Contents

ADR-009: Admin-Configurable Infrastructure Tiers

Status

Accepted

Context

Cato infrastructure costs range dramatically based on scale: - **DEV**: ~$350/month for development and testing - **STAGING**: ~$20-50K/month for pre-production - **PRODUCTION**: ~$700-800K/month for 10MM+ users

We need a system that allows admins to switch between infrastructure tiers at runtime without recompilation, with automatic resource provisioning and cleanup.

Requirements

1. **Runtime Configurable** — No recompilation. Admin changes a setting, infrastructure responds.
2. **Auto-Scale Up** — When tier increases, provision required resources automatically.
3. **Auto-Scale Down + Cleanup** — When tier decreases, terminate/delete unused resources to stop billing.
4. **Admin-Editable Configs** — All tier configurations (instance types, counts, etc.) are editable.
5. **Safety Guards** — Confirmation dialogs, cooldown periods, and audit logging.

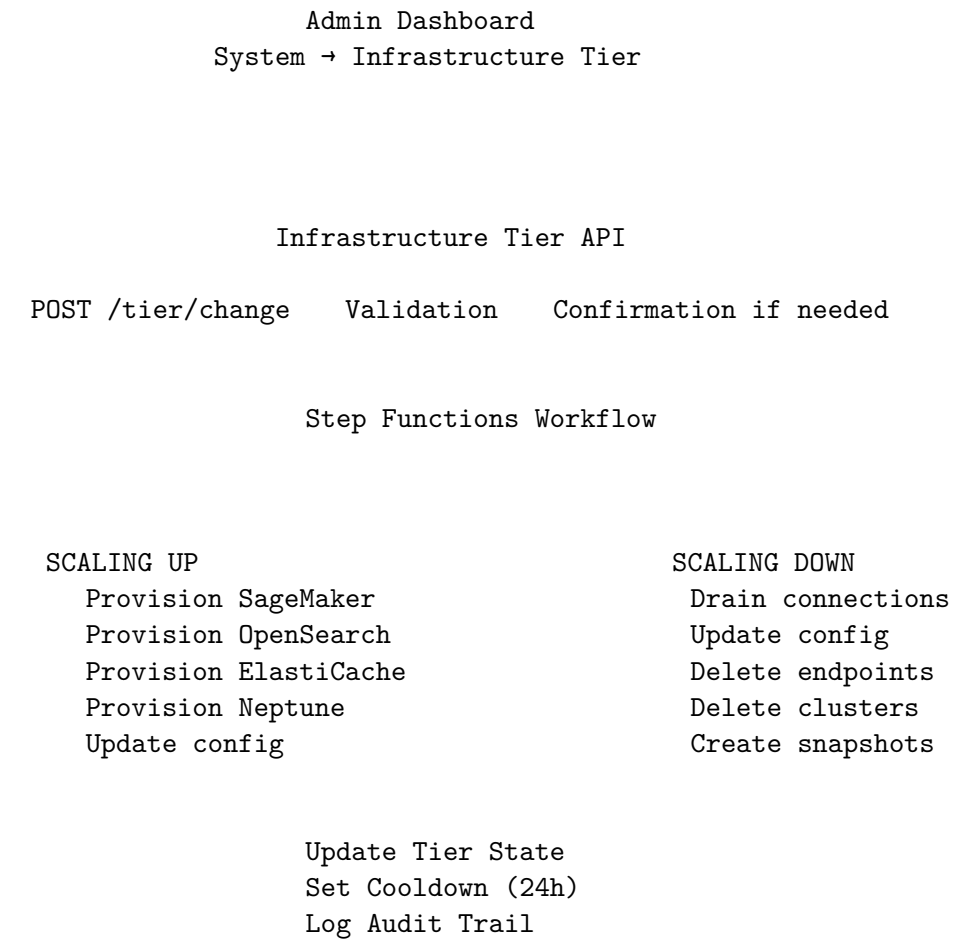
Decision

Implement a **3-tier infrastructure system** with full admin editability:

Tier Configurations

Tier	Est. Cost	SageMaker	OpenSearch	ElastiCache	Neptune
DEV	$350/mo	0-1 ml.g5.xlarge (scale-to-zero)	t3.small (provisioned)	Serverless	Serverless
STAGING	$35K/mo	2-20 ml.g5.2xlarge	r6g.large (provisioned)	cache.r7g.large	Serverless
PRODUCTION	$70K/mo	50-300 ml.g5.2xlarge	Serverless (50-500 OCUs)	cache.r7g.xlarge x6	db.r6g.2xlarge x3

Architecture



Database Schema

```
-- Current tier state per tenant
cato_infrastructure_tier
    current_tier (DEV|STAGING|PRODUCTION)
    target_tier (during transition)
    transition_status (STABLE|SCALING_UP|SCALING_DOWN|FAILED)
    cooldown_hours (default: 24)
    next_change_allowed_at
    estimated_monthly_cost
    actual_mtd_cost

-- Editable tier configurations
cato_tier_config
    tier_name
    display_name
    description
    estimated_monthly_cost
    sagemaker_shadow_self_* (instance type, min/max, scale-to-zero)
    opensearch_* (type, instance type, count)
    elasticache_* (type, node type, count)
    neptune_* (type, instance class, count)
    budget_* (monthly curiosity limit, daily exploration cap)
    features, limitations (JSON arrays for UI)

-- Audit trail
cato_tier_change_log
    from_tier, to_tier
    direction (SCALING_UP|SCALING_DOWN)
    status, duration
    changed_by, reason
    resources_provisioned, resources_cleaned_up
    errors (if any)
```

Safety Guards

1. **24-hour cooldown** between tier changes (configurable)
2. **Confirmation required** for PRODUCTION tier (both up and down)
3. **Audit logging** of all changes with who, when, why
4. **Super admin bypass** for cooldown in emergencies
5. **Rollback capability** if provisioning fails

Consequences

Positive

- Admins can switch tiers in ~5-15 minutes
- Resources are automatically cleaned up (no orphaned billing)
- Full visibility into costs before changes

- All configurations are editable without code changes
- Complete audit trail

Negative

- Step Functions adds complexity
- Tier transitions require ~5-15 minutes
- Risk of partial failures during transition

Mitigations

- Automatic rollback on provisioning failure
- Snapshots created before resource deletion
- Monitoring and alerts during transitions

Implementation

Files Created

File	Purpose
`migrations/121_infrastructure_tiers.sql`	Database schema
`lambda/shared/services/cato/infrastructure-tier.service.ts`	Core service
`lambda/admin/infrastructure-tier.ts`	Admin API endpoints
`apps/admin-dashboard/app/(dashboard)/system/infrastructure/page.tsx`	Admin UI

API Endpoints

Endpoint	Method	Description
`/tier`	GET	Get current tier status
`/tier/compare`	GET	Get tier comparison for UI
`/tier/configs`	GET	Get all tier configurations
`/tier/configs/:name`	GET/PUT	Get/update specific tier config
`/tier/change`	POST	Request tier change
`/tier/confirm`	POST	Confirm tier change
`/tier/transition-status`	GET	Get transition progress
`/tier/cooldown`	PUT	Update cooldown hours

UI Location

```
Admin Dashboard
  System
      Infrastructure Tier
          Current Status (tier, cost, status)
          Tier Selection (cards with cost comparison)
          Configuration Editor (edit any tier's resources)
          Change History (audit log)
```

Cost Optimization Notes

DEV Tier Optimizations

- SageMaker scale-to-zero when idle
- OpenSearch Provisioned (not Serverless $700 minimum)
- ElastiCache Serverless (cheap for low traffic)
- Neptune Serverless (1.0 NCU minimum)

PRODUCTION Tier

- Consider 3-year Savings Plans (64% discount on SageMaker)
- Bedrock Batch API for night-mode curiosity (50% discount)
- OpenSearch Serverless for true auto-scaling

References

- AWS Pricing Calculator
- SageMaker Pricing
- OpenSearch Serverless Pricing