Google Cloud - Professional Data Engineer Practice Exams 4.2 (185 ratings) !!!!!!

Notebook:	GCP EXAM		
Created:	12/2/2019 12:58 PM	**Updated:**	12/7/2019 10:31 PM
Author:	Venus		
Tags:	20191212, valuable, with explanation		
URL:	https://www.udemy.com/course/google-cloud-certified-professional-data-engineer-pra...		

Google Cloud Certified - Professional Data Engineer Practice Exam 1 - Results

Attempt 2

Question 1: Correct

You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old. What should you do?

A. Disable caching by editing the report settings.**(Correct)**

B. Disable caching in BigQuery by editing table details.

C. Refresh your browser tab showing the visualizations.

D. Clear your browser history for the past hour then reload the tab showing the visualizations.

Explanation

Correct answer is **A** as Data Studio caches data for performance and as the latest data is not shown, the caching can be disabled to fetch the latest data.

Refer GCP documentation - [Data Studio Caching]

Option B is wrong as BigQuery does not cache the data.

Options C & D are wrong this would not allow fetching of latest data.

Question 2: **Correct**

You company's on-premises Hadoop and Spark jobs have been migrated to Cloud Dataproc. When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through which proxy?

A. HTTPS

B. VPN

C. SOCKS **(Correct)**

D. HTTP

Explanation

Correct answer is **C** as the internal services can be accessed using the SOCKS proxy server.

Refer GCP documentation - [Dataproc - Connecting to web interfaces]

You can connect to web interfaces running on a Cloud Dataproc cluster using your project's Cloud Shell or the Cloud SDK gcloud command-line tool:

Cloud Shell: The Cloud Shell in the Google Cloud Platform Console has the Cloud SDK commands and utilities pre-installed, and it provides a Web Preview feature that allows you to quickly connect through an SSH tunnel to a web interface port on a cluster. However, a connection to the cluster from Cloud Shell uses local port forwarding, which

opens a connection to only one port on a cluster web interface—multiple commands are needed to connect to multiple ports. Also, Cloud Shell sessions automatically terminate after a period of inactivity (30 minutes).
`gcloud` command-line tool: The `gcloud compute ssh` command with dynamic port forwarding allows you to establish an SSH tunnel and run a SOCKS proxy server on top of the tunnel.
After issuing this command, you must configure your local browser to use the SOCKS proxy. This connection method allows you to connect to multiple ports on a cluster web interface.

Question 3: **Correct**

Your company is planning to migrate their on-premises Hadoop and Spark jobs to Dataproc. Which role must be assigned to a service account used by the virtual machines in a Dataproc cluster, so they can execute jobs?

A. Dataproc Worker **(Correct)**

B. Dataproc Viewer

C. Dataproc Runner

D. Dataproc Editor

Explanation

Correct answer is **A** as the compute engine should have Dataproc Worker role assigned.

Refer GCP documentation - [Dataproc Service Accounts]

Service accounts have [IAM roles] granted to them. Specifying a user-managed service account when creating a Cloud Dataproc cluster allows you to create and utilize clusters with fine-grained access and control to Cloud resources. Using multiple user-managed service accounts with different Cloud Dataproc clusters allows for clusters with different access to Cloud resources.

Service accounts used with Cloud Dataproc must have [Dataproc/Dataproc Worker] role (or have all the permissions granted by Dataproc Worker role).

Question 4: Correct

You currently have a Bigtable instance you've been using for development running a development instance type, using HDD's for storage. You are ready to upgrade your development instance to a production instance for increased performance. You also want to upgrade your storage to SSD's as you need maximum performance for your instance. What should you do?

A. Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD.

B. Export your Bigtable data into a new instance, and configure the new instance type as production with SSD's **(Correct)**

C. Run parallel instances where one instance is using HDD and the other is using SSD.

D. Use the Bigtable instance sync tool in order to automatically synchronize two different instances, with one having the new storage configuration.

Explanation

Correct answer is **B** as the storage for the cluster cannot be updated. You need to define the new cluster and copy or import the data to it.

Refer GCP documentation - Bigtable Choosing HDD vs SSD

Switching between SSD and HDD storage

When you create a Cloud Bigtable instance and cluster, your choice of SSD or HDD storage for the cluster is permanent. You cannot use the Google Cloud Platform Console to change the type of storage that is used for the cluster.

If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance. Alternatively, you can use a Cloud Dataflow or Hadoop MapReduce job to copy the data from one instance to another. Keep in mind that migrating an entire instance takes time, and you might need to add nodes to your Cloud Bigtable clusters before you migrate your instance.

Option A is wrong as storage type cannot be changed.

Options C & D are wrong as it would have two clusters running at the same time with same data, thereby increasing cost.

Question 5: **Correct**

You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type. Reload the data.

B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column DT for each row. Reference the column TS instead of the column DT from now on.

C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.

D. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, **(Correct)** new data is loaded into the table NEW_CLICK_STREAM.

Explanation

Correct answer is **D** as the column type cannot be changed and the column needs to casting loaded into a new table using either SQL Query or import/export.

Refer GCP documentation - BigQuery Changing Schema

Changing a column's data type is not supported by the GCP Console, the classic BigQuery web UI, the command-line tool, or the API. If you attempt to update a table by applying a schema that specifies a new data type for a column, the following error is returned: `BigQuery error in update operation: Provided Schema does not match Table [PROJECT_ID]:[DATASET].[TABLE].`

There are two ways to manually change a column's data type:

Using a SQL query — Choose this option if you are more concerned about simplicity and ease of use, and you are less concerned about costs.
Recreating the table — Choose this option if you are more concerned about costs, and you are less concerned about simplicity and ease of use.

Option 1: Using a query

Use a SQL query to select all the table data and to cast the relevant column as a different data type. You can use the query results to overwrite the table or to create a new destination table.

Option A is wrong as with this approach all the data would be lost and needs to be reloaded

Option B is wrong as numeric values cannot be used directly and would need casting.

Option C is wrong as view is not materialized views, so the future queries would always be taxed as the casting would be done always.

Question 6: **Correct**

Your company has a BigQuery dataset created, which is located near Tokyo. For efficiency reasons, the company now wants the dataset duplicated in Germany. How can be dataset be made available to the users in Germany?

A. Change the dataset from a regional location to multi-region location, specifying the regions to be included.

B. Export the data from BigQuery into a bucket in the new location, and import it into a new dataset at the new location.

C. Copy the data from the dataset in the source region to the dataset in the target region using BigQuery commands.

D. Export the data from BigQuery into nearby bucket in Cloud Storage. Copy to a new regional bucket in Cloud Storage in the new location and Import into the new dataset. **(Correct)**

Explanation

Correct answer is **D** as the dataset location cannot be changed once created. The dataset needs to be copied using Cloud Storage.

Refer GCP documentation - [BigQuery Exporting Data]

You cannot change the location of a dataset after it is created. Also, you cannot move a dataset from one location to another. If you need to move a dataset from one location to another, follow this process:

1. Export the data from your BigQuery tables to a regional or multi-region Cloud Storage bucket in the same location as your dataset. For example, if your dataset is in the EU multi-region location, export your data into a regional or multi-region bucket in the EU.There are no charges for exporting data from BigQuery, but you do incur charges for storing the exported data in Cloud Storage. BigQuery exports are subject to the limits on export jobs.
2. Copy or move the data from your Cloud Storage bucket to a regional or multi-region bucket in the new location. For example, if you are moving your data from the US multi-region location to the Tokyo regional location, you would transfer the data to a regional bucket in Tokyo. Note that transferring data between regions incurs network egress charges in Cloud Storage.

3. After you transfer the data to a Cloud Storage bucket in the new location, create a new BigQuery dataset (in the new location). Then, load your data from the Cloud Storage bucket into BigQuery.You are not charged for loading the data into BigQuery, but you will incur charges for storing the data in Cloud Storage until you delete the data or the bucket. You are also charged for storing the data in BigQuery after it is loaded. Loading data into BigQuery is subject to the limits on load jobs.

Question 7: **Correct**

A company has loaded its complete financial data for last year for analytics into BigQuery. A Data Analyst is concerned that a BigQuery query could be too expensive. Which methods can be used to reduce the number of rows processed by BigQuery?

A. Use the LIMIT clause to limit the number of values in the results.

B. Use the SELECT clause to limit the amount of data in the query. Partition data by date so the query can be more focused. **(Correct)**

C. Set the Maximum Bytes Billed, which will limit the number of bytes processed but still run the query if the number of bytes requested goes over the limit.

D. Use GROUP BY so the results will be grouped into fewer output values.

Explanation

Correct answer is **B** as SELECT with partition would limit the data for querying.

Refer GCP documentation - BigQuery Cost Best Practices

Best practice: Partition your tables by date.

If possible, partition your BigQuery tables by date. Partitioning your tables allows you to query relevant subsets of data which improves performance and reduces costs.

For example, when you query partitioned tables, use the `_PARTITIONTIME` pseudo column to filter for a date or a range of dates. The query processes data only in the partitions that are specified by the date or range.

Option A is wrong as LIMIT does not reduce cost as the amount of data queried is still the same.

Best practice: Do not use a `LIMIT` clause as a method of cost control.

Applying a `LIMIT` clause to a query does not affect the amount of data that is read. It merely limits the results set output. You are billed for reading all bytes in the entire table as indicated by the query.

The amount of data read by the query counts against your free tier quota despite the presence of a `LIMIT` clause.

Option C is wrong as the query would fail and would not execute if the Maximum bytes limit is exceeded by the query.

Best practice: Use the maximum bytes billed setting to limit query costs.

You can limit the number of bytes billed for a query using the maximum bytes billed setting. When you set maximum bytes billed, if the query will read bytes beyond the limit, the query fails without incurring a charge.

Option D is wrong as GROUP BY would return less output, but would still query the entire data.

Question 8: **Correct**

Your company receives streaming data from IoT sensors capturing various parameters. You need to calculate a running average for each of the parameter on streaming data, taking into account the data that can arrive late and out of order. How would you design the system?

A. Use Cloud Pub/Sub and Cloud Dataflow with Sliding Time Windows. **(Correct)**

B. Use Cloud Pub/Sub and Google Data Studio.

C. Cloud Pub/Sub can guarantee timely arrival and order.

D. Use Cloud Dataflow's built-in timestamps for ordering and filtering.

Explanation

Correct answer is **A** as Cloud Pub/Sub does not maintain message order and Dataflow can be used to order the messages and as well as calculate average using Sliding Time window.

Refer GCP documentation - Pub/Sub Subscriber

Cloud Pub/Sub delivers each message once and in the order in which it was published. However, messages may sometimes be delivered out of order or more than once. In general, accommodating more-than-once delivery requires your subscriber to be idempotent when processing messages. You can achieve exactly once processing of Cloud Pub/Sub message streams using Cloud Dataflow `PubsubIO`. `PubsubIO` de-duplicates messages on custom message identifiers or those assigned by Cloud Pub/Sub. You can also achieve ordered processing with Cloud Dataflow by using the standard sorting APIs of the service. Alternatively, to achieve ordering, the publisher of the topic to which you subscribe can include a sequence token in the message.

Option B is wrong as Data Studio is more of a visualization tool and does not help in analysis or ordering of messages.

Option C is wrong as Cloud Pub/Sub does not guarantee order and arrival.

Option D is wrong as Dataflow does not provide built-in timestamps for ordering and filtering. It needs to use the watermark/timestamp introduced either by the publisher source or Cloud Pub/Sub.

Question 9: **Correct**

You have developed a Machine Learning model to categorize where the financial transaction was a fraud or not. Testing the Machine Learning model with validation data returns 100% correct answers. What can you infer from the results?

A. The model is working extremely well, indicating the hyperparameters are set correctly.

B. The model is overfit. There is a problem. **(Correct)**

C. The model is underfit. There is a problem.

D. The model is perfectly fit. You do not need to continue training.

Explanation

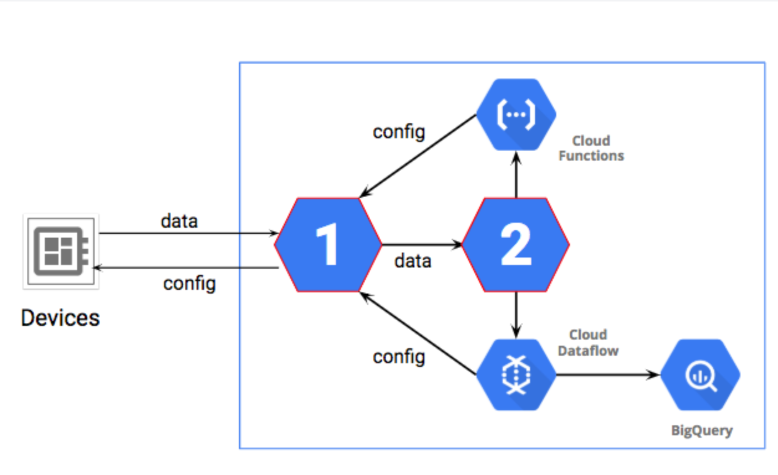
Correct answer is **B** as the 100% accuracy is an indicator that the validation data may have somehow gotten mixed in with the training data. You will need new validation data to generate recognizable error.

Overfitting results when a model performs well on the training set, generating only a small error, but struggles with new or unknown data. In other words, the model overfits itself to the data. Instead of training a model to pick out general features in a given type of data, an overtrained model learns only how to pick out specific features found in the training set.

Question 10: Correct

A company has a new IoT pipeline. Which services will make this design work?

Select the services that should be used to replace the icons with the number "1" and number "2" in the diagram.



A. Cloud IoT Core, Cloud Datastore

B. Cloud Pub/Sub, Cloud Storage

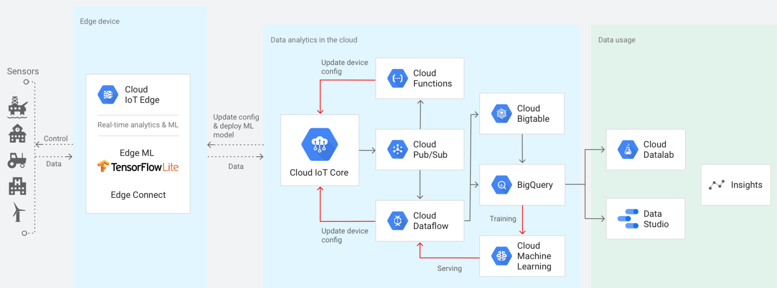
C. Cloud IoT Core, Cloud Pub/Sub **(Correct)**

D. App Engine, Cloud IoT Core

Explanation

Correct answer is **C** as device data captured by Cloud IoT Core gets published to Cloud Pub/Sub, which can then trigger Dataflow and Cloud Functions.

Refer GCP documentation - Cloud IoT Core



Cloud IoT Core is a fully managed service that allows you to easily and securely connect, manage, and ingest data from millions of globally dispersed devices. Cloud IoT Core, in combination with other services on Cloud IoT platform, provides a complete solution for collecting, processing, analyzing, and visualizing IoT data in real time to support improved operational efficiency.

Cloud IoT Core, using Cloud Pub/Sub underneath, can aggregate dispersed device data into a single global system that integrates seamlessly with Google Cloud data analytics services. Use your IoT data stream for advanced analytics, visualizations, machine learning, and more to help improve operational efficiency, anticipate problems, and build rich models that better describe and optimize your business.

Question 11: **Correct**

You are building storage for files for a data pipeline on Google Cloud. You want to support JSON files. The schema of these files will occasionally change. Your analyst teams will use running aggregate ANSI SQL queries on this data. What should you do?

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

B. Use BigQuery for storage. Select "Automatically detect" in the Schema section. **(Correct)**

C. Use Cloud Storage for storage. Link data as temporary tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

D. Use Cloud Storage for storage. Link data as permanent tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

Explanation

Correct answer is **B** as the requirement is to support occasionally (schema) changing JSON files and aggregate ANSI SQL queries: you need to use BigQuery, and it is quickest to use 'Automatically detect' for schema changes.

Refer GCP documentation - BigQuery 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 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.

To see the detected schema for a table:

Use the command-line tool's bq show command
Use the BigQuery web UI to view the table's schema
When enabled, BigQuery makes a best-effort attempt to automatically infer the schema for CSV and JSON files.

A is not correct because you should not provide format files: you can simply turn on the 'Automatically detect' schema changes flag.

C and D are not correct as Cloud Storage is not ideal for this scenario; it is cumbersome, adds latency and doesn't add value.

Question 12: **Correct**

You have 250,000 devices which produce a JSON device status event every 10 seconds. You want to capture this event data for outlier time series analysis. What should you do?

A. Ship the data into BigQuery. Develop a custom application that uses the BigQuery API to query the dataset and displays device outlier data based on your business requirements.

B. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.

C. Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements. **(Correct)**

D. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data based on your business requirements.

Explanation

Correct answer is **C** as the time series data with its data type, volume, and query pattern best fits BigTable capabilities.

Refer GCP documentation - Bigtable Time Series data and CBT

Options A & B are wrong as BigQuery is not suitable for the query pattern in this scenario.

Option D is wrong as you can use the simpler method of 'cbt tool' to support this scenario.

Question 13: **Correct**

You are building a data pipeline on Google Cloud. You need to select services that will host a deep neural network machine-learning model also hosted on Google Cloud. You also need to monitor and run jobs that could occasionally fail. What should you do?

A. Use Cloud Machine Learning to host your model. Monitor the status of the Operation object for 'error' results.

B. Use Cloud Machine Learning to host your model. Monitor the status of the Jobs object for 'failed' job states. **(Correct)**

C. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Jobs object for 'failed' job states.

D. Use a Kubernetes Engine cluster to host your model. Monitor the status of Operation object for 'error' results.

Explanation

Correct answer is **B** as the requirement is to host an Machine Learning Deep Neural Network job it is ideal to use the Cloud Machine Learning service. Monitoring works on Jobs object.

Refer GCP documentation - ML Engine Managing Jobs

You can use projects.jobs.get to get the status of a job. This method is also provided as `gcloud ml jobs describe`and in the **Jobs** page in the Google Cloud Platform Console. Regardless of how you get the status, the information is based on the members of the Job resource. You'll know the job is complete when `Job.state` in the response is equal to one of these values:

`SUCCEEDED`
`FAILED`
`CANCELLED`
Option A is wrong as monitoring should not be on Operation object to monitor failures.

Options C & D are wrong as you should not use a Kubernetes Engine cluster for Machine Learning jobs.

Question 14: Correct

You are developing an application on Google Cloud that will label famous landmarks in users' photos. You are under competitive pressure to develop the predictive model quickly. You need to keep service costs low. What should you do?

A. Build an application that calls the Cloud Vision API. Inspect the generated MID values to supply the image labels.

B. Build an application that calls the Cloud Vision API. Pass landmark locations as base64-encoded strings. **(Correct)**

C. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Pass landmark locations as base64-encoded strings.

D. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Inspect the generated MID values to supply the image labels.

Explanation

Correct answer is **B** as the requirement is to quickly develop a model that generates landmark labels from photos, it can be easily supported by Cloud Vision API.

Refer GCP documentation - [Cloud Vision]

Cloud Vision offers both pretrained models via an API and the ability to build custom models using AutoML Vision to provide flexibility depending on your use case.

Cloud Vision API enables developers to understand the content of an image by encapsulating powerful machine learning models in an easy-to-use REST API. It quickly classifies images into thousands of categories (such as, “sailboat”), detects individual objects and faces within images, and reads printed words contained within images. You can build metadata on your image catalog, moderate offensive content, or enable new marketing scenarios through image sentiment analysis.

Option A is wrong as you should not inspect the generated MID values; instead, you should simply pass the image locations to the API and use the labels, which are output.

Options C & D are wrong as you should not build a custom classification TF model for this scenario, as it would require time.

Question 15: **Correct**

You regularly use prefetch caching with a Data Studio report to visualize the results of BigQuery queries. You want to minimize service costs. What should you do?

A. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and direct the users to view the report only once per business day (24-hour period).

B. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is selected for the report. **(Correct)**

C. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and also set it up to be a 'view-only' report.

D. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is not selected for the report.

Explanation

Correct option is **B** as you must set Owner credentials to use the 'enable cache' option in BigQuery. It is also a Google best practice to use the ‘enable cache’ option when the business scenario calls for using prefetch caching.

Refer GCP documentation - Datastudio data caching

The prefetch cache 预取缓存 is only active for data sources that use owner's credentials to access the underlying data.

Options A, C, & D are wrong as cache auto-expires every 12 hours; a prefetch cache is only for data sources that use the Owner's credentials and not the Viewer's credentials

Question 16: **Correct**

Your customer is moving their corporate applications to Google Cloud Platform. The security team wants detailed visibility of all projects in the organization. You provision the Google Cloud Resource Manager and set up yourself as the org admin. What Google Cloud Identity and Access

Management (Cloud IAM) roles should you give to the security team?

A. Org viewer, project owner

B. Org viewer, project viewer **(Correct)**

C. Org admin, project browser

D. Project owner, network admin

Explanation

Correct answer is **B** as the security team only needs visibility to the projects, project viewer provides the same with the best practice of least privilege.

Refer GCP documentation - Organization & Project access control

Option A is wrong as project owner will provide access however it does not align with the best practice of least privilege.

Option C is wrong as org admin does not align with the best practice of least privilege.

Option D is wrong as the user needs to be provided organization viewer access to see the organization.

Question 17: Correct

You want to optimize the performance of an accurate, real-time, weather-charting application. The data comes from 50,000 sensors sending 10 readings a second, in the format of a timestamp and sensor reading. Where should you store the data?

A. Google BigQuery

B. Google Cloud SQL

C. Google Cloud Bigtable **(Correct)**

D. Google Cloud Storage

Explanation

Correct answer is **C** as Bigtable is a ideal solution for storing time series data. Storing time-series data in Cloud Bigtable is a natural fit. Cloud Bigtable stores data as unstructured columns in rows; each row has a row key, and row keys are sorted lexicographically.

Refer GCP documentation - Storage Options

Google Cloud Bigtable	A scalable, fully-managed NoSQL wide-column database that is suitable for both real-time access and analytics workloads.	Low-latency read/write access High-throughput analytics Native time series support	IoT, finance, adtech Personalization, recommendations Monitoring Geospatial datasets Graphs

Option A is wrong as Google BigQuery is a scalable, fully-managed Enterprise Data Warehouse (EDW) with SQL and fast response times. It is for analytics and OLAP workload, though it also provides storage capacity and price similar to GCS. It cannot handle the required real time ingestion of data.

Option B is wrong as Google Cloud SQL is a fully-managed MySQL and PostgreSQL relational database service for Structured data and OLTP workloads. It also won't stand for this type of high ingesting rate in real time.

Option D is wrong as Google Cloud Storage is a scalable, fully-managed, highly reliable, and cost-efficient object / blob store. It cannot stand for this amount of data streaming ingestion rate in real-time.

Question 18: Correct

You need to take streaming data from thousands of Internet of Things (IoT) devices, ingest it, run it through a processing pipeline, and store it for analysis. You want to

run SQL queries against your data for analysis. What services in which order should you use for this task?

A. Cloud Dataflow, Cloud Pub/Sub, BigQuery

B. Cloud Pub/Sub, Cloud Dataflow, Cloud Dataproc

C. Cloud Pub/Sub, Cloud Dataflow, BigQuery **(Correct)**

D. App Engine, Cloud Dataflow, BigQuery

Explanation

Correct answer is **C** as the need to ingest it, transform and store the Cloud Pub/Sub, Cloud Dataflow, BigQuery is ideal stack to handle the IoT data.

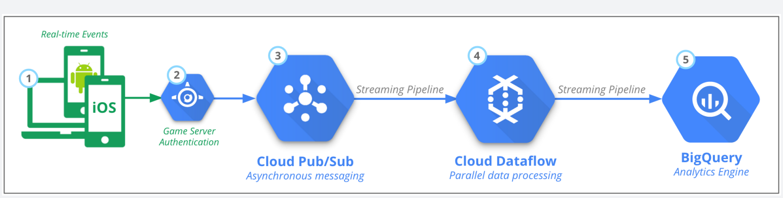
Refer GCP documentation - IoT

Google Cloud Pub/Sub provides a globally durable message ingestion service. By creating topics for streams or channels, you can enable different components of your application to subscribe to specific streams of data without needing to construct subscriber-specific channels on each device. Cloud Pub/Sub also natively connects to other Cloud Platform services, helping you to connect ingestion, data pipelines, and storage systems.

Google Cloud Dataflow provides the open Apache Beam programming model as a managed service for processing data in multiple ways, including batch operations, extract-transform-load (ETL) patterns, and continuous, streaming computation. Cloud Dataflow can be particularly useful for managing the high-volume data processing pipelines required for IoT scenarios. Cloud Dataflow is also designed to integrate seamlessly with the other Cloud Platform services you choose for your pipeline.

Google BigQuery provides a fully managed data warehouse with a familiar SQL interface, so you can store your IoT data alongside any of your other enterprise analytics and logs. The performance and cost of BigQuery means you might keep your valuable data longer, instead of deleting it just to save disk space.

Sample Arch - Mobile Gaming Analysis Telemetry



Option A is wrong as the stack is correct, however the order is not correct.

Option B is wrong as Dataproc is not an ideal tool for analysis. Cloud **Dataproc** is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.

Option D is wrong as App Engine is not an ideal ingestion tool to handle IoT data.

Question 19: Correct

Your company is planning the infrastructure for a new large-scale application that will need to store over 100 TB or a petabyte of data in NoSQL format for Low-latency read/write and High-throughput analytics. Which storage option should you use?

A. Cloud Bigtable **(Correct)**

B. Cloud Spanner

C. Cloud SQL

D. Cloud Datastore

Explanation

Correct answer is **A** as Bigtable is an ideal solution to provide low latency, high throughput data processing storage option with analytics

Refer GCP documentation - Storage Options

Cloud Bigtable	A scalable, fully managed NoSQL wide-column database that is suitable for both low-latency	Low-latency read/write access High-throughput data processing Time series support	IoT, finance, adtech Personalization, recommendations Monitoring Geospatial datasets Graphs

	single-point lookups and precalculated analytics.		

Options B & C are wrong as they are relational databases

Option D is wrong as Cloud Datastore is not ideal for analytics.

Question 20: **Correct**

You have hundreds of IoT devices that generate 1 TB of streaming data per day. Due to latency, messages will often be delayed compared to when they were generated. You must be able to account for data arriving late within your processing pipeline. How can the data processing system be designed?

A. Use Cloud SQL to process the delayed messages.

B. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Dataflow to process messages, and use windows, watermarks (timestamp), and triggers to process late data. **(Correct)**

C. Use SQL queries in BigQuery to analyze data by timestamp.

D. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Pub/Sub to process messages by timestamp and fix out of order issues.

Explanation

Correct answer is **B** as Cloud Pub/Sub can help handle the streaming data. However, Cloud Pub/Sub does not handle the ordering, which can be done using Dataflow and adding watermarks to the messages from the source.

Refer GCP documentation - Cloud Pub/Sub ordering & Subscriber

How do you assign an order to messages published from different publishers? Either the publishers themselves have to coordinate, or the message delivery service itself has to attach a notion of order to every incoming message. Each message would need to include the ordering information. The order information could be a timestamp (though it has to be a timestamp that all servers get from the same source in order to avoid issues of clock drift), or a sequence number (acquired from a single source with ACID guarantees). Other messaging systems that guarantee ordering of messages require settings that effectively limit the system to multiple publishers sending messages through a single server to a single subscriber.

Typically, Cloud Pub/Sub delivers each message once and in the order in which it was published. However, messages may sometimes be delivered out of order or more than once. In general, accommodating more-than-once delivery requires your subscriber to be idempotent when processing messages. You can achieve exactly once processing of Cloud Pub/Sub message streams using Cloud Dataflow `PubsubIO`. `PubsubIO` de-duplicates messages on custom message identifiers or those assigned by Cloud Pub/Sub. **You can also achieve ordered processing with Cloud Dataflow by using the standard sorting APIs of the service. Alternatively, to achieve ordering, the publisher of the topic to which you subscribe can include a sequence token in the message.**

Options A & C are wrong as SQL and BigQuery do not support ingestion and ordering of IoT data and would need other services like Pub/Sub.

Option D is wrong as Cloud Pub/Sub does not perform ordering of messages.

Question 21: **Correct**

Your company has data stored in BigQuery in Avro format. You need to export this Avro formatted data from BigQuery into Cloud Storage. What is the best method of doing so from the web console?

A. Convert the data to CSV format the BigQuery export options, then make the transfer.

B. Use the BigQuery Transfer Service to transfer Avro data to Cloud Storage.

C. Click on Export Table in BigQuery, and provide the Cloud Storage location to export to **(Correct)**

D. Create a Dataflow job to manage the conversion of Avro data to CSV format, then export to Cloud Storage.

Explanation

Correct answer is **C** as BigQuery can export Avro data natively to Cloud Storage.

Refer GCP documentation - [BigQuery Exporting Data]

After 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. When you export your data to multiple files, the size of the files will vary.

You cannot export data to a local file or to Google Drive, but you can save query results to a local file. The only supported export location is Google Cloud Storage.

For **Export format**, choose the format for your exported data: CSV, JSON (Newline Delimited), or Avro.

Option A is wrong as BigQuery can export Avro data natively to Cloud Storage and does not need to be converted to CSV format.

Option B is wrong as BigQuery Transfer Service is for moving BigQuery data to Google SaaS applications (AdWords, DoubleClick, etc.). You will want to do a normal export of data, which works with Avro formatted data.

Option D is wrong as Google Cloud Dataflow can be used to read data from BigQuery instead of manually exporting it, but doesn't work through console.

Question 22: Correct

Your company has its input data hosted in BigQuery. They have existing Spark scripts for performing analysis which they want to reuse. The output needs to be stored in BigQuery for future analysis. How can you set up your

Dataproc environment to use BigQuery as an input and output source?

A. Use the Bigtable syncing service built into Dataproc.

B. Manually use a Cloud Storage bucket to import and export to and from both BigQuery and Dataproc

C. Install the BigQuery connector on your Dataproc cluster **(Correct)**

D. You can only use Cloud Storage or HDFS for your Dataproc input and output.

Explanation

Correct answer is **C** as Dataproc has a BigQuery connector library which allows it directly interface with BigQuery.

Refer GCP documentation - [Dataproc BigQuery Connector]

You can use a BigQuery connector to enable programmatic read/write access to BigQuery. This is an ideal way to process data that is stored in BigQuery. No command-line access is exposed. The BigQuery connector is a Java library that enables Hadoop to process data from BigQuery using abstracted versions of the Apache Hadoop InputFormat and OutputFormat classes.

Option A is wrong Bigtable syncing service does not exist.

Options B & D are wrong as Dataproc can directly interface with BigQuery.

Question 23: **Correct**

You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

A. Include ORDER BY DESK on timestamp column and LIMIT to 1.

B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.

C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.

D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1. **(Correct)**

Explanation

Correct answer is **D** as the best approach is to ROW_NUMBER with PARTITION by the UNIQUE_ID and filter it by row_number = 1.

