

Google.Professional-Machine-Learning-Engineer.v2026-10-09.q179

Exam Code:	Professional-Machine-Learning-Engineer
Exam Name:	Google Professional Machine Learning Engineer
Certification Provider:	Google
Free Question Number:	179
Version:	v2026-10-09
# of views:	182
# of Questions views:	1790
https://www.freepdfdumps.com/Google.Professional-Machine-Learning-Engineer.v2026-10-09.q179.html	

NEW QUESTION: 1

You work for a manufacturing company. You need to train a custom image classification model to detect product defects at the end of an assembly line. Although your model is performing well, some images in your holdout set are consistently mislabeled with high confidence. You want to use Vertex AI to understand your model's results. What should you do?

- A. Create an index by using Vertex AI Matching Engine. Query the index with your mislabeled images.
- B. Configure feature-based explanations by using Integrated Gradients. Set visualization type to PIXELS, and set clip_percent_upperbound to 95.
- C. Configure example-based explanations. Specify the embedding output layer to be used for the latent space representation.
- D. Configure feature-based explanations by using XRAI. Set visualization type to OUTLINES, and set polarity to positive.

Answer: B (LEAVE A REPLY)

NEW QUESTION: 2

You have been asked to productionize a proof-of-concept ML model built using Keras. The model was trained in a Jupyter notebook on a data scientist's local machine. The notebook contains a cell that performs data validation and a cell that performs model analysis. You need to orchestrate the steps contained in the notebook and automate the execution of these steps for weekly retraining. You expect much more training data in the future. You want your solution to take advantage of managed services while minimizing cost. What should you do?

- A. Move the Jupyter notebook to a Notebooks instance on the largest N2 machine type, and schedule the execution of the steps in the Notebooks instance using Cloud Scheduler.
- B. Write the code as a TensorFlow Extended (TFX) pipeline orchestrated with Vertex AI Pipelines. Use standard TFX components for data validation and model analysis, and use Vertex AI Pipelines for model retraining.
- C. Rewrite the steps in the Jupyter notebook as an Apache Spark job, and schedule the execution of the job on ephemeral Dataproc clusters using Cloud Scheduler.
- D. Extract the steps contained in the Jupyter notebook as Python scripts, wrap each script in an Apache Airflow BashOperator, and run the resulting directed acyclic graph (DAG) in Cloud Composer.

Answer: B (LEAVE A REPLY)

The reason for this choice is that TFX and Vertex AI Pipelines provide a scalable and cost-effective solution for productionizing machine learning models. TFX is an end-to-end ML platform for building scalable and repeatable ML workflows, while Vertex AI Pipelines provides a fully managed service for orchestrating ML workflows at scale. By using TFX and Vertex AI Pipelines, you can automate the execution of the steps contained in the Jupyter notebook, and schedule the pipeline for weekly retraining. This approach also takes advantage of managed services, which helps to minimize cost.

NEW QUESTION: 3

You work for an organization that operates a streaming music service.

You have a custom production model that is serving a next song recommendation based on a user's recent listening history.

Your model is deployed on a Vertex AI endpoint. You recently retrained the same model by using fresh data.

The model received positive test results offline. You now want to test the new model in production while minimizing complexity.

What should you do?

- A.** Configure a model monitoring job for the existing Vertex AI endpoint. Configure the monitoring job to detect prediction drift and set a threshold for alerts. Update the model on the endpoint from the previous model to the new model. If you receive an alert of prediction drift, revert to the previous model.
- B.** Create a new Vertex AI endpoint for the new model and deploy the new model to that new endpoint. Build a service to randomly send 5% of production traffic to the new endpoint. Monitor end-user metrics such as listening time. If end-user metrics improve between models over time, gradually increase the percentage of production traffic sent to the new endpoint.
- C.** Deploy the new model to the existing Vertex AI endpoint. Use traffic splitting to send 5% of production traffic to the new model. Monitor end-user metrics, such as listening time. If end-user metrics improve between models over time, gradually increase the percentage of production traffic sent to the new model.
- D.** Capture incoming prediction requests in BigQuery. Create an experiment in Vertex AI Experiments. Run batch predictions for both models using the captured data. Use the user's selected song to compare the models performance side by side. If the new model's performance metrics are better than the previous model, deploy the new model to production.

Answer: C ([LEAVE A REPLY](#))

NEW QUESTION: 4

You work for a semiconductor manufacturing company. You need to create a real-time application that automates the quality control process. High-definition images of each semiconductor are taken at the end of the assembly line in real time. The photos are uploaded to a Cloud Storage bucket along with tabular data that includes each semiconductor's batch number, serial number, dimensions, and weight. You need to configure model training and serving while maximizing model accuracy. What should you do?

- A.** Use Vertex AI Data Labeling Service to label the images, and train an AutoML image classification model. Schedule a daily batch prediction job that publishes a Pub/Sub message when the job completes.

- B.** Import the tabular data into BigQuery, use Vertex AI Data Labeling Service to label the data and train an AutoML tabular classification model. Deploy the model, and configure Pub/Sub to publish a message when a semiconductor's data is categorized into the failing class.
- C.** Convert the images into an embedding representation. Import this data into BigQuery, and train a BigQuery ML K-means clustering model with two clusters. Deploy the model and configure Pub/Sub to publish a message when a semiconductor's data is categorized into the failing cluster.
- D.** Use Vertex AI Data Labeling Service to label the images, and train an AutoML image classification model. Deploy the model, and configure Pub/Sub to publish a message when an image is categorized into the failing class.

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 5

You are an AI engineer with an apparel retail company. The sales team has observed seasonal sales patterns over the past 5-6 years. The sales team analyzes and visualizes the weekly sales data stored in CSV files. You have been asked to estimate weekly sales for future seasons to optimize inventory and personnel workloads. You want to use the most efficient approach. What should you do?

- A.** Upload the files into Cloud Storage. Use Python to preprocess and load the tabular data into BigQuery. Use time series forecasting models to predict weekly sales.
- B.** Upload the files into Cloud Storage. Use Python to preprocess and load the tabular data into BigQuery. Train a logistic regression model by using BigQuery ML to predict each product's weekly sales as one of three categories: high, medium, or low.
- C.** Load the files into BigQuery. Preprocess data by using BigQuery SQL. Connect BigQuery to Looker. Create a Looker dashboard that shows weekly sales trends in real time and can slice and dice the data based on relevant filters.
- D.** Create a custom conversational application using Vertex AI Agent Builder. Include code that enables file upload functionality, and upload the files. Use few-shot prompting and retrieval-augmented generation (RAG) to predict future sales trends by using the Gemini large language model (LLM).

Answer: (SHOW ANSWER)

Uploading data to Cloud Storage and then into BigQuery for efficient querying and preprocessing enables scalable data management. Using time series forecasting models on this structured data is the most effective method to predict future weekly sales, capturing seasonal patterns and trends accurately.

NEW QUESTION: 6

You are building an application that extracts information from invoices and receipts. You want to implement this application with minimal custom code and training. What should you do?

- A.** Use the Cloud Vision API with TEXT_DETECTION type to extract text from the invoices and receipts, and use a pre-built natural language processing (NLP) model to parse the extracted text.
- B.** Use the Cloud Document AI API to extract information from the invoices and receipts.
- C.** Use Vertex AI Agent Builder with the pre-built Layout Parser model to extract information from the invoices and receipts.
- D.** Train an AutoML Natural Language model to classify and extract information from the invoices and receipts.

Answer: B ([LEAVE A REPLY](#))

The Cloud Document AI API is specifically designed for processing and extracting structured information from documents such as invoices and receipts. It provides pre-trained models that can recognize and extract relevant fields (e.g., invoice number, date, total amount) with minimal custom code or training. This solution requires minimal setup and is highly efficient for structured documents, making it the ideal choice for this use case.

NEW QUESTION: 7

You are building an ML model to detect anomalies in real-time sensor data. You will use Pub/Sub to handle incoming requests. You want to store the results for analytics and visualization. How should you configure the pipeline?

- A. 1 = Dataflow, 2 = AI Platform, 3 = BigQuery
- B. 1 = DataProc, 2 = AutoML, 3 = Cloud Bigtable
- C. 1 = BigQuery, 2 = AutoML, 3 = Cloud Functions
- D. 1 = BigQuery, 2 = AI Platform, 3 = Cloud Storage

Answer: ([SHOW ANSWER](#))

Dataflow for the pipeline, BigQuery for storing and visualization and Platform AI (now Vertex) to build the model.

<https://cloud.google.com/architecture/building-anomaly-detection-dataflow-bigqueryml-dlp>

NEW QUESTION: 8

A regulator requires your bank to provide, for each declined loan application, the specific applicant attributes that most influenced the decision. Your model is a deep neural network deployed to an online endpoint. You need to satisfy this requirement without replacing the model.

What should you do?

- A. Publish the model's global feature importance scores computed during training.
- B. Configure the deployed model with Vertex Explainable AI using integrated gradients, and return per-prediction feature attributions with each response.
- C. Train a separate logistic regression model on the same features, and use its coefficients to explain the neural network's decisions.
- D. Log all input features to BigQuery, and let analysts inspect the raw values for declined applications.

Answer: B ([LEAVE A REPLY](#))

Integrated gradients produce local, per-instance attributions that identify which input features drove an individual prediction, which is what the regulator requires. Global importance and surrogate models describe overall or approximate behavior and do not faithfully explain a specific decision.

NEW QUESTION: 9

You work for a large retailer and have been asked to segment your customers by their purchasing habits. The purchase history of all customers has been uploaded to BigQuery. You suspect that there may be several distinct customer segments, however you are unsure of how many, and you don't yet understand the commonalities in their behavior. You want to find the most efficient solution. What should you do?

- A. Create a k-means clustering model using BigQuery ML. Allow BigQuery to automatically optimize the number of clusters.
- B. Create a new dataset in Dataprep that references your BigQuery table. Use Dataprep to identify similarities within each column.

C. Use the Data Labeling Service to label each customer record in BigQuery. Train a model on your labeled data using AutoML Tables. Review the evaluation metrics to understand whether there is an underlying pattern in the data.

D. Get a list of the customer segments from your company's Marketing team. Use the Data Labeling Service to label each customer record in BigQuery according to the list. Analyze the distribution of labels in your dataset using Data Studio.

Answer: A ([LEAVE A REPLY](#))

when to use k-means : Your data may contain natural groupings or clusters of data. You may want to identify these groupings descriptively in order to make data-driven decisions. For example, a retailer may want to identify natural groupings of customers who have similar purchasing habits or locations. This process is known as customer segmentation.

<https://cloud.google.com/bigquery/docs/kmeans-tutorial>

NEW QUESTION: 10

You have developed a BigQuery ML model that predicts customer churn, and deployed the model to Vertex AI Endpoints. You want to automate the retraining of your model by using minimal additional code when model feature values change. You also want to minimize the number of times that your model is retrained to reduce training costs. What should you do?

A. 1. Enable request-response logging on Vertex AI Endpoints

2. Schedule a TensorFlow Data Validation job to monitor prediction drift

3. Execute model retraining if there is significant distance between the distributions

B. 1. Create a Vertex AI Model Monitoring job configured to monitor prediction drift

2. Configure alert monitoring to publish a message to a Pub/Sub queue when a monitoring alert is detected

3. Use a Cloud Function to monitor the Pub/Sub queue, and trigger retraining in BigQuery

C. 1. Create a Vertex AI Model Monitoring job configured to monitor training/serving skew

2. Configure alert monitoring to publish a message to a Pub/Sub queue when a monitoring alert is detected

3. Use a Cloud Function to monitor the Pub/Sub queue, and trigger retraining in BigQuery

D. 1. Enable request-response logging on Vertex AI Endpoints

2. Schedule a TensorFlow Data Validation job to monitor training/serving skew

3. Execute model retraining if there is significant distance between the distributions

Answer: ([SHOW ANSWER](#)**)**

NEW QUESTION: 11

You are developing a demand forecasting model for a large online retailer. The company has 20,000 SKUs that are sold through a single distribution center. You have three years of historical data. You need to provide a daily forecast at the SKU level for the next two weeks. You need to develop the first version quickly and minimize the development effort while delivering high forecasting accuracy. What should you do?

A. Train a multivariate BigQuery ML ARIMA_PLUS_XREG model on all SKUs by using the TIME_SERIES_ID_COL option.

B. Write a Python script in Agent Platform Workbench that trains a BigQuery ML ARIMA model for each SKU, and use a corresponding model for each SKU forecast.

C. For each SKU, send the last month's history of sales and the SKU's description to Gemini Pro to generate a prediction for the next two weeks. Instruct the LLM to act as a category specialist in retail and to focus on accuracy.

D. Fine-tune the Gemini Pro model on the historical data to predict the daily sales for each SKU. Use the fine-tuned model without prompt engineering.

Answer: ([SHOW ANSWER](#))

BigQuery ML can train and manage separate time-series forecasts for all 20,000 SKUs in a single SQL model by using the SKU identifier as TIME_SERIES_ID_COL. ARIMA_PLUS_XREG automates model selection, trend and seasonality handling, and supports additional explanatory variables, providing a scalable, accurate solution with minimal development effort.

Reference:

<https://docs.cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-syntax-create-multivariate-time-series>

<https://docs.cloud.google.com/bigquery/docs/arima-plus-xreg-multiple-time-series-forecasting-tutorial>

NEW QUESTION: 12

You work on a growing team of more than 50 data scientists who all use AI Platform. You are designing a strategy to organize your jobs, models, and versions in a clean and scalable way.

Which strategy should you choose?

A. Set up restrictive IAM permissions on the AI Platform notebooks so that only a single user or group can access a given instance.

B. Separate each data scientist's work into a different project to ensure that the jobs, models, and versions created by each data scientist are accessible only to that user.

C. Use labels to organize resources into descriptive categories.

Apply a label to each created resource so that users can filter the results by label when viewing or monitoring the resources.

D. Set up a BigQuery sink for Cloud Logging logs that is appropriately filtered to capture information about AI Platform resource usage. In BigQuery, create a SQL view that maps users to the resources they are using

Answer: C ([LEAVE A REPLY](#))

As IAM roles are given to the entire AI Notebook resource, not to a specific instance.

NEW QUESTION: 13

You developed a Vertex AI ML pipeline that consists of preprocessing and training steps and each set of steps runs on a separate custom Docker image. Your organization uses GitHub and GitHub Actions as CI/CD to run unit and integration tests. You need to automate the model retraining workflow so that it can be initiated both manually and when a new version of the code is merged in the main branch. You want to minimize the steps required to build the workflow while also allowing for maximum flexibility. How should you configure the CI/CD workflow?

A. Trigger a Cloud Build workflow to run tests, build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

B. Trigger GitHub Actions to run the tests, launch a job on Cloud Run to build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

C. Trigger GitHub Actions to run the tests, build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

D. Trigger GitHub Actions to run the tests, launch a Cloud Build workflow to build custom Docker images, push the images to Artifact Registry, and launch the pipeline in Vertex AI Pipelines.

Answer: C ([LEAVE A REPLY](#))

NEW QUESTION: 14

You need to train a ControlNet model with Stable Diffusion XL for an image editing use case. You want to train this model as quickly as possible. Which hardware configuration should you choose to train your model?

A. Configure one a2-highgpu-1g instance with an NVIDIA A100 GPU with 80 GB of RAM. Use float32precision during model training.

B. Configure one a2-highgpu-1g instance with an NVIDIA A100 GPU with 80 GB of RAM. Use bfloat16quantization during model training.

C. Configure four n1-standard-16 instances, each with one NVIDIA Tesla T4 GPU with 16 GB of RAM. Use float32precision during model training.

D. Configure four n1-standard-16 instances, each with one NVIDIA Tesla T4 GPU with 16 GB of RAM. Use float16quantization during model training.

Answer: B ([LEAVE A REPLY](#))

An A100 GPU with 80 GB of RAM is optimized for training large models like ControlNet with Stable Diffusion XL. Using bfloat16 precision accelerates training while maintaining model accuracy and is natively supported by A100 GPUs. This configuration offers the best performance for fast training without compromising quality.

NEW QUESTION: 15

You are working on a system log anomaly detection model for a cybersecurity organization. You have developed the model using TensorFlow, and you plan to use it for real-time prediction. You need to create a Dataflow pipeline to ingest data via Pub/Sub and write the results to BigQuery.

You want to minimize the serving latency as much as possible. What should you do?

A. Containerize the model prediction logic in Cloud Run, which is invoked by Dataflow.

B. Load the model directly into the Dataflow job as a dependency, and use it for prediction.

C. Deploy the model to a Vertex AI endpoint, and invoke this endpoint in the Dataflow job.

D. Deploy the model in a TF Serving container on Google Kubernetes Engine, and invoke it in the Dataflow job.

Answer: ([SHOW ANSWER](#)**)**

Using private endpoints to serve online predictions with Vertex AI provides a low-latency, secure connection to the Vertex AI online prediction service. This guide shows how to configure private endpoints on Vertex AI by using VPC Network Peering to peer your network with the Vertex AI online prediction service.

<https://cloud.google.com/vertex-ai/docs/predictions/using-private-endpoints>

NEW QUESTION: 16

You are building a TensorFlow text-to-image generative model by using a dataset that contains billions of images with their respective captions. You want to create a low maintenance, automated workflow that reads the data from a Cloud Storage bucket collects statistics, splits the dataset into training/validation/test datasets performs data transformations trains the model using the training/validation datasets, and validates the model by using the test dataset. What should you do?

- A.** Use the MLFlow SDK and deploy it on a Google Kubernetes Engine cluster. Create multiple components that use Dataflow and Vertex AI services.
- B.** Use the Apache Airflow SDK to create multiple operators that use Dataflow and Vertex AI services. Deploy the workflow on Cloud Composer.
- C.** Use the TensorFlow Extended (TFX) SDK to create multiple components that use Dataflow and Vertex AI services. Deploy the workflow on Vertex AI Pipelines.
- D.** Use the Kubeflow Pipelines (KFP) SDK to create multiple components that use Dataflow and Vertex AI services. Deploy the workflow on Vertex AI Pipelines.

Answer: C ([LEAVE A REPLY](#))

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF Special Discount: Freepdfdumps**)

NEW QUESTION: 17

You need to train a regression model based on a dataset containing 50,000 records that is stored in BigQuery. The data includes a total of 20 categorical and numerical features with a target variable that can include negative values. You need to minimize effort and training time while maximizing model performance. What approach should you take to train this regression model?

- A.** Create a custom TensorFlow DNN model
- B.** Use AutoML Tables to train the model without early stopping.
- C.** Use BQML XGBoost regression to train the model.
- D.** Use AutoML Tables to train the model with RMSLE as the optimization objective.

Answer: ([SHOW ANSWER](#)**)**

NEW QUESTION: 18

You started working on a classification problem with time series data and achieved an area under the receiver operating characteristic curve (AUC ROC) value of 99% for training data after just a few experiments. You haven't explored using any sophisticated algorithms or spent any time on hyperparameter tuning. What should your next step be to identify and fix the problem?

- A.** Address the model overfitting by using a less complex algorithm.
- B.** Address data leakage by applying nested cross-validation during model training.
- C.** Address data leakage by removing features highly correlated with the target value.
- D.** Address the model overfitting by tuning the hyperparameters to reduce the AUC ROC value.

Answer: ([SHOW ANSWER](#)**)**

In your case, you are getting a very high AUC ROC value on the training data. This is a sign that the model is overfitting. To address this, you should apply nested cross-validation during model training. This will help to prevent data leakage and ensure that the model is not overfitting to the training data.

NEW QUESTION: 19

You deployed an ML model into production a year ago. Every month, you collect all raw requests that were sent to your model prediction service during the previous month. You send a subset of these requests to a human labeling service to evaluate your model's performance. After a year, you notice that your model's performance sometimes degrades significantly after a month, while other times it takes several months to notice any decrease in performance. The labeling service is costly, but you also need to avoid large performance degradations. You want to determine how often you should retrain your model to maintain a high level of performance while minimizing cost.

What should you do?

- A.** Train an anomaly detection model on the training dataset, and run all incoming requests through this model. If an anomaly is detected, send the most recent serving data to the labeling service.
- B.** Identify temporal patterns in your model's performance over the previous year. Based on these patterns, create a schedule for sending serving data to the labeling service for the next year.
- C.** Compare the cost of the labeling service with the lost revenue due to model performance degradation over the past year. If the lost revenue is greater than the cost of the labeling service, increase the frequency of model retraining; otherwise, decrease the model retraining frequency.
- D.** Run training-serving skew detection batch jobs every few days to compare the aggregate statistics of the features in the training dataset with recent serving data. If skew is detected, send the most recent serving data to the labeling service.

Answer: D ([LEAVE A REPLY](#))

<https://cloud.google.com/blog/topics/developers-practitioners/monitor-models-training-serving-skew-vertex-ai>

<https://developers.google.com/machine-learning/guides/rules-of-ml>

NEW QUESTION: 20

You have recently used TensorFlow to train a classification model on tabular data. You have created a Dataflow pipeline that can transform several terabytes of data into training or prediction datasets consisting of TFRecords. You now need to productionize the model, and you want the predictions to be automatically uploaded to a BigQuery table on a weekly schedule. What should you do?

- A.** Import the model into BigQuery. Implement the data processing logic in a SQL query. On Vertex AI Pipelines create a pipeline that uses the BigQueryQueryJobOp and the BigQueryPredictModelJobOp components.
- B.** Import the model into Vertex AI and deploy it to a Vertex AI endpoint. On Vertex AI Pipelines, create a pipeline that uses the DataflowPythonJobOp and the ModelBatchPredictOp components.
- C.** Import the model into Vertex AI. On Vertex AI Pipelines, create a pipeline that uses the DataflowPythonJobOp and the ModelBatchPredictOp components.
- D.** Import the model into Vertex AI and deploy it to a Vertex AI endpoint. Create a Dataflow pipeline that reuses the data processing logic sends requests to the endpoint, and then uploads predictions to a BigQuery table.

