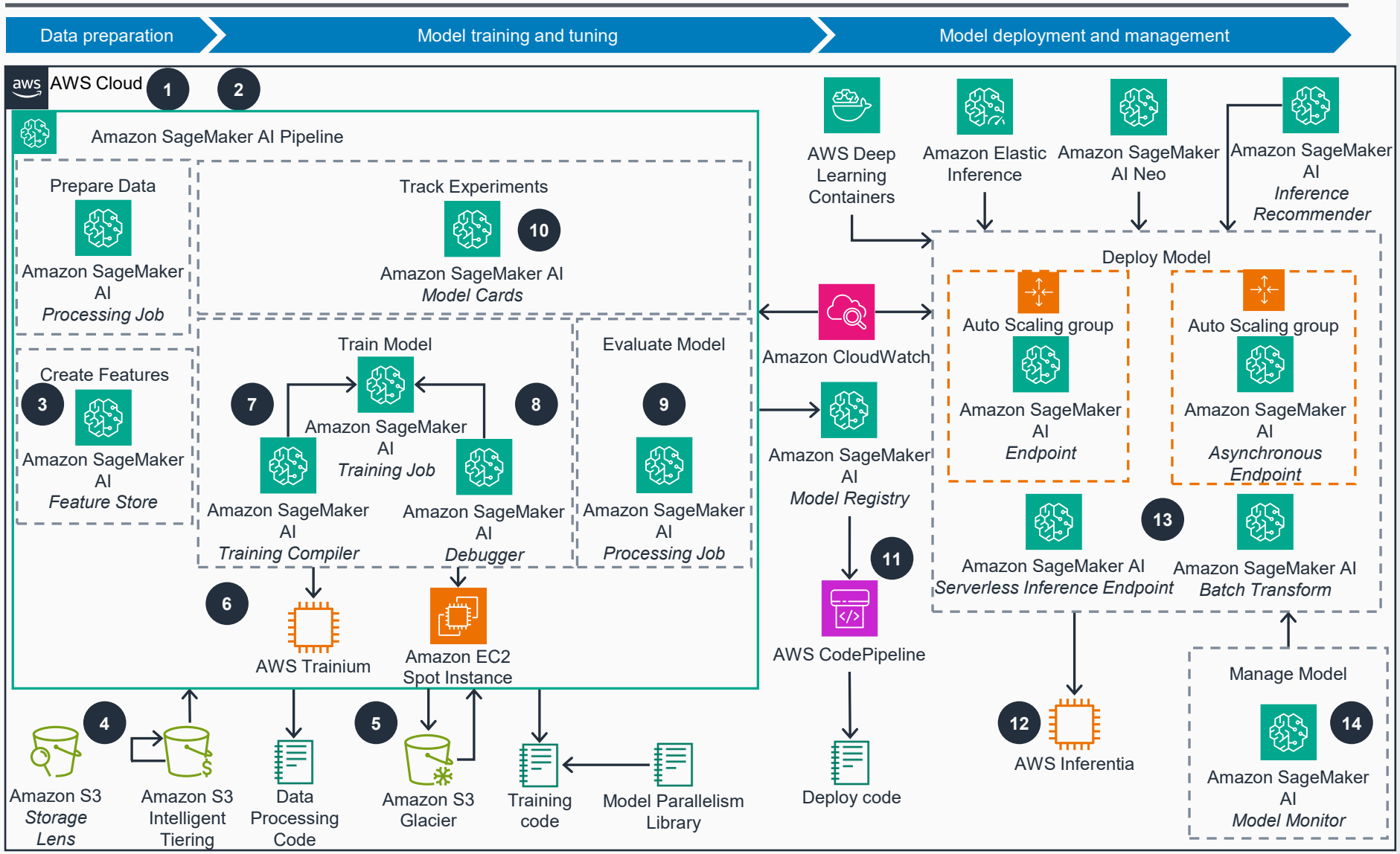
Guidance for Optimizing Machine Learning Operations for Sustainability on AWS

This architecture diagram illustrates how to reduce the environmental impact of an MLOps platform.



1 **Strategic Region Selection** optimizes performance and sustainability, as different AWS Regions have varying carbon footprints based on their energy sources.

