

Data Engineer Practice Exam

The Data Engineer practice exam will familiarize you with the format, level, and scope of questions you may encounter on the certification exam and help you determine your readiness or if you need more preparation and/or experience. Successful completion of the practice exam does not guarantee that you will pass the certification exam as the actual exam is longer and covers a wider range of topics.

Untitled Title



Do you have a candidate ID? *

If you completed a Google Cloud certification exam, you received a notification email with your candidate ID.

☐ Yes

☒ No

Enter your candidate ID *

.....



Exam Registration

Exam Registration

First Name *

Abhinav

Last Name *

Sodhani

Primary Email *

absodhani@deloitte.ca

Recovery Email

absodhani@deloitte.ca

Organization (Employer or School) *

Deloitte

Organization email (an email associated with your current organization)

absodhani@deloitte.ca



Country *

Canada



Primary Relationship to Google *

Customer



Send me offers, updates and useful tips for getting the most out of Google Cloud training and certification products and services.

*

No



✗ 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 BigQuery for storage. Provide format files for data load. Update the format files as needed.
- ☐ B. Use BigQuery for storage. Select "Automatically detect" in the Schema section.
- ☐ C. 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.
- ☒ D. 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. ✗

Correct answer

- ☒ B. Use BigQuery for storage. Select "Automatically detect" in the Schema section.

Feedback

A is not correct because you should not provide format files: you can simply turn on the 'Automatically detect' schema changes flag.

B 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.

C, D are not correct because you should not use Cloud Storage for this scenario: it is cumbersome and doesn't add value.





✗ You use a Hadoop cluster both for serving analytics and for processing and transforming data. The data is currently stored on HDFS in Parquet format. The data processing jobs run for 6 hours each night. Analytics users can access the system 24 hours a day. Phase 1 is to quickly migrate the entire Hadoop environment without a major re-architecture. Phase 2 will include migrating to BigQuery for analytics and to Cloud Dataflow for data processing. You want to make the future migration to BigQuery and Cloud Dataflow easier by following Google-recommended practices and managed services. What should you do?

- ☐ A. Lift and shift Hadoop/HDFS to Cloud Dataproc.
- ☐ B. Lift and shift Hadoop/HDFS to Compute Engine.
- ☒ C. Create a single Cloud Dataproc cluster to support both analytics and data processing, and point it at a Cloud Storage bucket that contains the Parquet files that were previously stored on HDFS. ✗
- ☐ D. Create separate Cloud Dataproc clusters to support analytics and data processing, and point both at the same Cloud Storage bucket that contains the Parquet files that were previously stored on HDFS.

Correct answer

- ☒ D. Create separate Cloud Dataproc clusters to support analytics and data processing, and point both at the same Cloud Storage bucket that contains the Parquet files that were previously stored on HDFS.

Feedback

A is not correct because it is not recommended to attach persistent HDFS to Cloud Dataproc clusters in GCP. (see references link)



B Is not correct because they want to leverage managed services which would mean Cloud Dataproc.

C is not correct because it is recommended that Cloud Dataproc clusters be job specific.

D Is correct because it leverages a managed service (Cloud Dataproc), the data is stored on GCS in Parquet format which can easily be loaded into BigQuery in the future and the Cloud Dataproc clusters are job specific.

 <https://cloud.google.com/sol...>

✓ You are building a new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

- ☐ A. Include ORDER BY DESC on timestamp column and LIMIT to 1.
- ☐ B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
- ☐ C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
- ☒ D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1. ✓

Feedback

A is not correct because this will just return one row.

B is not correct because this doesn't get you the latest value, but will get you a sum of the same event over time which doesn't make too much sense if you have duplicates.

C is not correct because if you have events that are not duplicated. it will be excluded.



D is correct because it will just pick out a single row for each set of duplicates.

