Google Certified Professional Data Engineer Exam

1. You are working on a sensitive project involving private user dat

A. Grant the consultant the Cloud Dataflow Developer role on the project.

B. Create an anonymized sample of the data for the consultant to work with in a different project.

C. Create a service account and allow the consultant to log on with it.

D. Grant the consultant the Viewer role on the project.

Answer(s): C

---

2. 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?

A. Schedule a Cloud Function to run hourly, pulling all avertable messages from the Pub/Sub topic and performing the necessary aggregations

B. Create a streaming Dataflow job to continually read from the Pub/Sub topic and perform the necessary aggregations using tumbling windows

C. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub-Sub topic and performing the necessary aggregations

D. 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.

Answer(s): B

---

3. You need to connect multiple applications with dynamic public IP addresses to a Cloud SQL instance. You configured users with strong passwords and enforced the SSL connection to your Cloud SOL instance. You want to use Cloud SQL public IP and ensure that you have secured connections. What should you do?

A. Add all application networks to Authorized Network and regularly update them.

B. Add CIDR 0.0.0.0/0 network to Authorized Network. Use Identity and Access Management (1AM) to add users.

C. Leave the Authorized Network empty. Use Cloud SQL Auth proxy on all applications.

D. Add CIDR 0.0.0.0/0 network to Authorized Network. Use Cloud SOL Auth proxy on all applications.

Answer(s): C

---

4. You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application.

A. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset

B. Create groups for your users and give those groups access to the dataset

C. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

D. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request

Answer(s): A

---

5. You have a BigQuery table that ingests data directly from a Pub/Sub subscription. The ingested data is encrypted with a Google-managed encryption key. You need to meet a new organization policy that requires you to use keys from a centralized Cloud Key Management Service (Cloud KMS) project to encrypt data at rest. What should you do?

A. Create a new BigOuory table by using customer-managed encryption keys (CMEK), and migrate the data from the old BigQuery table.

B. Create a new BigOuery table and Pub/Sub topic by using customer-managed encryption keys (CMEK), and migrate the data from the old Bigauery table.

C. Create a new Pub/Sub topic with CMEK and use the existing BigQuery table by using Google-managed encryption key.

D. Use Cloud KMS encryption key with Dataflow to ingest the existing Pub/Sub subscription to the existing BigQuery table.

Answer(s): A

---

6. Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the dat a. How should you deduplicate the data most efficiency?

A. Store each data entry as the primary key in a separate database and apply an index.

B. Compute the hash value of each data entry, and compare it with all historical data.

C. Assign global unique identifiers (GUID) to each data entry.

D. Maintain a database table to store the hash value and other metadata for each data entry.

Answer(s): D

---

7. You are running a Dataflow streaming pipeline, with Streaming Engine and Horizontal Autoscaling enabled. You have set the maximum number of workers to 1000. The input of your pipeline is Pub/Sub messages with notifications from Cloud Storage One of the pipeline transforms reads CSV files and emits an element for every CSV line. The Job performance is low. the pipeline is using only 10 workers, and you notice that the autoscaler is not spinning up additional workers. What should you do to improve performance?

A. Use Dataflow Prime, and enable Right Fitting to increase the worker resources.

B. Update the job to increase the maximum number of workers.

C. Enable Vertical Autoscaling to let the pipeline use larger workers.

D. Change the pipeline code, and introduce a Reshuffle step to prevent fusion.

Answer(s): A

---

8. You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.

A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.

B. Place the MariaDB instances in an Instance Group with a Health Check.

C. Install the StackDriver Agent and configure the MySQL plugin.

D. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.

Answer(s): D

---

9. You have a data pipeline with a Dataflow job that aggregates and writes time series metrics to Bigtable. You notice that data is slow to update in Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the dat a. What should you do?

A. Configure your Dataflow pipeline to use local execution.

B. Modify your Dataflow pipeline lo use the Flatten transform before writing to Bigtable.

C. Modify your Dataflow pipeline to use the CoGrcupByKey transform before writing to Bigtable.

D. Increase the maximum number of Dataflow workers by setting maxNumWorkers in PipelineOptions.

E. Increase the number of nodes in the Bigtable cluster.

Answer(s): D,E

---

10. Your business users need a way to clean and prepare data before using the data for analysis. Your business users are less technically savvy and prefer to work with graphical user interfaces to define their transformations. After the data has been transformed, the business users want to perform their analysis directly in a spreadsheet. You need to recommend a solution that they can use. What should you do?

