Google

Professional-Data-Engineer

Cloud Certified Professional Data Engineer

QUESTION & ANSWERS

Topics	Number of Questions	Question Sequance
Topic 1	4	1 - 4
Topic 2	0	5 - 4
Topic 3	2	5 - 6
Topic 4	1	7 - 7
Topic 5	7	8 - 14
Topic 6	18	15 - 32
Mix Questions	12	33 - 44
Total	44	

Topic 1

Case Study: Topic 1

Title : Main Questions Set A

QUESTION: 1

You are designing a basket abandonment system for an ecommerce company. The system will send a messageto a user based on these rules:No interaction by the user on the site for 1 hourHas added more than $30 worth of products to the basketHas not completed a transactionYou use Google Cloud Dataflow to process the data and decide if a message should be sent. How should youdesign the pipeline?

Option A : Use a fixed-time window with a duration of 60 minutes.

Option B : Use a sliding time window with a duration of 60 minutes.

Option C : Use a session window with a gap time duration of 60 minutes.

Option D : Use a global window with a time based trigger with a delay of 60 minutes.

Correct Answer: C

QUESTION: 2

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

Option A : Check the dashboard application to see if it is not displaying correctly.

Option B : Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.

Option C : Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.

Option D : Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

Correct Answer: B

QUESTION: 3

Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

Option A : Use Google Stackdriver Audit Logs to review data access.

Option B : Get the identity and access management IIAM) policy of each table

Option C : Use Stackdriver Monitoring to see the usage of BigQuery query slots.

Option D : Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Correct Answer: A

QUESTION: 4

You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

Option A : Linear regression

Option B : Logistic classification

Option C : Recurrent neural network

Option D : Feedforward neural network

Correct Answer: A

Topic 3
Case Study: Topic 3

Title : MJTelco Case Study

Company Overview

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world.

The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- Scale and harden their PoC to support significantly more data flows generated when they ramp to more
- than 50,000 installations.
- Refine their machine-learning cycles to verify and improve the dynamic models they use to control
- topology definition.

MJTelco will also use three separate operating environments – development/test, staging, and production – to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements

- Scale up their production environment with minimal cost, instantiating resources when and where
- needed in an unpredictable, distributed telecom user community.
- Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- Provide reliable and timely access to data for analysis from distributed research workers
- Maintain isolated environments that support rapid iteration of their machine-learning models without
- affecting their customers.

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

QUESTION: 5

You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery asits data source. It is company policy to ensure employees can view only the data associated with their region,so you create and populate a table for each region. You need to enforce the regional access policy to the data.Which two actions should you take? (Choose two.)

Option A : Ensure all the tables are included in global dataset.

Option B : Ensure each table is included in a dataset for a region.

Option C : Adjust the settings for each table to allow a related region-based security group view access.

Option D : Adjust the settings for each view to allow a related region-based security group view access.

Option E : Adjust the settings for each dataset to allow a related region-based security group view access.

Correct Answer: A,D

QUESTION: 6

MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

Option A : The zone

Option B : The number of workers

Option C : The disk size per worker

Option D : The maximum number of workers

Correct Answer: A

Topic 4

Case Study: Topic 4

Title : Main Questions Set B

QUESTION: 7

You work for a manufacturing plant that batches application log files together into a single log file once a day at 2:00 AM. You have written a Google Cloud Dataflow job to process that log file. You need to make sure the log file in processed once per day as inexpensively as possible. What should you do?

Option A : Change the processing job to use Google Cloud Dataproc instead

Option B : Manually start the Cloud Dataflow job each morning when you get into the office.

Option C : Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.

Option D : Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Correct Answer: C

Topic 5

Case Study: Topic 5

Title : Practice Questions

QUESTION: 8

Which of these statements about exporting data from BigQuery is false?

Option A : To export more than 1 GB of data, you need to put a wildcard in the destination filename.

Option B : The only supported export destination is Google Cloud Storage.

Option C : Data can only be exported in JSON or Avro format.

Option D : The only compression option available is GZIP.

Correct Answer: C

Explanation/Reference:

Explanation Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported. Reference: https://cloud.google.com/bigquery/docs/exporting-data

QUESTION: 9

Which of these are examples of a value in a sparse vector? (Select 2 answers.)

