

🔗 Google Discussions



Exam Professional Data Engineer All Questions

View all questions & answers for the Professional Data Engineer exam

[Go to Exam](#)



EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 179 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 179

Topic #: 1

[\[All Professional Data Engineer Questions\]](#)

You are building a real-time prediction engine that streams files, which may contain PII (personal identifiable information) data, into Cloud Storage and eventually into BigQuery. You want to ensure that the sensitive data is masked but still maintains referential integrity, because names and emails are often used as join keys.

How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the PII data is not accessible by unauthorized individuals?

- A. Create a pseudonym by replacing the PII data with cryptogenic tokens, and store the non-tokenized data in a locked-down bucket.
- B. Redact all PII data, and store a version of the unredacted data in a locked-down bucket.
- C. Scan every table in BigQuery, and mask the data it finds that has PII.
- D. Create a pseudonym by replacing PII data with a cryptographic format-preserving token.

[Show Suggested Answer](#)

by [AWSandeep](#) at *Sept. 2, 2022, 9:18 p.m.*

Comments

Type your comment...

[Submit](#)

 **zellick** Highly Voted 2 years, 5 months ago

Selected Answer: D

D is the answer.

<https://cloud.google.com/dlp/docs/pseudonymization#supported-methods>

Format preserving encryption: An input value is replaced with a value that has been encrypted using the FPE-FFX encryption algorithm with a cryptographic key, and then prepended with a surrogate annotation, if specified. By design, both the character set and the length of the input value are preserved in the output value. Encrypted values can be re-identified using the original cryptographic key and the entire output value, including surrogate annotation.

   upvoted 7 times

 **Pime13** Most Recent 3 months, 4 weeks ago

Selected Answer: D


D. Create a pseudonym by replacing PII data with a cryptographic format-preserving token.

A: Storing non-tokenized data in a locked-down bucket is not a standard practice and might not provide the necessary security.

B: Redacting all PII data and storing an unredacted version separately can lead to data management complexities and potential security risks.

C: Scanning every table in BigQuery and masking data post-ingestion is less efficient and might not ensure real-time protection.

   upvoted 1 times

 **ToiToi** 6 months ago

Selected Answer: D

Why other options are not as suitable:

A (Cryptogenic tokens and locked-down bucket): While this provides some protection, storing the non-tokenized data in a separate bucket adds complexity and risk.

B (Redaction and locked-down bucket): Redaction removes sensitive data entirely, which might limit its usefulness for analysis and other purposes.

C (Scanning and masking in BigQuery): This approach might be less efficient than masking the data during the streaming process before it reaches BigQuery.

   upvoted 1 times

 **SamuelTsch** 6 months, 1 week ago

Selected Answer: D

I would like to go with D. If the data could be deidentified later by the token, why should we store the data in a locked-down bucket?

   upvoted 1 times

 **GCP001** 1 year, 3 months ago

Selected Answer: D

D> Looks more suitable as it will handle Referential integrity. <https://cloud.google.com/dlp/docs/pseudonymization>

   upvoted 1 times

 **pss111423** 1 year, 5 months ago

answer A

<https://cloud.google.com/dlp/docs/transformations-reference> Replaces an input value with a token, or surrogate value, of the same length using AES in Synthetic Initialization Vector mode (AES-SIV). This transformation method, unlike format-preserving tokenization, has no limitation on supported string character sets, generates identical tokens for each instance of an identical input value, and uses surrogates to enable re-identification given the original encryption key.

   upvoted 2 times

 **akg001** 1 year, 8 months ago

Selected Answer: D

D is correct.

   upvoted 1 times

 **cetanx** 1 year, 10 months ago

Selected Answer: B

I've also asked to GPT but I had to remind the hard condition "names and emails are often used as join keys". It changed the answer to "B" after 3rd iteration.



masking all PII data may not satisfy the requirement of using names and emails as join keys, as the data is obfuscated and cannot be used for accurate join operations.

