GCP Mock Test -1

Question 1

Data analysts in your company use BigQuery as their structured database supporting SQL for querying. Recently, they started to study about machine learning and what possibilities they can achieve if they can apply some machine learning algorithms such as binary classification on existing data stored in BigQuery for customer behavior prediction. You are to assist preparing the required tools and resources for data analysts to apply machine learning algorithms on data. What should you do?

A. Use Vertex AI to build and deploy your binary classification model

B. Use BigQuery ML by building the binary classification model by specifying the model type and labels when creating the model.

C. Use Apache Spark MLLib with Dataproc to build the model. Use Dataproc-BigQuery connector to allow the cluster reading data.

D. Use Cloud ML Engine to build the classification model(s) using TensorFlow, allowing the model's service account permission to read the data from BigQuery.

Question 2

A coach line bus service company wants to predict how many passengers they expect to book for tickets on their buses for the upcoming months. This helps the company to know how many buses they need to be in service for maintenance and fuel and how many drivers to be available. The company has data sets of all booked tickets since its launch in 1968 and it allows private sharing of the data if this helps the prediction process.

You will build the machine learning model for the coach line company. Which technique you will use to predict the number of passengers in the next months?

A. Regression.

B. Association.

C. Classification.

D. Clustering.

Question 3

A company has over 25TB of data in Avro format stored in on-premise disks. You are migrating the tech stack used to Google Cloud. The current data pipeline built on-premise does the required data transformation and enrichment using Apache Spark. You decide to

use Dataproc for data processing. When the migration was approved by the management, one of the base requirements was for data to be highly available and cross-zone durability should be guaranteed. What should you do?

A. Use Google Storage to store data. Allow Dataproc cluster to access data from Google Storage.

B. Use BigQuery to store data. Install Dataproc-BigQuery connector to access data.

C. Use Dataproc cluster's HDFS namenodes to store data.

D. Use BigTable to store data. Use Dataproc-BigTable connector to access data.

Question 4

You want to build a system which uses a machine learning, image recognition model to detect customers' faces entering a retail shop, and based on the knowledge base it will return whether the customer is a new, returning or loyal customer. You are building the model using AutoML Vision. After training the model and testing it, you find the model's accuracy is lower due to overfitting. How can you solve this?

A. Images used should be taken from the same exact angle and resolution.

B. Instead of manually splitting samples to training and testing sets, allow AutoML Vision to split the sample set.

C. Samples used for training should be covering true positives only.

D. Images used should be taken from different angles, resolutions and points of view.

Question 5

You are preparing a dataset as a training set for a Machine Learning Model. You have the following columns chosen as features for the model:

- Zip code
- Age Group
- Income

Which feature type each column is:

A. 3 continuous.

B. 2 categorical, 1 continuous.

C. 2 continuous, 1 categorical.

D. 3 categorical.

Question 6

You need to deploy a machine learning model built by data science team in the firm you work for. As a data engineer, you will be responsible of monitoring the health and traffic of the hosted model on the cloud. Some jobs could fail due to several reasons and you should be able to alert data scientists of such failed jobs.

Which of the following approaches is best to implement on Google Cloud?

A. Use Vertex AI to host the model. Use Cloud Monitoring to monitor the status of jobs for 'failed' status.

B. Use Google Kubernetes Engine to host the model. Use Stackdriver to monitor the status of jobs for 'failed' status.

C. Use AutoML Vision to host the model. Use Stackdriver to monitor the status of jobs for 'failed' status.

D. Use Google Kubernetes Engine to host the model. Use Stackdriver to monitor the status of operations for 'error' status.

Question 7

You want to launch a Cloud Machine Learning Engine cluster to deploy a deep neural network model built by Tensorflow by data scientists of your company. Reviewing the standard tiers available by Google ML Engine, you could not find a tier that suits the requirements data scientists need for the cluster. Google allows you to specify custom cluster specification.

