Free Questions for Professional-Data-Engineer

Shared by Farrell on 09-08-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

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 2

Question Type: MultipleChoice

You want to schedule a number of sequential load and transformation jobs Data files will be added to a Cloud Storage bucket by an upstream process There is no fixed schedule for when the new data arrives Next, a Dataproc job is triggered to perform some transformations and write the data to BigQuery. You then need to run additional transformation jobs in BigQuery The transformation jobs are different for every table These jobs might take hours to complete You need to determine the most efficient and maintainable workflow to process hundreds of tables and provide the freshest data to your end users. What should you do?

Options:

A- 1Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage. Dataproc. and BigQuery operators
2 Use a single shared DAG for all tables that need to go through the pipeline
3 Schedule the DAG to run hourly

B- 1 Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators.
2 Create a separate DAG for each table that needs to go through the pipeline
3 Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG

C- 1 Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Cloud Storage, Dataproc. and BigQuery operators
2 Create a separate DAG for each table that needs to go through the pipeline
3 Schedule the DAGs to run hourly

D- 1 Create an Apache Airflow directed acyclic graph (DAG) in Cloud Composer with sequential tasks by using the Dataproc and BigQuery operators
2 Use a single shared DAG for all tables that need to go through the pipeline.
3 Use a Cloud Storage object trigger to launch a Cloud Function that triggers the DAG

Answer:

B

Explanation:

This option is the most efficient and maintainable workflow for your use case, as it allows you to process each table independently and trigger the DAGs only when new data arrives in the Cloud Storage bucket.By using the Dataproc and BigQuery operators, you can easily orchestrate the load and transformation jobs for each table, and leverage the scalability and performance of these services12.By creating a separate DAG for each table, you can customize the transformation logic and parameters for each table, and avoid the complexity and overhead of a single shared DAG3.By using a Cloud Storage object trigger, you can launch a Cloud Function that triggers the DAG for the corresponding table, ensuring that the data is processed as soon as possible and reducing the idle time and cost of running the DAGs on a fixed schedule4.

Option A is not efficient, as it runs the DAG hourly regardless of the data arrival, and it uses a single shared DAG for all tables, which makes it harder to maintain and debug. Option C is also not efficient, as it runs the DAGs hourly and does not leverage the Cloud Storage object trigger. Option D is not maintainable, as it uses a single shared DAG for all tables, and it does not use the Cloud Storage operator, which can simplify the data ingestion from the bucket.Reference:

1: Dataproc Operator | Cloud Composer | Google Cloud

2: BigQuery Operator | Cloud Composer | Google Cloud

3: Choose Workflows or Cloud Composer for service orchestration | Workflows | Google Cloud

4: Cloud Storage Object Trigger | Cloud Functions Documentation | Google Cloud

[5]: Triggering DAGs | Cloud Composer | Google Cloud

[6]: Cloud Storage Operator | Cloud Composer | Google Cloud

Question 3

Question Type: MultipleChoice

You are building an ELT solution in BigQuery by using Dataform. You need to perform uniqueness and null value checks on your final tables. What should you do to efficiently integrate these checks into your pipeline?

Options:

A- Build Dataform assertions into your code

B- Write a Spark-based stored procedure.

C- Build BigQuery user-defined functions (UDFs).

D- Create Dataplex data quality tasks.

Answer:

A

Explanation:

Dataform assertions are data quality tests that find rows that violate one or more rules specified in the query. If the query returns any rows, the assertion fails. Dataform runs assertions every time it updates your SQL workflow and alerts you if any assertions fail. You can create assertions for all Dataform table types: tables, incremental tables, views, and materialized views. You can add built-in assertions to the config block of a table, such as nonNull and rowConditions, or create manual assertions with SQLX for advanced use cases. Dataform automatically creates views in BigQuery that contain the results of compiled assertion queries, which you can inspect to debug failing assertions. Dataform assertions are an efficient way to integrate data quality checks into your ELT solution in BigQuery by using Dataform.Reference:Test tables with assertions | Dataform | Google Cloud,Test data quality with assertions | Dataform,Data quality tests and documenting datasets | Dataform,Data quality testing with SQL assertions | Dataform

Question 4

Question Type: MultipleChoice

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the dat

a. Which two actions should you take? (Choose two.)

