STREAMLINING MACHINE LEARNING WORKFLOWS WITH CONTROL-M AND AMAZON SAGEMAKER



In today's fast-paced digital landscape, the ability to harness the power of artificial intelligence (AI) and machine learning (ML) is crucial for businesses aiming to gain a competitive edge. Amazon SageMaker is a game-changing ML platform that empowers businesses and data scientists to seamlessly navigate the development of complex AI models. One of its standout features is its end-to-end ML pipeline, which streamlines the entire process from data preparation to model deployment. Amazon SageMaker's integrated Jupyter Notebook platform enables collaborative and interactive model development, while its data labeling service simplifies the often-labor-intensive task of data annotation.

It also boasts an extensive library of pre-built algorithms and deep learning frameworks, making it accessible to both newcomers and experienced ML practitioners. Amazon SageMaker's managed training and inference capabilities provide the scalability and elasticity needed for real-world AI deployments. Moreover, its automatic model tuning, and robust monitoring tools enhance the efficiency and reliability of AI models, ensuring they remain accurate and up-to-date over time. Overall, Amazon SageMaker offers a comprehensive, scalable, and user-friendly ML environment, making it a top choice for organizations looking to leverage the potential of AI.

Bringing Amazon SageMaker and Control-M together

Amazon SageMaker simplifies the entire ML workflow, making it accessible to a broader range of users, including data scientists and developers. It provides a unified platform for building, training, and deploying ML models. However, to truly harness the power of Amazon SageMaker, businesses

often require the ability to orchestrate and automate ML workflows and integrate them seamlessly with other business processes. This is where Control-M from BMC comes into play.

Control-M is a versatile application and data workflow orchestration platform that allows organizations to automate, monitor, and manage their data and AI-related processes efficiently. It can seamlessly integrate with SageMaker to create a bridge between AI modeling and deployment and business operations.

In this blog, we'll explore the seamless integration between Amazon SageMaker and Control-M and the transformative impact it can have on businesses.

Amazon SageMaker empowers data scientists and developers to create, train, and deploy ML models across various environments—on-premises, in the cloud, or on edge devices. An end-to-end data pipeline will include more than just Amazon SageMaker's AI and ML functionality, where data gets ingested from multiple sources, transformed, aggregated etc., before training a model and executing AI/ML pipelines with Amazon SageMaker. Control-M is often used for automating and orchestrating end-to-end data pipelines. A good example of end-to-end orchestration is covered in the blog, "Orchestrating a Predictive Maintenance Data Pipeline," co-authored by Amazon Web Services (AWS) and BMC.

Here, we will specifically focus on integrating Amazon SageMaker with Control-M. When you have Amazon SageMaker jobs embedded in your data pipeline or complex workflow orchestrated by Control-M, you can harness the capabilities of Control-M for Amazon SageMaker to efficiently execute an end-to-end data pipeline that it also includes Amazon SageMaker pipelines.

Key capabilities

Control-M for Amazon SageMaker provides:

- **Secure connectivity:** Connect to any Amazon SageMaker endpoint securely, eliminating the need to provide authentication details explicitly
- **Unified scheduling:** Integrate Amazon SageMaker jobs seamlessly with other Control-M jobs within a single scheduling environment, streamlining your workflow management
- **Pipeline execution:** Execute Amazon SageMaker pipelines effortlessly, ensuring that your ML workflows run smoothly
- **Monitoring and SLA management:** Keep a close eye on the status, results, and output of Amazon SageMaker jobs within the Control-M Monitoring domain and attach service level agreement (SLA) jobs to your Amazon SageMaker jobs for precise control
- **Advanced capabilities:** Leverage all Control-M capabilities, including advanced scheduling criteria, complex dependencies, resource pools, lock resources, and variables to orchestrate your ML workflows effectively
- **Parallel execution:** Run up to 50 Amazon SageMaker jobs simultaneously per agent, allowing for efficient job execution at scale

Control-M for Amazon SageMaker compatibility

Before diving into how to set up Control-M for Amazon SageMaker, it's essential to ensure that your environment meets the compatibility requirements:

- Control-M/EM: version 9.0.20.200 or higher

- Control-M/Agent: version 9.0.20.200 or higher
- Control-M Application Integrator: version 9.0.20.200 or higher
- Control-M Web: version 9.0.20.200 or higher
- Control-M Automation API: version 9.0.20.250 or higher

Please ensure you have the required installation files for each prerequisite available.

