Free Questions for Professional-Data-Engineer by certsdeals

Shared by Vargas on 29-01-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

You are loading CSV files from Cloud Storage to BigQuery. The files have known data quality issues, including mismatched data types, such as STRINGS and INT64s in the same column, and inconsistent formatting of values such as phone numbers or addresses. You need to create the data pipeline to maintain data quality and perform the required cleansing and transformation. What should you do?

Options:

A- Use Data Fusion to transform the data before loading it into BigQuery.

B- Load the CSV files into a staging table with the desired schema, perform the transformations with SQL. and then write the results to the final destination table.

C- Create a table with the desired schema, toad the CSV files into the table, and perform the transformations in place using SQL.

D- Use Data Fusion to convert the CSV files lo a self-describing data formal, such as AVRO. before loading the data to BigOuery.

Answer:

A

Explanation:

Data Fusion's advantages:

Visual interface: Offers a user-friendly interface for designing data pipelines without extensive coding, making it accessible to a wider range of users.

Built-in transformations: Includes a wide range of pre-built transformations to handle common data quality issues, such as:

Data type conversions

Data cleansing (e.g., removing invalid characters, correcting formatting)

Data validation (e.g., checking for missing values, enforcing constraints)

Data enrichment (e.g., adding derived fields, joining with other datasets)

Custom transformations: Allows for custom transformations using SQL or Java code for more complex cleaning tasks.

Scalability: Can handle large datasets efficiently, making it suitable for processing CSV files with potential data quality issues.

Integration with BigQuery: Integrates seamlessly with BigQuery, allowing for direct loading of transformed data.

Question 2

Question Type: **MultipleChoice**

You are implementing a chatbot to help an online retailer streamline their customer service. The chatbot must be able to respond to both text and voice inquiries. You are looking for a low-code or no-code option, and you want to be able to easily train the chatbot to provide answers to keywords. What should you do?

Options:

A- Use the Speech-to-Text API to build a Python application in App Engine.

B- Use the Speech-to-Text API to build a Python application in a Compute Engine instance.

C- Use Dialogflow for simple queries and the Speech-to-Text API for complex queries.

D- Use Dialogflow to implement the chatbot. defining the intents based on the most common queries collected.

Answer:

D

Explanation:

Dialogflow is a conversational AI platform that allows for easy implementation of chatbots without needing to code. It has built-in integration for both text and voice input via APIs like Cloud Speech-to-Text. Defining intents and entity types allows you to map common queries and keywords to responses. This would provide a low/no-code way to quickly build and iteratively improve the chatbot capabilities.

https://cloud.google.com/dialogflow/docs Dialogflow is a natural language understanding platform that makes it easy to design and integrate a conversational user interface into your mobile app, web application, device, bot, interactive voice response system, and so on. Using Dialogflow, you can provide new and engaging ways for users to interact with your product. Dialogflow can analyze multiple types of input from your customers, including text or audio inputs (like from a phone or voice recording). It can also respond to your customers in a couple of ways, either through text or with synthetic speech.

Question 3

Question Type: MultipleChoice

You are implementing workflow pipeline scheduling using open source-based tools and Google Kubernetes Engine (GKE). You want to use a Google managed service to simplify and automate the task. You also want to accommodate Shared VPC networking considerations. What should you do?

Options:

A- Use Dataflow for your workflow pipelines. Use Cloud Run triggers for scheduling.

B- Use Dataflow for your workflow pipelines. Use shell scripts to schedule workflows.

C- Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the host project.

D- Use Cloud Composer in a Shared VPC configuration. Place the Cloud Composer resources in the service project.

Answer:

D

Explanation:

Shared VPC requires that you designate a host project to which networks and subnetworks belong and a service project, which is attached to the host project. When Cloud Composer participates in a Shared VPC, the Cloud Composer environment is in the service project. Reference: https://cloud.google.com/composer/docs/how-to/managing/configuring-shared-vpc

Question 4

Question Type: MultipleChoice

You are developing a new deep teaming model that predicts a customer's likelihood to buy on your ecommerce site. Alter running an evaluation of the model against both the original training data and new test data, you find that your model is overfitting the dat

a. You want to improve the accuracy of the model when predicting new data. What should you do?

Options:

A- Increase the size of the training dataset, and increase the number of input features.

B- Increase the size of the training dataset, and decrease the number of input features.

C- Reduce the size of the training dataset, and increase the number of input features.

D- Reduce the size of the training dataset, and decrease the number of input features.

Answer:

B

Explanation:

https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estimates/

Question 5

Question Type: **MultipleChoice**

You issue a new batch job to Dataflow. The job starts successfully, processes a few elements, and then suddenly fails and shuts down. You navigate to the Dataflow monitoring interface where you find errors related to a particular DoFn in your pipeline. What is the most

likely cause of the errors?

Options:

A- Exceptions in worker code

B- Job validation

C- Graph or pipeline construction

D- Insufficient permissions

Answer:

A

Explanation:

https://cloud.google.com/dataflow/docs/guides/troubleshooting-your-pipeline#detect_an_exception_in_worker_code While your job is running, you might encounter errors or exceptions in your worker code. These errors generally mean that the DoFns in your pipeline code have generated unhandled exceptions, which result in failed tasks in your Dataflow job. Exceptions in user code (for example, your DoFn instances) are reported in the Dataflow monitoring interface.

Question 6

Question Type: MultipleChoice

You have a data processing application that runs on Google Kubernetes Engine (GKE). Containers need to be launched with their latest available configurations from a container registry. Your GKE nodes need to have GPUs. local SSDs, and 8 Gbps bandwidth. You want to efficiently provision the data processing infrastructure and manage the deployment process. What should you do?

