PROCESSING BIG DATA FOR ANALYTICS – MODEL AGGREGATION

Liu Fan
isslf@nus.edu.sg
Jan 2023

Total Slides 16:

Objectives

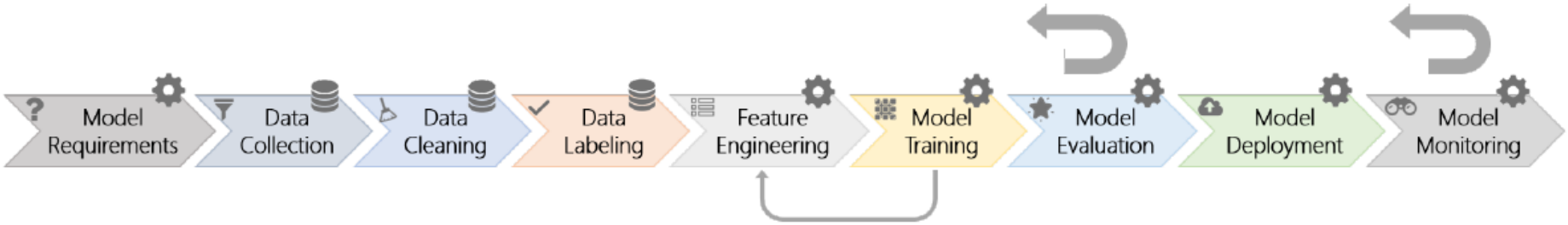
At the end of this module, you should be able to:

- Understand of the platform and business constraints
- Understand the aggregation opportunities

Agenda

- Different types of MLOPs platforms
- Business case and platform constraints understanding
- Model aggregation & it's needs

Machine Learning Lifecycle



- The machine learning lifecycle is the process of developing machine learning projects in an efficient manner.
- Building and training a model is a difficult, long process, but it's just one step of your whole task.
- There's a long process behind the machine learning lifecycle: data collection, data preparation, model training, model evaluation, model deployment, model monitoring.

MLOPs

- Data scientists and operation come together to take models to production.
- In many organizations data scientists work independently in a note book environment with no clear process defined for deployment and integration.
- Only very few organization shave tools to monitor model performance in production environment.
- MLOps stands for Machine Learning Operations. MLOps is focused on streamlining the process of deploying machine learning models to production, and then maintaining and monitoring them. MLOps is a collaborative function, often consisting of data scientists, ML engineers, and DevOps engineers.

Different MLOPs platforms

Open Source

- There are several open-source options available, like kubeflow, ModelDB, Seldon, and mlflow. You can pick the open-source products and stitch them together. Companies looking to build their own custom MLOps solution may choose this route.

Cloud Providers

- Most of the popular cloud providers also offer solutions to help manage the entire ML life cycle including MLOps: from data preparation to annotation to model training to deployment. The kitchen sink approach may make things overly complex, but it's worth considering if you are looking for a one-stop-shop solution.

https://insidebigdata.com/2021/03/19/how-to-choose-the-best-mlops-platform-for-your-organization/

Amazon SageMaker

Amazon SageMaker is an ML platform which helps you **build, train, manage, and deploy** machine learning models in a production-ready ML environment. SageMaker accelerates your experiments with purpose-built tools, including labeling, data preparation, training, tuning, hosting monitoring, and much more.

