

MSK STOCK retention: 28-day validated cost model

A decision-ready comparison of existing MSK Serverless, 7D/8D retention, two canned consumer groups, Provisioned Standard, and tiered storage across NV and SG.

RECOMMENDATION: Keep the existing NV and SG MSK Serverless clusters. Configure STOCK for 8 days when the canned consumers must start from exactly seven days ago: retention.ms=691200000 and retention.bytes=-1. Do not create a Provisioned cluster solely for this replay requirement.

Decision metric	7D target	8D recommended	Meaning
Retention-only monthly delta	\$70.98	\$82.81	Storage above the current 1D baseline
With two full groups / month	\$438.85	\$450.68	Storage plus two complete reads in each region
Recommended 7-day billing view	\$100.99	\$103.72	Steady-state cost prorated to 168 hours
Annual run-rate	\$5,266.16	\$5,408.12	Conditional; excludes taxes and unrelated base charges

Why 8D instead of 7D

A consumer starting exactly seven days behind a seven-day topic is reading at the deletion boundary. Eight days preserves the requested seven-day history plus one day for startup and catch-up. If the approved requirement is only 'keep the latest seven days' and replay margin is not required, configure seven days instead.

Important: the 8D recommendation costs only \$11.83/month more in retained storage across both regions in this model.

Selected 8D monthly split

Region	Retention storage	Two full readers	Monthly total	Source quality
NV us-east-1	\$38.29	\$166.37	\$204.66	Measured 28-day topic total
SG ap-southeast-1	\$44.52	\$201.50	\$246.02	Estimated from 7-day SG/NV ratio
Both regions	\$82.81	\$367.87	\$450.68	Conditional combined estimate

1. Source-data audit

Four monthly CSV files were supplied. They are exact duplicate downloads: each is 6,009,531 bytes with SHA-256 14f542793d821c7d8a48fcd907546d439eb787b9d1e873b0f922f938deb818c8. Only the unsuffixed file is used. It contains 8,064 contiguous five-minute observations - exactly 28 days - and 50 us-east-1 expression series.

Audit item	Result	Decision impact
Time coverage	2026/07/27 23:20:00 to 2026/08/24 23:15:00 (UTC+6 export)	Four full weeks; captures week-to-week variability
Missing timestamps	0	No gap-filling required
NV STOCK	e1 topic expression	Canonical measured series
NV STOCK_EXTENDED	e1 topic expression	Separate workload; not an SG counterpart
SG month series	Not present	SG result must remain estimated

Critical aggregation correction

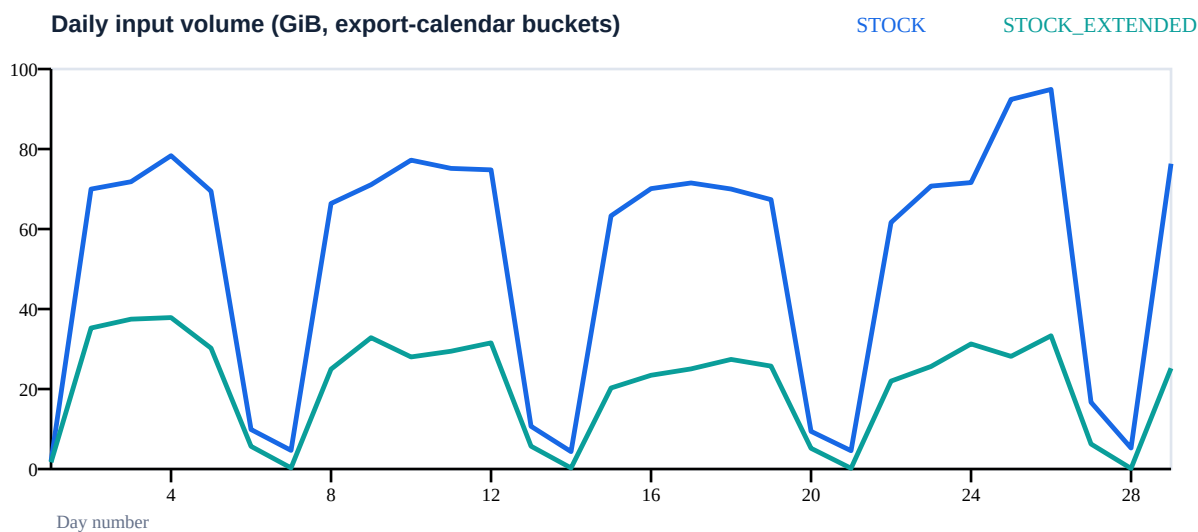
Across the legacy seven-day overlap, the new e1 STOCK series is 2.000x the explicit Average series; STOCK_EXTENDED is 3.000x. AWS documents that Serverless PerSec metrics are emitted per partition and should be combined with SUM. The new e1 values therefore replace the old raw Average values and are not multiplied again. The exact e1 widget expression still needs to be exported for final audit proof.

The older \$186/month document used the raw Average series and us-east-1 rates for SG. Its read-amplification ratios remain useful, but its absolute storage and reader costs are understated.

