20.0%

Great Start!

You did not pass the challenge on this attempt. This challenge is now locked and can be unlocked by using gems or by completing all of the recommended activities.

This challenge is now locked.

Unlock 0

Linux Academy

Go Back

Report Card

Expectations	Score
1. Google Cloud Data Engineer - Final Exam	20 %

Exam Breakdown

Google Cloud Data Engineer - Final Exam

1. **Which of these open source frameworks is best suited to process simultaneous batch and streaming in a single data pipeline?**

A. Apache Hadoop

B. Apache Kafka — ✕ **Your Answer**

Why is this incorrect?
Kafka is an open source framework that can handle streaming events, but does not do data processing.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/1/module/208)

C. Kubernetes

D. Apache Beam — ✓ **Correct**

Why is this correct?
Beam is able to process both stream and batch data in the same pipeline.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/1/module/208)

2. **Your production Bigtable instance is currently using four nodes. Due to the increased size of your table, you need to add additional nodes to offer better performance. How should you accomplish this without the risk of data loss?**

A. Power off your Bigtable instance, then increase the node count, then power back on. Be sure to schedule downtime in advance. — ✕ **Your Answer**

Why is this incorrect?
This is not necessary, (nor can you shut down a Bigtable instance). You can edit your instance size with no downtime. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208)

B. Export your Bigtable data as sequence files into Cloud Storage, then import the data into a new Bigtable instance with additional nodes added.

C. Use the node migration service to add additional nodes.

D. Edit instance details and increase the number of nodes. Save your changes. Data will re-distribute with no downtime. — ✓ **Correct**

Why is this correct?
You can add/remove nodes to Bigtable with no downtime necessary.
https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208)

3. **You are designing a relational data repository on Google Cloud to grow as needed. The data will be transactionally consistent and added from any location in the world. You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?**

A. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.

B. Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized. ✕ **Your Answer**

Why is this incorrect?
You should not use storage utilization as a scaling metric
https://linuxacademy.com/cp/courses/lesson/course/2113/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2113/lesson/1/module/208)

C. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

D. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span. ✓ **Correct**

Why is this correct?
This is correct because of the requirement for globally scalable transactions—use Cloud Spanner. CPU utilization is the recommended metric for scaling, per Google best practices, linked below.
https://linuxacademy.com/cp/courses/lesson/course/2113/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2113/lesson/1/module/208)

4. **You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud. You are using ANSI SQL to run queries for your analysts. You want to support complex aggregate queries and reuse existing code. How should you store and transform the input data?**

A. Use Cloud Storage for storage. Use Cloud Dataproc to run the transformations.

B. Use Cloud Storage for storage. Use Cloud Dataflow to run the transformations.

C. Use BigQuery for storage. Use Cloud Dataproc to run the transformations. ✓ **Correct**

D. Use BigQuery for storage. Use Cloud Dataflow to run the transformations.

5. What will happen to your data in a Bigtable instance if a node goes down?

A. Bigtable will attempt to rebuild the data from RAID disk configuration when the node comes back online.

B. Nothing, as the storage is separated from the node compute. ✓ **Correct**

Why is this correct?
Storage and compute are separate, so a node going down may affect performance, but not data integrity. Nodes only store pointers to storage as metadata.
https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208)

C. Lost data will automatically rebuild itself from Cloud Storage backups when the node comes back online.

D. Data will be lost, which makes regular backups to Cloud Storage necessary. ✗ **Your Answer**

Why is this incorrect?
While backups are a good idea, storage and compute are separate, so a node going down may affect performance, but not data integrity. Nodes only store pointers to storage as metadata.
https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208)

6. Your company needs to run analytics on their incoming inventory data. They need to use their existing Hadoop workloads to perform this task. What two steps must be performed to accomplish this? (Choose two answers)

A. Stream inventory data to Cloud Pub/Sub, process data with Cloud Dataflow into Bigtable and Cloud Storage.

B. Stream from Cloud Pub/Sub into Cloud Dataproc, which can then place relevant data in the appropriate storage location ✓ **Correct**

Why is this correct?
Dataproc can connect to Pub/Sub for the streaming ingest, when can then process the data and place in the correct location.

C. Use Spark to accept the streaming ingest on the Dataproc cluster, and then process jobs on HDFS. ✗ **Your Answer**

Why is this incorrect?
Pub/Sub is the correct service for streaming ingest.

