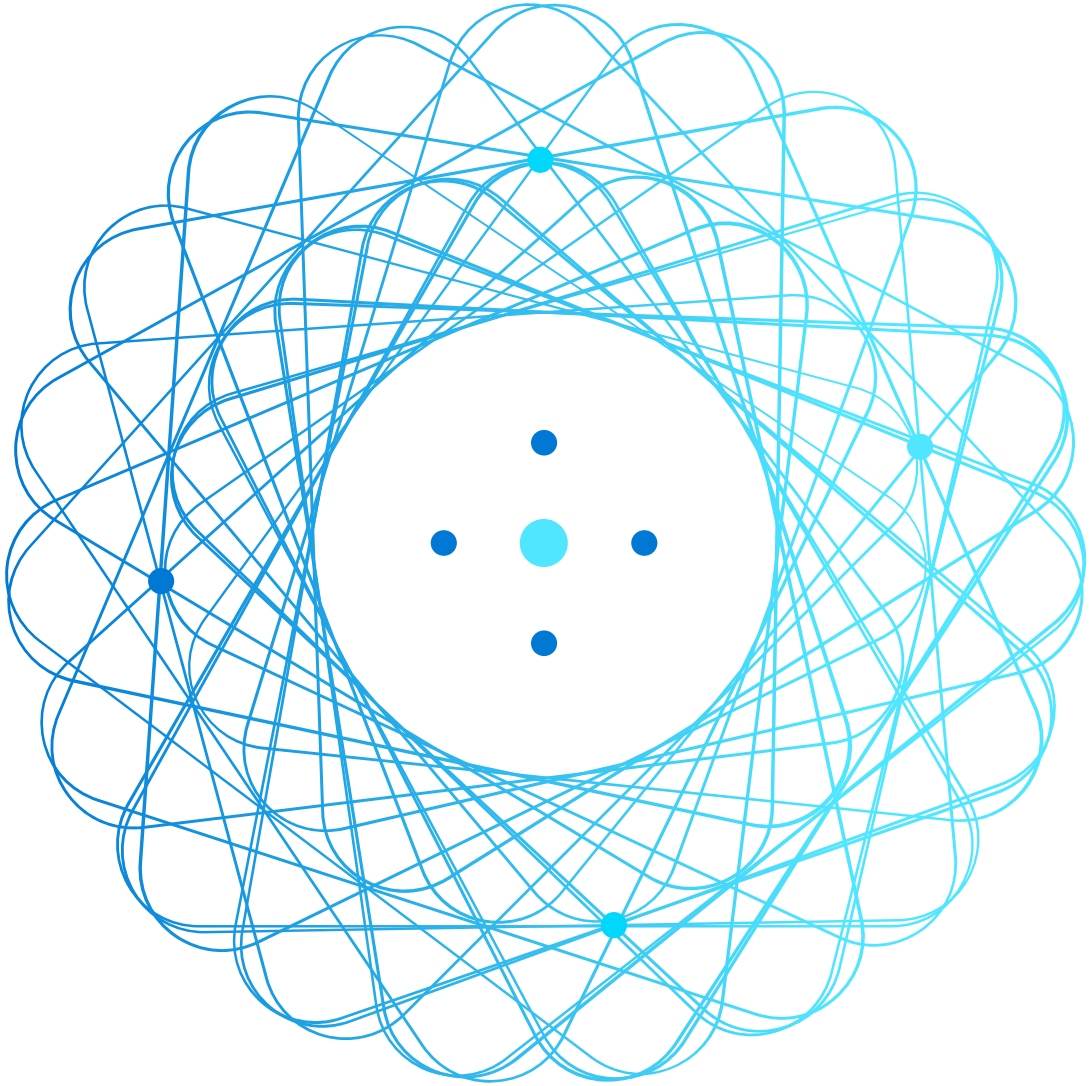
Microsoft Azure

Design a machine learning operations (MLOps) solution



Agenda

Design a machine learning operations (MLOps) solution

Design a machine learning operations (MLOps) solution

Understand MLOps

Machine Learning operations or **MLOps** help you to scale your model from a proof of concept or pilot project to production.

