Challenges and Opportunities for AI Industrialization

Hui Lei, PhD, FIEEE
VP and CTO, Cloud and Big Data

Futurewei Technologies
The Reality of AI Industrialization

- Less than 10% of AI pilot projects have reached full-scale production (IIA)
- Only 25% of companies have revenue-bearing AI projects in production (O’Reilly)
- By 2022, just 15% of projects for AI and IoT will be successful (Gartner)
Don’t We Have Many, Many Tools Already?

- AI tooling landscape is crowded and confusing
- Existing tools require deep expertise and skills
- Siloed tools result in overhead and technical debt

Source: https://mattturck.com/data2019
Existing ML Platforms Offer Partial Relief

- Commercial: Microsoft Azure ML Studio, Amazon SageMaker, IBM Watson Studio, Google TensorFlow Extended
- Open source: Kubeflow, MLflow
- Proprietary: Facebook FBLearner, Uber Michelangelo, AirBnb BigHead, LinkedIn Pro-ML
Challenges
Production AI Is More Than ML

- Business Processes
  - Infrastructure Management
  - AI Asset Management
  - DevOps
  - Hybrid Multicloud

- Data Pipelines

- Model Pipelines

- Insight Pipelines

Processes for producing datasets used by machine learning

Processes for training and optimizing machine learning models

Processes for deriving business insight through analytics and inference
Siloed, Distributed and Heterogeneous Data

Nearly 90% of companies report high or moderate degrees of data silos (CompTIA)

<table>
<thead>
<tr>
<th>Technology</th>
<th>Benefits</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Lakes</td>
<td>Centralized storage of disparate data</td>
<td>Incomplete and disintegrated data</td>
</tr>
<tr>
<td></td>
<td>Computation close to data</td>
<td>Data duplication and staleness</td>
</tr>
<tr>
<td>Virtual distributed File Systems</td>
<td>Global namespace</td>
<td>File abstraction too primitive</td>
</tr>
<tr>
<td></td>
<td>Local or memory caching for performance</td>
<td>Unaware of higher-level semantics</td>
</tr>
<tr>
<td>Distributed SQL Engines</td>
<td>Declarative interface</td>
<td>Structured content only</td>
</tr>
<tr>
<td></td>
<td>Masking data heterogeneity</td>
<td>Pre-defined data schemas</td>
</tr>
<tr>
<td>Dataflow Engines</td>
<td>Flexible data analysis and processing</td>
<td>Requirements on expertise and skills</td>
</tr>
<tr>
<td></td>
<td>Pushdown processing</td>
<td>Data orchestration left with applications</td>
</tr>
</tbody>
</table>
Complexity in Insight Generation

- Execution container selection
- Resource selection and scheduling
- Component replication and scaling
- Data caching and prefetching
- Request batching
- Pipeline orchestration
- Messaging and queuing
- Straggler mitigation
- Latency and throughput monitoring
- Operational health monitoring

- Data/model lineage and provenance
- Data quality and outlier management
- Data drift detection and mitigation
- Concept drift detection and mitigation
- Adversarial attack detection and mitigation
- Prediction bias detection and mitigation
- Data and model explanation
- Data subject rights management
- Metadata management
- Model re-training
Innovation Opportunities
Integral AI Development

Pipeline Lifecycle Management
- Define
- Build
- Test
- Deploy
- Operate
- Data Pipelines
- Model Pipelines
- Insight Pipelines

AI Asset Management
- Catalog
- Traceability
- Privacy
- Trust
- Quality
- Metadata Repository
- Dataset Repository
- Model Repository
- Pipeline Repository

Infrastructure Management
- Provision
- Install
- Optimize
- Scale
- Update
- Data Integration Tools
  - Bare Metal
  - On-Premises
- Data Analytics Tools
  - Virtual Machines
  - Private Cloud
- Machine Learning Tools
  - Docker Containers
  - Public Cloud
- Serverless Functions
  - Edge Cloud
- IoT Devices

- Integrated and simplified development of all AI pipelines
- Unified catalog and governance for all AI assets
- Automated management of hardware/software infrastructure
Data Virtualization

Data Access Abstraction

- Unified declarative and procedural interfaces

Engine Orchestration

- Plan Optimization
- Pushdown Processing
- Engine Orchestration
- Materialized Views
- Index Caching
- Shuffle IO Reduction
- Batch Processing
- Stream Processing
- Query Processing

Data Orchestration

- Intelligent Preloading
- Near-Data Processing
- Adaptive Partitioning

SQL DBs
NoSQL DBs
Graph DBs
Files / Objects
Streams

- Optimization across data processing engines
- Geographically distributed analytics capabilities
- Moving data and compute closer together
- Siloed and heterogeneous data storage
Assembling insight pipelines from reusable components

- Accuracy
- Throughput
- Simplicity

- Cost efficiency
- Robustness
- Fairness
## Summary

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Innovation Opportunity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Production AI is more than ML</td>
<td>Integral AI development</td>
</tr>
<tr>
<td>Siloed, distributed and heterogeneous data</td>
<td>Data virtualization</td>
</tr>
<tr>
<td>Complexity in insight generation</td>
<td>Composable insight</td>
</tr>
</tbody>
</table>