In this approach, you would redact or remove the sensitive PII data, such as names and emails, from the dataset that will be

used for real-time processing and analysis. The redacted data would be stored in the primary dataset to ensure that sensitive information is not accessible.

Additionally, you would create a copy of the original dataset with the PII data still intact, but this copy would be stored in a locked-down bucket with restricted access. This ensures that authorized individuals who need access to the unredacted data for specific purposes, such as join operations, can retrieve it from the secured location.

   upvoted 2 times

  **cetanx** 1 year, 10 months ago

made a typo up there, it has to be A

   upvoted 2 times

  **Oleksandr0501** 2 years ago

gpt:

The recommended approach for using the Cloud Data Loss Prevention API (DLP API) to protect sensitive PII data while maintaining referential integrity is to create pseudonyms by replacing the PII data with cryptographic format-preserving tokens.

This approach ensures that sensitive data is not accessible by unauthorized individuals, while still preserving the format and length of the original data, which is essential for maintaining referential integrity.

Replacing PII data with cryptogenic tokens, as mentioned in option A, is not recommended because cryptogenic tokens are not necessarily format-preserving, and this could affect the accuracy of data joins.

Therefore, option D is the best approach for using the DLP API to ensure that PII data is not accessible by unauthorized individuals while still maintaining referential integrity.

   upvoted 1 times

  **loicrichonnier** 1 year, 12 months ago

You shouldn't use ChatGPT as a source, the data used are not up to date and for such complex question a predicting text chatbot can help but, it's better to refer to the google documentation.


   upvoted 5 times

  **Oleksandr0501** 1 year, 12 months ago

that`s why i always mark "gpt", when copy from there... i know, thx

also, it might be A. Or D... Confusing question.

   upvoted 1 times

  **Prudvi3266** 2 years ago

Selected Answer: D

here catch is "cryptographic" key

   upvoted 3 times

  **musumusu** 2 years, 2 months ago

Answer D,

key word - "referential integrity" use format preserve option, it keeps same length of the value and last four digits of your value in column

   upvoted 1 times

  **tunstila** 2 years, 3 months ago

Selected Answer: D

The answer is D



   upvoted 1 times

  **nkit** 2 years, 4 months ago

Selected Answer: D

I believe "Format preserving token" in option D makes it easier choice for me



   upvoted 1 times

  **PrashantGupta1616** 2 years, 4 months ago

Selected Answer: D

D looks right

   upvoted 1 times

  **jkhong** 2 years, 4 months ago

Selected Answer: A



Question is super tricky, B and C are not the answers since they do not maintain referential integrity.

For D, it does preserve the length of input. But since we are only concerned with referencing during joins, there is no point of

maintaining the length anyway. Also, characters must be encoded as ASCII, this means that the name and email must be within the 256 character set. which is further limited to the alphabet characters, i.e. 94 characters.
(<https://cloud.google.com/dlp/docs/transformations-reference#crypto>)

Names nowadays do not just have ASCII characters but unicode as well, so D will not necessarily work all the time.

   upvoted 2 times

  **Atnafu** 2 years, 4 months ago

D is the answer

Pseudonymization is a de-identification technique that replaces sensitive data values with cryptographically generated tokens.

Keywords: You want to ensure that the sensitive data is masked but still maintains referential integrity

Part1- data is masked-Create a pseudonym by replacing PII data with a cryptographic token

Part 2 still maintains referential integrity- with a cryptographic format-preserving token

A Not an answer because

the locked-down button does not seem to google cloud word

   upvoted 4 times

  **julio** 2 years, 1 month ago

"button" is just a typo for "bucket"

   upvoted 1 times

  **dish11dish** 2 years, 5 months ago

Selected Answer: D

Though both option A nad D maintains referential integrity,question is why you wnat to keep untokenize data in GCS,best way is option D which even support Reversible feature which is not supported by option A refer chart in reference document.

reference:-

<https://cloud.google.com/dlp/docs/pseudonymization>

   upvoted 1 times

[Load full discussion...](#)



Platform

> Home

> All Exams

> Examtopics PRO

> Training Courses