✓ You are designing a streaming pipeline for ingesting player interaction data for a mobile game. You want the pipeline to handle out-of-order data delayed up to 15 minutes on a per-player basis and exponential growth in global users. What should you do?

- ☒ A. Design a Cloud Dataflow streaming pipeline with session windowing and a minimum gap duration of 15 minutes. Use "individual player" as the key. Use Cloud Pub/Sub as a message bus for ingestion. ✓
- ☐ B. Design a Cloud Dataflow streaming pipeline with session windowing and a minimum gap duration of 15 minutes. Use "individual player" as the key. Use Apache Kafka as a message bus for ingestion.
- ☐ C. Design a Cloud Dataflow streaming pipeline with a single global window of 15 minutes. Use Cloud Pub/Sub as a message bus for ingestion.
- ☐ D. Design a Cloud Dataflow streaming pipeline with a single global window of 15 minutes. Use Apache Kafka as a message bus for ingestion.

Feedback

A is correct because the question requires delay be handled on a per-player basis and session windowing will do that. PubSub handles the need to scale exponentially with traffic coming from around the globe.

B is not correct because Apache Kafka will not be able to handle an exponential growth in users globally as well as PubSub.

C is not correct because a global window does not meet the requirements of handling out-of-order delay on a per-player basis.

D is not correct because a global window does not meet the requirements of handling



out-of-order delay on a per-player basis.

<https://cloud.google.com/pu...>

<https://beam.apache.org/doc...>

✗ Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

- ☐ A. The CSV data loaded in BigQuery is not flagged as CSV.
- ☒ B. The CSV data had invalid rows that were skipped on import. ✗
- ☐ C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.
- ☐ D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Correct answer

- ☒ C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.

Feedback

A is not correct because if another data format other than CSV was selected then the data would not import successfully.

B is not correct because the data was fully imported meaning no rows were skipped.

C is correct because this is the only situation that would cause successful import.

D is not correct because whether the data has been previously transformed will not affect whether the source file will match the BigQuery table.



<https://cloud.google.com/big...>

✓ Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

- ☐ A. Create a Google Cloud Dataflow job to process the data.
- ☐ B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
- ☐ C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
- ☒ D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector. ✓
- ☐ E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Feedback

A is not correct because the goal is to re-use their Hadoop jobs and MapReduce and/or Spark jobs cannot simply be moved to Dataflow.

B is not correct because the goal is to persist the data beyond the life of the ephemeral clusters, and if HDFS is used as the primary attached storage mechanism, it will also disappear at the end of the cluster's life.

C is not correct because the goal is to use managed services as much as possible, and this is the opposite.

D is correct because it uses managed services, and also allows for the data to persist on GCS beyond the life of the cluster.

E is not correct because the goal is to use managed services as much as possible, and this is the opposite.



✓ You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible. What should you do?

- ☐ A. Load the data every 30 minutes into a new partitioned table in BigQuery.
- ☒ B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery. ✓
- ☐ C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore.
- ☐ D. Store the data in a file in a regional Google Cloud Storage bucket. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Feedback

A is not correct because it is not the cheapest way of accomplishing this task.

B is correct because regional storage is cheaper than BigQuery storage.

C is not correct because it is not the least expensive option. Using Dataflow to query BigQuery adds unnecessary cost to the deployment and will cost more than using BigQuery natively.

D is not correct because it is not the least expensive option. Using Dataflow to query BigQuery adds unnecessary cost to the deployment and will cost more than using BigQuery natively.



✓ 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 displays device outlier data based on your business requirements.
- ☐ B. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.
- ☒ C. Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements. ✓
- ☐ D. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data based on your business requirements.

Feedback

A & B are not correct because you do not need to use BigQuery for the query pattern in this scenario.

C is correct because the data type, volume, and query pattern best fits BigTable capabilities and also Google best practices as linked below.

D is not correct because you can use the simpler method of 'cbt tool' to support this scenario.

