



Developers CONFERENCE

Agenda

- SageMaker Overview
- SageMaker Workshop
- Bedrock Overview
- Bedrock Demo



Amazon SageMaker Studio

A single web-based interface for end-to-end
ML development

Amazon SageMaker overview

PREPARE DATA AND BUILD, TRAIN, AND DEPLOY ML MODELS FOR ANY USE CASE

PREPARE →

- Ground Truth:** Create high quality datasets for ML
- Data Wrangler:** Aggregate and prepare data for ML
- Processing:** Built-in Python, BYO R/Spark
- Feature Store:** Store, catalog, search, and reuse features
- Clarify:** Detect bias and understand model predictions



Data scientist



ML engineer



Business analyst

BUILD →

- Notebooks:** Fully managed Jupyter notebooks with elastic compute
- Built-In Algorithms:** Integrated tabular, NLP, and vision algorithms
- JumpStart:** UI-based discovery, training, and deployment of models, solutions, and examples
- Autopilot:** Automatically create ML models with full visibility
- Bring Your Own:** Bring your own container and algorithms
- Local Mode:** Test and prototype on your local machine

TRAIN AND TUNE →

- Fully Managed Training:** Broad hardware options, easy to set up and scale
- Distributed Training Libraries:** High-performance training for large datasets and models
- Automatic Model Tuning:** Hyperparameter optimization
- Managed Spot Training:** Reduce training cost by up to 90%
- Debugger and Profiler:** Debug and profile training runs
- Experiments:** Track, visualize, and share model artifacts across teams
- Customization Support:** Integrate with popular open-source frameworks and libraries

DEPLOY AND MANAGE →

- Fully Managed Deployment:** Ultra low-latency, high-throughput inference
- Real-Time Inference:** For steady traffic patterns
- Serverless Inference:** For intermittent traffic patterns
- Asynchronous Inference:** For large payloads or long processing times
- Batch Transform:** For offline inference on batches of large datasets
- Multi-Model Endpoints:** Reduce cost by hosting multiple models per instance
- Multi-Container Endpoints:** Reduce cost by hosting multiple containers per instance
- Inference Recommender:** Automatically select compute instance and configuration
- Model Monitor:** Maintain accuracy of deployed models
- Kubernetes and Kubeflow Integration:** Simplify Kubernetes-based ML
- Edge Manager:** Manage and monitor models on edge devices

SageMaker Studio

A single web-based interface for end-to-end ML

MLOps: Pipelines | Projects | Model Registry

Workflow automation, CI/CD for ML, central model catalog

SageMaker Canvas

Generate accurate ML predictions—no code required

Amazon SageMaker Studio

BRINGS TOOLS FOR EVERY STEP OF THE ML LIFECYCLE IN ONE UNIFIED VISUAL USER INTERFACE



SageMaker Studio gets models to production faster



Unified end-to-end ML: Prepare data and build, train, deploy, and monitor your models all in one place with managed IDEs of your choice, including JupyterLab, RStudio, and Code Editor based on Visual Studio Code – Open Source.



Team collaboration: Shared spaces offer collaboration and project organization between team members with data scientists, ML engineers, and business analysts.



Automate ML processes: MLOps tooling allows for consistency and efficiency for model deployment and monitoring.



Access to over 400 prebuilt models: Quickly spin up popular publicly available foundation models (FMs) through Amazon SageMaker JumpStart.



Easy cost management: Manage cost allocation by teams, users, and projects with automated tagging for each stage of the machine learning (ML) process.

Choice of IDEs

SAGEMAKER STUDIO OFFERS A WIDE RANGE OF FULLY MANAGED IDES FOR ML



JupyterLab

Launch fully managed JupyterLab in seconds for the interactive development environment (IDE) for notebooks, code, and data



Code Editor (based on Visual Studio – Code Open Source)

Use the lightweight and powerful code editor and boost productivity with its familiar shortcuts, terminal, debugger, and refactoring tools



RStudio

Use the fully managed IDE for R and seamlessly switch within RStudio and other IDEs for R and Python development

The screenshot shows the SageMaker Studio interface. On the left is a sidebar with a 'Applications' section containing icons for JupyterLab, Code Editor, RStudio, and Studio CLI. Below this are 'Quick actions' and a 'Home' button. The main area is the 'Home' dashboard, which includes sections for 'Overview', 'Getting started', and 'What's new'. It features cards for 'JupyterLab' (orange background), 'Code Editor' (blue background), 'JumpStart' (blue background), and 'AutoML' (purple background). At the bottom, there are sections for 'Workflows and tasks' (with 'Prepare data', 'Build, train, tune model', and 'Deploy model' sub-sections) and 'Collaborate'.

