Lessons learned from productising AI workloads with Open Source technology

Florian Moss, Automation Lead and Principal Architect
Financial Services, Red Hat

Use Case - GenAI for Equities and Risk Management

Investment Bank would like a chatbot that's trained on their own equities data so that they can avoid going to multiple sources (dashboards, applications, spreadsheets) to answer questions on portfolios, assets, and risk. They are open to RAG, Alignment Tuning, and Fine Tuning of LLMs.

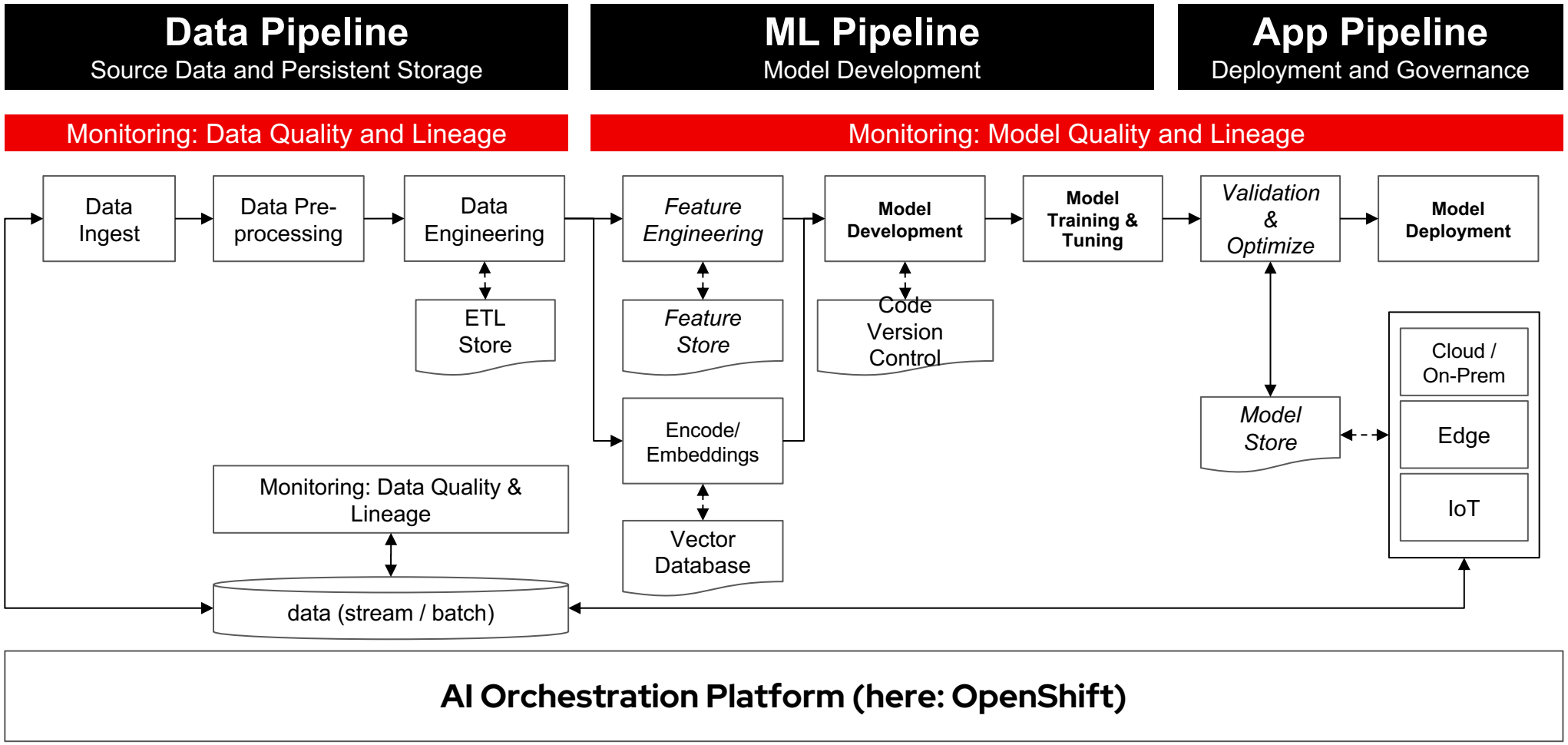
Equities and Risk Management Division

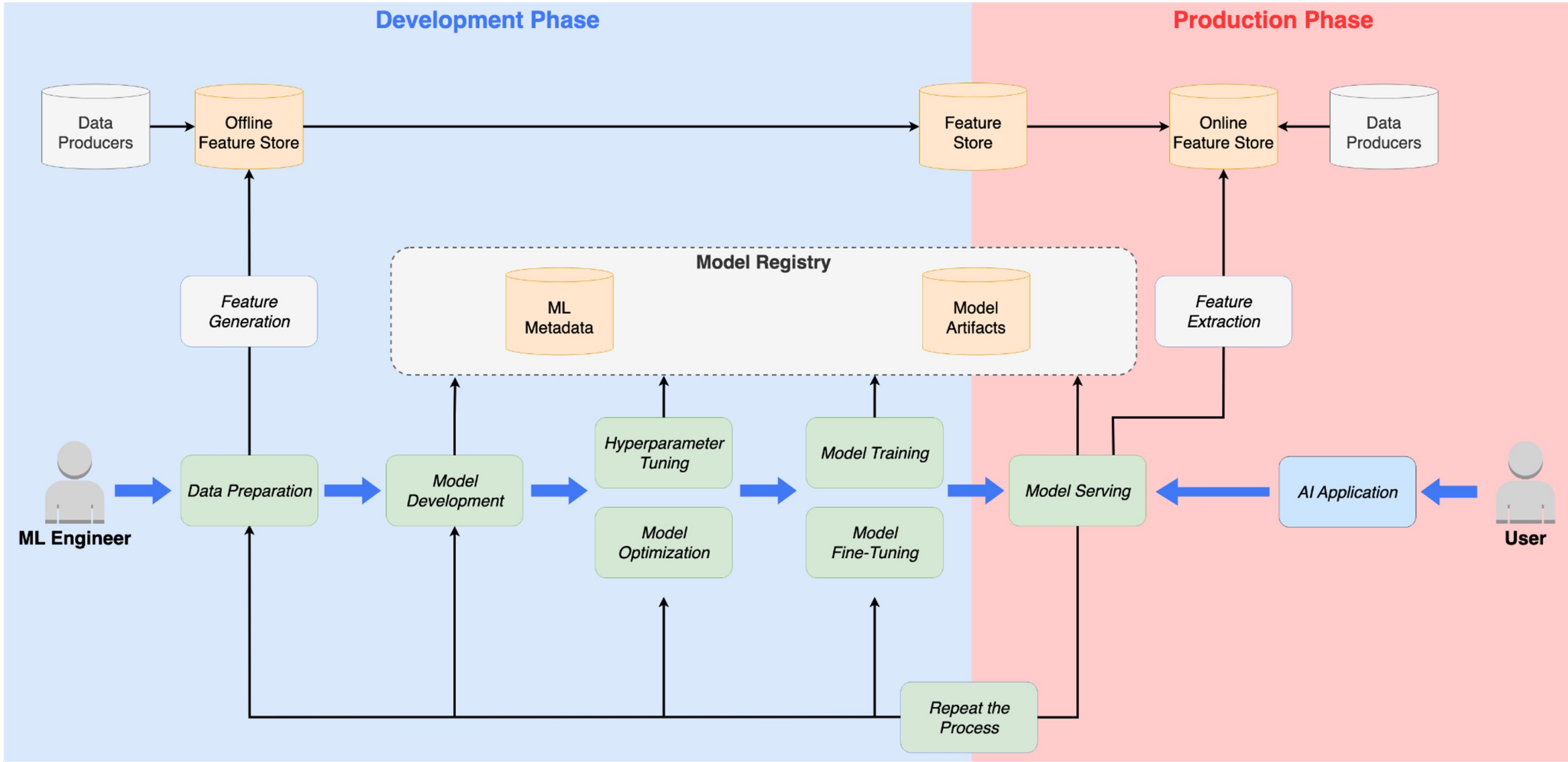
Chatbot Answering on Equities and Risk

- **Portfolio Monitoring:** Track portfolio performance, alerting users to significant changes or potential risks.
- **Risk Assessment:** Evaluate investment risk levels, suggesting adjustments based on market conditions.
- **Scenario Simulation:** Run hypothetical scenarios to predict outcomes and advise on risk mitigation strategies.
- **Regulatory Compliance:** Ensure adherence to regulations by providing guidance on compliance issues.
- **Automated Reporting:** Generate detailed reports on equities performance and risk metrics.
- **User Queries:** Answer questions about specific stocks, market conditions, and investment strategies.

Avoid lock-in, did some research: Kubeflow!

