Dezembro/2022

Data Masters – Case

Machine Learning Egineer

Treinamento e Deploy de Modelos de Deep Learning utilizando Databricks + Spark + TensorFlow + MLFlow em ambiente *Cross Cloud* – Microsoft Azure e Google Cloud Plataform

github: https://github.com/lisboavini/data_masters/ *

*O relatório juntamente com todos os códigos fontes e definições de ambiente encontram-se disponíveis dentro do repositório do GitHub.

Vinícius Lisboa

Pauta:

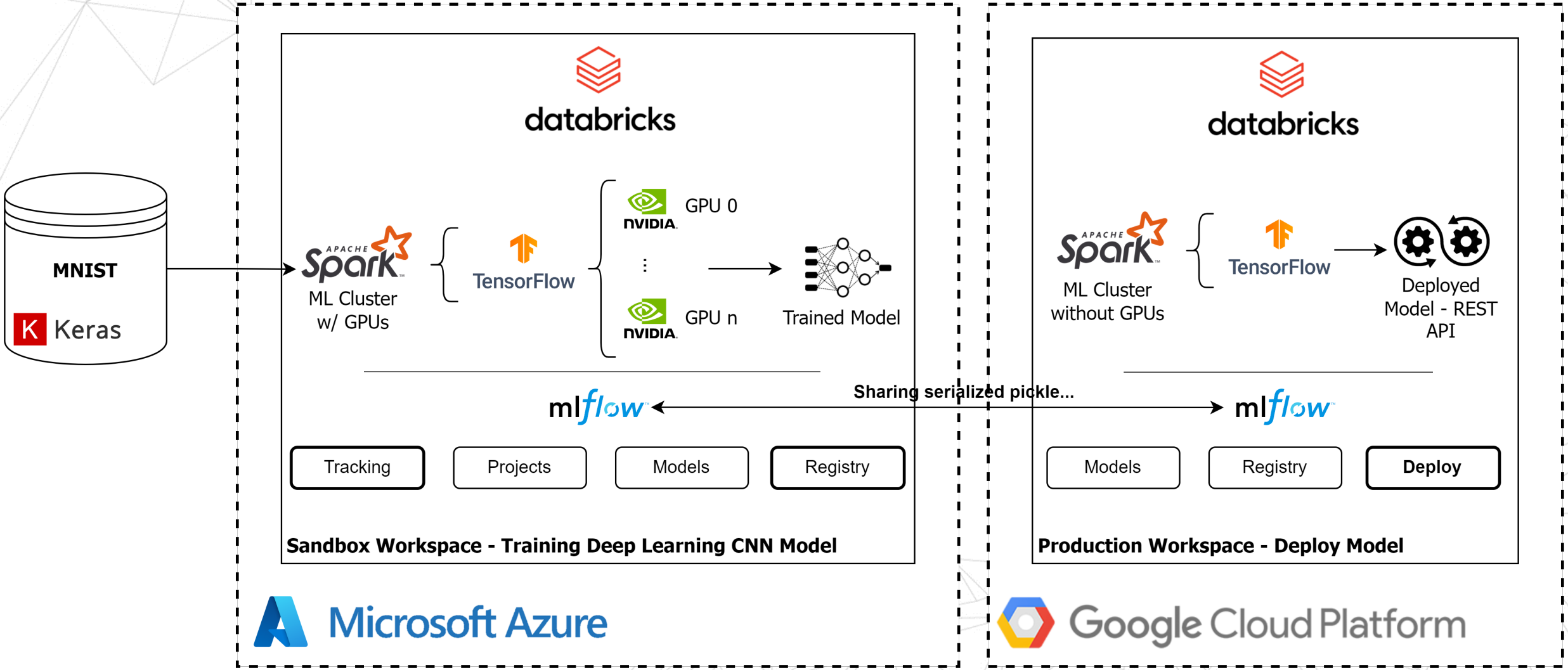
1. **Motivação**
   Desenho de solução completa cross-cloud
2. **Arquitetura de Solução – *Blue Prints***
   Desenho de solução completa cross-cloud
3. **Arquitetura Técnica/Dados**
   Proposta com arquitetura full cloud utilizando nossas plataformas de dados
4. ***Pipeline* de Treinamento Azure**
   Preparação e definições do ambiente de treinamento e experimentação
5. **Integração via MLFlow**
   Project, Logging, Registry e Deploy de Modelos
6. **Deploy GCP**
   Deploy do modelo via MLFlow, validação e promoção
7. **Teste API via *local request***
   Requisições REST locais para validar *endpoint* disponibilizado
8. **Proposta de Evolução**
   Possibilidades de incremento e melhorias para o case proposto

