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

- **A. Disable caching by editing the report settings.**
- B. Disable caching in BigQuery by editing table details.
- C. Refresh your browser tab showing the visualizations.
- D. Clear your browser history for the past hour then reload the tab showing the visualizations.

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

- A. HTTPS
- B. VPN
- **C. SOCKS**
- D. HTTP

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

- **A. Dataproc Worker**
- B. Dataproc Viewer
- C. Dataproc Runner
- D. Dataproc Editor

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

- A. Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD.
- **B. Export your Bigtable data into a new instance, and configure the new instance type as production with SSD's**
- 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.

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

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

- B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column DT for each row. Reference the column TS instead of the column DT from now on.
- C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.
- **D. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.**

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

A. Change the dataset from a regional location to multi-region location, specifying the regions to be included.
B. Export the data from BigQuery into a bucket in the new location, and import it into a new dataset at the new location.
C. Copy the data from the dataset in the source region to the dataset in the target region using BigQuery commands.
D. Export the data from BigQuery into nearby bucket in Cloud Storage. Copy to a new regional bucket in Cloud Storage in the new location and Import into the new dataset.

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

- A. Use the LIMIT clause to limit the number of values in the results.
- **B. Use the SELECT clause to limit the amount of data in the query. Partition data by date so the query can be more focused.**
- C. Set the Maximum Bytes Billed, which will limit the number of bytes processed but still run the query if the number of bytes requested goes over the limit.
- D. Use GROUP BY so the results will be grouped into fewer output values.

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

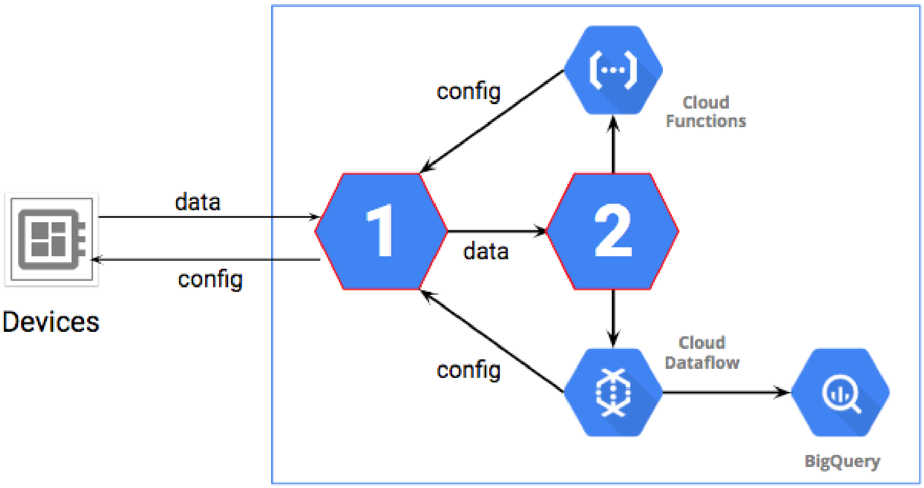
- **A. Use Cloud Pub/Sub and Cloud Dataflow with Sliding Time Windows.**
- B. Use Cloud Pub/Sub and Google Data Studio.
- C. Cloud Pub/Sub can guarantee timely arrival and order.
- D. Use Cloud Dataflow's built-in timestamps for ordering and filtering.

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

- A. The model is working extremely well, indicating the hyperparameters are set correctly.
- **B. The model is overfit. There is a problem.**
- C. The model is underfit. There is a problem.
- D. The model is perfectly fit. You do not need to continue training.

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

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



- A. Cloud IoT Core, Cloud Datastore
- B. Cloud Pub/Sub, Cloud Storage
- **C. Cloud IoT Core, Cloud Pub/Sub**
- D. App Engine, Cloud IoT Core

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

- A. Use BigQuery for storage. Provide format files for data load. Update the format files as needed.
- **B. Use BigQuery for storage. Select "Automatically detect" in the Schema section.**
- C. Use Cloud Storage for storage. Link data as temporary tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.
- D. Use Cloud Storage for storage. Link data as permanent tables in BigQuery and turn on the "Automatically detect" option in the Schema section of BigQuery.

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

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

- B. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.
- **C. Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements.**
- D. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data based on your business requirements.

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

- A. Use Cloud Machine Learning to host your model. Monitor the status of the Operation object for 'error' results.
- **B. Use Cloud Machine Learning to host your model. Monitor the status of the Jobs object for 'failed' job states.**
- C. Use a Kubernetes Engine cluster to host your model. Monitor the status of the Jobs object for 'failed' job states.
- D. Use a Kubernetes Engine cluster to host your model. Monitor the status of Operation object for 'error' results.

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

- A. Build an application that calls the Cloud Vision API. Inspect the generated MID values to supply the image labels.
- **B. Build an application that calls the Cloud Vision API. Pass landmark locations as base64-encoded strings.**
- C. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Pass landmark locations as base64-encoded strings.
- D. Build and train a classification model with TensorFlow. Deploy the model using Cloud Machine Learning Engine. Inspect the generated MID values to supply the image labels.

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

- A. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and direct the users to view the report only once per business day (24-hour period).
- **B. Set up the report to use the Owner's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is selected for the report.**
- C. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and also set it up to be a 'view-only' report.
- D. Set up the report to use the Viewer's credentials to access the underlying data in BigQuery, and verify that the 'Enable cache' checkbox is not selected for the report.

Your customer is moving their corporate applications to Google Cloud Platform. The security team wants detailed visibility of all projects in the organization. You provision the Google Cloud Resource

Manager and set up yourself as the org admin. What Google Cloud Identity and Access Management (Cloud IAM) roles should you give to the security team?

- A. Org viewer, project owner
- **B. Org viewer, project viewer**
- C. Org admin, project browser
- D. Project owner, network admin

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

- A. Google BigQuery
- B. Google Cloud SQL
- **C. Google Cloud Bigtable**
- D. Google Cloud Storage

You need to take streaming data from thousands of Internet of Things (IoT) devices, ingest it, run it through a processing pipeline, and store it for analysis. You want to run SQL queries against your data for analysis. What services in which order should you use for this task?

- A. Cloud Dataflow, Cloud Pub/Sub, BigQuery
- B. Cloud Pub/Sub, Cloud Dataflow, Cloud Dataproc
- **C. Cloud Pub/Sub, Cloud Dataflow, BigQuery**
- D. App Engine, Cloud Dataflow, BigQuery

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

- **A. Cloud Bigtable**
- B. Cloud Spanner
- C. Cloud SQL
- D. Cloud Datastore

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

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

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

- A. Convert the data to CSV format the BigQuery export options, then make the transfer.
- B. Use the BigQuery Transfer Service to transfer Avro data to Cloud Storage.
- **C. Click on Export Table in BigQuery, and provide the Cloud Storage location to export to**
- D. Create a Dataflow job to manage the conversion of Avro data to CSV format, then export to Cloud Storage.

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

- A. Use the Bigtable syncing service built into Dataproc.
- B. Manually use a Cloud Storage bucket to import and export to and from both BigQuery and Dataproc
- **C. Install the BigQuery connector on your Dataproc cluster**
- D. You can only use Cloud Storage or HDFS for your Dataproc input and output.

