

[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 250 DISCUSSION

Actual exam question from Google's Professional Data Engineer

Question #: 250

Topic #: 1

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

Your company's data platform ingests CSV file dumps of booking and user profile data from upstream sources into Cloud Storage. The data analyst team wants to join these datasets on the email field available in both the datasets to perform analysis. However, personally identifiable information (PII) should not be accessible to the analysts. You need to de-identify the email field in both the datasets before loading them into BigQuery for analysts. What should you do?

- A. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud Data Loss Prevention (Cloud DLP) with masking as the de-identification transformations type.
2. Load the booking and user profile data into a BigQuery table.
- B. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud DLP with format-preserving encryption with FFX as the de-identification transformation type.
2. Load the booking and user profile data into a BigQuery table.
- C. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.
2. Create a policy tag with the email mask as the data masking rule.
3. Assign the policy to the email field in both tables. A
4. Assign the Identity and Access Management bigquerydatapolicy.maskedReader role for the BigQuery tables to the analysts.
- D. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.
2. Create a policy tag with the default masking value as the data masking rule.
3. Assign the policy to the email field in both tables.
4. Assign the Identity and Access Management bigquerydatapolicy.maskedReader role for the BigQuery tables to the analysts



[Show Suggested Answer](#)

by  scaenruy at Jan. 3, 2024, 3:35 p.m.

Comments

[Submit](#)  **lipa31** Highly Voted  1 year, 3 months agoSelected Answer: B

Format-preserving encryption (FPE) with FFX in Cloud DLP is a strong choice for de-identifying PII like email addresses. FPE maintains the format of the data and ensures that the same input results in the same encrypted output consistently. This means the email fields in both datasets can be encrypted to the same value, allowing for accurate joins in BigQuery while keeping the actual email addresses hidden.

   upvoted 15 times  **Smakyl79** Highly Voted  1 year, 3 months ago

As it states "You need to de-identify the email field in both the datasets before loading them into BigQuery for analysts" data masking should not be an option as the data would stored unmasked in BigQuery?

   upvoted 5 times  **Anudeep58** Most Recent  10 months, 3 weeks agoSelected Answer: B

Option A:

Masking: Simple masking might not preserve the uniqueness and joinability of the email field, making it difficult to perform accurate joins between datasets.

Option C and D:

Dynamic Data Masking: These options involve masking the email field dynamically within BigQuery, which does not address the requirement to de-identify data before loading into BigQuery. Additionally, dynamic masking does not prevent access to the actual email data before it is loaded into BigQuery, potentially exposing PII during the data ingestion process.

   upvoted 4 times  **chrissamharris** 11 months, 3 weeks agoSelected Answer: B

format-preserving encryption with FFX is required as the analysts want to perform JOINS

   upvoted 2 times  **JyoGCP** 1 year, 2 months agoSelected Answer: B

Option B

<https://cloud.google.com/sensitive-data-protection/docs/pseudonymization>

   upvoted 3 times  **ML6** 1 year, 2 months agoSelected Answer: B


A) masking = replace with a surrogate character like # or * = output not unique, so cannot apply joins

C and D) question specifies to de-identify BEFORE loading into BQ, whereas these options perform dynamic masking IN BigQuery.

Therefore, only valid option is B.

   upvoted 4 times  **Matt_108** 1 year, 3 months agoSelected Answer: C

Option C. The need is to just mask the data to Analyst, without modifying the underlying data. Moreover, it's stored on 2 separate tables and the analysts need to be able to perform joins based on the masked data. Dynamic masking is the right module and the right masking rule is email mask (https://cloud.google.com/bigquery/docs/column-data-masking-intro#masking_options) which guarantees the join capabilities join

   upvoted 1 times  **task_7** 1 year, 3 months agoSelected Answer: B

A wouldn't preserve the email format

C&D maskedReader roles still grant access to the underlying values.
the only option is B

👍 ↩ 🚩 upvoted 5 times

🗄 👤 **alfguemat** 1 year, 3 months ago

I don't know why preserve email format is necessary to perform the join. A could be valid.

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **dduenas** 1 year, 3 months ago

masking only replace by specific characters, doing the field not unique and not ready for joins.

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **Sofia98** 1 year, 3 months ago

Selected Answer: C

I will go for C, because there is a separate type of masking for emails, so whe to use the dafault?
https://cloud.google.com/bigquery/docs/column-data-masking-intro#masking_options

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **GCP001** 1 year, 3 months ago

Selected Answer: C

data masking with BQ is correct with email masking rule.
Ref - <https://cloud.google.com/bigquery/docs/column-data-masking-intro>

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **tibuenoc** 1 year, 2 months ago

should be correct if they want to access tables and it's not valid for datasets

👍 ↩ 🚩 upvoted 1 times

🗄 👤 **Jordan18** 1 year, 4 months ago

why not B?

👍 ↩ 🚩 upvoted 2 times

🗄 👤 **raaad** 1 year, 4 months ago

Selected Answer: C

- The reason option C works well is that dynamic data masking in BigQuery allows the underlying data to remain unaltered (thus preserving the ability to join on this field), while also preventing analysts from viewing the actual PII.
- The analysts can query and join the data as needed for their analysis, but when they access the data, the email field will be masked according to the policy tag, and they will only see the masked version.

👍 ↩ 🚩 upvoted 2 times

🗄 👤 **scaenruy** 1 year, 4 months ago

Selected Answer: D

D. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.
2. Create a policy tag with the default masking value as the data masking rule.
3. Assign the policy to the email field in both tables.
4. Assign the Identity and Access Management bigquerydatapolicy.maskedReader role for the BigQuery tables to the analysts

👍 ↩ 🚩 upvoted 1 times



Platform

> Home

> Examtopics PRO

> All Exams

> Training Courses

