GCP Mock Test-2

Question 1

You have a system which writes its logs to a file. The system writes to a new log file every hour. You want to merge log files every 24 hours to a single file and upload it to a specific bucket in Google Storage. You created a Dataflow pipeline to process the logs files and you want to run this pipeline everyday at 3am.

Which of the following is the best approach to schedule this task?

A. Use Cloud Scheduler to create a cron job to run the Dataflow pipeline at 3am.

B. Create a Compute Engine VM and create a cron job to run Dataflow pipeline at 3am.

C. Configure Dataflow pipeline as a streaming job to process the data in real time.

D. Use Cloud Functions to run Dataflow pipeline at 3am.

Question 2

A company uses Airflow to orchestrate its data pipelines and DAGs (Directed Acyclic Graphs), installed and maintained on-premise by DevOps team. The company wants to migrate the data pipelines managed in Airflow to Google Cloud. The company is looking for a migration method which can make DAGs available and migrated without extra code modifications so the data pipelines can be available once migrated. Which service should you use?

A. App Engine

B. Cloud Function

C. Dataflow

D. Cloud Composer

Question 3

You use BigQuery as the main data warehouse. As you are checking the bill of Google Cloud usage for the past month, you noticed a hike in BigQuery costs. Upon more inspection, you found out it's due to ad-hoc queries scanning large amounts of data. You want to limit the size of data queries process to control costs and avoid excessive bills. What should you do?

A. Set monthly flat-rate pricing for BigQuery.

B. Change to using Google Storage as external table instead of using BigQuery's storage option.

C. Launch a Cloud SQL instance. Replicate BigQuery datasets to Cloud SQL. Allow users with ad-hoc queries to use Cloud SQL instead.

D. Set custom quotas on User-level with access on BigQuery based on their business requirements.

Question 4

You are using BigQuery as the data warehouse. Different departments are using BigQuery to read data. Upon checking the billing costs, you notice that there is a spike in running queries on BigQuery despite caching is enabled. You started scanning through the queries run on BigQuery trying to find out if some queries are not cached.

Which of the following can be reasons for queries not cached? (Choose 2)

A. SELECT queries without asterisk (*).

B. Queries select from authorized views on archive tables.

C. Queries multiple tables use wildcard.

D. Jobs use destination tables.

Question 5

You have a web service which allows users to create personal profiles and upload their photos online. The service will create thumbnails of various sizes and resolutions to be used on the service's different pages and widgets. The service uses Google Storage to store the original photos uploaded and thumbnails generated. The service usually uses the thumbnails for a certain period before generating new ones depends on user's activity and the age of the original photo the thumbnail is created from.

You want to remove all thumbnails created before 90 days. What should you do?

A. Use Cloud Scheduler to scan objects in thumbnails folder within the Google Storage bucket and delete all objects created before 90 days.

B. Use Cloud Function to scan objects in thumbnails folder within the Google Storage bucket and delete all objects created before 90 days.

C. Use object lifecycle management in Google Storage to delete thumbnail objects created before 90 days.

D. Use object lifecycle management in Google Storage to delete thumbnail objects older than the current thumbnail object version last created. Make sure the bucket used for storage has object versioning enabled.

Question 6

An e-commerce company uses BigQuery as its main data warehouse. One of the tables stores customers details such as name, address, email and phone number. Data team wants to modify the table's schema and add a new column called 'zipcode' which is previously included in address column. You are asked to modify the table's schema and do necessary changes. You need to perform the changes with minimal costs. What should you do?

A. Add a new column called 'zipcode' to customers table. Run an update statement to extract zip code from address column and set it to the new column.

B. Create a materialized view in Bigquery that extracts the zip code from the address as a new column

C. Export table data from BigQuery to Google Storage. Use Dataproc to transform data and extract the zip code from addresses and append it as a new column. Create a new table for customers with new column 'zipcode.' Import transformed data to new table.

D. Create a Dataflow pipeline to read data from BigQuery, extract zip code from address column, then write data to a newly created table in BigQuery with 'zipcode' column.

Question 7

You use BigQuery as your main data warehouse. By time, your tables start to get bigger and selecting from these tables result in scanning many rows which increases the cost of queries running on them. You want to find a way to reduce the costs of queries scanning through your big tables. What should you do? (Choose 2)

A. Use LIMIT when running SELECT statements on the tables.

B. Use partitioning to split data into partitions by columns most used for filtering data.

C. Set BigQuery to limit scanning data to certain size.

D. Use sharding to split data into several tables.

Question 8

An air-quality research facility monitors the quality of the air and alerts of possible high air pollution in a region. The facility receives event data from 25,000 sensors every 60 seconds. Event data is then used for time-series analysis per region. Cloud experts suggested using BigTable for storing event data.

What will you design the row key for each even in BigTable?

