(https://linuxacademy.com/cp)

90 ⬢ 2 🔔 Support ?(https://support.linuxacademy.com/hc/en-us)

shahrohit.1990@gmail.com

☰ Navigation

Data Engineer - Final Exam

🕓 2 hours ★ 50 Questions 🕓 2.4 Minutes per Question [Advanced (/search?type=Practice Exam Challenge&difficulty=Advanced&categories=Google Cloud)](/search?type=Practice Exam Challenge&difficulty=Advanced&categories=Google Cloud)

Go Back Start Challenge

84%

← Go Back

Congratulations!

You passed this challenge on this attempt.

Expectations Report Card

Google Cloud Data Engineer - Final Exam	84%

Exam Breakdown

Google Cloud Data Engineer - Final Exam

INCORRECT

1. 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.

Your Answer: C

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)

Correct Answer: D

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)

---

2. You are monitoring a streaming data pipeline that ingests streaming data into Cloud Pub/Sub, processed by Cloud Dataflow, and inserted into a BigQuery table. Your Pub/Sub topic has a substantially higher than acceptable number of undelivered messages. Choose two reasons why this may be happening.

A Your Publishers' message throughput is too low.

B The subscriber is not acknowledging messages as they are pulled.

C Your Dataflow subscriber is unable to keep up with the rate of incoming messages.

D Your Audit Logs do not have sufficient access to your Pub/Sub topic, causing delays in delivering.

Correct Answer: B

Why is this correct?

If your subscriber is not acknowledging pulled messages, Pub/Sub will not know to drop them from the queue.

Correct Answer: C

Why is this correct?

Your subscriber may be under provisioned to keep up with the compute needs necessary to keep the Pub/Sub backlog empty.

---

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.

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 Answer: D

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)

INCORRECT

4. 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

D Kubernetes Engine, Cloud Dataflow, Cloud Datastore

Your Answer: A

Why is this incorrect?

None of these services are ideal for the use case.

Correct Answer: C

Why is this correct?

Pub/Sub for streaming data ingest, Dataflow for processing streaming data, and BigQuery for storage and analysis.

---

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 Answer: D

Why is this correct?

BigQuery is able to store JSON formatted tables natively. It is the best choice for SQL style queries. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208)

---

6. 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.

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.

Correct Answer: A

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)

INCORRECT

7. 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.

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: A

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)

Correct Answer: C

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)

---

8. 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 Answer: D

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)

---

9. 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.

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.

Correct Answer: A

Why is this correct?

Combining co-dependent and redundant features allows you to reduce the total number of features trained without sacrificing accuracy.

---

10. In a Dataflow processing pipeline, which concept describes timestamps attached to incoming messages?

A Watermark

B ParDo

C PCollection

D Trigger

Correct Answer: A

Why is this correct?

Watermark describes the event time, which is what a timestamp designates.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/3/module/208)

---

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

A Leading with a non-reversed timestamp.

B Standard domain names (non-reversed).

C Reverse timestamps.

D Non-sequential numeric IDs.

Correct Answer: A

Why is this correct?

Like sequential IDs, timestamps will read and write from the same node, causing increased load.
https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/4/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/4/module/208)

Correct Answer: B

Why is this correct?

Non-reversed domain names at the start of a row key can lead to hotspotting. If you need to use domain names, reverse it. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/4/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/4/module/208)

---

12. 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.

B Apply custom encryption to the data before it goes through the pipeline.

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.

Correct Answer: A

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)

---

13. 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. Your current workflows for this task require using services that support Apache Spark and the Hadoop ecosystem. 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 BigQuery for storage. Use Cloud Dataflow to run the transformations.

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

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

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

Correct Answer: C

Why is this correct?

Cloud Dataflow does not support Apache Spark, but Cloud Dataproc does.
https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208)

---

14. 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

C Reduce the number of training examples

D Increase the number of training examples

E Increase the regularization parameters

F Decrease the regularization parameters

Correct Answer: B

Why is this correct?

Reducing the number of unneeded features can help reduce overfitting.

Correct Answer: D

Why is this correct?

More data is one of the best methods to increase the variety of samples and better generalize your model.

Correct Answer: E

Why is this correct?

Increasing your regularization parameters allows you to reduce 'noise' in your model to reduce overfitting.

---

15. You are developing an application on Google Cloud that will label famous landmarks in users' photos. You are under competitive pressure to develop the predictive model quickly. You need to keep service costs low. What should you do? 👍 👎

A Build and train a classification model with TensorFlow. Deploy the model using the Cloud Machine Learning Engine. Inspect the generated MID values to supply the image labels.

