1)Your retail company wants to predict customer churn using historical purchase data stored in BigQuery. The dataset includes customer demographics, purchase history, and a label indicating whether the customer churned or not. You want to build a machine learning model to identify customers at risk of churning. You need to create and train a logistic regression model for predicting customer churn, using the customer_data table with the churned column as the target label. Which BigQuery ML query should you use?

A
```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT *
FROM customer_data;
```

B
```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT * EXCEPT(churned),
       churned AS label
FROM customer_data;
```

C
```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT * EXCEPT(churned)
FROM customer_data;
```

D
```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT churned as label
FROM customer_data;
```

Correct Answer: *B*

5)You manage a Cloud Storage bucket that stores temporary files created during data processing. These temporary files are only needed for seven days, after which they are no longer needed. To reduce storage costs and keep your bucket organized, you want to automatically delete these files once they are older than seven days. What should you do?

A. Set up a Cloud Scheduler job that invokes a weekly Cloud Run function to delete files older than seven days.

B. Configure a Cloud Storage lifecycle rule that automatically deletes objects older than seven days.

C. Develop a batch process using Dataflow that runs weekly and deletes files based on their age.

D. Create a Cloud Run function that runs daily and deletes files older than seven days.

Correct Answer: *B*

https://shapingpixel.com

4)You want to process and load a daily sales CSV file stored in Cloud Storage into BigQuery for downstream reporting. You need to quickly build a scalable data pipeline that transforms the data while providing insights into data quality issues. What should you do?

A. Create a batch pipeline in Cloud Data Fusion by using a Cloud Storage source and a BigQuery sink.

B. Load the CSV file as a table in BigQuery, and use scheduled queries to run SQL transformation scripts.

C. Load the CSV file as a table in BigQuery. Create a batch pipeline in Cloud Data Fusion by using a BigQuery source and sink.

D. Create a batch pipeline in Dataflow by using the Cloud Storage CSV file to BigQuery batch template.

Correct Answer: *A*

www.shapingpixel.com

3)Your company is building a near real-time streaming pipeline to process JSON telemetry data from small appliances. You need to process messages arriving at a Pub/Sub topic, capitalize letters in the serial number field, and write results to BigQuery. You want to use a managed service and write a minimal amount of code for underlying transformations. What should you do?

A. Use a Pub/Sub to BigQuery subscription, write results directly to BigQuery, and schedule a transformation query to run every five minutes.

B. Use a Pub/Sub to Cloud Storage subscription, write a Cloud Run service that is triggered when objects arrive in the bucket, performs the transformations, and writes the results to BigQuery.

C. Use the “Pub/Sub to BigQuery” Dataflow template with a UDF, and write the results to BigQuery.

D. Use a Pub/Sub push subscription, write a Cloud Run service that accepts the messages, performs the transformations, and writes the results to BigQuery.

Correct Answer: *C*

www.shapingpixel.com

2)Your company has several retail locations. Your company tracks the total number of sales made at each location each day. You want to use SQL to calculate the weekly moving average of sales by location to identify trends for each store. Which query should you use?

A
```
SELECT store_id, date, total_sales, AVG(total_sales) OVER (
PARTITION BY store_id
ORDER BY total_sales RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

B
```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY date ORDER BY store_id ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
OVER (
PARTITION BY date ORDER BY store_id ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

C
```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY store_id
ORDER BY date ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

D
```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY total_sales
ORDER BY date RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

Correct Answer: *C*

www.shapingpixel.com

6)You work for a healthcare company that has a large on-premises data system containing patient records with personally identifiable information (PII) such as names, addresses, and medical diagnoses. You need a standardized managed solution that de-identifies PII across all your data feeds prior to ingestion to Google Cloud. What should you do?

A. Use Cloud Run functions to create a serverless data cleaning pipeline. Store the cleaned data in BigQuery.

B. Use Cloud Data Fusion to transform the data. Store the cleaned data in BigQuery.

C. Load the data into BigQuery, and inspect the data by using SQL queries. Use Dataflow to transform the data and remove any errors.

D. Use Apache Beam to read the data and perform the necessary cleaning and transformation operations. Store the cleaned data in BigQuery.

Correct Answer: *B*

https://shapingpixel.com

10)Another team in your organization is requesting access to a BigQuery dataset. You need to share the dataset with the team while minimizing the risk of unauthorized copying of data. You also want to create a reusable framework in case you need to share this data with other teams in the future. What should you do?

A. Create authorized views in the team's Google Cloud project that is only accessible by the team.

B. Create a private exchange using Analytics Hub with data egress restriction, and grant access to the team members.

C. Enable domain restricted sharing on the project. Grant the team members the BigQuery Data Viewer IAM role on the dataset.

D. Export the dataset to a Cloud Storage bucket in the team's Google Cloud project that is only accessible by the team.

Correct Answer: *B*

www.shapingpixel.com

9)You are designing a pipeline to process data files that arrive in Cloud Storage by 3:00 am each day. Data processing is performed in stages, where the output of one stage becomes the input of the next. Each stage takes a long time to run. Occasionally a stage fails, and you have to address the problem. You need to ensure that the final output is generated as quickly as possible. What should you do?

A. Design a Spark program that runs under Dataproc. Code the program to wait for user input when an error is detected. Rerun the last action after correcting any stage output data errors.

B. Design the pipeline as a set of PTransforms in Dataflow. Restart the pipeline after correcting any stage output data errors.

C. Design the workflow as a Cloud Workflow instance. Code the workflow to jump to a given stage based on an input parameter. Rerun the workflow after correcting any stage output data errors.

D. Design the processing as a directed acyclic graph (DAG) in Cloud Composer. Clear the state of the failed task after correcting any stage output data errors.

Correct Answer: *D*

https://shapingpixel.com

8)You work for an ecommerce company that has a BigQuery dataset that contains customer purchase history, demographics, and website interactions. You need to build a machine learning (ML) model to predict which customers are most likely to make a purchase in the next month. You have limited engineering resources and need to minimize the ML expertise required for the solution. What should you do?

A. Use BigQuery ML to create a logistic regression model for purchase prediction.

B. Use Vertex AI Workbench to develop a custom model for purchase prediction.

C. Use Colab Enterprise to develop a custom model for purchase prediction.

D. Export the data to Cloud Storage, and use AutoML Tables to build a classification model for purchase prediction.

Correct Answer: *A*

www.shapingpixel.com

7)You manage a large amount of data in Cloud Storage, including raw data, processed data, and backups. Your organization is subject to strict compliance regulations that mandate data immutability for specific data types. You want to use an efficient process to reduce storage costs while ensuring that your storage strategy meets retention requirements. What should you do?

A. Configure lifecycle management rules to transition objects to appropriate storage classes based on access patterns. Set up Object Versioning for all objects to meet immutability requirements.

B. Move objects to different storage classes based on their age and access patterns. Use Cloud Key Management Service (Cloud KMS) to encrypt specific objects with customer-managed encryption keys (CMEK) to meet immutability requirements.

C. Create a Cloud Run function to periodically check object metadata, and move objects to the appropriate storage class based on age and access patterns. Use object holds to enforce immutability for specific objects.

D. Use object holds to enforce immutability for specific objects, and configure lifecycle management rules to transition objects to appropriate storage classes based on age and access patterns.

https://shapingpixel.com

14)Your team wants to create a monthly report to analyze inventory data that is updated daily. You need to aggregate the inventory counts by using only the most recent month of data, and save the results to be used in a Looker Studio dashboard. What should you do?

A. Create a materialized view in BigQuery that uses the SUM( ) function and the DATE_SUB( ) function.

B. Create a saved query in the BigQuery console that uses the SUM( ) function and the DATE_SUB( ) function. Re-run the saved query every month, and save the results to a BigQuery table.

C. Create a BigQuery table that uses the SUM( ) function and the _PARTITIONDATE filter.

D. Create a BigQuery table that uses the SUM( ) function and the DATE_DIFF( ) function.

Correct Answer: *A*

www.shapingpixel.com

13)You need to create a new data pipeline. You want a serverless solution that meets the following requirements:
- Data is streamed from Pub/Sub and is processed in real-time.
- Data is transformed before being stored.
- Data is stored in a location that will allow it to be analyzed with SQL using Looker.

Which Google Cloud services should you recommend for the pipeline?

A. 1. Dataproc Serverless
2. Bigtable

B. 1. Cloud Composer
2. Cloud SQL for MySQL

C. 1. BigQuery
2. Analytics Hub

D. 1. Dataflow
2. BigQuery

Correct Answer: *D*

www.shapingpixel.com
https://shapingpixel.com

12)You recently inherited a task for managing Dataflow streaming pipelines in your organization and noticed that proper access had not been provisioned to you. You need to request a Google-provided IAM role so you can restart the pipelines. You need to follow the principle of least privilege. What should you do?

A. Request the Dataflow Developer role.

B. Request the Dataflow Viewer role.

C. Request the Dataflow Worker role.

D. Request the Dataflow Admin role.

Correct Answer: *A*

www.shapingpixel.com

11)Your company has developed a website that allows users to upload and share video files. These files are most frequently accessed and shared when they are initially uploaded. Over time, the files are accessed and shared less frequently, although some old video files may remain very popular.
You need to design a storage system that is simple and cost-effective. What should you do?

A. Create a single-region bucket with Autoclass enabled.

B. Create a single-region bucket. Configure a Cloud Scheduler job that runs every 24 hours and changes the storage class based on upload date.

C. Create a single-region bucket with custom Object Lifecycle Management policies based on upload date.

D. Create a single-region bucket with Archive as the default storage class.

Correct Answer: *C*

www.shapingpixel.com

15)You have a BigQuery dataset containing sales data. This data is actively queried for the first 6 months. After that, the data is not queried but needs to be retained for 3 years for compliance reasons. You need to implement a data management strategy that meets access and compliance requirements, while keeping cost and administrative overhead to a minimum. What should you do?

A. Use BigQuery long-term storage for the entire dataset. Set up a Cloud Run function to delete the data from BigQuery after 3 years.

B. Partition a BigQuery table by month. After 6 months, export the data to Coldline storage. Implement a lifecycle policy to delete the data from Cloud Storage after 3 years.

C. Set up a scheduled query to export the data to Cloud Storage after 6 months. Write a stored procedure to delete the data from BigQuery after 3 years.

D. Store all data in a single BigQuery table without partitioning or lifecycle policies.

Correct Answer: *B*

www.shapingpixel.com

19)Your company currently uses an on-premises network file system (NFS) and is migrating data to Google Cloud. You want to be able to control how much bandwidth is used by the data migration while capturing detailed reporting on the migration status. What should you do?

A. Use a Transfer Appliance.

B. Use Cloud Storage FUSE.

C. Use Storage Transfer Service.

D. Use gcloud storage commands.

Answer : **C**

www.shapingpixel.com

18)You work for a home insurance company. You are frequently asked to create and save risk reports with charts for specific areas using a publicly available storm event dataset. You want to be able to quickly create and re-run risk reports when new data becomes available. What should you do?

A. Export the storm event dataset as a CSV file. Import the file to Google Sheets, and use cell data in the worksheets to create charts.

B. Copy the storm event dataset into your BigQuery project. Use BigQuery Studio to query and visualize the data in Looker Studio.

C. Reference and query the storm event dataset using SQL in BigQuery Studio. Export the results to Google Sheets, and use cell data in the worksheets to create charts.

D. Reference and query the storm event dataset using SQL in a Colab Enterprise notebook. Display the table results and document with Markdown, and use Matplotlib to create charts.

Answer : **B**

17)You work for an online retail company. Your company collects customer purchase data in CSV files and pushes them to Cloud Storage every 10 minutes. The data needs to be transformed and loaded into BigQuery for analysis. The transformation involves cleaning the data, removing duplicates, and enriching it with product information from a separate table in BigQuery. You need to implement a low-overhead solution that initiates data processing as soon as the files are loaded into Cloud Storage. What should you do?

A. Use Cloud Composer sensors to detect files loading in Cloud Storage. Create a Dataproc cluster, and use a Composer task to execute a job on the cluster to process and load the data into BigQuery.

B. Schedule a direct acyclic graph (DAG) in Cloud Composer to run hourly to batch load the data from Cloud Storage to BigQuery, and process the data in BigQuery using SQL.

C. Use Dataflow to implement a streaming pipeline using an OBJECT_FINALIZE notification from Pub/Sub to read the data from Cloud Storage, perform the transformations, and write the data to BigQuery.

D. Create a Cloud Data Fusion job to process and load the data from Cloud Storage into BigQuery. Create an OBJECT_FINALI ZE notification in Pub/Sub, and trigger a Cloud Run function to start the Cloud Data Fusion job as soon as new files are loaded.

Answer : **C**

16)You have created a LookML model and dashboard that shows daily sales metrics for five regional managers to use. You want to ensure that the regional managers can only see sales metrics specific to their region. You need an easy-to-implement solution. What should you do?

A. Create a sales_region user attribute, and assign each manager's region as the value of their user attribute. Add an access_filter Explore filter on the region_name dimension by using the sales_region user attribute.

B. Create five different Explores with the sql_always_filter Explore filter applied on the region_name dimension. Set each region_name value to the corresponding region for each manager.

C. Create separate Looker dashboards for each regional manager. Set the default dashboard filter to the corresponding region for each manager.

D. Create separate Looker instances for each regional manager. Copy the LookML model and dashboard to each instance. Provision viewer access to the corresponding manager.

Correct Answer: *A*

https://shapingpixel.com

23)Your organization's business analysts require near real-time access to streaming dat
a. However, they are reporting that their dashboard queries are loading slowly. After investigating BigQuery query performance, you discover the slow dashboard queries perform several joins and aggregations. You need to improve the dashboard loading time and ensure that the dashboard data is as up-to-date as possible. What should you do?

A. Disable BiqQuery query result caching.

B. Modify the schema to use parameterized data types.

C. Create a scheduled query to calculate and store intermediate results.

D. Create materialized views.

Answer : **D**

www.shap

22)You need to create a data pipeline that streams event information from applications in multiple Google Cloud regions into BigQuery for near real-time analysis. The data requires transformation before loading. You want to create the pipeline using a visual interface. What should you do?

A. Push event information to a Pub/Sub topic. Create a Dataflow job using the Dataflow job builder.

B. Push event information to a Pub/Sub topic. Create a Cloud Run function to subscribe to the Pub/Sub topic, apply transformations, and insert the data into BigQuery.

C. Push event information to a Pub/Sub topic. Create a BigQuery subscription in Pub/Sub.

D. Push event information to Cloud Storage, and create an external table in BigQuery. Create a BigQuery scheduled job that executes once each day to apply transformations.

Answer : **A**

21)Your organization's ecommerce website collects user activity logs using a Pub/Sub topic. Your organization's leadership team wants a dashboard that contains aggregated user engagement metrics. You need to create a solution that transforms the user activity logs into aggregated metrics, while ensuring that the raw data can be easily queried. What should you do?

A. Create a Dataflow subscription to the Pub/Sub topic, and transform the activity logs. Load the transformed data into a BigQuery table for reporting.

B. Create an event-driven Cloud Run function to trigger a data transformation pipeline to run. Load the transformed activity logs into a BigQuery table for reporting.

C. Create a Cloud Storage subscription to the Pub/Sub topic. Load the activity logs into a bucket using the Avro file format. Use Dataflow to transform the data, and load it into a BigQuery table for reporting.

D. Create a BigQuery subscription to the Pub/Sub topic, and load the activity logs into the table. Create a materialized view in BigQuery using SQL to transform the data for reporting

Answer : **A**

https://shapingpixel.com

20)You are a Looker analyst. You need to add a new field to your Looker report that generates SQL that will run against your company's database. You do not have the Develop permission. What should you do?

A. Create a new field in the LookML layer, refresh your report, and select your new field from the field picker.

B. Create a calculated field using the Add a field option in Looker Studio, and add it to your report.

C. Create a table calculation from the field picker in Looker, and add it to your report.

D. Create a custom field from the field picker in Looker, and add it to your report.

Answer : **D**

www.shapingpixel.com

27)Your organization has a petabyte of application logs stored as Parquet files in Cloud Storage. You need to quickly perform a one-time SQL-based analysis of the files and join them to data that already resides in BigQuery. What should you do?

A. Create a Dataproc cluster, and write a PySpark job to join the data from BigQuery to the files in Cloud Storage.

B. Launch a Cloud Data Fusion environment, use plugins to connect to BigQuery and Cloud Storage, and use the SQL join operation to analyze the data.

C. Create external tables over the files in Cloud Storage, and perform SQL joins to tables in BigQuery to analyze the data.

D. Use the bq load command to load the Parquet files into BigQuery, and perform SQL joins to analyze the data.

Answer : **C**

https://shapingpixel.com

26)Your organization needs to implement near real-time analytics for thousands of events arriving each second in Pub/Sub. The incoming messages require transformations. You need to configure a pipeline that processes, transforms, and loads the data into BigQuery while minimizing development time. What should you do?

A. Use a Google-provided Dataflow template to process the Pub/Sub messages, perform transformations, and write the results to BigQuery.

B. Create a Cloud Data Fusion instance and configure Pub/Sub as a source. Use Data Fusion to process the Pub/Sub messages, perform transformations, and write the results to BigQuery.

C. Load the data from Pub/Sub into Cloud Storage using a Cloud Storage subscription. Create a Dataproc cluster, use PySpark to perform transformations in Cloud Storage, and write the results to BigQuery.

D. Use Cloud Run functions to process the Pub/Sub messages, perform transformations, and write the results to BigQuery.

Answer : **A**

https://shapingpixel.com

25)Your organization needs to store historical customer order data. The data will only be accessed once a month for analysis and must be readily available within a few seconds when it is accessed. You need to choose a storage class that minimizes storage costs while ensuring that the data can be retrieved quickly. What should you do?

A. Store the data in Cloud Storage using Nearline storage.

B. Store the data in Cloud Storage using Coldline storage.

C. Store the data in Cloud Storage using Standard storage.

D. Store the data in Cloud Storage using Archive storage.

Answer : **A**

www.shapingpixel.com

24)You have a Dataflow pipeline that processes website traffic logs stored in Cloud Storage and writes the processed data to BigQuery. You noticed that the pipeline is failing intermittently. You need to troubleshoot the issue. What should you do?

A. Use Cloud Logging to identify error groups in the pipeline's logs. Use Cloud Monitoring to create a dashboard that tracks the number of errors in each group.

B. Use Cloud Logging to create a chart displaying the pipeline's error logs. Use Metrics Explorer to validate the findings from the chart.

C. Use Cloud Logging to view error messages in the pipeline's logs. Use Cloud Monitoring to analyze the pipeline's metrics, such as CPU utilization and memory usage.

D. Use the Dataflow job monitoring interface to check the pipeline's status every hour. Use Cloud Profiler to analyze the pipeline's metrics, such as CPU utilization and memory usage.

Answer : **C**

https://shapingpixel.com

31) You manage a BigQuery table that is used for critical end-of-month reports. The table is updated weekly with new sales dat a. You want to prevent data loss and reporting issues if the table is accidentally deleted. What should you do?

A. Create a view of the table. On deletion, re-create the deleted table from the view and time travel data.
B. Schedule the creation of a new snapshot of the table once a week. On deletion, re-create the deleted table using the snapshot and time travel data.
C. Create a clone of the table. On deletion, re-create the deleted table by copying the content of the clone.
D. Configure the time travel duration on the table to be exactly seven days. On deletion, re-create the deleted table solely from the time travel data.

Correct Answer: B

30) You are a database administrator managing sales transaction data by region stored in a BigQuery table. You need to ensure that each sales representative can only see the transactions in their region. What should you do?

A. Create a data masking rule.
B. Create a row-level access policy.
C. Add a policy tag in BigQuery.
D. Grant the appropriate 1AM permissions on the dataset.

Correct Answer: B

www.shapingpixel.com

29) You created a customer support application that sends several forms of data to Google Cloud. Your application is sending:
1. Audio files from phone interactions with support agents that will be accessed during trainings.
2. CSV files of users' personally identifiable information (PII) that will be analyzed with SQL.
3. A large volume of small document files that will power other applications.
You need to select the appropriate tool for each data type given the required use case, while following Google-recommended practices. Which should you choose?

A.1. Cloud Storage
2. BigQuery
3. Firestore
B.1. Filestore
2. Cloud SQL for PostgreSQL
3. Datastore
C.1. Filestore
2. Bigtable
3. BigQuery
D.1. Cloud Storage
2. CloudSQL for PostgreSQL
3. Bigtable

Correct Answer: A

https://shapingpixel.com

28)You are developing a data ingestion pipeline to load small CSV files into BigQuery from Cloud Storage. You want to load these files upon arrival to minimize data latency. You want to accomplish this with minimal cost and maintenance. What should you do?

A. Use the bq command-line tool within a Cloud Shell instance to load the data into BigQuery.

B. Create a Cloud Composer pipeline to load new files from Cloud Storage to BigQuery and schedule it to run every 10 minutes.

C. Create a Cloud Run function to load the data into BigQuery that is triggered when data arrives in Cloud Storage.

D. Create a Dataproc cluster to pull CSV files from Cloud Storage, process them using Spark, and write the results to BigQuery.

Answer : **C**

www.shapingpixel.com

32) Your organization uses scheduled queries to perform transformations on data stored in BigQuery. You discover that one of your scheduled queries has failed. You need to troubleshoot the issue as quickly as possible. What should you do?

A. Set up a log sink using the gcloud CLI to export BigQuery audit logs to BigQuery. Query those logs to identify the error associated with the failed job ID.
B. Navigate to the Logs Explorer page in Cloud Logging. Use filters to find the failed job, and analyze the error details.
C. Request access from your admin to the BigQuery information_schema. Query the jobs view with the failed job ID, and analyze error details.
D. Navigate to the Scheduled queries page in the Google Cloud console. Select the failed job, and analyze the error details.

Correct Answer: D

www.shapingpixel.com

35) Your organization stores highly personal data in BigQuery and needs to comply with strict data privacy regulations. You need to ensure that sensitive data values are rendered unreadable whenever an employee leaves the organization. What should you do?

A. Use AEAD functions and delete keys when employees leave the organization.
B. Use customer-managed encryption keys (CMEK) and delete keys when employees leave the organization.
C. Use column-level access controls with policy tags and revoke viewer permissions when employees leave the organization.
D. Use dynamic data masking and revoke viewer permissions when employees leave the organization.

Correct Answer: B

www.shapingpixel.com

34) You need to design a data pipeline that ingests data from CSV, Avro, and Parquet files into Cloud Storage. The data includes raw user input. You need to remove all malicious SQL injections before storing the data in BigQuery. Which data manipulation methodology should you choose?

A. ETL
B. ELT
C. EL
D. ETLT

Correct Answer: A

www.shapingpixel.com

33) You are predicting customer churn for a subscription-based service. You have a 50 PB historical customer dataset in BigQuery that includes demographics, subscription information, and engagement metrics. You want to build a churn prediction model with minimal overhead. You want to follow the Google-recommended approach. What should you do?