A. Use event's timestamp as row key.

B. Use combination of sensor ID with timestamp as sensorID-timestamp.

C. Use combination of sensor ID with timestamp as timestamp-sensorID.

D. Use sensor ID as row key.

Question 9

Your team is planning to perform tests on Cloud BigTable instance to ensure the performance quality of the BigTable instance to be used in production. Which of the following conditions should be met to consider the performance testing valid? (Choose 3)

A. Use production instance.

B. Use at least 1TB of data.

C. Use at least 300GB of data.

D. Tests should run for at least 10 minutes.

E. Use development instance.

F. Storage type should be HDD.

Question 10

Your company hosts a gaming app which reaches over 30,000 players in a single minute. The app generates event data including information about players state, score, location coordinates and other stats. You need to find a storage solution which can support high read/write throughput with very low latency which doesn't exceed 10 milliseconds to ensure a quality performance experience for the players.

Which of the following is the best option for this scenario?

A. Cloud Spanner

B. BigQuery

C. BigTable

D. Datastore

Question 11

The development team decided to use BigTable to write thousands of incoming stream data. Their choice was based on BigTable's high performance and high throughput and low latency. However, the team is facing less than expected performance from the cluster. You are asked for advice on the reason for BigTable's instance performance issue.

Which of the following can be a reason for the performance issue of BigTable cluster?

Choose 2 Options.

A. Row key used is increased monotonically.

B. Rows are less than 10MB of size.

C. HDD disk is used in the BigTable cluster.

D. Cluster is launched in a region different than where users reside.

Question 12

Data team is looking for a database system which is highly available and supports atomic transactions. Database should have a flexible but semi-structured schema and supports querying using SQL-like language. Solution should be fully managed with no planned downtime.

Which of the following is the best choice for this scenario?

A. Cloud SQL

B. Cloud Spanner

C. BigTable

D. Datastore

Question 13

You have an existing data pipeline which uses Apache Spark to transform the data to be exported in CSV format, to be later loaded into MySQL database for further analysis. The CTO decides it's time to migrate the pipeline to Google Cloud. As a data architect, you are tasked to design the new pipeline with minimal changes to the current pipeline architecture for a smoother migration.

Which of the following approaches are best suitable for your CTO's requirement?

A. Use Dataproc for data transformation. Use Google Storage for storing data.

B. Use Dataflow for data transformation. Use BigQuery for storing data.

C. Use Dataproc for data transformation. Use BigQuery for storing data.

D. Use Dataflow for data transformation. Use Google Storage for storing data.

Question 14

A company decided to migrate their on-premise hadoop jobs to Google Cloud. As recommended by Google Cloud engineers, Dataproc is used to run Apache Hive jobs. Data residing in on-premise HDFS has been moved to Google Storage and connector was used for Dataproc to read the data. Upon monitoring the performance of Dataproc clusters running Hive jobs, you noticed the jobs are I/O intensive and use local disk to read/write data. This leads to performance issues. How can you solve this problem?

A. Increase persistent disk size for master node.

B. Increase persistent disk size for worker nodes.

C. Increase RAM capacity of Dataproc cluster's worker nodes.

D. Use local HDFS storage of Dataproc cluster nodes instead of Google Storage.

Question 15

Your team decided to use BigTable for storing event data. The engineer responsible of launching and testing the instance has reported a slower performance than expected by Google Cloud documentation. Which of the following could be a factor for the slow performance? (Choose 3)

A. The rows in the tables tested contain very few number of cells.

B. The rows in the tables have small data size.

C. The schema is not designed for the instance to evenly read and write data across the tables.

D. The instance uses HDD storage type.

E. The instance was scaled up recently.

F. The instance has too high number of nodes for the data size tested.

Question 16

An online learning platform wants to generate captions for its videos. The platform offers around 2,500 courses with topics about business, finance, cooking, development & science. The platform allows content with different languages such as French, German, Turkish and Thai. Thus, this can be very difficult for a single team to caption all available courses and they are looking for an approach which helps do such massive job.

Which product from Google Cloud will you suggest them to use?

A. Cloud Speech-to-Text.

B. Cloud Natural Language.

C. Vertex AI

D. AutoML Vision API.

Question 17

You need to build a machine learning model to recognize different animals for a pet shop. The purpose is to scan the photos on their twitter page and get stats about what pets people like sharing while tagging the pet shop brand the most. Due to cost constraints, the project should be as cost-effective as possible, and that includes work hours dedicated to the project.

Which approach will you consider to build the project?

A. Use Cloud ML Engine API and inspect the descriptions returned by the API. Consider the description with highest score.

B. Use Vision API and inspect the descriptions returned by the API. Consider the description with highest score.

C. Use Vision API and inspect the MID values returned by API to recognize the pets in photos.

D. Use Vision API and inspect the descriptions returned by the API. Consider the description with median score.

Question 18

You are working on building your own machine learning model and training it. When you tested the model on a testing set, you realized the error rate is very high and the model's output only matched 25% of expected output.