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 21

Your healthcare analytics team wants to use production support tickets that contain patient names, phone numbers, and medical record numbers to fine-tune a summarization model.

Company policy prohibits identifiable patient data from entering ML training datasets. You need to enable the project while complying with policy and minimizing manual review. What should you do?

- A.** Ask engineers to manually redact identifiers from a sample of tickets, and use only that sample for training.
- B.** Run the tickets through Sensitive Data Protection (Cloud DLP) to inspect and de-identify the detected infoTypes before writing the de-identified corpus to the training bucket.
- C.** Store the raw tickets in a bucket with customer-managed encryption keys, and restrict access with IAM before training.
- D.** Train on the raw tickets inside a VPC Service Controls perimeter so the data cannot leave the network.

Answer: B ([LEAVE A REPLY](#))

Sensitive Data Protection detects infoTypes such as person names, phone numbers, and medical record numbers and applies transformations like masking or tokenization at scale. Encryption and perimeter controls protect data in place but do not remove identifiers from the dataset, so they do not satisfy a policy that prohibits identifiable data in training corpora.

NEW QUESTION: 22

You are developing an ML model that predicts the cost of used automobiles based on data such as location, condition, model type, color, and engine/battery efficiency. The data is updated every night. Car dealerships will use the model to determine appropriate car prices. You created a Vertex AI pipeline that reads the data splits the data into training/evaluation/test sets performs feature engineering trains the model by using the training dataset and validates the model by using the evaluation dataset. You need to configure a retraining workflow that minimizes cost.

What should you do?

- A.** Compare the training and evaluation losses of the current run. If the losses are similar, deploy the model to a Vertex AI endpoint with training/serving skew threshold model monitoring. When the model monitoring threshold is triggered redeploy the pipeline.
- B.** Compare the results to the evaluation results from a previous run. If the performance improved deploy the model to a Vertex AI endpoint with training/serving skew threshold model monitoring. When the model monitoring threshold is triggered redeploy the pipeline.
- C.** Compare the results to the evaluation results from a previous run. If the performance improved deploy the model to a Vertex AI endpoint. Configure a cron job to redeploy the pipeline every night.
- D.** Compare the training and evaluation losses of the current run. If the losses are similar, deploy the model to a Vertex AI endpoint. Configure a cron job to redeploy the pipeline every night.

Answer: ([SHOW ANSWER](#)**)**

NEW QUESTION: 23

You work for a large online retailer. You are productionizing a demand forecasting model designed to predict future sales for inventory management. You need to retrain the model with fresh data every 30 days. The inventory planning team requires daily batch predictions for replenishment orders. You need to design a solution that generates daily predictions while minimizing resource costs. Which pipeline architecture should you implement?

- A.** Schedule a pipeline in Agent Platform Pipelines that generates batch predictions daily. Include training as a conditional step that runs if it is day 30 of the cycle.
- B.** Implement two pipelines in Agent Platform Pipelines: a continuous training pipeline that runs every 30 days, and a daily batch prediction pipeline that runs the batch prediction job.
- C.** Implement a pipeline in Agent Platform Pipelines for training that runs every 30 days to update a versioned model on an Agent Platform endpoint. Configure a Cloud Run function to send the daily batch dataset to the endpoint for predictions.
- D.** Implement a pipeline in Agent Platform Pipelines with sequential steps for data preparation, model training, and batch prediction. Schedule the pipeline to run daily.

Answer: B (LEAVE A REPLY)

Separating training and prediction into independently scheduled pipelines matches the different execution frequencies. The training pipeline updates the model every 30 days, while the batch prediction pipeline uses the latest approved model to generate daily forecasts. This avoids unnecessary daily retraining and minimizes compute costs while keeping predictions current.

NEW QUESTION: 24

You have developed an AutoML tabular classification model that identifies high-value customers who interact with your organization's website. You plan to deploy the model to a new Vertex AI endpoint that will integrate with your website application. You expect higher traffic to the website during nights and weekends. You need to configure the model endpoint's deployment settings to minimize latency and cost. What should you do?

- A.** Configure the model deployment settings to use an n1-standard-32machine type.
- B.** Configure the model deployment settings to use an n1-standard-4machine type. Set the minReplicaCountvalue to 1 and the maxReplicaCountvalue to 8.
- C.** Configure the model deployment settings to use an n1-standard-4machine type and a GPU accelerator. Set the minReplicaCountvalue to 1 and the maxReplicaCountvalue to 4.
- D.** Configure the model deployment settings to use an n1-standard-8machine type and a GPU accelerator.

Answer: B (LEAVE A REPLY)

This configuration allows you to optimize for both cost and performance. The n1-standard-4 machine type provides a good balance of resources for handling requests, while setting the minReplicaCount to 1 ensures that you always have at least one instance running to handle traffic, and setting the maxReplicaCount to 8 allows for scaling up during peak traffic times (nights and weekends) to minimize latency. This setup helps manage costs effectively while ensuring the application can handle higher loads when necessary.

NEW QUESTION: 25

You are a data scientist at an industrial equipment manufacturing company. You are developing a regression model to estimate the power consumption in the company's manufacturing plants based on sensor data collected from all of the plants. The sensors collect tens of millions of records every day. You need to schedule daily training runs for your model that use all the data collected up to the current date. You want your model to scale smoothly and require minimal development work. What should you do?

- A.** Develop a regression model using BigQuery ML.
- B.** Develop a custom TensorFlow regression model, and optimize it using Vertex AI Training.

- C. Develop a custom scikit-learn regression model, and optimize it using Vertex AI Training.
- D. Develop a custom PyTorch regression model, and optimize it using Vertex AI Training.

Answer: A ([LEAVE A REPLY](#))

NEW QUESTION: 26

You work for a gaming company that manages a popular online multiplayer game where teams with 6 players play against each other in 5-minute battles. There are many new players every day. You need to build a model that automatically assigns available players to teams in real time.

User research indicates that the game is more enjoyable when battles have players with similar skill levels. Which business metrics should you track to measure your model's performance?

- A. Average time players wait before being assigned to a team
- B. Rate of return as measured by additional revenue generated minus the cost of developing a new model
- C. Precision and recall of assigning players to teams based on their predicted versus actual ability
- D. User engagement as measured by the number of battles played daily per user

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 27

You need to quickly build and train a model to predict the sentiment of customer reviews with custom categories without writing code. You do not have enough data to train a model from scratch. The resulting model should have high predictive performance. Which service should you use?

- A. AutoML Natural Language
- B. Cloud Natural Language API
- C. AI Hub pre-made Jupyter Notebooks
- D. AI Platform Training built-in algorithms

Answer: A ([LEAVE A REPLY](#))

<https://cloud.google.com/natural-language/automl/docs/beginners-guide>

The Natural Language API discovers syntax, entities, and sentiment in text, and classifies text into a predefined set of categories.

NEW QUESTION: 28

You have created multiple versions of an ML model and have imported them to Vertex AI Model Registry. You want to perform A/B testing to identify the best performing model using the simplest approach. What should you do?

- A. Split incoming traffic to distribute prediction requests among the versions. Monitor the performance of each version using Vertex AI's built-in monitoring tools.
- B. Split incoming traffic among Google Kubernetes Engine (GKE) clusters, and use Traffic Director to distribute prediction requests to different versions. Monitor the performance of each version using Cloud Monitoring.
- C. Split incoming traffic to distribute prediction requests among the versions. Monitor the performance of each version using Looker Studio dashboards that compare logged data for each version.
- D. Split incoming traffic among separate Cloud Run instances of deployed models. Monitor the performance of each version using Cloud Monitoring.

Answer: ([SHOW ANSWER](#))

Vertex AI Model Registry supports traffic splitting across multiple model versions natively, allowing you to distribute prediction requests simply and monitor each version's performance using built-in Vertex AI monitoring tools. This approach is the simplest and most integrated for A/B testing within Vertex AI.

NEW QUESTION: 29

You are training an LSTM-based model on AI Platform to summarize text using the following job submission script:

```
gcloud ai-platform jobs submit training $JOB_NAME \
--package-path $TRAINER_PACKAGE_PATH \
--module-name $MAIN_TRAINER_MODULE \
--job-dir $JOB_DIR \
--region $REGION \
--scale-tier basic \
-- \
--epochs 20 \
--batch_size=32 \
--learning_rate=0.001 \
```

You want to ensure that training time is minimized without significantly compromising the accuracy of your model. What should you do?

- A. Modify the 'epochs' parameter.
- B. Modify the 'scale-tier' parameter.
- C. Modify the 'batch size' parameter.
- D. Modify the 'learning rate' parameter.

Answer: B (LEAVE A REPLY)

Changing the scale tier does not impact performance-only speeds up training time. Epochs, Batch size, and learning rate all are hyperparameters that might impact model accuracy.

NEW QUESTION: 30

You are training a deep learning model for semantic image segmentation with reduced training time. While using a Deep Learning VM Image, you receive the following error: The resource 'projects/deeplearning-platform/zones/europe-west4-c/acceleratorTypes/nvidia-tesla-k80' was not found. What should you do?

- A. Ensure that the selected GPU has enough GPU memory for the workload.
- B. Ensure that the required GPU is available in the selected region.
- C. Ensure that you have GPU quota in the selected region.
- D. Ensure that you have preemptible GPU quota in the selected region.

Answer: B (LEAVE A REPLY)

NEW QUESTION: 31

You are an AI engineer that works for a popular video streaming platform. You built a classification model using PyTorch to predict customer churn. Each week, the customer retention team plans to contact customers that have been identified

as at risk of churning with personalized offers. You want to deploy the model while minimizing maintenance effort. What should you do?

- A.** Use Vertex AI's prebuilt containers for prediction. Deploy the container on Cloud Run to generate online predictions.
- B.** Use Vertex AI's prebuilt containers for prediction. Deploy the model on Google Kubernetes Engine (GKE), and configure the model for batch prediction.
- C.** Deploy the model to a Vertex AI endpoint, and configure the model for batch prediction. Schedule the batch prediction to run weekly.
- D.** Deploy the model to a Vertex AI endpoint, and configure the model for online prediction. Schedule a job to query this endpoint weekly.

Answer: C (LEAVE A REPLY)

Deploying the model to a Vertex AI endpoint allows you to use Vertex AI's managed infrastructure, minimizing maintenance effort.

Batch prediction is ideal for this use case, as the churn predictions are required weekly, not in real-time. It is cost-efficient and straightforward to schedule batch jobs to generate predictions for the entire customer dataset at regular intervals. Scheduling the batch prediction weekly aligns perfectly with the customer retention team's plan to act on predictions weekly.

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF** Special Discount: **Freepdfdumps**)

NEW QUESTION: 32

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

Answer: D (LEAVE A REPLY)

One-hot is a good way to use categorical variables in regressions problems

<https://academic.oup.com/rheumatology/article/54/7/1141/1849688>

<https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-auto-preprocessing>

NEW QUESTION: 33

You have developed a fraud detection model for a large financial institution using Vertex AI. The model achieves high accuracy, but the stakeholders are concerned about the model's potential for bias based on customer demographics. You have been asked to provide insights into the model's decision-making process and identify any fairness issues. What should you do?

- A.** Create feature groups using Vertex AI Feature Store to segregate customer demographic features and non-demographic
- B.** Use feature attribution in Vertex AI to analyze model predictions and the impact of each feature on the model's predictions.
- C.** Enable Vertex AI Model Monitoring to detect training-serving skew. Configure an alert to send an email when the skew or drift for a modes feature exceeds a predefined threshold. Re-train the model by appending new data to existing raining data.
- D.** Compile a dataset of unfair predictions. Use Vertex AI Vector Search to identify similar data points in the model's predictions. Report these data points to the stakeholders.

Answer: (SHOW ANSWER)

Feature attribution in Vertex AI (using tools like SHAP or integrated gradients) provides transparency into which features most influence the model's predictions. This helps identify whether demographic features are contributing disproportionately, which is essential for detecting and addressing fairness concerns. It aligns with stakeholder needs for explainability and bias detection.

NEW QUESTION: 34

You need to train a natural language model to perform text classification on product descriptions that contain millions of examples and 100,000 unique words. You want to preprocess the words individually so that they can be fed into a recurrent neural network. What should you do?

- A.** Create a hot-encoding of words, and feed the encodings into your model.
- B.** Identify word embeddings from a pre-trained model, and use the embeddings in your model.
- C.** Sort the words by frequency of occurrence, and use the frequencies as the encodings in your model.
- D.** Assign a numerical value to each word from 1 to 100,000 and feed the values as inputs in your model.

Answer: B (LEAVE A REPLY)

<https://developers.google.com/machine-learning/guides/text-classification/> it is a Word Embedding case

NEW QUESTION: 35

Your e-commerce platform serves customers in 14 languages. You need to build semantic product search so that a query in Spanish returns relevant products whose descriptions are written in English or Japanese. You want to avoid maintaining a separate index per language.

What should you do?

- A.** Translate every query into English with the Cloud Translation API, and use a keyword-based BM25 index over English descriptions.

- B.** Generate embeddings for queries and product descriptions with a multilingual text embedding model, store them in a single Vector Search index, and retrieve by approximate nearest neighbor.
- C.** Train a separate embedding model per language, and query all 14 indexes in parallel, merging results by score.
- D.** Fine-tune a monolingual English embedding model on translated product descriptions.

Answer: B ([LEAVE A REPLY](#))

A multilingual embedding model maps semantically equivalent text from different languages into a shared vector space, so a Spanish query retrieves English and Japanese items from one index.

This avoids per-language indexes and the information loss and latency of a translation step in the query path.

NEW QUESTION: 36

Your organization's security policy states that prediction traffic for a patient-risk model must never traverse the public internet, and that data exfiltration from the project must be blocked even by users with valid credentials. You need to deploy the model accordingly. What should you do?

- A.** Deploy to a public endpoint, and restrict access with an IAM allow policy limited to the service account of the calling application.
- B.** Deploy to a Private Service Connect endpoint reachable from your VPC, and configure a VPC Service Controls perimeter around the project's Agent Platform and storage services.
- C.** Deploy to a public endpoint behind Cloud Armor with an IP allowlist for the corporate network range.
- D.** Deploy to a public endpoint, and encrypt request payloads in the application before sending them.

Answer: ([SHOW ANSWER](#)**)**

Private Service Connect keeps prediction traffic on Google's network with a private IP inside the consumer VPC, satisfying the no-public-internet requirement. VPC Service Controls adds a perimeter that blocks data movement to resources outside it even when the caller holds valid IAM credentials, which IAM and IP allowlists alone cannot do.

NEW QUESTION: 37

You are developing an ML pipeline using Vertex AI Pipelines. You want your pipeline to upload a new version of the XGBoost model to Vertex AI Model Registry and deploy it to Vertex AI Endpoints for online inference. You want to use the simplest approach. What should you do?

- A.** Use the Vertex AI SDK for Python within a custom component based on a python:3.10 image
- B.** Use the Vertex AI ModelEvaluationOp component to evaluate the model
- C.** Chain the Vertex AI ModelUploadOp and ModelDeployOp components together
- D.** Use the Vertex AI REST API within a custom component based on a vertex-ai/prediction/xgboost- cpu image

Answer: C ([LEAVE A REPLY](#))

NEW QUESTION: 38

Your company needs to generate product summaries for vendors. You evaluate a foundation model from Model Garden for text summarization and find the style of the summaries are not aligned with your company's brand voice. How should you improve this LLM-based summarization model to better meet your business objectives?

- A.** Replace the pre-trained model with another model in Model Garden.
- B.** Fine-tune the model using a company-specific dataset.
- C.** Increase the model's temperature parameter.

D. Tune the token output limit in the response.

Answer: B ([LEAVE A REPLY](#))

Fine-tuning the foundation model with a company-specific dataset helps align the generated summaries with the desired brand voice and style, tailoring the model's outputs to meet business objectives effectively.

NEW QUESTION: 39

You work at a mobile gaming startup that creates online multiplayer games. Recently, your company observed an increase in players cheating in the games, leading to a loss of revenue and a poor user experience. You built a binary classification model to determine whether a player cheated after a completed game session, and then send a message to other downstream systems to ban the player that cheated. Your model has performed well during testing, and you now need to deploy the model to production. You want your serving solution to provide immediate classifications after a completed game session to avoid further loss of revenue. What should you do?

A. Import the model into Vertex AI Model Registry. Use the Vertex Batch Prediction service to run batch inference jobs.

B. Save the model files in a Cloud Storage bucket. Create a Cloud Function to read the model files and make online inference requests on the Cloud Function.

C. Save the model files in a VM. Load the model files each time there is a prediction request, and run an inference job on the VM.

D. Import the model into Vertex AI Model Registry. Create a Vertex AI endpoint that hosts the model, and make online inference requests.

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 40

You are part of a team of ML engineers building a new churn prediction model using PyTorch.

Your team's raw data is stored in BigQuery. For the initial prototyping phase, your team needs a collaborative, cloud-based environment that meets the following criteria:

- It must provide access to GPU accelerators.
- It must allow seamless authentication to BigQuery without manual management of service account keys.
- It must be pre-configured with common ML frameworks.

You need to choose the most secure solution with the least setup overhead for your team to begin model development and experimentation. What should you do?

A. Configure a JupyterHub server on a cluster of Compute Engine instances. Install the necessary GPU drivers and other required libraries on each instance. Instruct each team member to create and download a service account key for BigQuery access.

B. Package the PyTorch model code in a Docker container, and upload it to Artifact Registry. Use Agent Platform Serverless Training to run a custom training job that runs on a GPU-enabled machine type. Instruct the team to analyze the training logs to iterate on the model.

C. Instruct team members to use the free version of Agent Platform Colab Enterprise. Use the Agent Platform SDK for Python to load the data into a shared Agent Platform Colab Enterprise notebook. Have each member perform analysis in the shared notebook.

D. Create Agent Platform Workbench instances for your team members. Select instance types with attached GPUs and the preinstalled PyTorch kernel. Use the BigQuery client library to access the data, and use the instances' managed identity for authentication.

Answer: D ([LEAVE A REPLY](#))

Agent Platform Workbench provides managed Jupyter environments with GPU-enabled machine types and preconfigured PyTorch kernels. Its managed identity enables secure access to BigQuery through Application Default Credentials, eliminating the need to create, distribute, or maintain service account keys while minimizing setup effort for collaborative experimentation.

NEW QUESTION: 41

Your team trains a transformer-based recommendation model written in JAX. Training runs involve very large matrix multiplications, use a fixed model architecture, and run for many hours at high utilization. You need to select the accelerator that minimizes cost per training run. What should you do?

- A.** Select a CPU-only machine with a high vCPU count, and enable parallel data loading.
- B.** Select TPU v5e slices, and shard the training workload across the slice.
- C.** Select multiple small GPUs with dynamic shape support for improved flexibility.
- D.** Select a single large-memory GPU, and enable gradient checkpointing.

Answer: B ([LEAVE A REPLY](#))

TPUs are optimized for large, dense matrix operations with static shapes, which matches transformer training and JAX's compilation model. For long-running, high-utilization training of a fixed architecture, TPU slices typically deliver better price-performance than equivalent GPU configurations.

NEW QUESTION: 42

You need to deploy a scikit-learn classification model to production. The model must be able to serve requests 24/7, and you expect millions of requests per second to the production application from 8 am to 7 pm. You need to minimize the cost of deployment. What should you do?

- A.** Deploy an online Vertex AI prediction endpoint. Set the max replica count to 1
- B.** Deploy an online Vertex AI prediction endpoint. Set the max replica count to 100
- C.** Deploy an online Vertex AI prediction endpoint with one GPU per replica. Set the max replica count to 100
- D.** Deploy an online Vertex AI prediction endpoint with one GPU per replica. Set the max replica count to 1

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 43

During batch training of a neural network, you notice that there is an oscillation in the loss. How should you adjust your model to ensure that it converges?

- A.** Decrease the size of the training batch.
- B.** Decrease the learning rate hyperparameter.
- C.** Increase the learning rate hyperparameter.
- D.** Increase the size of the training batch.

Answer: ([SHOW ANSWER](#)**)**

larger learning rates can reduce training time but may lead to model oscillation and may miss the optimal model parameter values.

NEW QUESTION: 44

You are working on a prototype of a text classification model in a managed Vertex AI Workbench notebook. You want to quickly experiment with tokenizing text by using a Natural Language Toolkit (NLTK) library. How should you add the library to your Jupyter kernel?

- A. Install the NLTK library from a Jupyter cell by using the `!pip install nltk --user` command.
- B. Write a custom Dataflow job that uses NLTK to tokenize your text and saves the output to Cloud Storage.
- C. Create a new Vertex AI Workbench notebook with a custom image that includes the NLTK library.
- D. Install the NLTK library from a terminal by using the `pip install nltk` command.

Answer: A ([LEAVE A REPLY](#))

NEW QUESTION: 45

Your production TensorFlow pipeline computes vocabulary mappings and z-score normalization statistics over the full training dataset. Engineers reimplemented the same transformations in the serving application, and subtle differences have caused prediction errors twice this year. You need to guarantee that serving transformations exactly match training transformations. What should you do?

