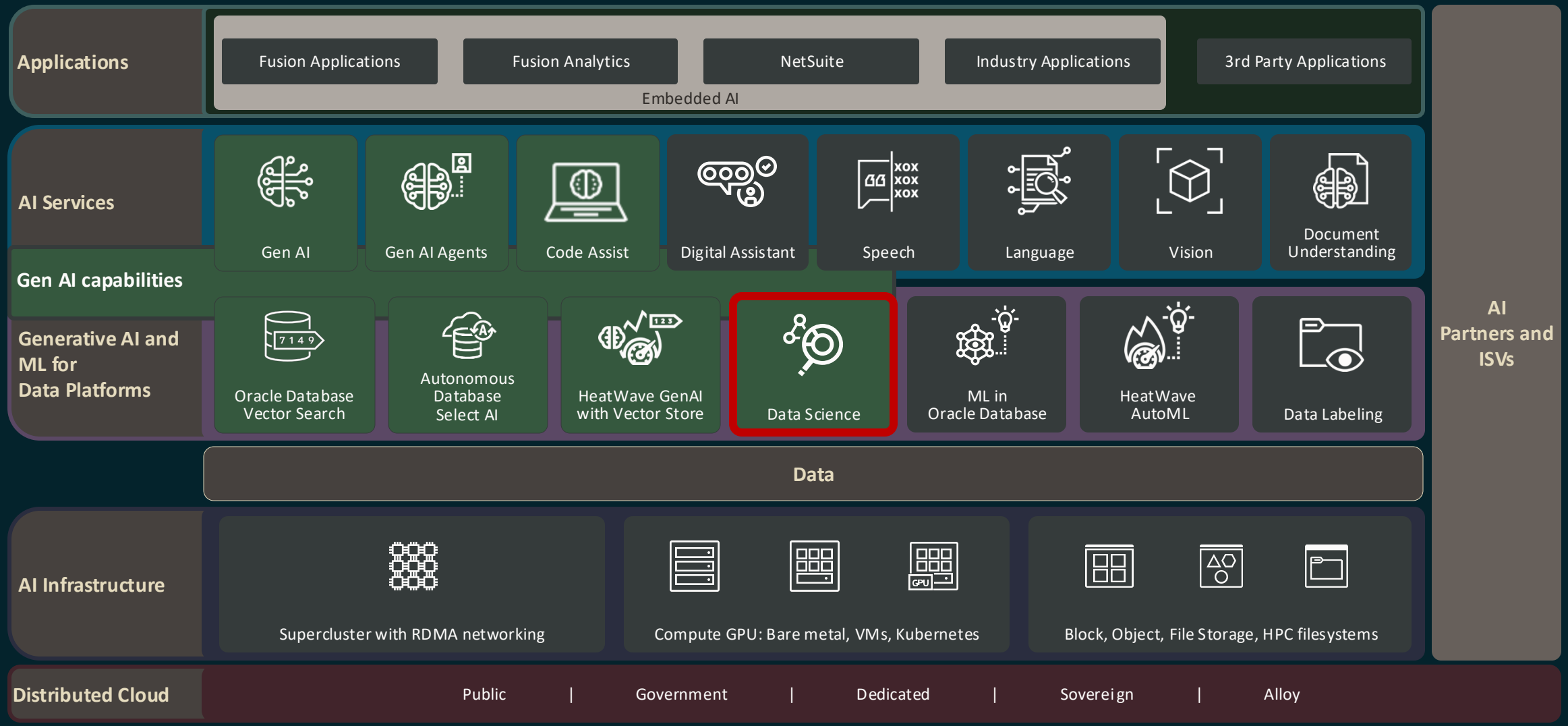
OCI Data Science

Oracle Data Science in action

Oracle AI

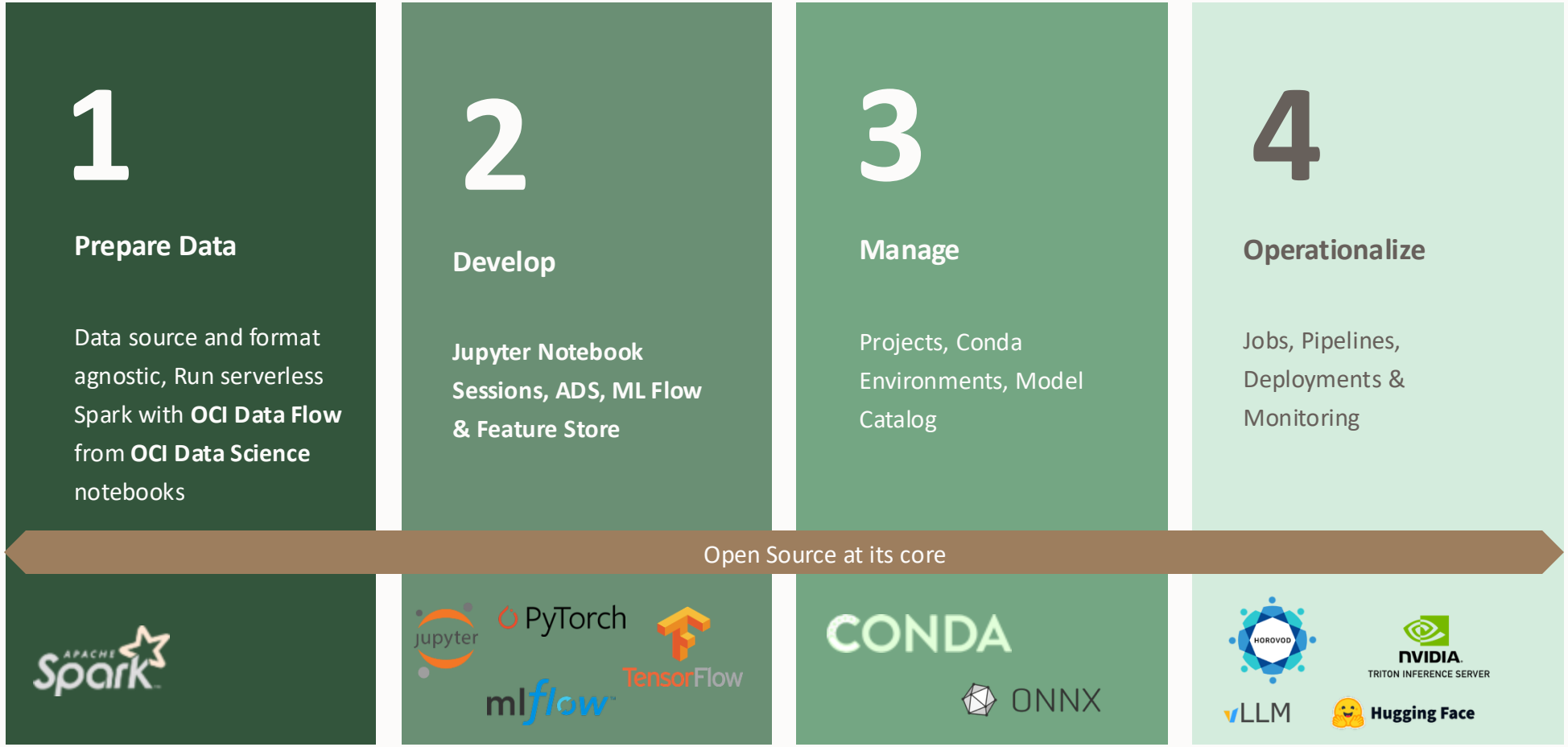
nfidential - Oracle Restricted

The Oracle AI stack – Generative AI services



Confidential - Oracle Restricted

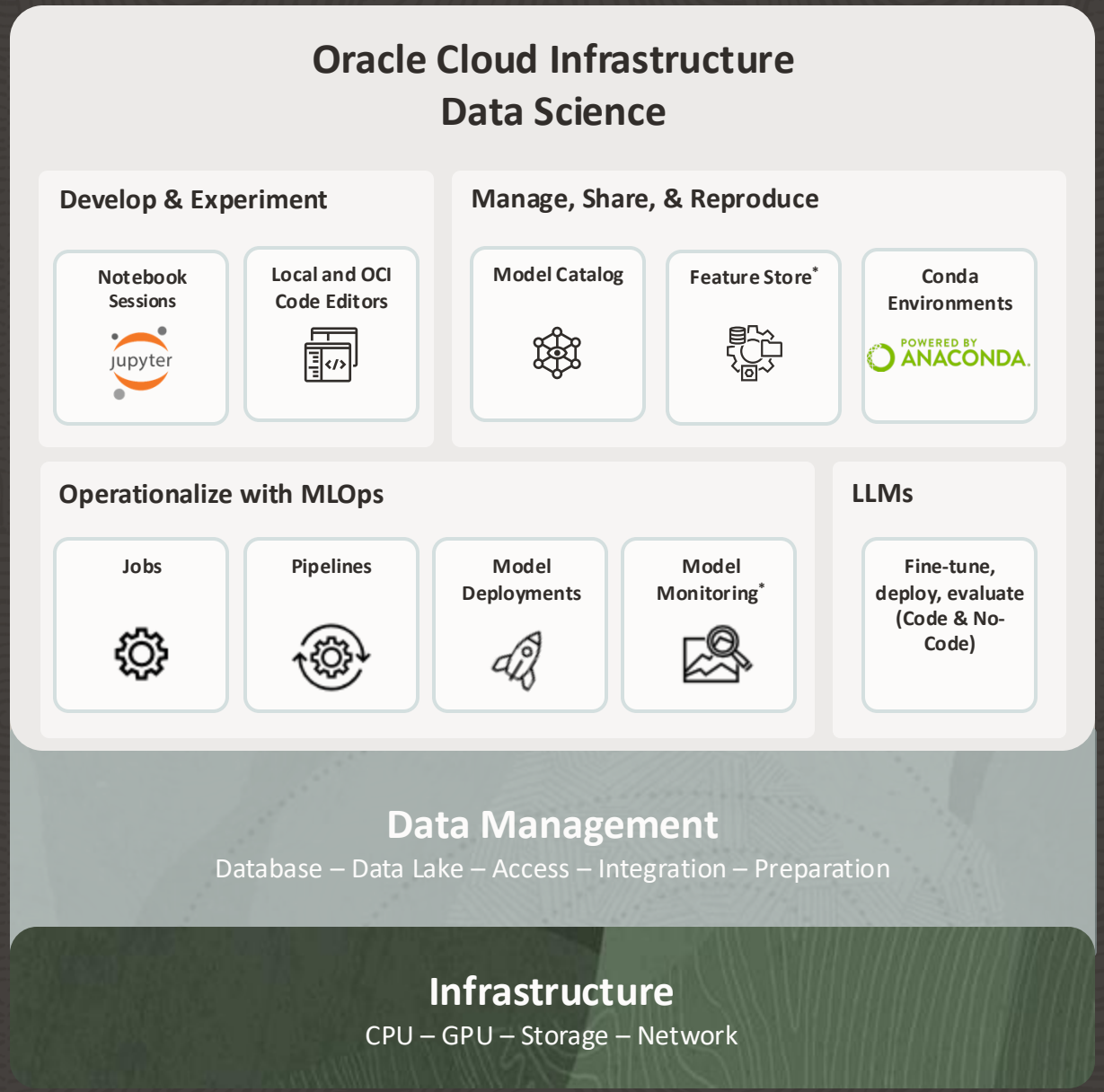
Data Science service – Full ML lifecycle Platform



Innovate with AI on Oracle Cloud

Develop effortlessly, operationalize at scale

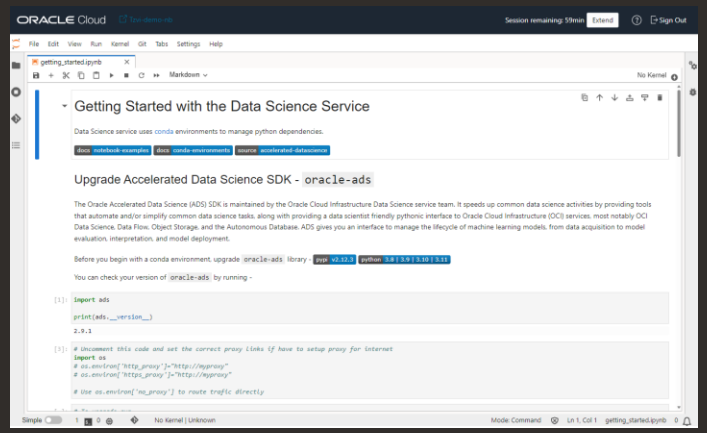
- Accelerate and automate the entire end-to-end data science lifecycle
- Use your favorite open-source Python tools and frameworks
- Fine-tune and deploy Large Language Models (LLMs) without writing code
- Enterprise-grade MLOps with flexible interfaces and unlimited scale
- Collaborate with teammates on shareable and reproducible data science assets
- Run large-scale workloads with access to GPUs and distributed data processing and model training
- Pay only for on demand infrastructure with no additional markup or overhead



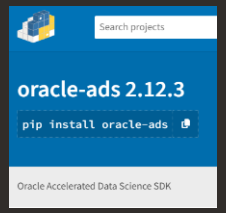
Enabling every team and every skillset

Boost data science productivity, Democratize AI

Jupyter Notebooks



Python based model training and inference



Open-source Accelerated Data Science (ADS) SDK to streamline data scientist work

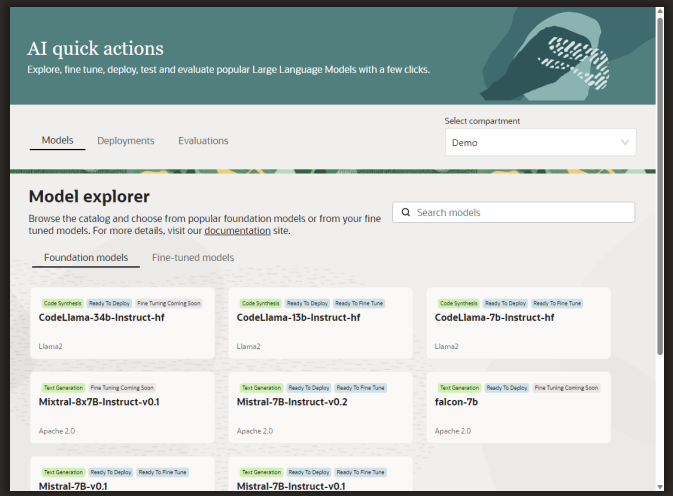
Code-first

No-Code / Low-Code

Share

- Projects
- Models
- Environments

AI Quick Actions



Low-code **operators** for specific use cases:

- Anomaly detection
- Time Series Forecasting
- PII detection

Confidential - Oracle Restricted

Why use OCI Data Science?

