greatlearning
Learning for Life

Model Deployment

Putting your ML system into production

greatlearning
Learning for Life

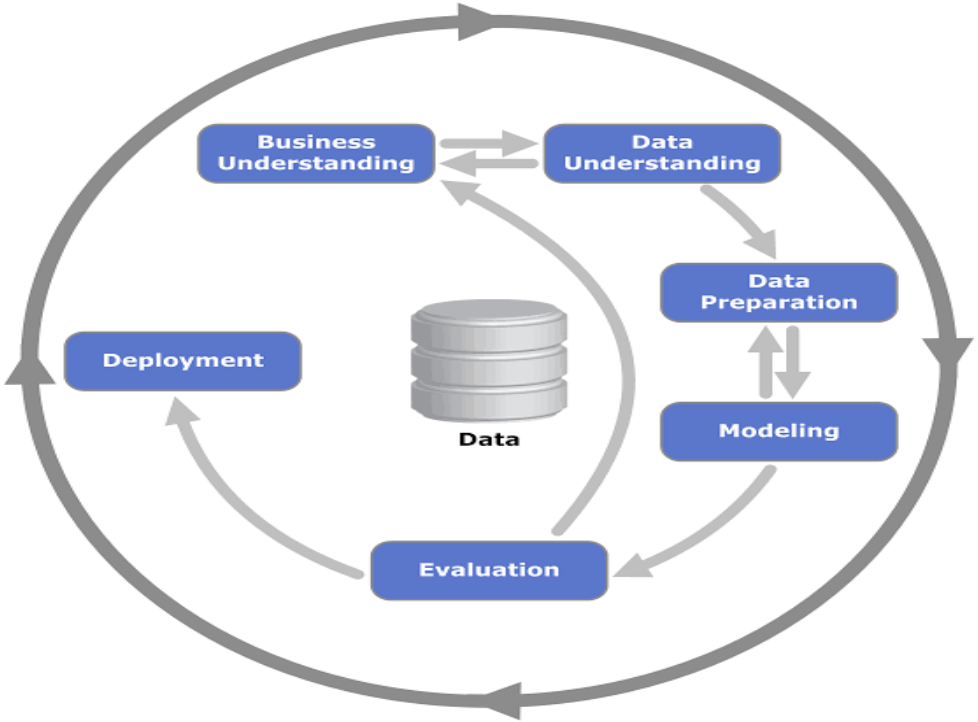
Agenda

- Serializing machine learning models
- Exposing the model through Rest APIs
- Packaging for reproducibility
- Create ML pipeline
- Scaling the model

What is model deployment?

ML model deployment is the process of publishing your model, which is currently in your local machine, to a larger user base.

