Cloud Bigtable

A single value in each row is indexed; this value is known as the row key

including a supported extension to the Apache HBase library for Java

Incredible scalability.

Simple administration.

Cluster resizing without downtime.

Good for

- **Time-series data,** such as CPU and memory usage over time for multiple servers.
- **Marketing data,** such as purchase histories and customer preferences.
- **Financial data,** such as transaction histories, stock prices, and currency exchange rates.
- **Internet of Things data,** such as usage reports from energy meters and home appliances.
- **Graph data,** such as information about how users are connected to one another.

You can manage security at the project level and instance level. Cloud Bigtable does not support table-level, row-level, column-level, or cell-level security restrictions

Cloud Bigtable is not a relational database; it does not support SQL queries or joins, nor does it support multi-row transactions. Also, it is not a good solution for storing less than 1 TB of data.

- If you need full SQL support for an online transaction processing (OLTP) system, consider Cloud Spanner or Cloud SQL.
- If you need interactive querying in an online analytical processing (OLAP) system, consider BigQuery.
- If you need to store immutable blobs larger than 10 MB, such as large images or movies, consider Cloud Storage.
- If you need to store highly structured objects in a document database, with support for ACID transactions and SQL-like queries, consider Cloud Datastore.

CLOUD SPANNER FEATURES

The first horizontally scalable, globally consistent, relational database service

Global Scale

Horizontally scalable across rows, regions, and continents, from 1 to hundreds or thousands of nodes.

Fully Managed

Ease of deployment at every scale and every stage. Synchronous replication and maintenance are automatic and built-in.

Relational Semantics

Everything you would expect from a relational database—schemas, ACID transactions, and SQL queries (ANSI 2011).

Multi-Language Support

Client libraries in C#, Go, Java, Node.js, PHP, Python, and Ruby. JDBC driver for connectivity with popular third-party tools.

Transactional Consistency

Purpose-built for external, strong, global transactional consistency.

Enterprise Grade Security

Data-layer encryption, IAM integration for access and controls, and audit logging.

Highly Available

Whenever, wherever, your data is highly available.

Cloud SQL features

Cloud SQL is a fully managed MySQL, PostgreSQL, and SQL Server database service.

Scalability

Easily scale up to 64 processor cores and more than 400 GB of RAM. Quickly scale out with read replicas.

High performance

Designed to scale from small development workloads up to performance-intensive workloads.

Integrated

Cloud SQL instances are accessible from just about any application, anywhere. Easily connect from App Engine, Compute Engine, and your workstation.

Availability protection

Live migration makes maintenance of our underlying infrastructure transparent. For isolation from failures, high availability provides continuous health-checking and automatically fails over if an instance is not healthy.

Fully managed

Replicated, managed, and backed-up, so you can make better use of your time.

Standard APIs

Build and deploy for the cloud faster because Cloud SQL offers standard MySQL, PostgreSQL, and SQL Server databases. Use standard connection drivers and built-in migration tools to get started quickly.

Security

Cloud SQL data is encrypted when on Google's internal networks and when stored in database tables, temporary files, and backups. Cloud SQL supports private connectivity with Virtual Private Cloud (VPC), and every Cloud SQL instance includes a network firewall, allowing you to control public network access to your database instance. Learn more about Google Cloud Platform's comprehensive security architecture.

Partnerships and integrations

Take advantage of our growing partner ecosystem and tools to make working with Cloud SQL even easier. Our partners can help you streamline the process of loading your data, create rich visualizations for meaningful insights, and monitor and manage your databases.

BigQuery Features

Serverless

With serverless data warehousing, Google does all resource provisioning behind the scenes, so you can focus on data and analysis rather than worrying about upgrading, securing, or managing the infrastructure.

Real-time analytics

BigQuery's high-speed streaming insertion API provides a powerful foundation for real-time analytics, making your latest business data immediately available for analysis.

Automatic high availability

BigQuery transparently and automatically provides highly durable, replicated storage in multiple locations and high availability with no extra charge and no additional setup.

Standard SQL

BigQuery supports a standard SQL dialect which is ANSI:2011 compliant, which thereby reduces the need for code rewrites. BigQuery also provides ODBC and JDBC drivers at no cost to ensure your current applications can interact with its powerful engine.

Federated query and logical data warehousing

Through powerful federated query, BigQuery can process external data sources in object storage (Cloud Storage), transactional databases (Cloud Bigtable), or spreadsheets in Drive — all without duplicating data.

Storage and compute separation

With BigQuery's separated storage and compute, you have the option to choose the storage and processing solutions that make sense for your business and control access and costs for each.

Automatic backup and easy restore

BigQuery automatically replicates data and keeps a seven-day history of changes, allowing you to easily restore and compare data from different times.

