

Contents

RADIANT v4.17.0 - Deployed System Guide	2
Table of Contents	2
1. System Overview	2
1.1 Platform Summary	2
1.2 Environment Tiers	3
1.3 Region Configuration	3
2. Infrastructure Components	3
2.1 VPC Architecture	3
2.2 Security Groups	4
2.3 IAM Roles	4
3. Deployment Architecture	4
3.1 CDK Stacks	4
3.2 Deployment Commands	4
3.3 Environment Variables	5
4. Service Endpoints	5
4.1 API Gateway Endpoints	5
4.2 Health Check Endpoints	5
4.3 Internal Endpoints	5
5. Database Operations	6
5.1 Aurora Configuration	6
5.2 Connection Management	6
5.3 Row-Level Security	6
5.4 Database Migrations	6
6. Lambda Functions	7
6.1 Function Inventory	7
6.2 Monitoring Lambda	7
6.3 Cold Start Optimization	7
7. AI Provider Integration	8
7.1 External Providers	8
7.2 Self-Hosted Models (SageMaker)	8
7.3 Model Thermal States	8
8. Monitoring & Observability	8
8.1 CloudWatch Dashboards	8
8.2 Key Metrics	9
8.3 Alerting	9
8.4 Log Groups	9
9. Security Configuration	9
9.1 Encryption	9
9.2 Secrets Management	9
9.3 WAF Rules	10
10. Backup & Recovery	10
10.1 Automated Backups	10
10.2 Recovery Procedures	10
10.3 Disaster Recovery	11
11. Maintenance Procedures	11
11.1 Routine Maintenance	11

11.2 Scaling Operations	11
11.3 Deployment Checklist	11
12. Troubleshooting	11
12.1 Common Issues	11
12.2 Emergency Procedures	12
Version History	12

RADIANT v4.17.0 - Deployed System Guide

Operations and infrastructure documentation for the deployed RADIANT platform

Last Updated: {{BUILD_DATE}} Version: {{RADIANT_VERSION}}

Table of Contents

1. System Overview
 2. Infrastructure Components
 3. Deployment Architecture
 4. Service Endpoints
 5. Database Operations
 6. Lambda Functions
 7. AI Provider Integration
 8. Monitoring & Observability
 9. Security Configuration
 10. Backup & Recovery
 11. Maintenance Procedures
 12. Troubleshooting
-

1. System Overview

1.1 Platform Summary

RADIANT is a multi-tenant AWS SaaS platform deployed across multiple AWS services:

Component	AWS Service	Purpose
Compute	Lambda	Serverless API handlers
Database	Aurora PostgreSQL	Primary data store
Cache	ElastiCache Redis	Session & response caching
Storage	S3	File storage, artifacts
Auth	Cognito	User authentication
API	API Gateway	REST & WebSocket APIs
AI	SageMaker, Bedrock	Model hosting
CDN	CloudFront	Static asset delivery
DNS	Route 53	Domain management

1.2 Environment Tiers

Environment	Purpose	Domain
Production	Live customer traffic	api.{{RADIANT_DOMAIN}}
Staging	Pre-release testing	staging-api.{{RADIANT_DOMAIN}}
Development	Development & testing	dev-api.{{RADIANT_DOMAIN}}

1.3 Region Configuration

Region	Role	Services
us-east-1	Primary	All services
us-west-2	DR Failover	Database replica, S3 replication
eu-west-1	EU Data Residency	Optional for GDPR compliance

2. Infrastructure Components

2.1 VPC Architecture

RADIANT VPC (10.0.0.0/16)

Public Subnets

us-east-1a	us-east-1b	us-east-1c
10.0.1.0/24	10.0.2.0/24	10.0.3.0/24
NAT Gateway	NAT Gateway	

Private Subnets (App)

us-east-1a	us-east-1b	us-east-1c
10.0.11.0/24	10.0.12.0/24	10.0.13.0/24
Lambda ENIs	Lambda ENIs	Lambda ENIs

