1.Your organization is streaming telemetry data into BigQuery for long-term storage (2 years) and analysis, at the rate of about 100 million records per day. They need to be able to run queries against certain time periods of data without incurring the costs of querying all available records. What is the preferred method for doing so?

A. Create a single table, but query only individual rows by data in the WHERE clause.
B. Use a LIMIT clause to limit the number of rows queried based on WHERE clause criteria.
C. Partition a single table by day, and run queries against individual partitions.
D. Create a new table, one for each day. Run queries against the groups of tables relevant to their needs.
Correct Answer: C

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

A. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Nearline storage bucket as a final destination.
B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Coldline storage bucket as a final destination.
C. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Nearline storage bucket as a final destination.
D. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Coldline bucket as a final destination.
Correct Answer: A

3.You need to deploy a TensorFlow machine-learning model to Google Cloud. You want to maximize the speed and minimize the cost of model prediction and deployment. What should you do?
A. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud Storage. Run 1 version on CPUs and another version on GPUs.
B. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud ML Engine. Run 1 version on CPUs and another version on GPUs.
C. Export your trained model to a SavedModel format. Deploy and run your model from a Kubernetes Engine cluster
D. Export your trained model to a SavedModel format. Deploy and run your model on Cloud ML Engine.
Correct Answer: D

4.What will happen to your data in a Bigtable instance if a node goes down?
A. Bigtable will attempt to rebuild the data from RAID disk configuration when the node comes back online.
B. Nothing, as the storage is separated from the node compute.
C. Lost data will automatically rebuild itself from Cloud Storage backups when the node comes back online.
D. Data will be lost, which makes regular backups to Cloud Storage necessary.
Correct Answer: B

5.Your organization is making the move to Google Cloud. You need to bring your existing big data processing workflows to the cloud without having to re-train employees on new products. Your organization uses the Apache Hadoop ecosystem for big data processing. Which Google Cloud managed service would your workflow move to?
A. Cloud Dataproc
B. Cloud Bigtable
C. Cloud Pub/Sub
D. Cloud Dataflow

Correct Answer: A

6.Your online shopping company needs to know when a user has not interacted with the site in 30 minutes. They need the website to alert the user once they have been idle for too long. You use Cloud Dataflow to process the interaction events and decide if an alert should be sent. How should you design the pipeline?
A. Implement a session window with a gap time duration of 30 minutes.
B. Implement a fixed-time window with a duration of 30 minutes.
C. Implement a global window with a time-based trigger with a delay of 30 minutes.
D. Implement a sliding time window with a duration of 30 minutes.
Correct Answer: A

7. When training a machine learning model, why do you need separate training and test data
A. Without different data, your model will not generalize for additional data, known as overfitting.
B. Both sets of data are necessary for deep and wide neural networks.
C. Your learning model will have an improper learning rate, making training difficult.
D. Without separate sets of data, your neural network will not have enough data to train with.
Correct Answer: A

8.You are developing an application on Google Cloud that will label famous landmarks in users' photos. You are under competitive pressure to develop the predictive model quickly. You need to keep service costs low. What should you do?
A. Build and train a classification model with TensorFlow. Deploy the model using the Cloud Machine Learning Engine. Inspect the generated MID values to supply the image labels.
B. Build an application that calls the Cloud Vision API. Pass client image locations as base64-encoded strings.
C. Build an application that calls the Cloud Vision API. Inspect the generated MID values to supply the image labels.
D. Build and train a classification model with TensorFlow. Deploy the model using the Cloud Machine Learning Engine. Pass client image locations as base64- encoded strings.
Correct Answer: B

9.You have 250,000 devices which produce a JSON device status event every 10 seconds. You want to capture this event data for outlier time series analysis. What should you do?
A. Ship the data into BigQuery. Develop a custom application that uses the BigQuery API to query the dataset and display a device's outlier data based on your business requirements.
B. Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements.
C. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for the device outlier data based on your business requirements.
D. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.
Correct Answer: B

10.Your company's aging Hadoop servers are nearing end of life. Instead of replacing your hardware, your CIO has decided to migrate the cluster to Google Cloud Dataproc. A direct lift and shift migration of the cluster would require 30 TB of disk space per individual node. There are cost concerns about using that much storage. How can you best minimize the cost of the migration?
A. Decouple storage from computer by placing the data in Cloud Storage
B. Place archived data in Cloud Storage, and only use 'hot' data in HDFS on the cluster disks.
C. Implement maximum data compression to reduce the amount of disk space your data uses.
D. Use preemptible VM's to save costs on cluster storage usage.
Correct Answer: A