[https://cloud.google.com/big...](https://cloud.google.com/bigtable/docs/quickstart-cbt)

[https://cloud.google.com/big...](https://cloud.google.com/bigtable/docs/quickstart-cbt)



✓ You are selecting a messaging 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 current status. You will be storing the data in a searchable repository. How should you set up the input messages?

- ☒ A. Use Cloud Pub/Sub 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 timestamp to every message in the publisher.
- ☐ D. Use Apache Kafka on Compute Engine for input. Attach a unique identifier to every message in the publisher.

Feedback

A is correct because of recommended Google practices; see the links below.

B is not correct because you should not attach a GUID to each message to support the scenario.

C & D are not correct because you should not use Apache Kafka for this scenario (it is overly complex compared to using Cloud Pub/Sub, which can support all of the requirements).

<https://cloud.google.com/pu...>

<http://www.jesse-anderson.c...>



✓ You want to publish system metrics to Google Cloud from a large number of on-prem hypervisors and VMs for analysis and creation of dashboards. You have an existing custom monitoring agent deployed to all the hypervisors and your on-prem metrics system is unable to handle the load. You want to design a system that can collect and store metrics at scale. You don't want to manage your own time series database. Metrics from all agents should be written to the same table but agents must not have permission to modify or read data written by other agents. What should you do?

- ☒ A. Modify the monitoring agent to publish protobuf messages to Cloud PubSub. Use a Dataproc cluster or Dataflow job to consume messages from Pubsub and write to BigTable. ✓
- ☐ B. Modify the monitoring agent to write protobuf messages directly to BigTable.
- ☐ C. Modify the monitoring agent to write protobuf messages to HBase deployed on GCE VM Instances
- ☐ D. Modify the monitoring agent to write protobuf messages to Cloud Pubsub. Use a Dataproc cluster or Dataflow job to consume messages from Pubsub and write to Cassandra deployed on GCE VM Instances.

Feedback

A Is correct because Bigtable can store and analyze time series data, and the solution is using managed services which is what the requirements are calling for.

B Is not correct because BigTable cannot limit access to specific tables.

C is not correct because it requires deployment of an HBase cluster

D is not correct because it requires deployment of an Cassandra cluster



✗ 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 intend to use ANSI SQL to run queries for your analysts. How should you transform the input data?

- ☐ A. Use BigQuery for storage. Use Cloud Dataflow to run the transformations.
- ☐ B. Use BigQuery for storage. Use Cloud Dataproc to run the transformations.
- ☐ C. Use Cloud Storage for storage. Use Cloud Dataflow to run the transformations.
- ☒ D. Use Cloud Storage for storage. Use Cloud Dataproc to run the transformations. ✗

Correct answer

- ☒ B. Use BigQuery for storage. Use Cloud Dataproc to run the transformations.

Feedback

A is not correct because Cloud Dataflow does not support Spark.

B is correct because of the requirement to use custom Spark transforms; use Cloud Dataproc. ANSI SQL queries require the use of BigQuery.

C & D are not correct because Cloud Storage does not support SQL, and you should not use Cloud Dataflow, either.

 <https://stackoverflow.com/qu...>



✓ 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 Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.
- ☒ B. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span. ✓
- ☐ C. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.
- ☐ D. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

Feedback

A is not correct because you should not use storage utilization as a scaling metric.

B is correct because of the requirement to globally scalable transactions—use Cloud Spanner. CPU utilization is the recommended metric for scaling, per Google best practices, linked below.

C & D are not correct because you should not use Cloud Bigtable for this scenario.

<https://cloud.google.com/sp...>

<https://cloud.google.com/big...>



✗ You have a Spark application that writes data to Cloud Storage in Parquet format. You scheduled the application to run daily using DataProcSparkOperator and Apache Airflow DAG by Cloud Composer. You want to add tasks to the DAG to make the data available to BigQuery users. You want to maximize query speed and configure partitioning and clustering on the table. What should you do?