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

- A. Include ORDER BY DESK on timestamp column and LIMIT to 1.
- B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
- C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
- **D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.**

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

- A. Load data into different partitions.
- **B. Load data into a different dataset for each client.**
- C. Put each client's BigQuery dataset into a different table.
- **D. Restrict a client's dataset to approved users.**
- E. Only allow a service account to access the datasets.
- **F. Use the appropriate identity and access management (IAM) roles for each client's users.**

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

- A. Run a local version of Jupiter on the laptop.
- B. Grant the user access to Google Cloud Shell.
- C. Host a visualization tool on a VM on Google Compute Engine.
- **D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.**

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

- A. Grant the consultant the Viewer role on the project.
- **B. Grant the consultant the Cloud Dataflow Developer role on the project.**
- C. Create a service account and allow the consultant to log on with it.
- D. Create an anonymized sample of the data for the consultant to work with in a different project.

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

- A. Check the dashboard application to see if it is not displaying correctly.
- B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.
- C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
- **D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.**

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

- A. Disable writes to certain tables.
- B. Restrict access to tables by role.
- C. Ensure that the data is encrypted at all times.
- **D. Restrict BigQuery API access to approved users.**
- **E. Segregate data across multiple tables or datasets.**
- **F. Use Google Stackdriver Audit Logging to determine policy violations.**

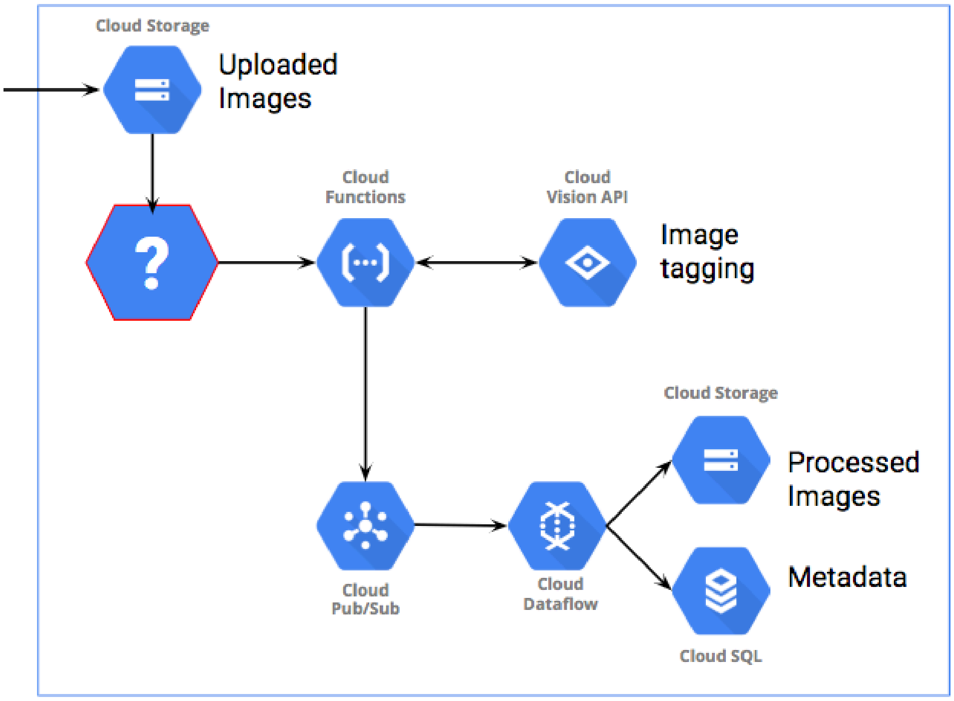
You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What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

- **A. Update the current pipeline and use the drain flag.**
- B. Update the current pipeline and provide the transform mapping JSON object.
- C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
- D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

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

- **A. They should use the Cloud Dataflow Cloud Runner.**
- B. They should upload the pipeline to Cloud Dataproc.
- C. They should use the local version of runner.
- D. Import the pipeline into BigQuery.

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



- A. Cloud Datastore
- B. Cloud Dataflow
- **C. Cloud Pub/Sub**
- D. Cloud Bigtable

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

- A. Use Primitive role on the dataset
- B. Use Predefined role on the dataset
- C. Use Authorized view with the same dataset with proper permissions
- **D. Use Authorized view with the different dataset with proper permissions**

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

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

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

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

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

- A. Utilize Cloud Storage to collect the inbound sensor data, analyze data with Dataproc and save the results to BigQuery.
- B. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to BigQuery.
- C. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Cloud SQL.
- **D. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Bigtable.**

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

- **A. Wide model is used for memorization, while the deep model is used for generalization.**
- B. Wide model is used for generalization, while the deep model is used for memorization.
- **C. A good use for the wide and deep model is a recommender system.**
- D. A good use for the wide and deep model is a small-scale linear regression problem.

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

- A. Bigtable
- B. Datastore
- C. Cloud Storage
- **D. Cloud Spanner**

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

- A. Classification
- B. Regression
- C. Association
- **D. Clustering**

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

- **A. Nearline**
- B. Standard
- **C. Multi-Regional**
- D. Dual-Regional
- E. Regional

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

- A. Create and Train a model using Tensorflow and Develop an REST based wrapper over it
- B. Use Cloud Image Intelligence API and Develop an REST based wrapper over it
- C. Use Cloud Natural Language API and Develop an REST based wrapper over it
- **D. Use Cloud Vision API and Develop an REST based wrapper over it**

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

- A. Use Cloud Vision API to identify video with inappropriate content and mark it for manual checks.
- B. Use Cloud Natural Language API to identify video with inappropriate content and mark it for manual checks.
- C. Use Cloud Speech-to-Text API to identify video with inappropriate content and mark it for manual checks.
- **D. Use Cloud Video Intelligence API to identify video with inappropriate content and mark it for manual checks.**

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

- A. Cloud Scheduler
- **B. Cloud Composer**
- C. Cloud Spanner
- D. Cloud Pipeline

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

- A. Grant only DATA VIEWER access to the third party team
- B. Grant fine grained DATA VIEWER access to the tables and columns within the dataset
- C. Create Authorized views for tables in a same project and grant access to the teams
- **D. Create Authorized views for tables in a separate project and grant access to the teams**

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

- A. Export the data access logs using aggregated sink to Cloud Storage in an existing project and grant VIEWER access to the project to the Auditor team
- B. Export the data access logs using project sink to BigQuery in an existing project and grant VIEWER access to the project to the Auditor team
- C. Export the data access logs using project sink to Cloud Storage in a separate project and grant VIEWER access to the project to the Auditor team
- **D. Export the data access logs using aggregated sink to Cloud Storage in a separate project and grant VIEWER access to the project to the Auditor team**

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

- A. Dataflow
- B. Dataproc
- C. Hadoop installation on Compute Engine
- **D. Dataprep**

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

- A. Dataflow
- **B. HBase on Dataproc**
- C. Bigtable
- D. BigQuery

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

- A. User
- **B. Jobuser**
- C. Owner
- D. Viewer

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

- A. Dataflow, BigQueryIO and PubSubIO, SideOutputs
- **B. Dataflow, BigQueryIO and PubSubIO, SideInputs**

- C. DataProc, BigQueryIO and PubSubIO, SideInputs
- D. DataProc, BigQueryIO and PubSubIO, SideOutputs

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

- A. Train Model using Tensorflow to identify PII and filter the information
- B. Store the data in BigQuery and create a Authorized view for the users
- **C. Use Data Loss Prevention APIs to identify the PII information and filter the information**
- D. Use Cloud Natural Language API to identify PII and filter the information

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

- A. Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.
- **B. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.**
- C. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.
- D. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

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

- A. Add the finance team members to the default IAM Owner role. Add the developers to a custom role that allows them to see their own spend only.
- **B. Add the finance team members to the Billing Administrator role for each of the billing accounts that they need to manage. Add the developers to the Viewer role for the Project.**
- C. Add the developers and finance managers to the Viewer role for the Project.
- D. Add the finance team to the Viewer role for the Project. Add the developers to the Security Reviewer role for each of the billing accounts.

