# Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

**https://www.2passeasy.com/dumps/Professional-Data-Engineer/**

**NEW QUESTION 1**
- (Exam Topic 1)
You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:

- No interaction by the user on the site for 1 hour
- Has added more than $30 worth of products to the basket
- 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.

**Answer:** D


**NEW QUESTION 2**
- (Exam Topic 1)
You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and analyze these very large datasets in real time. What should you do?

A. Send the data to Google Cloud Datastore and then export to BigQuery.
B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

**Answer:** B


**NEW QUESTION 3**
- (Exam Topic 1)
Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

A. Add a node to the MySQL cluster and build an OLAP cube there.
B. Use an ETL tool to load the data from MySQL into Google BigQuery.
C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.
D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

**Answer:** C


**NEW QUESTION 4**
- (Exam Topic 1)
Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks. What should you do?

A. Run a local version of Jupiter on the laptop.
B. Grant the user access to Google Cloud Shell.
C. Host a visualization tool on a VM on Google Compute Engine.
D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.

**Answer:** B


**NEW QUESTION 5**
- (Exam Topic 1)
Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiency?

A. Assign global unique identifiers (GUID) to each data entry.
B. Compute the hash value of each data entry, and compare it with all historical data.
C. Store each data entry as the primary key in a separate database and apply an index.
D. Maintain a database table to store the hash value and other metadata for each data entry.

**Answer:** D


**NEW QUESTION 6**
- (Exam Topic 1)
You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

A. Update the current pipeline and use the drain flag.

B. Update the current pipeline and provide the transform mapping JSON object.
C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

**Answer:** D

## NEW QUESTION 7
- (Exam Topic 1)
You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

A. Eliminate features that are highly correlated to the output labels.
B. Combine highly co-dependent features into one representative feature.
C. Instead of feeding in each feature individually, average their values in batches of 3.
D. Remove the features that have null values for more than 50% of the training records.

**Answer:** B

## NEW QUESTION 8
- (Exam Topic 2)
Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?

A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.
C. Use the NOW () function in BigQuery to record the event's time.
D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

**Answer:** B

## NEW QUESTION 9
- (Exam Topic 4)
You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.
You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

A. Redis
B. HBase
C. MySQL
D. MongoDB
E. Cassandra
F. HDFS with Hive

**Answer:** BDF

## NEW QUESTION 10
- (Exam Topic 4)
Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.
You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (choose two.)

A. Introduce data compression for each file to increase the rate file of file transfer.
B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.
D. Assemble 1,000 files into a tape archive (TAR) fil
E. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.
F. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premices data to the designated storage bucket.

**Answer:** CE

## NEW QUESTION 10
- (Exam Topic 5)
What is the HBase Shell for Cloud Bigtable?

A. The HBase shell is a GUI based interface that performs administrative tasks, such as creating and deleting tables.
B. The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables.
C. The HBase shell is a hypervisor based shell that performs administrative tasks, such as creating and deleting new virtualized instances.
D. The HBase shell is a command-line tool that performs only user account management functions to grant access to Cloud Bigtable instances.

**Answer:** B

**Explanation:**
The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables. The Cloud Bigtable HBase client for Java makes it possible to use the HBase shell to connect to Cloud Bigtable.
Reference: https://cloud.google.com/bigtable/docs/installing-hbase-shell

**NEW QUESTION 11**
- (Exam Topic 5)
When you design a Google Cloud Bigtable schema it is recommended that you .

A. Avoid schema designs that are based on NoSQL concepts
B. Create schema designs that are based on a relational database design
C. Avoid schema designs that require atomicity across rows
D. Create schema designs that require atomicity across rows

**Answer:** C

**Explanation:**
All operations are atomic at the row level. For example, if you update two rows in a table, it's possible that one row will be updated successfully and the other update will fail. Avoid schema designs that require atomicity across rows.
Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

**NEW QUESTION 12**
- (Exam Topic 5)
Why do you need to split a machine learning dataset into training data and test data?

