58.0%

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	58 %

Exam Breakdown

Google Cloud Data Engineer - Final Exam

1. **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**

Why is this correct?
Cloud Machine Learning Engine is the correct service for deep neural network models. You would correctly monitor Jobs for failures. https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

B. Use the Cloud Machine Learning Engine to host your model. Monitor the status of the Operation object for 'error' results. ✗ **Your Answer**

Why is this incorrect?
Cloud Machine Learning Engine is the correct service for deep neural network models. However, you would not monitor operation objects for failed jobs.
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. 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.

2. **You are designing storage for event data as part of building a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying individual values over time windows. Which storage service and schema design should you use?**

A. Use Cloud Bigtable for storage. Design tall and narrow tables, and use a new row for each single event version. ✓ **Correct**

B. Use Cloud Bigtable for storage. Design short and wide tables, and use a new column for each single event version.

C. Use Cloud Storage for storage. Join the raw file data with a BigQuery log table.

D. Use Cloud Storage for storage. Write a Cloud Dataprep job to split the data into partitioned tables.

3. **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.

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

4. **What is the difference between a deep and wide neural network? What would you use a deep AND wide neural network for? (Choose all that apply)**

A. Wide models are used for generalizations. Deep models are for memorization. ✕ **Your Answer**

Why is this incorrect?
This is backward. The reverse is true.
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. Deep and wide models are ideal for solving regression problems.

C. Wide models are used for memorization. Deep models are for generalization ✓ **Correct**

Why is this correct?
This is one of the correct answers. https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208)

D. Deep and wide models are ideal for a recommendation application. ✓ **Correct**

5. **You need to run analytical queries using SQL syntax against data formatted in JSON format. What should you do? Choose the best answer.**

A. Load your JSON data into Cloud SQL, and run queries against it in that service.

B. Load your JSON data into Cloud Storage. Add your JSON table as an external read source in BigQuery, since BigQuery is unable to store data in JSON format.

C. Import the data into Bigtable and use Bigtable for your queries.

D. Import the data in JSON format into BigQuery as a table, and run queries against it. ✓ **Correct**

6. **You are configuring your Cloud Pub/Sub subscription. Assuming that all requirements are met, which subscription delivery method offers better 'near real-time' delivery of messages?**

A. Pull

B. Push ✓ **Correct**

C. Cached

D. Instant

7. **You want to export your Cloud SQL tables into BigQuery for analysis. How can you do this?**

A. Convert your Cloud SQL data to JSON format, then import directly into BigQuery	
B. Export your Cloud SQL data to Cloud Storage, then import into BigQuery	✓ **Correct**
C. Import data from BigQuery directly from Cloud SQL.	
D. Use the BigQuery export function in Cloud SQL to manage exporting data into BigQuery.	

8. **You are setting up Cloud Dataproc to perform some data transformations using Apache Spark jobs. The data will be used for a new set of non-critical experiments in your marketing group. You want to set up a cluster that can transform a large amount of data in the most cost-effective way. What should you do?**

A. Set up a cluster in High Availability mode with default machine types. Add 10 additional Preemptible worker nodes.	
B. Set up a cluster in Standard mode with high-memory machine types. Add 10 additional Preemptible worker nodes.	✓ **Correct**
C. Set up a cluster in Standard mode with the default machine types. Add 10 additional local SSDs.	
D. Set up a cluster in High Availability mode with high-memory machine types. Add 10 additional local SSDs.	

9. **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**
D. Kubernetes Engine, Cloud Dataflow, Cloud Datastore	

10. 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

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 ✗ **Your Answer**

Why is this incorrect?
While the block size of parquet files does have an impact on performance in a complex Spark job, these are the same jobs and configurations that were run on the on-premises Hadoop cluster. The change in performance in two different environments with identical job configurations does not indicate a job configuration or file format issue.

11. 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

C. Clustering

D. Linear Regression ✓ **Correct**

12. What is the recommended minimum amount of data to store in Bigtable?

A. 500 GB — ✕ **Your Answer**

Why is this incorrect?
Google recommends that workloads of less than 1TB should not be used in Bigtable, especially from a cost/value perspective. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208)

B. 1 GB

C. 1 TB — ✓ **Correct**

Why is this correct?
Google recommends that workloads of less than 1TB should not be used in Bigtable, especially from a cost/value perspective. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208)

D. 500 TB

13. You need to choose a structure storage option for storing very large amounts of data with the following properties and requirements:

- The data has a single key
- You need very low latency Which solution should you choose?

A. Bigtable — ✓ **Correct**

Why is this correct?
Bigtable uses a single key and has very low latency (in milliseconds). It is the best choice.

