Architectural Decision Record (ADR) for AWS AI Model Deployment for Fish Watch System

Decision Summary

Utilize the architectural decisions for deploying AI models on AWS for fish watch system.

Context

The Fish Watch System, our real-time monitoring and analytics platform for fisheries management, requires the integration of AI models for various tasks such as fish species identification, behavior analysis, and anomaly detection. We need to decide on the approach for deploying these AI models within the AWS cloud environment.

Decision:

We have decided to deploy the AI models for the Fish Watch System on AWS using Amazon SageMaker, a fully managed service for building, training, and deploying machine learning models at scale.

Rationale:

Several factors influenced this decision:

1. **Managed Service**: Amazon SageMaker is a fully managed service that simplifies the end-to-end machine learning workflow, from data preparation and model training to deployment and monitoring. It abstracts away the complexities of infrastructure provisioning and management, allowing our team to focus on model development and experimentation.

2. **Scalability and Performance**: Amazon SageMaker is designed for scalability and high performance, enabling us to train and deploy AI models efficiently, even with large datasets and complex algorithms. It leverages AWS's underlying infrastructure to deliver low-latency predictions and handle varying workloads effectively.

3. **Built-in Algorithms and Frameworks**: Amazon SageMaker provides built-in algorithms and pre-configured environments for popular machine learning frameworks such as TensorFlow, PyTorch, and scikit-learn. This accelerates model development and experimentation, allowing us to leverage state-of-the-art algorithms and techniques for fisheries monitoring and analysis.

4. **Custom Model Training**: While Amazon SageMaker offers built-in algorithms, it also supports custom model training using custom containers or bring-your-own-algorithm

(BYOA) capabilities. This flexibility allows us to train custom AI models tailored to the specific requirements and domain expertise of the Fish Watch System.

5. **Deployment Options:** Amazon SageMaker offers multiple deployment options, including real-time inference endpoints for low-latency predictions, batch transform for offline processing, and edge deployment for running models on edge devices or IoT devices. This enables us to choose the most suitable deployment strategy based on our use case and performance requirements.

6. **Integration with AWS Ecosystem:** Amazon SageMaker seamlessly integrates with other AWS services, such as Amazon S3 for data storage, AWS Lambda for serverless computing, Amazon CloudWatch for monitoring, and AWS Identity and Access Management (IAM) for access control. This tight integration simplifies the development, deployment, and management of AI models within the AWS cloud environment.

7. **Cost-Effectiveness:** Amazon SageMaker offers pay-as-you-go pricing with no upfront costs, allowing us to scale our AI model deployments based on demand and optimize costs effectively. We can leverage cost-saving strategies such as spot instances for training and auto-scaling for inference endpoints to minimize expenses while maximizing resource.

Consequences:

By deploying AI models for the Fish Watch System on Amazon SageMaker, we anticipate the following consequences:

1. **Development Efficiency:** Using Amazon SageMaker streamlines the machine learning workflow, enabling our data scientists and developers to iterate quickly, experiment with different models, and deploy them into production faster.
2. **Operational Overhead:** While Amazon SageMaker abstracts away much of the infrastructure management, there may still be operational overhead associated with monitoring, optimizing, and maintaining the deployed AI models. We need to establish processes and best practices for model governance, versioning, and performance monitoring to ensure the reliability and effectiveness of the deployed models.
3. **Data Security and Privacy:** We must implement appropriate data security and privacy measures to protect sensitive data used for training and inference, ensuring compliance with regulatory requirements, and safeguarding the privacy rights of individuals. This includes encrypting data at rest and in transit, implementing access controls, and anonymizing or de-identifying sensitive information.
4. **Model Interpretability and Explainability:** It's essential to consider the interpretability and explainability of AI models deployed within the Fish Watch System, especially for decision-making in critical applications such as fisheries management. We need to employ techniques and tools for model interpretability, such as feature importance analysis and

model explainability frameworks, to enhance transparency and trust in the deployed models.

In summary, deploying AI models for the Fish Watch System on Amazon SageMaker offers a scalable, efficient, and cost-effective solution for integrating machine learning capabilities into our fisheries monitoring and analytics platform. With proper planning, execution, and governance, we believe Amazon SageMaker will enable us to derive valuable insights and make informed decisions to support sustainable fisheries management and marine conservation efforts.