11. What is the recommended minimum amount of data to store in Bigtable?

A. 500 GB
B. 1 GB
C. 1 TB
D. 500 TB
Correct Answer: C

12.You have hundreds of IoT devices that generate 1 TB of streaming data per day. Due to latency, messages will often be delayed compared to when they were generated. You must be able to account for data arriving late within your processing pipeline. What should you do?
A. Use Cloud SQL to process the delayed messages.
B. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Dataflow to process messages, and use windows, watermarks (timestamp), and triggers to process late data.
C. Use SQL queries in BigQuery to analyze data by timestamp.
D. Enable your IoT devices to generate a timestamp when sending messages. Use Cloud Pub/Sub to process messages by timestamp and fix out of order issues.
Correct Answer: B

13.You have a long-running, streaming Dataflow pipeline that you need to shut down. You do not need to preserve data currently in the processing pipeline and need it shut down as soon as possible. Which shutdown option should you use to complete the shutdown process?
A. Graceful shutdown
B. Cancel
C. Stop
D. Drain
Correct Answer: B

14.You need to run analytical queries using SQL syntax against data formatted in JSON format. What should you do? Choose the best answer.
A. Load your JSON data into Cloud SQL, and run queries against it in that service.
B. Load your JSON data into Cloud Storage. Add your JSON table as an external read source in BigQuery, since BigQuery is unable to store data in JSON format.
C. Import the data into Bigtable and use Bigtable for your queries.
D. Import the data in JSON format into BigQuery as a table, and run queries against it.
Correct Answer: D

15.Your organization has migrated their Hadoop workloads to Cloud Dataproc. To fully take advantage of the cloud, you want to decouple your Hadoop storage and compute, and be able to destroy your cluster when compute is complete in order to save costs while preserving your data. What should you do?
A. You must use another processing framework such as Apache Beam for this task.
B. Copy your data from HDFS to Cloud Storage. Update your scripts to point to the Cloud Storage location (gs://) instead of the HDFS location (hdfs://). Within your Dataproc job, configure output to output to Cloud Storage.
C. Use the Dataproc sync tool to synchronize HDFS with GCS.
D. You must leave your managed Dataproc cluster running in order to access computer data.
Correct Answer: B

16.You have data stored in a Cloud Storage bucket and also in a BigQuery dataset. You need to secure the data and provide 3 different types of access levels for your Google Cloud Platform users: administrator, read/write, and read-only. You want to follow Google-recommended practices. What should you do?
A. At the Organization level, add your administrator user accounts to the Owner role, add your read/write user accounts to the Editor role, and add your read-only user accounts to the Viewer role.

B. At the Project level, add your administrator user accounts to the Owner role, add your read/write user accounts to the Editor role, and add your read-only user accounts to the Viewer role.
C. Create 3 custom IAM roles with appropriate policies for the access levels needed for Cloud Storage and BigQuery. Add your users to the appropriate roles.
D. Use the appropriate pre-defined IAM roles for each of the access levels needed for Cloud Storage and BigQuery. Add your users to those roles for each of the services.
Correct Answer: D

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

18.You are monitoring a streaming data pipeline that ingests streaming data into Cloud Pub/Sub, processed by Cloud Dataflow, and inserted into a BigQuery table. Your Pub/Sub topic has a substantially higher than acceptable number of undelivered messages. Choose two reasons why this may be happening.
A. Your Publishers' message throughput is too low.
B. The subscriber is not acknowledging messages as they are pulled.
C. Your Dataflow subscriber is unable to keep up with the rate of incoming messages.
D. Your Audit Logs do not have sufficient access to your Pub/Sub topic, causing delays in delivering.
Correct Answer: B and C

19.You work at a very large organizations that has a very large analyst team. You use the default pricing model for BigQuery. During heavy usage, your analyst group occasionally runs out of the 2000 slots available for the BigQuery jobs. You do not want to create additional projects for the sole purpose of increasing slot count. What can you do to resolve this?
A. You must create an additional project to increase your slot count, then spread the BigQuery loads across both projects.
B. Force-enable the 'use cached results' option for all available queries.
C. Switch to flat rate pricing to enable a higher total slot quota for your project.
D. Use the quotas page to increase your BigQuery slot count to 3000 as needed..
Correct Answer: C

20.You want to display aggregate view counts for your YouTube channel data in Data Studio. You want to see the video tiles and view counts summarized over the last 30 days. You also want to segment the data by the Country Code using the fewest possible steps. What should you do?
A. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title and Country Code as report dimensions.
B. Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter.
C. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter.
D. Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title and Country Code as report dimensions.
Correct Answer: D

21. You are designing storage for event data as part of building a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying individual values over time windows. Which storage service and schema design should you use?
A. Use Cloud Bigtable for storage. Design tall and narrow tables, and use a new row for each single event version.
B. Use Cloud Bigtable for storage. Design short and wide tables, and use a new column for each single event version.
C. Use Cloud Storage for storage. Join the raw file data with a BigQuery log table.
D. Use Cloud Storage for storage. Write a Cloud Dataprep job to split the data into partitioned tables.
Correct Answer: A

