Google

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

NEW QUESTION 1
- (Exam Topic 1)
You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.
B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.
C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.
D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Answer: B

NEW QUESTION 2
- (Exam Topic 1)
Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

A. Check the dashboard application to see if it is not displaying correctly.
B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.
C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

Answer: B

NEW QUESTION 3
- (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 4
- (Exam Topic 1)
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.

Answer: D

NEW QUESTION 5
- (Exam Topic 1)
An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?

A. Use federated data sources, and check data in the SQL query.
B. Enable BigQuery monitoring in Google Stackdriver and create an alert.
C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.
D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Answer: D

NEW QUESTION 6
- (Exam Topic 1)
Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour.
The data scientists have written the following code to read the data for a new key features in the logs. BigQueryIO.Read
.named("ReadLogData")
.from("clouddataflow-readonly:samples.log_data")
You want to improve the performance of this data read. What should you do?

A. Specify the TableReference object in the code.
B. Use .fromQuery operation to read specific fields from the table.
C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
D. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.

Answer: D

NEW QUESTION 7
- (Exam Topic 1)
Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

A. Threading
B. Serialization
C. Dropout Methods
D. Dimensionality Reduction

Answer: C

Explanation:
Reference
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505

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 2)
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

Answer: C

NEW QUESTION 10
- (Exam Topic 3)
You need to compose visualization for operations teams with the following requirements:

❯ Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once
every minute)

❯ The report must not be more than 3 hours delayed from live data.

❯ The actionable report should only show suboptimal links.

❯ Most suboptimal links should be sorted to the top.

❯ Suboptimal links can be grouped and filtered by regional geography.

❯ User response time to load the report must be <5 seconds.

You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?

A. Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.
B. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.
C. Export the data to a spreadsheet, compose a series of charts and tables, one for each possible combination of criteria, and spread them across multiple tabs.
D. Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API.

Answer: B

NEW QUESTION 10
- (Exam Topic 3)
MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

A. The zone
B. The number of workers
C. The disk size per worker
D. The maximum number of workers

Answer: A

NEW QUESTION 12
- (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 16
- (Exam Topic 5)
The Dataflow SDKs have been recently transitioned into which Apache service?

A. Apache Spark
B. Apache Hadoop
C. Apache Kafka
D. Apache Beam

Answer: D

Explanation:
Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive Reference: https://cloud.google.com/dataflow/docs/

NEW QUESTION 21
- (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 24
- (Exam Topic 5)
You are planning to use Google's Dataflow SDK to analyze customer data such as displayed below. Your project requirement is to extract only the customer name from the data source and then write to an output PCollection.
Tom,555 X street Tim,553 Y street Sam, 111 Z street
Which operation is best suited for the above data processing requirement?

A. ParDo
B. Sink API
C. Source API
D. Data extraction

Answer: A

Explanation:
In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.
Reference: https://cloud.google.com/dataflow/model/par-do

NEW QUESTION 27
- (Exam Topic 5)
How would you query specific partitions in a BigQuery table?

A. Use the DAY column in the WHERE clause
B. Use the EXTRACT(DAY) clause
C. Use the __PARTITIONTIME pseudo-column in the WHERE clause

D. Use DATE BETWEEN in the WHERE clause

Answer: C

Explanation:
Partitioned tables include a pseudo column named _PARTITIONTIME that contains a date-based timestamp for data loaded into the table. To limit a query to particular partitions (such as Jan 1st and 2nd of 2017), use a clause similar to this:
WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02')
Reference: https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

NEW QUESTION 28
- (Exam Topic 5)
Which is not a valid reason for poor Cloud Bigtable performance?

A. The workload isn't appropriate for Cloud Bigtable.
B. The table's schema is not designed correctly.
C. The Cloud Bigtable cluster has too many nodes.
D. There are issues with the network connection.

Answer: C

Explanation:
The Cloud Bigtable cluster doesn't have enough nodes. If your Cloud Bigtable cluster is overloaded, adding more nodes can improve performance. Use the monitoring tools to check whether the cluster is overloaded.
Reference: https://cloud.google.com/bigtable/docs/performance

NEW QUESTION 33
- (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 34
- (Exam Topic 5)
Which Java SDK class can you use to run your Dataflow programs locally?