Geospatial data types and functions

BigQuery GIS brings SQL support for arbitrary points, lines, polygons, and multi-polygons in WKT and GeoJSON format. You can simplify your geospatial analyses, see your location-based data in new ways, or unlock entirely new lines of business.

Data transfer service

The BigQuery Data Transfer Service automatically transfers data from external data sources, like Google Marketing Platform, Google Ads, YouTube, and partner SaaS applications to BigQuery on a scheduled and fully managed basis. Users can also easily transfer data from Teradata and Amazon S3 to BigQuery.

Big data ecosystem integration

With Cloud Dataproc and Cloud Dataflow, BigQuery provides integration with the Apache Big Data ecosystem, allowing existing Hadoop/Spark and Beam workloads to read or write data directly from BigQuery.

Petabyte scale

Get great performance on your data, while knowing you can scale seamlessly to store and analyze petabytes more without having to buy more capacity.

Flexible pricing models

On-demand pricing lets you pay only for the storage and compute that you use. Flat-rate pricing enables high-volume users or enterprises to choose a stable monthly cost. For more information, see BigQuery pricing or cost controls.

Data governance and security

BigQuery makes it easy to maintain strong security with fine-grained identity and access management with Cloud Identity and Access Management, and your data is always encrypted at rest and in transit.

Geoexpansion

BigQuery gives you the option of geographic data control (in US, Asia, and European locations), without the headaches of setting up and managing clusters and other computing resources in region.

Foundation for AI

Besides bringing ML to your data with BigQuery ML, integrations with Cloud ML Engine and TensorFlow enable you to train powerful models on structured data in minutes with just SQL.

Foundation for BI

BigQuery forms the data warehousing backbone for modern BI solutions and enables seamless data integration, transformation, analysis, visualization, and reporting with tools from Google and our technology partners.

Flexible data ingestion

Load your data from Cloud Storage or stream it into BigQuery at thousands of rows per second to enable real-time analysis of your data. Use familiar data integration tools like Informatica, Talend, and others out of the box.

Programmatic interaction

BigQuery provides a REST API for easy programmatic access and application integration. Client libraries are available in Java, Python, Node.js, C#, Go, Ruby, and PHP. Business users can use Google Apps Script to access BigQuery from Sheets.

Rich monitoring and logging with Stackdriver

BigQuery provides rich monitoring, logging, and alerting through Stackdriver Audit Logs and it can serve as a repository for logs from any application or service using Stackdriver Logging.

Cloud Storage

Cloud data storage

Highly Scalable NoSQL Database. ACID transactions, SQL-like queries, indexes and much more.

QUESTION 84
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.

QUESTION 85
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.

QUESTION 86
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 fron 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 fron PubSub and write to GCS.

QUESTION 87
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.

QUESTION 88
Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data).

What should you do?

A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
B. Add a try... catch block to your *DoFn* that transforms the data, extract erroneous rows from logs.
C. Add a try... catch block to your *DoFn* that transforms the data, 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.

QUESTION 89
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 longtitude 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 longtitude as input vectors to your neural net.
B. Create a numeric column from a feature cross of latitude and longtitude.
C. Create a feature cross of latitude and longtitude, bucketize at the minute level and use L1 regularization during optimization.
D. Create a feature cross of latitude and longtitude, bucketize it at the minute level and use L2 regularization during optimization.

QUESTION 90
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.

QUESTION 91
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.

QUESTION 92
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

QUESTION 93
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.
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.

QUESTION 94
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

QUESTION 95
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 2 answers.

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.

QUESTION 96
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.

QUESTION 97
You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.
What should you do?

A. Use Cloud Dataflow with Beam to detect errors and perform transformations.
B. Use Cloud Dataprep with recipes to detect errors and perform transformations.
C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

QUESTION 98
Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing on-premises applications every day. What should they do?

A. Execute `gsutil rsync` from the on-premises servers.
B. Use Cloud Dataflow and write the data to Cloud Storage.
C. Write a job template in Cloud Dataproc to perform the data transfer.
D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

QUESTION 99
You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using `bq query - -dry_run` you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

A. Create a separate table for each ID.
B. Use the LIMIT keyword to reduce the number of rows returned.
C. Recreate the table with a partitioning column and clustering column.
D. Use the `bq query - -maximum_bytes_billed` flag to restrict the number of bytes billed.

QUESTION 100
You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

A. Use `bq load` to load a batch of sensor data every 60 seconds.
B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
C. Use the INSERT statement to insert a batch of data every 60 seconds.
D. Use the MERGE statement to apply updates in batch every 60 seconds.

QUESTION 101
You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
C. Export the records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.
D. Export the records from the database as an Avro file. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

QUESTION 102
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
C. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

QUESTION 103
You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?

