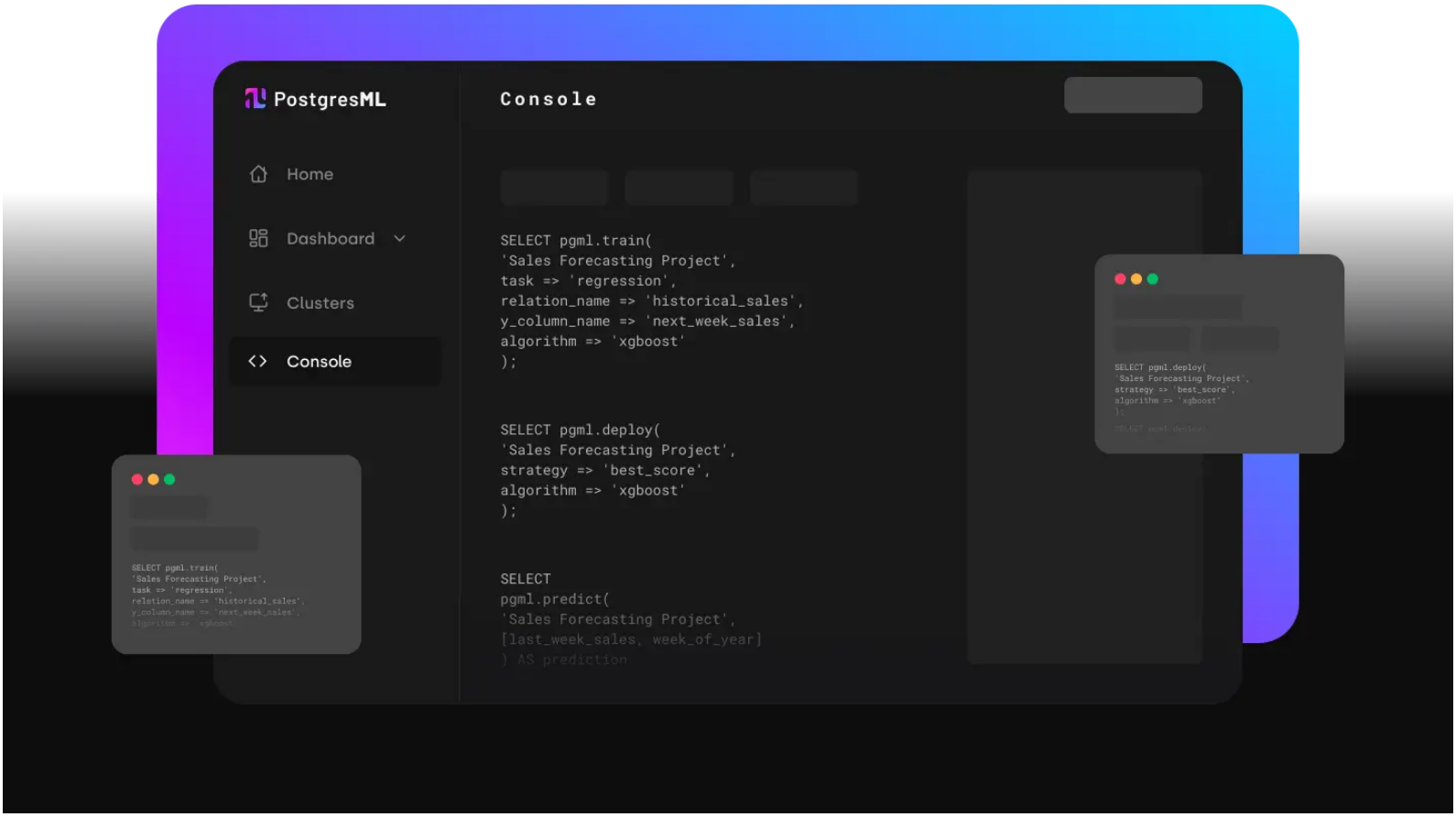
Postgres ML
BETA

on a Scalable Solution

Our open source software scales Postgres horizontally for your interactive AI powered applications, with the latest LLMs, vector operations, classical Machine Learning and good old Postgres application workloads.