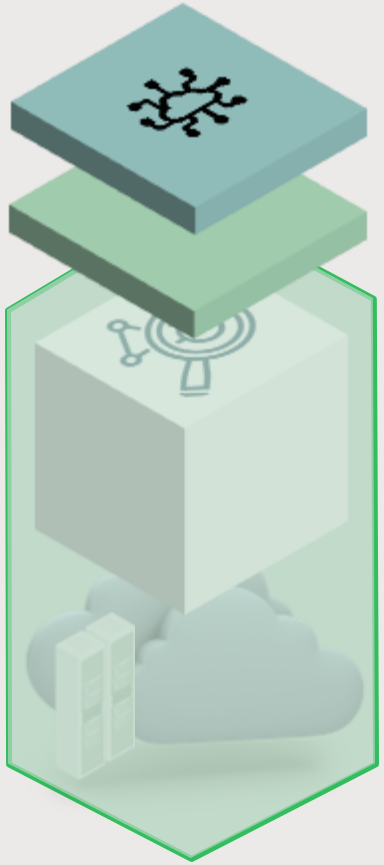
Skip the infrastructure management – focus on data science

Data scientists don't like to "mess" with cloud computing, they want to spend their time doing pure ML

OCI GPU + VM/OKE

- ML model
- Conda environment
- Container
- CUDA driver
- Linux Image
- NVIDIA GPU
- OCI Networking

OCI GPU + OCI Data Science



- ML model
- Conda environment / conatiner
- Managed by OCI Data Science

Ref:

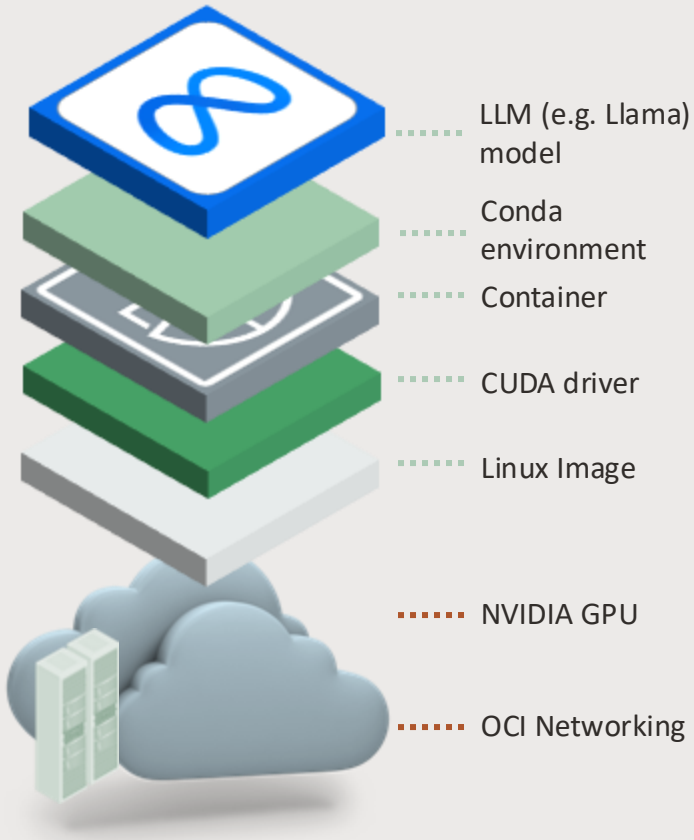
- https://docs.nvidia.com/datacenter/cloud-native/index.html
- https://www.oracle.com/cloud/compute/bare-metal/#rc30p7
- https://www.oracle.com/artificial-intelligence/data-science/

Confidential - Oracle Restricted

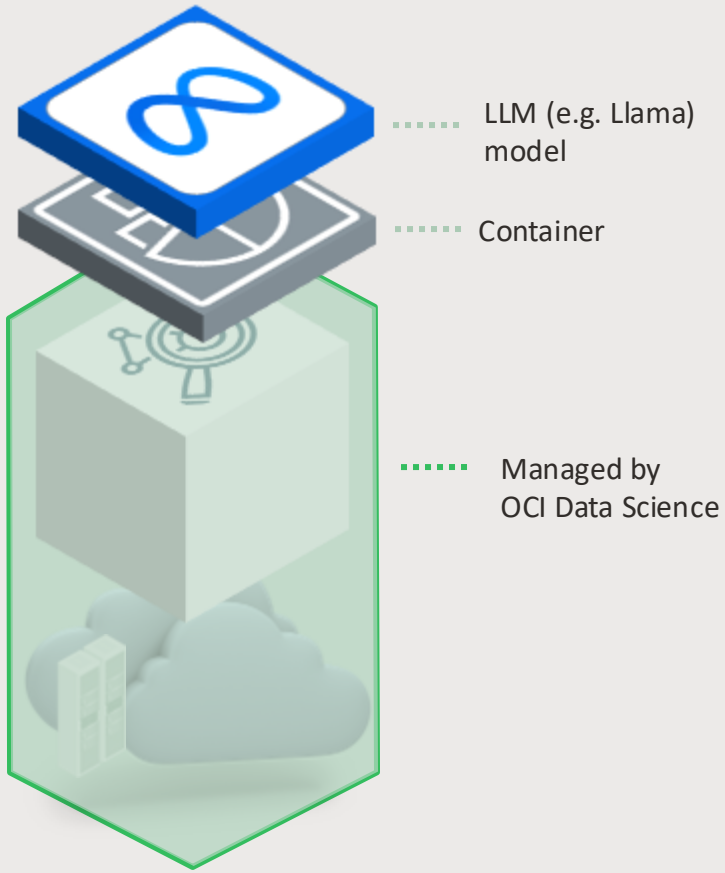
Custom Generative AI without coding – AI Quick Actions

With AI Quick Actions customers can fine tune and deploy LLMs from the UI

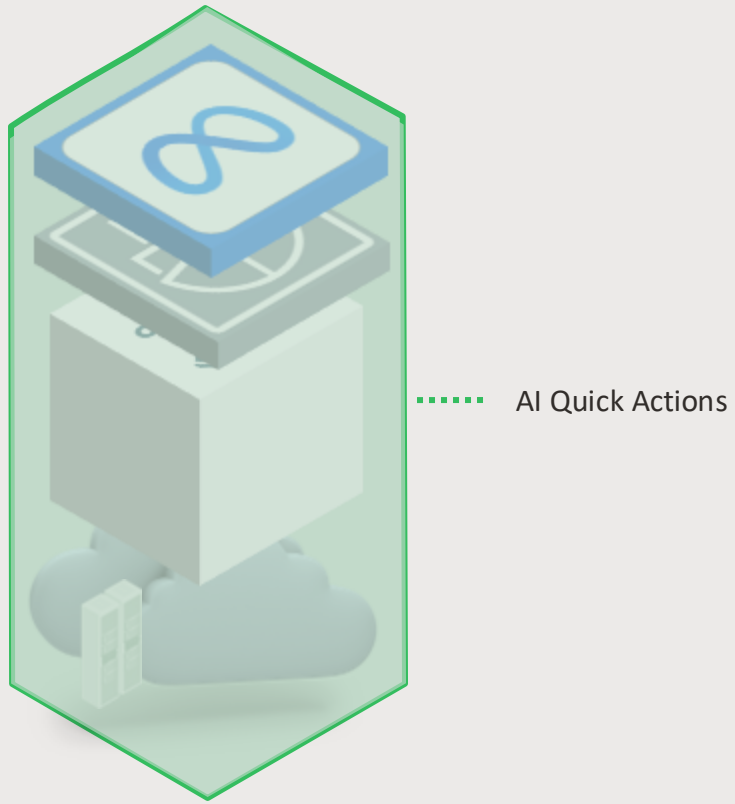
OCI GPU + VM/OKE



OCI GPU + OCI Data Science



OCI Data Science AI Quick Actions



Confidential - Oracle Restricted

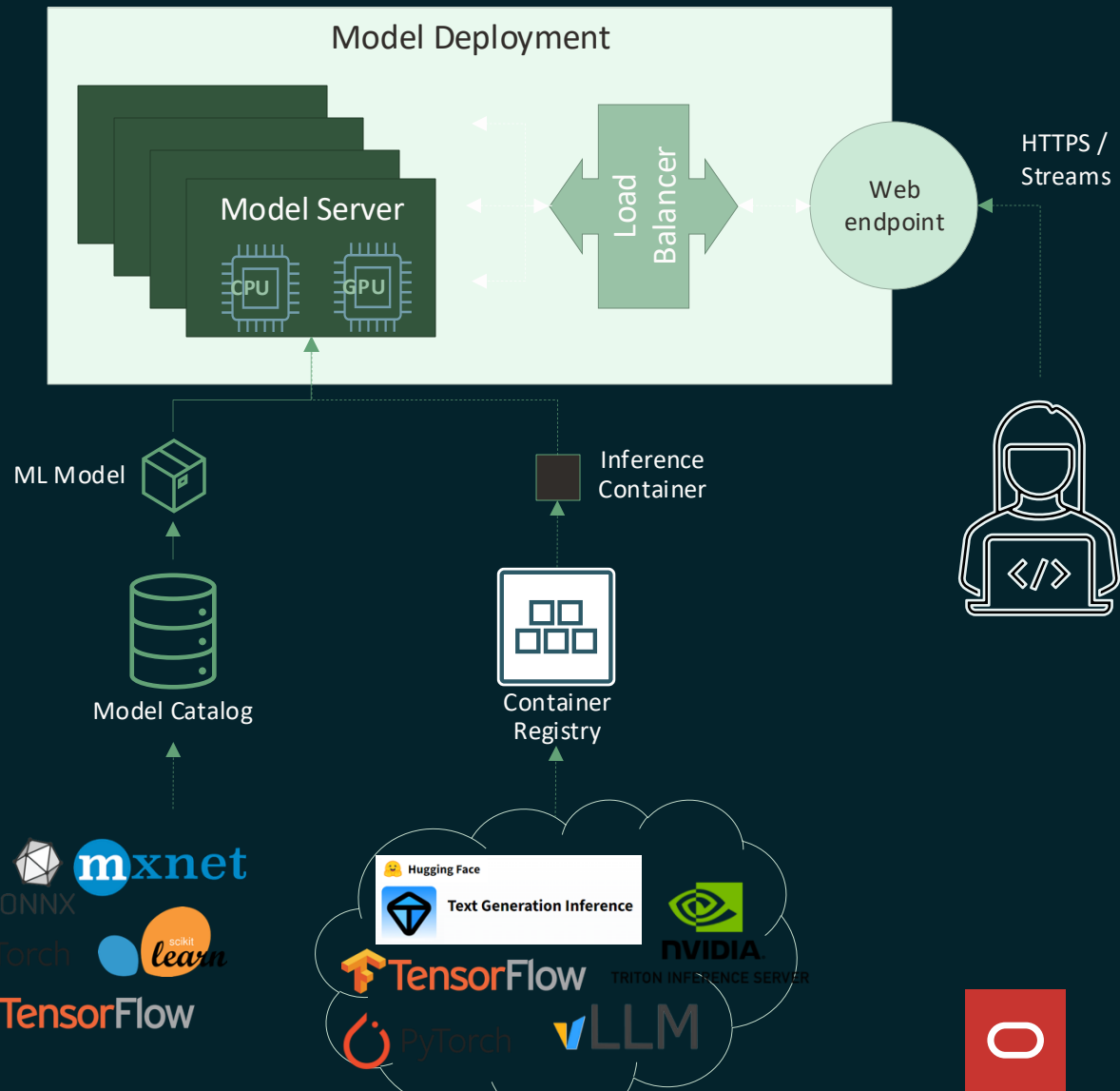
Example: Model Deployment – Flexible Inference at Scale

Full service-managed model inference

The service will automate the creation and management of all related infrastructure:

- Model Servers (VM/BM instances, OS Images, inference containers)
- Load Balancer, Storage, Networking
- Network Router and web endpoint

Additional support for: Autoscaling of instances, zero downtime on update, custom containers, and more



GenAI service vs. Data Science?

