Weights & Biases

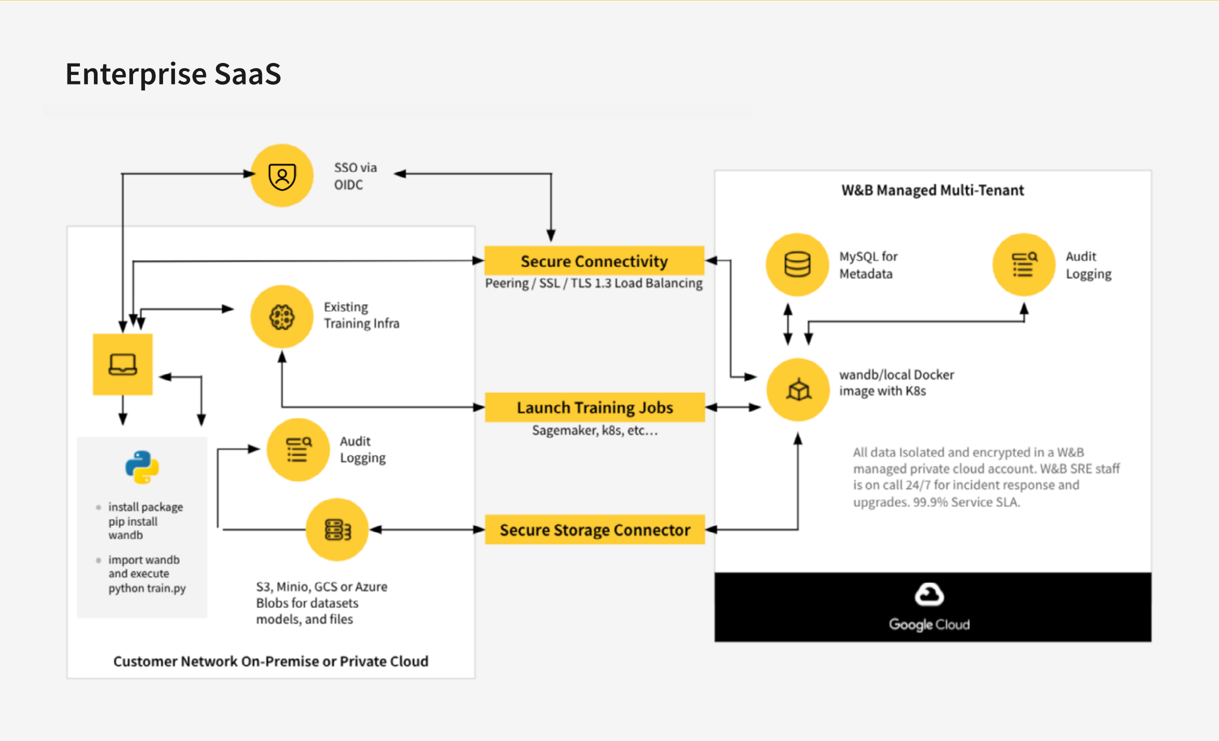
W&B Deployment **Options**

Weights & Biases offers flexible deployment options including a multi-tenant cloud, a single-tenant cloud, and customer-managed private deployments.

Shared Responsibility Model

	SaaS/ Multi-Tenant	Dedicated Cloud/ Single-Tenant	Customer-Managed
MySQL/DB management	Fully hosted & managed by W&B on GCP cloud	Fully hosted & managed by W&B on cloud & region of customer choice	Fully hosted & managed by customer
Object Storage (S3/GCS/Blob storage)	Fully hosted by W&B on GCP cloud	Option 1: Fully hosted by W&B Option 2: Customer can configure their own bucket using the Secure Storage Connector	Fully hosted & managed by customer
SSO Support	W&B managed via Auth0	Option 1: Customer Managed Option 2: Managed by W&B via Auth0	Fully managed by customer
W&B Service (App)	Fully managed by W&B	Fully managed by W&B	Fully managed by customer
App Security	W&B	W&B/Customer	Customer
Maintenance (upgrades, backups etc.)	Maintained by W&B every two weeks	Maintained by W&B every two weeks	Managed by Customer
Support	Support SLA	Support SLA	Support SLA
Supported Clouds	N/A	AWS, GCP (GA) Azure (coming soon)	AWS, GCP, Azure