22. How can you set up your Dataproc environment to use BigQuery as an input and output source?
A. Use the Bigtable syncing service built into Dataproc.
B. Manually use a Cloud Storage bucket to import and export to and from both BigQuery and Dataproc.
C. You can only use Cloud Storage or HDFS for your Dataproc input and output.
D. Install the BigQuery connector on your Dataproc cluster.
Correct Answer: D

23. Your team has decided to use Datalab for interactive machine learning exercises. You want your team members to share their work and progress with each other. How do you accomplish this?
A. Every team member will use their own Datalab notebook and synchronize changes to the shared Cloud Source Repository.
B. Use the team sync feature included in Datalab notebooks to synchronize each member's work.
C. Give your team members Compute Instance Admin and Service Account Actor roles to access a shared notebook.
D. Create a shared Datalab notebook, and assign the Datalab Editor role to your team members to access it.
Correct Answer: A

24.You are planning the architecture for a data pipeline. This pipeline will automatically take files from a private Cloud Storage bucket, transform them in Cloud Dataflow, and write the results to BigQuery. How should you execute this pipeline in line with Google's security best practices?
A. Grant Project Viewer role to a service account, and have the service account execute the nightly batch job.
B. Run a batch job with a service account with the Project Owner role.
C. Set a reminder to manually execute the job using your user account with Project Owner access.
D. Execute a batch job to run the pipeline. Grant a service account read access to the batch files in Cloud Storage, and editor roles in Dataflow and BigQuery to transform and load data into BigQuery
Correct Answer: D

25.You created a job which runs daily to import highly sensitive data from an on-premises location to Cloud Storage. You also set up a streaming data insert into Cloud Storage via a Kafka node that is running on a Compute Engine instance. You need to encrypt the data at rest and supply your own encryption key. Your key should not be stored in the Google Cloud. What should you do?
A. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in your Kafka node hosted on Compute Engine.
B. Create a dedicated service account, and use encryption at rest to reference your data stored in Cloud Storage and Compute Engine data as part of your API service calls.

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

26.In a Dataflow processing pipeline, which concept describes timestamps attached to incoming messages?
A. Watermark
B. ParDo
C. PCollection
D. Trigger
Correct Answer: A

27.You are building an application that needs to convert recorded customer service calls into text format, and will then examine call transcripts to determine customer sentiment. What is the most time effective method of doing this?
A. Use the Cloud Speech to Text and Cloud Translate pre-trained API's to perform both steps in the process.
B. Create a machine learning model and train it with both training and test data of your recorded audio logs on Cloud ML Engine. Do the same with analyzing customer sentiment from the transcribed calls.
C. Use the Cloud Speech to Text and Cloud Natural Language pre-trained API's to perform both steps in the process.
D. Hire an outside consulting firm to perform the process.
Correct Answer: C

28.What types of jobs does Cloud Dataproc support? (Choose all that apply)
A. Hive
B. Beam
C. Pig
D. Spark
Correct Answer: A and C and D

29.Your BiqQuery table needs to be accessed by team members who are not proficient in technology. You want to simplify the columns they need to query to avoid confusion. How can you do this while preserving all of the data in your table?
A. Create a query that uses the reduced number of columns they will access. Save this query as a view in a different dataset. Give your team members access to the new dataset and instruct them to query against the saved view instead of the main table.
B. Train your team members on how to query larger tables.
C. Apply column filtering to your table, and restrict the unfiltered view to yourself and those who need access to the full table.
D. Create a copy of your table in a different dataset, and remove the unneeded columns from the copy. Have your team members run queries against this copy.
Correct Answer: A

30. You are setting up multiple MySQL databases on Compute Engine. You need to collect logs from your MySQL applications for audit purposes. How should you approach this?
A. Configure Cloud Composer to monitor and report on instance performance metrics.
B. Install the Stackdriver Logging agent on your database instances and configure the fluentd plugin to read and export your MySQL logs into Stackdriver Logging.
C. Install the Stackdriver Monitoring agent on your instances, configure the MySQL plugin, and export logs to Stackdriver Monitoring.
D. Configure Stackdriver Logging to natively monitor application logs, which will appear in Stackdriver Logging.

Correct Answer: B.