Refer GCP documentation - [BigQuery Streaming Data - Removing Duplicates]

To remove duplicates, perform the following query. You should specify a destination table, allow large results, and disable result flattening.

```
#standardSQL SELECT * EXCEPT(row_number) FROM
( SELECT *, ROW_NUMBER() OVER (PARTITION BY I
D_COLUMN) row_number FROM `TABLE_NAME`) WHERE
row_number = 1
```

Question 24: Correct

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three)

A. Load data into different partitions.

B. Load data into a different dataset for each client. **(Correct)**

C. Put each client's BigQuery dataset into a different table.

D. Restrict a client's dataset to approved users. **(Correct)**

E. Only allow a service account to access the datasets.

F. Use the appropriate identity and access management (IAM) roles for each client's users. **(Correct)**

Explanation

Correct answers are **B, D & F**. As the access control can be done using IAM roles on the dataset only to the specific approved users.

Refer GCP documentation - BigQuery Access Control

BigQuery uses Identity and Access Management (IAM) to manage access to resources. The three types of resources available in BigQuery are organizations, projects, and datasets. In the IAM policy hierarchy, datasets are child resources of projects. Tables and views are child resources of datasets — they inherit permissions from their parent dataset.

To grant access to a resource, assign one or more roles to a user, group, or service account. Organization and project roles affect the ability to run jobs or manage the project's resources, whereas dataset roles affect the ability to access or modify the data inside of a particular dataset.

Options A & C are wrong as the access control can only be applied on dataset and views, not on partitions and tables.

Option E is wrong as service account is mainly for machines and would be a single account.

Question 25: **Correct**

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it

is slowing her down. You want to help her perform her tasks. What should you do?

A. Run a local version of Jupiter on the laptop.

B. Grant the user access to Google Cloud Shell.

C. Host a visualization tool on a VM on Google Compute Engine.

D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine. **(Correct)**

Explanation

Correct answer is **D** as Cloud Datalab provides a powerful interactive, scalable tool on Google Cloud with the ability to analyze, visualize data.

Refer GCP documentation - Datalab

Cloud Datalab is a powerful interactive tool created to explore, analyze, transform and visualize data and build machine learning models on Google Cloud Platform. It runs on Google Compute Engine and connects to multiple cloud services easily so you can focus on your data science tasks.

Cloud Datalab is built on Jupyter (formerly IPython), which boasts a thriving ecosystem of modules and a robust knowledge base. Cloud Datalab enables analysis of your data on Google BigQuery, Cloud Machine Learning Engine, Google Compute Engine, and Google Cloud Storage using Python, SQL, and JavaScript (for BigQuery user-defined functions).

Whether you're analyzing megabytes or terabytes, Cloud Datalab has you covered. Query terabytes of data in BigQuery, run local analysis on sampled data and run training jobs on terabytes of data in Cloud Machine Learning Engine seamlessly.

Use Cloud Datalab to gain insight from your data. Interactively explore, transform, analyze, and visualize your data using BigQuery, Cloud Storage and Python.

Go from data to deployed machine-learning (ML) models ready for prediction. Explore data, build, evaluate and optimize Machine Learning models using TensorFlow or Cloud Machine Learning Engine.

Options A, B & C do not provides all the abilities.

Question 26: **Correct**

You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

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

B. Grant the consultant the Cloud Dataflow Developer role on the project. **(Correct)**

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

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

Explanation

Correct answer is **B** as the Dataflow developer role would help provide the third-party consultant access to create and work on the Dataflow pipeline. However, it does not provide access to view the data, thus maintaining user's privacy.

Refer GCP documentation - [Dataflow roles]

`roles/dataflow.viewer`	`dataflow.<resource-type>.list` `dataflow.<resource-type>.get`	jobs, messages, metrics
`roles/dataflow.developer`	All of the above, as well as: `dataflow.jobs.create` `dataflow.jobs.drain` `dataflow.jobs.cancel`	jobs
`roles/dataflow.admin`	All of the above, as well as: `compute.machineTypes.get` `storage.buckets.get` `storage.objects.create` `storage.objects.get` `storage.objects.list`	NA

Option A is wrong as it would not allow the consultant to work on the pipeline.

Option C is wrong as the consultant cannot use the service account to login.

Option D is wrong as it does not enable collabaration.

Question 27: **Correct**

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?

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

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output. **(Correct)**

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

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

Explanation

Correct answer is **B** as the issue can be debugged by running a fixed dataset and checking the output.

Refer GCP documentation - Dataflow logging

Option A is wrong as the Dashboard uses data provided by Dataflow, the input source for Dashboard seems to be the issue

Option C is wrong as Monitoring would not help find missing messages in Cloud Pub/Sub.

Option D is wrong as Dataflow cannot be configured as Push endpoint with Cloud Pub/Sub.

Question 28: Correct

Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three)

A. Disable writes to certain tables.

B. Restrict access to tables by role.

C. Ensure that the data is encrypted at all times.

D. Restrict BigQuery API access to approved users. **(Correct)**

E. Segregate data across multiple tables or datasets. **(Correct)**

F. Use Google Stackdriver Audit Logging to determine policy violations. **(Correct)**

Explanation

Correct answers are **D, E & F**

Option D would help limit access to approved users only.

Option E as it would help segregate the data with the ability to provide access to users as per their needs.

Option F as it would help in auditing.

Refer GCP documentation - BigQuery Dataset Access Control & Access Control

You share access to BigQuery tables and views using project- level IAM roles and dataset-level access controls. Currently, you cannot apply access controls directly to tables or views.

Project-level access controls determine the users, groups, and service accounts allowed to access all datasets, tables, views, and table data within a project. Dataset-level access controls determine the users, groups, and service accounts allowed to access the tables, views, and table data in a specific dataset.

Option A is wrong as disabiling writes does not prevent the users from reading and does not align with the least privilege principle.

Option B is wrong as access cannot be control on tables.

Option C is wrong as data is encrypted by default, however it does not align with the least privilege principle.

Question 29: **Correct**

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

A. Update the current pipeline and use the drain flag. **(Correct)**

B. Update the current pipeline and provide the transform mapping JSON object.

C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.

D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Explanation

Correct answer is **A** as the key requirement is not to lose the data, the Dataflow pipeline can be stopped using the Drain option. Drain options would cause Dataflow to stop any new processing, but would also allow the existing processing to complete

Refer GCP documentation - Dataflow Stopping a Pipeline

Using the **Drain** option to stop your job tells the Cloud Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources. However, the Cloud Dataflow service will preserve any existing resources, such as worker instances, to finish processing and writing any buffered data in your pipeline.

When all pending processing and write operations are complete, the Cloud Dataflow service will clean up the GCP resources associated with your job.

Note: Your pipeline will continue to incur the cost of maintaining any associated GCP resources until all processing and writing has completed.

Use the Drain option to stop your job if you want to prevent data loss as you bring down your pipeline.

Effects of draining a job

When you issue the Drain command, Cloud Dataflow immediately closes any in-process windows and fires all triggers. The system **does not** wait for any outstanding time-based windows to finish. For example, if your pipeline is ten minutes into a two-hour window when you issue the Drain command, Cloud Dataflow won't wait for the remainder of the window to finish. It will close the window immediately with partial results.

Question 30: **Correct**

A client has been developing a pipeline based on PCollections using local programming techniques and is ready to scale up to production. What should they do?

A. They should use the Cloud Dataflow Cloud Runner. **(Correct)**

B. They should upload the pipeline to Cloud Dataproc.

C. They should use the local version of runner.

D. Import the pipeline into BigQuery.

Explanation

Correct answer is **A** as the PCollection indicates it is a Cloud Dataflow pipeline. And the Cloud Runner will enable the pipeline to scale to production levels.

Refer documentation - Dataflow Cloud Runner

The Google Cloud Dataflow Runner uses the Cloud Dataflow managed service. When you run your pipeline with the

Cloud Dataflow service, the runner uploads your executable code and dependencies to a Google Cloud Storage bucket and creates a Cloud Dataflow job, which executes your pipeline on managed resources in Google Cloud Platform.

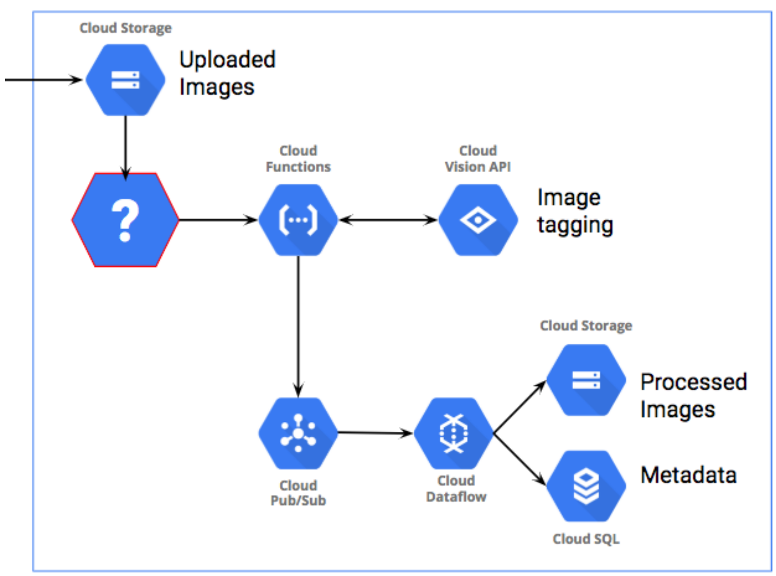
The Cloud Dataflow Runner and service are suitable for large scale, continuous jobs, and provide:

a fully managed service
autoscaling of the number of workers throughout the lifetime of the job
dynamic work rebalancing
Options B & D are wrong as PCollections are related to Dataflow

Option C is wrong as Local runner is execute the pipeline locally.

Question 31: Correct

A company is building an image tagging pipeline. Which service should be used in the icon with the question mark in the diagram?



A. Cloud Datastore

B. Cloud Dataflow

C. Cloud Pub/Sub **(Correct)**

D. Cloud Bigtable

Explanation

Correct answer is **C** as Cloud Storage upload events can push Cloud Pub/Sub to trigger a Cloud Function to ingest and process the image.

Refer GCP documentation - Cloud Storage Pub/Sub Notifications

Cloud Pub/Sub Notifications sends information about changes to objects in your buckets to Cloud Pub/Sub, where the information is added to a Cloud Pub/Sub topic of your choice in the form of messages. For example, you can track objects that are created and deleted in your bucket. Each notification contains information describing both the event that triggered it and the object that changed.

Cloud Pub/Sub Notifications are the recommended way to track changes to objects in your Cloud Storage buckets because they're faster, more flexible, easier to set up, and more cost-effective.

Options A, B & D are wrong as they cannot be configured for notifications from Cloud Storage.

Question 32: Correct

Your company is in a highly regulated industry. One of your requirements is to ensure external users have access only to the non PII fields information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which access control method would you use?

A. Use Primitive role on the dataset

B. Use Predefined role on the dataset

C. Use Authorized view with the same dataset with proper permissions

D. Use Authorized view with the different dataset with proper permissions **(Correct)**

Explanation

Correct answer is **D** as the controlled access can be granted using Authorized view. The Authorized view needs to be in a different dataset than the source.

Refer GCP documentation - BigQuery Authorized Views

Giving a view access to a dataset is also known as creating an authorized view in BigQuery. An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query.

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, your users would have access to both the view and the data.

Options A, B & C are wrong as they would provide access to the complete datasets with the source included.

Question 33: **Correct**

Your company is developing a next generation pet collar that collects biometric information to assist potential millions of families with promoting healthy lifestyles for their pets. Each collar will push 30kb of biometric data In JSON format every 2 seconds to a collection platform that will process and analyze the data providing health trending information back to the pet owners and veterinarians via a web portal. Management has tasked you to architect the collection platform ensuring the following requirements are met.

1. Provide the ability for real-time analytics of the inbound biometric data

2. Ensure processing of the biometric data is highly durable, elastic and parallel

3. The results of the analytic processing should be persisted for data mining

Which architecture outlined below win meet the initial requirements for the platform?

A. Utilize Cloud Storage to collect the inbound sensor data, analyze data with Dataproc and save the results to BigQuery.

B. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to BigQuery. **(Correct)**

C. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Cloud SQL.

D. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Bigtable.

Explanation

Correct answer is **B** as Cloud Pub/Sub provides elastic and scalable ingestion, Dataflow provides processing and BigQuery analytics.

Refer GCP documentation - IoT

Google Cloud Pub/Sub provides a globally durable message ingestion service. By creating topics for streams or channels, you can enable different components of your application to subscribe to specific streams of data without needing to construct subscriber-specific channels on each device. Cloud Pub/Sub also natively connects to other Cloud Platform services, helping you to connect ingestion, data pipelines, and storage systems.

Google Cloud Dataflow provides the open Apache Beam programming model as a managed service for processing data in multiple ways, including batch operations, extract-transform-load (ETL) patterns, and continuous, streaming computation. Cloud Dataflow can be particularly useful for managing the high-volume data processing pipelines required for IoT scenarios. Cloud Dataflow is also designed to integrate seamlessly with the other Cloud Platform services you choose for your pipeline.

Google BigQuery provides a fully managed data warehouse with a familiar SQL interface, so you can store your IoT data alongside any of your other enterprise analytics and logs. The performance and cost of BigQuery means you might keep your valuable data longer, instead of deleting it just to save disk space.

Option A is wrong as Cloud Storage is not an ideal ingestion service for real time high frequency data. Also Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.

Option C is wrong as Cloud SQL is a relational database and not suited for analytics data storage.

Option D is wrong as Bigtable is not ideal for long term analytics data storage.

Question 34: Correct

Which of the following statements about the Wide & Deep Learning model are true? (Choose two)

A. Wide model is used for memorization, while the deep model is used for generalization. **(Correct)**

B. Wide model is used for generalization, while the deep model is used for memorization.

C. A good use for the wide and deep model is a recommender system. **(Correct)**

D. A good use for the wide and deep model is a small-scale linear regression problem.

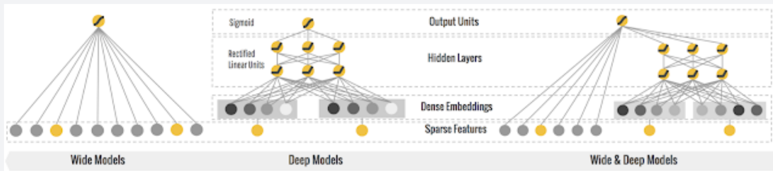
Explanation

Correct answers are **A & C** as Wide learning model is good for memorization and a Deep learning model is generalization. Both Wide and Deep learning model can help build good recommendation engine.

Refer Google blog - Wide Deep learning together

The human brain is a sophisticated learning machine, forming rules by memorizing everyday events ("sparrows can fly" and "pigeons can fly") and generalizing those learnings to apply to things we haven't seen before ("animals with wings can fly"). Perhaps more powerfully, memorization also allows us to further refine our generalized rules with exceptions ("penguins can't fly"). As we were exploring how to advance machine intelligence, we asked ourselves the question—can we teach computers to learn like humans do, by combining the power of memorization and generalization?

It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.



Question 35: Correct

A financial organization wishes to develop a global application to store transactions happening from different part of the world. The storage system must provide low latency transaction support and horizontal scaling. Which GCP service is appropriate for this use case?

A. Bigtable

B. Datastore

C. Cloud Storage

D. Cloud Spanner **(Correct)**

Explanation

Correct answer is **D** as Spanner provides Global scale, low latency and the ability to scale horizontally.

Refer GCP documentation - Storage Options

Cloud Spanner	Mission-critical, relational database service with transactional consistency,	Mission-critical applications High transactions Scale + consistency	Adtech Financial services Global supply chain Retail

	global scale, and high availability.	requirements	

Question 36: **Correct**

A retailer has 1PB of historical purchase dataset, which is largely unlabeled. They want to categorize the customer into different groups as per their spend. Which type of Machine Learning algorithm is suited to achieve this?

A. Classification

B. Regression

C. Association

D. Clustering **(Correct)**

Explanation

Correct answer is **D** as the data is unlabelled, unsupervised learning technique of Clustering can be applied to categorize the data.

Refer GCP documentation - Machine Learning

In unsupervised learning, the goal is to identify meaningful patterns in the data. To accomplish this, the machine must learn from an unlabeled data set. In other words, the model has no hints how to categorize each piece of data and must infer its own rules for doing so.

Options A & B are wrong as they are supervised learning techniques.

In **supervised machine learning**, you feed the features and their corresponding labels into an algorithm in a process called **training**. During training, the algorithm gradually determines the relationship between features and their corresponding labels. This relationship is called the **model**. Often times in machine learning, the model is very complex.

Option C is wrong as Association rules is mainly to identify relationship.

！！！！！！！ Question 37: **Correct**

Your company wants to host confidential documents in Cloud Storage. Due to compliance requirements, there is a need for the data to be highly available and resilient有弹力的 even in case of a regional outage. Which storage classes help meet the requirement? (Select THREE)

A. Nearline **(Correct)**

B. Standard **(Correct)**

C. Multi-Regional **(Correct)**

D. Dual-Regional

E. Regional

Explanation

Correct answers are **A, B & C** as Standard, Multi-Regional and Nearline storage classes provide multi-region geo-redundant deployment, which can sustain regional failure.

Update - There have been several changes in GCP storage classes. Standard Storage was newly introduced by Google Cloud with multi-regional capability. GCP supports now Standard, Nearline and Coldline storage classes. Multi-regional is only available, if you are already using it.

Circa Aug 14, 2019

Multi-Regional Storage and Regional Storage are now Standard Storage.

Combining these into a single Standard Storage class separates your storage class considerations from your location considerations.

Before that **Circa Oct 16, 2016** - Standard Storage class was changed.

Standard Storage class is now Multi-Regional Storage and Regional Storage.

The Multi-Regional Storage class provides the same price and performance along with geo-redundant copies of your data and a 99.95% availability SLA.

The Regional Storage class provides the same performance at a reduced price.

Refer GCP documentation - Cloud Storage Classes

Multi-Regional Storage is geo-redundant.

The geo-redundancy of Nearline Storage data is determined by the type of location in which it is stored: Nearline Storage data stored in multi-regional locations is redundant across multiple regions, providing higher availability than Nearline Storage data stored in regional locations.

Data that is geo-redundant is stored redundantly in at least two separate geographic places separated by at least 100 miles. Objects stored in multi-regional locations are geo-redundant, regardless of their storage class.

Geo-redundancy occurs asynchronously, but all Cloud Storage data is redundant within at least one geographic place as soon as you upload it.

Geo-redundancy ensures maximum availability of your data, even in the event of large-scale disruptions, such as natural disasters. For a dual-regional location, geo-redundancy is achieved using two specific regional locations. For other multi-regional locations, geo-redundancy is achieved using any combination of data centers within the specified multi-region, which may include data centers that are not explicitly available as regional locations.

Option D is wrong as dual-regional storage class does not exist.

Option E is wrong as Regional storage class is not geo-redundant. Data stored in a narrow geographic region and Redundancy is across availability zones

Question 38: **Incorrect**

Your company wants to develop an REST based application for image analysis. This application would help detect individual objects and faces within images, and reads printed words contained within images. You need to do a quick Proof of Concept (PoC) to implement and demo the same. How would you design your application?

A. Create and Train a model using Tensorflow and Develop an REST based wrapper over it

B. Use Cloud Image Intelligence API and Develop an REST based wrapper over it **(Incorrect)**

C. Use Cloud Natural Language API and Develop an REST based wrapper over it

D. Use Cloud Vision API and Develop an REST based wrapper over it **(Correct)**

Explanation

Correct answer is **D** as Cloud Vision API provide pre-built models to identify and detect objects and faces within images.

Refer GCP documentation - AI Products

Cloud Vision API enables you to derive insight from your images with our powerful pretrained API models or easily train custom vision models with AutoML Vision Beta. The API quickly classifies images into thousands of categories (such as “sailboat” or “Eiffel Tower”), detects individual objects and faces within images, and finds and reads printed words contained within images. AutoML Vision lets you build and train custom ML models with minimal ML expertise to meet domain-specific business needs.

没有image intelligence

Question 39: **Correct**

Your company is developing an online video hosting platform. Users can upload their videos, which would be available for all the other users to view and share. As a compliance requirement, the videos need to undergo content moderation before it is available for all the users. How would you design your application?

A. Use Cloud Vision API to identify video with inappropriate content and mark it for manual checks.

B. Use Cloud Natural Language API to identify video with inappropriate content and mark it for manual checks.

C. Use Cloud Speech-to-Text API to identify video with inappropriate content and mark it for manual checks.

D. Use Cloud Video Intelligence API to identify video with inappropriate content and mark it for manual checks. **(Correct)**

Explanation

Correct answer is **D** as Cloud Video Intelligence can be used to perform content moderation.

Refer GCP documentation - Cloud Video Intelligence

Google Cloud Video Intelligence makes videos searchable, and discoverable, by extracting metadata with an easy to use REST API. You can now search every moment of every video file in your catalog. It quickly annotates videos stored in Google Cloud Storage, and helps you identify key entities (nouns) within your video; and when they occur within the video. Separate signals from noise, by retrieving relevant information within the entire video, shot-by-shot, -or per frame.

Identify when inappropriate content is being shown in a given video. You can instantly conduct content moderation across petabytes of data and more quickly and efficiently filter your content or user-generated content.

Option A is wrong as Vision is for image analysis.

Option B is wrong as Natural Language is for text analysis

Option C is wrong as Speech-to-Text is for audio to text conversion.

Question 40: Correct

Your company has a variety of data processing jobs. Dataflow jobs to process real time streaming data using Pub/Sub. Data pipelines working with on-premises data. Dataproc spark batch jobs running weekly analytics with Cloud Storage. They want a single interface to manage and monitor the jobs. Which service would help implement a common monitoring and execution platform?

A. Cloud Scheduler

B. Cloud Composer **(Correct)**

C. Cloud Spanner

D. Cloud Pipeline

Explanation

Correct answer is **B** as Cloud Composer's managed nature allows you to focus on authoring, scheduling, and monitoring your workflows as opposed to provisioning resources.

Refer GCP documentation - [Cloud Composer]

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers. Built on the popular Apache Airflow open source project and operated using the Python programming language, Cloud Composer is free from lock-in and easy to use.

Cloud Composer's managed nature allows you to focus on authoring, scheduling, and monitoring your workflows as opposed to provisioning resources.

Option A is wrong as Cloud Scheduler is a fully managed enterprise-grade cron job scheduler. It is not an multi-cloud orchestration tool.

Option C is wrong as Google Cloud Spanner is relational database

Option D is wrong as Google Cloud Pipeline service does not exist.

Question 41: Correct

Your company hosts its analytical data in a BigQuery dataset for analytics. They need to provide controlled access to certain tables and columns within the tables to a third party. How do you design the access with least privilege?

A. Grant only DATA VIEWER access to the third party team

B. Grant fine grained DATA VIEWER access to the tables and columns within the dataset

C. Create Authorized views for tables in a same project and grant access to the teams

D. Create Authorized views for tables in a separate project and grant access to the teams **(Correct)**

Explanation

Correct answer is **D** as the controlled access can be provided using Authorized views created in a separate project.

Refer GCP documentation - BigQuery Authorized View

BigQuery is a petabyte-scale analytics data warehouse that you can use to run SQL queries over vast amounts of data in near realtime.

Giving a view access to a dataset is also known as creating an authorized view in BigQuery. An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query.

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, your data analysts would have access to both the view and the data.

Options A & B are wrong as access cannot be controlled over table, but only projects and datasets.

Option C is wrong as Authorized views should be created in a separate project. If they are created in the same project, the users would have access to the underlying tables as well.

Question 42: **Correct**

Your company is hosting its analytics data in BigQuery. All the Data analysts have been provided with the IAM owner role to their respective projects. As a compliance requirement, all the data access logs needs to be captured for audits. Also, the access to the logs needs to be limited

to the Auditor team only. How can the access be controlled?

A. Export the data access logs using aggregated sink to Cloud Storage in an existing project and grant VIEWER access to the project to the Auditor team

B. Export the data access logs using project sink to BigQuery in an existing project and grant VIEWER access to the project to the Auditor team

C. Export the data access logs using project sink to Cloud Storage in a separate project and grant VIEWER access to the project to the Auditor team

D. Export the data access logs using aggregated sink to Cloud Storage in a separate project and grant VIEWER access to the project to the Auditor team **(Correct)**

Explanation

Correct answer is **D** as the Data Analysts have OWNER roles to the projects, the logs need to be exported to a separate project which only the Auditor team has access to. Also, as there are multiple projects aggregated export sink can be used to export data access logs from all projects.

Refer GCP documentation - BigQuery Auditing and Aggregated Exports

You can create an aggregated export sink that can export log entries from all the projects, folders, and billing accounts of an organization. As an example, you might use this feature to export audit log entries from an organization's projects to a central location.

Options A & B are wrong as the export needs to be in separate project.

Option C is wrong as you need to use aggregated sink instead of project sink, as it would capture logs from all projects.

Question 43: **Correct**

Your company is building an aggregator, which receives feed from lot of other external data sources and companies. These dataset contain invalid & erroneous records, which need to be discarded. Your Data analysts should be able to perform the same without any programming or SQL knowledge. Which solution best fits the requirement?

A. Dataflow

B. Dataproc

C. Hadoop installation on Compute Engine

D. Dataprep **(Correct)**

Explanation

Correct answer is **D** as Dataprep provides the ability to detect, clean and transform data through a Graphical Interface without any programming knowledge.

Refer GCP documentation - Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.

Cloud Dataprep automatically detects schemas, datatypes, possible joins, and anomalies such as missing values, outliers, and duplicates so you get to skip the time-consuming work of profiling your data and go right to the data analysis.

Cloud Dataprep automatically identifies data anomalies and helps you to take corrective action fast. Get data transformation suggestions based on your usage pattern. Standardize, structure, and join datasets easily with a guided approach.

Options A, B & C are wrong as they all need programming knowledge.

Question 44: Correct

Your company is migrating to the Google cloud and looking for HBase alternative. Current solution uses a lot of custom code using the observer coprocessor. You are required to find the best alternative for migration while using managed services, is possible?

A. Dataflow

B. HBase on Dataproc **(Correct)**

C. Bigtable

D. BigQuery

Explanation

Correct answer is **B** as Bigtable is an HBase managed service alternative on Google Cloud. However, it does not support Coprocessors. So the best solution is to use HBase with Dataproc which can be installed using initialization actions.

Refer GCP documentation - Bigtable HBase differences

Coprocessors are not supported. You cannot create classes that implement the interface`org.apache.hadoop.hbase.coprocessor`.

Options A & D are wrong as Dataflow and BigQuery are not HBase alternative

Option C is wrong as Bigtable does not support Coprocessors.

Question 45: **Correct**

You have multiple Data Analysts who work with the dataset hosted in BigQuery within the same project. As a BigQuery Administrator, you are required to grant the data analyst only the privilege to create jobs/queries and an ability to cancel self-submitted jobs. Which role should assign to the user?

A. User

B. Jobuser **(Correct)**

C. Owner

D. Viewer

Explanation

Correct answer is **B** as JobUser access grants users permissions to run jobs and cancel their own jobs within the same project

Refer GCP documentation - BigQuery Access Control

`roles/bigquery.jobUser`	Permissions to run jobs, including queries, within the project. The jobUser role can get information about their own jobs and cancel their own jobs. Rationale: This role allows the separation of data access from the ability to run work in the project, which is useful when team members query data from multiple projects. This role does not allow access to any BigQuery data. If data access is required, grant dataset-level access controls. Resource Types: Organization Project

Option A is wrong as User would allow to run queries across projects.

Option C is wrong as Owner would give more privileges to the users

Option D is wrong as Viewer does not give user permissions to run jobs.

Question 46: **Correct**

You need to design a real time streaming data processing pipeline. The pipeline needs to read data from Cloud Pub/Sub, enrich it using Static reference data in BigQuery, transform it and store the results back in BigQuery for further analytics. How would you design the pipeline?

A. Dataflow, BigQueryIO and PubSubIO, SideOutputs

B. Dataflow, BigQueryIO and PubSubIO, SideInputs **(Correct)**

C. DataProc, BigQueryIO and PubSubIO, SideInputs

D. DataProc, BigQueryIO and PubSubIO, SideOutputs

Explanation

Correct answer is **B** as Dataflow is needed for real time streaming pipeline with the ability to enrich and transform using SideInputs. BigQueryIO and PubSubIO to interact with BigQuery and Pub/Sub.

Refer GCP documentation - Dataflow Use Case Patterns

In streaming mode, lookup tables need to be accessible by your pipeline. If the lookup table never changes, then the standard Cloud Dataflow `SideInput` pattern reading from a bounded source such as BigQuery is a perfect fit. However, if the lookup data changes over time, in streaming mode there are additional considerations and options. The pattern described here focuses on slowly-changing data — for example, a table that's updated daily rather than every few hours.

Options C & D are wrong as Dataproc is not ideal for handling real time streaming data.

Options A & D are wrong as the lookup tables can be referred using SideInputs.

Question 47: **Correct**

You are interacting with a Point Of Sale (PoS) terminal, which sends the transaction details only. Due to latest software update a bug was introduced in the terminal software that caused it to send individual PII and card details. As a security measure, you are required to implement a quick solution to prevent access to the PII. How would you design the solution?

A. Train Model using Tensorflow to identify PII and filter the information

B. Store the data in BigQuery and create a Authorized view for the users

C. Use Data Loss Prevention APIs to identify the PII information and filter the information **(Correct)**

D. Use Cloud Natural Language API to identify PII and filter the information

Explanation

Correct answer is **C** as Data Loss Prevention APIs can be used to quickly redact the sensitive information.

Refer GCP documentation - [Cloud DLP]

Cloud DLP helps you better understand and manage sensitive data. It provides fast, scalable classification and redaction for sensitive data elements like credit card numbers, names, social security numbers, US and selected international identifier numbers, phone numbers and GCP credentials. Cloud DLP classifies this data using more than 90 predefined detectors to identify patterns, formats, and checksums, and even understands contextual clues. You can optionally redact data as well using techniques like masking, secure hashing, bucketing, and format-preserving encryption.

Option A is wrong as building and training a model is not a quick and easy solution.

Option B is wrong as the data would still be stored in the base tables and accessible.

Option D is wrong as Cloud Natural APIs is for text analysis and does not handle sensitive information redaction.

Question 48: Correct

You are designing a relational data repository on Google Cloud to grow as needed. The data will be transactionally consistent and added from any location in the world. You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?

A. Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.

B. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span. **(Correct)**

C. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.

D. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

Explanation

Correct answer is **B** as the requirement is to support relational data service with transactionally consistently and globally scalable transactions, Cloud Spanner is an ideal choice. CPU utilization is the recommended metric for scaling, per Google best practices, linked below.

Refer GCP documentation -

Storage Options @ https://cloud.google.com/storage-options/ & Spanner Monitoring @ https://cloud.google.com/spanner/docs/monitoring

Option A is wrong as storage utilization is not a correct scaling metric for load.

Options C & D are wrong Bigtable is regional and not a relational data service.

Question 49: **Correct**

You are working on a project with two compliance requirements. The first requirement states that your developers should be able to see the Google Cloud Platform billing charges for only their own projects. The second requirement states that your finance team members can set budgets and view the current charges for all projects in the organization. The finance team should not be able to view the project contents. You want to set permissions. What should you do?

A. Add the finance team members to the default IAM Owner role. Add the developers to a custom role that allows them to see their own spend only.

B. Add the finance team members to the Billing Administrator role for each of the billing accounts that they need to manage. Add the developers to the Viewer role for the Project. **(Correct)**

C. Add the developers and finance managers to the Viewer role for the Project.

D. Add the finance team to the Viewer role for the Project. Add the developers to the Security Reviewer role for each of the billing accounts.