- Use the AWS Customer Carbon Footprint Tool to evaluate regional carbon impacts.
- Consider regions powered by AWS renewable energy projects.
- Implement data locality principles by co-locating data processing with storage and minimizing cross-region data transfer.

2 **Serverless Architecture** optimizes resource utilization and reduces idle computing capacity.

- Deploy Amazon SageMaker Pipelines for automated ML workflows.
- Use SageMaker Projects for MLOps automation.
- Implement SageMaker Model Monitor for production monitoring.
- Consider SageMaker Serverless Inference for auto-scaling endpoints.

3 **Reduce Duplication** and rerun of feature engineering code across teams and projects by using Amazon SageMaker Feature Store.

4 **Optimizing Storage Strategy** reduces the environmental impact of cloud workloads.

- Through intelligent use of Amazon Simple Storage Service (Amazon S3) Storage Classes, organizations can significantly reduce their carbon footprint while maintaining operational efficiency.
- Amazon S3 Intelligent-Tiering automatically moves data between access tiers based on usage patterns.
- Organizations can further optimize their storage footprint using Amazon S3 Storage Lens to identify opportunities for improved storage efficiency and reduced environmental impact.

5 **Data Storage Archives**

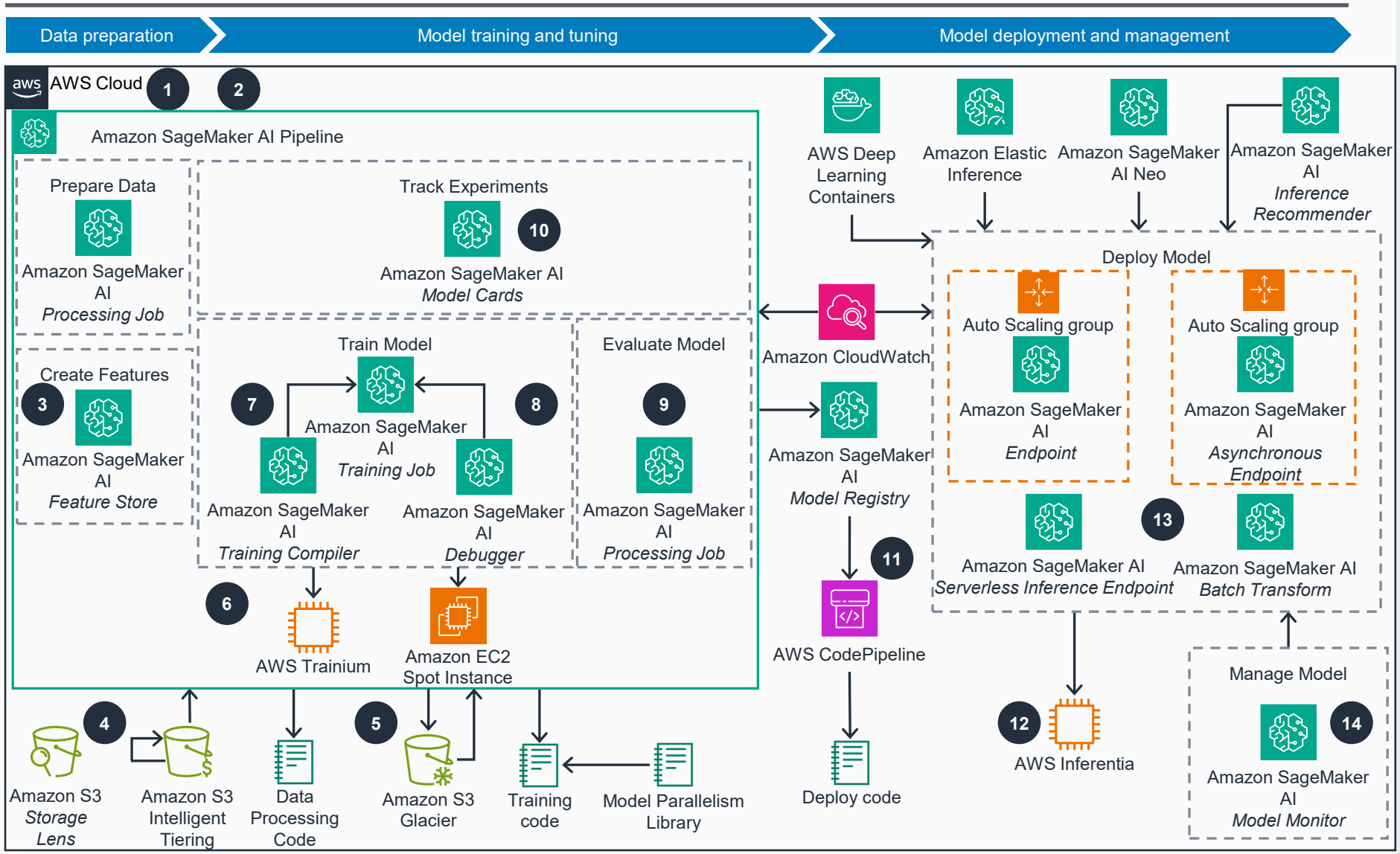
- Amazon S3 Glacier storage classes provide sustainable storage for rarely accessed data.
- Implementing Amazon S3 Lifecycle policies ensures automated data management through proper retention and deletion schedules, reducing unnecessary storage consumption.