Features

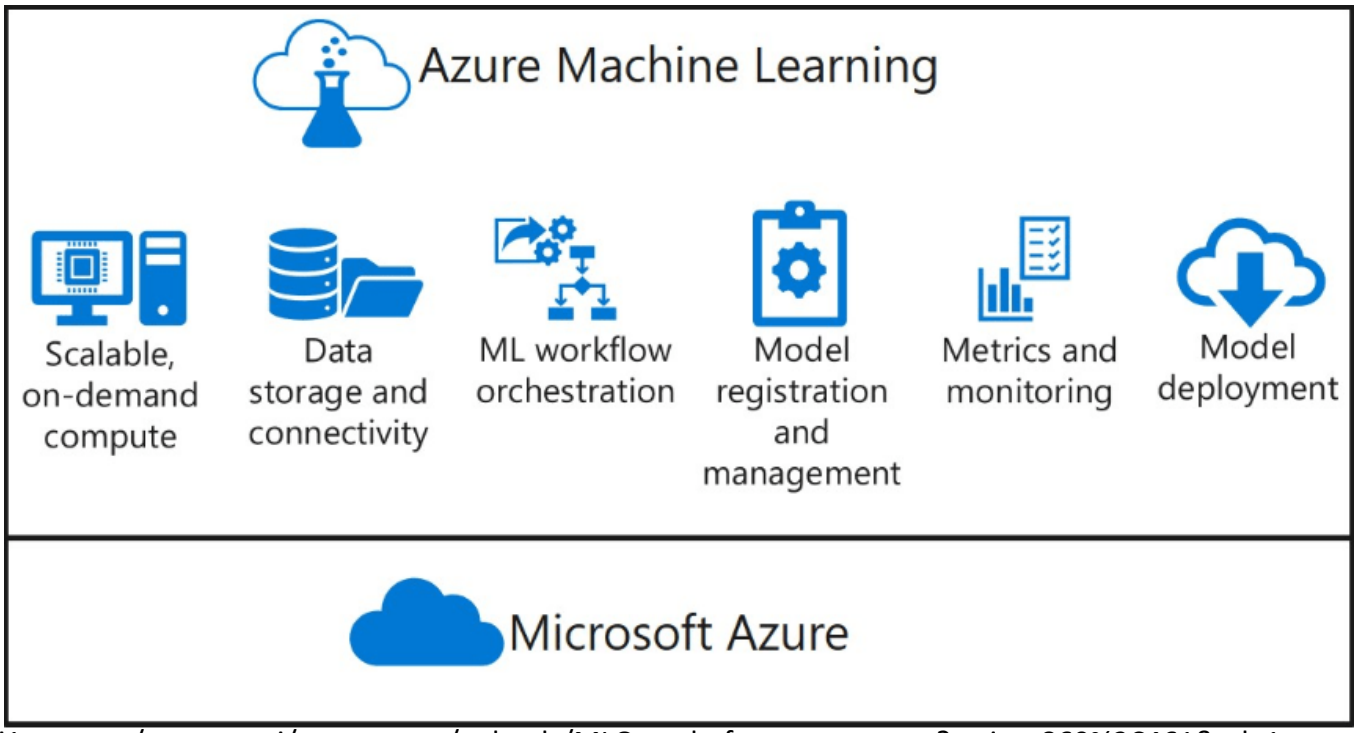
- Sagemaker comes with many ML Algorithms for training your dataset (big datasets)
- Sagemaker includes both supervised unsupervised ML algorithms
- The end-to-end ML platform speeds up the process of modeling, labeling, and deployment. AutoML features will automatically build, train and tune the best ML Model based on your data.
- SageMaker gives you the option to integrate APIs and SDKs, making it easy for setup, and you can use machine learning functions anywhere.

Pricing

- SageMaker is not free. If you're a new user, you might get the first two months free.
- Pricing is broken into ML storage, instances, and data processing.

Azure Machine Learning

Azure ML is a cloud-based platform which can be used to train, deploy, automate, manage, and monitor all your machine learning experiments. Just like SageMaker, it supports both supervised and unsupervised learning.



https://i1.wp.com/neptune.ai/wp-content/uploads/MLOps-platforms-azure.png?resize=860%2C461&ssl=1

Azure Machine Learning

Feature

- Azure ML platform supports both Python, R, Jupyter Lab and R studios with automated machine learning.
- The **drag-and-drop** feature provides a **code-free** machine learning environment which helps data scientists to collaborate easily.
- You get the option to train your model on your local machine or in the Azure machine learning cloud workspace.
- Azure machine learning supports many open source tools including Tensorflow, Scikit-learn, ONNX, Pytorch. It has its own open source platform for MLops (Microsoft MLOps).
- Some key features include collaborative notebooks, AutoML, data labeling, MLOps, Hybrid and multi-cloud support.

Pricing

- Azure Machine learning provides a 12-month free service and few credits to explore Azure. After the free credits/service is over, you pay for what you use.

Google Cloud Platform

Google Cloud is an **end-to-end fully managed platform** for machine learning and data science. It has features which help you manage service faster and seamlessly. Their ML workflow makes things easy for developers, scientists, and data engineers. The platform has many functions which support machine learning lifecycle management.