D. Connect Cloud Dataproc to Bigtable and Cloud Storage, running analytics on the data in both services. ✓ **Correct**

Why is this correct?
Dataproc can natively connect to both services and can run analytics on both.

7. **Your company's Kafka server cluster has been unable to scale to the demands of their data ingest needs. Streaming data ingest comes from locations all around the world. How can they migrate this functionality to Google Cloud to be able to scale for future growth?**

A. Create a separate Pub/Sub topic for each region. Configure endpoints to publish to the Pub/Sub topic closest to their location, and configure a new Cloud Dataflow pipeline in each region to subscribe to the equivalent Pub/Sub topic to process messages as they come in.

B. Create a single Pub/Sub topic. Configure endpoints to publish to the Pub/Sub topic, and configure Cloud Dataflow to subscribe to the same topic to process messages as they come in. ✓ **Correct**

Why is this correct?
This is the preferred managed and scalable solution for handling streaming ingest, especially at a global scale.

C. Create a Computer Engine managed instance group that is configured to autoscale to 150% of peak demand. Use a managed instance template with Kafka installed to automatically scale as needed, and direct traffic to this autoscaling cluster. ✗ **Your Answer**

Why is this incorrect?
This technically works; however, Pub/Sub is the far better option from an availability, management, and scalability standpoint.

D. Create a Kubernetes Engine cluster in each region needed. Install Kafka on the cluster. Use an HTTP load balancer to serve each Kubernetes cluster region. Configure a new Cloud Dataflow pipeline in each region to process requests forwarded from the Kubernetes cluster.

8. **You have a long-running, streaming Dataflow pipeline that you need to shut down. You do not need to preserve data currently in the processing pipeline and need it shut down as soon as possible. Which shutdown option should you use to complete the shutdown process?**

A. Graceful shutdown

B. Cancel ✓ **Correct**

Why is this correct?
Cancel will shut down the pipeline without allowing buffered jobs to complete.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/5/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/5/module/208)

C. Stop ✗ **Your Answer**

Why is this incorrect?
This is not a valid option. https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/5/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/5/module/208)

D. Drain

9. Your organization has migrated their Hadoop workloads to Cloud Dataproc. To fully take advantage of the cloud, you want to decouple your Hadoop storage and compute, and be able to destroy your cluster when compute is complete in order to save costs while preserving your data. What should you do?

A. You must use another processing framework such as Apache Beam for this task.
B. Copy your data from HDFS to Cloud Storage. Update your scripts to point to the Cloud Storage location *(gs://)* instead of the HDFS location *(hdfs://)*. Within your Dataproc job, configure output to output to Cloud Storage. ✓ **Correct**
C. Use the Dataproc sync tool to synchronize HDFS with GCS.
D. You must leave your managed Dataproc cluster running in order to access computer data.

10. You are building a data pipeline on Google Cloud. You need to select services that will host a deep neural network machine learning model also hosted on Google Cloud. You also need to monitor and run jobs that could occasionally fail. What should you do?

A. Use the Cloud Machine Learning Engine to host your model. Monitor the status of the Jobs object for 'failed' job states. ✓ **Correct**
B. Use the Cloud Machine Learning Engine to host your model. Monitor the status of the Operation object for 'error' results.
C. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Jobs object for 'failed' job states.
D. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Operation object for 'error' results.

11. Your company is making the move to Google Cloud and has chosen to use a managed database service to reduce overhead. Your existing database is used for a product catalog that provides real-time inventory tracking for a retailer. Your database is 500 GB in size. The data is semi-structured and does not need full atomicity. You are looking for a truly no-ops/serverless solution. What storage option should you choose?

A. Cloud Datastore ✓ **Correct**

Why is this correct?
Datastore is perfect for semi-structured data less than 1TB in size. Product catalogs are a recommended use case.
https://linuxacademy.com/cp/courses/lesson/course/2109/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2109/lesson/1/module/208)

B. Cloud Bigtable ✗ **Your Answer**

Why is this incorrect?
Bigtable is recommended for high-performance analytical workloads over 1 TB in size and is not the best fit.
https://linuxacademy.com/cp/courses/lesson/course/2109/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2109/lesson/1/module/208)

C. Cloud SQL

D. BigQuery

12. You are creating a machine learning model to predict the likelihood of fraud from credit card transaction data. The end result will be predicting the percent confidence of two results: "Fraud" and "Not Fraud". What type of learning model problem is this?

