Google Professional Data Engineer certification practice questions

Question 1: **Correct**
An insurance company wants to implement a chatbot service to help direct customers to the best customer support team for their questions. What GCP service would you recommend?

- **Text-to-Speech API**
- **Speech-to-Text API**
- **AutoML Tables**
- **Dialogflow (Correct)**

Explanation
The correct answer is Dialogflow is a service for creating conversational user interfaces. Speech-to-Text converts spoken words to written words. Text-to-Speech converts text words to human voice-like sound. AutoML Tables is a machine learning service for structured data. See https://cloud.google.com/dialogflow/docs

Question 2: **Correct**
You are training a deep learning model with a relatively small set of features and a large number of instances. The model is not performing as well as you like. You believe the model is underfitting. What technique would you try to improve performance?

- **AI**
-

Use feature crosses
(Correct)

- **Use gradient descent**

- **Use backpropagation**

Explanation
Since there are a large number of training instances and few features, this is a good candidate for adding synthetic features by using feature crosses. Dropout is a type of regularization used to prevent overfitting. Backpropagation is an algorithm for adjusting parameters in a neural network. Gradient descent algorithm is used in optimization problems. See https://developers.google.com/machine-learning/crash-course/feature-crosses/encoding-nonlinearity

Question 3: **Correct**
Machine learning engineers working in the us-central1 region have approximately 200 TB of data that will be used to train machine learning models. To train a model, only a small subset of that data is used. Data is organized into files that will be accessed about once per month. You would like to minimize storage costs but still have reliable and highly available storage. What would you recommend for storing this data?

- **Cloud Storage Multi-Region storage**

- **Cloud Storage Nearline storage**
(Correct)

- **SSD Persistent Disks**

- **Balanced Persistent Disks**

Explanation
Cloud Storage Nearline storage is the best option for highly available data that is accessed once per month. Cloud Storage Multi-region meets the need but is more expensive and since the team is working in one region, multi-region storage is not necessary. Persistent disks are used with virtual machines and would require additional cost and operational overhead to store this data on persistent disk. Cloud Storage is a managed service that does not require as much operational overhead and would therefore cost less. See https://cloud.google.com/storage/docs/storage-classes

Question 4: **Incorrect**
Your company is migrating an on-premises long-term storage archive to Google Cloud. The archived files are accessed on average about once every 30 days. You would like to minimize the cost of storage. What storage option would you recommend?

- **Nearline Storage (Correct)**
- **Coldline Storage (Incorrect)**
- **Multi-regional storage**
- **Persistent Disks**

Explanation
Nearline Storage is a class of Cloud Storage designed for objects that will be accessed at most once every 30 days. Coldline Storage is suitable for objects

accessed at most once per 90 days. Multi-regional storage is best suited for objects that should have low latency access from multiple regions. Persistent disks are used with Compute Engine instances and Kubernetes Engine clusters and should not be used for archival storage. See https://cloud.google.com/storage/docs/storage-classes

Question 5: **Correct**
A European health care company uses Cloud Pub/Sub as part of a data processing pipeline. The CTO of the company is concerned that data might accidentally be written to a region outside the European Union, which would violate the GDPR regulation. What would you recommend the company does to ensure data stays within Google Cloud regions in the European Union?

- **Set a Resource Location Restriction organization policy to ensure all topics are stored only in acceptable regions. (Correct)**
- **Set a Resource Location Restriction organization policy to ensure all buckets are stored only in acceptable regions.**
- **Only define Cloud Pub/Sub endpoints in acceptable regions when creating topics.**
- **Only define Cloud Pub/Sub endpoints in acceptable regions when creating subscriptions.**

Explanation
The correct answer is to set a Resource Location Restriction organization policy to ensure all topics are stored only in acceptable regions. Topics, not buckets, store messages in Cloud Pub/Sub. Users of Cloud Pub/Sub do not create endpoints; it is a globally managed service that does not require any endpoint configuration by users of Cloud Pub/Sub. see https://cloud.google.com/resource-manager/docs/organization-policy/defining-locations

Question 6: **Incorrect**

A group of data analysts has asked for your help setting up a Cloud Dataproc cluster to analyze large data sets using Spark. The cluster will run many jobs. You plan to follow Google recommended best practices. Which of the following would you do? (choose 2)

- **Use HDFS storage on persistent disks**
- **Use Cloud Storage for persistent storage (Correct)**
- **Use no more than 30% preemptible VMs for secondary workers (Correct)**
- **Use preemptible VMs only on workers that run HDFS (Incorrect)**
- **Disable autoscaling**

Explanation
Using Cloud Storage for persistent storage is recommended over using HDFS on local disks. This allows data to persist when a cluster is shut down without having to copy data from the cluster to Cloud Storage. Google recommends a maximum of 30% preemptible VMs as secondary nodes. Autoscaling is recommended when using Cloud Storage and the cluster runs many jobs. See https://cloud.google.com/blog/topics/developers-practitioners/dataproc-best-practices-guide

Question 7: **Correct**
A machine learning model is not performing as well in production as the validation tests would predict. You suspect the model is overfitting. What technique can you use during training to reduce the risk of overfitting?

L2 Regularization (Correct)

-

Gradient descent

-

Backpropagation

-

Label engineering

Explanation
L2 regularization is a technique for preventing overfitting. Gradient descent is a general approach to finding optimal values. Backpropagation is an algorithm to adjusting weights in a neural network. There is no such thing as label engineering in ML; feature engineering is a practice for determining additional features used to model training. See https://cloud.google.com/bigquery-ml/docs/preventing-overfitting

Question 8: **Incorrect**
The performance of an application that uses Bigtable is starting to degrade as volumes grow. You suspect your row key design may not be optimal. You want to review access patterns for a group of row keys. What tool would you use?

-

Cloud Monitoring

-

Cloud Logging

-

- **Key Visualizer (Correct)**
- **Cloud Key Trace (Incorrect)**

Explanation
Key Visualizer is a Bigtable tool for analyzing Bigtable usage patterns. Cloud Monitoring can be used for performance monitoring but it is not specifically designed to diagnose row key design problems. Cloud Logging can help when reviewing events in Bigtable but it is not designed to support row key design problems. There is no Google Cloud tool called Cloud Key Trace. See https://cloud.google.com/bigtable/docs/keyvis-overview

Question 9: **Incorrect**
You want to monitor a Cloud Dataflow job and know the maximum duration that an item has been waiting in the pipeline. What Cloud Monitoring metric would you use to track the maximum duration?

- **job/data_watermark_age**
- **job/system_lag (Correct)**
- **job/elapsed_time**
- **job/element_count (Incorrect)**

Explanation
The correct answer is job/system_lag. The job/data_watermark_age is the age of the most recent item that's been fully processed by the pipeline. Job/elapsed_time is the elapsed time of the pipeline run time. Job/element_count is the number of items processed in a PCollection for the Read_input and Process_element transforms. See https://cloud.google.com/architecture/building-production-ready-data-pipelines-using-dataflow-monitoring

Question 10: **Incorrect**
A developer tries to create a service account for a data pipeline but is unable to complete the operation. Which of the following could be the cause?

- **A policy has been applied to the resource hierarchy that enforces the constraints/iam.disableServiceAccountKeyCreation constraint.**

  (Correct)

- **The developer has not specified a properly configured deployment.yaml file. The yaml file should be corrected.**

- **The developer has not properly configured a Domain Name Services (DNS) A record. An A record should be added to DNS.**

- **A policy has been applied to an IAM group that disables the permission to create service accounts. That policy should be dropped.**
  (Incorrect)

Explanation
The constraints/iam.disableServiceAccountKeyCreation is enabled and that

prevents principals from creating user-managed service account keys. The constraint should be removed to allow developers to create a service account. Deployment.yaml files are used to configure Kubernetes deployments. DNS A records associate an IP address with a domain name. Policies are not attached to identity types, they are attached to entities in the resource hierarchy. See https://cloud.google.com/resource-manager/docs/organization-policy/org-policy-constraints

Question 11: **Incorrect**
You have a latency-sensitive application that uses Bigtable. You want to follow Google Cloud recommended best practices. What would you do?

- **Use a service account for all read operations**
- **Keep storage utilization per node below 60% (Correct)**
- **Use HDD storage instead of SSD storage**
- **Use a global load balancer in front of Bigtable (Incorrect)**

Explanation
For low latency applications, Google Cloud recommends keeping storage utilization below 60%. Using service accounts for read operations will not affect latency of write operations. HDD storage is less performant than SSD storage. A load balancer is not needed to distribute workload within a Bigtable cluster. Bigtable distributes operations based on row keys. See https://cloud.google.com/blog/products/databases/check-out-how-to-optimize-database-service-cloud-bigtable

Question 12: **Correct**
You have been tasked with ensuring the successful transfer of 100 TB of data from the AWS S3 object storage system. This is a one time transfer. A complete

and reliable transfer of all data is a top priority. How would you recommend loading this data into Cloud Storage?

- **gsutil cp**
- **Cloud Dataflow**
- **Cloud Storage Transfer Service (Correct)**
- **Transfer Appliance**

Explanation
Cloud Storage Transfer Service is designed to load terabytes of data using scheduled jobs and is well suited for transferring data from AWS. Gsutil is a command line utility for working with Cloud Storage but not designed to upload large volumes of data. Cloud Dataflow is a batch and stream processing system but is not designed for large data transfers from other public clouds. Transfer Appliance is designed for large data loads but requires attaching a storage device to the source system's network and so suitable only when you have physical access to the source system network. See https://cloud.google.com/storage-transfer-service

Question 13: **Correct**
A user of a Cloud Dataproc cluster needs permission to stop a cluster. They will also need to instantiate workflow templates and other common user tasks. You want to follow Google Cloud recommend best practices for security. What role would you use to grant permission to stop a cluster?

- **roles/dataproc.editor (Correct)**
-

roles/dataproc.admin

-

roles/dataproc.viewer

-

Create a custom role with only permissions to stop the cluster and imitate workflows.

Explanation
Roles/dataproc.editor will provide permissions to stop clusters, initiate workflow templates, and other common user tasks. Roles/dataproc.admin would provide more privileges than needed and violate the principle of Least Privilege. Roles.dataproc.viewer would not provide permission needed. There is no need to create a custom role and Google Cloud recommends using predefined roles when they meet your needs and only use custom roles when a predefined role does not exist that meets your needs. See https://cloud.google.com/dataproc/docs/concepts/iam/iam

Question 14: **Correct**
You are using Cloud Dataflow to process data that is represented as key value pairs. What Apache Beam construct will you likely use in your workflow?

