Data Science in production

Lecture 4: Model serving & deployment strategies

Alaa BAKHTI

Prediction strategies

Prediction strategies

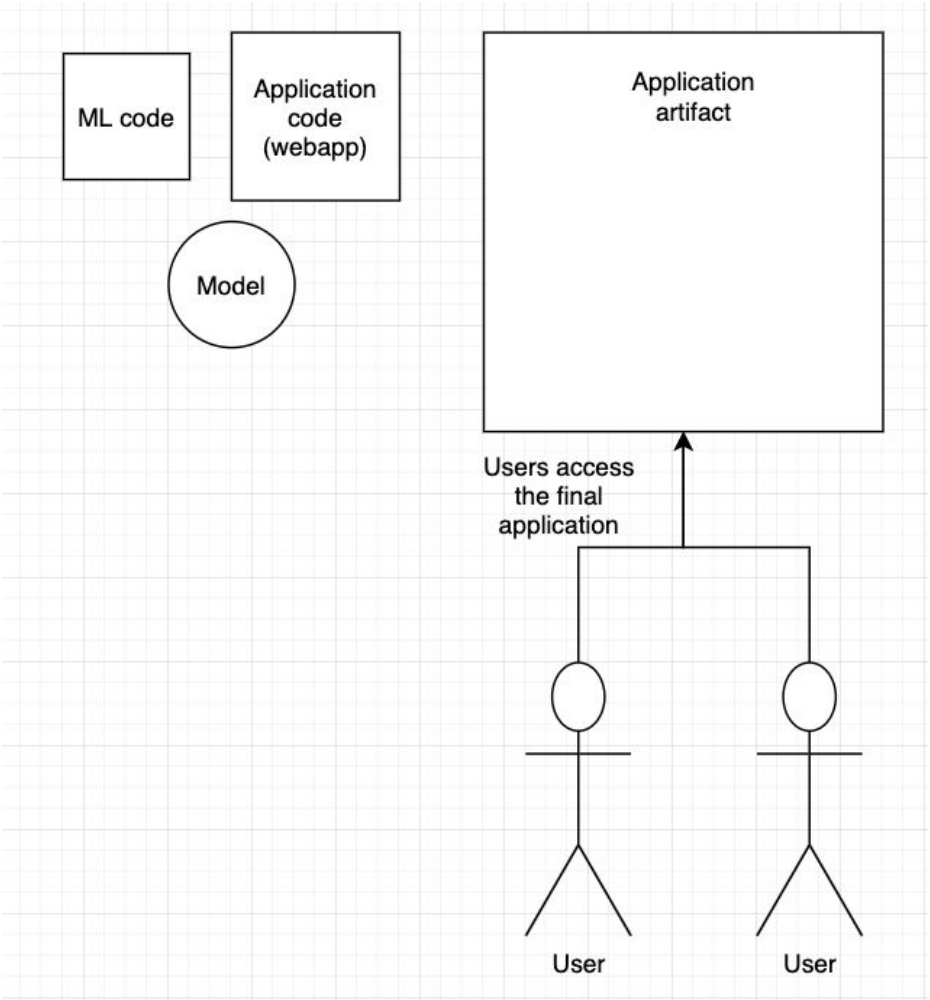
Batch prediction / inference

- The process of generating predictions on a batch of observations
- Scheduled at a recurrent time period (e.g. hourly, daily, etc)

Online prediction / inference (streaming)

