Free Test

Attempt

1

Completed on

Wednesday , 09 October 2019 , 04:58 PM

Marks Obtained

16 / 20

Time Taken

N/A

Your score is

80%

Result

Pass

Mode

Practice

Domains wise Quiz Performance Report

No	1
Domain	Design Data Processing Systems
Total Question	1
Correct	0
Incorrect	1
Unattempted	0
Marked for review	0
No	2
Domain	Other
Total Question	18
Correct	15
Incorrect	3
Unattempted	0
Marked for review	0
No	3
Domain	Operationalize Machine Learning Models
Total Question	1
Correct	1
Incorrect	0
Unattempted	0
Marked for review	0
Total	Total
All Domain	All Domain
Total Question	20
Correct	16
Incorrect	4
Unattempted	0
Marked for review	0

Review the Answers

Sorting by | All

Question 1 | Incorrect

Domain :Design Data Processing Systems

A multi-national company wants to unify their data sources by building a universal centralized data warehouse instead of their current architecture in which every branch has its own and branches from other regions cannot access it. They want to build a data analytics team to extract data from all branches and build daily reports and dashboards to visualize the metrics required for C-Level managers to take decisions. The current data warehouses are all MySQL databases and analytics team will use SQL for data reporting. The company is distributed among different continents (North America, Europe & Asia). Which of the following approach is best suits to satisfy the company's new data warehouse architecture?

A. **Use Cloud SQL to launch MySQL databases on each region. Enable cross-region read replication for each to sync between different regions.**

✓ B. **Use Cloud SQL to launch Multi-regional MySQL databases. Each in North America, Europe & Asia. Enable cross-region read replication for each to sync between different regions.** ✗

C. **Use BigQuery as a data warehouse and grant data analytics team editor roles.**

D. **Use Cloud Spanner by launching a multi-regional database to be the company's unified data warehouse.** ✓

Explanation:

The correct answer is Option D.

Cloud Spanner is a horizontally scalable, strongly consistent, relational database service. It's built to combine the benefits of relational database structure with non-relational horizontal scale. This delivers high-performance transactions and strong consistency across rows, regions and continents.

Option A is incorrect. Launching MySQL databases on each region defeats the purpose of having a unified data warehouse.

Option B is incorrect. Cloud SQL does not support multi-regional databases.

Option C is incorrect. While BigQuery is a strong and potential alternative, BigQuery does not have horizontal scaling and it only covers USA & Europe (*by the time of writing this question*). Another drawback is, BigQuery doesn't support full data manipulation language (DML) and it has limitations on how rows can be updated or deleted.

Option D is correct. Cloud Spanner is a relational database supports horizontal scaling across continents.

Source(s):

Cloud Spanner: https://cloud.google.com/spanner/

BigQuery dataset locations: https://cloud.google.com/bigquery/docs/locations

Bigquery: Data Manipulation Langauge: https://cloud.google.com/bigquery/docs/reference/standard- sql/data-

Ask our Experts

Rate this Question? ☺ ☹

Question 2 Correct

Domain : Other

A fast-food chain restaurant wants to detect the different meal photos its customers upload to the different social media platforms tagged with their name in order to know what meals customers like and share the most for better quality analysis. It asks your advice on developing such solution for them.
However, they want it to be available and in production the soonest possible because they expect a high activity on their social media pages by the next public holiday which is coming in 2 weeks and marketing team finds it a great opportunity to receive feedback based on what customers say online. What is the best approach for this?

A. Use AutoML Vision to build and train the model by using all the training photos you collected from food-chain's social media pages for better results.

✓ B. Use AutoML Vision to build and train the model by using 50-70% of training photos you collected from food-chain's social media pages while the rest of training set is to test and tune the model. ✓

C. Use Dataproc to build the model using SparkML. Use 50-70% of training photos you collected to train the model and the rest to test and tune the model. Deploy the model using Cloud ML Engine.

D. **Use Cloud ML Engine with TensorFlow to build the model. Use all training photos you collected to train the model. Deploy the model using Cloud ML Engine.**

Explanation:

The correct answer is Option B.

Since you have a very short time to build, train and deploy the model, building your own model can be time-consuming and not in your favor. Google provides a great ML service called AutoML to quickly build models for you. AutoML Vision is one of its products which you can start with a training set as little as a dozen photo samples and AutoML takes care of the rest.