A. LocalRunner
B. DirectPipelineRunner
C. MachineRunner
D. LocalPipelineRunner

Answer: B

Explanation:
DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization. Useful for small local execution and tests
Reference:
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun

NEW QUESTION 38
- (Exam Topic 5)
When creating a new Cloud Dataproc cluster with the projects.regions.clusters.create operation, these four values are required: project, region, name, and .

A. zone
B. node
C. label
D. type

Answer: A

Explanation:
At a minimum, you must specify four values when creating a new cluster with the projects.regions.clusters.create operation:
The project in which the cluster will be created The region to use
The name of the cluster
The zone in which the cluster will be created
You can specify many more details beyond these minimum requirements. For example, you can
also specify the number of workers, whether preemptible compute should be used, and the network settings. Reference:
https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create_a_new_cloud_dataproc_cluste

NEW QUESTION 39
- (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 44
- (Exam Topic 5)
Which role must be assigned to a service account used by the virtual machines in a Dataproc cluster so they can execute jobs?

A. Dataproc Worker
B. Dataproc Viewer
C. Dataproc Runner
D. Dataproc Editor

Answer: A

Explanation:
Service accounts used with Cloud Dataproc must have Dataproc/Dataproc Worker role (or have all the permissions granted by Dataproc Worker role).
Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

NEW QUESTION 45
- (Exam Topic 5)
All Google Cloud Bigtable client requests go through a front-end server they are sent to a Cloud Bigtable node.

A. before
B. after
C. only if
D. once

Answer: A

Explanation:
In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node.
The nodes are organized into a Cloud Bigtable cluster, which belongs to a Cloud Bigtable instance, which is a container for the cluster. Each node in the cluster handles a subset of the requests to the cluster.
When additional nodes are added to a cluster, you can increase the number of simultaneous requests that the cluster can handle, as well as the maximum throughput for the entire cluster.
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION 50
- (Exam Topic 5)
Google Cloud Bigtable indexes a single value in each row. This value is called the .

A. primary key
B. unique key
C. row key
D. master key

Answer: C

Explanation:
Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data.
A single value in each row is indexed; this value is known as the row key.
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION 53
- (Exam Topic 5)
What is the recommended action to do in order to switch between SSD and HDD storage for your Google Cloud Bigtable instance?

A. create a third instance and sync the data from the two storage types via batch jobs
B. export the data from the existing instance and import the data into a new instance
C. run parallel instances where one is HDD and the other is SDD
D. the selection is final and you must resume using the same storage type

Answer: B

Explanation:
When you create a Cloud Bigtable instance and cluster, your choice of SSD or HDD storage for the cluster is permanent. You cannot use the Google Cloud Platform Console to change the type of storage that is used for the cluster.
If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance. Alternatively, you can write
a Cloud Dataflow or Hadoop MapReduce job that copies the data from one instance to another. Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd–

NEW QUESTION 56
- (Exam Topic 5)
Which of the following statements is NOT true regarding Bigtable access roles?

A. Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.
B. To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.
C. You can configure access control only at the project level.
D. To give a user access to only one table in a project, you must configure access through your application.

Answer: B

Explanation:
For Cloud Bigtable, you can configure access control at the project level. For example, you can grant the ability to:
Read from, but not write to, any table within the project.
Read from and write to any table within the project, but not manage instances. Read from and write to any table within the project, and manage instances.
Reference: https://cloud.google.com/bigtable/docs/access-control

NEW QUESTION 57
- (Exam Topic 5)
When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied errors. What could be the reason for that?

A. Your gcloud does not have access to the BigQuery resources
B. BigQuery cannot be accessed from local machines
C. You are missing gcloud on your machine
D. Pipelines cannot be run locally

Answer: A

Explanation:
When reading from a Dataflow source or writing to a Dataflow sink using DirectPipelineRunner, the Cloud Platform account that you configured with the gcloud executable will need access to the corresponding source/sink
Reference:
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun

NEW QUESTION 60
- (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 61
- (Exam Topic 5)
The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?

