aws

Architecture Diagrams

Multi-Model Inference Workflow Orchestration

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Multi-Model Inference Workflow Orchestration: Architecture Diagrams

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Amazon's trademarks and trade dress may not be used in connection with any product or service that is not Amazon's, in any manner that is likely to cause confusion among customers, or in any manner that disparages or discredits Amazon. All other trademarks not owned by Amazon are the property of their respective owners, who may or may not be affiliated with, connected to, or sponsored by Amazon.

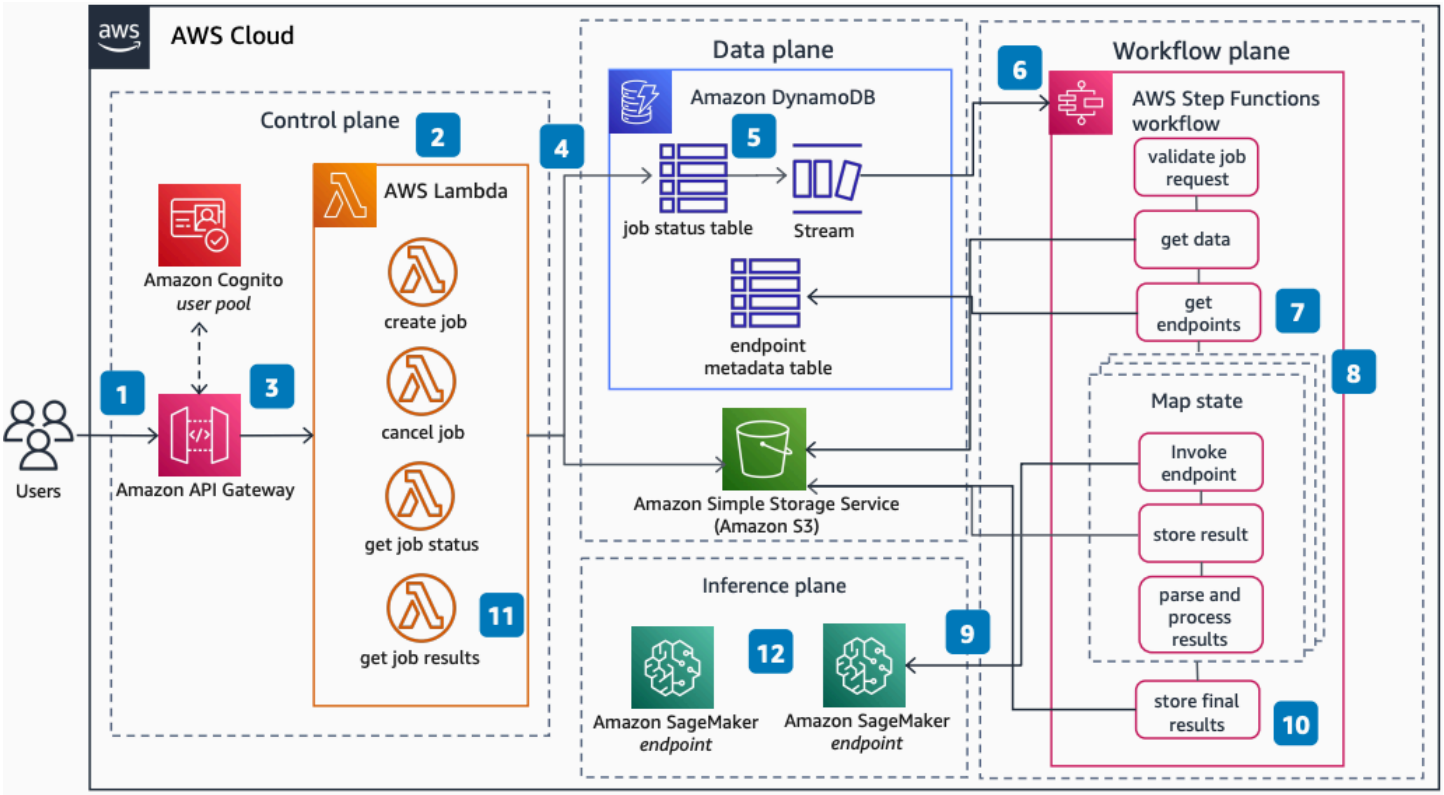
Table of Contents

Multi-Model Inference Workflow Orchestration

Publication date: **March 28, 2022 (Diagram history)**

This architecture shows how to orchestrate running multiple ML models for complex ML-driven insight. With AWS Step Functions and Amazon SageMaker AI, you can run parallel inference jobs and aggregate results.

Multi-Model Inference Workflow Orchestration



The following steps describe the architecture:

1. Amazon API Gateway provides a RESTful interface for users and administrators. Amazon Cognito user pools provide authentication.
2. A new job is created by a `POST` request to `/create-job` on the API, with the data to run insights on. This invokes the create job AWS Lambda function.
3. The function uploads the data to an Amazon Simple Storage Service bucket and adds job information to an Amazon DynamoDB table for tracking.

4. Lambda functions provide logic for API methods exposed through API Gateway. These allow jobs to be created, monitored, and facilitate retrieval of results.
5. The new item in the table triggers a DynamoDB stream which in turn triggers the Step Functions workflow for inference.
6. The Step Functions workflow orchestrates all steps required to run multiple ML inference jobs against the provided data object.
7. Metadata information about the ML endpoints is stored in a DynamoDB table for use in the workflow.
8. A map state in the step function calls each ML endpoint and stores the results in parallel. This allows the workflow to scale to any number of ML endpoint invocations.
9. SageMaker AI model endpoints host the ML models. The workflow invokes these endpoints per category.
10. The final results of all the ML insights gathered for the data are stored in Amazon S3.
11. After the workflow completes, you can retrieve the final results through a `GET` request to `/get-job-results`. This invokes the get job results Lambda function, which reads the Amazon S3 bucket.
12. Multi-model endpoints can extend each model type with extra optimizations, such as language-specific inference per category.

Further reading

For additional information, refer to the following resources:

- AWS Architecture Icons
- AWS Architecture Center
- AWS Well-Architected
- Amazon SageMaker AI product page
- AWS Step Functions product page

Diagram history

To be notified about updates to this reference architecture diagram, subscribe to the RSS feed.

Change	Description	Date
Initial publication	Reference architecture diagram first published.	March 28, 2022

Note

To subscribe to RSS updates, you must have an RSS plugin enabled for the browser you are using.