A real-world example:

The Abalone Dataset, sourced from the UCI Machine Learning Repository, has been frequently used in ML examples and tutorials to predict the age of abalones based on various attributes such as size, weight, and gender. The age of abalones is usually determined through a physical examination of their shells, which can be both tedious and intrusive. However, with ML, we can predict the age with considerable accuracy without resorting to physical examinations.

For this exercise, we used the Abalone tutorial provided by AWS. This tutorial efficiently walks users through the stages of data preprocessing, training, and model evaluation using Amazon SageMaker.

After understanding the tutorial's nuances, we trained the Amazon SageMaker model with the Abalone Dataset, achieving satisfactory accuracy. Further, we created a comprehensive continuous integration and continuous delivery (CI/CD) pipeline that automates model retraining and endpoint updates. This not only streamlined the model deployment process but also ensured that the Amazon SageMaker endpoint for inference was always up-to-date with the latest trained model.

Setting up Control-M for Amazon SageMaker

Now, let's walk through how to set up Control-M for Amazon SageMaker, which has three main steps:

1. Creating a connection profile that Control-M will use to connect to the Amazon SageMaker environment
2. Defining an Amazon SageMaker job in Control-M that will define what we want to run and monitor within Amazon SageMaker
3. Executing an Amazon SageMaker pipeline with Control-M

Step 1: Create a connection profile

To begin, you need to define a connection profile for Amazon SageMaker, which contains the necessary parameters for authentication and communication with SageMaker. Two authentication methods are commonly used, depending on your setup.

Example 1: Authentication with AWS access key and secret

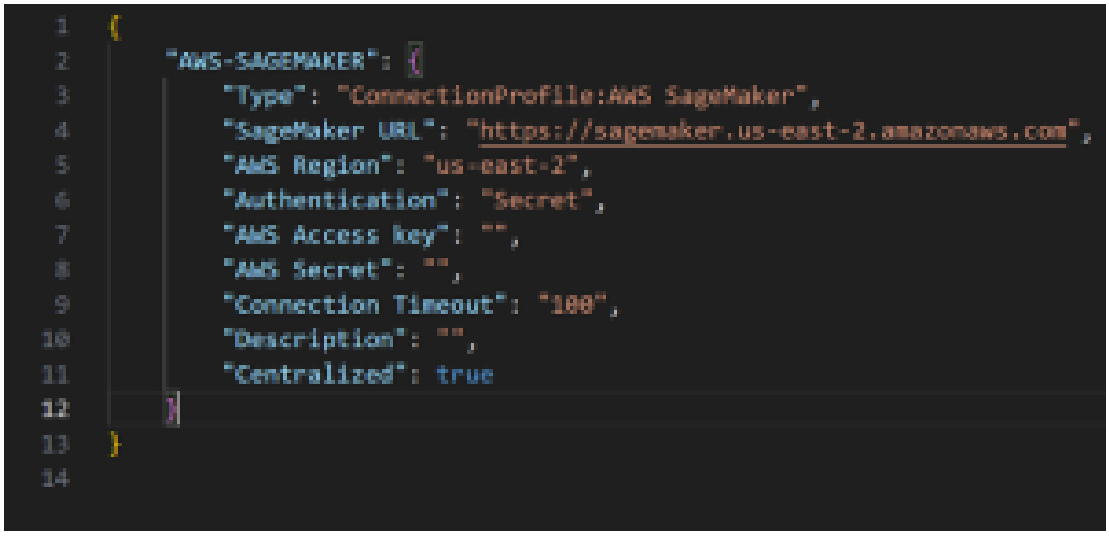


Figure 1. Authentication with AWS access key and secret.

