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 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 weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?

A. Issue a command to restart the database servers.
B. Retry the query with exponential backoff, up to a cap of 15 minutes.
C. Retry the query every second until it comes back online to minimize staleness of data.
D. Reduce the query frequency to once every hour until the database comes back online.

Answer: B

Explanation:
https://cloud.google.com/sql/docs/mysql/manage-connections#backoff

NEW QUESTION 3
- (Exam Topic 1)
Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)

A. Disable writes to certain tables.
B. Restrict access to tables by role.
C. Ensure that the data is encrypted at all times.
D. Restrict BigQuery API access to approved users.
E. Segregate data across multiple tables or databases.
F. Use Google Stackdriver Audit Logging to determine policy violations.

Answer: BDF

NEW QUESTION 4
- (Exam Topic 1)
You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

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

Answer: D

Explanation:
https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

NEW QUESTION 5
- (Exam Topic 1)
You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

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

Answer: D

NEW QUESTION 6
- (Exam Topic 1)
Your company’s on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

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

Answer: B

NEW QUESTION 7
- (Exam Topic 1)
Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

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

Answer: C

NEW QUESTION 8
- (Exam Topic 1)
You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP typ
B. Reload the data.
C. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each ro
D. Reference the column TS instead of the column DT from now on.
E. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP value
F. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.
G. Add two columns to the table CLICK STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN typ
H. Reload all data in append mod
I. For each appended row, set the value of IS_NEW to tru
J. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.
K. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP value
L. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP typ
M. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now o
N. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer: D

NEW QUESTION 9
- (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 10
- (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 10
- (Exam Topic 1)
Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

A. Supervised learning to determine which transactions are most likely to be fraudulent.
B. Unsupervised learning to determine which transactions are most likely to be fraudulent.

C. Clustering to divide the transactions into N categories based on feature similarity.
D. Supervised learning to predict the location of a transaction.
E. Reinforcement learning to predict the location of a transaction.
F. Unsupervised learning to predict the location of a transaction.

Answer: BCD

NEW QUESTION 11
- (Exam Topic 1)
You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old. What should you do?

A. Disable caching by editing the report settings.
B. Disable caching in BigQuery by editing table details.
C. Refresh your browser tab showing the visualizations.
D. Clear your browser history for the past hour then reload the tab showing the virtualizations.

Answer: A

Explanation:
Reference https://support.google.com/datastudio/answer/7020039?hl=en

NEW QUESTION 14
- (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 18
- (Exam Topic 2)
Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?

A. Export the data into a Google Sheet for virtualization.
B. Create an additional table with only the necessary columns.
C. Create a view on the table to present to the virtualization tool.
D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

Answer: C

NEW QUESTION 19
- (Exam Topic 2)
Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?

A. Store the common data in BigQuery as partitioned tables.
B. Store the common data in BigQuery and expose authorized views.
C. Store the common data encoded as Avro in Google Cloud Storage.
D. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.

Answer: B

NEW QUESTION 20
- (Exam Topic 3)
MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?

A. Rowkey: date#device_idColumn data: data_point
B. Rowkey: dateColumn data: device_id, data_point
C. Rowkey: device_idColumn data: date, data_point
D. Rowkey: data_pointColumn data: device_id, date
E. Rowkey: date#data_pointColumn data: device_id

Answer: D

NEW QUESTION 21
- (Exam Topic 3)
You need to compose visualizations for operations teams with the following requirements: Which approach meets the requirements?

A. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
C. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Answer: C

NEW QUESTION 23
- (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 27
- (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 30
- (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 33
- (Exam Topic 4)
Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?

A. Rewrite the job in Pig.

B. Rewrite the job in Apache Spark.
C. Increase the size of the Hadoop cluster.
D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Answer: A

NEW QUESTION 35
- (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 39
- (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 43
- (Exam Topic 5)
Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?

A. An hourly watermark
B. An event time trigger
C. The with Allowed Lateness method
D. A processing time trigger

Answer: D

Explanation:
When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.
Processing time triggers. These triggers operate on the processing time – the time when the data element is processed at any given stage in the pipeline.
Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam's default trigger is event time-based.
Reference: https://beam.apache.org/documentation/programming-guide/#triggers

NEW QUESTION 45
- (Exam Topic 5)
The for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

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

Answer: A

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

NEW QUESTION 49
- (Exam Topic 5)

Which TensorFlow function can you use to configure a categorical column if you don't know all of the possible values for that column?