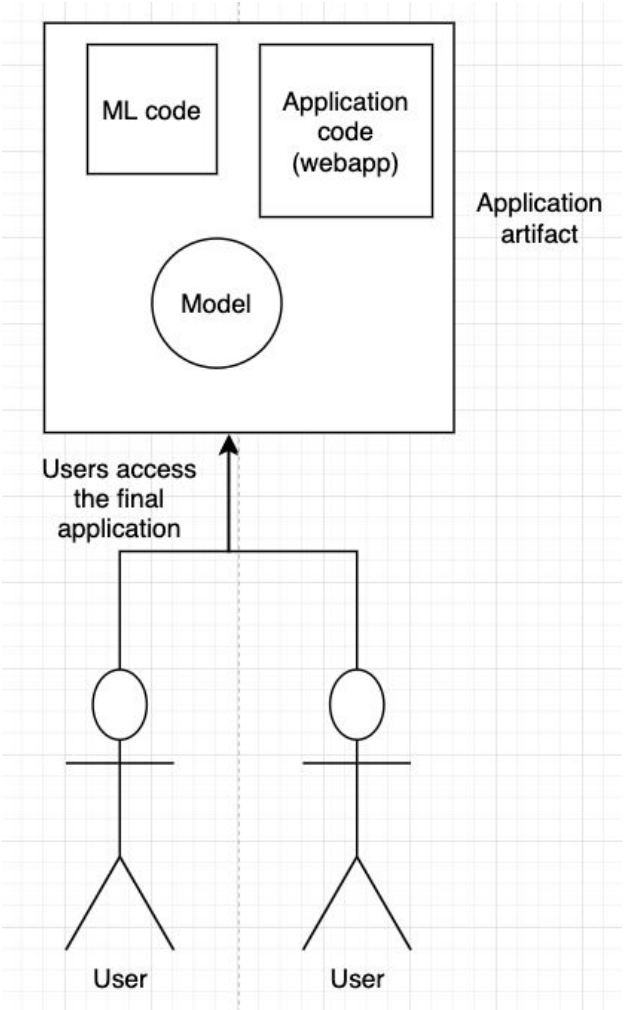
- The process of generating predictions in real time
- Typically, these predictions are generated on a single observation of data at runtime

Serving strategies


ML code
Application code (webapp)
Model
Application artifact
Users access the final application
User
User

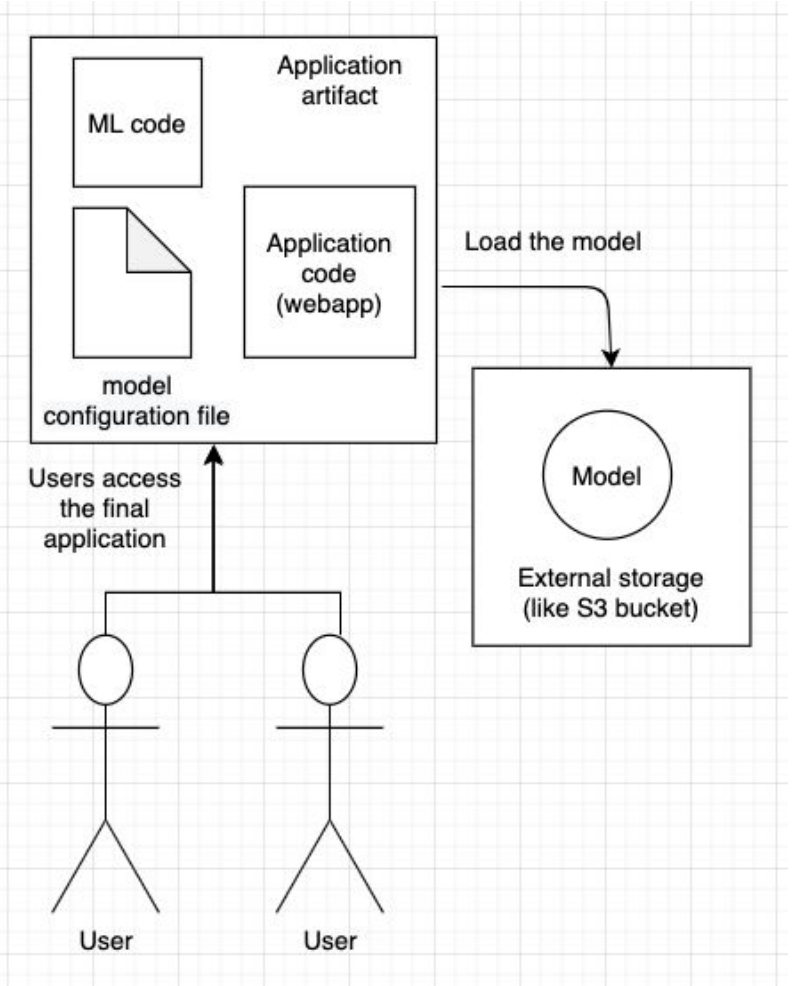
Embedded model

- Embed the model in the application
- The inference pipeline and model are included in the application artifact
- The ML part (code + model) and the application code are coupled