Docs

Get Started For Free



How It

Use the expressive power of SQL along with the most advanced machine learning algorithms and pretrained models in a high performance database.

PostgresML indexes application data alongside computed features like embedding vectors with your machine learning models all together in a seamless shared memory space. This eliminates network calls, process boundaries, data duplication and other unnecessary complexity, which makes it more reliable, efficient, fast and simple.

Our serverless platform is built around our custom Postgres pooler PgCat, which allows you to scale your inference layer to millions of predictions per second across multiple GPU accelerated machines.

Try Now For Free

Fast

Up to 40x performance improvement over traditional microservice architectures. Shared memory between data and models in a single process eliminates network calls, process boundaries and data duplication.

Comprehensive

PostgresML can download state-of-the-art open source models from HuggingFace, or train your own end to end. It also supports many algorithms like Torch, Tensorflow, XGBoost, LightGBM, and all the classical ones in Scikit.

Scalable

Scale your inference layer to millions of predictions per second in our GPU accelerated cloud. Start for

free with our serverless option, or use dedicated Postgres replicas.

Simple

PostgresML is your application and vector database, model store, feature store, inference server and ML deployment pipeline, with support for all major languages and application frameworks as clients.

Your Data

Efficiently load data or features from upstream sources with Postgres replication, or connect your application directly through our custom load balancer, PgCat.

Highly Available

Abstract everything behind a single connection string with smart query routing, sharding, and enterprise-grade managed infrastructure.

How to

Train

```
      pgml.train(
                    ,
  task =>            ,
  relation_name =>                ,
  y_column_name =>               ,
  algorithm =>
);
```

Learn About Training

Deploy

```
      pgml.deploy(
                  ,
  strategy =>          ,
  algorithm =>
);
```

Learn About Deployments

Predict

```
      pgml.predict(
                  ,
  ARRAY[
    last_week_sales,
    week_of_year
  ]
)    prediction
       prediction    ;
```

Learn About Predictions

Create Your First Project

What

"Absolutely brilliant" on PgCat, our sharded PostgreSQL proxy.

Carnegie Mellon

"The improvement is quite remarkable" on PostgresML v2.0.

Scaling Postgre

Start Now With Machine Learning

Get Started For Free

Discover All

Fastest Inference

PostgresML eliminates separation between your model server and

All Your Favorite Algorithms

datastore, minimizing latency and computation costs. You can even generate embeddings on the fly in queries. Our benchmarks show a 8x-40x improvement over Python HTTP microservices.

Overview

Whether you need a simple linear regression or extreme gradient boosting, PostgresML includes suppor for all classification and regression algorithms in **Scikit Learn**, **XGBoost**, **LightGBM** and pre-trained deep learnir models from **Hugging Face**.

Algorithms

8x-40x

Faster than Python

Prediction Latency

Algorithms

QPS on EC2

We Have the Perfect

PRIVATE ALPHA

Serverless

Starting at

$0

For startups
Free, without cache acceleration

- $0.25/hr per GB GPU cache
- Multi GPU burst capability
- Cache your models on the GPU
- Instant scalability up to 256GB GPU
- Scale further with advanced sharding functionality

Get Started For Free

PUBLIC BETA

Dedicated

Starting at **$0.60/hr**

For orgs of any size
Dedicated cluster with fixed hardware

- Choose CPU, RAM or GPU resources
- Horizontally scalable inference with replicas
- High availability for your production applications
- Multiple users
- Multiple databases
- Automated Backups
- Metrics
- Logs

Get Started Today

CUSTOM TIMELINES

Enterprise

Custom Pricing

--

For orgs with custom needs
Your hardware, your way

- Customized hardware
- Solution Architecture support
- Private VPC/On-prem deployments
- Access Control Lists
- Single Sign-on

Contact Us

PostgresML

PostgresML

Github

Documentation

Contact Us

Discord

Create an Account

Login

Terms of Service

Privacy Policy