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)
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: A

NEW QUESTION 3
- (Exam Topic 1)
You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

A. Grant the consultant the Viewer role on the project.
B. Grant the consultant the Cloud Dataflow Developer role on the project.
C. Create a service account and allow the consultant to log on with it.
D. Create an anonymized sample of the data for the consultant to work with in a different project.

Answer: C

NEW QUESTION 4
- (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 5
- (Exam Topic 1)
You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available.
How should you use this data to train the model?

A. Continuously retrain the model on just the new data.
B. Continuously retrain the model on a combination of existing data and the new data.
C. Train on the existing data while using the new data as your test set.
D. Train on the new data while using the existing data as your test set.

Answer: D

NEW QUESTION 6
- (Exam Topic 1)
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 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 prespecified date ranges.

C. Normalize the master patient-record table into the patient table and the visits table, and create othernecessary tables to avoid self-join.
D. Partition the table into smaller tables, with one for each clini
E. Run queries against the smaller table pairs, and use unions for consolidated reports.

Answer: B

NEW QUESTION 7
- (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 asingle row in the table.

Answer: D

NEW QUESTION 8
- (Exam Topic 1)
You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

A. Linear regression
B. Logistic classification
C. Recurrent neural network
D. Feedforward neural network

Answer: A

NEW QUESTION 9
- (Exam Topic 1)
You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?

A. Re-write the application to load accumulated data every 2 minutes.
B. Convert the streaming insert code to batch load for individual messages.
C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.

Answer: A

NEW QUESTION 10
- (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 10
- (Exam Topic 3)
MJTelco is building a custom interface to share data. They have these requirements:
- They need to do aggregations over their petabyte-scale datasets.
- They need to scan specific time range rows with a very fast response time (milliseconds). Which combination of Google Cloud Platform products should you recommend?

A. Cloud Datastore and Cloud Bigtable
B. Cloud Bigtable and Cloud SQL
C. BigQuery and Cloud Bigtable
D. BigQuery and Cloud Storage

Answer: C

NEW QUESTION 11

- (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 work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
B. Add a new column called FullName to the Users tabl
C. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.
D. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.
E. Use BigQuery to export the data for the table to a CSV fil
F. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullNam
G. Run a BigQuery load job to load the new CSV file into BigQuery.

Answer: C

NEW QUESTION 16
- (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 18
- (Exam Topic 4)
You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:

The user profile: What the user likes and doesn't like to eat

The user account information: Name, address, preferred meal times

The order information: When orders are made, from where, to whom
The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

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

Answer: A

NEW QUESTION 23
- (Exam Topic 4)
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 Datastor
D. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore
E. Store the data in a file in a regional Google Cloud Storage bucke
F. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Answer: A

NEW QUESTION 25
- (Exam Topic 4)
You work for a manufacturing plant that batches application log files together into a single log file once a day at 2:00 AM. You have written a Google Cloud Dataflow job to process that log file. You need to make sure the log file in processed once per day as inexpensively as possible. What should you do?

A. Change the processing job to use Google Cloud Dataproc instead.
B. Manually start the Cloud Dataflow job each morning when you get into the office.
C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.
D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Answer: C

NEW QUESTION 26
- (Exam Topic 5)
When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through a proxy.

A. HTTPS
B. VPN
C. SOCKS
D. HTTP

Answer: C

Explanation:
When using Cloud Dataproc clusters, configure your browser to use the SOCKS proxy. The SOCKS proxy routes data intended for the Cloud Dataproc cluster through an SSH tunnel.
Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces

NEW QUESTION 31
- (Exam Topic 5)
What are all of the BigQuery operations that Google charges for?

A. Storage, queries, and streaming inserts
B. Storage, queries, and loading data from a file
C. Storage, queries, and exporting data
D. Queries and streaming inserts

Answer: A

Explanation:
Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations.
Reference: https://cloud.google.com/bigquery/pricing

NEW QUESTION 35
- (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 36
- (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 40
- (Exam Topic 5)

Which of these is NOT a way to customize the software on Dataproc cluster instances?