A. Create a Looker dashboard that is connected to BigQuery. Use LookML to predict churn.
B. Export the data from BigQuery to a local machine. Use scikit-learn in a Jupyter notebook to build the churn prediction model.
C. Use the BigQuery Python client library in a Jupyter notebook to query and preprocess the data in BigQuery. Use the CREATE MODEL statement in BigQueryML to train the churn prediction model.
D. Use Dataproc to create a Spark cluster. Use the Spark MLlib within the cluster to build the churn prediction model.

Correct Answer: C

36) You are working with a large dataset of customer reviews stored in Cloud Storage. The dataset contains several inconsistencies, such as missing values, incorrect data types, and duplicate entries. You need to clean the data to ensure that it is accurate and consistent before using it for analysis. What should you do?

A. Use BigQuery to batch load the data into BigQuery. Use SQL for cleaning and analysis.
B. Use Storage Transfer Service to move the data to a different Cloud Storage bucket. Use event triggers to invoke Cloud Run functions to load the data into BigQuery. Use SQL for analysis.
C. Use Cloud Run functions to clean the data and load it into BigQuery. Use SQL for analysis.
D. Use the PythonOperator in Cloud Composer to clean the data and load it into BigQuery. Use SQL for analysis.

Correct Answer: A

www.shapingpixel.com

40) Your team uses the Google Ads platform to visualize metrics. You want to export the data to BigQuery to get more granular insights. You need to execute a one-time transfer of historical data and automatically update data daily. You want a solution that is low-code, serverless, and requires minimal maintenance. What should you do?

A. Export the historical data to BigQuery by using BigQuery Data Transfer Service. Use Cloud Composer for daily automation.

B. Export the historical data to Cloud Storage by using Storage Transfer Service. Use Pub/Sub to trigger a Dataflow template that loads data for daily automation.

C. Export the historical data as a CSV file. Import the file into BigQuery for analysis. Use Cloud Composer for daily automation.

D. Export the historical data to BigQuery by using BigQuery Data Transfer Service. Use BigQuery Data Transfer Service for daily automation.

Answer: D

39) Your organization has a BigQuery dataset that contains sensitive employee information such as salaries and performance reviews. The payroll specialist in the HR department needs to have continuous access to aggregated performance data, but they do not need continuous access to other sensitive dat

a. You need to grant the payroll specialist access to the performance data without granting them access to the entire dataset using the simplest and most secure approach. What should you do?

A. Use authorized views to share query results with the payroll specialist.
B. Create row-level and column-level permissions and policies on the table that contains performance data in the dataset. Provide the payroll specialist with the appropriate permission set.
C. Create a table with the aggregated performance data. Use table-level permissions to grant access to the payroll specialist.
D. Create a SQL query with the aggregated performance data. Export the results to an Avro file in a Cloud Storage bucket. Share the bucket with the payroll specialist.

Answer: A

Explanation:

Using authorized views is the simplest and most secure way to grant the payroll specialist access to aggregated performance data without exposing the entire dataset. Authorized views allow you to create a view in BigQuery that contains only the query results for the aggregated performance data. The payroll specialist can query the view without being granted access to the underlying sensitive data. This approach ensures security, adheres to the principle of least privilege, and eliminates the need to manage complex row-level or column-level permissions.

https://shapingpixel.com

38) You work for a global financial services company that trades stocks 24/7. You have a Cloud SGL for PostgreSQL user database. You need to identify a solution that ensures that the database is continuously operational, minimizes downtime, and will not lose any data in the event of a zonal outage. What should you do?

A. Continuously back up the Cloud SGL instance to Cloud Storage. Create a Compute Engine instance with PostgreSCL in a different region. Restore the backup in the Compute Engine instance if a failure occurs.
B. Create a read replica in another region. Promote the replica to primary if a failure occurs.
C. Configure and create a high-availability Cloud SQL instance with the primary instance in zone A and a secondary instance in any zone other than zone A.
D. Create a read replica in the same region but in a different zone.

Answer: C

Explanation:
Configuring a high-availability (HA) Cloud SQL instance ensures continuous operation, minimizes downtime, and prevents data loss in the event of a zonal outage. In this setup, the primary instance is located in one zone (e.g., zone A), and a synchronous secondary instance is located in a different zone within the same region. This configuration ensures that all data is replicated to the secondary instance in real-time. In the event of a failure in the primary zone, the system automatically promotes the secondary instance to primary, ensuring seamless failover with no data loss and minimal downtime. This is the recommended approach for mission-critical, highly available databases.

https://shapingpixel.com

37) You are migrating data from a legacy on-premises MySQL database to Google Cloud. The database contains various tables with different data types and sizes, including large tables with millions of rows and transactional data. You need to migrate this data while maintaining data integrity, and minimizing downtime and cost. What should you do?

A. Set up a Cloud Composer environment to orchestrate a custom data pipeline. Use a Python script to extract data from the MySQL database and load it to MySQL on Compute Engine.
B. Use Cloud Data Fusion to migrate the MySQL database to MySQL on Compute Engine.
C. Export the MySQL database to CSV files, transfer the files to Cloud Storage by using Storage Transfer Service, and load the files into a Cloud SQL for MySQL instance.
D. Use Database Migration Service to replicate the MySQL database to a Cloud SQL for MySQL instance.

Correct Answer: D

https://shapingpixel.com

Sample Questions

10. You are developing a data ingestion pipeline to load small CSV files into BigQuery from Cloud Storage. You want to load these files upon arrival to minimize data latency. You want to accomplish this with minimal cost and maintenance. What should you do?

A. Use the bq command-line tool within a Cloud Shell instance to load the data into BigQuery.

B. Create a Cloud Composer pipeline to load new files from Cloud Storage to BigQuery and schedule it to run every 10 minutes.

C. Create a Cloud Run function to load the data into BigQuery that is triggered when data arrives in Cloud Storage.

D. Create a Dataproc cluster to pull CSV files from Cloud Storage, process them using Spark, and write the results to BigQuery.

Answer: ?

Sample Questions

9. You have a Dataflow pipeline that processes website traffic logs stored in Cloud Storage and writes the processed data to BigQuery. You noticed that the pipeline is failing intermittently. You need to troubleshoot the issue. What should you do?

A. Use Cloud Logging to identify error groups in the pipeline's logs. Use Cloud Monitoring to create a dashboard that tracks the number of errors in each group.

B. Use Cloud Logging to create a chart displaying the pipeline's error logs. Use Metrics Explorer to validate the findings from the chart.

C. Use Cloud Logging to view error messages in the pipeline's logs. Use Cloud Monitoring to analyze the pipeline's metrics, such as CPU utilization and memory usage.

D. Use the Dataflow job monitoring interface to check the pipeline's status every hour. Use Cloud Profiler to analyze the pipeline's metrics, such as CPU utilization and memory usage.

Answer: ?

Sample Questions

8. Your organization's business analysts require near real-time access to streaming dat
a. However, they are reporting that their dashboard queries are loading slowly. After investigating BigQuery query performance, you discover the slow dashboard queries perform several joins and aggregations.
You need to improve the dashboard loading time and ensure that the dashboard data is as up-to-date as possible. What should you do?
A. Disable BiqQuery query result caching.
B. Modify the schema to use parameterized data types.
C. Create a scheduled query to calculate and store intermediate results.
D. Create materialized views.
Answer: ?

Sample Questions

7. You need to create a data pipeline that streams event information from applications in multiple Google Cloud regions into BigQuery for near real-time analysis. The data requires transformation before loading. You want to create the pipeline using a visual interface. What should you do?
A. Push event information to a Pub/Sub topic. Create a Dataflow job using the Dataflow job builder.
B. Push event information to a Pub/Sub topic. Create a Cloud Run function to subscribe to the Pub/Sub topic, apply transformations, and insert the data into BigQuery.
C. Push event information to a Pub/Sub topic. Create a BigQuery subscription in Pub/Sub.
D. Push event information to Cloud Storage, and create an external table in BigQuery. Create a BigQuery scheduled job that executes once each day to apply transformations.
Answer: ?

Sample Questions

6. You work for a home insurance company. You are frequently asked to create and save risk reports with charts for specific areas using a publicly available storm event dataset. You want to be able to quickly create and re-run risk reports when new data becomes available. What should you do?

A. Export the storm event dataset as a CSV file. Import the file to Google Sheets, and use cell data in the worksheets to create charts.

B. Copy the storm event dataset into your BigQuery project. Use BigQuery Studio to query and visualize the data in Looker Studio.

C. Reference and query the storm event dataset using SQL in BigQuery Studio. Export the results to Google Sheets, and use cell data in the worksheets to create charts.

D. Reference and query the storm event dataset using SQL in a Colab Enterprise notebook. Display the table results and document with Markdown, and use Matplotlib to create charts.

Answer: ?

Sample Questions

5. Your organization has a BigQuery dataset that contains sensitive employee information such as salaries and performance reviews. The payroll specialist in the HR department needs to have continuous access to aggregated performance data, but they do not need continuous access to other sensitive dat a. You need to grant the payroll specialist access to the performance data without granting them access to the entire dataset using the simplest and most secure approach. What should you do?

A. Use authorized views to share query results with the payroll specialist.

B. Create row-level and column-level permissions and policies on the table that contains performance data in the dataset. Provide the payroll specialist with the appropriate permission set.

C. Create a table with the aggregated performance data. Use table-level permissions to grant access to the payroll specialist.

D. Create a SQL query with the aggregated performance data. Export the results to an Avro file in a Cloud Storage bucket. Share the bucket with the payroll specialist.

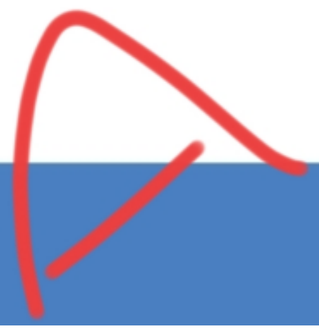
Answer: ?

Sample Questions

4. You recently inherited a task for managing Dataflow streaming pipelines in your organization and noticed that proper access had not been provisioned to you. You need to request a Googleprovided IAM role so you can restart the pipelines. You need to follow the principle of least privilege.
What should you do?

A. Request the Dataflow Developer role.
B. Request the Dataflow Viewer role.
C. Request the Dataflow Worker role.
D. Request the Dataflow Admin role.

Answer: ?



Sample Questions

3. Your organization uses a BigQuery table that is partitioned by ingestion time. You need to remove data that is older than one year to reduce your organization's storage costs. You want to use the most efficient approach while minimizing cost. What should you do?

A. Create a scheduled query that periodically runs an update statement in SQL that sets the "deleted" column to "yes" for data that is more than one year old. Create a view that filters out rows that have been marked deleted.

B. Create a view that filters out rows that are older than one year.

C. Require users to specify a partition filter using the alter table statement in SQL.

D. Set the table partition expiration period to one year using the ALTER TABLE statement in SQL.

Answer: ?

D

Sample Questions

2. You are a database administrator managing sales transaction data by region stored in a BigQuery table. You need to ensure that each sales representative can only see the transactions in their region. What should you do?

A. Add a policy tag in BigQuery.

B. Create a row-level access policy.

C. Create a data masking rule.

D. Grant the appropriate 1AM permissions on the dataset.

Answer: ?

Sample Questions

1. Your organization stores highly personal data in BigQuery and needs to comply with strict data privacy regulations. You need to ensure that sensitive data values are rendered unreadable whenever an employee leaves the organization. What should you do?

A. Use AEAD functions and delete keys when employees leave the organization.

B. Use dynamic data masking and revoke viewer permissions when employees leave the organization.

C. Use customer-managed encryption keys (CMEK) and delete keys when employees leave the organization.

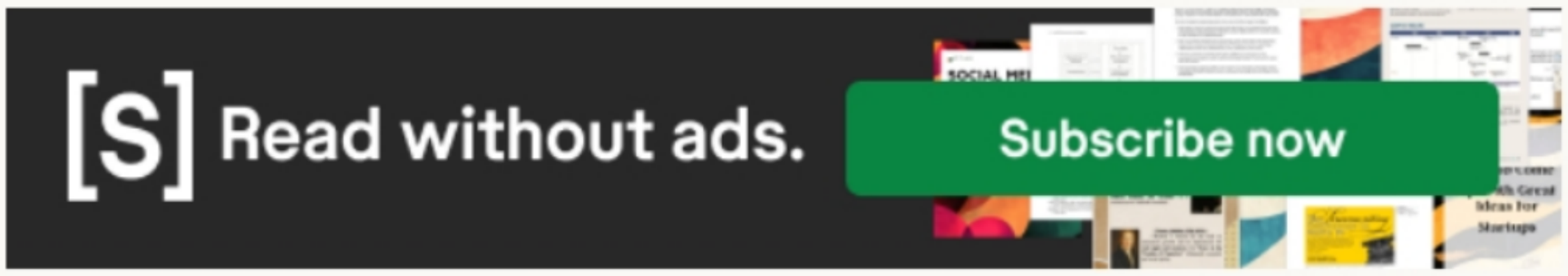
D. Use column-level access controls with policy tags and revoke viewer permissions when employees leave the organization.

Answer: ?

19.Your team needs to analyze large datasets stored in BigQuery to identify trends in user behavior. The analysis will involve complex statistical calculations, Python packages, and visualizations. You need to recommend a managed collaborative environment to develop and share the analysis.
What should you recommend?
A. Create a Colab Enterprise notebook and connect the notebook to BigQuery. Share the notebook with your team. Analyze the data and generate visualizations in Colab Enterprise.

Ad

Download to read ad-free



B. Create a statistical model by using BigQuery ML. Share the query with your team. Analyze the data and generate visualizations in Looker Studio.
C. Create a Looker Studio dashboard and connect the dashboard to BigQuery. Share the dashboard with your team. Analyze the data and generate visualizations in Looker Studio.
D. Connect Google Sheets to BigQuery by using Connected Sheets. Share the Google Sheet with your
team. Analyze the data and generate visualizations in Gooqle Sheets.
Answer: A
Explanation:
Using a Colab Enterprise notebook connected to BigQuery provides a managed, collaborative environment ideal for complex statistical calculations, Python packages, and visualizations. Colab Enterprise supports Python libraries for advanced analytics and offers seamless integration with BigQuery for querying large datasets. It allows teams to collaboratively develop and share analyses while taking advantage of its visualization capabilities. This approach is particularly suitable for tasks involving sophisticated computations and custom visualizations.

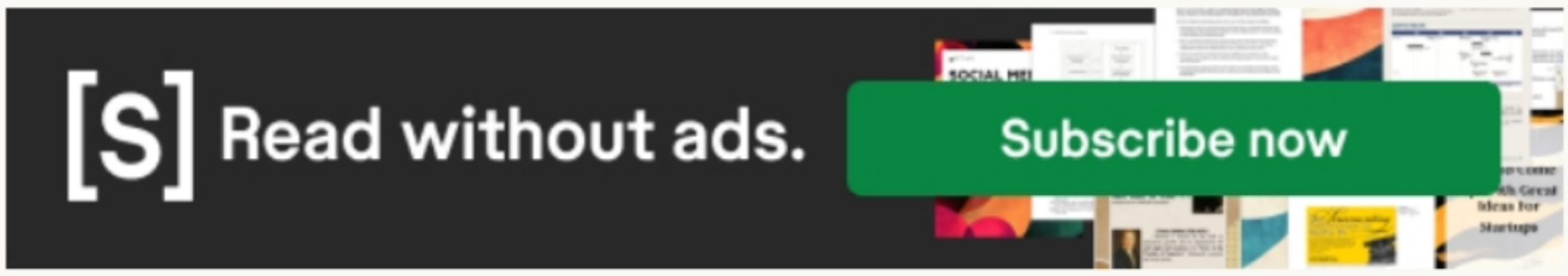
20. BigQuery

21.You are designing an application that will interact with several BigQuery datasets. You need to grant the application's service account permissions that allow it to query and update tables within the datasets, and list all datasets in a project within your application. You want to follow the principle of least privilege.
Which pre-defined IAM role(s) should you apply to the service account?
A. roles/bigquery.jobUser and roles/bigquery.dataOwner
B. roles/bigquery.connectionUser and roles/bigquery.dataViewer
C. roles/bigquery.admin
D. roles/bigquery.user and roles/bigquery.filteredDataViewer
Answer: A
Explanation:

17.You work for a financial organization that stores transaction data in BigQuery. Your organization has a regulatory requirement to retain data for a minimum of seven years for auditing purposes. You

Ad

Download to read ad-free



need to ensure that the data is retained for seven years using an efficient and cost-optimized approach.
What should you do?
A. Create a partition by transaction date, and set the partition expiration policy to seven years.
B. Set the table-level retention policy in BigQuery to seven years.
C. Set the dataset-level retention policy in BigQuery to seven years.
D. Export the BigQuery tables to Cloud Storage daily, and enforce a lifecycle management policy that has a seven-year retention rule.
Answer: B
Explanation:
Setting a table-level retention policy in BigQuery to seven years is the most efficient and cost-optimized solution to meet the regulatory requirement. A table-level retention policy ensures that the data cannot be deleted or overwritten before the specified retention period expires, providing compliance with auditing requirements while keeping the data within BigQuery for easy access and analysis. This approach avoids the complexity and additional costs of exporting data to Cloud Storage.

18.You want to build a model to predict the likelihood of a customer clicking on an online advertisement. You have historical data in BigQuery that includes features such as user demographics, ad placement, and previous click behavior. After training the model, you want to generate predictions on new data.
Which model type should you use in BigQuery ML?
A. Linear regression
B. Matrix factorization
C. Logistic regression
D. K-means clustering
Answer: C
Explanation:

15.Your organization consists of two hundred employees on five different teams. The leadership team is concerned that any employee can move or delete all Looker dashboards saved in the Shared folder. You need to create an easy-to-manage solution that allows the five different teams in your organization to view content in the Shared folder, but only be able to move or delete their team-specific dashboard.
What should you do?
A. 1. Create Looker groups representing each of the five different teams, and add users to their corresponding group.

16.Another team in your organization is requesting access to a BigQuery dataset. You need to share the dataset with the team while minimizing the risk of unauthorized copying of data. You also want to create a reusable framework in case you need to share this data with other teams in the future.
What should you do?
A. Create authorized views in the team's Google Cloud project that is only accessible by the team.
B. Create a private exchange using Analytics Hub with data egress restriction, and grant access to the team members.
C. Enable domain restricted sharing on the project. Grant the team members the BigQuery Data Viewer IAM role on the dataset.
D. Export the dataset to a Cloud Storage bucket in the team's Google Cloud project that is only accessible by the team.
Answer: B
Explanation:
Using Analytics Hub to create a private exchange with data egress restrictions ensures controlled sharing of the dataset while minimizing the risk of unauthorized copying. This approach allows you to provide secure, managed access to the dataset without giving direct access to the raw data. The egress restriction ensures that data cannot be exported or copied outside the designated boundaries. Additionally, this solution provides a reusable framework that simplifies future data sharing with other teams or projects while maintaining strict data governance.
Extract from Google Documentation: From "Analytics Hub Overview" (https://cloud.google.com/analytics-hub/docs): "Analytics Hub enables secure, controlled data sharing with private exchanges. Combine with organization policies like restrictDataEgress to prevent data copying, providing a reusable framework for sharing BigQuery datasets across teams."
Reference: Google Cloud Documentation - "Analytics Hub" (https://cloud.google.com/analytics-hub).

D. Use the historical sales data to train and create a BigQuery ML logistic regression model. Use the ML.PREDICT function call to output the predictions into a new BigQuery table.
Answer: A
Explanation:
Comprehensive and Detailed In-Depth
Forecasting product demand with seasonality requires a time series model, and BigQuery ML offers a scalable, serverless solution. Let's analyze:
Option A: BigQuery ML's time series models (e.g., ARIMA_PLUS) are designed for forecasting with seasonality and trends. The ML.FORECAST function generates predictions based on historical data, storing them in a table. This is scalable (no infrastructure) and integrates natively with BigQuery, ideal for ecommerce demand prediction.
Option B: Colab Enterprise with a custom Python model (e.g., Prophet) is flexible but requires coding, maintenance, and potentially exporting data, reducing scalability compared to BigQuery ML's in-place processing.
Option C: Linear regression predicts continuous values but doesn't handle seasonality or time series patterns effectively, making it unsuitable for demand forecasting.
Option D: Logistic regression is for binary classification (e.g., yes/no), not time series forecasting of demand quantities.
Why A is Best: ARIMA_PLUS in BigQuery ML automatically models seasonality and trends, requiring only SQL knowledge. It's serverless, scales with BigQuery's capacity, and keeps data in one place, minimizing complexity and cost. For example, CREATE MODEL ... OPTIONS(model_type='ARIMA_PLUS') followed by ML.FORECAST delivers accurate, scalable forecasts.
Extract from Google Documentation: From "BigQuery ML Time Series Forecasting" (https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-syntax-create-time-series): "The ARIMA_PLUS model type in BigQuery ML is designed for time series forecasting, accounting for seasonality and trends, making it ideal for predicting future values like product demand based on historical data."
Reference: Google Cloud Documentation - "BigQuery ML Time Series" (https://cloud.google.com/bigquery-ml/docs/time-series).

14.You need to design a data pipeline to process large volumes of raw server log data stored in Cloud Storage. The data needs to be cleaned, transformed, and aggregated before being loaded into BigQuery for analysis. The transformation involves complex data manipulation using Spark scripts that your team developed. You need to implement a solution that leverages your team's existing skillset, processes data at scale, and minimizes cost.
What should you do?
A. Use Dataflow with a custom template for the transformation logic.
B. Use Cloud Data Fusion to visually design and manage the pipeline.
C. Use Dataform to define the transformations in SQLX.
D. Use Dataproc to run the transformations on a cluster.
Answer: D
Explanation:
Comprehensive and Detailed In-Depth
The pipeline must handle large-scale log processing with existing Spark scripts, prioritizing skillset reuse, scalability, and cost. Let's break it down:
Option A: Dataflow uses Apache Beam, not Spark, requiring script rewrites (losing skillset leverage). Custom templates scale well but increase development cost and effort.
Option B: Cloud Data Fusion is a visual ETL tool, not Spark-based. It doesn't reuse existing scripts, requiring redesign, and is less cost-efficient for complex, code-driven transformations.
Option C: Dataform uses SQLX for BigQuery ELT, not Spark. It's unsuitable for pre-load

10. Bigtable
B. 1. Filestore

11.Your organization has decided to move their on-premises Apache Spark-based workload to Google Cloud. You want to be able to manage the code without needing to provision and manage your own cluster.
What should you do?
A. Migrate the Spark jobs to Dataproc Serverless.
B. Configure a Google Kubernetes Engine cluster with Spark operators, and deploy the Spark jobs.
C. Migrate the Spark jobs to Dataproc on Google Kubernetes Engine.
D. Migrate the Spark jobs to Dataproc on Compute Engine.
Answer: A
Explanation:
Migrating the Spark jobs to Dataproc Serverless is the best approach because it allows you to run Spark workloads without the need to provision or manage clusters. Dataproc Serverless automatically scales resources based on workload requirements, simplifying operations and reducing administrative overhead. This solution is ideal for organizations that want to focus on managing their Spark code without worrying about the underlying infrastructure. It is cost-effective and fully managed, aligning well with the goal of minimizing cluster management.