A. Workers
B. Masters, workers, and parameter servers
C. Workers and parameter servers
D. Parameter servers

Answer: C

Explanation:
The CUSTOM tier is not a set tier, but rather enables you to use your own cluster specification. When you use this tier, set values to configure your processing cluster according to these guidelines:
You must set TrainingInput.masterType to specify the type of machine to use for your master node. You may set TrainingInput.workerCount to specify the number of workers to use.
You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use.
You can specify the type of machine for the master node, but you can't specify more than one master node. Reference: https://cloud.google.com/ml-engine/docs/training-overview#job_configuration_parameters

NEW QUESTION 66

- (Exam Topic 5)
Which methods can be used to reduce the number of rows processed by BigQuery?

A. Splitting tables into multiple tables; putting data in partitions
B. Splitting tables into multiple tables; putting data in partitions; using the LIMIT clause
C. Putting data in partitions; using the LIMIT clause
D. Splitting tables into multiple tables; using the LIMIT clause

Answer: A

Explanation:
If you split a table into multiple tables (such as one table for each day), then you can limit your query to the data in specific tables (such as for particular days). A better method is to use a partitioned table, as long as your data can be separated by the day.
If you use the LIMIT clause, BigQuery will still process the entire table. Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

NEW QUESTION 69
- (Exam Topic 6)
You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this? Choose 2 answers.

A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
B. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
D. Write an application that uses Cloud Datastore client libraries to read all the entitie
E. Treat each entity as a BigQuery table row via BigQuery streaming inser
F. Assign an export timestamp for each export, and attach it as an extra column for each ro
G. Make sure that the BigQuery table is partitioned using the export timestamp column.
H. Write an application that uses Cloud Datastore client libraries to read all the entitie
I. Format the exported data into a JSON fil
J. Apply compression before storing the data in Cloud Source Repositories.

Answer: CE

NEW QUESTION 71
- (Exam Topic 6)
Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

A. Increase the CPU size on your server.
B. Increase the size of the Google Persistent Disk on your server.
C. Increase your network bandwidth from your datacenter to GCP.
D. Increase your network bandwidth from Compute Engine to Cloud Storage.

Answer: C

NEW QUESTION 75
- (Exam Topic 6)
You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:

> Each department should have access only to their data.

> Each department will have one or more leads who need to be able to create and update tables and provide them to their team.

> Each department has data analysts who need to be able to query but not modify data. How should you set access to the data in BigQuery?

A. Create a dataset for each departmen
B. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.
C. Create a dataset for each departmen
D. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.
E. Create a table for each departmen
F. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.
G. Create a table for each departmen
H. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

Answer: D

NEW QUESTION 78
- (Exam Topic 6)
You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

A. PigLatin using Pig
B. HiveQL using Hive
C. Java using MapReduce
D. Python using MapReduce

Answer: D

NEW QUESTION 80
- (Exam Topic 6)
You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.
What should you do?

A. Use Cloud Dataflow with Beam to detect errors and perform transformations.
B. Use Cloud Dataprep with recipes to detect errors and perform transformations.
C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

Answer: B

NEW QUESTION 81
- (Exam Topic 6)
You are updating the code for a subscriber to a Put/Sub feed. You are concerned that upon deployment the subscriber may erroneously acknowledge messages, leading to message loss. You subscriber is not set up to retain acknowledged messages. What should you do to ensure that you can recover from errors after deployment?

A. Use Cloud Build for your deployment if an error occurs after deployment, use a Seek operation to locate a tmestamp logged by Cloud Build at the start of the deployment
B. Create a Pub/Sub snapshot before deploying new subscriber cod
C. Use a Seek operation to re-deliver messages that became available after the snapshot was created
D. Set up the Pub/Sub emulator on your local machine Validate the behavior of your new subscriber togs before deploying it to production
E. Enable dead-lettering on the Pub/Sub topic to capture messages that aren't successful acknowledged if an error occurs after deployment, re-deliver any messages captured by the dead-letter queue

Answer: B

NEW QUESTION 85
- (Exam Topic 6)
You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:
SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country
You check the query plan for the query and see the following output in the Read section of Stage:1:

What is the most likely cause of the delay for this query?