A. Set initialization actions
B. Modify configuration files using cluster properties
C. Configure the cluster using Cloud Deployment Manager
D. Log into the master node and make changes from there

Answer: C

Explanation:
You can access the master node of the cluster by clicking the SSH button next to it in the Cloud Console.
You can easily use the --properties option of the dataproc command in the Google Cloud SDK to modify many common configuration files when creating a cluster.
When creating a Cloud Dataproc cluster, you can specify initialization actions in executables and/or scripts that Cloud Dataproc will run on all nodes in your Cloud Dataproc cluster immediately after the cluster is set up. [https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions]
Reference: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster-properties

NEW QUESTION 43
- (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 45
- (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 48
- (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 50
- (Exam Topic 5)
Scaling a Cloud Dataproc cluster typically involves .

A. increasing or decreasing the number of worker nodes
B. increasing or decreasing the number of master nodes
C. moving memory to run more applications on a single node
D. deleting applications from unused nodes periodically

Answer: A

Explanation:

After creating a Cloud Dataproc cluster, you can scale the cluster by increasing or decreasing the number of worker nodes in the cluster at any time, even when jobs are running on the cluster. Cloud Dataproc clusters are typically scaled to:
1) increase the number of workers to make a job run faster
2) decrease the number of workers to save money
3) increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage Reference: https://cloud.google.com/dataproc/docs/concepts/scaling-clusters

NEW QUESTION 53
- (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 58
- (Exam Topic 5)
If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?

A. Unsupervised learning
B. Regressor
C. Classifier
D. Clustering estimator

Answer: B

Explanation:
Regression is the supervised learning task for modeling and predicting continuous, numeric variables. Examples include predicting real-estate prices, stock price movements, or student test scores.
Classification is the supervised learning task for modeling and predicting categorical variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades.
Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset.
Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.
Reference: https://elitedatascience.com/machine-learning-algorithms

NEW QUESTION 63
- (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 65
- (Exam Topic 5)
Which of the following are feature engineering techniques? (Select 2 answers)

A. Hidden feature layers
B. Feature prioritization
C. Crossed feature columns
D. Bucketization of a continuous feature

Answer: CD

Explanation:
Selecting and crafting the right set of feature columns is key to learning an effective model. Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive
bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.
Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.
Reference: https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

NEW QUESTION 68
- (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 71
- (Exam Topic 5)
Which of the following IAM roles does your Compute Engine account require to be able to run pipeline jobs?

A. dataflow.worker
B. dataflow.compute
C. dataflow.developer
D. dataflow.viewer

Answer: A

Explanation:
The dataflow.worker role provides the permissions necessary for a Compute Engine service account to execute work units for a Dataflow pipeline
Reference: https://cloud.google.com/dataflow/access-control

NEW QUESTION 72
- (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 73
- (Exam Topic 5)
Which of the following is NOT a valid use case to select HDD (hard disk drives) as the storage for Google Cloud Bigtable?

A. You expect to store at least 10 TB of data.
B. You will mostly run batch workloads with scans and writes, rather than frequently executing random reads of a small number of rows.
C. You need to integrate with Google BigQuery.
D. You will not use the data to back a user-facing or latency-sensitive application.

Answer: C

Explanation:
For example, if you plan to store extensive historical data for a large number of remote-sensing devices and then use the data to generate daily reports, the cost savings for HDD storage may justify the performance tradeoff. On the other hand, if you plan to use the data to display a real-time dashboard, it probably would not make sense to use HDD storage—reads would be much more frequent in this case, and reads are much slower with HDD storage.
Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

NEW QUESTION 76
- (Exam Topic 5)
For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?

A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.

Answer: C

Explanation:
It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,
If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud

Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.
It is recommended to avoid creating your Compute Engine instance in a different region from
your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.
Reference: https://cloud.google.com/bigtable/docs/creating-compute-instance

NEW QUESTION 78
- (Exam Topic 5)
Cloud Dataproc is a managed Apache Hadoop and Apache service.

A. Blaze
B. Spark
C. Fire
D. Ignite

Answer: B

Explanation:
Cloud Dataproc is a managed Apache Spark and Apache Hadoop service that lets you use open source data tools for batch processing, querying, streaming, and machine learning.
Reference: https://cloud.google.com/dataproc/docs/

NEW QUESTION 79
- (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 82
- (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 86
- (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 89
- (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 91
- (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 92
- (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 96
- (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 97
- (Exam Topic 5)
Cloud Bigtable is a recommended option for storing very large amounts of _____ ?