Measured topic totals

Metric	STOCK	STOCK_EXTENDED	Combined NV
28-day input	1,531.5 GiB	600.4 GiB	2,131.9 GiB
Average day	54.7 GiB	21.4 GiB	76.1 GiB
7D logical footprint	382.9 GiB	150.1 GiB	533.0 GiB
730-hour projection	1,663.7 GiB	652.2 GiB	2,315.9 GiB

2. What STOCK and STOCK_EXTENDED are doing



STOCK_EXTENDED is 39.2% of STOCK input volume and has a different timing pattern. It should not be silently included in the STOCK retention change. The supplied data contains no SG STOCK_EXTENDED series, so there is no defensible SG match.

Traffic statistic	STOCK	STOCK_EXTENDED
Mean input	0.680 MB/s	0.266 MB/s
P95 input	1.630 MB/s	0.769 MB/s
Peak input	2.949 MB/s	1.032 MB/s
Share of STOCK	100.0%	39.2%

Week-to-week variability

The four exact seven-day STOCK blocks are 370.5, 376.8, 354.5, and 429.8 GiB. The last block is 21.3% above the third block. Using the complete 28-day average is materially safer than selecting one convenient week.

SG coverage

Every monthly full label is us-east-1. The SG month is not hidden inside topics whose names contain MSK-SG. The SG monthly curve in the dashboard is explicitly estimated by multiplying the NV curve by the measured seven-day SG/NV ratio of 96.9%. This preserves the observed regional relationship but remains medium confidence until an SG month export is supplied.

3. Cost calculation

The task is modeled as a marginal change to existing Serverless clusters. Existing cluster-hour, partition-hour, and producer data-in charges remain unchanged. New consumer groups add Data-Out. Serverless storage is logical metered storage; AWS-managed replication is not multiplied into the cost.

Formula	Definition
Measured GiB	$\text{Sum}(\text{BytesInPerSec} \times 300 \text{ seconds}) / 2^{30}$
1D to D-day storage delta	Average daily GiB $\times (D - 1) \times$ regional storage rate
Reader Data-Out	Projected monthly input GiB $\times$ full consumer groups $\times$ regional out rate
Seven-day billing cost	Monthly run-rate $\times 168 / 730$
Provisioned primary	Daily GiB $\times D \times \text{RF} \times$ headroom $\times$ EBS rate

Current AWS unit prices used

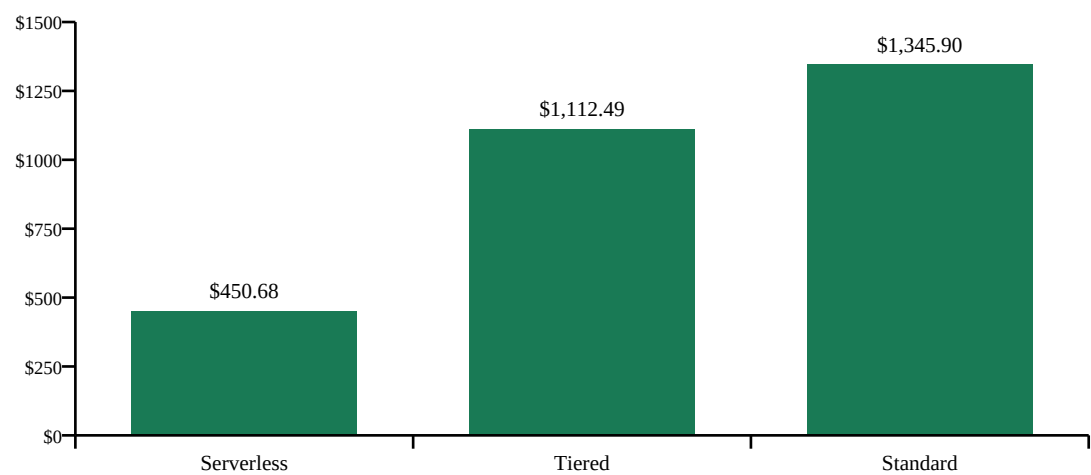
Price	NV us-east-1	SG ap-southeast-1
Serverless cluster-hour	\$0.7500	\$0.9375
Serverless partition-hour	\$0.001500	\$0.001875
Serverless data-in / GB	\$0.1000	\$0.1250
Serverless data-out / GB	\$0.0500	\$0.0625
Serverless storage / GB-month	\$0.1000	\$0.1200
Standard m7g.large / broker-hour	\$0.2040	\$0.2550
Standard EBS / GB-month	\$0.1000	\$0.1200
Tiered storage / GB-month	\$0.0600	\$0.0652

Scenario matrix - both regions

Target	New full groups / region	Retention storage	Total / month	Total / 7 days
7D	0	\$70.98	\$70.98	\$16.33
7D	1	\$70.98	\$254.91	\$58.66
7D	2	\$70.98	\$438.85	\$100.99
8D	0	\$82.81	\$82.81	\$19.06
8D	1	\$82.81	\$266.74	\$61.39
8D	2	\$82.81	\$450.68	\$103.72

