MLOps, meet OPI

Debo Dutta - VP Engineering
About me

Current: VP Engineering, Nutanix (multi cloud platform co)
   AI4Infra (P&L engg) & Infra4AI (opensource)
   Founding member, MLCommons.org
   MLOps/Kubeflow.org (c2017)
   Openstack (c2012)

Past:
   Visiting Scholar, Stanford
   Distinguished Engg.@Cisco
   Researcher in Computational Biology
   PhD CS, USC/ISI (NS2 group)
   Btech, IIT Kharagpur
AI fueling digital transformation

- **Data -> AI -> Insights -> Actions**
  - Enterprise data explosion - databases, conversations, logs, metrics, knowledge graphs
  - Digital transformation: use enterprise data to increase operational efficiency
  - AI accelerates digital transformation by providing learned insights which leads to better actions.

- **Key challenge: Accelerating traditional businesses with MLOps**
  - Finance, retail, pharma, healthcare, public sector
  - Manage data and ML

Enterprise AI/ML use cases (similar infra needs)

- 57% Improving customer experience
- 50% Generating customer insights and intelligence
- 46% Detecting fraud
- 44% Increasing long-term customer engagement
- 43% Supply chain optimization
- 40% Managing inventory
- 39% Reducing costs
- 33% Financial planning
- 32% Generating financial insights
- 34% Acquiring new customers
- 31% Managing logistics
- 27% Back office automation
- 22% Increasing customer loyalty
- 18% Recommender systems
- 13% Automating processes
- 11% Retaining customers

@2021 Algorithmia AI survey
Where is the “valuable” enterprise data?

Source: images.google.com
OPI Opportunity: MLOps

End to end Workflow Orchestration

- Repeatable
- Composable
- Re-usable

Survey: Kubeflow PM Team
MLOps Market Opportunity

According to Cognilytica, the global MLOps market will be worth $4 billion by 2025.

The industry was worth $350 million in 2019.

Source: https://askwonder.com/research/historical-global-ai-ml-ops-spend-ulshsljuf
MLOps ecosystem: confusion

Source: images.google.com
OPI Opportunity: Optimizing MLOps
Current ML Infra is disaggregated

Training is inefficient due to communication complexity

ETL and Pre-training is expensive

Data is often in a filer, or object store or in DBs

Ref: https://www.researchgate.net/figure/Example-of-TensorFlow-graph-some-parts-of-the-graph-are-enlarged-for-better-visibility__fig3_323362772
MLOps + OPI

- Apache Spark
- mlflow
- Kubeflow Pipelines
- TensorFlow
- PyTorch
- mxnet
- Docker
- Kubernetes
- IPDK Interface

MLOps

IPU Network Fabric

ML Training Cluster

STORAGE

ML Training Cluster

CPU

CPU

GPU

johnugeorge, Kubeflow/Tensorflow
Potential OPI directions:

Reference Arch for ML With OPI
MLPerf workloads
MLCommons “cubes”
MLPerf Storage-ML

ML Image classification training cluster
Imagenet DB
KIITS DB

Docker
MLPerf
Kubernetes

IPU Network Fabric

Resnet model

U-Net3d model
Research-y Futures: Push down ML with OPI

Collaboration opportunities:
- Distributed training
- Split learning
- Federated learning
- SPDK primitives for storage+ML
- Infra optimization - scheduling

Nutanix Distributed File System (Data Plane)
- Scale-out architecture
- Resource pooling
- Distributed metadata management
- Automatic data tiering
- Physical storage abstraction
- Distributed storage operations
- Tunable data resilience
- Robust service resilience

S3 API

Any Hypervisor
- X86 Node

...