- A. Use TensorFlow Transform to compute the statistics in the pipeline and emit a transform graph that is attached to the exported SavedModel for serving.
- B. Document the transformation logic in a shared design document that both teams review before each release.
- C. Compute normalization statistics on each incoming serving batch so the values are always current.
- D. Write unit tests in the serving application that compare its output against sample training records.

Answer: A ([LEAVE A REPLY](#))

TensorFlow Transform performs a full-pass analysis during preprocessing and exports the resulting constants inside a transform graph that becomes part of the serving signature. Because the same graph executes at training and inference, skew from duplicated logic is structurally eliminated rather than merely tested for.

NEW QUESTION: 46

You work for a company that captures live video footage of checkout areas in their retail stores.

You need to use the live video footage to build a model to detect the number of customers waiting for service in near real time. You want to implement a solution quickly and with minimal effort.

How should you build the model?

- A. Train an AutoML object detection model on an annotated dataset by using Vertex AutoML.
- B. Train a Seq2Seq+ object detection model on an annotated dataset by using Vertex AutoML.
- C. Use the Vertex AI Vision Occupancy Analytics model.
- D. Use the Vertex AI Vision Person/vehicle detector model.

Answer: ([SHOW ANSWER](#))

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF** Special Discount: **Freepdfdumps**)

NEW QUESTION: 47

You recently used BigQuery ML to train an AutoML regression model. You shared results with your team and received positive feedback. You need to deploy your model for online prediction as quickly as possible. What should you do?

- A. Retrain the model by using Vertex AI Deploy the model from Vertex AI Model Registry to a Vertex AI endpoint.
- B. Retrain the model by using BigQuery ML, and specify Vertex AI as the model registry. Deploy the model from Vertex AI Model Registry to a Vertex AI endpoint,
- C. Alter the model by using BigQuery ML, and specify Vertex AI as the model registry. Deploy the model from Vertex AI Model Registry to a Vertex AI endpoint.
- D. Export the model from BigQuery ML to Cloud Storage. Import the model into Vertex AI Model Registry. Deploy the model to a Vertex AI endpoint.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 48

You developed a Vertex AI pipeline that trains a classification model on data stored in a large BigQuery table. The pipeline has four steps, where each step is created by a Python function that uses the KubeFlow v2 API. The components have the following names:

```
dt=datetime.now().strftime("%Y%m%d%H%M%S")
f"export-{dt}.yaml", f"preprocess-{dt}.yaml", f"train-{dt}.yaml",
f"calibrate-{dt}.yaml"
```

You launch your Vertex AI pipeline as the following:

```
job = aip.PipelineJob(
    display_name="my-awesome-pipeline",
    template_path="pipeline.json",
    job_id=f"my-awesome-pipeline-{dt}",
    parameter_values=params,
    enable_caching=True,
    location="europe-west1"
)
```

You perform many model iterations by adjusting the code and parameters of the training step.

You observe high costs associated with the development, particularly the data export and preprocessing steps. You need to reduce model development costs. What should you do?

- A. Change the name of the pipeline to f"my-awesome-pipeline-{dt}".
- B. Add the {"kubeflow.v1.caching": True} parameter to the set of params provided to your PipelineJob.

C. Move the first step of your pipeline to a separate step, and provide a cached path to Cloud Storage as an input to the main pipeline.

D. Change the components' YAML filenames to export.yaml, preprocess.yaml, f "train-{dt}.yaml", f"calibrate-{dt}.yaml".

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 49

You work for an online grocery store. You recently developed a custom ML model that recommends a recipe when a user arrives at the website. You chose the machine type on the Vertex AI endpoint to optimize costs by using the queries per second (QPS) that the model can serve, and you deployed it on a single machine with 8 vCPUs and no accelerators. A holiday season is approaching and you anticipate four times more traffic during this time than the typical daily traffic. You need to ensure that the model can scale efficiently to the increased demand. What should you do?

A. 1. Maintain the same machine type on the endpoint. Configure the endpoint to enable autoscaling based on vCPU usage.

2. Set up a monitoring job and an alert for CPU usage.

3. If you receive an alert, investigate the cause.

B. 1. Change the machine type on the endpoint to have 32 vCPUs.

2. Set up a monitoring job and an alert for CPU usage.

3. If you receive an alert, scale the vCPUs further as needed.

C. 1. Maintain the same machine type on the endpoint.

2. Set up a monitoring job and an alert for CPU usage.

3. If you receive an alert, add a compute node to the endpoint.

D. 1. Change the machine type on the endpoint to have a GPU. Configure the endpoint to enable autoscaling based on the GPU usage.

2. Set up a monitoring job and an alert for GPU usage.

3. If you receive an alert, investigate the cause.

Answer: **A** ([LEAVE A REPLY](#))

NEW QUESTION: 50

Your organization wants to make its internal shuttle service route more efficient. The shuttles currently stop at all pick-up points across the city every 30 minutes between 7 am and 10 am.

The development team has already built an application on Google Kubernetes Engine that requires users to confirm their presence and shuttle station one day in advance. What approach should you take?

A. 1. Build a tree-based regression model that predicts how many passengers will be picked up at each shuttle station.

2. Dispatch an appropriately sized shuttle and provide the map with the required stops based on the prediction.

B. 1. Define the optimal route as the shortest route that passes by all shuttle stations with confirmed attendance at the given time under capacity constraints.

2. Dispatch an appropriately sized shuttle and indicate the required stops on the map.

C. 1. Build a tree-based classification model that predicts whether the shuttle should pick up passengers at each shuttle station.

2. Dispatch an available shuttle and provide the map with the required stops based on the prediction.

- D.** 1. Build a reinforcement learning model with tree-based classification models that predict the presence of passengers at shuttle stops as agents and a reward function around a distance-based metric.
2. Dispatch an appropriately sized shuttle and provide the map with the required stops based on the simulated outcome.

Answer: A ([LEAVE A REPLY](#))

NEW QUESTION: 51

You developed a Python module by using Keras to train a regression model. You developed two model architectures, linear regression and deep neural network (DNN), within the same module.

You are using the `training_method` argument to select one of the two methods, and you are using the `learning_rate` and `num_hidden_layers` arguments in the DNN. You plan to use Vertex AI's hypertuning service with a budget to perform 100 trials. You want to identify the model architecture and hyperparameter values that minimize training loss and maximize model performance. What should you do?

- A.** Run one hypertuning job for 100 trials. Set `num_hidden_layers` as a conditional hyperparameter based on its parent hyperparameter `training_method`, and set `learning_rate` as a non-conditional hyperparameter.
- B.** Run two separate hypertuning jobs, a linear regression job for 50 trials, and a DNN job for 50 trials. Compare their final performance on a common validation set, and select the set of hyperparameters with the least training loss.
- C.** Run one hypertuning job with `training_method` as the hyperparameter for 50 trials. Select the architecture with the lowest training loss, and further hypertune it and its corresponding hyperparameters for 50 trials.
- D.** Run one hypertuning job for 100 trials. Set `num_hidden_layers` and `learning_rate` as conditional hyperparameters based on their parent hyperparameter `training_method`.

Answer: (SHOW ANSWER)

This approach allows you to explore both model architectures (linear regression and DNN) within a single hypertuning job while effectively managing hyperparameters. By setting `num_hidden_layers` as a conditional hyperparameter based on `training_method`, you ensure that the DNN-specific hyperparameters are only considered when the DNN architecture is selected. The `learning_rate` can be tuned across both architectures as a non-conditional hyperparameter. This strategy maximizes the use of the 100 trials to identify the optimal architecture and hyperparameters while minimizing training loss and maximizing model performance.

NEW QUESTION: 52

You work for an auto insurance company. You are preparing a proof-of-concept ML application that uses images of damaged vehicles to infer damaged parts. Your team has assembled a set of annotated images from damage claim documents in the company's database. The annotations associated with each image consist of a bounding box for each identified damaged part and the part name. You have been given a sufficient budget to train models on Google Cloud. You need to quickly create an initial model. What should you do?

- A.** Train an object detection model in AutoML by using the annotated image data.
- B.** Train an object detection model in Vertex AI custom training by using the annotated image data.
- C.** Download a pre-trained object detection model from TensorFlow Hub. Fine-tune the model in Vertex AI Workbench by using the annotated image data.
- D.** Create a pipeline in Vertex AI Pipelines and configure the `AutoMLTrainingJobRunOp` component to train a custom object detection model by using the annotated image data.

Answer: A ([LEAVE A REPLY](#))

NEW QUESTION: 53

You developed a Transformer model in TensorFlow to translate text. Your training data includes millions of documents in a Cloud Storage bucket. You plan to use distributed training to reduce training time. You need to configure the training job while minimizing the effort required to modify code and to manage the cluster's configuration. What should you do?

- A.** Create a Vertex AI custom distributed training job with Reduction Server. Use N1 high-memory machine type instances for the first and second pools, and use N1 high-CPU machine type instances for the third worker pool.
- B.** Create a training job that uses Cloud TPU VMs. Use `tf.distribute.TPUStrategy` for distribution.
- C.** Create a Vertex AI custom training job with a single worker pool of A2 GPU machine type instances. Use `tf.distribute.MirroredStrategy` for distribution.
- D.** Create a Vertex AI custom training job with GPU accelerators for the second worker pool. Use `tf.distribute.MultiWorkerMirroredStrategy` for distribution.

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 54

You work for a large hotel chain and have been asked to assist the marketing team in gathering predictions for a targeted marketing strategy. You need to make predictions about user lifetime value (LTV) over the next 20 days so that marketing can be adjusted accordingly. The customer dataset is in BigQuery, and you are preparing the tabular data for training with AutoML Tables.

This data has a time signal that is spread across multiple columns. How should you ensure that AutoML fits the best model to your data?

- A.** Manually combine all columns that contain a time signal into an array.
Allow AutoML to interpret this array appropriately.
Choose an automatic data split across the training, validation, and testing sets.
- B.** Submit the data for training without performing any manual transformations.
Allow AutoML to handle the appropriate transformations.
Choose an automatic data split across the training, validation, and testing sets.
- C.** Submit the data for training without performing any manual transformations, and indicate an appropriate column as the Time column.
Allow AutoML to split your data based on the time signal provided, and reserve the more recent data for the validation and testing sets.
- D.** Submit the data for training without performing any manual transformations.
Use the columns that have a time signal to manually split your data.
Ensure that the data in your validation set is from 30 days after the data in your training set and that the data in your testing sets from 30 days after your validation set.

Answer: D ([LEAVE A REPLY](#))

<https://cloud.google.com/automl-tables/docs/data-best-practices#time>

The time signal that is spread across multiple columns so manual split is required.

- If the time information is not contained in a single column, you can use a manual data split to use the most recent data as the test data, and the earliest data as the training data.

NEW QUESTION: 55

You recently deployed a scikit-learn model to a Vertex AI endpoint. You are now testing the model on live production traffic. While monitoring the endpoint, you discover twice as many requests per hour than expected throughout the day. You want the endpoint to efficiently scale when the demand increases in the future to prevent users from experiencing high latency. What should you do?

- A.** Set the target utilization percentage in the autoscalingMetricSpecs configuration to a higher value
- B.** Configure an appropriate minReplicaCount value based on expected baseline traffic
- C.** Deploy two models to the same endpoint, and distribute requests among them evenly
- D.** Change the model's machine type to one that utilizes GPUs

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 56

You deployed a conversational application that uses Gemini. You recently collected thousands of rows of user feedback about the verbosity and accuracy of the model's responses. The feedback indicated that the responses are factually correct but users want different levels of verbosity depending on the type of question. You need to adjust the model to return responses that are more consistent with users' expectations. You want to use a scalable solution and minimize application maintenance. What should you do?

- A.** Experiment with other models in Model Garden. Perform A/B testing by setting each model as your application's default model. Choose a model based on the results.
- B.** Ask users to indicate all scenarios where they expect concise responses versus verbose responses. Modify the application's prompt to include these scenarios and their respective verbosity levels. Re-evaluate the verbosity of responses with updated prompts.
- C.** Implement a keyword-based routing layer. If the user's input contains the words "detailed" or "description," return a verbose response. If the user's input contains the word "fact," re-prompt Gemini to summarize the response and return a concise response.
- D.** Ask users to provide examples of responses with the appropriate verbosity as a list of question and answer pairs. Use this dataset to perform supervised fine-tuning (SFT) in Agent Platform. Re-evaluate the verbosity of responses with the tuned version.

Answer: **D** ([LEAVE A REPLY](#))

Supervised fine-tuning with representative question-and-answer examples teaches the model to associate different query types with the expected response length. This provides a scalable, consistent behavior change within the model itself and avoids maintaining extensive prompt rules or keyword-based routing logic.

NEW QUESTION: 57

You developed a BigQuery ML linear regressor model by using a training dataset stored in a BigQuery table. New data is added to the table every minute. You are using Cloud Scheduler and Vertex AI Pipelines to automate hourly model training, and use the model for direct inference. The feature preprocessing logic includes quantile bucketization and MinMax scaling on data received in the last hour. You want to minimize storage and computational overhead. What should you do?

- A.** Preprocess and stage the data in BigQuery prior to feeding it to the model during training and inference.
- B.** Use the TRANSFORM clause in the CREATE MODEL statement in the SQL query to calculate the required statistics.

C. Create a component in the Vertex AI Pipelines directed acyclic graph (DAG) to calculate the required statistics, and pass the statistics on to subsequent components.

D. Create SQL queries to calculate and store the required statistics in separate BigQuery tables that are referenced in the CREATE MODEL statement.

Answer: B ([LEAVE A REPLY](#))

Using the TRANSFORM clause allows you to apply preprocessing steps directly within the model training process. This means that the required statistics for quantile bucketization and MinMax scaling can be calculated on-the-fly during model training and inference, minimizing storage overhead by not requiring separate tables to store these statistics. This approach also streamlines the workflow by integrating the preprocessing into the model training SQL query, thus reducing computational overhead and complexity in your pipeline.

NEW QUESTION: 58

You built a deep learning-based image classification model by using on-premises data. You want to use Vertex AI to deploy the model to production. Due to security concerns, you cannot move your data to the cloud. You are aware that the input data distribution might change over time. You need to detect model performance changes in production. What should you do?

A. Create a Vertex AI Model Monitoring job. Enable training-serving skew detection for your model.

B. Use Vertex Explainable AI for model explainability. Configure example-based explanations.

C. Use Vertex Explainable AI for model explainability. Configure feature-based explanations.

D. Create a Vertex AI Model Monitoring job. Enable feature attribution skew and drift detection for your model.

Answer: ([SHOW ANSWER](#)**)**

NEW QUESTION: 59

You are the lead ML engineer for a large ecommerce company. The Data Analytics team has prepared a dataset of 50 million transaction rows, which is stored in a partitioned BigQuery table.

The team consists primarily of SQL experts with limited experience in Python or TensorFlow.

You need to build a binary classification model to predict the likelihood of a customer making a purchase in the next seven days. You need to provide daily batch predictions. You need to minimize development effort, data movement, and the overall complexity of the solution. What should you do?

A. Create a model with BigQuery ML that includes the MODEL_TYPE=AUTOML_CLASSIFIER settings. Use the ML.PREDICT function to generate predictions as part of your nightly data warehouse jobs.

B. Use the BigQuery ML TRANSFORM function to normalize and scale numerical features, one-hot encode categorical features, and train a MODEL_TYPE=LOGISTIC_REG mode. Use the ML.PREDICT function to generate predictions as part of your nightly data warehouse jobs.

C. Train a classification model with Agent Platform AutoML Tabular Workflows, deploy the trained model to an Agent Platform endpoint, and add a nightly Cloud Run job to trigger batch predictions.

D. Develop an XGBoost model on an Agent Platform Workbench notebook. Load the data into a pandas DataFrame, and pickle it to a Cloud Storage bucket. Create another notebook that loads the pickled model from Cloud Storage and saves predictions back to Cloud Storage. Schedule this notebook for nightly execution.

Answer: A ([LEAVE A REPLY](#))

BigQuery ML AutoML classification allows the SQL-focused team to train a high-quality binary classifier directly on the 50-million-row BigQuery table without moving the data or developing custom Python training code. ML.PREDICT can be integrated into existing nightly warehouse jobs to generate daily batch predictions entirely within BigQuery, minimizing operational complexity and development effort.

Reference:

<https://docs.cloud.google.com/bigquery/docs/reference/standard-sql/bigqueryml-syntax-predict>

<https://docs.cloud.google.com/bigquery/docs/create-machine-learning-model>

<https://docs.cloud.google.com/bigquery/docs/export-model-tutorial>

NEW QUESTION: 60

Your team is developing a customer support chatbot for a healthcare company that processes sensitive patient information. You need to ensure that all personally identifiable information (PII) captured during customer conversations is protected prior to storing or analyzing the data. What should you do?

- A. Use the Cloud Natural Language API to identify and redact PII in chatbot conversations.
- B. Use the Cloud Natural Language API to classify and categorize all data, including PII, in chatbot conversations.
- C. Use the DLP API to encrypt PII in chatbot conversations before storing the data.
- D. Use the DLP API to scan and de-identify PII in chatbot conversations before storing the data.

Answer: D (LEAVE A REPLY)

DLP API (Data Loss Prevention API) is specifically designed to identify, classify, and protect sensitive information like PII. It can scan chatbot conversations for PII and perform de-identification operations such as masking, tokenization, or redaction to ensure data privacy.

De-identifying PII before storage ensures compliance with data protection regulations such as HIPAA (for healthcare data) and enhances data security by removing or obfuscating sensitive information.

NEW QUESTION: 61

You are building a linear model with over 100 input features, all with values between -1 and 1.

You suspect that many features are non-informative. You want to remove the non-informative features from your model while keeping the informative ones in their original form. Which technique should you use?

- A. Use principal component analysis (PCA) to eliminate the least informative features.
- B. Use L1 regularization to reduce the coefficients of uninformative features to 0.
- C. After building your model, use Shapley values to determine which features are the most informative.
- D. Use an iterative dropout technique to identify which features do not degrade the model when removed.

Answer: (SHOW ANSWER)

L1 regularization it's good for feature selection

<https://www.quora.com/How-does-the-L1-regularization-method-help-in-feature-selection>

<https://developers.google.com/machine-learning/crash-course/regularization-for-sparsity/l1-regularization>

Learning-Engineer exam dumps, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF** Special Discount: **Freepdfdumps**)

NEW QUESTION: 62

You are an ML engineer on an agricultural research team working on a crop disease detection tool to detect leaf rust spots in images of crops to determine the presence of a disease. These spots, which can vary in shape and size, are correlated to the severity of the disease. You want to develop a solution that predicts the presence and severity of the disease with high accuracy.

What should you do?

- A. Create an object detection model that can localize the rust spots.
- B. Develop an image segmentation ML model to locate the boundaries of the rust spots.
- C. Develop a template matching algorithm using traditional computer vision libraries.
- D. Develop an image classification ML model to predict the presence of the disease.

Answer: B ([LEAVE A REPLY](#))

The shape of the spot is quite important for the severity of the disease, and image segmentation could help us to determine it in a more granular manner. And it is often used in the healthcare industry, for getting the shapes of all the cancerous cells.

NEW QUESTION: 63

You have been asked to build a model using a dataset that is stored in a medium-sized (~10 GB) BigQuery table. You need to quickly determine whether this data is suitable for model development. You want to create a one-time report that includes both informative visualizations of data distributions and more sophisticated statistical analyses to share with other ML engineers on your team. You require maximum flexibility to create your report. What should you do?

- A. Use Vertex AI Workbench user-managed notebooks to generate the report.
- B. Use the Google Data Studio to create the report.
- C. Use the output from TensorFlow Data Validation on Dataflow to generate the report.
- D. Use Dataprep to create the report.

Answer: A ([LEAVE A REPLY](#))

While Google Data Studio (Option B) is a powerful data visualization and reporting tool, it might not provide the same level of flexibility and sophistication for statistical analyses compared to a notebook environment.

NEW QUESTION: 64

You need to write a generic test to verify whether Dense Neural Network (DNN) models automatically released by your team have a sufficient number of parameters to learn the task for which they were built. What should you do?

- A. Train the model for a few iterations, and check for NaN values.
- B. Train the model for a few iterations, and verify that the loss is constant.
- C. Train a simple linear model, and determine if the DNN model outperforms it.
- D. Train the model with no regularization, and verify that the loss function is close to zero.

Answer: D ([LEAVE A REPLY](#))

A is not correct because the test does not check that the model has enough parameters to learn the task.

B is not correct because the loss should decrease if you have enough parameters to learn the task.

C is not correct because outperforming the linear model does not guarantee that the model has enough parameters to learn tasks with non-linear data representations. The option also doesn't quantify a metric to give an indication of how well the model performed.

D is correct because the test can check that the model has enough parameters to memorize the task.

<https://developers.google.com/machine-learning/testing-debugging/pipeline/deploying#testing-for-algorithmic-correctness>

NEW QUESTION: 65

You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app. You have reviewed old chat logs and tagged each conversation for intent based on each customer's stated intention for contacting customer service. About 70% of customer inquiries are simple requests that are solved within 10 intents. The remaining 30% of inquiries require much longer and more complicated requests. Which intents should you automate first?

A. Automate a blend of the shortest and longest intents to be representative of all intents.

B. Automate the more complicated requests first because those require more of the agents' time.

C. Automate the 10 intents that cover 70% of the requests so that live agents can handle the more complicated requests.

D. Automate intents in places where common words such as payment only appear once to avoid confusing the software.

Answer: C (LEAVE A REPLY)

A is incorrect because you should not automate the higher value requests.

B is incorrect because live agents are better suited to handle these complicated requests.

C is correct because it enables a machine to handle the most simple requests and gives the live agents more opportunity to handle higher value requests.

