Twitter Thread by Yan Cui is making the AppSync Masterclass

Yan Cui is making the AppSync Masterclass
@theburningmonk

I've gotten a few questions about Aurora Serverless v2 preview, so here's what I've learnt so far. Please feel free to chime in if I've missed anything important or got any of the facts wrong.

Alright, here goes the ■...

Q: does it replace the existing Aurora Serverless offering?
A: no, it lives side-by-side with the existing Aurora Serverless, which will still be available to you as "v1".

Q: Aurora Serverless v1 takes a few seconds to scale up, that's too much for our use case where we get a lot of spikes. Is that the same with v2?
A: no, v2 scales up in milliseconds, during preview the max ACU is only 32 though

Q: is the cold start for Aurora Serverless v2 still a few seconds?
A: yes, unfortunately...
Q: so if you want to avoid cold starts, what's the minimum ACU you have to run?
A: minimum ACU with v2 is 0.5

Q: does v2 still scale up in double increments, e.g. 4 ACU -> 8 ACU?
A: no, it scales up in increments of 0.5 ACUs, so it's a much tighter fit for your workload, so you'll waste less money on over-provisioned ACUs

Q: is there anything I can do with v2 that I can't do with v1?
A: yes, v2 supports all the Aurora features, including those that v1 is missing, such as global database, IAM auth and Lambda triggers

Q: wait, but it's twice as much per ACU!
A: yes, but v1 requires a lot of over-provisioning because it doubles ACU each time and takes 15 mins to scale down. v2 scales in 0.5 ACU increments and scales down in < 1 min. AND you get all the Aurora features!

US East (N. Virginia)	v2 (preview)
Aurora Capacity Unit	$0.12 per ACU Hour

Scale up latency	Instant scaling to hundreds-of-thousands of transactions per second
Scale down latency	Up to 15X faster (<1 minute)
Starting capacity	0.5 ACU
Capacity granularity	Fine-grained, with increments as small as 0.5 ACU
Read Replicas	Up to 15 Aurora Replicas for read scalability
Multi-AZ and SLA	Distribute read replicas in separate AZs for high availability. Refer Amazon Aurora SLA for details.
Aurora Global Database	Sub-Second Data Access in Any Region and Cross-Region Disaster Recovery. Refer Aurora Global Database for details.

eliminates over-provisioned ACUs from v1

Q: can you use "provisioned" and "serverless" instances in the same Aurora cluster?
A: yes you can! cool, right!?

Q: is data API supported on v2?
A: not in the preview, I'm guessing it'll be there in GA

Q: if using from Lambda, do I need to use RDS Proxy to manage the connections to the cluster? Data API kinda mitigated that for v1..
A: yes, you probably should, until data API is enabled on v2, otherwise, more connections = more ACUs, it can run you into trouble

did I miss anything? plz feel free to chimp in

@jeremy_daly has written a nice summary post on this too, with a nice experiment on the scaling behaviour of v2, so check it out if you haven't already https://t.co/DIGylN0HFX