Options:

A- Configure your Cloud Dataflow pipeline to use local execution

B- Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions

C- Increase the number of nodes in the Cloud Bigtable cluster

D- Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable

E- Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable

Answer:

B, C

Question 5

Question Type: MultipleChoice

You are troubleshooting your Dataflow pipeline that processes data from Cloud Storage to BigQuery. You have discovered that the Dataflow worker nodes cannot communicate with one another Your networking team relies on Google Cloud network tags to define firewall rules You need to identify the issue while following Google-recommended networking security practices. What should you do?

Options:

A- Determine whether your Dataflow pipeline has a custom network tag set.

B- Determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 for the Dataflow network tag.

C- Determine whether your Dataflow pipeline is deployed with the external IP address option enabled.

D- Determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 on the subnet used by Dataflow workers.

Answer:

D

Explanation:

Dataflow worker nodes need to communicate with each other and with the Dataflow service on TCP ports 12345 and 12346. These ports are used for data shuffling and streaming engine communication. By default, Dataflow assigns a network tag called dataflow to the worker nodes, and creates a firewall rule that allows traffic on these ports for the dataflow network tag. However, if you use a custom network tag for your Dataflow pipeline, you need to create a firewall rule that allows traffic on these ports for your custom network tag. Otherwise, the worker nodes will not be able to communicate with each other and the Dataflow service, and the pipeline will fail.

Therefore, the best way to identify the issue is to determine whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 for the Dataflow network tag. If there is no such firewall rule, or if the firewall rule does not match the network tag used by your Dataflow pipeline, you need to create or update the firewall rule accordingly.

Option A is not a good solution, as determining whether your Dataflow pipeline has a custom network tag set does not tell you whether there is a firewall rule that allows traffic on the required ports for that network tag. You need to check the firewall rule as well.

Option C is not a good solution, as determining whether your Dataflow pipeline is deployed with the external IP address option enabled does not tell you whether there is a firewall rule that allows traffic on the required ports for the Dataflow network tag. The external IP address option determines whether the worker nodes can access resources on the public internet, but it does not affect the internal communication between the worker nodes and the Dataflow service.

Option D is not a good solution, as determining whether there is a firewall rule set to allow traffic on TCP ports 12345 and 12346 on the subnet used by Dataflow workers does not tell you whether the firewall rule applies to the Dataflow network tag. The firewall rule should be based on the network tag, not the subnet, as the network tag is more specific and secure.Reference:Dataflow network tags | Cloud Dataflow | Google Cloud,Dataflow firewall rules | Cloud Dataflow | Google Cloud,Dataflow network configuration | Cloud Dataflow | Google Cloud,Dataflow Streaming Engine | Cloud Dataflow | Google Cloud.

Question 6

Question Type: MultipleChoice

You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SideInputs to join data You noticed that the pipeline is taking longer to complete than expected, what should you do to expedite the Dataflow job?

Options:

A- Switch to compressed Avro files

B- Reduce the batch size

C- Retry records that throw an error

D- Use CoGroupByKey instead of the SideInput

Answer:

B

Question 7

Question Type: MultipleChoice

You are designing a Dataflow pipeline for a batch processing job. You want to mitigate multiple zonal failures at job submission time. What should you do?

Options:

A- Specify a worker region by using the ---region flag.

B- Set the pipeline staging location as a regional Cloud Storage bucket.

C- Submit duplicate pipelines in two different zones by using the ---zone flag.

D- Create an Eventarc trigger to resubmit the job in case of zonal failure when submitting the job.

Answer:

B

Explanation:

By specifying a worker region, you can run your Dataflow pipeline in a multi-zone or multi-region configuration, which provides higher availability and resilience in case of zonal failures1.The ---region flag allows you to specify the regional endpoint for your pipeline, which determines the location of the Dataflow service and the default location of the Compute Engine resources1.If you do not specify a zone by using the ---zone flag, Dataflow automatically selects a zone within the region for your job workers1. This option is recommended over submitting duplicate pipelines in two different zones, which would incur additional costs and complexity.Setting the pipeline staging location as a regional Cloud Storage bucket does not affect the availability of your pipeline, as the staging location only stores the pipeline code and dependencies2. Creating an Eventarc trigger to resubmit the job in case of zonal failure is not a reliable solution, as it depends on the availability of the Eventarc service and the zonal resources at the time of resubmission.Reference:

1: Pipeline troubleshooting and debugging | Cloud Dataflow | Google Cloud

3: Regional endpoints | Cloud Dataflow | Google Cloud

Question 8

Question Type: MultipleChoice

You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the dat

a. They should only see certain tables based on their team membership. How should you set user permissions?