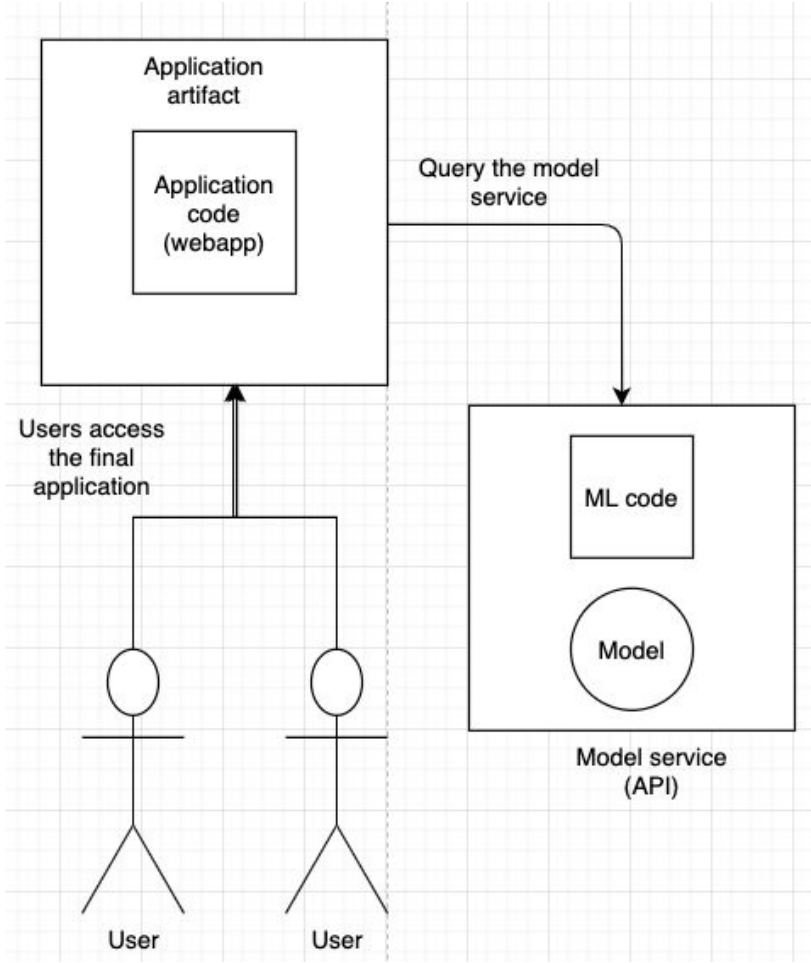
Model published as data

- The model is not stored in the application artifact but the ML code is. The model is stored in a File Storage
- To determine what model to use, the model configuration is used
- Example: the model is stored in a an s3 bucket or blob storage and its address and how to load it is in the application code.
- The model is loaded at the start of the application
- If we want to change the model, we need to update its configuration file and to restart the application



