ADA Project Presentation

Open Source Tools in Machine Learning Operations aka MLOps

Aniket Saha
MDS202207
anikets

Chennai Mathematical Institute

November 24, 2023

Contents

1. Introduction
   - Need of MLOps
   - Benefits
2. Machine Learning Pipeline
   - Stolen Words
   - Stolen Diagram
3. MLOps Using Open Source Tools
   - Overview
   - An Infographic
   - Brief Details of a Few Tools
4. Implementation
   - KNIME & MLflow
5. Conclusion
   - Problems
   - Comments
   - References

Intro to MLOps

Why do we need MLOps?

- MLOps, or **Machine Learning Operations**, is the amalgamation of Machine Learning, DevOps, and Data Engineering.
- The primary goal of MLOps is to **streamline and optimize the process of taking a machine learning model from development to production**.
- This methodology is required because **machine learning models are inherently different from traditional software applications**. They interact dynamically with the data they consume, requiring continuous monitoring, retraining, and maintenance to ensure they perform effectively when deployed.

Intro to MLOps

Benefits of MLOps

- **Improved Collaboration:** MLOps encourages collaboration across data scientists, ML engineers, and operations teams, fostering a culture of shared responsibility.
- **Efficient Scaling:** With MLOps, models can be scaled efficiently to handle increased loads and complex workloads without a proportional increase in operational issues.
- **Faster Deployment:** Automation of the ML pipeline enables faster deployment of models, reducing the time from development to production.
- **Better Model Management:** Version control for models and data ensures traceability, repeatability, and compliance.
- **Continuous Improvement:** Continuous integration and delivery allow for iterative improvements to models, leading to better performance and accuracy.

ML Project Pipeline

- **Design Phase**
  - **Planning:** Planning a project based on business needs.
  - **Data:** Identify source of data and collection of data.
  - **Data Management:** Versioning and preprocessing of datasets.
- **Development Phase**
  - **Experimentation:** Model development including training and tuning.
  - **Version Control:** Tracking of models and datasets versions.
  - **Testing:** Validating model performance for robustness.
- **Deployment Phase**
  - **Deployment:** Rolling out the model in a production environment.
  - **Monitoring and Operations:** Continuous performance and operational health checks.
  - **Feedback Loop:** Refining models with ongoing feedback.

MLOps Pipeline

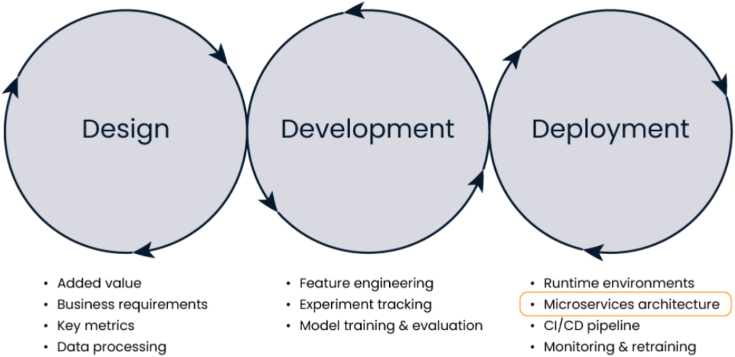

Figure 1: Overview of Pipeline

Reference: DataCamp

Open Source Tools in MLOps

- Amazon's AWS Sagemaker, Microsoft's AzureML and Google's Google Cloud AI provide end-to-end solutions for MLOps.
- However, here we will discuss some open source tools (or at least free to use) to do some similar tasks.
  - MLflow
  - ZenML
  - Kubeflow
  - Metaflow
  - Kedro
  - Ray
  - KNIME