A. multi-keyed data with very high latency
B. multi-keyed data with very low latency
C. single-keyed data with very low latency
D. single-keyed data with very high latency

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. Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION 99
- (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 101
- (Exam Topic 6)
You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

A. Increase the number of max workers
B. Use a larger instance type for your Cloud Dataflow workers
C. Change the zone of your Cloud Dataflow pipeline to run in us-central1
D. Create a temporary table in Cloud Bigtable that will act as a buffer for new dat
E. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery
F. Create a temporary table in Cloud Spanner that will act as a buffer for new dat
G. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

Answer: BE

NEW QUESTION 104
- (Exam Topic 6)
You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the
Google-recommended cloud native architecture for this scenario?

A. Edge TPUs as sensor devices for storing and transmitting the messages.
B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

Answer: C

NEW QUESTION 107
- (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 112
- (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 116
- (Exam Topic 6)
You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:

🔵 Executing the transformations on a schedule

🔵 Enabling non-developer analysts to modify transformations

🔵 Providing a graphical tool for designing transformations
What should you do?

A. Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis
B. Load each month's CSV data into BigQuery, and write a SQL query to transform the data to a standard schem
C. Merge the transformed tables together with a SQL query
D. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformatio
E. The Python code should be stored in a revision control system and modified as the incoming data's schema changes
F. Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe.Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

Answer: D

NEW QUESTION 120
- (Exam Topic 6)
You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time. Consumers will receive the data in the following ways:

🔵 Real-time event stream

🔵 ANSI SQL access to real-time stream and historical data

🔵 Batch historical exports
Which solution should you use?

A. Cloud Dataflow, Cloud SQL, Cloud Spanner
B. Cloud Pub/Sub, Cloud Storage, BigQuery
C. Cloud Dataproc, Cloud Dataflow, BigQuery
D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

Answer: A

NEW QUESTION 125
- (Exam Topic 6)
Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

A. Migrate the workload to Google Cloud Dataflow
B. Use pre-emptible virtual machines (VMs) for the cluster
C. Use a higher-memory node so that the job runs faster
D. Use SSDs on the worker nodes so that the job can run faster

Answer: A

NEW QUESTION 129
- (Exam Topic 6)
Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?

A. Cloud Bigtable
B. Google BigQuery
C. Google Cloud Storage
D. Google Cloud Datastore

Answer: A

Explanation:
Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series

NEW QUESTION 134
- (Exam Topic 6)
You are developing an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of data. What should you do?

A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results.Deploy the models using Cloud Datapro
B. Call the model from your application.
C. Build and train a classification model with Spark MLlib to generate label
D. Build and train a second classification model with Spark MLlib to filter results to match customer preference
E. Deploy theModels using Cloud Datapro
F. Call the models from your application.
G. Build an application that calls the Cloud Video Intelligence API to generate label
H. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.
I. Build an application that calls the Cloud Video Intelligence API to generate label
J. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Answer: C

NEW QUESTION 135
- (Exam Topic 6)
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 applicatio
B. Process the generated Entity Analysis as labels.
C. Call the Cloud Natural Language API from your applicatio
D. Process the generated Sentiment Analysis as labels.
E. Build and train a text classification model using TensorFlo
F. Deploy the model using Cloud Machine Learning Engin
G. Call the model from your application and process the results as labels.
H. Build and train a text classification model using TensorFlo
I. Deploy the model using a KubernetesEngine cluste
J. Call the model from your application and process the results as labels.

Answer: B

NEW QUESTION 139
- (Exam Topic 6)
You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?

A. Use Transfer Appliance to copy the data to Cloud Storage
B. Use gsutil cp –J to compress the content being uploaded to Cloud Storage
C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic

Answer: A

NEW QUESTION 141
- (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 143
- (Exam Topic 6)
You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?

A. cron
B. Cloud Composer
C. Cloud Scheduler
D. Workflow Templates on Cloud Dataproc

Answer: D

NEW QUESTION 146
- (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 150
- (Exam Topic 6)
You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
B. Create encryption keys in Cloud Key Management Servic
C. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
D. Create encryption keys locall
E. Upload your encryption keys to Cloud Key Management Servic
F. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
G. Create encryption keys in Cloud Key Management Servic
H. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

Answer: C

NEW QUESTION 152
- (Exam Topic 6)
Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

A. Cloud Dataflow
B. Cloud Composer
C. Cloud Dataprep
D. Cloud Dataproc

Answer: D

NEW QUESTION 154
- (Exam Topic 6)
You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

A. Increase the cluster size with more non-preemptible workers.
B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.
D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.