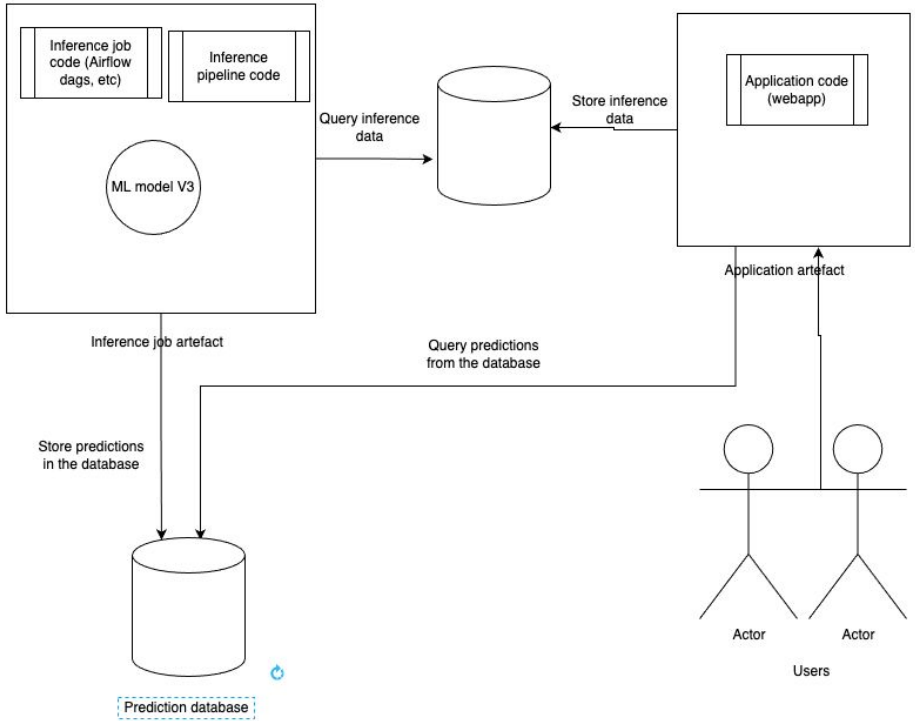
Model as a service

- The model is exposed via an API (Application Programming Interface)
- The application make post requests to the API to make predictions with the model



Exposing model predictions

- The model predictions are saved in a database
- The web application queries predictions from the database
- Useful when the model predictions are used by multiple applications (webapp, dashboard, monitoring, etc)



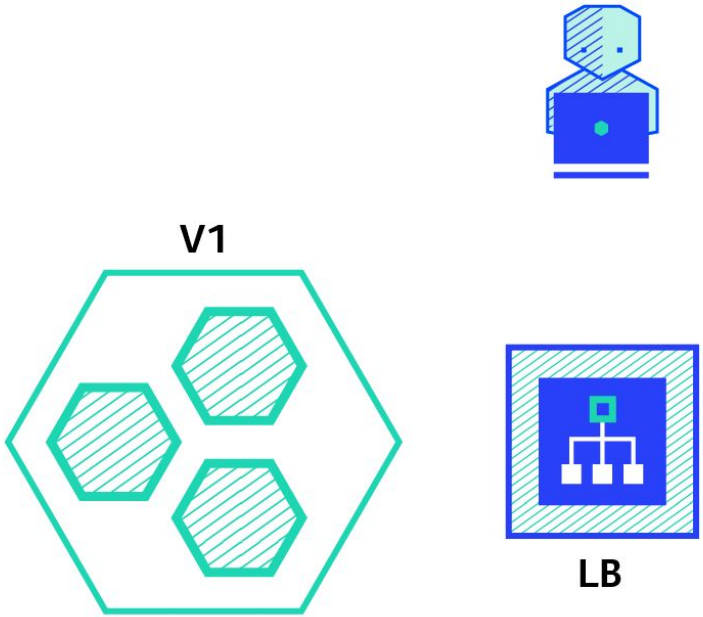
API

“In computing, an application programming interface (API) is an interface that defines interactions between multiple software applications or mixed hardware-software intermediaries. It defines the kinds of calls or requests that can be made, how to make them, the data formats that should be used, the conventions to follow, etc.” - *Wikipedia*

- API frameworks: Django, Flask, FastAPI, Requests...
- Categories: Full stack framework, micro framework, client library
- The components of an API framework