31. Your organization needs to develop their machine learning model to control topology definitions. There are a large number of possible configurations to achieve the best results. What components of their machine learning model would they adjust to account for increased complexity? (Choose two answers.)
A. Learning rate
B. Neurons
C. Epoch
D. Hidden layers
Correct Answer: B and D

32. You are building a machine learning model to predict the number of lightning strikes during a storm. Your model has thousands of input features to train on. You want to improve the training speed of the model by removing features, but do not want to negatively effect your model's accuracy. What action should you take?
A, Combine highly co-dependent and redundant features into one representative feature.
B. Implement L2 regularization to automatically 'prune' unneeded features
C. Remove the features that have null values for the majority of your records.
D. Remove features that have high correlation to your output labels.
Correct Answer: A

33. You need to export Avro formatted data from BigQuery into Cloud Storage. What is the best method of doing so from the web console?
A. Convert the data to CSV format the BigQuery export options, then make the transfer.
B. Use the BigQuery Transfer Service to transfer Avro data to Cloud Storage.
C. Click on Export Table in BigQuery, and provide the Cloud Storage location to export to.
D. Create a Dataflow job to manage the conversion of Avro data to CSV format, then export to Cloud Storage.
Correct Answer: C

34. You want to train your machine learning model on AI Platform while saving costs. Which scaling tier would you choose?
A. BASIC
B. CUSTOM
C. STANDARD_1
D. PREMIUM_1
Correct Answer: A

35. As part of your backup plan, you create regular boot-disk snapshots of Compute Engine instances that are running. You want to be able to restore these snapshots using the fewest possible steps for replacement instances. What should you do?
A. Export the snapshots to Cloud Storage. Create images from the exported snapshot files.
B. Use the snapshots to create replacement disks. Use the disks to create instances as needed.
C. Use the snapshots to create replacement instances as needed.
D. Export the snapshots to Cloud Storage. Create disks from the exported snapshot files. Create images from the new disks.
Correct Answer: C

36. You are working on a project with two compliance requirements. The first requirement states that your developers should be able to see the Google Cloud Platform billing charges for only their projects. The second requirement states that your finance team members can set budgets and view the current charges for all projects in the organization. The finance team should not be able to view the project contents. You want to set permissions. What should you do?
A. Add the finance team to the Viewer role for the Project. Add the developers to the Security Reviewer role for each of the billing accounts.
B. Add the developers and finance managers to the Viewer role for the Project.

C. Add the finance team members to the default IAM Owner role. Add the developers to a custom role that allows them to see their spending only.
D. Add the finance team members to the Billing Administrator role for each of the billing accounts that they need to manage. Add the developers to the Viewer role for the Project.
Correct Answer: D

37.Your organization is ready to migrate their Hadoop workloads to Google Cloud. For the data migration, they need a cost-effective 'data lake' that will scale to their growing data needs and be able to easily connect to their Hadoop workloads in the cloud. What two actions should they perform?
A. Install the Bigtable connector in the on-premises Hadoop cluster, then migrate data to Bigtable for long-term storage.
B. Add the Cloud Storage connector to their on-premises Hadoop environment, and transfer their data to a Cloud Storage bucket.
C. For the existing Hadoop jobs that are migrating to Dataproc, use the gs:// prefix instead of hdfs:// to access data from Cloud Storage.
D. Create a Dataproc cluster for long-term use, and transfer data to the HDFS partition on the cluster.
Correct Answer: B and C

38. What types of Bigtable row keys can lead to hotspotting? (Choose all that apply)
A. Leading with a non-reversed timestamp.
B. Standard domain names (non-reversed).
C. Reverse timestamps.
D. Non-sequential numeric IDs.
Correct Answer: A and B

39. You are an administrator for several organizations in the same company. Each organization has data in their own BigQuery table within a single project. For application access reasons, all of the tables must remain in the same project. You think each organization should be able to view and run queries against their own data without exposing the data of organizations to unauthorized viewers. What should you recommend?
A. You must separate the tables by project, and use a service account in your application to access data in each project. Give out project-wide roles to each organization.
B. Place the tables in a single dataset, and apply IAM roles to each table, limiting access per table to each organization.
C. Create a separate dataset for each organization in the same project. Place each organization's table in each dataset. Restrict access to the organization's dataset to only that company, from which they can view their table but no one else's.
D. Place all data in a single table, create authorized views restricting access by row based on the SESSION_USER() field. Add that same SESSION_USER() field with the same email addresses according to which company needs access to which roles.
Correct Answer: C

40. Your infrastructure includes two 100-TB enterprise file servers. You need to perform a one-way, one-time migration of this data to the Google Cloud securely. Only users in Germany will access this data. You want to create the most cost-effective solution. What should you do?
A. Use 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 storage bucket as a final destination.
C. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.
D. Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Multi-Regional bucket as a final destination.
Correct Answer: C

41. You have a Dataflow job that keeps failing due to errors in your input data. What steps can you take to improve pipeline reliability while at the same time, capturing failed data for reprocessing?
A. Implement a try-catch block that transforms the both good and bad data. Create an additional output to use a new PCollection that can be output to Pub/Sub for later analysis.
B. Filter out errors as they occur, and view error entries using Stackdriver Logging
C. Implement a try-catch block that transforms the both good and bad data, and extract the incorrect entries from Stackdriver Logging
D. Implement a try-catch block that transforms the both good and bad data. Publish the erroneous data to Pub/Sub, which can then be placed into GCS for further analysis.
Correct Answer: A