A. So you can try two different sets of features
B. To make sure your model is generalized for more than just the training data
C. To allow you to create unit tests in your code
D. So you can use one dataset for a wide model and one for a deep model

**Answer:** B

**Explanation:**
The flaw with evaluating a predictive model on training data is that it does not inform you on how well the model has generalized to new unseen data. A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely to have lower accuracy on an unseen test dataset. The reason is that the model is not as generalized. It has specialized to the structure in the training dataset. This is called overfitting.
Reference: https://machinelearningmastery.com/a-simple-intuition-for-overfitting/

**NEW QUESTION 13**
- (Exam Topic 5)
If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?

A. Do not use a production instance.
B. Run your test for at least 10 minutes.
C. Before you test, run a heavy pre-test for several minutes.
D. Use at least 300 GB of data.

**Answer:** A

**Explanation:**
If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you
plan and execute your test:
Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.
Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use 100 GB of data per node.
Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.
Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.
Reference: https://cloud.google.com/bigtable/docs/performance

**NEW QUESTION 16**
- (Exam Topic 5)
What Dataflow concept determines when a Window's contents should be output based on certain criteria being met?

A. Sessions
B. OutputCriteria
C. Windows
D. Triggers

**Answer:** D

**Explanation:**
Triggers control when the elements for a specific key and window are output. As elements arrive, they are put into one or more windows by a Window transform and its associated WindowFn, and then passed to the associated Trigger to determine if the Windows contents should be output.
Reference:
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/transforms/windowing/Tri

**NEW QUESTION 18**
- (Exam Topic 5)
Which of the following statements about Legacy SQL and Standard SQL is not true?

A. Standard SQL is the preferred query language for BigQuery.
B. If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
C. One difference between the two query languages is how you specify fully-qualified table names (i.
D. table names that include their associated project name).
E. You need to set a query language for each dataset and the default is Standard SQL.

**Answer:** D

**Explanation:**
You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.
Standard SQL has been the preferred query language since BigQuery 2.0 was released.
In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.
Due to the differences in syntax between the two query languages (such as with project-qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

**NEW QUESTION 22**
- (Exam Topic 5)
Which of the following is NOT one of the three main types of triggers that Dataflow supports?

A. Trigger based on element size in bytes
B. Trigger that is a combination of other triggers
C. Trigger based on element count
D. Trigger based on time

**Answer:** A

**Explanation:**
There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers. You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way
Reference: https://cloud.google.com/dataflow/model/triggers

**NEW QUESTION 23**
- (Exam Topic 5)
Which Google Cloud Platform service is an alternative to Hadoop with Hive?

A. Cloud Dataflow
B. Cloud Bigtable
C. BigQuery
D. Cloud Datastore

**Answer:** C

**Explanation:**
Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.
Google BigQuery is an enterprise data warehouse. Reference: https://en.wikipedia.org/wiki/Apache_Hive

**NEW QUESTION 26**
- (Exam Topic 5)
Which of these statements about exporting data from BigQuery is false?

A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.
B. The only supported export destination is Google Cloud Storage.
C. Data can only be exported in JSON or Avro format.
D. The only compression option available is GZIP.

**Answer:** C

**Explanation:**
Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported.
Reference: https://cloud.google.com/bigquery/docs/exporting-data

**NEW QUESTION 29**
- (Exam Topic 5)
What are two methods that can be used to denormalize tables in BigQuery?

A. 1) Split table into multiple tables; 2) Use a partitioned table
B. 1) Join tables into one table; 2) Use nested repeated fields
C. 1) Use a partitioned table; 2) Join tables into one table
D. 1) Use nested repeated fields; 2) Use a partitioned table

**Answer:** B

**Explanation:**
The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are dealing with sales transactions, you would write each individual fact to a record, along with the accompanying dimensions such as order and customer information. The other method for denormalizing data takes advantage of BigQuery's native support for nested and repeated structures in JSON or Avro input data. Expressing records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which would be represented as nested, repeated elements.
Reference: https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