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

- A. Output custom metrics to Stackdriver from the game servers, and create a Dashboard in Stackdriver Monitoring Console to view them.
- B. Schedule BigQuery load jobs to ingest analytics files uploaded to Cloud Storage every ten minutes, and visualize the results in Google Data Studio.

- **C. Store time-series data from the game servers in Google Bigtable, and view it using Google Data Studio.**
- D. Insert the KPIs into Cloud Datastore entities, and run ad hoc analysis and visualizations of them in Cloud Datalab.

Your infrastructure includes two 100-TB enterprise file servers. You need to perform a one-way, one-time migration of this data to the Google Cloud securely. Only users in Germany will access this data. You want to create the most cost-effective solution. What should you do?

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

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.

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

- A. Write an Apache Spark job with a series of steps for Cloud Dataflow. The first step will examine the source data, and the second and third steps step will perform data transformations.
- B. Write an Apache Spark job with a series of steps for Cloud Dataproc. The first step will examine the source data, and the second and third steps step will perform data transformations.
- C. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Write a recipe to transform the data and add it to the Cloud Dataprep job.
- **D. Use Cloud Dataprep to preview the data distributions in sample source data table columns. Click on each column name, click on each appropriate suggested transformation, and then click 'Add' to add each transformation to the Cloud Dataprep job.**

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

- A. Set up a cluster in High Availability mode with high-memory machine types. Add 10 additional local SSDs.
- B. Set up a cluster in High Availability mode with default machine types. Add 10 additional Preemptible worker nodes.
- **C. Set up a cluster in Standard mode with high-memory machine types. Add 10 additional Preemptible worker nodes.**
- D. Set up a cluster in Standard mode with the default machine types. Add 10 additional local SSDs.

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

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

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

- **A. Load logs into Google BigQuery.**
- B. Load logs into Google Cloud SQL.
- C. Import logs into Google Stackdriver.
- D. Insert logs into Google Cloud Bigtable.
- **E. Upload log files into Google Cloud Storage.**

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

- A. Transfer appliance and Dataprep to decrypt the data
- B. Google Transfer service using multiple VPN connections
- C. gustil with multiple VPN connections
- **D. Transfer appliance and rehydrator to decrypt the data**

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

- A. Use Dataprep with Data Studio
- B. Load in Cloud Storage and use Dataflow with Data Studio
- **C. Load in Cloud Storage and use Dataprep with Data Studio**
- D. Use Dataflow with Data Studio

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

- **A. Load data into Google BigQuery.**
- B. Insert data into Google Cloud SQL.
- C. Put flat files into Google Cloud Storage.
- D. Stream data into Google Cloud Datastore.

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

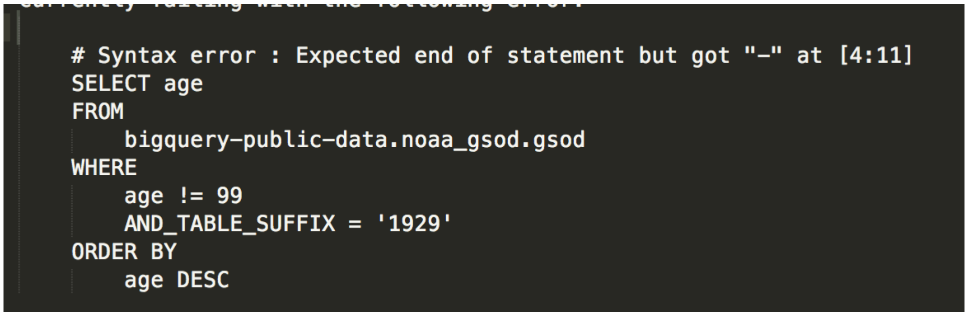
- A. Convert your Cloud SQL data to JSON format, then import directly into BigQuery
- **B. Export your Cloud SQL data to Cloud Storage, then import into BigQuery**
- C. Import data to BigQuery directly from Cloud SQL.
- D. Use the BigQuery export function in Cloud SQL to manage exporting data into BigQuery.

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

- A. Train your team members on how to query larger tables.
- **B. Create a query that uses the reduced number of columns they will access. Save this query as a view in a different dataset. Give your team members access to the new dataset and instruct them to query against the saved view instead of the main table.**
- 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.

