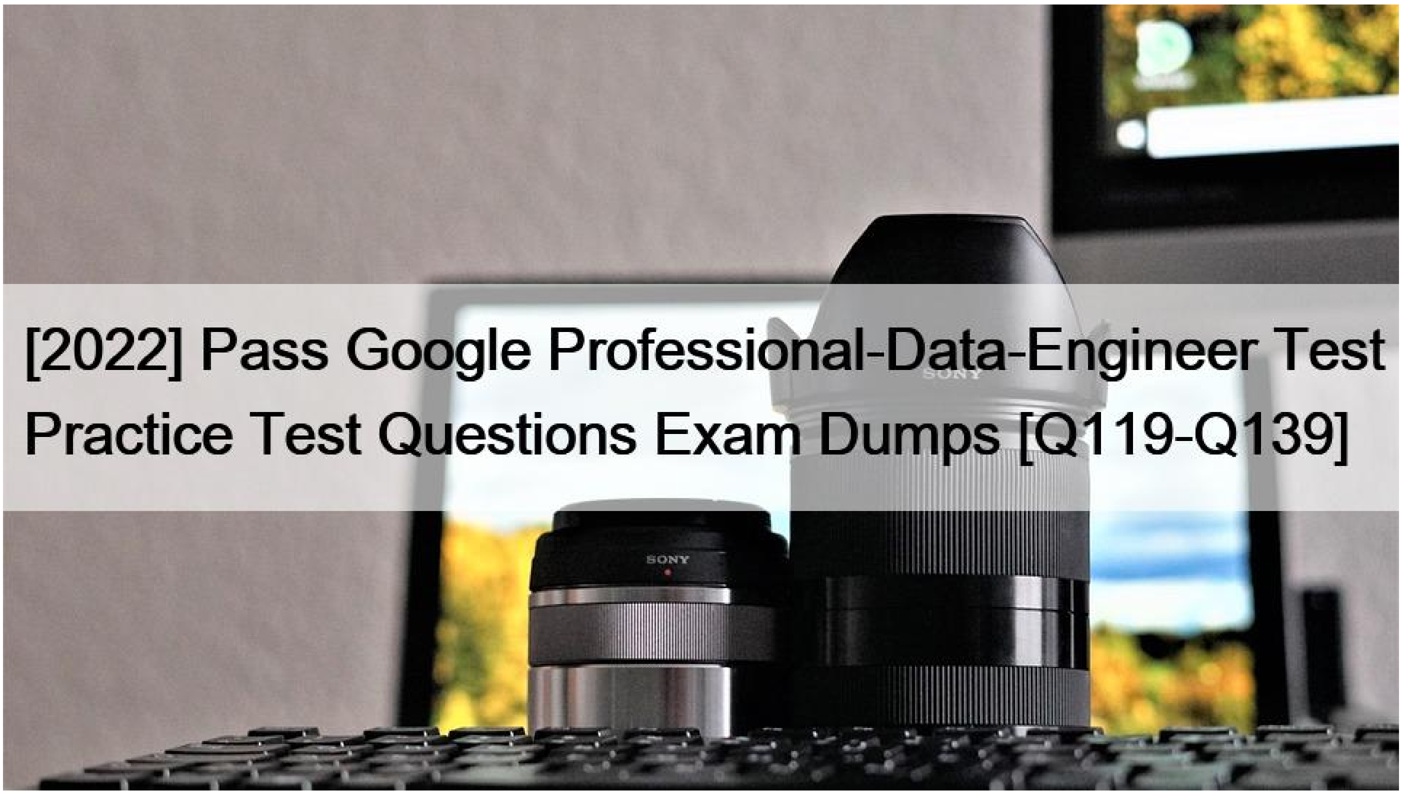
[2022 Pass Google Professional-Data-Engineer Test Practice Test Questions Exam Dumps [Q119-Q139



[2022] Pass Google Professional-Data-Engineer Test Practice Test Questions Exam Dumps
Verified Professional-Data-Engineer dumps Q&As - Professional-Data-Engineer dumps with Correct Answers

Understanding functional and technical aspects of Google Professional Data Engineer Exam Operationalizing machine learning models

The following will be discussed here:

- Continuous evaluation- Use of edge compute- Conversational experiences (e.g., Dialogflow)- Distributed vs. single machine - Hardware accelerators (e.g., GPU, TPU)- Customizing ML APIs (e.g., AutoML Vision, Auto ML text)- Leveraging pre-built ML models as a service- Measuring, monitoring, and troubleshooting machine learning models- Deploying an ML pipeline- ML APIs (e.g., Vision API, Speech API)- Ingesting appropriate data- Impact of dependencies of machine learning models- Machine learning terminology (e.g., features, labels, models, regression, classification, recommendation, supervised and unsupervised learning, evaluation metrics)- Common sources of error (e.g., assumptions about data)- Choosing the appropriate training and serving infrastructure- Operationalizing machine learning models

For more info visit:

Google-provided tutorials
Community-provided tutorials
Google-Data-Engineer-Practice-Test

NEW QUESTION 119

How would you query specific partitions in a BigQuery table?
* Use the DAY column in the WHERE clause
* Use the EXTRACT(DAY) clause
* Use the __PARTITIONTIME pseudo-column in the WHERE clause
* Use DATE BETWEEN in the WHERE clause
Partitioned tables include a pseudo column named _PARTITIONTIME that contains a date-based timestamp for data loaded into the table. To limit a query to particular partitions (such as Jan 1st and 2nd of 2017), use a clause similar to this:

WHERE _PARTITIONTIME BETWEEN TIMESTAMP(&#8216;2017-01-01&#8217;) AND TIMESTAMP(&#8216;2017-01-02&#8217;) Reference:
https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

NEW QUESTION 120

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)
* Denormalize the data as must as possible.
* Preserve the structure of the data as much as possible.
* Use BigQuery UPDATE to further reduce the size of the dataset.
* Develop a data pipeline where status updates are appended to BigQuery instead of updated.
* Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery&#8217;s support for external data sources to query.

NEW QUESTION 121

Which of the following statements about Legacy SQL and Standard SQL is not true?
* Standard SQL is the preferred query language for BigQuery.
* If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
* One difference between the two query languages is how you specify fully-qualified table names (i.e. table names that include their associated project name).
* You need to set a query language for each dataset and the default is Standard SQL.
You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.

Standard SQL has been the preferred query language since BigQuery 2.0 was released.

In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.

Due to the differences in syntax between the two query languages (such as with project- qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

Reference:

https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

NEW QUESTION 122

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients&#8217; personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems.

What should you do?
* Create an authorized view in BigQuery to restrict access to tables with sensitive data.
* Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
* Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
* Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API.

Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

NEW QUESTION 123

Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?
* Migrate the workload to Google Cloud Dataflow
* Use pre-emptible virtual machines (VMs) for the cluster
* Use a higher-memory node so that the job runs faster
* Use SSDs on the worker nodes so that the job can run faster

NEW QUESTION 124

Flowlogistic Case Study

Company Overview

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

* Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads

* Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

* Databases

&#8211; 8 physical servers in 2 clusters

&#8211; SQL Server &#8211; user data, inventory, static data

&#8211; 3 physical servers

&#8211; Cassandra &#8211; metadata, tracking messages

10 Kafka servers &#8211; tracking message aggregation and batch insert

* Application servers &#8211; customer front end, middleware for order/customs

&#8211; 60 virtual machines across 20 physical servers

&#8211; Tomcat &#8211; Java services

&#8211; Nginx &#8211; static content

&#8211; Batch servers

* Storage appliances

&#8211; iSCSI for virtual machine (VM) hosts

&#8211; Fibre Channel storage area network (FC SAN) &#8211; SQL server storage

Network-attached storage (NAS) image storage, logs, backups

* 10 Apache Hadoop /Spark servers

&#8211; Core Data Lake

&#8211; Data analysis workloads

* 20 miscellaneous servers

&#8211; Jenkins, monitoring, bastion hosts,

Business Requirements

* Build a reliable and reproducible environment with scaled panty of production.