Motivação

Extrair o máximo possível das ferramentas, utilizando do paralelismo, para habilitar capacidades integráveis aos *pipelines* do banco, sendo agnósticas ao provedor de Cloud e garantindo *deploys* de qualidade

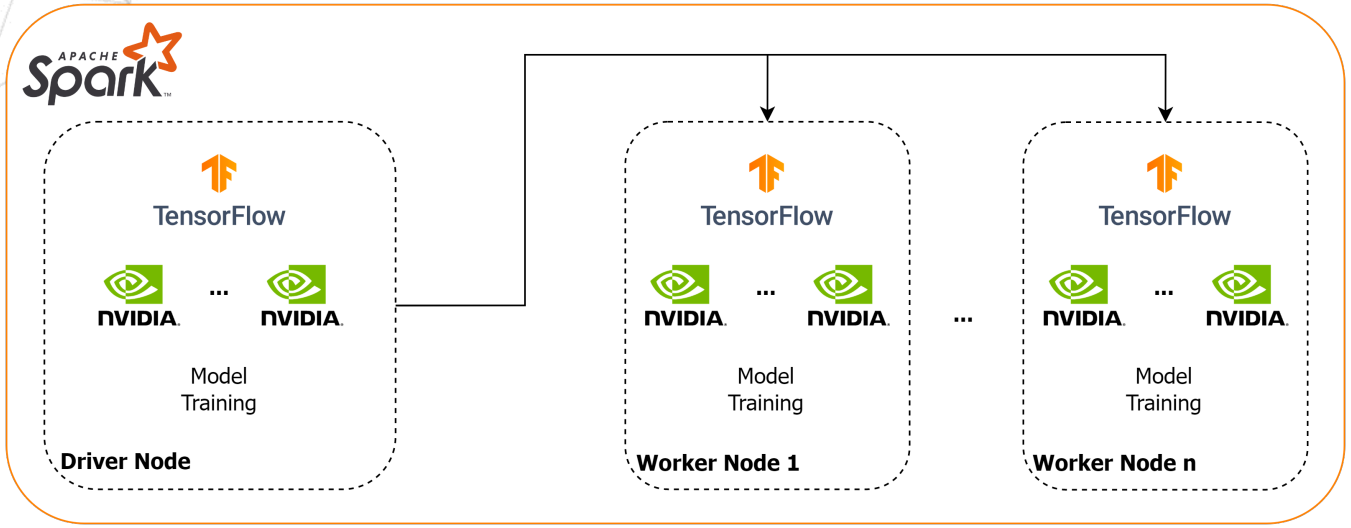
A apresentação será interativa, alternando entre slides e execução de código nos ambientes