Which of the following specifications you are allowed to set? (Choose 2)

A. workerCount

Correct selection

B. parameterServerCount

C. masterCount

D. workerMemory

Question 8

Currently, anyone can access and modify the data sets, as well as creating & deleting data sets. The security team wants to restrict access of users on BigQuery and assign the

minimum roles for each team based on their task requirements. You have the following teams:

- Data scientists: They should have read & write access to data sets. They may create & delete data sets.
- Data analysts: They have read access to data sets only.
- Development team: They need to create jobs to run queries for updating the website's stats and product details.

What roles are recommended for each team?

A. Assign 'roles/bigquery.dataOwner' role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.user'role to the development team.

B. Assign 'roles/bigquery.admin' role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.dataViewer' role to development team.

C. Assign owner role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.dataViewer' role to development team.

D. Assign admin role to data scientists. Assign 'roles/bigquery.dataViewer' role to data analysts. Assign 'roles/bigquery.user' role to development team.

Question 9

Marketing team in your company wants to access certain tables in BigQuery. These tables are stored along with other tables considered sensitive and it's not approved to be accessed by marketing team. You need to restrict marketing team's role to only read from the tables they are allowed to.

Which of the following actions will you take?

A. Assign marketing team's roles as viewers on these specific tables. They won't be able to select from other tables in the dataset.

B. Create authorized views on tables marketing team wants to read from on the same dataset tables reside in. Grant viewer role to marketing team on the views.

C. Create a new dataset in BigQuery. Grant viewer role to marketing team on the new dataset. Copy the tables from the current dataset to the new one.

D. Create a new dataset in BigQuery. Create authorized views on tables marketing team wants to read from. Grant viewer role to marketing team on new dataset.

Question 10

You have about 20TB of data which is not accessed and the data team decided to archive them in the cloud. The team is looking for a storage solution that is highly available with minimum costs. On the other hand, the data may be accessed a few times a year for reconciliation purposes. Reconciliation frequency is per month.

Which of the following choices best satisfy data team's requirements?

A. Google Storage Standard.

B. Google Storage Nearline.

C. Google Storage Coldline.

D. BigQuery.

Question 11

You have a complex data pipeline which has a combination of shell scripts, python code and spark jobs. These task are scheduled by cron jobs to run. The problem with this approach is, in case of failure, the whole pipeline breaks and failure control with stopping next tasks from running after a certain task fails and re-running the pipeline again is difficult and messy. You want a solution which can manage the pipeline's different jobs to be failure-resilient, scalable and easy to monitor.

What approach is best for this scenario?

A. Use Cloud Composer to orchestrate the pipeline workflow.

B. Use Dataproc for Apache Spark jobs and migrate all other tasks to use Apache Spark instead.

C. Use Cloud Scheduler to schedule pipeline's tasks.

D. Use Google Workflows to orchestrate the pipeline

Question 12

A dairy products company is using sensors installed around different areas in its farms to monitor employees activities and detect any intruders. Apache Kafka cluster is used to gather the events coming from sensors. Recently, Kafka cluster is becoming a bottleneck causing lag in receiving sensor events. Turns out sensors are sending more frequent events and due to the company expanding with more farms, more sensors are installed and this will cause extra load on the cluster. What is the most resilient approach to solve this issue?

A. Use pub/sub to ingest and stream sensor events.

B. Scale out Kafka cluster to withstand the continuously flowing event stream.

C. Spin up a new Kafka cluster and distribute sensors even streams between the two clusters.

D. Deploy Confluent's Managed Apache Kafka Cluster from the marketplace to scale the cluster according to workload

Question 13

You are building a streaming data pipeline for a VOD (Video-on-demand) service company. It receives event data from its player app sending details of what users are watching, video state (play, pause, loading), and other metrics that can be derived from the device used such as OS, brand, and screen resolution.

The event data collected should be analyzed by most recent data for quality check and further action in case of streaming issues. How can you ingest the stream data?

