PR OFESSIONAL-DATA-ENGINEER$^{Q\&As}$

Professional Data Engineer on Google Cloud Platform

Pass Google PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Following Questions and Answers are all new published by Google Official Exam Center

QUESTION 1

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

A. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.

B. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket. Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.

C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.

D. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

Correct Answer: B

QUESTION 2

You are running a streaming pipeline with Dataflow and are using hopping windows to group the data as the data arrives. You noticed that some data is arriving late but is not being marked as late data, which is resulting in inaccurate

aggregations downstream. You need to find a solution that allows you to capture the late data in the appropriate window.

What should you do?

A. Change your windowing function to session windows to define your windows based on certain activity.

B. Change your windowing function to tumbling windows to avoid overlapping window periods.

C. Expand your hopping window so that the late data has more time to arrive within the grouping.

D. Use watermarks to define the expected data arrival window Allow late data as it arrives.

Correct Answer: D

Watermarks are a way of tracking the progress of time in a streaming pipeline. They are used to determine when a window can be closed and the results emitted. Watermarks can be either event-time based or processing-time based. Event-time watermarks track the progress of time based on the timestamps of the data elements, while processing-time watermarks track the progress of time based on the system clock. Event-time watermarks are more accurate, but they require the data source to provide reliable timestamps. Processing-time watermarks are simpler, but they can be affected by system delays or backlogs. By using watermarks, you can define the expected data arrival window for each windowing function. You can also specify how to handle late data, which is data that arrives after the watermark has passed. You can either discard late data, or allow late data and update the results as new data arrives. Allowing late data requires you to use triggers to control when the results are emitted. In this case, using watermarks and allowing late data is the best solution to capture the late data in the appropriate window. Changing the windowing function to session windows or tumbling windows will not solve the problem of late data, as they still rely on watermarks to determine when to close the windows. Expanding the hopping window might reduce the amount of late data, but it will also change the semantics of the windowing function and the results. References: Streaming pipelines | Cloud Dataflow | Google Cloud Windowing | Apache Beam

QUESTION 3

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

A. Add a SideInput that returns a Boolean if the element is corrupt.

B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.

C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.

D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Correct Answer: B

QUESTION 4

What are the minimum permissions needed for a service account used with Google Dataproc?

A. Execute to Google Cloud Storage; write to Google Cloud Logging

B. Write to Google Cloud Storage; read to Google Cloud Logging

C. Execute to Google Cloud Storage; execute to Google Cloud Logging

D. Read and write to Google Cloud Storage; write to Google Cloud Logging

Correct Answer: D

Service accounts authenticate applications running on your virtual machine instances to other Google Cloud Platform services. For example, if you write an application that reads and writes files on Google Cloud Storage, it must first authenticate to the Google Cloud Storage API. At a minimum, service accounts used with Cloud Dataproc need permissions to read and write to Google Cloud Storage, and to write to Google Cloud Logging. Reference: https:// cloud.google.com/dataproc/docs/concepts/service- accounts#important_notes

QUESTION 5

Which of these operations can you perform from the BigQuery Web UI?

A. Upload a file in SQL format.

B. Load data with nested and repeated fields.

C. Upload a 20 MB file.

D. Upload multiple files using a wildcard.

Correct Answer: B

You can load data with nested and repeated fields using the Web UI. You cannot use the Web UI to:

-

Upload a file greater than 10 MB in size

-

Upload multiple files at the same time

-

Upload a file in SQL format

All three of the above operations can be performed using the "bq" command.

Reference: https://cloud.google.com/bigquery/loading-data

QUESTION 6

Your company\\'s customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

A. Add a node to the MySQL cluster and build an OLAP cube there.

B. Use an ETL tool to load the data from MySQL into Google BigQuery.

C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.

D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Correct Answer: C

QUESTION 7

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products of features of the platform. What should you do?

A. Export the information to Cloud Stackdriver, and set up an Alerting policy

B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver

C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs

D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

Correct Answer: B

QUESTION 8

Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.

Which approach should you take?

A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.

B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.

C. Use the NOW () function in BigQuery to record the event\\'s time.

D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

Correct Answer: B

QUESTION 9

You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand

updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate.

What should you do?

A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.

B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.

C. Use the BigQuery streaming the stream changes into a daily inventory movement table.Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

Correct Answer: A

QUESTION 10

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.

B. Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.

C. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.

D. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.

Correct Answer: C

QUESTION 11

Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of datA. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

A. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.

B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.

C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.

D. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.

Correct Answer: B

QUESTION 12

An aerospace company uses a proprietary data format to store its night data. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming

as few resources as possible.

What should you do?

A. Use a standard Dataflow pipeline to store the raw data in BigQuery and then transform the format later when the data is used.

B. Write a shell script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source

C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format

D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format

Correct Answer: D

QUESTION 13

You are responsible for writing your company\\'s ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

A. PigLatin using Pig

B. HiveQL using Hive

C. Java using MapReduce

D. Python using MapReduce

Correct Answer: D

QUESTION 14

You need (o give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system How should you design your pipeline to minimize that backpressure?

A. Call out to the service via HTTP

B. Create the pipeline statically in the class definition

C. Create a new object in the startBundle method of DoFn

D. Batch the job into ten-second increments

Correct Answer: A

QUESTION 15

If you\\'re running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?

A. Do not use a production instance.

B. Run your test for at least 10 minutes.

C. Before you test, run a heavy pre-test for several minutes.

D. Use at least 300 GB of data.

Correct Answer: A

If you\\'re running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test: Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load. Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use 100 GB of data per node. Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes. Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory. Reference: https://cloud.google.com/bigtable/docs/performance