Blue Prints – Arquitetura de Solução

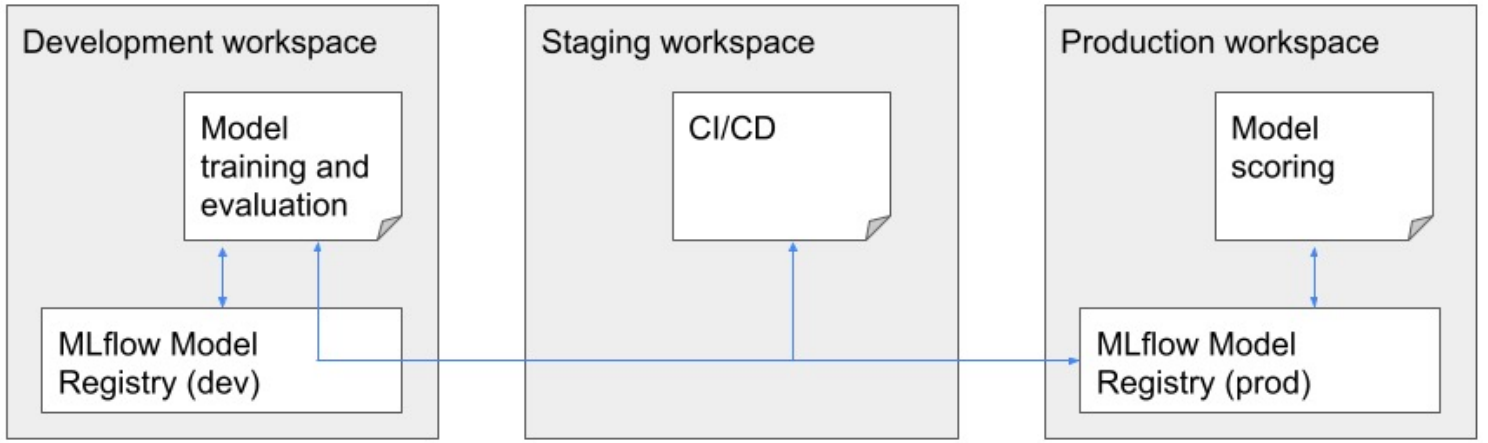


Arquitetura Técnica/Dados

Distributed Spark + TF Training



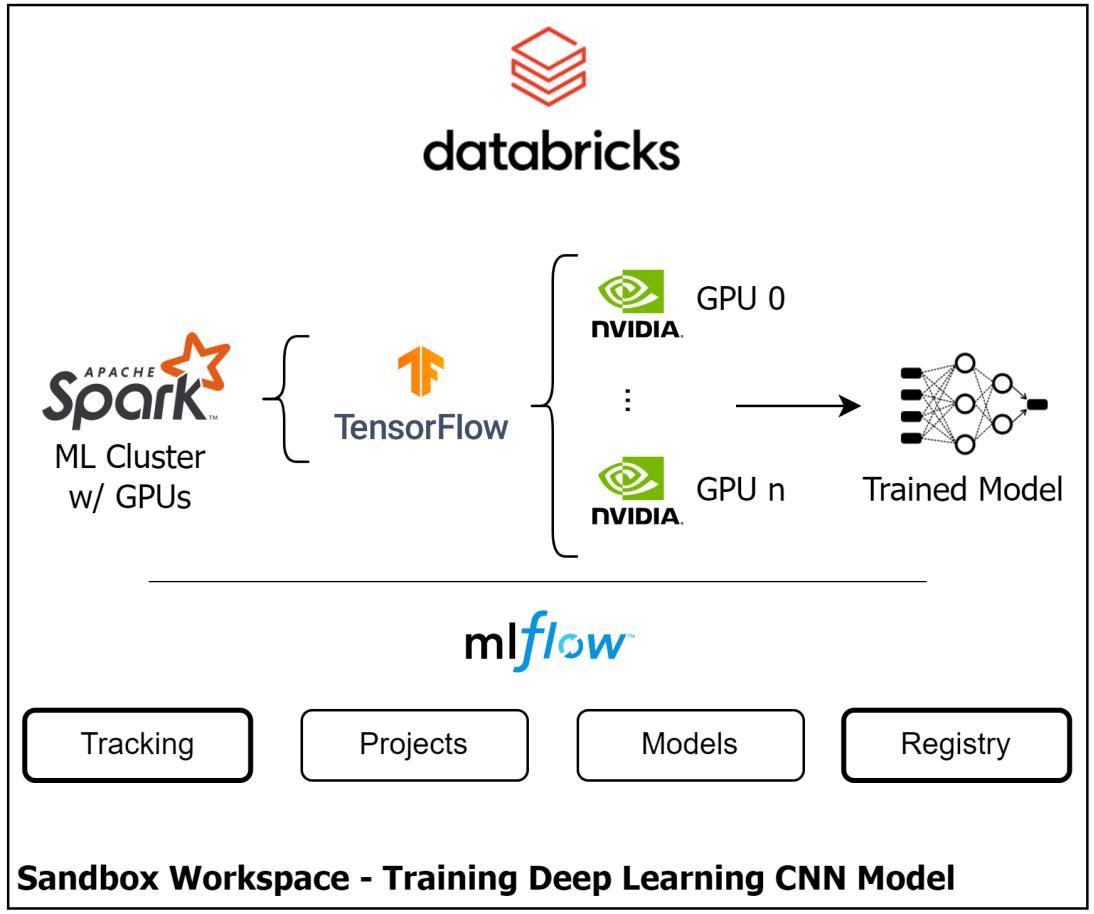
MLFlow Cross Workspace Integrations



Desafios

- Orquestrar o treinamento de modelos TF com Spark
  - Integração das libs + tools
- Comparar resultados de treinamento
  - Validação de eficiência e eficácia
- Integração segura entre workspaces
  - Criação e inclusão de secrets/scopes
- Registro de Modelos cross-cloud