-

Database connection

-

PCollections

(Correct)

-

User-defined function (UDF)

- **Watermark**

Explanation
The correct answer is a PCollection, which is a representation of key value pair data. A database connection may be used but it is not necessarily required. A watermark is used with streaming data but there is no indication it is needed for this use case. User defined functions may be used but there is no requirement for functionality not already provided by Cloud Dataflow. See https://cloud.google.com/dataflow/docs/concepts/beam-programming-model

Question 15: **Incorrect**
When using Cloud Data Fusion you receive an error that a Dataproc operation failed and the user is not authorized to act as a service account. What would you do to correct this problem?

- **Create a Cloud Dataproc cluster before starting Cloud Data Fusion.**

- **Grant the Service Account User role to Cloud Data Fusion. (Correct)**

- **Grant the Service Account User role to Cloud Dataproc. (Incorrect)**

- **Ensure both Cloud Data Fusion and Cloud Dataproc are running in the same zone.**

Explanation

The correct answer is to grant the Service Account User role to Cloud Data Fusion. Cloud Dataproc does not need that role assigned to it. This is an access control issue and not related to the location of clusters. Cloud Data Fusion will manage its use of Cloud Dataproc clusters. See https://cloud.google.com/data-fusion/docs/concepts/service-accounts

Question 16: **Correct**
A team of researchers is running a high performance distributed computing platform on premises but wants to migrate to Google Cloud. The platform uses virtual machines. The researchers want to be able to scale up the number of virtual machines in the cluster based on CPU load. What would you recommend they use?

- **Kubernetes cluster**
- **Managed instance groups (Correct)**
- **Unmanaged instance groups**
- **Cloud Run**

Explanation
The correct answer is managed instance groups, which is a way of deploying Compute Engine instances basesed on a template. Kubernetes is used for running containers, not virtual machines. Cloud Run is also used to run containers not virtual machines. Unmanaged instance groups run virtual machines but do not support autoscaling. See https://cloud.google.com/compute/docs/instance-groups

Question 17: **Correct**
A manufacturer of delivery drones has equipped drones with multiple sensors that send performance and environment data to the analytics pipeline. Temperature received over the past hour is analyzed and if any temperature reading is more than 2 standard deviations away from the mean for the past

hour, an alert is triggered. You would like to build the analysis pipeline using a managed service. What Google Cloud service would you recommend?

- **Cloud Dataflow (Correct)**
- **Cloud Dataproc**
- **Cloud Data Fusion**
- **Cloud Firestore**

Explanation
Cloud Dataflow is a managed Apache Beam runner used for stream and batch processing and is the best choice. Cloud Dataproc is a managed Hadoop/Spark cluster service. Cloud Data Fusion is an extraction, transformation, and load service typically used with data warehouses and related data analytic services. Cloud Firestore is a NoSQL document database service. See https://cloud.google.com/dataflow/docs/concepts

Question 18: **Correct**
A retailer has been using Kubernetes to deploy new applications built on microservices architectures. They now want to start building machine learning pipelines while leveraging their expertise in Kubernetes. What service would you recommend for running machine learning workflows on Kubernetes?

- **Vertex AI**
- **AutoML Tables**

- **Kubeflow (Correct)**
- **Cloud Composer**

Explanation
The correct answer is Kubeflow, an open source tool for running ML pipelines in Kubernetes. Vertex AI is a set of managed machine learning services in Google Cloud. AutoML Tables is a managed ML service specifically for structured data. Cloud Composer is a workflow orchestration service. See https://cloud.google.com/blog/products/ai-machine-learning/getting-started-kubeflow-pipelines

Question 19: **Correct**
A manufacturer of delivery drones has been using a PostgreSQL database running in Compute Engine to store data. The company is growing and the database is not able to keep up with the ingestion of telemetry data from the drones. The CTO would like to use a managed database service that will provide low latency writes and scale to petabytes of data. The top priority is scalability and the CTO is willing to invest development time in changing the application if needed. What managed Google Cloud database service would you recommend?

- **Cloud SQL using PostgreSQL**
- **Cloud Spanner**
- **Cloud Bigtable (Correct)**

BigQuery

Explanation
Cloud Bigtable is designed for low latency, high volume writes and scales to petabyte sized databases making it the best choice. There would likely be some changes to the application to work with Cloud Bigtable, which is a NoSQL database (PostgreSQL is a relational database). Cloud SQL using PostgreSQL will not scale to meet the requirements. BigQuery is an analytical database that can scale to petabytes but it is not designed for low latency writes such as needed in this application. Cloud Spanner can scale to petabytes and is a relational database but it is not designed for the kind of low latency writes needed here. Also, key features of Cloud Spanner, such as SQL query language and strong consistency are not mentioned in the requirements. See https://cloud.google.com/bigtable/docs/overview

Question 20: **Correct**
A North American retailer is planning to expand to Europe and specifically target individuals from ages 20 to 40 and living in Spain, France, Belgium, and Germany. The retailer plans to create detailed profiles about customer preferences so they can make recommendations. What regulations will the company need to comply with when it expands as planned? (Choose 2)

- **HIPAA**
- **SOX**
- **GDPR**
  (Correct)
- **PCI Data Security Standard**
  (Correct)

- **Expedited Funds Transfer Act**

Explanation
Retailers receive payment via payment cards and so are subject to the Payment Card Industry (PCI) Data Security Standard. Since the company will have data on European Union citizens, it must comply with GDPR. HIPAA is a healthcare regulation in the United States. Sarbanes Oxley (SOX) is a US regulation on public companies designed to prevent fraudulent accounting practices. There is no Expedited Funds Transfer Act. See https://cloud.google.com/security/compliance/pci-dss and https://cloud.google.com/privacy/gdpr

Question 21: **Incorrect**
A Cloud Spanner database is experiencing hot-spotting. You suggest changing the primary keys of tables in the database. What methods would you consider when defining new keys? (Choose 2).

- **Hash value of existing primary key**
(Correct)

- **Auto-incrementing values**

- **Big-reversed sequential values**
(Correct)

- **Timestamps**
(Incorrect)

Start primary key with low cardinality attribute

Explanation
Hash values of existing primary keys and bit-reversed sequential values would both provide well distributed keys and help avoid hot spotting. Auto-incrementing values, timestamps, and low cardinality attribute can all lead to hot spotting. See https://cloud.google.com/spanner/docs/schema-design

Question 22: **Correct**
You are uploading several hundred files to Google Cloud using gsutil rsynch. The set of files fails to fully upload. You'd rather not reload files that were successfully uploaded. What command would you use to resume the rsynch operation?

- **The same command that was used initially. Gsutil rsynch will automatically resume.**
  (Correct)

- **The gsutil rsynch resume command**

- **The same command that was used initially and the --resume parameter.**

- **The same command that was used initially and the --upload parameter.**

Explanation
The correct answer is the same command that was initially used. There is no need to specify other parameters to resume a gsutil rsynch command. See https://cloud.google.com/storage/docs/gsutil/commands/rsync

Question 23: **Correct**
Your company has created a new data analytics team. Data analysts will need to

read data from and write data to Cloud Storage and query data from BigQuery. Data engineers will also need to create Cloud Storage buckets and set data lifecycle management policies on buckets. You want to follow Google Cloud's recommended best practices. How would you manage access permission for the new team?

- **Grant roles to each user individually. Assign data engineers the same roles as data analysts and additional roles needed for their additional responsibilities.**

- **Create a group for data analysts and a group for data engineers. Add the identities of data analysts to the data analyst group. Add the identities of the data engineers to the data engineer group. Grant roles to the data analyst group to allow access needed by data analysts. Grant roles to the data engineer group needed by the data engineers.**
  (Correct)

- **Create a group for the data analytics team. Grant the group all roles needed by data analysts and data engineers to that group. Add the identities of all team members to the group.**

- **Grant roles to each user individually. Assign data engineers and data analysts the roles needed by either data analysts or data engineers.**

Explanation
Roles should be assigned to groups not individual identities. Each group should only have the roles needed to perform their job responsibilities in accordance with the Principle of Least Privilege. The correct answer is to crate two groups, assign data analysts to the data analyst group and data engineers to the data engineer group. Grant each group only the roles needed by that group. See https://cloud.google.com/iam/docs/recommender-best-practices and https://cloud.google.com/iam/docs/understanding-custom-roles

Question 24: **Incorrect**

A BigQuery data warehouse needs to access some data in Cloud Storage using external tables. Which of the following is not a supported file format for external tables based on files in Cloud Storage?

- **Avro**
  (Incorrect)
- **ORC**
- **Firestore export files**
- **Excel xlsx format**
  (Correct)

Explanation

The Excel xlsx file format is not supported for files stored in Cloud Storage. Avro, CSV, newline-delimited JSON, Datastore export files, Firestore export files, ORC, and Parquet files are supported. See https://cloud.google.com/bigquery/external-data-sources

Question 25: **Correct**

A data scientist is just learning to use Google Cloud for analytics. They would like to perform data quality checks and exploratory analysis on data sets stored in Cloud Storage. What Google Cloud service would you recommend they use?

- **Cloud Dataflow**
-

- **Cloud Dataprep (Correct)**

- **Cloud Dataproc**

- **Vertex AI**

Explanation
Cloud Dataprep is a managed service for preparing data for analysis and machine learning, including exploratory analysis. Cloud Dataflow is used for stream and batch processing. Cloud Dataproc is a managed Spark and Hadoop cluster service. Vertex AI is a comprehensive machine learning service. See https://cloud.google.com/dataprep/docs/concepts

Question 26: **Correct**
A team of data analysts want to run a series of jobs on large data sets. There is a complicated set of dependencies between the jobs. They want to use a managed service if possible. Which of the following would you recommend they try?

- **Write Airflow directed acyclic graphs in Python and execute them with Cloud Composer. (Correct)**

- **Write Airflow directed acyclic graphs in SQL and execute them with Cloud Composer.**

- **Write Airflow directed acyclic graphs in SQL and execute them**

with Cloud Workflows.