A. Clustering

B. Classification ✓ **Correct**

Why is this correct?
Categorical is for a set of finite categories, such as 'yes' or 'no'. Fraud is a yes/no output, so this fits.
https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

C. Regression ✗ **Your Answer**

Why is this incorrect?
Regression is for continuous variables. Fraud is either 'yes' or 'no', so this does not match.
https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

D. Hyperparameter

13. You are a consultant for several organizations. Each organization has data in their own BigQuery table within a single project. For application access reasons, all of the tables must remain in the same project. You want to give access to each organization to view and run queries against their own data without exposing the data of organizations to unauthorized viewers. What should you do?

A. You must separate the tables by project, and use a service account in your application to access data in each project. Give out project-wide roles to each organization.

B. Place the tables in a single dataset, and apply IAM roles to each table, limiting access per table to each organization.

C. Place all data in a single table, create authorized views restricting access by row based on the SESSION_USER() field. Add that same SESSION_USER() field with the same email addresses according to which company needs access to which roles. ✕ **Your Answer**

Why is this incorrect?
This might technically work, but is substantially more cumbersome than placing each table in a different dataset and is much more prone to error. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/1/module/208)

D. Create a separate dataset for each organization in the same project. Place each organization's table in each dataset. Restrict access to the organization's dataset to only that company, from which they can view their table but no one else's. ✓ **Correct**

Why is this correct?
You can assign roles at the dataset level. Placing tables in different datasets allows you to limit access per dataset. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/1/module/208)

14. How can you set up your Dataproc environment to use BigQuery as an input and output source?

A. Use the Bigtable syncing service built into Dataproc.

B. Manually use a Cloud Storage bucket to import and export to and from both BigQuery and Dataproc.

C. You can only use Cloud Storage or HDFS for your Dataproc input and output. ✕ **Your Answer**

Why is this incorrect?
This is not true. You can also use Bigtable and BigQuery after installing the appropriate connector. https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/4/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/4/module/208)

D. Install the BigQuery connector on your Dataproc cluster. ✓ **Correct**

Why is this correct?
You can install the BigQuery connector to your cluster for direct programmatic read/write access to BigQuery. Note that a Cloud Storage bucket is used between the two services, but you'll interact directly with BigQuery from Dataproc. https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/4/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/4/module/208)

15. You are training a machine learning model to predict the liklihood of rain based on an available dataset of weather data. In reviewing your input data, the amount of humidity in the air has a very strong influence on the chance of rain, especially compared to less relevant data. How can you incorporate this more important data to that it properly influences the model?

A. Create a feature from the humidity data point, and use L2 regularization to optimize the model.

B. Tune your hyperparameters to give greater weighting to the humidty feature over others.

C. Create a feature from the humidity data point, and use L1 regularization to optimize the model. ✓ **Correct**

D. Reduce your epochs except for humidity features.

16. You currently have a Bigtable instance you've been using for development running a development instance type, using HDD's for storage. You are ready to upgrade your development instance to a production instance for increased performance. You also want to upgrade your storage to SSD's as you need maximum performance for your instance. What should you do?

A. Export your Bigtable data into a new instance, and configure the new instance type as production with SSD's ✓ **Correct**

Why is this correct?
Since you cannot change the disk type on an existing Bigtable instance, you will need to export/import your Bigtable data into a new instance with the different storage type. You will need to export to Cloud Storage then back to Bigtable again. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208)

B. Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD.

C. Run parallel instances where one instance is using HDD and the other is using SSD. ✗ **Your Answer**

Why is this incorrect?
This is not a possible option. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208)

D. Use the Bigtable instance sync tool in order to automatically synchronize two different instances, with one having the new storage configuration.

17. As part of a complex rollout, you have hired a third party developer consultant to assist with creating your Dataflow processing pipeline. The data that this pipeline will process is very confidential, and the consultant cannot be allowed to view the data itself. What actions should you take so that they have the ability to help build the pipeline but cannot see the data it will process?

A. Assign the consultant the Dataflow Developer IAM role. ✓ **Correct**

Why is this correct?
With the Developer IAM role, the developer will be able to create and cancel Dataflow jobs. Without other Google Cloud IAM roles, they will not be able to view the data that will be going through the pipeline.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208)

B. Apply custom encryption to the data before it goes through the pipeline. ✗ **Your Answer**

Why is this incorrect?
This is not necessary as the Dataflow Developer IAM role does not grant access to the data it will be used on.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208)

C. Use a separate development project to construct the pipeline with example data, therefore not exposing the live data to the developer's work environment.

