Analytics Vidhya

# Introduction to LLMOps


Instructor

Kartik Nighania

MLOps Engineer at Typewise

Ex-Head of Engineering - YC-backed startup

Certified AWS Cloud and Kubernetes Engineer

Open source contributor

# Introduction

## As a GenAI Scientist / Specialist / Developer we want to:

- work with multiple foundation models
- do prompt engineering
- if needed, train on high-quality dataset
- if needed, access internal data with RAG

# Introduction

## As an AI/ML engineer, we aim to:

- work with multiple LLM APIs
- multi-GPU training
- manage vector databases
- deploy and scale ML models
- track and monitor all models

# Challenges in Managing LLM Projects

- Maintaining a single LLM project with multiple experiments can be challenging
- Scaling to numerous projects can create a bottleneck

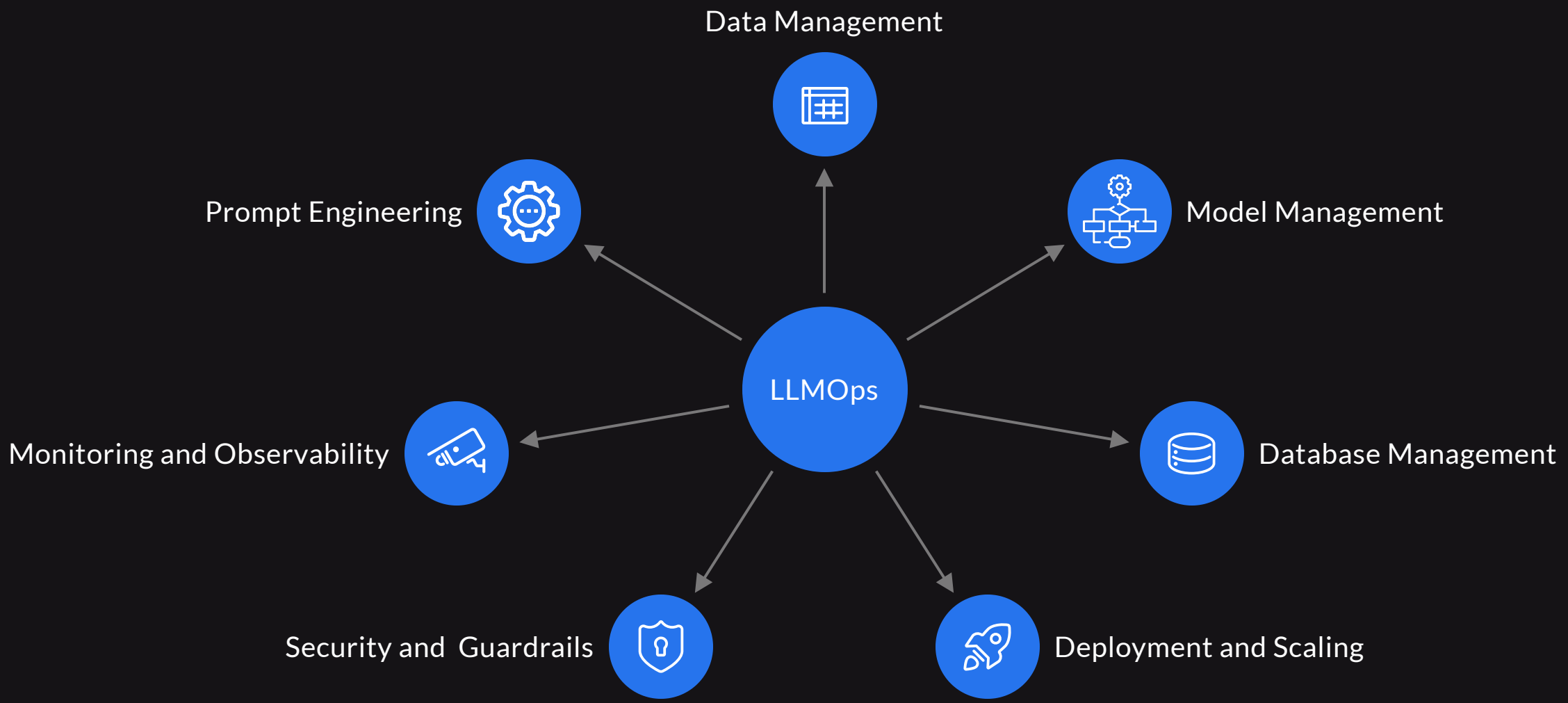
# Why LLMOps?

- Processes can be automated using LLMOps, thus preventing repetitive tasks and human errors.
- Security and other best practices can be incorporated into the automated system

# LLMOps

Practices and processes involved in managing and operating large language models (LLMs)

# Major LLMOps components



# Course Overview

SageMaker Notebook

LangChain

Langfuse

SageMaker Studio

Kubernetes

Load Testing (Locust)

# Thank You

Analytics Vidhya