- **Write Airflow directed acyclic graphs in Python and execute them with Cloud Workflows.**

Explanation
The correct answer is to write Airflow directed acyclic graphs in Python and execute them with Cloud Composer. Cloud Composer does not support writing directed acyclic graphs in SQL. Cloud Workflows is used with API workflows, not complicated batch job workflows. See https://cloud.google.com/composer

Question 27: **Correct**
A data analyst currently has the bigquery.dataViewer role and can successfully query a materialized view. They also want to be able to refresh the materialized view. You want to use a predefined role but not grant them any more permissions than needed to refresh the materialized view. What predefined role would you grant to the user?

- **bigquery.dataOwner**
- **bigquery.admin**
- **bigquery.dataEditor (Correct)**
- **bigquery.mvUpdater**

Explanation
The correct answer is bigquery.dataEditor, which can refresh a materialized

view. Both bigquery.admin and bigquery.dataAdmin grant more permissions than needed. There is no bigquery.mvUpdater role. See https://cloud.google.com/bigquery/docs/access-control

Question 28: **Correct**
A financial services company is using a single Bigtable cluster to store data about equity prices. There is a large volume of write operations during the trading day. There are also analytic batch jobs that run through the day. You have been hired to help optimize the performance of Bigtable. What would you recommend they do?

- **Isolate the write and batch workloads by adding a second cluster to the Bigtable instance and create two app profiles, one for write traffic and one for batch jobs.**
(Correct)

- **Continue to write data to Bigtable but create a Cloud Dataflow job to copy data to a Cloud Spanner data warehouse for batch operations.**

- **Continue to write data to Bigtable but create a Cloud Dataflow job to copy data to a Cloud Firestore data warehouse for batch operations.**

- **Isolate the write and batch workloads by adding a second set of tables to the Bigtable instance and write the data needed by batch jobs to the second set of tables.**

Explanation
To isolate batch analytics jobs from other operations in Bigtable, Google Cloud recommends using two clusters in a single instance and using app profile to route operations to the appropriate cluster. Cloud Spanner and Cloud Datastore are not designed for data warehousing. A second set of tables would be managed by the same set of nodes and not reduce the total workload on the nodes in the cluster. See https://cloud.google.com/bigtable/docs/performance

Question 29: **Correct**

A team of data analysts are proficient in using SQL but not programming in Java, Python, or other programming languages. They want to experiment with building machine learning models trained on relational data. They have approximately 1 TB of data to work with. What would you recommend they use?

- **Bigtable**
- **Cloud TPUs**
- **BigQuery ML (Correct)**
- **Cloud Fusion**

Explanation

The correct answer is BigQuery ML, which incorporates SQL functions to build, evaluate, and invoke machine learning models within SQL. Cloud TPUs are accelerators and are used with deep learning applications. Cloud Fusion is an ETL service. Bigtable is a NoSQL database for high volume, low latency write applications, such as IoT data ingestion and storage. See https://cloud.google.com/bigquery-ml/docs/introduction

Question 30: **Correct**

As a database administrator, you are finding that a Cloud SQL database using PostgreSQL is not meeting read operation SLAs. You want to improve performance with minimal changes to database applications. What would you try first to improve read performance?

- **Create a read replica.**

(Correct)

- **Use Cloud Memorystore to cache data to be read.**

- **Use change data capture to keep a second database in a different region in synch with the primary database while having read operations sent to the secondary database.**

- **Use PostgreSQL's explain plan feature to analyze queries and rewrite them to improve performance.**

Explanation
The correct answer is to create a read replica for read operations. Using a cache could improve read performance but would require changes to database applications. Using a secondary database would require changes to the application to read from the secondary database instead of the primary. Using explain plan could help with query optimization but that also requires changes to database applications. See https://cloud.google.com/sql/docs/mysql/replication

Question 31: **Correct**
You have a real-time monitoring application that streams data to Bigtable. It is not performing as well as expected. You use a row key that starts with a unique ID of each source system. Each source system sends 500K of data per minute and that is written to one row. There are approximately 200 column families, each having on average 10 columns. What could be the cause of the poor performance?

- **500K is more data than Bigtable can efficiently ingest per row**

-

- **200 column families exceeds the recommended 100 column family limit**
  (Correct)

- **Row keys should not start with a unique ID**

- **10 columns per column family is exceeds the recommended 5 columns maximum per column family**

Explanation
Google recommends limiting a table to no more than 100 column families otherwise performance can degrade. Bigtable can store up to 10MB per row so 500K is not too much data. Row keys should start with non-sequential prefixes to avoid hot spotting. 10 columns is not too many columns for a column family. See https://cloud.google.com/bigtable/docs/schema-design

Question 32: **Correct**
A retailer is building machine learning models to help predict the number of products that will be sold and therefore should be available in inventory. What kind of model should they build?

- **Regression**
  (Correct)

- **Classification**

- **Feature**

-

Reinforcement

Explanation
The correct answer is a regression model, which is used to predict a value based on a set of input parameters. Classification models categorize or label inputs, such as determining if a vehicle is a car or a truck. A feature is an attribute used in a machine learning model to describe the characteristics of an instance used in training. Reinforcement learning is a kind of machine learning in which an agent learns from the environment. See https://cloud.google.com/automl-tables/docs/problem-types

Question 33: **Correct**
What technique is used in backpropagation to update parameters of a model during training?`

- **L1 or Lasso Regression**
- **L2 or Ridge Regression**
- **Feature crosses**
- **Gradient descent (Correct)**

Explanation
Gradient descent is an optimization algorithm that minimizes a function by iteratively moving in the direction of the steepest descent. L1 or Lasso and L2 or Ridge Regression are regularization techniques. Feature crosses are a way to create synthetic features from two or more features. See https://builtin.com/data-science/gradient-descent

Question 34: **Correct**
A group of analysts is migrating a Hadoop cluster from on premises to GCP.

They want to follow Google Cloud recommended best practices. What should they do as part of the migration?

- **Use ephemeral clusters and Cloud Storage instead of HDFS on local storage.**
  (Correct)

- **Use ephemeral clusters and use HDFS on local storage.**

- **Continually run clusters and use Cloud Storage instead of HDFS on local storage.**

- **Continually run clusters and use HDFS on local storage.**

Explanation
Google Cloud recommends using ephemeral clusters. Since clusters start quickly you do not need to keep clusters running to avoid long startup time. Google Cloud also recommends using Cloud Storage to store persistent data so data does not have to be copied from a cluster before shutting down and then copied back to a cluster when starting a new cluster. https://cloud.google.com/blog/topics/developers-practitioners/dataproc-best-practices-guide

Question 35: **Correct**
A manufacturer of delivery drones is implementing a new data analysis pipeline to detect part failures before they occur. The drones have multiple sensors that send performance and environment data to an analytics pipeline. Currently, data is sent to a REST API endpoint. The REST API endpoint that receives data cannot always keep up with the pace data is arriving. When that happens, data is lost. Machine learning engineers have asked you to change the ingestion process to reduce this data loss. What would you do?

- **Write data to a Cloud Pub/Sub topic instead of a REST API endpoint and have the ingestion application read from the topic.**

(Correct)

- **Write data to a Cloud Storage bucket instead of a REST API endpoint and have the ingestion application read from the bucket.**

- **Write data to a Cloud SQL Postgres database endpoint and have the ingestion application query the database.**

- **Create a Hadoop cluster in Compute Engine using managed instance groups and write data to an Hbase database and have the application query the database.**

Explanation
Write data to a Cloud Pub/Sub topic instead of a REST API endpoint and have the ingestion application read from the topic. In the event of a spike in data, Cloud Pub/Sub will buffer the data until it can be processed. Cloud Storage is an object storage system and is often used for ingesting large objects, such as images, videos or documents but Cloud Pub/Sub is a better way to ingest small amounts of data, such as telemetry data from an IoT sensor. Cloud SQL is not sufficiently low latency or scalable enough for this use case. HBase on Hadoop would require more administrative overhead than using Cloud Pub/Sub and would not scale as well as Cloud Pub/Sub for this use case. See https://cloud.google.com/dataflow/docs/concepts/streaming-with-cloud-pubsub

Question 36: **Correct**
A machine learning model has been containerized and deployed on Kubernetes Engine. It is currently deployed with 2 replicas. You need to increase the number of replicas to 4. What command would you use?

- **kubectl scale deployment command with the --replicas 4 parameter.**

  (Correct)

-

- **kubectl scale deployment command with the --replicas 2 parameter.**

- **kubectl scale deployment command with the --deploy 4 parameter.**

- **kubectl scale deployment command with the --deploy 2 parameter.**

Explanation
The kubectl scale deployment with the --replicas 4 parameter will scale the deployment to 4 replicas. There is no --deploy parameter in the kubectl scale deployment command. See https://kubernetes.io/docs/concepts/workloads/controllers/deployment/

Question 37: **Correct**
A data pipeline uses Cloud Pub/Sub for ingesting data. The data is stored in topics and a Dataflow workflow reads from a subscription to that topic, processes the data, and writes output to BigQuery. What is the recommended way to authenticate when reading data from Cloud Pub/Sub?

- **Custom role**

- **Google Workspace Identity**

- **Use service accounts**
  (Correct)

-

Basic role

Explanation
Service accounts are the recommended way to authentic for most use cases when using Cloud Pub/Sub. Google Workspace Identity should be used by human users, service accounts are used for applications. Custom roles and basic roles are for authorization not authentication. See https://cloud.google.com/iam/docs/service-accounts

Question 38: **Correct**
Data at rest in Google Cloud is encrypted at the hardware, infrastructure, and platform levels. What encryption algorithm is used for encryption at the infrastructure level?

- **Blowfish**
- **AES256 (Correct)**
- **DES**
- **RSA**

Explanation
Advanced Encryption Standard (AES) with a 256-bit key is used for encrypting data at the infrastructure level in Google Cloud. Blowfish and RSA are also strong encryption algorithms but they are not used at the infrastructure level of GCP. Data Encryption Standard (DES) is a symmetric key algorithm that uses 56 bits and is considered weak encryption and should not be used for securing data. See https://cloud.google.com/security/encryption/default-encryption

Question 39: **Correct**
A database administrator would like to migrate a PostgreSQL database to a

managed service in Google Cloud with minimal changes. The database is used by a team of researchers all located in Spain and France. Which of the following services would you recommend?