42. Which of these is NOT a valid reason to choose an HDD storage type over SSD in a Bigtable instance?
A. You need to maintain costs.
B. You plan on running batch workloads instead of frequently executing random reads across a small number of rows.
C. You need to integrate Bigtable with Cloud Storage
D. You need to store over 10TB of data.
Correct Answer: C

43.Your company's Kafka server cluster has been unable to scale to the demands of their data ingest needs. Streaming data ingest comes from locations all around the world. How can they migrate this functionality to Google Cloud to be able to scale for future growth?
A. Create a separate Pub/Sub topic for each region. Configure endpoints to publish to the Pub/Sub topic closest to their location, and configure a new Cloud Dataflow pipeline in each region to subscribe to the equivalent Pub/Sub topic to process messages as they come in.
B. Create a single Pub/Sub topic. Configure endpoints to publish to the Pub/Sub topic, and configure Cloud Dataflow to subscribe to the same topic to process messages as they come in.
C. Create a Computer Engine managed instance group that is configured to autoscale to 150% of peak demand. Use a managed instance template with Kafka installed to automatically scale as needed, and direct traffic to this autoscaling cluster.
D. Create a Kubernetes Engine cluster in each region needed. Install Kafka on the cluster. Use an HTTP load balancer to serve each Kubernetes cluster region. Configure a new Cloud Dataflow pipeline in each region to process requests forwarded from the Kubernetes cluster.
Correct Answer: B

44.Which of these is NOT a type of trigger that applies to Dataflow?
A. Element size in bytes.
B. Element count.
C. Combinations of other triggers.
D. Timestamp
Correct Answer: A

45.You are building storage for files for a data pipeline on Google Cloud. You want to support JSON files. The schema of these files will occasionally change. Your analyst teams will use running aggregate ANSI SQL queries on this data. What should you do?
A. Use Cloud Storage for storage. Link data as permanent tables in BigQuery and turn on the Automatically detect option in the Schema section of BigQuery.
B. Use BigQuery for storage. Provide format files for data load. Update the format files as needed.
C. Use BigQuery for storage. Select Automatically detect in the Schema section.
D. Use Cloud Storage for storage. Link data as temporary tables in BigQuery and turn on the Automatically detect option in the Schema section of BigQuery.
Correct Answer: C

46.You are training a machine learning model to predict the likelihood of rain based on an available dataset of weather data. In reviewing your input data, the amount of humidity in the air has a very strong influence on the chance of rain, especially compared to less relevant data. How can you incorporate this more important data to that it properly influences the model?
A. Create a feature from the humidity data point, and use L2 regularization to optimize the model.
B. Tune your hyperparameters to give greater weighting to the humidity feature over others.
C. Create a feature from the humidity data point, and use L1 regularization to optimize the model.
D. Reduce your epochs except for humidity features.
Correct Answer: C

47. You are building a data pipeline on Google Cloud. You need to prepare source data for a machine-learning model. This involves quickly deduplicating rows from three input tables and also removing outliers from data columns where you do not know the data distribution. What should you do?
A. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Write a recipe to transform the data and add it to the Cloud Dataprep job.
B. Write an Apache Spark job with a series of steps for Cloud Dataflow. The first step will examine the source data, and the second and third steps will perform data transformations.
C. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Click on each column name, click on each appropriate suggested transformation, and then click Add to add each transformation to the Cloud Dataprep job.
D. Write an Apache Spark job with a series of steps for Cloud Dataproc. The first step will examine the source data, and the second and third steps will perform data transformations.
Correct Answer: C

48. Which of these open source frameworks is best suited to process simultaneous batch and streaming in a single data pipeline?
A. Apache Hadoop
B. Apache Kafka
C. Kubernetes
D. Apache Beam
Correct Answer: D

49.You have in your possession a database of financial transactions, which include a user's name, location, purchase location, and purchase amount. Each transaction is also labelled whether or not the transaction was fraudulent. With this data, what two types of machine learning can potentially applied to this dataset?
A. Apply reinforcement learning to predict the location of purchase.
B. Unsupervised learning to identify patterns (clustering) in the data to predict the location of future purchases.
C. Apply unsupervised learning to label which transactions are likely to be fraudulent.
D. Using the applied fraudulent/non-fraudulent labels, apply supervised classification learning to predict which future transactions are likely to be fraudulent
E. Apply supervised linear regression learning to label which transactions are likely to be fraudulent
Correct Answer: B and D