Prepare data seamlessly in SageMaker Studio



Low-code no-code data prep

Prepare data in a few steps using little to no code with our built-in data preparation tool Amazon SageMaker Data Wrangler



Native integration for Spark analysis

Create and manage EMR clusters and rapidly build, test, and run interactive data preparation and analytics applications with serverless AWS Glue



Data bias detection

Detect and limit potential bias during data preparation, after model training, and in your deployed model using Amazon SageMaker Clarify

```
%load_ext sagemaker_studio_analytics_extension.magics
%sm_analytics emr connect --cluster-id j-3L9V8NDCSQB15 --auth-type None --language python
Successfully read emr cluster(j-3L9V8NDCSQB15) details
Initiating EMR connection..
Starting Spark application
ID          YARN Application ID   Kind  State  Spark UI  Driver log  User  Current session?
0  application_166989018166_0002  pyspark  idle  Link  Link  None  ✓
SparkSession available as 'spark'.
{"namespace": "sagemaker-analytics", "cluster_id": "j-3L9V8NDCSQB15", "error_message": null, "success": true, "service": "emr", "operation": "connect"}
[3]: print(spark.version)
3.2.0-amzn-0
```

Generative AI and SageMaker Studio

Use

Generative AI features to increase data scientist efficiency within SageMaker Studio



Amazon CodeWhisperer and Jupyter AI extensions help data scientists generate, debug, and explain their source code, and the Amazon CodeGuru extension helps conduct security and code quality scans

Build

Large FMs for within SageMaker Studio



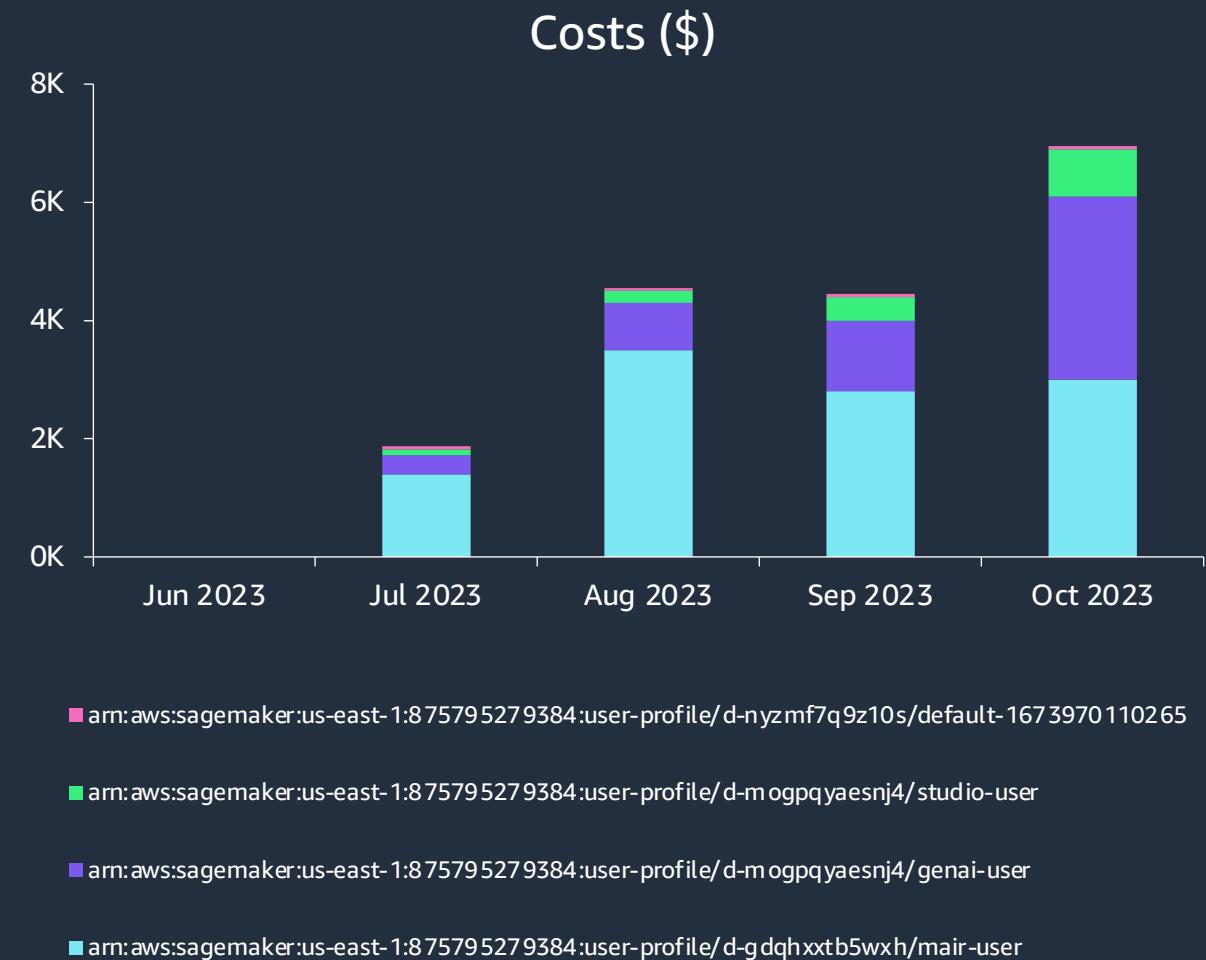
Have access to a wide range of FMs and notebooks backed by high performance compute to fine-tune large models, and have the ability to scale to distributed training directly from Studio notebooks

