Savings Report

Sample report — data platform, 2-account org

Generated August 5, 2026

Estimated monthly savings

$5,576.43

All figures are estimates based on on-demand us-east-1 pricing.

AI & ML — subtotal $3,525.77/mo

$1,050.00/mo — Bedrock on-demand spend above Provisioned Throughput breakeven (121212343434): Evaluate Bedrock Provisioned Throughput (1-month or 6-month model units) for the steady baseline; keep on-demand for burst overflow. Verify per-model utilization in the Bedrock console before committing.

$921.70/mo — On-demand SageMaker training that fits managed Spot (888899990000/us-east-1): Enable managed Spot training (EnableManagedSpotTraining) with S3 checkpointing for interruption tolerance. Typical savings run 60-70% vs on-demand; keep on-demand only for jobs that cannot checkpoint or must meet a hard deadline.

$881.01/mo — Sparse-traffic endpoint ep-scoring-batch (1x ml.g5.xlarge, 86% zero-traffic days): Convert to an asynchronous inference endpoint (queues requests, scales instances to zero between bursts) or a scheduled batch transform if scoring is periodic. Keep real-time only if interactive latency is genuinely required.

$673.06/mo — Idle endpoint ep-old-experiment (2x ml.m5.2xlarge, 0 invocations): Delete the endpoint (model artifacts remain in S3 and it can be recreated in minutes). If it is needed intermittently, move to a serverless or asynchronous inference endpoint that scales to zero.

All findings

Est. $/mo	Finding	Category	Effort	Recommendation
$1,050.00	Bedrock on-demand spend above Provisioned Throughput breakeven (121212343434)	ai_ml	Migration	Evaluate Bedrock Provisioned Throughput (1-month or 6-month model units) for the steady baseline; keep on-demand for burst overflow. Verify per-model utilization in the Bedrock console before committing.
$921.70	On-demand SageMaker training that fits managed Spot (888899990000/us-east-1)	ai_ml	Migration	Enable managed Spot training (EnableManagedSpotTraining) with S3 checkpointing for interruption tolerance. Typical savings run 60-70% vs on-demand; keep on-demand only for jobs that cannot checkpoint or must meet a hard deadline.
$881.01	Sparse-traffic endpoint ep-scoring-batch (1x ml.g5.xlarge, 86% zero-traffic days)	ai_ml	Migration	Convert to an asynchronous inference endpoint (queues requests, scales instances to zero between bursts) or a scheduled batch transform if scoring is periodic. Keep real-time only if interactive latency is genuinely required.

Est. $/mo	Finding	Category	Effort	Recommendation
$717.96	Uncovered EC2 on-demand spend (888899990000)	commitments	Config change	Purchase a 1-year no-upfront Compute Savings Plan sized to the steady baseline (start at ~70% of the uncovered spend after actioning the idle/rightsizing findings above, then ratchet up). Compute SPs follow the workload across instance families, sizes, and regions, so they stay safe through migrations.
$673.06	Idle endpoint ep-old-experiment (2x ml.m5.2xlarge, 0 invocations)	ai_ml	Config change	Delete the endpoint (model artifacts remain in S3 and it can be recreated in minutes). If it is needed intermittently, move to a serverless or asynchronous inference endpoint that scales to zero.
$620.00	Inter-AZ data transfer spend (888899990000)	network	Rearchitecture	Identify the chattiest cross-AZ flows (VPC Flow Logs / Cost Explorer usage-type drilldown), then co-locate tight client-server pairs in one AZ and enable topology/AZ-aware routing where the stack supports it. Keep multi-AZ for real failover paths -- this targets accidental cross-AZ chatter.
$345.00	S3 bucket datalake-raw-events has no lifecycle policy (50,000 GB Standard)	storage	Config change	Add a lifecycle rule transitioning objects to S3 Intelligent-Tiering (safe default -- no retrieval fees, per-object monitoring only) or to Standard-IA after 30 days for known-cold data. Expire incomplete multipart uploads while you're in there.
$180.00	NAT gateway data-processing spend (121212343434)	network	Migration	Add gateway VPC endpoints for S3 and DynamoDB (free) and interface endpoints for other high-volume AWS services so that traffic stops paying the NAT per-GB toll. Check VPC Flow Logs for the top talkers before choosing endpoints.
$124.83	Idle RDS instance rds-metastore-dev (db.m5.large, 0 connections)	database	Config change	Stop the instance (RDS restarts stopped instances after 7 days -- schedule the stop weekly, or snapshot and delete if the database is retired). Savings shown are instance hours only; storage keeps billing while stopped.

Est. $/mo	Finding	Category	Effort	Recommendation
$40.00	gp2 volume vol-kafka1 (2000 GB) -> gp3	storage	Config change	Migrate to gp3 via ModifyVolume (online, no downtime). gp3 baseline is 3000 IOPS / 125 MBps; provision extra only if the workload exceeded gp2 burst performance.
$10.22	Previous-generation EC2 instance i-emr-old (r4.xlarge -> r5.xlarge)	compute	Migration	Migrate to r5.xlarge (same architecture, drop-in for most workloads). If the stack can run on ARM, the equivalent Graviton family is cheaper still -- worth testing after the x86 move.
$9.00	Log group /data/emr-yarn never expires (600 GB stored)	observability	Config change	Set a retention policy (30-90 days for application logs; export to S3 with a lifecycle rule first if compliance needs longer). Existing bytes past the new retention age out automatically.
$3.65	Unassociated Elastic IP 203.0.113.60 (eipalloc-ml-idle)	network	Config change	Release the address if it is no longer needed, or associate it with an instance/ENI if it is being held deliberately. Check for DNS references to this IP first -- a released Elastic IP cannot be recovered.

Full findings detail (including resource identifiers) is available in findings.json alongside this report.