- **Cloud Spanner**
- **Cloud SQL**
  (Correct)
- **Cloud Bigtable**
- **Cloud Firestore**

Explanation
Cloud SQL is a regional SQL database managed service that supports PostgreSQL databases. Cloud Spanner is a global scale relational database which would cost more to operate than Cloud SQL. Cloud Bigtable and Cloud Firestore are both NoSQL databases and would not meet the requirements outlined here. See https://cloud.google.com/sql/docs/postgres

Question 40: **Incorrect**
The CIO of a online gaming company is concerned with the increasing cost of maintaining a MongoDB database used to store player game data. What managed service in Google Cloud would you recommend as an alternative option to MongoDB?

- **Cloud SQL**
- **Cloud Firestore**

(Correct)

- **Cloud Dataproc**
- **Bigtable**
  (Incorrect)

Explanation
MongoDB is a NoSQL database that uses a document model so Cloud Firestore is a good option for a managed service. Cloud SQL is a relational database. Cloud Dataproc is a managed Spark and Hadoop service, not a database. Bigtable is a NoSQL database but it is a wide column database, not a document database. See https://cloud.google.com/firestore

Question 41: **Incorrect**
You are creating a set of Cloud Storage buckets for storing data that will be accessed by several different teams in your organization. The teams have different access requirements. You want to follow Google Cloud's recommended best practices. How would you implement access controls to objects and buckets?

- **Uniform bucket level access**
  (Correct)
- **Fine-grained access controls**
  (Incorrect)
- **Signed URLs**
-

Signed policy documents

Explanation
Uniformed access control is the recommended method and uses IAM to apply permissions to buckets or groups of objects. Fine-grained access control is a legacy method based on access control lists and is not recommended. Signed URLs are used for time-limited access to an object. Signed policy documents are used to control what can be uploaded to a bucket. See https://cloud.google.com/storage/docs/uniform-bucket-level-access

Question 42: **Incorrect**
A financial services company is required to keep audit records for at least seven years. The data is unlikely to be accessed but must be kept anyway. The company has been storing this data in an on-premises file system but the CIO wants to a lower cost solution. The company is migrating several workloads to Google Cloud and is considering a Google Cloud-based solution. What would you recommend?

- **Cloud Storage Archive class storage (Correct)**
- **Cloud Storage Coldline storage (Incorrect)**
- **Cloud Storage Multi-Region storage**
- **Cloud Storage Dual-Region storage**

Explanation
Cloud Storage Archive class storage is the best choice for this kind of long term, low frequency access storage requirement. Coldline storage could be used but Archive class storage costs less. Multi-region and dual-region storage are not required according to the problem description and would cost more.

See https://cloud.google.com/storage/docs/storage-classes

Question 43: **Incorrect**
A Web hosting company has been using a custom built data store modeled on the sparse multidimensional array data structure. The CIO no longer wants to pay to develop and maintain a custom data store. Instead, the CIO wants a managed database service if possible and if not, they want to use a well supported open source database that is also based on sparse multidimensional array data structure. The Web hosting company is already using Google Cloud Compute Engine, Cloud Storage, and Kubernetes Engine. What would you recommend to the company?

- **Cassandra (Incorrect)**
- **Cloud Bigtable (Correct)**
- **BigQuery**
- **Cloud Spanner**

Explanation
Cloud Bigtable is a managed database service based on sparse multidimensional arrays and so meets the requirements, including using a managed database service. Cassandra is an open source database that is based on multidimensional sparse arrays but it is not a managed service and so not as good of a choice for the given requirements. BigQuery uses a compressed, columnar data model not a sparse multidimensional array and does not meet the requirements. Cloud Spanner is a global scale relational database and does not meet the requirements. See https://cloud.google.com/bigtable/docs/overview

Question 44: **Incorrect**

A group of data engineers will be working on several initiatives. Each initiative will have their own VMs, storage buckets, and sets of Cloud Functions. The initiatives will all be governed by the same set of constraints that are required to stay in compliance with regulations. How would you recommend the data engineers organize their Google Cloud resources?

- **Use a project for each initiative and place those projects in a folder. Attach policies to the folder to enforce constraints. (Correct)**
- **Use a folder for each initiative and place those folders in a project. Attach policies to the project to enforce constraints.**
- **Use a project for each initiative and place those projects in an organization. Attach policies to the organization to enforce constraints. (Incorrect)**
- **Use an organization for each initiative and place those folders in a project. Attach policies to the organization to enforce constraints.**

Explanation
Each initiative should have it's own project to isolate and manage it's resources. All projects should be in the same folder and policies should be attached to that folder so all projects in the folder will inherit them. Folders cannot be in a project and organizations cannot be in folders. See https://cloud.google.com/resource-manager/docs/cloud-platform-resource-hierarchy

Question 45: **Correct**
As the developer of a new Cloud Dataflow pipeline, you'd like to limit the processing resources used when testing a new pipeline. What parameter would you specify when executing your new Cloud Dataflow job?

--maxNumWorkers
(Correct)

- **--jobWorkers**

- **--maxVMs**

- **--maxContainers**

Explanation
The correct answer is --maxNumWorkers, the other options are not valid parameters when executing a workflow in Cloud Dataflow. See https://cloud.google.com/dataflow/docs/guides/deploying-a-pipeline

Question 46: **Correct**
A project sponsor wants to develop a machine learning model to classify potentially fraudulent transactions. They want to rank models based on a combination of precision and recall. What evaluation metric would you recommend?

- **F-score**
(Correct)

- **Root mean squared error**

- **Feature crosses**

- **Mean squared error**

Explanation
F-score is the harmonic mean of precision and recall and is often used to measure the overall performance of classification models. Root mean squared error and mean squared error are used to evaluate regression models. Feature crosses is a way to generate synthetic features, not measure the performance of machine learning models. See https://cloud.google.com/automl-tables/docs/evaluate

Question 47: **Incorrect**
A team of data scientists is becoming increasingly dependent on jobs running in a Cloud Dataproc cluster. They would like to increase the number of master nodes from 1 to 2 to improve availability. What command would they use?

- **gcloud dataproc master-nodes**
  (Incorrect)

- **gcloud dataproc nodes add-master**

- **gcloud dataproc create master-nodes**

- **The number of master nodes cannot be changed once a cluster is created.**
  (Correct)

Explanation
The correct answer is that the number of master nodes in a Cloud Dataproc cluster cannot be changed once the cluster is created. See https://cloud.google.com/dataproc/docs/guides/manage-cluster

Question 48: **Incorrect**
Messages are unexpectedly accumulating in service using Cloud Pub/Sub. A developer unfamiliar with Cloud Pub/Sub has asked for our help in diagnosing the problem. What would you point out with respect to how messages are removed from Cloud Pub/Sub topics?

- **Once at least one subscriber for each topic has acknowledged the message it will be deleted from storage.**
  (Incorrect)

- **Once at least one subscriber for each bucket has acknowledged the message it will be deleted from storage.**

- **Once at least one subscriber for any subscription has acknowledged the message it will be deleted from storage.**

- **Once at least one subscriber for each subscription has acknowledged the message it will be deleted from storage.**
  (Correct)

Explanation
The correct answer is that a message is deleted once at least one subscriber for a each subscription has acknowledged the message it will be deleted from storage. See https://cloud.google.com/pubsub/docs/subscriber

Question 49: **Correct**
Your BigQuery costs are higher than expected. You want to help data analysts using the BigQuery data warehouse reduce overall costs. Which of the following would you recommend? (choose 2)

- **Avoid using SELECT ***

(Correct)

- **Avoid using partitioned tables**
- **Avoid using clustered tables**
- **Use LIMIT only with clustered tables**
  (Correct)
- **Use the bq --estimate-bytes command to estimate the number of bytes read**

Explanation
SELECT * can scan large amounts of data and should be avoided. Using LIMIT only with clustered tables can reduce the amount of data scanned. Using LIMIT on non-clustered tables does not limit the number of bytes scanned. Partitioning and clustering can both help limit the amount of data scanned and therefore help reduce costs. The bq command to estimate the number of bytes scanned is bq --dry_run, not --estimate_bytes. See https://cloud.google.com/bigquery/docs/best-practices-performance-overview

Question 50: **Correct**
You are currently using Apache Kafka to ingest messages from IoT sensors. A data pipeline based on Apache Flink reads the data from Kafka and processes the data before writing results to long-term storage. If you wanted to migrate to Google Cloud and use managed services instead of Apache Kafka and Apache Flink, what services would you use? (Choose 2)

- **Cloud Pub/Sub**
  (Correct)

- **Cloud Dataflow (Correct)**
- **Cloud Firestore**
- **Cloud Data Fusion**
- **Cloud Composer**

Explanation
The correct answers are Cloud Pub/Sub as a replacement for Apache Kafka and Cloud Dataflow as a replacement for Apache Flink. Cloud Pub/Sub is a messaging service. Cloud Dataflow, like Apache Flink, implements an Apache Beam runner. Cloud Data Fusion is an ETL tool. Cloud Composer is a workflow orchestration tool based on Apache Airflow. Cloud Firestore is a NoSQL document database. See https://cloud.google.com/pubsub, https://cloud.google.com/dataflow, and https://cloud.google.com/blog/products/data-analytics/simplify-and-automate-data-processing-with-dataflow-prime

Question 1: **Correct**
You are consulting to a company developing an IoT application that analyzes data from sensors deployed on drones. The application depends on a database that can write large volumes of data at low latency. The company has used Hadoop HBase in the past but wants to migrate to a managed database service. What service would you recommend?

- **Cloud Firestore**

- **Cloud Spanner**
- **Bigtable (Correct)**
- **BigQuery**

Explanation
Bigtable is a wide column database with low latency writes that is well suited for IoT data storage and it has an HBase API. BigQuery is a data warehouse service. Cloud Dataproc is a managed Spark/Hadoop service. Cloud Firestore is a NoSQL document model database. See https://cloud.google.com/bigtable/docs/hbase-bigtable

Question 2: **Correct**
You have created a function that should run whenever a message is written to a Cloud Pub/Sub topic. What command would you use to deploy that function?