- ☐ A. Use "BashOperator" to call "bq insert".
- ☐ B. Use "BashOperator" to call "bq cp" with the "--append" flag.
- ☐ C. Use "GoogleCloudStorageToBigQueryOperator" with "schema_object" pointing to a schema JSON in Cloud Storage and "source_format" set to "PARQUET".
- ☒ D. Use "BigQueryCreateExternalTableOperator" with "schema_object" pointing to a schema JSON in Cloud Storage and "source_format" set to "PARQUET". ✗

Correct answer

- ☒ C. Use "GoogleCloudStorageToBigQueryOperator" with "schema_object" pointing to a schema JSON in Cloud Storage and "source_format" set to "PARQUET".

Feedback




A is not correct because bq insert will not set the partitioning and clustering and only supports JSON.

B is not correct because bq cp is for existing BigQuery tables only.

C is correct because it loads the data and sets partitioning and clustering.

D is not correct because an external table will not satisfy the query speed requirement.



 <https://cloud.google.com/big...> <https://cloud.google.com/big...> <https://airflow.incubator.apac...> <https://airflow.incubator.apac...> <https://cloud.google.com/big...> <https://cloud.google.com/big...>

✗ You have a website that tracks page visits for each user and then creates a Cloud Pub/Sub message with the session ID and URL of the page. You want to create a Cloud Dataflow pipeline that sums the total number of pages visited by each user and writes the result to BigQuery. User sessions timeout after 30 minutes. Which type of Cloud Dataflow window should you choose?

- ☐ A. A single global window
- ☐ B. Fixed-time windows with a duration of 30 minutes
- ☐ C. Session-based windows with a gap duration of 30 minutes
- ☒ D. Sliding-time windows with a duration of 30 minutes and a new window every 5 minute ✗

Correct answer

- ☒ C. Session-based windows with a gap duration of 30 minutes

Feedback

A is incorrect because a user-specific sum is never calculated, it just sums for arbitrary 30-min windows of time staggered by 5 minutes.

B is incorrect because there is no per-user metric being used so it's possible a sum will be created for some users while they are still browsing the site.

C is correct because it continues to sum user page visits during their browsing session and completes at the same time as the session timeout.

D is incorrect because if a user is still visiting the site when the 30-min window closes, the sum will be wrong.



<https://cloud.google.com/dat...>

✓ You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules: a). No interaction by the user on the site for 1 hour b). Has added more than \$30 worth of products to the basket c). Has not completed a transaction. You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

- ☐ A. Use a fixed-time window with a duration of 60 minutes.
- ☐ B. Use a sliding time window with a duration of 60 minutes.
- ☒ C. Use a session window with a gap time duration of 60 minutes. ✓
- ☐ D. Use a global window with a time based trigger with a delay of 60 minutes.

Feedback

A is not correct because assuming there is one key per user, a message will be sent every 60 minutes.

B is not correct because assuming there is one key per user, a message will be sent 60 minutes after they first started browsing even if they are still browsing.

C is correct because it will send a message per user after that user is inactive for 60 minutes.

D is not correct because it will cause messages to be sent out every 60 minutes to all users regardless of where they are in their current session.

<https://beam.apache.org/doc...>



✓ You need to stream time-series data in Avro format, and then write this to both BigQuery and Cloud Bigtable simultaneously using Cloud Dataflow. You want to achieve minimal end-to-end latency. Your business requirements state this needs to be completed as quickly as possible. What should you do?

- ☐ Create a pipeline and use ParDo transform.
- ☐ Create a pipeline that groups the data into a PCollection and uses the Combine transform.
- ☒ Create a pipeline that groups data using a PCollection and then uses Cloud Bigtable and BigQueryIO transforms. ✓
- ☐ Create a pipeline that groups data using a PCollection, and then use Avro I/O transform to write to Cloud Storage. After the data is written, load the data from Cloud Storage into BigQuery and Cloud Bigtable.