A. categorical_column_with_vocabulary_list
B. categorical_column_with_hash_bucket
C. categorical_column_with_unknown_values
D. sparse_column_with_keys

Answer: B

Explanation:
If you know the set of all possible feature values of a column and there are only a few of them, you can use categorical_column_with_vocabulary_list. Each key in the list will get assigned an auto-incremental ID starting from 0.
What if we don't know the set of possible values in advance? Not a problem. We can use categorical_column_with_hash_bucket instead. What will happen is that each possible value in the feature column occupation will be hashed to an integer ID as we encounter them in training.
Reference: https://www.tensorflow.org/tutorials/wide

NEW QUESTION 50
- (Exam Topic 5)
What are two methods that can be used to denormalize tables in BigQuery?

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

Answer: B

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

NEW QUESTION 54
- (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 55
- (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 59
- (Exam Topic 5)
Suppose you have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error.
SELECT person FROM `project1.example.table1` WHERE city = "London" How would you correct the error?

A. Add ", UNNEST(person)" before the WHERE clause.
B. Change "person" to "person.city".
C. Change "person" to "city.person".
D. Add ", UNNEST(city)" before the WHERE clause.

Answer: A

Explanation:
To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma. Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_resu

NEW QUESTION 63
- (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 66
- (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 67
- (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 71
- (Exam Topic 5)
Why do you need to split a machine learning dataset into training data and test data?

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

Answer: B

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

NEW QUESTION 72
- (Exam Topic 5)
Which of these statements about BigQuery caching is true?

A. By default, a query's results are not cached.
B. BigQuery caches query results for 48 hours.
C. Query results are cached even if you specify a destination table.

D. There is no charge for a query that retrieves its results from cache.

Answer: D

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

NEW QUESTION 77
- (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 82
- (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 83
- (Exam Topic 5)
Does Dataflow process batch data pipelines or streaming data pipelines?

A. Only Batch Data Pipelines
B. Both Batch and Streaming Data Pipelines
C. Only Streaming Data Pipelines
D. None of the above

Answer: B

Explanation:
Dataflow is a unified processing model, and can execute both streaming and batch data pipelines Reference: https://cloud.google.com/dataflow/

NEW QUESTION 85
- (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 89
- (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 90
- (Exam Topic 5)
Which of these operations can you perform from the BigQuery Web UI?

A. Upload a file in SQL format.
B. Load data with nested and repeated fields.
C. Upload a 20 MB file.
D. Upload multiple files using a wildcard.

Answer: B

Explanation:
You can load data with nested and repeated fields using the Web UI. You cannot use the Web UI to:
- Upload a file greater than 10 MB in size
- Upload multiple files at the same time
- Upload a file in SQL format
All three of the above operations can be performed using the "bq" command. Reference: https://cloud.google.com/bigquery/loading-data

NEW QUESTION 95
- (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 97
- (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 101
- (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 103
- (Exam Topic 5)
If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?

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

Answer: A

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

NEW QUESTION 108
- (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 111
- (Exam Topic 5)
Which Google Cloud Platform service is an alternative to Hadoop with Hive?

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

Answer: C

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

NEW QUESTION 112
- (Exam Topic 5)
Cloud Dataproc charges you only for what you really use with billing.

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

Answer: B

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

NEW QUESTION 114
- (Exam Topic 5)

Which of the following statements about Legacy SQL and Standard SQL is not true?

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

Answer: D

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

NEW QUESTION 115
- (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 120
- (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 124
- (Exam Topic 6)
An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application.
They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?

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

Answer: C

Explanation:
Reference: https://cloud.google.com/solutions/business-intelligence/

NEW QUESTION 125
- (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 128
- (Exam Topic 6)
You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

A. Deploy small Kafka clusters in your data centers to buffer events.
B. Have the data acquisition devices publish data to Cloud Pub/Sub.
C. Establish a Cloud Interconnect between all remote data centers and Google.
D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

Answer: B

NEW QUESTION 129
- (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 132
- (Exam Topic 6)
You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?