  - Disponibilização dos artefatos multi-workspace e multi-cloud

Pipeline de Treinamento Azure – *Workspace Sandbox*



Microsoft Azure

Clusters

Name	Type	RAM	*Cores*	*Workers*	GPU
data_masters_nc12_2_gpu	Standard_NC6s_v3	112 GB	6	2-8	Tesla P100
data_masters_standard_4_cores	Standard_DS3_v2	14 GB	4	2-8	-

Scopes

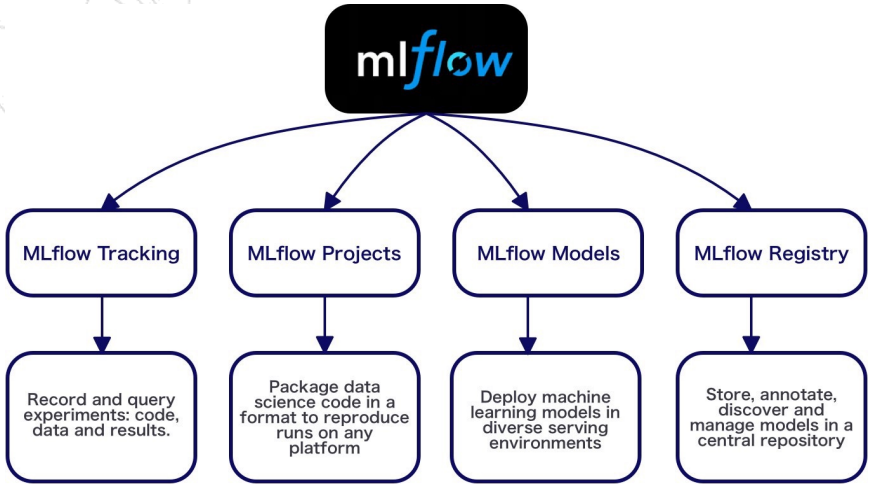
Scope	Key	Descrição
data_masters	data_masters_sandbox	Utilizada para teste de transferência de artefatos a partir do ambiente de deploy *