Example 2: Authentication with AWS IAM role from EC2 instance

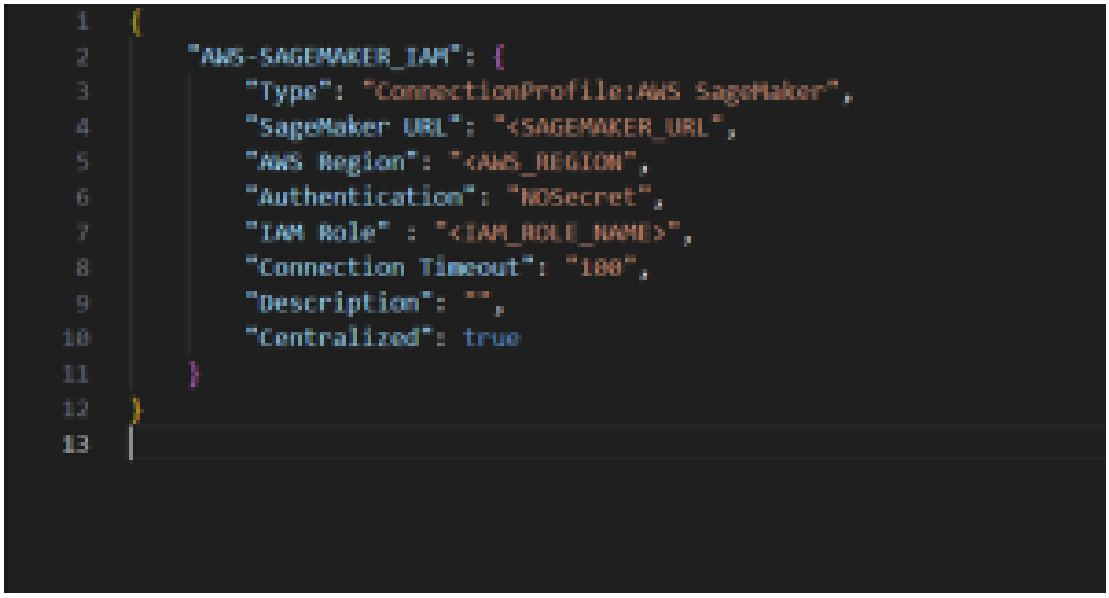
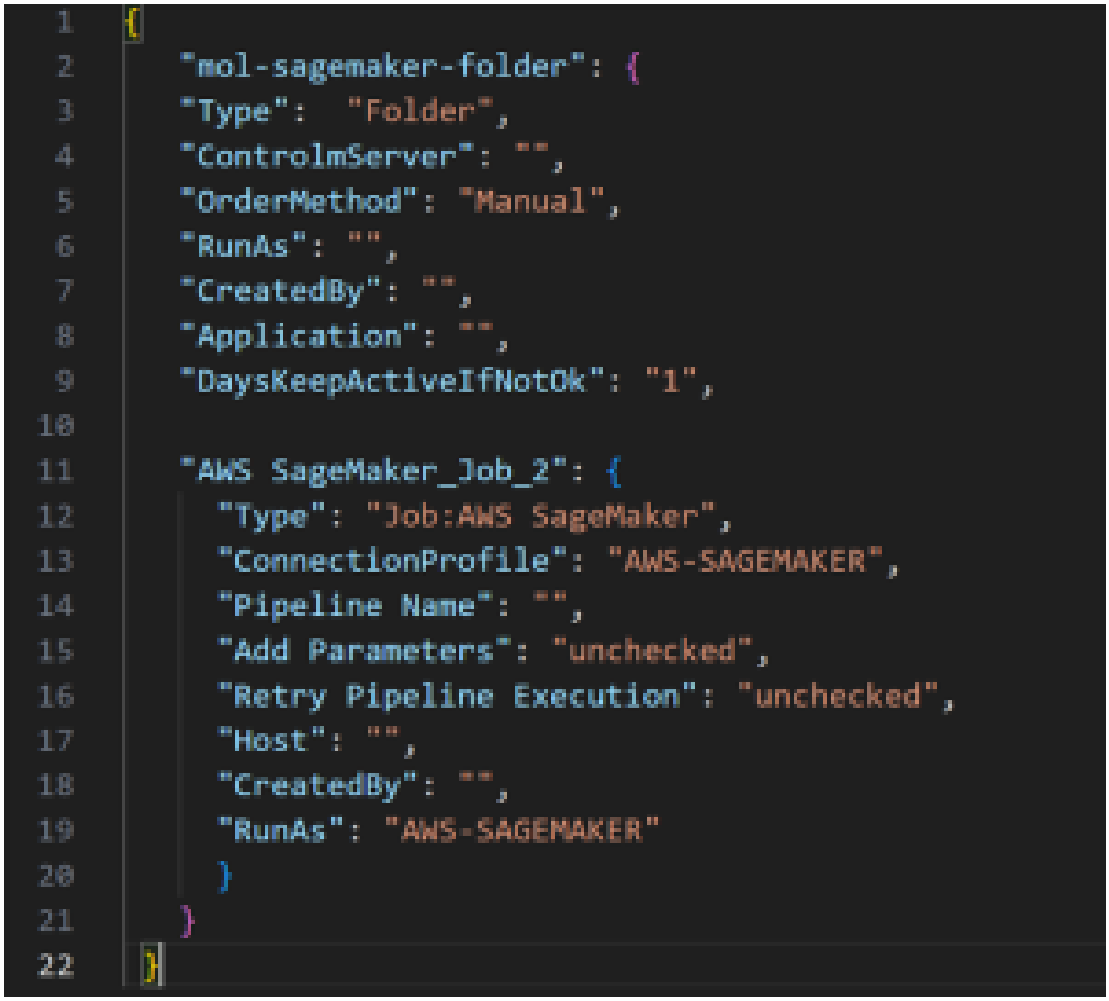