MLOps Tools

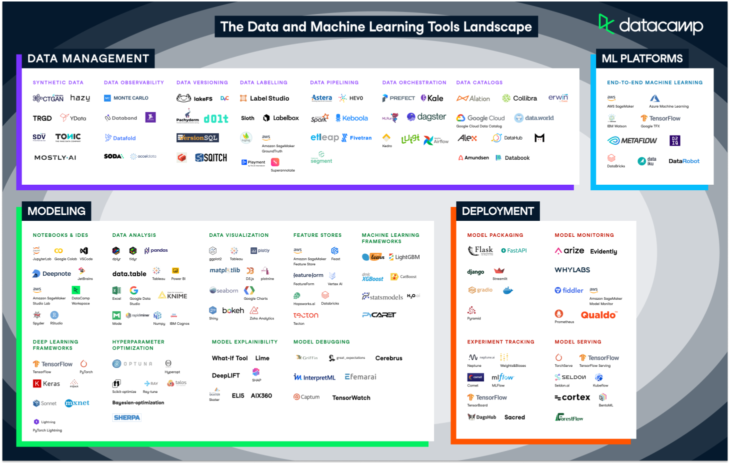

Figure 2: MLOps Tools

Reference: DataCamp

Machine Learning Tools Overview

- **MLflow**
  - MLflow is an open-source MLOps platform developed by Databricks, designed to simplify the management of Machine Learning workflows.
  - **Tracking:** Log experiments and metrics, and compare and evaluate models.
  - **Projects:** Organize and package code and dependencies, and simplify experiment reproduction.
  - **Website:** https://www.mlflow.org/
- **ZenML**
  - Developed by Maiot, it offers a comprehensive suite of tools for managing and scaling ML workflows. ZenML automates the end-to-end process of developing, deploying, and managing Machine Learning pipelines, from data collection to model deployment.
  - **Website:** https://zenml.io/

Machine Learning Tools Overview (Cont.)

- **Metaflow**
  - Metaflow is an open-source MLOps tool created by Netflix, designed to streamline the construction and management of data science workflows and projects.
  - Easy to learn, natively integrated with AWS whichhelps with data storage, computation, deployment.
  - **Website:** https://metaflow.org/
- **Kedro**
  - Kedro is an open-source MLOps tool developed by QuantumBlack, a McKinsey company, now part of the Anaconda ecosystem.
  - **Documentation:** https://kedro.readthedocs.io/
  - **GitHub Repository:** Kedro
  - **Using Kedro with Jupyter Notebook:** Kedro Notebooks

Machine Learning Tools Overview (Cont.)

- **Kubeflow**
  - Kubeflow is an open-source MLOps platform designed to simplify the deployment, scaling, and management of Machine Learning workflows on Kubernetes.
  - **Website:** https://kubeflow.org/
- **Ray**
  - Ray is an open-source unified compute framework allowing users to scale their AI and Python workloads quickly without needing complex infrastructures.
  - **Key Features:** Ray offers parallel and distributed execution primitives, enabling developers to scale AI and Python applications with minimal code changes using Pythonic APIs.
  - **Website:** https://ray.io/
- **KNIME**
  - **Purpose:** Data analytics, reporting, and integration platform.
  - **Key Features:** Visual workflow, extensibility, and collaboration.
  - **Website:** https://www.knime.com/

KNIME & MLflow

We will see implementations of a couple of open source MLOps tools.

- **KNIME**
  - KNIME is an open source tool which can be used for data management, data preprocessing, data analytics.
  - It has a well constructed GUI.
  - Useful for collaboration.
  - Has no-code ML features.
  - Can use many programming languages in it.
- **MLflow**
  - Open source platform for managing ML workflows.
  - Useful for tracking ML models.
  - Useful for version control of our ML Models.
  - We can show our analysis in a dashboard.

Problems

Does MLOps solve all problems?

- **Complex Data Management:** Handling massive datasets with privacy, security, and governance considerations.
- **Model Drift and Reproducibility:** Ensuring models remain accurate over time and can be reproduced or rolled back if necessary.
- **Infrastructure Complexity:** Deploying models across diverse environments requires robust infrastructure that can be complex to manage.

Conclusion

So we came across some tools that we can use to manage our MLOps needs. **So, why open source tools?**

- **Community Collaboration:** Leverages the power of a global community for continuous improvement, updates, and support.
- **Flexibility and Customization:** Provides flexibility and customization options for adapting tools to specific project requirements.
- **Cost-Effective:** Open source tools often provide cost-effective solutions compared to proprietary platforms.
- **Vendor Neutrality:** Avoids vendor lock-in, allowing users to choose and switch between cloud providers or on-premises solutions.

References

- DataCamp blogs. Link.
- Medium blogs. Link.
- MLflow Documentation. Link.
- Some GitHub repositories. (Source: Awesome MLOps) Link.
- ChatGPT. Link.
- YouTube lectures from Microsoft, Amazon, Google and other creators.