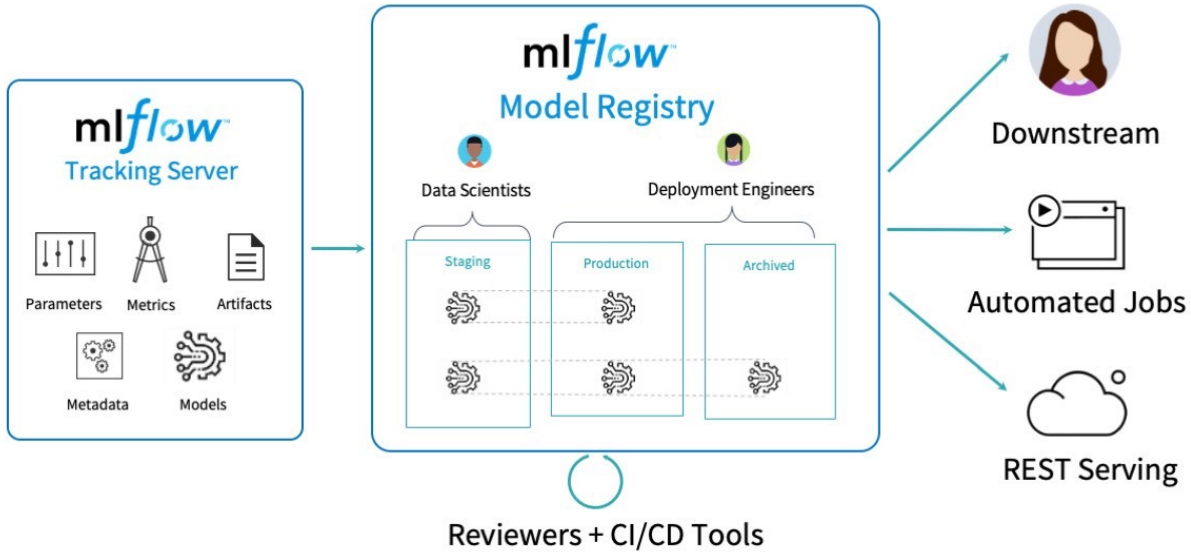
Abordagens de Treino

- Single Node
- Distributed: Distributed Mode
- Distributed: Local Mode
- Distributed: Custom Mode

* Grande gasto de tempo na tentativa de copiar artefatos direto para o repositório montado