- **gcloud pubsub topics pull**
- **gcloud pubsub subscription publish**
- **gcloud pubsub topics publish**
-

gcloud functions deploy
(Correct)

Explanation
The correct command is gcloud functions deploy. Gcloud pubsub topics publish publishes a message to a topic. The others are not valid gcloud pubsub commands. See https://cloud.google.com/sdk/gcloud/reference/functions/deploy

Question 3: **Incorrect**
A university research group has started a company to commercialize a laboratory management system. Their application uses a MongoDB database but the group would like to migrate to a managed database service in Google Cloud. What service would you recommend they use?

- **Cloud SQL**
  (Incorrect)
- **Cloud Firestore**
  (Correct)
- **Cloud Bigtable**
- **BigQuery**

Explanation
MongoDB is a document database so Cloud Firestore is the best option since it is also a document database. Cloud Bigtable is a wide-column NoSQL database and not a good replacement for MongoDB. BigQuery is an analytical database designed for data warehousing and data analysis. Cloud SQL is a relational database and not a good replacement for a NoSQL database. See https://firebase.google.com/docs/firestore/data-model

Question 4: **Correct**

As an analyst with a major metropolitan public transportation agency, you are tasked with monitoring data about passengers on all modes of transport provided by the agency. Since you know SQL, you would like to run a SQL query using Cloud Dataflow. What command allows you to run a SQL query and write results to a BigQuery table? (Assume all need parameters will be specified).

- **bq dataflow sql query**
- **gcloud bigquery sql query**
- **gcloud dataflow sql query (Correct)**
- **bq bigquery sql query**

Explanation
The correct answer is gcloud dataflow sql query. Bq is the command line tool for working with BigQuery but this calls for executing a Cloud Dataflow command, which requires a gcloud command. Gcloud bigquery is not a valid GCP command. See https://cloud.google.com/sdk/gcloud/reference/dataflow/sql/query

Question 5: **Incorrect**
A insurance claim review company provides expert opinion on contested insurance claims. The company uses Google Cloud for it's data analysis pipelines. Clients of the company upload documents to Cloud Storage. When a file is uploaded, the company wants to immediately move the files to a Classified Data bucket if the file contains personally identifying information. What method would you recommend to accomplish this?

- **Create a quarantine bucket for uploading, use Cloud Scheduler to run a job to run hourly that will call the Data Loss Prevention API to apply infotypes to detect PII. If PII is detected, move file to the**

Classified Data bucket.

- **Create a quarantine bucket for uploading, once a file is uploaded trigger a Cloud Function to call a custom built machine learning model trained to detect PII. If PII is detected, move the file to the Classified Data bucket.**

  (Incorrect)

- **Create a quarantine bucket for uploading, use Cloud Scheduler to run a job to run hourly that will call a custom built machine learning model trained to detect PII. If PII is detected, move file to the Classified Data bucket.**

- **Create a quarantine bucket for uploading, once a file is uploaded trigger a Cloud Function to call the Data Loss Prevention API to apply infotypes to detect PII. If PII is detected, move file to the Classified Data bucket.**
  (Correct)

Explanation
The correct solution is to use a quarantine bucket that triggers a Cloud Function on upload to invoke the DLP API and move the file if PII is found. Cloud Scheduler runs jobs at regular intervals but this calls for immediate processing of a file once uploaded so Cloud Functions should be used. You could train a custom machine learning model but that requires development time and maintenance. A managed service like DLP is a better option. See https://cloud.google.com/dlp/docs/reference/rest

Question 6: **Correct**
Auditors have informed the CIO of your company that all logs from applications running in Google Cloud will need to be retained for 60 days. You would also like to access logs from 3rd party tools up to 60 days old. What solution would you recommend to meet this requirement?

-

Use Cloud Logging and set up a Log Router to create a Bigtable sink to keep the logs 60 days. Create a data lifecycle policy to delete logs after 60 days.

- **Use Cloud Logging and set up a Log Router to create a Cloud Storage sink to keep the logs 60 days. Create a data lifecycle policy to delete logs after 60 days.**

  (Correct)

- **Use Cloud Logging and set up a Pub/Sub topic to receive log data and write that data to a Cloud Storage bucket to keep the logs 60 days. Create a data lifecycle policy to delete logs after 60 days.**

- **Use Cloud Logging and keep log data in the Cloud Firestore service for 60 days. Create a logging policy to delete the data after 60 days.**

Explanation
Cloud Logging keeps logs by default up to 30 days and could be stored for a custom time period. By setting up a log sink for Cloud Storage, you can route logs to Cloud Storage where it can be kept for 60 days and accessed by 3rd party tools during that time. If log data were written to Cloud Pub/Sub another service would have to read that data and write it to a long term storage system, such as Cloud Storage. Bigtable is not a Cloud Logging sink option. See https://cloud.google.com/logging/docs/routing/overview and https://cloud.google.com/logging/docs/buckets.

Question 7: **Correct**
A data engineer needs to load data stored in Avro files in Cloud Storage into Bigtable. They would like to have a reliable, easily monitored process for copying the data. What would you recommend they use to copy the

data?

- **Storage Transfer Service**
- **Cloud Dataflow, starting with a Cloud Storage Avro to Bigtable template.**
  (Correct)
- **gsutil**
- **Custom Python 3 program**

Explanation
The correct answer is to use Cloud Dataflow with a Cloud Storage Avro to Bigtable template. Using Python 3 would require more work than necessary. Gsutil is used to load data into Cloud Storage, not Bigtable. Storage Transfer Service is for copying data into Cloud Storage from other object storage system, such as AWS S3.
A custom Python 3 program would require more development effort than using Cloud Dataflow. See https://cloud.google.com/architecture/streaming-avro-records-into-bigquery-using-dataflow

Question 8: **Correct**
What data structure in the Cloud Firestore document data model is analogous to a row in a relational database?

- **Kinds**
- **Index**

- **Interleaved row**
- **Entity**
  (Correct)

Explanation
Entities are analogous to rows in relational data models, both of which describe a single modeled element. Kinds are collections of related entities and analogous to a table in relational data models. An index is used to implement efficient querying in both Cloud Firestore and relational databases. There is no such thing as an interleaved row; interleaved tables are a feature of Cloud Spanner which improves query performance by storing related data together. See https://firebase.google.com/docs/firestore/data-model

Question 9: **Incorrect**
Analysts are using Cloud Data Studio for analyzing data sets. They would like to improve the performance of the time required to update tables and charts when working with the data. What would you recommend they try to improve performance?

- **Use a live data source**
- **Use an extracted data source**

  (Correct)
- **Use a blended data source**
  (Incorrect)

Use an imported data source

Explanation
Extracted data sources are snapshots and can provide better performance than live data sources. Blended data sources are used to combine data from multiple data sources. There is no imported data source. See https://cloud.google.com/bigquery/external-data-sources

Question 10: **Correct**
A global transportation company is using Cloud Spanner for managing shipping orders. They have migrated an Oracle database to Cloud Spanner with minimal changes and are experiencing similar performance problems with joins. In particular, a one-to-many join between an orders table and an order items table is not performing as needed. What would you recommend?

- Use interleaved tables
  (Correct)
- Use Cloud Bigtable for better join performance
- Use interleaved hashes
- Use Cloud SQL for better join performance

Explanation
Interleaved tables store parent and children records together, such as orders and order items. This is more efficient than storing related items separately since the parent and child data can be read at the same time. Cloud Bigtable is a NoSQL database and would not meet requirements. Cloud SQL does not scale beyond regional-scale databases and would not meet requirements. There is no such thing as interleaved hashes in Cloud Spanner. See https://cloud.google.com/spanner/docs/schema-and-data-model and https://

cloud.google.com/spanner/docs/whitepapers/optimizing-schema-design

Question 11: **Correct**
A team of analysts is building machine learning models. They want to use managed services when possible but they would also like the ability to customize and tune their models. In particular, they want to be able to tune hyperparameters themselves. What managed AI service would you recommend they use?

- **Cloud TPUs**
- **Vertex AI AutoML training**
- **BigQuery ML**
- **Vertex AI custom training (Correct)**

Explanation
Vertex AI custom training allows for tuning hyperparameters. Vertex AI AutoML training tunes hyperparameters for you. BigQuery ML does not allow for hyperparameter tuning. Cloud TPUs are accelerators you can use to train large deep learning models. See https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform

Question 12: **Correct**
A consultant has recommended that you replace an existing messaging system with Cloud Pub/Sub. You are concerned that your existing system has a different delivery guarantee than Cloud Pub/Sub. What kind of message deliver semantics does Cloud Pub/Sub guarantee?

- **Deliver at least once (Correct)**

- **Deliver at most or deliver at least depending on configuration of the subscription**
- **Best effort but no guarantee**
- **Deliver at most once**

Explanation
Cloud Pub/Sub has a deliver at least once guarantee. It does not have deliver at most once guarantee. It is possible the Cloud Pub/Sub could deliver a message more than once. See https://cloud.google.com/pubsub/docs/subscriber

Question 13: **Correct**
A team of data scientists has been using an on-premises cluster running Hadoop and HBase. They want to migrate to a managed service in Google Cloud. They also want to minimize changes to programs that make extensive use of the HBase API. What GCP service would you recommend?

- **Cloud Spanner**
- **Bigtable**
  (Correct)
- **BigQuery**
-

Cloud Dataflow

Explanation
The correct answer is Bigtable, which is a data store providing an HBASE compatible API. BigQuery is a data warehouse service that supports SQL but does not have an HBASE compatible API. Cloud Spanner is a relational database and not a replacement for Hadoop and HBASE. Cloud Dataflow is a data pipeline service that includes an Apache Beam runner. See https://cloud.google.com/bigtable/docs/hbase-bigtable

Question 14: **Correct**
You use materialized views in BigQuery. You are incurring higher than expected charges for BigQuery and suspect it may be related to materialized views. What materialized view characteristic could increase your BigQuery costs? (Choose 2)

- **The datatypes used in materialized views**
- **The frequency of materialized view refresh (Correct)**
- **The total volume of data stored in materialized views (Correct)**
- **The number of users with read access to the materialized view**

Explanation
The amount of data stored and the frequency of refresh jobs can increase the cost of maintaining materialized views. The data types used in the materialized view do not affect the cost. The number of users reading a materialized view does not affect cost but the total amount of data scanned would impact cost. See https://cloud.google.com/bigquery/docs/materialized-views-intro

Question 15: **Incorrect**

A manufacturer has successfully migrated several data warehouses to BigQuery and is using Cloud Storage for machine learning data. ML engineers and data analysts are having difficulty finding data sets they need. The CTO of the company has asked for your advice on how to reduce the workload on ML engineers and analysts when they need to find data sets. What would you recommend?