B. Datastore — ✕ **Your Answer**

Why is this incorrect?
Datastore stores less data than Bigtable, and operates on multiple keys.

C. Cloud SQL

D. BigQuery

14. 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

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**

15. You have a Dataflow job that keeps failing due to errors in your input data. What steps can you take to improve pipeline reliabiity while at the same time, capturing failed data for reprocessing?

A. Implement a try-catch block that transforms the both good and bad data. Create an additional output to use a new PCollection that can be output to Pub/Sub for later analysis. ✓ **Correct**

Why is this correct?
Your pipeline needs to use an additional side output, that uses a PCollection to output erroneous data to Pub/Sub.

B. Filter out errors as they occur, and view error entries using Stackdriver Logging

C. Implement a try-catch block that transforms the both good and bad data, and extract the incorrect entries from Stackdriver Logging

D. Implement a try-catch block that transforms the both good and bad data. Publish the erroneous data to Pub/Sub, which can then be placed into GCS for further analysis. ✗ **Your Answer**

Why is this incorrect?
Almost correct, however the erroneous data needs to be exported via a side output via PCollection, not just written directly to Pub/Sub.

16. You work at a very large organizations that has a very large analyst team. You use the default pricing model for BigQuery. During heavy usage, your analyst group occasionally runs out of the 2000 slots avaiable for the BigQuery jobs. You do not want to create additional projects for the sole purpose of increasing slot count. What can you do to resolve this?

A. You must create an additional project to increase your slot count, then spread the BigQuery loads across both projects.

B. Force-enable the 'use cached results' option for all available queries.

C. Switch to flat rate pricing to enable a higher total slot quota for your project. ✓ **Correct**

D. Use the quotas page to increase your BigQuery slot count to 3000 as needed.

17. 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.
D. Use Cloud Pub/Sub for input. Attach a timestamp to every message in the publisher. ✓ **Correct**

18. You need to deploy a TensorFlow machine-learning model to Google Cloud. You want to maximize the speed and minimize the cost of model prediction and deployment. What should you do?

A. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud Storage. Run 1 version on CPUs and another version on GPUs.
B. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud ML Engine. Run 1 version on CPUs and another version on GPUs.
C. Export your trained model to a SavedModel format. Deploy and run your model from a Kubernetes Engine cluster
D. Export your trained model to a SavedModel format. Deploy and run your model on Cloud ML Engine. ✓ **Correct**

19. 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.

20. 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

B. Neurons ✓ **Correct**

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

C. Epoch ✗ **Your Answer**

Why is this incorrect?
An Epoch is a pass through the training dataset, not related to complexity.

D. Hidden layers ✓ **Correct**

21. 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. ✗ **Your Answer**

Why is this incorrect?
This is technically possible, but all data (hot and cold) can be placed in Cloud Storage and still perform well.

C. Implement maximum data compression to reduce the amount of disk space your data uses.

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

22. You are setting up multiple MySQL databases on Compute Engine. You need to collect logs from your MySQL applications for audit purposes. How should you approach this?

A. Configure Cloud Composer to monitor and report on instance performance metrics.	
B. Install the Stackdriver Logging agent on your database instances and configure the fluentd plugin to read and export your MySQL logs into Stackdriver Logging.	✓ **Correct**
C. Install the Stackdriver Monitoring agent on your instances, configure the MySQL plugin, and export logs to Stackdriver Monitoring.	
D. Configure Stackdriver Logging to natively monitor application logs, which will appear in Stackdriver Logging.	

23. Your organization is ready to migrate their Hadoop workloads to Google Cloud. For the data migration, they need a cost-effective 'data lake' that will scale to their growing data needs and be able to easily connect to their Hadoop workloads in the cloud. What two actions should they perform?

A. Install the Bigtable connector in the on-premises Hadoop cluster, then migrate data to Bigtable for long-term storage.	
B. Add the Cloud Storage connector to their on-premises Hadoop environment, and transfer their data to a Cloud Storage bucket.	✓ **Correct**
C. For the existing Hadoop jobs that are migrating to Dataproc, use the *gs://* prefix instead of *hdfs://* to access data from Cloud Storage.	✓ **Correct**
D. Create a Dataproc cluster for long-term use, and transfer data to the HDFS partition on the cluster.	

24. 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.	
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**

25. 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**
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).	
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.	

26. 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. **Why is this incorrect?** Hyperparameters deal with learning rate, which is not relevant for this question.	✗ **Your Answer**
C. Create a feature from the humidity data point, and use L1 regularization to optimize the model. **Why is this correct?** L1 regularization is able to reduce the weights of less important features to zero or near zero.	✓ **Correct**
D. Reduce your epochs except for humidity features.	

