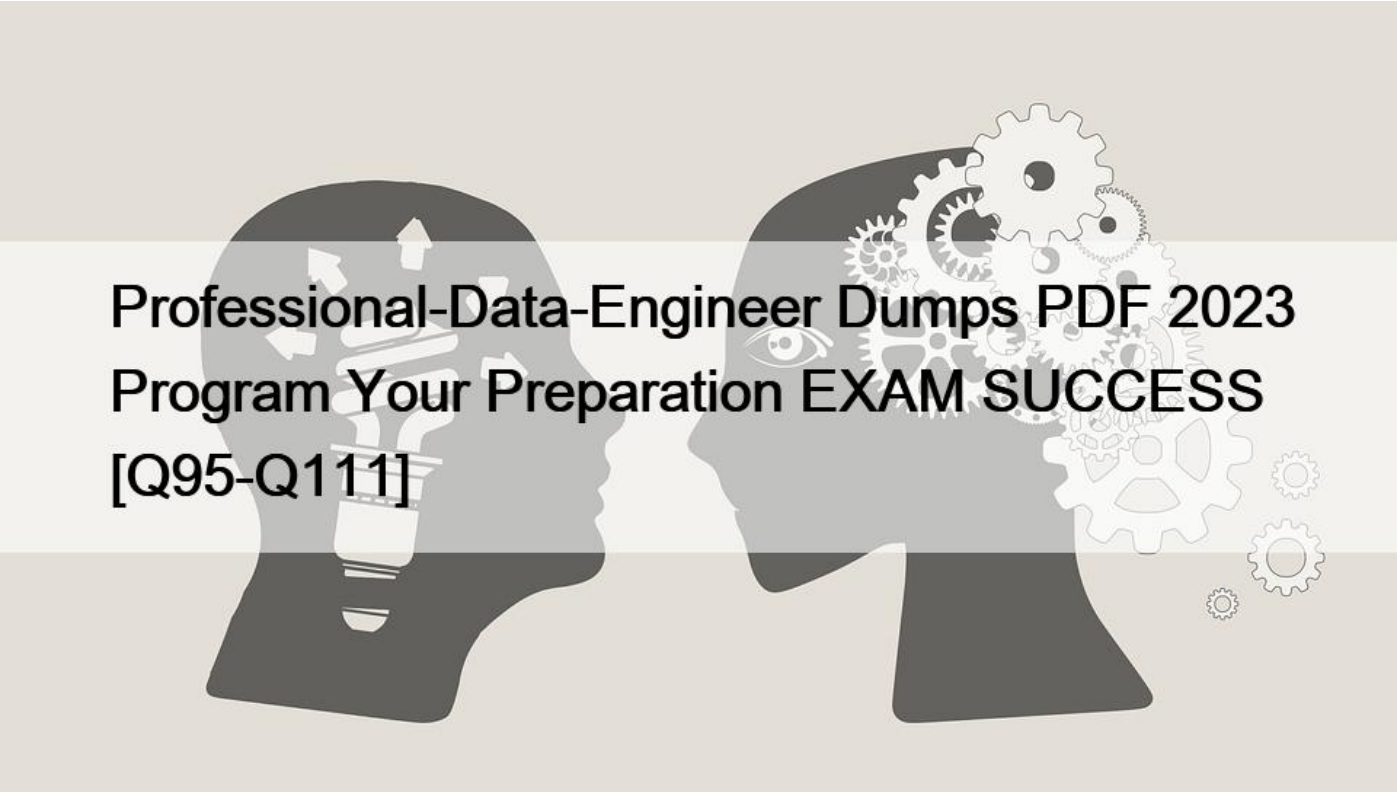
Professional-Data-Engineer Dumps PDF 2023 Program Your Preparation EXAM SUCCESS [Q95-Q111



Professional-Data-Engineer Dumps PDF 2023 Program Your Preparation EXAM SUCCESS Get Perfect Results with Premium Professional-Data-Engineer Dumps Updated 270 Questions NEW QUESTION 95

Which of these operations can you perform from the BigQuery Web UI?
* Upload a file in SQL format.
* Load data with nested and repeated fields.
* Upload a 20 MB file.
* Upload multiple files using a wildcard.

You can load data with nested and repeated fields using the Web UI.

You cannot use the Web UI to:

– Upload a file greater than 10 MB in size

– Upload multiple files at the same time

– Upload a file in SQL format

All three of the above operations can be performed using the “bq” command.

Reference: https://cloud.google.com/bigquery/loading-data

NEW QUESTION 96

Which of the following are feature engineering techniques? (Select 2 answers)
* Hidden feature layers
* Feature prioritization
* Crossed feature columns
* Bucketization of a continuous feature
Explanation

Selecting and crafting the right set of feature columns is key to learning an effective model.

Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.

Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.

Reference:

https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

NEW QUESTION 97

Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file.

What is the most likely cause of this problem?
* The CSV data loaded in BigQuery is not flagged as CSV.
* The CSV data has invalid rows that were skipped on import.
* The CSV data loaded in BigQuery is not using BigQuery’s default encoding.
* The CSV data has not gone through an ETL phase before loading into BigQuery.
Bigquery understands UTF-8 encoding anything other than that will result in data issues with schema.