Figure 2. Authentication with AWS IAM role.

Choose the authentication method that aligns with your environment. It is important to specify the Amazon SageMaker job type exactly as shown in the examples above. Please note that Amazon SageMaker is case-sensitive, so make sure to use the correct capitalization.

Step 2: Define an Amazon SageMaker job

Once you've set up the connection profile, you can define an Amazon SageMaker job type within Control-M, which type enables you to execute Amazon SageMaker pipelines effectively.



```
{
    "mol-sagemaker-folder": {
    "Type":  "Folder",
    "ControlmServer": "",
    "OrderMethod": "Manual",
    "RunAs": "",
    "CreatedBy": "",
    "Application": "",
    "DaysKeepActiveIfNotOk": "1",

    "AWS SageMaker_Job_2": {
      "Type": "Job:AWS SageMaker",
      "ConnectionProfile": "AWS-SAGEMAKER",
      "Pipeline Name": "",
      "Add Parameters": "unchecked",
      "Retry Pipeline Execution": "unchecked",
      "Host": "",
      "CreatedBy": "",
      "RunAs": "AWS-SAGEMAKER"
      }
    }
}
```

Figure 3. Example AWS SageMaker job definition.

In this example, we've defined an Amazon SageMaker job, specifying the connection profile to be used ("AWS-SAGEMAKER"). You can configure additional parameters such as the pipeline name, idempotency token, parameters to pass to the job, retry settings, and more. For a detailed understanding and code snippets, please refer to the BMC official documentation for Amazon SageMaker.

Step 3: Executing the Amazon SageMaker pipeline with Control-M

It's essential to note that the pipeline name and endpoint are mandatory JSON objects within the pipeline configuration. By executing the “ctm run” command on the *pipeline.json* file, it activates the pipeline's execution within AWS.

First, we run “*ctm build sagemakerjob.json*” to validate our JSON configuration and then the “*ctm run sagemakerjob.json*” command to execute the pipeline.