50. You are creating a machine learning model to predict the likelihood of fraud from credit card transaction data. The end result will be predicting the percent confidence of two results: "Fraud" and "Not Fraud". What type of learning model problem is this?
A. Clustering
B. Classification
C. Regression
D. Hyperparameter
Correct Answer: B

51.Your organization needs to be able to reliably handle ever-increasing amounts of streaming telemetry data, process it, and economically store analysed data. What services should they use for this task?
A. Stackdriver, Cloud Dataproc, Cloud Spanner
B. Cloud Pub/Sub, Cloud Dataproc, Bigtable
C. Cloud Pub/Sub, Cloud Dataflow, Bigquery
D. Kubernetes Engine, Cloud Dataflow, Cloud Datastore
Correct Answer: C

52. You are using a Compute Engine instance to manage your Cloud Dataflow processing workloads. What IAM role do you need to grant to the instance so that it has the necessary access?
A. Dataflow Viewer
B. Dataflow Developer
C. Dataflow Worker
D. Dataflow Computer
Correct Answer: C

53. Regarding Cloud Pub/Sub, which resource locations can have access controlled via IAM roles? (Choose all that apply)
A. Topics
B. Publisher
C. Project-wide predefined roles
D. Subscription
Correct Answer: A and C and D

54. You are selecting a streaming service for log messages that must include final result message ordering as part of building a data pipeline on Google Cloud. You want to stream input for 5 days and be able to query the most recent message value. You will be storing the data in a searchable repository. How should you set up the input messages?
A. Use Apache Kafka on Compute Engine for input. Attach a timestamp to every message in the publisher.
B. Use Cloud Pub/Sub for input. Attach a unique identifier to every message in the publisher.
C. Use Apache Kafka on Compute Engine for input. Attach a unique identifier to every message in the publisher.
D. Use Cloud Pub/Sub for input. Attach a timestamp to every message in the publisher.
Correct Answer: D

55. What open source software is Cloud Pub/Sub most similar to?
A. Apache Beam
B. Apache Kafka
C. HBase
D. Apache Hadoop
Correct Answer: B

56. You are working in the updated BigQuery interface. While conducting BigQuery queries against a large table with many columns, in the job Execution Details section, you notice you have an excessively high read time in the first stage of your query execution. How can you troubleshoot this to increase performance and reduce costs? (Choose all that apply)
A. Reduce the number of read operations by adding a LIMIT clause to your query.
B. Restrict the number of columns in your SELECT field for those needed. This will reduce read times on your query.
C. Reduce the number of write operations by optimizing the complexity of your query functions.
D. Partition or separate your large table into smaller pieces. Conduct a query against your smaller (or partitioned) tables to reduce read times.
Correct Answer: B and D

57. In order to protect live customer data, your organization needs to maintain separate operating environments —development/test, staging, and production— to meet the needs of running experiments, deploying new features, and serving production customers. What is the best practice for isolating these environments while at the same time maintaining operability?
A. Create separate organization accounts for each environment, and use domain wide IAM roles to allow access between each organization environment to share data as needed.
B. Create a separate project for dev/test, staging, and production. Migrate relevant data between projects when ready for the next stage.
C. Place all three environments in the same project, however, use separate Cloud Storage buckets, Cloud ML Engine clusters, and other services for each environment
D. Place resources into the same project. but use object versioning in Cloud Storage in order to separate data by environment.
Correct Answer: B.

58.Which of these statements do not apply to preemptible worker nodes on Cloud Dataproc? Choose two answers.
A. You must have a max of 2:1 ratio of preemptible to standard workers.
B. Preemptible workers only function as processing nodes.
C. Your cluster can be created with only preemptible workers
D. Preemptible workers can be added after the cluster is created.
Correct Answer: A and C

59.In AI Platform, what does the CUSTOM tier allow you to configure? Choose the best answer.
A. Custom number of master, worker, and parameter servers.
B. Custom number of workers and parameter servers. Machine type of master server
C. The number of workers.
D. Parameter servers.
Correct Answer: B

60. When using AI Platform to train machine learning models, how are online predictions different from batch predictions? (Choose all that apply)
A. Online prediction results are written to Cloud Storage as output.
B. Online predictions are returned in the response message.
C. Batch predictions are used to reduce latency in serving predictions.
D. Batch predictions are optimized to handle a high volume of prediction examples while running on more complex models.
Correct Answer: B and D

61. You are migrating a Hadoop cluster to Cloud Dataproc using GCS for storage. After migration, some of your existing, more complex Spark jobs (in parquet format) are performing noticeably worse than your on-premises cluster. You are using mostly preemptible VM's (with a few required non-preemptible) in order to save on costs.
A. Change your file format to CSV format
B. Increase the size of your cluster by twice as many preemptible VM's
C. Switch disks from HDD to SSD. Change the default preemptible VM settings to increase the size of the boot disk.
D. Switch your disks from HDD to SSD, run the job in HDFS before copying the results back to GCS
E. Ensure that your parquet files are at an optimized block size
Correct Answer: C

62.You are designing a relational data repository on Google Cloud to grow as needed. The data will be transactionally consistent and added from any location in the world. You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?
A. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.

B. Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.
C. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.
D. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.
Correct Answer: D

63.You need to replicate the logs that are ingested by your on-premises Apache Kafka cluster to Google Cloud to be stored for analysis in BigQuery. What should you do?
A. Create an identical Kafka cluster on Compute Engine in GCP. Configure your on-premises Kafka cluster to duplicate all data to the GCP Kafka cluster. Use a Dataflow job to process data from Kafka and insert into BigQuery.
B. Configure the Pub/Sub Kafka connector on your on-premises Kafka cluster, and configure Pub/Sub as a source connector. Use a Cloud Dataflow job to read from a subscribed Pub/Sub topic and write to BigQuery
C. Create a Cloud Composer workflow to manage the replication of data from your Kafka cluster directly into BigQuery.
D. Configure the Pub/Sub Kafka connector on your on-premises Kafka cluster, and configure Pub/Sub as a sink connector. Use a Cloud Dataflow job to read from a subscribed Pub/Sub topic and write to BigQuery
Correct Answer: D

64.You need to choose a structure storage option for storing very large amounts of data with the following properties and requirements:
The data has a single key
You need very low latency Which solution should you choose?
A. Bigtable
B. Datastore
C. Cloud SQL
D. BiqQuery
Correct Answer: A

65. When training a machine learning model on AI Platform on a distributed scaled tier, what types of machines are part of that distributed resource? (Choose all that apply)
A. Host
B. Worker
C. Master
D. Parameter server
Correct Answer: B and C and D

66. You are developing an application that will only recognize and tag specific business to business product logos in images. You do not have an extensive background working with machine learning models, but need to get your application working. What is the current best method to accomplish this task?
A. Create a custom machine learning model to recognize specific logos in photos, then train it on AI Platform.
B. Use the Cloud Vision API to recognize logos in the images.
C. Use the AutoML Vision service to train a custom model using the Vision API
D. Use the Cloud Vision API to recognize all logos in images, then use the Cloud Natural Language API to recognize specific logos by name.
Correct Answer: C

67.Your company needs to run analytics on their incoming inventory data. They need to use their existing Hadoop workloads to perform this task. What two steps must be performed to accomplish this? (Choose two answers)

A. Stream inventory data to Cloud Pub/Sub, process data with Cloud Dataflow into Bigtable and Cloud Storage.
B. Stream from Cloud Pub/Sub into Cloud Dataproc, which can then place relevant data in the appropriate storage location
C. Use Spark to accept the streaming ingest on the Dataproc cluster, and then process jobs on HDFS.
D. Connect Cloud Dataproc to Bigtable and Cloud Storage, running analytics on the data in both services.
Correct Answer: B and D

68. Your organization has just recently started using Google Cloud. Everyone in the company has access to all datasets in BigQuery, using it as they see fit without documenting their use cases. You need to implement a formal security policy, but need to first determine what everyone has been doing in BigQuery. What is your first step to do so?
A. Inspect the IAM policy of each table.
B. Export billing into Cloud Storage, and view BigQuery related records to determine user activity.
C. View the usage of BigQuery query slots in Stackdriver Monitoring.
D. View audit logs in Stackdriver Logging to review data access.
Correct Answer: D

70. You are training a facial detection machine learning model. Your model is suffering from overfitting your training data. Choose three steps you can take to solve this problem.
A. Use a larger set of features
B. Use a smaller set of features
C. Reduce the number of training examples
D. Increase the number of training examples
E. Increase the regularization parameters
F. Decrease the regularization parameters
Correct Answer: B and D and E

71.Your production Bigtable instance is currently using four nodes. Due to the increased size of your table, you need to add additional nodes to offer better performance. How should you accomplish this without the risk of data loss?
A. Power off your Bigtable instance, then increase the node count, then power back on. Be sure to schedule downtime in advance.
B. Export your Bigtable data as sequence files into Cloud Storage, then import the data into a new Bigtable instance with additional nodes added.
C Use the node migration service to add additional nodes.
D. Edit instance details and increase the number of nodes. Save your changes. Data will re-distribute with no downtime.
Correct Answer: D

72. You currently have a Bigtable instance you've been using for development running a development instance type, using HDD's for storage. You are ready to upgrade your development instance to a production instance for increased performance. You also want to upgrade your storage to SSD's as you need maximum performance for your instance. What should you do?
A. Export your Bigtable data into a new instance, and configure the new instance type as production with SSD's
B. Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD.
C. Run parallel instances where one instance is using HDD and the other is using SSD.
D. Use the Bigtable instance sync tool in order to automatically synchronize two different instances, with one having the new storage configuration.
Correct Answer: A