D. Anonymize the data before it gets to the Dataflow pipeline.

18. When training a machine learning model on AI Platform on a distributed scaled tier, what types of machines are part of that distributed resource? (Choose all that apply)

A. Host

B. Worker ✓ **Correct**

C. Master ✓ **Correct**

Why is this correct?
You have a single Master instance per scaled tier.
https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

D. Parameter server ✓ **Correct**

Why is this correct?
Parameter servers coordinate shared model states between the workers.
https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

19. Which of these is NOT a valid reason to choose an HDD storage type over SSD in a Bigtable instance?

A. You need to maintain costs.

B. You plan on running batch workloads instead of frequently executing random reads across a small number of rows. ✕ **Your Answer**

Why is this incorrect?
This is a valid reason for choosing HDD storage.

C. You need to integrate Bigtable with Cloud Storage ✓ **Correct**

Why is this correct?
This is not a valid reason for choosing HDD storage, therefore it is the correct answer.

D. You need to store over 10TB of data.

20. You are creating a machine learning model for predicting a person's income given a variety of factors such as age, race, occupation, and others. What type of problem are we trying to solve in our prediction values?

A. Classification

B. Unsupervised learning ✕ **Your Answer**

Why is this incorrect?
Unsupervised learning includes clustering, which does not fit our model.
https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

C. Clustering

D. Linear Regression ✓ **Correct**

Why is this correct?
A linear regression problem is a set of continuous values, such as income, stock prices, etc. By contrast, a logistic regression model is more similar to a classification model (yes/no, true/false, etc).
https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

21. Your BigQuery table needs to be accessed by team members who are not proficient in technology. You want to simplify the columns they need to query to avoid confusion. How can you do this while preserving all of the data in your table?

A. Create a query that uses the reduced number of columns they will access. Save this query as a view in a different dataset. Give your team members access to the new dataset and instruct them to query against the saved view instead of the main table. ✓ **Correct**

Why is this correct?
This is the preferred method of preserving your live data while at the same time restricting what an end user can access. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/2/module/208)

B. Train your team members on how to query larger tables. ✗ **Your Answer**

Why is this incorrect?
This would technically work, but using a view is the better answer. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/2/module/208)

C. Apply column filtering to your table, and restrict the unfiltered view to yourself and those who need access to the full table.

D. Create a copy of your table in a different dataset, and remove the unneeded columns from the copy. Have your team members run queries against this copy.

22. What is the purpose of hyperparameters in a machine learning training model?

A. Form the basis of labels on your training data.

B. Hyperparameters adjust the training process itself. ✓ **Correct**

Why is this correct?
Learning rate and hidden layers (hyperparameters) are variables that adjust the learning model but have no relation to the training data used. https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208)

C. Train for a regression machine learning problem.

D. They help your model learn from the training data. ✗ **Your Answer**

Why is this incorrect?
Learning rate and hidden layers (hyperparameters) are variables that adjust the learning model but have no relation to the training data used. https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208)

23. You have 250,000 devices which produce a JSON device status event every 10 seconds. You want to capture this event data for outlier time series analysis. What should you do?

A. Ship the data into BigQuery. Develop a custom application that uses the BigQuery API to query the dataset and display a device's outlier data based on your business requirements.

B. Ship the data into Cloud Bigtable. Use the Cloud Bigtable `cbt` tool to display device outlier data based on your business requirements. ✓ **Correct**

C. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for the device outlier data based on your business requirements.

D. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.

24. Your online shopping company needs to know when a user has not interacted with the site in 30 minutes. They need the website to alert the user once they have been idle for too long. You use Cloud Dataflow to process the interaction events and decide if an alert should be sent. How should you design the pipeline?

A. Implement a session window with a gap time duration of 30 minutes. ✓ **Correct**

Why is this correct?
You need a window to be based around the last activity event, which a session window provides.

B. Implement a fixed-time window with a duration of 30 minutes.

C. Implement a global window with a time-based trigger with a delay of 30 minutes. ✗ **Your Answer**

Why is this incorrect?
You need a window to be based around the last activity event, which a session window provides.