Option A : [0, 5, 0, 0, 0, 0]

Option B : [0, 0, 0, 1, 0, 0, 1]

Option C : [0, 1]

Option D : [1, 0, 0, 0, 0, 0, 0]

Correct Answer: C,D

Explanation/Reference:

Explanation Categorical features in linear models are typically translated into a sparse vector in which each possible value has a corresponding index or id. For example, if there are only three possible eye colors you can represent 'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and 'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words). [0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1. [0, 5, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s. Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

QUESTION: 10

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

Option A : Execute to Google Cloud Storage; write to Google Cloud Logging

Option B : Write to Google Cloud Storage; read to Google Cloud Logging

Option C : Execute to Google Cloud Storage; execute to Google Cloud Logging

Option D : Read and write to Google Cloud Storage; write to Google Cloud Logging

Correct Answer: D

Explanation/Reference:

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

QUESTION: 11

By default, which of the following windowing behavior does Dataflow apply to unbounded data sets?

Option A : Windows at every 100 MB of data

Option B : Single, Global Window

Option C : Windows at every 1 minute

Option D : Windows at every 10 minutes

Correct Answer: B

Explanation/Reference:

Explanation Dataflow's default windowing behavior is to assign all elements of a PCollection to a single, global window, even for unbounded PCollections Reference: https://cloud.google.com/dataflow/model/pcollection

QUESTION: 12

How would you query specific partitions in a BigQuery table?

Option A : Use the DAY column in the WHERE clause

Option B : Use the EXTRACT(DAY) clause

Option C : Use the __PARTITIONTIME pseudo-column in the WHERE clause

Option D : Use DATE BETWEEN in the WHERE clause

Correct Answer: C

Explanation/Reference:

Explanation Partitioned tables include a pseudo column named _PARTITIONTIME that contains a date-based timestamp for data loaded into the table. To limit a query to particular partitions (such as Jan 1st and 2nd of 2017), use a clause similar to this: WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02') Reference: https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

QUESTION: 13

Which Java SDK class can you use to run your Dataflow programs locally?

Option A : LocalRunner

Option B : DirectPipelineRunner

Option C : MachineRunner

Option D : LocalPipelineRunner

Correct Answer: B

Explanation/Reference:

DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization. Useful for small local execution and tests Reference: https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun

QUESTION: 14

When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?

Option A : 500 TB

Option B : 1 GB

Option C : 1 TB

Option D : 500 GB

Correct Answer: C

Explanation/Reference:

Explanation Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions. It is not a good solution for less than 1 TB of data. Reference:

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

Topic 6

Case Study: Topic 6

Title : Main Questions Set C

QUESTION: 15

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

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

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

Option C : Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.

Option D : Install the StackDriver Agent and configure the MySQL plugin.

Correct Answer: C

QUESTION: 16

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the “Trust No One” (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

Option A : Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.

Option B : Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket. Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.

Option C : Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.

Option D : Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

Correct Answer: B

QUESTION: 17

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

Option A : Add a SideInput that returns a Boolean if the element is corrupt.

Option B : Add a ParDo transform in Cloud Dataflow to discard corrupt elements

Option C : Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.

Option D : Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Correct Answer: B

QUESTION: 18

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

Option A : Increase the CPU size on your server.

Option B : Increase the size of the Google Persistent Disk on your server.

Option C : Increase your network bandwidth from your datacenter to GCP.

Option D : Increase your network bandwidth from Compute Engine to Cloud Storage.

Correct Answer: C

QUESTION: 19

As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects. Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take? Choose 2 answers.

Option A : Use Cloud Deployment Manager to automate access provision.

Option B : Introduce resource hierarchy to leverage access control policy inheritance.

Option C : Create distinct groups for various teams, and specify groups in Cloud IAM policies.

Option D : Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.

Option E : For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.

Correct Answer: A,C

QUESTION: 20

You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every thirdmonth the schema of the files changes. Your requirements for implementing these transformations include:Executing the transformations on a scheduleEnabling non-developer analysts to modify transformationsProviding a graphical tool for designing transformationsWhat should you do?

Option A : Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis

Option B : Load each month’s CSV data into BigQuery, and write a SQL query to transform the data to a standard schema. Merge the transformed tables together with a SQL query

Option C : Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformation. The Python code should be stored in a revision control system and modified as the incoming data’s schema changes

Option D : Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe. Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

Correct Answer: D

QUESTION: 21

You are designing an Apache Beam pipeline to enrich data from Cloud Pub/Sub with static reference data from BigQuery. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis. Which job type and transforms should this pipeline use?