Private Subnets (Data)

us-east-1a	us-east-1b	us-east-1c
10.0.21.0/24	10.0.22.0/24	10.0.23.0/24

Aurora, RDS ElastiCache Backup

2.2 Security Groups

Security Group	Inbound Rules	Purpose
sg-lambda	None (outbound only)	Lambda functions
sg-aurora	5432 from sg-lambda	Database access
sg-redis	6379 from sg-lambda	Cache access
sg-alb	443 from 0.0.0.0/0	Load balancer

2.3 IAM Roles

Role	Attached Policies	Used By
radiant-lambda-role	AWSLambdaVPCAccess, SecretsManager, S3, SageMaker	Lambda functions
radiant-api-gw-role	CloudWatchLogs, Lambda invoke	API Gateway
radiant-sagemaker-role	SageMakerFullAccess, S3	SageMaker endpoints

3. Deployment Architecture

3.1 CDK Stacks

Stack	Resources	Dependencies
NetworkingStack	VPC, Subnets, NAT	None
SecurityStack	IAM, KMS, Security Groups	Networking
DataStack	Aurora, ElastiCache, S3	Security
AuthStack	Cognito User Pool	Data
ApiStack	API Gateway, Lambda	Auth, Data
AIStack	SageMaker, Bedrock config	Api
MonitoringStack	CloudWatch, Alarms	All
AdminStack	Admin Dashboard hosting	Api

3.2 Deployment Commands

```
# Deploy all stacks to development
```

```
npm run deploy:dev
```

```
# Deploy to staging with approval
```

```
npm run deploy:staging
```

```
# Deploy to production (requires approval)
npm run deploy:prod

# Deploy specific stack
cd packages/infrastructure
cdk deploy RadiantApiStack --context environment=prod
```

3.3 Environment Variables

Variable	Description	Source
AURORA_CLUSTER_ARN	Database cluster ARN	CDK output
AURORA_SECRET_ARN	Database credentials	Secrets Manager
COGNITO_USER_POOL_ID	Auth pool ID	CDK output
REDIS_URL	Cache connection string	CDK output
S3_BUCKET	Storage bucket name	CDK output

4. Service Endpoints

4.1 API Gateway Endpoints

Endpoint	Method	Description
/v1/chat	POST	Send chat message
/v1/chat/stream	WebSocket	Streaming responses
/v1/models	GET	List available models
/v1/usage	GET	Usage statistics
/admin/v1/*	Various	Admin API

4.2 Health Check Endpoints

Endpoint	Expected Response	Checks
/health	{"status":"healthy"}	Lambda running
/health/db	{"status":"connected"}	Aurora connection
/health/cache	{"status":"connected"}	Redis connection
/health/providers	Provider status array	AI provider APIs

4.3 Internal Endpoints

Service	Endpoint	Port
Aurora PostgreSQL	radiant-cluster.xxx.us-east-1.rds.amazonaws.com	5432
ElastiCache Redis	radiant-cache.xxx.cache.amazonaws.com	6379
SageMaker Runtime	Via AWS SDK	-

5. Database Operations

5.1 Aurora Configuration

Setting	Value
Engine	PostgreSQL 15.4
Instance Type	Serverless v2
Min ACUs	2 (dev), 8 (prod)
Max ACUs	16 (dev), 128 (prod)
Storage	Auto-scaling to 128 TB
Encryption	AES-256 (KMS)

5.2 Connection Management

```
-- Check active connections
SELECT count(*) FROM pg_stat_activity;

-- View connection by application
SELECT application_name, count(*)
FROM pg_stat_activity
GROUP BY application_name;

-- Terminate idle connections
SELECT pg_terminate_backend(pid)
FROM pg_stat_activity
WHERE state = 'idle'
AND query_start < now() - interval '10 minutes';
```

5.3 Row-Level Security

All tenant data is protected by RLS:

```
-- Verify RLS is enabled
SELECT tablename, rowsecurity
FROM pg_tables
WHERE schemaname = 'public' AND rowsecurity = true;

-- Set tenant context (done automatically by Lambda)
SET app.current_tenant_id = 'tenant-uuid';

-- Queries automatically filter by tenant
SELECT * FROM users; -- Only returns current tenant's users
```