Our initial approach to model deployments




Development Phase
Production Phase
Data Producers
Offline Feature Store
Feature Store
Online Feature Store
Data Producers
Model Registry
ML Metadata
Model Artifacts
Feature Generation
Feature Extraction
ML Engineer
Data Preparation
Model Development
Hyperparameter Tuning
Model Optimization
Model Training
Model Fine-Tuning
Model Serving
AI Application
User
Repeat the Process

When going into production...

Deployment infrastructure and scalability

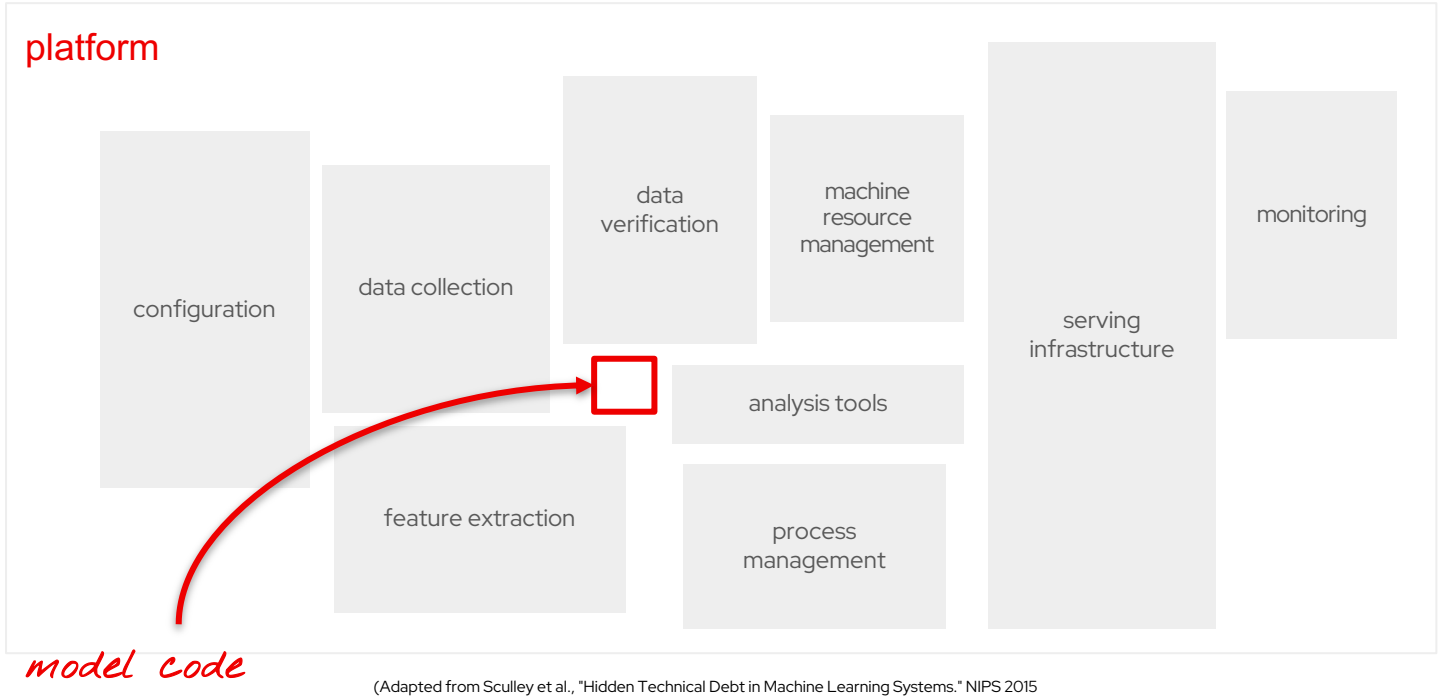
- ➔ All Cloud: Regulatory issues
- ➔ All on-prem: Resource constraints
- ➔ Both!

Kubeflow

- ➔ Keeping up with changes
- ➔ Resource Management
- ➔ Version Compatibility
- ➔ Bugs, bugs, bugs

We need an operations team and engineering partner!

Model code is one component of a larger system



(Adapted from Sculley et al., "Hidden Technical Debt in Machine Learning Systems." NIPS 2015

Hidden Technical Debt in Machine Learning Systems

"Only a small fraction of real-world ML systems is composed of the ML code…The required **surrounding infrastructure is vast and complex.**"

"Developing and deploying ML systems is relatively fast and cheap, but **maintaining them over time is difficult and expensive**"