What is the problem you are facing and how can you fix it?

A. The model is underfitting: You need to increase the features and use more training data.

B. The model is underfitting: You need to lower the features and use less training data.

C. The model is overfitting: You need to lower the features and use more training data.

D. The model is overfitting: You need to increase the features and use more training data.

Question 19

You are building a machine learning model to solve a binary classification problem. The model is going to predict the likelihood of a customer to be using a fraudulent credit card when purchasing online.

Since there is a very small fraction of purchase transactions are proved to be fraudulent, more than 99% of the purchase transactions are valid. You want to make sure the machine learning model is able to identify the fraudulent transactions.

What is the technique to examine the effectiveness of the model?

A. Gradient Descent

B. Recall

C. Feature engineering

D. Precision

Question 20

A Kafka cluster is receiving event data from outsourced sensors. The cluster is installed in a Compute Engine instance and it writes events to Google Storage. Due to the new security rules in the company, data written to Google Storage should be encrypted. Security team wants to be sure encryption key used is provided by them using on-premise vault and no keys generated by third-parties are used.

What should you do to follow security team's rules?

A. Reference the encryption key provided by security team when calling API service when writing data to Google Storage to encrypt the data.

B. Store the encryption key provided by security team in Compute Engine instance and reference it when calling API service when writing data to Google Storage to encrypt the data.

C. Store the encryption key provided by security team in Cloud Key Management Service (KMS) and reference it when calling API service when writing data to Google Storage to encrypt the data.

D. Create encryption keys using Cloud Key Management Service (KMS) and reference it when calling API service when writing data to Google Storage to encrypt the data.

Question 21

You have a PostgreSQL database instance on Cloud SQL. The database is used in production by the different micro-services writing and reading to it based on each micro-service's needs. Data analysts connect to the instance to run their analysis and reporting SQL queries which adds more load to the database with intensive I/O operations on it. You want to find a solution which can avoid intensive I/O operations on production database and allow data analysts to run their queries without interruptions may lead to late reports to management. What should you do?

A. Scale up Cloud SQL instance from standard to high CPU machine type.

B. Use Cloud SQL read replicas to replicate production instance. Share endpoint with data analysts to replace with existing endpoint in their SQL clients to use the replica.

C. Create a snapshot from production instance. Create a new Cloud SQL instance from the snapshot. Use Cloud Scheduler cron job to create a snapshot and a new instance daily.

D. Import data from Cloud SQL instance to BigQuery. Create necessary users on BigQuery dataset created for data analysts to run their SQL queries on.

Question 22

You receive a daily comma-separated (CSV) file which should be imported to BigQuery. You need to scan the file in case of incomplete or improperly aligned column values which will cause importing to BigQuery fail.

What should you do to handle invalid inputs?

A. Import data to BigQuery, then run queries to check if data skew exists among table fields.

B. Import file to BigQuery with setting &ndash;max_bad_records.

C. Use Google Stackdriver to monitor import status and create an alert for failed imports.

D. Trigger the Cloud Function on the GCS bucket and process the CSV file, insert the valid records in the Bigquery table, and push the invalid records to the error table in Bigquery for later analysis

Question 23

The data analysts in your company want to prepare data sets for reporting to upper management. While the current data pipeline does part of data modeling to the data sets, data analysts still want to perform extra data profiling on data such as detecting duplicates, count null values and other profiling techniques. They ask your advice on what tool to use.

Which of the following is recommended?

A. Cloud Dataprep

B. Dataproc

C. Cloud Composer

D. Cloud Function

Question 24

You are in need to restore a snapshot of a compute engine instance you have previously scheduled for regular daily snapshots. Which of the following are the steps you should do to perform the restoration?

A. You can simply create a replacement instance directly by selecting the snapshot from the list of snapshots available.

B. You need to create a persistent disk from the snapshot of your choice. Create a new compute engine instance and attach the persistent disk to it.

C. Create a new compute instance with the same exact machine type as the one the snapshot was created from. Create a persistent disk using the snapshot to be restored from. Attach the persistent disk to the compute engine instance.

D. Export snapshot to Google Storage. Create new compute engine instance, then using gsutil tool, copy the snapshot to the instance's persistent disk to be restored.

Question 25

Your data team is using BigQuery as their main data warehouse. There is no formal security policy implemented to track users activity on the data warehouse. A new security policy to be implemented which states any activity on cloud resources should be tracked and logged and BigQuery is one of these resources.

What action should be taken to log the data warehouse's activities?

A. Restrict users access to BigQuery's tables using Identity & Access Management (IAM).

B. You can list all query history from BigQuery UI.

C. Use Google Audit Logs to capture all the activities for Bigquery and review data warehouse access.

D. Enable caching on BigQuery to allow auditing users activity.