5.4 Database Migrations

```
# Run pending migrations
cd packages/infrastructure
npm run migrate:up
```

```
# Rollback last migration
```

```
npm run migrate:down
```

```
# View migration status
```

```
npm run migrate:status
```

6. Lambda Functions

6.1 Function Inventory

Function	Memory	Timeout	Trigger
radiant-api	1024 MB	30s	API Gateway
radiant-brain	2048 MB	60s	API Gateway
radiant-webhooks	512 MB	30s	EventBridge
radiant-scheduler	512 MB	300s	EventBridge (cron)
radiant-batch	3008 MB	900s	SQS

6.2 Monitoring Lambda

```
# View recent invocations
```

```
aws lambda get-function \
  --function-name radiant-api \
  --query 'Configuration.[FunctionName,MemorySize,Timeout,LastModified]'
```

```
# Check concurrent executions
```

```
aws cloudwatch get-metric-statistics \
  --namespace AWS/Lambda \
  --metric-name ConcurrentExecutions \
  --dimensions Name=FunctionName,Value=radiant-api \
  --start-time "$(date -u -d '1 hour ago' +%Y-%m-%dT%H:%M:%SZ) \
  --end-time $(date -u +%Y-%m-%dT%H:%M:%SZ) \
  --period 300 \
  --statistics Maximum
```

6.3 Cold Start Optimization

Strategy	Implementation
Provisioned Concurrency	10 instances for radiant-api
Keep-Warm	CloudWatch scheduled ping every 5 min
Connection Pooling	RDS Proxy for database
Lazy Loading	Defer non-critical imports

7. AI Provider Integration

7.1 External Providers

Provider	Models	Authentication
OpenAI	GPT-4, GPT-3.5	API Key
Anthropic	Claude 3, Claude 2	API Key
Google	Gemini Pro, PaLM 2	Service Account
Cohere	Command, Embed	API Key
Mistral	Mistral Large, Medium	API Key

7.2 Self-Hosted Models (SageMaker)

Model	Instance Type	Thermal Default
LLaMA-2-70B	ml.g5.48xlarge	COLD
Stable Diffusion XL	ml.g5.2xlarge	COLD
Whisper Large	ml.g4dn.xlarge	COLD
CodeLlama-34B	ml.g5.12xlarge	COLD

7.3 Model Thermal States

```
# Check endpoint status
aws sagemaker describe-endpoint \
  --endpoint-name radiant-llama-70b

# Scale endpoint to warm
aws sagemaker update-endpoint-weights-and-capacities \
  --endpoint-name radiant-llama-70b \
  --desired-weights-and-capacities '[{"VariantName":"AllTraffic","DesiredInstanceCount":1}]'

# Scale to zero (cold)
aws sagemaker update-endpoint-weights-and-capacities \
  --endpoint-name radiant-llama-70b \
  --desired-weights-and-capacities '[{"VariantName":"AllTraffic","DesiredInstanceCount":0}]'
```

8. Monitoring & Observability

8.1 CloudWatch Dashboards

Dashboard	Metrics
Platform Overview	Request rate, latency, errors
Database Performance	Connections, ACUs, query time
AI Provider Health	Success rate, latency by provider
Billing & Usage	Credits consumed, revenue

8.2 Key Metrics

Metric	Warning	Critical
API Error Rate	> 1%	> 5%
P99 Latency	> 5s	> 15s
Aurora ACU Usage	> 80%	> 95%
Lambda Errors	> 10/min	> 50/min

8.3 Alerting

List configured alarms

```
aws cloudwatch describe-alarms \  
  --alarm-name-prefix radiant
```

View alarm state

```
aws cloudwatch describe-alarm-history \  
  --alarm-name radiant-high-error-rate
```

8.4 Log Groups

Log Group	Retention	Content
/aws/lambda/radiant-api	30 days	API request logs
/aws/lambda/radiant-brain	30 days	Model routing logs
/radiant/audit	365 days	Security audit logs
/radiant/billing	2 years	Billing events