D. Implement a sliding time window with a duration of 30 minutes.

25. You have hundreds of IoT devices that generate 1 TB of streaming data per day. Due to latency, messages will often be delayed compared to when they were generated. You must be able to account for data arriving late within your processing pipeline. What should you do?

A. Use Cloud SQL to process the delayed messages.

B. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Dataflow to process messages, and use windows, watermarks (timestamp), and triggers to process late data. ✓ **Correct**

Why is this correct?
Dataflow is the service that corrects out of order messages.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208)

C. Use SQL queries in BigQuery to analyze data by timestamp.

D. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Pub/Sub to process messages by timestamp and fix out of order issues. ✗ **Your Answer**

Why is this incorrect?
Pub/Sub does not care about message order; you would use Dataflow to process out of order messages by timestamp. https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208)

26. Which of these is NOT a type of trigger that applies to Dataflow?

A. Element size in bytes. ✓ **Correct**

Why is this correct?
Element size is not a type of trigger, therefore it is our correct answer.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208)

B. Element count.

C. Combinations of other triggers. ✗ **Your Answer**

Why is this incorrect?
This is a valid trigger type. https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208)

D. Timestamp

27. You are developing an application that will only recognize and tag specific business to business product logos in images. What is the best method to accomplish this task?

A. Use the Cloud Vision API to recognize logos in the images.

B. Create a custom machine learning model to recognize specific logos in photos, then train it on Cloud ML Engine. ✓ **Correct**

C. Train your model on Kubernetes Engine to scale training as quickly as possible.

D. Use the Cloud Vision API to recognize all logos in images, then use the Cloud Natural Language API to recognize specific logos by name.

28. You need to replicate the logs that are ingested by your on-premises Apache Kafka cluster to Google Cloud to be stored for analysis in BigQuery. What should you do?

A. Create an identical Kafka cluster on Compute Engine in GCP. Configure your on-premises Kafka cluster to duplicate all data to the GCP Kafka cluster. Use a Dataflow job to process data from Kafka and insert into BigQuery.

B. Configure the Pub/Sub Kafka connector on your on-premises Kafka cluster, and configure Pub/Sub as a source connector. Use a Cloud Dataflow job to read from a subscribed Pub/Sub topic and write to BigQuery ✗ **Your Answer**

Why is this incorrect?
Almost correct, but Pub/Sub needs to be the sink connector, not source.

C. Create a Cloud Composer workflow to manage the replication of data from your Kafka cluster directly into BigQuery.

D. Configure the Pub/Sub Kafka connector on your on-premises Kafka cluster, and configure Pub/Sub as a sink connector. Use a Cloud Dataflow job to read from a subscribed Pub/Sub topic and write to BigQuery ✓ **Correct**

Why is this correct?
You can connect Kafka to GCP by using a connector. The 'downstream' service (Pub/Sub) will use a sink connector.

29. In machine learning, what is the difference between test and training data?

A. Training data is used for hyperparameter tuning, and test data is used for feature engineering.

B. Test data is used to tune parameters, like weights and biases.

C. Test data is labeled with the 'correct' answer; training data is not. ✕ **Your Answer**

Why is this incorrect?
Training data is your labeled data and used to 'train' your model. Test data is used to 'test' the model for accuracy without labels. https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

D. Training data has a label attached to train on features for the correct answer. Test data is used to test the trained model for accuracy when completed on new data. ✓ **Correct**

Why is this correct?
Training data has labels to act as the 'source of truth'. Both data types may have labels attached to them, but the training data is used to 'train' the model, and test data 'tests' the trained model for accuracy on new data. https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

30. You are building a data pipeline on Google Cloud. You need to prepare source data for a machine-learning model. This involves quickly deduplicating rows from three input tables and also removing outliers from data columns where you do not know the data distribution. What should you do?

A. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Write a recipe to transform the data and add it to the Cloud Dataprep job.

B. Write an Apache Spark job with a series of steps for Cloud Dataflow. The first step will examine the source data, and the second and third steps will perform data transformations.

C. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Click on each column name, click on each appropriate suggested transformation, and then click **Add** to add each transformation to the Cloud Dataprep job. ✓ **Correct**

D. Write an Apache Spark job with a series of steps for Cloud Dataproc. The first step will examine the source data, and the second and third steps will perform data transformations.