A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
C. Stop the Cloud Dataflow pipeline with the Cancel optio
D. Create a new Cloud Dataflow job with the updated code
E. Stop the Cloud Dataflow pipeline with the Drain optio
F. Create a new Cloud Dataflow job with the updated code

Answer: A

NEW QUESTION 133
- (Exam Topic 6)
You work on a regression problem in a natural language processing domain, and you have 100M labeled exmaples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

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

Answer: D

NEW QUESTION 135
- (Exam Topic 6)
Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?

A. Perform hyperparameter tuning
B. Train a classifier with deep neural networks, because neural networks would always beat SVMs
C. Deploy the model and measure the real-world AUC; it’s always higher because of generalization
D. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC

Answer: A

Explanation:
https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568

NEW QUESTION 138
- (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 139
- (Exam Topic 6)
You are building a teal-lime prediction engine that streams files, which may contain Pll (personal identifiable information) data, into Cloud Storage and eventually into BigQuery You want to ensure that the sensitive data is masked but still maintains referential Integrity, because names and emails are often used as join keys How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the Pll data is not accessible by unauthorized individuals?

A. Create a pseudonym by replacing the Pll data with cryptogenic tokens, and store the non-tokenized data in a locked-down button.
B. Redact all Pll data, and store a version of the unredacted data in a locked-down bucket
C. Scan every table in BigQuery, and mask the data it finds that has Pll
D. Create a pseudonym by replacing Pll data with a cryptographic format-preserving token

Answer: A

NEW QUESTION 144
- (Exam Topic 6)
Your company currently runs a large on-premises cluster using Spark Hive and Hadoop Distributed File System (HDFS) in a colocation facility. The duster is designed to support peak usage on the system, however, many jobs are batch n nature, and usage of the cluster fluctuates quite dramatically.
Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings. They are also hoping to modernize their existing infrastructure to use more servers offerings m order to take advantage of the cloud Because of the tuning of their contract renewal with the colocation facility they have only 2 months for their initial migration How should you recommend they approach thee upcoming migration strategy so they can maximize their cost savings in the cloud will still executing the migration in time?

A. Migrate the workloads to Dataproc plus HOPS, modernize later
B. Migrate the workloads to Dataproc plus Cloud Storage modernize later
C. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery
D. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery

Answer: D

NEW QUESTION 146
- (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 151
- (Exam Topic 6)
You need to choose a database for a new project that has the following requirements:
- Fully managed
- Able to automatically scale up
- Transactionally consistent
- Able to scale up to 6 TB
- Able to be queried using SQL Which database do you choose?

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

Answer: C

NEW QUESTION 154
- (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: B

NEW QUESTION 156
- (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 157
- (Exam Topic 6)
You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of you Cloud Bigtable cluster. Which two actions can you take to accomplish this? Choose 2 answers.

A. Review Key Visualizer metric
B. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
C. Review Key Visualizer metric
D. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
E. Monitor the latency of write operation
F. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
G. Monitor storage utilizatio
H. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
I. Monitor latency of read operation
J. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

Answer: AC

NEW QUESTION 161
- (Exam Topic 6)
An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and cat the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?

A. Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.
B. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.
C. Write an application that makes a queue in a NoSQL database
D. Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic

Answer: C

NEW QUESTION 165
- (Exam Topic 6)
Your company is migrating its on-premises data warehousing solution to BigQuery. The existing data warehouse uses trigger-based change data capture (CDC) to apply daily updates from transactional database sources Your company wants to use BigQuery to improve its handling of CDC and to optimize the performance of the data warehouse Source system changes must be available for query m near-real time using tog-based CDC streams You need to ensure that changes in the BigQuery reporting table are available with minimal latency and reduced overhead. What should you do? Choose 2 answers

A. Perform a DML INSERT UPDATE, or DELETE to replicate each CDC record in the reporting table m real time.
B. Periodically DELETE outdated records from the reporting tablePeriodically use a DML MERGE to simultaneously perform DML INSER
C. UPDATE, and DELETE operations in the reporting table
D. Insert each new CDC record and corresponding operation type into a staging table in real time
E. Insert each new CDC record and corresponding operation type into the reporting table in real time and use a materialized view to expose only the current version of each unique record.

Answer: BD

NEW QUESTION 167
- (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 regularizationduring optimization.
D. Create a feature cross of latitude and longtitude, bucketize it at the minute level and use L2 regularization during optimization.