Integração via MLFlow





Fluxo end-to-end

Log, Tracking and Evaluate

Acompanha todo o fluxo de treinamento

Registry Local

Criação do pickle do local, diretório com artefato do tensor

Registry Remoto

Comunicação segura via client, para registrar o pickle em outro diretório

Validação e Deploy

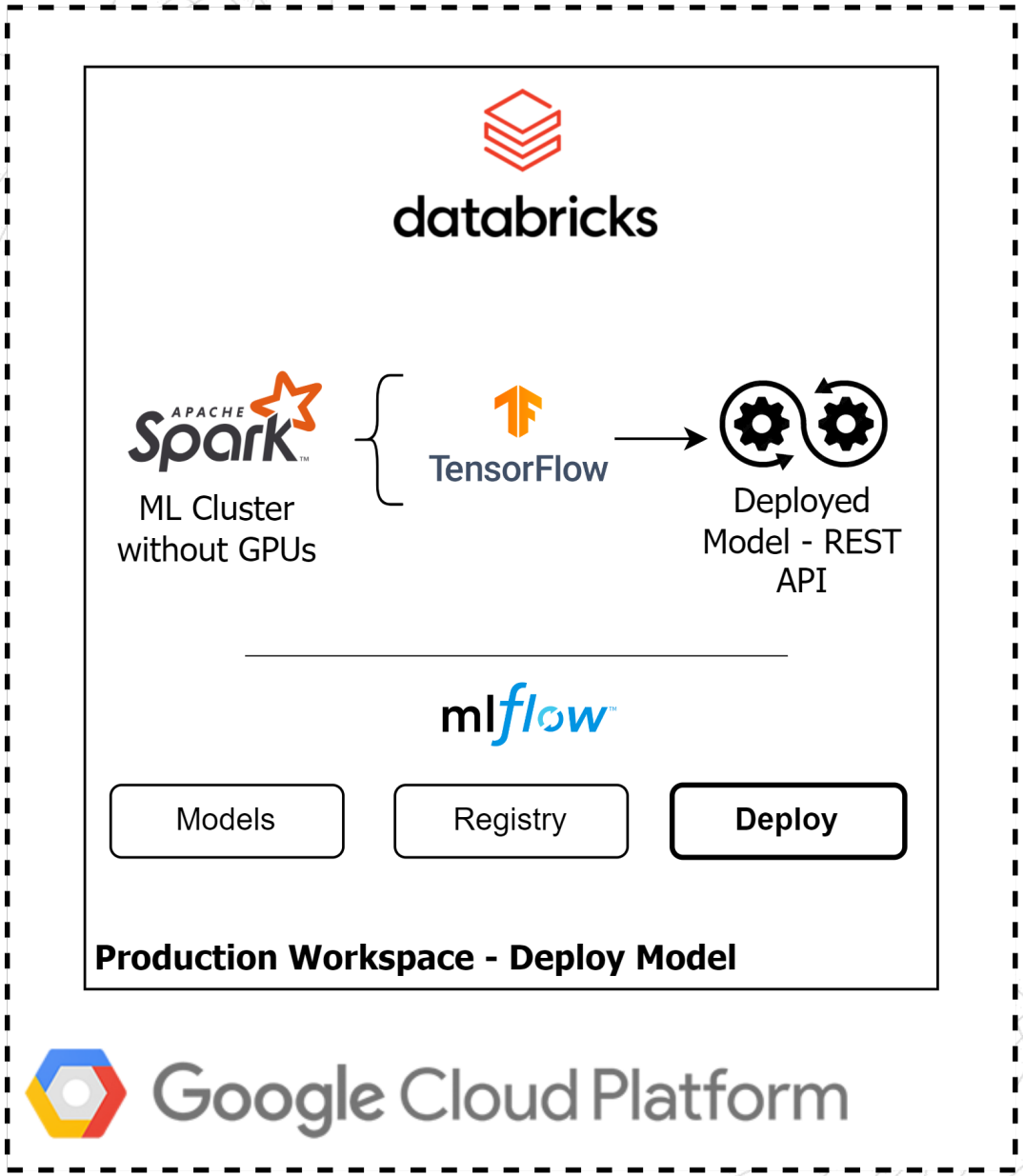
Load e Predict em Notebook, avanço do Stage e Deploy API REST

Requisitos

Permissão e Segurança

Acesso p/ criação de token, criação de secrets para acesso seguro

Deploy GCP



Clusters

Name	Type	RAM	*Cores*	*Workers*	GPU
data_masters_e2_8_cores_deploy	e2-standard-8	32 GB	8	2-4	-