12.You used BigQuery ML to build a customer purchase propensity model six months ago. You want to compare the current serving data with the historical serving data to determine whether you need to retrain the model.
What should you do?
A. Compare the two different models.
B. Evaluate the data skewness.
C. Evaluate data drift.
D. Compare the confusion matrix.
Answer: C
Explanation:
Evaluating data drift involves analyzing changes in the distribution of the current serving data compared to the historical data used to train the model. If significant drift is detected, it indicates that the data patterns have changed over time, which can impact the model's performance. This analysis helps determine whether retraining the model is necessary to ensure its predictions remain accurate and relevant. Data drift evaluation is a standard approach for monitoring machine learning models over time.

13.You manage an ecommerce website that has a diverse range of products. You need to forecast future product demand accurately to ensure that your company has sufficient inventory to meet customer needs and avoid stockouts. Your company's historical sales data is stored in a BigQuery table. You need to create a scalable solution that takes into account the seasonality and historical data to predict product demand.
What should you do?
A. Use the historical sales data to train and create a BigQuery ML time series model. Use the ML.FORECAST function call to output the predictions into a new BigQuery table.
B. Use Colab Enterprise to create a Jupyter notebook. Use the historical sales data to train a custom prediction model in Python.
C. Use the historical sales data to train and create a BigQuery ML linear regression model. Use the ML.PREDICT function call to output the predictions into a new BigQuery table.

What should you do?
A. Set up a Cloud Monitoring dashboard to track key Dataflow metrics, such as data throughput, error rates, and resource utilization.
B. Create a custom script to periodically poll the Dataflow API for job status updates, and send email alerts if any errors are identified.
C. Navigate to the Dataflow Jobs page in the Google Cloud console. Use the job logs and worker logs to identify the error.
D. Use the gcloud CLI tool to retrieve job metrics and logs, and analyze them for errors and performance bottlenecks.
Answer: C
Explanation:
To troubleshoot a failed Dataflow job as quickly as possible, you should navigate to the Dataflow Jobs page in the Google Cloud console. The console provides access to detailed job logs and worker logs, which can help you identify the cause of the failure. The graphical interface also allows you to visualize pipeline stages, monitor performance metrics, and pinpoint where the error occurred, making it the most efficient way to diagnose and resolve the issue promptly.
Extract from Google Documentation: From "Monitoring Dataflow Jobs" (https://cloud.google.com/dataflow/docs/guides/monitoring-jobs): "To troubleshoot a failed Dataflow job quickly, go to the Dataflow Jobs page in the Google Cloud Console, where you can view job logs and worker logs to identify errors and their root causes."
Reference: Google Cloud Documentation - "Dataflow Monitoring" (https://cloud.google.com/dataflow/docs/guides/monitoring-jobs).

8. Social media activity: CloudSQL for MySQL
D. 1. Online transactions: Cloud SQL for MySQL

9.You are building a batch data pipeline to process 100 GB of structured data from multiple sources for daily reporting. You need to transform and standardize the data prior to loading the data to ensure that it is stored in a single dataset. You want to use a low-code solution that can be easily built and managed.
What should you do?
A. Use Cloud Data Fusion to ingest data and load the data into BigQuery. Use Looker Studio to perform data cleaning and transformation.
B. Use Cloud Data Fusion to ingest the data, perform data cleaning and transformation, and load the data into BigQuery.
C. Use Cloud Data Fusion to ingest the data, perform data cleaning and transformation, and load the data into Cloud SQL for PostgreSQL.
D. Use Cloud Storage to store the data. Use Cloud Run functions to perform data cleaning and transformation, and load the data into BigQuery.
Answer: B
Explanation:
Comprehensive and Detailed in Depth
Why B is correct:Cloud Data Fusion is a fully managed, cloud-native data integration service for building and managing ETL/ELT data pipelines.
It provides a graphical interface for building pipelines without coding, making it a low-code solution.
Cloud data fusion is perfect for the ingestion, transformation and loading of data into BigQuery.
Why other options are incorrect:A: Looker studio is for visualization, not data transformation.
C: Cloud SQL is a relational database, not ideal for large-scale analytical data.
D: Cloud run is for stateless applications, not batch data processing.
Reference: Cloud Data Fusion: https://cloud.google.com/data-fusion/docs

[S] Read without ads. Subscribe now

5.Your retail company wants to analyze customer reviews to understand sentiment and identify areas for improvement. Your company has a large dataset of customer feedback text stored in BigQuery that includes diverse language patterns, emojis, and slang. You want to build a solution to classify customer sentiment from the feedback text.
What should you do?
A. Preprocess the text data in BigQuery using SQL functions. Export the processed data to AutoML Natural Language for model training and deployment.
B. Export the raw data from BigQuery. Use AutoML Natural Language to train a custom sentiment analysis model.
C. Use Dataproc to create a Spark cluster, perform text preprocessing using Spark NLP, and build a sentiment analysis model with Spark MLlib.
D. Develop a custom sentiment analysis model using TensorFlow. Deploy it on a Compute Engine instance.
Answer: B
Explanation:
Comprehensive and Detailed in Depth
Why B is correct:AutoML Natural Language is designed for text classification tasks, including sentiment analysis, and can handle diverse language patterns without extensive preprocessing.
AutoML can train a custom model with minimal coding.
Why other options are incorrect:A: Unnecessary extra preprocessing. AutoML can handle the raw data.
C: Dataproc and Spark are overkill for this task. AutoML is more efficient and easier to use.
D: Developing a custom TensorFlow model requires significant expertise and time, which is not efficient for this scenario.
Reference: AutoML Natural Language: https://cloud.google.com/natural-language/automl/docs

6.Following a recent company acquisition, you inherited an on-premises data infrastructure that needs to move to Google Cloud. The acquired system has 250 Apache Airflow directed acyclic graphs (DAGs) orchestrating data pipelines. You need to migrate the pipelines to a Google Cloud managed service with minimal effort.
What should you do?
A. Convert each DAG to a Cloud Workflow and automate the execution with Cloud Scheduler.
B. Create a new Cloud Composer environment and copy DAGs to the Cloud Composer dags/ folder.
C. Create a Google Kubernetes Engine (GKE) standard cluster and deploy Airflow as a workload. Migrate all DAGs to the new Airflow environment.
D. Create a Cloud Data Fusion instance. For each DAG, create a Cloud Data Fusion pipeline.
Answer: B
Explanation:
Comprehensive and Detailed in Depth
Why B is correct:Cloud Composer is a managed Apache Airflow service that provides a seamless migration path for existing Airflow DAGs.
Simply copying the DAGs to the Cloud Composer folder allows them to run directly on Google Cloud.
Why other options are incorrect:A: Cloud Workflows is a different orchestration tool, not compatible with Airflow DAGs.
C: GKE deployment requires setting up and managing a Kubernetes cluster, which is more complex.
D: Cloud Data Fusion is a data integration tool, not suitable for orchestrating existing pipelines.
Reference: Cloud Composer: https://cloud.google.com/composer/docs

7.Your organization uses Dataflow pipelines to process real-time financial transactions. You discover that one of your Dataflow jobs has failed. You need to troubleshoot the issue as quickly as possible.

your data relationships, allowing users to filter and drill into details like individual transactions directly within a dashboard."
Reference: Looker Documentation - "Explores and Dashboards" (https://cloud.google.com/looker/docs).

3.Your company's ecommerce website collects product reviews from customers. The reviews are loaded as CSV files daily to a Cloud Storage bucket. The reviews are in multiple languages and need to be translated to Spanish. You need to configure a pipeline that is serverless, efficient, and requires minimal maintenance.
What should you do?
A. Load the data into BigQuery using Dataproc. Use Apache Spark to translate the reviews by invoking the Cloud Translation API. Set BigQuery as the sink.U
B. Use a Dataflow templates pipeline to translate the reviews using the Cloud Translation API. Set BigQuery as the sink.
C. Load the data into BigQuery using a Cloud Run function. Use the BigQuery ML create model statement to train a translation model. Use the model to translate the product reviews within BigQuery.
D. Load the data into BigQuery using a Cloud Run function. Create a BigQuery remote function that invokes the Cloud Translation API. Use a scheduled query to translate new reviews.
Answer: D
Explanation:
Loading the data into BigQuery using a Cloud Run function and creating a BigQuery remote function that invokes the Cloud Translation API is a serverless and efficient approach. With this setup, you can use a scheduled query in BigQuery to invoke the remote function and translate new product reviews on a regular basis. This solution requires minimal maintenance, as BigQuery handles storage and querying, and the Cloud Translation API provides accurate translations without the need for custom ML model development.

4.Your company is building a near real-time streaming pipeline to process JSON telemetry data from small appliances. You need to process messages arriving at a Pub/Sub topic, capitalize letters in the serial number field, and write results to BigQuery. You want to use a managed service and write a minimal amount of code for underlying transformations.
What should you do?
A. Use a Pub/Sub to BigQuery subscription, write results directly to BigQuery, and schedule a transformation query to run every five minutes.
B. Use a Pub/Sub to Cloud Storage subscription, write a Cloud Run service that is triggered when objects arrive in the bucket, performs the transformations, and writes the results to BigQuery.
C. Use the "Pub/Sub to BigQuery" Dataflow template with a UDF, and write the results to BigQuery.
D. Use a Pub/Sub push subscription, write a Cloud Run service that accepts the messages, performs the transformations, and writes the results to BigQuery.
Answer: C
Explanation:
Using the "Pub/Sub to BigQuery" Dataflow template with a UDF (User-Defined Function) is the optimal choice because it combines near real-time processing, minimal code for transformations, and scalability. The UDF allows for efficient implementation of custom transformations, such as capitalizing letters in the serial number field, while Dataflow handles the rest of the managed pipeline seamlessly.

2.Your company uses Looker to visualize and analyze sales data. You need to create a dashboard that displays sales metrics, such as sales by region, product category, and time period. Each metric relies on its own set of attributes distributed across several tables. You need to provide users the ability to filter the data by specific sales representatives and view individual transactions. You want to follow the Google-recommended approach.
What should you do?
A. Create multiple Explores, each focusing on each sales metric. Link the Explores together in a dashboard using drill-down functionality.
B. Use BigQuery to create multiple materialized views, each focusing on a specific sales metric. Build the dashboard using these views.
C. Create a single Explore with all sales metrics. Build the dashboard using this Explore.
D. Use Looker's custom visualization capabilities to create a single visualization that displays all the sales metrics with filtering and drill-down functionality.
Answer: C
Explanation:
Creating a single Explore with all the sales metrics is the Google-recommended approach. This Explore should be designed to include all relevant attributes and dimensions, enabling users to analyze sales data by region, product category, time period, and other filters like sales representatives. With a well-structured Explore, you can efficiently build a dashboard that supports filtering and drill-down functionality. This approach simplifies maintenance, provides a consistent data model, and ensures users have the flexibility to interact with and analyze the data seamlessly within a unified framework.
Looker's recommended approach for dashboards is a single, unified Explore for scalability and usability, supporting filters and drill-downs.
Option A: Materialized views in BigQuery optimize queries but bypass Looker's modeling layer, reducing flexibility.
Option B: Custom visualizations are for specific rendering, not multi-metric dashboards with filtering/drill-down.
Option C: Multiple Explores fragment the data model, complicating dashboard cohesion and maintenance.
Option D: A single Explore joins tables in LookML, enabling all metrics, filters (e.g., sales rep), and drill-downs to transactions, per Looker's best practices.
Extract from Google Documentation: From "Building Dashboards in Looker"

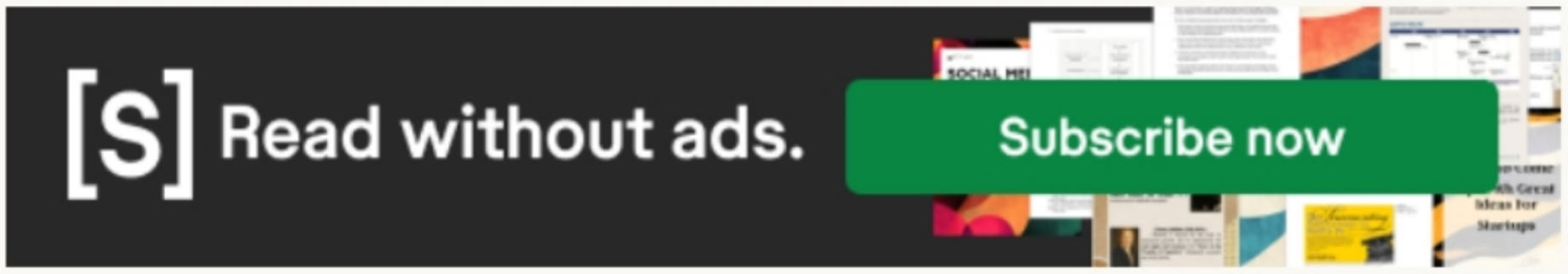
200-301 Implementing and Administering Cisco Solutions

Share some Associate Data Practitioner exam online questions below.

1.Your organization needs to implement near real-time analytics for thousands of events arriving each second in Pub/Sub. The incoming messages require transformations. You need to configure a pipeline that processes, transforms, and loads the data into BigQuery while minimizing development time.
What should you do?

Ad

Download to read ad-free



A. Use a Google-provided Dataflow template to process the Pub/Sub messages, perform transformations, and write the results to BigQuery.
B. Create a Cloud Data Fusion instance and configure Pub/Sub as a source. Use Data Fusion to process the Pub/Sub messages, perform transformations, and write the results to BigQuery.
C. Load the data from Pub/Sub into Cloud Storage using a Cloud Storage subscription. Create a Dataproc cluster, use PySpark to perform transformations in Cloud Storage, and write the results to BigQuery.
D. Use Cloud Run functions to process the Pub/Sub messages, perform transformations, and write the results to BigQuery.
Answer: A
Explanation:
Using a Google-provided Dataflow template is the most efficient and development-friendly approach to implement near real-time analytics for Pub/Sub messages. Dataflow templates are pre-built and optimized for processing streaming data, allowing you to quickly configure and deploy a pipeline with minimal development effort. These templates can handle message ingestion from Pub/Sub, perform necessary transformations, and load the processed data into BigQuery, ensuring scalability and low latency for near real-time analytics.

5. You need to ingest small CSV files into BigQuery from Cloud Storage as soon as they arrive, minimizing data latency while keeping costs and maintenance effort low. What is the best solution?

A. Use the bq command-line tool within a Cloud Shell instance to load the data into BigQuery.
B. Set up a Cloud Composer workflow to scan for new files and load them into BigQuery every 10 minutes.
C. Deploy a Cloud Run function that automatically triggers and loads the data into BigQuery when new files arrive.
D. Configure a Dataproc cluster to pull CSV files from Cloud Storage, process them with Spark, and write the results to BigQuery.

Event-Driven Data Ingestion with Cloud Run

Cloud Run enables serverless event-driven processing, where functions are triggered automatically when new files arrive in Cloud Storage. This ensures low-latency ingestion into BigQuery while minimizing operational overhead. It is a cost-effective and scalable solution for real-time data processing compared to scheduled batch jobs or manual ingestion processes.

4. You are designing a data processing pipeline that handles sensitive customer data stored in a Cloud Storage bucket. To ensure availability in the event of a single-zone failure, what should you do?

A. Enable Cloud CDN for the Cloud Storage bucket.
B. Turn on Object Versioning for the bucket.
C. Store the data in a multi-region bucket.
D. Move the data to Nearline storage for redundancy.

Cloud Storage Multi-Region Buckets and Disaster Recovery

Multi-region buckets in Google Cloud Storage distribute data across multiple geographical locations, ensuring redundancy and availability even in the event of a single-zone failure. This approach enhances disaster recovery, minimizes downtime, and is recommended for business-critical data that requires high availability.

3. Your organization stores highly sensitive personal data in BigQuery and must comply with strict data privacy regulations. You need to ensure that sensitive data values become unreadable when an employee leaves the organization. What is the best way to achieve this?

A. Encrypt the data using AEAD functions and delete encryption keys when employees leave.
B. Apply dynamic data masking and revoke the employee's access permissions.
C. Use customer-managed encryption keys (CMEK) and delete the keys upon employee departure.
D. Implement column-level security using policy tags and remove the employee's access permissions.

Customer-Managed Encryption Keys (CMEK) and Data Security in BigQuery 40 seconds

Customer-managed encryption keys (CMEK) provide full control over encryption and decryption processes. Organizations can encrypt sensitive data using CMEK and delete the keys when access should be revoked, rendering the data unreadable. This approach is crucial for regulatory compliance (e.g., GDPR, HIPAA) and ensures that former employees cannot access sensitive information after leaving.

2. Your organization's business analysts rely on real-time streaming data for decision-making, but they have reported that their dashboard queries are running slowly. After analyzing BigQuery query performance, you find that the queries involve complex joins and aggregations. What is the best way to improve dashboard performance while keeping the data as up-to-date as possible?

A. Disable query result caching in BigQuery.
B. Optimize the schema by using parameterized data types.
C. Schedule periodic queries to precompute and store intermediate results.
D. Create materialized views to store aggregated query results.

BigQuery Materialized Views and Query Optimization

Materialized views in BigQuery store precomputed query results, reducing query complexity and improving performance. They automatically refresh as new data arrives, making them ideal for real-time dashboards that involve expensive aggregations and joins. This approach minimizes query execution time and cost while keeping data as fresh as possible.

1. A database administrator needs to restrict access to sales transaction data stored in a BigQuery table so that each sales representative can only view transactions from their assigned region. What is the best approach to achieve this?

A. Apply a policy tag to the dataset in BigQuery.
B. Implement a row-level security policy.
C. Configure a data masking rule on the table.
D. Assign IAM permissions based on user roles at the dataset level.

Row-Level Security (RLS) in BigQuery

Row-level security (RLS) allows fine-grained access control by restricting access to specific rows in a table based on user attributes. It is implemented using Authorized Views or Row Access Policies, ensuring that users can only see data relevant to them. Unlike IAM permissions that work at the dataset or table level, RLS ensures that different users can query the same table but only access data they are authorized to see.

Associate Data Practitioner Google Cloud Associate Data Practitioner (ADP Exam) exam dumps questions are the best material for you to test all the related Google exam topics. By using the Associate Data Practitioner exam dumps questions and practicing your skills, you can increase your confidence and chances of passing the Associate Data Practitioner exam.

Features of Dumpsinfo's products

- Instant Download
- Free Update in 3 Months
- Money back guarantee
- PDF and Software
- 24/7 Customer Support

Besides, Dumpsinfo also provides unlimited access. You can get all Dumpsinfo files at lowest price.

Google Cloud Associate Data Practitioner (ADP Exam) Associate Data Practitioner exam free dumps questions are available below for you to study.

Full version: *Associate Data Practitioner Exam Dumps Questions*

1.Your company's customer support audio files are stored in a Cloud Storage bucket. You plan to analyze the audio files' metadata and file content within BigQuery to create inference by using BigQuery ML. You need to create a corresponding table in BigQuery that represents the bucket containing the audio files.
What should you do?
A. Create an external table.
B. Create a temporary table.
C. Create a native table.
D. Create an object table.
Answer: D
Explanation:

To analyze audio files stored in a Cloud Storage bucket and represent them in BigQuery, you should create an object table. Object tables in BigQuery are designed to represent objects stored in Cloud Storage, including their metadata. This enables you to query the metadata of audio files directly from BigQuery without duplicating the data. Once the object table is created, you can use it in conjunction with other BigQuery ML workflows for inference and analysis.

2.Your organization has decided to migrate their existing enterprise data warehouse to BigQuery. The existing data pipeline tools already support connectors to BigQuery. You need to identify a data migration approach that optimizes migration speed.
What should you do?
A. Create a temporary file system to facilitate data transfer from the existing environment to Cloud Storage. Use Storage Transfer Service to migrate the data into BigQuery.
B. Use the Cloud Data Fusion web interface to build data pipelines. Create a directed acyclic graph (DAG) that facilitates pipeline orchestration.
C. Use the existing data pipeline tool's BigQuery connector to reconfigure the data mapping.
D. Use the BigQuery Data Transfer Service to recreate the data pipeline and migrate the data into BigQuery.
Answer: C
Explanation:
Since your existing data pipeline tools already support connectors to BigQuery, the most efficient approach is to use the existing data pipeline tool's BigQuery connector to reconfigure the data mapping. This leverages your current tools, reducing migration complexity and setup time, while optimizing migration speed. By reconfiguring the data mapping within the existing pipeline, you can seamlessly direct the data into BigQuery without needing additional services or intermediary steps.

3.Your company uses Looker to generate and share reports with various stakeholders. You have a complex dashboard with several visualizations that needs to be delivered to specific stakeholders on a recurring basis, with customized filters applied for each recipient. You need an efficient and scalable solution to automate the delivery of this customized dashboard. You want to follow the Google-recommended approach.
What should you do?
A. Create a separate LookML model for each stakeholder with predefined filters, and schedule the dashboards using the Looker Scheduler.
B. Create a script using the Looker Python SDK, and configure user attribute filter values. Generate a new scheduled plan for each stakeholder.
C. Embed the Looker dashboard in a custom web application, and use the application's scheduling features to send the report with personalized filters.
D. Use the Looker Scheduler with a user attribute filter on the dashboard, and send the dashboard with personalized filters to each stakeholder based on their attributes.
Answer: D
Explanation:
Using the Looker Scheduler with user attribute filters is the Google-recommended approach to efficiently automate the delivery of a customized dashboard. User attribute filters allow you to dynamically customize the dashboard's content based on the recipient's attributes, ensuring each stakeholder sees data relevant to them. This approach is scalable, does not require creating separate models or custom scripts, and leverages Looker's built-in functionality to automate recurring deliveries effectively.