Feedback

A is not correct because ParDo doesn't write to BigQuery or BigTable.

B is not correct because Combine doesn't write to BigQuery or Bigtable.

C is correct because this is the right set of transformations that accepts and writes to the required data stores.

D is not correct because to meet the business requirements, it is much faster and easier using Dataflow (answer C).

<https://cloud.google.com/blo...>



✓ Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

- ☒ A. Put the data into Google Cloud Storage. ✓
- ☐ B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
- ☐ C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
- ☐ D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.


Feedback

A is correct because Google recommends using Google Cloud Storage instead of HDFS as it is much more cost effective especially when jobs aren't running.

B is not correct because this will decrease the compute cost but not the storage cost.

C is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Google Cloud Storage.

D is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Google Cloud Storage.

 <https://cloud.google.com/dat...>



✓ You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on non-key columns. What should you do?

- ☐ A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
- ☐ B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
- ☒ C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns. ✓
- ☐ D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Feedback

A is not correct because Cloud SQL does not natively scale horizontally.

B is not correct because Cloud SQL does not natively scale horizontally.

C is correct because Cloud Spanner scales horizontally, and you can create secondary indexes for the range queries that are required.

D is not correct because Cloud Dataflow is a data pipelining tool to move and transform data, but the use case is centered around querying.



✓ Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

- ☐ A. Use a row key of the form <timestamp>.
- ☐ B. Use a row key of the form <sensorid>.
- ☐ C. Use a row key of the form <timestamp>#<sensorid>.
- ☒ D. Use a row key of the form <sensorid>#<timestamp>.



Feedback


A is not correct because this will cause most writes to be pushed to a single node (known as hotspotting)

B is not correct because this will not allow for multiple readings from the same sensor as new readings will overwrite old ones.

C is not correct because this will cause most writes to be pushed to a single node (known as hotspotting)

D is correct because it will allow for retrieval of data based on both sensor id and timestamp but without causing hotspotting.

 <https://cloud.google.com/big...>

 <https://cloud.google.com/big...>



✓ You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

- ☒ A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels. ✓
- ☐ B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.
- ☐ C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.
- ☐ D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Feedback

A is correct because it provides a managed service and a fully trained model, and the user is pulling the entities, which is the right label.

B is not correct because sentiment is the incorrect label for this use case.

C is not correct because this requires experience with machine learning.

D is not correct because this requires experience with machine learning.



- ✓ Your company is using WILDCARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the error shown below. Which table name will make the SQL statement work correctly?

```
# Syntax error: Expected end of statement but got "-" at [4:11]
SELECT age
FROM
  bigquery-public-data.noaa_gsod.gsod
WHERE
  age != 99
  AND _TABLE_SUFFIX = '1929'
ORDER BY
  age DESC
```

- ☐ A. `bigquery-public-data.noaa_gsod.gsod`
- ☐ B. bigquery-public-data.noaa_gsod.gsod*
- ☐ C. 'bigquery-public-data.noaa_gsod.gsod*'
- ☒ D. `bigquery-public-data.noaa_gsod.gsod*` ✓


Feedback

A is not correct because this is not the correct wildcard syntax as there is no wildcard character present.

B is not correct because this is not the correct wildcard syntax since it's missing backticks.

C is not correct because this is not the correct wildcard syntax since it's not using a backtick as the last character

*D is correct because it follows the correct wildcard syntax of enclosing the table name in backticks and including the * wildcard character.*

 <https://cloud.google.com/big...>



✗ You are working on an ML-based application that will transcribe conversations between manufacturing workers. These conversations are in English and between 30-40 sec long. Conversation recordings come from old enterprise radio sets that have a low sampling rate of 8000 Hz, but you have a large dataset of these recorded conversations with their transcriptions. You want to follow Google-recommended practices. How should you proceed with building your application?