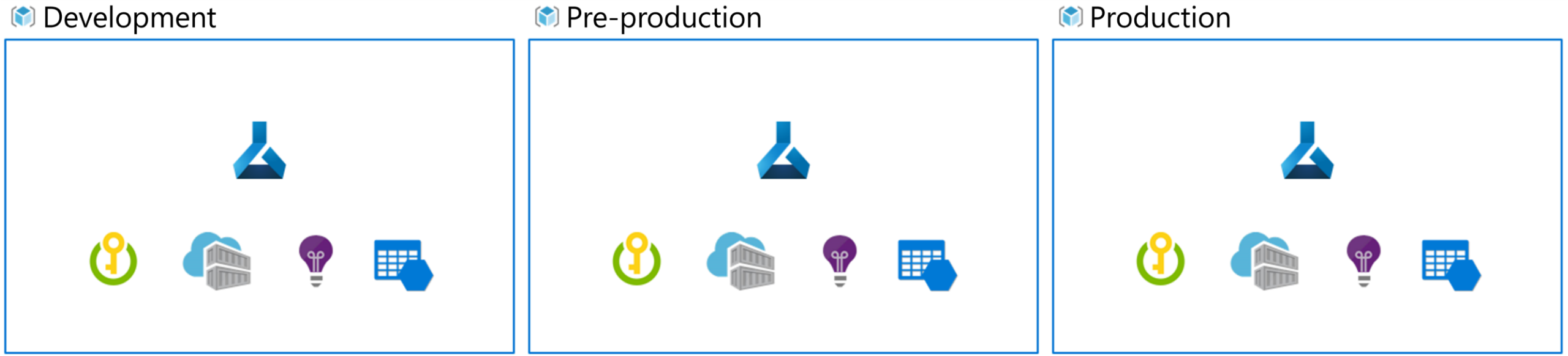
A model in production is ready for *large-scale deployment* and can be retrained and redeployed when necessary.

Implementing MLOps helps you to make your machine learning workloads *robust* and *reproducible*.

Set up environments for development and production

Within MLOps (and DevOps), an environment refers to a collection of resources.

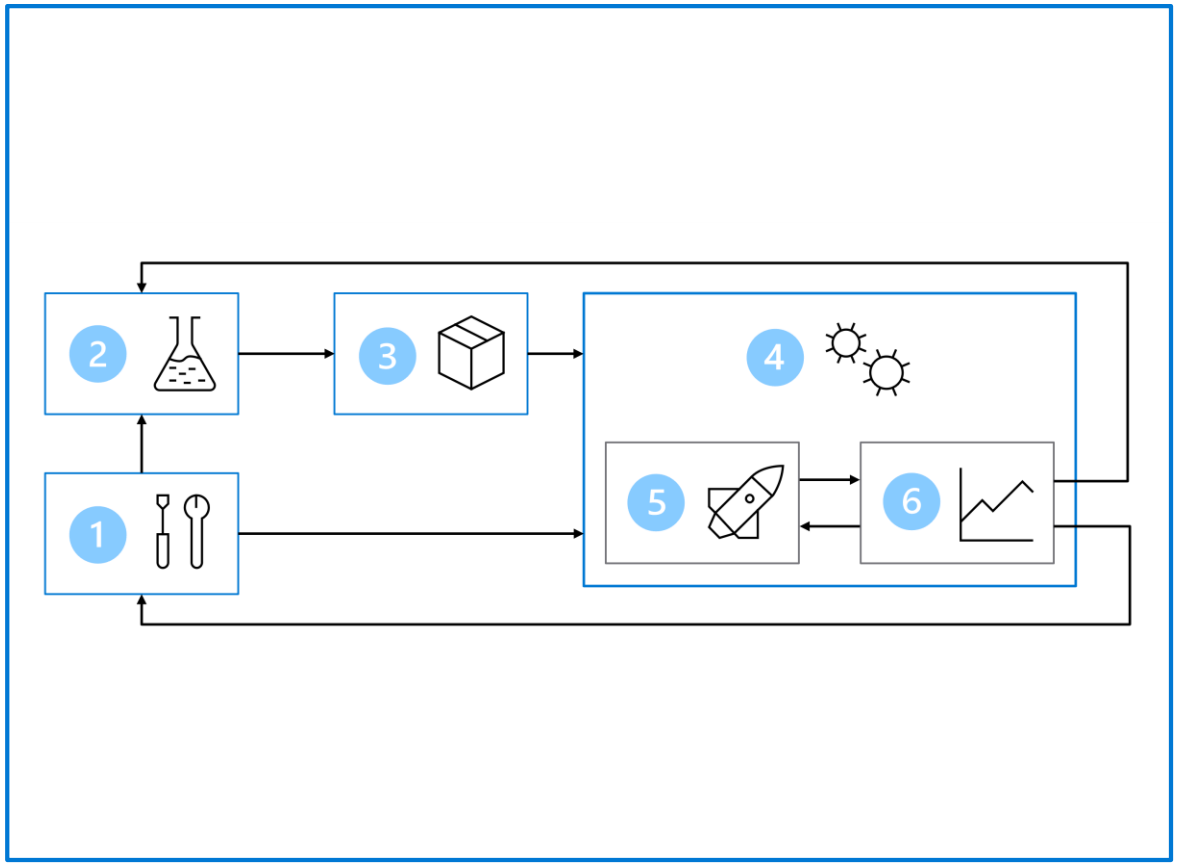
These resources are used to deploy an application, or with machine learning projects, to deploy a model. In an MLOps project, they refer to the Azure resources needed for a phase in the project.