D is incorrect because Dialogflow can handle the same word in multiple intents.

<https://cloud.google.com/blog/products/ai-machine-learning/discover-card-how-we-designed-an-experiment-to-evaluate-conversational-experience-platforms>

NEW QUESTION: 66

You developed a custom model by using Vertex AI to forecast the sales of your company's products based on historical transactional data. You anticipate changes in the feature distributions and the correlations between the features in the near future. You also expect to receive a large volume of prediction requests. You plan to use Vertex AI Model Monitoring for drift detection and you want to minimize the cost. What should you do?

A. Use the features and the feature attributions for monitoring. Set a monitoring-frequency value that is lower than the default.

B. Use the features for monitoring. Set a prediction-sampling-rate value that is closer to 1 than 0.

C. Use the features for monitoring. Set a monitoring-frequency value that is higher than the default.

D. Use the features and the feature attributions for monitoring. Set a prediction-sampling-rate value that is closer to 0 than 1.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 67

You work for a bank with strict data governance requirements. You recently implemented a custom model to detect fraudulent transactions. You want your training code to download internal data by using an API endpoint hosted in your project's network. You need the data to be accessed in the most secure way, while mitigating the risk of data exfiltration. What should you do?

- A.** Enable VPC Service Controls for peerings, and add Vertex AI to a service perimeter.
- B.** Create a Cloud Run endpoint as a proxy to the data. Use Identity and Access Management (IAM) authentication to secure access to the endpoint from the training job.
- C.** Download the data to a Cloud Storage bucket before calling the training job.
- D.** Configure VPC Peering with Vertex AI, and specify the network of the training job.

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 68

You work for a small company that has deployed an ML model with autoscaling on Vertex AI to serve online predictions in a production environment. The current model receives about 20 prediction requests per hour with an average response time of one second. You have retrained the same model on a new batch of data, and now you are canary testing it, sending ~10% of production traffic to the new model. During this canary test, you notice that prediction requests for your new model are taking between 30 and 180 seconds to complete. What should you do?

- A.** Remove your new model from the production environment. Compare the new model and existing model codes to identify the cause of the performance bottleneck.
- B.** Submit a request to raise your project quota to ensure that multiple prediction services can run concurrently.
- C.** Remove your new model from the production environment. For a short trial period, send all incoming prediction requests to BigQuery. Request batch predictions from your new model, and then use the Data Labeling Service to validate your model's performance before promoting it to production.
- D.** Turn off auto-scaling for the online prediction service of your new model. Use manual scaling with one node always available.

Answer: ([SHOW ANSWER](#)**)**

NEW QUESTION: 69

You have recently developed a new ML model in a Jupyter notebook. You want to establish a reliable and repeatable model training process that tracks the versions and lineage of your model artifacts. You plan to retrain your model weekly. How should you operationalize your training process?

- A.** 1. Create an instance of the CustomTrainingJob class with the Vertex AI SDK to train your model.
2. Using the Notebooks API, create a scheduled execution to run the training code weekly.
- B.** 1. Create an instance of the CustomJob class with the Vertex AI SDK to train your model.
2. Use the Metadata API to register your model as a model artifact.
3. Using the Notebooks API, create a scheduled execution to run the training code weekly.
- C.** 1. Create a managed pipeline in Vertex AI Pipelines to train your model by using a Vertex AI CustomTrainingJobOp component.
2. Use the ModelUploadOp component to upload your model to Vertex AI Model Registry.
3. Use Cloud Scheduler and Cloud Run functions to run the Vertex AI pipeline weekly.

- D.** 1. Create a managed pipeline in Vertex AI Pipelines to train your model using a Vertex AI HyperParameterTuningJobRunOp component.
2. Use the ModelUploadOp component to upload your model to Vertex AI Model Registry.
3. Use Cloud Scheduler and Cloud Run functions to run the Vertex AI pipeline weekly.

Answer: C (LEAVE A REPLY)

Creating a managed pipeline in Vertex AI Pipelines that includes training (via CustomTrainingJobOp) and model registration (via ModelUploadOp) enables tracking of model versions and lineage. Scheduling the pipeline with Cloud Scheduler and Cloud Run ensures reliable, repeatable weekly retraining following Google best practices.

NEW QUESTION: 70

You built a custom ML model using scikit-learn. Training time is taking longer than expected. You decide to migrate your model to Vertex AI Training, and you want to improve the model's training time. What should you try out first?

- A.** Migrate your model to TensorFlow, and train it using Vertex AI Training.
- B.** Train your model in a distributed mode using multiple Compute Engine VMs.
- C.** Train your model with DLVM images on Vertex AI, and ensure that your code utilizes NumPy and SciPy internal methods whenever possible.
- D.** Train your model using Vertex AI Training with GPUs.

Answer: C (LEAVE A REPLY)

GPU is not useful for scikit-learn model

<https://scikit-learn.org/stable/faq.html#will-you-add-gpu-support>

but DLVM did mention it supports scikit-learn framework

<https://cloud.google.com/deep-learning-vm>

NEW QUESTION: 71

You are training a ResNet-50 computer vision model on Gemini Enterprise Agent Platform using a custom PyTorch container. The training job is configured with a generic machine type and one NVIDIA A100 GPU. During the training, you monitor the resource utilization charts. You observe that the GPU utilization is volatile, frequently dropping to 0% for short intervals, while the CPU utilization remains consistently above 90%. You want to maximize GPU utilization to reduce training time. What should you do?

- A.** Increase the training batch size to saturate the GPU memory.
- B.** Update the DataLoader utility to increase the num_workers parameter and enable pin_memory.
- C.** Switch to a machine type with a higher number of GPUs, and implement the torch.nn.parallel.DistributedDataParallel strategy in the training code.
- D.** Enable automatic mixed precision (AMP) to speed up the matrix multiplications.

Answer: (SHOW ANSWER)

The consistently high CPU utilization and intermittent GPU idle periods indicate that data loading and preprocessing cannot supply batches fast enough. Increasing parallel data-loading workers and using pinned memory improves CPU-side throughput and accelerates host-to-GPU transfers, keeping the GPU fed more consistently and reducing training time.

NEW QUESTION: 72

You have deployed a deep learning model to an Agent Platform endpoint using a machine type with NVIDIA GPUs. You initially configured the endpoint to autoscale based on a target CPU utilization of 60%. During a load test, you observe that prediction latency increases significantly as traffic rises, but the number of replicas remains constant. Cloud Monitoring indicates that CPU utilization stays below 40%, while the GPU utilization consistently exceeds 90%. You need to ensure that the endpoint scales efficiently to handle the increased load. What should you do?

- A. Update the autoscaling configuration to scale based on the GPU duty cycle metric.
- B. Update the autoscaling configuration to decrease the target CPU utilization to 20%.
- C. Redeploy the model to a machine type that includes a TPU.
- D. Increase the minimum replica count for the endpoint to match the peak load that was observed during testing.

Answer: A (LEAVE A REPLY)

The endpoint is constrained by GPU capacity rather than CPU capacity. Configuring autoscaling to use GPU duty cycle allows replica count to increase when GPU utilization becomes high, directly responding to the resource causing prediction latency.

NEW QUESTION: 73

As the lead ML Engineer for your company, you are responsible for building ML models to digitize scanned customer forms. You have developed a TensorFlow model that converts the scanned images into text and stores them in Cloud Storage. You need to use your ML model on the aggregated data collected at the end of each day with minimal manual intervention. What should you do?

- A. Use the batch prediction functionality of AI Platform.
- B. Create a serving pipeline in Compute Engine for prediction.
- C. Use Cloud Functions for prediction each time a new data point is ingested.
- D. Deploy the model on AI Platform and create a version of it for online inference.

Answer: A (LEAVE A REPLY)

You need to use your ML model on the aggregated data collected at the end of each day with minimal manual intervention.

NEW QUESTION: 74

Your company stores a large number of audio files of phone calls made to your customer call center in an on-premises database. Each audio file is in wav format and is approximately 5 minutes long. You need to analyze these audio files for customer sentiment. You plan to use the Speech-to-Text API. You want to use the most efficient approach. What should you do?

- A. 1. Upload the audio files to Cloud Storage.
2. Call the `speech:longrunningrecognize` API endpoint to generate transcriptions
3. Create a Cloud Function that calls the Natural Language API by using the `analyzeSentiment` method

- B. 1. Upload the audio files to Cloud Storage
2. Call the `speech:longrunningrecognize` API endpoint to generate transcriptions
3. Call the `predict` method of an AutoML sentiment analysis model to analyze the transcriptions.

- C. 1. Iterate over your local files in Python
2. Use the Speech-to-Text Python Library to create a `speech.RecognitionAudio` object and set the content to the audio file data

3. Call the `speech:longrunningrecognize` API endpoint to generate transcriptions.

4. Call the Natural Language API by using the `analyzeSentiment` method

D. 1. Iterate over your local files in Python

2. Use the Speech-to-Text Python library to create a `speech.RecognitionAudio` object, and set the content to the audio file data

3. Call the `speech:recognize` API endpoint to generate transcriptions

4. Call the `predict` method of an AutoML sentiment analysis model to analyze the transcriptions.

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 75

You are in the exploratory phase of development of a demand forecasting model. Several terabytes of historical sales and inventory records are stored in a BigQuery table. You need to create a new, clean table to perform preliminary analyses on metrics, such as moving averages of sales figures and weekly aggregates, without additional data movement. You want to use the simplest approach and minimize overhead. What should you do?

A. Create a `PreprocessingOp` component in the Agent Platform Pipelines workflow to compute the moving averages and extract time-based features and write the processed data to a new destination table in BigQuery.

B. Read the data from BigQuery into a Dask DataFrame using Agent Platform Workbench. Perform exploratory data analysis (EDA) and feature engineering transformations, and write the processed data back to a new destination table in BigQuery.

C. Write a SQL query in the BigQuery console. Use the `ML.TRANSFORM` function in BigQuery ML to compute the moving averages and extract time-based features. Write the processed data to a new destination table in BigQuery.

D. Write a SQL query in the BigQuery console. Use standard SQL functions like `CAST`, `EXTRACT`, and window functions with `OVER()` clauses to perform the transformations and calculations.

Materialize the results into a new destination table in BigQuery.

Answer: ([SHOW ANSWER](#)**)**

BigQuery standard SQL can calculate moving averages with window functions, create weekly aggregates, and derive date-based features directly where the data resides. Materializing the query results into a new BigQuery table avoids data movement and additional ML or pipeline infrastructure, making it the simplest and lowest-overhead approach.

NEW QUESTION: 76

You work for a food product company. Your company's historical sales data is stored in BigQuery. You need to use Vertex AI's custom training service to train multiple TensorFlow models that read the data from BigQuery and predict future sales. You plan to implement a data preprocessing algorithm that performs mm-max scaling and bucketing on a large number of features before you start experimenting with the models. You want to minimize preprocessing time, cost, and development effort. How should you configure this workflow?

A. Write the transformations into Spark that uses the `spark-bigquery-connector`, and use Dataproc to preprocess the data.

B. Add the transformations as a preprocessing layer in the TensorFlow models.

C. Create a Dataflow pipeline that uses the `BigQueryIO` connector to ingest the data, process it, and write it back to BigQuery.

D. Write SQL queries to transform the data in-place in BigQuery.

Answer: ([SHOW ANSWER](#))

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF Special Discount: Freepdfdumps**)

NEW QUESTION: 77

You built a legal document assistant with Gemini. Each user session repeatedly asks questions about the same 300-page contract, and the contract text is resent with every request. Costs and time-to-first-token are both higher than acceptable. The contract does not change during a session. You need to reduce cost and latency. What should you do?

- A. Use context caching so the contract is processed once and reused across requests in the session.
- B. Summarize the contract to 5 pages, and send only the summary with each request.
- C. Lower max_output_tokens so the model returns shorter responses.
- D. Split the contract into 30 separate requests and merge the responses in the application.

Answer: ([SHOW ANSWER](#))

Context caching stores the tokenized, prefilled context so that repeated requests reference the cache instead of reprocessing the full document, which lowers input token cost and time-to-first-token. Summarization discards detail a legal assistant needs, and reducing output length does not address the input cost.

NEW QUESTION: 78

Your company manages a video sharing website where users can watch and upload videos. You need to create an ML model to predict which newly uploaded videos will be the most popular so that those videos can be prioritized on your company's website. Which result should you use to determine whether the model is successful?

- A. The model predicts videos as popular if the user who uploads them has over 10,000 likes.
- B. The model predicts 97.5% of the most popular clickbait videos measured by number of clicks.
- C. The model predicts 95% of the most popular videos measured by watch time within 30 days of being uploaded.
- D. The Pearson correlation coefficient between the log-transformed number of views after 7 days and 30 days after publication is equal to 0.

Answer: ([SHOW ANSWER](#))

<https://developers.google.com/machine-learning/problem-framing/framing#quantify-it>

NEW QUESTION: 79

Your team needs to build a model that predicts whether images contain a driver's license, passport, or credit card. The data engineering team already built the pipeline and generated a dataset composed of 10,000 images with driver's licenses, 1,000 images with passports, and 1,000 images with credit cards. You now have to train a model with the following label map:

`['drivers_license', 'passport', 'credit_card']`.

Which loss function should you use?

- A. Categorical hinge
- B. Binary cross-entropy
- C. Categorical cross-entropy
- D. Sparse categorical cross-entropy

Answer: C ([LEAVE A REPLY](#))

Use sparse categorical crossentropy when your classes are mutually exclusive (e.g. when each sample belongs exactly to one class) and categorical crossentropy when one sample can have multiple classes or labels are soft probabilities (like [0.5, 0.3, 0.2]).

NEW QUESTION: 80

Your company recently migrated several of its ML models to Google Cloud. You have started developing models in Vertex AI. You need to implement a system that tracks model artifacts and model lineage. You want to create a simple, effective solution that can also be reused for future models. What should you do?

- A. Use a combination of Vertex AI Pipelines and the Vertex AI SDK to integrate metadata tracking into the ML workflow.
- B. Use Vertex AI Pipelines for model artifacts and MLflow for model lineage.
- C. Use Vertex AI Experiments for model artifacts and use Vertex ML Metadata for model lineage.
- D. Implement a scheduled metadata tracking solution using Cloud Composer and Cloud Run functions.

Answer: A ([LEAVE A REPLY](#))

Using Vertex AI Pipelines with the Vertex AI SDK enables automatic tracking of model artifacts and lineage as part of the ML workflow. This approach integrates directly with the Vertex AI ecosystem, making it easy to reuse across different models and projects without introducing unnecessary complexity.

NEW QUESTION: 81

You work for a magazine publisher and have been tasked with predicting whether customers will cancel their annual subscription. In your exploratory data analysis, you find that 90% of individuals renew their subscription every year, and only 10% of individuals cancel their subscription. After training a NN Classifier, your model predicts those who cancel their subscription with 99% accuracy and predicts those who renew their subscription with 82% accuracy. How should you interpret these results?

- A. This is a good result because the accuracy across both groups is greater than 80%.
- B. This is a good result because predicting those who cancel their subscription is more difficult, since there is less data for this group.
- C. This is not a good result because the model is performing worse than predicting that people will always renew their subscription.
- D. This is not a good result because the model should have a higher accuracy for those who renew their subscription than for those who cancel their subscription.

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 82

You work on a team where the process for deploying a model into production starts with data scientists training different versions of models in a Kubeflow pipeline. The workflow then stores the new model artifact into the corresponding Cloud Storage bucket. You need to build the next steps of the pipeline after the submitted model is ready to be tested and deployed in production on AI Platform. How should you configure the architecture before deploying the model to production?

- A. Deploy model in test environment -> Evaluate and test model -> Create a new AI Platform model version
- B. Validate model -> Deploy model in test environment -> Create a new AI Platform model version
- C. Create a new AI Platform model version -> Evaluate and test model -> Deploy model in test environment
- D. Create a new AI Platform model version -> Deploy model in test environment -> Validate model

Answer: A (LEAVE A REPLY)

A is correct because the model can be validated after it is deployed to the test environment, and the release version is established before the model is deployed in production.

B is not correct because the model cannot be validated before being deployed to the test environment.

C is not correct because the model version is being set up for the release candidate before the model is validated.

Moreover, the model cannot be validated before being deployed to the test environment.

D is not correct because the model version is being set up for the release candidate before the model is validated.

<https://cloud.google.com/solutions/machine-learning/architecture-for-mlops-using-tfx-kubeflow-pipelines-and-cloud-build>

NEW QUESTION: 83

You work for a credit card company and have been asked to create a custom fraud detection model based on historical data using AutoML Tables. You need to prioritize detection of fraudulent transactions while minimizing false positives. Which optimization objective should you use when training the model?

- A. An optimization objective that minimizes Log loss
- B. An optimization objective that maximizes the Precision at a Recall value of 0.50
- C. An optimization objective that maximizes the area under the precision-recall curve (AUC PR) value
- D. An optimization objective that maximizes the area under the receiver operating characteristic curve (AUC ROC) value

Answer: C (LEAVE A REPLY)

AUC PR is the area under the precision-recall curve. It is a good metric to use for fraud detection because it takes into account the trade-off between precision and recall.

In this case, you need to prioritize detection of fraudulent transactions while minimizing false positives. This means that you need to maximize the precision of the model. AUC PR is a good metric to use for this because it takes into account the trade-off between precision and recall.

NEW QUESTION: 84

You are working on a binary classification ML algorithm that detects whether an image of a classified scanned document contains a company's logo. In the dataset, 96% of examples don't have the logo, so the dataset is very skewed. Which metric would give you the most confidence in your model?

- A. Precision
- B. Recall
- C. RMSE
- D. F1 score

Answer: D (LEAVE A REPLY)

<https://stephenallwright.com/imbalanced-data-metric/>

NEW QUESTION: 85

You work for a pet food company that manages an online forum. Customers upload photos of their pets on the forum to share with others. About 20 photos are uploaded daily. You want to automatically and in near real time detect whether each uploaded photo has an animal. You want to prioritize time and minimize cost of your application development and deployment. What should you do?

- A.** Download an object detection model from TensorFlow Hub. Deploy the model to a Vertex AI endpoint. Send new user-submitted images to the model endpoint to classify whether each photo has an animal.
- B.** Manually label previously submitted images with bounding boxes around any animals. Build an AutoML object detection model by using Vertex AI. Deploy the model to a Vertex AI endpoint. Send new user-submitted images to your model endpoint to detect whether each photo has an animal.
- C.** Send user-submitted images to the Cloud Vision API. Use object localization to identify all objects in the image and compare the results against a list of animals.
- D.** Manually label previously submitted images as having animals or not. Create an image dataset on Vertex AI. Train a classification model by using Vertex AutoML to distinguish the two classes. Deploy the model to a Vertex AI endpoint. Send new user-submitted images to your model endpoint to classify whether each photo has an animal.

Answer: C (LEAVE A REPLY)

NEW QUESTION: 86

You work for a large bank that serves customers through an application hosted in Google Cloud that is running in the US and Singapore. You have developed a PyTorch model to classify transactions as potentially fraudulent or not. The model is a three-layer perceptron that uses both numerical and categorical features as input, and hashing happens within the model.

You deployed the model to the us-central1 region on nl-highcpu-16 machines, and predictions are served in real time. The model's current median response latency is 40 ms. You want to reduce latency, especially in Singapore, where some customers are experiencing the longest delays.

What should you do?

- A.** Attach an NVIDIA T4 GPU to the machines being used for online inference.
- B.** Change the machines being used for online inference to nl-highcpu-32.
- C.** Deploy the model to Vertex AI private endpoints in the us-central1 and asia-southeast1 regions, and allow the application to choose the appropriate endpoint.
- D.** Create another Vertex AI endpoint in the asia-southeast1 region, and allow the application to choose the appropriate endpoint.

Answer: (SHOW ANSWER)

By having an endpoint in the asia-southeast1 region (Singapore), the data doesn't have to travel as far, significantly reducing the round-trip time. Allowing the application to choose the appropriate endpoint based on the user's location ensures that requests are handled by the nearest available server, optimizing response times for users in different regions.

NEW QUESTION: 87

You have a custom job that runs on Vertex AI on a weekly basis. The job is implemented using a proprietary ML workflow that produces the datasets, models, and custom artifacts, and sends them to a Cloud Storage bucket. Many different versions of the datasets and models were created. Due to compliance requirements, your company needs to track which model was used for making a particular prediction, and needs access to the artifacts for each model. How should you configure your workflows to meet these requirements?

- A.** Register each model in Vertex AI Model Registry, and use model labels to store the related dataset and model information.
- B.** Use the Vertex AI Metadata API inside the custom job to create context, execution, and artifacts for each model, and use events to link them together.
- C.** Create a Vertex AI experiment, and enable autologging inside the custom job.
- D.** Configure a TensorFlow Extended (TFX) ML Metadata database, and use the ML Metadata API.

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 88

Your organization manages an online message board. A few months ago, you discovered an increase in toxic language and bullying on the message board. You deployed an automated text classifier that flags certain comments as toxic or harmful. Now some users are reporting that benign comments referencing their religion are being misclassified as abusive. Upon further inspection, you find that your classifier's false positive rate is higher for comments that reference certain underrepresented religious groups. Your team has a limited budget and is already overextended. What should you do?

- A.** Raise the threshold for comments to be considered toxic or harmful.
- B.** Replace your model with a different text classifier.
- C.** Add synthetic training data where those phrases are used in non-toxic ways.
- D.** Remove the model and replace it with human moderation.