4.Your company has several retail locations. Your company tracks the total number of sales made at

each location each day. You want to use SQL to calculate the weekly moving average of sales by location to identify trends for each store.
Which query should you use?
A)

```
SELECT store_id, date, total_sales, AVG(total_sales) OVER (
PARTITION BY store_id
ORDER BY total_sales RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

B)

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY date ORDER BY store_id ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

C)

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY store_id
ORDER BY date ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

D)

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY total_sales
ORDER BY date RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

A. Option A
B. Option B
C. Option C
D. Option D
Answer: C
Explanation:
To calculate the weekly moving average of sales by location:
The query must group by store_id (partitioning the calculation by each store).
The ORDER BY date ensures the sales are evaluated chronologically.
The ROWS BETWEEN 6 PRECEDING AND CURRENT ROW specifies a rolling window of 7 rows (1 week if each row represents daily data).
The AVG(total_sales) computes the average sales over the defined rolling window.
Chosen query meets these requirements:
- PARTITION BY store_id groups the calculation by each store.
- ORDER BY date orders the rows correctly for the rolling average.

- ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ensures the 7-day moving average.

5.You are using your own data to demonstrate the capabilities of BigQuery to your organization's leadership team. You need to perform a one- time load of the files stored on your local machine into BigQuery using as little effort as possible.
What should you do?
A. Write and execute a Python script using the BigQuery Storage Write API library.
B. Create a Dataproc cluster, copy the files to Cloud Storage, and write an Apache Spark job using the spark-bigquery-connector.
C. Execute the bq load command on your local machine.
D. Create a Dataflow job using the Apache Beam FileIO and BigQueryIO connectors with a local runner.
Answer: C
Explanation:
Using the bq load command is the simplest and most efficient way to perform a one-time load of files from your local machine into BigQuery. This command-line tool is easy to use, requires minimal setup, and supports direct uploads from local files to BigQuery tables. It meets the requirement for minimal effort while allowing you to quickly demonstrate BigQuery's capabilities to your organization's leadership team.

6. Cloud SQL for MySQL
C. 1. BigQuery

7. Cloud SQL for PostgreSQL

8.You manage a large amount of data in Cloud Storage, including raw data, processed data, and backups. Your organization is subject to strict compliance regulations that mandate data immutability for specific data types. You want to use an efficient process to reduce storage costs while ensuring that your storage strategy meets retention requirements.
What should you do?
A. Configure lifecycle management rules to transition objects to appropriate storage classes based on access patterns. Set up Object Versioning for all objects to meet immutability requirements.
B. Move objects to different storage classes based on their age and access patterns. Use Cloud Key Management Service (Cloud KMS) to encrypt specific objects with customer-managed encryption keys (CMEK) to meet immutability requirements.
C. Create a Cloud Run function to periodically check object metadata, and move objects to the appropriate storage class based on age and access patterns. Use object holds to enforce immutability for specific objects.
D. Use object holds to enforce immutability for specific objects, and configure lifecycle management rules to transition objects to appropriate storage classes based on age and access patterns.
Answer: D
Explanation:
Using object holds and lifecycle management rules is the most efficient and compliant strategy for this scenario because:
Immutability: Object holds (temporary or event-based) ensure that objects cannot be deleted or overwritten, meeting strict compliance regulations for data immutability.
Cost efficiency: Lifecycle management rules automatically transition objects to more cost-effective storage classes based on their age and access patterns.
Compliance and automation: This approach ensures compliance with retention requirements while reducing manual effort, leveraging built-in Cloud Storage features.

9.Your organization plans to move their on-premises environment to Google Cloud. Your organization's network bandwidth is less than 1 Gbps. You need to move over 500 ?? of data to Cloud Storage securely, and only have a few days to move the data.
What should you do?
A. Request multiple Transfer Appliances, copy the data to the appliances, and ship the appliances back to Google Cloud to upload the data to Cloud Storage.
B. Connect to Google Cloud using VPN. Use Storage Transfer Service to move the data to Cloud Storage.
C. Connect to Google Cloud using VPN. Use the gcloud storage command to move the data to Cloud Storage.
D. Connect to Google Cloud using Dedicated Interconnect. Use the gcloud storage command to move the data to Cloud Storage.
Answer: A
Explanation:
Using Transfer Appliances is the best solution for securely and efficiently moving over 500 TB of data to Cloud Storage within a limited timeframe, especially with network bandwidth below 1 Gbps. Transfer Appliances are physical devices provided by Google Cloud to securely transfer large amounts of data. After copying the data to the appliances, they are shipped back to Google, where the data is uploaded to Cloud Storage. This approach bypasses bandwidth limitations and ensures the data is migrated quickly and securely.

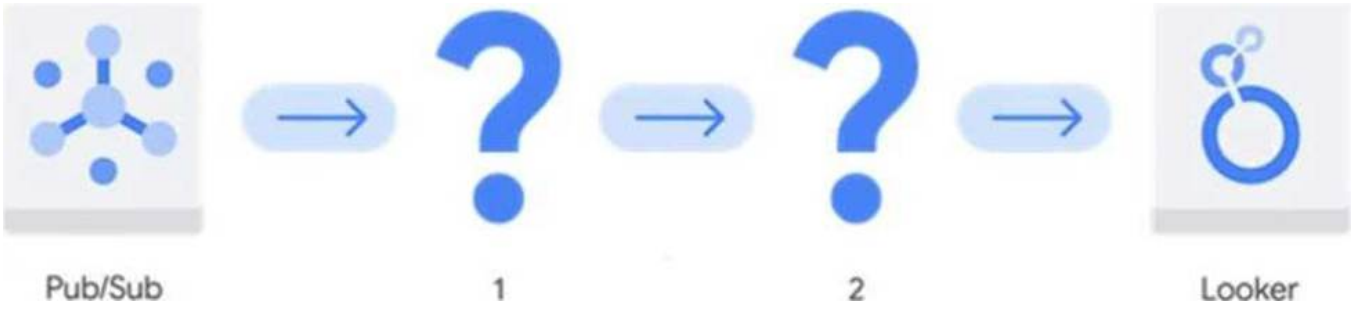
10.Your organization has decided to move their on-premises Apache Spark-based workload to Google Cloud. You want to be able to manage the code without needing to provision and manage your own cluster.
What should you do?
A. Migrate the Spark jobs to Dataproc Serverless.
B. Configure a Google Kubernetes Engine cluster with Spark operators, and deploy the Spark jobs.
C. Migrate the Spark jobs to Dataproc on Google Kubernetes Engine.
D. Migrate the Spark jobs to Dataproc on Compute Engine.
Answer: A
Explanation:
Migrating the Spark jobs to Dataproc Serverless is the best approach because it allows you to run Spark workloads without the need to provision or manage clusters. Dataproc Serverless automatically scales resources based on workload requirements, simplifying operations and reducing administrative overhead. This solution is ideal for organizations that want to focus on managing their Spark code without worrying about the underlying infrastructure. It is cost-effective and fully managed, aligning well with the goal of minimizing cluster management.

11. Audio files from phone interactions with support agents that will be accessed during trainings.

12.Your team wants to create a monthly report to analyze inventory data that is updated daily. You need to aggregate the inventory counts by using only the most recent month of data, and save the results to be used in a Looker Studio dashboard.
What should you do?
A. Create a materialized view in BigQuery that uses the SUM( ) function and the DATE_SUB( ) function.
B. Create a saved query in the BigQuery console that uses the SUM( ) function and the DATE_SUB( ) function. Re-run the saved query every month, and save the results to a BigQuery table.

C. Create a BigQuery table that uses the SUM( ) function and the _PARTITIONDATE filter.
D. Create a BigQuery table that uses the SUM( ) function and the DATE_DIFF( ) function.
Answer: A
Explanation:
Creating a materialized view in BigQuery with the SUM() function and the DATE_SUB() function is the best approach. Materialized views allow you to pre-aggregate and cache query results, making them efficient for repeated access, such as monthly reporting. By using the DATE_SUB() function, you can filter the inventory data to include only the most recent month. This approach ensures that the aggregation is up-to-date with minimal latency and provides efficient integration with Looker Studio for dashboarding.

13.You need to create a new data pipeline.
You want a serverless solution that meets the following
requirements:
• Data is streamed from Pub/Sub and is processed in real-time.
• Data is transformed before being stored.
• Data is stored in a location that will allow it to be analyzed with SQL using Looker.



Which Google Cloud services should you recommend for the pipeline?
A. 1. Dataproc Serverless

14.You are migrating data from a legacy on-premises MySQL database to Google Cloud. The database contains various tables with different data types and sizes, including large tables with millions of rows and transactional data. You need to migrate this data while maintaining data integrity, and minimizing downtime and cost.
What should you do?
A. Set up a Cloud Composer environment to orchestrate a custom data pipeline. Use a Python script to extract data from the MySQL database and load it to MySQL on Compute Engine.
B. Export the MySQL database to CSV files, transfer the files to Cloud Storage by using Storage Transfer Service, and load the files into a Cloud SQL for MySQL instance.
C. Use Database Migration Service to replicate the MySQL database to a Cloud SQL for MySQL instance.
D. Use Cloud Data Fusion to migrate the MySQL database to MySQL on Compute Engine.
Answer: C
Explanation:
Using Database Migration Service (DMS) to replicate the MySQL database to a Cloud SQL for MySQL instance is the best approach. DMS is a fully managed service designed for migrating databases to Google Cloud with minimal downtime and cost. It supports continuous data replication, ensuring data integrity during the migration process, and handles schema and data transfer efficiently. This solution is particularly suited for large tables and transactional data, as it maintains real-time synchronization between the source and target databases, minimizing downtime for the

migration.

15.You are developing a data ingestion pipeline to load small CSV files into BigQuery from Cloud Storage. You want to load these files upon arrival to minimize data latency. You want to accomplish this with minimal cost and maintenance.
What should you do?
A. Use the bq command-line tool within a Cloud Shell instance to load the data into BigQuery.
B. Create a Cloud Composer pipeline to load new files from Cloud Storage to BigQuery and schedule it to run every 10 minutes.
C. Create a Cloud Run function to load the data into BigQuery that is triggered when data arrives in Cloud Storage.
D. Create a Dataproc cluster to pull CSV files from Cloud Storage, process them using Spark, and write the results to BigQuery.
Answer: C
Explanation:
Using a Cloud Run function triggered by Cloud Storage to load the data into BigQuery is the best solution because it minimizes both cost and maintenance while providing low-latency data ingestion. Cloud Run is a serverless platform that automatically scales based on the workload, ensuring efficient use of resources without requiring a dedicated instance or cluster. It integrates seamlessly with Cloud Storage event notifications, enabling real-time processing of incoming files and loading them into BigQuery. This approach is cost-effective, scalable, and easy to manage.

Associate Data Practitioner Google Cloud Associate Data Practitioner (ADP Exam) exam dumps questions are the best material for you to test all the related Google exam topics. By using the Associate Data Practitioner exam dumps questions and practicing your skills, you can increase your confidence and chances of passing the Associate Data Practitioner exam.

Features of Dumpsinfo's products

- Instant Download
- Free Update in 3 Months
- Money back guarantee
- PDF and Software
- 24/7 Customer Support

Besides, Dumpsinfo also provides unlimited access. You can get all Dumpsinfo files at lowest price.

Google Cloud Associate Data Practitioner (ADP Exam) Associate Data Practitioner exam free dumps questions are available below for you to study.

Full version: *Associate Data Practitioner Exam Dumps Questions*

1. Customer feedback: Cloud Storage

2.Another team in your organization is requesting access to a BigQuery dataset. You need to share the dataset with the team while minimizing the risk of unauthorized copying of data. You also want to create a reusable framework in case you need to share this data with other teams in the future. What should you do?
A. Create authorized views in the team's Google Cloud project that is only accessible by the team.
B. Create a private exchange using Analytics Hub with data egress restriction, and grant access to the team members.
C. Enable domain restricted sharing on the project. Grant the team members the BigQuery Data Viewer IAM role on the dataset.

D. Export the dataset to a Cloud Storage bucket in the team’s Google Cloud project that is only accessible by the team.
Answer: B
Explanation:
Using Analytics Hub to create a private exchange with data egress restrictions ensures controlled sharing of the dataset while minimizing the risk of unauthorized copying. This approach allows you to provide secure, managed access to the dataset without giving direct access to the raw data. The egress restriction ensures that data cannot be exported or copied outside the designated boundaries. Additionally, this solution provides a reusable framework that simplifies future data sharing with other teams or projects while maintaining strict data governance.
Extract from Google Documentation: From "Analytics Hub Overview" (https://cloud.google.com/analytics-hub/docs): "Analytics Hub enables secure, controlled data sharing with private exchanges. Combine with organization policies like restrictDataEgress to prevent data copying, providing a reusable framework for sharing BigQuery datasets across teams."
Reference: Google Cloud Documentation - "Analytics Hub" (https://cloud.google.com/analytics-hub).

3.You are constructing a data pipeline to process sensitive customer data stored in a Cloud Storage bucket. You need to ensure that this data remains accessible, even in the event of a single-zone outage.
What should you do?
A. Set up a Cloud CDN in front of the bucket.
B. Enable Object Versioning on the bucket.
C. Store the data in a multi-region bucket.
D. Store the data in Nearline storaqe.
Answer: C
Explanation:
Storing the data in a multi-region bucket ensures high availability and durability, even in the event of a single-zone outage. Multi-region buckets replicate data across multiple locations within the selected region, providing resilience against zone-level failures and ensuring that the data remains accessible. This approach is particularly suitable for sensitive customer data that must remain available without interruptions.
A single-zone outage requires high availability across zones or regions.
Cloud Storage offers location-based redundancy options:
Option A: Cloud CDN caches content for web delivery but doesn’t protect against underlying storage outages?it’s for performance, not availability of the source data.
Option B: Object Versioning retains old versions of objects, protecting against overwrites or deletions, but doesn’t ensure availability during a zone failure (still tied to one location).
Option C: Multi-region buckets (e.g., us or eu) replicate data across multiple regions, ensuring accessibility even if a single zone or region fails. This provides the highest availability for sensitive data in a pipeline.
Option D: Nearline storage is a low-cost class with lower availability (single-region by default), not designed for outage resilience.
Why C is Best: Multi-region storage offers geo-redundancy, guaranteeing data access during outages with no additional configuration. It’s the simplest, most reliable choice for pipeline continuity.
Extract from Google Documentation: From "Cloud Storage Bucket Locations" (https://cloud.google.com/storage/docs/locations): "Multi-region buckets replicate data across multiple regions within a geography, providing high availability and durability, ensuring access even if a single zone or region experiences an outage."
Reference: Google Cloud Documentation - "Cloud Storage Locations" (https://cloud.google.com/storage/docs/locations).

4.You need to design a data pipeline that ingests data from CSV, Avro, and Parquet files into Cloud Storage. The data includes raw user input. You need to remove all malicious SQL injections before storing the data in BigQuery.
Which data manipulation methodology should you choose?
A. EL
B. ELT
C. ETL
D. ETLT
Answer: C
Explanation:
The ETL (Extract, Transform, Load) methodology is the best approach for this scenario because it allows you to extract data from the files, transform it by applying the necessary data cleansing (including removing malicious SQL injections), and then load the sanitized data into BigQuery. By transforming the data before loading it into BigQuery, you ensure that only clean and safe data is stored, which is critical for security and data quality.

5.You have a Cloud SQL for PostgreSQL database that stores sensitive historical financial data. You need to ensure that the data is uncorrupted and recoverable in the event that the primary region is destroyed. The data is valuable, so you need to prioritize recovery point objective (RPO) over recovery time objective (RTO). You want to recommend a solution that minimizes latency for primary read and write operations.
What should you do?
A. Configure the Cloud SQL for PostgreSQL instance for multi-region backup locations.
B. Configure the Cloud SQL for PostgreSQL instance for regional availability (HA). Back up the Cloud SQL for PostgreSQL database hourly to a Cloud Storage bucket in a different region.
C. Configure the Cloud SQL for PostgreSQL instance for regional availability (HA) with synchronous replication to a secondary instance in a different zone.
D. Configure the Cloud SQL for PostgreSQL instance for regional availability (HA) with asynchronous replication to a secondary instance in a different region.
Answer: A
Explanation:
Comprehensive and Detailed In-Depth
The priorities are data integrity, recoverability after a regional disaster, low RPO (minimal data loss), and low latency for primary operations. Let’s analyze:
Option A: Multi-region backups store point-in-time snapshots in a separate region. With automated backups and transaction logs, RPO can be near-zero (e.g., minutes), and recovery is possible post-disaster. Primary operations remain in one zone, minimizing latency.
Option B: Regional HA (failover to another zone) with hourly cross-region backups protects against zone failures, but hourly backups yield an RPO of up to 1 hour?too high for valuable data. Manual backup management adds overhead.
Option C: Synchronous replication to another zone ensures zero RPO within a region but doesn’t protect against regional loss. Latency increases slightly due to sync writes across zones.
Option D: Asynchronous replication to another region reduces RPO (e.g., seconds) but increases primary write latency (waiting for async confirmation) and doesn’t guarantee zero data loss in a crash.
Why A is Best: Multi-region backups balance low RPO (continuous logs + frequent backups) with no latency impact on primary operations (single-zone instance). For example, enabling backups to us-east1 from a us-west1 instance ensures recoverability without affecting performance, prioritizing RPO as requested.
Extract from Google Documentation: From "Cloud SQL Backup and Recovery"

(https://cloud.google.com/sql/docs/postgres/backup-recovery): "Configure multi-region backups to store data in a different region, ensuring recoverability after a regional failure with minimal RPO using automated backups and transaction logs, while keeping primary operations low-latency."
Reference: Google Cloud Documentation - "Cloud SQL Backups" (https://cloud.google.com/sql/docs/postgres/backup-recovery).

6.You need to create a data pipeline for a new application. Your application will stream data that needs to be enriched and cleaned. Eventually, the data will be used to train machine learning models. You need to determine the appropriate data manipulation methodology and which Google Cloud services to use in this pipeline.
What should you choose?
A. ETL; Dataflow -> BigQuery
B. ETL; Cloud Data Fusion -> Cloud Storage
C. ELT; Cloud Storage -> Bigtable
D. ELT; Cloud SQL -> Analytics Hub
Answer: A
Explanation:
Comprehensive and Detailed In-Depth
Streaming data requiring enrichment and cleaning before ML training suggests an ETL (Extract, Transform, Load) approach, with a focus on real-time processing and a data warehouse for ML.
Option A: ETL with Dataflow (streaming transformations) and BigQuery (storage/ML training) is Google's recommended pattern for streaming pipelines. Dataflow handles enrichment/cleaning, and BigQuery supports ML model training (BigQuery ML).
Option B: ETL with Cloud Data Fusion to Cloud Storage is batch-oriented and lacks streaming focus. Cloud Storage isn't ideal for ML training directly.
Option C: ELT (load then transform) with Cloud Storage to Bigtable is misaligned?Bigtable is for NoSQL, not ML training or post-load transformation.
Option D: ELT with Cloud SQL to Analytics Hub is for relational data and data sharing, not streaming or ML.
Reference: Google Cloud Documentation - "Dataflow: ETL Patterns" (https://cloud.google.com/dataflow/docs/guides), "BigQuery ML" (https://cloud.google.com/bigquery-ml).

7.Your company has several retail locations. Your company tracks the total number of sales made at each location each day. You want to use SQL to calculate the weekly moving average of sales by location to identify trends for each store.
Which query should you use?
A)

```
SELECT store_id, date, total_sales, AVG(total_sales) OVER (
PARTITION BY store_id
ORDER BY total_sales RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

B)

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY date ORDER BY store_id ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

C)

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY store_id
ORDER BY date ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

D)

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY total_sales
ORDER BY date RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

A. Option A
B. Option B
C. Option C
D. Option D
Answer: C
Explanation:
To calculate the weekly moving average of sales by location:
The query must group by store_id (partitioning the calculation by each store).
The ORDER BY date ensures the sales are evaluated chronologically.
The ROWS BETWEEN 6 PRECEDING AND CURRENT ROW specifies a rolling window of 7 rows (1 week if each row represents daily data).
The AVG(total_sales) computes the average sales over the defined rolling window.
Chosen query meets these requirements:
- PARTITION BY store_id groups the calculation by each store.
- ORDER BY date orders the rows correctly for the rolling average.
- ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ensures the 7-day moving average.
Extract from Google Documentation: From "Analytic Functions in BigQuery" (https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts): "Use ROWS BETWEEN n PRECEDING AND CURRENT ROW with ORDER BY a time column to compute moving averages over a fixed number of rows, such as a 7-day window, partitioned by a grouping key like store_id."
Reference: Google Cloud Documentation - "BigQuery Window Functions" (https://cloud.google.com/bigquery/docs/reference/standard-sql/window-function-calls).

8.You are working on a data pipeline that will validate and clean incoming data before loading it into BigQuery for real-time analysis. You want to ensure that the data validation and cleaning is performed

efficiently and can handle high volumes of data.
What should you do?
A. Write custom scripts in Python to validate and clean the data outside of Google Cloud. Load the cleaned data into BigQuery.
B. Use Cloud Run functions to trigger data validation and cleaning routines when new data arrives in Cloud Storage.
C. Use Dataflow to create a streaming pipeline that includes validation and transformation steps.
D. Load the raw data into BigQuery using Cloud Storage as a staging area, and use SQL queries in BigQuery to validate and clean the data.
Answer: C
Explanation:
Using Dataflow to create a streaming pipeline that includes validation and transformation steps is the most efficient and scalable approach for real-time analysis. Dataflow is optimized for high-volume data processing and allows you to apply validation and cleaning logic as the data flows through the pipeline. This ensures that only clean, validated data is loaded into BigQuery, supporting real-time analysis while handling high data volumes effectively.

9.You are migrating data from a legacy on-premises MySQL database to Google Cloud. The database contains various tables with different data types and sizes, including large tables with millions of rows and transactional data. You need to migrate this data while maintaining data integrity, and minimizing downtime and cost.
What should you do?
A. Set up a Cloud Composer environment to orchestrate a custom data pipeline. Use a Python script to extract data from the MySQL database and load it to MySQL on Compute Engine.
B. Export the MySQL database to CSV files, transfer the files to Cloud Storage by using Storage Transfer Service, and load the files into a Cloud SQL for MySQL instance.
C. Use Database Migration Service to replicate the MySQL database to a Cloud SQL for MySQL instance.
D. Use Cloud Data Fusion to migrate the MySQL database to MySQL on Compute Engine.
Answer: C
Explanation:
Using Database Migration Service (DMS) to replicate the MySQL database to a Cloud SQL for MySQL instance is the best approach. DMS is a fully managed service designed for migrating databases to Google Cloud with minimal downtime and cost. It supports continuous data replication, ensuring data integrity during the migration process, and handles schema and data transfer efficiently. This solution is particularly suited for large tables and transactional data, as it maintains real-time synchronization between the source and target databases, minimizing downtime for the migration.

10.You manage an ecommerce website that has a diverse range of products. You need to forecast future product demand accurately to ensure that your company has sufficient inventory to meet customer needs and avoid stockouts. Your company's historical sales data is stored in a BigQuery table. You need to create a scalable solution that takes into account the seasonality and historical data to predict product demand.
What should you do?
A. Use the historical sales data to train and create a BigQuery ML time series model. Use the ML.FORECAST function call to output the predictions into a new BigQuery table.
B. Use Colab Enterprise to create a Jupyter notebook. Use the historical sales data to train a custom prediction model in Python.
C. Use the historical sales data to train and create a BigQuery ML linear regression model. Use the

ML.PREDICT function call to output the predictions into a new BigQuery table.
D. Use the historical sales data to train and create a BigQuery ML logistic regression model. Use the ML.PREDICT function call to output the predictions into a new BigQuery table.
Answer: A
Explanation:
Comprehensive and Detailed In-Depth
Forecasting product demand with seasonality requires a time series model, and BigQuery ML offers a scalable, serverless solution. Let's analyze:
Option A: BigQuery ML's time series models (e.g., ARIMA_PLUS) are designed for forecasting with seasonality and trends. The ML.FORECAST function generates predictions based on historical data, storing them in a table. This is scalable (no infrastructure) and integrates natively with BigQuery, ideal for ecommerce demand prediction.
Option B: Colab Enterprise with a custom Python model (e.g., Prophet) is flexible but requires coding, maintenance, and potentially exporting data, reducing scalability compared to BigQuery ML's in-place processing.
Option C: Linear regression predicts continuous values but doesn't handle seasonality or time series patterns effectively, making it unsuitable for demand forecasting.
Option D: Logistic regression is for binary classification (e.g., yes/no), not time series forecasting of demand quantities.
Why A is Best: ARIMA_PLUS in BigQuery ML automatically models seasonality and trends, requiring only SQL knowledge. It's serverless, scales with BigQuery's capacity, and keeps data in one place, minimizing complexity and cost. For example, CREATE MODEL ... OPTIONS(model_type='ARIMA_PLUS') followed by ML.FORECAST delivers accurate, scalable forecasts.
Extract from Google Documentation: From "BigQuery ML Time Series Forecasting" (https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-syntax-create-time-series): "The ARIMA_PLUS model type in BigQuery ML is designed for time series forecasting, accounting for seasonality and trends, making it ideal for predicting future values like product demand based on historical data."
Reference: Google Cloud Documentation - "BigQuery ML Time Series" (https://cloud.google.com/bigquery-ml/docs/time-series).