B Build an application that calls the Cloud Vision API. Pass client image locations as base64-encoded strings.

C Build an application that calls the Cloud Vision API. Inspect the generated MID values to supply the image labels.

D Build and train a classification model with TensorFlow. Deploy the model using the Cloud Machine Learning Engine. Pass client image locations as base64- encoded strings.

Correct Answer: B

Why is this correct?

Cloud Vision API supports the ability to generate landmark labels from photos. You would want to pass along the images as base64 encoded strings, not MID. https://linuxacademy.com/cp/courses/lesson/course/2248/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2248/lesson/1/module/208)

---

16. 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.

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 Answer: B

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)

Correct Answer: D

Why is this correct?

This is correct. Batch predictions are used for larger loads and more complex models.
https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

17. 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 Answer: D

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)

18. 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.

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 Answer: D

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)

---

19. 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.

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.

Correct Answer: B

Why is this correct?

The Stackdriver Logging agent requires the fluentd plugin to be configured to read logs from your database application.

---

20. What is the recommended minimum amount of data to store in Bigtable? 👍 👎

A 500 GB

B 1 GB

C 1 TB

D 500 TB

Correct Answer: C

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)

INCORRECT

21. 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

C Master

D Parameter server

Your Answer: A

Why is this incorrect?

This is not one of the scale tier machine types. https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

Correct Answer: B

Why is this correct?

You can have multiple workers, which divide up the work of training the model. https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

Correct Answer: C

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)

Correct Answer: D

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)

---

22. Your organization has just recently started using Google Cloud. Everyone in the company has access to all datasets in BigQuery, using it as they see fit without documenting their use cases. You need to implement a formal security policy, but need to first determine what everyone has been doing in BigQuery. What is your first step to do so?

A Use Stackdriver Logging to review data access.

B View the usage of BigQuery query slots in Stackdriver Monitoring.

C Export billing into Cloud Storage, and view BigQuery related records to determine user activity.

D Inspect the IAM policy of each table.

Correct Answer: A

Why is this correct?

Stackdriver Logging will record the audit logs of jobs and queries of each individual user's actions.

---

23. Which of these statements is true regarding BigQuery caching? 👍 👎

A The BigQuery cache only lasts for 48 hours.

B Multiple users can use the same cached query.

C Cache is not enabled by default.

D Queries that retrieve results from the cache have no charge.

Correct Answer: D

Why is this correct?

Cached result have no charge. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/2/module/208)

---

24. 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 Answer: D

Why is this correct?

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

INCORRECT

25. 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.

C Use the snapshots to create replacement instances as needed.

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

Your Answer: A

Why is this incorrect?

Exporting an image to Cloud Storage is considered best practices for disaster recovery scenarios, but not for the above requirements. It is also more steps than necessary.

Correct Answer: C

Why is this correct?

Snapshots let you recreate instances in the fewest steps.

---

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

A Element size in bytes.

B Element count.

C Combinations of other triggers.

D Timestamp

Correct Answer: A

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)

---

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

A Apache Beam

B Apache Kafka

C HBase

D Apache Hadoop

Correct Answer: B

Why is this correct?

Kafka is the open source streaming ingest framework for creating a manual streaming pipeline.
https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208)

---

28. 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

B Cloud Bigtable

C Cloud Pub/Sub

D Cloud Dataflow

Correct Answer: A

Why is this correct?

Dataproc is a managed version of the entire Hadoop ecosystem and would be the best choice.
https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208)

---

29. Regarding Cloud Pub/Sub, which resource locations can have access controlled via IAM roles? (Choose all that apply)

A Topics

B Publisher

C Project-wide predefined roles

D Subscription

Correct Answer: A

Why is this correct?

You can control access by topics. https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208)

Correct Answer: C

Why is this correct?

You can apply project-wide predefined roles for all Pub/Sub components for a project. https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208)

Correct Answer: D

Why is this correct?

You can assign separate IAM roles by subscribers. https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208)

---

30. You are an administrator for several organizations in the same company. 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 think each organization should be able to view and run queries against their own data without exposing the data of organizations to unauthorized viewers. What should you recommend?

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 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.

D 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.

Correct Answer: C

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)

---

31. 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.

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.

Correct Answer: B

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)

---

32. 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 Answer: D

Why is this correct?