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

Larger image



```
currently failing with the following error:

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

Which table name will make the SQL statement work correctly?

- A. `bigquery-public-data.noaa_gsod.gsod`
- B. bigquery-public-data.noaa_gsod.gsod*
- C. `bigquery-public-data.noaa_gsod.gsod`*
- **D. `bigquery-public-data.noaa_gsod.gsod*`**

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

- A. Cloud SQL
- B. BigQuery
- C. Cloud Bigtable
- **D. Cloud Datastore**

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

- A. Send the data to Google Cloud Datastore and then export to BigQuery.
- **B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.**
- C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
- D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

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

- **A. Redefine the schema by evenly distributing reads and writes across the row space of the table.**
- B. The performance issue should be resolved over time as the size of the Bigtable cluster is increased.

- C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
- D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

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

- A. Create a Google Cloud Dataflow job to process the data.
- B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
- C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
- **D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.**
- E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

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

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

How would you correct the error?

- **A. Add ", UNNEST(person)" before the WHERE clause.**
- B. Change "person" to "person.city".
- C. Change "person" to "city.person".
- D. Add ", UNNEST(city)" before the WHERE clause.

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

- A. Preemptible workers cannot use persistent disk.
- **B. Preemptible workers cannot store data.**
- C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.
- **D. A Dataproc cluster cannot have only preemptible workers.**

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

- A. Cancel
- **B. Drain**
- C. Stop
- D. Pause

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

- A. Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD.
- B. Run parallel instances where one instance is using HDD and the other is using SSD.
- C. Use the Bigtable instance sync tool in order to automatically synchronize two different instances, with one having the new storage configuration.
- **D. Build a Dataflow pipeline or Dataproc job to copy the data to the new cluster with SSD storage type.**

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

- A. Rewrite the job in Pig.
- **B. Rewrite the job in Apache Spark.**
- C. Increase the size of the Hadoop cluster.
- D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

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

- A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
- B. Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.
- **C. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.**
- D. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.

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

- **A. Import the data into BigQuery for better performance.**
- B. Request more slots for greater capacity to improve performance.

- C. Divide the data into partitions based on date.
- D. Time to move to Cloud Bigtable; it is faster in all cases.

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

- A. Migrate to Cloud Spanner
- **B. Read replicas**
- C. Instance high availability configuration
- D. Migrate to Cloud Storage

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

- A. Implement a high-throughput Cloud VPN connection
- B. Cloud Router with VPN
- C. Dedicated Interconnect
- **D. Partner Interconnect**

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

- A. Upgrade to the latest version of HDFS. Change the settings in Hadoop components to optimize for the different kinds of work in the mix.
- B. Find more jobs to run so the cluster utilizations will cost-justify the expense.
- **C. Store persistent data off-cluster. Start a cluster for one kind of work then shut it down when it is not processing data.**
- D. Migrate from Cloud Dataproc to an open source Hadoop Cluster hosted on Compute Engine, because this is the only way to get all the Hadoop customizations needed for efficiency.

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

- **A. Storage, queries, and streaming inserts.**
- B. Storage, queries, and loading data from a file.
- C. Storage, queries, and exporting data.
- D. Queries and streaming inserts.

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

- A. Export the query results to a public Cloud Storage bucket.
- **B. Create a BigQuery Authorized View and assign a project-level user role to analyst three.**

- C. Assign the bigquery.resultsonly.viewer role to analyst three.
- D. Create a BigQuery Authorized View and assign an organizational level role to analyst three.

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. BigQuery

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

- A. Continuous features
- B. Input values
- **C. Weights**
- **D. Biases**

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

- A. Bigtable
- B. Cloud Datastore
- C. Cloud Storage
- **D. BigQuery**

Your company is planning to migrate their historical dataset into BigQuery. This data would be exposed to the data scientists for perform analysis using BigQuery ML. The data scientists would like to know which ML models does the BigQuery ML support. What would be your answer? (Choose 2)

- A. Random Forest
- **B. Linear Regression**
- C. K Means
- D. Principal Component Analysis
- **E. Multiclass logistic regression for Classification**

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

- A. Create and Train a model using Tensorflow and Develop an REST based wrapper over it
- B. Create and Train a model using BigQuery ML and Develop an REST based wrapper over it
- **C. Use Cloud Natural Language API and Develop an REST based wrapper over it**

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

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

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

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

- A. Cloud Dataproc
- B. Cloud Scheduler
- C. Schedule jobs on a single Compute Engine using Cron.
- **D. Cloud Composer**

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

Each team has multiple Team Leaders, who should have the ability to create, delete tables, but not delete dataset.

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

How would you design the access control?

- A. Grant Team leader group - OWNER and Data Analyst - WRITER
- B. Grant Team leader group - OWNER and Data Analyst - READER
- **C. Grant Team leader group - WRITER and Data Analyst - READER**
- D. Grant Team leader group - READER and Data Analyst - WRITER

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

- A. Create and Train a sentiment analysis model using Tensorflow
- B. Use Cloud Speech-to-Text API for sentiment analysis
- **C. Use Cloud Natural Language API for sentiment analysis**
- D. Use Cloud Vision API for sentiment analysis

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

1. No prior knowledge of any language with no coding

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

What solution best fits the need?

- A. Use Dataflow code and provide Data Analysts the access to the code. Store the schema externally to be easily modified.
- **B. Use Dataprep with transformation recipes.**
- C. Use Dataproc spark and provide Data Analysts the access to the code. Store the schema externally to be easily modified.
- D. Use DataLab with transformation recipes.

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

- A. App Engine, Dataproc, DataStudio
- B. Compute Engine, BigQuery Streaming Inserts, DataStudio
- **C. App Engine, BigQuery Streaming Inserts, DataStudio**
- D. App Engine, Dataflow, DataStudio

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

- A. Monitor the BigQuery Slots Used metric
- B. Monitor the BigQuery Slots Pending metric
- C. Monitor the BigQuery Slots Allocated metric
- **D. Monitor the BigQuery Slots Available metric**

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

- A. Use Dataflow and fixed windowing of 30 minutes
- **B. Use Dataflow and Session windowing with gap duration of 30 minutes**
- C. Use Dataflow and Global window with gap duration of 30 minutes
- D. Use Dataproc and store the data in BigQuery and aggregate the same

You have a real time data processing pipeline running in Dataflow. As a part of changed requirement you need to update the windowing and triggering strategy for the pipeline. You want to update the pipeline without any loss of in-flight messages. What is the best way to deploy the changes?

- A. Stop with pipeline using the drain option and use new Dataflow pipeline
- B. Stop with pipeline using the cancel option and use new Dataflow pipeline

- **C. Pass the --update option with --jobname parameter to the same name as the job you want to update**
- D. Pass the --update option with --jobname parameter to the new job name you want to update

Your company is planning to migrate its data first to Google Cloud Storage. You need to keep the contents of this bucket in sync with a new Google Cloud Storage bucket to support a backup storage destination. What is the best method to achieve this?

- A. Once per week, use a gsutil cp command to copy over newly modified files.
- **B. Use gsutil rsync commands to keep both locations in sync.**
- C. Use Storage Transfer Service to keep both the source and destination in sync.
- D. Use gsutil -m cp to keep both locations in sync.

Your company hosts a 2PB on-premises Hadoop cluster with sensitive data. They want to plan the migration of the cluster to Google Cloud as part of phase 1 activity before the jobs are moved. Current network speed between the colocation and cloud is 10Gbps. What is the efficient way to transfer the data?

- **A. Use Transfer appliance to transfer the data to Cloud Storage**
- B. Expose the data as a public URL and Storage Transfer Service to transfer it
- C. Use gsutil command to transfer the data to Cloud Storage
- D. Use hadoop distcp command to copy the data between cluster

You have migrated your Hadoop jobs with external dependencies on a Dataproc cluster. As a security requirement, the cluster has been setup using internal IP addresses only and does not have a direct Internet connectivity. How can the cluster be configured to allow the installation of the dependencies?

- A. Setup a SSH tunnel to Internet and route outbound requests through it.
