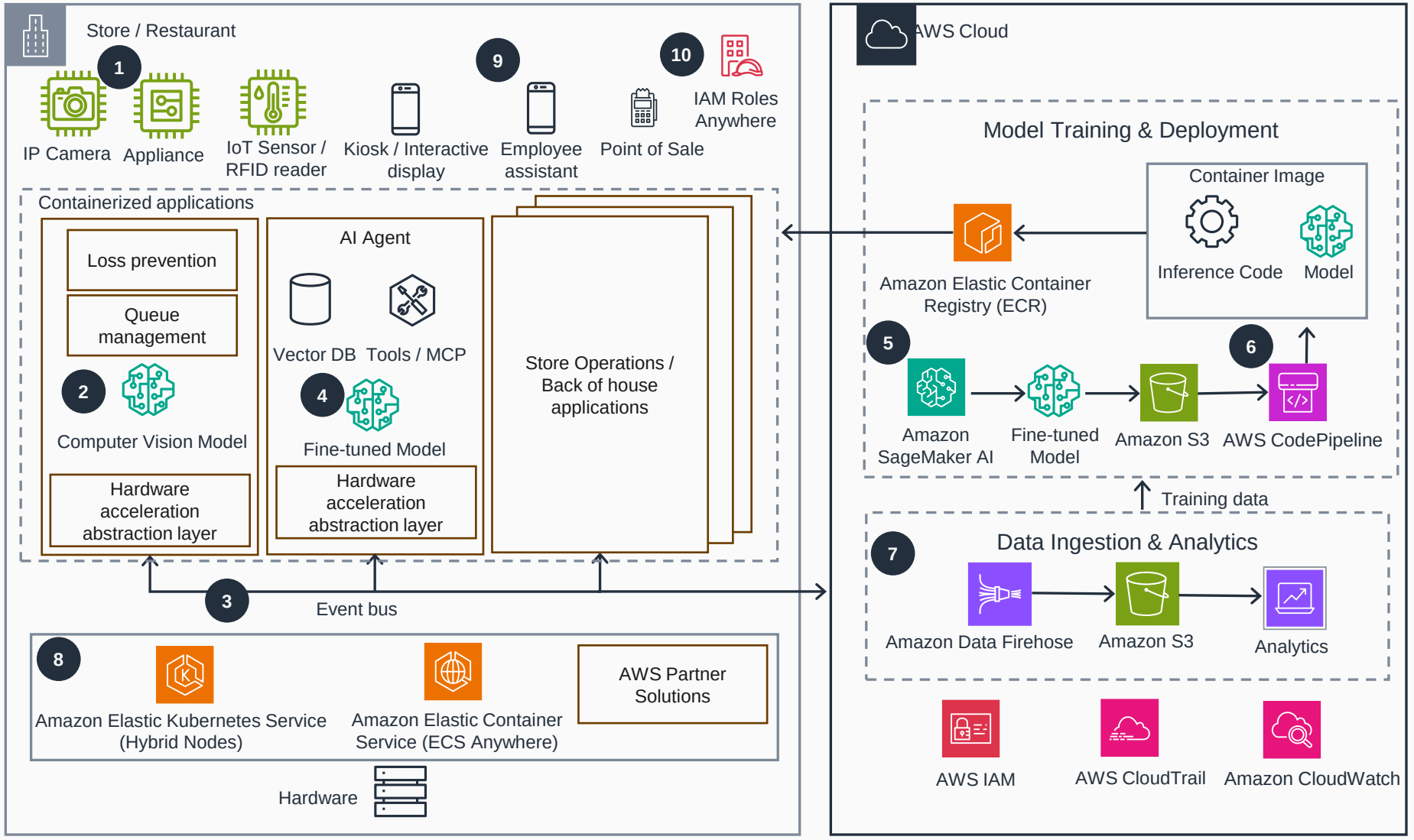
Guidance for AI at the edge for Retail on AWS

This architecture shows how to fine-tune lightweight models in the cloud for retail use cases—such as AI agents that guide customers or detect safety issues—and deploy them to on-premises retail facilities that don't require high-end, GPU-enabled hardware.



1. IoT sensors, IP cameras, RFID readers, point-of-sale systems, mobile clients, and appliances collect raw data in-store

2. Lightweight computer vision models detect high-traffic areas, safety issues, or long customer queues. Containers running inference code and AI models are deployed to retail locations.

3. Alerts and application events are shared between applications via an on-premises event bus. Common choices include **AWS IoT Core**, Apache Kafka or Redis.

4. In-store AI Agents use fine-tuned foundation models, vector databases for Retrieval Augmented Generation, Model Context Protocol (MCP) servers tool implementation, and the Strands SDK.

5. **Amazon SageMaker AI** is used to prepare training data and fine-tune / distill lightweight models for use in stores. Model artifacts are stored in **Amazon Simple Storage Service (S3)**

6. Model artifacts and inference code are combined in container images using **AWS CodePipeline**. Container images are pushed to **Amazon Elastic Container Registry** for later deployment.

7. User feedback and application data are ingested from the event but (step 3) into **Amazon Simple Storage Service (S3)** using **Amazon Data Firehose.** This data is used for training the lightweight models.

8. **Amazon Elastic Kubernetes Service (EKS), Amazon Elastic Container Service (ECS),** or one of several AWS partner solutions automate deployment and lifecycle management for container applications running in retail locations.

9. Customers and staff interact with AI agents via kiosks, interactive displays, or mobile applications.

10. In-store workloads are configured using **AWS IAM Anywhere,** allowing secure access to AWS cloud-based services like **Amazon CloudWatch** and **AWS CloudTrail** for logging and alerts.

Reviewed for technical accuracy August 21, 2025

© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Reference Architecture