W&B deployment guide

Weights & Biases overview

Weights & Biases is the AI developer platform that helps teams develop GenAI applications and fine-tune LLMs. Weights & Biases offers two products, W&B Weave and W&B Models.

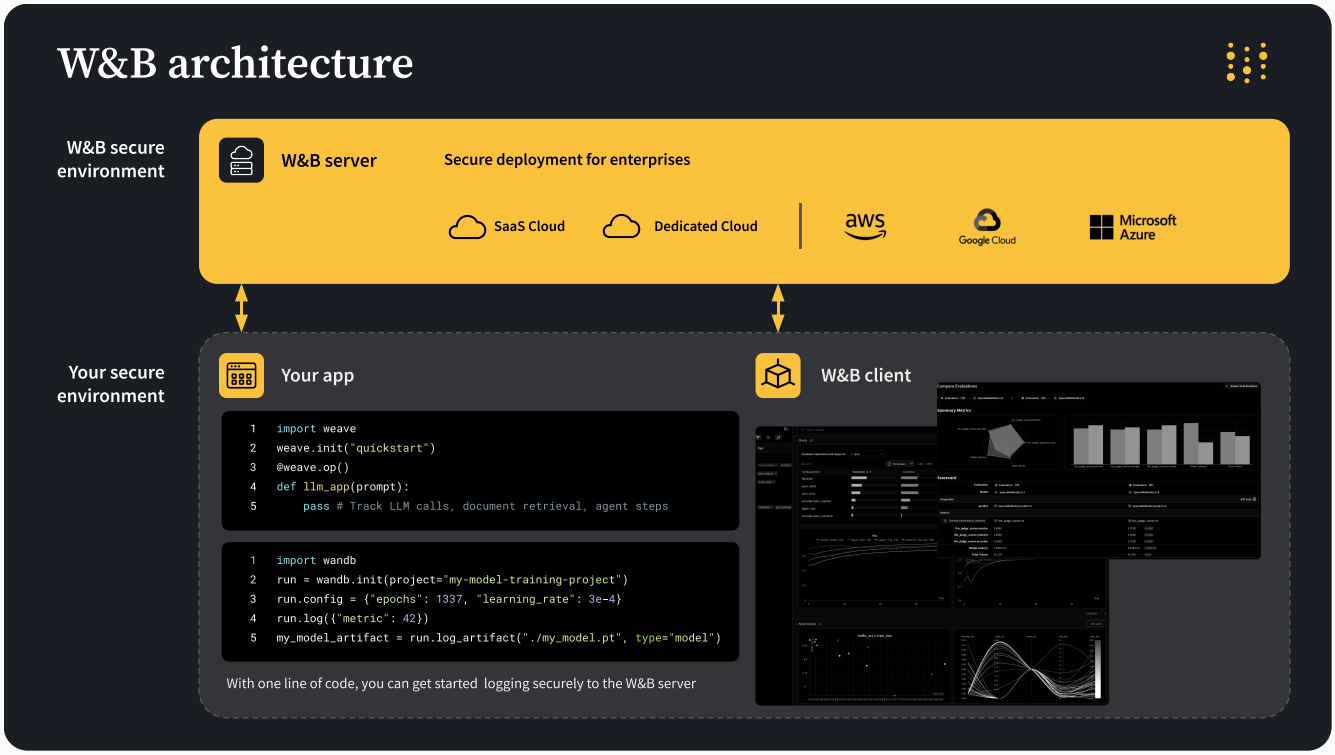
Weave helps developers evaluate, monitor, and iterate continuously to deliver generative AI applications with confidence. You can run robust application evaluations, keep pace with new LLMs, and monitor applications in production while collaborating securely. Weave is designed to overcome the barriers of traditional software development tools to meet the needs of non-deterministic LLM-powered applications. Weave is framework and LLM agnostic so you don't need to write any code to work with popular AI frameworks and LLMs such as OpenAI, Anthropic, Cohere, MistralAI, LangChain, LlamaIndex, DSPy, Cerebras, Google Gemini, Amazon Bedrock, Together AI, Groq, and more.

Models enables your ML teams to train, fine-tune, and manage AI models. Models boosts experiment speed and team collaboration to bring models to production faster while ensuring performance, data reliability, and security. Run more experiments, analyze them interactively, and quickly build higher-quality models. Centralize the tracking of models, datasets, metadata, and their lineage in the Registry to support governance, reproducibility, and CI/CD. Automate workflows for training, evaluation, and deployment to enable rapid iteration.

Weights & Biases architecture

No matter which product you use, Weave or Models, you benefit from a common, secure architecture. There are two components to the Weights & Biases architecture, the client and server. From the client, with just one line of code, you enable logging to the second component, the server. The W&B server ingests, stores, and processes data securely in the cloud, either traces and evaluations for Weave or experiment tracking artifacts for Models. The W&B server uses a secure Clickhouse cloud database to store Weave data and a secure cloud-native database to store Models data.

You have the flexibility to choose from either SaaS cloud (multi-tenant instance) or dedicated cloud (single-tenant instance) for the server. Once the data is logged in the server, you can access it via the Weights & Biases UIs or SDKs.



Cloud deployment options

You can choose between two deployment options, SaaS cloud or dedicated cloud, on any of the leading cloud providers including AWS, Google Cloud, or Microsoft Azure. Below we describe each cloud deployment option so you can decide which is right for you.