Option A is incorrect. AutoML Vision is the right choice. However, training the model with whole training set is not the right approach in Machine Learning because you ought to test the model before considering it accurate enough for production. Usually, training set is split into 70-30% sets, first for training while the second is for testing and tuning the model's parameters.

Option C is incorrect. Using any approach other than AutoML can be time-consuming and with such tight deadline, it's not the best approach.

Option D is incorrect. Using this approach can also be time-consuming and using the whole training set for training is not a best practice as explained before.

Thus, the best approach for this scenario is Option B.

Source(s):

Google Cloud AutoML: https://cloud.google.com/automl/

Cloud Machine Learning Engine: https://cloud.google.com/ml-engine/

Ask our Experts

Rate this Question?

Question 3 Correct

Domain :Operationalize Machine Learning Models

A financial services firm providing products such as credit cards and bank loans receives thousands of online applications from clients applying for their products. Because it takes a lot of effort to scan and check all applications if they meet the minimum requirements for the products they are applying for, they want to build a machine learning model takes application fields like annual income, marital status, date of birth, occupation and other attributes as input and finds out if the applicant is qualified for the product the client applied for. Which of the following the machine learning technique will help to build such model?

A. **Regression**

✓ B. **Classification**

C. **Clustering**

D. **Reinforcement learning**

Explanation:

The correct answer is Option B.

A regression problem is a problem which its output variable is of continuous value. Problems which finds out about variables such as weights, prices or age are considered regression problems. A classification problem is a problem which the output variable is a category. Examples of classification problems are finding a passenger's nationality, detect if a patient is diagnosed with a disease or if an applicant is qualified for a job interview. Regression and classification are supervised learning problems. It means, the machine learns from past experiences by training it on a labeled data set. A training set is a set of rows with input and output parameters. The machine then learns from the training set and improves its parameters for better detection.

Clustering is an unsupervised learning method. An unsupervised learning is a method to find references between input data without labeled output. The purpose is to find meaningful structure between the input sets with similar features and group them. Clustering is the method of grouping data points share similarities and separating dissimilar points to other groups. Examples of clustering applications are customer segmentation (new, frequent, loyal, ..), city land value and detecting anomalies in network traffic.

Reinforcement learning is a technique which a machine takes actions without training sets to reach the highest rewards possible. The agent learns from trial and decides what to do to perform a given task without supervision. The task punishes the agent for a wrong action and rewards it for achieving the task. Examples of reinforcement learning is asking an agent to play a maze game to reach the exit with traps along the way or making an agent play a video game and win a racing game.

From the explanation above, we can see the scenario problem which finding if a client is qualified for a product is a classification problem. So, option B is correct.

Ask our Experts

Rate this Question?

Question 4 Correct

Domain : Other

You have built a machine learning model to classify if a customer would buy a certain product when recommended by the company's website. You trained the model with a sample set. Upon testing the model, you found out only 28% of the testing sets are actually true positives and the model isn't very accurate. You figured out the model is over-fitted. How would you solve this?

A. Increase training data, increase feature parameters & increase regularization.

B. Decrease training data, decrease feature parameters & increase regularization.

✓ C. Increase training data, decrease feature parameters & increase regularization.

D. Increase training data, decrease feature parameters & decrease regularization.

Explanation:

Answer: C.

Description:

Overfitting happens when a model performs well on a training set, generating only a small error, while giving wrong output for the test set. This happens because the model is only picking up specific features input found in the training set instead of picking out general features of the given training set.

To solve overfitting, the following would help improving the model's quality:

• Increase the number of examples, the more data a model is trained with, the more use cases the model can be training on and better improves its predictions.

• Tune hyperparameters which is related to number and size of hidden layers (for neural networks), and regularization, which means using techniques to make your model simpler such as using dropout method to remove neuron networks or adding "penalty" parameters to the cost function.

• Remove features by removing irrelevant features. Feature engineering is a wide subject and feature selection is a critical part of building and training a model. Some algorithms have built- in feature selection, but in some cases, data scientists need to cherry-pick or manually select or remove features for debugging and finding the best model output.

From the brief explanation, to solve the overfitting problem in the scenario, you need to:

• Increase the training set.

• Decrease features parameters.

• Increase regularization.

Hence, answer C is correct.

