Free Questions for Professional-Data-Engineer by go4braindumps

Shared by Watson on 15-04-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

You work for a large real estate firm and are preparing 6 TB of home sales data lo be used for machine learning You will use SOL to transform the data and use BigQuery ML lo create a machine learning model. You plan to use the model for predictions against a raw dataset that has not been transformed. How should you set up your workflow in order to prevent skew at prediction time?

Options:

A- When creating your model, use BigQuerys TRANSFORM clause to define preprocessing stops. At prediction time, use BigQuery's ML. EVALUATE clause without specifying any transformations on the raw input data.

B- When creating your model, use BigQuery's TRANSFORM clause to define preprocessing steps Before requesting predictions, use a saved query to transform your raw input data, and then use ML. EVALUATE

C- Use a BigOuery to define your preprocessing logic. When creating your model, use the view as your model training data. At prediction lime, use BigQuery's ML EVALUATE clause without specifying any transformations on the raw input data.

D- Preprocess all data using Dataflow. At prediction time, use BigOuery's ML. EVALUATE clause without specifying any further transformations on the input data.

Answer:

A

Explanation:

https://cloud.google.com/bigquery-ml/docs/bigqueryml-transform Using the TRANSFORM clause, you can specify all preprocessing during model creation. The preprocessing is automatically applied during the prediction and evaluation phases of machine learning

Question 2

Question Type: MultipleChoice

You are designing a data warehouse in BigQuery to analyze sales data for a telecommunication service provider. You need to create a data model for customers, products, and subscriptions All customers, products, and subscriptions can be updated monthly, but you must maintain a historical record of all dat

a. You plan to use the visualization layer for current and historical reporting. You need to ensure that the data model is simple, easy-to-use. and cost-effective. What should you do?

Options:

A- Create a normalized model with tables for each entity. Use snapshots before updates to track historical data

B- Create a normalized model with tables for each entity. Keep all input files in a Cloud Storage bucket to track historical data

C- Create a denormalized model with nested and repeated fields Update the table and use snapshots to track historical data

D- Create a denormalized, append-only model with nested and repeated fields Use the ingestion timestamp to track historical data.

Answer:

D

Explanation:

- A denormalized, append-only model simplifies query complexity by eliminating the need for joins. - Adding data with an ingestion timestamp allows for easy retrieval of both current and historical states. - Instead of updating records, new records are appended, which maintains historical information without the need to create separate snapshots.

Question 3

Question Type: MultipleChoice

Your team is building a data lake platform on Google Cloud. As a part of the data foundation design, you are planning to store all the raw data in Cloud Storage You are expecting to ingest approximately 25 GB of data a day and your billing department is worried about the increasing cost of storing old dat

a. The current business requirements are:

* The old data can be deleted anytime

* You plan to use the visualization layer for current and historical reporting

* The old data should be available instantly when accessed

* There should not be any charges for data retrieval.

What should you do to optimize for cost?

Options:

A- Create the bucket with the Autoclass storage class feature.

B- Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 90 days to coldline. and 365 days to archive storage class. Delete old data as needed.

C- Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to coldline, 90 days to nearline. and 365 days to archive storage class Delete old data as needed.

D- Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearlme. 45 days to coldline. and 60 days to archive storage class Delete old data as needed.

Answer:

A

Explanation:

- Autoclass automatically moves objects between storage classes without impacting performance or availability, nor incurring retrieval costs. - It continuously optimizes storage costs based on access patterns without the need to set specific lifecycle management policies.

Question 4

Question Type: MultipleChoice

You are part of a healthcare organization where data is organized and managed by respective data owners in various storage services. As a result of this decentralized ecosystem, discovering and managing data has become difficult You need to quickly identify and implement a cost-optimized solution to assist your organization with the following

* Data management and discovery

* Data lineage tracking

* Data quality validation

How should you build the solution?

Options:

A- Use BigLake to convert the current solution into a data lake architecture.

B- Build a new data discovery tool on Google Kubernetes Engine that helps with new source onboarding and data lineage tracking.

C- Use BigOuery to track data lineage, and use Dataprep to manage data and perform data quality validation.

D- Use Dataplex to manage data, track data lineage, and perform data quality validation.