73. **What is the difference between a deep and wide neural network? What would you use a deep AND wide neural network for? (Choose all that apply)**
A. Wide models are used for generalizations. Deep models are for memorization.
B. Deep and wide models are ideal for solving regression problems.
C. Wide models are used for memorization. Deep models are for generalization
D. Deep and wide models are ideal for a recommendation application.
Correct Answer: C and D

74.You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud. Your current workflows for this task require using services that support Apache Spark and the Hadoop ecosystem. You are using ANSI SQL to run queries for your analysts. You want to support complex aggregate queries and reuse existing code. How should you store and transform the input data?
A. Use BigQuery for storage. Use Cloud Dataflow to run the transformations.
B. Use Cloud Storage for storage. Use Cloud Dataproc to run the transformations.
C. Use BigQuery for storage. Use Cloud Dataproc to run the transformations.
D. Use Cloud Storage for storage. Use Cloud Dataflow to run the transformations.
Correct Answer: C

75. You want to export your Cloud SQL tables into BigQuery for analysis. How can you do this?
A. Convert your Cloud SQL data to JSON format, then import directly into BigQuery
B. Export your Cloud SQL data to Cloud Storage, then import into BigQuery
C. Import data from BigQuery directly from Cloud SQL.
D. Use the BigQuery export function in Cloud SQL to manage exporting data into BigQuery.
Correct Answer: B

76.You are upgrading your existing (development) Cloud Bigtable instance for use in your production environment. The development Bigtable instance is using SSD storage. The instance contains a large amount of data that you want to make available for production immediately. You need to design for the fastest performance. What should you do?
A. Change your Cloud Bigtable instance type from Development to Production, and set the number of nodes to at least 3. Verify that the storage type is HDD.
B. Export the data from your current Cloud Bigtable instance to Cloud Storage. Create a new Cloud Bigtable Production instance type with at least 3 nodes. Select the SSD storage type. Import the data into the new instance from Cloud Storage.
C. Change your Cloud Bigtable instance type from Development to Production, and set the number of nodes to at least 3. Verify that the storage type is SSD.
D. Export the data from your current Cloud Bigtable instance to Cloud Storage. Create a new Cloud Bigtable Production instance type with at least 3 nodes. Select the HDD storage type. Import the data into the new instance from Cloud Storage.
Correct Answer: C

77. You are evaluating a storage solution for your data. Your data is in a structured, non-relational format, and will be used for analysis. You need the lowest latency read and write speeds possible. Your data is about 3 TB in size, predicted to grow to up to 5 TB. What solution should you use?
A. Use BigQuery to host your non-relational, structured data.
B. Use Cloud Bigtable using HDD storage.
C. Use Cloud Bigtable with SSD storage.
D. Use Cloud Datastore for your operations.
Correct Answer: C

78. What is the purpose of hyperparameters in a machine learning training model?
A. Form the basis of labels on your training data.
B. Hyperparameters adjust the training process itself.
C. Train for a regression machine learning problem.

D. They help your model learn from the training data.
Correct Answer: B

79. Which of these statements is true regarding BigQuery caching?
A. The BigQuery cache only lasts for 48 hours.
B. Multiple users can use the same cached query.
C. Cache is not enabled by default.
D. Queries that retrieve results from the cache have no charge.
Correct Answer: D

80. In machine learning, what is the difference between test and training data?
A. Training data is used for hyperparameter tuning, and test data is used for feature engineering.
B. Test data is used to tune parameters, like weights and biases.
C. Test data is labeled with the 'correct' answer; training data is not.
D. Training data has a label attached to train on features for the correct answer. Test data is used to test the trained model for accuracy when completed on new data
Correct Answer: D

81**. You need to process transactions in a point-of-sale application on Google Cloud Platform. You need to account for exponential user growth, but you do not want to deal with managing your infrastructure overhead. Which database service option should you use?**
A. Cloud Datastore
B. Cloud SQL
C. BigQuery
D. Cloud Memorystore
Correct Answer: A

82.You need to run analytical queries using SQL syntax against data formatted in JSON format. What should you do? Choose the best answer.
A. Load your JSON data into Cloud SQL, and run queries against it in that service.
B. Load your JSON data into Cloud Storage. Add your JSON table as an external read source in BigQuery, since BigQuery is unable to store data in JSON format.
C. Import the data into Bigtable and use Bigtable for your queries.
D. Import the data in JSON format into BigQuery as a table, and run queries against it.
Correct Answer: D

83. **Which of these options are adjusted by a machine learning neural network as it works with its training dataset? (Choose all that apply)**
(Possible Correct: 2)
A. Biases
B. Weights
C. Epochs
D. Features
Correct Answer: A and B