Explanation

Correct answer is **B** as there are 2 requirements, Finance team able to set budgets on project but not view project contents and developers able to only view billing charges of their projects. Finance with Billing Administrator role can set budgets and Developer with viewer role can view billing charges aligning with the principle of least privileges.

Refer GCP documentation - IAM Billing @ https://cloud.google.com/iam/docs/job-functions/billing

Option A is wrong as GCP recommends using pre-defined roles instead of using primitive roles and custom roles.

Option C is wrong as viewer role to finance would not provide them the ability to set budgets.

Option D is wrong as viewer role to finance would not provide them the ability to set budgets. Also, Security Reviewer role enables the ability to view custom roles but not administer them for the developers which they don't need.

!!!!!!!!!!Question 50: **Incorrect**

Your customer wants to capture multiple GBs of aggregate real-time key performance indicators (KPIs) from their game servers running on Google Cloud Platform and monitor the KPIs with low latency. How should they capture the KPIs?

A. Output custom metrics to Stackdriver from the game servers, and create a Dashboard in Stackdriver Monitoring Console to view them. **(Incorrect)**

B. Schedule BigQuery load jobs to ingest analytics files uploaded to Cloud Storage every ten minutes, and visualize the results in Google Data Studio.

C. Store time-series data from the game servers in Google Bigtable, and view it using Google Data Studio. **(Correct)**

D. Insert the KPIs into Cloud Datastore entities, and run ad hoc analysis and visualizations of them in Cloud Datalab.

Explanation

Correct answer is **C** as Bigtable is an ideal solution for storing time series data with the ability to provide analytics at real time at a very low latency. Data can be viewed using Google Data Studio.

Refer GCP documentation - Data lifecycle @ https://cloud.google.com/solutions/data-lifecycle-cloud-platform

Cloud Bigtable is a managed, high-performance NoSQL database service designed for terabyte- to petabyte-scale workloads. Cloud Bigtable is built on Google's internal Cloud Bigtable database infrastructure that powers Google Search, Google Analytics, Google Maps, and Gmail. The service provides consistent, low-latency, and high-throughput storage for large-scale NoSQL data. Cloud Bigtable is built for real-time app serving workloads, as well as large-scale analytical workloads.

Cloud Bigtable schemas use a single-indexed row key associated with a series of columns; schemas are usually structured either as tall or wide and queries are based on row key. The style of schema is dependent on the downstream use cases and it's important to consider data locality and distribution of reads and writes to maximize performance. Tall schemas are often used for storing time-series events, data that is keyed in some portion by a timestamp, with relatively fewer columns per row. Wide schemas follow the opposite approach, a simplistic identifier as the row key along with a large number of columns

Option A is wrong as Stackdriver is not an ideal solution for time series data and it does not provide analytics capability.

Option B is wrong as BigQuery does not provide low latency access and with jobs scheduled at every 10 minutes does not meet the real time criteria.

Option D is wrong as Datastore does not provide analytics capability.

Google Cloud Certified - Professional Data Engineer Practice Exam 2 - Results

Attempt 2

Question 1: **Correct**

Your infrastructure includes two 100-TB enterprise file servers. You need to perform a one-way, one-time migration of this data to the Google Cloud securely.

Only users in Germany will access this data. You want to create the most cost-effective solution. What should you do?

A. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination. **(Correct)**

B. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Multi-Regional bucket as a final destination.

C. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

D. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.

Explanation

Correct answer is **A** as the data is huge it can be transferred using Transfer Appliance in a time and cost effective way. Also, as the data is going to be accessed in a single region it can be hosted in a regional bucket.

Refer GCP documentation - Storage Classes

Multi-Regional Storage	>99.99% typical monthly availability	Storing data that is frequently accessed	$0.026

	99.95% availability SLA* Geo-redundant	("hot" objects) around the world, such as serving website content, streaming videos, or gaming and mobile applications. For Multi-Regional Storage data stored in dual-regional locations, you also get optimized performance when accessing Google Cloud Platform products that are located in one of the associated regions.	
Regional Storage	99.99% typical monthly availability 99.9% availability SLA* Lower cost per GB stored Data stored in a narrow	Storing frequently accessed data in the same region as your Google Cloud DataProc or Google Compute Engine instances that use it,	$0.020

	geographic region Redundant across availability zones	such as for data analytics.	

Option B is wrong as the data is accessed in a single region, it would be more cost effective storing it in a regional bucket.

Options C & D are wrong as the data is huge it is more time and cost effective to transfer the data Transfer Appliance.

Question 2: Correct

You are designing storage for event data as part of building a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying individual values over time windows. Which storage service and schema design should you use?

A. Use Cloud Bigtable for storage. Design tall and narrow tables, and use a new row for each single event version. **(Correct)**

B. Use Cloud Bigtable for storage. Design short and wide tables, and use a new column for each single event version.

C. Use Cloud Storage for storage. Join the raw file data with a BigQuery log table.

D. Use Cloud Storage for storage. Write a Cloud Dataprep job to split the data into partitioned tables.

Explanation

Correct answer is **A** as its an event data (time series) and need to be restricted to individual values over time windows, it is best to use Bigtable with tall and narrow tables.

Refer GCP documentation
- Bigtable Time series schema

For time series, you should generally use tall and narrow tables. *This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum.*

As an optimization, you can use short and wide tables, but avoid unbounded numbers of events. *For example, if you usually need to retrieve an entire month of events at once, the temperature table above is a reasonable optimization—the row is bounded in size to the number of days in a month.*

Option B is wrong as short and wide tables and are ideal for storing time series data.

Options C & D are wrong as you do not need to use GCS/BQ for this scenario.

Question 3: Correct

You are building a data pipeline on Google Cloud. You need to prepare source data for a machine-learning model. This involves quickly deduplicating rows from three input tables and also removing outliers from data columns where you do not know the data distribution. What should you do?

A. Write an Apache Spark job with a series of steps for Cloud Dataflow. The first step will examine the source data, and the second and third steps step will perform data transformations.

B. Write an Apache Spark job with a series of steps for Cloud Dataproc. The first step will examine the source data, and the second and third steps step will perform data transformations.

C. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Write a recipe to transform the data and add it to the Cloud Dataprep job.

D. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Click on each column name, click on each appropriate suggested transformation, and then click 'Add' to add each **(Correct)** transformation to the Cloud Dataprep job.

Explanation

Correct answer is **D** as the requirements is to prepare/clean source data, use Cloud Dataprep suggested transformations to quickly build a transformation job.

Refer GCP documentation - Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.

Cloud Dataprep automatically identifies data anomalies and helps you to take corrective action fast. Get data transformation suggestions based on your usage pattern. Standardize, structure, and join datasets easily with a guided approach.

Option C is wrong as you can simply use the suggested transformations instead of writing custom recipe in Cloud Dataprep

Options A & B are wrong as you should not use Apache Spark and Cloud Dataflow or Cloud Dataproc for this scenario.

Question 4: Correct

You are setting up Cloud Dataproc to perform some data transformations using Apache Spark jobs. The data will be used for a new set of non-critical experiments in your marketing group. You want to set up a cluster that can transform a large amount of data in the most cost-effective way. What should you do?

A. Set up a cluster in High Availability mode with high-memory machine types. Add 10 additional local SSDs.

B. Set up a cluster in High Availability mode with default machine types. Add 10 additional Preemptible worker nodes.

C. Set up a cluster in Standard mode with high-memory machine types. Add 10 additional Preemptible worker nodes. **(Correct)**

D. Set up a cluster in Standard mode with the default machine types. Add 10 additional local SSDs.

Explanation

Correct answer is **C** as Dataproc is a managed service which handles Spark and Hadoop jobs and Spark and **high-memory machines only need the Standard mode**. Also, using Preemptible nodes provides cost-efficiency as this is not mission-critical.

Refer GCP documentation
- Dataproc pricing

Note: Preemptible instances can be used to lower your Compute Engine costs for Cloud Dataproc clusters, but do not change the way you are billed for the Cloud Dataproc premium.

Options A & B are wrong as this scenario does not call for High Availability mode because it handles non-critical experiments.

Option D is wrong as local SSDs would cost more; instead, use Preemptible nodes to meet your objective of delivering a cost-effective solution.

Question 5: **Correct**

You want to display aggregate view counts for your YouTube channel data in Data Studio. You want to see the video tiles and view counts summarized over the last 30 days. You also want to segment the data by the Country Code using the fewest possible steps. What should you do?

A. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter.

B. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title **(Correct)**

and Country Code
as report
dimensions.

C. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter.

D. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title and Country Code as report dimensions.

Explanation

Correct answer is **B** as there is no need to export; you can use the existing YouTube data source. Country Code is a dimension because it's a string and should be displayed as such, that is, showing all countries, instead of filtering.

Refer GCP documentation - Data Studio Youtube connector

Option A is wrong as you cannot produce a summarized report that meets your business requirements using the options listed.

Options C & D are wrong as you do not need to export data from YouTube to Cloud Storage; you can simply use the existing YouTube data source.

Youtube + datastudio可以直连

Question 6: **Correct**

Your company wants to try out the cloud with low risk. They want to archive approximately 100 TB of their log data to the cloud and test the analytics features available to them there, while also retaining that data as a long-term disaster recovery backup. Which two steps should they take? (Choose two answers)

A. Load logs into Google BigQuery. **(Correct)**

B. Load logs into Google Cloud SQL.

C. Import logs into Google Stackdriver.

D. Insert logs into Google Cloud Bigtable.

E. Upload log files into Google Cloud Storage. **(Correct)**

Explanation

Correct answers are **A & E** as Google Cloud Storage can provide long term archival option and BigQuery provides analytics capabilities.

Option B is wrong as Cloud SQL is relational database and does not support the capacity required as well as not suitable for long term archival storage.

Option C is wrong as Stackdriver is a monitoring, logging, alerting and debugging tool. It is not ideal for long term retention of data and does not provide analytics capabilities.

Option D is wrong as Bigtable is a NoSQL solution and can be used for analytics. However it is ideal for data with low latency access and is expensive.

bigtable很贵，没说low latency尽量不用

Question 7: **Correct**

A company wants to transfer petabyte scale of data to Google Cloud for their analytics, however are constrained on their internet connectivity? Which GCP service can help them transfer the data quickly?

A. Transfer appliance and Dataprep to decrypt the data

B. Google Transfer service using multiple VPN connections

C. gustil with multiple VPN connections

D. Transfer appliance and rehydrator to decrypt the data **(Correct)**

Explanation

Correct answer is **D** as the data is huge it should be transferred using Transfer Appliance and use a Rehydrator to decrypt the data.

Refer GCP documentation - Data Rehydration

Once you capture your data onto the Google Transfer Appliance, ship the appliance to the Google upload facility for rehydration. Data rehydration is the process by which you fully reconstitute the files so you can access and use the transferred data.

To rehydrate data, the data is first copied from the Transfer Appliance to your Cloud Storage staging bucket. The data uploaded to your staging bucket is still compressed, deduplicated and encrypted. Data rehydration reverses this process and restores your data to a usable state. As the data is rehydrated, it is moved to the Cloud Storage destination bucket that you created.

To perform data rehydration, use a Rehydrator instance, which is a virtual appliance that runs as a Compute Engine instance on Google Cloud Platform.

The Transfer Appliance Rehydrator compares the CRC32C hash value of each file being rehydrated with the hash value computed when the file was captured. If the checksums don't match, the file is skipped and appears in the skip file list with the message "Data corruption detected".

Option A is wrong as Dataprep does not help is decrypting the data.

Option B is wrong as Google Transfer Service does not support importing data from on-premises data center. It only supports online imports.

Option C is wrong as the data is huge transferring it with gsutil would take a long time.

Question 8: Correct

A company has lot of data sources from multiple systems used for reporting. Over a period of time, a lot data is missing and you are asked to perform anomaly detection. How would you design the system?

A. Use Dataprep with Data Studio

B. Load in Cloud Storage and use Dataflow with Data Studio

C. Load in Cloud Storage and use Dataprep with Data Studio **(Correct)**

D. Use Dataflow with Data Studio

Explanation

Correct answer is **C** as Dataprep provides data cleaning and automatically identifies anomalies in the data. It can integrated with Cloud Storage and BigQuery

Refer GCP documentation - Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.

Cloud Dataprep automatically detects schemas, datatypes, possible joins, and anomalies such as missing values, outliers, and duplicates so you get to skip the time-consuming work of profiling your data and go right to the data analysis.

Cloud Dataprep automatically identifies data anomalies and helps you to take corrective action fast. Get data transformation suggestions based on your usage pattern. Standardize, structure, and join datasets easily with a guided approach.

Easily process data stored in Cloud Storage, BigQuery, or from your desktop. Export clean data directly into BigQuery for further analysis. Seamlessly manage user access and data security with Cloud Identity and Access Management.

Option A is wrong as Dataprep would not be able interact directly with local system.

Options B & D are wrong as Cloud Dataflow is a fully-managed service for transforming and enriching data in stream (real time) and batch (historical) modes with equal reliability and expressiveness -- no more complex workarounds or compromises needed. It does not provide anomaly detection.

dateprep可以与datastorage, bigquery直接相连，但不能与Local直接连接

Question 9: **Correct**

Your company plans to migrate a multi-petabyte data set to the cloud. The data set must be available 24hrs a day. Your business analysts have experience only with using a SQL interface. How should you store the data to optimize it for ease of analysis?

A. Load data into Google BigQuery. **(Correct)**

B. Insert data into Google Cloud SQL.

C. Put flat files into Google Cloud Storage.

D. Stream data into Google Cloud Datastore.

Explanation

Correct answer is **A** as BigQuery is the only of these Google products that supports an SQL interface and a high enough SLA (99.9%) to make it readily available.

Option B is wrong as Cloud SQL cannot support multi-petabyte data. Storage limit for Cloud SQL is 10TB

Option C is wrong as Cloud Storage does not provide SQL interface.

Option D is wrong as Datastore does not provide a SQL interface and is a NoSQL solution.

Question 10: Correct

Your company hosts its data into multiple Cloud SQL databases. You need to export your Cloud SQL tables into BigQuery for analysis. How can the data be exported?

A. Convert your Cloud SQL data to JSON format, then import directly into BigQuery

B. Export your Cloud SQL data to Cloud Storage, then import into BigQuery **(Correct)**

C. Import data to BigQuery directly from Cloud SQL.

D. Use the BigQuery export function in Cloud SQL to manage exporting data into BigQuery.

Explanation

Correct answer is **B** as BigQuery does not provide direct load from Cloud SQL. The data needs to be loaded through Cloud Storage.

Refer GCP documentation - BigQuery loading data

There are many situations where you can query data without loading it. For all other situations, you must first load your data into BigQuery before you can run queries.

You can load data:

From Cloud Storage
From other Google services, such as Google Ad Manager and Google Ads
From a readable data source (such as your local machine)

By inserting individual records using streaming inserts
Using DML statements to perform bulk inserts
Using a Google BigQuery IO transform in a Cloud Dataflow pipeline to write data to BigQuery
Options A, C & D are wrong as they are not supported options.

BQ可以支持多种导入，cloud storage, bigtable, local file,bigquery transfer等。不支持cloud sql直接导入

BQ只支持导出到cloud storage.

Question 11: Correct

Your BigQuery table needs to be accessed by team members who are not proficient in technology. You want to simplify the columns they need to query to avoid confusion. How can you do this while preserving all of the data in your table?

A. Train your team members on how to query larger tables.

B. Create a query that uses the reduced number of columns they will access. Save this query as a view in a different dataset. Give your team members access to the new dataset and instruct them to query against the saved view instead of the main table. **(Correct)**

C. Apply column filtering to your table, and restrict the unfiltered view to yourself and those who need access to the full table.

D. Create a copy of your table in a different dataset, and remove the unneeded columns from the copy. Have your team members run queries against this copy.

Explanation

Correct answer is **B** as the best way to limit and expose number of columns and access is to create a View. With BigQuery, the access can only be controlled on Datasets and Views, but not on tables.

Refer GCP documentation - BigQuery Views

Option A is wrong as it is not a feasible solution.

Option C is wrong as column filtering cannot be applied to Table and it can be done through Views.

Option D is wrong as it is not an ideal solution, as it results in duplication of data.
Also, deletion of Columns is not supported.

Question 12: Correct

Your company is using WILDCARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

```
# Syntax error : Expected end of statement but got "-" at [4:11]
SELECT age
FROM
    bigquery-public-data.noaa_gsod.gsod
WHERE
    age != 99
    AND_TABLE_SUFFIX = '1929'
ORDER BY
    age DESC
```

Which table name will make the SQL statement work correctly?

A. `bigquery-public-data.noaa_gsod.gsod`

B. bigquery-public-data.noaa_gsod.gsod*

C. `bigquery-public-data.noaa_gsod.gsod`*

D. `bigquery-public-data.noaa_gsod.gsod*` **(Correct)**

Explanation

Correct answer is **D** as the table name should include a * for the wildcard and it must be enclosed in backtick characters.

Refer GCP documentation - [BigQuery Wildcard table reference]

Wildcard tables enable you to query multiple tables using concise SQL statements. Wildcard tables are available only in standard SQL.

The wildcard character, "", represents one more characters of a table name. The wildcard character can appear only as the final character of a wildcard table name.*

The wildcard table name contains the special character (*), which means that you must enclose the wildcard table name in backtick (`) characters.

Question 13: **Incorrect**

You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling. Which Google database service should you use?

A. Cloud SQL **(Incorrect)**

B. BigQuery

C. Cloud Bigtable

D. Cloud Datastore **(Correct)**

Explanation

Correct answer is **D** as the payment transactions would need a transactional data service Datastore can support the same. Also it is fully managed with NoOps required.

Refer GCP documentation - Storage Options

Option A is wrong as **Cloud SQL would need infrastructure scaling. Although storage can be automatically scaled (upto a limit), instance type needs to be changed as per the load manually.**

Option B is wrong as BigQuery is an data warehousing option.

Option C is wrong as Bigtable is not a relational database but an NoSQL option.

Question 14: Correct

You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and analyze these very large datasets in real time. How should you design the system in Google Cloud?

A. Send the data to Google Cloud Datastore and then export to BigQuery.

B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery. **(Correct)**

C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.

D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

Explanation

Correct answer is **B** as the need to ingest it, transform and store the Cloud Pub/Sub, Cloud Dataflow, BigQuery is ideal stack to handle the IoT data.

Refer GCP documentation - IoT

Google Cloud Pub/Sub provides a globally durable message ingestion service. By creating topics for streams or channels, you can enable different components of your application to subscribe to specific streams of data without needing to construct subscriber-specific channels on each device. Cloud Pub/Sub also natively connects to other Cloud Platform services, helping you to connect ingestion, data pipelines, and storage systems.

Google Cloud Dataflow provides the open Apache Beam programming model as a managed service for processing data in multiple ways, including batch operations, extract-transform-load (ETL) patterns, and continuous, streaming computation. Cloud Dataflow can be particularly useful for managing the high-volume data processing pipelines required for IoT scenarios. Cloud Dataflow is also designed to integrate seamlessly with the other Cloud Platform services you choose for your pipeline.

Google BigQuery provides a fully managed data warehouse with a familiar SQL interface, so you can store your IoT data alongside any of your other enterprise analytics and logs. The performance and cost of BigQuery means you might keep your valuable data longer, instead of deleting it just to save disk space.

Sample Arch - [Mobile Gaming Analysis Telemetry]

Option A is wrong as the Datastore is not an ideal ingestion service.

Option C is wrong as Cloud Storage is not an ideal ingestion service and Dataproc is not a data warehousing solution.

Option D is wrong as Cloud SQL is not a data warehousing solution.

Question 15: **Correct**

Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

A. Redefine the schema by evenly distributing reads and writes across the row space of the table. **(Correct)**

B. The performance issue should be resolved over time as the size of the Bigtable cluster is increased.

C. Redesign the schema to use a

single row key to identify values that need to be updated frequently in the cluster.

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

Explanation

Correct answer is **A** as the schema needs to be redesigned to distribute the reads and writes evenly across each table.

Refer GCP documentation - Bigtable Performance

The table's schema is not designed correctly. *To get good performance from Cloud Bigtable, it's essential to design a schema that makes it possible to distribute reads and writes evenly across each table. See Designing Your Schema for more information.*

Option B is wrong as increasing the size of cluster would increase the cost.

Option C is wrong as single row key for frequently updated identifiers reduces performance

Frequently updated identifiers

Avoid using a single row key to identify a value that must be updated very frequently. For example, if you store memory-usage data once per second, do not use a single row key named `memusage` and update the row repeatedly. This type of operation overloads the tablet that stores the frequently used row. It can also cause a row to exceed its size limit, because a *cell's previous values take up space for a while.*

Instead, store one value per row, using a row key that contains the type of metric, a delimiter, and a timestamp.

Option D is wrong as sequential IDs would degrade the performance.

Sequential numeric IDs

Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, because new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes.

*A safer approach is to use a r**eversed version of the user's numeric ID**, which spreads traffic more evenly across all of the nodes for your Cloud Bigtable table.*

Question 16: **Correct**

Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

A. Create a Google Cloud Dataflow job to process the data.

B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.

C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.

D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector. **(Correct)**

E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Explanation

Correct answer is **D**. As the requirement is to reuse Hadoop 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.

Refer GCP documentation - Dataproc FAQs

Cloud Dataproc is a fast, easy-to-use, low-cost and fully managed service that lets you run the Apache Spark and Apache Hadoop ecosystem on Google Cloud Platform. Cloud Dataproc provisions big or small clusters rapidly, supports many popular job types, and is integrated with other Google Cloud Platform services, such as Cloud Storage and Stackdriver Logging, thus helping you reduce TCO.

Cloud Dataproc is a managed Spark/Hadoop service intended to make Spark and Hadoop easy, fast, and powerful. In a traditional

Hadoop deployment, even one that is cloud-based, you must install, configure, administer, and orchestrate work on the cluster. By contrast, Cloud Dataproc handles cluster creation, management, monitoring, and job orchestration for you.

Yes, Cloud Dataproc clusters automatically install the Cloud Storage connector. There are a number of benefits to choosing Cloud Storage over traditional HDFS including data persistence, reliability, and performance.

What happens to my data when a cluster is shut down?

Any data in Cloud Storage persists after your cluster is shut down. This is one of the reasons to choose Cloud Storage over HDFS since HDFS data is removed when a cluster is shut down (unless it is transferred to a persistent location prior to shutdown).

Option A is wrong as Dataflow is not suited to execute Hadoop jobs.

Option B is wrong as HDFS is associated with the Cluster. If the cluster is terminated, the data would be lost.

Option C is wrong as Cluster on Compute Engine would increase infrastructure management and persistent disks would not provide scalability.

Option E is wrong as Cluster on Compute Engine would increase infrastructure management and Local SSDs would not provide data durability.

Question 17: Correct

You have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error:

```
SELECT person FROM
`project1.example.table1`
WHERE city = "London"
```

How would you correct the error?

A. Add ", UNNEST(person)" before the WHERE clause. **(Correct)**

B. Change "person" to "person.city".

C. Change "person" to "city.person".

D. Add ", UNNEST(city)" before the WHERE clause.

Explanation

Correct answer is **A** as the person column needs to be UNNEST for the nested city field to be used directly in the WHERE clause. Also, note this is standard SQL query by the reference of the table.

Refer GCP documentation - BigQuery Nested Query

```
#standardSQL SELECT page.
title FROM `bigquery-publ
ic-data.samples.github_ne
sted`, UNNEST(payload.pag
es) AS page WHERE page.pa
```

```
ge_name IN ('db_jobskil
l', 'Profession');
```

Question 18: Correct

Your company's on-premises Spark jobs have been migrated to Cloud Dataproc. You are exploring the option to use Preemptible workers to increase the performance of the jobs, while cutting on costs. Which of these rules apply when you add preemptible workers to a Dataproc cluster? (Choose two)

A. Preemptible workers cannot use persistent disk.

B. Preemptible workers cannot store data. **(Correct)**

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

D. A Dataproc cluster cannot have only preemptible workers. **(Correct)**

Explanation

Correct answers are **B & D**.

Option B as Preemptible instances are disposable and should not be used to store data.

Option D as a Dataproc cluster cannot be with only preemptible instances. It needs to have **two** non-preemptible worker nodes.

Refer GCP documentation
- [Dataproc Preemptible VMs]

The following rules will apply when you use preemptible workers with a Cloud Dataproc cluster:

Processing only*—Since* ***preemptibles*** *can be reclaimed at any time, preemptible workers do* ***not store data****. Preemptibles added to a Cloud Dataproc cluster only function as processing nodes.*
No preemptible-only clusters*—To ensure clusters do not lose all workers, Cloud Dataproc cannot create preemptible-only clusters. If you use the* `gcloud dataproc clusters create` *command with* `--num-preemptible-workers`*, and you do not also specify a number of standard workers with* `--num-workers`*, Cloud Dataproc will automatically add* ***two*** *non-preemptible workers to the cluster.*
Persistent disk size*—As a default, all preemptible workers are created with the smaller of 100GB or the primary worker boot disk size. This disk space is used for local caching of data and is not available through HDFS. You can override the default disk size with the gcloud dataproc clusters create --preemptible-worker-boot-disk-sizecommand at cluster creation. This flag can be specified even if the cluster does not have any preemptible workers at creation time.*
Option A is wrong as preemptible nodes can have persistent disks.

Option C is wrong as Dataproc handles the addition and removal of preemptible nodes.

Question 19: **Correct**

You have a Dataflow job that you want to cancel. It is a streaming IoT pipeline, and you want to ensure that any data that is in-flight is processed and written to the output with no data loss. Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

A. Cancel

B. Drain **(Correct)**

C. Stop

D. Pause

Explanation

Correct answer is **B** as Drain command helps Dataflow process and complete in-flight messages and stops accepting any new ones.

Refer GCP documentation - Dataflow stopping a pipeline

If you need to stop a running Cloud Dataflow job, you can do so by issuing a command using either the Cloud Dataflow Monitoring Interface or the Cloud Dataflow Command-line Interface. There are two possible commands you can issue to stop your job: ***Cancel*** *and* ***Drain****.*

Note: *The* ***Drain*** *command is supported for streaming pipelines* *only.*

*Using the **Drain** option to stop your job tells the Cloud Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources. However, the Cloud Dataflow service will preserve any existing resources, such as worker instances, to finish processing and writing any buffered data in your pipeline. When all pending processing and write operations are complete, the Cloud Dataflow service will clean up the GCP resources associated with your job.*

***Note:** Your pipeline will continue to incur the cost of maintaining any associated GCP resources until all processing and writing has completed.*

Use the Drain option to stop your job if you want to prevent data loss as you bring down your pipeline.

Option A is wrong as Cancel does not handle in-flight messages and it might result in data loss.

Options C & D are wrong as Stop and Pause option do not exist.

Question 20: **Incorrect**

You currently have a Bigtable instance you've been using for development running a development instance type, using HDD's for storage. You are ready to upgrade your development instance to a production instance for

increased performance. You also want to upgrade your storage to SSD's as you need maximum performance for your instance. What should you do?

A. Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD. **(Incorrect)**

B. Run parallel instances where one instance is using HDD and the other is using SSD.

C. Use the Bigtable instance sync tool in order to automatically synchronize two different instances, with one having the new storage configuration.

D. Build a Dataflow pipeline or Dataproc job to copy the data to the new cluster with SSD storage type. **(Correct)**

Explanation

Correct answer is **D** as the storage for the cluster cannot be updated. You need to define the new cluster and copy or import the data to it.

Refer GCP documentation - Bigtable Choosing HDD vs SSD

Switching between SSD and HDD storage

When you create a Cloud Bigtable instance and cluster, your choice of SSD or HDD storage for the cluster is permanent. You cannot *use the Google Cloud Platform Console to change the type of storage that is used for the cluster.*

If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance. Alternatively, you can use a Cloud Dataflow or Hadoop MapReduce job to copy the data from one instance to another. Keep in mind that migrating an entire instance takes time, and you might need to add nodes to your Cloud Bigtable clusters before you migrate your instance.

Option A is wrong as storage type cannot be changed.

Options B & C are wrong as it would have two clusters running at the same time with same data, thereby increasing cost.

Question 21: **Correct**

Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?

A. Rewrite the job in Pig.

B. Rewrite the job in Apache Spark. **(Correct)**

C. Increase the size of the Hadoop cluster.

D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Explanation

Correct answer is **B** as Spark can improve the performance as it performs lazy in-memory execution.

Spark is important because it does part of its pipeline processing in memory rather than copying from disk. For some applications, this makes Spark extremely fast. With a Spark pipeline, you have two different kinds of operations, transforms and actions. Spark builds its pipeline used an abstraction called a directed graph. Each transform builds additional nodes into the graph but spark doesn't execute the pipeline until it sees an action.

Spark waits until it has the whole story, all the information. This allows Spark to choose the best way to distribute the work and run the pipeline. The process of waiting on transforms and executing on actions is called, lazy execution. For a transformation, the input is an RDD 弹性分布式数据集*and the output is an RDD. When Spark sees a transformation, it registers it in the directed graph and then it waits. An action triggers Spark* *to process the pipeline, the output is usually a result format, such as a text file, rather than an RDD.*

Option A is wrong as Pig is wrapper and would initiate Map Reduce jobs

Option C is wrong as it would increase the cost.

Option D is wrong Hive is wrapper and would initiate Map Reduce jobs. Also, reducing the size would reduce performance.

Question 22: **Correct**

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery, so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee. How can you make that data available while minimizing cost?

A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.

B. Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.

C. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery. **(Correct)**

D. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.

Explanation

Correct answer is **C** as the best option is to create a new table with the updated columns. Dataflow provides a serverless NoOps option to convert data.

Option A is wrong as it is better to create materialized tables instead of views as the query would be executed everytime. Refer [BigQuery Best Practices]

Best practice: If possible, materialize your query results in stages.

If you create a large, multi-stage query, each time you run it, BigQuery reads all the data that *is required by the query. You are billed for all the data that is read each time the query is run.*

Instead, break your query into stages where each stage materializes the query results by writing them to a destination table. Querying the smaller destination table reduces the amount of data that is read and lowers costs. The cost of storing the materialized results is much less than the cost of processing large amounts of data.

Option B is wrong as DML are limited by quotas.

Maximum number of combined UPDATE, DELETE, and MERGE statements per day per table — 200

Option D is wrong as Dataproc would need provisioning of servers and writing scripts.

Question 23: **Incorrect**

A company's BigQuery data is currently stored in external CSV files in Cloud Storage. As the data has increased over the period of time, the query performance has dropped. What steps can help improve the query performance maintaining the cost-effectiveness?

A. Import the data into BigQuery for better performance. **(Correct)**

B. Request more slots for greater capacity to improve performance.

C. Divide the data into partitions based on date. **(Incorrect)**

D. Time to move to Cloud Bigtable; it is faster in all cases.

Explanation

Correct answer is **A** as the performance issue is because the data is stored in a non-optimal format in an external storage medium.

Refer GCP documentation - BigQuery External Data Sources

Query performance for external data sources may not be as high as querying data in a native BigQuery table. If query speed is a priority, load the data into BigQuery instead of setting up an external data source. The performance of a query that includes an external data source depends on the external storage type. For example, querying data stored in Cloud Storage is faster than querying data stored in Google Drive. In general, query performance for external data sources should be equivalent to reading the data directly from the external storage.