Answer:

D

Explanation:

Dataplex is a Google Cloud service that provides a unified data fabric for data lakes and data warehouses. It enables data governance, management, and discovery across multiple data domains, zones, and assets. Dataplex also supports data lineage tracking, which shows the origin and transformation of data over time. Dataplex also integrates with Dataprep, a data preparation and quality tool that allows users to clean, enrich, and transform data using a visual interface. Dataprep can also monitor data quality and detect anomalies using machine learning. Therefore, Dataplex is the most suitable solution for the given scenario, as it meets all the requirements of data management and discovery, data lineage tracking, and data quality validation.Reference:

Dataplex overview

Automate data governance, extend your data fabric with Dataplex-BigLake integration

Dataprep documentation

Question 5

Question Type: MultipleChoice

Your chemical company needs to manually check documentation for customer order. You use a pull subscription in Pub/Sub so that sales agents get details from the order. You must ensure that you do not process orders twice with different sales agents and that you do not add more complexity to this workflow. What should you do?

Options:

A- Create a transactional database that monitors the pending messages.

B- Create a new Pub/Sub push subscription to monitor the orders processed in the agent's system.

C- Use Pub/Sub exactly-once delivery in your pull subscription.

D- Use a Deduphcate PTransform in Dataflow before sending the messages to the sales agents.

Answer:

C

Explanation:

Pub/Sub exactly-once delivery is a feature that guarantees that subscriptions do not receive duplicate deliveries of messages based on a Pub/Sub-defined unique message ID. This feature is only supported by the pull subscription type, which is what you are using in this scenario. By enabling exactly-once delivery, you can ensure that each order is processed only once by a sales agent, and that no order is lost or duplicated. This also simplifies your workflow, as you do not need to create a separate database or subscription to monitor the pending or processed messages.Reference:

Exactly-once delivery | Cloud Pub/Sub Documentation

Cloud Pub/Sub Exactly-once Delivery feature is now Generally Available (GA)

Question 6

Question Type: MultipleChoice

You have created an external table for Apache Hive partitioned data that resides in a Cloud Storage bucket, which contains a large number of files. You notice that queries against this table are slow You want to improve the performance of these queries What should you do?

Options:

A- Migrate the Hive partitioned data objects to a multi-region Cloud Storage bucket.

B- Create an individual external table for each Hive partition by using a common table name prefix Use wildcard table queries to reference the partitioned data.

C- Change the storage class of the Hive partitioned data objects from Coldline to Standard.

D- Upgrade the external table to a BigLake table Enable metadata caching for the table.

Answer:

D

Explanation:

BigLake is a Google Cloud service that allows you to query structured data in external data stores such as Cloud Storage, Amazon S3, and Azure Blob Storage with access delegation and governance. BigLake tables extend the capabilities of BigQuery to data lakes and enable a flexible, open lakehouse architecture. By upgrading an external table to a BigLake table, you can improve the performance of your queries by leveraging the BigQuery storage API, which supports data format conversion, predicate pushdown, column projection, and metadata caching. Metadata caching reduces the number of requests to the external data store and speeds up query execution. To upgrade an external table to a BigLake table, you can use theALTER TABLEstatement with theSET OPTIONSclause and specify theenable_metadata_cachingoption astrue. For example:

SQL

ALTER TABLE hive_partitioned_data

SET OPTIONS (

enable_metadata_caching=true

);

AI-generated code. Review and use carefully.More info on FAQ.

Introduction to BigLake tables

Upgrade an external table to BigLake

BigQuery storage API

Question 7

Question Type: **MultipleChoice**

You are creating a data model in BigQuery that will hold retail transaction dat

a. Your two largest tables, sales_transation_header and sales_transation_line. have a tightly coupled immutable relationship. These tables are rarely modified after load and are frequently joined when queried. You need to model the sales_transation_header and sales_transation_line tables to improve the performance of data analytics queries. What should you do?

Options:

A- Create a sal es_transaction table that Stores the sales_tran3action_header and sales_transaction_line data as a JSON data type.

B- Create a sale3_transaction table that holds the sales_transaction_header information as rows and the sales_transaction_line rows as nested and repeated fields.