Design an MLOps architecture

Together with other data scientists, data engineers and an infrastructure team, you may decide on using the following approach:

1. **Setup**: Create all necessary Azure resources for the solution.
2. **Model development (inner loop)**: Explore and process the data to train and evaluate the model.
3. **Continuous integration**: Package and register the model.
4. **Model deployment (outer loop)**: Deploy the model.
5. **Continuous deployment**: Test the model and promote to production environment.
6. **Monitoring**: Monitor model and endpoint performance.



Design for monitoring

Monitor the model:

To monitor a model in production, you can use the trained model to generate predictions on a small subset of new incoming data. By generating the **performance metrics** on that test data, you're able to verify whether the model is still achieving its goal.

Additionally, you may also want to monitor for any **responsible artificial intelligence** (AI) issues.

Monitor the data:

You typically train a machine learning model using a historical dataset that is representative of the new data that your model receives when deployed. However, over time there may be trends that change the profile of the data, making your model less accurate.

This change in data profiles between current and the training data is known as **data drift**, and it can be a significant issue for predictive models used in production.

Monitor the infrastructure

Within the workspace:

Review the compute utilization in the Azure Machine Learning workspace for your compute instance and compute cluster. Understanding how much of your compute you're using allows you to scale up to improve performance or scale down to minimize costs.

Use Azure Monitor:

Collect metrics of your Azure resources and set alerts to get notified when problems may arise.

You can monitor the workspace and the managed compute, the datastores, and the endpoints.

Case study - Design an MLOps solution（1/3）

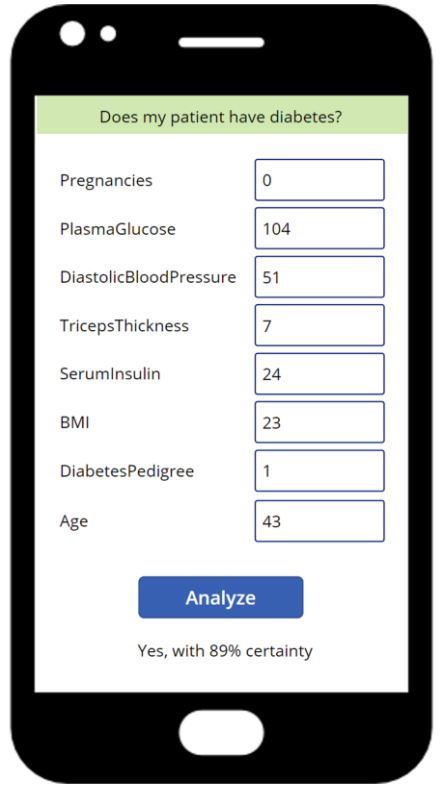
Welcome to Proseware! You've been hired as the lead data scientist to help us design a machine learning deployment solution.

At Proseware, we're developing a mobile application that will help doctors diagnose diseases in patients faster. A doctor can enter the patient's medical data into the app to get a diagnosis on the patient.

Our first planned feature is that the app will tell the doctor whether the patient should be further screened or treated for diabetes.

We need your help deciding how to design **for bringing the model to production**.

We're looking forward to your advice on how to design the machine learning operations (MLOps) solution!



Case study - Design an MLOps solution（2/3）

Consider the requirements:

- **Consider the environments**: Currently, we're working in a small team and you're the only data scientist involved. We want to see whether this project is successful before actually scaling up and getting a large team involved.
- **Consider the model**: As the model is used to help doctors, accuracy is important to us. The model should only be in use when we know it's performing as expected.
- **Consider the data**: We're starting small and will mostly use the deployed model to test our application. The data the deployed model generates predictions on shouldn't be used to retrain the model as it may be biased.

Case study - Design an MLOps solution (3/3)

How many Azure Machine Learning workspaces should the team create?

When should we retrain the model?

Learn more about MLOps

Explore the introduction to machine learning operations (MLOps)
https://learn.microsoft.com/training/paths/introduction-machine-learn-operations

Build your first MLOps automation pipeline with GitHub Actions
https://learn.microsoft.com/training/paths/build-first-machine-operations-workflow

Microsoft Azure