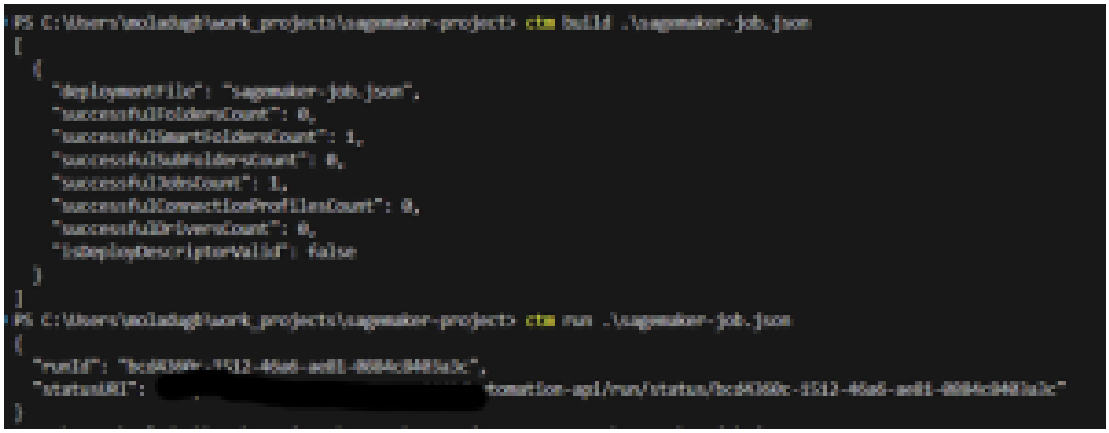


Figure 4. Launching Amazon SageMaker job.

As seen in the screenshot above the "*ctm run*" command has launched the Amazon SageMaker job. The next screenshot shows the pipeline running from the Amazon SageMaker console.

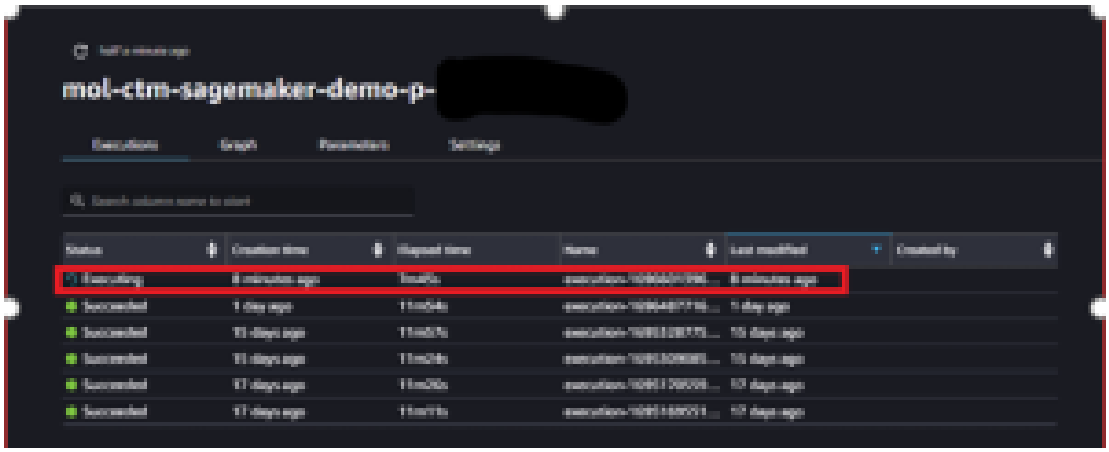


Figure 5. View of data pipeline running in Amazon SageMaker console.

In the Control-M monitoring domain, users have the ability to view job outputs. This allows for easy tracking of pipeline statuses and provides insights for troubleshooting any job failures.