AI-as-a-Service (“canned solution”)

No coding required

No AI skills required

“Ready to run” (zero setup)

Bring any model/container

Customizable to your needs

Available in all regions (inc. non-commercial)

Service Highlights

Open-Source as first-class citizen

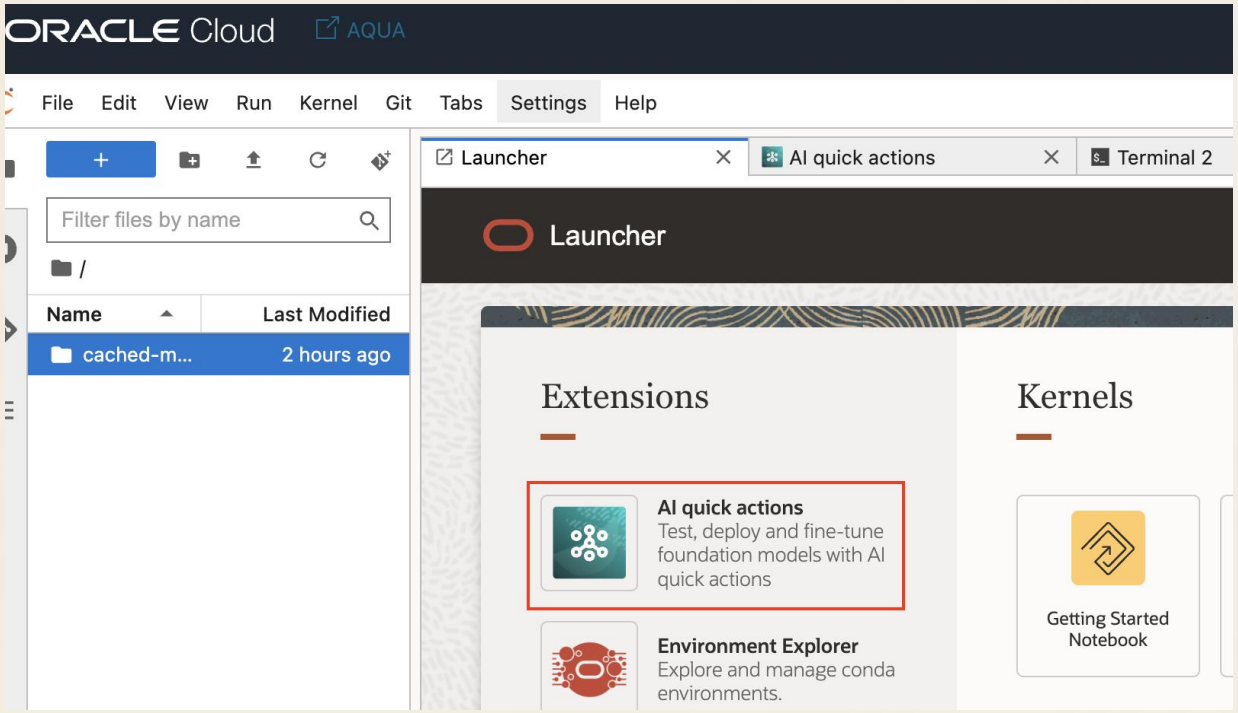
- Hosted **JupyterLab** notebooks
- Model development with open frameworks – **PyTorch, TensorFlow, SKLearn, XGBoost**, …
- Interoperable **ONNX** open model format
- Distributed training with **Horovod, Dask, PyTorch Distributed, TensorFlow**
- Experiment tracking with **MLFlow** and **TensorBoard**
- Model Inference with **NVIDIA Triton Inference Server**
- Built in support for LLM fine-tuning and inference with **vLLM, TGI, llama.cpp**
- LLM workflow management with **LangChain** integration
- Bring any LLM with **Hugging Face** integration
- NEW: NVIDIA NIMs in OCI Marketplace – deployed on OCI Data Science. Opportunity for partners to list their commercial models in OCI Marketplace.

AI Quick Actions

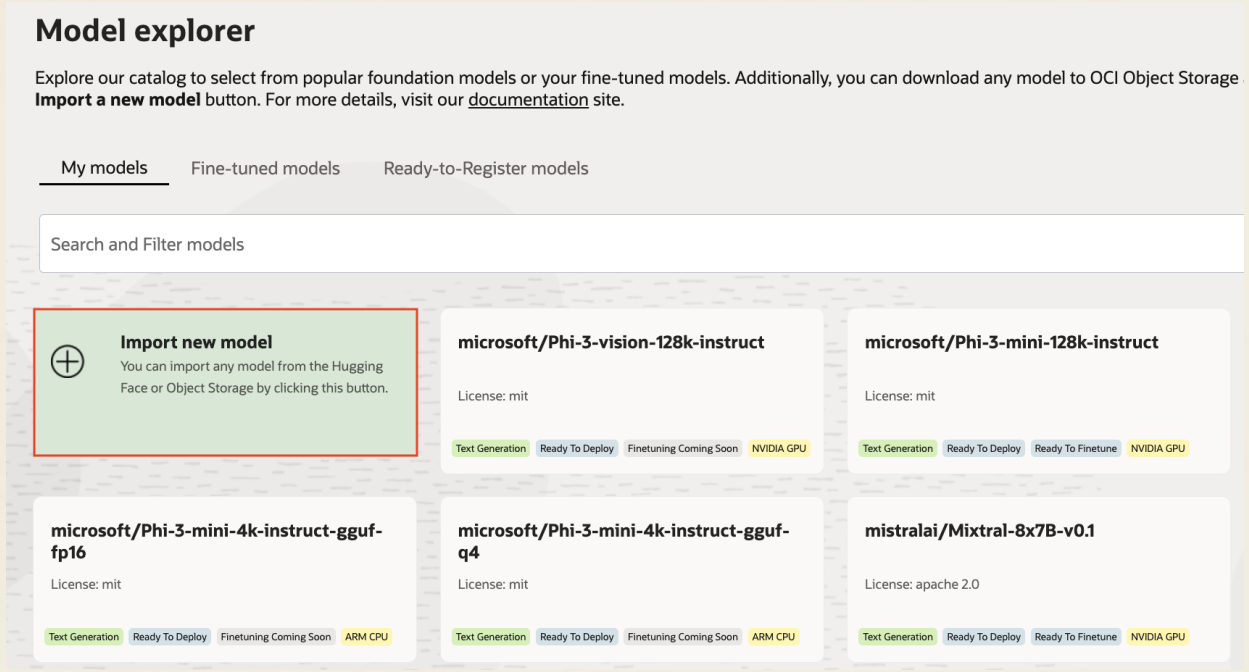
AI Quick Actions

No-code solution to fine-tune, deploy, and evaluate LLMs

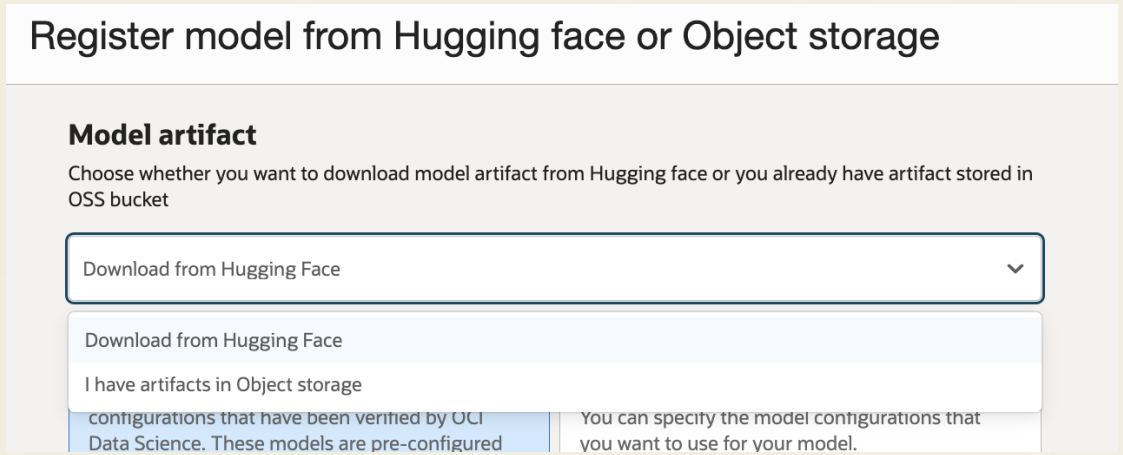
- Explore a curated list of popular Foundation Models
- Fine-tune on your data, in your tenancy, stored in your repository
- Deploy to real time web endpoint with a few clicks
- Test the model in the model's playground
- Generate evaluation reports to compare model quality
- Easily build applications on top of deployed LLMs
- Bring Any Hugging Face LLM



Explore a provided list of popular Foundation Models or Bring Your Own Model

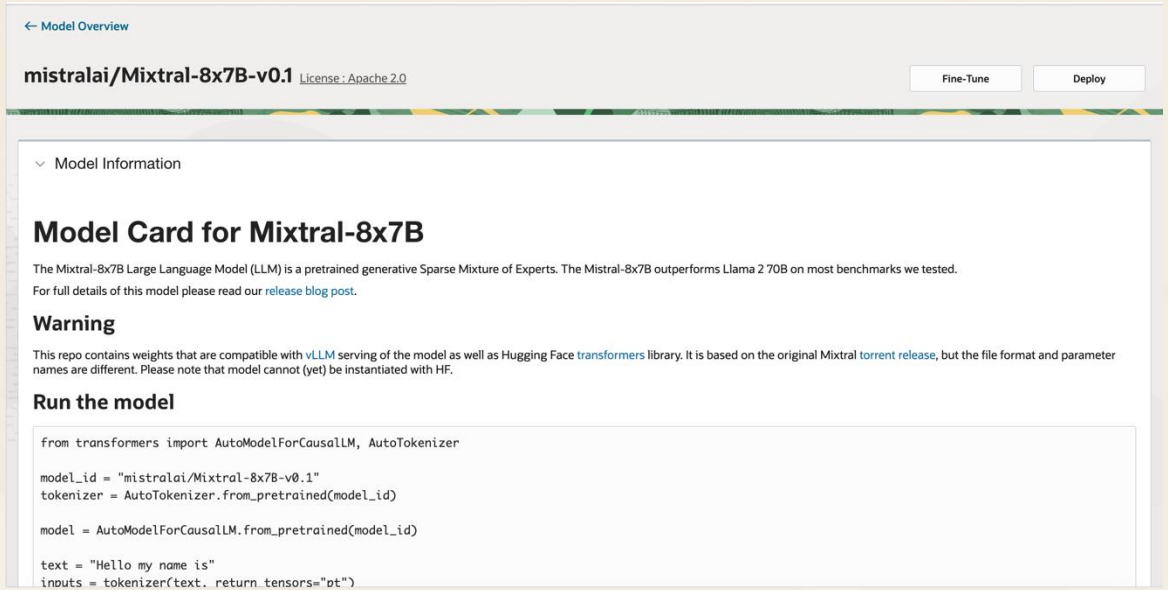


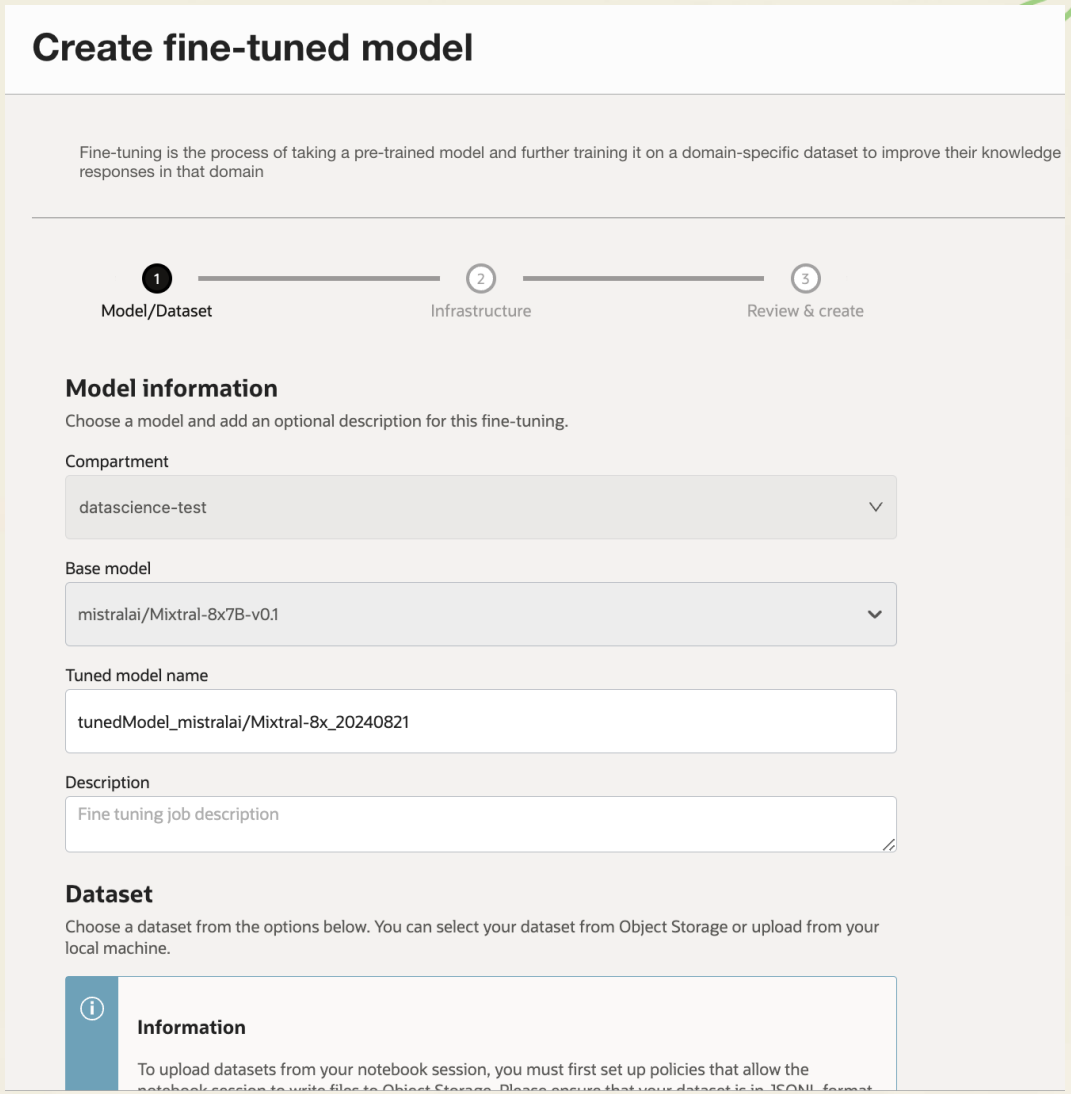
- Service provided models include Phi-3, Falcon, Mixtral and CodeLlama; text generation as well as multi-modal models
- Bring Your Own Model from Hugging Face or OCI Object Storage by registering the models



Fine tune

- Fine-tune on your data, the process runs in your tenancy
- The fine-tuned model is saved in your model repository