Reviewed for technical accuracy December 4, 2025

AWS Reference Architecture

Guidance for Optimizing Machine Learning Operations for Sustainability on AWS

This architecture diagram illustrates how to reduce the environmental impact of an MLOps platform.



6 **Hardware Optimization** maximizes computational efficiency while minimizing energy consumption.

- Leveraging purpose-built infrastructure like AWS Trainium, which offers up to 52% lower energy consumption compared to traditional **Amazon Elastic Compute Cloud (Amazon EC2)** instances, organizations can significantly reduce their carbon footprint during model training.
- Managed Spot Training in **Amazon SageMaker AI** takes advantage of unused **Amazon EC2** capacity, further enhancing resource efficiency and reducing idle infrastructure, making it a crucial component in sustainable ML practices.

7 **Model Training** using Amazon SageMaker AI Model Parallelism enables efficient distribution of large models across multiple GPUs, optimizing resource utilization.

8 **Resource Optimization** provides critical insights into resource utilization and opportunities for environmental impact reduction.

- Amazon SageMaker AI Debugger plays a pivotal role by automatically detecting resource underutilization and training inefficiencies, enabling real-time interventions that prevent waste.
- Integration with Amazon CloudWatch provides comprehensive metrics for right-sizing training jobs and optimizing resource allocation.

9 **Training Optimization Strategies** minimize the environmental impact of machine learning.

- Amazon SageMaker AI Automatic Model Tuning with Bayesian optimization significantly reduces the number of experimental training runs.
- Utilizing Amazon SageMaker AI Processing for efficient model evaluation and implementing systematic performance criteria, organizations can make informed trade-offs between model accuracy and carbon footprint.