- ☐ A. Use Cloud Speech-to-Text API, and send requests in a synchronous mode.
- ☐ B. Use Cloud Speech-to-Text API, and send requests in an asynchronous mode.
- ☒ C. Use Cloud Speech-to-Text API, but resample your captured recordings to a rate of 16000 Hz. ✗
- ☐ D. Train your own speech recognition model because you have an uncommon use case and you have a labeled dataset.

Correct answer

- ☒ A. Use Cloud Speech-to-Text API, and send requests in a synchronous mode.

Feedback

A is correct because synchronous mode is recommended for short audio files.


B is incorrect since the recommended way to process short audio files (shorter than 1 minutes) is a synchronous recognize request and not an asynchronous one.


C is incorrect since using the native sample rate is recommended over resampling.

D is incorrect since there is nothing in the question that suggests the off-the-shelf



is incorrect since there is nothing in the question that suggests the on the other model will not perform sufficiently.

 <https://cloud.google.com/sp...>

 <https://cloud.google.com/sp...>

 <https://cloud.google.com/sp...>



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

- ☒ A. Build an application that calls the Cloud Vision API. 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 and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Pass client image locations as base64-encoded strings.
- ☐ D. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Inspect the generated MID values to supply the image labels.

Correct answer

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

Feedback

A is not correct because you should not inspect the generated MID values; instead, you should simply pass the image locations to the API and use the labels, which are output.

B is correct because of the requirement to quickly develop a model that generates landmark labels from photos. This is supported in Cloud Vision API; see the link below.

C, D are not correct because you should not build a custom classification TF model for this scenario.

 <https://cloud.google.com/vision/>



✓ 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 Cloud Machine Learning to host your model. Monitor the status of the Operation object for 'error' results.
- ☒ B. Use Cloud Machine Learning to host your model. Monitor the status of the Jobs object for 'failed' job states. ✓
- ☐ 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 Operation object for 'error' results.

Feedback

A is not correct because you should not use the Operation object to monitor failures.

B is correct because of the requirement to host an ML DNN and Google-recommended monitoring object (Jobs); see the links below.

C, D are not correct because you should not use a Kubernetes Engine cluster for this scenario.

<https://cloud.google.com/ml-...>

<https://cloud.google.com/ku...>



✗ You work on a regression problem in a natural language processing domain, and you have 100M labeled examples in your dataset. You have randomly shuffled your data and split your dataset into training and test samples (in a 90/10 ratio). After you have trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

- ☒ A. Increase the share of the test sample in the train-test split. ✗
- ☐ B. Try to collect more data and increase the size of your dataset.
- ☐ C. Try out regularization techniques (e.g., dropout or batch normalization) to avoid overfitting.
- ☐ D. Increase the complexity of your model by, e.g., introducing an additional layer or increasing the size of vocabularies or n-grams used to avoid underfitting.

Correct answer

- ☒ D. Increase the complexity of your model by, e.g., introducing an additional layer or increasing the size of vocabularies or n-grams used to avoid underfitting.

Feedback

A is incorrect since test sample is large enough.


C is incorrect since regularization helps to avoid overfitting and we have a clear underfitting case.


B is incorrect since dataset is pretty large already, and having more data typically helps with overfitting and not with underfitting.



D is correct since increasing model complexity generally helps when you have an underfitting problem.

 <https://developers.google.co...>

 <https://towardsdatascience.c...>

 <https://developers.google.co...>

 <https://developers.google.co...>



✓ You are using Cloud Pub/Sub to stream inventory updates from many point-of-sale (POS) terminals into BigQuery. Each update event has the following information: product identifier "prodSku", change increment "quantityDelta", POS identification "termId", and "messageId" which is created for each push attempt from the terminal. During a network outage, you discovered that duplicated messages were sent, causing the inventory system to over-count the changes. You determine that the terminal application has design problems and may send the same event more than once during push retries. You want to ensure that the inventory update is accurate. What should you do?