Answer: C

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

NEW QUESTION 168
- (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 the users/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: A

NEW QUESTION 170
- (Exam Topic 6)
You need (o give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system How should you design your pipeline to minimize that backpressure?

A. Call out to the service via HTTP
B. Create the pipeline statically in the class definition
C. Create a new object in the startBundle method of DoFn
D. Batch the job into ten-second increments

Answer: A

NEW QUESTION 175
- (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: C

NEW QUESTION 179
- (Exam Topic 6)
You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query – -dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

A. Create a separate table for each ID.
B. Use the LIMIT keyword to reduce the number of rows returned.
C. Recreate the table with a partitioning column and clustering column.
D. Use the bq query - -maximum_bytes_billed flag to restrict the number of bytes billed.

Answer: C

NEW QUESTION 181
- (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: A

Explanation:
you can use dataprep for continuously changing target schema
In general, a target consists of the set of information required to define the expected data in a dataset. Often referred to as a "schema," this target schema

information can include:
Names of columns
Order of columns Column data types Data type format Example rows of data
A dataset associated with a target is expected to conform to the requirements of the schema. Where there are differences between target schema and dataset schema, a validation indicator (or schema tag) is displayed.
https://cloud.google.com/dataprep/docs/html/Overview-of-RapidTarget_136155049

NEW QUESTION 184
- (Exam Topic 6)
You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

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

Answer: C

NEW QUESTION 188
- (Exam Topic 6)
You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

A. Configure your Cloud Dataflow pipeline to use local execution
B. Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions
C. Increase the number of nodes in the Cloud Bigtable cluster
D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable
E. Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable

Answer: BC

NEW QUESTION 190
- (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 193
- (Exam Topic 6)
You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

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

Answer: C

NEW QUESTION 196
- (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 197
- (Exam Topic 6)
Your company needs to upload their historic data to Cloud Storage. The security rules don’t allow access from external IPs to their on-premises resources. After

an initial upload, they will add new data from existing
on-premises applications every day. What should they do?

A. Execute gsutil rsync from the on-premises servers.
B. Use Cloud Dataflow and write the data to Cloud Storage.
C. Write a job template in Cloud Dataproc to perform the data transfer.
D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

Answer: B

NEW QUESTION 201
- (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 205
- (Exam Topic 6)
You are working on a linear regression model on BigQuery ML to predict a customer's likelihood of purchasing your company's products. Your model uses a city name variable as a key predictive component in order to train and serve the model your data must be organized in columns. You want to prepare your data using the least amount of coding while maintaining the predictable variables. What should you do?

A. Use SQL in BigQuery to transform the stale column using a one-hot encoding method, and make each city a column with binary values.
B. Create a new view with BigQuery that does not include a column which city information.
C. Cloud Data Fusion to assign each city to a region that is labeled as 1, 2 3, 4, or 5, and then use that number to represent the city in the model.
D. Use TensorFlow to create a categorical variable with a vocabulary lis
E. Create the vocabulary file and upload that as part of your model to BigQuery ML.

Answer: C

NEW QUESTION 206
- (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 210
- (Exam Topic 6)
You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

A. Use Cloud Bigtable for storag
B. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
C. Use Cloud Bigtable for storag
D. Link as permanent tables in BigQuery for query.
E. Use Cloud Storage for storag
F. Link as permanent tables in BigQuery for query.
G. Use Cloud Storage for storag
H. Link as temporary tables in BigQuery for query.

Answer: A

NEW QUESTION 212
- (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 217
- (Exam Topic 6)
You have uploaded 5 years of log data to Cloud Storage A user reported that some data points in the log data are outside of their expected ranges, which indicates errors You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

A. Import the data from Cloud Storage into BigQuery Create a new BigQuery table, and skip the rows with errors.
B. Create a Compute Engine instance and create a new copy of the data in Cloud Storage Skip the rows with errors
C. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage
D. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage

Answer: D

NEW QUESTION 218
- (Exam Topic 6)
Your company is implementing a data warehouse using BigQuery, and you have been tasked with designing the data model You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?

A. Denormalize the data
B. Shard the data by customer ID
C. Materialize the dimensional data in views
D. Partition the data by transaction date

Answer: C

NEW QUESTION 221
- (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 226
- (Exam Topic 6)
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

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

Answer: A

NEW QUESTION 229
- (Exam Topic 6)
You are building a report-only data warehouse where the data is streamed into BigQuery via the streaming API Following Google's best practices, you have both a staging and a production table for the data How should you design your data loading to ensure that there is only one master dataset without affecting performance on either the ingestion or reporting pieces?