C- Create a sale_transaction table that holds the sales_transaction_header and sales_transaction_line information as rows, duplicating the sales_transaction_header data for each line.

D- Create separate sales_transation_header and sales_transation_line tables and. when querying, specify the sales transition line first in the WHERE clause.

Answer:

B

Explanation:

BigQuery supports nested and repeated fields, which are complex data types that can represent hierarchical and one-to-many relationships within a single table. By using nested and repeated fields, you can denormalize your data model and reduce the number of joins required for your queries. This can improve the performance and efficiency of your data analytics queries, as joins can be expensive and require shuffling data across nodes. Nested and repeated fields also preserve the data integrity and avoid data duplication. In this scenario, the sales_transaction_header and sales_transaction_line tables have a tightly coupled immutable relationship, meaning that each header row corresponds to one or more line rows, and the data is rarely modified after load. Therefore, it makes sense to create a single sales_transaction table that holds the sales_transaction_header information as rows and the

sales_transaction_line rows as nested and repeated fields. This way, you can query the sales transaction data without joining two tables, and use dot notation or array functions to access the nested and repeated fields. For example, the sales_transaction table could have the following schema:

Table

Field name

Type

Mode

id

INTEGER

NULLABLE

order_time

TIMESTAMP

NULLABLE

customer_id

INTEGER

NULLABLE

line_items

RECORD

REPEATED

line_items.sku

STRING

NULLABLE

line_items.quantity

INTEGER

NULLABLE

line_items.price

FLOAT

NULLABLE

To query the total amount of each order, you could use the following SQL statement:

SQL

```
SELECT id, SUM(line_items.quantity * line_items.price) AS total_amount
```

```
FROM sales_transaction

GROUP BY id;
```

AI-generated code. Review and use carefully.More info on FAQ.

Use nested and repeated fields

BigQuery explained: Working with joins, nested & repeated data

Arrays in BigQuery --- How to improve query performance and optimise storage

Question 8

Question Type: MultipleChoice

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?

Options:

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

Answer:

B

Explanation:

Cloud DLP is a service that helps you discover, classify, and protect your sensitive data. It supports various de-identification techniques, such as masking, redaction, tokenization, and encryption. Format-preserving encryption (FPE) with FFX is a technique that encrypts

sensitive data while preserving its original format and length. This allows you to join the encrypted data on the same field without revealing the actual values. FPE with FFX also supports partial encryption, which means you can encrypt only a portion of the data, such as the domain name of an email address. By using Cloud DLP to de-identify the email field with FPE with FFX, you can ensure that the analysts can join the booking and user profile data on the email field without accessing the PII. You can create a pipeline to de-identify the email field by using recordTransformations in Cloud DLP, which allows you to specify the fields and the de-identification transformations to apply to them. You can then load the de-identified data into a BigQuery table for analysis.Reference:

De-identify sensitive data | Cloud Data Loss Prevention Documentation

Format-preserving encryption with FFX | Cloud Data Loss Prevention Documentation

De-identify and re-identify data with the Cloud DLP API

De-identify data in a pipeline

Question 9

Question Type: **MultipleChoice**

You want to store your team's shared tables in a single dataset to make data easily accessible to various analysts. You want to make this data readable but unmodifiable by analysts. At the same time, you want to provide the analysts with individual workspaces in the same project, where they can create and store tables for their own use, without the tables being accessible by other analysts. What should you do?

Options:

A- Give analysts the BigQuery Data Viewer role at the project level Create one other dataset, and give the analysts the BigQuery Data Editor role on that dataset.

B- Give analysts the BigQuery Data Viewer role at the project level Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the project level.

C- Give analysts the BigQuery Data Viewer role on the shared dataset. Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the dataset level for their assigned dataset

D- Give analysts the BigQuery Data Viewer role on the shared dataset Create one other dataset and give the analysts the BigQuery Data Editor role on that dataset.