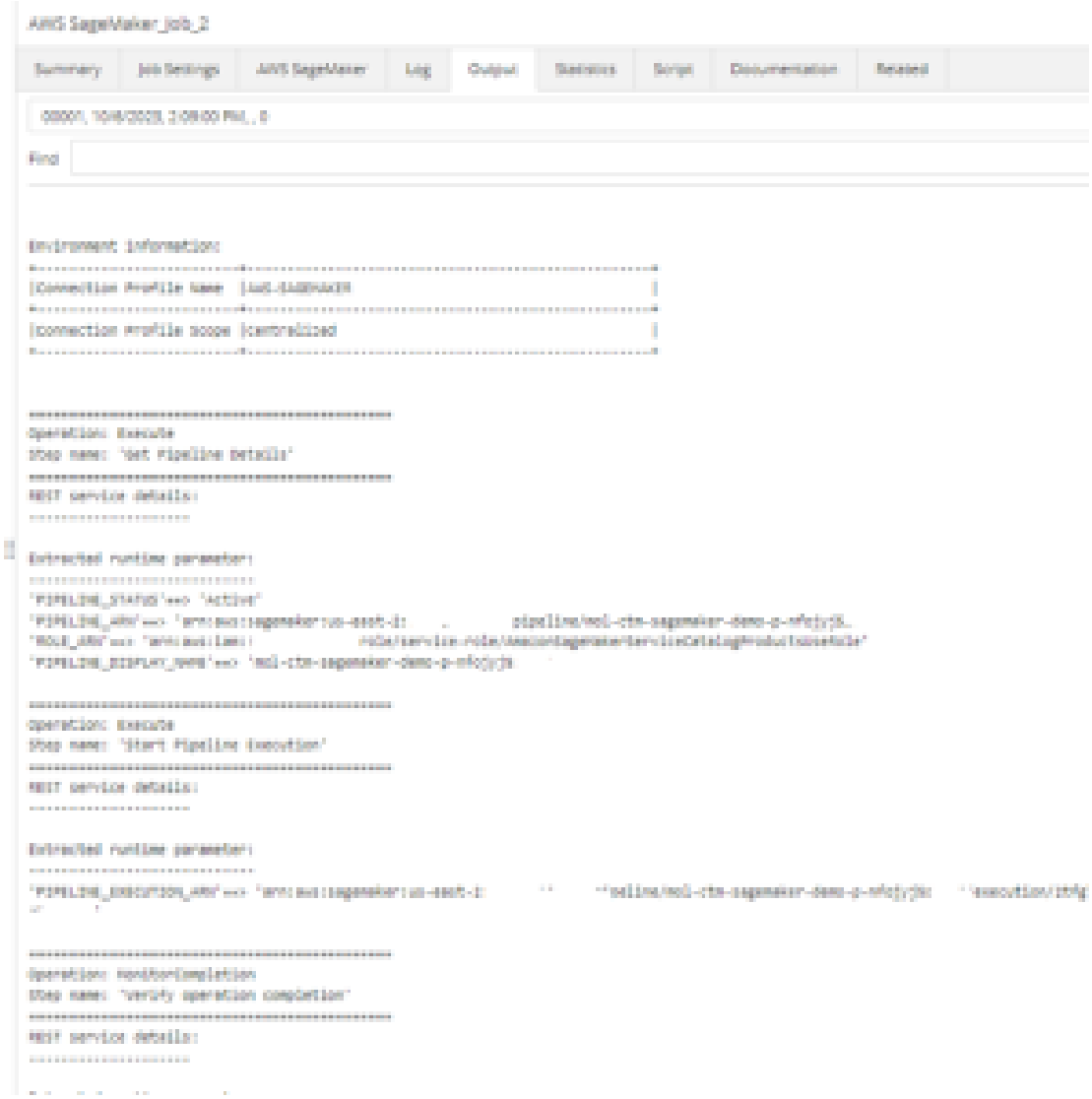


Figure 6. View of Amazon SageMaker job output from Control-M Monitoring domain.

This post walks through how the Control-M SageMaker integration enables teams to schedule, monitor, and govern ML training pipelines alongside enterprise data workflows — moving models from experimentation to reliable production operations.

Summary

In this blog, we demonstrated how to integrate Control-M with Amazon SageMaker to unlock the full potential of AWS ML services, orchestrating them effortlessly into your existing application and data workflows. This fusion not only eases the management of ML jobs but also optimizes your overall automation processes.

Stay tuned for more blogs on Control-M and BMC Control-M integrations! To learn more about Control-M integrations, visit our website.