31. Your company's aging Hadoop servers are nearing end of life. Instead of replacing your hardware, your CIO has decided to migrate the cluster to Google Cloud Dataproc. A direct lift and shift migration of the cluster would require 30 TB of disk space per individual node. There are cost concerns about using that much storage. How can you best minimize the cost of the migration?

A. Decouple storage from computer by placing the data in Cloud Storage ✓ **Correct**

Why is this correct?
Placing all input and output data in Cloud Storage allows you to 1. Treat clusters as ephemeral and 2. Use a much cheaper storage location compared to persistent disks without a noticeable impact on performance.

B. Place archived data in Cloud Storage, and only use 'hot' data in HDFS on the cluster disks.

C. Implement maximum data compression to reduce the amount of disk space your data uses. ✗ **Your Answer**

Why is this incorrect?
This is not necessary and will likely lead to performance degradation.

D. Use preemptible VM's to save costs on cluster storage usage.

32. Your organization needs to be able to reliably handle ever-increasing amounts of streaming telemetry data, process it, and economically store analyzed data. What services should they use for this task?

A. Stackdriver, Cloud Dataproc, Cloud Spanner

B. Cloud Pub/Sub, Cloud Dataproc, Bigtable

C. Cloud Pub/Sub, Cloud Dataflow, Bigquery ✓ **Correct**

Why is this correct?
Pub/Sub for streaming data ingest, Dataflow for processing streaming data, and BigQuery for storage and analysis.

D. Kubernetes Engine, Cloud Dataflow, Cloud Datastore ✗ **Your Answer**

Why is this incorrect?
Cloud Dataflow is correct, but Pub/Sub is used for streaming ingest, and Datastore would not be able to handle the size of data and is not ideal for analytics.

33. You are working on a project with two compliance requirements. The first requirement states that your developers should be able to see the Google Cloud Platform billing charges for only their projects. The second requirement states that your finance team members can set budgets and view the current charges for all projects in the organization. The finance team should not be able to view the project contents. You want to set permissions. What should you do?

A. Add the finance team to the Viewer role for the Project. Add the developers to the Security Reviewer role for each of the billing accounts.

B. Add the developers and finance managers to the Viewer role for the Project.

C. Add the finance team members to the default IAM Owner role. Add the developers to a custom role that allows them to see their spending only. ✕ **Your Answer**

Why is this incorrect?
Primitive roles are far too broad for this requirement. https://cloud.google.com/iam/docs/understanding-roles (https://cloud.google.com/iam/docs/understanding-roles)

D. Add the finance team members to the Billing Administrator role for each of the billing accounts that they need to manage. Add the developers to the Viewer role for the Project. ✓ **Correct**

Why is this correct?
This answer uses the principle of least privilege for IAM roles. https://cloud.google.com/iam/docs/understanding-roles (https://cloud.google.com/iam/docs/understanding-roles)

34. Which of these numbers are adjusted by a machine learning neural network as it works with its training dataset? (Choose all that apply)

A. Weights ✓ **Correct**

Why is this correct?
Weights are a parameter that adjusts for a neural network to learn from its training data. https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208)

B. Epochs ✕ **Your Answer**

Why is this incorrect?
An epoch is a single pass through the training dataset. https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208)

C. Biases ✓ **Correct**

D. Features

35. Pick two benefits of using denormalized data in BigQuery? (Choose all that apply)

A. Decreased query complexity — ✓ **Correct**

Why is this correct?
Not having to use JOIN clauses due to combined tables makes queries easier.
https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208)

B. Less storage space used — ✗ **Your Answer**

Why is this incorrect?
Denormalizing tables has no effect on storage amounts.
https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208)

C. Increased query performance — ✓ **Correct**

D. Reduces the amount of data processed

36. When training a machine learning model, why do you need separate training and test data?

A. Without different data, your model will not generalize for additional data, known as overfitting. — ✓ **Correct**

Why is this correct?
Without separate sets of data, your model will only learn from specifically the training data, and not new data.
https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

B. Both sets of data are necessary for deep and wide neural networks. — ✗ **Your Answer**

Why is this incorrect?
Neural network formation has nothing to do with separating test and training data.
https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/1/module/208)

C. Your learning model will have an improper learning rate, making training difficult.

D. Without separate sets of data, your neural network will not have enough data to train with.

37. What types of jobs does Cloud Dataproc support? (Choose all that apply)

A. Hive	✓ **Correct**
B. Beam	
C. Pig	✓ **Correct**
D. Spark	✓ **Correct**

38. You are training a facial detection machine learning model. Your model is suffering from overfitting your training data. Choose three steps you can take to solve this problem.

