Level: Intermediate

Google Cloud Certified Professional Data Engineer

← Back to the **Course**

Practice Test 4

Completed on **Sat, 06 Jul 2024**

2nd	30/31	96.77%
Attempt	Marks Obtained	Your Score

0h 49m 36s	PASS
Time Taken	Result

Share this Report in Social Media Share

Download Report

Domain wise Quiz Performance Report

No.	Domain	Total Question	Correct	Incorrect	Unattempted
1	Design Data Processing Systems	6	6	0	0
2	Store the data	3	3	0	0
3	Ingest and process the data	2	2	0	0
4	Prepare and use data for analysis	1	1	0	0
5	Design Data Processing Systems	19	18	1	0
Total	All Domains	31	30	1	0

Review the Answers

Filter By All Questions

Question 1

Correct

Domain: Other

A company has cloud data flow pipeline currently running in production. They have few changes in the pipeline to fix current bugs. Also, some of the transformation names have been updated for code ethics and therefore current data flow job needs to be updated. How should the cloud data flow job update?

A. Stop the existing pipeline. Replace the pipeline code for a prior job with the new updated code. Start the jobs again after code replacement.

B. Create another job with the same job name as the prior job. The previous job will complete after processing your inflight data. Once the previous job is completed, the new job will come into effect.

C. Create another job with the same job name as the prior job. Also, pass the –update option. The previous job will complete after processing your inflight data. Once the previous job is completed, the new job will come into effect.

D. Create another job with the same job name as the prior job. Also, pass the –update option. User needs to provide transform mapping using –transformNameMapping option. The replacement job preserves any intermediate state data from the prior job, as well as any buffered data records or metadata currently "in-flight" from the prior job. "In-flight" data will then be processed by the transforms in your new pipeline. right

Explanation:

The correct answer is option D

As per Google Cloud official documentation (https://cloud.google.com/dataflow/docs/guides/updating- a-pipeline)

When you update a job on the Cloud Dataflow service, you **replace** the existing job with a new job that runs your updated pipeline code. The Cloud Dataflow service retains the job name but runs the replacement job with an updated job.

The replacement job preserves any intermediate state data from the prior job, as well as any buffered data records or metadata currently "in-flight" from the prior job. For example, some records in your pipeline might be buffered while waiting for a window to resolve.

"In-flight" data will still be processed by the transforms in your new pipeline. However, additional transforms that you add in your replacement pipeline code may or may not take effect, depending on where the records are buffered.

Option A is incorrect. For updating an existing pipeline, a new pipeline needs to be created with the same job name.

Option B is incorrect. For updating an existing pipeline, a new pipeline needs to be created with the same job name, and also –update option should be passed. Also, as transform names are updated in the new pipeline, hence –transformNameMapping needs to be supplied while creation of this new job.

Option C in incorrect. For updating an existing pipeline, a new pipeline needs to be created with the same job name and also –update option should be passed. Also, as transform names are updated in the new pipeline, hence –transformNameMapping needs to be supplied while creation of this new job.

Option D is correct. For updating an existing pipeline, a new pipeline needs to be created with the same job name and also –update option should be passed. Also, as transform names are updated in the new pipeline, hence –transformNameMapping needs to be supplied while creation of this new job. This mapping will contain old transform jobs mapped to their new transform job. Eg. --transformNameMapping={"oldTransform1":"newTransform1","oldTransform2":"newTransform2",...}

Ask our Experts

Did you like this **Question?**

Question 2 Correct

Domain: Other

Your company has TB's of data stored on BigQuery. They have existing spark scripts for performing transformation and analysis on Cloud Data Proc. The output needs to be stored in BigQuery for future analysis. How can you set up BigQuery as an input and output source? Select 2 correct answers

A. Manually use a cloud storage bucket to import and export to and from both BigQuery and DataProc.

B. Install the BigQuery connector on your DataProc cluster. right

C. Only Cloud Storage and HDFS can be used as DataProc input and output.

D. Specify the BigQuery connector in the jars parameter when submitting a job. right

Explanation:

The correct answers are B & D.

BigQuery connector can be used as input and output source for DataProc cluster by using BigQuery connector. There are three ways BigQuery connectors can be used.

By installing BigQuery connector using initialization action. This will install the Big Query connector while the cluster starts.

Be specifying the BigQuery connector in the jars parameter when submitting a job. The jar could be placed on cloud storage and a path to the jar is provided.

BigQuery connector classes can be included as dependencies in your code.

Option A is incorrect. Big Query connector can be used to set up input and output for BigQuery on your Dataproc cluster.

Option B is correct. Big Query connector can be installed on Dataproc cluster using initialization action.

Option C is incorrect. Using a BiqQuery connector, Dataproc can use BigQuery as input and output source.

Ref URL: https://cloud.google.com/dataproc/docs/tutorials/bigquery-connector-spark-example

Option D is correct. Users can specify the BiqQuery connector in the jars parameter when submitting a job.

For eg: --jars=gs://hadoop-lib/bigquery/bigquery-connector-hadoop2-latest.jar

Ask our Experts

Did you like this **Question?**

Question 3

Incorrect Marked for review

Domain: Other

Your company is building storage for data pipelines. The input files are in JSON, the schema for th
JSON files can occasionally change. Once data is ingested into the Google cloud, the data scientist of

the company will be querying this data using ANSI SQL.

How should the data be ingested considering the above needs?

A. Use BigQuery for storage. Select “Automatically Detect” in the schema section. right

B. Use cloud storage for storage. Create Big Query temporary external data source tables and turn on the “Automatically Detect” option in the schema section of Big Query.

C. Use cloud storage for storage. Create permanent external data source tables and turn on the “Automatically Detect” option in the schema section of Big Query. wrong

D. Use BigQuery for storage. Provide format files for data load. Update the format files as needed.

Explanation:

The correct answer is A

Refer GCP documentation – BigQuery Auto Detection:- https://cloud.google.com/bigquery/docs/schema-detect;

Schema auto-detection is available when you load data into BigQuery, and when you query an external data source.

When auto-detection is enabled, BigQuery starts the inference process by selecting a random file in the data source and scanning up to 100 rows of data to use as a representative sample. BigQuery then examines each field and attempts to assign a data type to that field based on the values in the sample.

Option A is correct. As the requirement is to use ANSI SQL for querying purposes and to support occasional schema-changing JSON files; BigQuery will be an appropriate choice. Schema auto-detection is available in BigQuery when you load the data in BigQuery or if your query is from an external data source.

Option B & C is incorrect. Using cloud storage for storage is not ideal for this purpose as it would add latency and would be cumbersome.

Option D is incorrect. Providing format files is not correct, we can simply turn on the ‘Automatically Detect’ schema changes flag.

Ask our Expe

Did you like this **Question?**

Question 4 Correct Marked for review

Domain: Other

Your company has stored data in Big Query Avro format. You need to export this Avro formatted data to a cloud storage bucket. The data exported in cloud storage should also be in Avro format. Which of the following options is the best way to achieve the given requirement?

A. Convert the data in CSV format using the BigQuery export option, specify the cloud storage bucket as the destination.

B. Create spark job using cloud Dataproc, input to this job will be BigQuery. This job will fetch the data from BigQuery in CSV format and transform the data in Avro format before storing into cloud storage.

C. Use the BigQuery transfer service to transfer data in AVRO format in cloud storage.

D. Use the export table option in BigQuery, specify the format as Avro and destination as cloud storage bucket. right

Explanation:

Answer - D

Once you've loaded your data into BigQuery, you can export the data in several formats. BigQuery can export up to 1 GB of data to a single file. If you are exporting more than 1 GB of data, you must export your data to multiple files.

Option A is incorrect. The requirement is to export the data in Avro format in cloud storage and not in CSV format.

Option B is incorrect. Creating a spark job using cloud Dataproc is not correct as it would increase unnecessary complexity. We can directly export data in Avro format using the web console of BigQuery.

Option C is incorrect. BigQuery transfer service is used to automate data movement from Google application sources like Google Ads, Campaign Manager, Google Ad Manager and Youtube on a scheduled or managed basis.

Option D is correct. BiqQuery has native support for Avro format export to cloud storage. The only

supported export location is Google Cloud Storage. Data can be exported in CSV, JSON and AVRO format.

References:

https://cloud.google.com/bigquery/docs/exporting-data

https://cloud.google.com/bigquery-transfer/docs/cloud-storage-transfer

Ask our Experts

Did you like this **Question?**

Question 5 Correct

Domain: Other

Your company is building a multi-cloud data pipeline. Few of the batch jobs will be running on AWS cloud, upon completion of these batches Cloud Dataproc jobs should be triggered. You need to build an orchestration process to support multi-cloud dependencies. Which google cloud product can be used for this purpose?

A. Cloud Scheduler

B. Cloud Composer right

C. Apache Airflow

D. Oozie on Cloud Dataproc.

Explanation:

Correct Answer - B

Option A is incorrect. Cloud Schedule is fully managed cron job scheduler. It can virtually schedule any job, including batch, big data jobs. User can run their batch and big data jobs on a recurring schedule to make them more reliable and reduce manual toil.

Option B is correct. Cloud Composer is a fully managed workflow orchestration service that empowers users to schedule and monitor pipelines that span across clouds and on-premises data centres. As cloud composer is built on popular Apache Airflow, it can be used in this scenario to build dependencies from AWS cloud.

Option C is incorrect. Cloud Composer is built on Apache Airflow and is a fully managed workflow orchestration service. It is recommended to use Cloud Composer rather than Apache Airflow.

Option D is incorrect. Using Oozie on Cloud Dataproc does not support multi-cloud dependencies. Also, Cloud composer is a better option in this scenario due to its many benefits and is fully managed.

Ask our Experts

Did you like this **Question?**

Question 6 Correct

Domain: Other

You want to migrate your on-premise NoSQL database to Google Cloud Bigtable. After migration of your complete data, you are not getting the desired performance as per Bigtable standards. What could be the reasons for the performance issue? Select all that apply.

- [x] **A. The table schema is not designed correctly.** right
- [x] **B. Give some time to Bigtable to balance your data.** right
- [ ] **C. Adding nodes to a cluster and checking cluster performance after 1 hour.**
- [x] **D. Cloud Bigtable instance uses HDD disks.** right
- [ ] **E. Cloud Bigtable instance is a production instance.**

Explanation:

Correct answer is A,B and D.

Option A is correct. For getting good performance from Cloud Bigtable, it is important to design table schema efficiently to distribute reads and writes and avoid hot spotting. Selection of row_key is of very important and crucial for performance.

Option B is correct. If you immediately test Cloud Bigtable after loading of data, performance will be not as per standard. Cloud Bigtable won't be able to balance your data. It needs time to learn pattern from your data and create large enough shards for effectively use all of the nodes in your cluster.

Option C is incorrect. After addition of nodes it takes approximately 20 minutes before you see

performance improvement. In the option user query the data after 1 hour, so the performance improvement must be done by then.

Option D is correct. SDD disk has better performance that HDD disk. It is recommended to use SDD disk for performance sensitive applications.

Option E is incorrect. As Cloud Bigtable instance is already a production instance, there is no scope of further improvement by changing the instance type. Also, development instance has less performance than production instance. A development instance has a performance equivalent to one single node cluster.

Ask our Experts

Did you like this **Question?**

Question 7 Correct

Domain: Other

Your company has configured a streaming data pipeline, data from thousands of Internet of Things devices is processed using Pub-Sub and dataflow and then is ingested in BigQuery in real-time. The data in big query is stored into an ingestion-time partitioned table. You want to run SQL queries against your data for analysis. How would you run SQL queries against a particular partition?

A. Use the DATE column in the WHERE clause to filter results for a particular date.

B. Use the _PARTITIONTIME pseudo-column in the WHERE clause right

C. Use the column DAY_TIMESTAMP created during ingestion, filter the content of the table by applying the WHERE clause on this column

D. Use the _PARTITION_TIME pseudo-column in the WHERE clause

Explanation:

The correct answer is B

Option A is incorrect. As the table is an ingestion-time partitioned table, pseudo-column _PARTITTIONTIME is c+reated. No DATE column exists in the big query table for this data set.

Option B is correct. Ingestion-time partitioned table contains a pseudo column _PARTITTIONTIME. Refer

below explanation

When you create an ingestion-time partitioned table, two pseudo columns are added to the table: a

_PARTITIONTIME pseudo column and a _PARTITIONDATE pseudo column. The _PARTITIONTIME pseudo column contains a date-based timestamp for data that is loaded into the table. The _PARTITIONDATE pseudo column contains a date representation. Both pseudo column names are reserved, which means that you cannot create a column with either name in any of your tables.

_PARTITIONTIME and _PARTITIONDATE are available only in ingestion-time partitioned tables. Partitioned tables do not have pseudo columns. The _PARTITIONTIME pseudo column

The _PARTITIONTIME pseudo column contains a timestamp that is based on UTC time and represents the number of microseconds since the Unix epoch. For example, if data is appended to a table on April 15, 2016, all of the rows of data that are appended on that day contain the value TIMESTAMP("2016-04-15") in the _PARTITIONTIME column.

Option C is incorrect. As the table created, is an Ingestion-time partitioned table only two pseudo columns are created _PARTITIONTIME and _PARTITIONDATE. DAY_TIMESTAMP columns that do not exist are not created by the ingestion time partitioned table.

Option D is incorrect. Ingestion-time partitioned table creates two pseudo columns _PARTITIONTIME and _PARTITIONDATE. The syntax of the pseudo column _PARTITION_TIME is not correct.

Ask our Experts

Did you like this **Question?**

Question 8

Correct

Domain: Other

An application has the following requirement

1. It required strongly consistent
2. Total data as of now is 1GB, which can grow significantly in future therefore horizontal scaling, is required
3. It should be highly available.

Currently, this data is stored in a highly available relational database with a very high memory configuration, so performance is more desired than cost-benefit. Which data technology would be the best fit as per Google's recommendation?

A. Big Query

B. Big Table

C. Cloud SQL

D. Cloud Spanner right

Explanation:

The correct answer is D

Option A is incorrect. BigQuery is wrong, as it does not support consistent transactions. BigQuery is mostly used for analytical purposes.

Option B is incorrect. BigTable is wrong, as BigTable is not a relational database service.

Option C is incorrect. Although Cloud SQL support strong consistency and is highly available but due to horizontal scaling required by the customer Cloud SQL will not be an appropriate choice. Cloud SQL can only store up to 10TB of data and will not be able to scale for future needs.

Option D is correct. Cloud Spanner is a strongly consistent relational database which does not have a limitation on data storage. Also, customer needs performance over cost benefits so cloud spanner could the best choice in this case.

Cloud Spanner is the first scalable, enterprise-grade, globally distributed, and strongly consistent database service built for the cloud specifically to combine the benefits of relational database structure with non-relational horizontal scale. This combination delivers high-performance transactions and strong consistency across rows, regions, and continents with an industry-leading 99.999% availability SLA.

Ask our Experts

Did you like this **Question?**

Question 9 Correct Marked for review

Domain: Other

Your company has few spark jobs that are migrated to Cloud DataProc. The spark jobs need more nodes for processing the data so you are exploring the option of Preemptible workers to increase the performance of the jobs. Which of the below statement(s) regarding Preemptible workers is true?

A. In order to save cost, users can run the spark job using only Preemptible workers for non-critical batch jobs.

B. Preemptible workers cannot use the persistent disks.

C. Preemptible workers cannot store data. right

D. If a preemptible worker is reclaimed, then a replacement worker must be added manually.

Explanation:

Correct Answer - C

Option A is incorrect. To ensure Cluster does not lose all workers, Cloud Dataproc cannot create preemptible-only clusters. If the user does not specify the number of standard workers while creating a cluster, Cloud Dataproc will automatically add two non-preemptible workers to the cluster,

Option B is incorrect. Preemptible nodes can have persistent disks.

Option C is correct. Since preemptible can be reclaimed at any time, preemptible workers do not store data. Preemptible added to a Cloud Dataproc cluster only function as processing nodes.

Option D is incorrect. Cloud Dataproc handles the addition and removal of preemptible nodes.

Reference -

https://cloud.google.com/dataproc/docs/concepts/compute/preemptible-vms

Ask our Experts

Did you like this **Question?**

Question 10 Correct

Domain: Other

An online home product company wants to build a website where users can search for products by simply uploading the image of their product. The search engine will then match the product uploaded by the user with the feature of products the company has. The website will then show all the products to the user which match the product user has queried. Which below solution can be implemented efficiently in less time?

A. Train a tensor flow model by supplying all the images of the product company. This was model will learn from the feature of the product. This tensor flow model can be hosted on Google AI Platform.

B. Use Cloud Image Intelligence API and develop a REST based wrapper to fetch out results based on the user-uploaded images.

C. Create a deep neural network and train the models with the sample images. Train and test the model on sample data and minimize the loss function. Host the model on the Google Cloud AI Platform.

D. Use Cloud Vision API to create a product set each containing reference images. API can then be used to fetch out the matching product based on user input. right

Explanation:

Correct Answer is D

Option A is incorrect. Training and validating the tensor flow model could take more amount of time. As mentioned in the question we need to choose a solution that is both efficient and can be delivered in less amount of time.

Option B is incorrect - There is no service called Cloud Image Intelligence API. For this purpose, we should use Cloud Vision API.

Option C is incorrect. Training and validating a custom deep neural network models could take more amount of time. As mentioned in the question we need to choose a solution that is both efficient and can be delivered in less amount of time.

Option D is correct. Using Cloud Vision API we can easily train by providing reference images. The API will allow the user to query the product catalog based upon the new image as input and will fetch the best matching products.

Refer GCP documentation

https://cloud.google.com/vision/product-search/docs/tutorial

Ask our Experts

Did you like this **Question?**

Question 11

Correct

Domain: Design Data Processing Systems

You are a data engineer for your company. The company has chosen to use a serverless data warehouse service on Google Cloud. You are asked to analyze credit-card transactions in a Big query that currently sits in the Google Cloud SQL service. Select the best possible option provided by GCP to get your Cloud SQL data into Google Big Query.

A. In Cloud SQL, select Big Query as target and write all the data

B. Use the BigQuery connector in Cloud SQL, to export from Cloud SQL to Big Query.

C. In Big Query, use an external data source - Cloud SQL and establish the connection to get the

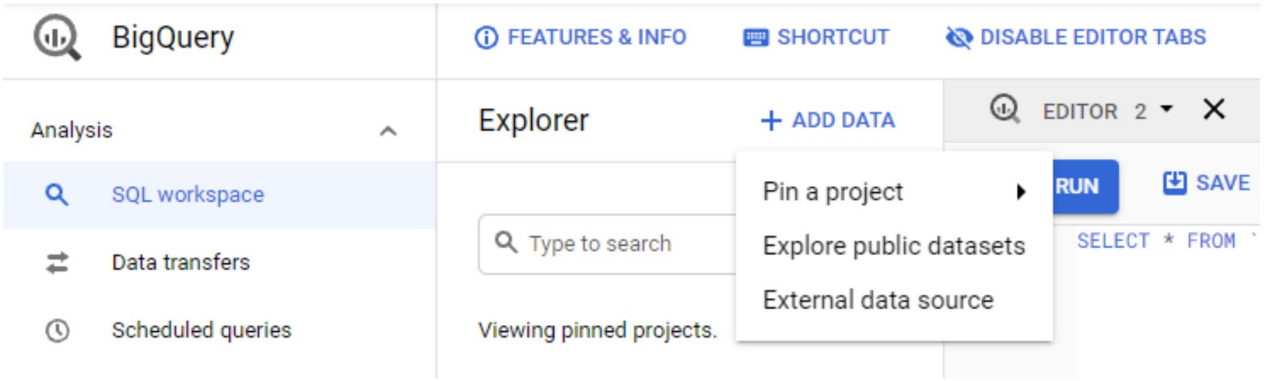
data. right

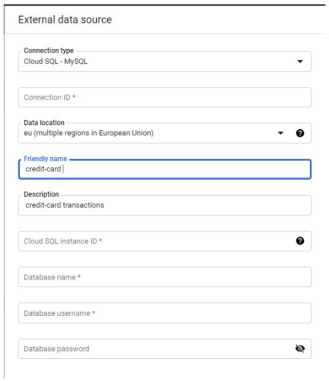
D. Create a Dataflow Job to copy data from Cloud SQL to BigQuery

Explanation:

Correct Answer: C

Option C is CORRECT because to get data from Cloud SQL, the best way to use the external data source option and select Cloud SQL and provide all the details regarding the database.





Option A is incorrect because we don't have any such option in cloud SQL.

Option B is incorrect because we don't have a big query connector to transfer data from Cloud SQL to Big Query.

For more information on the **Big Query**, please visit the below URL:

https://cloud.google.com/blog/topics/developers-practitioners/bigquery-explained-data-ingestion

Ask our Expe

Did you like this **Question?**

Question 12 Correct

Domain: Other

Data engineers of an ecommerce company wants to build and train an ML model to weigh customer feedback on a product but the management don't want to disclose the information regarding who submitted the feedback. Some of the information like delivery address and purchase history are critical for training of ML model. After the data is available your data exploration team needs to query the data so it is important to protect the sensitive data fields. The data is unstructured text-based dataset. Identify the best possible solution that can be quickly deployed:

A. Analyze the data using Cloud Data Prep, identify the sensitive data and remove sensitive data from the dataset. Create the recipe for the same. Once done Cloud Data Flow will be triggered which will store the data in Big Query for further analysis.

B. Use Google Cloud Data Loss Prevention API to identify sensitive information and mask the same before storing it for analysis. right

C. Analyze the data using Cloud Data Prep, identify the sensitive data and mask sensitive data from the dataset. Create the recipe for the same. Once done Cloud Data Flow will be triggered which will store the data in Big Query for further analysis.

D. Create a machine learning model to identify sensitive information based on past training. This model will analyze the data and will mask the same before supplying to data exploration team.

Explanation:

Correct answer is B.

Option A is incorrect. Building a solution using cloud data prep require some manual intervention and analysis. Identifying sensitive data hidden in unstructured data could be tricky and can lead to mistakes. Also, in the option removing sensitive data should not be done. Solution should encrypt or mask the sensitive data rather than completely removing from the dataset.

Option B is correct. Cloud DLP identify the data using more than 90 predefined detectors to identify patterns, format and checksums. Using cloud DLP sensitive data can be easily identified by the algorithm, Also, the algorithm can mask the data based on user input.

This solution can be easily deployed and will be most accurate.

Option C is incorrect. Building a solution using cloud data prep require some manual intervention and analysis. Identifying sensitive data hidden in unstructured data could be tricky and can lead to mistakes.

Option D is incorrect. Creating a custom machine learning model will take significant amount of time. Also, training the model will take lots of input. As sensitive data analysis can be a bit tricky and require lots of training for model. Using CLoud DLP can be an effective solution.

Ask our Experts

Did you like this **Question?**

Question 13 Correct

Domain: Other

Your company is migrating their 40 nodes Apache Hadoop cluster to the cloud. Company has few Spark and Pig jobs that they have already created and want to re-run the same on Google Cloud. Also, the company wants to minimize the management of cluster as much as possible. The data needs to be persisted beyond the life of the cluster. What should you do?

A. Create Cloud Dataflow job to process the data.

B. Create Hadoop cluster on Google Compute engine and use Local SSD disk to persist the data.

C. Create Cloud Dataproc cluster and use persistent disk for HDFS.

D. Use ephemeral Dataproc cluster with preemptible VMs to process the data and Store data in Google Cloud Storage with object lifecycle management policy. right

Explanation:

Correct answer is D.

Option A is incorrect. Dataflow job is not suited for Hadoop jobs.

Option B is incorrect. Creating Hadoop cluster on compute engine would increase infrastructure management cost. Also, persistent disks would not provide scalability.

Option C is incorrect. As Dataproc cluster is associated with persistent disk for HDFS, if the cluster terminated the data would be lost.

Option D is Correct. As the requirement is to reuse Spark and Hive jobs with minimizing the infrastructure management with the ability to store data in a durable external storage, Dataproc with cloud storage would be an ideal solution.

Ask our Experts

Did you like this **Question?**

Question 14 Correct

Domain: Design Data Processing Systems

A regional auto dealership is migrating its business applications to Google Cloud. The CTO of this company asked their data engineer to find the possible ways you can ingest data into BigQuery?

A. Batch Ingestion & Streaming Ingestion

B. Data Transfer Service

C. Query Materialization

D. Partner Integrations

E. All of the above right

Explanation:

Correct Answer: E

Option E is CORRECT because these all are the possible ways to load data into BigQuery. Lets understand these one by one -

1. Batch Ingestion - This involves ingesting large, bounded datasets that don't have to be processed in real-time.

To implement batch ingestion, one can use Google Cloud Storage, Cloud Dataflow, Cloud Dataproc, Cloud Data Fusion etc.

1. Stream Ingestion - This involves ingesting large, unbounded data that is processed in real-time.

To implement Stream Ingestion, one can use Apache Kafka to BQ connector, Cloud Pub/Sub, Cloud

Dataflow, Cloud Dataproc etc.

1. Data Transfer Service (DTS) - This is a fully managed service to load data from external cloud storage providers such as Amazon S3, Google SaaS applications such as Google Ads, and transferring data from data warehouse technologies such as Teradata etc.
2. Partner Integrations - These are the data integration alternatives from Google Cloud Partners.

This includes, Confluent, Informatica, snapLogic, Talend and many more.

1. Query Materialization - This is the best way to simplify extract, transform and load patterns in BigQuery. Using federated queries in BigQuery, one can persist their analysis results in Big Query to derive any insights.

Please find the image attached below having details about all the possible ways to ingest data into BigQuery.

For more information on the **Big Query**, please visit the below URL:

https://cloud.google.com/blog/topics/developers-practitioners/bigquery-explained-data-ingestion

Ask our Experts

Did you like this **Question?**

Question 15

Correct

Domain: Other

Choose all statement(s) which is/are correct for cloud pub/sub.

- [x] **A. Cloud PubSub has a default subscription message retention period of 7 days.** right

B. The subscription message retention period for cloud pubsub is configurable and can be configured to a maximum of 28 days and a minimum of 10 minutes.

C. The subscription message retention period for cloud pubsub subscription is configurable and can be configured to a maximum of 7 days and a minimum of 10 minutes. right

D. Subscription message retention period for cloud pub/sub is not configurable and is set to 7 days by default.

Explanation:

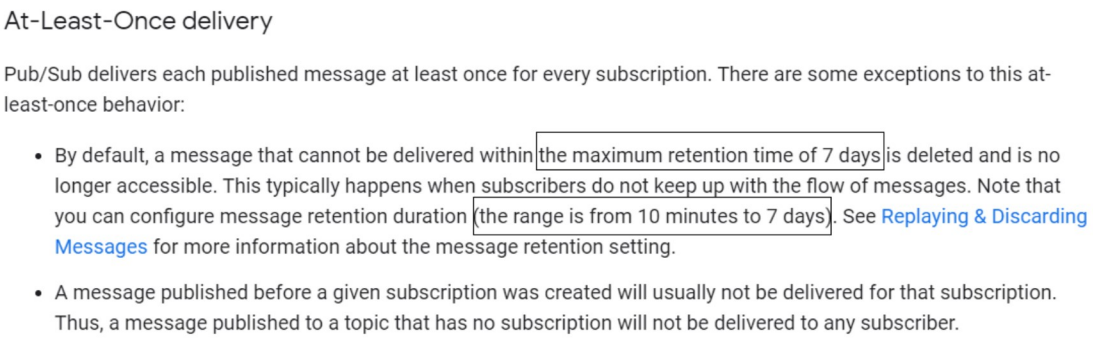
Correct Answer: A and C

Option A is correct. Cloud pub/sub has a default subscription message retention period of 7 days.

Option B is incorrect. Cloud pub/sub retention period is configurable and can be configured to a maximum of 7 days and a minimum of 10 minutes.

Option C is correct. Cloud pub/sub subscription message retention period is configurable and can be configured to a maximum of 7 days and a minimum of 10 minutes

Option D is incorrect. Cloud pub/sub retention period is configurable.



At-Least-Once delivery

Pub/Sub delivers each published message at least once for every subscription. There are some exceptions to this at-least-once behavior:

- By default, a message that cannot be delivered within the maximum retention time of 7 days is deleted and is no longer accessible. This typically happens when subscribers do not keep up with the flow of messages. Note that you can configure message retention duration (the range is from 10 minutes to 7 days). See Replaying & Discarding Messages for more information about the message retention setting.
- A message published before a given subscription was created will usually not be delivered for that subscription. Thus, a message published to a topic that has no subscription will not be delivered to any subscriber.

Reference:

https://cloud.google.com/pubsub/docs/subscriber

Ask our Experts

Did you like this **Question?**

Question 16 Correct

Domain: Other

Below are some statements regarding pub/sub delivery mechanism. Select the correct ones.

- [x] A. Pull based delivery is preferred when there is a large volume of messages. right
- [ ] B. Push based delivery is preferred when efficiency and throughput of message processing is critical.
- [x] C. In pull based delivery multiple subscriber can make pull calls to the same "shared" subscription. right
- [ ] D. Push based allows for batched delivery, therefore higher throughput can be achieved at low CPU.

Explanation:

Option A, C are correct.

Option A is correct. Pull based delivery is faster than push based delivery. Also, pull based can handle large volume of messages (more than 1/second).

Option B is incorrect. Pull based delivery is preferred if efficiency and throughput of message processing is required. Subscriber can call the pub/sub server for the message and pub/sub server respond with the message and ackId. In Push based delivery one message per request is sent.

Option C is correct. Multiple consumers/subscribers can subscribe and can make calls simultaneously in pull based model.

Option D is incorrect. One message per request is delivered in push based delivery. Batch delivery can be done in pull based model.

Ask our Experts

Did you like this **Question?**

Question 17 Correct Marked for review

Domain: Other

You are a data architect for your company. The company has chosen to use a managed database service on Google Cloud. Your existing database is used as a product catalogue that provides real-time inventory tracking. The data is semi-structured, also full atomicity is not a requirement. Current on-premise data is around 700GB in size. Choose the best serverless/no-ops solution for this purpose.

A. Cloud SQL.

B. Big Query

C. Cloud Datastore right

D. Cloud Big Table

Explanation:

Correct Answer - C

Options A and D are incorrect. These solutions are not completely No-Ops solutions. Cloud SQL is not suited for semi-structured data.

Option B is incorrect. Big Query is ideal for analytics solutions.

Option C is correct. Cloud Datastore offers NoOps, a NoSql solution suited for semistructured data and is also ideal for product catalogues.

Ask our Experts

Did you like this **Question?**

Question 18 Correct Marked for review

Domain: Other

Your company uses Google Analytics for tracking. You need to export the data from Google Analytics 360 on scheduled basis into BigQuery for further analysis. How can the data be exported?

A. Export Google Analytics data to cloud storage using gsutil, then import the data into Big Query.

B. Import data to BigQuery directly from Google Analytics using cron.

C. Use Cloud Scheduler to create job in Google Analytics to convert the data in AVRO format, import the data directly to BigQuery using bq command line.

D. Use BigQuery Data Transfer Service to import the data from Google Analytics. right

Explanation:

The correct answer is D.

Options A, B and C are incorrect. These options are not supported.

Option D is correct. BigQuery Data Transfer Service helps to automate importing of data from Google Analytics. Refer below GCP documentation

https://cloud.google.com/bigquery/docs/loading-data#data_transfer_service

Ask our Experts

Did you like this **Question?**

Question 19

Correct

Domain: Other

External users have joined your company, initially, these user needs to have the least privilege with the company. These users have been provided with the Cloud Dataproc Viewer role.

Which action can these external users perform?

A. Submit a Job.

B. List the jobs. right

C. Create a job.

D. Delete the cluster.

Explanation:

Correct Answer - B

Option B is correct. As Data Viewer role would allow users to only perform get and list operations. Users can list cloud Dataproc clusters, jobs, operations in a project.

Option A is incorrect as : The Viewer role does not include the necessary permissions to submit new

jobs. This requires higher privileges, such as the Editor or Job Creator roles.

Option C is incorrect as : Similar to submitting a job, creating a new job requires more extensive permissions than those granted by the Viewer role.

Option D is incorrect as : Deleting a cluster is a destructive action that requires elevated privileges, such as the Administrator role. The Viewer role is specifically designed for read-only operations.

Refer to the below documentation:

https://cloud.google.com/dataproc/docs/concepts/iam/iam

Ask our Experts

Did you like this **Question?**

Question 20 Correct

Domain: Other

You receive bank transactions data through Cloud Pub/Sub and you need to analyze the data using Cloud data flow. The transactions are in the below format:

2INDEL3465, JACK, 34627,DOLLAR,20191205234251000,D

1USCHG5627, SAM, 1276, DOLLAR, 20191205234252562,C

Currently the requirement is to extract customer name from the transaction and store the results in an output PCollection. Select the operation which is best suited for this processing.

A. Regex.find

B. Pardo right

C. Extract

D. Transform

Explanation:

Option B is correct.

Option A is incorrect. Regex.find will output Regex group containing all the lines that matches the

regex. In this case we need the customer name to be extracted and placed into another PCollection for further processing.

Option B is correct. As ParDo helps in extracting parts from elements. We can use ParDo for filtering a dataset. ParDo can be used to consider each element in PCollection and either output that element to a new collection or discard it.

Option C is incorrect. Extract option does not exist in Cloud Dataflow.

Option D is incorrect. Transform is a step in your pipeline and it represents data processing operation.

Ask our Experts

Did you like this **Question?**

Question 21 Correct

Domain: Other

You have some Hadoop jobs on-premise which the management of the company has decided to bring to Google Cloud Dataproc. Few of the jobs will still be running from on-premise hadoop cluster while other will be running from Cloud Dataproc. You need to orchestrate these jobs and add required dependencies between your on-premises and dataproc jobs. However, the company doesn't want any vendor lock-in and can move to AWS as well in future. The orchestration framework should be chosen which can accept all these future changes as well. Select the best possible option provided by GCP with very little overhead.

- A. Cloud Composer right
- B. Apache Airflow
- C. Apache Oozie
- D. Cloud Scheduler

Explanation:

Correct answer is A.

Option A is correct. Cloud Composer allows you to pull workflows together from wherever they liv supporting a fully-functioning and connected cloud environment. Since Cloud Composer is built on

Apache Airflow – an open-source technology – it provides freedom from vendor lock-in as well as integration with a wide variety of platforms. We can connect to on-premise database from Cloud Composer. For connecting to on-premise database refer below link:

https://www.progress.com/tutorials/cloud-and-hybrid/connect-to-on-premises-databases-from-google-composer

Option B is incorrect. As per question GCP service is required, Cloud Composer is built on top of Apache Airflow. Cloud Composer should be correct answer.

Option C is incorrect. Orchestration job using Apache Oozie does not support connecting to on-premises and dataproc at the same time. Also, Oozie can only be run with Hadoop.

There is no managed service for Oozie in GCP.

Option D is incorrect. Cloud Scheduler is fully managed cron job scheduler. In this case, we need orchestration framework where dependencies between jobs will be present hence Cloud Scheduler is not the correct answer.

Ask our Experts

Did you like this **Question?**

Question 22 Correct

Domain: Store the data

Your organization is migrating enterprise data from on-premises to Google Cloud. They have 1 Petabyte of archive data to transfer from on-premise servers.

Which of the following methods would you follow to securely migrate large volumes of data to Google Cloud Platform without disrupting business operations?

A. gsutil

B. Storage Transfer Service

C. Transfer Appliance Service right

D. gcloud

Explanation:

Correct Answer: C

Option C is CORRECT because Transfer Appliance is an offline migration service & can be used to securely migrate 1 Petabyte of data to Google Cloud Platform.

Option A is incorrect because the gsutil tool can not be used to transfer 1 Petabyte of data between on-premise and cloud storage. This is recommended if transferring less than 1 TB from on-premises.

Option B is incorrect because Storage Transfer Service can be used when transferring more than 1 TB from another Cloud Storage region.

Option D is incorrect because the gcloud is used to create and manage Google Cloud resources and not used to transfer the data.

For more information on the **Storage transfer**, please visit the below URL:

https://cloud.google.com/storage-transfer/docs/overview

Ask our Experts

Did you like this **Question?**

Question 23 Correct

Domain: Prepare and use data for analysis

You work as a data engineer in an organization with a large amount of data. They use Google Big Table to store their web service's activity logs for faster retrieval and update. What will happen if the BigTable node fails?

A. Data will be lost

B. Recover data from Cloud Storage when the node comes back online

C. Data will not be lost right

D. Data will be transferred automatically to new node

Explanation:

Correct Answer: C

Option C is CORRECT because data is never stored in Big Tables nodes.

Recovery of the Bigtable node from the failure is very fast because metadata information only needs to be replicated to the new node.

Option A is incorrect because storage and compute are separate so data will not be lost.

Option B is incorrect because Big Table does not store its data in Cloud storage.

For more information on the **Cloud BigTable**, please visit the below URL:

https://cloud.google.com/bigtable/docs/overview

Ask our Experts

Did you like this **Question?**

Question 24 Correct

Domain: Ingest and process the data

An eCommerce company wants to predict prices using historical data. They are interested in using Google Cloud and like to explore ML models.

As a Google Cloud ML engineer, can you suggest to them which AI / ML model can be used for this purpose? (Select 2)

- [x] A. Linear regression right
- [x] B. Decision tree right
- [ ] C. Dimensionality reduction
- [ ] D. Logistic Regression

Explanation:

Correct Answers: A & B

Linear Regression : Predicts a continuous outcome based on a linear relationship with input features. Minimizes error between predicted and actual values using a "best fit" line.

Decision Tree : Classifies data by splitting it into branches based on feature values.Each branch represents a rule, leading to leaf nodes with predicted classes. More flexible than linear regression

Options A & B are CORRECT because Linear regression and Decision tree can help to predict the value using the historical data based on the above explanation.

Option C is incorrect because it is part of feature engineering and cannot be used to predict the value.

Option D is incorrect as Logistic regression is a statistical method used to predict the probability of a binary outcome (e.g., yes or no, true or false). It utilizes a logistic function to transform the linear combination of input features into a probability between 0 and 1, representing the likelihood of the target outcome. Logistic regression is widely employed in various applications, including medical diagnosis, customer churn prediction, and spam filtering.

For more information on **Machine Learning**, please visit the below URL:

https://developers.google.com/machine-learning/crash-course/descending-into-ml/linear-regression

https://developers.google.com/machine-learning/glossary#d and refer to the Decision tree.

Ask our Experts

Did you like this **Question?**

Question 25

Correct

Domain: Design Data Processing Systems

You work as a data engineer in a gaming organization with a large amount of data. They use Cloud Pub/Sub topics to store the events from each player every 2 seconds. Which command you will use to retrieve messages with your subscription named “alertlogSub1”?

A. gcloud pubsub pull --auto-ack alertlogSub1

B. gcloud pubsub subscriptions pull --auto-ack alertlogSub1 right

C. gcloud pubsub subscriptions pull --ack alertlogSub1

D. gcloud pubsub pull subscriptions --auto-ack alertlogSub1

Explanation:

Correct Answer: B

Option B is CORRECT because the correct command is :

gcloud pubsub subscriptions pull --auto-ack alertlogSub1

Sample output :

DATA	MESSAGEID	ATTRIBUTES
Hi This is Whizlabs	1425642542586412	

--auto-ack used to automatically acknowledge every message pulled from the subscriptions named as “alertlogSub1”

Option A is incorrect because it's missing the subscriptions subcommand. The correct command structure is gcloud pubsub subscriptions pull.

Option C is incorrect because the --ack flag is used to explicitly acknowledge messages. The --auto-ack flag is used to automatically acknowledge messages, which is more efficient for high-volume workloads.

Option D is incorrect because pull is not a subcommand of gcloud pubsub command. The correct command structure is gcloud pubsub subscriptions pull.

For more information on the **Cloud PubSub**, please visit the below URL:

https://cloud.google.com/sdk/gcloud/reference/pubsub/subscriptions/pull

Ask our Experts

Did you like this **Question?**

Question 26 Correct

Domain: Design Data Processing Systems

Your organization is looking for a fully managed, cloud-native data integration service. Their engineers are very familiar with CDAP (Cask Data Application Platform).

Which managed service in Google Cloud would you recommend? (Single option)

A. Cloud Dataproc

B. Cloud Composer

C. Cloud Dataflow

D. Cloud Data Fusion right

Explanation:

Correct Answer: D

Option D is CORRECT because Cloud Data Fusion is built with an open-source core (CDAP) for pipeline portability.

This also provides end-to-end data lineage for root cause analysis.

Option A is incorrect because Cloud Dataproc is a fully managed and highly scalable service for running Apache Flink, Apache Spark, and other applications.

Option B is incorrect because Cloud Composer is a fully managed data-workflow orchestration service.

Option C is incorrect because Cloud Dataflow is a fully managed streaming analytics service.

For more information on the **Google Cloud Data fusion**, please visit the below URL:

https://cloud.google.com/data-fusion

Ask our Exp

Did you like this **Question?**

Question 27 Correct

Domain: Ingest and process the data

An online food delivery platform wants to generate texts from the menus cards, they are already using Google Cloud for running their web applications. They want to use Cloud Vision API for text detection from the image.

CTO asked you to verify if any image size limit with this service, so accordingly they can suggest to their image gathering team?

A. 4MB

B. 20MB right

C. 100MB

D. 10MB

Explanation:

Correct Answer: B

Option B is CORRECT because images files sent to the Cloud Vision API should not exceed 20 MB.

If you will try to add images of size more than 20 MB then you will get the following error message :

For more information on the **Cloud Vision API**, please visit the below URL:

https://cloud.google.com/vision/quotas

Ask our Experts

Did you like this **Question?**

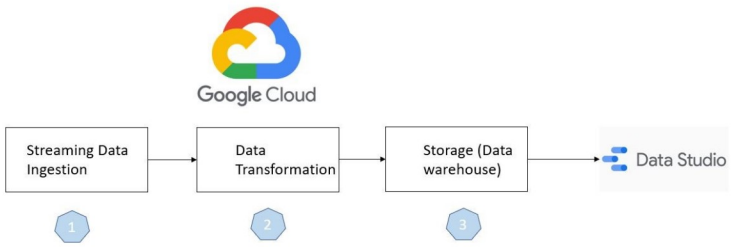
Question 28

Correct

Domain: Design Data Processing Systems

You need to design a data architecture to bring together all your data at any scale and provide insights into all your users.

You need to complete the below architecture.



Which of the following can be used in place of number 3 i.e. Storage (Data Warehouse)?

A. Cloud Pub/Sub

B. Cloud SQL

C. Big Query right

D. Cloud Spanner

Explanation:

Correct Answer: C

Option C is CORRECT because Big Query is recommended as a SQL-complaint analytics data warehouse.

This is a serverless, highly scalable, and cost-effective multi-cloud SQL-complaint data warehouse.

Option A is incorrect because Cloud Pub/Sub is recommended for Streaming data ingestion

Option B is incorrect because Cloud SQL is a fully-managed database service but not a data

warehouse service.

You can use Cloud SQL with MySQL, PostgreSQL, or SQL Server

Option D is incorrect because Cloud Spanner is a distributed SQL database management and storage service.

For more information on the **Big Query**, please visit the below URL:

https://cloud.google.com/bigquery/docs

Ask our Experts

Did you like this **Question?**

Question 29 Correct Marked for review

Domain: Store the data

Data engineers in your company use Big Query as their structured database supporting SQL for querying.

You were asked to select the data format for importing to Big Query considering the best performance and cost-efficient storage solution for multi-TB databases with millions of rows?

A. CSV

B. JSON

C. Parquet

D. AVRO right

Explanation:

Correct Answer: D

Option D is CORRECT because Avro is recommended for Big Query as it is cost-efficient and gives the best performance results.

Parquet format gives better performance than CSV and JSON.

For more information on the **Cloud Big Query**, please visit the below URL:

https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-avro

Ask our Experts

Did you like this **Question?**

Question 30 Correct

Domain: Store the data

A US-based Insurance company is migrating enterprise data from AWS Cloud provider to Google Cloud. They have 75 TB of archive data to transfer from AWS Cloud.

Which of the following methods would you follow to securely migrate large volumes of data to the Google Cloud Platform without disrupting business operations? (Single option)

A. gsutil

B. Storage Transfer Service right

C. Transfer Appliance Service

D. All of the above

Explanation:

Correct Answer: B

Option B is CORRECT because Storage Transfer Service can be used when transferring more than 1 TB from another Cloud Storage service.

Option A is incorrect because the gsutil tool can be used with on-premise and cloud storage.

This is recommended if transferring less than 1 TB from on-premises.

Option C is incorrect because Transfer Appliance is an offline migration service & can be used to securely migrate 1 Petabyte of data to Google Cloud Platform.

For more information on the **Storage transfer**, please visit the below URL:

https://cloud.google.com/storage-transfer/docs/overview

Ask our Experts

Did you like this **Question?**

Question 31

Correct Marked for review

Domain: Design Data Processing Systems

As a GCP data engineer, you were asked to use Cloud shell to load the Big query dataset with the data available in Google Cloud Storage.

Google Cloud Storage path: gs://whizlabs-public-dataset/department/cloudcertcount.csv

Big query Table: whizlabs_dataset.cloud_cert_dataset

Data format : CSV

Schema : Name:STRING,CERTIFICATION:STRING,AGE:INTEGER

Can you suggest which one of the following is correct? (select one)

- A. bq load --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER right
- B. bq extract --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER
- C. bq load --schema Name:STRING,CERTIFICATION:STRING,AGE:INTEGER --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv
- D. bq load --target_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER

Explanation:

Correct Answer: A

Options A is CORRECT because the current syntax is :
bq load --source_format = [format] [destination dataset].[table] [source_path] [schema]

bq load --source_format = CSV whizlabs_dataset.cloud_cert_dataset gs://whizlabs-public-dataset/department/cloudcertcount.csv Name:STRING,CERTIFICATION:STRING,AGE:INTEGER

bq load command is to load data into a Big Query table.

Option B is incorrect because the bq extract command is used to export table data to Google Cloud Storage.

All the other options are invalid.

For more information on the **Big Query-bq load**, please visit the below URL:

https://cloud.google.com/bigquery/docs/reference/bq-cli-reference#bq_load

https://cloud.google.com/bigquery/docs/reference/bq-cli-reference#bq_extract

Ask our Experts

Did you like this **Question?**

Finish Review

Refund Policy
Programs Guarantee

Need help? Please or +91 6364678444