A. Have a staging table that is an append-only model, and then update the production table every three hours with the changes written to staging
B. Have a staging table that is an append-only model, and then update the production table every ninety minutes with the changes written to staging
C. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every three hours

D. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every thirty minutes

Answer: D

NEW QUESTION 230
- (Exam Topic 6)
Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

A. Create a new view over events using standard SQL
B. Create a new partitioned table using a standard SQL query
C. Create a new view over events_partitioned using standard SQL
D. Create a service account for the ODBC connection to use for authentication
E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared “events”

Answer: AE

NEW QUESTION 234
- (Exam Topic 6)
After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You’ve loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.
What should you do?

A. Select random samples from the tables using the RAND() function and compare the samples.
B. Select random samples from the tables using the HASH() function and compare the samples.
C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sortin
D. Compare the hashes of each table.
E. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Answer: B

NEW QUESTION 237
- (Exam Topic 6)
You work for a mid-sized enterprise that needs to move its operational system transaction data from an on-premises database to GCP. The database is about 20 TB in size. Which database should you choose?

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

Answer: A

NEW QUESTION 241
- (Exam Topic 6)
You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?

A. Set the BigQuery dataset to be regiona
B. In the event of an emergency, use a point-in-time snapshot to recover the data.
C. Set the BigQuery dataset to be regiona
D. Create a scheduled query to make copies of the data to tables suffixed with the time of the backu
E. In the event of an emergency, use the backup copy of the table.
F. Set the BigQuery dataset to be multi-regiona
G. In the event of an emergency, use a point-in-time snapshot to recover the data.
H. Set the BigQuery dataset to be multi-regiona
I. Create a scheduled query to make copies of the data to tables suffixed with the time of the backu
J. In the event of an emergency, use the backup copy of the table.

Answer: B

NEW QUESTION 244
- (Exam Topic 6)
You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.
You have the following requirements:
- You will batch-load the posts once per day and run them through the Cloud Natural Language API.
- You will extract topics and sentiment from the posts.
- You must store the raw posts for archiving and reprocessing.
- You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

A. Store the social media posts and the data extracted from the API in BigQuery.

B. Store the social media posts and the data extracted from the API in Cloud SQL.
C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.
D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Answer: D

NEW QUESTION 246
- (Exam Topic 6)
You've migrated a Hadoop job from an on-premises cluster to Dataproc and Good Storage. Your Spark job is a complex analytical workload fiat consists of many shuffling operations, and initial data are parquet toes (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. Your organization is very cost-sensitive so you'd Idee to continue using Dataproc on preemptibles (with 2 non-preemptibles workers only) for this workload. What should you do?

A. Switch from HODs to SSDs override the preemptible VMs configuration to increase the boot disk size
B. Increase the see of your parquet files to ensure them to be 1 GB minimum
C. Switch to TFRecords format (appr 200 MB per We) instead of parquet files
D. Switch from HDDs to SSD
E. copy initial data from Cloud Storage to Hadoop Distributed File System (HDFS) run the Spark job and copy results back to Cloud Storage

Answer: A

NEW QUESTION 250
- (Exam Topic 6)
You want to build a managed Hadoop system as your data lake. The data transformation process is composed of a series of Hadoop jobs executed in sequence. To accomplish the design of separating storage from compute, you decided to use the Cloud Storage connector to store all input data, output data, and intermediary data. However, you noticed that one Hadoop job runs very slowly with Cloud Dataproc, when compared with the on-premises bare-metal Hadoop environment (8-core nodes with 100-GB RAM). Analysis shows that this particular Hadoop job is disk I/O intensive. You want to resolve the issue. What should you do?

A. Allocate sufficient memory to the Hadoop cluster, so that the intermediary data of that particular Hadoop job can be held in memory
B. Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS
C. Allocate more CPU cores of the virtual machine instances of the Hadoop cluster so that the networking bandwidth for each instance can scale up
D. Allocate additional network interface card (NIC), and configure link aggregation in the operating system to use the combined throughput when working with Cloud Storage

Answer: A

NEW QUESTION 255
.....

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