Source(s):

Building a serverless Machine learning model:
https://cloud.google.com/solutions/building-a- serverless-ml-model

Ask our Experts

Rate this Question? ☺ ☹

Question 5 Correct

Domain : Other

A coach line bus service company wants to predict how many passengers they expect to book for tickets on their buses for the upcoming months. This helps the company to know how many buses they need to be in service for maintenance and fuel and how many drivers to be available. The company has data sets of all booked tickets since its launch in 1968 and it allows private sharing of the data if this helps the prediction process. You will build the machine learning model for the coach line company. Which technique you will use to predict the number of passengers in the next months?

✓ A. **Regression.** ✅

B. **Association.**

C. **Classification.**

D. **Clustering.**

Explanation:

Answer: A.

Description:

A regression problem is a problem which its output variable is of continuous value. Problems which finds out about variables such as weights, prices or age are considered regression problems. A classification problem is a problem which the output variable is a category. Examples of classification problems are finding a passenger's nationality, detect if a patient is diagnosed with a disease or if an applicant is qualified for a job interview. Regression and classification are supervised learning problems. It means, the machine learns from past experiences by training it on a labeled data set. A training set is a set of rows with input and output parameters. The machine then learns from the training set and improves its parameters for better detection.

Association is a rule-learning technique for discovering interesting relations between variables in large data sets. Example of association rules is discovering regularities between products in large-scale transaction data recorded by point-of-sales for a retail chain store.

Clustering is an unsupervised learning method. An unsupervised learning is a method to find references between input data without labeled output. The purpose is to find meaningful structure between the input sets with similar features and group them. Clustering is the method of grouping data points share similarities and separating dissimilar points to other groups. Examples of clustering applications are customer segmentation (new, frequent, loyal, ..), city land value and detecting anomalies in network traffic.

From the explanation above, the technique to help solving the scenario is Answer A: Regression.

Ask our Experts

Rate this Question? ☺ ☹

Question 6 Correct

Domain : Other

A video-on-demand company wants to generate subtitles for its content on the web. They have over 20,000 hours of content to be subtitled and their current subtitle team cannot catch up with the every- growing video hours the content team keep adding to the website library. They want a solution to automate this as man power can be expensive and may take long time.
Which service of the following can greatly help the automation of video subtitles?

A. **Cloud Natural Language.**

✓ B. **Cloud Speech-to-Text.**

C. **AutoML Vision API.**

D. **Machine Learning Engine.**

Explanation:

Answer: B.

Description:

Answer A is incorrect: Cloud natural language service is to derive insights from unstructured text revealing meaning of the documents and

categorize articles. It won't help extracting captions from videos.

Answer B is correct: Cloud Speech-to-Text is a service to generate captions from videos by detecting speakers language and speech.

Answer C is incorrect: AutoML Vision API is a service to recognize and derive insights from images by either using pre-trained models or training a custom model based on a set of photographics.

Answer D is incorrect: Machine Learning Engine is a managed service letting developers and scientists build their own models and run them in production. This means, you have to build your own model to generate text from videos which needs much effort and experience to build such model. So, it's not a practical solution for this scenario.

Source(s):

Google NLP: https://cloud.google.com/natural-language/

Google Machine Learning Engine: https://cloud.google.com/ml-engine/

Google Vision API: https://cloud.google.com/vision

Google Speech-to-Text API: https://cloud.google.com/speech-to-text/

Ask our Experts

Rate this Question? ☺ ☹

Question 7 Incorrect

Domain : Other

An online learning platform wants to generate captions for its videos. The platform offers around 2,500 courses with topics about business, finance,

cooking, development & science. The platform allows content with different languages such as French, German, Turkish and Thai. Thus, this can be very difficult for a single team to caption all available courses and they are looking for an approach which helps do such massive job. Which product from Google Cloud will you suggest them to use?

A. **Cloud Speech-to-Text.** 

B. **Cloud Natural Language.**

C. **Machine Learning Engine.**

D. **AutoML Vision API.** 

Explanation:

Answer: A.

Description:

Answer A is correct: Cloud Speech-to-Text is a service to generate captions from videos by detecting speakers language and speech.

Answer B is incorrect: Cloud natural language service is to derive insights from unstructured text revealing meaning of the documents and categorize articles. It won't help extracting captions from videos.

Answer C is incorrect: Machine Learning Engine is a managed service letting developers and scientists build their own models and run them in production. This means, you have to build your own model to generate text from videos which needs much effort and experience to build such model. So, it's not a practical solution for this scenario.