Answer: D

Explanation:
Reference https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex

NEW QUESTION 156
- (Exam Topic 6)
You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

A. Batch job, PubSubIO, side-inputs
B. Streaming job, PubSubIO, JdbcIO, side-outputs
C. Streaming job, PubSubIO, BigQueryIO, side-inputs
D. Streaming job, PubSubIO, BigQueryIO, side-outputs

Answer: A

NEW QUESTION 160
- (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 165
- (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 166
- (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 167
- (Exam Topic 6)
You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery.
How should you securely run this workload?

A. Restrict the Google Cloud Storage bucket so only you can see the files
B. Grant the Project Owner role to a service account, and run the job with it
C. Use a service account with the ability to read the batch files and to write to BigQuery
D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Answer: B

NEW QUESTION 168
- (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 170
- (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 172
- (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 174
- (Exam Topic 6)
You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.
Which service do you select for storing and serving your data?

A. Cloud Spanner
B. Cloud Bigtable
C. Cloud Firestore
D. Cloud SQL

Answer: D

NEW QUESTION 177
- (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 178
- (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 181
- (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 185
- (Exam Topic 6)
You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Answer: A

NEW QUESTION 187
- (Exam Topic 6)
You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload.
What should you do?

A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

Answer: B

NEW QUESTION 189
- (Exam Topic 6)
You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC). All ORC files have been successfully copied to a Cloud Storage bucket. You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance. What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDF
B. Mount the Hive tables locally.
C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluste
D. Mount the Hive tables locally.
E. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluste
F. Then run the Hadoop utility to copy them do HDF
G. Mount the Hive tables from HDFS.
H. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive table
I. Replicate external Hive tables to the native ones.
J. Load the ORC files into BigQuer
K. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive table
L. Replicate external Hive tables to the native ones.

Answer: BC

NEW QUESTION 193
- (Exam Topic 6)
Your neural network model is taking days to train. You want to increase the training speed. What can you do?

A. Subsample your test dataset.
B. Subsample your training dataset.
C. Increase the number of input features to your model.
D. Increase the number of layers in your neural network.

Answer: D

Explanation:
Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

NEW QUESTION 194
- (Exam Topic 6)
You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.
What should you do?

A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
B. Place the MariaDB instances in an Instance Group with a Health Check.
C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.
D. Install the StackDriver Agent and configure the MySQL plugin.

Answer: C

NEW QUESTION 197
- (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 201
- (Exam Topic 6)
You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

A. Store and process the entire dataset in BigQuery.
B. Store and process the entire dataset in Cloud Bigtable.
C. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.
D. Store the warm data as files in Cloud Storage, and store the active data in BigQuer
E. Keep this ratio as 80% warm and 20% active.

Answer: D

NEW QUESTION 202
- (Exam Topic 6)
You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

A. Publisher throughput quota is too small.
B. Total outstanding messages exceed the 10-MB maximum.
C. Error handling in the subscriber code is not handling run-time errors properly.
D. The subscriber code cannot keep up with the messages.
E. The subscriber code does not acknowledge the messages that it pulls.

Answer: CD

NEW QUESTION 204
- (Exam Topic 6)
You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

A. Add a SideInput that returns a Boolean if the element is corrupt.
B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Answer: B

NEW QUESTION 205
- (Exam Topic 6)
You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants to add cameras to the delivery lines to detect and track any visual damage to the packages in transit. You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit. Which solution should you choose?

A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.
B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.
C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Function
D. Integrate the package tracking applications with this function.
E. Use TensorFlow to create a model that is trained on your corpus of image
F. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

Answer: A

NEW QUESTION 208
- (Exam Topic 6)
You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the data. They should only see certain tables based on their team membership. How should you set user permissions?

A. Assign the users/groups data viewer access at the table level for each table
B. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views
C. Create authorized views for each team in the same dataset in which the data resides, and assign theusers/groups data viewer access to the authorized views
D. Create authorized views for each team in datasets created for each tea
E. Assign the authorized views data viewer access to the dataset in which the data reside
F. Assign the users/groups data viewer access to the datasets in which the authorized views reside

Answer: C

NEW QUESTION 211
- (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 216
- (Exam Topic 6)
You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?

A. Create an API using App Engine to receive and send messages to the applications
B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them
C. Create a table on Cloud SQL, and insert and delete rows with the job information
D. Create a table on Cloud Spanner, and insert and delete rows with the job information

Answer: A

NEW QUESTION 218
- (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 219
......

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