- **B. Store the external dependencies in Cloud Storage and modify the initialization scripts**
- C. Setup a SOCKS proxy and route outbound requests through it.
- D. Setup the Dataproc master node is public subnet to be able to download external dependencies

You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud. You are using ANSI SQL to run queries for your analysts. You want to support complex aggregate queries and reuse existing code. How should you transform the input data?

- A. Use BigQuery for storage. Use Cloud Dataflow to run the transformations.
- **B. Use BigQuery for storage. Use Cloud Dataproc to run the transformations.**
- C. Use Cloud Storage for storage. Use Cloud Dataflow to run the transformations.
- D. Use Cloud Storage for storage. Use Cloud Dataproc to run the transformations.

As part of your backup plan, you set up regular snapshots of Compute Engine instances that are running. You want to be able to restore these snapshots using the fewest possible steps for replacement instances. What should you do?

- A. Export the snapshots to Cloud Storage. Create disks from the exported snapshot files. Create images from the new disks. Use the image to create instances as needed.
- B. Export the snapshots to Cloud Storage. Create images from the exported snapshot files. Use the image to create instances as needed.
- C. Use the snapshots to create replacement disks. Use the disks to create instances as needed.
- **D. Use the snapshots to create replacement instances as needed.**

You are asked to design next generation of smart helmet for accident detection and reporting system. Each helmet will push 10kb of biometric data In JSON format every 1 second to a collection platform that will process and use trained machine learning model to predict and detect if an accident happens and send notification. Management has tasked you to architect the platform ensuring the following requirements are met:

- Provide the ability for real-time analytics of the inbound biometric data
- Ensure ingestion and processing of the biometric data is highly durable. Elastic and parallel
- The results of the analytic processing should be persisted for data mining to improve the accident detection ML model in the future.

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

- A. Utilize Cloud Storage to collect the inbound sensor data, analyze data with Dataproc and save the results to BigQuery.
- **B. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to BigQuery.**
- C. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Cloud SQL.
- D. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Bigtable.

Your company processes high volumes of IoT data that are time-stamped. The total data volume can be several petabytes. The data needs to be written and changed at a high speed. You want to use the most performant storage option for your data. Which product should you use?

- A. Cloud Datastore
- B. Cloud Storage
- **C. Cloud Bigtable**
- D. BigQuery

A startup plans to use a data processing platform, which supports both batch and streaming applications. They would prefer to have a hands-off/serverless data processing platform to start with. Which GCP service is suited for them?

- A. Dataproc
- B. Dataprep
- **C. Dataflow**
- D. BigQuery

Your infrastructure runs on AWS and includes a set of multi-TB enterprise databases that are backed up nightly on the S3. You need to create a redundant backup to Google Cloud. You are responsible for performing scheduled monthly disaster recovery drills. You want to create a cost-effective solution. What should you do?

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

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

- **A. Use Cloud Pub/Sub for input. Attach a timestamp to every message in the publisher.**
- B. Use Cloud Pub/Sub for input. Attach a unique identifier to every message in the publisher.
- C. Use Apache Kafka on Compute Engine for input. Attach a timestamp to every message in the publisher.
- D. Use Apache Kafka on Compute Engine for input. Attach a unique identifier to every message in the publisher.

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

- **A. Export your trained model to a SavedModel format. Deploy and run your model on Cloud ML Engine.**
- B. Export your trained model to a SavedModel format. Deploy and run your model from a Kubernetes Engine cluster.
- C. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud Storage. Run 1 version on CPUs and another version on GPUs.
- D. Export 2 copies of your trained model to a SavedModel format. Store artifacts in Cloud ML Engine. Run 1 version on CPUs and another version on GPUs.

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

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

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

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

- A. Create a dedicated service account, and use encryption at rest to reference your data stored in Cloud Storage and Compute Engine data as part of your API service calls.
- B. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in Cloud Storage. Use your uploaded encryption key and reference it as part of your API service calls to encrypt your data in the Kafka node hosted on Compute Engine.
- C. Upload your own encryption key to Cloud Key Management Service, and use it to encrypt your data in your Kafka node hosted on Compute Engine.
- **D. Supply your own encryption key, and reference it as part of your API service calls to encrypt your data in Cloud Storage and your Kafka node hosted on Compute Engine.**

You have been asked to select the storage system for the click-data of your company's large portfolio of websites. This data is streamed in from a custom website analytics package at a typical rate of 6,000 clicks per minute. With bursts of up to 8,500 clicks per second. It must have been stored for future analysis by your data science and user experience teams. Which storage infrastructure should you choose?

- A. Google Cloud SQL
- **B. Google Cloud Bigtable**
- C. Google Cloud Storage
- D. Google Cloud Datastore

You are asked to design next generation of smart helmet for accident detection and reporting system. Each helmet will push 10kb of biometric data In JSON format every 1 second to a collection platform that will process and use trained machine learning model to predict and detect if an accident happens and send notification. Management has tasked you to architect the platform ensuring the following requirements are met:

- Provide the ability for real-time analytics of the inbound biometric data
- Ensure ingestion and processing of the biometric data is highly durable. Elastic and parallel
- The results of the analytic processing should be persisted for data mining to improve the accident detection ML model in the future.

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

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

- **B. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to BigQuery.**
- C. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Cloud SQL.
- D. Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Bigtable.

A company has its data distributed across multiple projects. They want to enable users to be able to execute BigQuery queries across dataset owned by the different projects. However, to optimize cost they want the billing to a single separate account. How should the access be controlled?

- **A. Add users to groups. Groups to have BigQuery User role for billing project and data viewer role to projects with dataset**
- B. Add users to groups. Groups to have BigQuery jobUser role for billing project and data viewer role to projects with dataset
- C. Add users to groups. Groups to have BigQuery User role for projects with dataset billing project and data viewer role to billing project
- D. Add users to groups. Groups to have BigQuery jobUser role for projects with dataset billing project and data viewer role to billing project

As part of a complex rollout, you have hired a third party developer consultant to assist with creating your Dataflow processing pipeline. The data that this pipeline will process is very confidential, and the consultant cannot be allowed to view the data itself. What actions should you take so that they have the ability to help build the pipeline but cannot see the data it will process?

- **A. Assign the consultant the Dataflow Developer IAM role.**
- B. Apply custom encryption to the data before it goes through the pipeline.
- C. Use a separate development project to construct the pipeline with example data, therefore not exposing the live data to the developer's work environment.
- D. Anonymize the data before it gets to the Dataflow pipeline.

Your company uses Google Analytics for tracking. You need to export the session and hit data from a Google Analytics 360 reporting view on scheduled basis into BigQuery for analysis. How can the data be exported?

- A. Configure a scheduler in Google Analytics to convert the Google Analytics data to JSON format, then import directly into BigQuery using bq command line.
- B. Use gsutil to export the Google Analytics data to Cloud Storage, then import into BigQuery and schedule it using Cron.
- C. Import data to BigQuery directly from Google Analytics using Cron
- **D. Use BigQuery Data Transfer Service to import the data from Google Analytics**

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

When using Cloud ML Engine to train machine learning models, how are online predictions different from batch predictions? (Choose TWO)

- A. Online prediction results are written to Cloud Storage as output.
- B. Batch predictions are used to reduce latency in serving predictions.
- **C. Online predictions are returned in the response message.**
- **D. Batch predictions are optimized to handle a high volume of prediction examples while running on more complex models.**

You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

- A. The message body for the sensor event is too large.
- B. Your custom endpoint has an out-of-date SSL certificate.
- C. Cloud Pub/Sub topic has too many messages published to it.
- **D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.**

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

- A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.
- B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.
- C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.
- **D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.**

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.**

You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity ‘Movie’ the property ‘actors’ and the property ‘tags’ have multiple values but the property ‘date released’ does not. A typical query would ask for all movies with actor= ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

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