Analyze and manage costs within SageMaker Studio

With SageMaker Studio, you only pay for what you use—no licenses

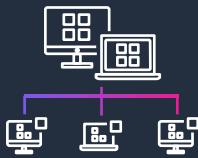
SageMaker Studio automatically tags resources by user profile, space, and domains, allowing administrators to granularly review costs attribution within Cost Explorer

Use lifecycle configurations to automatically shut down idle resources



Security on SageMaker Studio

BUILT-IN FEATURES HELP YOU GO FROM IDEA TO PRODUCTION FASTER, WITHOUT COMPROMISING SECURITY



Infrastructure and network isolation

Amazon Virtual Private Cloud (Amazon VPC) support, AWS PrivateLink support, and disabling internet access



Authentication and authorization

AWS IAM Identity Center, AWS Identity and Access Management (IAM), and IAM SourceIP restrictions



Data protection

Ensure automatic data encryption at rest and in transit with flexibility to bring your own keys



Auditability and monitoring

Track, trace, and audit all API calls, events, data access, or interactions down to the user and IP level to ensure quick remediation



Compliance certifications

ML workflows, continuous integration and continuous delivery (CI/CD), lineage tracking, and catalog

Get started with SageMaker Studio

Visit SageMaker Studio
to learn more



SageMaker Workshop

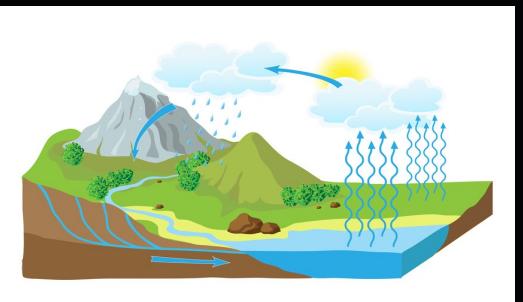
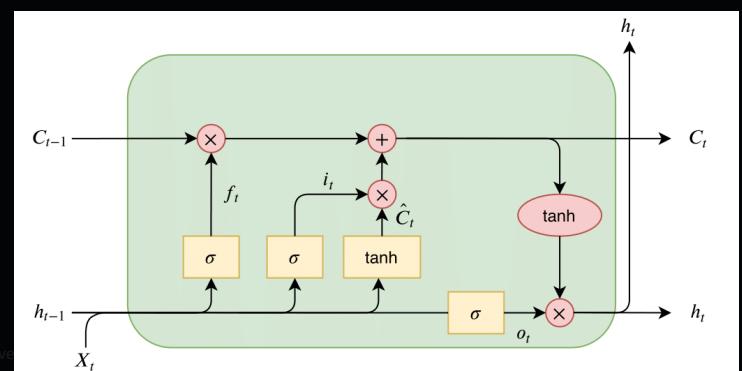
- <https://d3srcxjz0b50d1.cloudfront.net>

Learn SageMaker

- An overview of ML model development in SageMaker.
- Learn how to develop a simple Long-Short-Term Memory (LSTM) on SageMaker to post-process NWM data.
- Learn how to use different CPUs and GPUs.
- Learn how to use Git, Conda, and Amazon S3 buckets in our model development process.