A. Use Cloud Pub/Sub to ingest the events and attach a unique ID to every event in the publisher.

B. Use Cloud Pub/Sub to ingest the events and attach timestamp to every event in the publisher.

C. Use Cloud Pub/Subb to ingest the events and store them in BigQuery using a Pub/Sub Bigquery subscription.

D. Launch a compute engine and install Apache Kafka to ingest the event stream.

Question 14

Data analysts are using Google Data Studio to build dashboards reading data from BigQuery as a data source. The CTO wants to minimize the costs of BigQuery queries run by dashboards. You suggested enabling predictive (pre-fetch) caching.

Which of the following will minimize the costs?

A. Restrict data fetch to be once every 24 hours and make sure Data Studio report has view credentials on the BigQuery dataset.

B. Enable pre-fetch caching for the report and make sure Data Studio report has view credentials on the BigQuery dataset.

C. Enable pre-fetch caching for the report and make sure Data Studio report is an owner on the BigQuery dataset.

D. Restrict data fetch to be once every 24 hours and make sure Data Studio report is an owner on the BigQuery dataset.

Question 15

You have deployed a Tensorflow machine learning model using Cloud Machine Learning Engine. The model should be able to handle high volume of instances in a job to run complex models. The model should also write the output to Google Storage.

Which of the following approaches is recommended?

A. Use online prediction when using the model. Batch prediction supports asynchronous requests.

B. Use batch prediction when using the model. Batch prediction supports asynchronous requests.

C. Use batch prediction when using the model to return the results as soon as possible.

D. Use online prediction when using the model to return the results as soon as possible.

Question 16

A weather station facility which receives events from sensors installed on different pods distributed around the region return the current weather temperature based on sensor's location.

You are asked to build a pipeline to aggregate the incoming events to get the average temperature every 60 seconds for each region.

A. Tumbling window with a duration of 60 seconds.

B. Hopping window with a duration of 60 seconds.

C. Session window with a time gap duration of 60 seconds.

D. Global window with the time-based trigger of 60 seconds.

Question 17

A gaming app allows up to 32 players to compete in battle royale mode in a single gaming session. Recently, players are sending feedback complaining some users are idle and not competing in the session which breaks the experience for them. The development team decided to end the session for players who are idle for more than 60 seconds to solve this problem.

Gaming app sends events every second contain player's state (active, idle, pending) and other details. You want to build a Dataflow pipeline which aggregates these events so idle players can be detected in the time frame specified by development team.

Which windowing function you should choose to design the pipeline?

A. Tumbling window with a duration of 60 seconds.

B. Hopping window with a duration of 60 seconds.

C. Session window with a time gap duration of 60 seconds.

D. Global window with the time-based trigger of 60 seconds.

Question 18

Your company uses Google Cloud as its main platform. The lead architect decided to migrate the architecture's relational databases to Cloud Spanner for horizontal scaling and multi-regional availability. Upon using Cloud Spanner for a while after migration and monitoring its performance, it was reported Cloud Spanner instance's performance is not as expected as in the planning phase. What could be a possible reason for this? (Choose 2).

A. UUID is used as primary keys for the tables.

B. Primary keys are monotonically increased.

C. Primary keys are randomly generated 16-byte alphanumericals.

D. Primary key used is a combination of timestamp and original primary key as #timestamp-pk.

Question 19

You have a Dataflow pipeline which streams data to be stored to BigTable after it has been transformed and enriched. Development team needs to modify the transformation code based on client's needs. The pipeline is in production which keeps streaming and any interruption to the pipeline may lead to data loss or unexpected output.

How can you make sure the pipeline can be stopped without any consequences?

A. Turn off Dataflow pipeline with 'cancel' option.

B. Create a new Dataflow pipeline with the new transformation code, then switch data stream to the new pipeline.

C. Transfer Dataflow pipeline to write data to Google Storage. Perform the needed changes then transfer pipeline back to write to BigTable and re-process the data written in Google Storage.

D. Turn off Dataflow pipeline with 'drain' option.

Question 20

You are using Dataflow SDK to analyze data related to customer segmentation. You need to extract certain fields from the data files to be processed for further transformation.Which operation is used to perform the operation required?