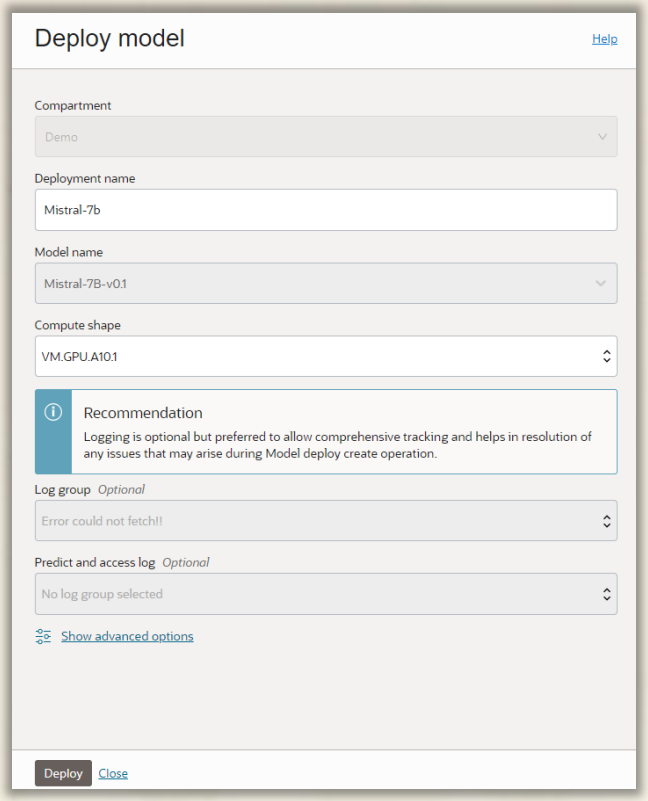


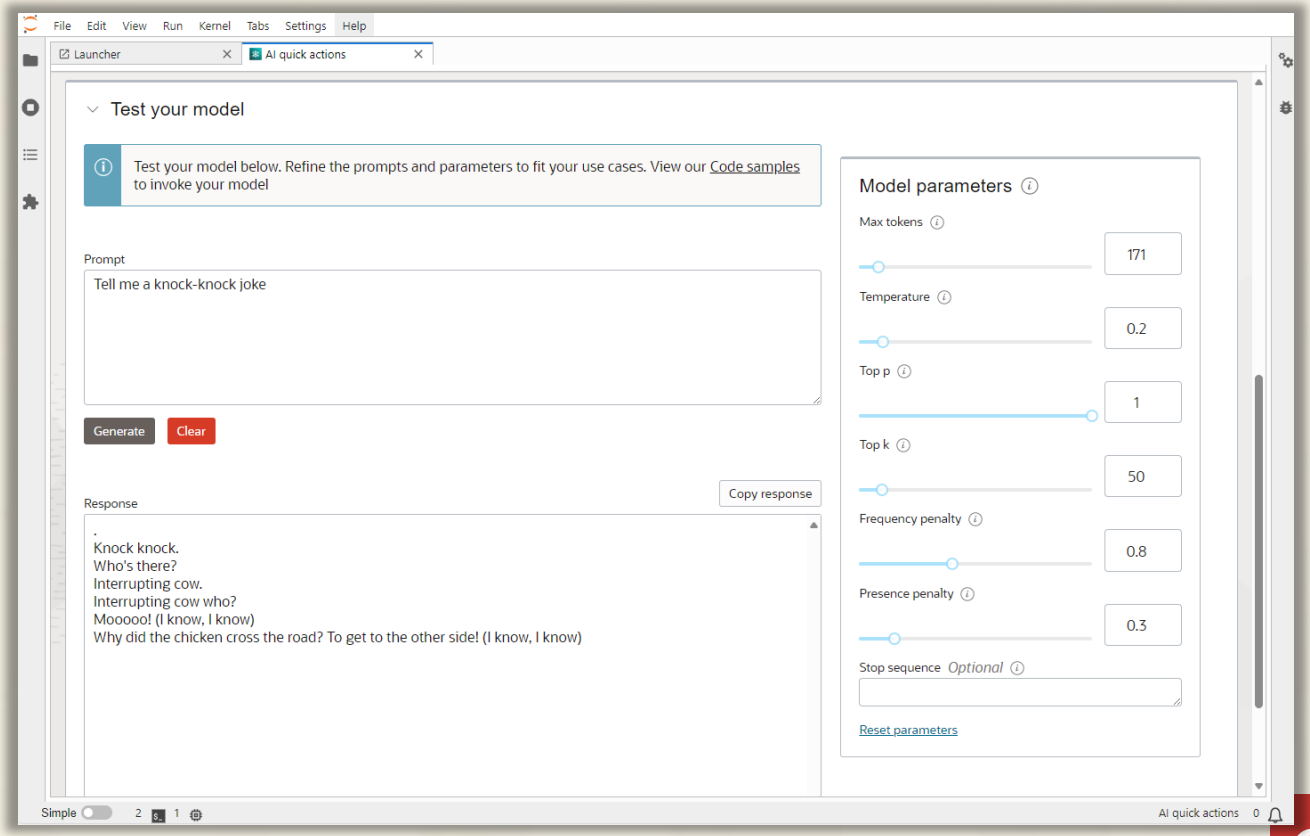


Confidential - Oracle Restricted

Deploy and test

- Deploy LLMs to any scale, using specialized inference servers like TGI (Hugging Face), vLLM and llama.cpp (for models in GGUF format)
- Test the model in real time after deployment
- Integration with Langchain

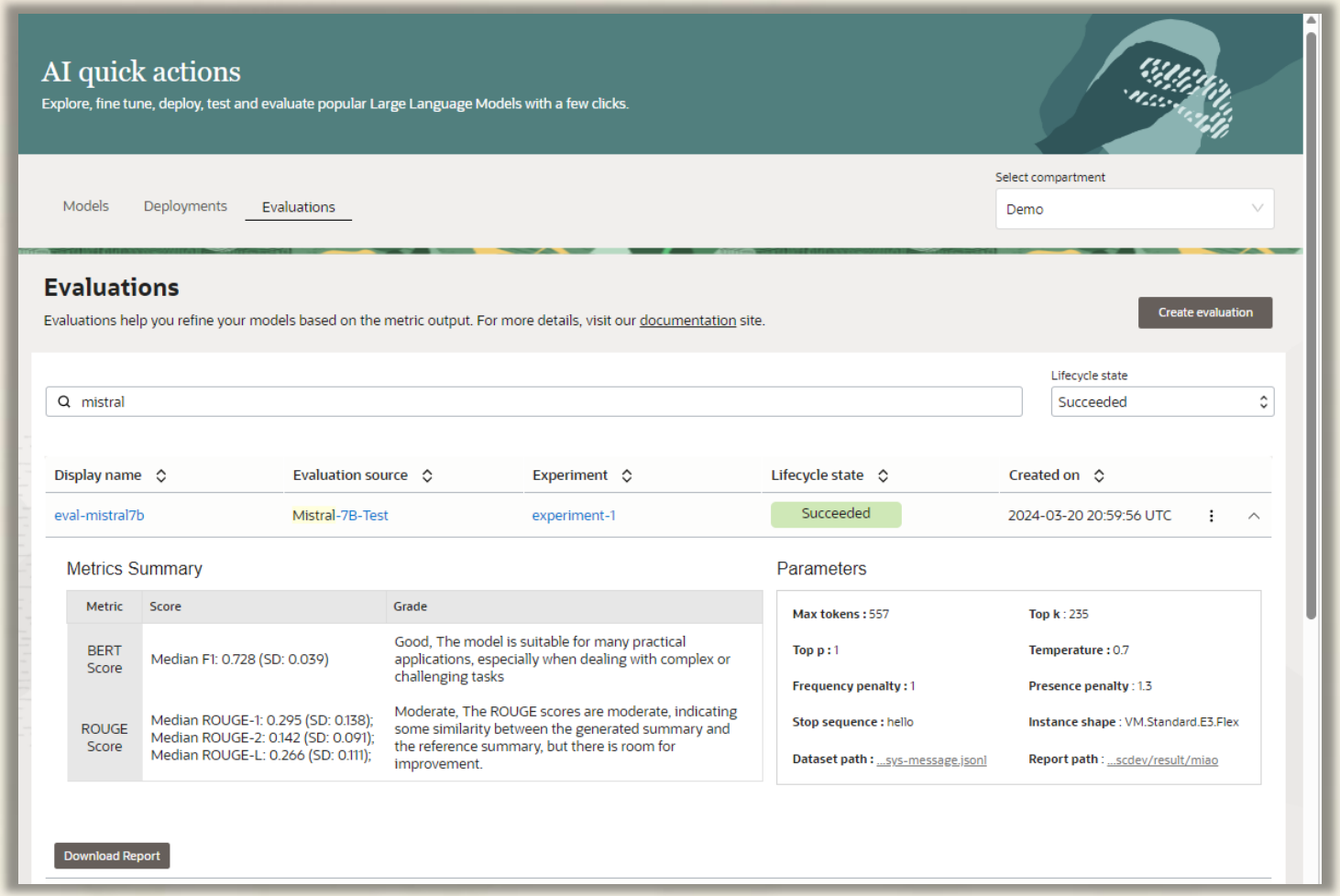




Confidential - Oracle Restricted

Evaluation

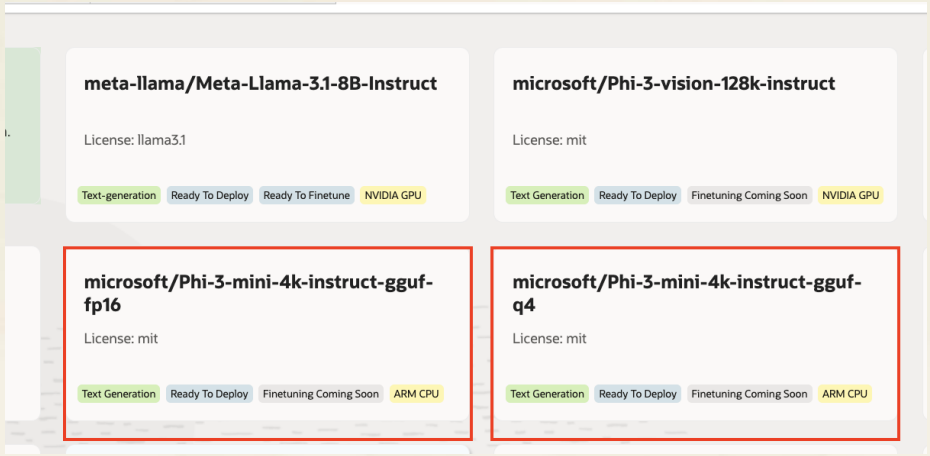
Compare models with detailed performance reports, using BERTScore, ROUGE, and others

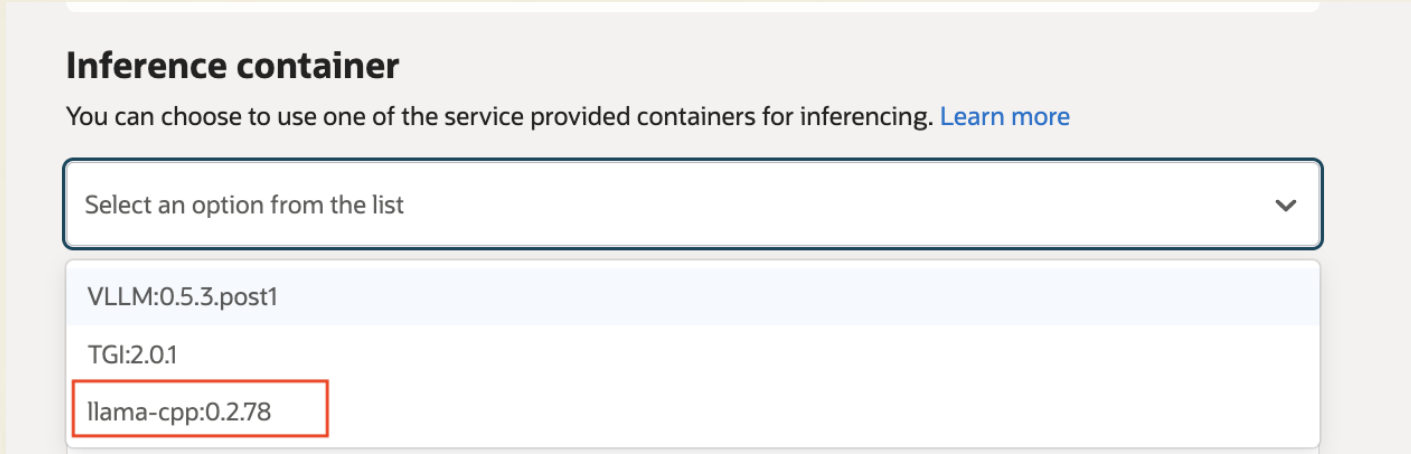


Confidential - Oracle Restricted

LLM inferencing with Ampere A1

- LLM inferencing with Ampere A1 shapes for models in GGUF format
- Service managed container with llama.cpp