A. Use Dataprep to clean the data, and write the results to BigQuery Analyze the data by using Connected Sheets.

B. Use Dataprep to clean the data, and write the results to BigQuery Analyze the data by using Looker Studio.

C. Use Dataflow to clean the data, and write the results to BigQuery. Analyze the data by using Connected Sheets.

D. Use Dataflow to clean the data, and write the results to BigQuery. Analyze the data by using Looker Studio.

Answer(s): A

---

11. You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.

A. Increase the size of your parquet files to ensure them to be 1 GB minimum.

B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.

C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.

D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

Answer(s): C

---

12. Your company's customer_order table in BigOuery stores the order history for 10 million customers, with a table size of 10 PB. You need to create a dashboard for the support team to view the order history. The dashboard has two filters, countryname and username. Both are string data types in the BigQuery table. When a filter is applied, the dashboard fetches the order history from the table and displays the query results. However, the dashboard is slow to show the results when applying the filters to the following query:

A. Cluster the table by country field, and partition by username field.

B. Partition the table by country and username fields.

C. Cluster the table by country and username fields

D. Partition the table by _PARTITIONTIME.

Answer(s): C

---

13. Your organization has two Google Cloud projects, project A and project B. In project A, you have a Pub/Sub topic that receives data from confidential sources. Only the resources in project A should be able to access the data in that topic. You want to ensure that project B and any future project cannot access data in the project A topic. What should you do?

A. Configure VPC Service Controls in the organization with a perimeter around the VPC of project A.

B. Add firewall rules in project A so only traffic from the VPC in project A is permitted.

C. Configure VPC Service Controls in the organization with a perimeter around project A.

D. Use Identity and Access Management conditions to ensure that only users and service accounts in project A can access resources in project.

Answer(s): C

---

14. You need (o give new website users a globally unique identifier (GUID) using a service that takes in data points and returns a GUID This data is sourced from both internal and external systems via HTTP calls that you will make via microservices within your pipeline There will be tens of thousands of messages per second and that can be multithreaded, and you worry about the backpressure on the system How should you design your pipeline to minimize that backpressure?

A. Create a new object in the startBundle method of DoFn

B. Call out to the service via HTTP

C. Create the pipeline statically in the class definition

D. Batch the job into ten-second increments

Answer(s): B

---

15. You are architecting a data transformation solution for BigQuery. Your developers are proficient with SOL and want to use the ELT development technique. In addition, your developers need an intuitive coding environment and the ability to manage SQL as code. You need to identify a solution for your developers to build these pipelines. What should you do?

A. Use Cloud Composer to load data and run SQL pipelines by using the BigQuery job operators.

B. Use Dataflow jobs to read data from Pub/Sub, transform the data, and load the data to BigQuery.

C. Use Dataform to build, manage, and schedule SQL pipelines.

D. Use Data Fusion to build and execute ETL pipelines

Answer(s): C

---

16. You have some data, which is shown in the graphic below. The two dimensions are X and Y, and the shade of each dot represents what class it is. You want to classify this data accurately using a linear algorithm.

A. X^2

B. cos(X)

C. X^2+Y^2

D. Y^2

Answer(s): B

---

17. You've migrated a Hadoop job from an on-premises cluster to Dataproc and Good Storage. Your Spark job is a complex analytical workload fiat consists of many shuffling operations, and initial data are parquet toes (on average 200-400 MB size each) You see some degradation in performance after the migration to Dataproc so you'd like to optimize for it. Your organization is very cost-sensitive so you'd Idee to continue using Dataproc on preemptibles (with 2 non-preemptibles workers only) for this workload. What should you do?

A. Switch from HODs to SSDs override the preemptible VMs configuration to increase the boot disk size

B. Switch from HDDs to SSDs. copy initial data from Cloud Storage to Hadoop Distributed File System (HDFS) run the Spark job and copy results back to Cloud Storage

C. Increase the see of your parquet files to ensure them to be 1 GB minimum

D. Switch to TFRecords format (appr 200 MB per We) instead of parquet files

Answer(s): A

---

18. 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?

A. The performance issue should be resolved over time as the site of the BigDate cluster is increased.

B. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.

D. Redefine the schema by evenly distributing reads and writes across the row space of the table.

Answer(s): D

---

19. You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

A. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.

B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.

C. Create a cron schedule in Cloud Dataprep.

D. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.

Answer(s): A

---

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

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..

Answer(s): D