- **Use Cloud Logging to track files uploaded to Cloud Storage and data sets to BigQuery.**
- **Use Cloud Fusion to tracking both files uploaded to Cloud Storage and data sets loaded into BigQuery.**
- **Use Cloud Data Catalog to automatically extract metadata from Cloud Storage objects and BigQuery data.**
  (Correct)
- **Query the metadata catalog of BigQuery and Cloud Storage and write the results to a BigQuery table where the ML engineers and data analysts can query the data with SQL.**

  (Incorrect)

Explanation

The correct answer is to use Cloud Data Catalog, which can automatically extract metadata from sources including Cloud Storage, BigQuery, Cloud Bigtable, Cloud Pub/Sub, and Google Sheets. Cloud Logging is used for recording data about events and is not the best way to collect metadata. Cloud Fusion is an ETL tool, not a metadata extraction tool. Developing your own metadata extraction tool, such as one that queries BigQuery metadata, requires more work and maintenance than using a managed service. See https://cloud.google.com/data-catalog/docs/concepts/overview

Question 16: **Correct**

You would like to set a maximum number of concurrent jobs in Cloud Dataproc. How would you do that?

- **Set dataproc:dataproc.scheduler.max-concurrent-jobs property when creating a cluster.**

  (Correct)

- **Set dataproc:dataproc.scheduler.max-concurrent-jobs property when adding worker nodes.**

- **Set computengine:mgi.scheduler.max-concurrent-jobs property when creating a managed instance group for Cloud Dataproc cluster.**

- **Use Cloud Monitoring to detect the number of jobs running and when the maximum threshold is exceeded, trigger a Cloud Function to terminate the most recently created job.**

Explanation

The correct answer is to set the dataproc:dataproc.scheduler.max-concurrent-jobs property when creating a cluster. That property is not set when adding worker nodes. Properties of a Dataproc cluster that are specific to Dataproc are not set in Compute Engine. You do not need to monitor and terminate jobs using ad hoc procedures like triggering a Cloud Function after a maximum threshold is exceeded. See https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster-properties

Question 17: **Correct**

The CTO of your company is concerned about the costs of running data pipelines, especially some large batch processing jobs. The jobs do not have to be run on a fixed schedule and the CTO is willing to wait longer for jobs to

complete if it can reduce costs. You are using Cloud Dataflow for most pipelines and would like to cut costs but not make any more changes than necessary. What would you recommend?

- **Use Dataflow FlexRS**

  (Correct)

- **Use Dataflow Streaming Engine**

- **Use a different Apache Beam Runner**

- **Use Dataflow Shuffle**

Explanation
The correct answer is to use Cloud Dataflow flexible resource scheduling (FlexRS) which reduces batch processing costs using scheduling techniques and preemptible VMs along with regular VMs. Streaming Engine is an optimization for stream, not batch, processing. Dataflow Shuffle provides for faster execution of batch jobs but does not necessarily reduce costs. Using a different Apache Beam runner would require more management overhead, for example, by running Apache Flink in Compute Engine. See https://cloud.google.com/dataflow/docs/guides/flexrs

Question 18: **Correct**
You are in the process of creating lifecycle policies to manage objects stored in Cloud storage. Which of the following are lifecycle conditions you can use in your policies? (Choose 3)

- **File size**

Matches Storage Class
(Correct)

- **Is Live**
  (Correct)
- **File type**
- **Age**
  (Correct)

Explanation
The correct answers are age, matches storage class, and is live. File type and file size are not conditions available in lifecycle management policies. See https://cloud.google.com/storage/docs/lifecycle

Question 19: **Correct**
A multi-national financial services company is creating a new service to facilitate cross-currency transactions. The database must provide strong consistency for transactions that may be initiated by any customer. Customers are initially located in Europe but the company plans to expand to Asia, Africa, North America and South America within a year. The database must support normalized data models. What Google Cloud managed database service would you use?

- **Cloud Bigtable**
- **BigQuery**

- **Cloud Spanner**
  (Correct)

- **Cloud SQL**

Explanation
Cloud Spanner is the correct choice, it provides global scale relational database services, including strong consistency. Cloud SQL is appropriate for regional-scale databases. BigQuery is an analytical database designed for data warehousing and analytics. Cloud Bigtable is a NoSQL database and does not meet the specified requirements. https://cloud.google.com/blog/topics/developers-practitioners/what-cloud-spanner

Question 20: **Incorrect**
Your team is deploying a new data pipeline. Developers who will maintain the pipeline will need permissions granted by three different roles. Those roles also have permissions that are not needed by the maintainers. Following Google Cloud recommended practices, what would you recommend?

- **Create a custom role with only the permissions needed. This follows the principal of least privilege.**
  (Correct)

- **Assign the Owner role instead of the three roles to minimize role management overhead.**

- **Create a custom group with all the permissions in the three different roles. This follows the principle of maximum privilege.**

(Incorrect)

- **Assign the three existing roles to the maintainers in order to minimize role management overhead.**

Explanation
Creating a custom role with only the permissions needed is the correct answer. This follows the principle of least privilege. Permissions are assignee to roles not groups. The Owner role is a primitive role that grants excessive privileges and should only be used in limited cases when security risks are minimal. Assigning the three existing roles would grant more permissions than needed and would violate the principle of least privilege. There is no principle of maximum privilege. https://cloud.google.com/blog/products/identity-security/dont-get-pwned-practicing-the-principle-of-least-privilege

Question 21: **Incorrect**
A manufacturer of delivery drones has a monitoring system built on an Apache Beam runner. Temperature received over the past hour is analyzed and if any temperature reading is more than 2 standard deviations away from the mean for the past hour, an alert is triggered. What kind of windowing functions would you use to implement this operation?

- **concurrent windows**

- **fixed windows (also called tumbling windows)**
(Incorrect)

- **sliding window (also called hopping windows)**
(Correct)

- **session windows**

Explanation
Sliding windows (also called hopping windows) model a consistent time interval in a stream so it is the best option for continuously averaging the temperature for the past hour. Fixed windows (also known as tumbling windows) model a consistent, disjoint time interval in the stream. Session windows can contain a gap in duration and are used to model non-continuous streams of data. There is no concurrent window type of functions in Apache Beam runners such as Cloud Dataflow. See https://cloud.google.com/dataflow/docs/concepts/streaming-pipelines

Question 22: **Correct**
You are migrating a data warehouse from on-premises to Google Cloud. Users of the data warehouse are concerned that they will not have access to highly performant, in memory analysis. What service would you suggest to have comparable features and performance in Google Cloud?

- **BigQuery BI Engine (Correct)**
- **Bigtable with BI Engine**
- **BigQuery Cloud Memorystore with memcached**
- **BigQuery with Cloud Memorystore using Redis**

Explanation
BigQuery BI Engine is an in-memory analytics engine. Cloud Memorystore is a cache and better suited to storing key-value data for applications that need low latency access to data. There is no Bigtable BI Engine service. See https://cloud.google.com/bigquery/docs/bi-engine-intro

Question 23: **Incorrect**
A startup is providing a streaming service for cricket fans around the world. The service will provide both live streams and videos of previously played matches.

The architect of the startup wants to ensure all users have the same experience regardless of where they are located. What GCP service could the startup use to help ensure a consistent experience for previously played matches?

- **Cloud CDN**

  (Correct)

- **Cloud Storage using multiple regions**

- **Cloud Firestore**

  (Incorrect)

- **Cloud Storage using Nearline storage**

Explanation
Cloud CDN is a content delivery network service designed to store copies of data close to end users. Cloud Storage using multiple regions would require more management than Cloud CDN and does not have the automatic caching features of Cloud CDN. Cloud Storage Nearline is for storing objects that are accessed less than once in 30 days. Cloud Firestore is a NoSQL database and not appropriate for storing and streaming videos. See https://cloud.google.com/cdn

Question 24: **Incorrect**
A team of analysts working with healthcare data have analyzed data in a BigQuery dataset for personally identifiable information. They want to store the results of the analysis in a managed service that will make it easy for them to retrieving information about the PII analysis at later times. What service would you recommend?

Data Loss Prevention

- **Cloud Spanner**

- **BigQuery**
  (Incorrect)

- **Data Catalog**
  (Correct)

Explanation
The correct answer is Data Catalog, a metadata management service designed for data discovery and metadata management. BigQuery Cloud Spanner could be used by Data Catalog is specifically designed to support metadata management and the types of queries that are typically used for metadata management. Data Loss Prevention is a service to identify types of information and estimate re-identification risk, it is not a service to persistently store data. See https://cloud.google.com/data-catalog/docs/concepts/overview

Question 25: **Incorrect**
You are developing a distributed system and want to decouple two services. You want to ensure messages use a standard format and you plan to use Cloud Pub/Sub. What schema types are supported by Cloud Pub/Sub? (Choose 2)

- **CSV**

- **Protocol Buffer**
  (Correct)

- **Parquet**
  (Incorrect)
- **Avro**
  (Correct)
- **Thrift**

Explanation
Cloud Pub/Sub supports Avro and Protocol Buffer schemas. Thrift is an alternative to Protocol Buffers but is not supported for schemas. Parquet is an open source file format used in Hadoop. CSV is a file format often used when sharing data between applications.
See https://cloud.google.com/pubsub/docs/schemas

Question 26: **Incorrect**
What types of indexes are automatically created in Cloud Firestore? (Choose 2).

- **Hash indexes**
- **Composite indexes, single value**
- **Composite indexes, multi-value**
  (Incorrect)

- **Atomic values, ascending (Correct)**

- **Atomic values, descending (Correct)**

Explanation
Cloud Firestore automatically creates atomic value ascending and descending indexes. A composite index is made up of two or more values and are not created manually. There is no single valued composite index; all composite indexes have multiple values. There isn't a hash index type in Cloud Firestore. sEe https://firebase.google.com/docs/firestore/query-data/index-overview

Question 27: **Correct**
A developer is deploying a Cloud SQL database to production and wants to follow Google Cloud recommended best practices. What should the developer use for authentication?