Option A : Batch job, PubSubIO, side-inputs

Option B : Streaming job, PubSubIO, JdbcIO, side-outputs

Option C : Streaming job, PubSubIO, BigQueryIO, side-inputs

Option D : Streaming job, PubSubIO, BigQueryIO, side-outputs

Correct Answer: C

QUESTION: 22

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

Option A : Create an authorized view in BigQuery to restrict access to tables with sensitive data.

Option B : Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information

Option C : Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information

Option D : Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

Correct Answer: D

QUESTION: 23

You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants. What should you do?

Option A : Increase the size of the dataset by collecting additional data.

Option B : Train a linear regression to predict a credit default risk score.

Option C : Remove the bias from the data and collect applications that have been declined loans.

Option D : Match loan applicants with their social profiles to enable feature engineering.

Correct Answer: B

QUESTION: 24

Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?

Option A : They have not assigned the timestamp, which causes the job to fail

Option B : They have not set the triggers to accommodate the data coming in late, which causes the job to fail

Option C : They have not applied a global windowing function, which causes the job to fail when the pipeline is created

Option D : They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created

Correct Answer: C

QUESTION: 25

You are migrating a table to BigQuery and are deeding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and state in which the store is located You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state city and individual store. You want to model this table to minimize query time and cost. What should you do?

Option A : Partition by transaction time; cluster by state first, then city then store ID

Option B : Partition by transaction tome cluster by store ID first, then city, then stale

Option C : Top-level cluster by stale first, then city then store

Option D : Top-level cluster by store ID first, then city then state.

Correct Answer: C

QUESTION: 26

An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and cat the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?

Option A : Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.

Option B : Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.

Option C : Write an application that makes a queue in a NoSQL database

Option D : Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic

Correct Answer: C

QUESTION: 27

Your new customer has requested daily reports that show their net consumption of Google Cloud compute

resources and who used the resources. You need to quickly and efficiently generate these daily reports. What

should you do?

Option A :

Do daily exports of Cloud Logging data to BigQuery. Create views filtering by project, log type, resource, and user.

Option B :

Filter data in Cloud Logging by project, resource, and user; then export the data in CSV format.

Option C :

Filter data in Cloud Logging by project, log type, resource, and user, then import the data into BigQuery.

Option D :

Export Cloud Logging data to Cloud Storage in CSV format. Cleanse the data using Dataprep, filtering by project, resource, and user.

Correct Answer: B

Explanation/Reference:

Explanation https://cloud.google.com/logging/docs/view/logs-explorer-interface?cloudshell=true

QUESTION: 28

You are designing a data warehouse in BigQuery to analyze sales data for a telecommunication service

provider. You need to create a data model for customers, products, and subscriptions All customers, products,

and subscriptions can be updated monthly, but you must maintain a historical record of all data. You plan to

use the visualization layer for current and historical reporting. You need to ensure that the data model is

simple, easy-to-use. and cost-effective. What should you do?

Option A :

Create a normalized model with tables for each entity. Use snapshots before updates to track historical data

Option B :

Create a normalized model with tables for each entity. Keep all input files in a Cloud Storage bucket to track historical data

Option C :

Create a denormalized model with nested and repeated fields Update the table and use snapshots to track historical data

Option D :

Create a denormalized, append-only model with nested and repeated fields Use the ingestion timestamp to track historical data.

Correct Answer: D

Explanation/Reference:

- A denormalized, append-only model simplifies query complexity by eliminating the need for joins. - Adding data with an ingestion timestamp allows for easy retrieval of both current and historical states. - Instead of updating records, new records are appended, which maintains historical information without the need to create separate snapshots.

QUESTION: 29

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

Option A :

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

Option B :

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

Option C :

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

Option D :

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

Correct Answer: A

Explanation/Reference:

To use CMEK for BigQuery, you need to create a key ring and a key in Cloud KMS, and then specify the key resource name when creating or updating a BigQuery table. You cannot change the encryption type of an existing table, so you need to create a new table with CMEK and copy the data from the old table with Google-managed encryption key.

References:

Customer-managed Cloud KMS keys | BigQuery | Google Cloud

Creating and managing encryption keys | Cloud KMS Documentation | Google Cloud

QUESTION: 30

You are designing a Dataflow pipeline for a batch processing job. You want to mitigate multiple zonal failures at job submission time. What should you do?

Option A :

Specify a worker region by using the —region flag.