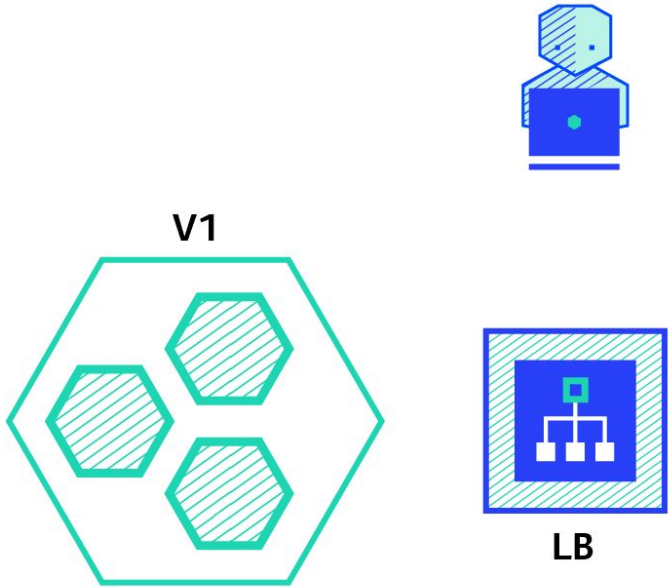
Deployment strategies

Recreate



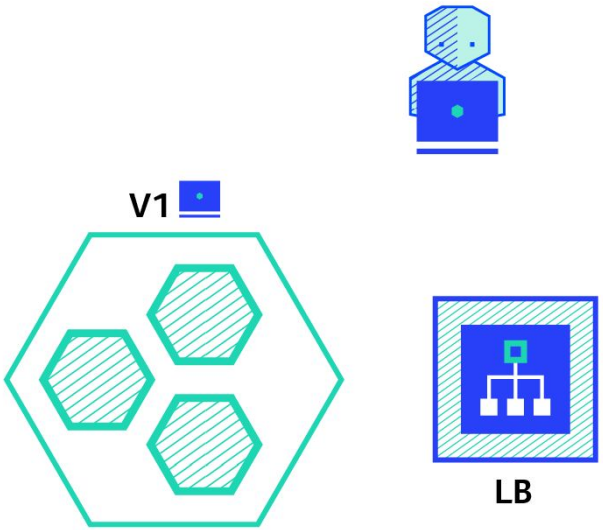
Source: https://thenewstack.io/deployment-strategies/

Canary deployment



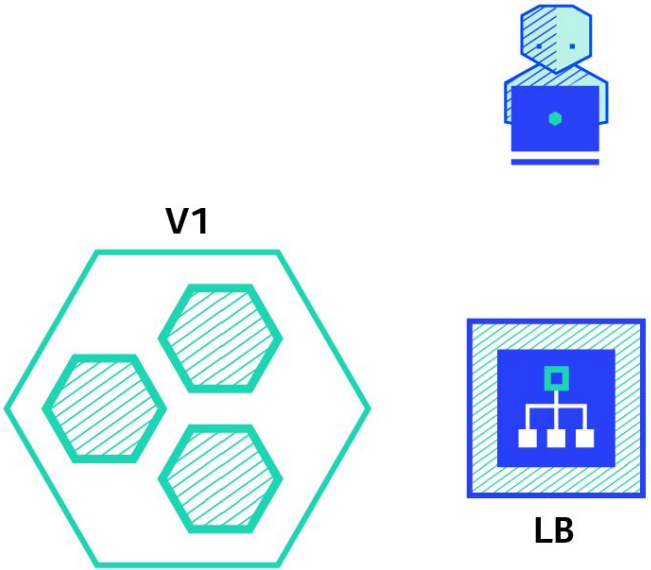
Source: https://thenewstack.io/deployment-strategies/

A/B testing



Source: https://thenewstack.io/deployment-strategies/

Shadow production



Source: https://thenewstack.io/deployment-strategies/

Practical work

Practical work

Implement the serving strategies:

- Embedded model
- Model published as data
- Model as a service

Frameworks to use: FastAPI, Streamlit, requests

Ressources

- Overview of the different approaches to putting Machine Learning (ML) models in production
- Six Strategies for Application Deployment – The New Stack