MMM Trainer
Cost Estimates & Technical Requirements

Pricing Guide for Self-Hosted Deployment

Cost Analysis & Requirements Documentation

January 12, 2026

1 Overview

This document provides comprehensive cost estimates for deploying and operating the MMM Trainer platform on Google Cloud Platform. All costs are based on actual production measurements from January 2026, running on the current infrastructure configuration (8 vCPU, 32GB RAM, full core utilization).

1.1 Key Takeaways

- **Training jobs** account for 90-99% of total variable costs
- **Fixed infrastructure** costs approximately $2/month
- **Per-second billing** means faster machines don't necessarily cost more
- **Actual production runs** take 80-120 minutes (not linear scaling from benchmark)

2 Training Job Costs

All costs are based on actual production measurements using Google Cloud Platform services in the `europe-west1` region with 8 vCPU and 32GB RAM configuration.

2.1 Job Performance and Costs

Workload	Iterations × Trials	Duration	Cost/Job	Use Case
Test Run	200 × 3 = 600	0.8 min	$0.01	Quick validation
Benchmark	**2,000 × 5 = 10K**	**12 min**	**$0.20**	**Standard testing**
Production	**10,000 × 5 = 50K**	**80-120 min**	**$1.33-$2.00**	**Production runs**
Large Production	10,000 × 10 = 100K	160-240 min	$2.67-$4.00	High-quality runs

Table 1: Training job durations and costs

Important Notes:

- Benchmark runs (12 minutes) are ideal for experimentation and testing
- Production runs (80-120 minutes) provide more thorough model exploration and are recommended for final models
- Large production runs with 10 trials offer highest quality results but take longer
- All durations based on verified production measurements with full 8-core utilization

2.2 Cost Calculation Details

Cloud Run Pricing (europe-west1):

- CPU: $0.000024 per vCPU-second
- Memory: $0.0000025 per GiB-second
- Per-second billing (no minimum charge)

Example: Benchmark Workload (2,000 × 5)

```
Time: 720 seconds (12 minutes)
CPUs: 8 vCPU
Memory: 32 GiB

CPU cost:    720 sec × 8 vCPU × $0.000024 = $0.138
Memory cost: 720 sec × 32 GiB × $0.0000025 = $0.058
Total: ~$0.20 per job
```

Example: Production Workload (10,000 × 5)

```
Low estimate (80 minutes):
Time: 4,800 seconds
CPU cost:    4,800 sec × 8 vCPU × $0.000024 = $0.922
Memory cost: 4,800 sec × 32 GiB × $0.0000025 = $0.384
Total: ~$1.33 per job

High estimate (120 minutes):
```

```
Time: 7,200 seconds
CPU cost:    7,200 sec × 8 vCPU × $0.000024 = $1.382
Memory cost: 7,200 sec × 32 GiB × $0.0000025 = $0.576
Total: ~$2.00 per job
```

3 Monthly Cost Scenarios

3.1 Usage-Based Estimates

Usage Level	Web Calls	Training Jobs	Benchmark Cost	Production Cost
Light	100	10	$4	$15-22
Moderate	500	50	$12	$69-102
Heavy	1,000	100	$22	$135-202
Very Heavy	5,000	500	$102	$667-1,002

Table 2: Monthly cost estimates by usage volume

Assumptions:

- Training job ratio: 1 job per 10 web requests (adjust based on your usage patterns)
- Fixed costs included: $2/month for infrastructure
- Web service costs are negligible (~$0.002 per request)
- Snowflake costs are separate and depend on your warehouse configuration

3.2 Cost Comparison: Benchmark vs Production

Jobs per Month	Benchmark Cost	Production Cost
10	$4	$15-22
25	$7	$35-52
50	$12	$69-102
100	$22	$135-202
250	$52	$335-502
500	$102	$667-1,002

Table 3: Monthly costs for different job volumes (includes $2 fixed costs)

