PR OFESSIONAL-DATA-ENGINEER Q&As

Professional Data Engineer on Google Cloud Platform

Pass Google PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.passapply.com/professional-data-engineer.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Google Official Exam Center

QUESTION 1

You need to modernize your existing on-premises data strategy. Your organization currently uses.

1.

Apache Hadoop clusters for processing multiple large data sets, including on-premises Hadoop Distributed File System (HDFS) for data replication.

2.

Apache Airflow to orchestrate hundreds of ETL pipelines with thousands of job steps.

You need to set up a new architecture in Google Cloud that can handle your Hadoop workloads and requires minimal changes to your existing orchestration processes. What should you do?

A. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases Convert your ETL pipelines to Dataflow.

B. Use Bigtable for your large workloads, with connections to Cloud Storage to handle any HDFS use cases Orchestrate your pipelines with Cloud Composer.

C. Use Dataproc to migrate your Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Use Cloud Data Fusion to visually design and deploy your ETL pipelines.

D. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Orchestrate your pipelines with Cloud Composer..

Correct Answer: D

Dataproc is a fully managed service that allows you to run Apache Hadoop and Spark workloads on Google Cloud. It is compatible with the open source ecosystem, so you can migrate your existing Hadoop clusters to Dataproc with minimal changes. Cloud Storage is a scalable, durable, and cost-effective object storage service that can replace HDFS for storing and accessing data. Cloud Storage offers interoperability with Hadoop through connectors, so you can use it as a data source or sink for your Dataproc jobs. Cloud Composer is a fully managed service that allows you to create, schedule, and monitor workflows using Apache Airflow. It is integrated with Google Cloud services, such as Dataproc, BigQuery, Dataflow, and Pub/Sub, so you can orchestrate your ETL pipelines across different platforms. Cloud Composer is compatible with your existing Airflow code, so you can migrate your existing orchestration processes to Cloud Composer with minimal changes. The other options are not as suitable as Dataproc and Cloud Composer for this use case, because they either require more changes to your existing code, or do not meet your requirements. Dataflow is a fully managed service that allows you to create and run scalable data processing pipelines using Apache Beam. However, Dataflow is not compatible with your existing Hadoop code, so you would need to rewrite your ETL pipelines using Beam. Bigtable is a fully managed NoSQL database service that can handle large and complex data sets. However, Bigtable is not compatible with your existing Hadoop code, so you would need to rewrite your queries and applications using Bigtable APIs. Cloud Data Fusion is a fully managed service that allows you to visually design and deploy data integration pipelines using a graphical interface. However, Cloud Data Fusion is not compatible with your existing Airflow code, so you would need to recreate your orchestration processes using Cloud Data Fusion UI. References: Dataproc overview Cloud Storage connector for Hadoop Cloud Composer overview

QUESTION 2

You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

A. Linear regression

B. Logistic classification

C. Recurrent neural network

D. Feedforward neural network

Correct Answer: A

QUESTION 3

You have designed an Apache Beam processing pipeline that reads from a Pub/Sub topic. The topic has a message retention duration of one day, and writes to a Cloud Storage bucket. You need to select a bucket location and processing strategy to prevent data loss in case of a regional outage with an RPO of 15 minutes. What should you do?

A. 1 Use a regional Cloud Storage bucket 2 Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs 3 Seek the subscription back in time by one day to recover the acknowledged messages 4 Start the Dataflow job in a secondary region and write in a bucket in the same region

B. 1 Use a multi-regional Cloud Storage bucket 2 Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs 3 Seek the subscription back in time by 60 minutes to recover the acknowledged messages 4 Start the Dataflow job in a secondary region

C. 1. Use a dual-region Cloud Storage bucket.

2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs

3 Seek the subscription back in time by 15 minutes to recover the acknowledged messages

4 Start the Dataflow job in a secondary region

D. 1. Use a dual-region Cloud Storage bucket with turbo replication enabled 2 Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs 3 Seek the subscription back in time by 60 minutes to recover the acknowledged messages 4 Start the Dataflow job in a secondary region.

Correct Answer: C

A dual-region Cloud Storage bucket is a type of bucket that stores data redundantly across two regions within the same continent. This provides higher availability and durability than a regional bucket, which stores data in a single region. A dual-region bucket also provides lower latency and higher throughput than a multi-regional bucket, which stores data across multiple regions within a continent or across continents. A dual-region bucket with turbo replication enabled is a premium option that offers even faster replication across regions, but it is more expensive and not necessary for this scenario. By using a dual-region Cloud Storage bucket, you can ensure that your data is protected from regional outages, and that you can access it from either region with low latency and high performance. You can also monitor the Dataflow metrics with Cloud Monitoring to determine when an outage occurs, and seek the subscription back in time by 15 minutes to recover the acknowledged messages. Seeking a subscription allows you to replay the messages from a Pub/Sub topic that were published within the message retention duration, which is one day in this case. By seeking the subscription back in time by 15 minutes, you can meet the RPO of 15 minutes, which means the maximum amount of data loss that is acceptable for your business. You can then start the Dataflow job in a secondary region and write to the same dual-region bucket, which will resume the processing of the messages and prevent data loss. Option A is not a good solution, as using a regional Cloud Storage bucket does not provide any redundancy or protection from regional outages. If the region where the bucket is located experiences an outage, you will not be able to access your data or write new data to the bucket. Seeking the subscription back in time by one day is also unnecessary and inefficient, as it will replay all the messages from the past day, even though you only need to recover the messages from the past 15

minutes. Option B is not a good solution, as using a multi-regional Cloud Storage bucket does not provide the best performance or cost-efficiency for this scenario. A multi-regional bucket stores data across multiple regions within a continent or across continents, which provides higher availability and durability than a dual-region bucket, but also higher latency and lower throughput. A multi-regional bucket is more suitable for serving data to a global audience, not for processing data with Dataflow within a single continent. Seeking the subscription back in time by 60 minutes is also unnecessary and inefficient, as it will replay more messages than needed to meet the RPO of 15 minutes. Option D is not a good solution, as using a dual-region Cloud Storage bucket with turbo replication enabled does not provide any additional benefit for this scenario, but only increases the cost. Turbo replication is a premium option that offers faster replication across regions, but it is not required to meet the RPO of 15 minutes. Seeking the subscription back in time by 60 minutes is also unnecessary and inefficient, as it will replay more messages than needed to meet the RPO of 15 minutes. References: Storage locations | Cloud Storage | Google Cloud, Dataflow metrics | Cloud Dataflow | Google Cloud, Seeking a subscription | Cloud Pub/Sub | Google Cloud, Recovery point objective (RPO) | Acronis.

QUESTION 4

When you design a Google Cloud Bigtable schema it is recommended that you _____.

A. Avoid schema designs that are based on NoSQL concepts

B. Create schema designs that are based on a relational database design

C. Avoid schema designs that require atomicity across rows

D. Create schema designs that require atomicity across rows

Correct Answer: C

All operations are atomic at the row level. For example, if you update two rows in a table, it\\'s possible that one row will be updated successfully and the other update will fail. Avoid schema designs that require atomicity across rows. Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

QUESTION 5

You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output. Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

A. Cancel

B. Drain

C. Stop

D. Finish

Correct Answer: B

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

Latest PROFESSIONAL-DATA-ENGINEER Dumps | PROFESSIONAL-DATA-ENGINEER PDF Dumps | PROFESSIONAL-DATA-ENGINEER VCE Dumps