Answer: A ([LEAVE A REPLY](#))

NEW QUESTION: 89

You are developing a TensorFlow Extended (TFX) pipeline with standard TFX components. The pipeline includes data preprocessing steps. After the pipeline is deployed to production, it will process up to 100 TB of data stored in BigQuery. You need the data preprocessing steps to scale efficiently, publish metrics and parameters to Vertex AI Experiments, and track artifacts by using Vertex ML Metadata. How should you configure the pipeline run?

- A.** Run the TFX pipeline in Vertex AI Pipelines. Configure the pipeline to use Vertex AI Training jobs with distributed processing.
- B.** Run the TFX pipeline in Vertex AI Pipelines. Set the appropriate Apache Beam parameters in the pipeline to run the data preprocessing steps in Dataflow.
- C.** Run the TFX pipeline in Dataproc by using the Apache Beam TFX orchestrator. Set the appropriate Vertex AI permissions in the job to publish metadata in Vertex AI.
- D.** Run the TFX pipeline in Dataflow by using the Apache Beam TFX orchestrator. Set the appropriate Vertex AI permissions in the job to publish metadata in Vertex AI.

Answer: B (LEAVE A REPLY)

This configuration allows you to leverage the scalability of Dataflow for processing large datasets (up to 100 TB) efficiently while running the TFX pipeline within Vertex AI Pipelines. By using Apache Beam parameters, you can optimize the data preprocessing steps to handle large volumes of data effectively. Additionally, integrating with Vertex AI Pipelines ensures that you can publish metrics and parameters to Vertex AI Experiments and track artifacts using Vertex ML Metadata, fulfilling all your requirements for scalability, monitoring, and metadata management.

NEW QUESTION: 90

You are using Vertex AI to manage your ML models and datasets. You recently updated one of your models. You want to track and compare the new version with the previous one and incorporate dataset versioning. What should you do?

- A.** Use Vertex AI TensorBoard to visualize the training metrics of the new model version, and use Data Catalog to manage dataset versioning.
- B.** Use Vertex AI Model Monitoring to monitor the performance of the new model version, and use Vertex AI Training to manage dataset versioning.
- C.** Use Vertex AI Experiments to track and compare model artifacts and versions, and use Vertex ML Metadata to manage dataset versioning.
- D.** Use Vertex AI Experiments to track and compare model artifacts and versions, and use Vertex AI managed datasets to manage dataset versioning.

Answer: C (LEAVE A REPLY)

Vertex AI Experiments is designed to track and compare model versions and their training artifacts, while Vertex ML Metadata manages dataset versioning and lineage, providing an integrated approach to model and dataset version control.

NEW QUESTION: 91

You work for a logistics company that receives thousands of PDF invoices each day from various freight carriers. The invoices are scanned and saved as images in a Cloud Storage bucket. Each invoice contains important information such as invoice number, invoice date, supplier name, and invoice amount. However, the layout and format vary significantly from carrier to carrier.

You need to build a pipeline that extracts the fields from all incoming invoices and loads the structured data into a BigQuery table for analysis. You want the solution to be scalable while also minimizing the need for manual data entry and custom parsing logic. What should you do?

- A.** Create a Cloud Run service that uses the Gemini Pro model to read documents to identify and return the required values in JSON format. Create another Cloud Run service to load the JSON files into a Big Query table. Run the Cloud Run services sequentially.
- B.** Use the Cloud Vision API's OCR feature on a Compute Engine VM to extract all raw text from each document. Write a Python script with regular expressions to search for keywords and parse the extracted text to find the required field values.
- C.** Schedule a notebook in an Agent Platform Workbench instance to periodically train a custom named entity recognition (NER) model on a labeled dataset of invoice documents. Deploy the model to an Agent Platform endpoint, and post-process the model output to write results to a BigQuery table.

D. Configure a processor using Document AI's pre-trained invoice processor, and send each document to its endpoint. Extract the structured entities from the processor's JSON output.

Answer: D (LEAVE A REPLY)

Document AI's pre-trained Invoice Parser is designed to process invoices with varied layouts and extract structured entities such as invoice number, invoice date, supplier name, and invoice amount. It provides scalable document processing and structured JSON output without requiring custom parsing rules or a manually trained model.

Reference:

<https://docs.cloud.google.com/document-ai/docs/pretrained-overview>

<https://docs.cloud.google.com/document-ai/docs/overview>

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF Special Discount: Freepdfdumps**)

NEW QUESTION: 92

You have deployed a model on Vertex AI for real-time inference. During an online prediction request, you get an "Out of Memory" error. What should you do?

- A.** Use batch prediction mode instead of online mode.
- B.** Send the request again with a smaller batch of instances.
- C.** Use base64 to encode your data before using it for prediction.
- D.** Apply for a quota increase for the number of prediction requests.

Answer: B (LEAVE A REPLY)

429 - Out of Memory

<https://cloud.google.com/ai-platform/training/docs/troubleshooting>

NEW QUESTION: 93

Your organization's employee onboarding team wants you to build an interactive self-help tool for new employees. The tool needs to receive queries from users and provide answers from the organization's internal documentation. This documentation is spread across standalone documents such as PDF files. You want to build a solution quickly while minimizing maintenance overhead. What should you do?

- A.** Create a custom chatbot user interface hosted on App Engine. Use Vertex AI to fine-tune a Gemini model on the organization's internal documentation. Send users' queries to the fine-tuned model by using the custom chatbot and return the model's responses to the users.
- B.** Deploy an internal website to a Google Kubernetes Engine (GKE) cluster. Build a search index by ingesting all of the organization's internal documentation. Use Vertex AI Vector Search to implement a semantic search that retrieves results from the search index based on the query entered into the search box.

C. Use Vertex AI Agent Builder to create an agent. Securely index the organization's internal documentation to the agent's datastore. Send users' queries to the agent and return the agent's grounded responses to the users.

D. Deploy an internal website to a Google Kubernetes Engine (GKE) cluster. Organize the relevant internal documentation into sections. Collect user feedback on website content and store it in BigQuery. Request that the onboarding team regularly update the links based on user feedback.

Answer: ([SHOW ANSWER](#))

Vertex AI Agent Builder is specifically designed for building conversational AI agents quickly with minimal overhead. It allows for secure indexing of internal documents, making them accessible for semantic query resolution. By indexing the internal documentation in the agent's datastore, the system can provide grounded and accurate responses to user queries based on the organization's documentation. This approach minimizes the maintenance overhead as the Agent Builder handles much of the complexity, including semantic search and conversational interface.

NEW QUESTION: 94

You are training an ML model using data stored in BigQuery that contains several values that are considered Personally Identifiable Information (PII). You need to reduce the sensitivity of the dataset before training your model. Every column is critical to your model. How should you proceed?

A. Using Dataflow, ingest the columns with sensitive data from BigQuery, and then randomize the values in each sensitive column.

B. Use the Cloud Data Loss Prevention (DLP) API to scan for sensitive data, and use Dataflow with the DLP API to encrypt sensitive values with Format Preserving Encryption.

C. Use the Cloud Data Loss Prevention (DLP) API to scan for sensitive data, and use Dataflow to replace all sensitive data by using the encryption algorithm AES-256 with a salt.

D. Before training, use BigQuery to select only the columns that do not contain sensitive data.

Create an authorized view of the data so that sensitive values cannot be accessed by unauthorized individuals.

Answer: ([SHOW ANSWER](#))

https://cloud.google.com/dlp/docs/transformations-reference#types_of_de-identification_techniques

<https://cloud.google.com/dlp/docs/transformations-reference#crypto>

NEW QUESTION: 95

You are developing a model to detect fraudulent credit card transactions. You need to prioritize detection, because missing even one fraudulent transaction could severely impact the credit card holder. You used AutoML to train a model on users' profile information and credit card transaction data. After training the initial model, you notice that the model is failing to detect many fraudulent transactions. How should you adjust the training parameters in AutoML to improve model performance? (Choose two.)

A. Increase the score threshold

B. Add more positive examples to the training set

C. Decrease the score threshold.

D. Add more negative examples to the training set

E. Reduce the maximum number of node hours for training

Answer: B,C ([LEAVE A REPLY](#))

NEW QUESTION: 96

You work for the AI team of an automobile company, and you are developing a visual defect detection model using TensorFlow and Keras. To improve your model performance, you want to incorporate some image augmentation functions such as translation, cropping, and contrast tweaking. You randomly apply these functions to each training batch. You want to optimize your data processing pipeline for run time and compute resources utilization. What should you do?

- A. Embed the augmentation functions dynamically in the tf.Data pipeline.
- B. Embed the augmentation functions dynamically as part of Keras generators.
- C. Use Dataflow to create all possible augmentations, and store them as TFRecords.
- D. Use Dataflow to create the augmentations dynamically per training run, and stage them as TFRecords.

Answer: (SHOW ANSWER)

https://www.tensorflow.org/tutorials/load_data/images#using_tfdata_for_finer_control

NEW QUESTION: 97

You need to analyze user activity data from your company's mobile applications. Your team will use BigQuery for data analysis, transformation, and experimentation with ML algorithms. You need to ensure real-time ingestion of the user activity data into BigQuery. What should you do?

- A. Configure Pub/Sub to stream the data into BigQuery.
- B. Run an Apache Spark streaming job on Dataproc to ingest the data into BigQuery.
- C. Run a Dataflow streaming job to ingest the data into BigQuery.
- D. Configure Pub/Sub and a Dataflow streaming job to ingest the data into BigQuery.

Answer: A (LEAVE A REPLY)

<https://cloud.google.com/blog/products/data-analytics/pub-sub-launches-direct-path-to-bigquery-for-streaming-analytics>

NEW QUESTION: 98

You are tasked with building an MLOps pipeline to retrain tree-based models in production. The pipeline will include components related to data ingestion, data processing, model training, model evaluation, and model deployment. Your organization primarily uses PySpark-based workloads for data preprocessing. You want to minimize infrastructure management effort. How should you set up the pipeline?

- A. Set up a TensorFlow Extended (TFX) pipeline on Vertex AI Pipelines to orchestrate the MLOps pipeline. Write a custom component for the PySpark-based workloads on Dataproc.
- B. Set up a Vertex AI Pipelines to orchestrate the MLOps pipeline. Use the predefined Dataproc component for the PySpark-based workloads.
- C. Set up Kubeflow Pipelines on Google Kubernetes Engine to orchestrate the MLOps pipeline. Write a custom component for the PySparkbased workloads on Dataproc.
- D. Set up Cloud Composer to orchestrate the MLOps pipeline. Use Dataproc workflow templates for the PySpark-based workloads in Cloud Composer.

Answer: B (LEAVE A REPLY)

This approach minimizes infrastructure management effort by leveraging Vertex AI Pipelines, a managed service designed for orchestrating machine learning workflows, which simplifies pipeline management. By using the predefined

Dataproc component, you can easily integrate PySpark-based workloads running on Dataproc without the need to write custom components.

This setup provides a scalable and efficient solution for orchestrating the entire MLOps pipeline, from data ingestion to deployment, with minimal manual intervention for infrastructure management.

NEW QUESTION: 99

You work for a telecommunications company. You're building a model to predict which customers may fail to pay their next phone bill. The purpose of this model is to proactively offer at-risk customers assistance such as service discounts and bill deadline extensions. The data is stored in BigQuery and the predictive features that are available for model training include:

- Customer_id
- Age
- Salary (measured in local currency)
- Sex
- Average bill value (measured in local currency)
- Number of phone calls in the last month (integer)
- Average duration of phone calls (measured in minutes)

You need to investigate and mitigate potential bias against disadvantaged groups, while preserving model accuracy. What should you do?

- A.** Train a BigQuery ML boosted trees classification model with all features. Use the `ML.GLOBAL_EXPLAIN` method to calculate the global attribution values for each feature of the model. If the feature importance value for any of the sensitive features exceeds a threshold, discard the model and train without this feature.
- B.** Define a fairness metric that is represented by accuracy across the sensitive features. Train a BigQuery ML boosted trees classification model with all features. Use the trained model to make predictions on a test set. Join the data back with the sensitive features, and calculate a fairness metric to investigate whether it meets your requirements.
- C.** Train a BigQuery ML boosted trees classification model with all features. Use the `ML.EXPLAIN_PREDICT` method to calculate the attribution values for each feature for each customer in a test set. If for any individual customer, the importance value for any feature exceeds a predefined threshold, discard the model and train the model again without this feature.
- D.** Determine whether there is a meaningful correlation between the sensitive features and the other features. Train a BigQuery ML boosted trees classification model and exclude the sensitive features and any meaningfully correlated features.

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 100

You work for a biotech startup that is experimenting with deep learning ML models based on properties of biological organisms. Your team frequently works on early-stage experiments with new architectures of ML models, and writes custom TensorFlow ops in C++. You train your models on large datasets and large batch sizes. Your typical batch size has 1024 examples, and each example is about 1 MB in size. The average size of a network with all weights and embeddings is 20 GB. What hardware should you choose for your models?

- A.** A cluster with 2 n1-highcpu-64 machines, each with 8 NVIDIA Tesla V100 GPUs (128 GB GPU memory in total), and a n1-highcpu-64 machine with 64 vCPUs and 58 GB RAM
- B.** A cluster with 2 a2-megagpu-16g machines, each with 16 NVIDIA Tesla A100 GPUs (640 GB GPU memory in total), 96 vCPUs, and 1.4 TB RAM
- C.** A cluster with an n1-highcpu-64 machine with a v2-8 TPU and 64 GB RAM
- D.** A cluster with 4 n1-highcpu-96 machines, each with 96 vCPUs and 86 GB RAM

Answer: D (LEAVE A REPLY)

"writes custom TensorFlow ops in C++" -> use CPUs when "Models that contain many custom TensorFlow operations written in C++": https://cloud.google.com/tpu/docs/intro-to-tpu#when_to_use_tpus

NEW QUESTION: 101

You work for an ecommerce company that wants to automatically classify products in images to improve user experience. You have a substantial dataset of labeled images depicting various unique products. You need to implement a solution for identifying custom products that is scalable, effective, and can be rapidly deployed. What should you do?

- A.** Develop a rule-based system to categorize the images.
- B.** Use a TensorFlow deep learning model that is trained on the image dataset.
- C.** Use a pre-trained object detection model from Model Garden.
- D.** Use AutoML Vision to train a model using the image dataset.

Answer: D (LEAVE A REPLY)

AutoML Vision is an ideal choice for this use case because it provides a scalable and easy-to-use platform for training custom machine learning models on your labeled image dataset. It eliminates the need for deep technical expertise in building and fine-tuning models while leveraging advanced techniques to achieve high accuracy. With AutoML Vision, you can train a custom model specifically tailored to classify your unique products, ensuring better performance compared to a generic pre-trained model.

NEW QUESTION: 102

You are an ML engineer at a retail company. You have built a model that predicts a coupon to offer an ecommerce customer at checkout based on the items in their cart. When a customer goes to checkout, your serving pipeline, which is hosted on Google Cloud, joins the customer's existing cart with a row in a BigQuery table that contains the customers' historic purchase behavior and uses that as the model's input. The web team is reporting that your model is returning predictions too slowly to load the coupon offer with the rest of the web page. How should you speed up your model's predictions?

- A.** Attach an NVIDIA P100 GPU to your deployed model's instance.
- B.** Use a low latency database for the customers' historic purchase behavior.
- C.** Deploy your model to more instances behind a load balancer to distribute traffic.
- D.** Create a materialized view in BigQuery with the necessary data for predictions.

Answer: (SHOW ANSWER)

NEW QUESTION: 103

You have built a custom model that performs several memory-intensive preprocessing tasks before it makes a prediction. You deployed the model to a Vertex AI endpoint, and validated that results were received in a reasonable amount of

time. After routing user traffic to the endpoint, you discover that the endpoint does not autoscale as expected when receiving multiple requests.

What should you do?

- A. Increase the CPU utilization target in the autoscaling configurations.
- B. Use a machine type with more memory
- C. Decrease the number of workers per machine
- D. Decrease the CPU utilization target in the autoscaling configurations

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 104

You have a high-traffic ecommerce application that uses a recommendation model deployed on an Agent Platform endpoint. During flash sales, traffic quickly spikes from 1,000 to 50,000 queries per second (QPS). During these spikes, the application returns 503: Service Unavailable errors because the autoscaler cannot provision new resources quickly enough to meet the sudden demand. Your need to eliminate service disruptions during the start of the sale. How should you configure the endpoint?

- A. Change the machine type to an a2-highgpu-1g instance with an NVIDIA A100 GPU.
- B. Increase the min_replica_count to a baseline that covers a significant portion of the expected spike, and set the max_replica_count to 100.
- C. Set the min_replica_count to 1 and the max_replica_count to 100 in the deployed model's automatic scaling configuration.
- D. Deploy the model to a private Agent Platform endpoint, and autoscale based on GPU utilization.

Answer: B ([LEAVE A REPLY](#))

Increasing the minimum replica count pre-provisions enough serving capacity to absorb the initial traffic surge before autoscaling can add more replicas. Setting a sufficiently high maximum replica count then allows the endpoint to continue scaling as demand grows, preventing startup- period 503 errors during flash sales.

NEW QUESTION: 105

You are an ML engineer at a large grocery retailer with stores in multiple regions. You have been asked to create an inventory prediction model. Your model's features include region, location, historical demand, and seasonal popularity. You want the algorithm to learn from new inventory data on a daily basis. Which algorithms should you use to build the model?

- A. Classification
- B. Reinforcement Learning
- C. Recurrent Neural Networks (RNN)
- D. Convolutional Neural Networks (CNN)

Answer: C ([LEAVE A REPLY](#))

Recurrent Neural Networks (RNN) are a type of neural network that are used to process sequential data. They are well-suited for this task because the inventory data is sequential in nature. The RNN can learn from the historical demand data and the seasonal popularity data to predict the future inventory levels.

NEW QUESTION: 106

You are optimizing the training of a 175-billion parameter LLM on Gemini Enterprise Agent Platform. You have provisioned a TPU v5p Pod slice with 32 chips for the job. You initially ran the training using standard data parallelism, but the job immediately failed with an out-of-memory (OOM) error. You need to implement a training strategy that resolves the memory issue and minimizes training latency. What should you do?

- A.** Configure the training to use a parameter server architecture where the model weights are stored in the host CPU memory.
- B.** Enable gradient checkpointing, and continue using the data parallelism strategy.
- C.** Configure the training to shard the model parameters and optimizer states, and continue using the data parallelism strategy.
- D.** Scale up the TPU Pod slice to include more chips, and continue using the data parallelism strategy.

Answer: C ([LEAVE A REPLY](#))

Sharding the model parameters and optimizer states across the TPU chips distributes the memory footprint instead of replicating the full 175-billion-parameter model on every device. This resolves the out-of-memory failure while retaining data parallelism for high-throughput training and lower overall training latency.

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF** Special Discount: **Freepdfdumps**)

NEW QUESTION: 107

You are training and deploying updated versions of a regression model with tabular data by using Vertex AI Pipelines, Vertex AI Training, Vertex AI Experiments, and Vertex AI Endpoints. The model is deployed in a Vertex AI endpoint, and your users call the model by using the Vertex AI endpoint. You want to receive an email when the feature data distribution changes significantly, so you can retrigger the training pipeline and deploy an updated version of your model. What should you do?

- A.** Use Vertex AI Model Monitoring. Enable prediction drift monitoring on the endpoint, and specify a notification email.
- B.** In Cloud Monitoring create a logs-based metric and a threshold alert for the metric. Configure Cloud Monitoring to send an email when the alert is triggered.
- C.** Export the container logs of the endpoint to BigQuery. Create a Cloud Function to run a SQL query over the exported logs and send an email. Use Cloud Scheduler to trigger the Cloud Function.
- D.** In Cloud Logging, create a logs-based alert using the logs in the Vertex AI endpoint. Configure Cloud Logging to send an email when the alert is triggered.

Answer: A ([LEAVE A REPLY](#))

NEW QUESTION: 108

You are an ML engineer at a bank. The bank's leadership team wants to reduce the number of loan defaults. The bank has labeled historic data about loan defaults stored in BigQuery. You have been asked to use AI to support the loan application process. For compliance reasons, you need to provide explanations for loan rejections. What should you do?

- A.** Import the historic loan default data into AutoML. Train and deploy a linear regression model to predict default probability. Report the probability of default for each loan application.
- B.** Create a custom application that uses the Gemini large language model (LLM). Provide the historic data as context to the model, and prompt the model to predict customer defaults. Report the prediction and explanation provided by the LLM for each loan application.
- C.** Train and deploy a BigQuery ML classification model trained on historic loan default data. Enable feature-based explanations for each prediction. Report the prediction, probability of default, and feature attributions for each loan application.
- D.** Load the historic loan default data into a Vertex AI Workbench instance. Train a deep learning classification model using TensorFlow to predict loan default. Run inference for each loan application, and report the predictions.

Answer: ([SHOW ANSWER](#))

BigQuery ML supports training classification models directly on data stored in BigQuery and provides built-in support for feature-based explanations using `ML.EXPLAIN_PREDICT`. This satisfies the compliance requirement for explainability while enabling accurate predictions and seamless integration with existing data infrastructure.

NEW QUESTION: 109

You have trained a text classification model in TensorFlow using AI Platform. You want to use the trained model for batch predictions on text data stored in BigQuery while minimizing computational overhead.

What should you do?

- A.** Export the model to BigQuery ML.
- B.** Deploy and version the model on AI Platform.
- C.** Use Dataflow with the SavedModel to read the data from BigQuery.
- D.** Submit a batch prediction job on AI Platform that points to the model location in Cloud Storage.