```
indexes:
 - kind: Movie
 properties:
 - name: actors
 - name: date_released
 - kind: Movie
 properties:
 - name: tags
 - name: date_released
```

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

```
indexes:
 - kind: Movie
 properties:
 - name: actors
 - name: tags
 - name: date_released
```

- C. Set the following in your entity options: exclude_from_indexes = 'actors, tags'
- D. Set the following in your entity options: exclude_from_indexes = 'date_published'

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

- **A. Put the data into Google Cloud Storage.**
- B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
- C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
- D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

As a mandate for least privilege within the company, the external users have been provided the Cloud Dataproc Viewer role. Which action these external users perform?

- A. Submit a job.
- B. Create a cluster.
- C. Delete a cluster.
- **D. List the jobs.**

You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

- A. PCollection

- **B. Transform**
- C. Pipeline
- D. Sink API

You receive customer transactions through Cloud Pub/Sub and you are planning to use Google's Dataflow SDK to analyze customer data such as displayed below:

1, Tom, 555 X street
2, Tim, 553 Y street
3, Sam, 111 Z street

Your project requirement is to extract only the customer name from the data source and then write to an output PCollection. Which operation is best suited for the above data processing requirement?

- A. Sink API
- **B. ParDo**
- C. Transform
- D. Extract

Your BigQuery dataset contains 1500 tables. When conducting a query, you are limited to a maximum of 1000 tables that you can query at once. You need to query data across all 1500 tables. What should you do?

- A. Place tables into separate datasets.
- **B. If possible, merge the 1500 tables to bring the total number below 1000. You may still partition single tables to divide data for queries.**
- C. Export the data to Bigtable, and conduct your query inside of Bigtable.
- D. Create multiple views of chunks of the 1500 tables, then query the multiple views.

An application that relies on Cloud SQL to read infrequently changing data is predicted to grow dramatically. How can you increase capacity for more read-only clients?

- A. Configure high availability on the master node
- B. Establish an external replica in the customer's data center
- C. Use backups so you can restore if there's an outage
- **D. Configure read replicas.**

You have configured streaming data pipelines to ingest data from thousands of Internet of Things (IoT) devices, ingest it into BigQuery. The data is stored into ingestion-time partitioned table. You want to run SQL queries against your data for analysis. How would you query specific partitions in a BigQuery table?

- A. Use the DATE column in the WHERE clause
- B. Use the EXTRACT(DATE) clause
- **C. Use the _PARTITIONTIME pseudo-column in the WHERE clause**
- D. Use DATE BETWEEN in the WHERE clause

You are deploying 10,000 new IoT devices to collect temperatures in your warehouses globally. However, the source data is streamed in bursts and is not periodical and must be transformed before it can be used. How should you design the system in Google Cloud?

- A. Use Cloud Bigtable for fast input and cbt for ETL.
- B. Ingest data to Cloud Storage. Use Cloud Dataproc for ETL.
- C. Use Cloud Pub/Sub to buffer the data, and then use BigQuery for ETL.
- **D. Use Cloud Pub/Sub to buffer the data, and then use Cloud Dataflow for ETL.**

A client wants to store files from one location and retrieve them from another location. Security requirements are that no one should be able to access the contents of the file while it is hosted in the cloud. What is the best option?

- A. Default encryption should be sufficient
- B. Customer-Supplied Encryption Keys (CSEK)
- C. Customer Managed Encryption Keys (CMEK)
- **D. Client-side encryption**

A company wants to connect cloud applications to an Oracle database in its data center. Requirements are a maximum of 20 Gbps of data and a Service Level Agreement (SLA) of 99%. Which option best suits the requirements?

- A. Implement a high-throughput Cloud VPN connection
- B. Cloud Router with VPN
- **C. Dedicated Interconnect**
- D. Partner Interconnect

An application has the following data requirements. 1. It requires strongly consistent transactions. 2. Total data will be less than 500 GB. 3. The data does not need to be streaming or real time. Which data technology would fit these requirements?

- A. BigQuery
- B. Cloud Bigtable
- **C. Cloud SQL**
- D. Cloud Memorystore

Your company has hired an external consultant to help import its relational data into BigQuery for analysis. The consultant mentions the data needs to be denormalized in BigQuery. What are the two benefits of using denormalized data structures in BigQuery?

- A. Reduces the amount of data processed, reduces the amount of storage required.
- **B. Increases query speed, makes queries simpler.**
- C. Reduces the amount of storage required, increases query speed.
- D. Reduces the amount of data processed, increases query speed.

Your company is forecasting a sharp increase in the number and size of Apache Spark and Hadoop jobs being run on your local datacenter. You want to utilize the cloud to help you scale this upcoming demand with the least amount of operations work and code change. Which product should you use?

- A. Google Cloud Dataflow
- **B. Google Cloud Dataproc**
- C. Google Compute Engine

- D. Google Container Engine

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

- **A. Supervised learning to determine which transactions are most likely to be fraudulent.**
- B. Unsupervised learning to determine which transactions are most likely to be fraudulent.
- **C. Clustering to divide the transactions into N categories based on feature similarity.**
- D. Supervised learning to divide the transactions into N categories based on feature similarity.

You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

- **A. Linear regression**
- B. Logistic classification
- C. Recurrent neural network
- D. Feedforward neural network

A retailer wishes to identify the products, which are bought together. It already has a historical dataset containing a customer id, receipt id and the products bought. Which type of Machine Learning algorithm is suited to achieve this?

- A. Classification
- B. Regression
- **C. Association**
- D. Clustering

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

- A. Standard
- B. Regional
- **C. Coldline**
- D. Dual-Regional
- **E. Multi-Regional**

You are designing a photo sharing mobile app. Users will upload pictures from their mobile device directly and will be able to share pictures with others. As a compliance requirement, no image with offensive content should be allowed to be uploaded. How would you design your application?

- **A. Use Cloud Vision API to identify image with offensive content and mark it for manual checks.**
- B. Use Cloud Natural Language API to identify image with offensive content and mark it for manual checks.
- C. Use Cloud Image Intelligence API to identify image with offensive content and mark it for manual checks.

- D. Use Cloud Video Intelligence API to identify image with offensive content and mark it for manual checks.

Your company is working on a multi-cloud initiative. The data processing pipelines requires creating workflows that connect data, transfer data, processing, and using services across clouds. What cloud native tool should be used for orchestration?

- A. Cloud Scheduler
- B. Cloud Dataflow
- **C. Cloud Composer**
- D. Cloud Dataproc

You are responsible for security and access control to a BigQuery dataset hosted within a project. Multiple users from multiple teams need to have access to the different tables within the dataset. How can the access be control?

- **A. Create Authorized views for tables in a separate project and grant access to the teams**
- B. Create Authorized views for tables in a same project and grant access to the teams
- C. Create Materialized views for tables in a separate project and grant access to the teams
- D. Create Materialized views for tables in a same project and grant access to the teams

An organization wishes to automate data movement from Software as a Service (SaaS) applications such as Google Ads and Google Ad Manager on a scheduled, managed basis. This data is further needed for analytics and generate reports. How can the process be automated?

- A. Use Storage Transfer Service to move the data to Cloud Storage
- B. Use Storage Transfer Service to move the data to BigQuery
- **C. Use BigQuery Data Transfer Service to move the data to BigQuery**
- D. Use Transfer Appliance to move the data to Cloud Storage

Your company wants to develop a robotic car. The car needs to figure out a best way to traverse a path by it owns and the best way is the one with limited hurdles and shortest path. They aim to maximum a cumulative measure (say a reward) based on interactions with a given system. What type of machine learning needs to be applied?

- A. Supervised learning
- B. Unsupervised learning
- **C. Reinforcement learning**
- D. Dimensionality Reduction Technique

You are planning to embed a online customer support service within your website. Which GCP Service would allow you to design and integrate a conversational user interface into a mobile app, web application, device, bot, interactive voice response systems, and so on?

- A. Cloud Video Intelligence
- B. Cloud Vision
- C. Cloud Natural language
- **D. Dialogflow**

A bank wishes to predict that a given loan application will default in future. Given a dataset containing customer demographic information, loan application information, credit score and saving balance account information and a label containing default indicator (Y – Will Default, N – Will Not Default). Which type of Machine Learning algorithm is suited to achieve this?

- A. **Classification**