* Aggregate data in a centralized Data Lake for analysis

* Use historical data to perform predictive analytics on future shipments

* Accurately track every shipment worldwide using proprietary technology

* Improve business agility and speed of innovation through rapid provisioning of new resources

* Analyze and optimize architecture for performance in the cloud

* Migrate fully to the cloud if all other requirements are met

Technical Requirements

* Handle both streaming and batch data

* Migrate existing Hadoop workloads

* Ensure architecture is scalable and elastic to meet the changing demands of the company.

* Use managed services whenever possible

* Encrypt data flight and at rest

Connect a VPN between the production data center and cloud environment

SEO Statement

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO&#8217; s tracking technology.

CFO Statement

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don&#8217;t want to commit capital to building out a server environment.

Flowlogistic&#8217;s management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

* Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
* Cloud Pub/Sub, Cloud Dataflow, and Local SSD
* Cloud Pub/Sub, Cloud SQL, and Cloud Storage
* Cloud Load Balancing, Cloud Dataflow, and Cloud Storage
* Cloud Dataflow, Cloud SQL, and Cloud Storage

Explanation

NEW QUESTION 125

Which of the following job types are supported by Cloud Dataproc (select 3 answers)?

* Hive
* Pig
* YARN
* Spark

Cloud Dataproc provides out-of-the box and end-to-end support for many of the most popular job types, including Spark, Spark SQL, PySpark, MapReduce, Hive, and Pig jobs.

Reference:

https://cloud.google.com/dataproc/docs/resources/faq#what_type_of_jobs_can_i_run

NEW QUESTION 126

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

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

Explanation

NEW QUESTION 127

Which TensorFlow function can you use to configure a categorical column if you don&#8217;t know all of the possible values for that column?

* categorical_column_with_vocabulary_list
* categorical_column_with_hash_bucket
* categorical_column_with_unknown_values
* sparse_column_with_keys

Explanation

If you know the set of all possible feature values of a column and there are only a few of them, you can use categorical_column_with_vocabulary_list. Each key in the list will get assigned an auto-incremental ID starting from 0.

What if we don&#8217;t know the set of possible values in advance? Not a problem. We can use categorical_column_with_hash_bucket instead. What will happen is that each possible value in the feature column occupation will be hashed to an integer ID as we encounter them in training.

Reference: https://www.tensorflow.org/tutorials/wide

NEW QUESTION 128

Your company is using WHILECARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

# Syntax error : Expected end of statement but got &#8220;-&#8221; at [4:11]

SELECT age

FROM

bigquery-public-data.noaa_gsod.gsod

WHERE

age != 99

AND_TABLE_SUFFIX = &#8216;1929&#8217;

ORDER BY

age DESC

Which table name will make the SQL statement work correctly?
* &#8216;bigquery-public-data.noaa_gsod.gsod&#8217;
* bigquery-public-data.noaa_gsod.gsod*
* &#8216;bigquery-public-data.noaa_gsod.gsod&#8217;*
* &#8216;bigquery-public-data.noaa_gsod.gsod*`

NEW QUESTION 129

Which of the following is NOT true about Dataflow pipelines?
* Dataflow pipelines are tied to Dataflow, and cannot be run on any other runner
* Dataflow pipelines can consume data from other Google Cloud services
* Dataflow pipelines can be programmed in Java
* Dataflow pipelines use a unified programming model, so can work both with streaming and batch data sources
Dataflow pipelines can also run on alternate runtimes like Spark and Flink, as they are built using the Apache Beam SDKs
Reference: https://cloud.google.com/dataflow/

NEW QUESTION 130

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

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

Import and Export to Bigquery from Cloud Storage is FREE. Also, when u store the csv files, Cloud Storage is cheaper than Bigquery. For processing Dataproc is cheaper than Dataflow.

NEW QUESTION 131

You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.

Which Google database service should you use?

* Cloud SQL
* BigQuery
* Cloud Bigtable
* Cloud Datastore

NEW QUESTION 132

You are building a model to make clothing recommendations. You know a user&#8217;s fashion pis likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

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

NEW QUESTION 133

Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?

* An hourly watermark
* An event time trigger
* The with Allowed Lateness method
* A processing time trigger

When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.

Processing time triggers. These triggers operate on the processing time ?the time when the data element is processed at any given stage in the pipeline. Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam&#8217;s default trigger is event time-based.