This is the preferred method to fulfill the requirement to minimize costs. https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

---

33. 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 Answer: D

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)

---

34. 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

C Regression

D Hyperparameter

Correct Answer: B

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)

---

35. You need to export Avro formatted data from BigQuery into Cloud Storage. What is the best method of doing so from the web console? 👍 👎

A Convert the data to CSV format the BigQuery export options, then make the transfer.

B Use the BigQuery Transfer Service to transfer Avro data to Cloud Storage.

C Click on Export Table in BigQuery, and provide the Cloud Storage location to export to.

D Create a Dataflow job to manage the conversion of Avro data to CSV format, then export to Cloud Storage.

Correct Answer: C

Why is this correct?

BigQuery can export Avro data natively to Cloud Storage.
https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/3/module/208)

INCORRECT

36. You are using a Compute Engine instance to manage your Cloud Dataflow processing workloads. What IAM role do you need to grant to the instance so that it has the necessary access?

A Dataflow Viewer

B Dataflow Developer

C Dataflow Worker

D Dataflow Computer

Your Answer: B

Why is this incorrect?

Service accounts need the Dataflow Worker role.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208)

Correct Answer: C

Why is this correct?

Dataflow Worker is assigned to the Compute Engine service account for necessary access.
https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2243/lesson/2/module/208)

---

37. 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.

B Deep and wide models are ideal for solving regression problems.

C Wide models are used for memorization. Deep models are for generalization

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

Correct Answer: C

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)

Correct Answer: D

Why is this correct?

Both models combined are good at providing recommendations based on previous selections.
https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2246/lesson/2/module/208)

38. 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.

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.

Correct Answer: B

Why is this correct?

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

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

Question List Show All Answers

‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 ›

A Hive

B Beam

C Pig

D Spark

Correct Answer: A

Why is this correct?

Hive is part of the Hadoop ecosystem, therefore it is part of Dataproc.
https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208)

Correct Answer: C

Why is this correct?

Pig is part of the Hadoop ecosystem, therefore it is part of Dataproc.
https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208)

Correct Answer: D

Why is this correct?

Spark is part of the Hadoop ecosystem, therefore it is part of Dataproc.
https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2237/lesson/1/module/208)

---

40. 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.

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.

Correct Answer: A

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)

INCORRECT

41. 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.

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.

Your Answer: D

Why is this incorrect?

You should not use Cloud Storage or BigQuery for this scenario. Bigtable is the better option.
https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/3/module/208

(https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/3/module/208)

Correct Answer: A

Why is this correct?

You will want to use Bigtable for this 'values over time' scenario. Using tall and narrow tables is the best practice for this use case. https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/3/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2111/lesson/3/module/208)

INCORRECT

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

A Decreased query complexity

B Less storage space used

C Increased query performance

D Reduces the amount of data processed

Your Answer: D

Why is this incorrect?

The amount of data is the same. However, performance is increased by not having to query from multiple tables. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208)

Correct Answer: A

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)

Correct Answer: C

Why is this correct?

Denormalizing data increases performance on denormalized data since all of the data is in a single table instead of relying on JOIN's to combine multiple tables' data. https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2238/lesson/4/module/208)

43. 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.

C Save costs.

D Restrict access to other parties.

Correct Answer: B

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)

Correct Answer: C

Why is this correct?

Training locally does not cost money on the cloud when running.
https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208
(https://linuxacademy.com/cp/courses/lesson/course/2247/lesson/1/module/208)

---

44. 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

C Stop

D Drain

Correct Answer: B

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)

---

45. 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

C Import data from BigQuery directly from Cloud SQL.

D Use the BigQuery export function in Cloud SQL to manage exporting data into BigQuery.

Correct Answer: B

Why is this correct?

You cannot import data into BigQuery directly from Cloud SQL. You need to export your data to a Cloud Storage bucket first. https://linuxacademy.com/cp/courses/lesson/course/2109/lesson/4/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2109/lesson/4/module/208)

---

46. 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

C Cached

D Instant

Correct Answer: B

Why is this correct?

Push deliver has Pub/Sub initiate the transfer of messages to the subscriber, and has overall better performance. Be aware that push delivery has more requirements than pull. https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208 (https://linuxacademy.com/cp/courses/lesson/course/2241/lesson/2/module/208)

---

47. 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

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.

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

Correct Answer: A

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)

---

48. 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.

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 Answer: D

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)

---

49. 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.

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.

Correct Answer: B

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)

---

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

A Biases

B Weights

C Epochs

D Features

Correct Answer: A

Why is this correct?

Biases are a parameter that is adjusted 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)

Correct Answer: B

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)