Answer D is incorrect: AutoML Vision API is a service to recognize and derive insights from images by either using pre-trained models or training a custom model based on a set of photographics.

Source(s):

Google NLP: https://cloud.google.com/natural-language/

Google Machine Learning Engine: https://cloud.google.com/ml-engine/

Google Vision API: https://cloud.google.com/vision

Google Speech-to-Text API: https://cloud.google.com/speech-to-text/

Ask our Experts

Rate this Question? ☺ ☹

Question 8 Correct

Domain : Other

You have a dataflow pipeline reads a CSV file daily at 6am, applies the needed cleansing & transformation on it, then loads it to BigQuery. Occassionally, the CSV file might be modified within the day due to human error or incomplete data. This causes you to manually re-run dataflow pipeline again. Is there a way to fix this by automatically re-run the pipeline if file has been modified?

A. **Use Cloud Scheduler to re-run dataflow after 6am. Check what is the average time the file is modified and schedule based on it.**

B. **Use Dataproc to reprocess the file after 6am. You can use Cloud Functions to launch a Dataproc cluster.**

✓ C. **Use Cloud Composer to rerun dataflow and reprocess the file. Create a custom sensor to detect file condition if changed.** 

D. **Use a compute engine to schedule a cron job to run every 10 minutes to check if the file was modified to rerun dataflow.**

Explanation:

Answer: C

Cloud Composer is a fully managed workflow orchestration service built on Apache Airflow. Cloud composer is built specifically to schedule and monitor workflows and take required actions. You can use Cloud Composer to orchestrate dataflow pipeline and create a custom sensor to detect file's condition if any changes occurred, then it triggers the dataflow pipeline to run again.

Answer A is incorrect: Guessing what time scheduler should rerun dataflow is not efficient.

Answer B is incorrect: Dataproc is unnecessary in this scenario. **Answer D is incorrect:** This solution is viable, but Answer C has a better and more efficient design.

Source(s):

Cloud Composer: https://cloud.google.com/composer/

Ask our Experts

Rate this Question? ☺ ☹

Question 9 Correct

Domain : Other

A dairy products company is using sensors installed around different areas in its farms to monitor employees activities and detect any intruders. Apache Kafka cluster is used to gather the events coming from sensors. Recently, Kafka cluster is becoming a bottleneck causing lag in receiving sensor events. Turns out sensors are sending more frequent events and due to the company expanding with more farms, more sensors are installed and this will cause extra load on the cluster. What is the most resilient approach to solve this issue?

✓ A. **Use pub/sub to ingest and stream sensor events.** ✓

B. **Scale out Kafka cluster to withstand the continuously flowing event stream.**

C. **Spin up a new Kafka cluster and distribute sensors even streams between the two clusters.**

D. **Build a Dataflow pipeline to ingest the events stream.**

Explanation:

Answer: A.

Description: Cloud Pub/Sub is a service to ingest event streams at any scale. It's scalable and reliable for stream analytics and event-driven

computing systems. So it's the most reliable Google product for such scenario.

Answers B & C are wrong because these are not scalable solutions.

Answer D is wrong because Dataflow cannot ingest event streams. It needs Pub/Sub service to do so.

Source(s):

Google Pub/Sub: https://cloud.google.com/pubsub/docs/overview

---

Ask our Experts

Rate this Question? ☺ ☹

Question 10 Correct

Domain : Other

A social media platform stores various details of their platform users such as session login time, URLs visited, activities on platform and other logs. With GDPR (General Data Protection Regulation) compliance to be officially implemented, the platform now allows users to download their activity logs from their profile settings which they can click a button to call an API to generate a full report.
Recently, users are complaining timeouts after 60 seconds of requesting to download their activity logs at peak hours when the platform has the most traffic. They have to try for several minutes or even hours for the API to return their report available for download.
How can you solve this issue?

A. Increase timeout for API at peak times to 120 seconds. If it keeps failing, try increasing the timeout until the issue is resolved.

B. Build a Dataflow pipeline to generate daily reports of users' activity logs. Users can download those daily reports whenever they want to.

C. Migrate data source to Cloud Spanner for horizontal scaling to avoid query timeouts.

✓ D. Use Pub/sub to pull the requests for activity logs from users. Send a link to users by their email addresses with a temporary download link for them to access their report.

Explanation:

Answer: D.