A. Set the BigQuery dataset to be regional. In the event of an emergency, use a point-in-time snapshot to recover the data.
B. Set the BigQuery dataset to be regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.
C. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.
D. Set the BigQuery dataset to be multi-regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.

QUESTION 104
You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

A. Create a cron schedule in Cloud Dataprep.
B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.
C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.
D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.

QUESTION 105
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

QUESTION 107
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.

QUESTION 108
You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.
B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.
C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.
D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

QUESTION 109
You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.

A. Publisher throughput quota is too small.
B. Total outstanding messages exceed the 10-MB maximum.
C. Error handling in the subscriber code is not handling run-time errors properly.
D. The subscriber code cannot keep up with the messages.
E. The subscriber code does not acknowledge the messages that it pulls.

QUESTION 110
You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

A. Add a SideInput that returns a Boolean if the element is corrupt.
B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

QUESTION 111
You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the Data Science team runs a query filtered on a date column and limited to 30–90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries. What should you do?

A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type.
B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.
C. Modify your pipeline to maintain the last 30–90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.
D. Write an Apache Beam pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

QUESTION 112
You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

A. Deploy small Kafka clusters in your data centers to buffer events.
B. Have the data acquisition devices publish data to Cloud Pub/Sub.
C. Establish a Cloud Interconnect between all remote data centers and Google.
D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

QUESTION 113
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

QUESTION 114
Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

A. Cloud Dataflow
B. Cloud Composer
C. Cloud Dataprep
D. Cloud Dataproc

QUESTION 116
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

A. Implement clustering in BigQuery on the ingest date column.
B. Implement clustering in BigQuery on the package-tracking ID column.
C. Tier older data onto Cloud Storage files, and leverage extended tables.
D. Re-create the table using data partitioning on the package delivery date.

QUESTION 120
You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

A. An alert based on a decrease of `subscription/num_undelivered_messages` for the source and a rate of change increase of `instance/storage/used_bytes` for the destination
B. An alert based on an increase of `subscription/num_undelivered_messages` for the source and a rate of change decrease of `instance/storage/used_bytes` for the destination

---

C. An alert based on a decrease of `instance/storage/used_bytes` for the source and a rate of change increase of `subscription/num_undelivered_messages` for the destination
D. An alert based on an increase of `instance/storage/used_bytes` for the source and a rate of change decrease of `subscription/num_undelivered_messages` for the destination

QUESTION 121
You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

A. Edge TPUs as sensor devices for storing and transmitting the messages.
B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

QUESTION 122
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 2 answers.

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

QUESTION 123
You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? Choose 2 answers.

A. Denormalize the data as must as possible.
B. Preserve the structure of the data as much as possible.
C. Use BigQuery UPDATE to further reduce the size of the dataset.
D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

QUESTION 124
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 too including Cloud Dataproc, BigQuery, and Compute Engine.
- A streaming data pipeline stores new data daily.
- Peformance 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 or Cloud Dataproc and Compute Engine.
C. Store the data in a regional Cloud Storage bucket. Aceess 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.

QUESTION 125
You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?

A. Store and process the entire dataset in BigQuery.
B. Store and process the entire dataset in Cloud Bigtable.
C. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.
D. Store the warm data as files in Cloud Storage, and store the active data in BigQuery. Keep this ratio as 80% warm and 20% active.

QUESTION 126
You work for a manufacturing company that sources up to 750 different components, each from a different supplier. You've collected a labeled dataset that has on average 1000 examples for each unique component. Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (as Proof-Of-Concept) within a few working days. What should you do?

A. Use Cloud Vision AutoML with the existing dataset.
B. Use Cloud Vision AutoML, but reduce your dataset twice.
C. Use Cloud Vision API by providing custom labels as recognition hints.
D. Train your own image recognition model leveraging transfer learning techniques.

QUESTION 127
You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?

A. Use Cloud TPUs without any additional adjustment to your code.
B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.
C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.
D. Stay on CPUs, and increase the size of the cluster you're training your model on.

QUESTION 128
You work on a regression problem in a natural language processing domain, and you have 100M labeled exmaples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test

---

set. How should you improve the performance of your model?

A. Increase the share of the test sample in the train-test split.
B. Try to collect more data and increase the size of your dataset.
C. Try out regularization techniques (e.g., dropout of batch normalization) to avoid overfitting.
D. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.

QUESTION 129
You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

A. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

QUESTION 130
The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a `quotaExceeded` error. What should you do?

A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.
B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.
C. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.
D. Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.

QUESTION 133

A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: `SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features)`. How should you create the ML pipeline?

A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.

B. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.

C. Create a Cloud Dataflow pipeline using `BigQueryIO` to read results from the query. Grant the Dataflow Worker role to the application service account.

D. Create a Cloud Dataflow pipeline using `BigQueryIO` to read predictions for all users from the query. Write the results to Cloud Bigtable using `BigtableIO`. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.