**NEW QUESTION 31**
- (Exam Topic 5)
Cloud Bigtable is Google's Big Data database service.

A. Relational
B. mySQL
C. NoSQL
D. SQL Server

**Answer:** C

**Explanation:**
Cloud Bigtable is Google's NoSQL Big Data database service. It is the same database that Google uses for services, such as Search, Analytics, Maps, and Gmail. It is used for requirements that are low latency and high throughput including Internet of Things (IoT), user analytics, and financial data analysis.
Reference: https://cloud.google.com/bigtable/

**NEW QUESTION 32**
- (Exam Topic 5)
Which of these sources can you not load data into BigQuery from?

A. File upload
B. Google Drive
C. Google Cloud Storage
D. Google Cloud SQL

**Answer:** D

**Explanation:**
You can load data into BigQuery from a file upload, Google Cloud Storage, Google Drive, or Google Cloud Bigtable. It is not possible to load data into BigQuery directly from Google Cloud SQL. One way to get data from Cloud SQL to BigQuery would be to export data from Cloud SQL to Cloud Storage and then load it from there.
Reference: https://cloud.google.com/bigquery/loading-data

**NEW QUESTION 36**
- (Exam Topic 5)
Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?

A. A sequential numeric ID
B. A timestamp followed by a stock symbol
C. A non-sequential numeric ID
D. A stock symbol followed by a timestamp

**Answer:** AB

**Explanation:**
using a timestamp as the first element of a row key can cause a variety of problems.
In brief, when a row key for a time series includes a timestamp, all of your writes will target a single node; fill that node; and then move onto the next node in the cluster, resulting in hotspotting.
Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, since new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes.
[https://cloud.google.com/bigtable/docs/schema-design]
Reference:
https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotti

**NEW QUESTION 38**
- (Exam Topic 5)
Which of these statements about BigQuery caching is true?

A. By default, a query's results are not cached.
B. BigQuery caches query results for 48 hours.
C. Query results are cached even if you specify a destination table.
D. There is no charge for a query that retrieves its results from cache.

**Answer:** D

**Explanation:**
When query results are retrieved from a cached results table, you are not charged for the query. BigQuery caches query results for 24 hours, not 48 hours.
Query results are not cached if you specify a destination table.
A query's results are always cached except under certain conditions, such as if you specify a destination table. Reference:
https://cloud.google.com/bigquery/querying-data#query-caching

**NEW QUESTION 43**
- (Exam Topic 5)
What is the general recommendation when designing your row keys for a Cloud Bigtable schema?

A. Include multiple time series values within the row key
B. Keep the row keep as an 8 bit integer
C. Keep your row key reasonably short
D. Keep your row key as long as the field permits

**Answer:** C

**Explanation:**
A general guide is to, keep your row keys reasonably short. Long row keys take up additional memory and storage and increase the time it takes to get responses from the Cloud Bigtable server.
Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

**NEW QUESTION 47**
- (Exam Topic 5)
You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output.
Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

A. Cancel
B. Drain
C. Stop
D. Finish

**Answer:** B

**Explanation:**
Using the Drain option to stop your job tells the Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources, but the Dataflow
service will preserve any existing resources (such as worker instances) to finish processing and writing any buffered data in your pipeline.
Reference: https://cloud.google.com/dataflow/pipelines/stopping-a-pipeline

**NEW QUESTION 51**
- (Exam Topic 5)
You want to use a BigQuery table as a data sink. In which writing mode(s) can you use BigQuery as a sink?

A. Both batch and streaming
B. BigQuery cannot be used as a sink
C. Only batch
D. Only streaming

**Answer:** A

**Explanation:**
When you apply a BigQueryIO.Write transform in batch mode to write to a single table, Dataflow invokes a BigQuery load job. When you apply a BigQueryIO.Write transform in streaming mode or in batch mode using a function to specify the destination table, Dataflow uses BigQuery's streaming inserts
Reference: https://cloud.google.com/dataflow/model/bigquery-io

**NEW QUESTION 56**
- (Exam Topic 5)
If a dataset contains rows with individual people and columns for year of birth, country, and income, how many of the columns are continuous and how many are categorical?