Confidential - Oracle Restricted

Demo and informational videos

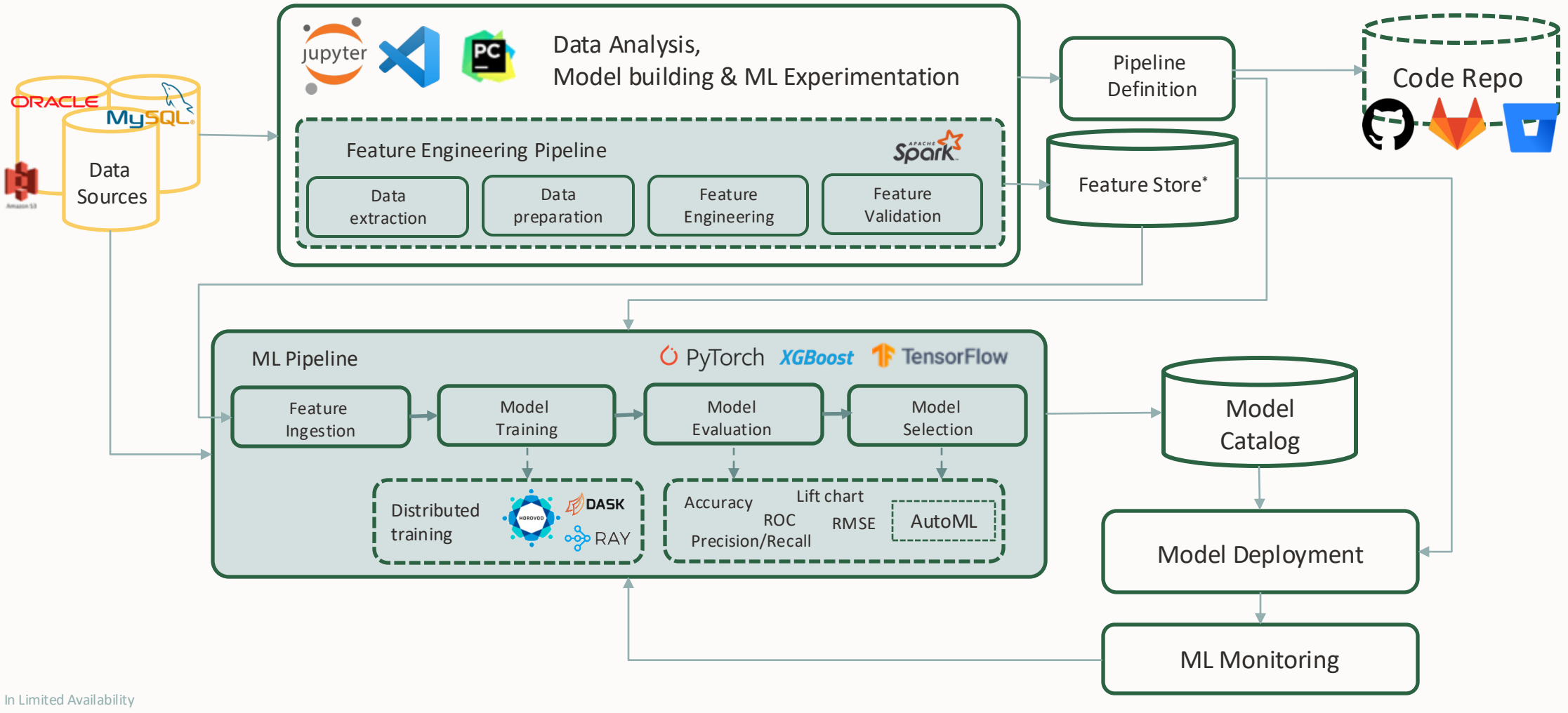
OCI Data Science quick overview

https://www.youtube.com/watch?v=H_omgbX5UUw

Day One and Beyond: Discover the Power of AI with Oracle's Data Science and AI Quick Actions

https://www.youtube.com/watch?v=7gp0lm3MMS0

Data Science Workflow (Experimentation & MLOps)



* In Limited Availability

Confidential - Oracle Restricted

Data Access

Configurable networking and built-in Python connectors make data access flexible and easy

- Data Source-Agnostic
  - Oracle Cloud, other clouds, on-premises
- Data Format-Agnostic
  - Structured, unstructured, semi-structured
  - CSV, TSV, Parquet, libsvm, json, Excel, HDF5, SQL, xml, apache server log files (clf, log), arff, etc.

Data Processing at scale

Run serverless Spark with **OCI Data Flow** from OCI Data Science notebooks.

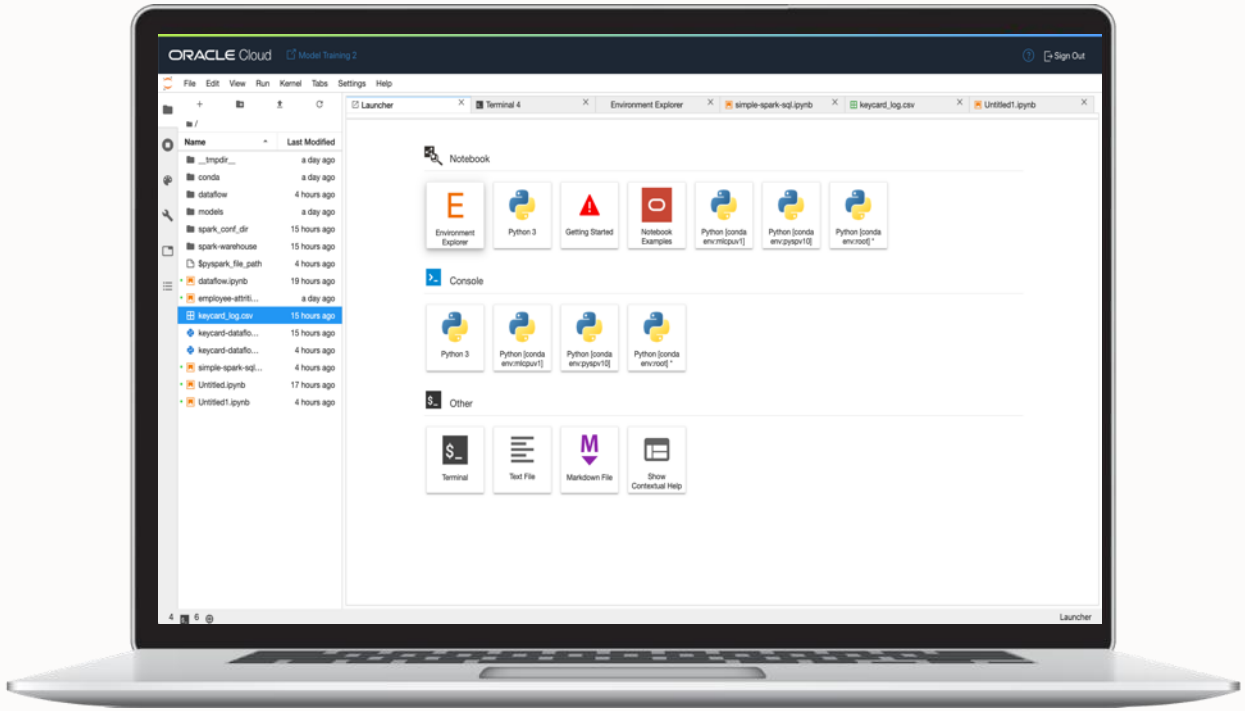
- **In Batch Mode:** Develop and iterate on a Spark job in a notebook, then execute at scale in Data Flow using the same environment
- **In Interactive Mode:** Seamlessly submit large-scale interactive Spark workloads from Data Science to fully-managed serverless Spark clusters with Data Flow's Livy integration, processing up to petabytes of data for large-scale data preparation and model training.

Development and Experimentation

Notebook Sessions

JupyterLab Interface for Building and Training your models

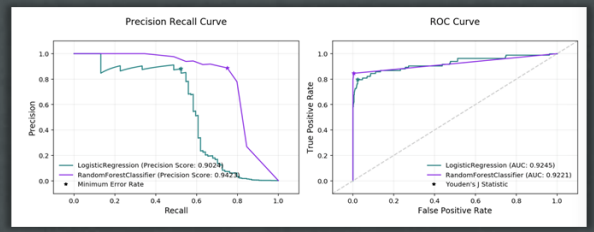
- Fully managed in the cloud
- Support for CPU & GPU shapes
- Persistent session storage for data, notebooks, and environments
- Easy to use out-of-the-box and custom Conda environments
- Git integration for remote version control
- OCI Vault integration for secret management
- Private networking support – Isolate your resource from the public internet
- Execute scripts during lifecycle changes

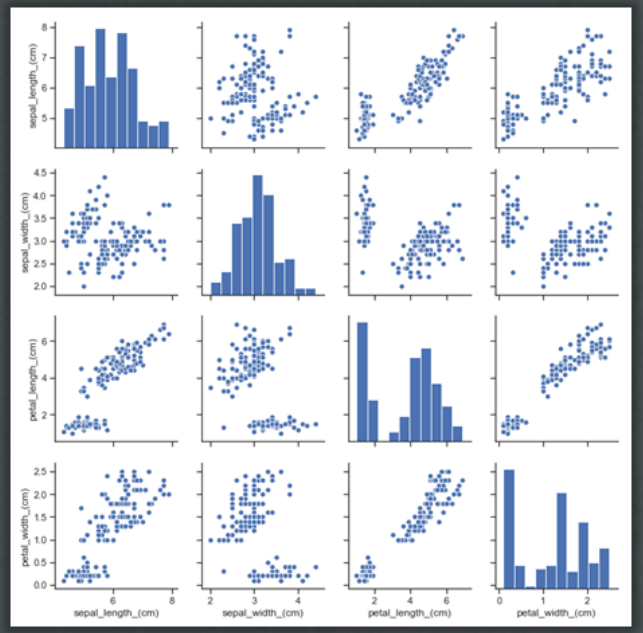


Oracle Accelerated Data Science (ADS) SDK

A Python toolkit to **increase the data scientist's productivity**, covering the end-to-end lifecycle of ML models from data acquisition to model evaluation, interpretation, and model deployment

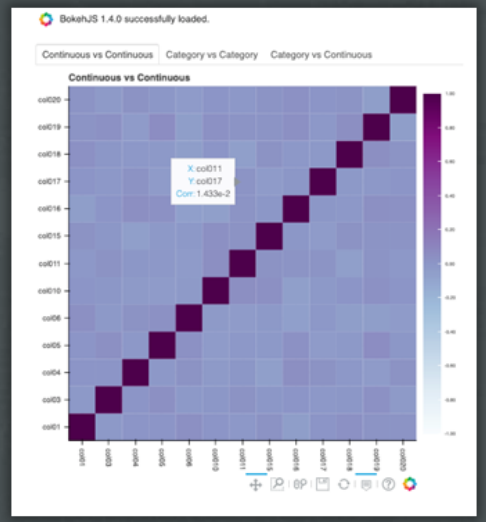
```
>>> import ads
>>> ads.hello()
```



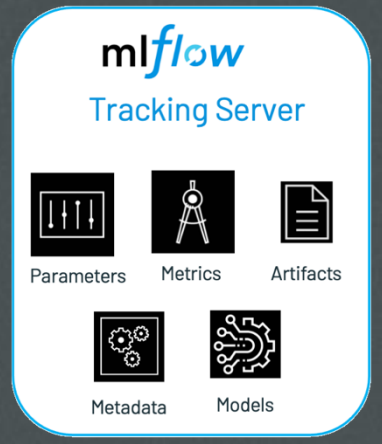


Feature highlights

Data connectors	Data Profiling	Create, Manage, Track Pipelines	Distributed Model Training
Auto tuning	Model Evaluation	Distributed ETL on Spark	Large Language Models



Experiment Tracking with mlflow

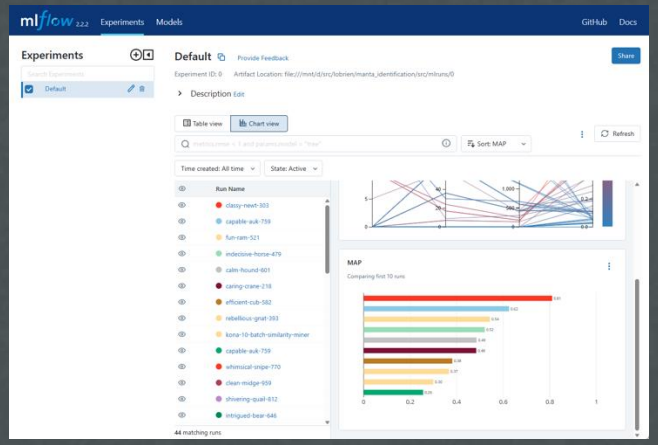


mlflow tracking server

Available as a docker image

Deploy on OCI Container Instance or Oracle Kubernetes Engine (OKE)

Support for Oracle MySQL database as tracking database



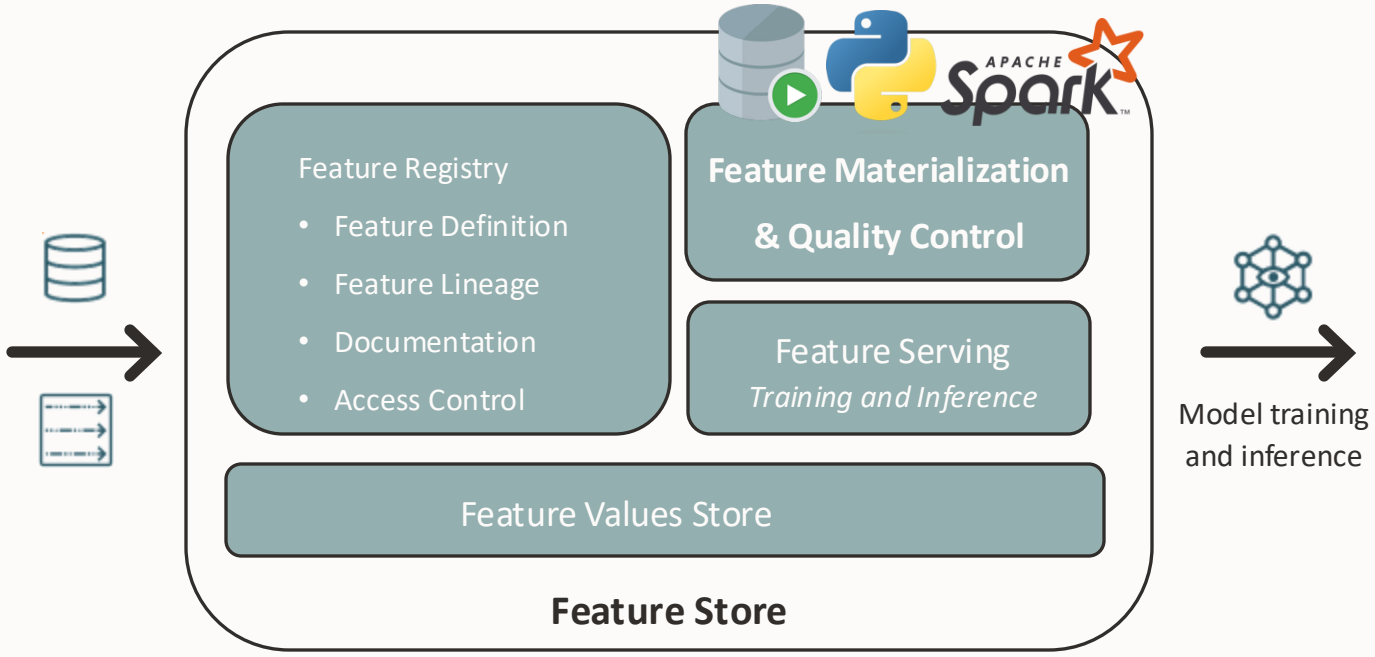
mlflow web UI

mlflow SDK & CLI

Support in Data Science Notebook session, Jobs, Pipelines

Available to download from PyPi

Feature Store (in Preview)



