Google Professional Data Engineer Exam Actual Questions

---

1) Your company built a TensorFlow neutral-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

- A. Threading
- B. Serialization
- C. Dropout Methods
- D. Dimensionality Reduction

Dropout Methods :
* What is the dropout method? -> The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much.
Other options : Early Stopping / L1 and L2 Regularization / Dropout / Max-Norm Regularization / Data Augmentation

Option A - wrong (as Threading is to make training faster)
Option B - wrong (Serialization) used while saving the model
Option D - Wrong (Dimensionality Reduction) - This is though core parameter, but in question it's mention model works okay on Training Data.. So dimension is not the issue here

Answer is **Dropout Methods**

There are various ways to prevent overfitting when dealing with DNNs. In this post, we'll review these techniques and then apply them specifically to TensorFlow models:
- Early Stopping
- L1 and L2 Regularization
- Dropout
- Max-Norm Regularization
- Data Augmentation

D is not correct here because it is used for normal ml models whereas dropout methods is used for neural networks.

Reference:
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877

---

2) 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.
- 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.

In Recommendation System - Matrix stored in database - which store the users rating and other details like how much time user spent on that page. We generally recommend on basis of matrix. And In production that matrix updated (or model get retrained once a day or once a week) - Because there could be new users rating. or we can say new data.

So model retrained fully i.e. on Existing Data + New Data => and generate a new matrix (user/item matrix)

Reference:
https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html
Stream new Data back to the model as it becomes available

---

3) You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

- A. Add capacity (memory and disk space) to the database server by the order of 200.
- B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
- C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.
- D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

C is correct because this option provides the least amount of inconvenience over using pre-specified date ranges or one table per clinic while also increasing performance due to avoiding self-joins.
A is not correct because adding additional compute resources is not a recommended way to resolve database schema problems.
B is not correct because this will reduce the functionality of the database and make running reports more difficult.
D is not correct because this will likely increase the number of tables so much that it will be more difficult to generate reports vs. the correct option.
https://cloud.google.com/bigquery/docs/best-practices-performance-patterns

https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax#explicit-alias-visibility

---

4) 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.

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 virtualizations.

correct answer -> Disable caching by editing the report settings.

A cache is a temporary data storage system. Fetching cached data can be much faster than fetching it directly from the underlying data set, and helps reduce the number of queries sent, minimizing costs for paid data access.

Reference:

https://support.google.com/datastudio/answer/7020039?hl=en#zippy=%2Cin-this-article

---

5) An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values(CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?

A. Use federated data sources, and check data in the SQL query.

B. Enable BigQuery monitoring in Google Stackdriver and create an alert.

C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.

D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

The answer is D. An ETL pipeline will be implemented for this scenario. Check out handling invalid inputs in cloud data flow https://cloud.google.com/blog/products/gcp/handling-invalid-inputs-in-dataflow

Handling invalid inputs in Dataflow using Side Outputs as a "Dead Letter" file. The structure is known as a dead letter pattern.

---

6) Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?

A. Issue a command to restart the database servers.

B. Retry the query with exponential backoff, up to a cap of 15 minutes.

C. Retry the query every second until it comes back online to minimize staleness of data.

D. Reduce the query frequency to once every hour until the database comes back online.

Correct answer is B. App engine create applications that use Cloud SQL database connections effectively. Below is what is written in google cloud documentation.

If your application attempts to connect to the database and does not succeed, the database could be temporarily unavailable. In this case, sending too many simultaneous connection requests might waste additional database resources and increase the time needed to recover. Using exponential backoff prevents your application from sending an unresponsive number of connection requests when it can't connect to the database.

This retry only makes sense when first connecting, or when first grabbing a connection from the pool. If errors happen in the middle of a transaction, the application must do the retrying, and it must retry from the beginning of a transaction. So even if your pool is configured properly, the application might still see errors if connections are lost.

reference link is https://cloud.google.com/sql/docs/mysql/manage-connections

---

7) 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 regression

B. Logistic classification

C. Recurrent neural network

D. Feedforward neural network

Correct answer -> Linear regression is a statistical method that allows to summarize and study relationships between two continuous (quantitative) variables: One variable, denoted X, is regarded as the independent variable. The other variable denoted y is regarded as the dependent variable. Linear regression uses one independent variable X to explain or predict the outcome of the dependent variable y. Whenever you are told to predict some future value of a process which is currently running, you can go with a regression algorithm. Reference: https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60

A tip here to decide when a liner regression should be used or logistics regression needs to be used. If you are forecasting that is the values in the column that you are predicting is numeric, it is always **linear regression**. If you are classifying, that is buy or no buy, yes or no, you will be using **logistics regression.**

---

8) 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 answer -> Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

You can use the ROW_NUMBER() to turn non-unique rows into unique rows and then delete the duplicate rows.

Reference:

https://www.mysqltutorial.org/mysql-window-functions/mysql-row_number-function/

A is not correct because this will just return one row.

B is not correct because this doesn't get you the latest value, but will get you a sum of the same event over time which doesn't make too much sense if you have duplicates.

C is not correct because if you have events that are not duplicated, it will be excluded.

---

9) 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 answer : D BUT, it should be ` as a first character and not "

---

10) 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.

E. Segregate data across multiple tables or databases.

F. Use Google Stackdriver Audit Logging to determine policy violations.

correct option -> B. Restrict access to tables by role. Reference: https://cloud.google.com/bigquery/docs/table-access-controls-intro

correct option-> D. Restrict BigQuery API access to approved users. ***Only approved users will have access which means other users will have minimum amount of information required to do their job.*** Reference: https://cloud.google.com/bigquery/docs/access-control

correct option -> F. Use Google Stackdriver Audit Logging to determine policy violations. Reference: https://cloud.google.com/bigquery/docs/table-access-controls-intro#logging

A. Disable writes to certain tables. ---> Read is still available(not minimal access) C. Ensure that the data is encrypted at all times. ---> Data is encrypted by default. E. Segregate data across multiple tables or databases. ---> Normalization is of no help here.

---

11) You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:

☞ No interaction by the user on the site for 1 hour

☞ Has added more than $30 worth of products to the basket

☞ 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.

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

The correct answer is C. There are 3 windowing concepts in dataflow and each can be used for below use case
1) Fixed window
2) Sliding window
3) Session window.

Fixed window = any aggregation use cases, any batch analysis of data, relatively simple use cases.
Sliding window = Moving averages of data
Session window = user session data, click data and real time gaming analysis. The question here is about user session data and hence session window.

---

12) 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.

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

D. Restrict a client's dataset to approved users.

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.

---

13) 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

B. BigQuery

C. Cloud Bigtable

D. Cloud Datastore

---

14) 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.

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.

E. You already have labels for which samples are mutated and which are normal in the database.

Anomaly detection has two basic assumptions:

*Anomalies only occur very rarely in the data.