Answer: A ([LEAVE A REPLY](#))

https://cloud.google.com/bigquery-ml/docs/making-predictions-with-imported-tensorflow-models#importing_models

NEW QUESTION: 110

You work for an online retail company that is creating a visual search engine. You have set up an end-to-end ML pipeline on Google Cloud to classify whether an image contains your company's product. Expecting the release of new products in the near future, you configured a retraining functionality in the pipeline so that new data can be fed into your ML models. You also want to use AI Platform's continuous evaluation service to ensure that the models have high accuracy on your test dataset. What should you do?

- A.** Keep the original test dataset unchanged even if newer products are incorporated into retraining.
- B.** Extend your test dataset with images of the newer products when they are introduced to retraining.
- C.** Replace your test dataset with images of the newer products when they are introduced to retraining.
- D.** Update your test dataset with images of the newer products when your evaluation metrics drop below a pre-decided threshold.

Answer: B ([LEAVE A REPLY](#))

By extending your test dataset with images of the newer products, you can ensure that your ML models are still able to classify images of the older products accurately, even after they have been retrained with data from the newer products. This will help you to maintain a high level of accuracy for your visual search engine.

NEW QUESTION: 111

You are developing a process for training and running your custom model in production. You need to be able to show lineage for your model and predictions. What should you do?

- A.** 1. Create a Vertex AI managed dataset.
2. Use a Vertex AI training pipeline to train your model.
3. Generate batch predictions in Vertex AI.
- B.** 1. Upload your dataset to BigQuery.
2. Use a Vertex AI custom training job to train your model.
3. Generate predictions by using Vertex AI SDK custom prediction routines.
- C.** 1. Use Vertex AI Experiments to train your model.
2. Register your model in Vertex AI Model Registry.
3. Generate batch predictions in Vertex AI.
- D.** 1. Use a Vertex AI Pipelines custom training job component to train your model.
2. Generate predictions by using a Vertex AI Pipelines model batch predict component.

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 112

You have trained a model by using data that was preprocessed in a batch Dataflow pipeline. Your use case requires real-time inference. You want to ensure that the data preprocessing logic is applied consistently between training and serving. What should you do?

- A.** Refactor the transformation code in the batch data pipeline so that it can be used outside of the pipeline. Share this code with the end users of the endpoint.
- B.** Batch the real-time requests by using a time window and then use the Dataflow pipeline to preprocess the batched requests. Send the preprocessed requests to the endpoint.
- C.** Perform data validation to ensure that the input data to the pipeline is the same format as the input data to the endpoint.
- D.** Refactor the transformation code in the batch data pipeline so that it can be used outside of the pipeline. Use the same code in the endpoint.

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 113

You are building an ML pipeline to process and analyze both streaming and batch datasets. You need the pipeline to handle data validation, preprocessing, model training, and model deployment in a consistent and automated way. You want to design an efficient and scalable solution that captures model training metadata and is easily reproducible. You want to be able to reuse custom components for different parts of your pipeline. What should you do?

- A.** Use Cloud Composer for distributed processing of batch and streaming data in the pipeline.
- B.** Use Dataflow for distributed processing of batch and streaming data in the pipeline.

- C. Use Cloud Build to build and push Docker images for each pipeline component.
- D. Implement an orchestration framework such as Kubeflow Pipelines or Vertex AI Pipelines.

Answer: D (LEAVE A REPLY)

Vertex AI Pipelines (or Kubeflow Pipelines) are designed for orchestrating end-to-end ML workflows. They support consistent execution of validation, preprocessing, training, and deployment steps, capture metadata automatically, allow for easy reproducibility, and promote reuse of custom components. This makes them the most efficient and scalable solution for your use case.

NEW QUESTION: 114

You built a custom Vertex AI pipeline job that preprocesses images and trains an object detection model. The pipeline currently uses 1 n1-standard-8 machine with 1 NVIDIA Tesla V100 GPU.

You want to reduce the model training time without compromising model accuracy. What should you do?

- A. Reduce the number of layers in your object detection model.
- B. Train the same model on a stratified subset of your dataset.
- C. Update the WorkerPoolSpec to use a machine with 24 vCPUs and 1 NVIDIA Tesla V100 GPU.
- D. Update the WorkerPoolSpec to use a machine with 24 vCPUs and 3 NVIDIA Tesla V100 GPUs.

Answer: D (LEAVE A REPLY)

Increasing the number of GPUs while scaling CPU resources allows parallel training, which reduces training time without compromising accuracy. Using 3 NVIDIA Tesla V100 GPUs provides more compute power for efficient model training in Vertex AI Pipelines.

NEW QUESTION: 115

You work at an ecommerce startup. You need to create a customer churn prediction model. Your company's recent sales records are stored in a BigQuery table. You want to understand how your initial model is making predictions. You also want to iterate on the model as quickly as possible while minimizing cost. How should you build your first model?

- A. Export the data to a Cloud Storage bucket. Create a `tf.data.Dataset` to read the data from Cloud Storage. Implement a deep neural network in TensorFlow.
- B. Prepare the data in BigQuery and associate the data with a Vertex AI dataset. Create an `AutoMLTabularTrainingJob` to train a classification model.
- C. Create a `tf.data.Dataset` by using the TensorFlow `BigQueryClient`. Implement a deep neural network in TensorFlow.
- D. Export the data to a Cloud Storage bucket. Load the data into a pandas `DataFrame` on Vertex AI Workbench and train a logistic regression model with scikit-learn.

Answer: B (LEAVE A REPLY)

NEW QUESTION: 116

You work on a team that builds state-of-the-art deep learning models by using the TensorFlow framework. Your team runs multiple ML experiments each week, which makes it difficult to track the experiment runs. You want a simple approach to effectively track, visualize, and debug ML experiment runs on Google Cloud while minimizing any overhead code. How should you proceed?

- A. Set up Vertex AI Experiments to track metrics and parameters. Configure Vertex AI TensorBoard for visualization.

B. Set up a Vertex AI Workbench notebook instance. Use the instance to save metrics data in a Cloud Storage bucket and to host TensorBoard locally for visualization.

C. Set up a Cloud Function to write and save metrics files to a Cloud Storage bucket. Configure a Google Cloud VM to host TensorBoard locally for visualization.

D. Set up a Cloud Function to write and save metrics files to a BigQuery table. Configure a Google Cloud VM to host TensorBoard locally for visualization.

Answer: A ([LEAVE A REPLY](#))

NEW QUESTION: 117

You are an ML engineer at a global shoe store. You manage the ML models for the company's website. You are asked to build a model that will recommend new products to the user based on their purchase behavior and similarity with other users. What should you do?

A. Build a classification model

B. Build a knowledge-based filtering model

C. Build a collaborative-based filtering model

D. Build a regression model using the features as predictors

Answer: ([SHOW ANSWER](#)**)**

Collaborative filtering is a technique that recommends items to users based on the ratings of other users. It works by finding users who have similar ratings to the current user and then recommending items that those users have liked.

https://cloud.google.com/architecture/recommendations-using-machine-learning-on-compute-engine#filtering_the_data

NEW QUESTION: 118

You built and manage a production system that is responsible for predicting sales numbers.

Model accuracy is crucial, because the production model is required to keep up with market changes. Since being deployed to production, the model hasn't changed; however the accuracy of the model has steadily deteriorated. What issue is most likely causing the steady decline in model accuracy?

A. Poor data quality

B. Lack of model retraining

C. Too few layers in the model for capturing information

D. Incorrect data split ratio during model training, evaluation, validation, and test

Answer: ([SHOW ANSWER](#)**)**

Retraining is needed as the market is changing. Its how the Model keep updated and predictions accuracy.

NEW QUESTION: 119

You manage an ML workflow for a fraud detection model. The source code for data processing and model training is stored in a version control repository. You need to implement a continuous training (CT) pipeline that automatically runs unit tests and launches a new training run on Agent Platform Pipelines when code is updated in the repository. You want to use a fully managed service to orchestrate this process. What should you do?

A. Create a Cloud Run function that is triggered by repository commits. The function runs the unit tests, launches a custom training job, and uses Model Monitoring to validate the new model's performance before deployment.

B. Define a Cloud Workflow to monitor the repository for changes. Configure the workflow on Agent Platform Pipelines to orchestrate and run the unit tests. Use the model evaluation component to validate the new model's performance before deployment.

C. Configure Agent Platform Pipelines to poll the repository for updates. Add a conditional step in the pipeline that uses Model Monitoring to validate the new model's performance before deployment.

D. Create a Cloud Build trigger associated with the repository. Define a build configuration file that runs the unit tests and triggers a job on Agent Platform Pipelines. Use the model evaluation component to validate the model performance before deployment.

Answer: ([SHOW ANSWER](#))

Cloud Build provides fully managed, repository-triggered CI automation. A build trigger can run unit tests whenever code changes are committed and, after successful validation, invoke the Agent Platform Pipeline for retraining. The pipeline can then evaluate the new model before allowing deployment.

NEW QUESTION: 120

You are developing an image recognition model using PyTorch based on ResNet50 architecture.

Your code is working fine on your local laptop on a small subsample. Your full dataset has 200k labeled images. You want to quickly scale your training workload while minimizing cost. You plan to use 4 V100 GPUs. What should you do?

A. Create a Google Kubernetes Engine cluster with a node pool that has 4 V100 GPUs. Prepare and submit a TFJob operator to this node pool.

B. Create a Vertex AI Workbench user-managed notebooks instance with 4 V100 GPUs, and use it to train your model.

C. Configure a Compute Engine VM with all the dependencies that launches the training. Train your model with Vertex AI using a custom tier that contains the required GPUs.

D. Package your code with Setuptools, and use a pre-built container. Train your model with Vertex AI using a custom tier that contains the required GPUs.

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 121

You are training a TensorFlow model on a structured dataset with 100 billion records stored in several CSV files. You need to improve the input/output execution performance. What should you do?

A. Load the data into BigQuery, and read the data from BigQuery.

B. Load the data into Cloud Bigtable, and read the data from Bigtable.

C. Convert the CSV files into shards of TFRecords, and store the data in Cloud Storage.

D. Convert the CSV files into shards of TFRecords, and store the data in the Hadoop Distributed File System (HDFS).

Answer: ([SHOW ANSWER](#))

Cloud Bigtable is typically used to process unstructured data, such as time-series data, logs, or other types of data that do not conform to a fixed schema. However, Cloud Bigtable can also be used to store structured data if necessary, such as in the case of a key-value store or a database that does not require complex relational queries.

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF** Special Discount: **Freepdfdumps**)

NEW QUESTION: 122

You are developing a model to help your company create more targeted online advertising campaigns. You need to create a dataset that you will use to train the model. You want to avoid creating or reinforcing unfair bias in the model. What should you do? (Choose two.)

- A. Include a comprehensive set of demographic features
- B. Collect a stratified sample of production traffic to build the training dataset
- C. Conduct fairness tests across sensitive categories and demographics on the trained model
- D. Include only the demographic groups that most frequently interact with advertisements
- E. Collect a random sample of production traffic to build the training dataset

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 123

You work on the data science team for a multinational beverage company. You need to develop an ML model to predict the company's profitability for a new line of naturally flavored bottled waters in different locations. You are provided with historical data that includes product types, product sales volumes, expenses, and profits for all regions. What should you use as the input and output for your model?

- A. Use latitude, longitude, and product type as features. Use profit as model output.
- B. Use latitude, longitude, and product type as features. Use revenue and expenses as model outputs.
- C. Use product type and the feature cross of latitude with longitude, followed by binning, as features. Use profit as model output.
- D. Use product type and the feature cross of latitude with longitude, followed by binning, as features. Use revenue and expenses as model outputs.

Answer: ([SHOW ANSWER](#))

<https://developers.google.com/machine-learning/crash-course/feature-crosses/video-lecture>

<https://developers.google.com/machine-learning/crash-course/regularization-for-sparsity/l1-regularization>

NEW QUESTION: 124

Your organization's call center has asked you to develop a model that analyzes customer sentiments in each call. The call center receives over one million calls daily, and data is stored in Cloud Storage. The data collected must not leave the region in which the call originated, and no Personally Identifiable Information (PII) can be stored or analyzed. The data science team has a third-party tool for visualization and access which requires a SQL ANSI-2011 compliant interface.

You need to select components for data processing and for analytics. How should the data pipeline be designed?



- A. 1 = Dataflow, 2 = BigQuery
- B. 1 = Pub/Sub, 2 = Datastore
- C. 1 = Dataflow, 2 = Cloud SQL
- D. 1 = Cloud Function, 2 = Cloud SQL

Answer: A (LEAVE A REPLY)

Dataflow - Unified stream and batch data processing that's serverless, fast, and cost-effective BigQuery - Good for analytics and dashboards

<https://github.com/GoogleCloudPlatform/dataflow-contact-center-speech-analysis>

NEW QUESTION: 125

You are building an internal assistant that answers employee questions about a 12,000-page HR policy corpus. The corpus is updated weekly, and answers must cite the specific policy section.

Hallucinated answers have been reported during testing. You need to improve factual accuracy while keeping the content current. What should you do?

- A. Fine-tune Gemini on the policy corpus each week so the model memorizes the latest policies.
- B. Increase the model temperature to 0 and instruct the model in the system prompt to answer only from HR policies.
- C. Chunk and embed the corpus into Vector Search, retrieve the top-k relevant chunks at query time, and pass them to Gemini as grounding context with their source identifiers.
- D. Place the entire corpus in the system instruction of every request, and instruct the model to quote the relevant section.

Answer: C (LEAVE A REPLY)

Retrieval-augmented generation grounds the model in retrieved passages, which reduces hallucination and allows the response to cite the originating chunk. Re-embedding only changed documents keeps the index current at low cost, whereas weekly fine-tuning is expensive and does not reliably provide citations.

NEW QUESTION: 126

While running a model training pipeline on Vertex AI, you discover that the evaluation step is failing because of an out-of-memory error. You are currently using TensorFlow Model Analysis (TFMA) with a standard Evaluator TensorFlow Extended (TFX) pipeline component for the evaluation step. You want to stabilize the pipeline without downgrading the evaluation quality while minimizing infrastructure overhead. What should you do?

- A. Include the flag `-runner=DataflowRunner` in `beam_pipeline_args` to run the evaluation step on Dataflow.
- B. Move the evaluation step out of your pipeline and run it on custom Compute Engine VMs with sufficient memory.

C. Migrate your pipeline to Kubeflow hosted on Google Kubernetes Engine, and specify the appropriate node parameters for the evaluation step.

D. Add `tfma.MetricsSpec()` to limit the number of metrics in the evaluation step.

Answer: A ([LEAVE A REPLY](#))

<https://blog.tensorflow.org/2020/03/tensorflow-extended-tfx-using-apache-beam-large-scale-data-processing.html>

NEW QUESTION: 127

You recently developed a wide and deep model in TensorFlow. You generated training datasets using a SQL script that preprocessed raw data in BigQuery by performing instance-level transformations of the data. You need to create a training pipeline to retrain the model on a weekly basis. The trained model will be used to generate daily recommendations. You want to minimize model development and training time. How should you develop the training pipeline?

A. Use the Kubeflow Pipelines SDK to implement the pipeline. Use the `BigQueryJobOp` component to run the preprocessing script and the `CustomTrainingJobOp` component to launch a Vertex AI training job.

B. Use the TensorFlow Extended SDK to implement the pipeline. Use the `ExampleGen` component with the BigQuery executor to ingest the data the `Transform` component to preprocess the data, and the `Trainer` component to launch a Vertex AI training job.

C. Use the Kubeflow Pipelines SDK to implement the pipeline. Use the `DataflowPythonJobOp` component to preprocess the data and the `CustomTrainingJobOp` component to launch a Vertex AI training job.

D. Use the TensorFlow Extended SDK to implement the pipeline. Implement the preprocessing steps as part of the `input_fn` of the model. Use the `ExampleGen` component with the BigQuery executor to ingest the data and the `Trainer` component to launch a Vertex AI training job.

Answer: B ([SHOW ANSWER](#))

NEW QUESTION: 128

You recently deployed a model to a Vertex AI endpoint. Your data drifts frequently, so you have enabled request-response logging and created a Vertex AI Model Monitoring job. You have observed that your model is receiving higher traffic than expected. You need to reduce the model monitoring cost while continuing to quickly detect drift. What should you do?

A. Replace the monitoring job with a DataFlow pipeline that uses TensorFlow Data Validation (TFDV)

B. Increase the `monitor_interval` parameter in the `ScheduleConfig` of the monitoring job

C. Replace the monitoring job with a custom SQL script to calculate statistics on the features and predictions in BigQuery

D. Decrease the `sample_rate` parameter in the `RandomSampleConfig` of the monitoring job

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 129

You have created a Vertex AI pipeline that automates custom model training. You want to add a pipeline component that enables your team to most easily collaborate when running different executions and comparing metrics both visually and programmatically. What should you do?

A. Add a component to the Vertex AI pipeline that logs metrics to a BigQuery table. Load the table into a pandas DataFrame to compare different executions of the pipeline. Use Matplotlib to visualize metrics.

- B.** Add a component to the Vertex AI pipeline that logs metrics to Vertex ML Metadata. Use Vertex AI Experiments to compare different executions of the pipeline. Use Vertex AI TensorBoard to visualize metrics.
- C.** Add a component to the Vertex AI pipeline that logs metrics to Vertex ML Metadata. Load the Vertex ML Metadata into a pandas DataFrame to compare different executions of the pipeline. Use Matplotlib to visualize metrics.
- D.** Add a component to the Vertex AI pipeline that logs metrics to a BigQuery table. Query the table to compare different executions of the pipeline. Connect BigQuery to Looker Studio to visualize metrics.

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 130

Your team is training a large number of ML models that use different algorithms, parameters, and datasets. Some models are trained in Vertex AI Pipelines, and some are trained on Vertex AI Workbench notebook instances. Your team wants to compare the performance of the models across both services. You want to minimize the effort required to store the parameters and metrics. What should you do?

- A.** Implement all models in Vertex AI Pipelines Create a Vertex AI experiment, and associate all pipeline runs with that experiment.
- B.** Create a Vertex AI experiment. Submit all the pipelines as experiment runs. For models trained on notebooks log parameters and metrics by using the Vertex AI SDK.
- C.** Store all model parameters and metrics as model metadata by using the Vertex AI Metadata API.
- D.** Implement an additional step for all the models running in pipelines and notebooks to export parameters and metrics to BigQuery.

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 131

You are developing an ML model intended to classify whether X-ray images indicate bone fracture risk. You have trained a ResNet architecture on Vertex AI using a TPU as an accelerator, however you are unsatisfied with the training time and memory usage. You want to quickly iterate your training code but make minimal changes to the code. You also want to minimize impact on the model's accuracy. What should you do?

- A.** Reduce the number of layers in the model architecture.
- B.** Reduce the global batch size from 1024 to 256.
- C.** Reduce the dimensions of the images used in the model.
- D.** Configure your model to use bfloat16 instead of float32.

Answer: ([SHOW ANSWER](#)**)**

<https://cloud.google.com/tpu/docs/bfloat16>

NEW QUESTION: 132

You are working on a binary classification ML algorithm that detects whether an image of a classified scanned document contains a company's logo. In the dataset, 96% of examples don't have the logo, so the dataset is very skewed. Which metrics would give you the most confidence in your model?

- A.** F-score where recall is weighed more than precision
- B.** RMSE

C. F1 score

D. F-score where precision is weighed more than recall

Answer: A ([LEAVE A REPLY](#))

Even a model which always says that don't have the logo will have a good precision because is the most common. What we need is improve recall.

NEW QUESTION: 133

You recently developed a deep learning model. To test your new model, you trained it for a few epochs on a large dataset. You observe that the training and validation losses barely changed during the training run. You want to quickly debug your model. What should you do first?

A. Verify that your model can obtain a low loss on a small subset of the dataset

B. Add handcrafted features to inject your domain knowledge into the model

C. Use the Vertex AI hyperparameter tuning service to identify a better learning rate

D. Use hardware accelerators and train your model for more epochs

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 134

You are training an object detection machine learning model on a dataset that consists of three million X-ray images, each roughly 2 GB in size. You are using Vertex AI Training to run a custom training application on a Compute Engine instance with 32-cores, 128 GB of RAM, and 1 NVIDIA P100 GPU. You notice that model training is taking a very long time. You want to decrease training time without sacrificing model performance. What should you do?

A. Replace the NVIDIA P100 GPU with a v3-32 TPU in the training job.

B. Use the `tf.distribute.Strategy` API and run a distributed training job.

C. Enable early stopping in your Vertex AI Training job.

D. Increase the instance memory to 512 GB and increase the batch size.

Answer: ([SHOW ANSWER](#))

NEW QUESTION: 135

You are a SQL analyst. You need to utilize a TensorFlow customer segmentation model stored In Cloud Storage. You want to use the simplest and most efficient approach. What should you do?

A. Import the model into Vertex AI Model Registry. Deploy the model to a Vertex AI endpoint, and use SQL for inference in BigQuery.

B. Deploy the model by using TensorFlow Serving, and call for inference from BigQuery.

C. Convert the model into a BigQuery ML model, and use SQL for inference.

D. Import the model into BigQuery, and use SQL for inference.

Answer: A ([LEAVE A REPLY](#))

Importing the TensorFlow model into Vertex AI Model Registry and deploying it to a Vertex AI endpoint allows you to perform scalable, managed online predictions. BigQuery supports making calls to Vertex AI endpoints via SQL functions, enabling simple and efficient inference directly from SQL queries.

NEW QUESTION: 136

You recently joined an enterprise-scale company that has thousands of datasets. You know that there are accurate descriptions for each table in BigQuery, and you are searching for the proper BigQuery table to use for a model you are building on AI Platform. How should you find the data that you need?

