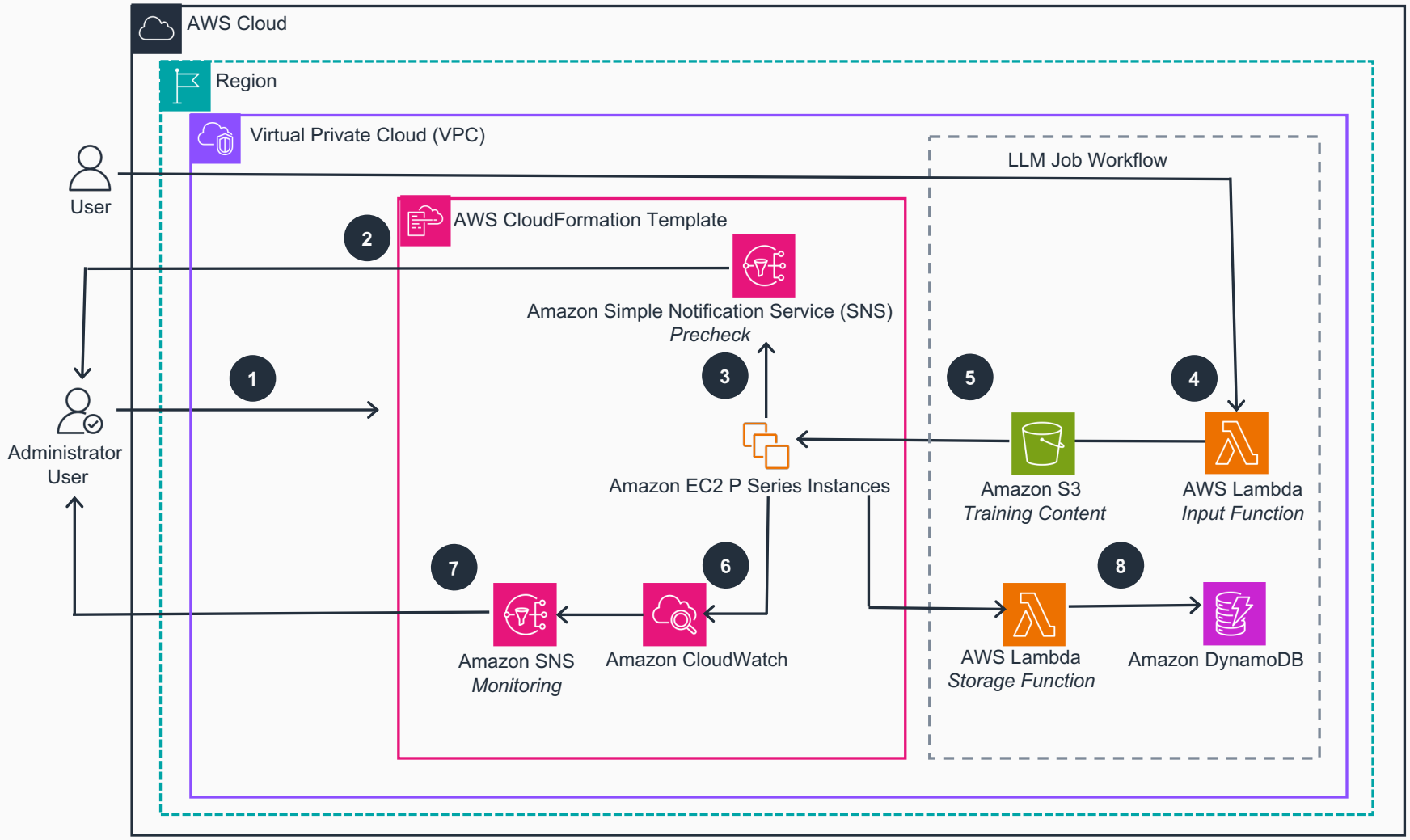
Guidance for LLM Training Operations on AWS

This Guidance helps organizations optimize their LLM training operations through a comprehensive pre-flight testing and monitoring framework. By implementing specialized validation checks and continuous performance monitoring, it reduces costly training interruptions and ensures optimal resource utilization across distributed computing environments.



1. Administrator deploys **AWS CloudFormation** template with custom settings for CPU, memory, and disk thresholds, along with email address for **Amazon Simple Notification Services (SNS) Topics** notifications
2. **AWS CloudFormation** template creates **Amazon EC2** instances and executes pre-flight checks validating GPU health, CUDA drivers, NCCL testing, EFA connectivity, and CPU/memory/disk performance, while configuring security groups, permissions, CloudWatch agent, and notification channels
3. **Amazon SNS** topic sends a notifications for any failed pre-flight checks to administrator with specific failure details
4. User Initiates the LLM training job by invoking the **Amazon Lambda** Input function for fetching the training data stored in **Amazon S3**
5. **Amazon EC2** instances loads data from **Amazon S3** bucket and runs the LLM training process
6. Monitor system health continuously through **Amazon CloudWatch** by tracking real-time CPU usage against defined thresholds, memory consumption, disk space utilization and I/O performance, network throughput and connectivity status, plus GPU utilization and temperature for ML workloads
7. Send runtime monitoring alerts through **Amazon SNS** when operational thresholds are exceeded during training, including specific triggering metrics and current system status in each notification
8. **Amazon Lambda** Storage Function stores the queryable training metadata in the **Amazon DynamoDB**

Reviewed for technical accuracy Aug 6, 2025

AWS Reference Architecture