9. Security Configuration

9.1 Encryption

Data	Encryption
Database at rest	AES-256 (KMS CMK)
S3 objects	AES-256 (SSE-S3)
Secrets	KMS envelope encryption
API traffic	TLS 1.3

9.2 Secrets Management

List RADIANT secrets

```
aws secretsmanager list-secrets \  
  --filters Key=name,Values=radiant
```

Rotate database credentials

```
aws secretsmanager rotate-secret \  
  --secret-id radiant-db-credentials
```

```

--secret-id radiant/aurora-credentials

# View secret metadata
aws secretsmanager describe-secret \
  --secret-id radiant/openai-api-key

```

9.3 WAF Rules

Rule	Action	Purpose
Rate Limiting	Block	> 1000 req/min per IP
SQL Injection	Block	SQLi pattern detection
XSS	Block	Cross-site scripting
Geo Blocking	Block	Sanctioned countries

10. Backup & Recovery

10.1 Automated Backups

Resource	Frequency	Retention
Aurora Snapshots	Daily	35 days
S3 Objects	Continuous	Versioned
DynamoDB	Point-in-time	35 days
Secrets	Auto-versioned	30 versions

10.2 Recovery Procedures

Database Point-in-Time Recovery:

```

aws rds restore-db-cluster-to-point-in-time \
  --source-db-cluster-identifier radiant-cluster \
  --db-cluster-identifier radiant-cluster-restored \
  --restore-to-time 2024-01-15T10:00:00Z

```

S3 Object Recovery:

```

# List object versions
aws s3api list-object-versions \
  --bucket radiant-storage \
  --prefix tenant/123/

# Restore specific version
aws s3api copy-object \
  --bucket radiant-storage \
  --copy-source radiant-storage/file.txt?versionId=xxx \
  --key file.txt

```

10.3 Disaster Recovery

RPO	RTO	Strategy
1 hour	4 hours	Cross-region Aurora replica, S3 CRR

11. Maintenance Procedures

11.1 Routine Maintenance

Task	Frequency	Procedure
Security patches	Weekly	Automated via CDK
Log rotation	Daily	Automated
Database vacuum	Weekly	Automated
Certificate renewal	60 days	ACM auto-renewal

11.2 Scaling Operations

```
# Scale Aurora
aws rds modify-db-cluster \
  --db-cluster-identifier radiant-cluster \
  --serverless-v2-scaling-configuration MinCapacity=16,MaxCapacity=64

# Update Lambda memory
aws lambda update-function-configuration \
  --function-name radiant-api \
  --memory-size 2048
```

11.3 Deployment Checklist

- ☐ Run tests in staging
 - ☐ Review CloudWatch for anomalies
 - ☐ Announce maintenance window
 - ☐ Deploy with `--require-approval`
 - ☐ Verify health checks
 - ☐ Monitor error rates for 30 min
 - ☐ Update status page
-

12. Troubleshooting

12.1 Common Issues

High Latency:

```
# Check database performance
aws cloudwatch get-metric-statistics \
```

```

--namespace AWS/RDS \
--metric-name DatabaseConnections \
--dimensions Name=DBClusterIdentifier,Value=radiant-cluster

# Check Lambda duration
aws logs filter-log-events \
  --log-group-name /aws/lambda/radiant-api \
  --filter-pattern "REPORT Duration"

```

Connection Errors:

```

# Verify security group rules
aws ec2 describe-security-groups --group-ids sg-xxx

# Check VPC endpoints
aws ec2 describe-vpc-endpoints \
  --filters Name=vpc-id,Values=vpc-xxx

```

12.2 Emergency Procedures

Rollback Deployment:

```

# List recent deployments
aws cloudformation list-stacks \
  --stack-status-filter UPDATE_COMPLETE

# Rollback to previous version
cdk deploy --rollback

```

Disable Problematic Feature:

```

# Update feature flag
aws ssm put-parameter \
  --name /radiant/features/new-feature \
  --value "false" \
  --overwrite

```

Version History

Version	Date	Changes
4.17.0	{{BUILD_DATE}}	Initial system guide

This documentation is automatically generated as part of the RADIANT build process.