- A.** Use Data Catalog to search the BigQuery datasets by using keywords in the table description.
- B.** Tag each of your model and version resources on AI Platform with the name of the BigQuery table that was used for training.
- C.** Maintain a lookup table in BigQuery that maps the table descriptions to the table ID. Query the lookup table to find the correct table ID for the data that you need.
- D.** Execute a query in BigQuery to retrieve all the existing table names in your project using the INFORMATION_SCHEMA metadata tables that are native to BigQuery. Use the result to find the table that you need.

Answer: ([SHOW ANSWER](#))

<https://cloud.google.com/data-catalog/docs/concepts/overview>

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF** Special Discount: **Freepdfdumps**)

NEW QUESTION: 137

You need to build an object detection model for a small startup company to identify if and where the company's logo appears in an image. You were given a large repository of images, some with logos and some without. These images are not yet labelled. You need to label these pictures, and then train and deploy the model. What should you do?

- A.** Use Google Cloud's Data Labelling Service to label your data. Use AutoML Object Detection to train and deploy the model.
- B.** Use Vision API to detect and identify logos in pictures and use it as a label. Use AI Platform to build and train a convolutional neural network.
- C.** Create two folders: one where the logo appears and one where it doesn't. Manually place images in each folder. Use AI Platform to build and train a convolutional neural network.
- D.** Create two folders: one where the logo appears and one where it doesn't. Manually place images in each folder. Use AI Platform to build and train a real time object detection model.

Answer: ([SHOW ANSWER](#))

A is correct as this will allow you to easily create a request for a labelling task and deploy a high- performance model.

B is not correct because Vision API is not guaranteed to work with any company logos, and in the statement it explicitly mentions a small startup, which will further decrease the chance of success.

C is not correct because the task of manually labelling the data is time consuming and should be avoided if possible.

D is not correct because the task of labelling object detection data is very tedious, and real time object detection is designed detecting objects in videos rather than in images.

<https://cloud.google.com/ai-platform/data-labeling/docs>

NEW QUESTION: 138

You work as an ML engineer at a social media company, and you are developing a visual filter for users' profile photos. This requires you to train an ML model to detect bounding boxes around human faces. You want to use this filter in your company's iOS-based mobile phone application.

You want to minimize code development and want the model to be optimized for inference on mobile phones. What should you do?

- A. Train a model using AutoML Vision and use the "export for Core ML" option.
- B. Train a model using AutoML Vision and use the "export for Coral" option.
- C. Train a model using AutoML Vision and use the "export for TensorFlow.js" option.
- D. Train a custom TensorFlow model and convert it to TensorFlow Lite (TFLite).

Answer: (SHOW ANSWER)

<https://cloud.google.com/vision/automl/docs/export-edge>

Core ML -> iOS and macOS

Coral -> Edge TPU-based device

TensorFlow.js -> web

NEW QUESTION: 139

You work as an ML researcher at an investment bank, and you are experimenting with the Gemma large language model (LLM). You plan to deploy the model for an internal use case. You need to have full control of the model's underlying infrastructure and minimize the model's inference time. Which serving configuration should you use for this task?

- A. Deploy the model on a Vertex AI endpoint manually by creating a custom inference container.
- B. Deploy the model on a Google Kubernetes Engine (GKE) cluster by using the deployment options in Model Garden.
- C. Deploy the model on a Vertex AI endpoint by using one-click deployment in Model Garden.
- D. Deploy the model on a Google Kubernetes Engine (GKE) cluster manually by creating a custom yaml manifest.

Answer: B (LEAVE A REPLY)

Deploying the Gemma LLM on a GKE cluster using Model Garden deployment options gives you full control over the underlying infrastructure while benefiting from optimized configurations. This setup minimizes inference time through fine-tuned resource allocation and allows customization of the serving environment, which is ideal for internal, performance-sensitive use cases.

NEW QUESTION: 140

You work for a retail company. You have been tasked with building a model to determine the probability of churn for each customer. You need the predictions to be interpretable so the results can be used to develop marketing campaigns that target at-risk customers. What should you do?

- A. Build an AutoML tabular regression model. Configure the model to generate explanations when it makes predictions.
- B. Build a custom TensorFlow neural network by using Vertex AI custom training. Configure the model to generate explanations when it makes predictions.
- C. Build a random forest regression model in a Vertex AI Workbench notebook instance. Configure the model to generate feature importances after the model is trained.

D. Build a random forest classification model in a Vertex AI Workbench notebook instance. Configure the model to generate feature importances after the model is trained.

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 141

Your team is building a convolutional neural network (CNN)-based architecture from scratch. The preliminary experiments running on your on-premises CPU-only infrastructure were encouraging, but have slow convergence. You have been asked to speed up model training to reduce time-to-market. You want to experiment with virtual machines (VMs) on Google Cloud to leverage more powerful hardware. Your code does not include any manual device placement and has not been wrapped in Estimator model-level abstraction. Which environment should you train your model on?

- A.** AVM on Compute Engine and 1 TPU with all dependencies installed manually.
- B.** AVM on Compute Engine and 8 GPUs with all dependencies installed manually.
- C.** A Deep Learning VM with an n1-standard-2 machine and 1 GPU with all libraries pre-installed.
- D.** A Deep Learning VM with more powerful CPU e2-highcpu-16 machines with all libraries pre-installed.

Answer: C ([LEAVE A REPLY](#))

To support CNN, you should use GPU.

For preliminary experiment, pre-installed pkgs/libs are good choice.

https://cloud.google.com/deep-learning-vm/docs/cli#creating_an_instance_with_one_or_more_gpus

https://cloud.google.com/deep-learning-vm/docs/introduction#pre-installed_packages

NEW QUESTION: 142

You recently developed a deep learning model using Keras, and now you are experimenting with different training strategies. First, you trained the model using a single GPU, but the training process was too slow. Next, you distributed the training across 4 GPUs using `tf.distribute.MirroredStrategy` (with no other changes), but you did not observe a decrease in training time. What should you do?

- A.** Distribute the dataset with `tf.distribute.Strategy.experimental_distribute_dataset`
- B.** Create a custom training loop.
- C.** Use a TPU with `tf.distribute.TPUStrategy`.
- D.** Increase the batch size.

Answer: D ([LEAVE A REPLY](#))

https://www.tensorflow.org/guide/gpu_performance_analysis

NEW QUESTION: 143

You lead a data science team at a large international corporation. Most of the models your team trains are large-scale models using high-level TensorFlow APIs on AI Platform with GPUs. Your team usually takes a few weeks or months to iterate on a new version of a model. You were recently asked to review your team's spending. How should you reduce your Google Cloud compute costs without impacting the model's performance?

- A.** Use AI Platform to run distributed training jobs with checkpoints.
- B.** Use AI Platform to run distributed training jobs without checkpoints.
- C.** Migrate to training with Kuberflow on Google Kubernetes Engine, and use preemptible VMs with checkpoints.
- D.** Migrate to training with Kuberflow on Google Kubernetes Engine, and use preemptible VMs without checkpoints.

Answer: C (LEAVE A REPLY)

<https://cloud.google.com/blog/products/ai-machine-learning/reduce-the-costs-of-ml-workflows-with-preemptible-vms-and-gpus?hl=en>

NEW QUESTION: 144

You are deploying a new version of a model to a production Vertex AI endpoint that is serving traffic. You plan to direct all user traffic to the new model. You need to deploy the model with minimal disruption to your application. What should you do?

- A.** 1. Create a new model. Set the parentModel parameter to the model ID of the currently deployed model. Upload the model to Vertex AI Model Registry.
2. Deploy the new model to the existing endpoint, and set the new model to 100% of the traffic
- B.** 1. Create a new model. Set it as the default version. Upload the model to Vertex AI Model Registry
2. Deploy the new model to the existing endpoint
- C.** 1. Create a new endpoint
2. Create a new model. Set it as the default version. Upload the model to Vertex AI Model Registry
3. Deploy the new model to the new endpoint
4. Update Cloud DNS to point to the new endpoint
- D.** 1. Create a new endpoint
2. Create a new model. Set the parentModel parameter to the model ID of the currently deployed model and set it as the default version. Upload the model to Vertex AI Model Registry
3. Deploy the new model to the new endpoint, and set the new model to 100% of the traffic.

Answer: (SHOW ANSWER)

NEW QUESTION: 145

You work for a magazine distributor and need to build a model that predicts which customers will renew their subscriptions for the upcoming year. Using your company's historical data as your training set, you created a TensorFlow model and deployed it to AI Platform. You need to determine which customer attribute has the most predictive power for each prediction served by the model. What should you do?

- A.** Use AI Platform notebooks to perform a Lasso regression analysis on your model, which will eliminate features that do not provide a strong signal.
- B.** Stream prediction results to BigQuery. Use BigQuery's CORR(X1, X2) function to calculate the Pearson correlation coefficient between each feature and the target variable.
- C.** Use the AI Explanations feature on AI Platform. Submit each prediction request with the 'explain' keyword to retrieve feature attributions using the sampled Shapley method.
- D.** Use the What-If tool in Google Cloud to determine how your model will perform when individual features are excluded. Rank the feature importance in order of those that caused the most significant performance drop when removed from the model.

Answer: C (LEAVE A REPLY)

<https://cloud.google.com/ai-platform/prediction/docs/ai-explanations/overview> AI Explanations helps you understand your model's outputs for classification and regression tasks. Whenever you request a prediction on AI Platform, AI Explanations tells you how much each feature in the data contributed to the predicted result.

NEW QUESTION: 146

You are an ML engineer at a media company. You need to build an ML model to analyze video content, identify objects, and alert users if there is inappropriate content. Which Google Cloud products should you use to build this project?

- A. Pub/Sub, Cloud Function, Cloud Vision API
- B. Pub/Sub, Cloud IoT, Dataflow, Cloud Vision API, Cloud Logging
- C. Pub/Sub, Cloud Function, Video Intelligence API, Cloud Logging
- D. Pub/Sub, Cloud Function, AutoML Video Intelligence, Cloud Logging

Answer: (SHOW ANSWER)

A is not correct as there is no tool for alerting and notifying.

B is not correct as it uses Cloud Vision API for processing videos.

C is correct as Video Intelligence API can find inappropriate components and other components satisfy the requirements of real-time processing and notification. (<https://cloud.google.com/video-intelligence>) D is not correct as AutoML Video intelligence should be only used in case of customization.

<https://cloud.google.com/solutions/processing-user-generated-content-using-video-intelligence>

NEW QUESTION: 147

You have recently created a proof-of-concept (POC) deep learning model. You are satisfied with the overall architecture, but you need to determine the value for a couple of hyperparameters.

You want to perform hyperparameter tuning on Vertex AI to determine both the appropriate embedding dimension for a categorical feature used by your model and the optimal learning rate.

You configure the following settings:

- For the embedding dimension, you set the type to INTEGER with a minValue of 16 and maxValue of 64.
- For the learning rate, you set the type to DOUBLE with a minValue of 10e-05 and maxValue of 10e-02.

You are using the default Bayesian optimization tuning algorithm, and you want to maximize model accuracy. Training time is not a concern. How should you set the hyperparameter scaling for each hyperparameter and the maxParallelTrials?

- A. Use UNIT_LOG_SCALE for the embedding dimension, UNIT_LINEAR_SCALE for the learning rate, and a small number of parallel trials.
- B. Use UNIT_LOG_SCALE for the embedding dimension, UNIT_LINEAR_SCALE for the learning rate, and a large number of parallel trials.
- C. Use UNIT_LINEAR_SCALE for the embedding dimension, UNIT_LOG_SCALE for the learning rate, and a large number of parallel trials.
- D. Use UNIT_LINEAR_SCALE for the embedding dimension, UNIT_LOG_SCALE for the learning rate, and a small number of parallel trials.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 148

You are developing a recommendation engine for an online clothing store. The historical customer transaction data is stored in BigQuery and Cloud Storage. You need to perform exploratory data analysis (EDA), preprocessing and model training. You plan to rerun these EDA, preprocessing, and training steps as you experiment with different types of algorithms. You want to minimize the cost and development effort of running these steps as you experiment. How should you configure the environment?

- A.** Create a Vertex AI Workbench user-managed notebook on a Dataproc Hub, and use the `%%bigquery` magic commands in Jupyter to query the tables.
- B.** Create a Vertex AI Workbench user-managed notebook using the default VM instance, and use the `%%bigquery` magic commands in Jupyter to query the tables.
- C.** Create a Vertex AI Workbench managed notebook to browse and query the tables directly from the JupyterLab interface.
- D.** Create a Vertex AI Workbench managed notebook on a Dataproc cluster, and use the `spark-bigquery-connector` to access the tables.

Answer: C ([LEAVE A REPLY](#))

NEW QUESTION: 149

You were asked to investigate failures of a production line component based on sensor readings.

After receiving the dataset, you discover that less than 1% of the readings are positive examples representing failure incidents.

You have tried to train several classification models, but none of them converge. How should you resolve the class imbalance problem?

- A.** Use the class distribution to generate 10% positive examples.
- B.** Use a convolutional neural network with max pooling and softmax activation.
- C.** Downsample the data with upweighting to create a sample with 10% positive examples.
- D.** Remove negative examples until the numbers of positive and negative examples are equal.

Answer: C ([LEAVE A REPLY](#))

- less than 1% of the readings are positive
- none of them converge.

Downsampling (in this context) means training on a disproportionately low subset of the majority class examples.

<https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data#downsampling-and-upweighting>

NEW QUESTION: 150

You work for an international bank. Your team is developing a generative AI application that summarizes complex financial reports. You need to implement an automated, scalable solution to evaluate and quantify the quality of generated summaries against the previous version of the solution. What should you do?

- A.** Use Model Monitoring to track the prediction drift of the token output length.
- B.** Run an automatic side-by-side (AutoSxS) evaluation job in Agent Platform to compare outputs from the new and previous versions of the solution.
- C.** Deploy both versions of the solution, and use A/B testing to route a portion of production traffic to the new version and route the remaining traffic to the previous version. Compare user click-through rates.

D. Run an evaluation job in Agent Platform, and calculate a BLEU score to compare the new and previous versions of the solution.

Answer: B ([LEAVE A REPLY](#))

Automatic side-by-side evaluation is designed to compare the generated outputs of two model or application versions at scale. It uses an automated evaluator to assess relative response quality across the same prompt set, providing quantitative comparison results without requiring manual review or production traffic experiments.

NEW QUESTION: 151

You need to build classification workflows over several structured datasets currently stored in BigQuery. Because you will be performing the classification several times, you want to complete the following steps without writing code: exploratory data analysis, feature selection, model building, training, and hyperparameter tuning and serving. What should you do?

- A.** Run a logistic regression job on BigQuery ML.
- B.** Train a TensorFlow model on Vertex AI.
- C.** Use scikit-learn in Notebooks with pandas library.
- D.** Train a classification Vertex AutoML model.

Answer: D ([LEAVE A REPLY](#))

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF Special Discount: Freepdfdumps**)

NEW QUESTION: 152

Your data science team needs to rapidly experiment with various features, model architectures, and hyperparameters. They need to track the accuracy metrics for various experiments and use an API to query the metrics over time. What should they use to track and report their experiments while minimizing manual effort?

- A.** Use Kubeflow Pipelines to execute the experiments.

Export the metrics file, and query the results using the Kubeflow Pipelines API.

- B.** Use AI Platform Training to execute the experiments.

Write the accuracy metrics to BigQuery, and query the results using the BigQuery API.

- C.** Use AI Platform Training to execute the experiments.

Write the accuracy metrics to Cloud Monitoring, and query the results using the Monitoring API.

- D.** Use AI Platform Notebooks to execute the experiments.

Collect the results in a shared Google Sheets file, and query the results using the Google Sheets API.

Answer: A ([LEAVE A REPLY](#))

<https://codelabs.developers.google.com/codelabs/cloud-kubeflow-pipelines-gis> Kubeflow Pipelines (KFP) helps solve these issues by providing a way to deploy robust, repeatable machine learning pipelines along with monitoring, auditing, version tracking, and reproducibility. Cloud AI Pipelines makes it easy to set up a KFP installation.

NEW QUESTION: 153

You are a data analyst on a marketing team. You have 40 GB of customer transaction data already stored in BigQuery, and you need to produce a churn-propensity model quickly to validate whether the project is worth further investment. Your team has strong SQL skills but limited Python experience. You want to minimize infrastructure setup. What should you do?

- A. Export the data to Cloud Storage, and train a scikit-learn model in a Workbench instance using a custom training job.
- B. Create a BigQuery ML logistic regression model directly on the transaction table, and evaluate it with ML.EVALUATE.
- C. Build an AutoML tabular training job in Agent Platform by importing the BigQuery table as a managed dataset.
- D. Deploy a Dataproc cluster, and use Spark MLlib to train a gradient-boosted tree classifier.

Answer: ([SHOW ANSWER](#))

BigQuery ML lets the team train and evaluate a model with SQL where the data already resides, avoiding data movement and infrastructure provisioning. For a fast feasibility check with SQL- oriented staff, it delivers a baseline model with the least effort, and the model can later be exported or promoted to a managed service if the project proceeds.

NEW QUESTION: 154

You work for a magazine distributor and need to build a model that predicts which customers will renew their subscriptions for the upcoming year. Using your company's historical data as your training set, you created a TensorFlow model and deployed it to Vertex AI. You need to determine which customer attribute has the most predictive power for each prediction served by the model. What should you do?

- A. Use the What-If tool in Google Cloud to determine how your model will perform when individual features are excluded. Rank the feature importance in order of those that caused the most significant performance drop when removed from the model.
- B. Stream prediction results to BigQuery. Use BigQuery's CORR(X1, X2) function to calculate the Pearson correlation coefficient between each feature and the target variable.
- C. Use Vertex AI Workbench user-managed notebooks to perform a Lasso regression analysis on your model, which will eliminate features that do not provide a strong signal.
- D. Use Vertex Explainable AI. Submit each prediction request with the explain' keyword to retrieve feature attributions using the sampled Shapley method.

Answer: D ([LEAVE A REPLY](#))

NEW QUESTION: 155

You work for a ride-sharing company. Your team trains a demand-forecasting model in a batch pipeline using aggregated features such as 30-minute rolling trip counts per zone. The same features must be served to an online endpoint with single-digit millisecond latency. During production testing, you notice the online predictions differ significantly from offline evaluation results. You need to eliminate the discrepancy while minimizing the amount of code you maintain.

What should you do?

- A.** Recompute the aggregated features inside the model serving container at request time by querying BigQuery directly.
- B.** Store the features in Feature Store, and use the offline store for training data retrieval with point-in-time correctness and the online store for low-latency serving.
- C.** Export the training feature table to Cloud Storage as CSV files, and load them into a Memorystore instance that the endpoint queries at prediction time.
- D.** Increase the frequency of the batch pipeline to run every five minutes, and write the results to a Cloud SQL table read by the endpoint.

Answer: B (LEAVE A REPLY)

Feature Store maintains a single feature definition that is materialized to both an offline store for training and an online store for serving, which removes training-serving skew. Point-in-time lookups also prevent label leakage when building training sets, and the managed service eliminates the custom synchronization code required by the other options.

NEW QUESTION: 156

You work at a bank. You have a custom tabular ML model that was provided by the bank's vendor. The training data is not available due to its sensitivity. The model is packaged as a Vertex AI Model serving container, which accepts a string as input for each prediction instance. In each string, the feature values are separated by commas. You want to deploy this model to production for online predictions and monitor the feature distribution over time with minimal effort. What should you do?

- A.** 1. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
2. Create a Vertex AI Model Monitoring job with feature drift detection as the monitoring objective, and provide an instance schema
- B.** 1. Refactor the serving container to accept key-value pairs as input format
2. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
3. Create a Vertex AI Model Monitoring job with feature skew detection as the monitoring objective
- C.** 1. Refactor the serving container to accept key-value pairs as input format
2. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
3. Create a Vertex AI Model Monitoring job with feature drift detection as the monitoring objective.
- D.** 1. Upload the model to Vertex AI Model Registry, and deploy the model to a Vertex AI endpoint
2. Create a Vertex AI Model Monitoring job with feature skew detection as the monitoring objective, and provide an instance schema

Answer: A (LEAVE A REPLY)

NEW QUESTION: 157

You are developing models to classify customer support emails. You created models with TensorFlow Estimators using small datasets on your on-premises system, but you now need to train the models using large datasets to ensure high performance. You will port your models to Google Cloud and want to minimize code refactoring and infrastructure overhead for easier migration from on-prem to cloud. What should you do?

- A.** Use AI Platform for distributed training.
- B.** Create a cluster on Dataproc for training.
- C.** Create a Managed Instance Group with autoscaling.
- D.** Use Kubeflow Pipelines to train on a Google Kubernetes Engine cluster.

Answer: A (LEAVE A REPLY)

AI platform provides lower infrastructure overhead and allows you to not have to refactor your code too much (no containerization and such, like in KubeFlow).

NEW QUESTION: 158

You work for a global footwear retailer and need to predict when an item will be out of stock based on historical inventory data. Customer behavior is highly dynamic since footwear demand is influenced by many different factors. You want to serve models that are trained on all available data, but track your performance on specific subsets of data before pushing to production. What is the most streamlined and reliable way to perform this validation?

- A.** Use the TFX ModelValidator tools to specify performance metrics for production readiness.
- B.** Use k-fold cross-validation as a validation strategy to ensure that your model is ready for production.
- C.** Use the last relevant week of data as a validation set to ensure that your model is performing accurately on current data.
- D.** Use the entire dataset and treat the area under the receiver operating characteristics curve (AUC ROC) as the main metric.