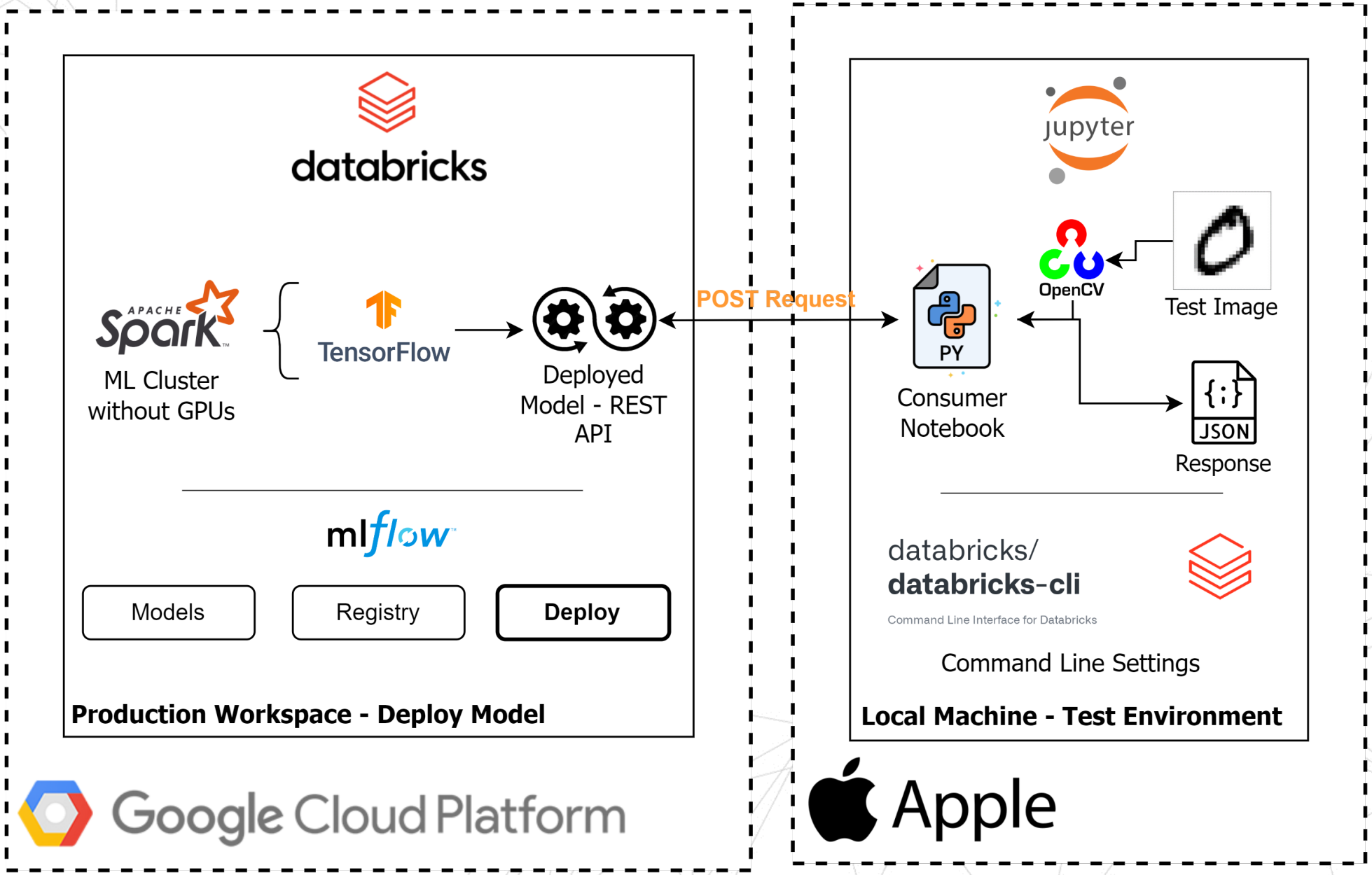
Scopes

Scope	Key	Descrição
data_masters_gcp	data_masters_deploy	Utilizada para registro de artefatos criados no sandbox

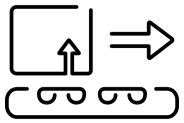
MLFlow Ciclo de Validação e Deploy

- Verificação do modelo registrado remotamente
- Adição de descrição e versão utilizando client
- Mudança de Stage (Stagging, Production) com client
- Load e Predict em notebook p/ validação de funcionamento
- Deploy usando MLFlow embedded serving

Teste API via *local request*

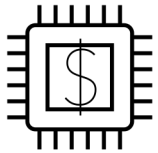


Proposta de Evolução



Acoplamento com Schedulagem e Esteira

Possibilitar o deploy dos modelos de produção usando o recurso de registry remoto combinado com esteira Maestro



Comparativo de performance GPU/CPU

Necessário um processo de treinamento mais longo para avaliar o real ganho com uso alavancado de multi-gpu, contrastado com o aumento de custos

Treinamento com arquiteturas de rede diferentes

Utilização da estratégia custom para treinar diferentes arquiteturas de modelos comparando resultados simultaneamente

Utilização de Terraform para criação de Cluster

Automatização da criação de clusters em todos os ambientes, principalmente com enfoque para o ambiente de deploy utilizando API RESt