- B. Regression
- C. Association
- D. Clustering

Your company currently hosts an AWS S3 bucket. You need to keep the contents of this bucket in sync with a new Google Cloud Storage bucket to support a backup storage destination. What is the best method to achieve this?

- A. Once per week, use a gsutil cp command to copy over newly modified files.
- B. Use gsutil rsync commands to keep both locations in sync.
- **C. Use Storage Transfer Service to keep both the source and destination in sync.**
- D. Use gsutil -m cp to keep both locations in sync.

You are part of your company's migration team to transfer 1PB of data to Google Cloud. The network speed between the on-premises data center and Google Cloud is 100Mbps. The migration activity has a timeframe of 6 months. What is the efficient way to transfer the data?

- A. Use BigQuery Data Transfer Service to transfer the data to Cloud Storage
- B. Expose the data as a public URL and Storage Transfer Service to transfer it
- **C. Use Transfer appliance to transfer the data to Cloud Storage**
- D. Use gsutil command to transfer the data to Cloud Storage

You have migrated your Hadoop jobs with external dependencies on a Dataproc cluster. As a security requirement, the cluster has been setup using internal IP addresses only and does not have a direct Internet connectivity. How can the cluster be configured to allow the installation of the dependencies? (Choose two)

- A. Setup a SSH tunnel to Internet and route outbound requests through it.
- **B. Use Cloud Dataproc custom images instead of initialization actions to set up job dependencies.**
- C. Setup a SOCKS proxy and route outbound requests through it.
- **D. Setup a NAT Gateway to allow Dataproc cluster to download external dependencies**

You have a dataset in BigQuery storing transaction data with details of product and date purchased. Query fired on the data for a product using --dry-run shows that is performs a complete scan. How can the performance of the query be improved?

- A. Dry run always shows complete scan and the result would be different when the actual query is fired
- B. Use the limit parameter to limit the data queried
- C. Set maximum bytes on the query to limit the amount of data queried
- **D. Use Partitioning and clustering on the table**

Your company is building a package tracking application to track the complete lifecycle of the package. The data is stored in a BigQuery time partitioned table. Over the period of time the data in the table and grown manifold and Data Scientists are complaining of slowness in their package tracking queries. How can the table be modified to improve the performance and maintaining cost effectiveness?

- A. Import the table data to Bigtable
- B. Change the partitioned table column from time to date
- **C. Update to table to perform clustering on package id**
- D. Ask the Data Scientists to use LIMIT parameter on the queries

You are using Dataflow for running a real time streaming data processing pipeline. The pipeline currently uses 3 workers and is running on n1-standard-2 compute engine machine types. The pipeline is currently running slow and you want to increase its performance. How can you update the pipeline to improve the performance? (Choose 2)

- **A. Change workerMachineType machine type from n1-standard-2 to n1-standard-4**
- B. Move the Dataflow pipeline is a dedicated network
- **C. Modify the maxNumWorkers parameter to increase the worker nodes**
- D. Change workerMachineType machine type from n1-standard-2 to n1-standard-1

You are running your Spark jobs on Google Cloud using Dataproc. The jobs are running very slow and upon investigating you find out that only 1 core of the available 16 cores are being used for the jobs. How do you tune the cluster to use all the cores and improve the job performance?

- A. Pass the spark.driver.cores to parameter tune the number of cores used
- B. Increase the number of task nodes
- **C. Pass the spark.executor.cores parameter to tune the number of cores used**
- D. Update the job application to use all of the available cores

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

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

Your company has successfully migrated to the cloud and wants to analyze their data stream to optimize operations. They do not have any existing code for this analysis, so they are exploring all their options. These options include a mix of batch and stream processing, as they are running some hourly jobs and live processing some data as it comes in. Which technology should they use for this?

- A. Google Cloud Dataproc
- **B. Google Cloud Dataflow**
- C. Google Container Engine with Bigtable
- D. Google Compute Engine with Google BigQuery

Your company is developing an Online training platform. The platform allows instructors to host their courses for users. You need a build a feature that would generate automated transcript for the videos upload, which the instructor can further fine tune. What is the best and quick method to build this feature?

- A. Extract audio and Use Cloud Vision API to generate the transcript
- **B. Extract audio and Use Cloud Speech-to-Text API to generate the transcript**
- C. Extract audio and Use Cloud Natural Language API to generate the transcript
- D. Extract audio and Use Cloud Video Intelligence API to generate the transcript

A company has its data stored within a single project acme-company-project. Users across teams need to be able to access various tables within that dataset. Each team has a separate project acme-company-team-00x created. How can the access be control while billing only the team querying the dataset?

- A. Create Authorized views for tables required by the team in their respective project. Grant BigQuery User role for acme-company-team-00x and data viewer role to acme-company-project dataset
- **B. Create Authorized views for tables required by the team in their respective project. Grant BigQuery User role for acme-company-team-00x and data viewer role to acme-company-team-00x dataset - MORE**
- **C. Create Authorized views for tables required by the team in their respective project. Grant BigQuery JobUser role for acme-company-team-00x and data viewer role to acme-company-team-00x dataset - LESS**
- D. Create Authorized views for tables required by the team in the acme-company-project project. Grant BigQuery User role for acme-company-team-00x and data viewer role to acme-company-team-00x dataset

You are tasked with building an online analytical processing (OLAP) marketing analytics and reporting tool. This requires a relational database that can operate on hundreds of terabytes of data. What is the Google recommended tool for such applications?

- A. Cloud Spanner, because it is globally distributed
- B. Cloud SQL, because it is a fully managed relational database
- C. Cloud Firestore, because it offers real-time synchronization across devices
- **D. BigQuery, because it is designed for large-scale processing of tabular data**

You work for a manufacturing plant that batches application log files together into a single log file once a day at 2:00 AM. You have written a Google Cloud Dataflow job to process that log file. You need to make sure the log file in processed once per day as inexpensively as possible. What should you do?

- A. Change the processing job to use Google Cloud Dataproc instead.
- B. Manually start the Cloud Dataflow job each morning when you get into the office.
- **C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.**
- D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?

- A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.
- B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
- C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
- **D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.**

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

- **A. There are very few occurrences of mutations relative to normal samples.**
- B. There are roughly equal occurrences of both normal and mutated samples in the database.
- C. You expect future mutations to have different features from the mutated samples in the database.
- **D. You expect future mutations to have similar features to the mutated samples in the database.**
- E. You already have labels for which samples are mutated and which are normal in the database.

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)

- A. Create a new view over events using standard SQL
- B. Create a new partitioned table using a standard SQL query

- C. Create a new view over events_partitioned using standard SQL
- **D. Create a service account for the ODBC connection to use for authentication**
- **E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"**

You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery. How should you securely run this workload?

- A. Restrict the Google Cloud Storage bucket so only you can see the files
- B. Grant the Project Owner role to a service account, and run the job with it
- **C. Use a service account with the ability to read the batch files and to write to BigQuery**
- D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

- A. Add a node to the MySQL cluster and build an OLAP cube there.
- **B. Use an ETL tool to load the data from MySQL into Google BigQuery.**
- C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.
- D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.

You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

- **A. Get more training examples**
- B. Reduce the number of training examples
- **C. Use a smaller set of features**
- D. Use a larger set of features
- **E. Increase the regularization parameters**
- F. Decrease the regularization parameters

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

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

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

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour. The data scientists have written the following code to read the data for a new key features in the logs.