- ☐ A. Inspect the "publishTime" of each message. Make sure that messages whose "publishTime" values match rows in the BigQuery table are discarded.
- ☐ B. Inspect the "messageId" of each message. Make sure that any messages whose "messageId" values match corresponding rows in the BigQuery table are discarded.
- ☐ C. Instead of specifying a change increment for "quantityDelta", always use the derived inventory value after the increment has been applied. Name the new attribute "adjustedQuantity".
- ☒ D. Add another attribute orderId to the message payload to mark the unique check-out order across all terminals. Make sure that messages whose "orderId" and "prodSku" values match corresponding rows in the BigQuery table are discarded. ✓

Feedback

A is not correct because publishTime cannot uniquely identify a message and it does not address push retries.


B is not correct because duplication in this case could be caused by a terminal re-try, in which case messageId could be different for the same event.



C is not correct because there are many terminals. Calculating the projected inventory values on the terminal introduces a race condition where multiple terminals could update the inventory data simultaneously.

D is correct because the client application must include a unique identifier to disambiguate possible duplicates due to push retries.

 <https://cloud.google.com/pu...>

 <https://cloud.google.com/pu...>



✓ You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database table must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

- ☐ A. Add capacity (memory and disk space) to the database server by the order of 200.
- ☐ B. Shard the tables into smaller ones based on date ranges, and only generate reports with pre-specified date ranges.
- ☒ C. Normalize the master patient-record table into the patients table and the visits table, and create other necessary tables to avoid self-join. ✓
- ☐ D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Feedback

A is not correct because adding additional compute resources is not a recommended way to resolve database schema problems.

B is not correct because this will reduce the functionality of the database and make running reports more difficult.

C is correct because this option provides the least amount of inconvenience over using pre-specified date ranges or one table per clinic while also increasing performance due to avoiding self-joins.

D is not correct because this will likely increase the number of tables so much that it



will be more difficult to generate reports vs. the correct option.

 <https://cloud.google.com/big...>

 <https://cloud.google.com/big...>



✗ Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have the freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

- ☐ A. Use Google Stackdriver Audit Logs to review data access.
- ☐ B. Get the identity and access management (IAM) policy of each table.
- ☒ C. Use Stackdriver Monitoring to see the usage of BigQuery query slots. ✗
- ☐ D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Correct answer

- ☒ A. Use Google Stackdriver Audit Logs to review data access.

Feedback

A is correct because this is the best way to get granular access to data showing which users are accessing which data.

B is not correct because we already know that all users already have access to all data, so this information is unlikely to be useful. It will also not show what users have done, just what they can do.

C is not correct because slot usage will not inform security policy.

D is not correct because a billing account is typically shared among many people and will only show the amount of data queried and stored.



✓ 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. 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.
- ☐ B. 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.
- ☐ C. 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.
- ☒ 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. ✓

Feedback

A is not correct because the scenario states that you must supply your own encryption key instead of using one generated by Google Cloud Platform.

B is not correct because the scenario states that you should use, but not store, your own key with Google Cloud Platform services.

C is not correct because it does not meet the scenario requirement to reference, but not store, your own key with Google Cloud Platform services.

D is correct because the scenario requires you to use your own key and also to not



B is correct because the scenario requires you to use your own key and also to not store your key on Compute Engine, and also this is a Google recommended practice.


✓ 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 own 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 members to the default IAM Owner role. Add the developers to a custom role that allows them to see their own spend only.
- ☒ B. 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. ✓
- ☐ C. Add the developers and finance managers to the Viewer role for the Project.
- ☐ D. 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.

Feedback

B is correct because it uses the principle of least privilege for IAM roles; use the Billing Administrator IAM role for that job function.

A, C, & D are not correct because it is a best practice to use pre-defined IAM roles when they exist and match your business scenario; see the link below.

 <https://cloud.google.com/ia...>



This form was created inside of Google.com. [Privacy & Terms](#)

Google Forms

