Cluster Choice Decision Matrix

November 05, 2025

Pick the right Databricks compute in one minute. Default to Serverless for 80 percent of workloads. Carve out Dedicated only when it clearly wins on strict latency, long stable usage, or specialized configs.

Decision Tree

- Q1: Is the workload ad hoc, bursty, or nightly but short? → Choose Serverless Standard.
- Q2: Does startup time affect an SLA? → If yes, try Serverless Performance for that job. If no, stay on Serverless Standard.
- Q3: Is the pipeline long running and stable, or near continuous? → Profile Dedicated jobs cluster.
- Q4: Do you need custom runtime, pinned libraries, private networking, GPUs, or special drivers? → Use Dedicated with a small, documented footprint.
- Q5: Is the cost difference within 10 percent? → Choose the lower-ops option. Saved engineer time becomes real money.

Latency Cut Lines

- Start latency: under 1 minute for interactive, under 3 minutes for batch.
- End-to-end latency: must land inside the defined business window (for example, 10 minutes).

Mini Lab Latency Measure, Don’t Guess

- Create two jobs with the same notebook and data slice.
- Test three modes: Serverless Standard, Serverless Performance, Dedicated (min workers 1).
- Run 3 times per mode in a quiet window. Record Start latency and Run duration.
- Guardrails: avoid cached reads if testing I/O; keep scheduling consistent; note if Dedicated terminated between runs.

Mini Lab Cost Know What You Pay For

- Pull DBUs per run from Billing for each job.
- Multiply by your DBU rate to get cost per run.
- Multiply by runs per day and days per month.
- Add storage and egress if applicable.

Cold Start Playbook

Serverless

- Use Performance mode for a small set of latency-sensitive jobs.
- Pre-warm with a tiny scheduled run before the morning spike.
- Split heavy init so it does not block the first query.

Dedicated

- Use instance pools to reduce VM provision time.

- Tune auto termination to keep clusters warm only during burst windows.
- Bake dependencies into the image. Avoid heavy installs at job start.

Top 5 Cost Traps

- No tags or ownership. Fix with enforced tags in cluster policies.
- Autoscale max set huge. Cap by tier and review usage.
- Long auto termination windows that keep clusters warm between short jobs.
- Heavy library installs on every run instead of pre-baked images.
- Silent retries and timeouts that multiply DBUs.

Policy Pack and Hygiene

- Cluster policies: cap max workers, restrict instance types, enforce tags owner, cost_center, env.
- Budgets and alerts: monthly team budgets; alert on 7-day DBU drift above 20 percent.
- Cleanup: auto-terminate dev after inactivity; sweep orphaned tables and temp paths weekly; rotate logs to cheaper storage after 14 days.

Scoreboard Template

Metric	Serverless Standard	Serverless Performance	Dedicated Cold	Dedicated Warm
Start latency				
Run duration				
DBUs per run				

Verdict: TTFB winner, steady throughput winner, cost winner by pattern.

One Minute Quick Start

- Default to Serverless Standard.
- If SLA sensitive, try Serverless Performance for that one job.
- If long and stable, profile a small Dedicated jobs cluster.
- Run the mini labs. If the cost gap is within 10 percent, choose the lower-ops option.
- Apply the policy pack to lock in savings.