Answer: C (LEAVE A REPLY)

<https://www.tensorflow.org/tfx/guide/evaluator>

NEW QUESTION: 159

You work for a healthcare organization and need to automate the patient triage process. You want to build a telephony agent solution using the Gemini Live API to perform human-like conversations with patients and gather necessary patient information. You plan to pilot the solution in Japanese. You need to minimize latency and development time, and ensure high, native-like fluency in the model's responses. What should you do?

- A.** Use Gemini Live API Speech-to-Text functionality to capture the user's input followed by Text-to-Speech streaming processing with Chirp to communicate with the patient in their language.
- B.** Use the Gemini Live API, and configure the language_code field to ja-JP. Include a clear system instruction to respond in the patient's language.
- C.** Fine-tune the Gemini Flash model on a dataset of Japanese medical transcripts to ensure native fluency. Deploy this custom model behind the Gemini Live API.
- D.** Use the Gemini Pro model to automatically select the appropriate language.

Answer: (SHOW ANSWER)

The Gemini Live API provides native, real-time audio interaction without a separate speech-to-text and text-to-speech pipeline, minimizing conversational latency and development effort.

Configuring Japanese as the language and adding a clear system instruction guides the model to maintain fluent, natural Japanese throughout the patient interaction.

Reference:

<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/live-api>

<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/live-api/configure-language-voice>

<https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/gemini/2-5-flash-live-api>

NEW QUESTION: 160

Your team is using a TensorFlow Inception-v3 CNN model pretrained on ImageNet for an image classification prediction challenge on 10,000 images. You will use AI Platform to perform the model training. What TensorFlow distribution strategy and AI Platform training job configuration should you use to train the model and optimize for wall-clock time?

- A. Default Strategy; Custom tier with a single master node and four v100 GPUs.
- B. One Device Strategy; Custom tier with a single master node and four v100 GPUs.
- C. One Device Strategy; Custom tier with a single master node and eight v100 GPUs.
- D. MirroredStrategy; Custom tier with a single master node and four v100 GPUs.

Answer: D (LEAVE A REPLY)

A is not correct because Default Strategy does not distribute training across multiple devices.

B is not correct because One Device Strategy does not distribute training across multiple devices.

C is not correct because One Device Strategy does not distribute training across multiple devices.

D is correct because this is the only strategy that can perform distributed training; albeit there is only a single copy of the variables on the CPU host.

https://www.tensorflow.org/guide/distributed_training

NEW QUESTION: 161

You are training a Resnet model on AI Platform using TPUs to visually categorize types of defects in automobile engines. You capture the training profile using the Cloud TPU profiler plugin and observe that it is highly input-bound. You want to reduce the bottleneck and speed up your model training process. Which modifications should you make to the tf.data dataset? (Choose two.)

- A. Increase the buffer size for the shuttle option.
- B. Use the interleave option for reading data.
- C. Decrease the batch size argument in your transformation.
- D. Reduce the value of the repeat parameter.
- E. Set the prefetch option equal to the training batch size.

Answer: B,C (LEAVE A REPLY)

NEW QUESTION: 162

You work for a company that manages a ticketing platform for a large chain of cinemas.

Customers use a mobile app to search for movies they're interested in and purchase tickets in the app. Ticket purchase requests are sent to Pub/Sub and are processed with a Dataflow streaming pipeline configured to conduct the following steps:

1. Check for availability of the movie tickets at the selected cinema.
2. Assign the ticket price and accept payment.
3. Reserve the tickets at the selected cinema.
4. Send successful purchases to your database.

Each step in this process has low latency requirements (less than 50 milliseconds). You have developed a logistic regression model with BigQuery ML that predicts whether offering a promo code for free popcorn increases the chance of a ticket purchase, and this prediction should be added to the ticket purchase process. You want to identify the simplest way to deploy this model to production while adding minimal latency. What should you do?

- A. Run batch inference with BigQuery ML every five minutes on each new set of tickets issued.

- B.** Export your model in TensorFlow format, and add a `tfx_bsl.public.beam.RunInference` step to the Dataflow pipeline.
- C.** Export your model in TensorFlow format, deploy it on Vertex AI, and query the prediction endpoint from your streaming pipeline.
- D.** Convert your model with TensorFlow Lite (TFLite), and add it to the mobile app so that the promo code and the incoming request arrive together in Pub/Sub.

Answer: ([SHOW ANSWER](#))

<https://www.tensorflow.org/lite/guide>

NEW QUESTION: 163

You are building a real-time prediction engine that streams files which may contain Personally Identifiable Information (PII) to Google Cloud. You want to use the Cloud Data Loss Prevention (DLP) API to scan the files. How should you ensure that the PII is not accessible by unauthorized individuals?

A. Stream all files to Google Cloud, and then write the data to BigQuery.

Periodically conduct a bulk scan of the table using the DLP API.

B. Stream all files to Google Cloud, and write batches of the data to BigQuery.

While the data is being written to BigQuery, conduct a bulk scan of the data using the DLP API.

C. Create two buckets of data: Sensitive and Non-sensitive.

Write all data to the Non-sensitive bucket.

Periodically conduct a bulk scan of that bucket using the DLP API, and move the sensitive data to the Sensitive bucket.

D. Create three buckets of data: Quarantine, Sensitive, and Non-sensitive.

Write all data to the Quarantine bucket.

Periodically conduct a bulk scan of that bucket using the DLP API, and move the data to either the Sensitive or Non-Sensitive bucket.

Answer: **D** ([LEAVE A REPLY](#))

https://cloud.google.com/architecture/automating-classification-of-data-uploaded-to-cloud-storage#building_the_quarantine_and_classification_pipeline

NEW QUESTION: 164

You are designing an ML recommendation model for shoppers on your company's ecommerce website. You will use Recommendations AI to build, test, and deploy your system. How should you develop recommendations that increase revenue while following best practices?

A. Use the "Other Products You May Like" recommendation type to increase the click-through rate.

B. Use the "Frequently Bought Together" recommendation type to increase the shopping cart size for each order.

C. Import your user events and then your product catalog to make sure you have the highest quality event stream.

D. Because it will take time to collect and record product data, use placeholder values for the product catalog to test the viability of the model.

Answer: **B** ([LEAVE A REPLY](#))

<https://cloud.google.com/recommendations-ai/docs/placements#rps>

NEW QUESTION: 165

You are developing a model to detect fraudulent credit card transactions. You need to prioritize detection, because missing even one fraudulent transaction could severely impact the credit card holder. You used AutoML to train a model on users' profile information and credit card transaction data. After training the initial model, you notice that the model is failing to detect many fraudulent transactions. How should you increase the number of fraudulent transactions that are detected?

- A. Add more non-fraudulent examples to the training set.
- B. Reduce the maximum number of node hours for training.
- C. Increase the probability threshold to classify a fraudulent transaction.
- D. Decrease the probability threshold to classify a fraudulent transaction.

Answer: D (LEAVE A REPLY)

Lowering the probability threshold makes the model more sensitive to identifying positive cases (fraudulent transactions), increasing the true positive rate. This approach helps prioritize detection, even if it may lead to more false positives - which is acceptable in high-risk use cases like fraud detection.

NEW QUESTION: 166

Your company manages an application that aggregates news articles from many different online sources and sends them to users. You need to build a recommendation model that will suggest articles to readers that are similar to the articles they are currently reading. Which approach should you use?

- A. Create a collaborative filtering system that recommends articles to a user based on the user's past behavior.
- B. Encode all articles into vectors using word2vec, and build a model that returns articles based on vector similarity.
- C. Build a logistic regression model for each user that predicts whether an article should be recommended to a user.
- D. Manually label a few hundred articles, and then train an SVM classifier based on the manually classified articles that categorizes additional articles into their respective categories.

Answer: (SHOW ANSWER)

<https://cloud.google.com/blog/topics/developers-practitioners/meet-ais-multitool-vector-embeddings>

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF Special Discount: Freepdfdumps**)

NEW QUESTION: 167

You are training an ML model on a large dataset. You are using a TPU to accelerate the training process. You notice that the training process is taking longer than expected. You discover that the TPU is not reaching its full capacity. What should you do?

- A. Increase the number of epochs
- B. Increase the batch size

- C. Increase the learning rate
- D. Decrease the learning rate

Answer: B ([LEAVE A REPLY](#))

NEW QUESTION: 168

You are an ML engineer at a bank. You need to build a solution that provides transparent and understandable explanations for AI-driven decisions for loan approvals, credit limits, and interest rates. You want to build this system to require minimal operational overhead. What should you do?

- A. Deploy the Learning Interpretability Tool (LIT) on App Engine to provide explainability and visualization of the output.
- B. Use Vertex Explainable AI to generate feature attributions, and use feature-based explanations for your models.
- C. Use AutoML Tables with built-in explainability features, and use Shapley values for explainability.
- D. Deploy pre-trained models from TensorFlow Hub to provide explainability using visualization tools.

Answer: B ([LEAVE A REPLY](#))

Vertex Explainable AI provides robust, built-in tools for generating feature attributions and explanations for machine learning models deployed on Vertex AI. It allows you to understand which features contributed most to the AI's decision-making, making it suitable for use cases requiring transparency, such as loan approvals, credit limits, and interest rates. This approach integrates seamlessly into the Vertex AI ecosystem, minimizing operational overhead while delivering feature-based explanations that meet regulatory and business requirements for interpretability.

NEW QUESTION: 169

You recently trained an XGBoost model on tabular data. You plan to expose the model for internal use as an HTTP microservice. After deployment, you expect a small number of incoming requests. You want to productionize the model with the least amount of effort and latency. What should you do?

- A. Deploy the model to BigQuery ML by using CREATE MODEL with the BOOSTED_TREE_REGRESSOR statement, and invoke the BigQuery API from the microservice.
- B. Build a Flask-based app. Package the app in a custom container on Vertex AI, and deploy it to Vertex AI Endpoints.
- C. Build a Flask-based app. Package the app in a Docker image, and deploy it to Google Kubernetes Engine in Autopilot mode.
- D. Use a prebuilt XGBoost Vertex container to create a model, and deploy it to Vertex AI Endpoints.

Answer: ([SHOW ANSWER](#)**)**

NEW QUESTION: 170

You lead an ML team at a financial technology company. Your team is building a system to detect potentially fraudulent credit card transactions in real-time. The model must return a prediction with very low latency (under 200ms) to approve or deny the transaction. The system must also scale to handle millions of transactions per day. Your dataset consists of structured, tabular transaction data stored in BigQuery. Your team is proficient in data science but wants to avoid the complexity of managing and optimizing a real-time model serving infrastructure. What should you do?

- A. In an Agent Platform Workbench notebook, build a deep neural network (DNN) using PyTorch. Provision a cluster of GPU-enabled Compute Engine instances, and create an API to serve the model for real-time predictions.

- B.** Use BigQuery ML to train a boosted tree classification model. Schedule a batch prediction job using the ML.PREDICT function that runs every minute and scores recent transactions.
- C.** Use Forecasting with Agent Platform AutoML to train a time-series model on the sequence of transactions. Deploy the model as a tabular workflow to predict which future time windows are likely to contain fraudulent activity.
- D.** Use Agent Platform AutoML to train using your tabular dataset from BigQuery. Deploy the trained model to an Agent Platform endpoint, and integrate the endpoint's API into your transaction processing application.

Answer: D ([LEAVE A REPLY](#))

Agent Platform AutoML can train an accurate classification model directly from structured BigQuery data while minimizing custom model-development effort. Deploying the trained model to a managed Agent Platform endpoint provides low-latency online predictions, autoscaling, and managed serving infrastructure suitable for millions of real-time transaction requests.

NEW QUESTION: 171

You work for a bank and are building a random forest model for fraud detection. You have a dataset that includes transactions, of which 1% are identified as fraudulent. Which data transformation strategy would likely improve the performance of your classifier?

- A.** Write your data in TFRecords.
- B.** Z-normalize all the numeric features.
- C.** Oversample the fraudulent transaction 10 times.
- D.** Use one-hot encoding on all categorical features.

Answer: C ([LEAVE A REPLY](#))

<https://swarit.medium.com/detecting-fraudulent-consumer-transactions-through-machine-learning-25b1f2cabbb4>

NEW QUESTION: 172

You are responsible for managing and monitoring a Vertex AI model that is deployed in production. You want to automatically retrain the model when its performance deteriorates. What should you do?

- A.** Create a Vertex AI Model Monitoring job to track the model's performance with production data, and trigger retraining when specific metrics drop below predefined thresholds.
- B.** Collect feedback from end users, and retrain the model based on their assessment of its performance.
- C.** Configure a scheduled job to evaluate the model's performance on a static dataset, and retrain the model if the performance drops below predefined thresholds.
- D.** Use Vertex Explainable AI to analyze feature attributions and identify potential biases in the model. Retrain when significant shifts in feature importance or biases are detected.

Answer: (SHOW ANSWER)

Vertex AI Model Monitoring automatically tracks model performance on production data and can be configured to trigger retraining workflows when key metrics fall below defined thresholds, enabling proactive, automated model maintenance.

NEW QUESTION: 173

You are implementing a batch inference ML pipeline in Google Cloud. The model was developed using TensorFlow and is stored in SavedModel format in Cloud Storage. You need to apply the model to a historical dataset containing 10 TB of data that is stored in a BigQuery table. How should you perform the inference?

- A.** Import the TensorFlow model by using the CREATE MODEL statement in BigQuery ML. Apply the historical data to the TensorFlow model
 - B.** Export the historical data to Cloud Storage in CSV format. Configure a Vertex AI batch prediction job to generate predictions for the exported data
 - C.** Configure a Vertex AI batch prediction job to apply the model to the historical data in BigQuery
 - D.** Export the historical data to Cloud Storage in Avro format. Configure a Vertex AI batch prediction job to generate predictions for the exported data
- Answer: C** ([LEAVE A REPLY](#))

NEW QUESTION: 174

You work with a data engineering team that has developed a pipeline to clean your dataset and save it in a Cloud Storage bucket. You have created an ML model and want to use the data to refresh your model as soon as new data is available. As part of your CI/CD workflow, you want to automatically run a Kubeflow Pipelines training job on Google Kubernetes Engine (GKE). How should you architect this workflow?

- A.** Configure your pipeline with Dataflow, which saves the files in Cloud Storage.
After the file is saved, start the training job on a GKE cluster.
- B.** Use App Engine to create a lightweight python client that continuously polls Cloud Storage for new files.
As soon as a file arrives, initiate the training job.
- C.** Configure a Cloud Storage trigger to send a message to a Pub/Sub topic when a new file is available in a storage bucket.
Use a Pub/Sub-triggered Cloud Function to start the training job on a GKE cluster.
- D.** Use Cloud Scheduler to schedule jobs at a regular interval.
For the first step of the job, check the timestamp of objects in your Cloud Storage bucket.
If there are no new files since the last run, abort the job.

Answer: C ([LEAVE A REPLY](#))

<https://cloud.google.com/architecture/architecture-for-mlops-using-tfx-kubeflow-pipelines-and-cloud-build#triggering-and-scheduling-kubeflow-pipelines>

NEW QUESTION: 175

You trained a text classification model. You have the following SignatureDefs:

```
signature_def['serving_default']:
  The given SavedModel SignatureDef contains the following input(s):
    inputs['text'] tensor_info:
      dtype: DT_STRING
      shape: (-1, 2)
      name: serving_default_text:0
  The given SavedModel SignatureDef contains the following output(s):
    outputs ['Softmax'] tensor_info:
      dtype: DT_FLOAT
      shape: (-1, 2)
      name: StatefulPartitionedCall:0
  Method name is: tensorflow/serving/predict
```

You started a TensorFlow-serving component server and tried to send an HTTP request to get a prediction using:

```
headers = {"content-type": "application/json"}
json_response = requests.post('http:
//localhost:8501/v1/models/text_model:predict', data=data,
headers=headers)
```

What is the correct way to write the predict request?

- A. data = json.dumps({"signature_name": "seving_default", "instances" [['ab', 'bc', 'cd']]})
- B. data = json.dumps({"signature_name": "serving_default", "instances" [['a', 'b', 'c', 'd', 'e', 'f']]})
- C. data = json.dumps({"signature_name": "serving_default", "instances" [['a', 'b', 'c'], ['d', 'e', 'f']]})
- D. data = json.dumps({"signature_name": "serving_default", "instances" [['a', 'b'], ['c', 'd'], ['e', 'f']]})

Answer: D (LEAVE A REPLY)

A negative number in the shape enables auto expand

(<https://stackoverflow.com/questions/37956197/what-is-the-negative-index-in-shape-arrays-used-for-tensorflow>).

Then the first number -1 out of the shape (-1, 2) speaks the number of 1 dimensional arrays within the tensor (and it can autoexpand) while the second number (2) sets the number of elements in the inner array at 2.

NEW QUESTION: 176

You are training models in Vertex AI by using data that spans across multiple Google Cloud projects. You need to find, track, and compare the performance of the different versions of your models. Which Google Cloud services should you include in your ML workflow?

- A. Vertex AI Pipelines, Vertex AI Experiments, and Vertex AI Metadata
- B. Dataplex, Vertex AI Experiments, and Vertex AI ML Metadata
- C. Dataplex, Vertex AI Feature Store, and Vertex AI TensorBoard
- D. Vertex AI Pipelines, Vertex AI Feature Store, and Vertex AI Experiments

Answer: (SHOW ANSWER)

NEW QUESTION: 177

You work for a media company. You are developing a solution that requires transcription and diarizing (speaker identification) of live audio streams in near real-time. You are using an optimized WhisperX model built on the PyTorch framework that uses several libraries and weights sourced from the Hugging Face Hub. The deployment must include proprietary Python post-processing logic including speaker diarization and alignment correction that executes immediately after model inference. You need your solution to be efficient, scalable, handle high throughput, and minimize serving latency. How should you deploy this model and its required logic?

- A.** Deploy the WhisperX model artifacts to an Agent Platform endpoint using a pre-built PyTorch serving container configured with a GPU. Host the custom post-processing scripts in a Cloud Run function that receives the endpoint's prediction output and performs the final diarization.
- B.** Use a self-managed Google Kubernetes Engine (GKE) cluster for deployment. Package the model and logic into a container, and configure the cluster's horizontal pod autoscaler (HPA) to manage throughput.
- C.** Package the WhisperX model artifacts, the specific Python post-processing scripts, and the necessary dependencies (including Hugging Face libraries) into a single, custom serving container image. Deploy this custom image to a GPU-enabled Agent Platform endpoint.
- D.** Locate the WhisperX model in the Model Garden and use the "1-click deploy" feature to deploy it to an Agent Platform endpoint. Route the output to a Dataflow pipeline that uses a custom DoFn function to handle the post-processing and logging.

Answer: ([SHOW ANSWER](#))

A custom serving container can bundle the optimized WhisperX model, Hugging Face dependencies, model weights, and proprietary diarization and alignment logic into one inference service. Deploying that container to a GPU-enabled Agent Platform endpoint provides managed autoscaling and high-throughput serving while keeping post-processing in the same request path, which minimizes data transfer and end-to-end latency.

NEW QUESTION: 178

You are developing an AI text generator that will be able to dynamically adapt its generated responses to mirror the writing style of the user and mimic famous authors if their style is detected. You have a large dataset of various authors' works, and you plan to host the model on a custom VM. You want to use the most effective model. What should you do?

- A.** Deploy Llama 3 from Model Garden, and use prompt engineering techniques.
- B.** Fine-tune a BERT-based model from TensorFlow Hub.
- C.** Fine-tune Llama 3 from Model Garden on Vertex AI Pipelines.
- D.** Use the Gemini 1.5 Flash foundational model to build the text generator.

Answer: ([SHOW ANSWER](#))

Gemini 1.5 Flash is a state-of-the-art foundational model optimized for generative AI tasks like dynamic text generation and style adaptation. Using such a powerful foundational model provides better effectiveness than fine-tuning older architectures like BERT or Llama 3, and it integrates well with Google Cloud's Vertex AI ecosystem for scalability and performance.

NEW QUESTION: 179

You are training a custom language model for your company using a large dataset. You plan to use the Reduction Server strategy on Vertex AI. You need to configure the worker pools of the distributed training job. What should you do?

- A.** Configure the machines of the first two pools to have TPUs, and to use a container image where your training code runs. Configure the third pool to have TPUs, and use the reductionserver container image.
- B.** Configure the machines of the first two worker pools to have TPUs and to use a container image where your training code runs. Configure the third worker pool without accelerators, and use the reductionserver container image without accelerators, and choose a machine type that prioritizes bandwidth.
- C.** Configure the machines of the first two worker pools to have GPUs, and to use a container image where your training code runs. Configure the third worker pool to have GPUs, and use the reductionserver container image.
- D.** Configure the machines of the first two worker pools to have GPUs and to use a container image where your training code runs. Configure the third worker pool to use the reductionserver container image without accelerators, and choose a machine type that prioritizes bandwidth.

Answer: D ([LEAVE A REPLY](#))

Valid Professional-Machine-Learning-Engineer Dumps shared by Actual4test.com for Helping Passing Professional-Machine-Learning-Engineer Exam! Actual4test.com now offer the **newest Professional-Machine-Learning-Engineer exam dumps**, the Actual4test.com Professional-Machine-Learning-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** Actual4test.com Professional-Machine-Learning-Engineer dumps with Test Engine here: https://www.actual4test.com/Professional-Machine-Learning-Engineer_examcollection.html (412 Q&As Dumps, **30%OFF** Special Discount: **Freepdfdumps**)