SaaS/Multi-Tenant Architecture


Enterprise SaaS
SSO via OIDC
Secure Connectivity
Peering / SSL / TLS 1.3 Load Balancing
Existing Training Infra
Launch Training Jobs
Sagemaker, k8s, etc…
Audit Logging
Secure Storage Connector
install package pip install wandb
import wandb and execute python train.py
S3, Minio, GCS or Azure Blobs for datasets models, and files
Customer Network On-Premise or Private Cloud
W&B Managed Multi-Tenant
MySQL for Metadata
Audit Logging
wandb/local Docker image with K8s
All data Isolated and encrypted in a W&B managed private cloud account. W&B SRE staff is on call 24/7 for incident response and upgrades. 99.9% Service SLA.
Google Cloud

OVERVIEW

Our most popular deployment option. A Multi-Tenant SaaS offering that allows you access to a fast, secure version of W&B with all of the latest features.

Infrastructure

- Traditional SaaS with multi-tenancy hosted on GCP in `**us-central**` region W&B service runs on a highly available Kubernetes cluster powered by Google Kubernetes Engine
- The service connects to a Cloud SQL database instance (MySQL 5.7) maintained with regular backups
- The service also uses a scalable GCS object storage bucket for file storag

Reliability & Maintenance

- All updates, maintenance and service availability are taken care of by W&B
- All resources are monitored and SRE team is alerted if a resource exceeds/falls below set threshold value

Migration

- Migration to and from SaaS to other deployment types is currently not possible

Single Sign-On (SSO)

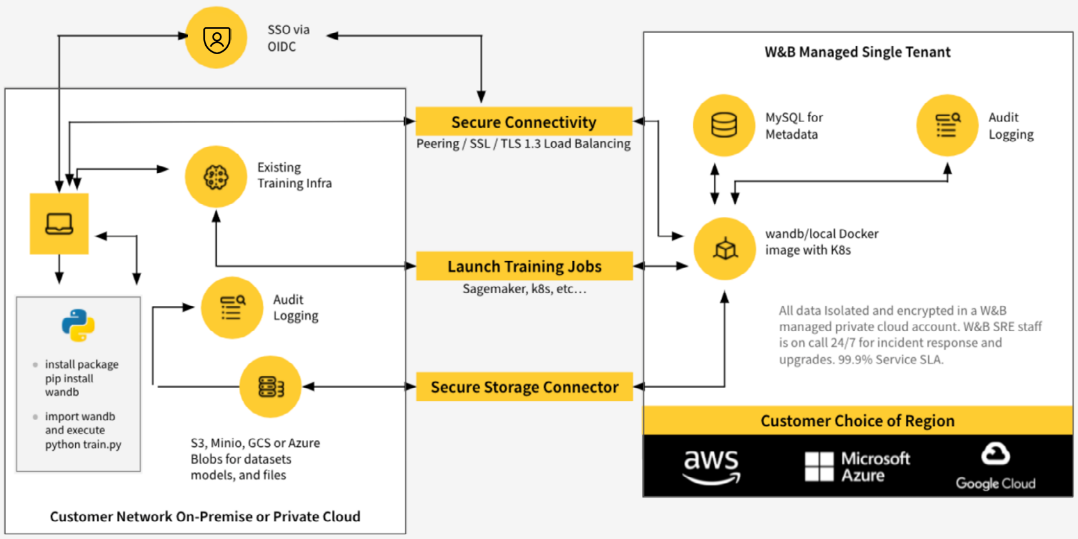
- W&B can configure SSO with Auth0 and customer’s Identity Provider