A. 1 continuous and 2 categorical
B. 3 categorical
C. 3 continuous
D. 2 continuous and 1 categorical

**Answer:** D

**Explanation:**
The columns can be grouped into two types—categorical and continuous columns:
A column is called categorical if its value can only be one of the categories in a finite set. For example, the native country of a person (U.S., India, Japan, etc.) or the education level (high school, college, etc.) are categorical columns.
A column is called continuous if its value can be any numerical value in a continuous range. For example, the capital gain of a person (e.g. $14,084) is a continuous column.
Year of birth and income are continuous columns. Country is a categorical column.
You could use bucketization to turn year of birth and/or income into categorical features, but the raw columns are continuous.
Reference: https://www.tensorflow.org/tutorials/wide#reading_the_census_data

**NEW QUESTION 60**
- (Exam Topic 5)
When a Cloud Bigtable node fails, is lost.

A. all data
B. no data

C. the last transaction
D. the time dimension

**Answer:** B

**Explanation:**
A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.
Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:
Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud Bigtable simply updates the pointers for each node.
Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.
When a Cloud Bigtable node fails, no data is lost Reference: https://cloud.google.com/bigtable/docs/overview

**NEW QUESTION 63**
- (Exam Topic 5)
The for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

A. Cloud Dataflow connector
B. DataFlow SDK
C. BiqQuery API
D. BigQuery Data Transfer Service

**Answer:** A

**Explanation:**
The Cloud Dataflow connector for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline. You can use the connector for both batch and streaming operations.
Reference: https://cloud.google.com/bigtable/docs/dataflow-hbase

**NEW QUESTION 66**
- (Exam Topic 5)
Cloud Dataproc charges you only for what you really use with billing.

A. month-by-month
B. minute-by-minute
C. week-by-week
D. hour-by-hour

**Answer:** B

**Explanation:**
One of the advantages of Cloud Dataproc is its low cost. Dataproc charges for what you really use with minute-by-minute billing and a low, ten-minute-minimum billing period.
Reference: https://cloud.google.com/dataproc/docs/concepts/overview

**NEW QUESTION 70**
- (Exam Topic 5)
Which action can a Cloud Dataproc Viewer perform?

A. Submit a job.
B. Create a cluster.
C. Delete a cluster.
D. List the jobs.

**Answer:** D

**Explanation:**
A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.
Reference: https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations_summary

**NEW QUESTION 75**
- (Exam Topic 5)
You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

A. PCollection
B. Transform
C. Pipeline
D. Sink API

**Answer:** B

**Explanation:**
In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.
Reference: https://cloud.google.com/dataflow/model/programming-model

**NEW QUESTION 78**
- (Exam Topic 5)
Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

A. The wide model is used for memorization, while the deep model is used for generalization.
B. A good use for the wide and deep model is a recommender system.
C. The wide model is used for generalization, while the deep model is used for memorization.
D. A good use for the wide and deep model is a small-scale linear regression problem.

**Answer:** AB

**Explanation:**
Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.
Reference: https://research.googleblog.com/2016/06/wide-deep-learning-better-together-with.html

**NEW QUESTION 79**
- (Exam Topic 5)
Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?

A. Field promotion
B. Randomization
C. Salting
D. Hashing

**Answer:** A

**Explanation:**
By default, prefer field promotion. Field promotion avoids hotspotting in almost all cases, and it tends to make it easier to design a row key that facilitates queries.
Reference:
https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotti

**NEW QUESTION 81**
- (Exam Topic 5)
What are two of the characteristics of using online prediction rather than batch prediction?

A. It is optimized to handle a high volume of data instances in a job and to run more complex models.
B. Predictions are returned in the response message.
C. Predictions are written to output files in a Cloud Storage location that you specify.
D. It is optimized to minimize the latency of serving predictions.

