Dumpsengine

Google

Professional-Data-Engineer

Cloud Certified Professional Data Engineer

QUESTION & ANSWERS

DUMPSENGINE

D. Pipelines can share data between instances

Answer: D Explanation:

The data and transforms in a pipeline are unique to, and owned by, that pipeline. While your program can create multiple pipelines, pipelines cannot share data or transforms
Reference: https://cloud.google.com/dataflow/model/pipelines

Question No : 2

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

A. Hive
B. Pig
C. YARN
D. Spark

Answer: A,B,D Explanation:

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

Question No : 3

C. Classifier
D. Clustering estimator

Answer: B Explanation:

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

Question No : 4

Which of the following is not possible using primitive roles?

A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.
B. Give UserA owner access and UserB editor access for all datasets in a project. C. Give a user access to view all datasets in a project, but not run queries on them. D. Give GroupA owner access and GroupB editor access for all datasets in a project.
Answer: C Explanation:

Primitive roles can be used to give owner, editor, or viewer access to a user or group, but they can't be used to separate data access permissions from job-running permissions.

A. You expect to store at least 10 TB of data.
B. You will mostly run batch workloads with scans and writes, rather than frequently executing random reads of a small number of rows.
C. You need to integrate with Google BigQuery.
D. You will not use the data to back a user-facing or latency-sensitive application.

Answer: C Explanation:

For example, if you plan to store extensive historical data for a large number of remote- sensing devices and then use the data to generate daily reports, the cost savings for HDD storage may justify the performance tradeoff. On the other hand, if you plan to use the data to display a real-time dashboard, it probably would not make sense to use HDD storage—reads would be much more frequent in this case, and reads are much slower with HDD storage.
Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

Question No : 6

The YARN ResourceManager and the HDFS NameNode interfaces are available on a
Cloud Dataproc cluster .

A. application node B. conditional node C. master node
D. worker node

Question No : 7

What are the minimum permissions needed for a service account used with Google Dataproc?

A. Execute to Google Cloud Storage; write to Google Cloud Logging
B. Write to Google Cloud Storage; read to Google Cloud Logging
C. Execute to Google Cloud Storage; execute to Google Cloud Logging
D. Read and write to Google Cloud Storage; write to Google Cloud Logging

Answer: D Explanation:

Service accounts authenticate applications running on your virtual machine instances to other Google Cloud Platform services. For example, if you write an application that reads and writes files on Google Cloud Storage, it must first authenticate to the Google Cloud Storage API. At a minimum, service accounts used with Cloud Dataproc need permissions to read and write to Google Cloud Storage, and to write to Google Cloud Logging.
Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

Question No : 8

Which of these is not a supported method of putting data into a partitioned table?

Answer: D Explanation:

You cannot change an existing table into a partitioned table. You must create a partitioned table from scratch. Then you can either stream data into it every day and the data will automatically be put in the right partition, or you can load data into a specific partition by using "$YYYYMMDD" at the end of the table name.
Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

Question No : 9

Which of the following is NOT one of the three main types of triggers that Dataflow supports?

A. Trigger based on element size in bytes
B. Trigger that is a combination of other triggers
C. Trigger based on element count
D. Trigger based on time

Answer: A Explanation:

There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers. You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way
Reference: https://cloud.google.com/dataflow/model/triggers

A. Use the DAY column in the WHERE clause
B. Use the EXTRACT(DAY) clause
C. Use the PARTITIONTIME pseudo-column in the WHERE clause
D. Use DATE BETWEEN in the WHERE clause

Answer: C Explanation:

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('2017-01-01') AND TIMESTAMP('2017-01-02')
Reference: https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

Dumpsengine