Options:

A- Use Compute Engi.no startup scriots to pull container Images, and use gloud commands to provision the infrastructure.

B- Use GKE to autoscale containers, and use gloud commands to provision the infrastructure.

C- Use Cloud Build to schedule a job using Terraform build to provision the infrastructure and launch with the most current container images.

D- Use Dataflow to provision the data pipeline, and use Cloud Scheduler to run the job.

Answer:

C

Explanation:

https://cloud.google.com/architecture/managing-infrastructure-as-code

Question 7

Question Type: MultipleChoice

You want to create a machine learning model using BigQuery ML and create an endpoint foe hosting the model using Vertex Al. This will enable the processing of continuous streaming data in near-real time from multiple vendors. The data may contain invalid values. What should you do?

Options:

A- Create a new BigOuery dataset and use streaming inserts to land the data from multiple vendors. Configure your BigQuery ML model to use the 'ingestion' dataset as the training data.

B- Use BigQuery streaming inserts to land the data from multiple vendors whore your BigQuery dataset ML model is deployed.

C- Create a Pub'Sub topic and send all vendor data to it Connect a Cloud Function to the topic to process the data and store it in BigQuery.

D- Create a Pub/Sub topic and send all vendor data to it Use Dataflow to process and sanitize the Pub/Sub data and stream it to BigQuery.

Answer:

D

Explanation:

Dataflow provides a scalable and flexible way to process and clean the incoming data in real-time before loading it into BigQuery.

Question 8

Question Type: MultipleChoice

You need to migrate a Redis database from an on-premises data center to a Memorystore for Redis instance. You want to follow Google-recommended practices and perform the migration for minimal cost. time, and effort. What should you do?

Options:

A- Make a secondary instance of the Redis database on a Compute Engine instance, and then perform a live cutover.

B- Write a shell script to migrate the Redis data, and create a new Memorystore for Redis instance.

C- Create a Dataflow job to road the Redis database from the on-premises data center. and write the data to a Memorystore for Redis

instance

D- Make an RDB backup of the Redis database, use the gsutil utility to copy the RDB file into a Cloud Storage bucket, and then import the RDB tile into the Memorystore for Redis instance.

Answer:

D

Explanation:

The import and export feature uses the native RDB snapshot feature of Redis to import data into or export data out of a Memorystore for Redis instance. The use of the native RDB format prevents lock-in and makes it very easy to move data within Google Cloud or outside of Google Cloud. Import and export uses Cloud Storage buckets to store RDB files. Reference: https://cloud.google.com/memorystore/docs/redis/import-export-overview

Question 9

Question Type: MultipleChoice

You are designing a system that requires an ACID-compliant database. You must ensure that the system requires minimal human intervention in case of a failure. What should you do?

Options:

A- Configure a Cloud SQL for MySQL instance with point-in-time recovery enabled.

B- Configure a Cloud SQL for PostgreSQL instance with high availability enabled.

C- Configure a Bigtable instance with more than one cluster.

D- Configure a BJgQuery table with a multi-region configuration.

Answer:

B

Explanation:

The best option to meet the ACID compliance and minimal human intervention requirements is to configure a Cloud SQL for PostgreSQL instance with high availability enabled. Key reasons: Cloud SQL for PostgreSQL provides full ACID compliance, unlike Bigtable which provides only atomicity and consistency guarantees. Enabling high availability removes the need for manual failover as Cloud SQL will automatically failover to a standby replica if the leader instance goes down. Point-in-time recovery in MySQL requires manual intervention to restore data if needed. BigQuery does not provide transactional guarantees required for an ACID database. Therefore, a Cloud SQL for PostgreSQL instance with high availability meets the ACID and minimal intervention requirements best. The automatic failover will ensure availability and uptime without administrative effort.

Question 10

Question Type: MultipleChoice

Your startup has a web application that currently serves customers out of a single region in Asi

a. You are targeting funding that will allow your startup lo serve customers globally. Your current goal is to optimize for cost, and your post-funding goat is to optimize for global presence and performance. You must use a native JDBC driver. What should you do?

Options:

A- Use Cloud Spanner to configure a single region instance initially. and then configure multi-region C oud Spanner instances after securing funding.

B- Use a Cloud SQL for PostgreSQL highly available instance first, and 8gtable with US. Europe, and Asia
replication alter securing funding

C- Use a Cloud SQL for PostgreSQL zonal instance first and Bigtable with US. Europe, and Asia after securing funding.

D- Use a Cloud SOL for PostgreSQL zonal instance first, and Cloud SOL for PostgreSQL with highly available configuration after securing funding.

Answer:

A

Explanation:

https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs_regional_versus_multi-region_configurations

Question 11

Question Type: MultipleChoice

You have a data pipeline with a Dataflow job that aggregates and writes time series metrics to Bigtable. You notice that data is slow to update in 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 dat

a. What should you do?

Choose 2 answers

Options:

A- Configure your Dataflow pipeline to use local execution.

B- Modify your Dataflow pipeline lo use the Flatten transform before writing to Bigtable.

C- Modify your Dataflow pipeline to use the CoGrcupByKey transform before writing to Bigtable.

D- Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions.

E- Increase the number of nodes in the Bigtable cluster.

Answer:

D, E

Explanation:

https://cloud.google.com/bigtable/docs/performance#performance-write-throughput

https://cloud.google.com/dataflow/docs/reference/pipeline-options