**Answer:** BD

**Explanation:**
 Online prediction
Optimized to minimize the latency of serving predictions. Predictions returned in the response message.
Batch prediction
Optimized to handle a high volume of instances in a job and to run more complex models. Predictions written to output files in a Cloud Storage location that you specify.
Reference:
https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

**NEW QUESTION 82**
- (Exam Topic 5)
Which SQL keyword can be used to reduce the number of columns processed by BigQuery?

A. BETWEEN
B. WHERE
C. SELECT
D. LIMIT

**Answer:** C

**Explanation:**
SELECT allows you to query specific columns rather than the whole table.
LIMIT, BETWEEN, and WHERE clauses will not reduce the number of columns processed by BigQuery.
Reference:
https://cloud.google.com/bigquery/launch-checklist#architecture_design_and_development_checklist

**NEW QUESTION 84**
- (Exam Topic 5)
Which software libraries are supported by Cloud Machine Learning Engine?

A. Theano and TensorFlow
B. Theano and Torch
C. TensorFlow
D. TensorFlow and Torch

**Answer:** C

**Explanation:**
Cloud ML Engine mainly does two things:
Enables you to train machine learning models at scale by running TensorFlow training applications in the cloud.
Hosts those trained models for you in the cloud so that you can use them to get predictions about new data.
Reference: https://cloud.google.com/ml-engine/docs/technical-overview#what_it_does

**NEW QUESTION 86**
- (Exam Topic 6)
You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app_events_YYYYMMDD. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?

A. Use the TABLE_DATE_RANGE function
B. Use the WHERE_PARTITIONTIME pseudo column
C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
D. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD

**Answer:** A

**Explanation:**
Reference:
https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understandyour-mobile-ap

**NEW QUESTION 89**
- (Exam Topic 6)
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

A. Implement clustering in BigQuery on the ingest date column.
B. Implement clustering in BigQuery on the package-tracking ID column.
C. Tier older data onto Cloud Storage files, and leverage extended tables.
D. Re-create the table using data partitioning on the package delivery date.

**Answer:** A

**NEW QUESTION 91**
- (Exam Topic 6)
You're training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longtitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency.
What should you do?

A. Provide latitude and longtitude as input vectors to your neural net.
B. Create a numeric column from a feature cross of latitude and longtitude.
C. Create a feature cross of latitude and longtitude, bucketize at the minute level and use L1 regularization during optimization.
D. Create a feature cross of latitude and longtitude, bucketize it at the minute level and use L2 regularization during optimization.

**Answer:** B

**Explanation:**
Reference https://cloud.google.com/bigquery/docs/gis-data

**NEW QUESTION 92**
- (Exam Topic 6)
Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

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

**Answer:** B

**NEW QUESTION 96**
- (Exam Topic 6)
You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

A. Cloud Speech-to-Text API
B. Cloud Natural Language API
C. Dialogflow Enterprise Edition
D. Cloud AutoML Natural Language

**Answer:** D

**NEW QUESTION 98**
- (Exam Topic 6)
You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account.
What should you do?

A. Convert your batch BQ queries into interactive BQ queries.
B. Create an additional project to overcome the 2K on-demand per-project quota.
C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.
D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

**Answer:** C

**Explanation:**
Reference https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery

**NEW QUESTION 99**
- (Exam Topic 6)
You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.
B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

**Answer:** C

**NEW QUESTION 104**
- (Exam Topic 6)
You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

A. Transform text files to compressed Avro using Cloud Dataflo
B. Use BigQuery for storage and query.
C. Transform text files to compressed Avro using Cloud Dataflo
D. Use Cloud Storage and BigQuerypermanent linked tables for query.
E. Compress text files to gzip using the Grid Computing Tool
F. Use BigQuery for storage and query.
G. Compress text files to gzip using the Grid Computing Tool
H. Use Cloud Storage, and then import into Cloud Bigtable for query.