Answer:

C

Explanation:

The BigQuery Data Viewer role allows users to read data and metadata from tables and views, but not to modify or delete them. By giving analysts this role on the shared dataset, you can ensure that they can access the data for analysis, but not change it. The BigQuery Data Editor role allows users to create, update, and delete tables and views, as well as read and write data. By giving analysts this role at the dataset level for their assigned dataset, you can provide them with individual workspaces where they can store their own tables and views, without affecting the shared dataset or other analysts' datasets. This way, you can achieve both data protection and

data isolation for your team.Reference:

BigQuery IAM roles and permissions

Basic roles and permissions

Question 10

Question Type: MultipleChoice

You are migrating a large number of files from a public HTTPS endpoint to Cloud Storage. The files are protected from unauthorized access using signed URLs. You created a TSV file that contains the list of object URLs and started a transfer job by using Storage Transfer Service. You notice that the job has run for a long time and eventually failed Checking the logs of the transfer job reveals that the job was running fine until one point, and then it failed due to HTTP 403 errors on the remaining files You verified that there were no changes to the source system You need to fix the problem to resume the migration process. What should you do?

Options:

A- Set up Cloud Storage FUSE, and mount the Cloud Storage bucket on a Compute Engine Instance Remove the completed files from the TSV file Use a shell script to iterate through the TSV file and download the remaining URLs to the FUSE mount point.

B- Update the file checksums in the TSV file from using MD5 to SHA256. Remove the completed files from the TSV file and rerun the

Storage Transfer Service job.

C- Renew the TLS certificate of the HTTPS endpoint Remove the completed files from the TSV file and rerun the Storage Transfer Service job.

D- Create a new TSV file for the remaining files by generating signed URLs with a longer validity period. Split the TSV file into multiple smaller files and submit them as separate Storage Transfer Service jobs in parallel.

Answer:

D

Explanation:

A signed URL is a URL that provides limited permission and time to access a resource on a web server. It is often used to grant temporary access to protected files without requiring authentication. Storage Transfer Service is a service that allows you to transfer data from external sources, such as HTTPS endpoints, to Cloud Storage buckets. You can use a TSV file to specify the list of URLs to transfer. In this scenario, the most likely cause of the HTTP 403 errors is that the signed URLs have expired before the transfer job could complete. This could happen if the signed URLs have a short validity period or the transfer job takes a long time due to the large number of files or network latency. To fix the problem, you need to create a new TSV file for the remaining files by generating new signed URLs with a longer validity period. This will ensure that the URLs do not expire before the transfer job finishes. You can use the Cloud Storage tools or your own program to generate signed URLs. Additionally, you can split the TSV file into multiple smaller files and submit them as separate Storage Transfer Service jobs in parallel. This will speed up the transfer process and reduce the risk of errors.Reference:

Signed URLs | Cloud Storage Documentation

V4 signing process with Cloud Storage tools

V4 signing process with your own program

Using a URL list file

What Is a 403 Forbidden Error (and How Can I Fix It)?

Question 11

Question Type: MultipleChoice

You have data located in BigQuery that is used to generate reports for your company. You have noticed some weekly executive report fields do not correspond to format according to company standards for example, report errors include different telephone formats and different country code identifiers. This is a frequent issue, so you need to create a recurring job to normalize the dat

a. You want a quick solution that requires no coding What should you do?

Options:

A- Use Cloud Data Fusion and Wrangler to normalize the data, and set up a recurring job.

B- Use BigQuery and GoogleSQL to normalize the data, and schedule recurring quenes in BigQuery.

C- Create a Spark job and submit it to Dataproc Serverless.

D- Use Dataflow SQL to create a job that normalizes the data, and that after the first run of the job, schedule the pipeline to execute recurrently.

Answer:

A

Explanation:

Cloud Data Fusion is a fully managed, cloud-native data integration service that allows you to build and manage data pipelines with a graphical interface. Wrangler is a feature of Cloud Data Fusion that enables you to interactively explore, clean, and transform data using a spreadsheet-like UI. You can use Wrangler to normalize the data in BigQuery by applying various directives, such as parsing, formatting, replacing, and validating data. You can also preview the results and export the wrangled data to BigQuery or other destinations. You can then set up a recurring job in Cloud Data Fusion to run the Wrangler pipeline on a schedule, such as weekly or daily. This way, you can create a quick and code-free solution to normalize the data for your reports.Reference:

Cloud Data Fusion overview

Wrangler overview

Wrangle data from BigQuery

[Scheduling pipelines]