NEW QUESTION 98

Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of dat

a. They want to improve this performance while minimizing cost. What should they do?
* Redefine the schema by evenly distributing reads and writes across the row space of the table.
* The performance issue should be resolved over time as the site of the BigDate cluster is increased.
* Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
* Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

NEW QUESTION 99

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)
* Create a new view over events using standard SQL
* Create a new partitioned table using a standard SQL query
* Create a new view over events_partitioned using standard SQL
* Create a service account for the ODBC connection to use for authentication
* Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared “events”

NEW QUESTION 100

For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?
* Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
* Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
* Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
* Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.
It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,

If it’s not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.

It is recommended to avoid creating your Compute Engine instance in a different region from

your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.

NEW QUESTION 101

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the “Trust No One” (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?
* Use gcloud kms keys createto create a symmetric key. Then use gcloud kms encryptto encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
* Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encryptto encrypt each archival file with the key. Use gsutil cpto upload each encrypted file to the Cloud Storage bucket.

Manually destroy the key previously used for encryption, and rotate the key once.
* Specify customer-supplied encryption key (CSEK) in the .botoconfiguration file. Use gsutil cpto upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
* Specify customer-supplied encryption key (CSEK) in the .botoconfiguration file. Use gsutil cpto upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

NEW QUESTION 102

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

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

https://cloud.google.com/sql/docs/mysql/high-availability

NEW QUESTION 103

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

* Add a SideInput that returns a Boolean if the element is corrupt.
* Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
* Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
* Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

NEW QUESTION 104

Case Study 1 – 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 – user data, inventory, static data

3 physical servers

– Cassandra – metadata, tracking messages

10 Kafka servers – tracking message aggregation and batch insert

* Application servers – customer front end, middleware for order/customs

60 virtual machines across 20 physical servers

– Tomcat – Java services

– Nginx – 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

* 10 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’ 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’t want to commit capital to building out a server environment.

Flowlogistic’s CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they’ve purchased a visualization tool to simplify the creation of BigQuery reports. However, they’ve been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?
* Export the data into a Google Sheet for virtualization.
* Create an additional table with only the necessary columns.
* Create a view on the table to present to the virtualization tool.
* Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

NEW QUESTION 105

If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?
* Unsupervised learning
* Regressor

* Classifier
* Clustering estimator

Regression is the supervised learning task for modeling and predicting continuous, numeric variables. Examples include predicting real-estate prices, stock price movements, or student test scores.

Classification is the supervised learning task for modeling and predicting categorical variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades.

Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset. Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.

Reference: https://elitedatascience.com/machine-learning-algorithms

NEW QUESTION 106

You are designing a pipeline that publishes application events to a Pub/Sub topic. You need to aggregate events across hourly intervals before loading the results to BigQuery for analysis. Your solution must be scalable so it can process and load large volumes of events to BigQuery. What should you do?

* Create a streaming Dataflow job to continually read from the Pub/Sub topic and perform the necessary aggregations using tumbling windows
* Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub-Sub topic and performing the necessary aggregations
* Schedule a Cloud Function to run hourly, pulling all avertable messages from the Pub/Sub topic and performing the necessary aggregations
* Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.

NEW QUESTION 107

What are two methods that can be used to denormalize tables in BigQuery?

* 1) Split table into multiple tables; 2) Use a partitioned table
* 1) Join tables into one table; 2) Use nested repeated fields
* 1) Use a partitioned table; 2) Join tables into one table
* 1) Use nested repeated fields; 2) Use a partitioned table

The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are dealing with sales transactions, you would write each individual fact to a record, along with the accompanying dimensions such as order and customer information.

The other method for denormalizing data takes advantage of BigQuery’s native support for nested and repeated structures in JSON or Avro input data. Expressing records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which would be represented as nested, repeated elements.

NEW QUESTION 108

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

* Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
* Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.
* Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
* Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.
Avro is compressed format and dataflow for parallel pipeline and bigquery for storage.

NEW QUESTION 109

Does Dataflow process batch data pipelines or streaming data pipelines?
* Only Batch Data Pipelines
* Both Batch and Streaming Data Pipelines
* Only Streaming Data Pipelines
* None of the above
Explanation

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

NEW QUESTION 110

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?
* Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
* Use Cloud Dataproc to run your transformations. Use the diagnosecommand to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
* Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
* Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs.

Configure the job to use non-default Compute Engine machine types when needed.

NEW QUESTION 111

How can you get a neural network to learn about relationships between categories in a categorical feature?
* Create a multi-hot column
* Create a one-hot column
* Create a hash bucket
* Create an embedding column
There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead of having just one value, like a continuous feature, it has many values, or dimensions. This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn’t encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.

Both of these problems can be solved by representing a categorical feature with an embedding column. The idea is that each category has a smaller vector with, let’s say, 5 values in it.

But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic

features in a neural network. The difference is that each category has a set of weights (5 of them in this case).

You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.

Reference:
https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/a-wide-and-deep-model.html