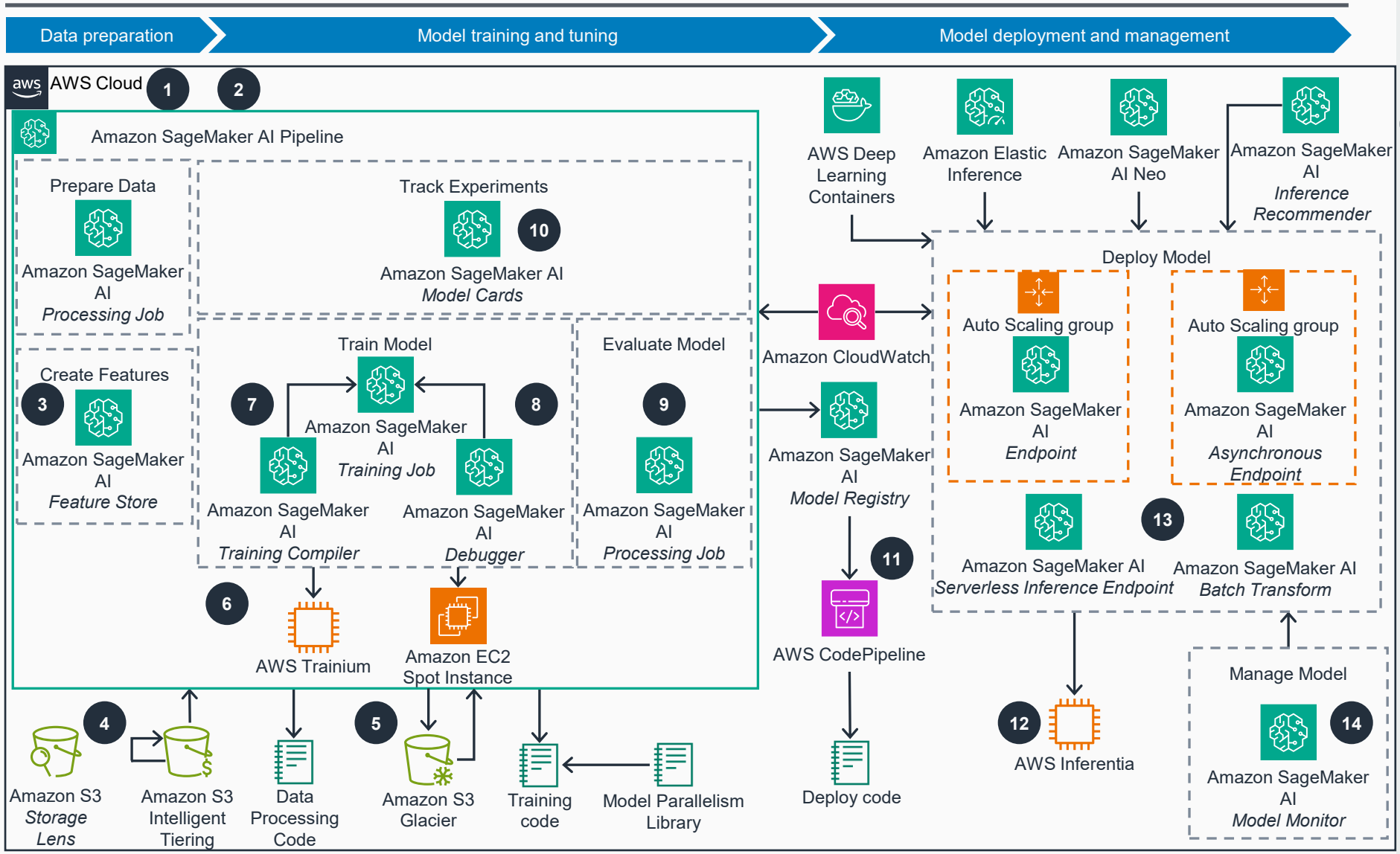
10 **Documentation** through Amazon SageMaker AI Model Cards enables tracking of environmental impact metrics, promoting transparency and accountability in sustainable ML practices.

Reviewed for technical accuracy December 4, 2025

AWS Reference Architecture

Guidance for Optimizing Machine Learning Operations for Sustainability on AWS

This architecture diagram illustrates how to reduce the environmental impact of an MLOps platform.



11 **Automated Deployment Infrastructure** optimizes resource utilization and reduces manual intervention.

- By implementing Amazon SageMaker AI Model Registry alongside AWS CodePipeline and Amazon SageMaker AI Pipelines, organizations can create efficient, repeatable deployment processes that minimize resource waste and operational overhead.

12 **Energy-efficient Deployment** options reduce the environmental impact of machine learning inference workloads.

- AWS designed AWS Inferentia chips to deliver high performance at the lowest cost in **Amazon EC2** for deep learning (DL) and generative AI inference applications.

13 **Scalable Endpoint Solutions** optimize resource utilization and minimize environmental impact in machine learning deployments.

- Amazon SageMaker AI Serverless Inference automatically manages compute resources based on workload demands, eliminating idle resource waste for intermittent traffic patterns.
- Amazon Asynchronous Endpoints optimize resource usage for latency-tolerant applications.
- For batch processing needs, Amazon SageMaker AI Batch Transform provides resource-efficient inference by automatically decommissioning clusters upon job completion.

14 **Production Monitoring** ensures optimal resource utilization and model performance over time.

- Amazon SageMaker AI Model Monitor provides comprehensive monitoring capabilities that detect model drift, assess data quality, and track resource utilization, enabling organizations to make data-driven decisions about model retraining and resource allocation.

Reviewed for technical accuracy December 4, 2025

AWS Reference Architecture