Support

- On-call SREs and 5*15*7 technical support team
- Option to setup slack channel for suppor

Dedicated Cloud/Single-Tenant Architecture

Single Tenant Managed Cloud


SSO via OIDC
Secure Connectivity
Peering / SSL / TLS 1.3 Load Balancing
Existing Training Infra
Launch Training Jobs
Sagemaker, k8s, etc…
Audit Logging
install package pip install wandb
import wandb and execute python train.py
Secure Storage Connector
S3, Minio, GCS or Azure Blobs for datasets models, and files
Customer Network On-Premise or Private Cloud
W&B Managed Single Tenant
MySQL for Metadata
Audit Logging
wandb/local Docker image with K8s
Customer Choice of Region
aws
Microsoft Azure
Google Cloud

All data Isolated and encrypted in a W&B managed private cloud account. W&B SRE staff is on call 24/7 for incident response and upgrades. 99.9% Service SLA.

OVERVIEW

For customers with sensitive use cases or stringent Enterprise security requirements, Dedicated Cloud provides an isolated environment where your data lives alone.

Infrastructure

Hosted by W&B:

- SaaS offering with Single Tenancy hosted by W&B (currently supports AWS & GCP) and a region of customer's choice
- W&B service (Docker container) runs in a highly available Kubernetes cluster
- The service connects to a database (supported versions: MySQL 5.7, MySQL 8) maintained with regular backups

Hosted by W&B or Customer:

- The service also connects to a scalable object storage bucket (S3/GCS) for file storage,
  - Option 1: W&B can setup this object storage bucket on behalf of the customer in W&B's cloud account **OR**
  - Option 2: Customers can bring their own bucket hosted in customer's cloud account (BYOB)

Data Access

- Most communication between Python client, the W&B application and the object storage bucket happens via encrypted pre-signed URLs
- W&B server needs read access to the bucket in order to generate these signed URLs for use by Python client
- W&B server needs write access to the bucket to update file metadata

Dedicated Cloud/Single-Tenant Architecture

Reliability & Maintenance

- All updates, maintenance and service availability are taken care of by W&B
- All resources are monitored and SRE team is alerted if a resource exceeds/falls below set threshold value

Migration

- Migration to and from Dedicated Cloud to Private Cloud & On-prem is supported

Single Sign-On (SSO)

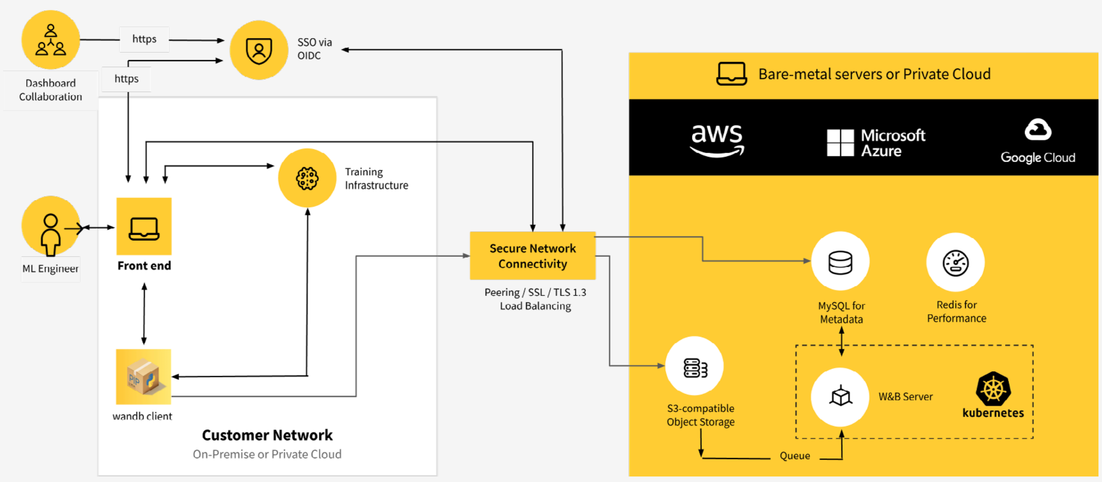
- W&B can configure SSO with Auth0 and customer’s Identity Provider

Support

- On-call SREs and 5*15*7 technical support team
- Option to setup slack channel for support

On-Prem & Private Cloud Architecture

Customer-Managed On-prem



OVERVIEW

Whether it is in your own customer-managed cloud or on-prem servers, deploy and manage Weights & Biases your way.