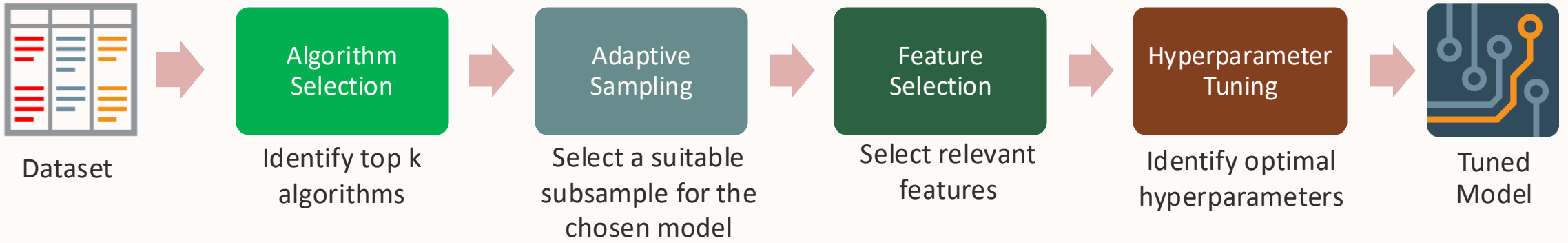
A solution focused on the lifecycle of features:

- Define feature engineering pipelines and build features with fully-managed execution
- Version and document features and feature pipelines
- Share, govern, and control access to features
- Consume features for both batch and real-time inference scenarios

Automated Model Training with AutoML

ADS offers Oracle's AutoML engine, developed over years of R&D in Oracle Labs

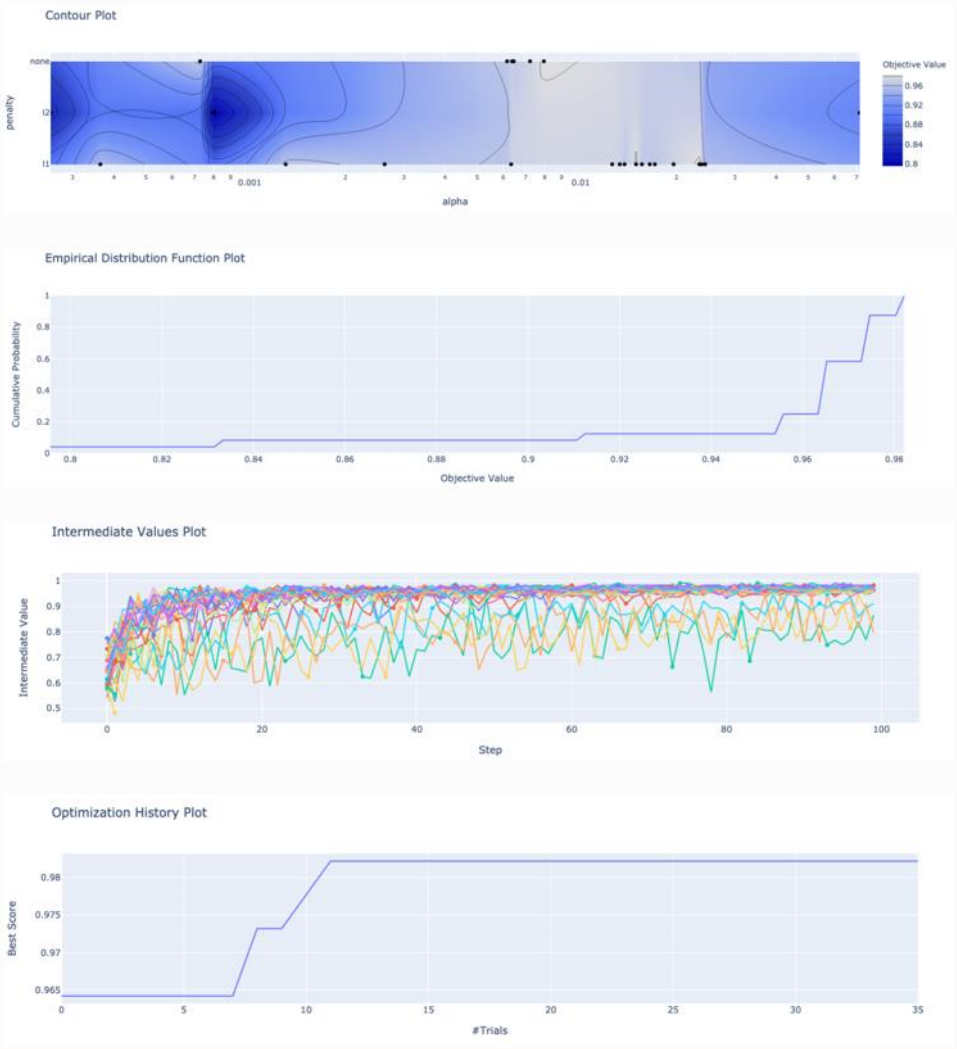
- Automated algorithm selection, data and feature selection, and hyperparameter tuning
- Optimized for data scientist expertise and time, runtime, and model performance



Automated Hyperparameter Tuning

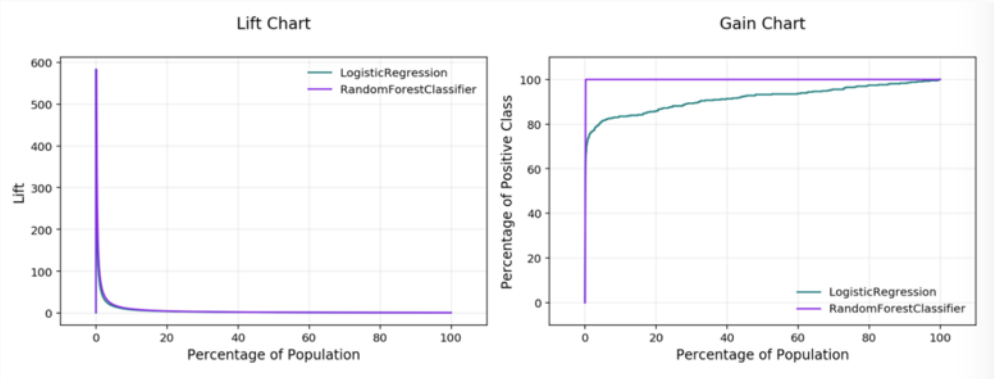
Automated Hyperparameter Tuner in ADS

- In addition to the other services for training models, ADS now includes a hyperparameter tuning framework called ADSTuner.
- Supports several hyperparameter search strategies out of the box, as well as user-defined search spaces and strategies
- A valuable add on to ML libraries which do not include hyperparameter tuning

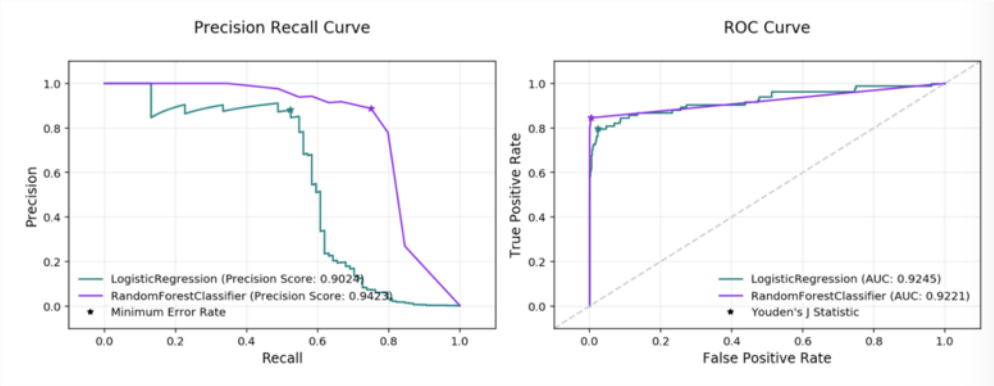


Model Validation

ADS Evaluator helps data scientists understand their models' accuracy and performance



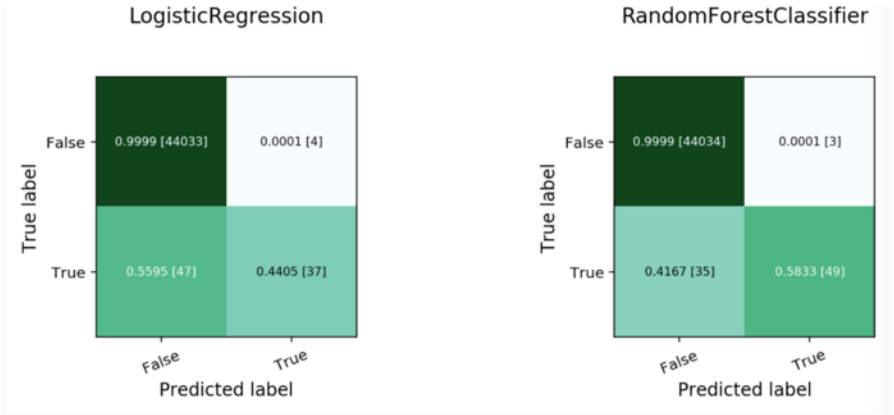
Lift & Gain Chart



PR & ROC Curves

Evaluation Metrics (testing data):

	LogisticRegression	RandomForestClassifier
accuracy	0.9988	0.9991
hamming_loss	0.001156	0.0008839
kappa_score_	0.5915	0.7268
precision	0.9024	0.8814
recall	0.4405	0.619
f1	0.592	0.7273
auc	0.9245	0.9042

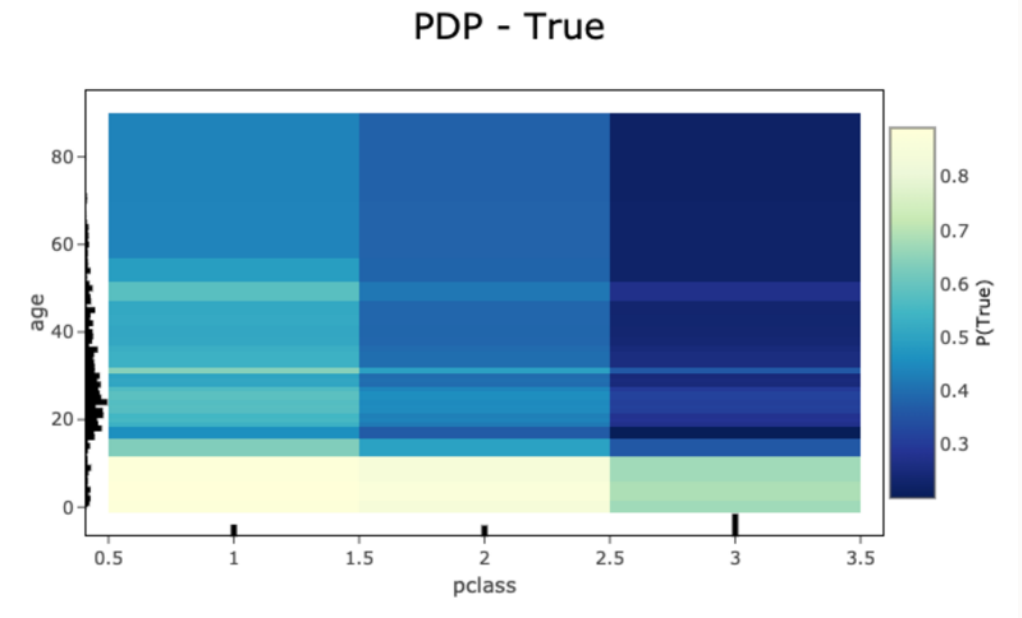


Normalized Confusion Matrix

Model Explanation

ADS offers Oracle's MLX for Model Explanation, developed over years of R&D in Oracle Labs

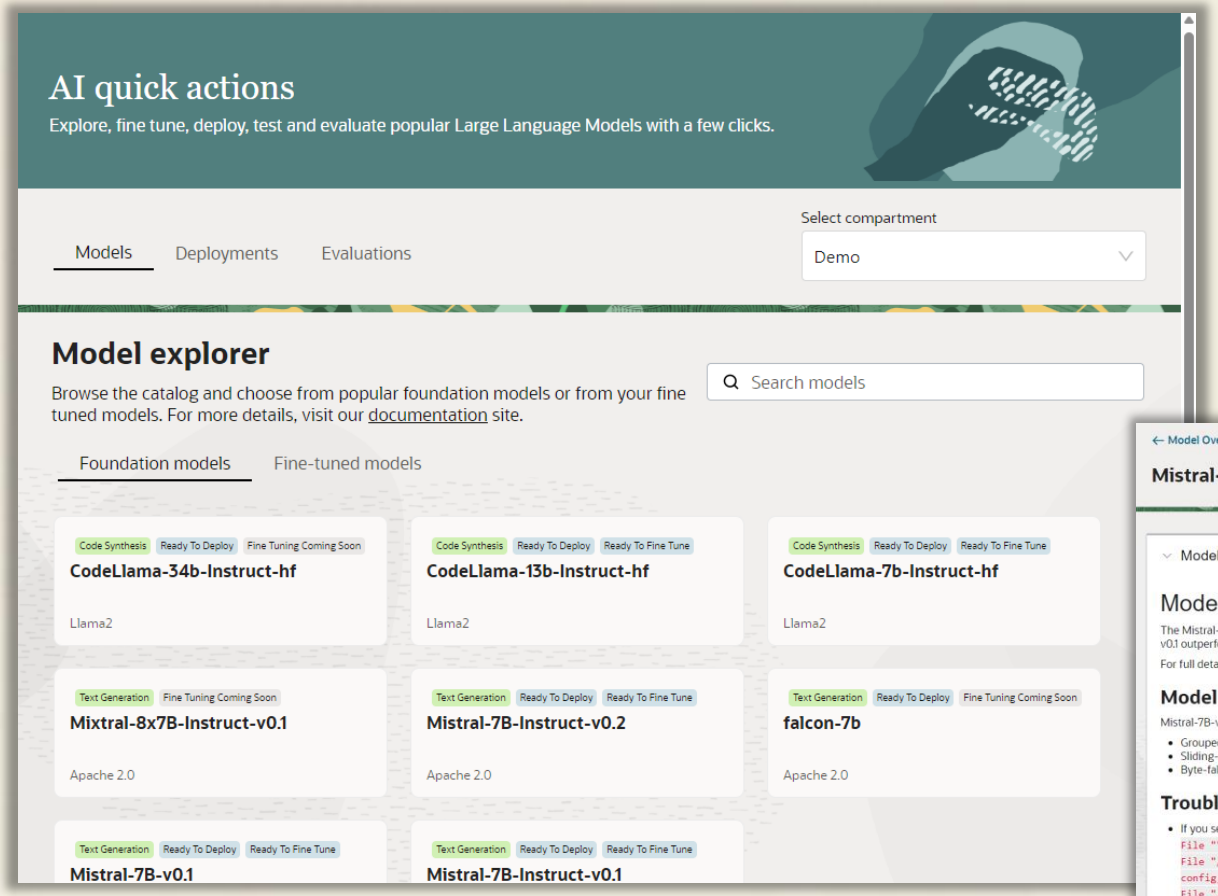
- Automated model-agnostic explanations improve understanding and trust, address regulatory needs, and increase speed of ML adoption
- Global explanations help explain the overall behavior of a model and local explanations explain specific model predictions