Feature

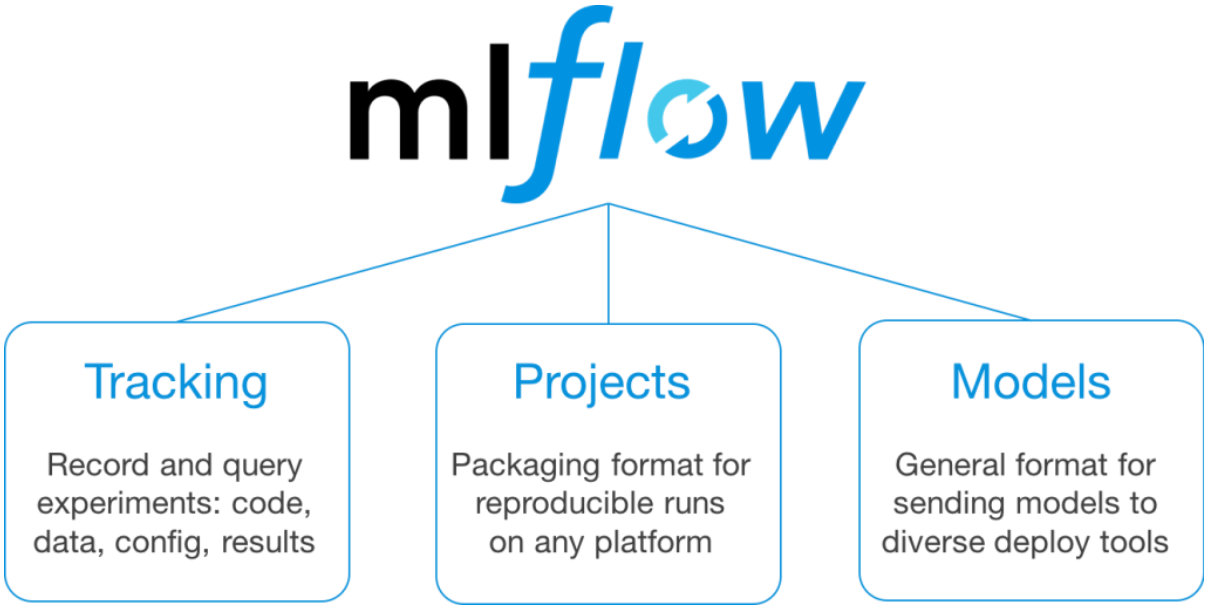
- Cloud storage and bigquery helps you prepare and store your datasets
- You can perform your task without writing any code by using the Auto ML feature with an easy-to-use UI. You can use Google Colab where you can run your notebook for free.
- Deployment can be done with Auto ML features, and it can perform real-time actions on your model.
- Manage and monitor your model and end-to-end workflow with pipelines.

Pricing

- You'll get a 300$ free credit when you start using GCP. When the free trial is over, you will be charged monthly based on what tools you have used.

MLFlow

MLflow is an **open-source** platform for managing the machine learning lifecycle – experiments, deployment and central model registry. It was designed to work with any machine learning library, algorithm and deployment tool.



https://i2.wp.com/neptune.ai/wp-content/uploads/MLflow-components.png?resize=1024%2C512&ssl=1

MLFlow

Feature

- It can work with any machine learning library, language or any existing code. It runs in the same manner in any cloud.
- MLflow mainly consists of four components, MLflow tracking, MLflow projects, MLflow models and MLflow registry.
- MLflow tracking is all about recording and querying your code and data experiments.
- MLflow projects is a package of data science which provides code in reusable and reproducible format. It also includes an API and cmd tool for running ML and data science projects.
- MLflow models help you deploy different types of machine learning models. Each model is saved as a dir containing arbitrary files.

TensorFlow Extended (TFX)

Tensorflow Extended is a Google-production-scale ML Platform. It provides shared libraries and frameworks to integrate to your machine learning workflow.

Feature

- TensorFlow extended lets you orchestrate your machine learning workflow on many platforms like Apache, Beam, KubeFlow, etc..
- It provides full production ML pipeline. TFX Pipeline includes ExampleGen, StatisticsGen, SchemaGen, ExampleValidator, Transform, Trainer, Tuner, Evaluator, InfraValidator, Pusher, BulkInferrer.
- TFX components provide functions to help you get started developing machine learning processes easily.

Essential criteria to choose an MLOps

- **User-friendly** - The platform should work out of the box with popular ML frameworks so you are not hindering innovation by restricting your data science teams to work on specific tools and frameworks.
- **Ease of use** –Look for these three aspects when evaluating a platform: easy to install, easy to set up, and easy to customize.
- **Interoperability** – MLOps is a new layer in your software stack, and the platform should play well with your existing ecosystem of model training, deployment pipeline, monitoring, and approval workflow tools.
- **Reproducibility** – Managing a model life cycle begins with the ability to version and make them accurately reproducible. Reproducibility is critical whether you are collaborating with team members, debugging a production failure, or iterating an existing model.
- **Scalability** – When we talk about production operations, a scalable platform is a must-have. Choose a platform that not only meets your current needs but can scale for the future. Look for a platform that can elegantly scale for both real-time and batch workloads, serve high throughput scenarios, scale automatically with the increasing traffic, manage cost versus user experience efficiently, follow safe deployment and release best practices.

Summary

- Model aggregation is important, as eventually we need to deploy our ML model into production environment.
- The two types of integration platforms are: open-source and Cloud Providers.
- Which platform to choose is depend on the budget, business requirements, services, etc..
- There are some essential criteria to consider when you choose a platform.

References

- https://neptune.ai/blog/improving-machine-learning-deep-learning-models
- https://www.forbes.com/sites/cognitiveworld/2020/12/13/the-five-major-platforms-for-machine-learning-model-development/?sh=a6c45244e371
- https://www.xenonstack.com/blog/mlops
- https://www.datacenterdynamics.com/en/opinions/constraints-cloud-why-we-need-machine-learning-edge/