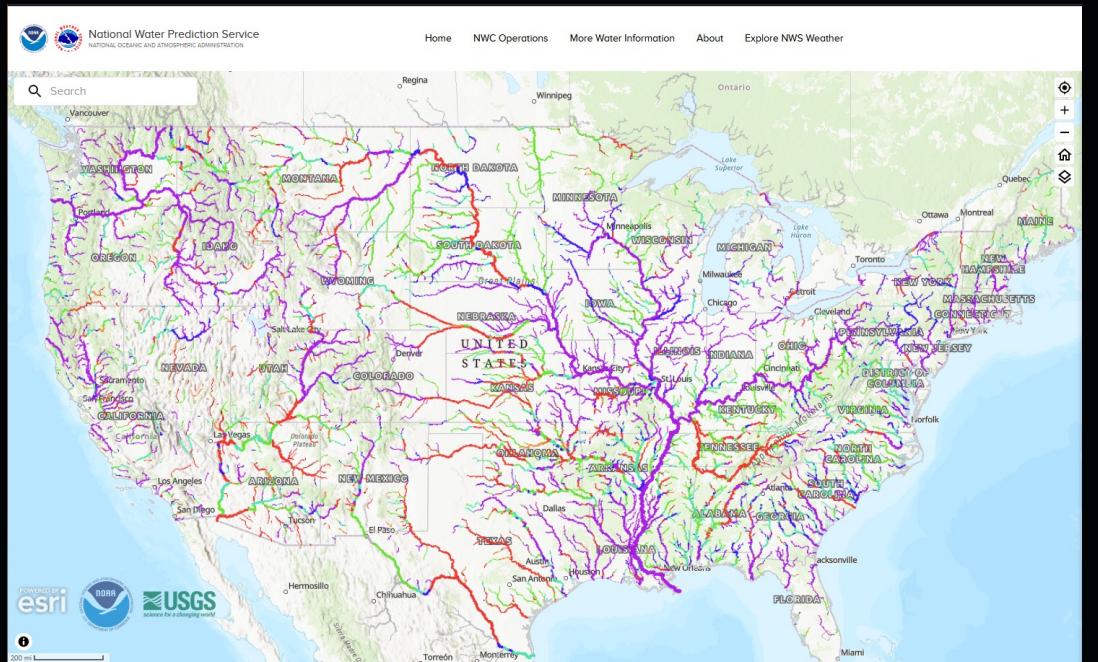
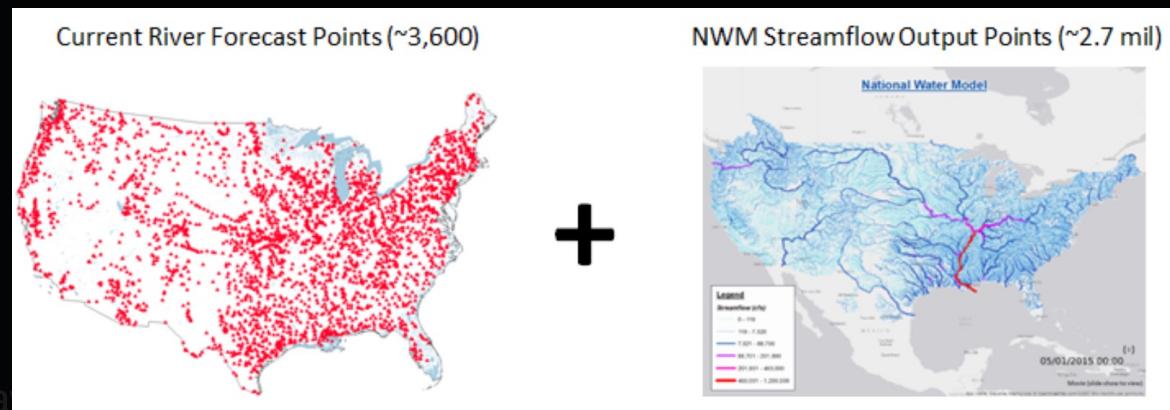