4 Cost Breakdown

4.1 Fixed Monthly Costs

Total Fixed: ~$2/month

Service	Cost/Month	Notes
GCS Storage	$0.50-$2.00	Depends on data retention
Secret Manager	$0.36	6 secrets × $0.06
Cloud Scheduler	$0.30	Covered by free tier
Artifact Registry	$0.50	Container image storage

Table 4: Fixed infrastructure costs

4.2 Variable Costs

Per Training Job:

- Benchmark (2K × 5): $0.20
- Production (10K × 5): $1.33-$2.00
- Large Production (10K × 10): $2.67-$4.00

Per Web Request:

- Web service: ~$0.002 (negligible at typical volumes)
- Secret access: ~$0.00003 per request

Storage Growth:

- ~2GB per training result
- Lifecycle policies move data to cheaper storage after 30/90 days
- Nearline: $0.010/GB/month (after 30 days)
- Coldline: $0.004/GB/month (after 90 days)

4.3 Key Cost Drivers

1. **Training Jobs (90-99% of costs):**
   - Compute resources (CPU + memory) for 80-120 minutes
   - Number of jobs per month is the primary cost driver
   - Production workloads cost 6.7-10× more than benchmark per job
2. **Storage (Minor):**
   - Base storage: ~80GB for historical data
   - Growth: ~2GB per training run
   - Lifecycle policies reduce costs by 50-80% over time
3. **Web Service (Negligible):**

- Minimal cost due to short request durations
- `min_instances=0` eliminates idle costs

4. **Snowflake (Separate):**
   - Billed separately by Snowflake
   - Depends on warehouse size and query volume
   - 70% cache hit rate reduces warehouse usage

5 Cost Optimization Strategies

5.1 Immediate Cost Savings

1. **Use Benchmark Runs for Testing:**
   - 5× cheaper than production runs ($0.20 vs $1.33-$2.00)
   - 6-10× faster (12 min vs 80-120 min)
   - Ideal for experimentation and parameter tuning
2. **Set min_instances=0:**
   - Saves ˜$43/month in idle costs
   - Trade-off: 10-30 second cold start delay
   - Recommended for most deployments
3. **Implement Lifecycle Policies:**
   - Move data to Nearline after 30 days (50% cost reduction)
   - Move data to Coldline after 90 days (80% cost reduction)
   - Automatically applied to GCS bucket

5.2 Advanced Optimizations

1. **Adjust Resource Allocation:**
   - Monitor actual memory usage
   - Consider reducing from 32GB to 16GB if usage is low
   - 15% cost savings if memory can be reduced
2. **Compress Results:**
   - Reduce storage and egress costs by 50%
   - Requires code changes in R scripts
   - Minimal impact on total costs (storage is minor driver)
3. **Optimize Snowflake Caching:**
   - Current 70% cache hit rate already saves significant costs
   - Further optimization possible with query patterns
   - Snowflake costs are separate from GCP

6 Cost Calculator

6.1 Simple Cost Estimation Formula

$$\text{Monthly Cost} = \text{Fixed} + (\text{Jobs} \times \text{Cost per Job}) \tag{1}$$

Where:

- Fixed = \$2-\$3/month (infrastructure)
- Jobs = Number of training jobs per month
- Cost per Job = \$0.20 (benchmark) or \$1.67 avg (production)

Examples:

- 50 benchmark jobs/month: \$2 + (50 × \$0.20) = **\$12/month**
- 50 production jobs/month: \$2 + (50 × \$1.67) = **\$86/month**
- 100 production jobs/month: \$2 + (100 × \$1.67) = **\$169/month**
- 200 production jobs/month: \$2 + (200 × \$1.67) = **\$336/month**

6.2 Detailed Cost Estimation

For more precise estimates, consider:

1. **Training Job Mix:**
   - Percentage of benchmark vs production runs
   - Average duration based on your data complexity
   - Number of trials per job (5 vs 10)
2. **Usage Patterns:**
   - Peak vs steady-state usage
   - Seasonal variations in modeling needs
   - Team size and concurrent users
3. **Storage Requirements:**
   - Data retention policies
   - Archive older results to cheaper storage
   - Estimate based on 2GB per training result
4. **Snowflake Costs:**
   - Warehouse size (SMALL, MEDIUM, LARGE)
   - Query frequency and cache hit rate
   - Data transfer volumes

7 Regional Pricing

7.1 GCP Region Selection

Recommended Regions:

Region	Location	Notes
`europe-west1`	Belgium	Low cost, EU compliance
`europe-west4`	Netherlands	Low cost, EU compliance
`us-central1`	Iowa	Lowest US cost
`us-east1`	South Carolina	Low US cost

Table 5: Recommended GCP regions for deployment

Important Considerations:

- Choose region closest to your Snowflake instance for optimal performance
- EU regions required for GDPR compliance if processing EU data
- Pricing varies by region (typically ±10-20% from `europe-west1`)
- Network egress charges apply for data leaving the region

8 Pricing References

8.1 Official GCP Pricing

- **Cloud Run:** `https://cloud.google.com/run/pricing`
- **Cloud Storage:** `https://cloud.google.com/storage/pricing`
- **Secret Manager:** `https://cloud.google.com/secret-manager/pricing`
- **Artifact Registry:** `https://cloud.google.com/artifact-registry/pricing`
- **Cloud Scheduler:** `https://cloud.google.com/scheduler/pricing`

8.2 Current Infrastructure Configuration

Web Service:

- CPU: 2 vCPU
- Memory: 4GB
- Min instances: 0 (cost-optimized)
- Max instances: 10

Training Jobs:

- CPU: 8 vCPU
- Memory: 32GB
- Timeout: 6 hours
- Max retries: 1
- Actual cores used: 8 (full utilization achieved)

8.3 Verified Performance Data

All cost estimates in this document are based on:

- **Benchmark runs:** Verified January 9, 2026 (12.0 minutes with 8 cores)
- **Production runs:** Actual observed durations of 80-120 minutes
- **Infrastructure:** 8 vCPU, 32GB RAM, full core utilization
- **Region:** `europe-west1` (Belgium)

9 FAQ

9.1 Common Cost Questions

Q: Why don't production runs scale linearly from benchmark?

A: Production runs (10K × 5 iterations) are 5× larger than benchmark (2K × 5) but take 6.7-10× longer (80-120 min vs 12 min) due to non-linear scaling and overhead factors in the Robyn algorithm.

Q: Can I reduce costs by using fewer vCPUs?

A: While 4 vCPU would be cheaper per second, the job would take 2× longer, resulting in similar total cost due to per-second billing. 8 vCPU provides optimal balance of speed and cost.

Q: What's the cost difference between min_instances=0 and min_instances=1?

A: `min_instances=1` costs ~$43/month for always-on service but eliminates cold starts. `min_instances=0` saves this cost but adds 10-30 second startup delay.

Q: How much does Snowflake add to total costs?

A: Snowflake costs are separate and depend on your warehouse configuration. With 70% cache hit rate, only 30% of queries hit Snowflake. Estimate ~$0.10-0.20 per query on SMALL warehouse.

Q: Are there free tier benefits?

A: Yes, Cloud Scheduler jobs are free (first 3 jobs), and Cloud Run has a small monthly free tier. These are included in the $2/month fixed cost estimate.

Q: What causes cost variations month-to-month?

A: Primary variation comes from number of training jobs executed. Storage costs grow gradually (2GB per job), but this is minor compared to compute costs.