- **Cloud SQL Auth proxy (Correct)**

- **Strong encryption**

- **IAM**

- **Cloud Identity**

Explanation
Cloud SQL Auth proxy is the recommended way to connect to Cloud SQL.

Cloud Identity is an Identity as a Service provided by Google Cloud. IAM is Identity and Access Management service for managing identities and their authorizations. Strong encryption is used to protect the confidentiality and integrity of data, not to perform authentication. See https://cloud.google.com/sql/docs/mysql/sql-proxy

Question 28: **Correct**
To comply with industry regulations, you will need to capture logs of all changes made to IAM roles and identities. Logs must be kept for 3 years. How would you meet this requirement?

- **Use Cloud Audit Logs and keep the logs in Cloud Monitoring. Specify a three year retention policy in Cloud Logging that automatically deletes the logs after three years.**

- **Use Cloud Audit Logs and keep the logs in Cloud Logging. Specify a three year retention policy in Cloud Logging that automatically deletes the logs after three years.**

- **Use Cloud Audit Logs and export them to Bigtable. Create a retention policy and retention policy lock to prevent the logs from being deleted prior to them reaching 3 years of age. Define a lifecycle policy to delete the logs after three years.**

- **Use Cloud Audit Logs and export them to Cloud Storage. Create a retention policy and retention policy lock to prevent the logs from being deleted prior to them reaching 3 years of age. Define a lifecycle policy to delete the logs after three years.**
  (Correct)

Explanation
Cloud Audit log captures changes to IAM entities and keeps logs for 30 days.

To keep them longer, export them to Cloud Storage. Use a retention policy to define how long the logs should be kept and using a retention policy lock to prevent changes to the retention period. Cloud Logging does not keep logs beyond 30 days and does not support retention policies. Cloud Monitoring collects and displays metrics, it does not store logs. Bigtable is not a good storage option for logs, it is designed for low latency writes at high volumes and provides for key lookups and queries that require range scanning. See https://cloud.google.com/logging/docs/audit

Question 29: **Incorrect**
In order to comply with industry regulations, you will need to use customer managed keys when analyzing data using Cloud Dataproc. You will be managing Cloud Dataproc clusters using command line tools. What command would you use with the --gce-pd-kms-key parameter to specify a Cloud KMS resource ID to use with the cluster?

- **gcloud dataproc clusters kms**
- **gcloud clusters dataproc create (Incorrect)**
- **gcloud dataproc clusters create (Correct)**
- **gcloud clusters dataproc kms**

Explanation
The correct answer is gcloud dataproc clusters create. The other options are not valid gcloud commands. See https://cloud.google.com/sdk/gcloud/reference/dataproc/clusters/create

Question 30: **Incorrect**
An industry regulation requires that when analyzing personal identifying information (PII), you must not run analysis on physical servers that are shared

with other cloud customers. You plan to use Cloud Dataproc for analyzing data with PII. What will you need to do when creating a Cloud Dataproc Cluster to ensure you are in compliance with this regulation?

- **Create an unmanaged instance group and specify that instance group when creating the cluster.**
  (Incorrect)

- **Create a sole-tenant node group and specify that node group when creating the cluster.**
  (Correct)

- **Disable autoscaling to prevent the addition of non-sole tenant VMs.**

- **You cannot configure Cloud Dataproc to use sole tenant nodes. You will need to run Spark in a Compute Engine managed instance group that you manage yourself.**

Explanation
The correct answer is to create a sole-tenant node group and specify that node group when creating the cluster. Cloud Dataproc does support sole tenants so you don't need to run a self-managed Spark cluster. You can use autoscaling with sole tenant node groups. Unmanaged instance groups are not required and not recommend except for legacy, heterogeneous clusters migrating to Compute Engine. https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/sole-tenant-nodes

Question 31: **Correct**
A data warehouse team is concerned that some data sources may have poor quality controls. They do not want to bring incorrect or invalid data into the data warehouse. What could they do to understand the scope of the problem before starting to write ETL code?

-

Load all source data into a data lake and then load it to the data warehouse.

- **Perform a data quality assessment on the source data after it is extracted from the source system. These should include checks for ranges of values in each attribute, distribution of values in each attribute, counts of the number of invalid and missing values, and other checks on source data.**

  (Correct)

- **Load the data into the data warehouse and log any records that fail integrity or consistency checks.**

- **Have administrators of the source systems produce a data quality verification before exporting the data.**

Explanation
The correct answer is performing a data quality assessment on data extracted from the source system. Loading data from a data lake to a data warehouse will not provide an assessment of the range of the problem. Loading data into the data warehouse and logging failed checks is less efficient because it will provide log messages but not aggregate statistics on the full scope of the problem. The source systems may not have the ability to perform data quality assessments and if they do, you may get different kinds of reports from different systems. By performing a data quality assessment on extracted data you can produce a consistent set of reports for all data sources. See https://cloud.google.com/blog/products/data-analytics/principles-and-best-practices-for-data-governance-in-the-cloud

Question 32: **Correct**
As part of the ingestion process, you want to ensure any messages written to a

Cloud Pub/Sub topic all have a standard structure. What is the recommended way to ensure messages have the standard structure?

- **Create a schema and assign it to a subscription during subscription creation.**

- **Use a data quality function in Cloud Function to check the structure as it is written to Cloud Pub/Sub.**

- **Create a schema and assign it to a topic during topic creation.**

  (Correct)

- **Use a data quality function in Cloud Function to reformat the message if needed before it is read from a subscription.**

Explanation
Schemas are used to define a standard message structure and they are assigned to topics during creation. Schemas are not assigned to subscription. Cloud Functions should not be used to implement a feature that is available in Cloud Pub/Sub. Cloud Functions support only one type of Pub/Sub event, google.pubsub.topic.publish. See https://cloud.google.com/pubsub/docs/schemas

Question 33: **Incorrect**
A financial services company wants to use BigQuery for data warehousing and analytics. The company is required to ensure encryption keys are stored and managed in a key management system that's deployed outside of a public cloud. They want to minimize the management overhead of key management while remaining in compliance. What would you recommend they do?

-

Use Data Catalog for external data management, specifically keys

-

Use Dataproc for external data management, specifically keys

-

Use external data sources with BigQuery and encrypt the external data sources outside of Google Cloud
(Incorrect)

-

Use Cloud EKM for external key management
(Correct)

Explanation
The correct answer is to use External Key Management, it allows the company to maintain separation between data in BigQuery and their encryption keys. Data Catalog is a metadata and data discovery service, not a key management service. BigQuery external data sources allow for accessing data not stored in BigQuery and do not address the requirements. Cloud Dataproc is a managed Spark and Hadoop service, not a key management service. See https://cloud.google.com/kms/docs/ekm

Question 34: **Correct**
You have developed a DoFn function for a Cloud Dataflow workflow. You discover that the PCollection does not have all the data needed to perform a necessary computation. You want to provide additional input each time an element of a PCollection is processed. What kind of Apache Beam construct would you use?

-

Partition

-

Watermark

- **Side input (Correct)**
- **Custom window**

Explanation
The correct answer is a side input, which is an additional input for DoFn. A partition in Apache Beam separates elements of a collection into multiple output collections. A Watermark is used to indicate no data with timestamps earlier than the watermark will arrive in the future. A custom window is created using WindowFn functions to implement windows based on data-driven gaps. See https://cloud.google.com/architecture/e-commerce/patterns/slow-updating-side-inputs

Question 35: **Correct**
As a consultant to a multi-national company, you are tasked with helping design a service to support an inventory management system that is strongly consistent, supports SQL, and can scale to support hundreds of users in North America, Asia, and Europe. What Google Cloud service would you recommend for this service?

- **BigQuery**
- **Cloud Firestore**
- **Cloud SQL**
-

Cloud Spanner (Correct)

Explanation
Cloud Spanner is a global, horizontally scalable relational database with strong consistency and is the best option. Cloud SQL is not scalable beyond a single region. Cloud Firestore does not support SQL. BigQuery is an analytical database for data warehousing not an OLTP system such as an inventory management system. See https://cloud.google.com/blog/products/gcp/introducing-cloud-spanner-a-global-database-service-for-mission-critical-applications

Question 36: **Correct**
Epidemiology and infectious disease researchers are collecting data on the genomic sequences of several pathogens. The data is stored in a bioinformatics-specific format called FASTQ and are tens of gigabytes in size. They will eventually store several terabytes of FASTQ data. The data will be processed by Cloud Dataflow and results will be written to BigQuery. What is a good option for storing FASTQ data?

- **Bigtable**
- **Cloud Firestore**
- **Cloud Storage (Correct)**
- **BigQuery**

Explanation
The specialized data format in this scenario makes object storage a good option so Cloud Storage is the best choice. Cloud Firestore is a good option for document storage, such as JSON structures. BigQuery and Bigtable are not suited to store large objects. See https://cloud.google.com/blog/topics/

developers-practitioners/map-storage-options-google-cloud

Question 37: **Incorrect**
A team of socio-economic researchers is analyzing documents as part of a research study. The documents have had personally identifying information redacted. The researchers are concerned that someone with access to the data may be able to use quasi-identifiers, such as age and postal code, to re-identify some individuals. How can the researchers quantify that risk?

- **Apply the re-identification infotype to each document with quasi-identifiers to calculate the level of risk.**
- **Run a custom machine learning model trained to estimate the re-identification risk.**
  (Incorrect)
- **Use counts of the number of occurrences of quasi-identifiers identified using Data Loss Prevention infotypes.**
- **Run a re-identification risk analysis using the Data Loss Prevention service.**
  (Correct)

Explanation
A re-identification risk analysis job using DLP will provide the information needed by the researchers. Using a custom trained machine learning program to estimate risk would take longer, require maintenance, and assumes the researchers are also proficient in machine learning. DLP uses infotypes but there is no re-identification risk infotype. Counting specific infotypes may provide some indication of re-identification risk but it is unlikely that a simple linear model of risk will give accurate or useful information. See https://cloud.google.com/blog/products/identity-security/taking-charge-of-your-data-understanding-re-identification-risk-and-quasi-identifiers-with-cloud-dlp

Question 38: **Incorrect**

You are developing a deep learning model and have training data with a large number of features. You are not sure which features are important. You'd like to use a regularization technique that will drive the parameter for the least important features toward zero. What regularization technique would you use?

