General AI/ML

Unit 4: Productionizing with Docker

4.1.2

Model Deployment and Integration

On-Device Deployment & Edge-Computing

Introduction

- **On-device** refers to the capability of running AI algorithms and processing data directly on an end-user device, without needing to connect to a server or cloud

- **Edge computing** is a distributed computing paradigm that pushes computation and data storage away from centralized data centers and brings them closer to the network's periphery or "edge"

- Synergy between on-device AI and edge computing enables (near) real-time processing and decision-making at the data source

Importance of On-device Deployment

- Speed: Processing data directly on the device drastically reduces the need for internet bandwidth and cloud computing resources, leading to faster response times
- Privacy: On-device AI processes data locally, ensuring sensitive information does not leave the device, which is crucial for compliance with data protection laws
- Reliability: By operating independently of the cloud, on-device AI ensures functionality even in areas with poor internet connectivity
- Cost: Reduced data transfer costs over time

Challenges of On-device Deployment

- Hardware limitations: Device power and memory constraints pose a challenge to running complex AI models

- Model optimization: Converting large AI models for efficient on-device execution

- Energy consumption: Balancing performance with battery life

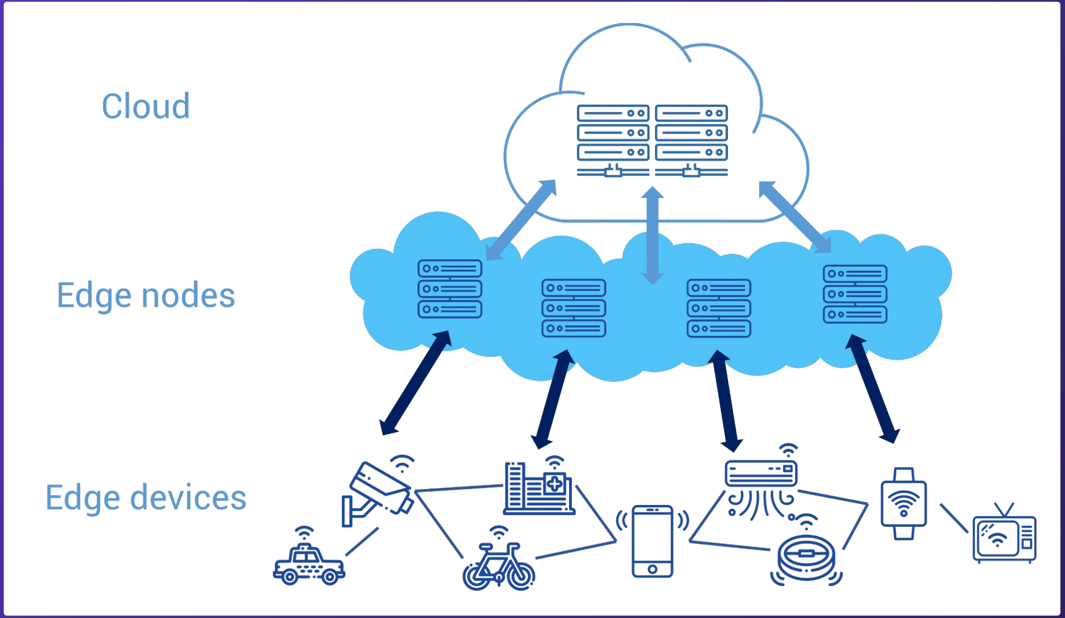
- Development complexity: Managing models and updates across diverse devices

- Security concerns: Protecting devices from cyber threats becomes increasingly complex as more devices process sensitive data

Edge Computing

- A distributed computing architecture that brings computation and data storage closer to the source of data

- Facilitates faster decision-making by analyzing data at its source, reducing the latency typically associated with cloud computing.



TIL-AI
TODAY I LEARNED-AI

Use-cases of On-device & Edge Computing

- Smartphones: Scene and object recognition; image enhancement
- Augmented reality: Immersive AR experiences without latency
- Smart home devices: Voice command recognition
- Wearables: Monitoring heart rate, steps, sleep patterns in real time
- Smart cities: Localized processing of data from sensors and cameras to optimize traffic flow
- Autonomous Vehicles: Real-time decisions about navigation and safety