SaaS cloud

SaaS cloud is the Weights & Biases multi-tenant, fully managed option hosted in Google Cloud North America. SaaS cloud provides access to a secure version of Weave or Models with all of the latest features. You get:

- Automatic updates, maintenance, and scaling
- Single sign-on (SSO)
- Team-level access control
- Restricted projects for sensitive data processing
- Team-level bring your own bucket (ByoB) to bring sensitive data in your own cloud or on-premise infrastructure (applicable for Models data)
- Identity federation through the SDK
- Automated user, team, and role management using SCIM API
- Org and team-level privacy controls
- SOC 2 Type 2 certification

Why choose SaaS cloud?

SaaS cloud is the most popular deployment option, because its the quickest and most cost efficient to start with. Customers do not have to think about DevOps, site reliability engineering (SRE), security audits, or any product infrastructure.

Dedicated cloud

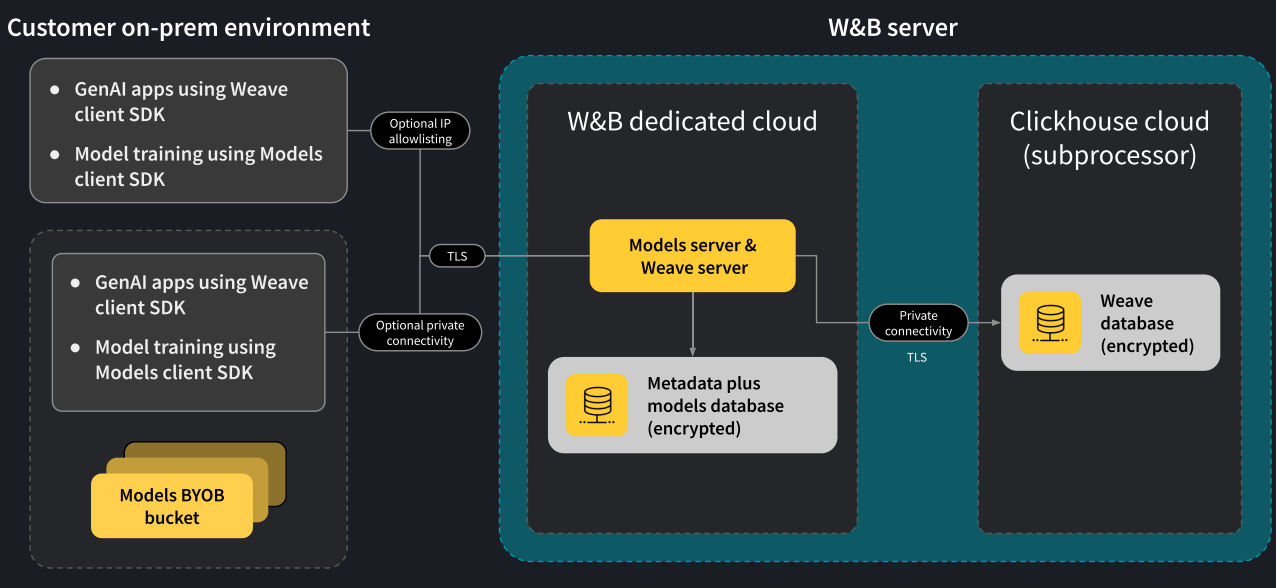
Dedicated cloud is the Weights & Biases single-tenant, fully managed option hosted in the global regions of AWS, Google Cloud, and Microsoft Azure. You get all the benefits of SaaS cloud and more including:

- Data residency
- Unique compute & storage for isolation (including for Clickhouse applicable to Weave)
- Deployment-level bring your own bucket (ByoB) to complement the capability at team-level (applicable for Models data)
- Advanced controls with enterprise SSO
- Private connectivity (inbound and outbound)
  - For Weave, we use outbound private connectivity from server component to the instance's unique Clickhouse cluster.
  - For Models, we provide optional outbound private connectivity from the server to buckets in your cloud account.
  - We also provide optional inbound private connectivity from your environment to the Weave or Models server enabling a completely private instance.
- Data-at-rest encryption
  - We use Weights & Biases managed unique keys for each instance's Clickhouse cluster and MySQL by default. We offer an exception for you to provide your own key to encrypt the data in our managed Clickhouse cluster and MySQL (AWS & GCP).
- HIPAA compliance - Applicable for Models with ByoB, and coming soon for Weave

Why choose dedicated cloud?

Dedicated cloud is a good fit for enterprise and digital native customers who value data residency and isolation and have advanced security needs. Customers find dedicated cloud much easier and highly cost efficient in the long term compared to a self-managed platform.

How dedicated cloud works



Self-managed option for Models

While its not recommended by Weights & Biases because it typically creates greater customer costs and less efficient support compared to dedicated cloud, some customers opt to deploy the Weights & Biases server component on-prem or in their own cloud account for Models. With the self-managed option, your IT/DevOps/MLOps team is responsible for provisioning your deployment, managing upgrades, round-the-clock observability, and continuously maintaining your self- managed W&B server instance. Weights & Biases updates are released once every month, and we offer tools to aid administration.

You need to configure several infrastructure components in order to set up W&B server in your own managed infrastructure. Some of those components include include, but are not limited to:

- Kubernetes cluster
- MySQL 8 database cluster
- Amazon S3-compatible object storage
- Redis cache cluster

Next Steps

Ready to deploy? Reach out to use https://wandb.ai/site/contact and we'll be in touch.