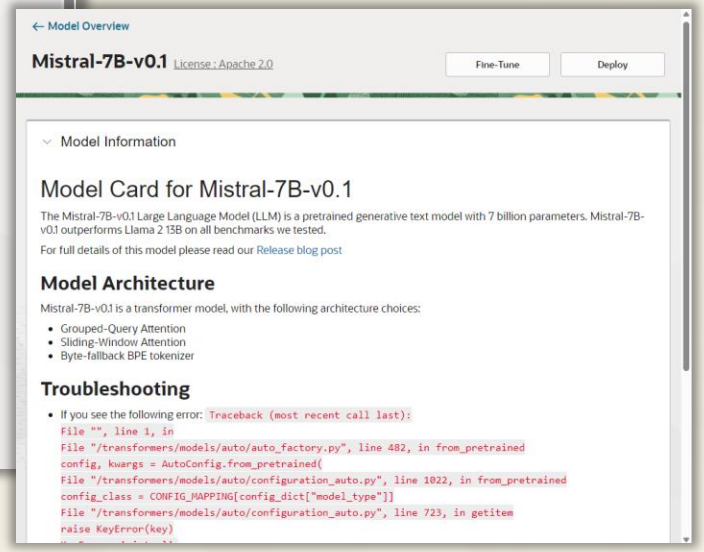


Generative AI

AI Quick Actions

No-code solution to fine-tune, deploy, and evaluate LLMs





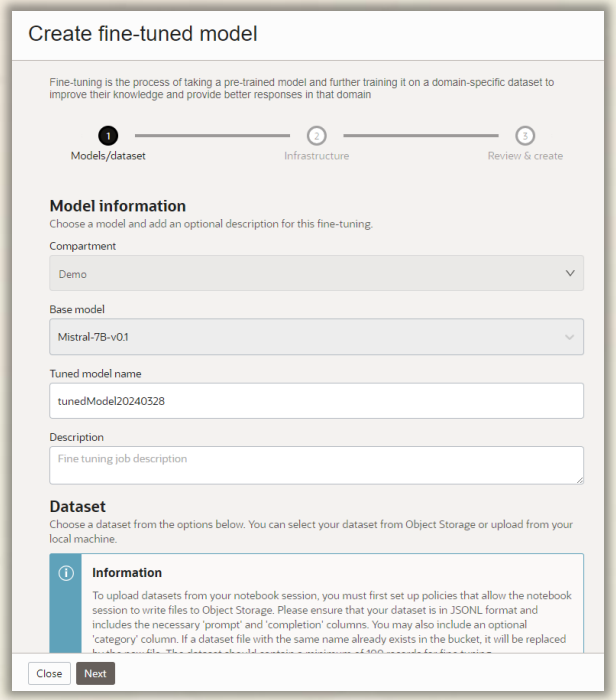
Explore a curated list of popular Foundation Models

AI Quick Actions

No-code solution to fine-tune, deploy, and evaluate LLMs

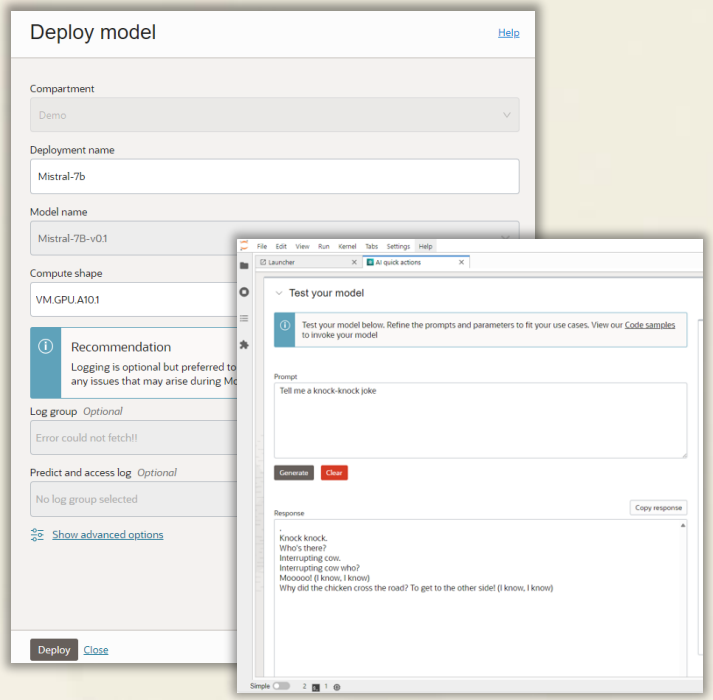
Fine Tune

On your data to specialize the model. Fine tuned model is saved in your model repository



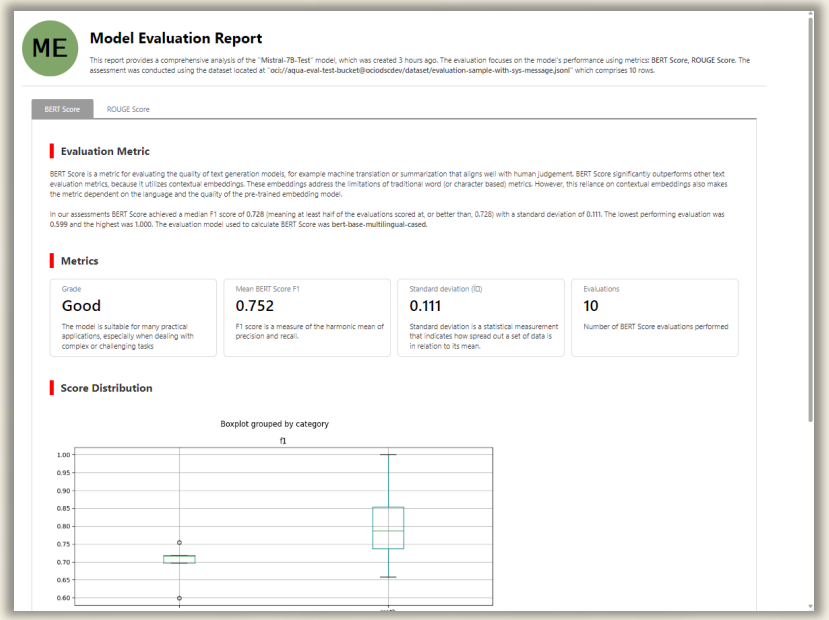
Deploy and test

To a real-time inference endpoint. Use the playground to interact with the model



Evaluate

Compare models with detailed performance reports, using BERTScore, ROUGE, and others



Confidential - Oracle Restricted

Management, Shareability, Reproducibility

Data Science is a team work

Projects

Collaborative workspace for teams of data scientists

- Organize your work
- All resources are created within Projects. Data scientists can create, name, and describe their projects.
- Leverage granular access control

Conda Environments

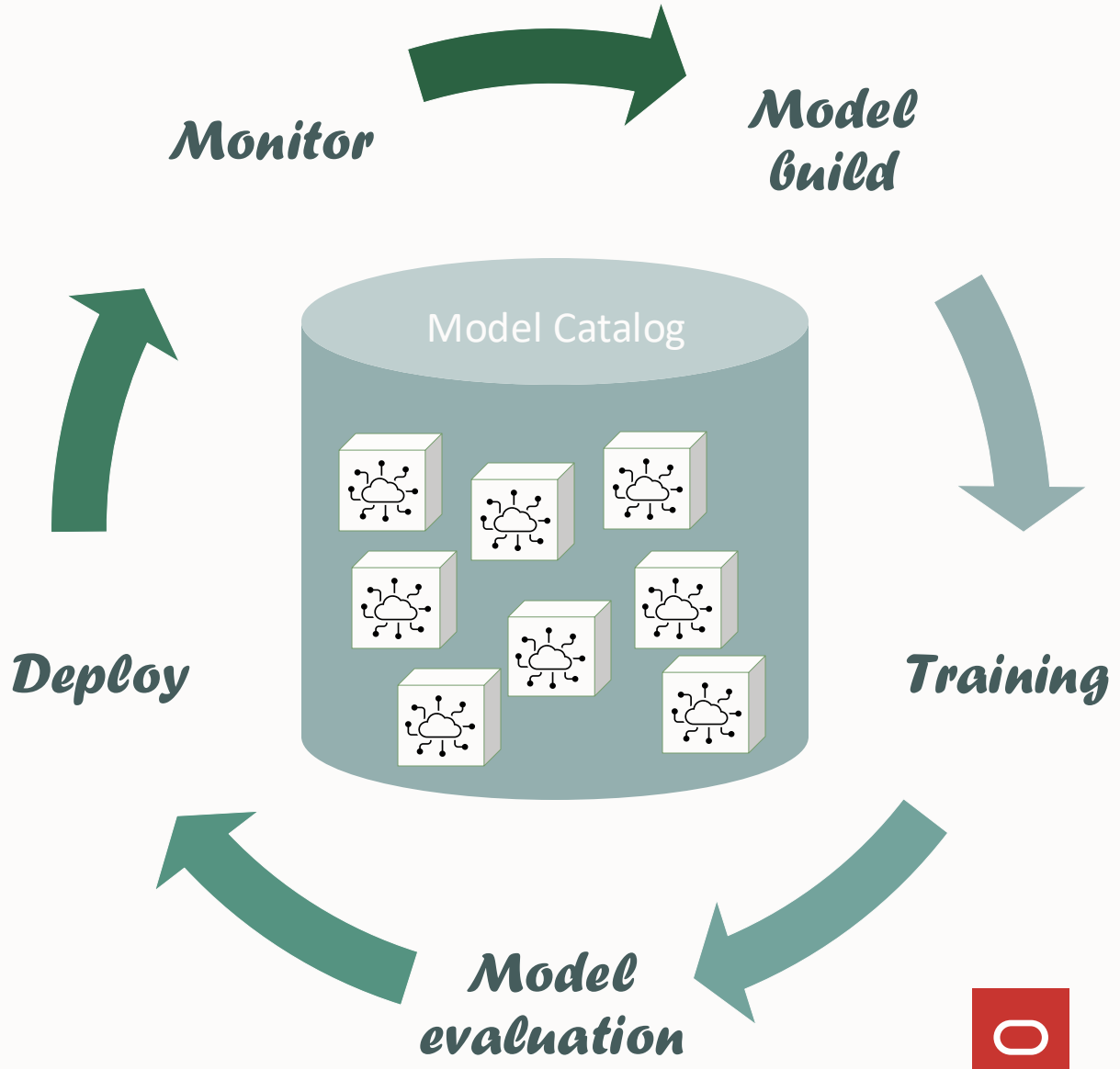
Dependency management for reproducibility

- Pre-built, curated conda environments, addressing a variety of use cases and tools like NLP, Graph Analytics, Spark, etc.
- Publish your own custom environment and share with colleagues
- Ensure reproducibility for training and inferencing

Model Management Through the Model Catalog

The Model Catalog fosters collaboration and ensures model auditability and reproducibility amongst the data science team

- Track model metadata:
  - Model Provenance
  - Model Taxonomy
  - Custom/user-defined Metadata
  - Input and Output Data schema
  - Model Introspection Results
- Version models
- Support up to 400GB in model artifact size
- Easily deploy models from the catalog