*Their features differ from the normal instances significantly.

Anomaly detection involves identifying rare data instances (anomalies) that come from a different class or distribution than the majority (which are simply called "normal" instances). Given a training set of only normal data, the semi-supervised anomaly detection task is to identify anomalies in the future. Good solutions to this task have applications in fraud and intrusion detection. The unsupervised anomaly detection task is different: Given unlabeled, mostly-normal data, identify the anomalies among them. https://www.science.gov/topicpages/u/unsupervised+anomaly+detection

A because "Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal", B is for Supervised anomaly detection https://en.wikipedia.org/wiki/Anomaly_detection

---

15) 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, design the application 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.

Description: The data is first comes to buffer and then written to Storage. If we are running queries in buffer we will face above mentioned issues. If we wait for the bigquery to write the data to storage then we won't face the issue. So We need to wait till it's written to storage

D - speaks about near real-time approach. None other.

---

16) 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.

B. Get the identity and access management IIAM) 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.

A is correct because this is the best way to get granular access to data showing which users are accessing which data.

B is not correct because we already know that all users already have access to all data, so this information is unlikely to be useful. It will also not show what users have done, just what they can do.

C is not correct because slot usage will not inform security policy.

D is not correct because a billing account is typically shared among many people and will only show the amount of data queried and stored
https://cloud.google.com/bigquery/docs/reference/auditlogs/#mapping-audit-entries-to-log-streams

https://cloud.google.com/bigquery/docs/monitoring#slots-available

---

17) 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.

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

D is correct because it uses managed services, and also allows for the data to persist on GCS beyond the life of the cluster.

A is not correct because the goal is to re-use their Hadoop jobs and MapReduce and/or Spark jobs cannot simply be moved to Dataflow.

B is not correct because the goal is to persist the data beyond the life of the ephemeral clusters, and if HDFS is used as the primary attached storage mechanism, it will also disappear at the end of the cluster's life.

C is not correct because the goal is to use managed services as much as possible, and this is the opposite.

E is not correct because the goal is to use managed services as much as possible, and this is the opposite.

Dataproc is used to migrate Hadoop and Spark jobs on GCP. Dataproc with GCS connected through Google Cloud Storage connector helps store data after the life of the cluster. When the job is high I/O intensive, then we need to create a small persistent disk.

---

18) Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use? (Choose three.)

A. Supervised learning to determine which transactions are most likely to be fraudulent.

B. Unsupervised learning to determine which transactions are most likely to be fraudulent.

C. Clustering to divide the transactions into N categories based on feature similarity.

D. Supervised learning to predict the location of a transaction.

E. Reinforcement learning to predict the location of a transaction.

F. Unsupervised learning to predict the location of a transaction.

B - Not labelled as Fraud or not. So Unsupervised.

C - Clustering can be done based on location, amount etc.

D - Location is already given. So labelled. Hence supervised.

Answer is **B-C-D**

Fraud is not a feature, so unsupervised, location is given so supervised, Clustering can be done looking at the done with same features.

Say the model predict a location, guessing US or Sweden are both wrong when the answer is Canada. But US is closer, the distance from the correct location can be used to calculate a reward. Through reinforcement learning (E) the model could guess a location with better accuracy than supervised (D).

Supervised : With labels / Unsupervised : Without labels

---

19) Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc**. A like-for-like migration** of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

A. Put the data into Google Cloud Storage.

B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.

C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.

D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

A: First rule of dataproc is to keep data in GCS

B: Wrong eVM won't solve the problem of larger storage prices.

C: May be, but nothing mentioned in terms of what to tune in the question, also this is like-for-like migration so tuning may not be part of the migration.

D: Again, this is like-for-like so need to define what is hot data and which is cold data, also persistent disk costlier than cloud storage.

---

20) You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous

events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

A. The message body for the sensor event is too large.

B. Your custom endpoint has an out-of-date SSL certificate.

C. The Cloud Pub/Sub topic has too many messages published to it.

D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

The Answer should be D. The custom endpoint is not acknowledging the message, that is the reason for Pub/Sub to send the message again and again. Not B.

---

21) 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.

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.

Description: Using Hash values we can remove duplicate values from a database. Hash values will be the same for duplicate data and thus can be easily rejected.

---

22) 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 answer -> Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine. Google says: Use Cloud Datalab to easily explore, visualize, analyze, and transform data using familiar languages, such as Python and SQL, interactively. Reference: https://cloud.google.com/datalab/docs

This is exactly the tool needed in this scenario But what about her laptop slowing her down? Well, Cloud Datalab is packaged as a container and run in a VM (Virtual Machine) instance. Reference: https://cloud.google.com/datalab/docs/concepts/key-concepts

Cloud Datalab is **packaged as a container and run in a VM (Virtual Machine) instance.** VM creation**, running the container in that VM, and establishing a connection from your browser to the Cloud Datalab container, which allows you to open existing Cloud Datalab notebooks and create new notebooks**. Read through the introductory notebooks in the `/docs/intro` directory to get a sense of how a notebook is organized and executed.

Cloud Datalab uses **notebooks instead of the text files containing code. Notebooks bring together code, documentation written as markdown, and the results of code execution—whether as text, image or, HTML/JavaScript**.

Reference:
https://cloud.google.com/datalab/docs/quickstarts

---

23) 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. What should you do?

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.

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.

correct answer is Cloud Pub/Sub ---> Dataflow ---> Bigquery Collected messages containing temperature values will be published to a topic on Cloud Pub/Sub, Messages will be read in streaming mode by Cloud Dataflow, a simplified stream and batch data processing solution, Cloud Datastore will save data to be displayed directly into the UI of the App Engine application, while BigQuery will act as a data warehouse that will enable the execution of more in depth analysis.
Reference: https://cloud.google.com/blog/products/iot-devices/quick-and-easy-way-set-end-end-iot-solution-google-cloud-platform

---

24) 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 TS 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. Add two columns to the table CLICK STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEW to true. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.

E. 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.

E : Export and load data into a new table. You can also change a column's data type by exporting your table data to Cloud Storage, and then loading the data into a new table with a schema definition that specifies the correct data type for the column. You can also use the load job to overwrite the existing table. Advantages You are not charged for the export job or the load job. Currently, BigQuery load and export jobs are free. If you use the load job to overwrite the original table, you incur storage costs for one table instead of two, but you lose the original data. Disadvantages If you load the data into a new table, you incur storage costs for the original table and the new table (unless you delete the old one). You incur costs for storing the exported data in Cloud Storage.

Creating a new table from existing table in BigQuery with new transformed column will be simple and will not involve and migration effort. Also future query performance will improve.

---

25) You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.

B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.

C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.

D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

correct answer -> Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.
Option C is also most likely right answer but it doesn't have the filter. We don't want all the tables. We only want one. So the correct answer is D. Logging sink - Using a Logging sink, you can direct specific log entries to your business logic. In this example, you can use Cloud Audit logs for Compute Engine which use the resource type gce_firewall_rule to filter for the logs of interest. You can also add an event type GCE_OPERATION_DONE to the filter to capture only the completed log events. Here is the Logging filter used to identify the logs. You can try out the query in the Logs Viewer. Pub/Sub topic – In Pub/Sub, you can create a topic to which to direct the log sink and use the Pub/Sub message to trigger a cloud function. Reference:
https://cloud.google.com/blog/products/management-tools/automate-your-response-to-a-cloud-logging-event

---

26) 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.

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.

The answer should be D. We are asking the consultant to implement the code, not to run it in the real infrastructure. By providing a sample anonymised dataset, they will be able to do the work. They don't even need any DataFlow permission in the new project, as they can use a DirectRunner to run a pipeline locally: https://cloud.google.com/dataflow/docs/concepts/security-and-permissions#security_and_permissions_for_local_pipelines/

https://cloud.google.com/dataflow/docs/guides/setting-pipeline-options#LocalExecution

OR

Answer is **Grant the consultant the Cloud Dataflow Developer role on the project.**

The Dataflow developer role will not provide access to the underlying data.

Reference:
https://cloud.google.com/dataflow/docs/concepts/access-control#example_role_assignment

---

27) You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

A. Eliminate features that are highly correlated to the output labels.

B. Combine highly co-dependent features into one representative feature.

C. Instead of feeding in each feature individually, average their values in batches of 3.

D. Remove the features that have null values for more than 50% of the training records.

This method is called Data Engineering, that you combine two or more values to get a custom info, this will avoid that the model read an extra column on the training and probably increase its accuracy.

A: correlated to output means that feature can contribute a lot to the model. so not a good idea.
C: you need to run with almost same number, but you will iterate twice, once for averaging and second time to feed the averaged value.
D: removing features even if it 50% nulls is not good idea, unless you prove that it is not at all correlated to output. But this is nowhere so can remove.

---

28) 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.

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.

B is correct: I think both read.from and read.fromQuery will use GCS as the intermediate step, but fromQuery can greatly reduce the amount of data (and is a Best Practice in BQ to query only the columns that are required), so the processing performance should be increased.

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).

---

29) 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>.

A. Use a row key of the form <timestamp>. ---> Incorrect, because google says don't use a timestamp by itself or at the beginning of a row key.
B. Use a row key of the form <sensorid>. ---> Incorrect, because google says Include a timestamp as part of your row key.
C. Use a row key of the form <timestamp>#<sensorid>. ---> Incorrect, because google says don't use a timestamp by itself or at the beginning of a row key.

D. Use a row key of the form >#<sensorid>#<timestamp>. ---> Correct answer, because of option A,B,C reasons. - Timestamp isn't by itself, neither at the beginning. - Timestamp is included.

Best practices of bigtable states that rowkey should not be only timestamp or have timestamp at starting. It's better to have sensorid and timestamp as rowkey

---

30) 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.

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.

D Is 0 load for the operating mysql, because you are loading already done backup and processing it using your spark analytics in dataproc. But B will have a big impact in production, but will be the best option for future OLAP analysis over the past data. The problem with that option is that it doesn't specify if the ETL is performed daily, hourly, or in low load periods, so the MYSQL server will be overloaded meanwhile the ETL is loading the data to bigquery.

A: OLAP on MySQL performs poorly.
B: ETL consumes lot of MySQL resources, to read the data, as per question MySQL is under pressure already.
C: Similar to B.
D: By mounting backup can avoid reading from MySQL, data freshness is not an issue as per the question (and is not mention in the question).

Reference:
https://cloud.google.com/blog/products/data-analytics/genomics-data-analytics-with-cloud-pt2

---

31) 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.

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.

Answer is B. You can update the pipeline. Submit a new job with same job name and pass the transform mapping JSON object as a parameter to the new job. Dataflow will take care of draining

the first job and starting the updated job.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

OR (preferable) :

Question mentions not to lose any data. All other options may lead to some data loss. If you want to prevent data loss as you bring down your streaming pipelines. 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.

Reference:
https://cloud.google.com/dataflow/docs/guides/stopping-a-pipeline#drain

---

32) 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.

B. The performance issue should be resolved over time as the site of the BigDate 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.

correct answer -> Redefine the schema by evenly distributing reads and writes across the row space of the table. Make sure you're reading and writing many different rows in your table. Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Bigtable distribute the workload across all of the nodes in your cluster. If reads and writes cannot be spread across all of your Bigtable nodes, performance will suffer. If you find that you're reading and writing only a small number of rows, you might need to redesign your schema so that reads and writes are more evenly distributed. Reference:
https://cloud.google.com/bigtable/docs/performance#troubleshooting

---

33) 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.

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.

Answer is **Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.**

Stack driver could tell us about performance but not logging of missing data.

B as the issue can be debugged by running a fixed dataset and checking the output. Refer GCP documentation - Dataflow logging: https://cloud.google.com/dataflow/docs/guides/logging
A is wrong as the Dashboard uses data provided by Dataflow, the input source for Dashboard seems to be the issue
C is wrong as Monitoring will not help find missing messages in Cloud Pub/Sub
D is wrong as Dataflow cannot be configured as Push endpoint with Cloud Pub/Sub.

Flowlogistic Case Study -

Company Overview -

Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.

Company Background -

The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

☞ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads

☞ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

☞ Databases

8 physical servers in 2 clusters

- SQL Server "" user data, inventory, static data

3 physical servers

- Cassandra "" metadata, tracking messages

10 Kafka servers "" tracking message aggregation and batch insert

☞ Application servers "" customer front end, middleware for order/customs

60 virtual machines across 20 physical servers

- Tomcat "" Java services

- Nginx "" static content

- Batch servers

☞ Storage appliances

- iSCSI for virtual machine (VM) hosts

- Fibre Channel storage area network (FC SAN) "" SQL server storage

- Network-attached storage (NAS) image storage, logs, backups

☞ 10 Apache Hadoop /Spark servers

- Core Data Lake

- Data analysis workloads

☞ 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements -

☞ Build a reliable and reproducible environment with scaled panty of production.

☞ Aggregate data in a centralized Data Lake for analysis

☞ Use historical data to perform predictive analytics on future shipments

☞ Accurately track every shipment worldwide using proprietary technology

☞ Improve business agility and speed of innovation through rapid provisioning of new resources

☞ Analyze and optimize architecture for performance in the cloud

Migrate fully to the cloud if all other requirements are met

Technical Requirements -

☞ Handle both streaming and batch data

☞ Migrate existing Hadoop workloads

☞ Ensure architecture is scalable and elastic to meet the changing demands of the company.

☞ Use managed services whenever possible

☞ Encrypt data flight and at rest

☞ Connect a VPN between the production data center and cloud environment

SEO Statement -

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement -

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.

Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.

You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage

B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD

C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage

D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage

A. Only A can manage a lot's of data. Target is Cloud Storage (obviously not SSD) Input is Pub/Sub to replace Kafka Cloud SQL + Storage has no sense in this context
B is incorrect, because local SSD wouldn't satisfy the needs.
C is incorrect, because one of the requirements is 'Global', Cloud SQL is well suited for regional applications. Cloud Spanner is a better suit in that regard.
D is incorrect, because Load Balancer is for web traffic, not messages.

(Question 42 on the exam topic, page 9/49)

34) 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.

C. Increase the size of the Hadoop cluster.

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

Answer: B Description: Spark performs in-memory processing and faster, which results in optimization of job's processing time

---

35) 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.

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.

There are two ways to manually rename a column:

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.

Similarly, if you need to create a new column, using Update query is not cost-effective. Creating a new table is cost-effective way.

Reference:
https://cloud.google.com/bigquery/docs/manually-changing-schemas?hl=en#changing_a_columns_name

---

36) You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity "~Movie' the property "~actors' and the property"~tags' have multiple values but the

property "~date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

A. Manually configure the index in your index config as follows:

```
Indexes:
-kind: Movie
    Properties:
    -name: actors
    name: date_released
-kind: Movie
    Properties:
    -name: tags
    name: date_released
```

B. Manually configure the index in your index config as follows:

```
Indexes:
    -kind: Movie
        Properties:
        -name: actors
        -name: tags
-name: date_published
```

C. Set the following in your entity options: exclude_from_indexes = "~actors, tags'

D. Set the following in your entity options: exclude_from_indexes = "~date_published'

Correct Answer: A as it is better to manually configure index with separate properties. Refer GCP documentation - Datastore Indexes:
https://cloud.google.com/datastore/docs/concepts/indexes#index_limits

When the same property' is repeated multiple times, Datastore mode can detect exploding indexes and suggest an alternative index. However, in all other circumstances (such as the query defined in this example), a Datastore mode database will generate an exploding index. In this case, you can circumvent the exploding index by manually configuring an index in your index configuration file: indexes: - kind: Task properties: - name: tags - name: created - kind: Task properties: - name: collaborators - name: created

---

37) 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 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.

Description: Scheduler for adhoc jobs – 3 jobs free and $0.10 per job

https://cloud.google.com/appengine/docs/flexible/nodejs/scheduling-jobs-with-cron-yaml

38) You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible.

What should you do?

A. Load the data every 30 minutes into a new partitioned table in BigQuery.

B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery

C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore

D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

B is correct. we need it as cheaper as possible so cloud storage + external tables. Data will be overwritten every 30 min https://cloud.google.com/blog/products/gcp/accessing-external-federated-data-sources-with-bigquerys-data-access-layer

regional storage is cheaper than BigQuery storage.

Use cases for external data sources include:

Loading and cleaning your data in one pass by querying the data from an external data source (a location external to BigQuery) and writing the cleaned result into BigQuery storage.
Having a small amount of frequently changing data that you join with other tables. As an external data source, the frequently changing data does not need to be reloaded every time it is updated.

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

39) You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:

☞ The user profile: What the user likes and doesn't like to eat

☞ The user account information: Name, address, preferred meal times

☞ The order information: When orders are made, from where, to whom

The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

A. BigQuery

B. Cloud SQL

C. Cloud Bigtable

D. Cloud Datastore

Datastore is NoSQL hence Schemaless, so good luck optimizing a schema with it. BQ allows both schema optimization and ML
https://cloud.google.com/datastore/docs/concepts/overview#:~:text=Unlike%20traditional%20relational%20databases%20which%20enforce%20a%20schema%2C%20Datastore%20is%20schemaless.

BUT:

Answer is **Cloud SQL**

The database will be used to store all the transactional data of the product. what we need is the database only for store transactional data, not for analysis and ML. so the answer should be "the database that stores transactional data", which means, Cloud SQL. if you want to analyze or do ML you just specify Cloud SQL as a federated data source.

A: it's good for analysis but it costs too much to input/output data frequently.
C: BigTable is not good for transactional data.
D: okay datastore supports transactions, but it is weaker than RDB, and also, in this case, the data schema has already defined , you should use RDB.

---

40) 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 is not using BigQuery's default encoding.

D. The CSV data has not gone through an ETL phase before loading into BigQuery.

A is not correct because if another data format other than CSV was selected then the data would not import successfully.
B is not correct because the data was fully imported meaning no rows were skipped.
C is correct because this is the only situation that would cause successful import.
D is not correct because whether the data has been previously transformed will not affect whether the source file will match the BigQuery table.

---

41) Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low. You are told that due to seasonality, your

company expects the number of files to double for the next three months. Which two actions should you take? (Choose two.)

A. Introduce data compression for each file to increase the rate file of file transfer.

B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.

C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.

D. Assemble 1,000 files into a tape archive (TAR) file. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.

E. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.

20k files * 24 hours = 480k files x2 * 4kilobyte= 3.8gb everyday and must be processed in 10 hours - either C or E (not both)= total size is less than 1tb per day, so C (gsutil) is the right tool - in gsutil, the process can be paralleled (so you can utilize bandwidth throughput) and using rsync it is also possible to compress (to increase the transfer rate). and you do not need to decompress at target (to decrease process time at target such as untar or decompress). So A is better than E.! the following cmd does the job. gsutil -m rsync -az sourceDir gs://targetBucket

A - would minimize over all size during seasonal times C - gcp standard for skinny files

Answers are; **A. Introduce data compression for each file to increase the rate file of file transfer. C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.**

B - wrong (we need to provide solution without changing internet speed)
D - if we TAR 1000 files, its okay, but Volume is getting increased continously..How we define the number ?
E - Bandwidth already low, so storage Transfer service will not help here.

Follow these rules of thumb when deciding whether to use gsutil or Storage Transfer Service:

Transfer scenario Recommendation
Transferring from another cloud storage provider Use Storage Transfer Service.
Transferring less than 1 TB from on-premises Use gsutil.
Transferring more than 1 TB from on-premises Use Transfer service for on-premises data.
Transferring less than 1 TB from another Cloud Storage region Use gsutil.
Transferring more than 1 TB from another Cloud Storage region Use Storage Transfer Service.

---

42) 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

C. MySQL

D. MongoDB

E. Cassandra

F. HDFS with Hive

Answer is BDE -
A. Redis - Redis is an in-memory non-relational key-value store. Redis is a great choice for implementing a highly available in-memory cache to decrease data access latency, increase throughput, and ease the load off your relational or NoSQL database and application. Since the question does not ask cache, A is discarded.
B. HBase - Meets reqs
C. MySQL - they do not need ACID, so not needed.
D. MongoDB - Meets reqs
E. Cassandra - Apache Cassandra is an open source NoSQL distributed database trusted by thousands of companies for scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data.
F. HDFS with Hive - Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data. HIVE IS NOT A DATABSE.

---

43) 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

B. Reduce the number of training examples

C. Use a smaller set of features

D. Use a larger set of features

E. Increase the regularization parameters

F. Decrease the regularization parameters

The tools to prevent overfitting: less variables, regularization, early ending on the training.
- Adding more training data will increase the complexity of the training set and help with the variance problem.
- Reducing the feature set will ameliorate the overfitting and help with the variance problem.
- Increasing the regularization parameter will reduce overfitting and help with the variance problem.

---

44) 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

D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Data owners can't create jobs or queries. -> B out
We need service Account -> D out
Access only granting me does not solve the problem -> A out
The answer is C. (Minimum rights to perform the job)
Service Account is the best way 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.
https://cloud.google.com/bigquery/docs/authentication

---

45) 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

D Image says that average(dark) and maximum(light) have difference in few times, this it is a skew
https://cloud.google.com/bigquery/query-plan-explanation

Purple is reading, Blue is writing. so majority is reading.

Partition skew, sometimes called data skew, is when data is partitioned into very unequally sized partitions. This creates an imbalance in the amount of data sent between slots. You can't share partitions between slots, so if one partition is especially large, it can slow down, or even crash the slot that processes the oversized partition.

Reference:
https://cloud.google.com/bigquery/docs/best-practices-performance-patterns

---

46) 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.

A and C is out since it does not give the real-time feature.
If we think about the business scenario, we need to give the bid to who published it first. D is given to the who processed first. Technically it can be implemented because DataFlow is only support pull subscription with cloud pub-sub. In answer B, events are pushed to the endpoint. Explicitly they haven't mentioned that it pushes to the "Cloud Data Flow". It may be a custom API endpoint. Some people have a misunderstanding about this point. There is no clue about that. It may be a REST API endpoint or application. The event consists of the timestamp field itself. So by inserting it into the D B we can find out who is the winner. Based on that we can Mark answer as B.

OR

Answer D : The need is to collate the messages in real-time. We need to de-dupe the messages based on timestamp of when the event occurred. This can be done by publishing ot Pub-Sub and consuming via Dataflow.

Reference:
https://stackoverflow.com/questions/62997414/push-vs-pull-for-gcp-dataflow

---

47) 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

E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"

C = A standard SQL query cannot reference a view defined using legacy SQL syntax.
D = For the ODBC drivers is needed a service account which will get a standard Bigquery role.

---

48) 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

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

Answer: A Description: Legacy sql uses table date range whereas standard sql uses table_sufix for wildcard

---

49) 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

Answer: D Description: Caution: 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 PCollections. 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. See Setting your PCollection's 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. So it looks like D

---

50) You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.

B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.

C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.

D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Should go with B. Two reasons, it is a cleaner approach with single job to handle the calibration before the data is used in the pipeline. Second, doing this step in later stages can be complex and maintenance of those jobs in the future will become challenging.

---

51) An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application. They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?

A. BigQuery

B. Cloud SQL

C. Cloud BigTable

D. Cloud Datastore

B - Cloud SQL should be the only correct answer. Required solution needs to support transactions as well as analysis through a BI tool. Firestore/Datastore does not support SQL syntax typically needed to do analysis done by a BI tool. BigQuery is not suitable for transactional use case. BigTable does not support SQL.

---

52) 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

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

Description: Google says that when you have multiple wildcard tables, best option is to shard it into single partitioned table. Time and cost efficient
Partitioning > table sharding: https://cloud.google.com/bigquery/docs/partitioned-tables#dt_partition_shard

---

53) 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

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

Preemptible workers are the default secondary worker type. They are reclaimed and removed from the cluster if they are required by Google Cloud for other tasks. Although the potential removal of preemptible workers can affect job stability, 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 https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms

Hadoop/Spark jobs are run on Dataproc, and the pre-emptible machines cost 80% less

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

---

54) 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.

D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

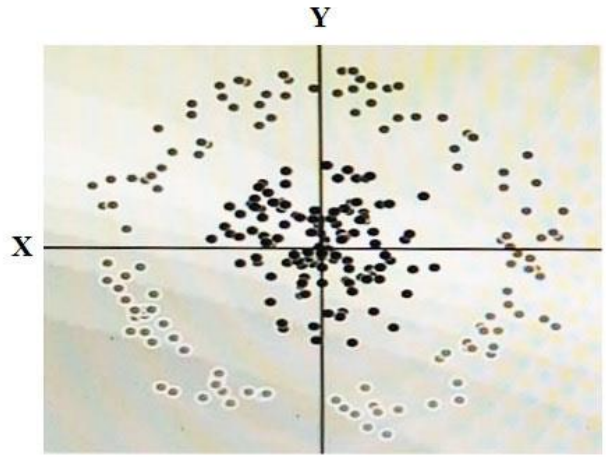
A is a direct No, if data don't have timestamp, we'll only have the processing time and not the "event time".
B is not either, sliding windows are not for this. Hopping|sliding windowing is useful for taking running averages of data, but not to process late data.
D looks correct but has one concept missing, the watermark to know if the process time is ok with the event time or not. I'm not 100% sure is incorrect. If, since we have a "predictable time period", might be this will do. I mean, if our dashboard is shown after the last input data has arrived (single global window), this should be ok. We'd have a "perfect watermark". Anyway we'd need triggering .

C is, I think, the correct answer: Watermark is different from late data. Watermark in implementation is a monotonically increasing timestamp. When Beam/Dataflow see a record with an event

timestamp that is earlier than the watermark, the record is treated as late data. I'll try to explain: Late data is inherent to Beam's model for out-of-order processing. What does it mean for data to be late? The definition and its properties are intertwined with watermarks that track the progress of each computation across the event time domain. The simple intuition behind handling lateness is this: only late input should result in late data anywhere in the pipeline. So, is not easy to decide between C and D. If you ask me I'd say C since for D we ought to make some suppositions.

---

55) You have some data, which is shown in the graphic below. The two dimensions are X and Y, and the shade of each dot represents what class it is. You want to classify this data accurately using a linear algorithm. To do this you need to add a synthetic feature. What should the value of that feature be?



A. X^2+Y^2

B. X^2

C. Y^2

D. cos(X)

For fitting a linear classifier when the data is in a circle use A.

---

56) 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

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

The answer is C. When access data through application, Google recommendation is using service account.

---

57) 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.

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 0 using a custom script.

Dataprep is the tool. A or B. Since they need to have a real-valued cannot be null N/A or empty, have to be "0", so it has to be B.

---

58) 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.

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.

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.

A makes no sense, you need to use your own keys. You don't create keys locally and upload them, you should import it to make it work..using the kms public key…not just "uploading" it.
C is also out.
IT's between B and D Cloud KMS is a cloud-hosted key management service that lets you manage cryptographic keys for your cloud services the same way you do on-premises, You can generate, use, rotate, and destroy cryptographic keys from there. Since you want to encrypt data at rest, is B, you don't use them for any API calls. https://cloud.google.com/compute/docs/disks/customer-managed-encryption

---

59) 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 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.

C. The cloud video intelligence API does the label generation without the need of building any model, A and B are excluded. Now, the bbdd most suitable for this is bigtable and not SQL (this big joins would be anything but fast). https://cloud.google.com/video-intelligence/docs/feature-label-detection

---

60) 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.

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.

C. Dataproc does not seem to be a good solution in this case as it always require a manual intervention to adjust resources. Autoscaling with dataflow will automatically handle changing data volumes with no manual intervention, and monitoring through Stackdriver can be used to spot bottleneck. Total execution time is not good there as it does not provide a precise view on potential bottleneck.

---

61) 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 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.

Destination is GCS and having multi-regional so A is the best option available.

Even since BigQuery Data Transfer Service initially supports Google application sources like Google Ads, Campaign Manager, Google Ad Manager and YouTube but it does not support destination anything other than bq data set

> Currently, you cannot use the BigQuery Data Transfer Service to transfer data out of BigQuery. The data has to be analysed "worldwide", says the brief. To have perfomant queries, storage should be worldwide https://cloud.google.com/bigquery/external-data-sources?hl=en#data-locations

> When you query data in an external data source such as Cloud Storage, the data you're querying must be in the same location as your BigQuery dataset. So, A

---

62) 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.

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.

If you use gzip compression BigQuery cannot read the data in parallel....C and D are out since you want to paralelize. BigQuery supports standard ANSI SQL 2011, but also BigQuery is capable of making SQL queries against external data sources, such as log files in Cloud Storage, transactional records in Cloud Bigtable, and many other kinds of data outside BigQuery. I would say A since we don't have a data lake strategy and there is no advantage in having data in cloud storage. https://cloud.google.com/bigquery/docs/loading-data

Correct answer : A or B

A and B are correct, but B is the best answer

The advantages of creating external tables are that they are fast to create so you skip the part of importing data and no additional monthly billing storage costs are accrued to your account since you only get charged for the data that is stored in the data lake, which is comparatively cheaper than storing it in BigQuery

A : Importing data into BigQuery may take more time compared to creating external tables on data. Additional storage costs by BigQuery is another issue which can be more expensive than Google Storage.

---

63) You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.

B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.

C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.

D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Correct Answer : A Entity analysis -> Identify entities within documents receipts, invoices, and contracts and label them by types such as date, person, contact information, organization, location, events, products, and media.
Sentiment analysis -> Understand the overall opinion, feeling, or attitude sentiment expressed in a block of text. -- Avoid Custom models

---

66) 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 for query.

D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

BigQuery can access data in external sources, known as federated sources. Instead of first loading data into BigQuery, you can create a reference to an external source. External sources can be Cloud Bigtable, Cloud Storage, and Google Drive.

When accessing external data, you can create either permanent or temporary external tables. Permanent tables are those that are created in a dataset and linked to an external source. Dataset-level access controls can be applied to these tables. When you are using a temporary table, a table is created in a special dataset and will be available for approximately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse.

---

67) 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.

D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

A is not correct because Cloud SQL does not natively scale horizontally.
Same for B.
C is correct because Cloud Spanner scales horizontally, and you can create secondary indexes for the range queries that are required.
D is not correct because Dataflow is a data pipelining tool to move and transform data, but the use case is centered around querying.

Answer is **Use Cloud Spanner for storage. Add secondary indexes to support query patterns.**

Spanner allows transaction tables to scale horizontally and secondary indexes for range queries

Reference:
https://cloud.google.com/spanner/docs/secondary-indexes

---

68) Your financial services company is moving to cloud technology and wants to store 50 TB of financial time-series 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

B. Google BigQuery

C. Google Cloud Storage

D. Google Cloud Datastore

Answer is **Cloud Bigtable**

Bigtable is GCP's managed wide-column database. It is also a good option for migrating on-premises Hadoop HBase databases to a managed database because Bigtable has an HBase interface.
Cloud Bigtable is a wide-column NoSQL database used for high-volume databases that require low millisecond (ms) latency. Cloud Bigtable is used for IoT, time-series, finance, and similar applications.

Reference:
https://cloud.google.com/blog/products/databases/getting-started-with-time-series-trend-predictions-using-gcp

---

69) An organization maintains a Google BigQuery dataset that contains tables with user-level data. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

A. Create and share an authorized view that provides the aggregate results.

B. Create and share a new dataset and view that provides the aggregate results.

C. Create and share a new dataset and table that contains the aggregate results.

D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

B will not minimize storage cost,
neither will C
I dont think is D: dataviewer role will allow this: – Read the dataset's metadata and to list tables in the dataset. – Read data and metadata from the dataset's tables. This is not what we want. They'll see our users data! A looks good: An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables. and also enforcing row-level access, which is what we want. Should be created with something like this: SELECT name, lastname FROM `dataset.view` WHERE allowed_viewer = SESSION_USER() And also, the cost will be charged to the user(s) performing the query https://cloud.google.com/bigquery/docs/share-access-views/

https://binx.io/blog/2018/12/06/bigquery-authorized-view/

---

70) 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 AppEngine service that collects user information and logs the access before providing a link to the bucket.

Keywords here are 1. "Archived": Immutable and hence, BQ and Cloud SQL are ruled out 2. "Auditable": Means track any changes done. Only D can provide the audibility piece! I will go with D

---

71) Your neural network model is taking days to train. You want to increase the training speed. What can you do?

A. Subsample your test dataset.

B. Subsample your training dataset.

C. Increase the number of input features to your model.

D. Increase the number of layers in your neural network.

increase the number of layers will make the training slower https://www.quora.com/Is-there-a-specific-reason-why-a-neural-network-with-more-layers-might-perform-worse-than-a-network-with-fewer-layers

https://towardsdatascience.com/how-to-train-neural-network-faster-with-optimizers-d297730b3713

---

72) You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?

A. PigLatin using Pig

B. HiveQL using Hive

C. Java using MapReduce

D. Python using MapReduce

Answer: A Description: Pig is scripting language which can be used for checkpointing and splitting pipelines

---

73) 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.

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

Answer: C Description : Speed of data transfer depends on Bandwidth Few things in computing highlight the hardware limitations of networks as transferring large amounts of data. Typically you can transfer 1 GB in eight seconds over a 1 Gbps network. If you scale that up to a huge dataset (for example, 100 TB), the transfer time is 12 days. Transferring huge datasets can test the limits of your infrastructure and potentially cause problems for your business.

---

74) After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison. What should you do?

A. Select random samples from the tables using the RAND() function and compare the samples.

B. Select random samples from the tables using the HASH() function and compare the samples.

C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.

D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Answer: C Description: Full comparison with this option, rest are comparison on sample which does not ensure all the data will be ok

---

75) 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.

D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

B is not, since the requisite is “avoid introducing new projects” (come on!)
As it says in the url directly: For such situations, you can use the flat-rate service. In this model, a certain number of slots are dedicated to your project(s), and you can establish a hierarchical priority model amongst the projects. The flat-rate model is especially suitable for large enterprises with multiple business units and workloads with varying priorities and budgets.

---

76) You have an Apache Kafka cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins. What should you do?

A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.

B. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.

C. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connector. Use a Dataflow job to read from PubSub and write to GCS.

D. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector. Use a Dataflow job to read from PubSub and write to GCS.

"A" is the answer which complies with the requirements (specifically, "The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins"). Indeed, one of the uses of what is called "Geo-Replication" (or Cross-Cluster Data Mirroring) in Kafka is precisely cloud migrations:
https://kafka.apache.org/documentation/#georeplication
However I agree with Ganshank, and the optimal "Google way" way would be "D", installing the Pub/Sub Kafka connector to move the data from on-prem to GCP.

---

77) You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing

operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload. What should you do?

A. Increase the size of your parquet files to ensure them to be 1 GB minimum.

B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.

C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.

D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

D: 1)Switch to SSD disks If you perform many shuffling operations or partitioned writes, switch to SSDs to boost performance. 2) Use preemptible VMs As a default, preemptible VMs are created with a smaller boot disk size, and you might want to override this configuration if you are running shuffle-heavy workloads https://cloud.google.com/architecture/hadoop/migrating-apache-spark-jobs-to-cloud-dataproc#optimize_performance

---

78) 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. that transforms the data, extract erroneous rows from logs. . that can be stored to PubSub later.

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, 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.

C says nothing about fixing the errors, just sending them back to Pub/Sub. Write all the errors to a new Pcollection and send that to Pub/Sub for handling seems better.
Therefore D. with side output Based on below blog
https://cloud.google.com/blog/products/gcp/handling-invalid-inputs-in-dataflow

OR :
The error records are directly written to PubSub from the DoFn (it's equivalent in python).
You cannot directly write a PCollection to PubSub. You have to extract each record and write one at a time. Why do the additional work and why not write it using PubSubIO in the DoFn itself?
You can write the whole PCollection to Bigquery though, as explained in

Reference:
https://medium.com/google-cloud/dead-letter-queues-simple-implementation-strategy-for-cloud-pub-sub-80adf4a4a800

---

79) You're training a model to predict housing prices based on an available dataset with real estate properties. Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency.

What should you do?

A. Provide latitude and longitude as input vectors to your neural net.

B. Create a numeric column from a feature cross of latitude and longitude.

C. Create a feature cross of latitude and longitude, bucketize at the minute level and use L1 regularization during optimization.

D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.

A and B don't make sense
Now; the crossed feature represents a well defined city block. If the model learns that certain city blocks (within range of latitudes and longitudes) are more likely to be more expensive than others, it is a stronger signal than two features considered individually. BUT we'll have way too many dimensions (and that is bad). Would L2 regularization accomplish this task? Unfortunately not. L2 regularization encourages weights to be small, but doesn't force them to exactly 0.0. However, there is a regularization term called L1 regularization that serves as an approximation to L0, but has the advantage of being convex and thus efficient to compute. So we can use L1 regularization to encourage many of the uninformative coefficients in our model to be exactly 0, and thus reap RAM savings at inference time. https://developers.google.com/machine-learning/crash-course/regularization-for-sparsity/l1-regularization

Use L1 regularization when you need to assign greater importance to more influential features. It shrinks less important feature to 0.
L2 regularization performs better when all input features influence the output & all with the weights are of equal size.

Reference:
https://developers.google.com/machine-learning/crash-course/regularization-for-sparsity/l1-regularization

---

80) 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.

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

It is definitely A. B: can't be because Health Checks just checks that machine is online C: StackDriver Logging is for Logging. Here we talk of Monitoring / Alerting D: StackDriver Agent monitors default metrics of VMs and some Database stuff with the MySQL Plugin. Here you want to monitor some more custom stuff like Replication of MariaDB (I didn't find anything of this sort in the plugin page), and you may want to use Custom Metrics rather than default metrics. "Cloud Monitoring automatically collects more than 1,500 built-in metrics from more than 100 monitored resources. But those metrics cannot capture application-specific data or client-side system data. Those metrics can give you information on backend latency or disk usage, but they can't tell you how many background routines your application spawned." https://cloud.google.com/monitoring/custom-metrics/open-census#monitoring_opencensus_metrics_quickstart-python

Answer is **Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.**

MariaDB needs costume metrics, and stackdriver built-in monitoring tools will not provide these metrics. OpenCensus Agent will do this for you

Reference:
https://cloud.google.com/monitoring/custom-metrics/open-census

---

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

What should you do?

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

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

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

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

A and B dont make sense - labelled dataset.
C doesnt either, What should we do with the applications?
D looks correct, in order to find new features for the model.

---

82) 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

Answer is **Cloud SQL**
Cloud SQL supports MySQL 5.6 or 5.7, and provides up to 624 GB of RAM and 30 TB of data storage, with the option to automatically increase the storage size as needed.

Reference:
https://cloud.google.com/sql/docs/features

---

83) You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes. You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload. What should you do?

A. Export Bigtable dump to GCS and run your analytical job on top of the exported files. profile for the analytics workload. profile for the analytics workload.

B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.

C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.

D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

You need to perform an hourly batch job on the cluster that already has high workload. in such cases, the best option is to replicate the cluster with single cluster routing. The original cluster can continue its read and writes. the replicated cluster can be used for analytical workload without impacting original cluster. Multi cluster routing is beneficial in cases where high availability is needed but requirement is only to isolate analytical workload from existing cluster.

Reference:
https://cloud.google.com/bigtable/docs/replication-overview#use-cases

---

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

A. Batch job, PubSubIO, side-inputs

B. Streaming job, PubSubIO, JdbcIO, side-outputs

C. Streaming job, PubSubIO, BigQueryIO, side-inputs

D. Streaming job, PubSubIO, BigQueryIO, side-outputs

Static reference data from BigQuery will go as side-inputs and data from pub-sub will go as streaming data using PubSubIO and finally BigQueryIO is required to push the final data to BigQuery
You need pubsubIO and BigQueryIO for streaming data and writing enriched data back to BigQuery.
side-inputs are a way to enrich the data

Reference:
https://cloud.google.com/architecture/e-commerce/patterns/slow-updating-side-inputs

85) You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of you Cloud Bigtable cluster. Which two actions can you take to accomplish this? (Choose two.)

A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.

B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.

C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.

D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.

E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

D: In general, do not use more than 70% of the hard limit on total storage, so you have room to add more data. If you do not plan to add significant amounts of data to your instance, you can use up to 100% of the hard limit
C: If this value is frequently at 100%, you might experience increased latency. Add nodes to the cluster to reduce the disk load percentage. The key visualizer metrics options, suggest other things other than increase the cluster size. https://cloud.google.com/bigtable/docs/monitoring-instance

86) You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps. You have the following requirements:
- You will batch-load the posts once per day and run them through the Cloud Natural Language API.
- You will extract topics and sentiment from the posts.
- You must store the raw posts for archiving and reprocessing.
- You will create dashboards to be shared with people both inside and outside your organization.
You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving.

What should you do?

A. Store the social media posts and the data extracted from the API in BigQuery.

B. Store the social media posts and the data extracted from the API in Cloud SQL.

C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.

D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Answer is **Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.**

Social media posts can images/videos which cannot be stored in bigquery

---

87) You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day.

Which tool should you use?

A. cron

B. Cloud Composer

C. Cloud Scheduler

D. Workflow Templates on Cloud Dataproc

Answer is **Cloud Composer**

Cloud Composer is an Apache Airflow managed service, it serves well when orchestrating interdependent pipelines, and Cloud Scheduler is just a managed Cron service.

Reference:
https://stackoverflow.com/questions/59841146/cloud-composer-vs-cloud-scheduler

---

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

What should you do?

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

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

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

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

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

Protection of sensitive data, like personally identifiable information (PII), is critical to your business. Deploy de-identification in migrations, data workloads, and real-time data collection and processing.

Reference:
https://cloud.google.com/dlp

---

89) 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

D. Cloud AutoML Natural Language

Answer is **Dialogflow Enterprise Edition**

since we need to recognize both voice and intent

Reference:
https://cloud.google.com/blog/products/gcp/introducing-dialogflow-enterprise-edition-a-new-way-to-build-voice-and-text-conversational-apps

---

90) You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once and must be ordered within windows of 1 hour. How should you design the solution?

A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.

B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.

C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.

D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.

Answer is **Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.**

Dataflow has autoscaling feature and pubsub is best solution

---

91) 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.

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.

Answer : B : The permissions are required at dataset levels hence READER, WRITER & OWNER which are the primitive roles for dataset to be used.

Reference:
https://cloud.google.com/bigquery/docs/access-control-primitive-roles#dataset-primitive-roles

---

92) You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time.

Which two methods can accomplish this?

(Choose two.)

A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.

B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.

C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.

D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.

E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.

Option A; Cheap storage and it is a supported method
https://cloud.google.com/datastore/docs/export-import-entities

Option B; Data exported from one Datastore mode database can be imported into another Datastore mode database, even one in another project.
https://cloud.google.com/datastore/docs/export-import-entities

---

93) You are designing a cloud-native historical data processing system to meet the following conditions:
- The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.
- A streaming data pipeline stores new data daily.
- Performance is not a factor in the solution.
- The solution design should maximize availability.

How should you design data storage for this solution?

A.Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.

B.Store the data in BigQuery. Access the data using the BigQuery Connector on Cloud Dataproc and Compute Engine.

C.Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.

D.Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

Answer : D : Multi-region increases high availability and pdf can be stored in Google Cloud Storage

---

94) Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per second. Many third parties use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:
- Single global endpoint
- ANSI SQL support
- Consistent access to the most up-to-date data

What should you do?

A.Implement BigQuery with no region selected for storage or processing.

B.Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.

C.Implement Cloud SQL for PostgreSQL with the master in Norht America and read replicas in Asia and Europe.

D.Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.

Answer : B : Cloud Spanner has three types of replicas: read-write replicas, read-only replicas, and witness replicas. Bigquery cannot support 250K data ingestion/second , as ANSI SQL support is required , no other options left except Spanner.

---

95) You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time. Consumers will receive the data in the following ways:
- Real-time event stream

- ANSI SQL access to real-time stream and historical data
- Batch historical exports

Which solution should you use?

A. Cloud Dataflow, Cloud SQL, Cloud Spanner

B. Cloud Pub/Sub, Cloud Storage, BigQuery

C. Cloud Dataproc, Cloud Dataflow, BigQuery

D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

"Processing logs at scale using Cloud Dataflow"
As far as this document is concerned.

- Cloud Storage for storing exported logs in batch mode.
- Pub/Sub for streaming exported logs in streaming mode.
- Dataflow for processing log data.
- BigQuery for storing processing output and supporting rich queries on that output.

It says the data is collected from the market, and the problem is that the methods are defined as requirements. Therefore, a close answer is B.

Reference:
https://cloud.google.com/solutions/processing-logs-at-scale-using-dataflow?hl=ja

---

96) You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must scale to accommodate increases in usage and must accommodate the addition of new applications without negatively affecting the performance of existing ones.

What should you do?

A. Create an API using App Engine to receive and send messages to the applications

B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them

C. Create a table on Cloud SQL, and insert and delete rows with the job information

D. Create a table on Cloud Spanner, and insert and delete rows with the job information

Pub/sub will be used to streaming data between application

---

97) You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?

A. Use Transfer Appliance to copy the data to Cloud Storage

B. Use gsutil cp ""J to compress the content being uploaded to Cloud Storage

C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage

D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic

Huge amount of data with log network bandwidth, Transfer Appliance is best for moving data over 100TB

---

98) You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:
- Executing the transformations on a schedule
- Enabling **non-developer analysts to modify transformations**
- Providing a graphical tool for designing transformations

What should you do?

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

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

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

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

Answer : A : Dataprep is used by non developers

---

99) You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants to add cameras to the delivery lines to detect and track any visual damage to the packages in transit. You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit.

Which solution should you choose?

A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.

B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.

C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Functions. Integrate the package tracking applications with this function.

D. Use TensorFlow to create a model that is trained on your corpus of images. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

Answer : B : AutoML is used to train model and do damage detection.
Auto Vision is used is a pre trained model used to detect objects in images.

---

100) You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second.

A. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.

B. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a fixed time window of 1 hour. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.

C. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtable. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hour. If that number falls below 4000, send an alert.

D. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuery. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour. If that number falls below 4000, send an alert.

Answer : A : Kafka IO and Dataflow is a valid option for interconnect (needless where Kafka is located - On Prem/Google Cloud/Other cloud)
Sliding Window will help to calculate average.

-