- **Backpropagation**
- **L1 or Lasso Regression (Correct)**
- **L2 or Ridge Regression**
- **Dropout (Incorrect)**

Explanation

L1 or Lassos Regression adds an absolute value of magnitude penalty which drives the parameters (or coefficients) of least useful features toward zero. L2 or Ridge Regression adds a squared magnitude penalty that penalizes large parameters. Dropout is another form of regularization that ignores some features at some steps of the training process. Backpropagation is an algorithm for assigning error penalties to nodes in a neural network. See https://cloud.google.com/bigquery-ml/docs/preventing-overfitting and https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Question 39: **Correct**

A new workload has been deployed to Cloud Dataproc, which is configured with an autoscaling policy. You are noticing a FetchFailedException is occurring intermittently. What would be the most likely cause of this problem?

The autoscaling policy is scaling down and shuffle data is lost when a node is decommissioned.

(Correct)

- **You are using a GCS bucket with improper access controls.**

- **You are using Google Cloud Storage instead of local storage for persistent storage.**

- **The autoscaling policy is adding nodes too fast and data is being dropped.**

Explanation
The FetchFailedException can occur when shuffle data is lost when a node is decommissioned. The autopolicy should be configured based on the longest running job in the cluster. Adding nodes will not cause a loss of data. Cloud Storage is the preferred persistent storage method for Dataproc clusters. While FetchFailedException can be caused by network issues, that is not likely to be a problem when using Cloud Storage for a Cloud Dataproc cluster. If the storage bucket had improper access controls then errors would occur consistently, not intermittently. See https://cloud.google.com/blog/topics/developers-practitioners/dataproc-best-practices-guide

Question 40: **Correct**
A Cloud Dataproc cluster is experiencing a higher than normal workload and you'd like to add several preemptible VMs as worker nodes. What command would you use?

-

The number of preemptible nodes in a Cloud Dataproc cluster cannot be changed once the cluster is created.

- **gcloud dataproc clusters update with the --preemptible-vms parameter**

- **gcloud dataproc clusters update with the --num-secondary-workers parameter**
  (Correct)

- **gcloud dataproc clusters update with the --num-workers parameter**

Explanation
The number of preemptible nodes can be updated using gcloud dataproc clusters update with the --num-secondary-workers parameter. The --num-workers parameter is used to change the number of primary (non-preemptible) workers. There is no --preemptible-vms parameter in the gcloud dataproc command. The number of preemptible (secondary) workers can be changed after creating a cluster. See https://cloud.google.com/sdk/gcloud/reference/dataproc/clusters/update

Question 41: **Incorrect**
An IoT service uses Bigtable to store timeseries data. You have noticed that write operations tend to happen on one node at a time rather than being evenly distributed across nodes. What could be the cause of this problem?

- **Using the wrong type of GCP load balancer in front of Bigtable**

  (Incorrect)

-

- **Using a row key that causes data that arrives close in time to be written to a single node, rather than evenly distributed.**

  (Correct)

- **Misconfiguring replication**

- **Using too many columns in your data model**

Explanation
This is an example of hot spotting, where workload is skewed toward a small number of nodes instead of evenly distributed. In Bigtable, this can be caused by row keys that are lexically close to each other and generated close in time. Bigtable distributes write operations based on the row key, not one of the GCP load balancers. Replication does not impact where data is originally written. Bigtable is a wide column database and can support a large number of columns and the number of columns does not affect the distribution of data across nodes. See https://cloud.google.com/bigtable/docs/performance

Question 42: **Incorrect**
You are developing a data pipeline that will run several data transformation programs on Compute Engine virtual machines. You do not want to use your credentials for authenticating and authorizing these programs. You want to follow Google Cloud recommended practices, how would you authenticate and authorize the data transformation programs?

- **Create a Gmail account and use that account to create an IAM group. Store the password for the group in Secret Manager.**

- **Create a Gmail account and use that account to create an IAM user. Store the password for the account in Secret Manager.**

-

- **Create a service account and assign roles to the service account that are needed to execute the data transformation programs. Use Google managed keys to store both public and private portion of the service account keys.**
  (Correct)

- **Create a service account and assign roles to the service account that are needed to execute the data transformation programs. Use Secret Manager to store service account keys.**
  (Incorrect)

Explanation
Service accounts should be uses, not a user identity or a group. A service account should be created and assigned necessary roles. Google managed keys should be used for managing service accounts, not Secret Manager, which is used for secrets such as usernames and passwords. See https://cloud.google.com/docs/authentication/production

Question 43: **Incorrect**
To avoid hot-spotting in your Bigtable clusters, you have designed a row key that uses a UUID prefix. This is not working as expected and there is hot-spotting when writing data to Bigtable. What could be the cause of the hot-spotting?

- **You have chosen a type of UUID that has sequentially ordered strings.**

  (Correct)

- **Secondary indexes are slowing write operations.**

-

You have incorrectly configured column families.

- **The name of the row key column is too long.**

  (Incorrect)

Explanation
This could be caused by UUIDs that are sequentially generated. You should use UUID version 4 that uses a random number generator. Column families structure do not affect hot spotting. The name of a row key does not cause hot spotting. Bigtable does not support secondary indexes. See https://cloud.google.com/bigtable/docs/performance

Question 44: **Correct**
You work for a game developer that is using Cloud Firestore and needs to regularly create backups. You'd like to issue a command and have it return immediately while the backup runs in the background. You want the backup file to be stored in a Cloud Storage bucket named game-ds-backup. What command would you use?

- **gsutil datastore export gs://game-ds-backup --async**
- **gsutil datastore export gs://game-ds-backup**
- **gcloud datastore backup gs://game-ds-backup**
- **gcloud datastore export gs://game-ds-backup --async**
  (Correct)

Explanation
The correct command is gcloud datastore export gs://game-ds-backup --async. Export, not backup, is the datastore command to save data to a Cloud Storage bucket. Gsutil is used to manage Cloud Storage, not Cloud Datastore. See https://cloud.google.com/datastore/docs/export-import-entities and https://cloud.google.com/sdk/gcloud/reference/datastore/export

Question 45: **Incorrect**
An online gaming company has used a normalized database to manage players' in-game possessions but it is difficult to maintain because the schema has to change frequently to support new game features and types of possessions. What kind of data model would you recommend instead of a normalized data model?

- **Network model**
- **Document model (Correct)**
- **Star schema**
- **Snowflake schema (Incorrect)**

Explanation
A document model supports semi-structured schemas that frequently change. Both a star schema and snowflake schema are denormalized relational data models used in data warehousing but would not meet the needs of an interactive game. A network model is used to model graph-like structures such as transportation networks and is not as good a fit for the requirements as a document model. See https://firebase.google.com/docs/firestore/data-model

Question 46: **Incorrect**
You have to migrate a large volume of data from an on-premises data store to

Cloud Storage. You want to add metadata tags to objects with personally identifiable information. What two Google Cloud managed services could you use to accomplish this?

- **Data Catalog and Cloud Firestore**
- **Data Loss Prevention and Data Catalog (Correct)**
- **Compute Engine and Data Catalog (Incorrect)**
- **Data Loss Prevention and Compute Engine**

Explanation
Data Loss Prevention can identify personal identifiable information and Data Catalog can assign and store metadata tags to objects. Compute Engine could be used with custom applications but it is not a managed service. Cloud Firestore is a document database and could be used for storage but Data Catalog is a better option because it is designed specifically for this kind of use case. See https://cloud.google.com/dlp and https://cloud.google.com/data-catalog

Question 47: **Incorrect**
You are training a deep learning model for a classification task. The precision and recall of the model is quite low. What could you do to improve the precision and recall scores?

- **Use dropout**
-

Use L1 regularization
(Incorrect)

- **Use more training instances**
  (Correct)

- **Use L2 regularization**

Explanation
The correct answer is to use more training instances. This is an example of underfitting. The other options are all regularizations used in cases of overfitting. See https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Question 48: **Incorrect**
A developer is creating a dashboard to monitor a service that uses Cloud Pub/Sub. They want to know when applications that read data from a pull subscription in Cloud Pub/Sub are not keeping up with the messages being ingested. What metric would you recommend they monitor?

- **subscription/excess_ingestion_volume**
  (Incorrect)

- **subscription/num_undelivered_messages**
  (Correct)

- **topic/num_undelivered_messages**

topic/excess_ingestion_volume

Explanation
The subscription/num_undelivered_messages is the count of undelivered messages an one metric to indicate how well subscribers are keeping up with ingestion. The metric is tracked for subscriptions not topics. There is no metric called excess_ingestion_rate. See https://cloud.google.com/pubsub/docs/monitoring

Question 49: **Incorrect**
You are designing a Bigtable schema and have several groups of columns that are frequently used together. You want to optimize read performance and follow Google Cloud recommended best practices. How would you treat these groups of columns?

- **Put related columns in column family**
  (Correct)

- **Create secondary indexes that include all columns in a group.**
  Create one secondary index for each group.
  (Incorrect)

- **Put only one set of related columns in a table and use one table for each group**

- **Define a separate row key for each group.**

Explanation
Related columns should be placed in a column family. A single table can have multiple column families. Related data should be in one table, not multiple tables. Bigtable does not support secondary indexes. Row keys are specified for each row, not for each column family. See https://cloud.google.com/bigtable/docs/schema-design#best-practices

Question 50: **Incorrect**

You have created a Compute Engine instance with an attached GPU but the GPU is not used when you train a Tensorflow model. What might you do to ensure the GPU can be used for training your models? (Choose 2)

- **Use Pytorch instead of Tensorflow**
- **Install GPU drivers (Correct)**
- **Grant the Owner basic role to the VM service account (Incorrect)**
- **Update Python 3 on the VM**
- **Use a Deep Learning VM image (Correct)**

Explanation

GPU drivers need to be installed if they are not installed already when using GPUs. Deep Learning VM images have GPU drivers installed. Using Pytorch instead of Tensorflow will require work to recode and Pytorch would not be able to use GPUs either if the drivers are not installed. Updating Python will not address the problem of missing drivers. Granting a new role to the service account of the VM will not address the need to install GPU drivers. See https://cloud.google.com/compute/docs/gpus/install-drivers-gpu