11.Your organization's website uses an on-premises MySQL as a backend database. You need to migrate the on-premises MySQL database to Google Cloud while maintaining MySQL features. You want to minimize administrative overhead and downtime.
What should you do?
A. Install MySQL on a Compute Engine virtual machine. Export the database files using the mysqldump command. Upload the files to Cloud Storage, and import them into the MySQL instance on Compute Engine.
B. Use Database Migration Service to transfer the data to Cloud SQL for MySQL, and configure the on premises MySQL database as the source.
C. Use a Google-provided Dataflow template to replicate the MySQL database in BigQuery.
D. Export the database tables to CSV files, and upload the files to Cloud Storage. Convert the MySQL schema to a Spanner schema, create a JSON manifest file, and run a Google-provided Dataflow template to load the data into Spanner.
Answer: B
Explanation:
Comprehensive and Detailed in Depth
Why B is correct:Database Migration Service (DMS) is designed for migrating databases to Cloud SQL with minimal downtime and administrative overhead.
Cloud SQL for MySQL is a fully managed MySQL service, which aligns with the requirement to

minimize administrative overhead.
Why other options are incorrect:A: Installing MySQL on Compute Engine requires manual management of the database instance, which increases administrative overhead.
C: BigQuery is not a direct replacement for a relational MySQL database. It's an analytical data warehouse.
D: Spanner is a globally distributed, scalable database, but it requires schema conversion and is not a direct replacement for MySQL, and it is also much more complex than cloud SQL.
Reference: Database Migration Service: https://cloud.google.com/database-migration
Cloud SQL for MySQL: https://cloud.google.com/sql/docs/mysql

12.You manage a Cloud Storage bucket that stores temporary files created during data processing. These temporary files are only needed for seven days, after which they are no longer needed. To reduce storage costs and keep your bucket organized, you want to automatically delete these files once they are older than seven days.
What should you do?
A. Set up a Cloud Scheduler job that invokes a weekly Cloud Run function to delete files older than seven days.
B. Configure a Cloud Storage lifecycle rule that automatically deletes objects older than seven days.
C. Develop a batch process using Dataflow that runs weekly and deletes files based on their age.
D. Create a Cloud Run function that runs daily and deletes files older than seven days.
Answer: B
Explanation:
Configuring a Cloud Storage lifecycle rule to automatically delete objects older than seven days is the best solution because:
Built-in feature: Cloud Storage lifecycle rules are specifically designed to manage object lifecycles, such as automatically deleting or transitioning objects based on age.
No additional setup: It requires no external services or custom code, reducing complexity and maintenance.
Cost-effective: It directly achieves the goal of deleting files after seven days without incurring additional compute costs.

13.Your organization consists of two hundred employees on five different teams. The leadership team is concerned that any employee can move or delete all Looker dashboards saved in the Shared folder. You need to create an easy-to-manage solution that allows the five different teams in your organization to view content in the Shared folder, but only be able to move or delete their team-specific dashboard.
What should you do?
A. 1. Create Looker groups representing each of the five different teams, and add users to their corresponding group.

14.Your company currently uses an on-premises network file system (NFS) and is migrating data to Google Cloud. You want to be able to control how much bandwidth is used by the data migration while capturing detailed reporting on the migration status.
What should you do?
A. Use a Transfer Appliance.
B. Use Cloud Storage FUSE.
C. Use Storage Transfer Service.
D. Use gcloud storage commands.
Answer: C
Explanation:

Using the Storage Transfer Service is the best solution for migrating data from an on-premises NFS to Google Cloud. This service allows you to control bandwidth usage by configuring transfer speed limits and provides detailed reporting on the migration status. Storage Transfer Service is specifically designed for large-scale data migrations and supports scheduling, monitoring, and error handling, making it an efficient and reliable choice for your use case.

15.You manage data at an ecommerce company. You have a Dataflow pipeline that processes order data from Pub/Sub, enriches the data with product information from Bigtable, and writes the processed data to BigQuery for analysis. The pipeline runs continuously and processes thousands of orders every minute. You need to monitor the pipeline's performance and be alerted if errors occur.
What should you do?
A. Use Cloud Monitoring to track key metrics. Create alerting policies in Cloud Monitoring to trigger notifications when metrics exceed thresholds or when errors occur.
B. Use the Dataflow job monitoring interface to visually inspect the pipeline graph, check for errors, and configure notifications when critical errors occur.
C. Use BigQuery to analyze the processed data in Cloud Storage and identify anomalies or inconsistencies. Set up scheduled alerts based when anomalies or inconsistencies occur.
D. Use Cloud Logging to view the pipeline logs and check for errors. Set up alerts based on specific keywords in the logs.
Answer: A
Explanation:
Comprehensive and Detailed in Depth
Why A is correct:Cloud Monitoring is the recommended service for monitoring Google Cloud services, including Dataflow.
It allows you to track key metrics like system lag, element throughput, and error rates.
Alerting policies in Cloud Monitoring can trigger notifications based on metric thresholds.
Why other options are incorrect:B: The Dataflow job monitoring interface is useful for visualization, but Cloud Monitoring provides more comprehensive alerting.
C: BigQuery is for analyzing the processed data, not monitoring the pipeline itself. Also Cloud Storage is not where the data resides during processing.
D: Cloud Logging is useful for viewing logs, but Cloud Monitoring is better for metric-based alerting.
Reference: Cloud Monitoring for Dataflow: https://cloud.google.com/dataflow/docs/guides/using-monitoring
Cloud Monitoring: https://cloud.google.com/monitoring/docs

16.You are responsible for managing Cloud Storage buckets for a research company. Your company has well-defined data tiering and retention rules. You need to optimize storage costs while achieving your data retention needs.
What should you do?
A. Configure the buckets to use the Archive storage class.
B. Configure a lifecycle management policy on each bucket to downgrade the storage class and remove objects based on age.
C. Configure the buckets to use the Standard storage class and enable Object Versioning.
D. Configure the buckets to use the Autoclass feature.
Answer: B
Explanation:
Configuring a lifecycle management policy on each Cloud Storage bucket allows you to automatically transition objects to lower-cost storage classes (such as Nearline, Coldline, or Archive) based on their age or other criteria. Additionally, the policy can automate the removal of objects once they are no longer needed, ensuring compliance with retention rules and optimizing storage costs. This approach

aligns well with well-defined data tiering and retention needs, providing cost efficiency and automation.
Extract from Google Documentation: From "Object Lifecycle Management" (https://cloud.google.com/storage/docs/lifecycle): "Use lifecycle management policies to automatically transition objects to lower-cost storage classes (e.g., Nearline, Coldline) and delete them based on age, optimizing costs according to your specific tiering and retention requirements."
Reference: Google Cloud Documentation - "Cloud Storage Lifecycle" (https://cloud.google.com/storage/docs/lifecycle).

17.You have created a LookML model and dashboard that shows daily sales metrics for five regional managers to use. You want to ensure that the regional managers can only see sales metrics specific to their region. You need an easy-to-implement solution.
What should you do?
A. Create a sales_region user attribute, and assign each manager's region as the value of their user attribute. Add an access_filter Explore filter on the region_name dimension by using the sales_region user attribute.
B. Create five different Explores with the sql_always_filter Explore filter applied on the region_name dimension. Set each region_name value to the corresponding region for each manager.
C. Create separate Looker dashboards for each regional manager. Set the default dashboard filter to the corresponding region for each manager.
D. Create separate Looker instances for each regional manager. Copy the LookML model and dashboard to each instance. Provision viewer access to the corresponding manager.
Answer: A
Explanation:
Using a sales_region user attribute is the best solution because it allows you to dynamically filter data based on each manager's assigned region. By adding an access_filter Explore filter on the region_name dimension that references the sales_region user attribute, each manager sees only the sales metrics specific to their region. This approach is easy to implement, scalable, and avoids duplicating dashboards or Explores, making it both efficient and maintainable.

18.You want to process and load a daily sales CSV file stored in Cloud Storage into BigQuery for downstream reporting. You need to quickly build a scalable data pipeline that transforms the data while providing insights into data quality issues.
What should you do?
A. Create a batch pipeline in Cloud Data Fusion by using a Cloud Storage source and a BigQuery sink.
B. Load the CSV file as a table in BigQuery, and use scheduled queries to run SQL transformation scripts.
C. Load the CSV file as a table in BigQuery. Create a batch pipeline in Cloud Data Fusion by using a BigQuery source and sink.
D. Create a batch pipeline in Dataflow by using the Cloud Storage CSV file to BigQuery batch template.
Answer: A
Explanation:
Using Cloud Data Fusion to create a batch pipeline with a Cloud Storage source and a BigQuery sink is the best solution because:
Scalability: Cloud Data Fusion is a scalable, fully managed data integration service.
Data transformation: It provides a visual interface to design pipelines, enabling quick transformation of data.
Data quality insights: Cloud Data Fusion includes built-in tools for monitoring and addressing data

quality issues during the pipeline creation and execution process.

19. Bigtable
B. 1. Filestore

20.Your company has an on-premises file server with 5 TB of data that needs to be migrated to Google Cloud. The network operations team has mandated that you can only use up to 250 Mbps of the total available bandwidth for the migration. You need to perform an online migration to Cloud Storage.
What should you do?
A. Use Storage Transfer Service to configure an agent-based transfer. Set the appropriate bandwidth limit for the agent pool.
B. Use the gcloud storage cp command to copy all files from on-premises to Cloud Storage using the - -daisy-chain option.
C. Request a Transfer Appliance, copy the data to the appliance, and ship it back to Google Cloud.
D. Use the gcloud storage cp command to copy all files from on-premises to Cloud Storage using the - -no-clobber option.
Answer: A
Explanation:
Comprehensive and Detailed in Depth
Why A is correct: Storage Transfer Service with agent-based transfer allows for online migrations and provides the ability to set bandwidth limits.
Agents are installed on-premises and can be configured to respect network constraints.
Why other options are incorrect: B: The --daisy-chain option is not related to bandwidth control.
C: Transfer Appliance is for offline migrations and is not suitable for online transfers with bandwidth constraints.
D: The --no-clobber option prevents overwriting existing files but does not control bandwidth.
Reference: Storage Transfer Service: https://cloud.google.com/storage-transfer-service/docs
Storage Transfer Service Agents: https://cloud.google.com/storage-transfer-service/docs/agent-overview
gcloud storage cp: https://cloud.google.com/storage/docs/gsutil/commands/cp

21. Create five subfolders inside the Shared folder. Grant each team member the Manage Access, Edit access level to their corresponding subfolder.
Answer: C
Explanation:
Comprehensive and Detailed in Depth
Why C is correct:Setting the Shared folder to "View" ensures everyone can see the content.
Creating Looker groups simplifies access management.
Subfolders allow granular permissions for each team.
Granting "Manage Access, Edit" allows teams to modify only their own content.
Why other options are incorrect:A: Grants View access only, so teams can't edit.
B: Moving content to personal folders defeats the purpose of sharing.
D: Grants edit access to all members of the team, not the team as a whole, which is not ideal.
Reference: Looker Access Control: https://cloud.google.com/looker/docs/access-control
Looker Groups: https://cloud.google.com/looker/docs/groups

From: Snapshot-Content-Location: https://www.itexams.com/exam/Associate-Data-Practitioner/2 Subject: Associate Data Practitioner Exam - Free Google Questions and Answers - ITExams.com Date: Thu, 19 Jun 2025 09:32:01 +0530 MIME-Version: 1.0 Content-Type: multipart/related; type="text/html"; boundary="----MultipartBoundary--Ik8bAKr07umiNImJUhDf5jljXWLHvKbn08acUL7hpk----" ------MultipartBoundary--Ik8bAKr07umiNImJUhDf5jljXWLHvKbn08acUL7hpk---- Content-Type: text/html Content-ID: Content-Transfer-Encoding: binary Content-Location: https://www.itexams.com/exam/Associate-Data-Practitioner/2

Google Associate Data Practitioner - Google Cloud Certified - Associate Data Practitioner Exam

Question #6 (Topic: Exam A)

You work for a healthcare company that has a large on-premises data system containing patient records with personally identifiable information (PII) such as names, addresses, and medical diagnoses. You need a standardized managed solution that de-identifies PII across all your data feeds prior to ingestion to Google Cloud. What should you do?

A. Use Cloud Run functions to create a serverless data cleaning pipeline. Store the cleaned data in BigQuery.

B. Use Cloud Data Fusion to transform the data. Store the cleaned data in BigQuery.

C. Load the data into BigQuery, and inspect the data by using SQL queries. Use Dataflow to transform the data and remove any errors.

D. Use Apache Beam to read the data and perform the necessary cleaning and transformation operations. Store the cleaned data in BigQuery.

Answer: **B**

Next Question

Question #7 (Topic: Exam A)

You manage a large amount of data in Cloud Storage, including raw data, processed data, and backups. Your organization is subject to strict compliance regulations that mandate data immutability for specific data types. You want to use an efficient process to reduce storage costs while ensuring that your storage strategy meets retention requirements. What should you do?

A. Configure lifecycle management rules to transition objects to appropriate storage classes based on access patterns. Set up Object Versioning for all objects to meet immutability requirements.

B. Move objects to different storage classes based on their age and access patterns. Use Cloud Key Management Service (Cloud KMS) to encrypt specific objects with customer-managed encryption keys (CMEK) to meet immutability requirements.

C. Create a Cloud Run function to periodically check object metadata, and move objects to the appropriate storage class based on age and access patterns. Use object holds to enforce immutability for specific objects.

D. Use object holds to enforce immutability for specific objects, and configure lifecycle management rules to transition objects to appropriate storage classes based on age and access patterns.

Answer: **D**

Next Question

Question #8 (Topic: Exam A)

You work for an ecommerce company that has a BigQuery dataset that contains customer purchase history, demographics, and website interactions. You need to build a machine learning (ML) model to predict which customers are most likely to make a purchase in the next month. You have limited engineering resources and need to minimize the ML expertise required for the solution. What should you do?

A. Use BigQuery ML to create a logistic regression model for purchase prediction.

B. Use Vertex AI Workbench to develop a custom model for purchase prediction.

C. Use Colab Enterprise to develop a custom model for purchase prediction.

D. Export the data to Cloud Storage, and use AutoML Tables to build a classification model for purchase prediction.

Answer: **A**

Next Question

Question #9 (Topic: Exam A)

You are designing a pipeline to process data files that arrive in Cloud Storage by 3:00 am each day. Data processing is performed in stages, where the output of one stage becomes the input of the next. Each stage takes a long time to run. Occasionally a stage fails, and you have to address
the problem. You need to ensure that the final output is generated as quickly as possible. What should you do?

A. Design a Spark program that runs under Dataproc. Code the program to wait for user input when an error is detected. Rerun the last action after correcting any stage output data errors.

B. Design the pipeline as a set of PTransforms in Dataflow. Restart the pipeline after correcting any stage output data errors.

C. Design the workflow as a Cloud Workflow instance. Code the workflow to jump to a given stage based on an input parameter. Rerun the workflow after correcting any stage output data errors.

D. Design the processing as a directed acyclic graph (DAG) in Cloud Composer. Clear the state of the failed task after correcting any stage output data errors.

Answer: **D**

Next Question

Question #10 (Topic: Exam A)

Another team in your organization is requesting access to a BigQuery dataset. You need to share the dataset with the team while minimizing the risk of unauthorized copying of data. You also want to create a reusable framework in case you need to share this data with other teams in the future. What should you do?

A. Create authorized views in the team's Google Cloud project that is only accessible by the team.

B. Create a private exchange using Analytics Hub with data egress restriction, and grant access to the team members.

C. Enable domain restricted sharing on the project. Grant the team members the BigQuery Data Viewer IAM role on the dataset.

D. Export the dataset to a Cloud Storage bucket in the team's Google Cloud project that is only accessible by the team.

Answer: **B**

CertyIQ

CertyIQ

Premium exam material

Get certification quickly with the CertyIQ Premium exam material.
Everything you need to prepare, learn & pass your certification exam easily. Lifetime free updates
First attempt guaranteed success.

https://www.CertyIQ.com

About CertyIQ

We here at CertyIQ eventually got enough of the industry's greedy exam paid for. Our team of IT professionals comes with years of experience in the IT industry Prior to training CertIQ we worked in test areas where we observed the horrors of the paywall exam preparation system.

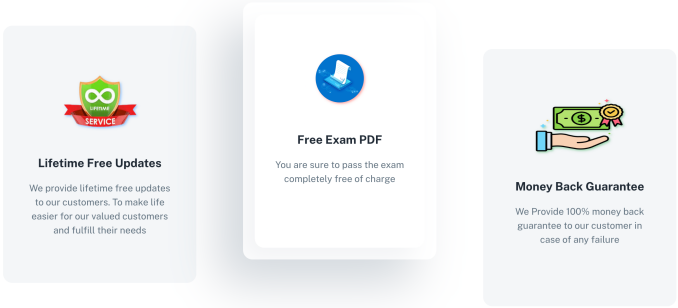
The misuse of the preparation system has left our team disillusioned. And for that reason, we decided it was time to make a difference. We had to make In this way, CertyIQ was created to provide quality materials without stealing from everyday people who are trying to make a living.

Doubt Support

We have developed a very scalable solution using which we are able to solve 400+ doubts every single day with an average rating of 4.8 out of 5.

https://www.certyiq.com

Mail us on - certyiqofficial@gmail.com



John

October 19, 2022

★★★★★

Thanks you so much for your help. I scored 972 in my exam today. More than 90% were from your PDFs!

Dana

September 04, 2022

★★★★★

Thanks a lot for this updated AZ-900 Q&A. I just passed my exam and got 974, I followed both of your Az-900 videos and the 6 PDF, the PDFs are very much valid, all answers are correct. Could you please create a similar video/PDF for DP900, your content/PDF's is really awesome. The team did a really good job. Thank You 😊.

Ahamed Shibly

2 months ago

★★★★★

Customer support is realy fast and helpful, I just finished my exam and this video along with the 6 PDF helped me pass! Definitely recommend getting the PDFs. Thank you!

October 22, 2022

★★★★★

Passed my exam today with 891 marks. Out of 52 questions, 51 were from certyiq PDFs including Contoso case study.
Thank You certyiq team!

Henry Rome

2 months ago

★★★★★

These questions are real and 100 % valid. Thank you so much for your efforts, also your 4 PDFs are awesome, I passed the DP900 exam on 1 Sept. With 968 marks. Thanks a lot, buddy!

Esmaria

2 months ago

★★★★★

Simple easy to understand explanations. To anyone out there wanting to write AZ900, I highly recommend 6 PDF's.Thank you so much, appreciate all your hard work in having such great content. Passed my exam Today - 3 September with 942 score.

Google

(Associate Data Practitioner)

Google Cloud Certified - Associate Data Practitioner

Total: **72 Questions**

Link: https://certyiq.com/papers/google/associate-data-practitioner

Question: 1 CertyIQ

Your retail company wants to predict customer churn using historical purchase data stored in BigQuery. The dataset includes customer demographics, purchase history, and a label indicating whether the customer churned or not. You want to build a machine learning model to identify customers at risk of churning. You need to create and train a logistic regression model for predicting customer churn, using the customer_data table with the churned column as the target label. Which BigQuery ML query should you use?

A.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT *
FROM customer_data;
```

B.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT * EXCEPT(churned),
       churned AS label
FROM customer_data;
```

C.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT * EXCEPT(churned)
FROM customer_data;
```

D.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT churned as label
FROM customer_data;
```

Answer: B

Explanation:

B.Option B is best because it correctly selects all columns as features except the churned column, and then explicitly designates the churned column as the label for BigQuery ML, which is required for training. Option A (SELECT *) is incorrect because while it includes the churned column, it doesn't explicitly define it as the label. Option C (SELECT * EXCEPT(churned)) is incorrect because it excludes the target churned column, which is essential for training a supervised model. Option D (SELECT churned as label) is incorrect because it only selects the label and excludes all feature columns needed for prediction. Therefore, Option B is the only query that correctly sets up the data for BigQuery ML logistic regression training.

Question: 2 CertyIQ

Your company has several retail locations. Your company tracks the total number of sales made at each location each day. You want to use SQL to calculate the weekly moving average of sales by location to identify trends for each store. Which query should you use?