Description:

Cloud Pub/Sub is a service to ingest event streams at any scale. It's scalable and reliable for stream analytics and event-driven computing systems.

Pub/sub is a good product to de-couple a system's components so they communicate with each other asymmetrically. From the scenario shown here, instead of directly calling the API to export required report which puts great loads on the API and hence the timeouts faced by users. Instead, the platform can "publish" messages to a "topic" related to exporting activity log reports sending the required parameters such as user ID and custom settings such as date range and what data to export. The API can be switched to be a "subscriber" which receives the messages

sent and processes each message asymmetrically to generate the report, then sends the download link to the user's mailbox when ready.

Hence, answer D is correct.

Answer A is incorrect: Increasing timeout isn't a scalable solution and it may keep occurring eventually when more and more users join the platform.

Answer B is incorect: While this would solve the timeout issues, generating daily reports for users can be costly as more users join, knowing that requesting activity log reports are a non-frequent action and this costs both compute and storage resources. This solution also doesn't provide flexibility with what parameters the report is generated on such as date range and other custom metrics.

Answer C is incorrect: This solution has several issues. First, we're assuming the data source is a relational database, which can be unlikely since NoSQL databases better perform for massive log input which uses the user ID as a key to reach the data. Second, Cloud Spanner isn't a cheap solution for a service not frequently used.

Source(s):

Google Pub/Sub: https://cloud.google.com/pubsub/docs/overview

Question 11 Correct

Domain : Other

A company decides to migrate its on-premise data infrastructure to the cloud mainly for high availability of cloud services and to lower the high costs of storing data on-premise. The infrastructure uses HDFS to store data and be processed and transformed using Apache Hive & Spark. The company wants to migrate the infrastructure and DevOps team still wants to administrate the infrastructure in the cloud. As a data architect, which of the following is the approach recommended by Google?

- A. **Use Dataproc to process the data. Store data in Google Storage.** ✓
- B. **Build a Dataflow pipeline. Store the data in Google Storage. Use Cloud Compute to launch instances and install the required dependencies for processing the data.**
- C. **Use Dataproc to process the data. Store data in Dataproc's HDFS.**
- D. **Build a Dataflow pipeline. Store the data in persistent disks in HDFS. Execute the code in Spark framework provided by Dataflow.**

Explanation:

Answer: A.

Description: Dataproc is cloud-native Apache Hadoop & Apache Spark service. Dataproc is a fully-managed service from Google to run Apache Hadoop & Spark clusters. Dataflow is a simplified streaming/batching data processing service. With Apache Beam, it provides rich set of windowing and session analysis primitives as well as an ecosystem of source & sink connectors.

Answer B is incorect: Dataflow is serverless which may not suit DevOps requirement to fully manage the pipeline and it's unnecessary to use

Cloud Compute for installing dependencies. **Answer C is incorrect**: Dataproc's HDFS is volatile, means it will be removed when the cluster is deleted. Dataproc clusters can be kept up indefinitely but this may lead to high costs which defeats the purpose of migration.

Answer D: In addition to what discussed in answer B, storing data using persistent disks can be only accessible by Compute engines and it's more expensive than storing in Google Storage.

Answer A fulfills the requirements for migrating the on-premise infra to the cloud with high availability, minimum costs and full control by DevOps.

Source(s):

Cloud Dataproc: https://cloud.google.com/dataproc/

Cloud Dataflow & Dataflow vs. Dataproc:

Ask our Experts

Rate this Question? ☺ ☹

Question 12 Correct

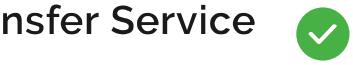
Domain : Other

A multi-international company has multiple Google Storage buckets in different regions around the world. Each branch has its own set of buckets in the region nearest to them to avoid latencies. However, this led to a problem for analytics team to reach and do the necessary reports on the data using BigQuery since they need to create they need to create the tables in the same region either to import the data or create external tables to access the data in different regions. The head of data decided to sync the data daily from different Google Storage buckets scattered in

different regions to a single multi-regional bucket to do the necessary data analysis and reporting.
Which service could help with this approach?

A. **Appliance Transfer Service**

B. **gsutil**

✓ C. **Storage Transfer Service** 

D. **Dataflow**

Explanation:

Answer: C.

Description:

Storage Transfer Service allows you to quickly import *ONLINE* data into Cloud Storage. You can also set up a repeating schedule for transferring data, as well as transfer data within Cloud Storage, from one bucket to another.