**Answer:** D

**NEW QUESTION 105**
- (Exam Topic 6)
You need to create a new transaction table in Cloud Spanner that stores product sales data. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?

A. The current epoch time
B. A concatenation of the product name and the current epoch time
C. A random universally unique identifier number (version 4 UUID)
D. The original order identification number from the sales system, which is a monotonically increasing integer

**Answer:** C

**NEW QUESTION 107**
- (Exam Topic 6)
You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing operations and initial data are parquet files (on average
200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.
What should you do?

A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
B. Switch to TFRecords formats (app
C. 200MB per file) instead of parquet files.
D. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
E. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

**Answer:** C

**NEW QUESTION 109**
- (Exam Topic 6)
Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000

records per second. Many third parties use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:

- Single global endpoint
- ANSI SQL support
- Consistent access to the most up-to-date data What should you do?

A. Implement BigQuery with no region selected for storage or processing.
B. Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.
C. Implement Cloud SQL for PostgreSQL with the master in Norht America and read replicas in Asia and Europe.
D. Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.

**Answer:** B


**NEW QUESTION 111**
- (Exam Topic 6)
You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS_yyyymmdd. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

A. Convert all daily log tables into date-partitioned tables
B. Convert the sharded tables into a single partitioned table
C. Enable query caching so you can cache data from previous months
D. Create separate views to cover each month, and query from these views

**Answer:** A


**NEW QUESTION 112**
- (Exam Topic 6)
Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

- The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured
- Support for publish/subscribe semantics on hundreds of topics
- Retain per-key ordering Which system should you choose?

A. Apache Kafka
B. Cloud Storage
C. Cloud Pub/Sub
D. Firebase Cloud Messaging

**Answer:** A


**NEW QUESTION 117**
- (Exam Topic 6)
You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products of features of the platform. What should you do?

A. Export the information to Cloud Stackdriver, and set up an Alerting policy
B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver
C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs
D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

**Answer:** B


**NEW QUESTION 119**
- (Exam Topic 6)
You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?

A. Cloud Scheduler
B. Cloud Dataflow
C. Cloud Functions
D. Cloud Composer

**Answer:** A


**NEW QUESTION 120**
- (Exam Topic 6)
You work for an advertising company, and you've developed a Spark ML model to predict click-through rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

A. Use Cloud ML Engine for training existing Spark ML models
B. Rewrite your models on TensorFlow, and start using Cloud ML Engine
C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery

D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

**Answer:** A

**NEW QUESTION 123**
- (Exam Topic 6)
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

A. Use gcloud kms keys create to create a symmetric ke
B. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
C. Use gcloud kms keys create to create a symmetric ke
D. Then use gcloud kms encrypt to encrypt each archival file with the ke
E. Use gsutil cp to upload each encrypted file to the Cloud Storage bucke
F. Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.
G. Specify customer-supplied encryption key (CSEK) in the .boto configuration fil
H. Use gsutil cp to upload each archival file to the Cloud Storage bucke
I. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
J. Specify customer-supplied encryption key (CSEK) in the .boto configuration fil
K. Use gsutil cp to upload each archival file to the Cloud Storage bucke
L. Save the CSEK in a different project that only the security team can access.

**Answer:** B

**NEW QUESTION 126**
......

# THANKS FOR TRYING THE DEMO OF OUR PRODUCT

Visit Our Site to Purchase the Full Set of Actual Professional-Data-Engineer Exam Questions With Answers.

We Also Provide Practice Exam Software That Simulates Real Exam Environment And Has Many Self-Assessment Features. Order the Professional-Data-Engineer Product From:

## https://www.2passeasy.com/dumps/Professional-Data-Engineer/

## Money Back Guarantee

### Professional-Data-Engineer Practice Exam Features:

* Professional-Data-Engineer Questions and Answers Updated Frequently

* Professional-Data-Engineer Practice Questions Verified by Expert Senior Certified Staff

* Professional-Data-Engineer Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* Professional-Data-Engineer Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year