Amazon SageMaker



The National Water Model

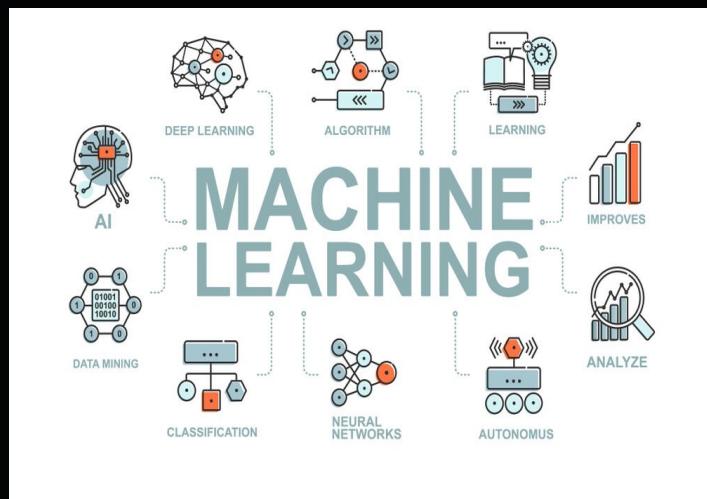
- Addresses the need for a consistent, large-scale forecast.
- Created by NOAA's Office of Water Prediction.
- Developed based on WRF-Hydro.
- Provides predictions for 2.7 million reaches.



Post-processing Hydrological Predictions



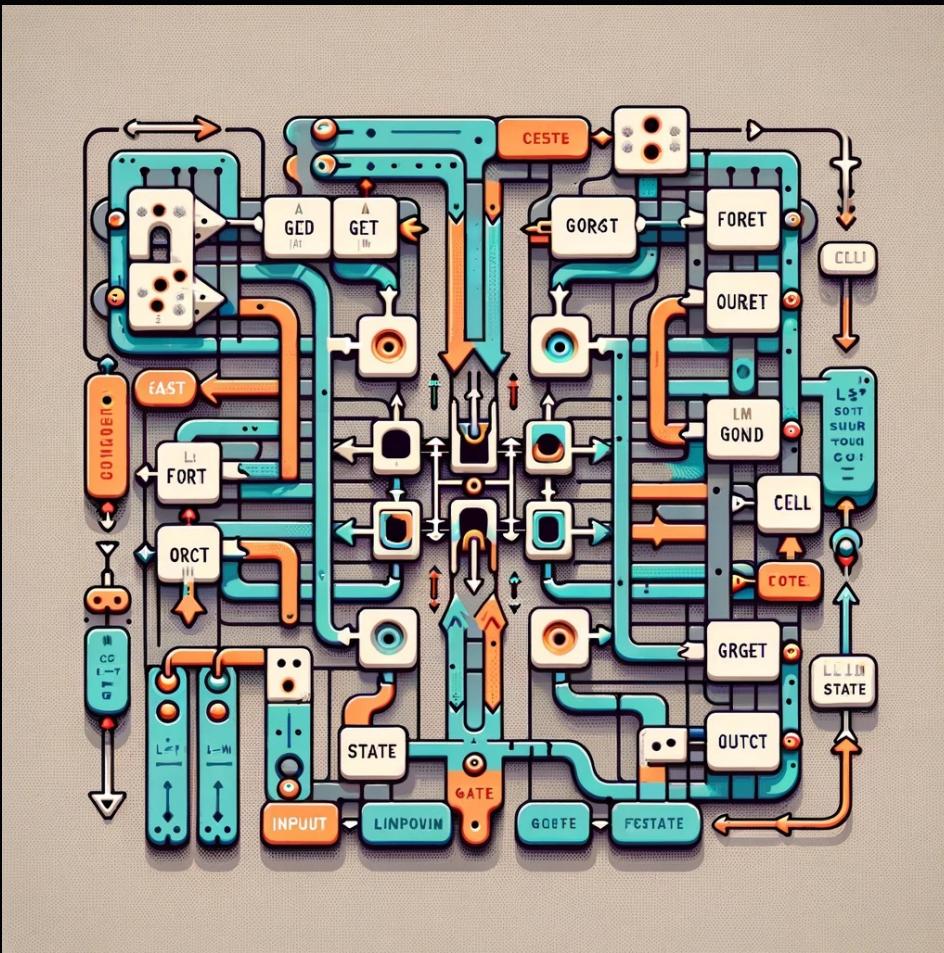
- Accurate prediction of future streamflow is critical for effective water management.
- There are different ways to improve the predictions, including post-processing.
- Post-processing corrects biases by transforming model outputs based on the relationship between observations and the model.
- ML models proved to be useful in post-processing.