4. Architecture recommendation

Monthly cost comparison - both regions, 8D, two full readers



Architecture	Monthly lower bound	What is included	Decision
Reuse existing Serverless	\$450.68	8D storage + two full reads, both regions	Recommended
New Provisioned Tiered	\$1,112.49	3 m7g.large/region, RF3, local EBS, remote tier, retrieval	Not economical
New Provisioned Standard	\$1,345.90	3 m7g.large/region, RF3, 20% storage headroom	Not economical

Provisioned figures are lower bounds

They omit producer/copier development, MSK Replicator, cross-region transfer, private connectivity, migration effort, and operations. Even before those items, the existing Serverless option is the lowest-cost design. A full-cluster Serverless-to-Provisioned migration is a separate decision and cannot be made from a file that contains NV topic input but no cluster-wide output, partition inventory, broker utilization, or SG month.

Replication factor

Do not expose an RF multiplier for Serverless: AWS manages leader plus replicas and bills the logical Serverless storage quantity. Apply RF only to Provisioned Standard physical EBS planning. The comparison uses RF3 and 20% headroom.

MSK Replicator

Replicator is not a substitute for retention. It adds an hourly charge, per-GB processing, source Data-Out, target Data-In/storage, cross-region transfer, and private-connectivity costs. Use it only when a second cluster is required for topology or disaster recovery, and price it from ReplicatorBytesInPerSec rather than STOCK producer input.

5. Capacity and operational guardrails

The 28-day NV STOCK peak topic input is 2.949 MB/s. MSK Serverless limits are 5 MB/s ingress and 10 MB/s egress per partition, but the exported e1 topic expression does not prove hot-partition distribution or one-minute peaks. Go-live needs per-partition one-minute validation.

Guardrail	Current evidence	Required action
Retention time	7D history requested; canned reader starts exactly one week back	Use 8D: retention.ms=691200000
Retention size	Max rolling 8D NV STOCK = 489.6 GiB aggregate	Set retention.bytes=-1; confirm skew and current config
Serverless default size	250 GiB per partition	Time retention can lose to size retention if not disabled
Ingress capacity	Topic-total 5-minute peak only	Export 1-minute per-partition Sum/Max and throttling
Egress capacity	Legacy 5-minute scenario reached ~8.26 MB/s NV	Validate hot partition and canary before go-live
Consumer topology	Two applications named; group IDs not supplied	Count distinct full-topic consumer groups, not pods

Required read-only evidence before approval

1. CloudWatch widget source JSON to prove the e1 query and statistic.
2. A full SG month export for STOCK and STOCK_EXTENDED using the same topic-total expression.
3. BytesInPerSec, BytesOutPerSec, and MessagesInPerSec at one-minute resolution for both regions.
4. kafka-topics describe for partition count and RF; kafka-configs describe for retention.ms and retention.bytes.
5. Consumer group IDs and confirmation of whether each group reads the full topic.
6. MSK Replicator direction, topic list, and copyTopicConfigurations setting.
7. CUR/Cost Explorer quantities for Serverless In-Bytes, Out-Bytes, StorageHours, ClusterHours, and PartitionHours.

6. Sources, assumptions, and confidence

Primary AWS sources

<https://aws.amazon.com/msk/pricing/>

<https://pricing.us-east-1.amazonaws.com/offers/v1.0/aws/AmazonMSK/20260729211408/us-east-1/index.json>

<https://pricing.us-east-1.amazonaws.com/offers/v1.0/aws/AmazonMSK/20260729211408/ap-southeast-1/index.json>

<https://docs.aws.amazon.com/msk/latest/developerguide/serverless-monitoring.html>

<https://docs.aws.amazon.com/msk/latest/developerguide/serverless-config.html>

<https://docs.aws.amazon.com/msk/latest/developerguide/limits.html>

Model assumptions

Assumption	Value	Confidence / effect
Monthly period	730 hours	AWS standard planning month
SG monthly STOCK	NV 28-day curve x 0.968920	Medium; replace with SG month export
New reader	One distinct consumer group reads every STOCK byte once	High only after group topology is confirmed
Existing Serverless fixed charges	Excluded from marginal comparison	Correct if no new cluster/partition is required
Provisioned broker floor	3 x kafka.m7g.large in each region	Lower bound; workload validation still required
Provisioned RF/headroom	RF3 and 20%	Planning assumption
Byte unit	GiB = bytes / 2 ³⁰	Matches prior analysis; reconcile final invoice to CUR usage units

Confidence statement

High confidence: NV 28-day input totals, current regional price-list rates, option ranking, and the need to separate STOCK_EXTENDED. Medium confidence: combined NV+SG dollars because the SG month series and exact e1 source expression are absent. Conditional confidence: two-reader cost because consumer group IDs and full-topic read behavior are not supplied.

Final decision

Reuse existing Serverless. Set 8D only because the consumer must retrieve data from exactly seven days earlier; otherwise 7D is the literal retention target. Set retention.bytes=-1. Keep STOCK_EXTENDED out unless the applications explicitly subscribe to it. Do not create a Provisioned cluster for this requirement.