27. Your organization is making the move to Google Cloud. You need to bring your existing big data processing workflows to the cloud without having to re-train employees on new products. Your organization uses the Apache Hadoop ecosystem for big data processing. Which Google Cloud managed service would your workflow move to?

A. Cloud Dataproc	✓ **Correct**
B. Cloud Bigtable	
C. Cloud Pub/Sub	
D. Cloud Dataflow	

28. In order to protect live customer data, your organization needs to maintain separate operating environments —development/test, staging, and production— to meet the needs of running experiments, deploying new features, and serving production customers. What is the best practice for isolating these environments while at the same time maintaining operability?

A. Create separate organization accounts for each environment, and use domain wide IAM roles to allow access between each organization environment to share data as needed.

B. Create a separate project for dev/test, staging, and production. Migrate relevant data between projects when ready for the next stage. ✓ **Correct**

C. Place all three environments in the same project, however, use separate Cloud Storage buckets, Cloud ML Engine clusters, and other services for each environment

D. Place resources into the same project. but use object versioning in Cloud Storage in order to separate data by environment.

29. 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.

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. ✗ **Your Answer**

Why is this incorrect?
You cannot produce a summarized report that meets your business requirements using the options listed. Using Views as the metric and setting Video Title and Country Code as the report dimensions is the better option.
https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2250/lesson/2/module/208)

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)

30. Your BigQuery dataset contains 1500 tables. When conducting a query, you are limited to a maximum of 1000 tables that you can query at once. You need to query data across all 1500 tables. What should you do?

A. Place tables into separate datasets.

B. If possible, merge the 1500 tables to bring the total number below 1000. You may still partition single tables to divide data for queries. ✓ **Correct**

Why is this correct?
If you have over 1000 tables, you need to bring that number to below 1000 to query all of them at once. Merge tables, then use table partitioning to divide single tables into segments (called partitions), as long they are partitioned by time. 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. Export the data to Bigtable, and conduct your query inside of Bigtable.

D. Create multiple views of chunks of the 1500 tables, then query the multiple views. ✗ **Your Answer**

Why is this incorrect?
This will not work as it will still limit to 1000 tables per query, even if hidden behind views. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208)

31. 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.

32. 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.

D. Timestamp ✗ **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)

33. What types of Bigtable row keys can lead to hotspotting? (Choose all that apply)

A. Leading with a non-reversed timestamp. ✓ **Correct**

B. Standard domain names (non-reversed). ✓ **Correct**

C. Reverse timestamps.

D. Non-sequential numeric IDs.

34. 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.

35. 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.

36. You created a job which runs daily to import highly sensitive data from an on-premises location to Cloud Storage. You also set up a streaming data insert into Cloud Storage via a Kafka node that is running on a Compute Engine instance. You need to encrypt the data at rest and supply your own encryption key. Your key should not be stored in the Google Cloud. What should you do?

A. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in your Kafka node hosted on Compute Engine.
B. Create a dedicated service account, and use encryption at rest to reference your data stored in Cloud Storage and Compute Engine data as part of your API service calls.
C. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in Cloud Storage. Use your uploaded encryption key and reference it as part of your API service calls to encrypt your data in the Kafka node hosted on Compute Engine. ✕ **Your Answer** **Why is this incorrect?** The scenario states that you should use, but not store, your own key with Google Cloud Platform services. https://cloud.google.com/storage/docs/encryption/customer-supplied-keys (https://cloud.google.com/storage/docs/encryption/customer-supplied-keys)
D. Supply your own encryption key, and reference it as part of your API service calls to encrypt your data in Cloud Storage and your Kafka node hosted on Compute Engine. ✓ **Correct** **Why is this correct?** The question requires you to use your own key and also not store your key on Google Cloud. https://cloud.google.com/storage/docs/encryption/customer-supplied-keys (https://cloud.google.com/storage/docs/encryption/customer-supplied-keys)

37. You have in your possession a database of financial transactions, which include a user's name, location, purchase location, and purchase amount. With this data, what two types of machine learning can potentially applied to this dataset?

A. Apply supervised regressing learning to label which transactions are likely to be fraudulent

B. Apply unsupervised learning to label which transactions are likely to be fraudulent.

C. Unsupervised learning to identify patterns (clustering) in the data to predict the location of future purchases. ✓ **Correct**

Why is this correct?
Unsupervised learning does not use labels but does look for patterns (or clustering) of data in order to make predictions based on the patterns it learns.

D. Apply reinforcement learning to predict the location of purchase. ✕ **Your Answer**

Why is this incorrect?
Reinforcement learning uses reward systems to complete a task. Predicting the location would be an example of using unsupervised learning to find patterns.