A.

```
SELECT store_id, date, total_sales, AVG(total_sales) OVER (
PARTITION BY store_id
ORDER BY total_sales RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

B.

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY date ORDER BY store_id ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

C.

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY store_id
ORDER BY date ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

D.

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY total_sales
ORDER BY date RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

Answer: C

Explanation:

C. Option C is best because it correctly partitions by store_id to calculate the moving average for each store individually, and orders by date to ensure the moving average is calculated chronologically over the preceding 7 days (including the current day, thus 6 preceding). ROWS BETWEEN 6 PRECEDING AND CURRENT ROW then correctly defines the 7-day window for the average. Option A is incorrect because it orders by total_sales

instead of date, making the rolling average based on sales value order, not time, which is illogical for a weekly trend. Option B is incorrect because it partitions by date instead of store_id, calculating a daily moving average across all stores, not per store. Option D is incorrect because it partitions by total_sales, which is nonsensical for analyzing trends by location, and orders by date within that illogical partition. Therefore, Option C is the only query that correctly calculates a weekly moving average of sales by store location.

Question: 3 CertyIQ

Your company is building a near real-time streaming pipeline to process JSON telemetry data from small appliances. You need to process messages arriving at a Pub/Sub topic, capitalize letters in the serial number field, and write results to BigQuery. You want to use a managed service and write a minimal amount of code for underlying transformations. What should you do?

A.Use a Pub/Sub to BigQuery subscription, write results directly to BigQuery, and schedule a transformation query to run every five minutes.

B.Use a Pub/Sub to Cloud Storage subscription, write a Cloud Run service that is triggered when objects arrive in the bucket, performs the transformations, and writes the results to BigQuery.

C.Use the “Pub/Sub to BigQuery” Dataflow template with a UDF, and write the results to BigQuery.

D.Use a Pub/Sub push subscription, write a Cloud Run service that accepts the messages, performs the transformations, and writes the results to BigQuery.

Answer: C

Explanation:

Here's a detailed justification for why option C is the most appropriate solution, along with supporting explanations and links:

The core requirement is to process a stream of JSON data, perform a transformation (capitalizing the serial number), and load the results into BigQuery, minimizing code and utilizing managed services.

Option A: Pub/Sub to BigQuery subscription with scheduled query: While Pub/Sub can directly load to BigQuery, it doesn't offer transformation capabilities. A scheduled query would introduce latency and isn't a true near real-time solution. Scheduled queries also add operational overhead.

Option B: Pub/Sub to Cloud Storage with Cloud Run: This approach introduces an intermediary Cloud Storage bucket, adding unnecessary complexity. Cloud Run requires managing containers and scaling, increasing operational overhead compared to Dataflow. The transformation logic needs to be written and maintained in Cloud Run.

Option C: "Pub/Sub to BigQuery" Dataflow template with a UDF: Dataflow is a fully managed, scalable data processing service ideal for streaming data. The "Pub/Sub to BigQuery" template specifically addresses this use case. User-Defined Functions (UDFs) in Dataflow allow you to write custom transformation logic (capitalizing the serial number in this case) without writing extensive code. This offers a balance between ease of use and customizability. Dataflow handles the scaling and infrastructure management, reducing operational burden.

Option D: Pub/Sub push subscription with Cloud Run: Similar to option B, this requires managing a Cloud Run service. While it processes messages directly, it necessitates more code for handling Pub/Sub messages, transformation logic, and writing to BigQuery compared to the Dataflow template.

Why Dataflow is the superior choice:

Dataflow shines because it's a fully managed, scalable, and cost-effective service specifically designed for data processing pipelines. Using the template and a UDF allows you to achieve near real-time processing with

minimal code and operational overhead. This aligns perfectly with the problem statement's requirements.

Authoritative Links:

Dataflow: https://cloud.google.com/dataflow
Pub/Sub to BigQuery Template: https://cloud.google.com/dataflow/docs/templates/provided-streaming#pubsub-to-bigquery
Dataflow UDFs: https://cloud.google.com/dataflow/docs/guides/templates/using-udfs

Question: 4 CertyIQ

You want to process and load a daily sales CSV file stored in Cloud Storage into BigQuery for downstream reporting. You need to quickly build a scalable data pipeline that transforms the data while providing insights into data quality issues. What should you do?

A.Create a batch pipeline in Cloud Data Fusion by using a Cloud Storage source and a BigQuery sink.

B.Load the CSV file as a table in BigQuery, and use scheduled queries to run SQL transformation scripts.

C.Load the CSV file as a table in BigQuery. Create a batch pipeline in Cloud Data Fusion by using a BigQuery source and sink.

D.Create a batch pipeline in Dataflow by using the Cloud Storage CSV file to BigQuery batch template.

Answer: A

Explanation:

The recommended approach is to use Cloud Data Fusion to create a batch pipeline with a Cloud Storage source and a BigQuery sink.

Here's why:

Scalability: Cloud Data Fusion is designed for building scalable data pipelines without needing to manage infrastructure. It can handle large daily CSV files efficiently.
Transformation Capabilities: Data Fusion provides a visual interface with various transformation plugins to clean, transform, and enrich the data as it moves from Cloud Storage to BigQuery.
Data Quality Insights: Data Fusion allows you to integrate data quality checks within the pipeline, enabling you to identify and handle data quality issues during the transformation process.
Ease of Use: Data Fusion's visual interface simplifies pipeline development compared to writing custom code with Dataflow or complex SQL queries.
Speed of Development: The visual nature of Cloud Data Fusion helps you build the pipeline rapidly.

Options B and C have drawbacks:

Option B: Loading into BigQuery and using scheduled queries for transformation limits the ability to handle complex transformation logic and incorporate data quality checks within a visual pipeline. It's more suitable for simpler transformations.
Option C: Using BigQuery as a source in Cloud Data Fusion after loading offers minimal benefits. The file is already in Cloud Storage, which is a more natural starting point.

Option D is also suitable, but Cloud Data Fusion is favored because it is easier to visualize the data and includes data quality checks.

Authoritative Links:

Cloud Data Fusion: https://cloud.google.com/data-fusion
BigQuery: https://cloud.google.com/bigquery

Cloud Storage: https://cloud.google.com/storage

Question: 5 CertyIQ

You manage a Cloud Storage bucket that stores temporary files created during data processing. These temporary files are only needed for seven days, after which they are no longer needed. To reduce storage costs and keep your bucket organized, you want to automatically delete these files once they are older than seven days. What should you do?

A.Set up a Cloud Scheduler job that invokes a weekly Cloud Run function to delete files older than seven days.

B.Configure a Cloud Storage lifecycle rule that automatically deletes objects older than seven days.

C.Develop a batch process using Dataflow that runs weekly and deletes files based on their age.

D.Create a Cloud Run function that runs daily and deletes files older than seven days.

Answer: B

Explanation:

The correct answer is B, configuring a Cloud Storage lifecycle rule. This is the most efficient and cost-effective way to automatically delete objects based on their age in Cloud Storage.

Here's why:

Cloud Storage Lifecycle Management: Cloud Storage offers built-in lifecycle management rules specifically designed for tasks like automatically deleting objects after a certain period. This feature directly addresses the stated requirement of deleting files older than seven days.
https://cloud.google.com/storage/docs/lifecycle

Cost Efficiency: Lifecycle rules are server-side operations performed by Cloud Storage itself. This eliminates the need to run external compute resources like Cloud Functions, Cloud Run, or Dataflow, leading to significant cost savings. You only pay for the storage of the objects until they are deleted.

Simplicity and Reduced Complexity: Configuring a lifecycle rule is straightforward and requires minimal code or scripting. You can define the rule directly in the Cloud Console or using the gsutil command-line tool. This simplifies the management of your Cloud Storage bucket.

Alternative A (Cloud Scheduler + Cloud Run): While possible, this approach introduces unnecessary complexity and cost. It requires creating and managing a scheduled job and a Cloud Run function, both of which consume resources.

Alternative C (Dataflow): Dataflow is designed for large-scale data processing and is overkill for a simple deletion task. It's also significantly more expensive than using lifecycle rules.

Alternative D (Cloud Run): Similar to Option A, this involves unnecessary infrastructure and costs compared to the built-in lifecycle rule functionality.

In summary, lifecycle rules are the best choice because they provide a native, serverless, cost-effective, and simple solution for automatically managing object lifecycles within Cloud Storage. They directly address the problem statement without requiring additional compute resources or complex configurations.

Question: 6 CertyIQ

You work for a healthcare company that has a large on-premises data system containing patient records with personally identifiable information (PII) such as names, addresses, and medical diagnoses. You need a standardized

managed solution that de-identifies PII across all your data feeds prior to ingestion to Google Cloud. What should you do?

A.Use Cloud Run functions to create a serverless data cleaning pipeline. Store the cleaned data in BigQuery.

B.Use Cloud Data Fusion to transform the data. Store the cleaned data in BigQuery.

C.Load the data into BigQuery, and inspect the data by using SQL queries. Use Dataflow to transform the data and remove any errors.

D.Use Apache Beam to read the data and perform the necessary cleaning and transformation operations. Store the cleaned data in BigQuery.

Answer: B

Explanation:

The correct answer is **B. Use Cloud Data Fusion to transform the data. Store the cleaned data in BigQuery.**

Here's why: The core requirement is a standardized, managed solution for de-identifying PII before ingestion into Google Cloud. Cloud Data Fusion perfectly fits this need. It's a fully managed, cloud-native data integration service that helps build and manage data pipelines without coding. It offers pre-built connectors and transformations ideal for data cleansing and masking PII. The de-identified data can then be seamlessly stored in BigQuery.

Here's why the other options are less suitable:

A. Use Cloud Run functions to create a serverless data cleaning pipeline. Store the cleaned data in BigQuery: While Cloud Run is serverless and scalable, building and maintaining a complete data cleaning pipeline from scratch requires significant coding and management effort. It lacks the pre-built transforms and connectors that Data Fusion provides, making it less standardized and managed.

C. Load the data into BigQuery, and inspect the data by using SQL queries. Use Dataflow to transform the data and remove any errors: Loading data into BigQuery before de-identification violates the requirement of de-identifying PII prior to ingestion. This approach also creates a potential compliance issue.

D. Use Apache Beam to read the data and perform the necessary cleaning and transformation operations. Store the cleaned data in BigQuery: Apache Beam is a powerful framework for batch and stream data processing, and it's the basis for Dataflow. However, using Beam directly requires more coding and infrastructure management than using the fully managed Data Fusion. Cloud Data Fusion provides a visual interface and pre-built transformations based on Beam, simplifying the process.

In summary, Cloud Data Fusion offers the best balance of being a standardized, managed solution that handles PII de-identification before data lands in BigQuery.

Authoritative Links for further research:

Cloud Data Fusion Overview: https://cloud.google.com/data-fusion/docs/concepts/overview
Data Loss Prevention (DLP) with Cloud Data Fusion: https://cloud.google.com/data-fusion/docs/how-to/dlp
BigQuery: https://cloud.google.com/bigquery

Question: 7 CertyIQ

You manage a large amount of data in Cloud Storage, including raw data, processed data, and backups. Your organization is subject to strict compliance regulations that mandate data immutability for specific data types. You want to use an efficient process to reduce storage costs while ensuring that your storage strategy meets retention requirements. What should you do?

A.Configure lifecycle management rules to transition objects to appropriate storage classes based on access patterns. Set up Object Versioning for all objects to meet immutability requirements.

B.Move objects to different storage classes based on their age and access patterns. Use Cloud Key Management Service (Cloud KMS) to encrypt specific objects with customer-managed encryption keys (CMEK) to meet immutability requirements.

C.Create a Cloud Run function to periodically check object metadata, and move objects to the appropriate storage class based on age and access patterns. Use object holds to enforce immutability for specific objects.

D.Use object holds to enforce immutability for specific objects, and configure lifecycle management rules to transition objects to appropriate storage classes based on age and access patterns.

Answer: D

Explanation:

The correct answer is D because it directly addresses both requirements: data immutability and cost reduction.

Object Holds for Immutability: Object holds are specifically designed to prevent the deletion or modification of objects in Cloud Storage, ensuring immutability. This directly satisfies the compliance regulations regarding data immutability. There are two types of object holds: Retention holds for temporarily preventing deletion and Legal holds for indefinite prevention of deletion until the hold is removed.
https://cloud.google.com/storage/docs/object-holds
Lifecycle Management for Cost Optimization: Lifecycle management rules automate the process of transitioning objects between different storage classes (e.g., Standard, Nearline, Coldline, Archive) based on their age and access patterns. This can significantly reduce storage costs by moving infrequently accessed data to cheaper storage classes while keeping frequently accessed data readily available.
https://cloud.google.com/storage/docs/lifecycle

Option A is incorrect because while object versioning helps in data recovery, it doesn't inherently enforce immutability. Older versions can still be deleted or modified depending on configurations. Also, relying solely on storage classes for immutability is not a reliable solution.

Option B is incorrect because CMEK, while enhancing security, doesn't guarantee immutability. Encryption keys can be managed, but the underlying data is still potentially mutable. Furthermore, manually moving objects and encrypting them isn't efficient for large datasets.

Option C is incorrect because using a Cloud Run function for moving objects, while possible, adds unnecessary complexity and overhead. Lifecycle management provides a native and more efficient way to achieve the same result. Moreover, while object holds enforce immutability as in the correct answer, using a Cloud Run function to determine storage classes is a less efficient approach to lifecycle management.

Question: 8 CertyIQ

You work for an ecommerce company that has a BigQuery dataset that contains customer purchase history, demographics, and website interactions. You need to build a machine learning (ML) model to predict which customers are most likely to make a purchase in the next month. You have limited engineering resources and need to minimize the ML expertise required for the solution. What should you do?

A.Use BigQuery ML to create a logistic regression model for purchase prediction.

B.Use Vertex AI Workbench to develop a custom model for purchase prediction.

C.Use Colab Enterprise to develop a custom model for purchase prediction.

D.Export the data to Cloud Storage, and use AutoML Tables to build a classification model for purchase prediction.

Answer: A

Explanation:

The correct answer is A, using BigQuery ML to create a logistic regression model for purchase prediction.

Here's a detailed justification:

BigQuery ML allows users to create and execute machine learning models directly within BigQuery using SQL queries. This approach is advantageous because it avoids the need to move large datasets, simplifying the process and reducing costs associated with data transfer and storage duplication. Logistic regression is a common and easily interpretable algorithm for binary classification problems like purchase prediction. It's well-suited for situations where you want to estimate the probability of an event occurring (in this case, a customer making a purchase).

Option B, Vertex AI Workbench, involves more engineering effort to develop and deploy a custom model, which contradicts the requirement to minimize engineering resources. Custom models also require more ML expertise. Option C, Colab Enterprise, shares the same drawback as Vertex AI Workbench. While Colab Enterprise is good for experimentation, it's generally not the best option for deploying a production-ready ML model without significant engineering. Option D, AutoML Tables, offers automated model building, but it still requires exporting the data to Cloud Storage. This extra step of data export adds complexity and potential costs, which is not ideal considering the limited engineering resources.

Therefore, using BigQuery ML provides a straightforward and efficient solution, minimizing the need for extensive ML expertise and reducing data transfer overhead. It leverages the existing BigQuery infrastructure for data storage and processing, making it a suitable choice given the constraints.

Refer to the following documentation for more information:

BigQuery ML: https://cloud.google.com/bigquery-ml/docs
Logistic Regression in BigQuery ML: https://cloud.google.com/bigquery-ml/docs/reference/standard-sql/bigqueryml-syntax-create-logistic-regression

Question: 9 CertyIQ

You are designing a pipeline to process data files that arrive in Cloud Storage by 3:00 am each day. Data processing is performed in stages, where the output of one stage becomes the input of the next. Each stage takes a long time to run. Occasionally a stage fails, and you have to address the problem. You need to ensure that the final output is generated as quickly as possible. What should you do?

A.Design a Spark program that runs under Dataproc. Code the program to wait for user input when an error is detected. Rerun the last action after correcting any stage output data errors.

B.Design the pipeline as a set of PTransforms in Dataflow. Restart the pipeline after correcting any stage output data errors.

C.Design the workflow as a Cloud Workflow instance. Code the workflow to jump to a given stage based on an input parameter. Rerun the workflow after correcting any stage output data errors.

D.Design the processing as a directed acyclic graph (DAG) in Cloud Composer. Clear the state of the failed task after correcting any stage output data errors.

Answer: D

Explanation:

The correct answer is **D. Design the processing as a directed acyclic graph (DAG) in Cloud Composer. Clear the state of the failed task after correcting any stage output data errors.**

Here's a detailed justification:

Cloud Composer is a fully managed workflow orchestration service built on Apache Airflow. It's specifically

designed for building, scheduling, and monitoring complex data pipelines. Designing the data processing as a Directed Acyclic Graph (DAG) in Cloud Composer is ideal because:

1. **Orchestration:** DAGs in Composer allow defining the dependencies between different stages of data processing. This ensures that stages are executed in the correct order and only when their dependencies are met.
2. **Resilience:** Composer provides built-in mechanisms for handling failures. When a task fails, Composer can automatically retry it. More importantly, when you need to manually intervene, you can clear the state of the failed task after fixing the underlying issue (like corrupted data), allowing the pipeline to resume from that point without re-running all previous successful stages. This significantly reduces the processing time, which addresses the prompt's requirement to get the final output as quickly as possible.
3. **Monitoring:** Composer provides a web interface for monitoring the progress of DAGs and tasks. This allows you to quickly identify and diagnose failures.
4. **Scalability:** Composer can scale to handle large data volumes and complex workflows.

The other options are less suitable:

A (Spark on Dataproc): While Dataproc is suitable for data processing, the suggestion of waiting for user input after an error is impractical for automated pipelines intended to run unattended. It requires constant human intervention and slows down the overall processing.

B (Dataflow): Dataflow is excellent for parallel data processing, but simply restarting the pipeline after correcting errors would reprocess all data from the beginning, which is inefficient and contradicts the requirement of minimizing processing time. While Dataflow has fault tolerance, it doesn't give you the same fine-grained control over restarting specific failed tasks as Composer.

C (Cloud Workflows): Cloud Workflows are useful for orchestrating serverless functions and API calls. While it can orchestrate data processing steps, it's not specifically designed for complex data pipelines with dependencies like Cloud Composer. The requirement to code the workflow to jump to a specific stage based on an input parameter adds unnecessary complexity to the solution. Composer's built-in DAG management and task state management is more straightforward.

In summary, Cloud Composer with a DAG design provides the best combination of orchestration, resilience, monitoring, and the ability to resume from a specific failed task after correction, which meets the problem's requirements.

Relevant links:

Cloud Composer Documentation
Apache Airflow Documentation

Question: 10 CertyIQ

Another team in your organization is requesting access to a BigQuery dataset. You need to share the dataset with the team while minimizing the risk of unauthorized copying of data. You also want to create a reusable framework in case you need to share this data with other teams in the future. What should you do?

A.Create authorized views in the team's Google Cloud project that is only accessible by the team.

B.Create a private exchange using Analytics Hub with data egress restriction, and grant access to the team members.

C.Enable domain restricted sharing on the project. Grant the team members the BigQuery Data Viewer IAM role on the dataset.

D.Export the dataset to a Cloud Storage bucket in the team's Google Cloud project that is only accessible by the team.

Answer: A

Explanation:

The correct answer is **A. Create authorized views in the team's Google Cloud project that is only accessible by the team.**

Here's why:

Authorized Views Minimize Copying: Authorized views allow you to share specific columns or rows of a dataset without granting direct access to the underlying data. This inherently reduces the risk of unauthorized data copying because the team can only access the data exposed through the view.

Reusability: Authorized views can be easily cloned or adapted for sharing data with other teams in the future, providing a reusable framework. The view definition remains consistent, and you only need to adjust the authorization to grant access to new teams.

Data Security: You retain control over what data is exposed through the view. Sensitive columns can be excluded, or data can be aggregated or masked within the view's definition, enhancing data security.

Team Isolation: Creating the authorized view in the team's Google Cloud project ensures that the view and its underlying BigQuery queries run within their project's context, preventing potential cross-project billing or IAM issues.

Here's why the other options are less ideal:

B. Create a private exchange using Analytics Hub with data egress restriction, and grant access to the team members. While Analytics Hub facilitates data sharing, it's more complex than necessary for a simple intra-organization data sharing scenario. Data egress restrictions do help with preventing unauthorized copying, but introducing Analytics Hub adds unnecessary overhead.

C. Enable domain restricted sharing on the project. Grant the team members the BigQuery Data Viewer IAM role on the dataset. Domain-restricted sharing helps control who can be granted access but doesn't inherently prevent unauthorized copying. Giving the BigQuery Data Viewer role grants access to the entire dataset, increasing the risk of unintentional or malicious data extraction.

D. Export the dataset to a Cloud Storage bucket in the team's Google Cloud project that is only accessible by the team. Exporting the data duplicates it, consuming additional storage space and increasing the attack surface. It doesn't address the need for a reusable framework. Exporting also requires more manual management and is less efficient than querying a view directly.

In summary, authorized views provide a secure, reusable, and efficient way to share data with other teams in your organization while minimizing the risk of unauthorized data copying.

Relevant links for further research:

Authorized Views: https://cloud.google.com/bigquery/docs/authorized-views
BigQuery IAM roles: https://cloud.google.com/bigquery/docs/access-control-basic-roles
Analytics Hub: https://cloud.google.com/analytics-hub/docs

Question: 11 CertyIQ

Your company has developed a website that allows users to upload and share video files. These files are most frequently accessed and shared when they are initially uploaded. Over time, the files are accessed and shared less

frequently, although some old video files may remain very popular.
You need to design a storage system that is simple and cost-effective. What should you do?

A.Create a single-region bucket with Autoclass enabled.

B.Create a single-region bucket. Configure a Cloud Scheduler job that runs every 24 hours and changes the storage class based on upload date.

C.Create a single-region bucket with custom Object Lifecycle Management policies based on upload date.

D.Create a single-region bucket with Archive as the default storage class.

Answer: A

Explanation:

The correct answer is **A. Create a single-region bucket with Autoclass enabled.**

Here's why:

The scenario describes data with an access pattern of "hot" immediately after upload and gradually cooling down over time, with the occasional surge in popularity for older files. This makes Autoclass the ideal choice.

Autoclass in Google Cloud Storage automatically transitions objects between storage classes (Standard, Nearline, Coldline, Archive) based on access patterns. It intelligently analyzes object usage and moves objects to more cost-effective storage classes when they are accessed less frequently, and back to Standard if they become popular again.

Cost Optimization: It reduces storage costs by automatically placing infrequently accessed objects in cheaper storage classes like Nearline, Coldline, or Archive.
Performance Optimization: When objects are accessed frequently, Autoclass ensures they reside in the Standard storage class for optimal performance.
Simplicity: It eliminates the need for manual lifecycle management rules or scripts, greatly simplifying storage administration. This aligns with the requirement for a "simple" solution.
Handles Fluctuations: Autoclass dynamically adapts to changes in access patterns, moving objects back to Standard class if they become popular again. This is crucial for the scenario where older files may occasionally regain popularity.

Let's examine why other options are less suitable:

B. Configure a Cloud Scheduler job: This requires manual configuration and scripting to determine the appropriate storage class based on upload date. This is a more complex solution and does not dynamically adapt to changing access patterns.
C. Custom Object Lifecycle Management policies: This also requires manual configuration, defining specific rules for transitioning objects based on their age. While effective, it does not adapt to changes in access frequency. It's also less flexible than Autoclass, which continuously monitors access patterns.
D. Archive as the default storage class: This would be extremely slow and unsuitable for the initial "hot" phase when files are most frequently accessed and shared. Archive storage is intended for long-term data retention with infrequent access.

Therefore, Autoclass is the most cost-effective and simplest solution because it automatically optimizes storage costs based on access patterns, including re-promoting objects to a faster storage class when required.

Authoritative Links:

Google Cloud Storage Autoclass: https://cloud.google.com/storage/docs/autoclass
Storage Classes: https://cloud.google.com/storage/docs/storage-classes
Object Lifecycle Management: https://cloud.google.com/storage/docs/lifecycle

Question: 12 CertyIQ

You recently inherited a task for managing Dataflow streaming pipelines in your organization and noticed that proper access had not been provisioned to you. You need to request a Google-provided IAM role so you can restart the pipelines. You need to follow the principle of least privilege. What should you do?

A.Request the Dataflow Developer role.
B.Request the Dataflow Viewer role.
C.Request the Dataflow Worker role.
