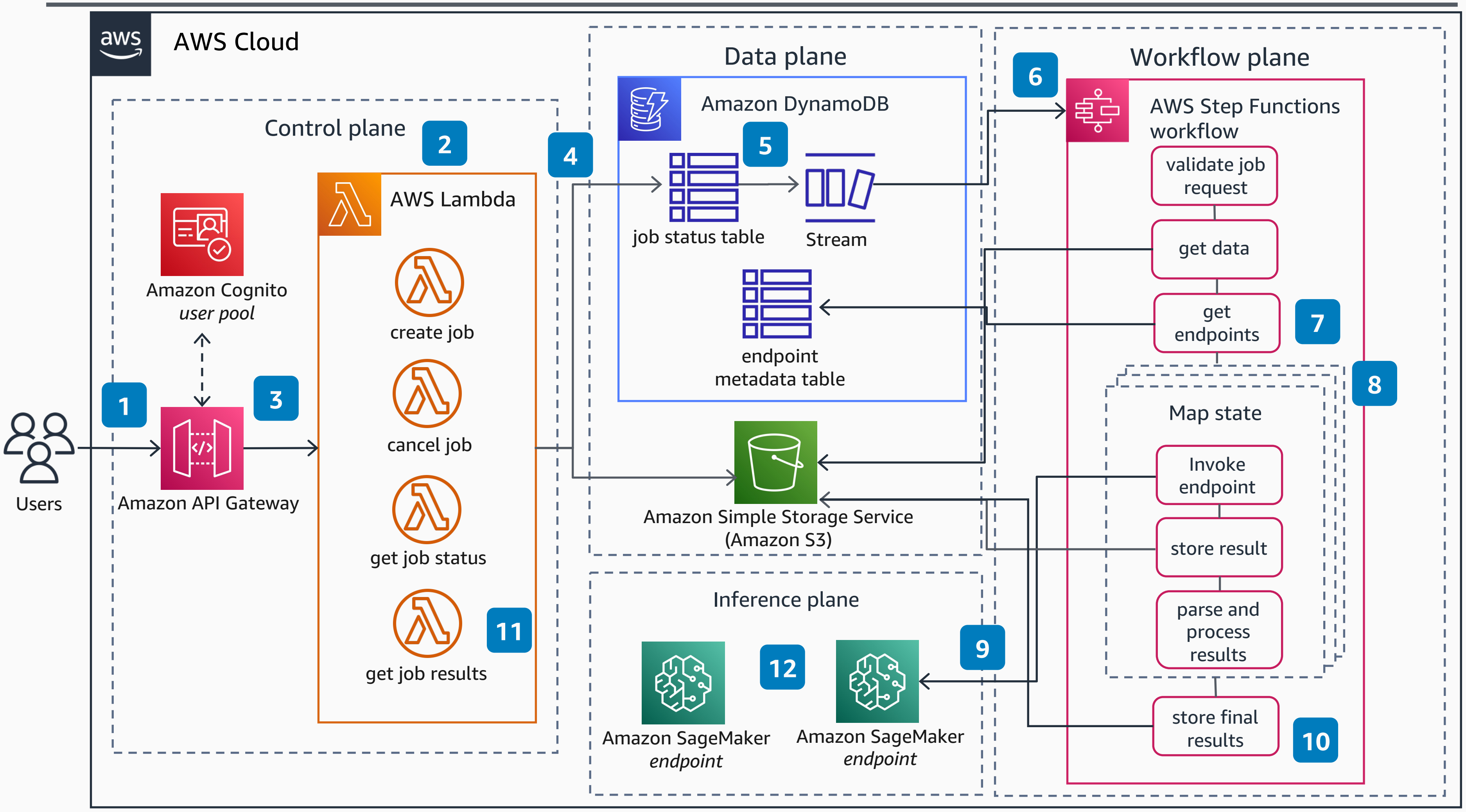
Multi-Model Inference Workflow Orchestration

This architecture helps you orchestrate running multiple machine learning (ML) models for complex ML-driven insight. Use this architecture for multi-category labelling scenarios per object.



1 **Amazon API Gateway** provides a RESTful interface for users and administrators. Authentication is provided by **Amazon Cognito** user pools.

2 **AWS Lambda** functions provide logic for API methods exposed via **Amazon API Gateway**. These allow jobs to be created and monitored, and facilitate the retrieval of results.

3 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* **Lambda** function.

4 The function uploads the data to an **Amazon S3** bucket, and adds job information to an **Amazon DynamoDB** table for tracking.

5 The new item in the table triggers a **DynamoDB** stream which in turn triggers the **AWS Step Functions** workflow for inference.

6 The **AWS Step Functions** workflow orchestrates all the steps required 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 is used in the step function to call each ML endpoint and store the results in parallel, allowing the workflow to scale to any number of ML endpoint invocations.

9 **Amazon SageMaker** model endpoints are used to host the ML models. These are invoked by the workflow per category.

10 The final results of all the ML insights gathered for the data are stored for future use in **Amazon S3**.

11 Once the workflow is complete, the final results can be retrieved by users through a GET request to /get-job-results. This invokes the *get job Results* **Lambda** function, which reads the **S3** bucket and retrieves the inference results.

12 Multi-model endpoints can be used to extend each model type with extra optimizations, such language-specific inference per category.

Reviewed for technical accuracy March 28, 2022

AWS Reference Architecture