9.2 Cost Predictability

Highly Predictable:

- Training job costs ($0.20 or $1.33-$2.00 per job)
- Fixed infrastructure ($2/month)
- Per-second billing eliminates surprises

Variables to Monitor:

- Number of training jobs per month (primary driver)
- Storage accumulation over time (minor impact)
- Snowflake warehouse usage (separate billing)

Recommended Budget Alerts:

- Set at 50% and 90% of expected monthly spend
- Monitor job counts weekly
- Review storage growth monthly

10 Minimum Technical Requirements

10.1 Required Skills and Knowledge

The technical team maintaining this application should have:

10.1.1 Essential Skills

- **Google Cloud Platform:**
  - Basic understanding of Cloud Run, GCS, and IAM
  - Ability to navigate GCP Console
  - Understanding of billing and cost management
- **Infrastructure as Code:**
  - Basic Terraform knowledge for infrastructure changes
  - Ability to read and modify Terraform configuration files
- **Version Control:**
  - Git and GitHub workflows
  - Understanding of CI/CD concepts
- **Container Technology:**
  - Basic Docker concepts
  - Understanding of container registries

10.1.2 Recommended Skills

- **Programming Languages:**
  - Python (for Streamlit application modifications)
  - R (for Robyn MMM customizations)
- **Data Warehouse:**
  - Snowflake query optimization
  - SQL for data extraction
- **Monitoring and Debugging:**
  - Cloud Logging for troubleshooting
  - Performance monitoring and optimization

10.2 Required Tools

All team members should have access to:

Tool	Version	Purpose
Google Cloud SDK	Latest	GCP CLI operations
Terraform	$\geq 1.5.0$	Infrastructure management
Docker	Latest	Container testing (optional)
Git	Latest	Version control
Python	≥ 3.11	Local development (optional)

Table 6: Required development tools

10.3 Access Requirements

10.3.1 Google Cloud Platform

- **For Monitoring:** Viewer role
- **For Deployments:** Editor or specific roles:
  - Cloud Run Admin
  - Storage Admin
  - Secret Manager Admin
  - Service Account Admin
- **For Debugging:** Logs Viewer, Monitoring Viewer

10.3.2 GitHub Repository

- **For Development:** Write access
- **For Releases:** Maintain or Admin access
- **For Secrets Management:** Admin access

10.3.3 Snowflake

- Read access to source data tables
- Access to a dedicated warehouse for queries
- Appropriate role (not ACCOUNTADMIN in production)

10.4 Team Structure Recommendations

For successful deployment and maintenance, we recommend:

Minimum Team:

- 1 DevOps Engineer (GCP, Terraform, CI/CD expertise)
- 1 Data Scientist/Analyst (familiar with MMM and Robyn)

Recommended Team:

- 1-2 DevOps Engineers (deployment, maintenance, monitoring)
- 2-3 Data Scientists/Analysts (model development, analysis)
- 1 Data Engineer (Snowflake integration, data pipelines)

Time Commitment:

- Initial deployment: 3-5 hours (one-time)
- Ongoing maintenance: 2-4 hours/month
- Model development: Varies by business needs

10.5 Training and Onboarding

Recommended Training Topics:

1. **GCP Fundamentals (4-8 hours):**
   - Cloud Run architecture and deployment
   - IAM and service accounts
   - Cloud Storage and lifecycle policies
   - Billing and cost management
2. **Platform-Specific Training (2-4 hours):**
   - MMM Trainer web interface walkthrough
   - Training job configuration and execution
   - Results interpretation and visualization
   - Troubleshooting common issues
3. **Robyn MMM Framework (8-16 hours):**
   - MMM concepts and methodology
   - Robyn-specific features and parameters
   - Model interpretation and validation
   - Best practices for production use

Learning Resources:

- Google Cloud documentation and tutorials
- Robyn GitHub repository and documentation
- Platform-specific documentation in repository
- Community forums and support channels

10.6 Support Requirements

Internal Support:

- Designated DevOps contact for infrastructure issues
- Data science lead for modeling questions
- Documentation maintenance and updates

External Dependencies:

- GCP support plan (optional but recommended)

- Snowflake support for data warehouse issues
- Community support for Robyn framework questions

Recommended Support Plan:

- GCP Standard Support: $100/month minimum
- Response times: 4 hours for production issues
- 24/7 support availability
- Technical account management for larger deployments