Option B :

Set the pipeline staging location as a regional Cloud Storage bucket.

Option C :

Submit duplicate pipelines in two different zones by using the —zone flag.

Option D :

Create an Eventarc trigger to resubmit the job in case of zonal failure when submitting the job.

Correct Answer: A

Explanation/Reference:

By specifying a worker region, you can run your Dataflow pipeline in a multi-zone or multi-region configuration, which provides higher availability and resilience in case of zonal failures1. The —region flag allows you to specify the regional endpoint for your pipeline, which determines the location of the Dataflow service and the default location of the Compute Engine resources1. If you do not specify a zone by using the —zone flag, Dataflow automatically selects a zone within the region for your job workers1. This option is recommended over submitting duplicate pipelines in two different zones, which would incur additional costs and complexity. Setting the pipeline staging location as a regional Cloud Storage bucket does not affect the availability of your pipeline, as the staging location only stores the pipeline code and dependencies2. Creating an Eventarc trigger to resubmit the job in case of zonal failure is not a reliable solution, as it depends on the availability of the Eventarc service and the zonal resources at the time of resubmission. References:

1: Pipeline troubleshooting and debugging | Cloud Dataflow | Google Cloud

3: Regional endpoints | Cloud Dataflow | Google Cloud

QUESTION: 31

You are loading CSV files from Cloud Storage to BigQuery. The files have known data quality issues,

including mismatched data types, such as STRINGS and INT64s in the same column, and inconsistent

formatting of values such as phone numbers or addresses. You need to create the data pipeline to maintain data quality and perform the required cleansing and transformation. What should you do?

Option A :

Use Data Fusion to transform the data before loading it into BigQuery.

Option B :

Load the CSV files into a staging table with the desired schema, perform the transformations with SQL. and then write the results to the final destination table.

Option C :

Create a table with the desired schema, toad the CSV files into the table, and perform the transformations in place using SQL.

Option D :

Use Data Fusion to convert the CSV files lo a self-describing data formal, such as AVRO. before loading the data to BigOuery.

Correct Answer: A

Explanation/Reference:

Data Fusion's advantages:

Visual interface: Offers a user-friendly interface for designing data pipelines without extensive coding, making

it accessible to a wider range of users.

Built-in transformations: Includes a wide range of pre-built transformations to handle common data quality issues, such as:

Data type conversions

Data cleansing (e.g., removing invalid characters, correcting formatting)

Data validation (e.g., checking for missing values, enforcing constraints)

Data enrichment (e.g., adding derived fields, joining with other datasets)

Custom transformations: Allows for custom transformations using SQL or Java code for more complex cleaning tasks.

Scalability: Can handle large datasets efficiently, making it suitable for processing CSV files with potential data quality issues.

Integration with BigQuery: Integrates seamlessly with BigQuery, allowing for direct loading of transformed data.

QUESTION: 32

Your startup has a web application that currently serves customers out of a single region in Asia. You are targeting funding that will allow your startup lo serve customers globally. Your current goal is to optimize for cost, and your post-funding goat is to optimize for global presence and performance. You must use a native JDBC driver. What should you do?

Option A :

Use Cloud Spanner to configure a single region instance initially. and then configure multi-region C oud Spanner instances after securing funding.

Option B :

Use a Cloud SQL for PostgreSQL highly available instance first, and 8»gtable with US. Europe, and Asia replication alter securing funding

Option C :

Use a Cloud SQL for PostgreSQL zonal instance first and Bigtable with US. Europe, and Asia after securing funding.

Option D :

Use a Cloud SOL for PostgreSQL zonal instance first, and Cloud SOL for PostgreSQL with highly available configuration after securing funding.

Correct Answer: A

Explanation/Reference:

Explanation

https://cloud.google.com/spanner/docs/instance-configurations#tradeoffs_regional_versus_multi-region_configu

MIXED Questions

QUESTION: 33

In order to securely transfer web traffic data from your computer's web browser to the Cloud Dataproc cluster you should use a(n) _____.

Option A : VPN connection

Option B : Special browser

Option C : SSH tunnel

Option D : FTP connection

Correct Answer: C

Explanation/Reference:

Explanation To connect to the web interfaces, it is recommended to use an SSH tunnel to create a secure connection to the master node. Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#connecting_to_the_web_interfaces

QUESTION: 34

You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query: SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country You check the query plan for the query and see the following output in the Read section of Stage:1:What is the most likely cause of the delay for this query?