A. ParDo

B. PCollection

C. Transform

D. Pipeline

Question 21

An e-wallet company is designing a relational database solution for their e-payment transactions. Database will face high read/write transactions and accessed from different parts of Europe and may be expanded to other continents in the future. The database should be scalable and able to scale out to meet high demands.

What is the best approach for this scenario?

A. Use Cloud BigTable as a database. For scaling out, monitor CPU utilization and increase nodes when more than 75% of CPU is utilized in a 15-minute timespan.

B. Use Cloud SQL as a database. For scaling out, monitor disk utilization and increase nodes when more than 85% of storage is utilized in a 15-minute timespan.

C. Use Cloud BigTable as a database. For scaling out, monitor memory utilization and increase nodes when more than 80% of memory is utilized in a 15-minute timespan.

D. Use Cloud Spanner as a database. For scaling out, monitor CPU utilization and increase nodes when more than 65% of CPU is utilized in a 15-minute timespan.

Question 22

An e-payment service generates thousands of gigabytes of logs every month which are streamed to a Dataflow pipeline, transformed, and stored in a data warehouse for further analysis. These raw logs are not accessed once the transformation is done by Dataflow. CTO suggested that log files should be archived after a month of the transformation and after making sure that the data is not required for debugging. Which of the following storage types are recommended?

A. Google Storage Standard

B. Google Storage Nearline

C. Google Storage Coldline

D. Google Storage Archive

Question 23

As a solution for a serverless data warehouse, you decided to use BigQuery to store and query data. You built a Dataflow pipeline to read data from Google Storage and import it to BigQuery. You added a few users to access BigQuery for reporting purposes. You want to monitor the activity on BigQuery by getting details about query count and execution time. You want such metrics to appear on a dashboard to be shared later with other stakeholders. What should you do?

A. Build a script to use gcloud command to extract queries execution time and data size scanned every 1 hour. Send the stats to Operation Suite and create a dashboard showing the metrics.

B. Use Cloud Monitoring to create a dashboard and graphs showing query metrics.

C. You need to contact Google Cloud support in order to enable metrics on BigQuery UI.

D. From BigQuery UI, you can view run queries and execution time. You can share it by exporting the stats to a file.

Question 24

A company decides to migrate its on-premise data infrastructure to the cloud mainly for high availability of cloud services and to lower the high costs of storing data on-premise. The infrastructure uses HDFS to store data and be processed and transformed using Apache Hive & Spark. The company wants to migrate the infrastructure and DevOps team still wants to administrate the infrastructure in the cloud. As a data architect, which of the following is the approach recommended by Google?

A. Use ephemeral Dataproc cluster with preemptible VMs to process the data and Store data in Google Cloud Storage with an object lifecycle management policy.

B. Build a Dataflow pipeline. Store the data in Google Storage. Use Cloud Compute to launch instances and install the required dependencies for processing the data.

C. Use Dataproc to process the data. Store data in Dataproc's HDFS.

D. Build a Dataflow pipeline. Store the data in persistent disks in HDFS. Execute the code in Spark framework provided by Dataflow.

Question 25

Your team was working on a development BigTable instance for some time experimenting on it to stream event data coming from hundreds of sensors sending events frequently. The team lead considered instead of deleting the instance and losing all events collected since building the pipeline, it would be a better idea to use the instance in production with the required changes to ensure high availability and best performance.

Which of the following approaches is best to satisfy the team lead's requirements, given that you are currently using HDD for development purposes?

A. Export the data from BigTable development instance to Google Storage, launch a new BigTable production instance with SSD storage type, then load the data from Google Storage to the new BigTable instance.

B. Export the data from BigTable development instance to Google Storage, launch a new BigTable production instance with HDD storage type, then load the data from Google Storage to the new BigTable instance.

C. Change BigTable instance type from development to production, scale up number of nodes and ensure the storage type is HDD.

D. Change BigTable instance type from development to production, scale up number of nodes and ensure the storage type is SSD.