D.Request the Dataflow Admin role.

Answer: A

Explanation:

The correct answer is A, requesting the Dataflow Developer role. Here's why:

The principle of least privilege dictates granting only the minimum necessary permissions to perform a task. In this scenario, the requirement is to restart Dataflow pipelines.

Dataflow Admin (D) grants full control over Dataflow jobs and resources. This is excessive for simply restarting pipelines. It allows creating, updating, and deleting pipelines, which is beyond the needed scope.

Dataflow Viewer (B) only allows read-only access. This role is suitable for monitoring pipelines but not for making changes like restarting them.

Dataflow Worker (C) is designed for the Compute Engine instances executing the Dataflow pipeline, not for users managing the pipelines. This role lacks the permissions to initiate restart operations.

Dataflow Developer (A) provides the necessary permissions to manage and execute Dataflow jobs, including the ability to start and restart pipelines. This role grants the ability to create, update, and delete Dataflow templates and launch jobs from these templates.

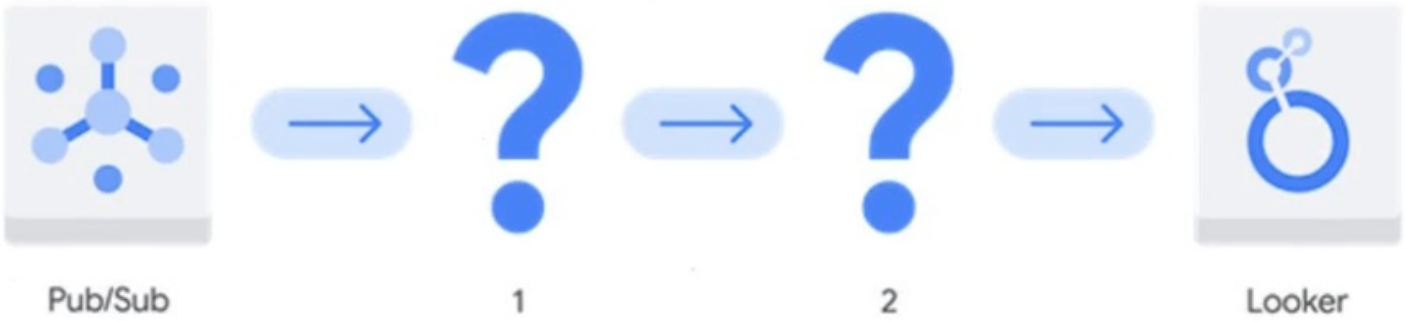
Therefore, requesting the Dataflow Developer role adheres to the principle of least privilege by granting the user sufficient permissions to restart pipelines without granting broader, unnecessary access to other Dataflow resources. This minimizes the potential security risk associated with over-provisioned access.

Further research:

Dataflow roles
IAM best practices using least privilege

Question: 13 CertyIQ

You need to create a new data pipeline. You want a serverless solution that meets the following requirements:
• Data is streamed from Pub/Sub and is processed in real-time.
• Data is transformed before being stored.
• Data is stored in a location that will allow it to be analyzed with SQL using Looker.



Which Google Cloud services should you recommend for the pipeline?

A.1. Dataproc Serverless
2. Bigtable

B.1. Cloud Composer
2. Cloud SQL for MySQL

C.1. BigQuery
2. Analytics Hub

D.1. Dataflow
2. BigQuery

Answer: D

Explanation:

D. 1. Dataflow 2. BigQuery.

You need to recommend Google Cloud services for a data pipeline, which typically involves:

Streaming or batch data processing.

Efficient storage and querying for analytics.

Dataflow is a serverless, fully managed service designed for ETL (Extract, Transform, Load) and data processing. It supports:

Real-time (streaming) and batch data processing.

Integration with BigQuery, Pub/Sub, and Cloud Storage.

Apache Beam-based pipelines for scalability and efficiency.

BigQuery is a fully managed, serverless data warehouse for:

Fast analytics on large datasets.

Seamless integration with Dataflow, Cloud Storage, and other services.

SQL-based querying without infrastructure management.

Dataflow + BigQuery is the ideal solution for an efficient and scalable data pipeline in Google Cloud.

Question: 14 CertyIQ

Your team wants to create a monthly report to analyze inventory data that is updated daily. You need to aggregate the inventory counts by using only the most recent month of data, and save the results to be used in a Looker Studio dashboard. What should you do?

A.Create a materialized view in BigQuery that uses the SUM( ) function and the DATE_SUB( ) function.

B.Create a saved query in the BigQuery console that uses the SUM( ) function and the DATE_SUB( ) function. Re-run the saved query every month, and save the results to a BigQuery table.

C.Create a BigQuery table that uses the SUM( ) function and the _PARTITIONDATE filter.

D.Create a BigQuery table that uses the SUM( ) function and the DATE_DIFF( ) function.

Answer: A

Explanation:

The correct answer is **A. Create a materialized view in BigQuery that uses the SUM( ) function and the DATE_SUB( ) function.**

Here's why:

Materialized Views: Materialized views in BigQuery are pre-computed views that periodically cache the results of a query. This is ideal for scenarios requiring consistent and fast access to aggregated data, like the monthly report in this case. The data is automatically refreshed based on a user-defined schedule or when the underlying data changes, ensuring the report always reflects the most recent month's inventory. https://cloud.google.com/bigquery/docs/materialized-views-intro

SUM() Function: The SUM() function is necessary for aggregating the inventory counts, providing a total count for each inventory item within the specified time frame.

DATE_SUB() Function: The DATE_SUB() function is crucial for filtering the data to include only the most recent month's data. It allows you to subtract a specific interval (in this case, one month) from the current date. Using DATE_SUB() in the materialized view's query ensures that only data from the past month is considered in the aggregation. https://cloud.google.com/bigquery/docs/reference/standard-sql/date_functions#date_sub

Why other options are less suitable:

B. Saved Query: While a saved query can calculate the aggregate, it needs to be manually re-run each month and the results saved to a table. This approach is not automated and requires manual intervention, which is less efficient and prone to human error.

C. BigQuery Table with _PARTITIONDATE: While partitioning based on _PARTITIONDATE is good practice for data organization, it doesn't automatically aggregate data for only the most recent month. A query would still be needed to filter and aggregate. Additionally, creating a table directly doesn't inherently involve aggregation functions like SUM().

D. BigQuery Table with DATE_DIFF(): Similar to option C, creating a BigQuery table using DATE_DIFF() would require an additional query to filter and aggregate. Directly creating a table using only DATE_DIFF() wouldn't produce the desired aggregated results. And also like Option C, creating a table directly doesn't inherently involve aggregation functions like SUM().

In summary, a materialized view provides an automated, efficient, and up-to-date solution for aggregating the most recent month's inventory data and making it readily available for the Looker Studio dashboard. It combines the aggregation logic with a filtering mechanism and stores the results in a pre-computed manner, which makes it more performant and reliable than the other options.

Question: 15 CertyIQ

You have a BigQuery dataset containing sales data. This data is actively queried for the first 6 months. After that,

the data is not queried but needs to be retained for 3 years for compliance reasons. You need to implement a data management strategy that meets access and compliance requirements, while keeping cost and administrative overhead to a minimum. What should you do?

A.Use BigQuery long-term storage for the entire dataset. Set up a Cloud Run function to delete the data from BigQuery after 3 years.

B.Partition a BigQuery table by month. After 6 months, export the data to Coldline storage. Implement a lifecycle policy to delete the data from Cloud Storage after 3 years.

C.Set up a scheduled query to export the data to Cloud Storage after 6 months. Write a stored procedure to delete the data from BigQuery after 3 years.

D.Store all data in a single BigQuery table without partitioning or lifecycle policies.

Answer: B

Explanation:

The correct answer is B. Here's a detailed justification:

Option B is the most cost-effective and efficient solution for managing the sales data in BigQuery. Partitioning the BigQuery table by month enables efficient querying for the first 6 months because BigQuery only scans relevant partitions. After 6 months, when the data is no longer actively queried, exporting the data to Cloud Storage Coldline offers significant cost savings compared to storing it in BigQuery, especially for data that's rarely accessed. Coldline is designed for infrequent access with lower storage costs. Implementing a lifecycle policy in Cloud Storage to delete the data after 3 years automatically handles the compliance requirement without any manual intervention, reducing administrative overhead.

Option A is less efficient because BigQuery long-term storage still incurs higher costs than Coldline for infrequently accessed data. Additionally, using a Cloud Run function to delete data introduces more complexity and operational overhead than a Cloud Storage lifecycle policy.

Option C involves setting up a scheduled query which adds to the complexity and potential cost. Moreover, the stored procedure to delete data after 3 years would require maintenance and monitoring.

Option D is not cost-effective as it stores all data in BigQuery without leveraging cheaper storage options for infrequently accessed data, leading to higher storage costs unnecessarily. It also lacks automated deletion, increasing administrative burden for compliance.

Therefore, option B strikes the optimal balance between cost, performance, and manageability, making it the best solution.

Further Reading:

BigQuery Pricing: https://cloud.google.com/bigquery/pricing
Cloud Storage Pricing: https://cloud.google.com/storage/pricing
Cloud Storage Lifecycle Management: https://cloud.google.com/storage/docs/lifecycle

Question: 16 CertyIQ

You have created a LookML model and dashboard that shows daily sales metrics for five regional managers to use. You want to ensure that the regional managers can only see sales metrics specific to their region. You need an easy-to-implement solution. What should you do?

A.Create a sales_region user attribute, and assign each manager's region as the value of their user attribute. Add an access_filter Explore filter on the region_name dimension by using the sales_region user attribute.

B.Create five different Explores with the sql_always_filter Explore filter applied on the region_name dimension. Set each region_name value to the corresponding region for each manager.

C.Create separate Looker dashboards for each regional manager. Set the default dashboard filter to the

corresponding region for each manager.

D.Create separate Looker instances for each regional manager. Copy the LookML model and dashboard to each instance. Provision viewer access to the corresponding manager.

Answer: A

Explanation:

The correct answer is A. Here's why:

Option A offers the most efficient and maintainable solution for data access control in Looker, aligning with the principle of least privilege and utilizing Looker's built-in features for personalized data experiences. User attributes are a powerful feature in Looker that allow you to define user-specific values. By creating a sales_region user attribute and assigning each manager's region to it, you establish a link between the user and their data scope. The access_filter parameter in LookML then leverages this user attribute to dynamically filter the data shown in the Explore. This ensures that each manager only sees sales metrics related to their assigned region. This approach avoids code duplication and centralizes access control within the LookML model.Options B, C, and D are less optimal for various reasons. Option B, creating separate Explores with sql_always_filter, requires maintaining multiple code blocks and lacks flexibility. Option C, creating separate dashboards with default filters, is also less scalable and harder to manage as the number of users increases. Moreover, default dashboard filters can be overridden by users, potentially exposing data they shouldn't see. Option D, creating separate Looker instances, is the most extreme and resource-intensive approach, leading to significant overhead in terms of maintenance, cost, and administration.

Key benefits of using user attributes and access_filter:

Centralized Access Control: Access rules are defined in the LookML model, making them easy to manage and audit.
Scalability: The solution scales easily as you add more users or regions.
Maintainability: Changes to access rules only need to be made in one place.
Security: Ensures that users only have access to the data they are authorized to see.
Personalized User Experience: Delivers a tailored data experience to each user based on their assigned region.

Relevant documentation:

Looker User Attributes: Explains how to define and use user attributes.
Looker Access Filters: Describes how to use access filters to restrict data access based on user attributes.

Question: 17 CertyIQ

You need to design a data pipeline that ingests data from CSV, Avro, and Parquet files into Cloud Storage. The data includes raw user input. You need to remove all malicious SQL injections before storing the data in BigQuery. Which data manipulation methodology should you choose?

A.EL
B.ELT
C.ETL
D.ETLT

Answer: C

Explanation:

The correct methodology for this scenario is ETL (Extract, Transform, Load). Here's why:

The requirement is to ingest data from various sources (CSV, Avro, Parquet) in Cloud Storage, and importantly, remove malicious SQL injections before loading the data into BigQuery. This necessitates a transformation step.

ETL is designed precisely for this. First, data is Extracted from the various sources (Cloud Storage in this case). Then, the crucial step of Transformation takes place. This is where the malicious SQL injections are removed through data cleansing, validation, and sanitization techniques. Finally, the cleaned and transformed data is Loaded into BigQuery.

ELT (Extract, Load, Transform) would not be suitable here. ELT loads the raw data directly into the data warehouse (BigQuery), and then performs the transformation. Loading potentially malicious SQL injections into BigQuery poses a significant security risk. BigQuery would have to run the transformation on the loaded malicious code potentially leading to an exploit.

ELB (Elastic Load Balancing) is a service for distributing network traffic and does not pertain to data manipulation methodologies.

ELTC is not a standard data processing methodology.

Therefore, ETL is the optimal approach because it allows for data cleansing and security measures (SQL injection removal) to be implemented before the data is loaded into BigQuery, mitigating the security risk. The transformation step ensures data quality and integrity.

For further research:

ETL (Extract, Transform, Load): https://cloud.google.com/learn/what-is-etl
Data Security in BigQuery: https://cloud.google.com/bigquery/docs/security-overview
SQL Injection Prevention: OWASP (Open Web Application Security Project) guidelines on SQL injection prevention techniques will be useful in designing the transformation step within the ETL pipeline. (https://owasp.org/www-community/attacks/SQL_Injection)

Question: 18 CertyIQ

You are working with a large dataset of customer reviews stored in Cloud Storage. The dataset contains several inconsistencies, such as missing values, incorrect data types, and duplicate entries. You need to clean the data to ensure that it is accurate and consistent before using it for analysis. What should you do?

A.Use the PythonOperator in Cloud Composer to clean the data and load it into BigQuery. Use SQL for analysis.

B.Use BigQuery to batch load the data into BigQuery. Use SQL for cleaning and analysis.

C.Use Storage Transfer Service to move the data to a different Cloud Storage bucket. Use event triggers to invoke Cloud Run functions to load the data into BigQuery. Use SQL for analysis.

D.Use Cloud Run functions to clean the data and load it into BigQuery. Use SQL for analysis.

Answer: A

Explanation:

Option A is the most suitable approach for cleaning and analyzing a large dataset of customer reviews with inconsistencies stored in Cloud Storage.

Here's why:

Data Cleaning with PythonOperator in Cloud Composer: Cloud Composer is a managed Apache Airflow service that allows you to orchestrate complex data pipelines. Using the PythonOperator within a Cloud Composer workflow lets you execute Python scripts to perform data cleaning tasks. Python provides extensive libraries (e.g., pandas) for handling missing values, data type conversions, duplicate removal, and

other data quality issues.
Loading Cleaned Data into BigQuery: After cleaning the data using Python, the PythonOperator can load the cleaned data directly into BigQuery. BigQuery is a fully-managed, serverless data warehouse that is ideal for storing and analyzing large datasets.
SQL for Analysis: Once the data is in BigQuery, you can use SQL, the standard language for querying and analyzing data in relational databases, to gain insights from the customer reviews.
Orchestration and Scalability: Cloud Composer provides a centralized platform for managing and monitoring the entire data pipeline. It can scale to handle large datasets and complex data cleaning workflows.

Why other options are less ideal:

B. BigQuery for Cleaning: While BigQuery supports some data manipulation using SQL, complex cleaning operations with Python would be more effective. Batch loading might not address real-time issues.
C. Storage Transfer Service and Cloud Run: Storage Transfer Service is primarily for moving data between storage systems, not for data transformation. Using Cloud Run functions for data cleaning could be expensive and complex for large datasets.
D. Cloud Run Functions for cleaning: Using Cloud Run functions for data cleaning could be expensive and complex for large datasets compared to using Cloud Composer, which is designed specifically for these types of data pipelines.

Supporting Information:

Cloud Composer: https://cloud.google.com/composer/docs
BigQuery: https://cloud.google.com/bigquery/docs
PythonOperator: https://airflow.apache.org/docs/apache-airflow/stable/operators/python.html

Therefore, leveraging Cloud Composer with the PythonOperator allows for robust data cleaning, seamless integration with BigQuery for storage, and the power of SQL for insightful analysis. This makes option A the optimal solution.

Question: 19 CertyIQ

Your retail organization stores sensitive application usage data in Cloud Storage. You need to encrypt the data without the operational overhead of managing encryption keys. What should you do?

A.Use Google-managed encryption keys (GMEK).

B.Use customer-managed encryption keys (CMEK).

C.Use customer-supplied encryption keys (CSEK).

D.Use customer-supplied encryption keys (CSEK) for the sensitive data and customer-managed encryption keys (CMEK) for the less sensitive data.

Answer: A

Explanation:

The correct answer is A: Use Google-managed encryption keys (GMEK). Here's why:

The primary requirement is to encrypt data at rest in Cloud Storage while minimizing operational overhead related to key management. Google-managed encryption keys (GMEK) offer the simplest solution. With GMEK, Google handles the key generation, rotation, storage, and protection. This offloads the entire key management burden from the organization.

Customer-managed encryption keys (CMEK), while providing more control over encryption keys, require the organization to use Cloud Key Management Service (KMS) or a similar service to manage the keys. This

introduces operational overhead, including key creation, rotation policies, access control, and monitoring. While CMEK might be suitable in scenarios where very granular control and auditing of key usage are paramount, they are not the ideal choice when minimizing operational effort is the goal.

Customer-supplied encryption keys (CSEK) require the organization to generate, manage, and provide the encryption keys themselves with each request to read or write the data. This option is the most operationally intensive, as it necessitates managing and securely transmitting the encryption keys for every operation. Furthermore, CSEK provides the least protection since Google never stores the keys.

Therefore, GMEK fulfills the data encryption requirement while eliminating the operational burden of key management, making it the most appropriate solution for this scenario. GMEK allows the organization to focus on their core business functions instead of complex key management procedures. It also utilizes Google's security infrastructure and expertise, providing a secure and reliable solution.

Further reading:

Google Cloud Storage Encryption Options: https://cloud.google.com/storage/docs/encryption
Google-managed encryption keys (GMEK): https://cloud.google.com/kms/docs/cmek/concepts (Mentions GMEK as the default and simplest option)

Question: 20 CertyIQ

You work for a financial organization that stores transaction data in BigQuery. Your organization has a regulatory requirement to retain data for a minimum of seven years for auditing purposes. You need to ensure that the data is retained for seven years using an efficient and cost-optimized approach. What should you do?

A.Create a partition by transaction date, and set the partition expiration policy to seven years.

B.Set the table-level retention policy in BigQuery to seven years.

C.Set the dataset-level retention policy in BigQuery to seven years.

D.Export the BigQuery tables to Cloud Storage daily, and enforce a lifecycle management policy that has a seven-year retention rule.

Answer: B

Explanation:

The correct answer is B: Set the table-level retention policy in BigQuery to seven years.

Here's a detailed justification:

The requirement is to retain transaction data in BigQuery for seven years efficiently and cost-effectively due to regulatory requirements. BigQuery offers several features to manage data retention, and the optimal choice depends on the granularity and scope of the retention policy.

Option A, creating a partition by transaction date and setting partition expiration, works but may not be the most straightforward solution for a seven-year minimum retention. Partitions are suitable for optimizing query performance and cost by targeting specific data subsets. While partition expiration can delete old data, ensuring data is retained for a minimum period requires careful management of partition creation and expiration. It can get complicated if you need to backfill data older than expected.

Option B, setting the table-level retention policy, is the most direct and suitable approach. BigQuery's table-level retention policy allows you to specify a minimum retention period for all data within a table. This guarantees data won't be deleted before seven years, satisfying the regulatory need. This is a simple configuration, and easily audited.

Option C, dataset-level retention, applies the retention policy to all tables within the dataset. While valid, it

might be too broad if only certain tables require the seven-year retention. Applying it dataset-wide might unnecessarily retain data that doesn't need it, increasing storage costs.

Option D, exporting to Cloud Storage and using lifecycle management, is a viable backup and archival strategy, but adds significant overhead. Exporting the data incurs egress charges from BigQuery, and managing lifecycle policies adds operational complexity. BigQuery's storage costs are already optimized for long-term storage, making this approach less cost-effective unless specific archival requirements exist (e.g., storing data in a cheaper storage class). Also, it adds complexity to restore the data to BigQuery if needed for auditing.

Therefore, setting the table-level retention policy is the most efficient, straightforward, and cost-optimized way to satisfy the seven-year retention requirement. It directly addresses the need without unnecessary overhead or complexity.

Relevant Links:

BigQuery Table Expiration: https://cloud.google.com/bigquery/docs/managing-table-storage#table_expiration
BigQuery Dataset Expiration: https://cloud.google.com/bigquery/docs/managing-table-storage#dataset_expiration
Cloud Storage Object Lifecycle Management: https://cloud.google.com/storage/docs/lifecycle

Question: 21 CertyIQ

You need to create a weekly aggregated sales report based on a large volume of data. You want to use Python to design an efficient process for generating this report. What should you do?

A.Create a Cloud Run function that uses NumPy. Use Cloud Scheduler to schedule the function to run once a week.

B.Create a Colab Enterprise notebook and use the bigframes.pandas library. Schedule the notebook to execute once a week.

C.Create a Cloud Data Fusion and Wrangler flow. Schedule the flow to run once a week.

D.Create a Dataflow directed acyclic graph (DAG) coded in Python. Use Cloud Scheduler to schedule the code to run once a week.

Answer: B

Explanation:

The correct answer is B. Here's a detailed justification:

Requirement: The question emphasizes efficient processing of a large volume of data using Python to generate a weekly sales report.

Option A: Cloud Run with NumPy: While Cloud Run provides a scalable container execution environment and NumPy is suitable for numerical operations, it's not optimized for handling large datasets typically encountered in data warehousing or analytics. NumPy primarily works with in-memory data, making it less efficient for substantial data volumes.

Option B: Colab Enterprise with BigFrames: Colab Enterprise offers an interactive notebook environment with scalable compute. The key advantage is bigframes.pandas. This library allows you to use familiar Pandas-like syntax to interact with BigQuery, Google's fully-managed, serverless data warehouse. This is ideal for large datasets because the computation happens in BigQuery's scalable infrastructure, rather than in the notebook's memory. Scheduling execution of Colab notebooks is easily achievable. Google Cloud Colab Enterprise. BigFrames documentation helps understand usage scenarios BigFrames

Option C: Cloud Data Fusion with Wrangler: Cloud Data Fusion is an excellent ETL (Extract, Transform, Load) tool, particularly when you need a no-code/low-code approach for building data pipelines. Wrangler within Cloud Data Fusion helps clean and transform data through a visual interface. However, if you specifically want to use Python and aim for maximum control over the aggregation logic, this visual, no-code approach is not the most appropriate. Additionally, using code within Data Fusion usually involves writing UDFs (User Defined Functions), which adds complexity compared to simply using Python in a notebook connected to BigQuery via BigFrames.

Option D: Dataflow DAG with Python: Dataflow is a powerful data processing service well-suited for both batch and stream processing. Coding a Dataflow DAG in Python gives you precise control over the transformation logic. However, for a simple weekly aggregation report on existing data within BigQuery, Dataflow represents more overhead in terms of setup and code complexity than leveraging BigQuery directly through BigFrames. Dataflow excels in continuous, streaming scenarios.

Therefore, **Option B (Colab Enterprise with bigframes.pandas)** provides the best balance of Python programmability, scalability via BigQuery, and ease of use for the specified weekly reporting requirement involving a large volume of data. It allows for efficient execution by offloading computations to BigQuery's infrastructure and is simpler to set up than Dataflow.

Question: 22 CertyIQ

Your organization has decided to move their on-premises Apache Spark-based workload to Google Cloud. You want to be able to manage the code without needing to provision and manage your own cluster. What should you do?

A.Migrate the Spark jobs to Dataproc Serverless.
B.Configure a Google Kubernetes Engine cluster with Spark operators, and deploy the Spark jobs.
C.Migrate the Spark jobs to Dataproc on Google Kubernetes Engine.
D.Migrate the Spark jobs to Dataproc on Compute Engine.

Answer: A

Explanation:

The correct answer is A: Migrate the Spark jobs to Dataproc Serverless. Here's why:

Dataproc Serverless is a fully managed, auto-scaling service that allows you to run Spark workloads without managing the underlying infrastructure. This aligns perfectly with the requirement of managing the code without provisioning and managing a cluster. It automatically provisions resources as needed, scales up or down based on workload demands, and handles cluster management tasks. This simplifies operations and reduces administrative overhead.

Option B, configuring a Google Kubernetes Engine (GKE) cluster with Spark operators, requires significant effort in setting up, configuring, and maintaining the GKE cluster itself, as well as the Spark operators. This contradicts the requirement of avoiding cluster management.

Option C, migrating Spark jobs to Dataproc on GKE, still involves some level of Kubernetes cluster management, although less than Option B. Dataproc on GKE allows you to use Kubernetes as the underlying infrastructure for Dataproc, but still demands managing the GKE cluster resources.

Option D, migrating Spark jobs to Dataproc on Compute Engine, involves managing the underlying virtual machines (VMs) and the Dataproc cluster itself. This is the most hands-on approach and directly contradicts the requirement of avoiding cluster management.

Dataproc Serverless is designed specifically for users who want to focus on their Spark jobs without the operational overhead of managing clusters. It streamlines the process and minimizes the need for infrastructure management. It seamlessly integrates with other Google Cloud services and offers cost optimization through its pay-per-use model.

For further reading, explore the following resources:

Dataproc Serverless: https://cloud.google.com/dataproc-serverless
Dataproc Overview: https://cloud.google.com/dataproc/docs/concepts/overview

Thank you

Thank you for being so interested in the premium exam material.
I'm glad to hear that you found it informative and helpful.

But Wait

I wanted to let you know that there is more content available in the full version.
The full paper contains additional sections and information that you may find helpful, and I encourage you to download it to get a more comprehensive and detailed view of all the subject matter.

Download Full Version Now

Future is Secured

100% Pass Guarantee

24/7 Customer Support

Mail us - certyiqofficial@gmail.com

Free Updates

Lifetime Free Updates!

Total: **72 Questions**

Link: https://certyiq.com/papers/google/associate-data-practitioner

Question 1 (Exam A)

Your retail company wants to predict customer churn using historical purchase data stored in BigQuery. The dataset includes customer demographics, purchase history, and a label indicating whether the customer churned or not. You want to build a machine learning model to identify customers at risk of churning. You need to create and train a logistic regression model for predicting customer churn, using the customer_data table with the churned column as the target label. Which BigQuery ML query should you use?

A.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT *
FROM customer_data;
```