A. Use a larger set of features	
B. Use a smaller set of features	✓ **Correct**
C. Reduce the number of training examples **Why is this incorrect?** More data is one of the best methods to increase the variety of samples and better generalize your model.	✕ **Your Answer**
D. Increase the number of training examples	✓ **Correct**
E. Increase the regularization parameters **Why is this correct?** Increasing your regularization parameters allows you to reduce 'noise' in your model to reduce overfitting.	✓ **Correct**
F. Decrease the regularization parameters	

39. You are selecting a streaming service for log messages that must include final result message ordering as part of building a data pipeline on Google Cloud. You want to stream input for 5 days and be able to query the most recent message value. You will be storing the data in a searchable repository. How should you set up the input messages?

A. Use Apache Kafka on Compute Engine for input. Attach a timestamp to every message in the publisher.

B. Use Cloud Pub/Sub for input. Attach a unique identifier to every message in the publisher.

C. Use Apache Kafka on Compute Engine for input. Attach a unique identifier to every message in the publisher. ✕ **Your Answer**

Why is this incorrect?
Apache Kafka is overly complex compared to using Cloud Pub/Sub, which can support all of the requirements. https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208)

D. Use Cloud Pub/Sub for input. Attach a timestamp to every message in the publisher. ✓ **Correct**

Why is this correct?
Adding a timestamp is necessary for making sure that the final result messaging is in the correct order. https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208)

40. You want to display aggregate view counts for your YouTube channel data in Data Studio. You want to see the video tiles and view counts summarized over the last 30 days. You also want to segment the data by the Country Code using the fewest possible steps. What should you do?

A. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title and Country Code as report dimensions.

B. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter. ✕ **Your Answer**

Why is this incorrect?
You do not need to export data from YouTube to Cloud Storage; you can simply use the existing YouTube data source. https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/2/module/208)

C. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter.

D. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title and Country Code as report dimensions. ✓ **Correct**

Why is this correct?
There is no need to export; you can use the existing YouTube data source. Country Code is a dimension because it's a string and should be displayed as such, that is, showing all countries, instead of filtering. https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/2/module/208)

41. Why do you want to train a machine learning model locally before training on cloud resources? (Choose all that apply)

A. Faster training with scaling resources.

B. Faster iteration. ✓ **Correct**

Why is this correct?
Local training allows you to make faster adjustments.
https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

C. Save costs. ✓ **Correct**

D. Restrict access to other parties.

42. Your infrastructure includes two 100-TB enterprise file servers. You need to perform a one-way, one-time migration of this data to the Google Cloud securely. Only users in Germany will access this data. You want to create the most cost-effective solution. What should you do?

A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.

B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination. ✗ **Your Answer**

Why is this incorrect?
Storage Transfer Service is not for data stored on-premises, but for AWS/Google Cloud/online locations.
https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208)

C. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination. ✓ **Correct**

Why is this correct?
This answer is correct because you are performing a one-time (rather than an ongoing series) data transfer from on-premises to Google Cloud Platform for users in a single region (Germany). Using a Regional storage bucket will reduce cost and also conform to regulatory requirements.
https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208)

D. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Multi-Regional bucket as a final destination.

43. As part of your backup plan, you create regular boot-disk snapshots of Compute Engine instances that are running. You want to be able to restore these snapshots using the fewest possible steps for replacement instances. What should you do?

A. Export the snapshots to Cloud Storage. Create images from the exported snapshot files.

B. Use the snapshots to create replacement disks. Use the disks to create instances as needed. ✕ **Your Answer**

Why is this incorrect?
This is more steps than needed. You can recreate instances directly from a boot-disk snapshot.

C. Use the snapshots to create replacement instances as needed. ✓ **Correct**

Why is this correct?
Snapshots let you recreate instances in the fewest steps.

D. Export the snapshots to Cloud Storage. Create disks from the exported snapshot files. Create images from the new disks.