A. Users are running too many concurrent queries in the system
B. The [myproject:mydataset.mytable] table has too many partitions
C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Answer: A

NEW QUESTION 90
- (Exam Topic 6)
You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

A. Use bq load to load a batch of sensor data every 60 seconds.
B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
C. Use the INSERT statement to insert a batch of data every 60 seconds.
D. Use the MERGE statement to apply updates in batch every 60 seconds.

Answer: C

NEW QUESTION 91
- (Exam Topic 6)
A TensorFlow machine learning model on Compute Engine virtual machines (n2-standard -32) takes two days to complete framing. The model has custom TensorFlow operations that must run partially on a CPU You want to reduce the training time in a cost-effective manner. What should you do?

A. Change the VM type to n2-highmem-32
B. Change the VM type to e2 standard-32
C. Train the model using a VM with a GPU hardware accelerator
D. Train the model using a VM with a TPU hardware accelerator

Answer: C

NEW QUESTION 93
- (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 98
- (Exam Topic 6)
An organization maintains a Google BigQuery dataset that contains tables with user-level datA. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

A. Create and share an authorized view that provides the aggregate results.
B. Create and share a new dataset and view that provides the aggregate results.
C. Create and share a new dataset and table that contains the aggregate results.
D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Answer: D

Explanation:
Reference: https://cloud.google.com/bigquery/docs/access-control

NEW QUESTION 99
- (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 104
- (Exam Topic 6)
Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data).
What should you do?

A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
B. Add a try… catch block to your DoFn that transforms the data, extract erroneous rows from logs.
C. Add a try… catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.
D. Add a try… catch block to your DoFn that transforms the data, use a sideOutput to create a PCollectionthat can be stored to PubSub later.

Answer: C

NEW QUESTION 108
- (Exam Topic 6)
You are designing a pipeline that publishes application events to a Pub/Sub topic. You need to aggregate events across hourly intervals before loading the results to BigQuery for analysis. Your solution must be scalable so it can process and load large volumes of events to BigQuery. What should you do?

A. Create a streaming Dataflow job to continually read from the Pub/Sub topic and perform the necessary aggregations using tumbling windows
B. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub-Sub topic and performing the necessary aggregations
C. Schedule a Cloud Function to run hourly, pulling all avertable messages from the Pub/Sub topic and performing the necessary aggregations
D. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.

Answer: A

NEW QUESTION 112
- (Exam Topic 6)
Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?

A. Create a file on a shared file and have the application servers write all bid events to that fil
B. Process the file with Apache Hadoop to identify which user bid first.
C. Have each application server write the bid events to Cloud Pub/Sub as they occu
D. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
E. Set up a MySQL database for each application server to write bid events int
F. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
G. Have each application server write the bid events to Google Cloud Pub/Sub as they occu
H. Use a pull subscription to pull the bid events using Google Cloud Dataflo
I. Give the bid for each item to the userin the bid event that is processed first.

Answer: C

NEW QUESTION 117
- (Exam Topic 6)
Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

A. Enable data access logs in each Data Analyst's projec
B. Restrict access to Stackdriver Logging via Cloud IAM roles.
C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' project
D. Restrict access to the Cloud Storage bucket.
E. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit log
F. Restrict access to the project with the exported logs.
G. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit log
H. Restrict access to the project that contains the exported logs.

Answer: D

NEW QUESTION 118
- (Exam Topic 6)
You are testing a Dataflow pipeline to ingest and transform text files. The files are compressed gzip, errors are written to a dead-letter queue, and you are using SideInputs to join data You noticed that the pipeline is taking longer to complete than expected, what should you do to expedite the Dataflow job?

