Google Discussions

EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 65 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 65
Topic #: 1
[All Professional Data Engineer Questions]

You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataproc job.

B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.

C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataprep job.

D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 0 using a custom script.

Show Suggested Answer

by madhu1171 at *March 13, 2020, 1:45 p.m.*

Comments

Type your comment...

Submit

jvg637 Highly Voted 5 years, 1 month ago
real-valued can not be null N/A or empty, have to be “0”, so it has to be B.
upvoted 40 times

upvoted 40 times

[Removed] Highly Voted 5 years, 1 month ago

Should be B

upvoted 16 times

monyu Most Recent 1 month, 1 week ago

Selected Answer: B

Usually, None values are converted to 0 in data cleaning and preparation process. The key point here is, we don't require any other tool than DataPrep to identify and modify the value

upvoted 1 times

Parandhaman_Margan 1 month, 3 weeks ago

Selected Answer: D

Cloud Dataflow is ideal for scalable data processing and allows for real-time transformations.
Logistic regression requires numerical (real-valued) inputs, and null values cannot remain as they are.

upvoted 1 times

Erg_de 6 months ago

Selected Answer: D

Option D:
Using null value conversion to 0 is the most correct practice for this case. Accompanying it with a script allows us to implement the necessary logic to handle null cases properly, adapting to the model while maintaining data integrity.

upvoted 2 times

certs4pk 5 months ago

y use a data flow job when it can b done via data prep (much simpler & straight forward, less resource intensive)..

upvoted 1 times

AjoeT 1 year, 1 month ago

Selected Answer: B

B. Dataprep has the feature to convert it into 0.

upvoted 2 times

niru12376 1 year, 2 months ago

0 is still a value, which can add bias in the model and the model will take that into account while making predictions so 'none'

upvoted 1 times

Nandababy 1 year, 4 months ago

Why not D? keyword is Monitor, B would replace all empty fields and also cause unintended bias.

upvoted 1 times

Nandababy 1 year, 4 months ago

However, Sergiomujica is right. If we need to prepare data using a casual method then its B "Dataprep".

upvoted 1 times

sergiomujica 1 year, 8 months ago

The questions says "You need to prepare data using a casual method ", thats dataprep and values should be 0 so the right answer is B

upvoted 1 times

Mathew106 1 year, 9 months ago

Selected Answer: B

No brainer. We need a real value and Dataprep is made for this. Dataflow is mainly for pre-processing before BigQuery ingests the data.

upvoted 2 times

theseawillclaim 1 year, 9 months ago

Selected Answer: B

Dataprep is made for this kind of stuff, no reason to use a streaming service such as Dataflow.

upvoted 2 times

Oleksandr0501 2 years ago

Selected Answer: B

gpt:Cloud Dataprep is a data preparation service that can be used to transform, clean and shape data in a visually interactive way. It provides an easy-to-use interface to find and replace null values.

Cloud Dataflow is a fully-managed service for executing data processing pipelines, which allows for parallel execution of data processing tasks. However, it requires more expertise to set up and operate than Cloud Dataprep, and is usually used for

more complex data processing needs.

Therefore, option B is the most suitable approach for the given requirements.

upvoted 1 times

samdhimal 2 years, 3 months ago

Seems to me like Answers are both B and D.
B is faster to implement while D takes time.
Doesnt mean that it's wrong though. I m not sure why everyone has picked just B. Why not D? D works and does the same job. And also having custom script provides more flexibility and control over the data processing tasks and it allows you to handle missing values in a more flexible and efficient way.

upvoted 2 times

rajm893 1 year, 11 months ago

The "casual way" or easy way to convert to to 0 is using Dataprep job rather than using the custom script.

upvoted 2 times

AmmarFasih 1 year, 11 months ago

A simple rule. Whenever any service is available by GCP for a task, always recommend to use GCP service over any other.

upvoted 1 times

GCPpro 2 years, 3 months ago

B is the correct answer.

upvoted 1 times

AzureDP900 2 years, 4 months ago

Answer is Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.

Key phrases are "casual method", "need to replace null with real values", "logistic regression". Logistic regression works on numbers. Null need to be replaced with a number. And Cloud dataprep is best casual tool out of given options.

upvoted 3 times

DGames 2 years, 4 months ago

Selected Answer: B

real value 0

upvoted 1 times

byash1 3 years, 3 months ago

Selected Answer: B

It is B

upvoted 2 times

Load full discussion...