Option A : Users are running too many concurrent queries in the system

Option B : The [myproject:mydataset.mytable] table has too many partitions

Option C : Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values

Option D : Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

Correct Answer: A

QUESTION: 35

A team of data scientists has been using an on-premises cluster running Hadoop and HBase. They want to migrate to a managed service in Google Cloud. They also want to minimize changes to programs that make extensive use of the HBase API. What GCP service would you recommend?

Option A :

Bigtable

Option B :

BigQuery

Option C :

Cloud Spanner

Option D :

Cloud Dataflow

Correct Answer: A

Explanation/Reference:

The correct answer is Bigtable, which is a data store providing an HBASE compatible API. BigQuery is a data warehouse service that supports SQL but does not have an HBASE compatible API. Cloud Spanner is a relational database and not a replacement for Hadoop and HBASE. Cloud Dataflow is a data pipeline service that includes an Apache Beam runner. See https://cloud.google.com/bigtable/docs/hbase-bigtable

QUESTION: 36

You are training a deep learning model for a classification task. The precision and recall of the model is quite low. What could you do to improve the precision and recall scores?

Option A :

Use L1 regularization

Option B :

Use L2 regularization

Option C :

Use more training instances

Option D :

Use dropout

Correct Answer: C

Explanation/Reference:

The correct answer is to use more training instances. This is an example of underfitting. The other options are all regularizations used in cases of overfitting. See https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

QUESTION: 37

Analysts are using Cloud Data Studio for analyzing data sets. They would like to improve the performance of the time required to update tables and charts when working with the data. What would you recommend they try to improve performance?

Option A :

Use an imported data source

Option B :

Use a blended data source

Option C :

Use a live data source

Option D :

Use an extracted data source

Correct Answer: D

Explanation/Reference:

Extracted data sources are snapshots and can provide better performance than live data sources. Blended data sources are used to combine data from multiple data sources. There is no imported data source. See

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

QUESTION: 38

You are developing a data pipeline that will run several data transformation programs on Compute Engine virtual machines. You do not want to use your credentials for authenticating and authorizing these programs. You want to follow Google Cloud recommended practices, how would you authenticate and authorize the data transformation programs?

Option A :

Create a Gmail account and use that account to create an IAM group. Store the password for the group in Secret Manager.

Option B :

Create a service account and assign roles to the service account that are needed to execute the data transformation programs. Use Google managed keys to store both public and private portion of the service account keys.

Option C :

Create a Gmail account and use that account to create an IAM user. Store the password for the account in Secret Manager.

Option D :

Create a service account and assign roles to the service account that are needed to execute the data transformation programs. Use Secret Manager to store service account keys.

Correct Answer: B

Explanation/Reference:

Service accounts should be uses, not a user identity or a group. A service account should be created and assigned necessary roles. Google managed keys should be used for managing service accounts, not Secret Manager, which is used for secrets such as usernames and passwords. See https://cloud.google.com/docs/authentication/production

QUESTION: 39

A insurance claim review company provides expert opinion on contested insurance claims. The company uses Google Cloud for it's data analysis pipelines. Clients of the company upload documents to Cloud Storage. When a file is uploaded, the company wants to immediately move the files to a Classified Data bucket if the file contains personally identifying information. What method would you recommend to accomplish this?

Option A :

Create a quarantine bucket for uploading, once a file is uploaded trigger a Cloud Function to call a custom built machine learning model trained to detect PII. If PII is detected, move the file to the Classified Data bucket.

Option B :

Create a quarantine bucket for uploading, once a file is uploaded trigger a Cloud Function to call the Data Loss Prevention API to apply infotypes to detect PII. If PII is detected, move file to the Classified Data bucket.

Option C :

Create a quarantine bucket for uploading, use Cloud Scheduler to run a job to run hourly that will call a custom built machine learning model trained to detect PII. If PII is detected, move file to the Classified Data bucket.

Option D :

Create a quarantine bucket for uploading, use Cloud Scheduler to run a job to run hourly that will call the Data Loss Prevention API to apply infotypes to detect PII. If PII is detected, move file to the Classified Data bucket.