E. Apply labels to the data based on whether it is fraudulent or not-fraudulent. Then apply supervised classification learning to predict which future transactions are likely to be fraudulent ✓ **Correct**

38. What open source software is Cloud Pub/Sub most similar to?

A. Apache Beam

B. Apache Kafka ✓ **Correct**

C. HBase

D. Apache Hadoop

39. 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. ✕ **Your Answer**

Why is this incorrect?
This is not a valid Bigtable function. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/1/module/208)

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.

40. 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.

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**

41. 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

42. Your infrastructure runs on another cloud and includes a set of multi-TB enterprise databases that are backed up nightly both on-premises and also to that cloud. You need to create a redundant backup to Google Cloud. You are responsible for performing scheduled monthly disaster recovery drills. You want to create a cost-effective solution. What should you do?

A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Nearline storage bucket as a final destination. — ✓ **Correct**

Why is this correct?
This is correct because you will need to access your backup data monthly to test your disaster recovery process, so you should use a Nearline bucket; also, because you will be performing ongoing, regular data transfers, so you should use the storage transfer service.
https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208)

B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Coldline storage bucket as a final destination.

C. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Nearline storage bucket as a final destination.

D. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Coldline bucket as a final destination. — ✗ **Your Answer**

Why is this incorrect?
Transfer Appliance is used for on-premises transfers, not cloud-to-cloud, and is not used for repeated/scheduled transfers. Also, Coldline buckets need to stay un-modified for 3 months (90 days) to avoid additional charges, and your scenario calls for once a month access.
https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2103/lesson/3/module/208)

43. Your team has decided to use Datalab for interactive machine learning exercises. You want your team members to share their work and progress with each other. How do you accomplish this?

A. Every team member will use their own Datalab notebook and synchronize changes to the shared Cloud Source Repository.	✓ **Correct**
B. Use the team sync feature included in Datalab notebooks to synchronize each member's work.	
C. Give your team members Compute Instance Admin and Service Account Actor roles to access a shared notebook.	
D. Create a shared Datalab notebook, and assign the Datalab Editor role to your team members to access it.	

44. 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.	
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**

45. 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**

Why is this correct?
The data type, volume, and query pattern best fits BigTable capabilities and also Google best practices. Also, the `cbt` tool is a simpler method for access.
https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/2/module/208)

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. ✗ **Your Answer**

Why is this incorrect?
Using the `cbt` tool is a simpler method of querying your data than installing an HBase shell.
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. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.

46. 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**

C. Save costs. ✓ **Correct**

D. Restrict access to other parties.

47. You are building storage for files for a data pipeline on Google Cloud. You want to support JSON files. The schema of these files will occasionally change. Your analyst teams will use running aggregate ANSI SQL queries on this data. What should you do?

A. Use Cloud Storage for storage. Link data as permanent tables in BigQuery and turn on the *Automatically detect* option in the Schema section of BigQuery.

B. Use BigQuery for storage. Provide format files for data load. Update the format files as needed.

C. Use BigQuery for storage. Select **Automatically detect** in the Schema section. ✓ **Correct**

Why is this correct?
This is correct because of the requirement to support occasionally (schema) changing JSON files and aggregate ANSI SQL queries; you need to use BigQuery, and it is quickest to use *Automatically detect* for schema changes. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208)

D. Use Cloud Storage for storage. Link data as temporary tables in BigQuery and turn on the *Automatically detect* option in the Schema section of BigQuery. ✗ **Your Answer**

Why is this incorrect?
This is not correct because you should not use Cloud Storage for this scenario; it is cumbersome and doesn't add value. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208)

48. 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**

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.

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.

49. 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 **Why is this incorrect?** Reducing the number of unneeded features can help reduce overfitting.	✕ **Your Answer**
B. Use a smaller set of features **Why is this correct?** Reducing the number of unneeded features can help reduce overfitting.	✓ **Correct**
C. Reduce the number of training examples	
D. Increase the number of training examples	✓ **Correct**
E. Increase the regularization parameters	✓ **Correct**
F. Decrease the regularization parameters	

50. You are building a machine learning model to predict the number of lightning strikes during a storm. Your model has thousands of input features to train on. You want to improve the training speed of the model by removing features, but do not want to negatively effect your model's accuracy. What action should you take?

A. Combine highly co-dependent and redundant features into one representative feature.	✓ **Correct**
B. Implement L2 regularization to automatically 'prune' unneeded features	
C. Remove the features that have null values for the majority of your records.	
D. Remove features that have high correlation to your output labels.	