Extreme Gradient Boosting (XGBoost) Algorithm



- LSTMs are a Recurrent Neural Network (RNN) type that can effectively learn long-term dependencies in sequential data.
 - LSTMs are good for tasks where past information and context affect current output, such as language translation, speech recognition, and time series forecasting.
 - LSTM cells have three gates that regulate info flow: input, forget, and output.





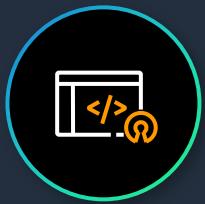
Amazon Bedrock

Build and scale generative AI applications
with foundation models

Building generative AI applications is challenging



Accessing multiple FMs
and newer versions



Customizing FMs
is complicated



Maintaining data privacy
and security



Getting FMs
to execute tasks

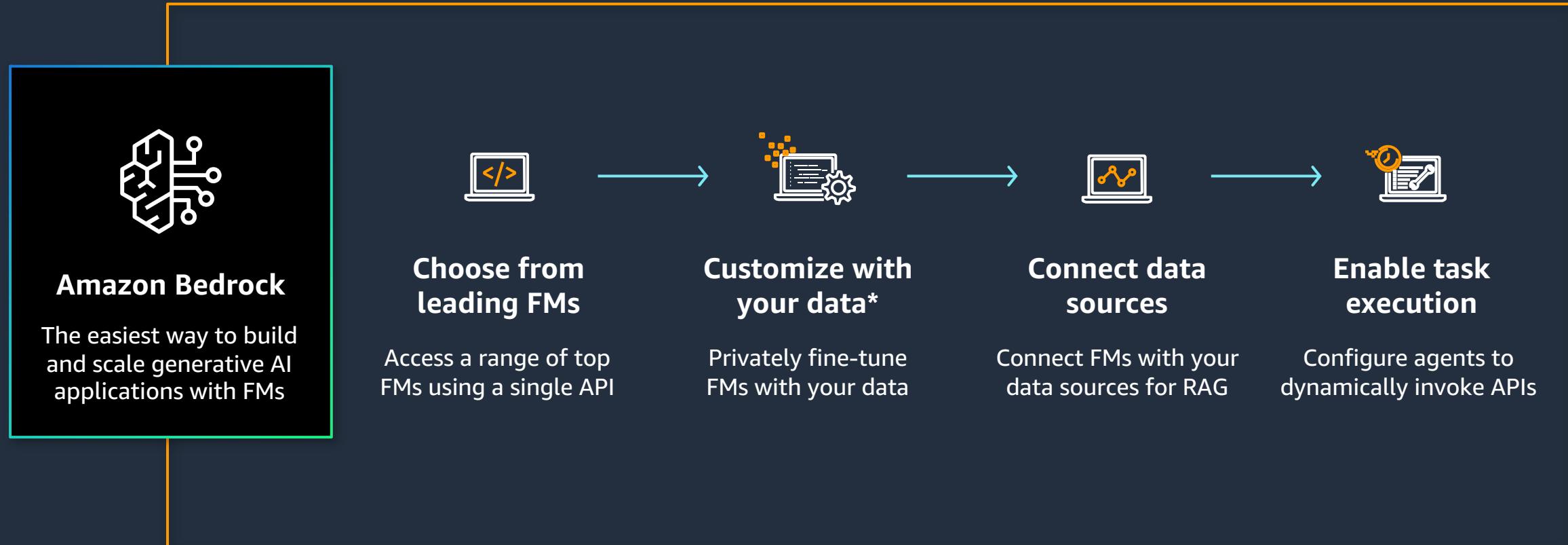


Connecting to
data sources



Managing infrastructure
can be challenging

How it works



* Your data is not used for service improvements and is not shared with third-party model providers.

Knowledge base for retrieval augmented generation (RAG)

Connect FMs to data sources including vector engine for Amazon OpenSearch Serverless, Pinecone, and Redis Enterprise Cloud

Enable automatic data source detection

Provide citations



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Step 1

Provide details

Step 2

Set up data source

Step 3

Review and create

Set up data source [Info](#)

Set up your data source by specifying the S3 location of your data, choosing an embeddings model to convert the data, and providing details for a vector database in which Bedrock can store, manage, and update your embeddings.