Reference: https://beam.apache.org/documentation/programming-guide/#triggers

NEW QUESTION 134

An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application.

They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?
* BigQuery
* Cloud SQL
* Cloud BigTable
* Cloud Datastore
ference: https://cloud.google.com/solutions/business-intelligence/

NEW QUESTION 135

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?
* Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
* Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
* Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.
* Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

NEW QUESTION 136

Case Study: 1 &#8211; Flowlogistic

Company Overview

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background

The company started as a regional trucking company, and then expanded into other logistics market.

Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

Databases

8 physical servers in 2 clusters

SQL Server &#8211; user data, inventory, static data

3 physical servers

Cassandra &#8211; metadata, tracking messages

10 Kafka servers &#8211; tracking message aggregation and batch insert

Application servers &#8211; customer front end, middleware for order/customs 60 virtual machines across 20 physical servers Tomcat &#8211; Java services Nginx &#8211; static content Batch servers Storage appliances iSCSI for virtual machine (VM) hosts Fibre Channel storage area network (FC SAN) ?SQL server storage Network-attached storage (NAS) image storage, logs, backups Apache Hadoop /Spark servers Core Data Lake Data analysis workloads

20 miscellaneous servers

Jenkins, monitoring, bastion hosts,

Business Requirements

Build a reliable and reproducible environment with scaled panty of production. Aggregate data in a centralized Data Lake for analysis Use historical data to perform predictive analytics on future shipments Accurately track every shipment worldwide using proprietary technology Improve business agility and speed of innovation through rapid provisioning of new resources Analyze and optimize architecture for performance in the cloud Migrate fully to the cloud if all other requirements are met Technical Requirements Handle both streaming and batch data Migrate existing Hadoop workloads Ensure architecture is scalable and elastic to meet the changing demands of the company.

Use managed services whenever possible

Encrypt data flight and at rest

Connect a VPN between the production data center and cloud environment SEO Statement We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO&#8217; s tracking technology.

CFO Statement

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability.

Additionally, I don&#8217;t want to commit capital to building out a server environment.

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

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

NEW QUESTION 137

The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?

* Workers
* Masters, workers, and parameter servers
* Workers and parameter servers
* Parameter servers

The CUSTOM tier is not a set tier, but rather enables you to use your own cluster specification. When you use this tier, set values to configure your processing cluster according to these guidelines:

You must set TrainingInput.masterType to specify the type of machine to use for your master node. You may set TrainingInput.workerCount to specify the number of workers to use. You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use. You can specify the type of machine for the master node, but you can&#8217;t specify more than one master node.

Reference: https://cloud.google.com/ml-engine/docs/training-overview#job_configuration_parameters

NEW QUESTION 138

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

* Delete the table CLICK_STREAM, and then re-create it such that the column DTis of the TIMESTAMPtype.

Reload the data.

* Add a column TSof the TIMESTAMPtype to the table CLICK_STREAM, and populate the numeric values from the column TSfor each row. Reference the column TSinstead of the column DTfrom now on.
* Create a view CLICK_STREAM_V, where strings from the column DTare cast into TIMESTAMPvalues.

Reference the view CLICK_STREAM_Vinstead of the table CLICK_STREAMfrom now on.

* Add two columns to the table CLICK STREAM: TSof the TIMESTAMPtype and IS_NEWof the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEWto true. For future queries, reference the column TSinstead of the column DT, with the WHEREclause ensuring that the value of IS_NEWmust be true.

* Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DTinto TIMESTAMPvalues. Run the query into a destination table NEW_CLICK_STREAM, in which the column TSis the TIMESTAMPtype. Reference the table NEW_CLICK_STREAMinstead of the table CLICK_STREAMfrom now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.

NEW QUESTION 139

Your company is implementing a data warehouse using BigQuery and you have been tasked with designing the data model You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days Based on Google&#8217;s recommended practices, what should you do to speed up the query without increasing storage costs?

* Denormalize the data
* Shard the data by customer ID
* Materialize the dimensional data in views
* Partition the data by transaction date

Professional-Data-Engineer certification guide Q&A from Training Expert Actualtests4sure:
https://www.actualtests4sure.com/Professional-Data-Engineer-test-questions.html]