```
BigQueryIO.Read
        .named("ReadLogData")
        .from("clouddataflow-readonly:samples.log_data")
```

You want to improve the performance of this data read. What should you do?

- A. Specify the TableReference object in the code.
- B. Use .fromQuery operation to read specific fields from the table.
- **C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.**
- D. Call a transform that returns TableRow objects, where each element in the PCollection represents a single row in the table.

You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on non-key columns. What should you do?

- A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
- B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
- **C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.**
- D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

- A. Use a row key of the form <timestamp>.
- B. Use a row key of the form <sensorid>.
- C. Use a row key of the form <timestamp>#<sensorid>.
- **D. Use a row key of the form <sensorid>#<timestamp>.**

Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

- A. Set a single global window to capture all the data.
- B. Set sliding windows to capture all the lagged data.
- **C. Use watermarks and timestamps to capture the lagged data.**
- D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?

- A. They have not assigned the timestamp, which causes the job to fail
- B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
- C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
- **D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created**

You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, the application was designed to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?

- A. Re-write the application to load accumulated data every 2 minutes.
- B. Convert the streaming insert code to batch load for individual messages.
- C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
- **D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.**

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

- A. Continuously retrain the model on just the new data.
- **B. Continuously retrain the model on a combination of existing data and the new data.**
- C. Train on the existing data while using the new data as your test set.
- D. Train on the new data while using the existing data as your test set.

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?

- **A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.**
- B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.

- C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
- D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

A. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.

- B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
- **C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.**
- D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app_events_YYYYMMDD. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?

- **A. Use the TABLE_DATE_RANGE function**
- B. Use the WHERE _PARTITIONTIME pseudo column
- C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
- D. Use SELECT IF(date >= YYYY-MM-DD AND date <= YYYY-MM-DD)

Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly. How should you optimize the cluster for cost?

- A. Migrate the workload to Google Cloud Dataflow
- **B. Use pre-emptible virtual machines (VMs) for the cluster**
- C. Use a higher-memory node so that the job runs faster
- D. Use SSDs on the worker nodes so that the job can run faster

You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?

- A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to ‘none’ using a Cloud Dataproc job.
- **B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.**
- C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to ‘none’ using a Cloud Dataprep job.

- D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.

You are developing an application that uses a recommendation engine on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of data. What should you do?

- A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc. Call the model from your application.
- B. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.
- **C. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.**
- D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application. What should you do?

- A. Create groups for your users and give those groups access to the dataset
- B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
- **C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset**
- D. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

- A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
- **B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.**
- C. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
- D. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

- A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
- B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
- **C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.**
- D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.

Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

- **A. Use Google Stackdriver Audit Logs to review data access.**
- B. Get the identity and access management (IAM) policy of each table
- C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
- D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.

Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiency?

- **A. Assign global unique identifiers (GUID) to each data entry.**
- B. Compute the hash value of each data entry, and compare it with all historical data.
- C. Store each data entry as the primary key in a separate database and apply an index.
- D. Maintain a database table to store the hash value and other metadata for each data entry.

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

- **A. PigLatin using Pig**
- B. HiveQL using Hive
- C. Java using MapReduce
- D. Python using MapReduce

Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?

- **A. Cloud Bigtable**
- B. Google BigQuery
- C. Google Cloud Storage
- D. Google Cloud Datastore

Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of data. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?

- A. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.
- B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.
- C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.
- **D. In a bucket on Cloud Storage that is accessible only by an App Engine service that collects user information and logs the access before providing a link to the bucket.**

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

- A. Increase the CPU size on your server.
- B. Increase the size of the Google Persistent Disk on your server.
- **C. Increase your network bandwidth from your datacenter to GCP.**
- D. Increase your network bandwidth from Compute Engine to Cloud Storage

You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

- **A. Linear regression**
- B. Logistic classification
- C. Recurrent neural network
- D. Feedforward neural network

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

A. No interaction by the user on the site for 1 hour

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

C. Has not completed a transaction

You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

- A. Use a fixed-time window with a duration of 60 minutes.

- B. Use a sliding time window with a duration of 60 minutes.
- **C. Use a session window with a gap time duration of 60 minutes.**
- D. Use a global window with a time based trigger with a delay of 60 minutes.

By default, which of the following windowing behavior does Dataflow apply to unbounded data sets?

- A. Windows at every 100 MB of data.
- **B. Single, Global Window.**
- C. Windows at every 1 minute.
- D. Windows at every 10 minutes.

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

You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of- Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required. You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

- A. Redis
- **B. HBase**
- C. MySQL
- **D. MongoDB**
- **E. Cassandra**
- F. HDFS with Hive

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**

Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

- A. The CSV data loaded in BigQuery is not flagged as CSV.

- B. The CSV data has invalid rows that were skipped on import.
- **C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.**
- D. The CSV data has not gone through an ETL phase before loading into BigQuery.

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

- A. Increase the cluster size with more non-preemptible workers.
- B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
- C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.
- **D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.**

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.**

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.

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.**

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.

You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS_yyyymmdd. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

- A. Convert all daily log tables into date-partitioned tables
- **B. Convert the sharded tables into a single partitioned table**
- C. Enable query caching so you can cache data from previous months
- D. Create separate views to cover each month, and query from these views

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.

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.**

You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:

- Each department should have access only to their data.
- Each department will have one or more leads who need to be able to create and update tables and provide them to their team.
- Each department has data analysts who need to be able to query but not modify data.

How should you set access to the data in BigQuery?

- A. Create a dataset for each department. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.
- **B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.**
- C. Create a table for each department. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.
- D. Create a table for each department. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

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.

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?

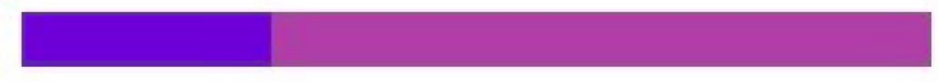
- A. Convert your batch BQ queries into interactive BQ queries.
- B. Create an additional project to overcome the 2K on-demand per-project quota.
- **C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.**
- D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:

SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country

You check the query plan for the query and see the following output in the Read section of Stage:1:

Larger image



What is the most likely cause of the delay for this query?

- **A. Users are running too many concurrent queries in the system**
- B. The [myproject:mydataset.mytable] table has too many partitions
- C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
- **D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew**