Google

Exam Questions Professional-Data-Engineer

Google Professional Data Engineer Exam

About Exambible

Your Partner of IT Exam

Found in 1998

Exambible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, Exambible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

All examinations will be up to date.

* 24/7 Quality Support

We will provide service round the clock.

* 100% Pass Rate

Our guarantee that you will pass the exam.

* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

NEW QUESTION 1
- (Exam Topic 1)
Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

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

Answer: C

NEW QUESTION 2
- (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 numericvalues 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 3
- (Exam Topic 1)
Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three.)

A. Load data into different partitions.
B. Load data into a different dataset for each client.
C. Put each client's BigQuery dataset into a different table.
D. Restrict a client's dataset to approved users.
E. Only allow a service account to access the datasets.
F. Use the appropriate identity and access management (IAM) roles for each client's users.

Answer: BDF

NEW QUESTION 4
- (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 5
- (Exam Topic 1)
Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

A. Redefine the schema by evenly distributing reads and writes across the row space of the table.
B. The performance issue should be resolved over time as the site of the BigDate cluster is increased.
C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

Answer: A

NEW QUESTION 6
- (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 7
- (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 8
- (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 9
- (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 10
- (Exam Topic 3)
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)

A. Ensure all the tables are included in global dataset.
B. Ensure each table is included in a dataset for a region.
C. Adjust the settings for each table to allow a related region-based security group view access.
D. Adjust the settings for each view to allow a related region-based security group view access.
E. Adjust the settings for each dataset to allow a related region-based security group view access.

Answer: BD

NEW QUESTION 10
- (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 11
- (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 14
- (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 19
- (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 21
- (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 25
- (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 29
- (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 30
- (Exam Topic 5)
To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?

A. gcloud ml-engine local train
B. gcloud ml-engine jobs submit training
C. gcloud ml-engine jobs submit training local
D. You can't run a TensorFlow program on your own computer using Cloud ML Engine .

Answer: A

Explanation:
gcloud ml-engine local train - run a Cloud ML Engine training job locally
This command runs the specified module in an environment similar to that of a live Cloud ML Engine Training Job.
This is especially useful in the case of testing distributed models, as it allows you to validate that you are
properly interacting with the Cloud ML Engine cluster configuration. Reference: https://cloud.google.com/sdk/gcloud/reference/ml-engine/local/train

NEW QUESTION 33
- (Exam Topic 5)
Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?

A. Use K-means Clustering to detect faces in the pixels.
B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.

Answer: C

Explanation:
Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as “deep” learning. So deep is a strictly defined, technical term that means more than one hidden layer.
In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer’s output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the
previous layer.
A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.
Feature engineering is difficult to perform on raw image data.
K- means Clustering is an unsupervised learning method used to categorize unlabeled data. Reference: https://deeplearning4j.org/neuralnet-overview

NEW QUESTION 38
- (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 43
- (Exam Topic 5)
How would you query specific partitions in a BigQuery table?

A. Use the DAY column in the WHERE clause
B. Use the EXTRACT(DAY) clause
C. Use the PARTITIONTIME pseudo-column in the WHERE clause
D. Use DATE BETWEEN in the WHERE clause

Answer: C

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

NEW QUESTION 45
- (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 49
- (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 50
- (Exam Topic 5)
You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output.
Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

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

Answer: B

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

NEW QUESTION 51
- (Exam Topic 5)
When a Cloud Bigtable node fails, is lost.

A. all data
B. no data

C. the last transaction
D. the time dimension

Answer: B

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

NEW QUESTION 56
- (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 61
- (Exam Topic 5)
Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

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

Answer: AB

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

NEW QUESTION 64
- (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 67
- (Exam Topic 5)
Which SQL keyword can be used to reduce the number of columns processed by BigQuery?

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

Answer: C

Explanation:

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

NEW QUESTION 68
- (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 71
- (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 76
- (Exam Topic 6)
As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects. Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take? Choose 2 answers.

A. Use Cloud Deployment Manager to automate access provision.
B. Introduce resource hierarchy to leverage access control policy inheritance.
C. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
E. For each Cloud Storage bucket or BigQuery dataset, decide which projects need acces
F. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.

Answer: AC

NEW QUESTION 77
- (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 82
- (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 83
- (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 85
- (Exam Topic 6)
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

A. Create a table in BigQuery, and append the new samples for CPU and memory to the table
B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second
C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second
D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.

Answer: D

NEW QUESTION 89
- (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 90
- (Exam Topic 6)
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

A. Use Cloud Dataproc to run your transformation
B. Monitor CPU utilization for the cluste
C. Resize the number of worker nodes in your cluster via the command line.
D. Use Cloud Dataproc to run your transformation
E. Use the diagnose command to generate an operational output archiv
F. Locate the bottleneck and adjust cluster resources.
G. Use Cloud Dataflow to run your transformation
H. Monitor the job system lag with Stackdrive
I. Use the default autoscaling setting for worker instances.
J. Use Cloud Dataflow to run your transformation
K. Monitor the total execution time for a sampling of job
L. Configure the job to use non-default Compute Engine machine types when needed.

Answer: B

NEW QUESTION 92
- (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 97
- (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 98
- (Exam Topic 6)
You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the Data Science team runs a query filtered on a date column and limited to 30–90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. What should you do?

A. Re-create the tables using DD
B. Partition the tables by a column containing a TIMESTAMP or DATE Type.
C. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.
D. Modify your pipeline to maintain the last 30–90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.
E. Write an Apache Beam pipeline that creates a BigQuery table per da
F. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

Answer: C

NEW QUESTION 102
- (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 105
- (Exam Topic 6)
Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects. What should you do?

A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned_bytes
B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated_for_project
C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric
D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

Answer: D

NEW QUESTION 107
- (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 111
- (Exam Topic 6)
You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? Choose 2 answers.

A. Denormalize the data as must as possible.
B. Preserve the structure of the data as much as possible.
C. Use BigQuery UPDATE to further reduce the size of the dataset.
D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro fil
F. Use BigQuery'ssupport for external data sources to query.

Answer: DE

NEW QUESTION 112
- (Exam Topic 6)
You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:
- Decoupling producer from consumer
- Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
- Near real-time SQL query
- Maintain at least 2 years of historical data, which will be queried with SQ

Which pipeline should you use to meet these requirements?

A. Create an application that provides an AP
B. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
C. Create an application that writes to a Cloud SQL database to store the dat
D. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
E. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
F. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.

Answer: A

NEW QUESTION 115
- (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 120
- (Exam Topic 6)
You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

A. Create a cron schedule in Cloud Dataprep.
B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.

C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.
D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.

Answer: C

NEW QUESTION 122
- (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 125
- (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 126
- (Exam Topic 6)
Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

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

Answer: D

NEW QUESTION 129
- (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 130
- (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 134
- (Exam Topic 6)
Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

A. Set a single global window to capture all the data.
B. Set sliding windows to capture all the lagged data.
C. Use watermarks and timestamps to capture the lagged data.
D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

Answer: B

NEW QUESTION 138
- (Exam Topic 6)
You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?

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

Answer: A

NEW QUESTION 142
- (Exam Topic 6)
You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.
What should you do?

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

Answer: A

NEW QUESTION 143
- (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 146
- (Exam Topic 6)
You have an Apache Kafka Cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins.
What should you do?

A. Deploy a Kafka cluster on GCE VM Instance
B. Configure your on-prem cluster to mirror your topics tothe cluster running in GC
C. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
D. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connecto
E. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
F. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connecto
G. Use a Dataflow job to read fron PubSub and write to GCS.
H. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connecto
I. Use a Dataflow job to read fron PubSub and write to GCS.

Answer: A

NEW QUESTION 148
- (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 149
.....

Relate Links

100% Pass Your Professional-Data-Engineer Exam with Exambible Prep Materials

https://www.exambible.com/Professional-Data-Engineer-exam/

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - https://www.exambible.com/