ML Process Overview

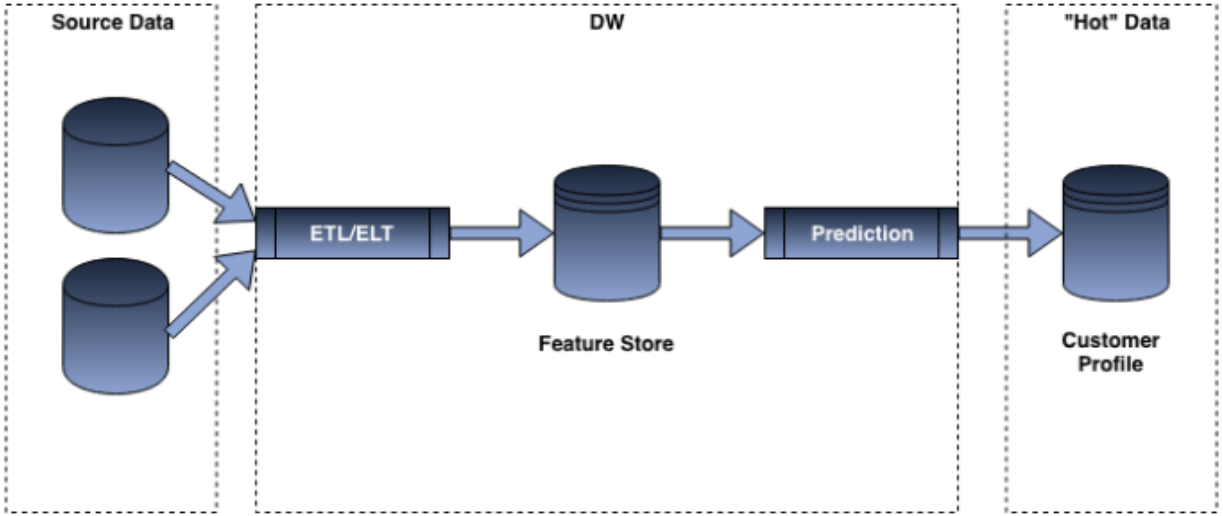


Modes of training and serving the models

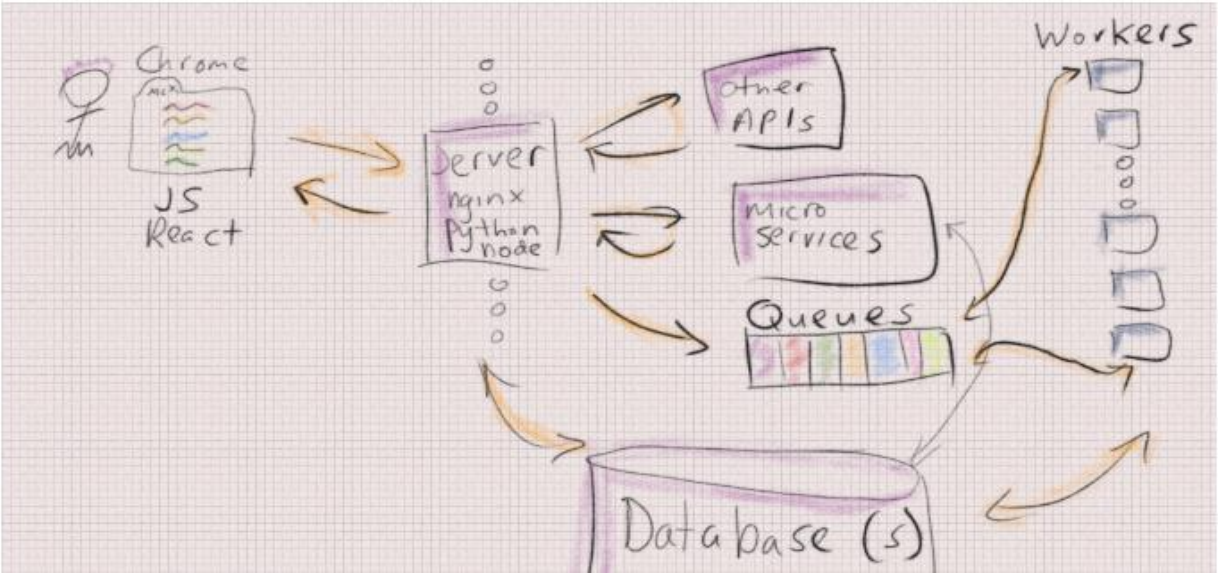
- **Train:** one off, batch and real-time/online training

- **Serve:** batch, real-time (web service, in-app, database trigger)

Batch prediction



Real-time prediction



Model serialization aka pickling

greatlearning
Learning for Life

Create REST API using flask

Docker - what problem it solves?

- Build once and run anywhere with Docker
- No environment issues
- No OS issues
- Preconfigured environment

Source: developermemes

Kubernetes - why to use?

Kubernetes is an orchestration platform which enables -

- Fault tolerance
- Auto-Scale
- Load Balancing
- Rolling service updates

"Google runs all software in containers and they run around 2 billion containers every week."

greatlearning
Learning for Life

Deploy and scale docker with kubernetes

greatlearning
Learning for Life

Any Questions?

greatlearning
Learning for Life

Thank you!

Happy Learning :)