A. Switch to compressed Avro files
B. Reduce the batch size
C. Retry records that throw an error
D. Use CoGroupByKey instead of the SideInput

Answer: B

NEW QUESTION 120
- (Exam Topic 6)
You are using BigQuery and Data Studio to design a customer-facing dashboard that displays large quantities of aggregated data. You expect a high volume of concurrent users. You need to optimize tie dashboard to provide quick visualizations with minimal latency. What should you do?

A. Use BigQuery BI Engine with materialized views
B. Use BigQuery BI Engine with streaming data.
C. Use BigQuery BI Engine with authorized views
D. Use BigQuery BI Engine with logical reviews

Answer: B

NEW QUESTION 123
- (Exam Topic 6)
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
C. Use the BigQuery streaming the stream changes into a daily inventory movement tabl
D. Calculate balances in a view that joins it to the historical inventory balance tabl
E. Update the inventory balance table nightly.
F. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table.Calculate balances in a view that joins it to the historical inventory balance tabl
G. Update the inventory balance table nightly.

Answer: A

NEW QUESTION 126
- (Exam Topic 6)
An aerospace company uses a proprietary data format to store its night data. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming as few resources as possible. What should you do?

A. Use a standard Dataflow pipeline to store the raw data m BigQuery and then transform the format later when the data is used
B. Write a she script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source
C. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format
D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format

Answer: D

NEW QUESTION 131
- (Exam Topic 6)
You are designing a cloud-native historical data processing system to meet the following conditions:

➤ The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.

➤ A streaming data pipeline stores new data daily.

➤ Peformance is not a factor in the solution.

> The solution design should maximize availability.

How should you design data storage for this solution?

A. Create a Cloud Dataproc cluster with high availabilit
B. Store the data in HDFS, and peform analysis as needed.
C. Store the data in BigQuer
D. Access the data using the BigQuery Connector or Cloud Dataproc and Compute Engine.
E. Store the data in a regional Cloud Storage bucke
F. Aceess the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
G. Store the data in a multi-regional Cloud Storage bucke
H. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Answer: D

NEW QUESTION 136
- (Exam Topic 6)
You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

A. Use Cloud TPUs without any additional adjustment to your code.
B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.
C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.
D. Stay on CPUs, and increase the size of the cluster you're training your model on.

Answer: B

NEW QUESTION 141
- (Exam Topic 6)
You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?

A. Create a Cloud Dataproc Workflow Template
B. Create an initialization action to execute the jobs
C. Create a Directed Acyclic Graph in Cloud Composer
D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

Answer: C

NEW QUESTION 144
- (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 148
- (Exam Topic 6)
Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?

A. They have not assigned the timestamp, which causes the job to fail
B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

Answer: C

NEW QUESTION 149
- (Exam Topic 6)
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 nonkey columns. What should you do?

A. Use Cloud SQL for storag
B. Add secondary indexes to support query patterns.

C. Use Cloud SQL for storag
D. Use Cloud Dataflow to transform data to support query patterns.
E. Use Cloud Spanner for storag
F. Add secondary indexes to support query patterns.
G. Use Cloud Spanner for storag
H. Use Cloud Dataflow to transform data to support query patterns.

Answer: D

Explanation:

Reference: https://cloud.google.com/solutions/data-lifecycle-cloud-platform

NEW QUESTION 151
- (Exam Topic 6)
You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once, and must be ordered within windows of 1 hour. How should you design the solution?

A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.
B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.
C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.
D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.

Answer: D

NEW QUESTION 155
- (Exam Topic 6)
You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

A. Use Cloud Dataprep to find null values in sample source dat
B. Convert all nulls to 'none' using a Cloud Dataproc job.
C. Use Cloud Dataprep to find null values in sample source dat
D. Convert all nulls to 0 using a Cloud Dataprep job.
E. Use Cloud Dataflow to find null values in sample source dat
F. Convert all nulls to 'none' using a Cloud Dataprep job.
G. Use Cloud Dataflow to find null values in sample source dat
H. Convert all nulls to using a custom script.

Answer: C

NEW QUESTION 157
......

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

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