Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 38 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 38
Topic #: 1
[All Professional Data Engineer Questions]

MJTelco Case Study -

Company Overview -
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
✑ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
✑ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments `" development/test, staging, and production `" to meet the needs of

running experiments, deploying new features, and serving production customers.

Business Requirements -
✑ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
✑ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
✑ Provide reliable and timely access to data for analysis from distributed research workers
✑ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -
✑ Ensure secure and efficient transport and storage of telemetry data
✑ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
✑ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
✑ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

A. The zone

B. The number of workers

C. The disk size per worker

D. The maximum number of workers

Show Suggested Answer

by jvg637 at *March 15, 2020, 1:16 p.m.*

Comments

Type your comment...

Submit

Radhika7983 Highly Voted 3 years, 11 months ago

The correct answer is D. Please look for the details in below
https://cloud.google.com/dataflow/docs/guides/specifying-exec-params
We need to specify and set execution parameters for cloud data flow .

Also, to enable autoscaling, set the following execution parameters when you start your pipeline:

--autoscaling_algorithm=THROUGHPUT_BASED
--max_num_workers=N
The objective of autoscaling streaming pipelines is to minimize backlog while maximizing worker utilization and throughput, and quickly react to spikes in load. By enabling autoscaling, you don't have to choose between provisioning for peak load and fresh results. Workers are added as CPU utilization and backlog increase and are removed as these metrics come down. This way, you're paying only for what you need, and the job is processed as efficiently as possible.

upvoted 27 times

jvg637 Highly Voted 4 years, 7 months ago

D. The maximum number of workers answers to the scale question

upvoted 26 times

cqrm3n Most Recent 3 months ago

Selected Answer: D

The answer is D because Google Dataflow is serverless and auto scales based on demand. To allow it to scale up compute power dynamically, we need to set the maximum number of workers.

upvoted 1 times

I__SHA1234567 7 months, 3 weeks ago

Selected Answer: D

Cloud Dataflow dynamically scales the number of workers based on the amount of data being processed and the processing requirements. By updating the maximum number of workers, you allow Dataflow to scale up the compute power as needed to handle the workload efficiently. This ensures that the pipeline can adapt to changes in data volume and processing demands.

upvoted 2 times

rtcpost 1 year ago

Selected Answer: D

D. The maximum number of workers

By increasing the maximum number of workers, you ensure that Cloud Dataflow can scale its compute power to handle the increased data processing load efficiently.

upvoted 2 times

vaga1 1 year, 5 months ago

Selected Answer: D

dataflow auto-scales, then if it is not scaling is because it has reached the maximum number of workers that have been set.

upvoted 2 times

abi01a 1 year, 6 months ago

A is the correct answer. Dataflow is Serverless. Specify your Region, autoscaling and other 'knobing' activities that are 'under the hood' will be taken care for you. Remember the company cannot afford to staff an Operations team to monitor data feeds so rely on ...

upvoted 2 times

bha11111 1 year, 7 months ago

Selected Answer: D

this is correct

upvoted 2 times

GCPpro 1 year, 9 months ago

D . is the correct answer

upvoted 1 times

jkh_goh 1 year, 9 months ago

Answer A provided is definitely wrong. Who comes up with these answers?

upvoted 1 times

Ender_H 2 years, 1 month ago

Selected Answer: D

Correct Answer: D

A: The zone has nothing to do with scaling computer power.

B: The key word here is, "Scale its compute power up AS REQUIRED", with this answer, the number of workers would immediately scale the computer power.

C: we need to scale compute power, not storage

D: is the correct answer, changing the Number of Maximum workers will allow Dataflow to add up to that number of workers if required.

https://cloud.google.com/dataflow/docs/reference/pipeline-options#resource_utilization

upvoted 2 times

sraakesh95 2 years, 9 months ago

Selected Answer: D

@Radhika7983

upvoted 2 times

medeis_jar 2 years, 10 months ago

Selected Answer: D

The correct answer is D.
https://cloud.google.com/dataflow/docs/guides/specifying-exec-params
We need to specify and set execution parameters for cloud data flow .

Also, to enable autoscaling, set the following execution parameters when you start your pipeline:

--autoscaling_algorithm=THROUGHPUT_BASED
--max_num_workers=N

upvoted 2 times

maurodipa 2 years, 11 months ago

Answer is A: Dataflow is serverless, so no need to specify neither the number of workers, nor the max number of workers.
https://cloud.google.com/dataflow

upvoted 5 times

Jarek7 1 year, 6 months ago

Have you ever use it? You pay for workers processing, so you specify max number of workers. Here is the doc:
https://cloud.google.com/sdk/gcloud/reference/dataflow/jobs/run

upvoted 1 times

anji007 3 years ago

Ans: D

upvoted 1 times

sumanshu 3 years, 3 months ago

Vote for D

upvoted 3 times

Lodu_Lalit 3 years, 7 months ago

D, thats because scalability is directly corerlated to max number of workers, size determines the speed of functioning

upvoted 3 times

Load full discussion...