Option B is wrong as there is feature to request more slots.

Option C is wrong as partitioning of data at source would not improve query time for all use cases. - 没说一定按时间查询

Option D is wrong as Bigtable is more ideal for NoSQL data type and can get very expensive - 没说要低延迟

BQ技能存贮也能分析，要求速度的话，都放在一起最快

Question 24: **Incorrect**

A client is using Cloud SQL database to serve infrequently changing lookup tables that host data used by applications. The applications will not modify the tables. As they expand into other geographic regions they want to ensure good performance. What do you recommend?

A. Migrate to Cloud Spanner **(Correct)**

B. Read replicas **(Incorrect)**

C. Instance high availability configuration

D. Migrate to Cloud Storage

Explanation

Correct answer is **A** as Cloud Spanner provides a globally distributed relational database.`

Refer GCP documentation - Cloud Spanner

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.

Option B is wrong Cloud SQL, currently, does not support

read replicas in different geographic regions. 读取副本必须与主实例位于同一个区域。

Read replicas must be in the same region as the master instance.

Option C is wrong as high availability is for failover and not for performance.

Option D is wrong as Cloud Storage is not ideal storage for relational data.

Question 25: **Correct**

A company wants to connect cloud applications to an Oracle database in its data center. Requirements are a maximum of 9 Gbps of data and a Service Level Agreement (SLA) of 99%. Which option best suits the requirements?

A. Implement a high-throughput Cloud VPN connection

B. Cloud Router with VPN

C. Dedicated Interconnect

D. Partner Interconnect **(Correct)**

Explanation

Correct answer is **D** as Partner Interconnect is useful for data up to 10 Gbps and is offered by ISPs with SLAs.

Refer GCP documentation - Interconnect Options

Flexible capacity options with a minimum of 50 Mbps. More points of connectivity through one of our supported service providers. Traffic between networks flows through a service provider, not through the public Internet.

Google provides an SLA for the connection between Google and service provider. Whether an end-to-end SLA for the connection is offered, depends on your service provider. Check with them for more information.

Option A is wrong as Cloud VPN is over the internet through IPSec VPN at a low cost for your data bandwidth needs up to 3.0 Gbps.

Option B is wrong as Cloud Router helps only in dynamic routing.

Option C is wrong as Dedicated Interconnect is suitable for High bandwidth connections with a minimum of 10 Gbps. Traffic flows directly between networks, not through the public Internet.

Question 26: **Correct**

A company has migrated their Hadoop cluster to the cloud and is now using Cloud Dataproc with the same settings and same methods as in the data center. What would you advise them to do to make better use of the cloud environment?

A. Upgrade to the latest version of HDFS. Change the settings in Hadoop components to optimize for the different kinds of work in the mix.

B. Find more jobs to run so the cluster utilizations will cost-justify the expense.

C. Store persistent data off-cluster. Start a cluster for one kind of work then shut it down when it is not processing data. **(Correct)**

D. Migrate from Cloud Dataproc to an open source Hadoop Cluster hosted on Compute Engine, because this is the only way to get all the Hadoop customizations needed for efficiency.

Explanation

Correct answer is **C** as Storing persistent data off the cluster allows the cluster to be shut down when not processing data. And it allows separate clusters to be started per job or per kind of work, so tuning is less important.

Refer GCP documentation - Dataproc Cloud Storage

Direct data access *– Store your data in Cloud Storage and access it directly, with no need to transfer it into HDFS first.*
HDFS compatibility *– You can easily access your data in Cloud Storage using the* `gs://` *prefix instead of* `hdfs://`.
Interoperability 互操作性；互用性*– Storing data in Cloud Storage enables seamless interoperability between Spark, Hadoop, and Google services.*

__Data accessibility__ – When you shut down a Hadoop cluster, you still have access to your data in Cloud Storage, unlike HDFS.
__High data availability__ – Data stored in Cloud Storage is highly available and globally replicated without a loss of performance.
__No storage management overhead__ – Unlike HDFS, Cloud Storage requires no routine maintenance such as checking the file system, upgrading or rolling back to a previous version of the file system, etc.
__Quick startup__ – In HDFS, a MapReduce job can't start until the `NameNode` is out of safe mode —a process that can take from a few seconds to many minutes depending on the size and state of your data. With Cloud Storage, you can start your job as soon as the task nodes start, leading to significant cost savings over time.

Question 27: **Correct**

Your company is planning to migrate their analytics data into BigQuery. There is a need to handle both batch and streaming data. You are assigned the role to determine the costs that would be incurred for different operations. What are all of the BigQuery operations that Google charges for?

A. Storage, queries, and streaming inserts. **(Correct)**

B. Storage, queries, and loading data from a file.

C. Storage, queries, and exporting data.

D. Queries and streaming inserts.

Explanation

Correct answer is **A** as BigQuery charges for Storage, Queries and Streaming inserts. Loading and Exporting of data are free operations and not charged by BigQuery.

Refer GCP documentation - BigQuery Pricing

BigQuery offers scalable, flexible pricing options to help fit your project and your budget.

BigQuery storage costs are based solely on the amount of data you store. Storage charges can be: -
不常用的存储便宜

Active — A monthly charge for data stored in tables you have modified in the last 90 days.
Long-term — A lower monthly charge for data stored in tables that have not been modified in the last 90 days.
Query costs are based on the amount of data processed by the query. Query charges can be:

On-demand — The most flexible option. On-demand query pricing is based solely on usage.
Flat-rate — Enterprise customers generally prefer ***flat-rate pricing*** *for queries because it offers predictable, fixed month-to-month costs.*
Sample Pricing for US (multi-region)

Active storage	*$0.020 per GB*	*The first 10 GB is free each month.*

		See [Storage pricing](#) for details.
Long-term storage	*$0.010 per GB*	*The first 10 GB is free each month. See [Storage pricing](#) for details.*
Streaming Inserts	*$0.010 per 200 MB*	*You are charged for rows that are successfully inserted. Individual rows are calculated using a 1 KB minimum size. See [Streaming pricing](#) for details.*
Queries (analysis)	*$5.00 per TB*	*First 1 TB per month is free, see [On-demand pricing](#) for details. [Flat-rate pricing](#) is also available for high-volume customers.*

Options B & C are wrong as Loading and Exporting data are not charged.

Option D is wrong as Storage is also charged.

Question 28: **Correct**

Your company is in a highly regulated industry. You have 2 groups of analysts, who perform the initial analysis and sanitization of the data. You now need to provide analyst three secure access to these BigQuery query results, but not the underlying tables or datasets. How would you share the data?

A. Export the query results to a public Cloud Storage bucket.

B. Create a BigQuery Authorized View and assign a project-level user role to analyst three. **(Correct)**

C. Assign the bigquery.resultsonly.viewer role to analyst three.

D. Create a BigQuery Authorized View and assign an organizational level role to analyst three.

Explanation

Correct answer is **B** as you need to copy or store the query results in a separate dataset and provide authorization to view and/or use that dataset. The other solutions are not secure.

Refer GCP documentation - [BigQuery Authorized Views]

Giving a view access to a dataset is also known as creating an authorized view in BigQuery. An authorized view allows you to share query results with particular users and groups without giving them access to the underlying *tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query.*

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, your users would have access to both the view and the data.

Option A is wrong as a public Cloud Storage bucket is accessible to all.

Option C is wrong as there is no resultsonly viewer role.

Option D is wrong as an Organizational role would provide access to the underlying data as well.

Question 29: **Correct**

Your company is making the move to Google Cloud and has chosen to use a managed database service to reduce overhead. Your existing database is used for a product catalog that provides real-time inventory tracking for a retailer. Your database is 500 GB in size. The data is semi-structured and does not need full atomicity. You are looking for a truly no-ops/serverless solution. What storage option should you choose?

A. Cloud Datastore **(Correct)**

B. Cloud Bigtable

C. Cloud SQL

D. BigQuery

Explanation

Correct answer is **A** as Cloud Datastore offers NoOps NoSQL solution which is suited for Semistructured data and ideal for product catalogs.

Refer GCP documentation - Storage Options

Cloud Datastore	A scalable, fully managed NoSQL document database for your web and mobile applications.	Semistructured application data Hierarchical data Durable key-value data	User profiles Product catalogs Game state

Options B & C are wrong as they are not complete NoOps solution. Also Cloud SQL is not suited for Semi Structured data.

Option D is wrong as BigQuery is ideal for analytics solution

Question 30: Correct

Which of these numbers are adjusted by a neural network as it learns from a training dataset? (Choose two)

A. Continuous features

B. Input values

C. Weights (Correct)

D. Biases (Correct)

Explanation

Correct answers are **C & D** as weights and bias are the parameters learned by the computer from the training datasets.

Refer Google Cloud blog - Understanding Neural Network

As you can see a neural network is a simple mechanism that's implemented with basic math. The only difference between the traditional programming and neural network is, again, that you let the computer determine the parameters (weights and bias) by learning from training datasets. In other words, the trained weight pattern in our example wasn't programmed by humans.

Question 31: Correct

A user wishes to generate reports on petabyte scale data using a Business Intelligence (BI) tools. Which storage option provides integration with BI tools and supports OLAP workloads up to petabyte-scale?

A. Bigtable

B. Cloud Datastore

C. Cloud Storage

D. BigQuery (Correct)

Explanation

Correct answer is **D** as BigQuery is fully managed data warehouse and is fast and easy to use on data of any size. With BigQuery, you'll get great performance on your data, while knowing you can scale seamlessly to store and analyze petabytes more without having to buy more capacity.

Refer GCP documentation - [Storage Options]

[BigQuery]	A scalable, fully managed enterprise data warehouse (EDW) with SQL and fast ad-hoc queries.	OLAP workloads up to petabyte scale Big data exploration and processing Reporting via business intelligence (BI) tools	Analytical reporting on large data Data science and advanced analyses Big data processing using SQL

Options A & B are wrong as Bigtable & Datastore are NoSQL solution and not suitable for OLAP data warehouse work loads.

Option C is wrong as Cloud Storage provides object storage only.

Question 32: **Correct**

Your company is planning to migrate their historical dataset into BigQuery. This data would be exposed to the data scientists for perform analysis

using BigQuery ML. The data scientists would like to know which ML models does the BigQuery ML support. What would be your answer? (Choose 2)

A. Random Forest

B. Linear Regression **(Correct)**

C. K Means

D. Principal Component Analysis

E. Multiclass logistic regression for Classification **(Correct)**

Explanation

Correct answers are **B & E** as BigQuery ML supports Linear regression, Binary Logistic regression and Multiclass logistic regression.

Refer GCP documentation - BigQuery ML

BigQuery ML currently supports the following types of models:

Linear regression — These models can be used for predicting a numerical value.
Binary logistic regression — These models can be used for predicting one of two classes (such as identifying whether an email is spam). - 是或者不是
Multiclass logistic regression for classification — These models can be used to predict more than two classes such as whether an input is "low-value", "medium-value", or "high-value".

- Linear regression for forecasting; for example, the sales of an item on a given day. Labels are real-valued (they cannot be +/- infinity or NaN).

- Binary logistic regression for classification; for example, determining whether a customer will make a purchase. Labels must only have two possible values.
- Multiclass logistic regression for classification. These models can be used to predict multiple possible values such as whether an input is "low-value," "medium-value," or "high-value." Labels can have up to 50 unique values. In BigQuery ML, multiclass logistic regression training uses a multinomial classifier with a cross entropy loss function.
- K-means clustering for data segmentation (beta); for example, identifying customer segments. K-means is an unsupervised learning technique, so model training does not require labels nor split data for training or evaluation.
- TensorFlow model importing. This feature allows you to create BigQuery ML models from previously-trained TensorFlow models, then perform prediction in BigQuery ML. See the CREATE MODEL statement for importing TensorFlow models for more information.

Question 33: **Correct**

Your company wants to develop an REST based application for text analysis to identify entities and label by types such as person, organization, location, events, products, and media from within a text. You need to do a quick Proof of Concept (PoC) to implement and demo the same. How would you design your application?

A. Create and Train a model using Tensorflow and Develop an REST based wrapper over it

B. Create and Train a model using BigQuery ML and Develop an REST based wrapper over it

C. Use Cloud Natural Language API and Develop an REST based wrapper over it **(Correct)**

D. Use Cloud Vision API and Develop an REST based wrapper over it

Explanation

Correct answer is **C** as the solution needs to developed quickly, the Cloud Natural Language API can be used to perform text analysis.

Refer GCP documentation - [AI Products]

Cloud Natural Language API reveals the structure and meaning of text by offering powerful machine learning models in an easy-to-use REST API. And with AutoML Natural Language Beta you can build and train ML models easily, without extensive ML expertise. You can use Natural Language to extract information about people, places, events, and much more mentioned in text documents, news articles, or blog posts. You can also use it to understand sentiment about your product on social media or parse intent from customer conversations happening in a call center or a messaging app.

Options A & B are wrong as they do not provide quick results.

Option D is wrong as Cloud Vision is for image analysis and not text analysis.

Question 34: **Correct**

Your company wants to transcribe the conversations between the manufacturing employees at real time. The conversations are recorded using old radio systems in the 8000Hz frequency. They are in English with a short duration of 35-40 secs. You need to design the system inline with Google recommended best practice. How would you design the application?

A. Use Cloud Speech-to-Text API

(Correct)

in synchronous mode

B. Use Cloud Speech-to-Text API in asynchronous mode

C. Re-sample the audio using 16000Hz frequency and Use Cloud Speech-to-Text API in synchronous mode

D. Re-sample the audio using 16000Hz frequency and Use Cloud Speech-to-Text API in asynchronous mode

Explanation

Correct answer is **A** as Speech-to-Text can be used to convert short duration audio in synchronous calls. As well as it is recommended not to re-sample the data, if it is coming at a lower sampling rate from the source.

Refer GCP documentation - Speech-to-Text Sync & Best Practices

Lower sampling rates may reduce accuracy. However, avoid re-sampling. For example, in telephony the native rate is commonly 8000 Hz, which is the rate that should be sent to the service.

Synchronous speech recognition returns the recognized text for short audio (less than ~1 minute) in the response as soon as it is processed. To process a speech recognition request for long audio, useAsynchronous Speech Recognition.

Question 35: **Correct**

You have lot of Spark jobs. Some jobs need to run independently while others can run parallelly. There is also inter-dependency between the jobs and the dependent jobs should not be triggered unless the previous ones are completed. How do you orchestrate the pipelines?

A. Cloud Dataproc

B. Cloud Scheduler

C. Schedule jobs on a single Compute Engine using Cron.

D. Cloud Composer **(Correct)**

Explanation

Correct answer is **D** as Cloud Composer can help create workflows that connect data, processing, and services across clouds, giving you a unified data environment.

Refer GCP documentation - Cloud Composer

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers. Built on the popular Apache Airflow open source project and operated using the Python programming language, Cloud Composer is free from lock-in and easy to use.

Cloud Composer pipelines are configured as directed acyclic graphs (DAGs) using Python, making it easy for users of any experience level to author and schedule a workflow. One-click deployment yields instant access to a rich library of connectors and multiple graphical representations of your workflow in action, increasing pipeline reliability by making troubleshooting easy. Automatic synchronization of your directed acyclic graphs ensures your jobs stay on schedule.

Option A is wrong as Google Cloud Dataproc is a fast, easy to use, managed Spark and Hadoop service for distributed data processing. It does not help easy orchestration.

Option B is wrong as Cloud Scheduler is a fully managed enterprise-grade cron job scheduler. It is not an orchestration tool.

Option C is wrong as it does not help orchestrate the dependency between jobs, but merely schedule them.

Question 36: **Correct**

Your company is planning to host its analytics data in BigQuery. You are required to control access to the dataset with least privilege meeting the following guidelines

Each team has multiple Team Leaders, who should have the

ability to create, delete tables, but not delete dataset.

Each team has Data Analysts, who should be able to query data, but not modify it

How would you design the access control?

A. Grant Team leader group - OWNER and Data Analyst - WRITER

B. Grant Team leader group - OWNER and Data Analyst - READER

C. Grant Team leader group - WRITER and Data Analyst - READER **(Correct)**

D. Grant Team leader group - READER and Data Analyst - WRITER

Explanation

Correct answer is **C** as Team leader group should be provider the WRITER access and the Data Analysts should be provided only the reader access.

Refer GCP documentation - BigQuery Dataset Primitive Roles

Dataset role	Capabilities
READER	Can read, query, copy or export tables in the dataset Can call get on the dataset Can call get and list on tables in the dataset Can call list on table data for tables in the dataset

WRITER	Same as READER, plus: Can edit or append data in the dataset Can call insert, insertAll, update or delete Can use tables in the dataset as destinations for load, copy or query jobs
OWNER	Same as WRITER, plus: Can call update on the dataset Can call delete on the dataset Note: A dataset must have at least one entity with the OWNER role. A user with the OWNER role can't remove their own OWNER role.

Options A & D are wrong as Data Analyst should not have the WRITER permissions

Options A & B are wrong as Team leader should not have the OWNER permission

Question 37: **Correct**

Your company wants to develop a system to measure the feedback of their products from the reviews posted by people on various Social media platforms. The reviews are mainly text based. You need to do a quick Proof of Concept (PoC) to implement and demo the same. How would you design your application?

A. Create and Train a sentiment analysis model using Tensorflow

B. Use Cloud Speech-to-Text API for sentiment analysis

C. Use Cloud

Natural Language API for sentiment analysis **(Correct)**

D. Use Cloud Vision API for sentiment analysis

Explanation

Correct answer is **C** as Natural Language processing provides pre-model to perform sentiment analysis.

Refer GCP documentation - Cloud Natural Language

You can use Cloud Natural Language to extract information about people, places, events, and much more mentioned in text documents, news articles, or blog posts. You can use it to understand sentiment about your product on social media or parse intent from customer conversations happening in a call center or a messaging app. You can analyze text uploaded in your request or integrate with your document storage on Google Cloud Storage.

Option A is wrong as building and training a senetiment analysis model using Tensorflow would take time and effort.

Option B is wrong as Speech-to-Text API is for audio to text conversion.

Option D is wrong as Cloud Vision is for image analysis.

Question 38: Correct

Your company receives a lot of financial data in CSV files. The files need to be processed, cleaned and transformed before they are made available for analytics. The schema of the data also changes every third month. The Data analysts should be able to perform the tasks

1. No prior knowledge of any language with no coding

2. Provided a GUI tool to build and modify the schema

What solution best fits the need?

A. Use Dataflow code and provide Data Analysts the access to the code. Store the schema externally to be easily modified.

B. Use Dataprep with transformation recipes. **(Correct)**

C. Use Dataproc spark and provide Data Analysts the access to the code. Store the schema externally to be easily modified.

D. Use DataLab with transformation recipes.

Explanation

Correct answer is **B** as Dataprep can be used to handle schema changes by Data Analysts without any programming knowledge, but through an easy to use GUI.

Refer GCP documentation - Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually *exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.*

Visually explore and interact with data in seconds. Instantly understand data distribution and patterns. You don't need to write code. You can prepare data with a few clicks.

Process diverse datasets — structured and unstructured. Transform data stored in CSV, JSON, or relational table formats. Prepare datasets of any size, megabytes to terabytes, with equal ease.

Options A, C & D are wrong as they would need programming knowledge.

Question 39: **Correct**

An organization wishes to enable real time analytics on user interactions on their web application. They estimate that there will be 1000 interactions per second and wishes to use services, which are ops free. Which combination of services can be used in this case?

A. App Engine, Dataproc, DataStudio

B. Compute Engine, BigQuery Streaming Inserts, DataStudio

C. App Engine,

BigQuery Streaming Inserts, DataStudio **(Correct)**

D. App Engine, Dataflow, DataStudio

Explanation

Correct answer is **C** as the focus is more on **NoOps**, the **App Engine** can be used to capture and insert the data into **BigQuery** using streaming inserts. The data can then be analyzed and visualized using **DataStudio**.

Options A & D are wrong as Dataflow and Dataproc would need processing and storage.

Option B is wrong as Compute Engine would not be Ops free.

compute engine 需要OPS

Question 40: Correct

Your company has assigned fixed number for slots to each project for BigQuery. Each project wants to monitor the number of available slots. How can the monitoring be configured?

A. Monitor the BigQuery Slots Used metric

B. Monitor the BigQuery Slots Pending metric

C. Monitor the BigQuery Slots Allocated metric

D. Monitor the

BigQuery Slots Available metric	**(Correct)**

Explanation

Correct answer is **D** as BigQuery provides 2 metrics for Slots. Slots Allocated to the project and Slots Available for the project.

Refer GCP documentation - [BigQuery Metrics]

BigQuery	Slots available	slots	Total number of slots available to the project. If the project shares a reservation of slots with other projects the slots being used by the other projects is not depicted.

Question 41: **Correct**

Your company is working on real time click stream analysis. They want to implement a feature to capture user click during a session and aggregate the count for that session. Session timeout is 30 mins. How

would you design the data processing?

A. Use Dataflow and fixed windowing of 30 minutes

B. Use Dataflow and Session windowing with gap duration of 30 minutes **(Correct)**

C. Use Dataflow and Global window with gap duration of 30 minutes

D. Use Dataproc and store the data in BigQuery and aggregate the same

Explanation

Correct answer is **B** as Dataflow would help in performing real time analytics and data count aggregation over a window. Session windows to track the session for the aggregate click count by the user.

Refer GCP documentation - Beam Windowing Basics

*A **session window** function defines windows that contain elements that are within a certain gap duration of another element. Session windowing applies on a per-key basis and is useful for data that is irregularly distributed with respect to time. For example, a data stream representing user mouse activity may have long periods of idle time interspersed with high concentrations of clicks. If data arrives after the minimum specified gap duration time, this initiates the start of a new window.*

Options A & C are wrong as Fixed and Global windowing would not work.

Option D is wrong as Dataproc and BigQuery would not provide real time analytics.

!!!!!!!Question 42: Correct

You have a real time data processing pipeline running in Dataflow. As a part of changed requirement you need to update the windowing and triggering strategy for the pipeline. You want to update the pipeline without any loss of in-flight messages. What is the best way to deploy the changes?

A. Stop with pipeline using the drain option and use new Dataflow pipeline

B. Stop with pipeline using the cancel option and use new Dataflow pipeline

C. Pass the --update option with --jobname parameter to the same name as the job you want to update **(Correct)**

D. Pass the --update option with --jobname parameter to the new job name you want to update

Explanation

Correct answer is **C** as Dataflow allows updates to the existing pipeline in case of compatible changes while saving the intermediate state data.

Refer GCP documentation - [Dataflow 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** `jobId`.*

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.

You can change windowing and trigger strategies for the `PCollection`s in your replacement pipeline, but use caution. Changing the windowing or trigger strategies will not affect data that is already buffered or otherwise in-flight.

We recommend that you attempt only smaller changes to your pipeline's windowing, such as changing the duration of fixed- or sliding-time windows. Making major changes to windowing or triggers, like changing the windowing algorithm, might have unpredictable results on your pipeline output.

To update your job, you'll need to launch a new job to replace the ongoing job. When you launch your replacement job, you'll need to set the following pipeline options to perform the update *process in addition to the job's regular options:*

Pass the `--update` option.
Set the `--jobName` option in PipelineOptions to the ***same name as the job you want to update****.*
If any transform names in your pipeline have changed, you must supply a transform mapping and pass it using the `--transformNameMapping` option.
Option A is wrong as with Drain option the windows and triggers would closed immediately.

When you issue the Drain command, Cloud Dataflow immediately closes any in-process windows and fires all triggers*. The system* ***does not*** *wait for any outstanding time-based windows to finish. For example, if your pipeline is ten minutes into a two-hour window when you issue the Drain command, Cloud Dataflow won't wait for the remainder of the window to finish. It will close the window immediately with partial results.*

Option B is wrong as Cancel immediately halts processing, you may lose any "in-flight" data.

Option D is wrong as the job name should be the same.

!!!!!Question 43: **Correct**

Your company is planning to migrate its data first to Google Cloud Storage. You need to keep the contents of this bucket

in sync with a new Google Cloud Storage bucket to support a backup storage destination. What is the best method to achieve this?

A. Once per week, use a gsutil cp command to copy over newly modified files.

B. Use gsutil rsync commands to keep both locations in sync. **(Correct)**

C. Use Storage Transfer Service to keep both the source and destination in sync.

D. Use gsutil -m cp to keep both locations in sync.

Explanation

Correct answer is **B** as the data transfer is between on-premises and Google Cloud, the gsutil **rsync** can be used to keep the source and destination in sync.

gsutil rsync command makes the contents under dst_url the same as the contents under src_url, by copying any missing files/objects (or those whose data has changed), and (if the -d option is specified) deleting any extra files/objects. src_url must specify a directory, bucket, or bucket subdirectory.

Options A & D are wrong as copy can be used to copy, however there needs to be more handling to keep it in sync.

Option C is wrong as the data is not available in an online location.

Question 44: Correct

Your company hosts a 2PB on-premises Hadoop cluster with sensitive data. They want to plan the migration of the cluster to Google Cloud as part of phase 1 activity before the jobs are moved. Current network speed between the colocation and cloud is 10Gbps. What is the efficient way to transfer the data?

A. Use Transfer appliance to transfer the data to Cloud Storage **(Correct)**

B. Expose the data as a public URL and Storage Transfer Service to transfer it

C. Use gsutil command to transfer the data to Cloud Storage

D. Use hadoop distcp command to copy the data between cluster

Explanation

Correct answer is **A** as even with 10Gbps of transfer speed it would take minimum 24 days (assuming consistent speed and no interruption) to transfer the complete data. So the best option is to use Google Transfer Appliance.

Refer GCP documentation - Data Transfer

Google Transfer Appliance - Securely capture, ship, and upload *your data to Google Cloud Storage using the Transfer Appliance 100 TB or 480 TB models.*

Options B, C & D are wrong as they would still route the request through Internet.

Question 45: Correct

You have migrated your Hadoop jobs with external dependencies on a Dataproc cluster. As a security requirement, the cluster has been setup using internal IP addresses only and does not have a direct Internet connectivity. How can the cluster be configured to allow the installation of the dependencies?

A. Setup a SSH tunnel to Internet and route outbound requests through it.

B. Store the external dependencies in Cloud Storage and modify the initialization scripts **(Correct)**

C. Setup a SOCKS proxy and route outbound requests through it.

D. Setup the Dataproc master node is public subnet to be able to download external dependencies

Explanation

Correct answer is **B** as the Dataproc cluster is configured with internal IP addresses only, the dependencies can be stored in Cloud Storage so that they can be accessed using internal IPs.

Refer GCP documentation - Dataproc Init Actions

*If you create a Cloud Dataproc cluster with internal IP addresses only, attempts to access the Internet in an initialization action will fail unless you have configured routes to direct the traffic through **a NAT or a VPN gateway**. Without access to the Internet, you can enable Private Google Access, and place job dependencies in Cloud Storage; cluster nodes can download the dependencies from Cloud Storage from internal IPs.*

Options A, C & D are wrong as they would not allow secure outbound connection.

Question 46: **Correct**

You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud. You are using ANSI SQL to run queries for your analysts. You want to support complex aggregate queries and reuse existing code. How should you transform the input data?

A. Use BigQuery for storage. Use Cloud Dataflow to run the transformations.

B. Use BigQuery for storage. Use Cloud Dataproc to run the transformations. **(Correct)**

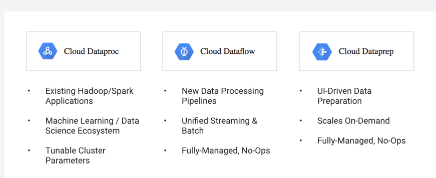
C. Use Cloud Storage for storage. Use Cloud Dataflow to run the transformations.

D. Use Cloud Storage for storage. Use Cloud Dataproc to run the transformations.

Explanation

Correct answer is **B** as there are 2 requirements to reuse existing Spark code and support ANSI SQL queries. Dataproc helps reuse the Spark jobs as is and ANSI SQL queries require the use of BigQuery. Google Cloud Dataproc is a fast, easy to use, managed Spark and Hadoop service for distributed data processing.

Refer GCP documentation - Data lifecycle @ https://cloud.google.com/solutions/data-lifecycle-cloud-platform#processing_large-scale_data



Option A is wrong as Dataflow does not support Spark jobs. Google Cloud Dataflow is a fully managed service for strongly consistent, parallel data-processing pipelines.

Options C & D are wrong as Cloud Storage directly does not support ANSI SQL queries and Cloud Dataflow does not support Spark.

Question 47: Correct

As part of your backup plan, you set up regular snapshots of Compute Engine instances that are running. You want to be able to restore these snapshots using the fewest possible steps for replacement instances. What should you do?

A. Export the snapshots to Cloud Storage. Create disks from the exported snapshot files. Create images from the new disks. Use the image to create instances as needed.

B. Export the snapshots to Cloud Storage. Create images from the exported snapshot files. Use the image to create instances as needed.

C. Use the snapshots to create replacement disks. Use the disks to create instances as needed.

D. Use the snapshots to create replacement instances as needed. **(Correct)**

Explanation

Correct answer is **D** as the question focuses on minimal steps and the snapshot is available, an instance can be directly created from the snapshot.

Refer GCP documentation - Compute Engine - Create Instance @ https://cloud.google.com/compute/docs/instances/create-start-instance

Creating an instance from an image

Creating an instance from a public image

Creating an instance from a custom image

Creating an instance with an image shared with you

Creating an instance from a snapshot

Creating an instance from a container image

Options A, B & C are wrong as it is possible, however they are multi-step process.

Question 48: **Correct**

You are asked to design next generation of smart helmet for accident detection and reporting system. Each helmet will push 10kb of biometric data In JSON format every 1 second to a collection platform that will process and use trained machine learning model to predict and detect if an accident happens and send notification. Management has tasked you to architect the platform ensuring the following requirements are met:

· Provide the ability for real-time analytics of the inbound biometric data

· Ensure ingestion and processing of the biometric data is highly durable. Elastic and parallel

· The results of the analytic processing should be persisted for data mining to improve the accident detection ML model in the future.

Which architecture outlined below win meet the initial requirements for the platform?

A. Utilize Cloud Storage to collect the inbound sensor data, analyze data with Dataproc and save the results to BigQuery.

B. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to BigQuery. **(Correct)**

C. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Cloud SQL.

D. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Bigtable.

Explanation

Correct answer is **B** as Cloud Pub/Sub provides elastic and scalable ingestion, Dataflow provides processing and BigQuery analytics.

Refer GCP documentation - IoT @ https://cloud.google.com/solutions/iot-overview

Google Cloud Pub/Sub provides a globally durable message ingestion service. By creating topics for streams or channels, you can enable different components of your application to subscribe to specific streams of data without needing to construct subscriber-specific channels on each device. Cloud Pub/Sub also natively connects to other Cloud Platform services, helping you to connect ingestion, data pipelines, and storage systems.

Google Cloud Dataflow provides the open Apache Beam programming model as a managed service for processing data in multiple ways, including batch operations, extract-transform-load (ETL) patterns, and continuous, streaming computation. Cloud Dataflow can be particularly useful for managing the high-volume data processing pipelines required for IoT scenarios. Cloud Dataflow is also designed to integrate seamlessly with the other Cloud Platform services you choose for your pipeline.

Google BigQuery provides a fully managed data warehouse with a familiar SQL interface, so you can store your IoT data alongside any of your other enterprise analytics and logs. The performance and cost of BigQuery means you might keep your valuable data longer, instead of deleting it just to save disk space.