Infrastructure

Hosted by Customer on their On-Prem infrastructure:

- Can be hosted in any of the major cloud providers (GCP, AWS, Azure etc.) or on customer's bare-metal servers
- W&B service (Docker container) - recommended to run on Kubernetes
- Setup an external MySQL5.7/MySQL8 database instance to store metadata from the application
- Setup an external scalable object storage solution to store artifact data logged from the application

(W&B provides Terraform templates for all major private clouds that customers can use to spin up all the required infrastructure)

Data Access

- All communication happens within the customer's private network
- Anonymized and aggregated telemetry data will be sent to wandb servers for improving the product. Specific telemetry information collected are listed below:
    - userCount
    - teamCount
    - sweepCount
    - runCount
    - projectCount
    - sweepRunCount
    - artifactCount
    - artifactStorageBytes
    - artifactReferenceBytes
    - artifactCollectionCount
    - artifactTypeCount
    - storageBytes
    - reportCount
    - reportCommentCount
    - computeHours
    - userStat

On-Prem & Private Cloud Architecture

Scalability

- W&B's docker container currently scales vertically, meaning it scales the resource requirements as opposed to the replica count
- Kubernetes Pod configuration requirements:
    - Minimum configuration (xlarge instances)
        - **4 VCPUs**
        - **8 GB RAM**
    - Recommended scaling configuration (2xlarge instances)
        - **8 VCPUs**
        - **32 GB RAM**

Reliability & Maintenance

- W&B releases a new version of their `wandb/local` docker container every 2 weeks
- Customer's IT team would be responsible for upgrading to the latest version at that cadence
- Since the service would be hosted on Kubernetes, upgrades do not require any downtime
- Regular database snapshots and backups enabled

Advanced Reliability Settings

Redis

(A reference Redis terraform module can be found here.)

Configuring an external redis server will improve the reliability of the service and enable caching which will decrease load times especially in large projects. We recommend using a managed redis service (ex: ElastiCache) with HA and the following specs:

- Minimum 4GB of memory, suggested 8GB of memory
- Redis version 6.x
- In transit encryption
- Authentication enabled

Rate Limiting

By default, the W&B server does not enforce any rate limits. You can enable rate limiting by setting the following environment variables:

- GORILLA_DEFAULT_RATE_LIMITS_FILESTREAM=20
- GORILLA_DEFAULT_RATE_LIMITS_GRAPHQL=200

This will ensure that no single API key can overwhelm the server and our client library will automatically backoff and retry requests that hit the rate limit.

On-Prem & Private Cloud Architecture

State Management

- State management in W&B is done via two object stores
- MySQL5.7 database
  - Recommended instance size - db.r5.large
  - Scale resources if the database CPU load consistently is > 70%
- An S3 object storage bucket

Monitoring Recommendations

- Database CPU Load
- Kubernetes Pod Health Check (the deployment provides a livenessProbe with /healthz checkpoint to check for pod health)
- Audit logging enabled on the bucket for monitoring access to the bucket
- Enabling out of the box Cloud watch logs for different services

Single Sign-On (SSO)

- Customers can configure SSO directly with the application for implicit and pkce grant types and with an OIDC compliant Identity Provider
- For SAML connections, SSO needs to be configured to route via W&B's Auth0 as it's not directly supported out of the box by the application

Support

- Support SLA
- Option to setup slack channel