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.

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.

Ans: B

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.

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.

Ans: A

Your infrastructure runs on another cloud and includes a set of multi-TB enterprise databases that are backed up nightly both on premises and also to that cloud. You need to create a redundant backup to Google Cloud. You are responsible for performing scheduled monthly disaster recovery drills. You want to create a cost-effective solution. What should you do?

A. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Nearline storage bucket as a final destination.

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

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

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

Ans: C

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.

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.

Ans: B

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.

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.

Ans: C

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.

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.

Ans: A

You are selecting a streaming service for log messages that must include final result message ordering as part of building a data pipeline on Google Cloud. You want to stream input for 5 days and be able to query the most recent message value. You will be storing the data in a searchable repository. How should you set up the input messages?

A. Use Cloud Pub/Sub for input. Attach a timestamp to every message in the publisher.

B. Use Cloud Pub/Sub for input. Attach a unique identifier to every message in the publisher.

C. Use Apache Kafka on Compute Engine for input. Attach a timestamp to every message in the publisher.

D. Use Apache Kafka on Compute Engine for input. Attach a unique identifier to every message in the publisher.

Ans: A

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.

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.

Ans: B

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.

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.

Ans: B

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 transformation to the Cloud Dataprep job.

Ans: D

You need to deploy a TensorFlow machine-learning model to Google Cloud. You want to maximize the speed and minimize the cost of model prediction and deployment. What should you do?

A. Export your trained model to a SavedModel format. Deploy and run your model on Cloud ML Engine.

B. Export your trained model to a SavedModel format. Deploy and run your model from a Kubernetes Engine cluster.

C. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud Storage. Run 1 version on CPUs and another version on GPUs.

D. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud ML Engine. Run 1 version on CPUs and another version on GPUs.

Ans: A

You have data stored in a Cloud Storage dataset and also in a BigQuery dataset. You need to secure the data and provide 3 different types of access levels for your Google Cloud Platform users: administrator, read/write, and read-only. You want to follow Google-recommended practices.What should you do?

A. Create 3 custom IAM roles with appropriate policies for the access levels needed for Cloud Storage and BigQuery. Add your users to the appropriate roles.

B. At the Organization level, add your administrator user accounts to the Owner role, add your read/write user accounts to the Editor role, and add your read-only user accounts to the Viewer role.

C. At the Project level, add your administrator user accounts to the Owner role, add your read/write user accounts to the Editor role, and add your read-only user accounts to the Viewer role.

D. Use the appropriate pre-defined IAM roles for each of the access levels needed for Cloud Storage and BigQuery. Add your users to those roles for each of the services.

Ans: D

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.

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.

Ans: B

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.

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

Ans: C

You are upgrading your existing (development) Cloud Bigtable instance for use in your production environment. The instance contains a large amount of data that you want to make available for production immediately. You need to design for fastest performance. What should you do?

A. Change your Cloud Bigtable instance type from Development to Production, and set the number of nodes to at least 3. Verify that the storage type is HDD.

B. Change your Cloud Bigtable instance type from Development to Production, and set the number of nodes to at least 3. Verify that the storage type is SSD.

C. Export the data from your current Cloud Bigtable instance to Cloud Storage. Create a new Cloud Bigtable Production instance type with at least 3 nodes. Select the HDD storage type. Import the data into the new instance from Cloud Storage.

D. Export the data from your current Cloud Bigtable instance to Cloud Storage. Create a new Cloud Bigtable Production instance type with at least 3 nodes. Select the SSD storage type. Import the data into the new instance from Cloud Storage.

Ans: B

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.

B. Export the snapshots to Cloud Storage. Create images from the exported snapshot files.

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.

Ans: D

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.

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.

Ans: B

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 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.

Ans: B

You created a job which runs daily to import highly sensitive data from an on-premises location to Cloud Storage. You also set up a streaming data insert into Cloud Storage via a Kafka node that is running on a Compute Engine instance. You need to encrypt the data at rest and supply your own encryption key. Your key should not be stored in the Google Cloud. What should you do?

A. Create a dedicated service account, and use encryption at rest to reference your data stored in Cloud Storage and Compute Engine data as part of your API service calls.

B. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in Cloud Storage. Use your uploaded encryption key and reference it as part of your API service calls to encrypt your data in the Kafka node hosted on Compute Engine.

C. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in your Kafka node hosted on Compute Engine.

D. Supply your own encryption key, and reference it as part of your API service calls to encrypt your data in Cloud Storage and your Kafka node hosted on Compute Engine.

Ans: D

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.

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.

Ans: B

<https://github.com/jorwalk/data-engineering-gcp/blob/master/practice-exam/data-engineer.md>