44. While conducting BigQuery queries against a large table with many columns, you notice in the details section that you have a very large purple bar in the first stage of your query execution. How can you troubleshoot this to increase performance and reduce costs? (Choose all that apply)

A. Restrict the number of columns in your SELECT field for those needed. This will reduce read times on your query. ✓ **Correct**

Why is this correct?
The purple bar indicates the number of read operations. Limiting columns read will reduce the read time of your query. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208)

B. Partition or separate your large table into smaller pieces. Conduct a query against your smaller (or partitioned) tables to reduce read times. ✓ **Correct**

C. Reduce the number of read operations by adding a LIMIT clause to your query.

D. Reduce the number of write operations by optimizing the complexity of your query functions.

45. When using AI Platform to train machine learning models, how are online predictions different from batch predictions? (Choose all that apply)

A. Online prediction results are written to Cloud Storage as output.

B. Online predictions are returned in the response message. ✓ **Correct**

Why is this correct?
Online predictions create near real-time feedback with small, inline predictions.
https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

C. Batch predictions are used to reduce latency in serving predictions.

D. Batch predictions are optimized to handle a high volume of prediction examples while running on more complex models. ✓ **Correct**

46. You are migrating a Hadoop cluster to Cloud Dataproc using GCS for storage. After migration, some of your existing, more complex Spark jobs (in parquet format) are performing noticably worse than your on-premises cluster. You are using mostly preemptible VM's (with a few required non-preemptible) in order to save on costs.

A. Change your file format to CSV format

B. Increase the size of your cluster by twice as many preemptible VM's ✗ **Your Answer**

Why is this incorrect?
This would work, but would also substantially increase costs.

C. Switch disks from HDD to SSD. Change the default preemptible VM settings to increase the size of the boot disk. ✓ **Correct**

Why is this correct?
By default, preemptible node disk sizes are limited to 100GB or the size of the non-preemptible node disk sizes, whichever is smaller. However you can override the default preemptible disk size to any requested size. Since the majority of our cluster is using preemptible nodes, the size of the disk used for caching operations will see a noticeable performance improvement using a larger disk. Also, SSD's will perform better than HDD. This will increase costs slightly, but is the best option available while maintaining costs.

D. Switch your disks from HDD to SSD, run the job in HDFS before copying the results back to GCS

E. Ensure that your parquet files are at an optimized block size

47. You are evaluating a storage solution for your data. Your data is in a structured, non-relational format, and will be used for analysis. You need the lowest latency read and write speeds possible. Your data is about 3 TB in size, predicted to grow to up to 5 TB. What solution should you use?

A. Use BigQuery to host your non-relational, structured data.
B. Use Cloud Bigtable using HDD storage.
C. Use Cloud Bigtable with SSD storage. ✓ **Correct**
D. Use Cloud Datastore for your operations.

48. Your organization is streaming telemetry data into BigQuery for long-term storage (2 years) and analysis, at the rate of about 100 million records per day. They need to be able to run queries against certain time periods of data without incurring the costs of querying all available records. What is the preferred method for doing so?

A. Create a single table, but query only individual rows by data in the WHERE clause.
B. Use a LIMIT clause to limit the number of rows queried based on WHERE clause criteria.
C. Partition a single table by day, and run queries against individual partitions. ✓ **Correct**
D. Create a new table, one for each day. Run queries against the groups of tables relevant to their needs.

49. You regularly use prefetch caching with a Data Studio report to visualize the results of BigQuery queries. You want to minimize service costs. What should you do?

A. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and verify that the *Enable cache* checkbox is selected for the report. ✓ **Correct**

Why is this correct?
You must set Owner credentials to use the *enable cache* option in BigQuery. It is also a Google best practice to use the *enable cache* option when the business scenario calls for using prefetch caching.
https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/1/module/208)

B. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and direct the users to view the report only once per business day (24-hour period). ✗ **Your Answer**

Why is this incorrect?
Cache auto-expires after 12 hours. 24-hour cache is not a valid option.
https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/1/module/208)

C. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and verify that the *Enable cache* checkbox is not selected for the report.

D. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and also set it up to be a *view-only* report.

50. Your organization needs to develop their machine learning model to control topology definitions. There are a large number of possible configurations to achieve the best results. What components of their machine learning model would they adjust to account for increased complexity? (Choose two answers.)

A. Learning rate ✗ **Your Answer**

Why is this incorrect?
Learning rate is a hyperparameter, not related to adjusting to training data.

B. Neurons ✓ **Correct**

Why is this correct?
Adding additional neurons allows combining more input values.

C. Epoch

D. Hidden layers ✓ **Correct**