Data source

Specify the S3 location of your data. The vector database will ingest your data and convert it into an embedding.

Data source name

Enter name

Valid characters are a-z, A-Z, 0-9, _ (underscore) and - (hyphen). The name can have up to 50 characters.

S3 URI

Search



[View](#)

[Browse S3](#)

KMS Key - optional

Provide the KMS key to allow Bedrock to decrypt and encrypt your data.

Choose an AWS key or enter an ARN

KMS Key for transient data storage - optional

Provide a KMS key to store the transient data while we are in process of converting your data into embeddings.

Choose an AWS key or enter an ARN



[Create an AWS KMS key](#)

Embeddings model

Select an embeddings model to convert your data into an embedding. Pricing depends on the model.



Titan G1 Embeddings - Text v0.02

Vector database

Select your previously created database to allow Bedrock to store, update, and manage embeddings. [Learn more](#)

Select an existing database [Info](#)

Vector engine Amazon

OpenSearch Serverless

If you are a first time user, visit [OpenSearch](#) to create a vector database.



Pinecone

If you are a first time user, visit [Pinecone](#) to create a vector database.



Redis Enterprise Cloud

If you are a first time user, visit [Redis Enterprise Cloud](#) to create a vector database.



Amazon Bedrock supports leading foundation models



Amazon Titan

Text summarization, generation, classification, open-ended Q&A, information extraction, embeddings, and search



Jurassic-2

Multilingual LLMs for text generation in Spanish, French, German, Portuguese, Italian, and Dutch



Claude 2

LLM for conversations, question answering, and workflow automation based on research into training honest and responsible AI systems



Command

Text-generation model for business applications and embeddings model for search, clustering, or classification in 100+ languages



Llama 2 (coming soon)

Fine-tuned models ideal for dialogue use cases and language tasks



Stable Diffusion

Generation of unique, realistic, high-quality images, art, logos, and designs

Security and privacy

Private connectivity between Amazon Bedrock and your Amazon Virtual Private Cloud (Amazon VPC)

Your data is encrypted in transit and at rest

Support for standards, including GDPR compliance and HIPAA eligibility



Bedrock Demo

The screenshot shows the Amazon Bedrock interface with the "Chat playground" selected. On the left, a sidebar lists various services like "Getting started", "Foundation models", "Playgrounds", "Safeguards", "Orchestration", and "Assessment & deployment". The main area displays a model configuration for "Claude 3 Haiku". The configuration panel includes fields for "Temperature" (set to 1), "Top P" (set to 0.999), "Top K" (set to 250), and "Maximum length" (set to 4096). It also features sections for "Image" and "Run". Below the configuration is a summary of NWM v2.0 historical data, mentioning its application to hydrological processes and its transparency for the public. At the bottom, there's a "Model metrics" section with a link to "Model evaluation".

Amazon Bedrock <

Amazon Bedrock > Chat playground

Chat playground Info

All Claude 3 Haiku Change

v1 | ODT

An example prompt for long document Q&A supplemented by An example prompt for long document q&a supplemented by An example prompt to set up in-character roleplaying as a career

forecasts, for the appropriate reasons.

All models have ingrained assumptions (stated or unstated) that influence their performance. Most of these models are based on hydrological processes developed for pristine headwater basins in a particular location and for a specific event types. These assumptions imply that no single model is best everywhere, or, for all types of events. A framework like the one presented here offers a unique way to compare model results (either model-to-model or model-to-observation) that directly target questions related to model parametrization; process representation; and the presence of conditional and unconditional biases. Future research could use this decomposition framework to further diagnose error contributions from the entire modeling cycle including forcings, parameter estimation, process selection and calibration/ regionalization. Moreover, this approach can be applied to other model development and intercomparison efforts. Its application to the NWM v2.0 historical data provides increased transparency for the public, catering to those seeking to use and improve NWM model outputs.

</article>. Please summarize the top 10 key points of this content.

Configurations Reset

Temperature: 1

Top P: 0.999

Top K: 250

Length

Maximum length: 4096

Stop sequences: Human:

Model metrics

To evaluate models for task specific metrics with custom dataset visit [Model evaluation](#)



Thank you!