B.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT * EXCEPT(churned),
       churned AS label
FROM customer_data;
```

C.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT * EXCEPT(churned)
FROM customer_data;
```

D.

```
CREATE OR REPLACE MODEL churn_prediction_model
OPTIONS(model_type='logistic_reg') AS
SELECT churned as label
FROM customer_data;
```

EXPOSE CORRECT ANSWER

Answer : **B**

NEXT QUESTION

Question 2 (Exam A)

Your company has several retail locations. Your company tracks the total number of sales made at each location each day. You want to use SQL to calculate the weekly moving average of sales by location to identify trends for each store. Which query should you use?

A.

```
SELECT store_id, date, total_sales, AVG(total_sales) OVER (
PARTITION BY store_id
ORDER BY total_sales RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolli
FROM store_sales_daily
```

B.

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY date ORDER BY store_id ROWS BETWEEN 6 PRECEDING AND CURRENT
FROM store_sales_daily
```

C.

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY store_id
ORDER BY date ROWS BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

D.

```
SELECT store_id, date, total_sales, AVG(total_sales)
OVER (
PARTITION BY total_sales
ORDER BY date RANGE BETWEEN 6 PRECEDING AND CURRENT ROW ) as rolling_avg
FROM store_sales_daily
```

EXPOSE CORRECT ANSWER

Answer : **C**

NEXT QUESTION

Question 3 (Exam A)

Your company is building a near real-time streaming pipeline to process JSON telemetry data from small appliances. You need to process messages arriving at a Pub/Sub topic, capitalize letters in the serial number field, and write results to BigQuery. You want to use a managed service and write a minimal amount of code for underlying transformations. What should you do?

A. Use a Pub/Sub to BigQuery subscription, write results directly to BigQuery, and schedule a transformation query to run every five minutes.
B. Use a Pub/Sub to Cloud Storage subscription, write a Cloud Run service that is triggered when objects arrive in the bucket, performs the transformations, and writes the results to BigQuery.
C. Use the “Pub/Sub to BigQuery” Dataflow template with a UDF, and write the results to BigQuery.
D. Use a Pub/Sub push subscription, write a Cloud Run service that accepts the messages, performs the transformations, and writes the results to BigQuery.

EXPOSE CORRECT ANSWER

Answer : **C**

NEXT QUESTION

Question 4 (Exam A)

You want to process and load a daily sales CSV file stored in Cloud Storage into BigQuery for downstream reporting. You need to quickly build a scalable data pipeline that transforms the data while providing insights into data quality issues. What should you do?

A. Create a batch pipeline in Cloud Data Fusion by using a Cloud Storage source and a BigQuery sink.
B. Load the CSV file as a table in BigQuery, and use scheduled queries to run SQL transformation scripts.
C. Load the CSV file as a table in BigQuery. Create a batch pipeline in Cloud Data Fusion by using a BigQuery source and sink.
D. Create a batch pipeline in Dataflow by using the Cloud Storage CSV file to BigQuery batch template.

EXPOSE CORRECT ANSWER

Answer : **A**

NEXT QUESTION

Question 5 (Exam A)

You manage a Cloud Storage bucket that stores temporary files created during data processing. These temporary files are only needed for seven days, after which they are no longer needed. To reduce storage costs and keep your bucket organized, you want to automatically delete these files once they are older than seven days. What should you do?

A. Set up a Cloud Scheduler job that invokes a weekly Cloud Run function to delete files older than seven days.
B. Configure a Cloud Storage lifecycle rule that automatically deletes objects older than seven days.
C. Develop a batch process using Dataflow that runs weekly and deletes files based on their age.
D. Create a Cloud Run function that runs daily and deletes files older than seven days.

EXPOSE CORRECT ANSWER

Answer : **B**

NEXT QUESTION

Question 6 (Exam A)

You work for a healthcare company that has a large on-premises data system containing patient records with personally identifiable information (PII) such as names, addresses, and medical diagnoses. You need a standardized managed solution that de-identifies PII across all your data feeds prior to ingestion to Google Cloud. What should you do?

A. Use Cloud Run functions to create a serverless data cleaning pipeline. Store the cleaned data in BigQuery.
B. Use Cloud Data Fusion to transform the data. Store the cleaned data in BigQuery.
C. Load the data into BigQuery, and inspect the data by using SQL queries. Use Dataflow to transform the data and remove any errors.
D. Use Apache Beam to read the data and perform the necessary cleaning and transformation operations. Store the cleaned data in BigQuery.

EXPOSE CORRECT ANSWER

Answer : **B**

NEXT QUESTION

Question 7 (Exam A)

You manage a large amount of data in Cloud Storage, including raw data, processed data, and backups. Your organization is subject to strict compliance regulations that mandate data immutability for specific data types. You want to use an efficient process to reduce storage costs while ensuring that your storage strategy meets retention requirements. What should you do?

A. Configure lifecycle management rules to transition objects to appropriate storage classes based on access patterns. Set up Object Versioning for all objects to meet immutability requirements.
B. Move objects to different storage classes based on their age and access patterns. Use Cloud Key Management Service (Cloud KMS) to encrypt specific objects with customer-managed encryption keys (CMEK) to meet immutability requirements.
C. Create a Cloud Run function to periodically check object metadata, and move objects to the appropriate storage class based on age and access patterns. Use object holds to enforce immutability for specific objects.
D. Use object holds to enforce immutability for specific objects, and configure lifecycle management rules to transition objects to appropriate storage classes based on age and access patterns.

EXPOSE CORRECT ANSWER

Answer : **D**

NEXT QUESTION

Question 8 (Exam A)

You work for an ecommerce company that has a BigQuery dataset that contains customer purchase history, demographics, and website interactions. You need to build a machine learning (ML) model to predict which customers are most likely to make a purchase in the next month. You have limited engineering resources and need to minimize the ML expertise required for the solution. What should you do?

A. Use BigQuery ML to create a logistic regression model for purchase prediction.
B. Use Vertex AI Workbench to develop a custom model for purchase prediction.
C. Use Colab Enterprise to develop a custom model for purchase prediction.
D. Export the data to Cloud Storage, and use AutoML Tables to build a classification model for purchase prediction.

EXPOSE CORRECT ANSWER

Answer : **A**

NEXT QUESTION

Question 9 (Exam A)

You are designing a pipeline to process data files that arrive in Cloud Storage by 3:00 am each day. Data processing is performed in stages, where the output of one stage becomes the input of the next. Each stage takes a long time to run. Occasionally a stage fails, and you have to address the problem. You need to ensure that the final output is generated as quickly as possible. What should you do?

A. Design a Spark program that runs under Dataproc. Code the program to wait for user input when an error is detected. Rerun the last action after correcting any stage output data errors.
B. Design the pipeline as a set of PTransforms in Dataflow. Restart the pipeline after correcting any stage output data errors.

C. Design the workflow as a Cloud Workflow instance. Code the workflow to jump to a given stage based on an input parameter. Rerun the workflow after correcting any stage output data errors.
D. Design the processing as a directed acyclic graph (DAG) in Cloud Composer. Clear the state of the failed task after correcting any stage output data errors.

EXPOSE CORRECT ANSWER

Answer : **D**

NEXT QUESTION

Question 10 (Exam A)

Another team in your organization is requesting access to a BigQuery dataset. You need to share the dataset with the team while minimizing the risk of unauthorized copying of data. You also want to create a reusable framework in case you need to share this data with other teams in the future. What should you do?

A. Create authorized views in the team's Google Cloud project that is only accessible by the team.
B. Create a private exchange using Analytics Hub with data egress restriction, and grant access to the team members.
C. Enable domain restricted sharing on the project. Grant the team members the BigQuery Data Viewer IAM role on the dataset.
D. Export the dataset to a Cloud Storage bucket in the team's Google Cloud project that is only accessible by the team.

EXPOSE CORRECT ANSWER

Answer : **B**

NEXT QUESTION

Talk to us!

Have any questions or issues ? Please dont hesitate to contact us

support@certlibrary.com

Certlibrary.com is owned by MBS Tech Limited: Room 1905 Nam Wo Hong Building, 148 Wing Lok Street, Sheung Wan, Hong Kong. Company registration number: 2310926

Certlibrary doesn't offer Real Microsoft Exam Questions. Certlibrary Materials do not contain actual questions and answers from Cisco's Certification Exams.

CFA Institute does not endorse, promote or warrant the accuracy or quality of Certlibrary. CFA® and Chartered Financial Analyst® are registered trademarks owned by CFA Institute.

Terms & Conditions | Privacy Policy

Question #:1

Your organization has decided to move their on-premises Apache Spark-based workload to Google Cloud. You want to be able to manage the code without needing to provision and manage your own cluster. What should you do?

A. Migrate the Spark jobs to Dataproc Serverless.

B. Configure a Google Kubernetes Engine cluster with Spark operators, and deploy the Spark jobs.

C. Migrate the Spark jobs to Dataproc on Google Kubernetes Engine.

D. Migrate the Spark jobs to Dataproc on Compute Engine.

Answer: A

Explanation

Migrating the Spark jobs toDataproc Serverlessis the best approach because it allows you to run Spark workloads without the need to provision or manage clusters. Dataproc Serverless automatically scales resources based on workload requirements, simplifying operations and reducing administrative overhead. This solution is ideal for organizations that want to focus on managing their Spark code without worrying about the underlying infrastructure. It is cost-effective and fully managed, aligning well with the goal of minimizing cluster management.

Question #:2

Your company has developed a website that allows users to upload and share video files. These files are most frequently accessed and shared when they are initially uploaded. Over time, the files are accessed and shared less frequently, although some old video files may remain very popular. You need to design a storage system that is simple and cost-effective. What should you do?

A. Create a single-region bucket with custom Object Lifecycle Management policies based on upload date.

B. Create a single-region bucket with Autoclass enabled.

C. Create a single-region bucket. Configure a Cloud Scheduler job that runs every 24 hours and changes the storage class based on upload date.

D. Create a single-region bucket with Archive as the default storage class.

Answer: B

Explanation

The storage system must balance cost, simplicity, and access patterns: high initial access, decreasing over time, with some files remaining popular. Google Cloud Storage offers tailored options for this:

- **Option A**: Custom Object Lifecycle Management (OLM) policies (e.g., transition to Nearline after 30 days, Archive after 90 days) are effective but static. They don't adapt to actual usage, so popular old files in Archive would incur high retrieval costs.

- **Option B**: Autoclass automatically adjusts storage classes (Standard, Nearline, Coldline, Archive) based on object access patterns, not just age. It keeps frequently accessed files in Standard (low latency /cost for access) and moves inactive ones to cheaper classes, minimizing costs while preserving simplicity. This fits the "some files remain popular" nuance.

- **Option C**: A Cloud Scheduler job to manually change classes daily is complex (requires scripting, monitoring), error-prone, and less cost-effective than automated solutions like Autoclass or OLM.

Question #:3

You have a BigQuery dataset containing sales data. This data is actively queried for the first 6 months. After that, the data is not queried but needs to be retained for 3 years for compliance reasons. You need to implement a data management strategy that meets access and compliance requirements, while keeping cost and administrative overhead to a minimum. What should you do?

A. Use BigQuery long-term storage for the entire dataset. Set up a Cloud Run function to delete the data from BigQuery after 3 years.

B. Partition a BigQuery table by month. After 6 months, export the data to Coldline storage. Implement a lifecycle policy to delete the data from Cloud Storage after 3 years.

C. Set up a scheduled query to export the data to Cloud Storage after 6 months. Write a stored procedure to delete the data from BigQuery after 3 years.

D. Store all data in a single BigQuery table without partitioning or lifecycle policies.

Answer: B

Explanation

Partitioning the BigQuery table by month allows efficient querying of recent data for the first 6 months, reducing query costs. After 6 months, exporting the data toColdline storageminimizes storage costs for data that is rarely accessed but needs to be retained for compliance. Implementing a lifecycle policy in Cloud Storage automates the deletion of the data after 3 years, ensuring compliance while reducing administrative overhead. This approach balances cost efficiency and compliance requirements effectively.

Question #:4

Your organization needs to store historical customer order data. The data will only be accessed once a month for analysis and must be readily available within a few seconds when it is accessed. You need to choose a storage class that minimizes storage costs while ensuring that the data can be retrieved quickly. What should you do?

A. Store the data in Cloud Storaqe usinq Nearline storaqe.

B. Store the data in Cloud Storaqe usinq Coldline storaqe.

C. Store the data in Cloud Storage using Standard storage.

D. Store the data in Cloud Storage using Archive storage.

Answer: A

Explanation

UsingNearline storagein Cloud Storage is the best option for data that is accessed infrequently (such as once a month) but must be readily available within seconds when needed. Nearline offers a balance between low storage costs and quick retrieval times, making it ideal for scenarios like monthly analysis of historical data. It is specifically designed for infrequent access patterns while avoiding the higher retrieval costs and longer access times of Coldline or Archive storage.

Question #:5

You are responsible for managing Cloud Storage buckets for a research company. Your company has well-defined data tiering and retention rules. You need to optimize storage costs while achieving your data retention needs. What should you do?

A. Configure the buckets to use the Archive storage class.

B. Configure a lifecycle management policy on each bucket to downgrade the storage class and remove objects based on age.

C. Configure the buckets to use the Standard storage class and enable Object Versioning.

D. Configure the buckets to use the Autoclass feature.

Answer: B

Explanation

Configuring alifecycle management policyon each Cloud Storage bucket allows you to automatically transition objects to lower-cost storage classes (such as Nearline, Coldline, or Archive) based on their age or other criteria. Additionally, the policy can automate the removal of objects once they are no longer needed, ensuring compliance with retention rules and optimizing storage costs. This approach aligns well with well-defined data tiering and retention needs, providing cost efficiency and automation.

Extract from Google Documentation: From "Object Lifecycle Management" (https://cloud.google.com /storage/docs/lifecycle):*"Use lifecycle management policies to automatically transition objects to lower-cost storage classes (e.g., Nearline, Coldline) and delete them based on age, optimizing costs according to your specific tiering and retention requirements."*

Question #:6

Following a recent company acquisition, you inherited an on-premises data infrastructure that needs to move to Google Cloud. The acquired system has 250 Apache Airflow directed acyclic graphs (DAGs) orchestrating data pipelines. You need to migrate the pipelines to a Google Cloud managed service with minimal effort. What should you do?

A. Convert each DAG to a Cloud Workflow and automate the execution with Cloud Scheduler.

B. Create a new Cloud Composer environment and copy DAGs to the Cloud Composer dags/ folder.

C. Create a Google Kubernetes Engine (GKE) standard cluster and deploy Airflow as a workload. Migrate all DAGs to the new Airflow environment.

D. Create a Cloud Data Fusion instance. For each DAG, create a Cloud Data Fusion pipeline.

Answer: B

Explanation

Comprehensive and Detailed in Depth Explanation:

Why B is correct:Cloud Composer is a managed Apache Airflow service that provides a seamless migration path for existing Airflow DAGs.

Simply copying the DAGs to the Cloud Composer folder allows them to run directly on Google Cloud.

Why other options are incorrect:A: Cloud Workflows is a different orchestration tool, not compatible with Airflow DAGs.

C: GKE deployment requires setting up and managing a Kubernetes cluster, which is more complex.

D: Cloud Data Fusion is a data integration tool, not suitable for orchestrating existing pipelines.

Question #:7

Your company uses Looker as its primary business intelligence platform. You want to use LookML to visualize the profit margin for each of your company's products in your Looker Explores and dashboards. You need to implement a solution quickly and efficiently. What should you do?

A. Apply a filter to only show products with a positive profit margin.

B. Define a new measure that calculates the profit margin by using the existing revenue and cost fields.

C. Create a new dimension that categorizes products based on their profit margin ranges (e.g., high, medium, low).

D. Create a derived table that pre-calculates the profit margin for each product, and include it in the Looker model.

Answer: B

Explanation

Comprehensive and Detailed in Depth Explanation:

Why B is correct:Defining a new measure in LookML is the most efficient and direct way to calculate and visualize aggregated metrics like profit margin.

Measures are designed for calculations based on existing fields.

Why other options are incorrect:A: Filtering doesn't calculate or visualize the profit margin itself.

C: Dimensions are for categorizing data, not calculating aggregated metrics.

D: Derived tables are more complex and unnecessary for a simple calculation like profit margin, which can be done using a measure.

Question #:8

Your organization’s ecommerce website collects user activity logs using a Pub/Sub topic. Your organization’s leadership team wants a dashboard that contains aggregated user engagement metrics. You need to create a solution that transforms the user activity logs into aggregated metrics, while ensuring that the raw data can be easily queried. What should you do?

A. Create a Dataflow subscription to the Pub/Sub topic, and transform the activity logs. Load the transformed data into a BigQuery table for reporting.

B. Create an event-driven Cloud Run function to trigger a data transformation pipeline to run. Load the transformed activity logs into a BigQuery table for reporting.

C. Create a Cloud Storage subscription to the Pub/Sub topic. Load the activity logs into a bucket using the Avro file format. Use Dataflow to transform the data, and load it into a BigQuery table for reporting.

D. Create a BigQuery subscription to the Pub/Sub topic, and load the activity logs into the table. Create a materialized view in BigQuery using SQL to transform the data for reporting

Answer: A

Explanation

UsingDataflowto subscribe to the Pub/Sub topic and transform the activity logs is the best approach for this scenario. Dataflow is a managed service designed for processing and transforming streaming data in real time. It allows you to aggregate metrics from the raw activity logs efficiently and load the transformed data into a BigQuery table for reporting. This solution ensures scalability, supports real-time processing, and enables querying of both raw and aggregated data in BigQuery, providing the flexibility and insights needed for the dashboard.

Question #:9

You are working on a data pipeline that will validate and clean incoming data before loading it into BigQuery for real-time analysis. You want to ensure that the data validation and cleaning is performed efficiently and can handle high volumes of data. What should you do?

A. Write custom scripts in Python to validate and clean the data outside of Google Cloud. Load the cleaned data into BigQuery.

B. Use Cloud Run functions to trigger data validation and cleaning routines when new data arrives in Cloud Storage.

C. Use Dataflow to create a streaming pipeline that includes validation and transformation steps.

D. Load the raw data into BigQuery using Cloud Storage as a staging area, and use SQL queries in BigQuery to validate and clean the data.

Answer: C

Explanation

UsingDataflowto create a streaming pipeline that includes validation and transformation steps is the most efficient and scalable approach for real-time analysis. Dataflow is optimized for high-volume data processing and allows you to apply validation and cleaning logic as the data flows through the pipeline. This ensures that only clean, validated data is loaded into BigQuery, supporting real-time analysis while handling high data volumes effectively.

Question #:10

Your team needs to analyze large datasets stored in BigQuery to identify trends in user behavior. The analysis will involve complex statistical calculations, Python packages, and visualizations. You need to recommend a managed collaborative environment to develop and share the analysis. What should you recommend?

A. Create a Colab Enterprise notebook and connect the notebook to BigQuery. Share the notebook with your team. Analyze the data and generate visualizations in Colab Enterprise.

B. Create a statistical model by using BigQuery ML. Share the query with your team. Analyze the data and generate visualizations in Looker Studio.

C. Create a Looker Studio dashboard and connect the dashboard to BigQuery. Share the dashboard with your team. Analyze the data and generate visualizations in Looker Studio.

D. Connect Google Sheets to BigQuery by using Connected Sheets. Share the Google Sheet with your team. Analyze the data and generate visualizations in Gooqle Sheets.

Answer: A

Explanation

Using a Colab Enterprise notebook connected to BigQuery provides a managed, collaborative environment ideal for complex statistical calculations, Python packages, and visualizations. Colab Enterprise supports Python libraries for advanced analytics and offers seamless integration with BigQuery for querying large datasets. It allows teams to collaboratively develop and share analyses while taking advantage of its visualization capabilities. This approach is particularly suitable for tasks involving sophisticated computations and custom visualizations.