Correct Answer: B

Explanation/Reference:

The correct solution is to use a quarantine bucket that triggers a Cloud Function on upload to invoke the DLP API and move the file if PII is found. Cloud Scheduler runs jobs at regular intervals but this calls for immediate processing of a file once uploaded so Cloud Functions should be used. You could train a custom machine learning model but that requires development time and maintenance. A managed service like DLP is a better option. See https://cloud.google.com/dlp/docs/reference/rest

QUESTION: 40

A European health care company uses Cloud Pub/Sub as part of a data processing pipeline. The CTO of the company is concerned that data might accidentally be written to a region outside the European Union, which would violate the GDPR regulation. What would you recommend the company does to ensure data stays within Google Cloud regions in the European Union?

Option A :

Set a Resource Location Restriction organization policy to ensure all topics are stored only in acceptable regions.

Option B :

Set a Resource Location Restriction organization policy to ensure all buckets are stored only in acceptable regions.

Option C :

Only define Cloud Pub/Sub endpoints in acceptable regions when creating topics.

Option D :

Only define Cloud Pub/Sub endpoints in acceptable regions when creating subscriptions.

Correct Answer: A

Explanation/Reference:

The correct answer is to set a Resource Location Restriction organization policy to ensure all topics are stored only in acceptable regions. Topics, not buckets, store messages in Cloud Pub/Sub. Users of Cloud Pub/Sub do not create endpoints; it is a globally managed service that does not require any endpoint configuration by users of Cloud Pub/Sub. see https://cloud.google.com/resource-manager/docs/organization-policy/defining-locations

QUESTION: 41

A team of researchers is running a high performance distributed computing platform on premises but wants to migrate to Google Cloud. The platform uses virtual machines. The researchers want to be able to scale up the number of virtual machines in the cluster based on CPU load. What would you recommend they use?

Option A :

Kubernetes cluster

Option B :

Managed instance groups

Option C :

Unmanaged instance groups

Option D :

Cloud Run

Correct Answer: B

Explanation/Reference:

The correct answer is managed instance groups, which is a way of deploying Compute Engine instances basesed on a template. Kubernetes is used for running containers, not virtual machines. Cloud Run is also used to run containers not virtual machines. Unmanaged instance groups run virtual machines but do not support autoscaling. See https://cloud.google.com/compute/docs/instance-groups

QUESTION: 42

A data analyst currently has the bigquery.dataViewer role and can successfully query a materialized view. They also want to be able to refresh the materialized view. You want to use a predefined role but not grant them any more permissions than needed to refresh the materialized view. What predefined role would you grant to the user?

Option A :

bigquery.dataOwner

Option B :

bigquery.admin

Option C :

bigquery.dataEditor

Option D :

bigquery.mvUpdater

Correct Answer: C

Explanation/Reference:

The correct answer is bigquery.dataEditor, which can refresh a materialized view. Both bigquery.admin and bigquery.dataAdmin grant more permissions than needed. There is no bigquery.mvUpdater role. See https://cloud.google.com/bigquery/docs/access-control

QUESTION: 43

A data pipeline uses Cloud Pub/Sub for ingesting data. The data is stored in topics and a Dataflow workflow reads from a subscription to that topic, processes the data, and writes output to BigQuery. What is the recommended way to authenticate when reading data from Cloud Pub/Sub?

Option A :

Custom role

Option B :

Google Workspace Identity

Option C :

Use service accounts

Option D :

Basic role

Correct Answer: C

Explanation/Reference:

Service accounts are the recommended way to authentic for most use cases when using Cloud Pub/Sub. Google Workspace Identity should be used by human users, service accounts are used for applications. Custom roles and basic roles are for authorization not authentication. See https://cloud.google.com/iam/docs/service-accounts

QUESTION: 44

Messages are unexpectedly accumulating in service using Cloud Pub/Sub. A developer unfamiliar with Cloud Pub/Sub has asked for our help in diagnosing the problem. What would you point out with respect to how messages are removed from Cloud Pub/Sub topics?

Option A :

Once at least one subscriber for each topic has acknowledged the message it will be deleted from storage.

Option B :

Once at least one subscriber for each bucket has acknowledged the message it will be deleted from storage.

Option C :

Once at least one subscriber for any subscription has acknowledged the message it will be deleted from storage.

Option D :

Once at least one subscriber for each subscription has acknowledged the message it will be deleted from storage.

Correct Answer: D

Explanation/Reference:

The correct answer is that a message is deleted once at least one subscriber for a each subscription has acknowledged the message it will be deleted from storage. See https://cloud.google.com/pubsub/docs/subscriber