Options:

A- Assign the users/groups data viewer access at the table level for each table

B- Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views

C- Create authorized views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the authorized views

D- Create authorized views for each team in datasets created for each team. Assign the authorized views data viewer access to the dataset in which the data resides. Assign the users/groups data viewer access to the datasets in which the authorized views reside

Answer:

A

Question 9

Question Type: MultipleChoice

You are designing a system that requires an ACID-compliant database. You must ensure that the system requires minimal human intervention in case of a failure. What should you do?

Options:

A- Configure a Cloud SQL for MySQL instance with point-in-time recovery enabled.

B- Configure a Cloud SQL for PostgreSQL instance with high availability enabled.

C- Configure a Bigtable instance with more than one cluster.

D- Configure a BJgQuery table with a multi-region configuration.

Answer:

B

Explanation:

The best option to meet the ACID compliance and minimal human intervention requirements is to configure a Cloud SQL for PostgreSQL instance with high availability enabled. Key reasons: Cloud SQL for PostgreSQL provides full ACID compliance, unlike Bigtable which provides only atomicity and consistency guarantees. Enabling high availability removes the need for manual failover as Cloud SQL will automatically failover to a standby replica if the leader instance goes down. Point-in-time recovery in MySQL requires manual intervention to restore data if needed. BigQuery does not provide transactional guarantees required for an ACID database. Therefore, a Cloud SQL for PostgreSQL instance with high availability meets the ACID and minimal intervention requirements best. The automatic failover will ensure availability and uptime without administrative effort.

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

Different teams in your organization store customer and performance data in BigOuery. Each team needs to keep full control of their collected data, be able to query data within their projects, and be able to exchange their data with other teams. You need to implement an organization-wide solution, while minimizing operational tasks and costs. What should you do?

Options:

A- Create a BigQuery scheduled query to replicate all customer data into team projects.
B- Enable each team to create materialized views of the data they need to access in their projects.
C- Ask each team to publish their data in Analytics Hub. Direct the other teams to subscribe to them.
D- Ask each team to create authorized views of their data. Grant the biquery. jobUser role to each team.

Answer:

C

Explanation:

To enable different teams to manage their own data while allowing data exchange across the organization, using Analytics Hub is the best approach. Here's why option C is the best choice:

Analytics Hub:

Analytics Hub allows teams to publish their data as data exchanges, making it easy for other teams to discover and subscribe to the data they need.

This approach maintains each team's control over their data while facilitating easy and secure data sharing across the organization.

Data Publishing and Subscribing:

Teams can publish datasets they control, allowing them to manage access and updates independently.

Other teams can subscribe to these published datasets, ensuring they have access to the latest data without duplicating efforts.

Minimized Operational Tasks and Costs:

This method reduces the need for complex replication or data synchronization processes, minimizing operational overhead.

By centralizing data sharing through Analytics Hub, it also reduces storage costs associated with duplicating large datasets.

Steps to Implement:

Set Up Analytics Hub:

Enable Analytics Hub in your Google Cloud project.

Provide training to teams on how to publish and subscribe to data exchanges.

Publish Data:

Each team publishes their datasets in Analytics Hub, configuring access controls and metadata as needed.

Subscribe to Data:

Teams that need access to data from other teams can subscribe to the relevant data exchanges, ensuring they always have up-to-date data.

Analytics Hub Documentation

Publishing Data in Analytics Hub

Subscribing to Data in Analytics Hub

Question 12

Question Type: MultipleChoice

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured

Support for publish/subscribe semantics on hundreds of topics

Retain per-key ordering

Which system should you choose?

Options:

A- Apache Kafka

B- Cloud Storage

C- Cloud Pub/Sub

D- Firebase Cloud Messaging

Answer:

A