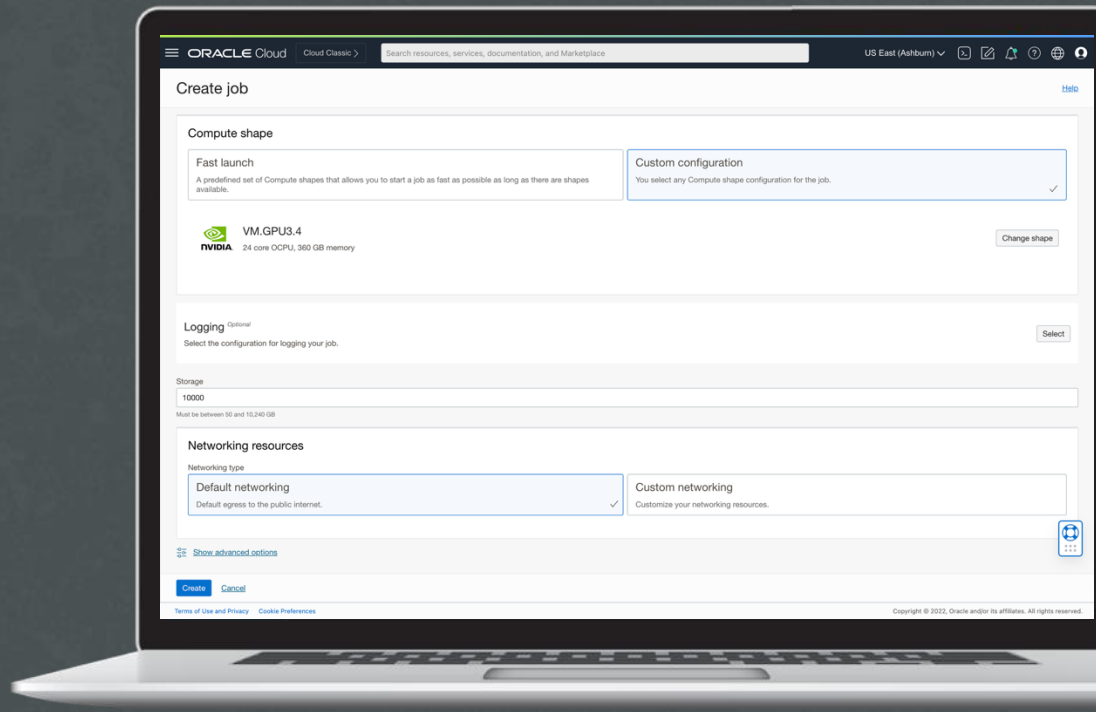


Operate at Scale with MLOps

Cloud-Scale execution with ML Jobs

Operationalize tasks and execute at scale in the cloud

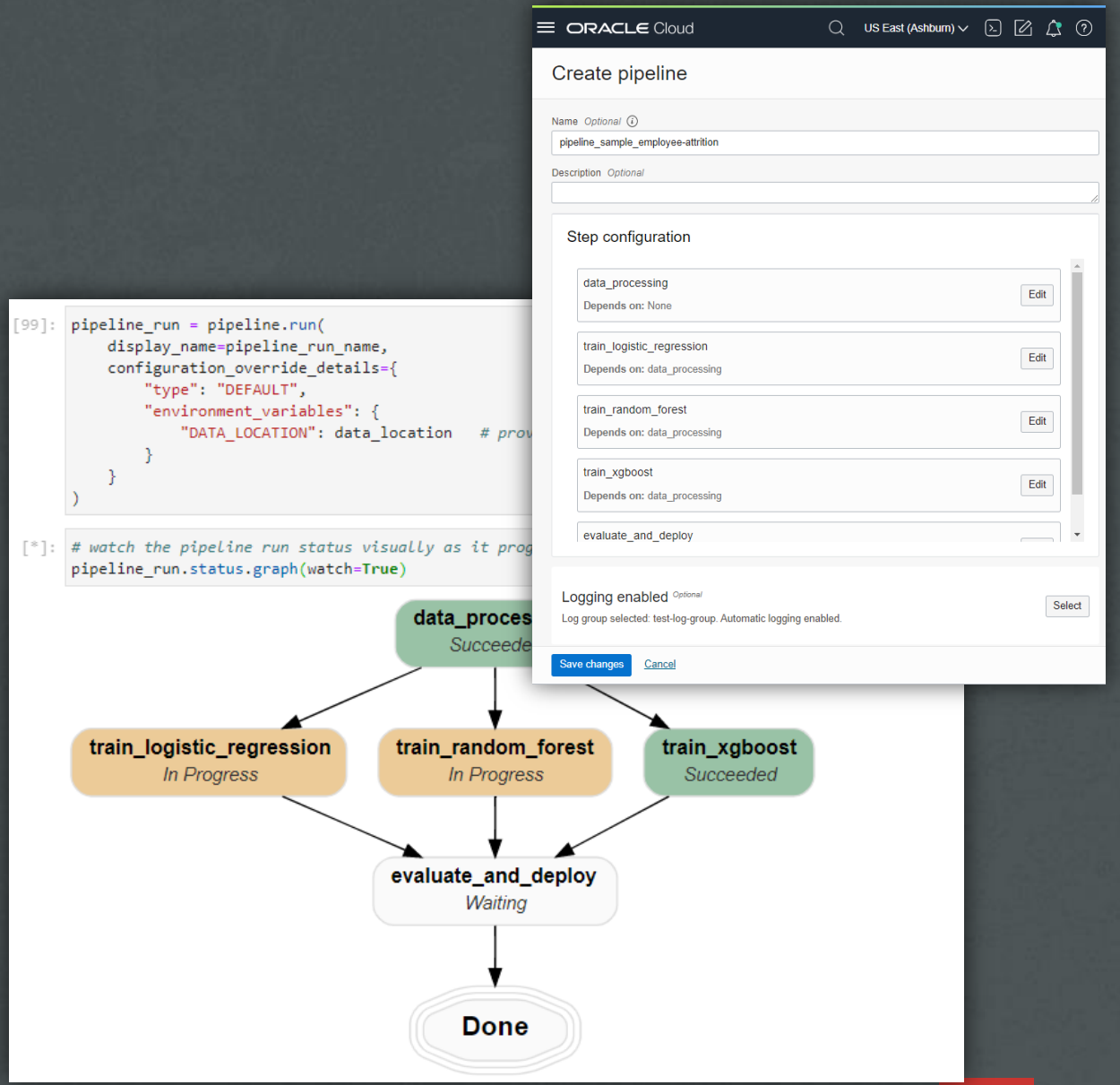
- Deploy large-scale, repeatable ML tasks like data processing, model training, and batch scoring
- Fully service-managed
- Support for CPU & GPU shapes
- Supports containers and Pyton/Java/Bash scripts
- Distributed, multi-node training (with Horovod, PyTorch Distributed, TensorFlow Distributed, Dask)
- Mount persistent storage (OCI File Storage, OCI Object Storage)



ML Pipelines

Operationalize and automate your model development, training, and deployment workflows with a fully-managed service to author, debug, track, manage, and execute ML pipelines.

- Create reusable tasks and use as steps in a pipeline – import and prep data, train, evaluate and deploy
- Steps can run sequentially or in parallel
- Scale infra with variety of supported VMs, including GPUs
- Create pipelines via code or configuration (YAML)



Flexible Inference at Scale

Model Deployment:

Full service managed model inference

Deploy models as managed web endpoints

Real-Time inference through HTTPS requests and data streams [In Preview]

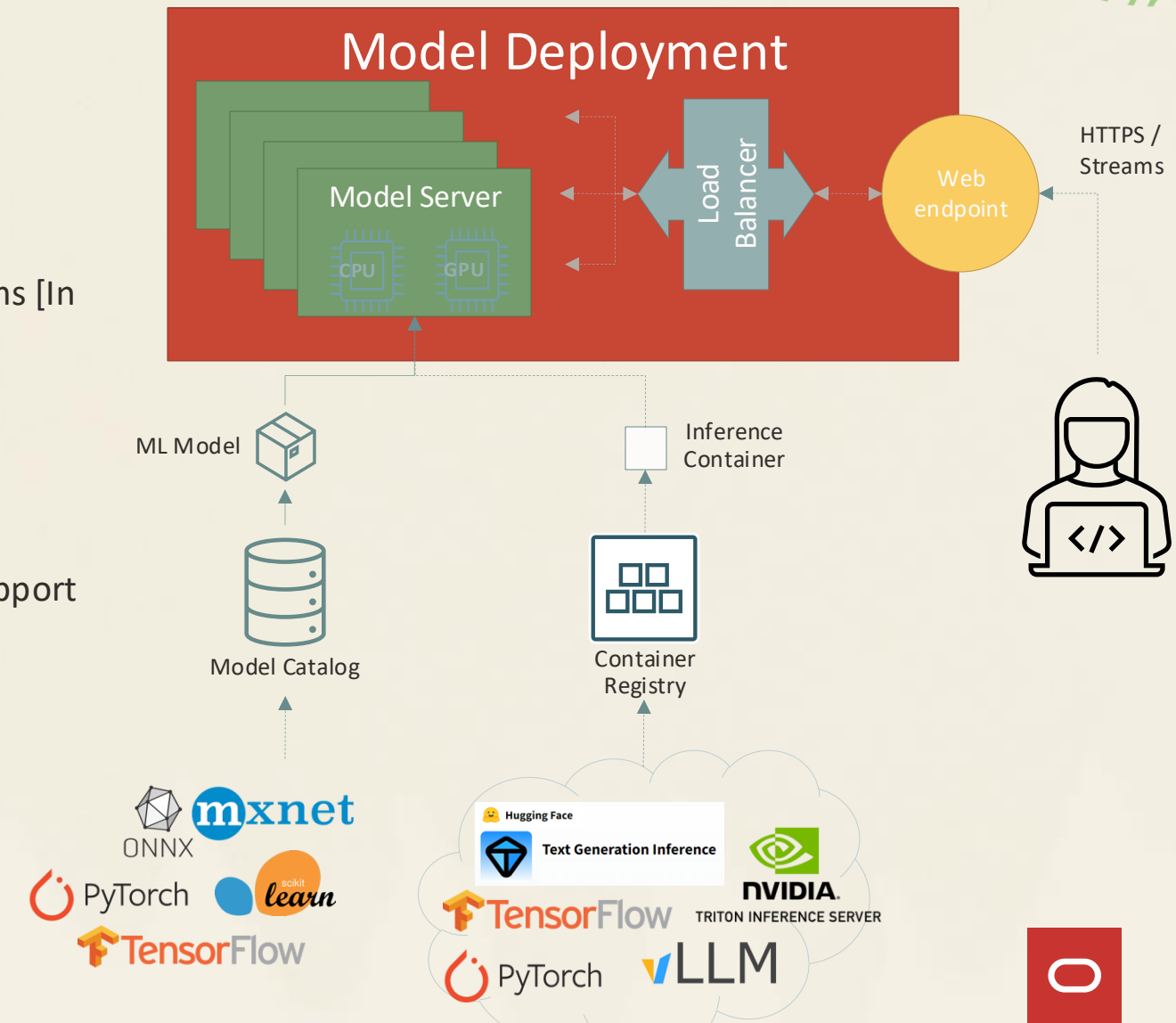
Zero management, Zero downtime

Deploy on CPUs or GPUs

Containers support – Deploy any inference server. Built in support for NVIDIA Triton Inference Server

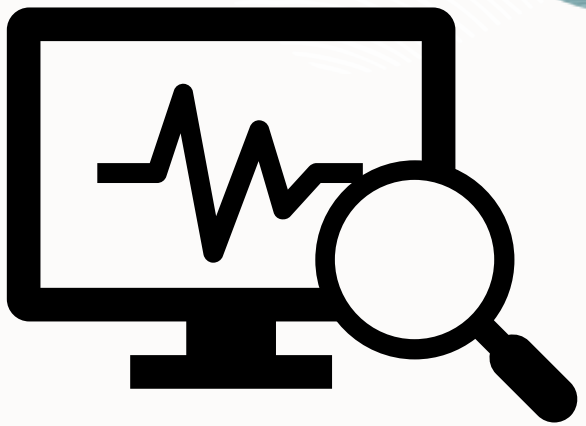
Autoscaling of instances

Bring your own networking (for public internet access or networking restrictions)

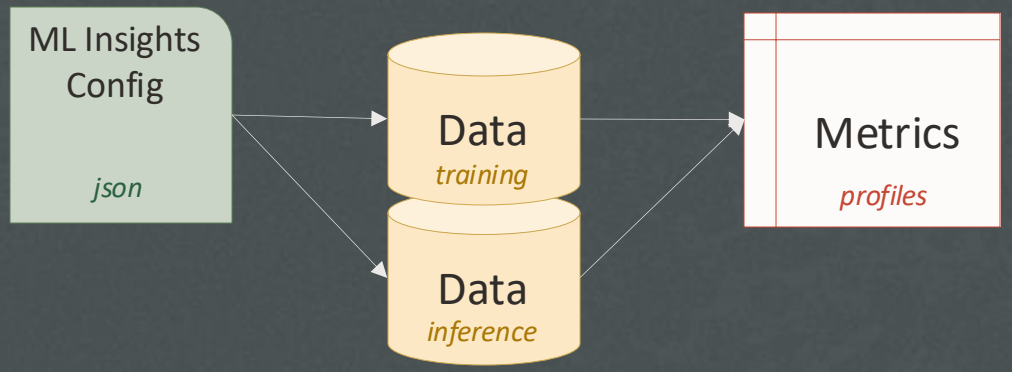


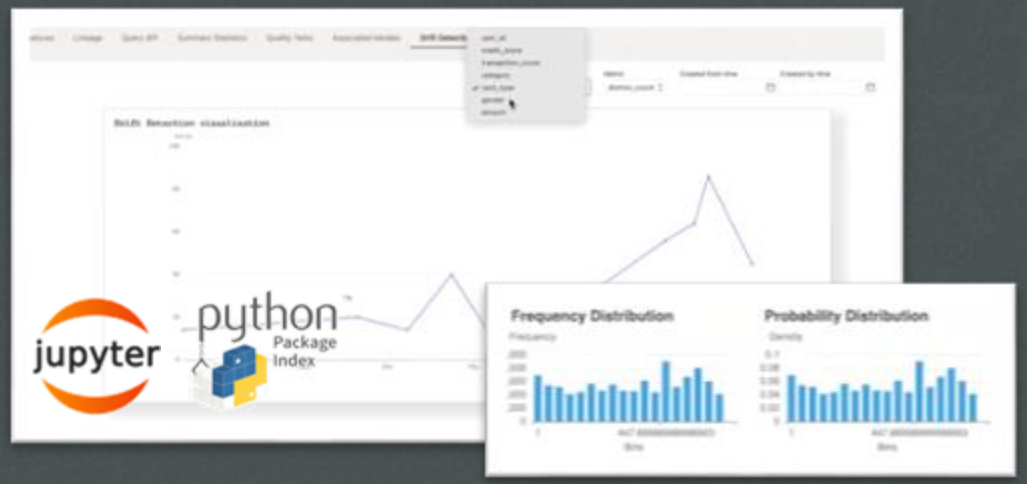
ML Monitoring

- Track model metrics through validation and production
- Ensure models remain healthy in production
- Monitor any model I/O data, including very large scale datasets
- Receive alerts when metrics cross predefined thresholds
- Take actions on alerts, such as retraining a model or updating a deployment



ML Insights Library





For Data Scientists who need to quickly evaluate their data to decide on their ML Monitoring use cases and set up long running monitoring process to continuously evaluate their models and data.

Configurable Metrics for monitoring:

- Data Integrity
- Data Quality/ Summary
- Feature and Prediction Drift Detection
- Model Performance for both classification and Regression Models

Extendable with:

- Custom Metrics
- Conditional Features & Transformers
- Data readers
- Post Processing
- Test/Test Suites

Thank you

oracle.com/data-science