Transfer Appliance is an *OFFLINE* secure, high capacity storage server that you set up in your datacenter. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage.

So, answer C is correct, while answer

A is incorrect.

Answer B is incorrect: gsutil tool is good for programmatic usage by developers and may be useful to copy and move megabytes/gigabytes of data. Not so practical for Terabytes of data. It's also not reliable data

transfer technique as it is related to the machine's connectivity with Google Cloud.

Answer D is incorrect: Dataflow as a solution may be viable, but you need to build a pipeline to migrate data from bucket to another. Storage Transfer Service can do it without the extra effort.

Source(s):

Google Cloud Storage Transfer Service: https://cloud.google.com/storage-transfer/docs/

Google Appliance Transfer Service: https://cloud.google.com/transfer-appliance/

Ask our Experts

Rate this Question? ☺ ☹

Question 13 Correct

Domain : Other

A company is migrating its current infrastructure from on-premise to Google cloud. It stores over 280TB of data on its on-premise HDFS servers. You were tasked to move data from HDFS to Google Storage in a secure and efficient manner. Which of the following approaches are best to fulfill this task?

A. **Install Google Storage gsutil tool on servers and copy the data from HDFS to Google Storage.**

B. **Use Cloud Data Transfer Service to migrate the data to Google Storage.**

C. **Import the data from HDFS to BigQuery. Then, export the data to Google Storage in AVRO format.**

D. **Use Transfer Appliance Service to migrate the data to Google Storage.**

Explanation:

Answer: D.

Description:

Storage Transfer Service allows you to quickly import *ONLINE* data into Cloud Storage. You can also set up a repeating schedule for transferring data, as well as transfer data within Cloud Storage, from one bucket to another.

Transfer Appliance is an *OFFLINE* secure, high capacity storage server that you set up in your datacenter. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage.

So, answer D is the correct one, while **B is incorrect.**

Answer A is incorrect: gsutil tool is good for programmatic usage by developers and may be useful to copy and move megabytes/gigabytes of data. Not so practical for Terabytes of data. It's also not reliable data transfer technique as it is related to the machine's connectivity with Google Cloud.

Answer C is incorrect: In order to migrate to BigQuery, you need to migrate data to Google Storage. This is a useless approach as the main challenge is migrating data from HDFS to Google Storage and BigQuery won't help solving it.

Source(s):

Google Cloud Storage Transfer Service: https://cloud.google.com/storage-transfer/docs/

Google Appliance Transfer Service: https://cloud.google.com/transfer-appliance/

Migrate HDFS to Google Storage: https://cloud.google.com/solutions/migration/hadoop/hadoop-gcp-migration-data

Ask our Experts

Rate this Question?

Question 14 Correct

Domain : Other

A company is moving its data center from its on-premise servers to the cloud. It was estimated that they have about 2 Petabytes of data to be moved and security team is very concerned the data should be migrated securely and project manager has a timeline of 6 months for the whole migration to be done.
Which of the following approaches is best to do the job?

- ✓ A. **Appliance Transfer Service.**
- B. **Google Storage (Coldline).**
- C. **Cloud Transfer Service.**
- D. **Datastore.**

Explanation:

Answer: A.

Description:

Storage Transfer Service allows you to quickly import *ONLINE* data into Cloud Storage. You can also set up a repeating schedule for transferring data, as well as transfer data within Cloud Storage, from one bucket to another.

Transfer Appliance is an *OFFLINE* secure, high capacity storage server that you set up in your datacenter. You fill it with data and ship it to an ingest location where the data is uploaded to Google Cloud Storage. Data is encrypted automatically, and remains safe until you decrypt it.

So, answer A suggests using Transfer Appliance is the best approach.

Answer C is incorrect: It's an online data transfer service. Won't help for on-premise infrastructure.

Answer B is incorrect: Google Storage Coldline is a way to store archive data not frequently accessed cheaply in the cloud. It won't help migrating data to the cloud.

Answer D is incorrect: Datastore is a NoSQL database built for automatic scaling and high performance. It won't help with this scenario.

Source(s):

Google Cloud Storage Transfer Service:
https://cloud.google.com/storage-transfer/docs/

Google Appliance Transfer Service: https://cloud.google.com/transfer-appliance/

Google Datastore:
https://cloud.google.com/datastore/docs/concepts/overview

Google Storage Classes:
https://cloud.google.com/storage/docs/storage-classes

Ask our Experts

Rate this Question? ☺ ☹

Question 15 Incorrect

Domain : Other

You have the following legacy SQL query in BigQuery:

```
# legacy SQL
SELECTorder_date,
    COUNT(DISTINCT customer_id)) AS customers
FROM
    [my-project:orders.orders_2018]
GROUP BY
    order_date
ORDER BY
    order_date;
```

How can you convert this query to standard SQL?

A. **Change table syntax to `my-project:orders.orders_2018`** ✗

B. **Change table syntax to `my-project.orders.orders_2018`** ✓

C. **Change table syntax to [my-project.orders.orders_2018]**

D. **No change required. This works fine if standard SQL is enabled.**

Explanation:

Answer: B.

Description:

The correct way to represent a BigQuery table in standard SQL is:

`project-id.dataset-id.table-id`

So, answer B is the correct one.

Source(s):

Migrating from Legacy to Standard SQL:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#subqueries_in_more_places

Ask our Experts

Rate this Question? ☺ ☹

Question 16 Correct

Domain : Other

What is the keyword in BigQuery standard SQL used when selecting from multiple tables with wildcard by their suffices?

A. _WILDCARD_SUFFIX

B. _TABLES_SUFFIX

C. _SUFFIX

✓ D. _TABLE_SUFFIX

Explanation:

Answer: D.

Description:

To restrict the query so that it scans an arbitrary set of tables, use the *_TABLE_SUFFIX* pseudo column in the WHERE clause. The *_TABLE_SUFFIX* pseudo column contains the values matched by the table wildcard.

Answer D is the correct one. Other answers are incorrect because such keywords do not exist.

Source(s):

Bigquery – Querying wildcard tables:
https://cloud.google.com/bigquery/docs/querying-wildcard-tables

Ask our Experts

Rate this Question? ☺ ☹

Question 17 Correct

Domain : Other

You have the following BigQuery legacy SQL query :
SELECT SUM (amount)
FROM TABLE_DATE_RANGE ([some-dataset.orders_],
TIMESTAMP ('2017-06-01'),
TIMESTAMP ('2017-09-01'));
How can you convert it to standard SQL?

A.
```
SELECT SUM(amount)
    FROM `some-dataset.orders_*`
    WHERE TABLE_DATE_RANGE BETWEEN '20170601'
AND '20170901';
```

B.
```
SELECT SUM(amount)
    FROM `some-dataset.orders_*`
    WHERE _TABLE_SUFFIX BETWEEN '20170601'
AND '20170901';
```

C.
```
SELECT SUM(amount)
    FROM `some-dataset.orders_`
    WHERE _TABLE_SUFFIX BETWEEN '20170601' AND
'20170901';
```

D.
```
SELECT SUM(amount)
    FROM `some-dataset.orders_*`
    WHERE _TABLE_DATE_RANGE BETWEEN '20170601'
AND '20170901';
```

Explanation:

Answer: B.

Description:

To restrict the query so that it scans an arbitrary set of tables, use the *_TABLE_SUFFIX* pseudo column in the WHERE clause. The *_TABLE_SUFFIX* pseudo column contains the values matched by the table wildcard.

TABLE_DATE_RANGE queries daily tables that overlap with the time range between and . This function is a *LEGACY* SQL function.

Answer A is incorrect: TABLE_DATE_RANGE is a legacy SQL function.

Answer C is incorrect: To use wildcard table, the asterisk symbol "*" should be used at the end of the wildcard table.

Answer D is incorrect: This is no function called "_TABLE_DATE_RANGE" (starts with underscore "_" sign).

So, answer B is the accurate standard SQL syntax.

Source(s):

Bigquery – Querying wildcard tables:https://cloud.google.com/bigquery/docs/querying-wildcard-tables

BigQuery Legacy SQL: https://cloud.google.com/bigquery/docs/reference/legacy-sql#table-date-range

Ask our Experts

Rate this Question? ☺ ☹

Question 18 Correct

Domain : Other

A weather forecasting facility receives events from its 25,000 sensors every 10 seconds. Those events are stored in Google Storage in JSON format. Events can have different attributes based on purpose, location

and brand. Data Science team wants to apply their SQL-queries on this data for further transformation and forecasting analysis.
Which of the following approaches is best to satisfy Data Scientists request?

A. **Load the data directly to BigQuery with enabling "auto-detect" option.**

B. **Build a dataflow pipeline to read JSON data and transform it to a structured format like CSV. Then, load the data to BigQuery.**

C. **Import the data to BigTable. Choose combination #eventType-location-brand to differentiate between different events.**

D. **Use Dataproc cluster and create Hive external clusters on the data for data scientists to query data.**

Explanation:

Answer: A.

Description:

Schema auto-detection: 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 the 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. BigQuery makes a best-effort attempt to automatically infer the schema for CSV and JSON files.

So, answer A is the correct answer. The other answers are complicated and unnecessary approaches for this scenario.

Source(s):

BigQuery – Auto-detect schema:
https://cloud.google.com/bigquery/docs/schema-detect

Ask our Experts

Rate this Question? ☺ ☹

Question 19 Incorrect

Domain : Other

An environment safety facility receives thousands of events every 60 seconds from its sensors assembled in different sectors monitoring air pollution in the region. Scientists want to access and query the data for observation and daily reporting. Due to current funding state, their budget is limited and they seek a cost-effective, highly available and ACID-compliant solution supports SQL querying.
Which approach would you recommend for such scenario?

A. **Use BigQuery to store and query the event data. Enable streaming on BigQuery for data to be loaded in real-time.**

B. **Batch-load data into BigTable with launching 10 nodes to allow high performance.**

✓ C. **Use Cloud SQL to load events into a relational database and allow access to scientists to query.** ✗

D. **Use BigQuery to store and query event data. Batch load the data to BigQuery using its API.** ✓

Explanation:

Answer: D.

Description:

BigQuery supports both batch & streaming data. However, due to mentioned budget restrictions, the solution would choose the cheaper approach, which is batching data to BigQuery. Batching data to BigQuery is free of charge. Streaming data on the other hand is charged by size.

So, answer D is correct.

Answer A is incorrect for this scenario.

Answer B is incorrect: BigTable does not support SQL querying.

Answer C is incorrect: Cloud SQL needs administration and not easily scalable. Cloud SQL does not provide batching tools. BigQuery is a better approach for such scenario.

Source(s):

Bigquery: Streaming data: https://cloud.google.com/bigquery/streaming-data-into-bigquery

Bigquery: Batch data: https://cloud.google.com/bigquery/batch

Bigquery Pricing: https://cloud.google.com/bigquery/pricing

Question 20 Correct

Domain : Other

Analytics team receives data from different data sources stored in Google Storage. The team wants to query the data for required ETL operations which they will fully take care of using SQL. They want your advice on what is the best approach recommended by Google to do it. What would you suggest?

A. Batch load the data from Google Storage into BigQuery using its batch API, run cleansing and transformation queries on data and insert the transformed rows to another BigQuery table.

B. Batch load the data from Google Storage into BigQuery using its batch API, run cleansing and transformation queries on data and export the data to Google Storage. Launch Dataproc cluster and use Hive to query the transformed data.

✓ C. Create external tables on data using BigQuery, apply the cleansing and transformation queries on data then load the output to an internal BigQuery table for reporting and visualization.

D. Create external tables on data using BigQuery, apply the cleansing and transformation queries on data then load the output to BigTable for reporting and visualization.

Explanation:

Answer: C.

Description:

An external data source (also known as a federated data source) is a data source that you can query directly even though the data is not stored in BigQuery. Instead of loading or streaming the data, you create a table that references the external data source.

Querying an external data source using a temporary table is useful for one-time, ad-hoc queries over external data, or for extract, transform, and load (ETL) processes.

In summary, using external tables in BigQuery is useful for such cases:

• Perform ETL operations on data.

• Frequently changed data.

• Data is being ingested periodically.

Answer C is the correct answer based on above explanation and using BigQuery for reporting and visualization is a better approach.

Answer D is incorrect because BigTable isn't a practical (and cheap) approach to report and visualize data.

Answers A & B are incorrect: Based on Google's best practices, using external tables for ETL is better than loading data to BigQuery.

Source(s):

BigQuery external tables: https://cloud.google.com/bigquery/external-data-sources

BigQuery – Define external tables: https://cloud.google.com/bigquery/external-table-definition

Ask our Experts

Rate this Question? 

Finish Review

Certification

Cloud Certification

Java Certification

PM Certification

Big Data Certification

Company

Support

Discussions

Blog

Follow us