Option A is wrong as Cloud Storage is not an ideal ingestion service for real time high frequency data. Also Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.

Option C is wrong as Cloud SQL is a relational database and not suited for analytics data storage.

Option D is wrong as Bigtable is not ideal for long term analytics data storage.

BIGTABLE不适合长期存储分析

Question 49: **Correct**

Your company processes high volumes of IoT data that are time-stamped. The total data volume can be several petabytes. The data needs to be written and changed at a high speed. You want to use the most performant storage option for your data. Which product should you use?

A. Cloud Datastore

B. Cloud Storage

C. Cloud Bigtable **(Correct)**

D. BigQuery

Explanation

Correct answer is **C** as Cloud Bigtable is the most performant storage option to work with IoT and time series data. Google Cloud Bigtable is a fast, fully managed, highly-scalable NoSQL database service. It is designed for the collection and retention of data from 1TB to hundreds of PB.

Refer GCP documentation - Bigtable Time series data @ https://cloud.google.com/bigtable/docs/schema-design-time-series

Option A is wrong as Cloud Datastore is not the most performant product for frequent writes or timestamp-based queries.

Option B is wrong as Cloud Storage is designed for object storage not for this type of data ingestion and collection.

Option D is wrong as BigQuery is more of an a scalable, fully managed enterprise data warehousing solution and not ideal fast changing data.

BQ不适合频繁的修改数据

Question 50: **Correct**

A startup plans to use a data processing platform, which supports both batch and streaming applications. They would prefer to have a hands-off/serverless data processing platform to start with. Which GCP service is suited for them?

A. Dataproc

B. Dataprep

C. Dataflow (Correct)

D. BigQuery

Explanation

Correct answer is **C** as Dataflow helps design data processing pipelines and *is a fully managed service for strongly consistent, parallel data-processing pipelines. It provides an SDK for Java with composable primitives for building data-processing pipelines for batch or continuous processing. This service manages the life cycle of Google Compute Engine resources of the processing pipeline(s). It also provides a monitoring user interface for understanding pipeline health.*

Refer GCP documentation - Dataflow @ https://cloud.google.com/dataflow/

Cloud Dataflow is a fully-managed service for transforming and enriching data in stream(real time) and batch (historical) modes with equal reliability and expressiveness -- no more complex workarounds or compromises needed. And with its serverless approach to resource provisioning and management, you have access to virtually limitless capacity to solve your biggest data processing challenges, while paying only for what you use.

Cloud Dataflow unlocks transformational use cases across industries, including:

Clickstream, Point-of-Sale, and segmentation analysis in retail

Fraud detection in financial services

Personalized user experience in gaming

IoT analytics in manufacturing, healthcare, and logistics

Option A is wrong as Google Cloud Dataproc is a fast, easy to use, managed Spark and Hadoop service for distributed data processing. It is not serverless and more suited for batch processing.

Option B is wrong as Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. It does not help process batch and streaming data.

Option D is wrong as BigQuery is an analytics data warehousing solution.

Google Cloud Certified - Professional Data Engineer Practice Exam 1 - Results

Attempt 3

Question 1: Correct

You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than

1 hour old. What should you do?

A. Disable caching by editing the report settings. (Correct)

B. Disable caching in BigQuery by editing table details.

C. Refresh your browser tab showing the visualizations.

D. Clear your browser history for the past hour then reload the tab showing the visualizations.

Explanation

Correct answer is A as Data Studio caches data for performance and as the latest data is not shown, the caching can be disabled to fetch the latest data.

Refer GCP documentation - Data Studio Caching

Option B is wrong as BigQuery does not cache the data.

Options C & D are wrong this would not allow fetching of latest data.

Question 2: Correct

You company's on-premises Hadoop and Spark jobs have been migrated to Cloud Dataproc. When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through which proxy?

A. HTTPS

B. VPN

C. SOCKS (Correct)

D. HTTP

Explanation

Correct answer is **C** as the internal services can be accessed using the SOCKS proxy server.

Refer GCP documentation - Dataproc - Connecting to web interfaces

You can connect to web interfaces running on a Cloud Dataproc cluster using your project's Cloud Shell or the Cloud SDK gcloud command-line tool:

Cloud Shell: The Cloud Shell in the Google Cloud Platform Console has the Cloud SDK commands and utilities pre-installed, and it provides a Web Preview feature that allows you to quickly connect through an SSH tunnel to a web interface port on a cluster. However, a connection to the cluster from Cloud Shell uses local port forwarding, which opens a connection to only one port on a cluster web interface—multiple commands are needed to connect to multiple ports. Also, Cloud Shell sessions automatically terminate after a period of inactivity (30 minutes).

`gcloud` command-line tool: The `gcloud compute ssh` command with dynamic port forwarding allows you to establish an SSH tunnel and run a SOCKS proxy server on top of the tunnel. After issuing this command, you must configure your local browser to use the SOCKS proxy. This connection method allows you to connect to multiple ports on a cluster web interface.

Question 3: Correct

Your company is planning to migrate their on-premises Hadoop and Spark jobs to Dataproc. Which role must be assigned to a service account used by the virtual machines in a Dataproc cluster, so they can execute jobs?

A. Dataproc Worker (Correct)

B. Dataproc Viewer

C. Dataproc Runner

D. Dataproc Editor

Explanation

Correct answer is A as the compute engine should have Dataproc Worker role assigned.

Refer GCP documentation - [Dataproc Service Accounts]

Service accounts have [IAM roles] granted to them. Specifying a user-managed *service account when creating a Cloud Dataproc cluster allows you to create and utilize clusters with fine-grained access and control to Cloud resources. Using multiple user-managed service accounts with different Cloud Dataproc clusters allows for clusters with different access to Cloud resources.*

Service accounts used with Cloud Dataproc must have [Dataproc/Dataproc Worker] role (or have all the permissions granted by Dataproc Worker role).

Question 4: Correct

You currently have a Bigtable instance you've been using for development running a development instance type, using HDD's for storage. You are ready to upgrade your development instance to a production instance for increased performance. You also want to upgrade your storage to SSD's as you need maximum performance for your instance. What should you do?

A. Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD.

B. Export your Bigtable data into a new instance, and configure the new instance type as (Correct)

production with
SSD's

C. Run parallel instances where one instance is using HDD and the other is using SSD.

D. Use the Bigtable instance sync tool in order to automatically synchronize two different instances, with one having the new storage configuration.

Explanation

Correct answer is B as the storage for the cluster cannot be updated. You need to define the new cluster and copy or import the data to it.

Refer GCP documentation - [Bigtable Choosing HDD vs SSD]()

Switching between SSD and HDD storage

When you create a Cloud Bigtable instance and cluster, your choice of SSD or HDD storage for the cluster is permanent. You cannot use the Google Cloud Platform Console to change the type of storage that is used for the cluster.

If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance. Alternatively, you can use a Cloud Dataflow or Hadoop MapReduce job to copy the data from one instance to another. Keep in mind that migrating an entire instance takes time, and you *might need to add nodes to your Cloud Bigtable clusters before you migrate your instance.*

Option A is wrong as storage type cannot be changed.

Options C & D are wrong as it would have two clusters running at the same time with same data, thereby increasing cost.

Question 5: Correct

You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type. Reload the data.

B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column DT for each row. Reference the column TS instead of the column DT from now on.

C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.

D. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM. (Correct)

Explanation

Correct answer is D as the column type cannot be changed and the column needs to casting loaded into a new table using either SQL Query or import/export.

Refer GCP documentation - BigQuery Changing Schema

Changing a column's data type is not supported by the GCP Console, the classic BigQuery web UI, the command-line tool, or the API. If you attempt to update a table by applying a schema that specifies a new data type for a column, the following error is returned: `BigQuery error in update operation: Provided Schema does not match Table [PROJECT_ID]:[DATASET].[TABLE].`

There are two ways to manually change a column's data type:

Using a SQL query — Choose this option if you are more concerned about simplicity and ease of use, and you are less concerned about costs.
Recreating the table — Choose this option if you are more concerned about costs, and you are less concerned about simplicity and ease of use.

Option 1: Using a query

Use a SQL query to select all the table data and to cast the relevant column as a different data type. You can use the query results to overwrite the table or to create a new destination table.

Option A is wrong as with this approach all the data would be lost and needs to be reloaded

Option B is wrong as numeric values cannot be used directly and would need casting.

Option C is wrong as view is not materialized views, so the future queries would always be taxed as the casting would be done always.

Question 6: Correct

Your company has a BigQuery dataset created, which is located near Tokyo. For efficiency reasons, the company now wants the dataset duplicated in Germany. How can be dataset be made available to the users in Germany?

A. Change the dataset from a regional location to multi-region location, specifying the regions to be included.

B. Export the data from BigQuery into a bucket in the new location, and import it into a new dataset at the new location.

C. Copy the data from the dataset in the source region to the dataset in the target region using BigQuery commands.

D. Export the data from BigQuery into nearby bucket in Cloud Storage. Copy to a new regional bucket in Cloud Storage in the new location and Import into the new dataset. (Correct)

Explanation

Correct answer is D as the dataset location cannot be changed once created. The dataset needs to be copied using Cloud Storage.

Refer GCP documentation - BigQuery Exporting Data

You cannot change the location of a dataset after it is created. Also, you cannot move a dataset from one location to another. If you need to move a dataset from one location to another, follow this process:

1. *Export the data from your BigQuery tables to a regional or multi-region Cloud Storage bucket in the same location as your dataset. For example, if your dataset is in the EU multi-region location, export your data into a regional or multi-region bucket in the EU.There are no charges for exporting data from BigQuery, but you do incur charges for storing the exported data in Cloud Storage. BigQuery exports are subject to the limits on export jobs.*
2. *Copy or move the data from your Cloud Storage bucket to a regional or multi-region bucket in the new location. For example, if you are moving your data from the US multi-region location to the Tokyo regional location, you would transfer the data to a regional bucket in Tokyo. Note that*

transferring data between regions incurs network egress charges in Cloud Storage.

3. *After you transfer the data to a Cloud Storage bucket in the new location, create a new BigQuery dataset (in the new location). Then, load your data from the Cloud Storage bucket into BigQuery.You are not charged for loading the data into BigQuery, but you will incur charges for storing the data in Cloud Storage until you delete the data or the bucket. You are also charged for storing the data in BigQuery after it is loaded. Loading data into BigQuery is subject to the limits on load jobs.*

Question 7: Correct

A company has loaded its complete financial data for last year for analytics into BigQuery. A Data Analyst is concerned that a BigQuery query could be too expensive. Which methods can be used to reduce the number of rows processed by BigQuery?

A. Use the LIMIT clause to limit the number of values in the results.

B. Use the SELECT clause to limit the amount of data in the query. Partition data by date so the query can be more focused. (Correct)

C. Set the Maximum Bytes Billed, which will limit the number of bytes processed but still run the query if the number of bytes requested goes over the limit.

D. Use GROUP BY so the results will be grouped into fewer output values.

Explanation

Correct answer is **B** as SELECT with partition would limit the data for querying.

Refer GCP documentation - BigQuery Cost Best Practices

Best practice: *Partition your tables by date.*

If possible, partition your BigQuery tables by date. Partitioning your tables allows you to query relevant subsets of data which improves performance and reduces costs.

For example, when you query partitioned tables, use the `_PARTITIONTIME` pseudo column to filter for a date or a range of dates. The query processes data only in the partitions that are specified by the date or range.

Option A is wrong as LIMIT does not reduce cost as the amount of data queried is still the same.

Best practice: *Do not use a `LIMIT` clause as a method of cost control.*

Applying a `LIMIT` clause to a query does not affect the amount of data that is read. It merely limits the results set output. You are billed for reading all bytes in the entire table as indicated by the query.

The amount of data read by the query counts against your free tier quota despite the presence of a `LIMIT` clause.

Option C is wrong as the query would fail and would not execute if the Maximum bytes limit is exceeded by the query.

Best practice: *Use the maximum bytes billed setting to limit query costs.*

You can limit the number of bytes billed for a query using the maximum bytes billed setting. When you set maximum bytes billed, if the query will read bytes beyond the limit, the query fails without incurring a charge.

Option D is wrong as GROUP BY would return less output, but would still query the entire data.

Question 8: Correct

Your company receives streaming data from IoT sensors capturing various parameters. You need to calculate a running average for each of the parameter on streaming data, taking into account the data that can arrive late and out of order. How would you design the system?

A. Use Cloud Pub/Sub and Cloud Dataflow with Sliding Time Windows. (Correct)

B. Use Cloud Pub/Sub and Google Data Studio.

C. Cloud Pub/Sub can guarantee timely arrival and order.

D. Use Cloud Dataflow's built-in timestamps for ordering and filtering.

Explanation

Correct answer is A as Cloud Pub/Sub does not maintain message order and Dataflow can be used to order the messages and as well as calculate average using Sliding Time window.

Refer GCP documentation - Pub/Sub Subscriber

Cloud Pub/Sub delivers each message once and in the order in which it was published. However, messages may sometimes be delivered out of order or more than once. In general, accommodating more-than-once delivery requires your subscriber to be idempotent when processing messages. You can achieve exactly once processing of Cloud Pub/Sub *message streams using Cloud Dataflow `PubsubIO`. `PubsubIO` de-duplicates messages on custom message identifiers or those assigned by Cloud Pub/Sub. You can also achieve ordered processing with Cloud Dataflow by using the standard sorting APIs of the service. Alternatively, to achieve ordering, the publisher of the topic to which you subscribe can include a sequence token in the message.*

Option B is wrong as Data Studio is more of a visualization tool and does not help in analysis or ordering of messages.

Option C is wrong as Cloud Pub/Sub does not guarantee order and arrival.

Option D is wrong as Dataflow does not provide built-in timestamps for ordering and filtering. It needs to use the watermark/timestamp introduced either by the publisher source or Cloud Pub/Sub.

Question 9: Correct

You have developed a Machine Learning model to categorize where the financial transaction was a fraud or not. Testing the Machine Learning model with validation data returns 100% correct answers. What can you infer from the results?

A. The model is working

extremely well, indicating the hyperparameters are set correctly.

B. The model is overfit. There is a problem. (Correct)

C. The model is underfit. There is a problem.

D. The model is perfectly fit. You do not need to continue training.

Explanation

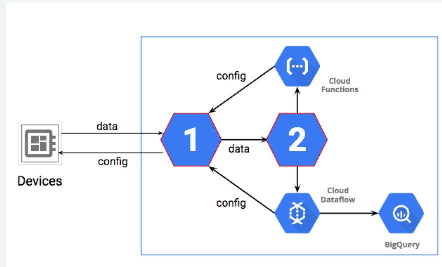
Correct answer is B as the 100% accuracy is an indicator that the validation data may have somehow gotten mixed in with the training data. You will need new validation data to generate recognizable error.

Overfitting results when a model performs well on the training set, generating only a small error, but struggles with new or unknown data. In other words, the model overfits itself to the data. Instead of training a model to pick out general features in a given type of data, an overtrained model learns only how to pick out specific features found in the training set.

Question 10: Correct

A company has a new IoT pipeline. Which services will make this design work?

Select the services that should be used to replace the icons with the number "1" and number "2" in the diagram.



A. Cloud IoT Core, Cloud Datastore

B. Cloud Pub/Sub, Cloud Storage

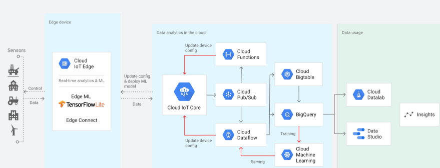
C. Cloud IoT Core, Cloud Pub/Sub (Correct)

D. App Engine, Cloud IoT Core

Explanation

Correct answer is **C** as device data captured by Cloud IoT Core gets published to Cloud Pub/Sub, which can then trigger Dataflow and Cloud Functions.

Refer GCP documentation - Cloud IoT Core



Cloud IoT Core is a fully managed service that allows you to easily and securely connect, manage, and ingest data from millions of globally dispersed devices. Cloud IoT Core, in combination with other services on Cloud IoT platform, provides a complete *solution for collecting, processing, analyzing, and visualizing IoT data in real time to support improved operational efficiency.*

Cloud IoT Core, using Cloud Pub/Sub underneath, can aggregate dispersed device data into a single global system that integrates seamlessly with Google Cloud data analytics services. Use your IoT data stream for advanced analytics, visualizations, machine learning, and more to help improve operational efficiency, anticipate problems, and build rich models that better describe and optimize your business.

Question 11: Correct

You are building storage for files for a data pipeline on Google Cloud. You want to support JSON files. The schema of these files will occasionally change. Your analyst teams will use running aggregate ANSI SQL queries on this data. What should you do?

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

B. Use BigQuery for storage. Select "Automatically detect" in the Schema section. (Correct)

C. Use Cloud Storage for storage. Link data as temporary tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

D. Use Cloud Storage for storage. Link data as permanent tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

Explanation

没说要省钱的话，尽量把
Correct answer is B as the requirement is to support occasionally (schema) changing JSON files and aggregate ANSI SQL queries: you need to use BigQuery, and it is quickest to use 'Automatically detect' for schema changes.

Refer GCP documentation - BigQuery 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 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.

To see the detected schema for a table:

Use the command-line tool's bq show command

Use the BigQuery web UI to view the table's schema When enabled, BigQuery makes a best-effort attempt to automatically infer the schema for CSV and JSON files.

A is not correct because you should not provide format files: you can simply turn on the 'Automatically detect' schema changes flag.

C and D are not correct as Cloud Storage is not ideal for this scenario; it is cumbersome, adds latency and doesn't add value.

Question 12: Correct

You have 250,000 devices which produce a JSON device status event every 10 seconds. You want to capture this event data for outlier time series analysis. What should you do?

A. Ship the data into BigQuery. Develop a custom application that uses the BigQuery API to query the dataset and displays device outlier data based on your business requirements.

B. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.

C. Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements. (Correct)

D. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data based on your business requirements.

Explanation

Correct answer is **C** as the time series data with its data type, volume, and query pattern best fits BigTable capabilities.

Refer GCP documentation - Bigtable Time Series data and CBT

Options A & B are wrong as BigQuery is not suitable for the query pattern in this scenario.

Option D is wrong as you can use the simpler method of 'cbt tool' to support this scenario.

Question 13: Correct

You are building a data pipeline on Google Cloud. You need to select services that will host a deep neural network machine-learning model also hosted on Google Cloud. You also need to monitor and run jobs that could occasionally fail. What should you do?

A. Use Cloud Machine Learning

to host your model. Monitor the status of the Operation object for 'error' results.

B. Use Cloud Machine Learning to host your model. Monitor the status of the Jobs object for 'failed' job states. (Correct)

C. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Jobs object for 'failed' job states.

D. Use a Kubernetes Engine cluster to host your model. Monitor the status of Operation object for 'error' results.

Explanation

Correct answer is B as the requirement is to host an Machine Learning Deep Neural Network job it is ideal to use the Cloud Machine Learning service. Monitoring works on Jobs object.

Refer GCP documentation - [ML Engine Managing Jobs]

You can use [projects.jobs.get] to get the status of a job. This method is also provided as `gcloud ml jobs describe` and in the [Jobs page] in the Google Cloud Platform Console. Regardless of how you get the status, the information is based on the members of the [Job resource]. You'll know the job is complete when `Job.state` in *the response is equal to one of these values:*

`SUCCEEDED`
`FAILED`
`CANCELLED`

Option A is wrong as monitoring should not be on Operation object to monitor failures.

Options C & D are wrong as you should not use a Kubernetes Engine cluster for Machine Learning jobs.

Question 14: Correct

You are developing an application on Google Cloud that will label famous landmarks in users' photos. You are under competitive pressure to develop the predictive model quickly. You need to keep service costs low. What should you do?

A. Build an application that calls the Cloud Vision API. Inspect the generated MID values to supply the image labels.

B. Build an application that calls the Cloud Vision API. Pass landmark locations as base64-encoded strings. (Correct)

C. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Pass landmark locations as base64-encoded strings.

D. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Inspect the generated MID values to supply the image labels.

Explanation

Correct answer is B as the requirement is to quickly develop a model that generates landmark labels from photos, it can be easily supported by Cloud Vision API.

Refer GCP documentation - [Cloud Vision]

Cloud Vision offers both pretrained models via an API and the ability to build custom models using AutoML Vision to provide flexibility depending on your use case.

***Cloud Vision API** enables developers to understand the content of an image by encapsulating powerful machine learning models in an easy-to-use REST API. It quickly classifies images into thousands of categories (such as, "sailboat"), detects individual objects and faces within images, and reads printed words contained within images. You can build metadata on your image catalog, moderate offensive content, or enable new marketing scenarios through image sentiment analysis.*

Option A is wrong as you should not inspect the generated MID values; instead, you should simply pass the image locations to the API and use the labels, which are output.

Options C & D are wrong as you should not build a custom classification TF model for this scenario, as it would require time.

Question 15: Correct

You regularly use prefetch caching with a Data Studio report to visualize the results of BigQuery queries. You want to minimize service costs. What should you do?

A. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and direct the users to view the report only once per business day (24-hour period).

B. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is selected for the report. (Correct)

C. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and also set it up to be a 'view-only' report.

D. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is not selected for the report.

Explanation

Correct option is B as you must set Owner credentials to use the 'enable cache' option in BigQuery. It is also a Google best practice to use the ‘enable cache’ option when the business scenario calls for using prefetch caching.

Refer GCP documentation - Datastudio data caching

The prefetch cache is only active for data sources that use owner's credentials to access the underlying data.

Options A, C, & D are wrong as cache auto-expires every 12 hours; a prefetch cache is only for data sources that use the Owner's credentials and not the Viewer's credentials

Question 16: Correct

Your customer is moving their corporate applications to Google Cloud Platform. The security team wants detailed visibility of all projects in the organization. You provision the Google Cloud Resource

Manager and set up yourself as the org admin. What Google Cloud Identity and Access Management (Cloud IAM) roles should you give to the security team?

A. Org viewer, project owner

B. Org viewer, project viewer (Correct)

C. Org admin, project browser

D. Project owner, network admin

Explanation

Correct answer is B as the security team only needs visibility to the projects, project viewer provides the same with the best practice of least privilege.

Refer GCP documentation - Organization & Project access control

Option A is wrong as project owner will provide access however it does not align with the best practice of least privilege.

Option C is wrong as org admin does not align with the best practice of least privilege.

Option D is wrong as the user needs to be provided organization viewer access to see the organization.

Question 17: Correct

You want to optimize the performance of an accurate, real-time, weather-charting application. The data comes from 50,000 sensors sending 10 readings a second, in the format of a timestamp and sensor reading. Where should you store the data?

A. Google BigQuery

B. Google Cloud SQL

C. Google Cloud Bigtable (Correct)

D. Google Cloud Storage

Explanation

Correct answer is **C** as Bigtable is a ideal solution for storing time series data. *Storing time-series data in Cloud Bigtable is a natural fit. Cloud Bigtable stores data as unstructured columns in rows; each row has a row key, and row keys are sorted lexicographically.*

Refer GCP documentation - Storage Options

Google Cloud Bigtable	A scalable, fully-managed NoSQL wide-column database that is suitable for both real-time access and analytics workloads.	Low-latency read/write access High-throughput analytics Native time series support	IoT, finance, adtech Personalization, recommendations Monitoring Geospatial datasets Graphs

Option A is wrong as Google BigQuery is a scalable, fully-managed Enterprise Data Warehouse (EDW) with SQL and fast response times. It is for analytics and OLAP workload, though it also provides storage capacity and price similar to GCS. It cannot handle the required real time ingestion of data.

Option B is wrong as Google Cloud SQL is a fully-managed MySQL and PostgreSQL relational database service for Structured data and OLTP workloads. It also won't stand for this type of high ingesting rate in real time.

Option D is wrong as Google Cloud Storage is a scalable, fully-managed, highly reliable, and cost-efficient object / blob store. It cannot stand for this amount of data streaming ingestion rate in real-time.

Question 18: Correct

You need to take streaming data from thousands of Internet of Things (IoT) devices, ingest it, run it through a processing pipeline, and store it for analysis. You want to run SQL queries against your data for analysis. What services in which order should you use for this task?

A. Cloud Dataflow, Cloud Pub/Sub, BigQuery

B. Cloud Pub/Sub, Cloud

Dataflow, Cloud Dataproc

C. Cloud Pub/Sub, Cloud Dataflow, BigQuery (Correct)

D. App Engine, Cloud Dataflow, BigQuery

Explanation

Correct answer is **C** as the need to ingest it, transform and store the Cloud Pub/Sub, Cloud Dataflow, BigQuery is ideal stack to handle the IoT data.

Refer GCP documentation - IoT

Google Cloud Pub/Sub provides a globally durable message ingestion service. By creating topics for streams or channels, you can enable different components of your application to subscribe to specific streams of data without needing to construct subscriber-specific channels on each device. Cloud Pub/Sub also natively connects to other Cloud Platform services, helping you to connect ingestion, data pipelines, and storage systems.

Google Cloud Dataflow provides the open Apache Beam programming model as a managed service for processing data in multiple ways, including batch operations, extract-transform-load (ETL) patterns, and continuous, streaming computation. Cloud Dataflow can be particularly useful for managing the high-volume data processing pipelines required for IoT scenarios.

Cloud Dataflow is also designed to integrate seamlessly with the other Cloud Platform services you choose for your pipeline.

Google BigQuery provides a fully managed data warehouse with a familiar SQL interface, so you can store your IoT data alongside any of your other enterprise analytics and logs. The performance and cost of BigQuery means you might keep your valuable data longer, instead of deleting it just to save disk space.

Sample Arch - [Mobile Gaming Analysis Telemetry]

Option A is wrong as the stack is correct, however the order is not correct.

Option B is wrong as Dataproc is not an ideal tool for analysis. Cloud **Dataproc** is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.

Option D is wrong as App Engine is not an ideal ingestion tool to handle IoT data.

Question 19: Correct

Your company is planning the infrastructure for a new large-scale application that will need to store over 100 TB or a petabyte of data in NoSQL format for Low-latency read/write and High-throughput analytics. Which storage option should you use?

A. Cloud Bigtable (Correct)

B. Cloud Spanner

C. Cloud SQL

D. Cloud Datastore

Explanation

Correct answer is A as Bigtable is an ideal solution to provide low latency, high throughput data processing storage option with analytics

Refer GCP documentation - Storage Options

Cloud Bigtable	A scalable, fully managed NoSQL wide-column database that is suitable for both low-latency single-point lookups and precalculated analytics.	Low-latency read/write access High-throughput data processing Time series support	IoT, finance, adtech Personalization, recommendations Monitoring Geospatial datasets Graphs

Options B & C are wrong as they are relational databases

Option D is wrong as Cloud Datastore is not ideal for analytics.

Question 20: Correct

You have hundreds of IoT devices that generate 1 TB of streaming data per day. Due to latency, messages will often be delayed compared to when they were generated. You must be able to account for data arriving late within your processing pipeline. How can the data processing system be designed?

A. Use Cloud SQL to process the delayed messages.

B. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Dataflow to process messages, and use windows, watermarks (timestamp), and triggers to process late data. (Correct)

C. Use SQL queries in BigQuery to analyze data by timestamp.

D. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Pub/Sub to process messages by timestamp and fix out of order issues.

Explanation

Correct answer is B as Cloud Pub/Sub can help handle the streaming data. However, Cloud Pub/Sub does not handle the ordering, which can be done using Dataflow and adding watermarks to the messages from the source.

Refer GCP documentation - Cloud Pub/Sub ordering & Subscriber

How do you assign an order to messages published from different publishers? Either the publishers themselves have to coordinate, or the message delivery service itself has to attach a notion of order to every incoming message. Each message would need to include the ordering information. The order information could be a timestamp (though it has to be a timestamp that all servers get from the same source in order to avoid issues of clock drift), or a sequence number (acquired from a single source with ACID guarantees). Other messaging systems that guarantee ordering of messages require settings that effectively limit the system to multiple publishers sending messages through a single server to a single subscriber.

Typically, Cloud Pub/Sub delivers each message once and in the order in which it was published. However, messages may sometimes be delivered out of order or more than once. In general, accommodating more-than-once delivery requires your subscriber to be idempotent when processing messages. You can achieve exactly once processing of Cloud Pub/Sub message streams using Cloud Dataflow `PubsubIO`. `PubsubIO` *de-duplicates messages on custom message identifiers or* *those assigned by Cloud Pub/Sub. You can also achieve ordered processing with Cloud Dataflow by using the standard sorting APIs of the service. Alternatively, to achieve ordering, the publisher of the topic to which you subscribe can include a sequence token in the message.*

Options A & C are wrong as SQL and BigQuery do not support ingestion and ordering of IoT data and would need other services like Pub/Sub.

Option D is wrong as Cloud Pub/Sub does not perform ordering of messages.

Question 21: Correct

Your company has data stored in BigQuery in Avro format. You need to export this Avro formatted data from BigQuery into Cloud Storage. What is the best method of doing so from the web console?

A. Convert the data to CSV format the BigQuery export options, then make the transfer.

B. Use the BigQuery Transfer Service to transfer Avro data to Cloud Storage.

C. Click on Export Table in BigQuery, and provide the Cloud Storage (Correct)

location to export
to

D. Create a Dataflow job to manage the conversion of Avro data to CSV format, then export to Cloud Storage.

Explanation

Correct answer is **C** as BigQuery can export Avro data natively to Cloud Storage.

Refer GCP documentation - BigQuery Exporting Data

After 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. When you export your data to multiple files, the size of the files will vary.

You cannot export data to a local file or to Google Drive, but you can save query results to a local file. The only supported export location is Google Cloud Storage.

*For **Export format**, choose the format for your exported data: CSV, JSON (Newline Delimited), or Avro.*

Option A is wrong as BigQuery can export Avro data natively to Cloud Storage and does not need to be converted to CSV format.

Option B is wrong as BigQuery Transfer Service is for moving BigQuery data to Google SaaS applications (AdWords, DoubleClick, etc.). You will want to do a normal export of data, which works with Avro formatted data.

Option D is wrong as Google Cloud Dataflow can be used to read data from BigQuery instead of manually exporting it, but doesn't work through console.

Question 22: Correct

Your company has its input data hosted in BigQuery. They have existing Spark scripts for performing analysis which they want to reuse. The output needs to be stored in BigQuery for future analysis. How can you set up your Dataproc environment to use BigQuery as an input and output source?

A. Use the Bigtable syncing service built into Dataproc.

B. Manually use a Cloud Storage bucket to import and export to and from both BigQuery and Dataproc

C. Install the BigQuery connector on your Dataproc cluster (Correct)

D. You can only use Cloud Storage or HDFS for your Dataproc input and output.

Explanation

Correct answer is **C** as Dataproc has a BigQuery connector library which allows it directly interface with BigQuery.

Refer GCP documentation - [Dataproc BigQuery Connector]()

You can use a BigQuery connector to enable programmatic read/write access to BigQuery. This is an ideal way to process data that is stored in BigQuery. No command-line access is exposed. The BigQuery connector is a Java library that enables Hadoop to process data from BigQuery using abstracted versions of the Apache Hadoop InputFormat and OutputFormat classes.

Option A is wrong Bigtable syncing service does not exist.

Options B & D are wrong as Dataproc can directly interface with BigQuery.

Question 23: Correct

You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

A. Include ORDER BY DESK on timestamp column and LIMIT to 1.

B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.

C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.

D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1. (Correct)

Explanation

Correct answer is **D** as the best approach is to ROW_NUMBER with PARTITION by the UNIQUE_ID and filter it by row_number = 1.

Refer GCP documentation - [BigQuery Streaming Data - Removing Duplicates]

To remove duplicates, perform the following query. You should specify a destination table, allow large results, and disable result flattening.

```
#standardSQL SELECT * EXCEPT(row_number) FROM ( SELECT *, ROW_NUMBER() OVER (PARTITION BY ID_COLUMN) row_number FROM `TABLE_NAME`) WHERE row_number = 1
```

Question 24: Correct

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data. Which three steps should you take? (Choose three)

A. Load data into different partitions.

B. Load data into a different dataset for each client. (Correct)

C. Put each client's BigQuery dataset into a different table.

D. Restrict a client's dataset to approved users. (Correct)

E. Only allow a service account to access the datasets.

F. Use the appropriate identity and access management (IAM) roles for each client's users. (Correct)

Explanation

Correct answers are B, D & F.
As the access control can be done using IAM roles on the

dataset only to the specific approved users.

Refer GCP documentation - [BigQuery Access Control]

BigQuery uses Identity and Access Management (IAM) to manage access to resources. The three types of resources available in BigQuery are organizations, projects, and datasets. In the IAM policy hierarchy, datasets are child resources of projects. Tables and views are child resources of datasets — they inherit permissions from their parent dataset.

To grant access to a resource, assign one or more roles to a user, group, or service account. Organization and project roles affect the ability to run jobs or manage the project's resources, whereas dataset roles affect the ability to access or modify the data inside of a particular dataset.

Options A & C are wrong as the access control can only be applied on dataset and views, not on partitions and tables.

Option E is wrong as service account is mainly for machines and would be a single account.

Question 25: Correct

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a

Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks. What should you do?

A. Run a local version of Jupiter on the laptop.

B. Grant the user access to Google Cloud Shell.

C. Host a visualization tool on a VM on Google Compute Engine.

D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine. (Correct)

Explanation

Correct answer is D as Cloud Datalab provides a powerful interactive, scalable tool on Google Cloud with the ability to analyze, visualize data.

Refer GCP documentation - Datalab

Cloud Datalab is a powerful interactive tool created to explore, analyze, transform and visualize data and build machine learning models on Google Cloud Platform. It runs on Google Compute Engine and connects to multiple cloud *services easily so you can focus on your data science tasks.*

Cloud Datalab is built on Jupyter (formerly IPython), which boasts a thriving ecosystem of modules and a robust knowledge base. Cloud Datalab enables analysis of your data on Google BigQuery, Cloud Machine Learning Engine, Google Compute Engine, and Google Cloud Storage using Python, SQL, and JavaScript (for BigQuery user-defined functions).

Whether you're analyzing megabytes or terabytes, Cloud Datalab has you covered. Query terabytes of data in BigQuery, run local analysis on sampled data and run training jobs on terabytes of data in Cloud Machine Learning Engine seamlessly.

Use Cloud Datalab to gain insight from your data. Interactively explore, transform, analyze, and visualize your data using BigQuery, Cloud Storage and Python.

Go from data to deployed machine-learning (ML) models ready for prediction. Explore data, build, evaluate and optimize Machine Learning models using TensorFlow or Cloud Machine Learning Engine.

Options A, B & C do not provides all the abilities.

Question 26: Correct

You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

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

B. Grant the consultant the Cloud Dataflow Developer role on the project. (Correct)

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

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

Explanation

Correct answer is B as the Dataflow developer role would help provide the third-party consultant access to create and work on the Dataflow pipeline. However, it does not provide access to view the data, thus maintaining user's privacy.

Refer GCP documentation - Dataflow roles

`roles/dataflow.viewer`	`dataflow.<resource-type>.list` `dataflow.<resource-type>.get`	jobs, messages, metrics
`roles/dataflow.developer`	All of the above, as well as: `dataflow.jobs.create` `dataflow.jobs.drain` `dataflow.jobs.cancel`	jobs
`roles/dataflow.admin`	All of the above, as well as: `compute.machineTypes.get` `storage.buckets.get` `storage.objects.create` `storage.objects.get` `storage.objects.list`	NA

Option A is wrong as it would not allow the consultant to work on the pipeline.

Option C is wrong as the consultant cannot use the service account to login.

Option D is wrong as it does not enable collabaration.

Question 27: Correct

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?

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

B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output. (Correct)

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

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

Explanation

Correct answer is B as the issue can be debugged by running a fixed dataset and checking the output.

Refer GCP documentation - Dataflow logging

Option A is wrong as the Dashboard uses data provided by Dataflow, the input source for Dashboard seems to be the issue

Option C is wrong as Monitoring would not help find missing messages in Cloud Pub/Sub.

Option D is wrong as Dataflow cannot be configured as Push endpoint with Cloud Pub/Sub.

Question 28: Correct

Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three)

A. Disable writes to certain tables.

B. Restrict access to tables by role.

C. Ensure that the data is encrypted at all times.

D. Restrict BigQuery API access to approved users. (Correct)

E. Segregate data across multiple tables or datasets. (Correct)

F. Use Google Stackdriver Audit Logging to determine policy violations. (Correct)

Explanation

Correct answers are **D, E & F**

Option D would help limit access to approved users only.

Option E as it would help segregate the data with the ability to provide access to users as per their needs.

Option F as it would help in auditing.

Refer GCP documentation - [BigQuery Dataset Access Control](#) & [Access Control](#)

You share access to BigQuery tables and views using project-level IAM roles and [dataset-level access controls](#). Currently, you cannot apply access controls directly to tables or views.

Project-level access controls determine the users, groups, and service accounts allowed to access all datasets, tables, views, and table data within a project. Dataset-level access controls determine the users, groups, and service accounts allowed to access the tables, views, and table data in a specific dataset.

Option A is wrong as disabiling writes does not prevent the users from reading and does not align with the least privilege principle.

Option B is wrong as access cannot be control on tables.

Option C is wrong as data is encrypted by default, however it does not align with the least privilege principle.

Question 29: Correct

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an

update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

A. Update the current pipeline and use the drain flag. (Correct)

B. Update the current pipeline and provide the transform mapping JSON object.

C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.

D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Explanation

Correct answer is A as the key requirement is not to lose the data, the Dataflow pipeline can be stopped using the Drain option. Drain options would cause Dataflow to stop any new processing, but would also allow the existing processing to complete

Refer GCP documentation - Dataflow Stopping a Pipeline

*Using the **Drain** option to stop your job tells the Cloud Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources. However, the Cloud Dataflow* *service will preserve any existing resources, such as worker instances, to finish processing and writing any buffered data in your pipeline. When all pending processing and write operations are complete, the Cloud Dataflow service will clean up the GCP resources associated with your job.*

Note: *Your pipeline will continue to incur the cost of maintaining any associated GCP resources until all processing and writing has completed.*

Use the Drain option to stop your job if you want to prevent data loss as you bring down your pipeline.

Effects of draining a job

*When you issue the Drain command, Cloud Dataflow immediately closes any in-process windows and fires all triggers. The system **does not** wait for any outstanding time-based windows to finish. For example, if your pipeline is ten minutes into a two-hour window when you issue the Drain command, Cloud Dataflow won't wait for the remainder of the window to finish. It will close the window immediately with partial results.*

Question 30: Correct

A client has been developing a pipeline based on

PCollections using local programming techniques and is ready to scale up to production. What should they do?

A. They should use the Cloud Dataflow Cloud Runner. (Correct)

B. They should upload the pipeline to Cloud Dataproc.

C. They should use the local version of runner.

D. Import the pipeline into BigQuery.

Explanation

Correct answer is A as the PCollection indicates it is a Cloud Dataflow pipeline. And the Cloud Runner will enable the pipeline to scale to production levels.

Refer documentation - Dataflow Cloud Runner

The Google Cloud Dataflow Runner uses the Cloud Dataflow managed service. When you run your pipeline with the Cloud Dataflow service, the runner uploads your executable code and dependencies to a Google Cloud Storage bucket and creates a Cloud Dataflow job, which executes your pipeline on managed resources in Google Cloud Platform.

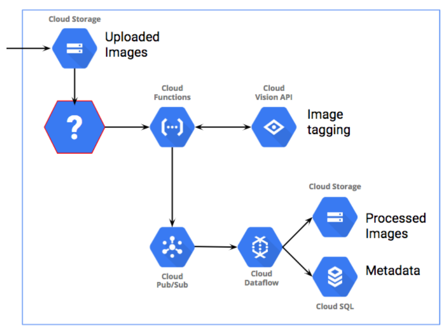
The Cloud Dataflow Runner and service are suitable for large scale, continuous jobs, and provide:

a fully managed service
autoscaling of the number of workers throughout the lifetime of the job
dynamic work rebalancing
Options B & D are wrong as PCollections are related to Dataflow

Option C is wrong as Local runner is execute the pipeline locally.

Question 31: Correct

A company is building an image tagging pipeline. Which service should be used in the icon with the question mark in the diagram?



A. Cloud Datastore

B. Cloud Dataflow

C. Cloud Pub/Sub (Correct)

D. Cloud Bigtable

Explanation

Correct answer is **C** as Cloud Storage upload events can push Cloud Pub/Sub to trigger a Cloud Function to ingest and process the image.

Refer GCP documentation
- [Cloud Storage Pub/Sub Notifications]

Cloud Pub/Sub Notifications sends information about changes to objects in your buckets to Cloud Pub/Sub, where the information is added to a Cloud Pub/Sub topic of your choice in the form of messages. For example, you can track objects that are created and deleted in your bucket. Each notification contains information describing both the event that triggered it and the object that changed.

Cloud Pub/Sub Notifications are the recommended way to track changes to objects in your Cloud Storage buckets because they're faster, more flexible, easier to set up, and more cost-effective.

Options A, B & D are wrong as they cannot be configured for notifications from Cloud Storage.

Question 32: Correct

Your company is in a highly regulated industry. One of your requirements is to ensure external users have access only to the non PII fields information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which access control method would you use?

A. Use Primitive role on the dataset

B. Use Predefined role on the dataset

C. Use Authorized view with the same dataset with proper permissions

D. Use Authorized view with the different dataset with proper permissions (Correct)

Explanation

Correct answer is D as the controlled access can be granted using Authorized view. The Authorized view needs to be in a different dataset than the source.

Refer GCP documentation - BigQuery Authorized Views

Giving a view access to a dataset is also known as creating an authorized view in BigQuery. An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query.

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, *your users would have access to both the view and the data.*

Options A, B & C are wrong as they would provide access to the complete datasets with the source included.

Question 33: Correct

Your company is developing a next generation pet collar that collects biometric information to assist potential millions of families with promoting healthy lifestyles for their pets. Each collar will push 30kb of biometric data In JSON format every 2 seconds to a collection platform that will process and analyze the data providing health trending information back to the pet owners and veterinarians via a web portal. Management has tasked you to architect the collection platform ensuring the following requirements are met.

1. Provide the ability for real-time analytics of the inbound biometric data

2. Ensure processing of the biometric data is highly durable, elastic and parallel

3. The results of the analytic processing should be persisted for data mining

Which architecture outlined below win meet the initial requirements for the platform?

A. Utilize Cloud Storage to

collect the inbound sensor data, analyze data with Dataproc and save the results to BigQuery.

B. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to BigQuery. (Correct)

C. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Cloud SQL.

D. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Bigtable.

Explanation

Correct answer is B as Cloud Pub/Sub provides elastic and scalable ingestion, Dataflow provides processing and BigQuery analytics.

Refer GCP documentation - IoT

Google Cloud Pub/Sub provides a globally durable message ingestion service. By creating topics for streams or channels, you can enable different components of your application to subscribe to specific streams of data without needing to construct subscriber-specific channels on each device. Cloud Pub/Sub also natively connects to other Cloud Platform services, *helping you to connect ingestion, data pipelines, and storage systems.*

Google Cloud Dataflow provides the open Apache Beam programming model as a managed service for processing data in multiple ways, including batch operations, extract-transform-load (ETL) patterns, and continuous, streaming computation. Cloud Dataflow can be particularly useful for managing the high-volume data processing pipelines required for IoT scenarios. Cloud Dataflow is also designed to integrate seamlessly with the other Cloud Platform services you choose for your pipeline.

Google BigQuery provides a fully managed data warehouse with a familiar SQL interface, so you can store your IoT data alongside any of your other enterprise analytics and logs. The performance and cost of BigQuery means you might keep your valuable data longer, instead of deleting it just to save disk space.

Option A is wrong as Cloud Storage is not an ideal ingestion service for real time high frequency data. Also Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.

Option C is wrong as Cloud SQL is a relational database and not suited for analytics data storage.

Option D is wrong as Bigtable is not ideal for long term analytics data storage.

Question 34: Correct

Which of the following statements about the Wide & Deep Learning model are true? (Choose two)

A. Wide model is used for memorization, while the deep model is used for generalization. (Correct)

B. Wide model is used for generalization, while the deep model is used for memorization.

C. A good use for the wide and deep model is a recommender system. (Correct)

D. A good use for the wide and deep model is a small-scale linear regression problem.

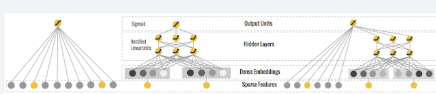
Explanation

Correct answers are A & C as Wide learning model is good for memorization and a Deep learning model is generalization. Both Wide and Deep learning model can help build good recommendation engine.

Refer Google blog - Wide Deep learning together

The human brain is a sophisticated learning machine, forming rules by memorizing everyday events (“sparrows can fly” and “pigeons can fly”) and generalizing those learnings to apply to things we haven't seen before (“animals with wings can fly”). Perhaps more powerfully, memorization also allows us to further refine our generalized rules with exceptions (“penguins can't fly”). As we were exploring how to advance machine intelligence, we asked ourselves the question—can we teach computers to learn like humans do, by combining the power of memorization and generalization?

It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.



Question 35: Correct

A financial organization wishes to develop a global application to store transactions happening from different part of the world. The storage system must provide low latency transaction support and horizontal scaling. Which GCP service is appropriate for this use case?

A. Bigtable

B. Datastore

C. Cloud Storage

D. Cloud Spanner (Correct)

Explanation

Correct answer is D as Spanner provides Global scale, low latency and the ability to scale horizontally.

Refer GCP documentation - Storage Options

Cloud Spanner	Mission-critical, relational database service with transactional consistency, global scale, and high availability.	Mission-critical applications High transactions Scale + consistency requirements	Adtech Financial services Global supply chain Retail

Question 36: Correct

A retailer has 1PB of historical purchase dataset, which is largely unlabeled. They want to categorize the customer into different groups as per their spend. Which type of Machine Learning algorithm is suited to achieve this?

A. Classification

B. Regression

C. Association

D. Clustering (Correct)

Explanation

Correct answer is D as the data is unlabelled, unsupervised learning technique of Clustering can be applied to categorize the data.

Refer GCP documentation - Machine Learning

In unsupervised learning, the goal is to identify meaningful patterns in the data. To accomplish this, the machine must learn from an unlabeled data set. In other words, the model has no hints how to categorize each piece of data and must infer its own rules for doing so.

Options A & B are wrong as they are supervised learning techniques.

In supervised machine learning, you feed the features and their corresponding labels *into an algorithm in a process called [training](). During training, the algorithm gradually determines the relationship between features and their corresponding labels. This relationship is called the [model](). Often times in machine learning, the model is very complex.*

Option C is wrong as Association rules is mainly to identify relationship.

Question 37: Correct

Your company wants to host confidential documents in Cloud Storage. Due to compliance requirements, there is a need for the data to be highly available and resilient even in case of a regional outage. Which storage classes help meet the requirement? (Select THREE)

A. Nearline (Correct)

B. Standard (Correct)

C. Multi-Regional (Correct)

D. Dual-Regional

E. Regional

Explanation

Correct answers are **A, B & C** as Standard, Multi-Regional and Nearline storage classes provide multi-region geo-redundant deployment, which can sustain regional failure.

Update - There have been several changes in GCP storage classes. Standard Storage was newly introduced by Google Cloud with multi-regional capability. GCP supports now Standard, Nearline and Coldline storage classes. Multi-regional is only available, if you are already using it.

Circa Aug 14, 2019

Multi-Regional Storage and Regional Storage are now Standard Storage.

Combining these into a single Standard Storage class separates your storage class considerations from your location considerations.

Before that **Circa Oct 16, 2016** - Standard Storage class was changed.

Standard Storage class is now Multi-Regional Storage and Regional Storage.

The Multi-Regional Storage class provides the same price and performance along with geo-redundant copies of your data and a 99.95% availability SLA.

The Regional Storage class provides the same performance at a reduced price.

Refer GCP documentation - Cloud Storage Classes

Multi-Regional Storage is geo-redundant.

The geo-redundancy of Nearline Storage data is *determined by the type of location in which it is stored: Nearline Storage data stored in multi-regional locations is redundant across multiple regions, providing higher availability than Nearline Storage data stored in regional locations.*

Data that is geo-redundant is stored redundantly in at least two separate geographic places separated by at least 100 miles. Objects stored in multi-regional locations are geo-redundant, regardless of their storage class.

Geo-redundancy occurs asynchronously, but all Cloud Storage data is redundant within at least one geographic place as soon as you upload it.

Geo-redundancy ensures maximum availability of your data, even in the event of large-scale disruptions, such as natural disasters. For a dual-regional location, geo-redundancy is achieved using two specific regional locations. For other multi-regional locations, geo-redundancy is achieved using any combination of data centers within the specified multi-region, which may include data centers that are not explicitly available as regional locations.

Option D is wrong as dual-regional storage class does not exist.

Option E is wrong as Regional storage class is not geo-redundant. Data stored in a narrow geographic region and Redundancy is across availability zones

Question 38: Correct

Your company wants to develop an REST based application for image analysis. This application would help detect individual objects and faces within images, and reads printed words contained within images. You need to do a quick Proof of Concept (PoC) to implement and demo the same. How would you design your application?

A. Create and Train a model using Tensorflow and Develop an REST based wrapper over it

B. Use Cloud Image Intelligence API and Develop an REST based wrapper over it

C. Use Cloud Natural Language API and Develop an REST based wrapper over it

D. Use Cloud Vision API and Develop an REST based wrapper over it (Correct)

Explanation

Correct answer is D as Cloud Vision API provide pre-built models to identify and detect objects and faces within images.

Refer GCP documentation - [AI Products]

Cloud Vision API enables you to derive insight from your images with our powerful pretrained API models or easily train custom vision models with AutoML Vision Beta. The API quickly classifies images into thousands of categories (such as "sailboat" or "Eiffel Tower"), detects individual objects and faces within images, and finds and reads printed words contained within images. AutoML Vision lets you build and train custom ML models with minimal ML expertise to meet domain-specific business needs.

Question 39: Correct

Your company is developing an online video hosting platform. Users can upload their videos, which would be available for all the other users to view and share. As a compliance requirement, the videos need to undergo content moderation before it is available for all the users. How would you design your application?

A. Use Cloud Vision API to identify video with inappropriate content and mark it for manual checks.

B. Use Cloud Natural Language API to identify video with inappropriate content and mark it for manual checks.

C. Use Cloud Speech-to-Text

API to identify video with inappropriate content and mark it for manual checks.

D. Use Cloud Video Intelligence API to identify video with inappropriate content and mark it for manual checks. (Correct)

Explanation

Correct answer is D as Cloud Video Intelligence can be used to perform content moderation.

Refer GCP documentation - [Cloud Video Intelligence]()

Google Cloud Video Intelligence makes videos searchable, and discoverable, by extracting metadata with an easy to use REST API. You can now search every moment of every video file in your catalog. It quickly annotates videos stored in Google Cloud Storage, and helps you identify key entities (nouns) within your video; and when they occur within the video. Separate signals from noise, by retrieving relevant information within the entire video, shot-by-shot, -or per frame.

Identify when inappropriate content is being shown in a given video. You can instantly conduct content moderation across petabytes of data and more quickly and efficiently filter your content or user-generated content.

Option A is wrong as Vision is for image analysis.

Option B is wrong as Natural Language is for text analysis

Option C is wrong as Speech-to-Text is for audio to text conversion.

Question 40: Correct

Your company has a variety of data processing jobs. Dataflow jobs to process real time streaming data using Pub/Sub. Data pipelines working with on-premises data. Dataproc spark batch jobs running weekly analytics with Cloud Storage. They want a single interface to manage and monitor the jobs. Which service would help implement a common monitoring and execution platform?

A. Cloud Scheduler

B. Cloud Composer (Correct)

C. Cloud Spanner

D. Cloud Pipeline

Explanation

Correct answer is B as Cloud Composer's managed nature allows you to focus on authoring, scheduling, and monitoring your workflows as opposed to provisioning resources.

Refer GCP documentation
- Cloud Composer

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers. Built on the popular Apache Airflow open source project and operated using the Python programming language, Cloud Composer is free from lock-in and easy to use.

Cloud Composer's managed nature allows you to focus on authoring, scheduling, and monitoring your workflows as opposed to provisioning resources.

Option A is wrong as Cloud Scheduler is a fully managed enterprise-grade cron job scheduler. It is not an multi-cloud orchestration tool.

Option C is wrong as Google Cloud Spanner is relational database

Option D is wrong as Google Cloud Pipeline service does not exist.

Question 41: Correct

Your company hosts its analytical data in a BigQuery dataset for analytics. They need to provide controlled access to certain tables and columns within the tables to a third party. How do you design the access with least privilege?

A. Grant only DATA VIEWER access to the third party team

B. Grant fine grained DATA VIEWER access to the tables and columns within the dataset

C. Create Authorized views for tables in a same project and grant access to the teams

D. Create Authorized views for tables in a separate project and grant access to the teams (Correct)

Explanation

Correct answer is D as the controlled access can be provided using Authorized views created in a separate project.

Refer GCP documentation - BigQuery Authorized View

BigQuery is a petabyte-scale analytics data warehouse that you can use to run SQL queries over vast amounts of data in near realtime.

Giving a view access to a dataset is also known as creating an authorized view in BigQuery. An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query.

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, your data analysts would have access to both the view and the data.

Options A & B are wrong as access cannot be controlled over table, but only projects and datasets.

Option C is wrong as Authorized views should be created in a separate project. If they are created in the same project, the users would have access to the underlying tables as well.

Question 42: Correct

Your company is hosting its analytics data in BigQuery. All the Data analysts have been provided with the IAM owner role to their respective projects. As a compliance requirement, all the data access logs needs to be captured for audits. Also, the access to the logs needs to be limited to the Auditor team only. How can the access be controlled?

A. Export the data access logs using aggregated sink to Cloud Storage in an existing project and grant VIEWER access to the project to the Auditor team

B. Export the data access logs using project sink to BigQuery in an existing project and grant VIEWER access to the project to the Auditor team

C. Export the data access logs using project sink to Cloud Storage in a separate project and grant VIEWER access to the project to the Auditor team

D. Export the data access logs using aggregated sink to Cloud Storage in a separate project and grant VIEWER access to the project to the Auditor team (Correct)

Explanation

Correct answer is D as the Data Analysts have OWNER roles to the projects, the logs need to be exported to a separate project which only the Auditor team has access to. Also, as there are multiple projects aggregated export sink can be used to export data access logs from all projects.

Refer GCP documentation - BigQuery Auditing and Aggregated Exports

You can create an aggregated export sink that can export log entries from all the projects, folders, and billing accounts of an organization. As an example, you might use this feature to export audit log entries from an organization's projects to a central location.

Options A & B are wrong as the export needs to be in separate project.

Option C is wrong as you need to use aggregated sink instead of project sink, as it would capture logs from all projects.

Question 43: Correct

Your company is building an aggregator, which receives feed from lot of other external data sources and companies. These dataset contain invalid & erroneous records, which need to be discarded. Your Data analysts should be able to perform the same without any programming or SQL knowledge. Which solution best fits the requirement?

A. Dataflow

B. Dataproc

C. Hadoop installation on Compute Engine

D. Dataprep (Correct)

Explanation

Correct answer is D as Dataprep provides the ability to detect, clean and transform data through a Graphical Interface without any programming knowledge.

Refer GCP documentation - Dataprep

Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.

Cloud Dataprep automatically detects schemas, datatypes, possible joins, and anomalies such as missing values, outliers, and duplicates so you get to skip the time-consuming work of profiling your data and go right to the data analysis.

Cloud Dataprep automatically identifies data anomalies and helps you to take corrective action fast. Get data transformation suggestions based on your usage pattern. Standardize, structure, and join datasets easily with a guided approach.

Options A, B & C are wrong as they all need programming knowledge.

Question 44: Correct

Your company is migrating to the Google cloud and looking for HBase alternative. Current solution uses a lot of custom code using the observer coprocessor. You are required to find the best alternative for migration while using managed services, is possible?

A. Dataflow

B. HBase on Dataproc (Correct)

C. Bigtable

D. BigQuery

Explanation

Correct answer is **B** as Bigtable is an HBase managed service alternative on Google Cloud. However, it does not support Coprocessors. So the best solution is to use HBase with Dataproc which can be installed using initialization actions.

Refer GCP documentation - [Bigtable HBase differences]

Coprocessors are not supported. You cannot create classes that implement the interface`org.apache.hadoop.hbase.coprocessor`.

Options A & D are wrong as Dataflow and BigQuery are not HBase alternative

Option C is wrong as Bigtable does not support Coprocessors.

Question 45: Correct

You have multiple Data Analysts who work with the dataset hosted in BigQuery within the same project. As a BigQuery Administrator, you are required to grant the data analyst only the privilege to create jobs/queries and an ability to cancel self-submitted jobs. Which role should assign to the user?

A. User

B. Jobuser (Correct)

C. Owner

D. Viewer

Explanation

Correct answer is B as JobUser access grants users permissions to run jobs and cancel their own jobs within the same project

Refer GCP documentation - BigQuery Access Control

`roles/bigquery.jobUser`	Permissions to run jobs, including queries, within the project. The jobUser role can get information about their own jobs and cancel their own jobs. Rationale: This role allows the separation of data access from the ability to run work in the project, which is useful when team

	members query data from multiple projects. This role does not allow access to any BigQuery data. If data access is required, grant dataset-level access controls. Resource Types: Organization Project

Option A is wrong as User would allow to run queries across projects.

Option C is wrong as Owner would give more privileges to the users

Option D is wrong as Viewer does not give user permissions to run jobs.

Question 46: Correct

You need to design a real time streaming data processing pipeline. The pipeline needs to read data from Cloud Pub/Sub, enrich it using Static reference data in BigQuery, transform it and store the results back in BigQuery for further analytics. How would you design the pipeline?

A. Dataflow, BigQueryIO and PubSubIO, SideOutputs

B. Dataflow, BigQueryIO and PubSubIO, SideInputs (Correct)

C. DataProc, BigQueryIO and PubSubIO, SideInputs

D. DataProc, BigQueryIO and PubSubIO, SideOutputs

Explanation

Correct answer is B as Dataflow is needed for real time streaming pipeline with the ability to enrich and transform using SideInputs. BigQueryIO and PubSubIO to interact with BigQuery and Pub/Sub.

Refer GCP documentation - Dataflow Use Case Patterns

In streaming mode, lookup tables need to be accessible by your pipeline. If the lookup table never changes, then the standard Cloud Dataflow `SideInput` *pattern reading from a bounded source such as BigQuery is a perfect fit. However, if the lookup data changes over time, in streaming mode there are additional considerations and options. The pattern described here focuses on slowly-changing data — for example, a table that's updated daily rather than every few hours.*

Options C & D are wrong as Dataproc is not ideal for handling real time streaming data.

Options A & D are wrong as the lookup tables can be referred using SideInputs.

Question 47: Correct

You are interacting with a Point Of Sale (PoS) terminal, which sends the transaction details only. Due to latest software update a bug was introduced in the terminal software that caused it to send individual PII and card details. As a security measure, you are required to implement a quick solution to prevent access to the PII. How would you design the solution?

A. Train Model using Tensorflow to identify PII and filter the information

B. Store the data in BigQuery and create a Authorized view for the users

C. Use Data Loss Prevention APIs to identify the PII information and filter the information (Correct)

D. Use Cloud Natural Language API to identify PII and filter the information

Explanation

Correct answer is **C** as Data Loss Prevention APIs can be used to quickly redact the sensitive information.

Refer GCP documentation - Cloud DLP

Cloud DLP helps you better understand and manage sensitive data. It provides fast, scalable classification and redaction for sensitive data elements like credit card numbers, names, social security numbers, US and selected international identifier numbers, phone numbers and GCP credentials. Cloud DLP classifies this data using more than 90 predefined detectors to identify patterns, formats, and checksums, and even understands contextual clues. You can optionally redact data as well using techniques like masking, secure hashing, bucketing, and format-preserving encryption.

Option A is wrong as building and training a model is not a quick and easy solution.

Option B is wrong as the data would still be stored in the base tables and accessible.

Option D is wrong as Cloud Natural APIs is for text analysis and does not handle sensitive information redaction.

Question 48: Correct

You are designing a relational data repository on Google Cloud to grow as needed. The data will be transactionally consistent and added from any location in the world. You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?

A. Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.

B. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span. (Correct)

C. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.

D. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

Explanation

Correct answer is B as the requirement is to support relational data service with transactionally consistently and globally scalable transactions, Cloud Spanner is an ideal choice. CPU utilization is the recommended metric for scaling, per Google best practices, linked below.

Refer GCP documentation -

Storage Options @ https://cloud.google.com/storage-

options/ & Spanner Monitoring @ https://cloud.google.com/spanner/docs/monitoring

Option A is wrong as storage utilization is not a correct scaling metric for load.

Options C & D are wrong Bigtable is regional and not a relational data service.

Question 49: Correct

You are working on a project with two compliance requirements. The first requirement states that your developers should be able to see the Google Cloud Platform billing charges for only their own projects. The second requirement states that your finance team members can set budgets and view the current charges for all projects in the organization. The finance team should not be able to view the project contents. You want to set permissions. What should you do?

A. Add the finance team members to the default IAM Owner role. Add the developers to a custom role that allows them to see their own spend only.

B. Add the finance team members to the Billing Administrator role for each of the billing accounts (Correct) that they need to manage. Add the developers to the Viewer role for the Project.

C. Add the developers and finance managers to the Viewer role for the Project.

D. Add the finance team to the Viewer role for the Project. Add the developers to the Security Reviewer role for each of the billing accounts.

Explanation

Correct answer is **B** as there are 2 requirements, Finance team able to set budgets on project but not view project contents and developers able to only view billing charges of their projects. Finance with Billing Administrator role can set budgets and Developer with viewer role can view billing charges aligning with the principle of least privileges.

Refer GCP documentation - IAM Billing @ https://cloud.google.com/iam/docs/job-functions/billing

Option A is wrong as GCP recommends using pre-defined roles instead of using primitive roles and custom roles.

Option C is wrong as viewer role to finance would not provide them the ability to set budgets.

Option D is wrong as viewer role to finance would not provide them the ability to set budgets. Also, Security

Reviewer role enables the ability to view custom roles but not administer them for the developers which they don't need.

Question 50: Correct

Your customer wants to capture multiple GBs of aggregate real-time key performance indicators (KPIs) from their game servers running on Google Cloud Platform and monitor the KPIs with low latency. How should they capture the KPIs?

A. Output custom metrics to Stackdriver from the game servers, and create a Dashboard in Stackdriver Monitoring Console to view them.

B. Schedule BigQuery load jobs to ingest analytics files uploaded to Cloud Storage every ten minutes, and visualize the results in Google Data Studio.

C. Store time-series data from the game servers in Google Bigtable, and view it using Google Data Studio. (Correct)

D. Insert the KPIs into Cloud Datastore entities, and run ad hoc analysis and visualizations of them in Cloud Datalab.

Explanation

Correct answer is **C** as Bigtable is an ideal solution for storing time series data with the ability to provide analytics at real time at a very low latency. Data can be viewed using Google Data Studio.

Refer GCP documentation - Data lifecycle @ https://cloud.google.com/solutions/data-lifecycle-cloud-platform

Cloud Bigtable is a managed, high-performance NoSQL database service designed for terabyte- to petabyte-scale workloads. Cloud Bigtable is built on Google's internal Cloud Bigtable database infrastructure that powers Google Search, Google Analytics, Google Maps, and Gmail. The service provides consistent, low-latency, and high-throughput storage for large-scale NoSQL data. Cloud Bigtable is built for real-time app serving workloads, as well as large-scale analytical workloads.

Cloud Bigtable schemas use a single-indexed row key associated with a series of columns; schemas are usually structured either as tall or wide and queries are based on row key. The style of schema is dependent on the downstream use cases and it's important to consider data locality and distribution of reads and writes to maximize performance. Tall schemas are often used for storing time-series events, data that is keyed in some portion by a timestamp, with relatively fewer columns per row. Wide *schemas follow the opposite approach, a simplistic identifier as the row key along with a large number of columns*

Option A is wrong as Stackdriver is not an ideal solution for time series data and it does not provide analytics capability.

Option B is wrong as BigQuery does not provide low latency access and with jobs scheduled at every 10 minutes does not meet the real time criteria.

Option D is wrong as Datastore does not provide analytics capability.

Google Cloud Certified - Professional Data Engineer Practice Exam 4 - Results

Attempt 3

Question 1: **Incorrect**

A company has its data stored within a single project acme-

company-project. Users across teams need to be able to access various tables within that dataset. Each team has a separate project acme-company-team-00x created. How can the access be control while billing only the team querying the dataset?

A. Create Authorized views for tables required by the team in their respective project. Grant BigQuery User role for acme-company-team-00x and data viewer role to acme-company-project dataset

B. Create Authorized views for tables required by the team in their respective project. Grant BigQuery User role for acme-company-team-00x and data viewer role to acme-company-team- **(Correct)**

00x dataset

C. Create Authorized views for tables required by the team in their respective project. Grant BigQuery JobUser role for acme-company-team-00x and data viewer role to acme-company-team-00x dataset **(Incorrect)**

D. Create Authorized views for tables required by the team in the acme-company-project project. Grant BigQuery User role for acme-company-team-00x and data viewer role to acme-company-team-00x dataset

Explanation

同样 **dataset** 但是不同 **project - user role**

Correct answer is **B** as the controlled access can be provided using Authorized views created in a separate project. The Users should be provided with the BigQuery User role on the project to query and Data Viewer role to the dataset to be able to view the dataset within the project.

Refer GCP documentation - [BigQuery Authorized View]

Giving a view access to a dataset is also known as creating an authorized view in BigQuery. An authorized view allows you to share query *results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query.*

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because *you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, your data analysts would have access to both the view and the data.*

In order to query the view, your data analysts need permission to run query jobs. The bigquery.userrole includes permissions to run jobs, including query *jobs, within the project. If you grant a user or group the bigquery.user role at the project level, the user can create datasets and can run query jobs against tables in those datasets. The bigquery.user role does not give users permission to query data, view table data, or view table schema details for datasets the user* *did not create.*

Assigning your data analysts the project-level bigquery.user role does not give them the ability to view or query table data in the dataset containing the tables queried by the view. Most individuals (data scientists, business intelligence analysts, data analysts) in an enterprise should be assigned the project-level bigquery.user role.

In order for your data analysts to query the view, they need READER access to the dataset containing the view. The bigquery.user role gives your data analysts the permissions required to create query jobs, but they cannot successfully query the view unless they also have at least READER access to the dataset containing *the view.*

Option A is wrong as viewer role should be provided to the dataset within the respective team project.

Option C is wrong as the user should be provided with the User role.

Option D is wrong as Authorized views should be created in a separate project. If they are created in the same project, the users would have access to the underlying tables as well.

Question 2: **Correct**

You are tasked with building an online analytical processing (OLAP) marketing analytics and reporting tool. This requires a relational database that can operate on hundreds of terabytes of data. What is the Google recommended

tool for such applications?

A. Cloud Spanner, because it is globally distributed

B. Cloud SQL, because it is a fully managed relational database

C. Cloud Firestore, because it offers real-time synchronization across devices

D. BigQuery, because it is designed for large-scale processing of tabular data **(Correct)**

Explanation

Correct answer is **D** as BigQuery is a fully managed data warehouse solution with analytics and reporting capability and able to handle large amounts of data.

Refer GCP documentation - Storage Options

BigQuery A scalable, fully managed enterprise data warehouse (EDW) with SQL and fast ad-hoc queries.OLAP workloads up to petabyte scaleBig data exploration and processingReporting via business intelligence (BI) toolsAnalytical reporting on large dataData science and advanced analysesBig data processing using SQL

Options A & B are wrong as they are relational databases and suitable for OLTP workloads.

Option C is wrong as Cloud Firestore is a shared file system to be attached to the virtual machines. It does not provide analytics capabilities.

Question 3: **Correct**

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?

A. Change the processing job to use Google Cloud Dataproc instead.

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

C. Create a cron job with Google App Engine **(Correct)** Cron Service to run the Cloud Dataflow job.

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

Explanation

Correct answer is **C** as the Cloud Dataflow job can be triggering using a cron job hosted on the GCP infrastructure.

Refer GCP documentation - Scheduling Dataflow pipelines using App Engine

[Cron Service]

App Engine Cron Service allows you to configure and run cron jobs at regular intervals. These cron jobs are a little different from regular Linux cron jobs in that they cannot run any script or command. They can only invoke a URL defined as part of your App Engine app via HTTP GET. In return, you don't have to worry *about how or where the cron job is running. App Engine infrastructure takes care of making sure that your cron job runs at the interval that you want it to run.*

Option A is wrong as Dataproc is more suitable for existing hadoop or spark jobs and it not an inexpensive approach.

Option B is wrong as manually triggering the pipeline is not an efficient approach.

Option D is wrong as Cloud Dataflow Streaming job only supports Cloud Pub/Sub

What data sources and sinks are supported in streaming mode?

You can read streaming data from Cloud Pub/Sub, and you can write streaming data to Cloud Pub/Sub or BigQuery..

Question 4: Correct

Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single

location in real time to determine which user bid first. What should you do?

A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.

B. Have each application server write the bid events to Cloud

Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.

C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.

D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first. **(Correct)**

需要dataflow在中间做处理

Explanation

Correct answer is **D** as Cloud Pub/Sub with Cloud Dataflow can be used to buffer the bids and process them as per the order.

Refer GCP documentation - Cloud Pub/Sub Subscriber

Cloud Pub/Sub provides a highly-available, scalable message delivery service.

The tradeoff for having these properties is that the order in which messages are received by subscribers is not guaranteed. While the lack of ordering may sound burdensome, there are very few use cases that actually require strict ordering.

Typically, Cloud Pub/Sub delivers each message once and in the order in which it was published. However, messages *may sometimes be delivered out of order or more than once. In general, accommodating more-than-once delivery requires your subscriber to be [idempotent] when processing messages. You can achieve exactly once processing of Cloud Pub/Sub message streams using Cloud Dataflow `PubsubIO`. `PubsubIO` de-duplicates messages on custom message identifiers or those assigned by Cloud Pub/Sub. You* *can also achieve ordered processing with Cloud Dataflow by using the standard sorting APIs of the service. Alternatively, to achieve ordering, the publisher of the topic to which you subscribe can include a sequence token in the message.*

Options A, B & C are wrong as they do not provide a scalable approach at the real time to collate and determine which user bid first.

Question 5: **Correct**

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support

this method? (Choose two.)

A. There are very few occurrences of mutations relative to normal samples. **(Correct)**

B. There are roughly equal occurrences of both normal and mutated samples in the database.

C. You expect future mutations to have different features from the mutated samples in the database.

D. You expect

future mutations to have similar features to the mutated samples in the database. **(Correct)**

E. You already have labels for which samples are mutated and which are normal in the database.

Explanation

Correct answers are **A & D** as Unsupervised Anomaly Detection would need the data to have fewer occurrences of mutation as compared to normal data and expect future mutations to have similar features.

Unsupervised Anomaly Detection - *These techniques do not need training data. As alternative, they based on two basic assumptions. First, they presume that most of the network connections are normal traffic and only a very small traffic percentage is abnormal. Second, they* *anticipate that malicious traffic is statistically various from normal traffic. According to these two assumptions, data groups of similar instances which appear frequently are assumed to be normal traffic, while infrequently instances which considerably various from the majority of the instances are regarded to be malicious*

Option B is wrong as an equal number of

mutations to normal data would not allow anomaly detection.

Option C is wrong as with different features for future mutations, the anomaly direction would not work.

Option E is wrong as it would be best to use supervised learning, as we already have labels for samples.

Supervised Anomaly Detection - *Supervised methods (also known as classification *methods) required a labeled training set containing both normal and anomalous samples to construct the predictive model. Theoretically, supervised methods provide better detection rate than semi-supervised and unsupervised methods, since they have access to more information. However, there exist some technical issues, which make these methods seem not accurate* *as they are supposed to be .*

Question 6: **Correct**

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which

queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

A. Create a new view over

events using standard SQL

B. Create a new partitioned table using a standard SQL query

C. Create a new view over events_partitioned using standard SQL

D. Create a service account for the ODBC connection to use for authentication **(Correct)**

E. Create a Google Cloud Identity and Access Management (Cloud **(Correct)**

IAM) role for the ODBC connection and shared "events"

Explanation

Correct answers are **D & E** as BigQuery supports authentication using Service Accounts and User accounts.

Refer GCP documentation - [BigQuery with ODBC driver]()

You'll need to provide credentials, either with a service account key or user authentication.

Service accounts* - *A service account

is a Google account that is associated with your GCP project. Use a service account to access the BigQuery API if your application can run jobs associated with service credentials rather than an end-user's credentials, such as a batch processing pipeline.

User accounts - *Use user credentials to ensure that your application has access only to BigQuery tables that* *are available to the end user. A user credential can run queries against only the end user's Cloud Platform project rather than the application's project, meaning the user is billed for queries instead of the application.*

Options A, B & C are wrong as the applications can connect to Legacy SQL using ODBC using service account key or user authentication.

Question 7: Correct

You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on

a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery. How should you securely run this workload?

A. Restrict the Google Cloud Storage bucket so only you can see the files

B. Grant the Project Owner role to a service account, and run the job with it

C. Use

a service account with the ability to read the batch files and to write to BigQuery **(Correct)**

D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Explanation

Correct answer is **C** as the best practice is to use a service account with least privilege.

Refer GCP documentation - IAM Best Practices - Service Accounts

A service account is a special type of Google account intended to represent a non-human user that needs to authenticate and be authorized to access data in Google APIs.

Typically, service accounts are used in scenarios such as:

Running workloads

on virtual machines (VMs).

Running workloads on on-premises workstations or data centers that call Google APIs.

Running workloads which are not tied to the lifecycle of a human user.

Option A is wrong as the best practice is to use a service account i.e. non human user for jobs.

Option B is wrong as Project Owner role does not align with the IAM best practices of least privilege.

All editor permissions and permissions for the following actions:

Manage roles and permissions for a project and all resources within the project.

Set up billing for a project.

Option D is wrong as the Project Viewer role does not grant access to write to BigQuery.

Permissions for read-only actions that do not affect state, such as viewing (but not modifying) existing resources or data.

Question 8: **Correct**

Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a

MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

A. Add a node to the MySQL cluster and build an OLAP cube there.

B. Use an ETL tool to load the data from MySQL into Google BigQuery. **(Correct)**

C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.

D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

Explanation

Correct answer is **B** as as moving data to BigQuery would reduce the load on the MySQL instances and allow data to be queried using the same SQLs.

Options A & C is wrong as this does not reduce the load on the existing MySQL instance.

Option D is wrong as backups cannot be mounted to Google Cloud SQL, but have to be restored or imported. Also, it needs operational effort.

Question 9: Correct

You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

A. Get more training examples **(Correct)**

B. Reduce the number of training examples

C. Use a smaller set of features **(Correct)**

D. Use a larger set of features

E. Increase the regularization parameters **(Correct)**

F. Decrease the regularization parameters

Explanation

Correct answers are **A, C & E**

Refer documentation - [Tensorflow Overfit vs Underfit]()

Overfitting is a phenomenon where a machine learning model models the training data too well but fails to perform well on the *testing data.*

If you train for too long though, the model will start to overfit and learn patterns from the training data that don't generalize to the test data. We need to strike a balance. Understanding how to train for an appropriate number of epochs as we'll explore below is a useful skill.

To prevent overfitting, the best solution *is to use more training data. A model trained on more data will naturally generalize better. When that is no longer possible, the next best solution is to use techniques like regularization. These place constraints on the quantity and type of information your model can store. If a network can only afford to memorize a small number* *of patterns, the optimization process will force it to focus on the most prominent patterns, which have a better chance of generalizing well.*

Train with more data - *It won't work every time, but training with more data can help algorithms detect the signal better.*

Remove features - *Some algorithms have built-in feature selection. For those* *that don’t, you can manually improve their generalizability by removing irrelevant input features.*

Regularization* - *Regularization refers to a broad range of techniques for artificially forcing your model to be simpler. The method will depend on the type of learner you’re using. For example, you could prune a decision tree, use dropout on a neural network, or add a

penalty parameter to the cost function in regression. Oftentimes, the regularization method is a hyperparameter as well, which means it can be tuned through cross-validation.

Question 10: **Correct**

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel

data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

A. Use Storage Transfer Service to

transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.

B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.

C. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage **(Correct)**

Multi-Regional storage bucket as a final destination.

D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

Explanation

Correct answer is **C** as BigQuery Data Transfer Service provides integration with youtube to transfer data to Cloud Storage. Using

Multi-Regional storage bucket would allow storage and querying data from across global.

Refer GCP documentation - BigQuery Transfer Service & Dataset Locations

BigQuery Data Transfer Service automates data movement from Software as a Service (SaaS) applications such as Google Ads and Google Ad Manager on a scheduled, managed basis. Your analytics team can lay *the foundation for a data warehouse without writing a single line of code.*

Like BigQuery, the BigQuery Data Transfer Service is a multi-regional resource.

Data locality is specified when you create a dataset to store your BigQuery Data Transfer Service core customer data. When you set up a transfer, the transfer configuration is set to the same *locality as the dataset. The BigQuery Data Transfer Service processes and stages data in the same location as the target BigQuery dataset.*

If your BigQuery dataset is in a multi-regional location, the Cloud Storage bucket containing the data you're loading must be in a regional or multi-regional bucket in the same location.

When you export

data, the regional or multi-regional Cloud Storage bucket must be in the same location as the BigQuery dataset.

Options A & B are wrong as Storage Transfer Service transfers data from an online *data source* to a *data sink*. Your *data source* can be an Amazon Simple Storage Service (Amazon S3) bucket, an HTTP/HTTPS location, or a Cloud Storage bucket. Your *data sink* (the destination) is always a Cloud Storage bucket.

Option D is wrong as Multi-regional storage should be preferred over Regional storage.

Question 11: **Correct**

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being generated

during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour. The data scientists have written the following code to read the data for a new key features in the logs.

```
BigQueryIO.Read
.named("ReadLogData")
.from("clouddataflow-readonly:samples.log_data")
```

You want to improve the performance

of this data read. What should you do?

A. Specify the TableReference object in the code.

B. Use `.fromQuery` operation to read specific fields from the table. **(Correct)**

C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.

D. Call a transform that returns TableRow objects, where each element in the PCollection represents a single row in the table.

Explanation

Correct answer is **B** as best practice is to limit the data queried.

`BigQueryIO.read.from()` directly reads the whole table from BigQuery. This function exports the whole table to temporary files in Google Cloud Storage, where it will later be read from. This requires almost no computation, as it only performs

an export job, and later Dataflow reads from GCS (not from BigQuery).

`BigQueryIO.read.fromQuery()` *executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs).*

Refer GCP

documentation - [BigQuery Best Practices]

Best practice: *Control projection — Query only the columns that you need.*

Projection refers to the number of columns that are read by your query. Projecting excess columns incurs additional (wasted) I/O and materialization (writing results).

Using `SELECT *` is the most expensive way to query data. When you use `SELECT *`, BigQuery *does a full scan of every column in the table.*

If you are experimenting with data or exploring data, use one of the data preview options instead of `SELECT *`.

Applying a `LIMIT` clause to a `SELECT *` query does not affect the amount of data read. You are billed for reading all bytes in the entire table, and the query counts against your *free tier quota.*

Instead, query only the columns you need. For example, use `SELECT * EXCEPT` to exclude one or more columns from the results.

If you do require queries against every column in a table, but only against a subset of data, consider:

Materializing results in a destination table and querying that table instead

Partitioning your tables by date and querying the relevant partition; for example, `WHERE _PARTITIONDATE="2017-01-01"` only scans the January 1, 2017 partition

Querying a subset of data or using `SELECT * EXCEPT` can greatly reduce the amount of data that is read by a query. In addition to the cost savings, performance is improved by reducing the amount of data

I/O and the amount of materialization that is required for the query results.

Options A & C are wrong as they do not improve query performance

Option D is wrong as performing inline transformation is not recommended and would reduce the performance.

Question 12: **Correct**

You are designing storage for two relational tables that

are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on non-key columns. What should you do?

A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.

B. Use Cloud SQL for storage.

Use Cloud Dataflow to transform data to support query patterns.

C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns. **(Correct)**

D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Explanation

Correct answer is **C** as Cloud Spanner provides the ability to scale horizontally and Secondary Indexes help to query non-key fields effectively.

Refer GCP documentation - Spanner & Secondary Indexes

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, no planned downtime, and enterprise-grade security. Cloud Spanner revolutionizes database administration and management and makes application development more efficient.

In a Cloud Spanner database, Cloud Spanner automatically creates an index *for each table's primary key column.*

You can also create secondary indexes for other columns. Adding a secondary index on a column makes it more efficient to look up data in that column.

Options A & B are wrong as Cloud SQL does not provide the ability to scale horizontally.

Option D is wrong as using Dataflow

is not an effective approach.

Question 13: **Correct**

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-

time dashboards?

A. Use a row key of the form `<timestamp>`.

B. Use a row key of the form `<sensorid>`.

C. Use a row key of the form `<timestamp>#<sensorid>`.

D. Use a row key of the form `<sensorid>#<timestamp>`. **(Correct)**

Explanation

Correct answer is **D** as the data is time-series data, it is recommended to use tall and narrow tables with a combination of both sensorid and timestamp. Also, it is recommended to not use timestamp at the start of the row key as most writes would be pushed to a single node.

Refer GCP documentation - Bigtable Schema Design & Time-Series Schema Design

A tall and narrow table has a small number of events per row, which could be just one *event, whereas a short and wide table has a large number of events per row.*

For time series, you should generally use tall and narrow tables. *This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will* *exceed the recommended maximum*

if you often need to retrieve data based on the time when it was recorded, it's a good idea to include a timestamp as part of your row key. ***Using the timestamp by itself as the row key is not recommended, as most writes would be pushed onto a single node. For the same reason, avoid placing a***

timestamp at the start of the row key.

For example, your application might need to record performance-related data, such as CPU and memory usage, once per second for a large number of machines. Your row key for this data could combine an identifier for the machine with a timestamp for the data (for example, `machine_4223421#1425330757685`).

Options A & B are

wrong as they would not querying based on sensor and time together to build the dashboard.

Option C is wrong as it is recommended to NOT have timestamp at the start of the row key.

Question 14: Correct

Your company receives both batch- and stream-based event data. You want to process

the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

A. Set a single global window to capture all the data.

B. Set sliding

windows to capture all the lagged data.

C. Use watermarks and timestamps to capture the lagged data. **(Correct)**

D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

Explanation

Correct answer is **C** as you would need both watermarks to identify the time period.

Refer GCP documentation - Dataflow Streaming Basics & Beam Windowing

In any data processing system, there is a certain amount of lag between the time a data event occurs (the "event time", determined by the timestamp on the data element itself) and the time the actual data element gets processed at any stage in *your pipeline (the "processing time", determined by the clock on the system processing the element). In addition, there are no guarantees that data events will appear in your pipeline in the same order that they were generated.*

Watermarks are the notion of when the system expects that all data in a certain window has arrived in the pipeline.

Cloud Dataflow tracks watermarks because data is not guaranteed to arrive in time order or at predictable intervals. In addition, there are no guarantees that data events appear in the pipeline in the same order that they were generated. After the watermark progresses past the end of a window, any further elements that arrive with a timestamp in that *window are considered late data.*

However, data isn't always guaranteed to arrive in a pipeline in time order, or to always arrive at predictable intervals. Beam tracks a watermark, which is the system's notion of when all data in a certain window can be expected to have arrived in the pipeline. Once the watermark progresses past the end of a window, *any further element that arrives with a timestamp in that window is considered **late data**.*

Option A is wrong as for unbounded data you need to choose non-global window.

Option B is wrong as Sliding windows do not catch late data.

Hopping windowing also represents time intervals in the data stream; however, hopping windows can overlap.

For example, each window might capture five minutes worth of data, but a new window starts every ten seconds. The frequency with which hopping windows begin is called the period. Therefore, our example would have a window duration of five minutes and a period of ten seconds.

Because multiple windows overlap, most elements in a dataset belong *to more than one window. Hopping windowing is useful for taking running averages of data; in our example, you can compute a running average of the past minutes' worth of data, updated every thirty seconds.*

Option D is wrong as you would need watermarks to identify late data.

Question 15: **Correct**

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?

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

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

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

D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created **(Correct)**

Explanation

Correct answer is **D** as

with unbounded Pub/Sub collection you need to apply the non-global windowing function.

Refer GCP documentation - Dataflow Streaming Pipeline Basics & Beam Windowing

Windowing enables grouping over unbounded collections by dividing the collection into windows according to the timestamps of the individual elements. Each window contains a finite number of elements. Grouping operations work *implicitly on a per-window basis; grouping operations process each collection as a succession of multiple, finite windows, though the entire collection might be of unbounded size.*

If you are using unbounded `PCollection`s, you must use either [non-global windowing]() or an [aggregation trigger]() in order to perform a `GroupByKey` or [CoGroupByKey](). This is because a bounded `GroupByKey` or `CoGroupByKey` must wait for all the data with a certain *key to be collected, but with unbounded collections, the data is unlimited. Windowing and/or triggers allow grouping to operate on logical, finite bundles of data within the unbounded data streams.*

If you do apply `GroupByKey` or `CoGroupByKey` to a group of unbounded `PCollection`s without setting either a non-global windowing strategy, a trigger strategy, or both for each collection, Beam generates

an IllegalStateException error at pipeline construction time.

Option A is wrong as PubsubIO will read the message from Pub/Sub and assign the message publish time to the element as the record timestamp.

Option B is wrong as trigger and watermarks are not mandatory. *A related concept, called* ***triggers****, determines when to emit the results of aggregation* *as unbounded data arrives. You can use triggers to refine the windowing strategy for your `PCollection`. Triggers allow you to deal with late-arriving data or to provide early results.*

Option C is wrong as with unbounded collection you need to apply non-global windowing function.

Question 16: **Correct**

You need

to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, the application was designed to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts

do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?

A. Re-write the application to load accumulated data every 2 minutes.

B. Convert the streaming insert code to batch load for individual messages.

C. Load the original message to

Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.

D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long. **(Correct)**

Explanation

Correct answer is **D** as the application can be adjusted to check the average latency and wait for a variable time.

Refer GCP documentation - [BigQuery Streaming Inserts]

Streamed data is available for real-time analysis within a few seconds of the first streaming insertion into a table. In rare circumstances (such as an outage), data in the streaming buffer may be temporarily unavailable. When data is unavailable, queries continue to run successfully, but *they skip some of the data that is still in the streaming buffer. These queries will contain a warning in the `errors` field of `bigquery.jobs.getQueryResults`, in the response to `bigquery.jobs.query` or in the `status.errors` field of `bigquery.jobs.get`.*

Data can take up to 90 minutes to become available for copy and export operations. Also, when streaming to a partitioned table, data in the streaming buffer has a

NULL value for the `_PARTITIONTIME` pseudo column. To see whether data is available for copy and export, check the `tables.get` response for a section named `streamingBuffer`. If that section is absent, your data should be available for copy or export, and should have a non-null value for the `_PARTITIONTIME` pseudo column. Additionally, the `streamingBuffer.oldestEntryTime` field can be leveraged to identify the age of records in the *streaming buffer*.

Option A is wrong as the data availability is variable, fixed time would not address the problem.

Option B is wrong as Batch load is not ideal for individual messages.

Option C is wrong as Cloud SQL is not ideal choice to support streaming data inserts.

Question 17: Correct

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

A. Continuously

retrain the model on just the new data.

B. Continuously retrain the model on a combination of existing data and the new data. (Correct)

C. Train on the existing data while using the new data as your test set.

D. Train on the new data while using the existing data as your test set.

Explanation

Correct answer is **B** as the preference is going to change over period of time, it is more logical to retrain the models on the new data and existing data.

Another way to keep your models up-to-date is to have an automated system to continuously evaluate and retrain your models. This type of system is often *referred to as continuous learning, and may look something like this:*

Save new training data as you receive it.

When you have enough new data, test its accuracy against your machine learning model.

If you see the accuracy of your model degrading over time, use the new data, or a combination of the new data and old training *data to build and deploy a new model.*

The benefit to a continuous learning system is that it can be completely automated.

Option A is wrong as the model can be improved taking into account the new and old data which would change over a period of time.

Options C & D are wrong as the training needs to happen on both new and old data. Training of one set of data and using on other set would result in an inaccurate model and results.

Question 18: Correct

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You

also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query. **(Correct)**

B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud

Storage and BigQuery permanent linked tables for query.

C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.

D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

Explanation

Correct answer is **A** as BigQuery can be used to store and query the text data. BigQuery natively supports Avro and can work with compressed blocks.

Refer GCP documentation - [BigQuery Loading Data]

The Avro binary format is the preferred format for loading compressed data. Avro data is faster to load because the data can be read in *parallel, even when the data blocks are compressed. Compressed Avro files are not supported, but compressed data blocks are. BigQuery supports the DEFLATE and Snappy codecs for compressed data blocks in Avro files.*

Option B is wrong as although it works, Google recommends using BigQuery for storage, if possible, as it results is better performance.

没说要省钱

Query performance for external data sources may not be as high as querying data in a native BigQuery table. If query speed is a priority, load the data into BigQuery instead of setting up an external data source. The performance of a query that includes an external data source depends on the external storage type.

For example, querying data stored in Cloud Storage is faster than querying data stored in Google Drive. In general, query performance for external data sources should be equivalent to reading the data directly from the external storage.

Options C & D are wrong Grid Computing Tools are not needed and Dataflow can work fine. Also, for text files (CSV and JSON) BigQuery can load uncompressed files faster.

For other data formats such as CSV and JSON, BigQuery can load uncompressed files significantly faster than compressed files because uncompressed files can be read in parallel. Because uncompressed files are larger, using them can lead to bandwidth limitations and higher

Cloud Storage costs for data staged in Cloud Storage prior to being loaded into BigQuery. You should also note that line ordering is not guaranteed for compressed or uncompressed files. It's important to weigh these tradeoffs depending on your use case.

In general, if bandwidth is limited, compress your CSV and JSON files *using gzip before uploading them to Cloud Storage. Currently, when loading data into BigQuery, gzip is the only supported file compression type for CSV and JSON files. If loading speed is important to your app and you have a lot of bandwidth to load your data, leave your files uncompressed.*

Question 19: **Correct**

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should

you use?

A. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.

B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.

C. Use Cloud Storage for storage. Link as permanent tables in BigQuery **(Correct)** for query.

D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

Explanation

Correct answer is **C** as Cloud Storage provides a cost-effective solution to store data and BigQuery Permanent tables can use Cloud Storage as an external data store and be shared.

Refer GCP documentation - BigQuery Temporary vs Permanent Tables

Permanent versus temporary external tables

You can query an external data source in BigQuery by using a permanent table or a temporary table. When you use a permanent table, you create a table in a BigQuery dataset that is linked to your external data source. Because the table is permanent, you can use dataset-

level access controls to share the table with others who also have access to the underlying external data source, and you can query the table at any time.

When you query an external data source using a temporary table, you submit a command that includes a query and creates a non-permanent table linked to the *external data source. When you use a temporary table, you do not create a table in one of your BigQuery datasets. Because the table is not permanently stored in a dataset, it cannot be shared with others. 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.*

Options A & B are wrong as Bigtable is not a cost-effective storage solution.

Option D is wrong as BigQuery temporary tables for useful for one-time jobs and cannot be shared with others.

Question 20: Correct

You have enabled the free

integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format `app_events_YYYYMMDD`. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?

A. Use the `TABLE_DATE_RANGE` function **(Correct)**

B. Use the `WHERE _PARTITIONTIME` pseudo column

C. Use `WHERE date BETWEEN YYYY-MM-DD AND`

`YYYY-MM-DD`

D. Use `SELECT IF(date >= YYYY-MM-DD AND date <= YYYY-MM-DD)`

Explanation

Correct answer is **A** as the data is already created by data, it would be best to use *`TABLE_DATE_RANGE`* to filter based on range of dates.

Refer GCP documentation - BigQuery with Firebase Analytics & Legacy SQL Reference

`TABLE_DATE_RANGE()` Queries multiple daily tables that span a date range.

What if we want to run a query across both platforms of our app over a specific date range? Since Firebase Analytics data is split into tables for each day, we can do this using BigQuery's [TABLE_DATE_RANGE] function. This query returns a count of the cities users are coming from over a one week period:

```
SELECT
```

```
user_
dim.g
eo_in
fo.ci
ty,
COUNT
(user
_dim.
geo_i
nfo.c
ity)
as ci
ty_co
unt
FROM
TABLE
_DATE
_RANG
E([fi
rebas
e-ana
lytic
s-sam
ple-d
ata:x
x.app
_even
ts_],
DATE_
ADD
('201
6-06-
07',
-7,
'DA
Y'),
CURRE
NT_TI
MESTA
MP
()),
GROUP
BY
user_
dim.g
eo_in
fo.ci
ty
```

```
ORDER BY city_count DESC
```

Option B is wrong as _PARTITIONTIME is valid only for ingestion streaming data.

Options C & D are wrong as they are not valid wildcard date functions for Legacy SQL.

Question 21: Correct

Your analytics team wants to build a simple statistical model to determine which

customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

A. Migrate the workload to Google Cloud Dataflow

B. Use pre-emptible virtual machines (VMs) for the cluster (Correct)

C. Use a higher-memory node so that the job runs faster

D. Use SSDs on the worker nodes so that the job can run faster

Explanation

Correct answer is **B** as the key requirement is to optimize cost, pre-emptible VMs can be used with Dataproc.

Refer GCP documentation - [Dataproc Preemptible-VMs]

In addition to using

standard Compute Engine virtual machines (VMs), Cloud Dataproc clusters can use preemptible VM instances, also known as preemptible VMs. You may decide to use preemptible instances to lower per-hour compute costs for non-critical data processing or to create very large clusters at a lower total cost.

All preemptible instances added to a

cluster use the machine type of the cluster's non-preemptible worker nodes. For example, if you create a cluster with workers that use `n1-standard-4` machine types, all preemptible instances added to the cluster will also use `n1-standard-4` machines. The addition or removal of preemptible workers from a cluster does not affect the number of non-preemptible *workers in the cluster.*

Because preemptible instances are reclaimed if they are required for other tasks, Cloud Dataproc adds preemptible instances as secondary workers in a managed instance group, which contains only preemptible workers. The managed group automatically re-adds workers lost due to reclamation as capacity permits. For example, if two preemptible machines are *reclaimed and removed from a cluster, these instances will be re-added to the cluster if and when capacity is available to re-add them.*

Option A is wrong as Dataflow would need the redesign of the application, as it cannot reuse the Spark scripts.

Options C & D are wrong as they would not reduce the cost.

Question 22: **Correct**

You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot

be removed. What should you do?

A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to ‘none’ using a Cloud Dataproc job.

B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job. **(Correct)**

C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataprep job.

D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.

Explanation

Correct answer is **B** as Cloud Dataprep would help find null values as well as help convert the null values as required.

Refer GCP documentation - DataPrep Manage Null values

Option A is wrong as Dataproc is not efficient to convert nulls values.

Options C & D are wrong as Dataflow is not efficient in finding nulls in the data.

Question 23: Correct

You are developing

an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from

other customer preferences on several TB of data. What should you do?

A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc. Call the model from your application.

B. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.

C. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and **(Correct)** filter the predicted labels to match the user's viewing history to generate preferences.

D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Explanation

Correct answer is **C** as Cloud Video Intelligence API provides an out of the box solution to generate labels from videos. Storing data in Bigtable would provide low latency and very fast filtering capability of TBs of data.

Options A & B are wrong as building a model for label extraction is cumbersome as compared to using already available Cloud Video Intelligence service.

Option D is wrong as Cloud SQL is not ideal for low latency access on TBs of data.

Question 24: **Correct**

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual

users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application. What should you do?

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

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

C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset (Correct)

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

Explanation

Correct answer is **C** as the Application needs to access BigQuery, it can be configured to use Service Account.

Refer GCP documentation - [BigQuery Service Account File]

A *service account* is a Google account that is associated with your GCP project. Use a service account to access the BigQuery API if your application can run jobs associated with service credentials rather than an end-user's credentials, such as a batch processing pipeline.

Manually create and obtain service account credentials to use BigQuery when an application is deployed on-premises or to other public *clouds. You can set the environment variable to load the credentials using Application Default Credentials, or you can specify the path to load the credentials manually in your application code.*

Options A, B & D are wrong as either they are not best practices or would provide users access to the dataset.

Question 25: **Incorrect**

You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

A. Create

a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.

B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances. (Correct)

C. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances. **(Incorrect)**

D. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

Explanation

Correct answer is **B** as encryptions keys in Cloud KMS can be used by Compute Engine to encrypt data and provides an ability to create, rotate, and destroy as needed

Refer GCP documentation - Compute Engine Encryption & Encryption at Rest

By default, Compute Engine encrypts customer *content at rest. Compute Engine handles and manages this encryption for you without any additional actions on your part. However, if you want to control and manage this encryption yourself, you can use key encryption keys. Key encryption keys do not directly encrypt your data but are used to encrypt the data encryption keys that encrypt your data.*

You have two options for key encryption keys in Compute Engine:

Use Cloud Key Management Service to create and manage key encryption keys. For more information, see Key management. This topic provides details about this option, known as customer-managed encryption keys (CMEK).

Create and manage your own key encryption keys. For information *about this option, known as customer-supplied encryption keys (CSEK), see [Encrypting Disks with Customer-Supplied Encryption Keys].*

After you create a Compute Engine resource that is protected by Cloud KMS, you do not need to specify the key because Compute Engine knows which KMS key was used. This is different from how Compute Engine accesses *resources protected by customer-supplied keys. For that access, you need to specify the customer-supplied key.*

Option A is wrong as the default encryption provided by Compute Engine does not allow creation, management and rotation.

Option C is wrong as CSEK does not need to be uploaded to Cloud KMS.

Option D is wrong

as the approach does not encrypt the data.

Question 26: **Correct**

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input

data volume that will vary in size with minimal manual intervention. What should you do?

A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.

B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.

C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances. **(Correct)**

D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.

Explanation

Correct answer is **C** as Dataflow, provides a cost-effective solution to perform transformations on the streaming data, with auto-scaling provides scaling without any intervention. System lag with Stackdriver provides monitoring for the streaming data.

Refer GCP

documentation - [Dataflow Monitoring]

With autoscaling enabled, the Cloud Dataflow service automatically chooses the appropriate number of worker instances required to run your job. The Cloud Dataflow service may also dynamically re-allocate more workers or fewer workers during runtime to account for the characteristics of your job. Certain parts of your pipeline may be computationally heavier than others, and the Cloud Dataflow service may automatically spin up additional workers during these phases of your job (and shut them down when they're no longer needed).

Stackdriver provides powerful monitoring, logging, and diagnostics. Cloud Dataflow integration with Stackdriver Monitoring allows you to access Cloud Dataflow job metrics such as Job

Status, Element Counts, System Lag (for streaming jobs), and User Counters from the Stackdriver dashboards. You can also employ Stackdriver alerting capabilities to be notified of a variety of conditions, such as long streaming system lag or failed jobs.

Options A & B are wrong as Dataproc does not provide a cost-effective solution as the machine needs to be configured.

Option D is wrong as using non-default Compute Engine machine types as needed would need manual intervention.

Question 27: **Correct**

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?

A. Use Google Stackdriver Audit Logs to review data access. **(Correct)**

B. Get the identity and access management (IAM) policy of each table

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

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

Explanation

Correct answer is **A** as Stackdriver BigQuery Data Access audit logs can provide the information what users are accessing what BigQuery datasets.

Refer GCP documentation - BigQuery Audit Logs

Cloud Audit Logs are a collection of logs provided by Google Cloud Platform that provide insight into operational concerns related to your use of Google Cloud services. This page provides details about BigQuery specific *log information, and it demonstrates how to use BigQuery to analyze logged activity.*

Option B is wrong as IAM policy is not attached to the tables.

Option C is wrong as Stackdriver only provides info for available and allocated Query Slots

Option D is wrong as billing does not provide information of what users are accessing which tables.

Question 28: **Correct**

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

How should you deduplicate the data most efficiency?

A. Assign global unique identifiers (GUID) to each data entry. **(Correct)**

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

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

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

Explanation

Correct answer is **A** as a global unique identifier would allow one to detect duplicates when the message is retransmitted.

Refer GCP documentation - Pub/Sub Duplicates

Cloud Pub/Sub assigns a unique `message_id` *to each message, which can be used to detect duplicate messages received by the subscriber. This will not, however, allow you to detect duplicates resulting from multiple publish requests on the same data.*

Option B is wrong as using the hash with timestamp of the transmission, it would never match.

Options C & D are wrong as using

database would not be cost effective solution.

can arrive late or out of order. How should yo

Question 30: Correct

Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries data in the cloud. This data is updated

frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?

A. Cloud Bigtable (Correct)

B. Google BigQuery

C. Google Cloud Storage

D. Google Cloud Datastore

Explanation

Correct answer is **A** as Bigtable is ideal for storing time-series data, data with frequent updates.

Refer GCP documentation - [Big data products]

[Cloud Bigtable] provides a massively scalable NoSQL database suitable for low-latency and high-throughput workloads. It integrates easily with popular big-

data tools like Hadoop and Spark, and it supports the open-source, industry-standard HBase API. Cloud Bigtable is a great choice for both operational and analytical applications, including IoT, user analytics, and financial data analysis.

Option B is wrong as BigQuery is not suitable for data with frequent updates.

Option C is wrong as Cloud Storage is not ideal for time-series data with frequent updates.

Option D is wrong as Datastore is not ideal for analytics time-series workload.

Question 31: Correct

Government regulations in your industry mandate that you have to maintain an auditable record of access

to certain types of data. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

A. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.

B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.

C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.

D. In a bucket on Cloud Storage that is accessible only by an App Engine service that **(Correct)** collects user information and logs the access before providing a link to the bucket.

Explanation

Correct answer is **D** as Cloud Storage is an ideal storage option for logs. The access can be controlled using an App Engine with access to the bucket and logging all access events.

Option A is wrong as

encryption can help protect data, however it does not help capture data access.

Options B & C are wrong as BigQuery and Cloud SQL are not an ideal storage option for logs.

Question 32: **Correct**

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?

A. Increase the CPU size on your server.

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

C. Increase your network bandwidth from your datacenter to GCP. **(Correct)**

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

Explanation

Correct answer is **C** as to improve data transfer speed the network bandwidth between the data center and GCP needs to be increased. Take into account parallel uploads are already being performed.

Refer GCP documentation - [Transferring Big Data sets to GCP]

Increase network bandwidth

Methods to increase your network bandwidth depends on how *you choose to connect to GCP. You can connect to GCP in three main ways:*

Public internet connection

Direct peering

Cloud Interconnect

Options A & B are wrong as they do not help increase transfer speeds.

Option D is wrong as you cannot increase network bandwidth from Compute Engine to Cloud Storage. Also, private access can be used to enable data transfer from Compute Engine to Cloud Storage using internal network.

Question 33: **Correct**

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?

A.

Linear (Correct) regression

B. Logistic classification

C. Recurrent neural network

D. Feedforward neural network

Explanation

Correct answer is **A** as linear regression can help predict housing prices and also run on a single resource-constrained virtual machine.

Refer documentation - [Machine learning]

Option B is wrong as the housing

price needs to be predicted, classification cannot be used.

Options C & D are wrong as neural network are resource intensive and would not be able to execute on single resource-constrained virtual machine.

Question 34: **Correct**

You are designing a basket abandonment system for an ecommerce company. The system will

send a message to a user based on these rules:

A. No interaction by the user on the site for 1 hour

B. Has added more than $30 worth of products to the basket

C. Has not completed a transaction

You use Google Cloud Dataflow to process the data and decide if a message should be sent. How

should you design the pipeline?

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

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

C. Use a session window with a gap time duration of 60 minutes. **(Correct)**

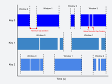
D. Use a global window with a time based trigger with a delay of 60 minutes.

Explanation

Correct answer is **C** as the key here is to track user inactivity for an hour. Session windows can be easily used to track the activity and trigger events based on the conditions.

Refer Beam documentation - Windowing

*A **session window** function defines windows that contain elements that are within a certain gap duration* *of another element. Session windowing applies on a per-key basis and is useful for data that is irregularly distributed with respect to time. For example, a data stream representing user mouse activity may have long periods of idle time interspersed with high concentrations of clicks. If data arrives after the minimum specified gap duration time, this* *initiates the start of a new window.*



Options A, B & D are wrong as they would not be able to track and reset the window based on user activity.

Question 35: Correct

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

A. Windows

at every 100 MB of data.

B. Single, Global Window. (Correct)

C. Windows at every 1 minute.

D. Windows at every 10 minutes.

Explanation

Correct answer is **B** as Dataflow, based on Apache Beam, by default applies a single, global window to unbounded datasets.

Refer Beam documentation - Windowing

Beam's default windowing behavior is to assign all elements of a `PCollection` to a single, global window and discard late data, even for unbounded `PCollection`s. Before you use a grouping transform such as `GroupByKey` on an unbounded `PCollection`, you must do at least one of the following:

Set a non-global windowing function.

Set a non-default trigger. This allows the *global window to emit results under other conditions, since the default windowing behavior (waiting for all data to arrive) will never occur.*

If you don't set a non-global windowing function or a non-default trigger for your unbounded `PCollection` and subsequently use a grouping transform such as `GroupByKey` or `Combine`, your pipeline will generate an error upon construction and your *job will fail.*

Question 36: **Correct**

You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

A. Cloud Speech-to-Text API

B. Cloud Natural Language API

C. Dialogflow Enterprise Edition **(Correct)**

D. Cloud AutoML Natural Language

Explanation

Correct answer is **C** as Dialogflow Enterprise Edition would provide an ideal solutionas the key requirement is to interpret voice commands and fire events.

Refer GCP documentation

- [AI Products](#)

Dialogflow is an end-to-end, build-once deploy-everywhere development suite for creating conversational interfaces for websites, mobile applications, popular messaging platforms, and IoT devices. You can use it to build interfaces (such as chatbots and conversational IVR) that enable natural and rich interactions between your users and your business. Dialogflow

Enterprise Edition users have access to Google Cloud Support and a service level agreement (SLA) for production deployments.

You can expand your conversational interface to recognize voice interactions and generate a voice response, all with a single API call. Powered by [Google Cloud Speech-to-Text]() and [Cloud Text-to-Speech](), it supports real-time streaming *and synchronous modes.*

Option A is wrong as Cloud Speech-to-Text API just provides speech-to-text conversion powered by ML.

Option B as Cloud Natural Language API help derive insights from unstructured text.

Option D is wrong as AutoML helps reveal the structure and meaning of text through machine learning.

[GCP PDE]()

[Question feedback]

Question 37: **Correct**

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does

not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required. You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

A. Redis

B. HBase (Correct)

C. MySQL

D. MongoDB (Correct)

E. Cassandra (Correct)

F. HDFS with Hive

Explanation

Correct answers are **B, D & E** as HBase, MongoDb and Cassandra are NoSQL options for storing data and provide low latency access to the data with an ability to scale horizontally and being highly available.

Option A is wrong as Redis is more of a caching engine.

Option C is wrong as MySQL is a relational database and would not scale.

Option E is wrong as HDFS with Hive is more ideal for batch jobs and do not provide low latency access to the data.

Question 38: Correct

You need to migrate a 2TB relational database to Google

Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern. Which service do you select for storing and serving your data?

A. Cloud Spanner

B. Cloud Bigtable

C. Cloud Firestore

D. Cloud SQL (Correct)

Explanation

Correct answer is **D** as Cloud SQL provides relational database.

Refer GCP documentation - Databases & Migrating from MySQL to Cloud Spanner

Option A is wrong as although Cloud Spanner provides relation database capability. However, the migration is not seamless and would need modification to the application.

Cloud Spanner uses certain concepts differently from *other enterprise database management tools, so you might need to adjust your application's architecture to take full advantage of its capabilities. You might also need to supplement Cloud Spanner with other services from Google Cloud Platform (GCP) to meet your needs.*

Options B & C are wrong as Bigtable and Firestore are NoSQL/ Non-relational database types and would require modification of the application.

Question 39: **Correct**

Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

A. The CSV data loaded in BigQuery is not flagged as CSV.

B. The CSV data has invalid rows that were skipped on import.

C. The CSV data loaded in BigQuery **(Correct)** is not using BigQuery's default encoding.

D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Explanation

Correct answer is **C** as the data imported fine, the mismatch would be due to the CSV file having a different encoding than BigQuery's default encoding of UTF-8.

Refer GCP documentation - [BigQuery Load CSV]()

CSV encoding

BigQuery expects CSV data to be UTF-8 encoded. If you have CSV files with data encoded in ISO-

8859-1 (also known as Latin-1) format, you should explicitly specify the encoding when you load your data so it can be converted to UTF-8.

Delimiters in CSV files can be any ISO-8859-1 single-byte character. To use a character in the range 128-255, you must encode the character as UTF-8. BigQuery converts the *string to ISO-8859-1 encoding and uses the first byte of the encoded string to split the data in its raw, binary state.*

Question 40: **Correct**

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should

you do?

A. Increase the cluster size with more non-preemptible workers.

B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.

C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.

D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning. **(Correct)**

Explanation

Correct answer is **D** as Dataproc cluster can be scaled using preemptible worker nodes, configured with graceful decommissioning to prevent losing in-progress work.

Refer GCP documentation - Dataproc Scaling Clusters

After creating a Cloud Dataproc cluster, you can adjust ("scale") the cluster by increasing or decreasing the number of primary or secondary worker nodes in the cluster. You can scale a Cloud Dataproc cluster at any time, even when jobs are running on the cluster.

Why scale a Cloud Dataproc cluster?

to increase the number of

workers to make a job run faster

to decrease the number of workers to save money (see [Graceful Decommissioning] as an option to use when downsizing a cluster to avoid losing work in progress).

to increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage

Because clusters can be scaled more than once, *you might want to increase/decrease the cluster size at one time, and then decrease/increase the size later.*

When you downscale a cluster, work in progress may terminate before completion. If you are using Cloud Dataproc v 1.2 or later, you can use Graceful Decommissioning, which incorporates Graceful Decommission of YARN Nodes to finish work in progress on a worker *before it is removed from the Cloud Dataproc cluster.*

Option A is wrong as non-preemptible workers would increase cost.

Option B & C are wrong as the approaches would lead to losing in-progress work.

Question 41: **Correct**

You have Cloud Functions written in Node.js that pull messages from

Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

A. Publisher throughput quota

is too small.

B. Total outstanding messages exceed the 10-MB maximum.

C. Error handling in the subscriber code is not handling run-time errors properly. **(Correct)**

D. The subscriber code cannot keep up with the messages.

E. The subscriber code does not acknowledge the messages that it pulls. **(Correct)**

Explanation

Correct answers are **C & E** as the handling is more than anticipated, the possible reasons are the messages are being redelivered either due to subscriber not acknowledging the message within the ack time or it not handling runtime errors.

Refer GCP documentation - [Pub/Sub Troubleshooting]

Dealing with duplicates and forcing retries - When you do not acknowledge a *message before its acknowledgement deadline has expired, Cloud Pub/Sub resends the message. As a result, Cloud Pub/Sub can send duplicate messages. Use Stackdriver to monitor acknowledge operations with the `expired` response code to detect this condition. To get this data, select the **Acknowledge message operations** metric, then group or filter it by the `response_code` label. Note that `response_code` is a system label* *on a metric - it is not a metric.*

Options A & D are wrong as the Cloud Function is processing more than anticipated without any errors.

Option B is wrong as this would lead into errors.

Question 42: Correct

You need to copy millions of sensitive patient records from a relational database

to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console. **(Correct)**

C. Export the

records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.

D. Export the records from the database as an Avro file.

Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

Explanation

Correct answer is **B** as exporting the files in Avro file provides compression of data. Using Transfer Appliance to transfer data from on-premises to Cloud Storage is both secure and time-efficient. The data can be loaded using BigQuery web UI.

Refer GCP documentation - Transfer Appliance & BigQuery Avro

Transfer Appliance is a high-capacity storage device that enables you to transfer and securely ship your data to a Google upload facility, where *we upload your data to Google Cloud Storage.*

Avro is the preferred format for loading data into BigQuery. Loading Avro files has the following advantages over CSV and JSON (newline delimited):

The Avro binary format:Is faster to load. The data can be read in parallel, even if the data blocks are compressed.Doesn't require typing or serialization.Is *easier to parse because there are no encoding issues found in other formats such as ASCII.*

When you load Avro files into BigQuery, the table schema is automatically retrieved from the self-describing source data.

Options A, C & D are wrong as all of the options would still use public internet to transfer the data to Cloud

Storage which is neither time-efficient and secure.

Question 43: **Incorrect**

Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to

reprocess all failing data). What should you do?

A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.

B. Add a `try… catch` block to your `DoFn` that transforms the data, extract erroneous rows from logs.

C. Add a `try… catch` block to your `DoFn` that transforms the data, **(Incorrect)** write erroneous rows to PubSub directly from the `DoFn`.

D. Add a `try… catch` block to your `DoFn` that transforms the data, use a sideOutput to create a PCollection that can be stored to PubSub later. **(Correct)**

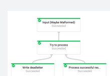
Explanation

Correct answer is **D** as the reliability of the Dataflow can be increased by handling the errors using the try... catch block and using sideOutput to store the failed records to a PubSub topic, acting as a Dead Letter Queue.

Refer GCP documentation - [Dataflow Handling Input Errors]

If the failure is within the processing code of a `DoFn`, one way to handle this is to catch the exception, log an error, and then *drop the input. The rest of the elements in the pipeline will be processed successfully, so progress can be made as normal. But just logging the elements isn't ideal because it doesn't provide an easy way to see these malformed inputs and reprocess them later.*

A better way to solve this would be to have a dead letter file *where all of the failing inputs are written for later analysis and reprocessing. We can use a side output in Dataflow to accomplish this goal. For example:*



Question 44: **Correct**

You have historical data covering the last three years in BigQuery and a data pipeline that

delivers new data to BigQuery daily. You have noticed that when the Data Science team runs a query filtered on a date column and limited to 30–90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. What should you do?**

A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type. **(Correct)**

B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.

C. Modify your pipeline to maintain the last 30–90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.

D. Write an Apache Beam

pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

Explanation

Correct answer is **A** as the table can be partitioned by TIMESTAMP or DATE. This would limit the number of records queried based on the predicate filters.

Refer GCP documentation - [BigQuery Partitioned Tables]

BigQuery also allows partitioned tables. Partitioned tables allow you to bind the partitioning scheme to a specific `TIMESTAMP` *or* `DATE` *column. Data written to a partitioned table is automatically delivered to the appropriate partition based on the date value (expressed in UTC) in the partitioning column.*

Partitioning versus sharding

As an alternative to partitioned tables, you can shard tables using a time-based naming approach such as `[PREFIX]_YYYYMMDD`. This is referred to as creating date-sharded tables. Using either standard SQL or legacy SQL, you can specify a query with a `UNION` operator to limit the tables scanned by the query.

Partitioned tables perform better than tables sharded by date. When *you create date-named tables, BigQuery must maintain a copy of the schema and metadata for each date-named table. Also, when date-named tables are used, BigQuery might be required to verify permissions for each queried table. This practice also adds to query overhead and impacts query performance. The recommended best* *practice is to use partitioned tables instead of date-sharded tables.*

Option B is wrong as exporting the data to CSV is not a cumbersome approach and does not provide the SQL querying capability

Option C is wrong as limiting the table to 30-90 would work, however it is still not cost-effective as the whole table will be always scanned. Also, there is a overhead

Option D is wrong as although sharding is a valid option, partitioning is preferred over sharding.

Question 45: Correct

You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google

BigQuery table with the table name format `LOGS_yyyymmdd`. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

A. Convert all daily log tables into date-partitioned tables

B. Convert the sharded tables into a single partitioned table **(Correct)**

C. Enable query caching so you can cache data from previous months

D. Create separate views to cover each month, and query from these views

Explanation

Correct answer is **B** as Google Cloud recommends using partitioned tables instead of sharded tables, which would help query a single table and improve performance.

Refer GCP documentation - [BigQuery Partitioned Tables]

BigQuery also allows partitioned tables. Partitioned tables allow you to bind the partitioning scheme to a specific `TIMESTAMP` *or* `DATE` *column. Data written to a partitioned table is* *automatically delivered to the appropriate partition based on the date value (expressed in UTC) in the partitioning column.*

Partitioning versus sharding

As an alternative to partitioned tables, you can shard tables using a time-based naming approach such as `[PREFIX]_YYYYMMDD`. This is referred to as creating date-sharded tables. Using either standard SQL or legacy SQL, you can specify a query *with a `UNION` operator to limit the tables scanned by the query.*

Partitioned tables perform better than tables sharded by date. When you create date-named tables, BigQuery must maintain a copy of the schema and metadata for each date-named table. Also, when date-named tables are used, BigQuery might be required to verify *permissions for each queried table. This practice also adds to query overhead and impacts query performance. The recommended best practice is to use partitioned tables instead of date-sharded tables.*

Option A is wrong as the tables are already sharded, creating the date partition would not help.

Option C is wrong as query caching

does not work for wildcard queries

*Currently, cached results are not supported for queries against multiple tables using a wildcard even if the **Use Cached Results** option is checked. If you run the same wildcard query multiple times, you are billed for each query*

Option D is wrong as the daily reports would still fail.

Question 46: **Correct**

A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

A. Implement clustering in BigQuery on the ingest date column.

B. Implement clustering in BigQuery on the package-tracking ID column. **(Correct)**

C. Tier older data onto Cloud Storage files, and leverage extended tables.

D. Re-create the table using data partitioning on the package delivery date.

Explanation

Correct answer is **B** as the tables are already partitioned and the analysts want to query for a package, Clustering on the package-tracking ID would help improve the query performance.

Refer GCP documentation - [BigQuery Cluster Tables]

When you create a clustered table in BigQuery, the table data is automatically organized based on the contents of one or more columns in the table's schema. The columns you specify are used to colocate related data. When you cluster a table using multiple columns, the order of *columns you specify is important. The order of the specified columns determines the sort order of the data.*

Clustering can improve the performance of certain types of queries such as queries that use filter clauses and queries that aggregate data. When data is written to a clustered table by a query job or a load job, BigQuery sorts the *data using the values in the clustering columns. These values are used to organize the data into multiple blocks in BigQuery storage. When you submit a query containing a clause that filters data based on the clustering columns, BigQuery uses the sorted blocks to eliminate scans of unnecessary data.*

Similarly, when you submit

a query that aggregates data based on the values in the clustering columns, performance is improved because the sorted blocks colocate rows with similar values.

When to use clustering

Currently, BigQuery supports clustering over a partitioned table. Use clustering over a partitioned table when:

Your data is already partitioned on a date or timestamp column.

You commonly use filters or aggregation against particular columns in your queries.

Table clustering is supported for both [ingestion time](#) partitioned tables and for tables [partitioned](#) on a `DATE` or `TIMESTAMP` column. Currently, clustering is not supported for non-partitioned tables.

Option A is wrong as clustering needs to be on the column queried, which is the package identifier.

Option C is wrong as

extended tables reduce performance and it is recommended to host the data within BigQuery.

Option D is wrong as partitioning on package delivery date would not improve the performance for queries for a package.

!!!!!Question 47: **Incorrect**

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?

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

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

C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs. (Incorrect)

D. Install the StackDriver Agent and configure the MySQL plugin. (Correct)

Explanation

Correct answer is **D** as MariaDB provides a drop in replacement for MySQL, the MySQL plugin can be used with Stackdriver agent seamlessly to capture network connections, disk IO and replication status for monitoring and alerting

Option A is wrong as the approach does not have minimal development effort.

Option B is wrong as placing in an Instance group with health check does not provide metrics.

Option C is wrong as Stackdriver Logging agent would only capture MariaDB logs.

Question 48: Correct

You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:

Each department should have access only to their data.

Each department will have one or more leads who need to be able to create and update tables and provide them to their team.

Each department has data analysts who need to be able to query but not modify data.

How should you set access to the data in BigQuery?

A. Create a dataset for each department.

Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.

B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset. **(Correct)**

C. Create

a table for each department. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.

D. Create a table for each department. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

Explanation

Correct answer is **B**. Each department needs to have a separate dataset and BigQuery access control works on dataset and not on tables. Data Analysts should be given the VIEWER role to query, but not modify data. Leads should be provided with EDITOR access to create and update tables and provide them to their team.

Refer GCP documentation - [BigQuery Access Control]

`READER` Can read, query, copy or export tables in the dataset. Can read routines in the datasetCan call get on the datasetCan call get and list on tables in the datasetCan call get and list on routines in the datasetCan call list on table data for tables in the datasetMaps to the `bigquery.dataViewer` predefined role`WRITER` Same as `READER`, plus:Can edit or append data in the datasetCan call insert, insertAll, update or delete on tablesCan use tables in the dataset as destinations for load, copy or query jobsCan call insert, update, or delete on routinesMaps to the `bigquery.dataEditor` predefined role

Option A is wrong as WRITER access to data analysts would enable them to modify the data.

Options C & D are wrong as BigQuery access control works at the dataset level only.

Question 49: Correct

You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results

to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows

and manually execute them when needed. What should you do?

A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs. **(Correct)**

B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.

C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.

D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

Explanation

Correct answer is **A** as Cloud Composer allows you schedule and monitor jobs as well as the ability to manually execute them when needed.

Refer GCP documentation - [Cloud Composer]()

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers. Built on the popular Apache Airflow open source project and operated using the Python programming language, Cloud

Composer is free from lock-in and easy to use.

Cloud Composer pipelines are configured as directed acyclic graphs (DAGs) using Python, making it easy for users of any experience level to author and schedule a workflow

Cloud Composer is deeply integrated within the Google Cloud Platform, giving users the ability to orchestrate their *full pipeline. Cloud Composer has robust, built-in integration with many products, including Google BigQuery, Cloud Dataflow, Cloud Dataproc, Cloud Datastore, Cloud Storage, Cloud Pub/Sub, and Cloud ML Engine.*

Cloud Composer gives you the ability to connect your pipeline through a single orchestration tool whether your workflow lives on-premises, in multiple *clouds, or fully within GCP. The ability to author, schedule, and monitor your workflows in a unified manner means you can break down the silos in your environment and focus less on infrastructure.*

Options B, C & D are wrong as they do not satisfy all the requirements.

Question 50: **Correct**

You are a head of BI at a large

enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing

new projects to your account. What should you do?

A. Convert your batch BQ queries into interactive BQ queries.

B. Create an additional project to overcome the 2K on-demand per-project quota.

C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects. **(Correct)**

D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

Explanation

Correct answer is **C** as if more slots are needed, flat-rate pricing can be checked. Flat-rate pricing offers predictable and consistent month-to-month costs.

Refer GCP documentation - BigQuery Slots

Maximum concurrent slots per project for on-demand pricing — 2,000

The default number of slots for on-demand queries is shared among all queries in a single project. As a rule, if you're processing less than 100 GB of queries at once, you're unlikely to be using all 2,000 slots.

To check how many slots you're

using, see Monitoring BigQuery using Stackdriver. If you need more than 2,000 slots, contact your sales representative to discuss whether flat-rate pricing meets your needs.

BigQuery offers flat-rate pricing for customers who prefer a stable monthly cost for queries rather than paying the on-demand price per TB of data processed.

When you enroll in flat-rate pricing, you purchase dedicated query processing capacity which is measured in BigQuery slots. The cost of all bytes processed is included in the monthly flat-rate price. If your queries exceed your flat-rate capacity, your queries will run proportionally more slowly until more of your flat-rate resources *become available.*

Option A is wrong as concurrent slots limit apply for both batch and interactive queries

Option B is wrong as it does not meet the requirement of avoiding introducing new projects to the account.

Option D is wrong as you cannot increase the amount of concurrent slots per project beyond 2000.

Question 51: **Correct**

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?

A. Users are running too many concurrent queries in the system

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

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

D. Most rows in the `[myproject:mydataset.mytable]` table have the same value in the country column, causing data skew **(Correct)**

Explanation

Correct answer is **D** as the query plan indicates the average time spent in reading data and the time taken by the slowest worker. The difference is huge and the reason is mostly skewed data.

Refer GCP documentation - [BigQuery Query Plan Execution]

The query stages also provide stage timing classifications, in both relative and absolute form. As each stage of execution represents work undertaken by one or more independent workers, information is provided in both average and worst-case times, representing the average performance for all workers in a stage as well as the *long-tail slowest worker performance for a given classification. The average and max times are furthermore broken down into absolute and relative representations. For the ratio-based statistics, the data is provided as a fraction of the longest time spent by any worker in any segment.*

`readRatioAvg` `readMsAvg`

Time the average worker spent reading input data.`readRatioMax` `readMsMax`

Time the slowest worker spent reading input data.