

 Custom View Settings**Topic 1 - Single Topic**

Question #1

Topic 1

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

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

Correct Answer: C

Reference:

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

Please note that there are tons of ways of further improving this result: design of layers and neurons, choosing different initialization and activation schemes, introduction of dropout layers of neurons, early stopping and so on. Furthermore, different types of deep learning models, such as recurrent neural networks might achieve better performance on this task. However, this is not the scope of this introductory post.

Community vote distribution

C (100%)

Question #2

Topic 1

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.** Most Voted
- 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.

Correct Answer: B*Community vote distribution*

B (90%)

10%

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

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

Correct Answer: C

Community vote distribution

C (100%)

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.

Correct Answer: A

Reference:

<https://support.google.com/datastudio/answer/7020039?hl=en>

How to tell if report data is cached

You can see if data is coming from the cache by viewing the report and looking in the bottom left corner. When all the charts on the current page are being served from the cache, you'll see a lightning bolt icon along with the time and date of the last update .

Blending and cached data

For a blended data source, the cache will use the setting that satisfies the desired refresh times for all of the data sources included in the blend.

For example, if you blend a Sheets data source having a refresh time of 15 minutes, with a BigQuery data source having a refresh time of 4 hours, the resulting blended data source will have a refresh time of 15 minutes.

Community vote distribution

A (64%)

C (27%)

9%

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

- A. Use federated data sources, and check data in the SQL query.
- B. Enable BigQuery monitoring in Google Stackdriver and create an alert.
- C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.
- D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Most Voted

Correct Answer: D

Community vote distribution

D (100%)

Next Questions ➔

 Custom View Settings

Question #6

Topic 1

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

- A. Issue a command to restart the database servers.
- B. Retry the query with exponential backoff, up to a cap of 15 minutes.
- C. Retry the query every second until it comes back online to minimize staleness of data.
- D. Reduce the query frequency to once every hour until the database comes back online.

Correct Answer: B*Community vote distribution*

B (100%)

Question #7

Topic 1

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

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

Correct Answer: A*Community vote distribution*

A (100%)

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

Correct Answer: D

Community vote distribution

D (100%)

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

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

Which table name will make the SQL statement work correctly?

- A. 'bigquery-public-data.noaa_gsod.gsod'
- B. bigquery-public-data.noaa_gsod.gsod*
- C. 'bigquery-public-data.noaa_gsod.gsod'*
- D. 'bigquery-public-data.noaa_gsod.gsod*' Most Voted**

Correct Answer: D

Reference:

<https://cloud.google.com/bigquery/docs/wildcard-tables>

Filtering selected tables using _TABLE_SUFFIX

To restrict a query so that it scans only a specified set of tables, use the `_TABLE_SUFFIX` pseudo column in a `WHERE` clause with a condition that is a constant expression.

The `_TABLE_SUFFIX` pseudo column contains the values matched by the table wildcard. For example, the previous sample query, which scans all tables from the 1940s, uses a table wildcard to represent the last digit of the year:

```
FROM
    `bigquery-public-data.noaa_gsod.gsod194*`
```

Community vote distribution

D (71%)

B (29%)

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. Most Voted Most Voted
- C. Ensure that the data is encrypted at all times.
- D. Restrict BigQuery API access to approved users. Most Voted Most Voted
- E. Segregate data across multiple tables or databases. Most Voted
- F. Use Google Stackdriver Audit Logging to determine policy violations. Most Voted

Correct Answer: *BDF*

Community vote distribution

BDE (52%) BDF (41%) 7%

[◀ Previous Questions](#)

[Next Questions ➔](#)

 Custom View Settings

Question #11

Topic 1

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

- No interaction by the user on the site for 1 hour
- Has added more than \$30 worth of products to the basket
-
- Has not completed a transaction

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

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

Correct Answer: C*Community vote distribution*

C (100%)

Question #12

Topic 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. Most Voted
- C. Put each client's BigQuery dataset into a different table.
- D. Restrict a client's dataset to approved users. Most Voted
- 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. Most Voted

Correct Answer: BDF*Community vote distribution*

BDF (64%)

BEF (36%)

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 Most Voted

Correct Answer: A

Community vote distribution

| | | |
|---------|---------|----|
| D (47%) | A (44%) | 9% |
|---------|---------|----|

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. Most Voted Most Voted
- 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. Most Voted
- D. You expect future mutations to have similar features to the mutated samples in the database. Most Voted
- E. You already have labels for which samples are mutated and which are normal in the database.

Correct Answer: BC

Community vote distribution

| | | |
|----------|----------|-----|
| AC (50%) | AD (39%) | 11% |
|----------|----------|-----|

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

- A. Re-write the application to load accumulated data every 2 minutes.
- B. Convert the streaming insert code to batch load for individual messages. Most Voted
- 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. Most Voted

Correct Answer: D

Community vote distribution

| | |
|---------|---------|
| D (65%) | B (32%) |
|---------|---------|

 Custom View Settings

Question #16

Topic 1

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.

Correct Answer: A*Community vote distribution*

A (85%)

C (15%)

Question #17

Topic 1

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.

Correct Answer: D*Community vote distribution*

D (100%)

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

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

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

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

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

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

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

Correct Answer: BCE

Community vote distribution

BCD (85%)

Other

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. Most Voted

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.

Correct Answer: A

Community vote distribution

A (100%)

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

HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

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

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

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

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

Correct Answer: B

Community vote distribution

D (87%)

13%

[← Previous Questions](#)

[Next Questions →](#)

 Custom View Settings

Question #21

Topic 1

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 efficiently?

- A. Assign global unique identifiers (GUID) to each data entry. Most Voted
- 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. Most Voted

Correct Answer: D*Community vote distribution*

| | | |
|---------|---------|---------|
| A (48%) | D (37%) | B (15%) |
|---------|---------|---------|

Question #22

Topic 1

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. Most Voted

Correct Answer: B*Community vote distribution*

| |
|----------|
| D (100%) |
|----------|

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

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

Correct Answer: B

Community vote distribution

B (100%)

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

- A. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type. Reload the data.
- B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each row. Reference the column TS instead of the column DT from now on.
- C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.
- D. Add two columns to the table CLICK_STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEW to true. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true. Most Voted
- E. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM. Most Voted

Correct Answer: D

Community vote distribution

E (68%)

C (27%)

5%

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. Most Voted

Correct Answer: B

Community vote distribution

D (100%)

◀ Previous Questions

Next Questions ➔

 Custom View Settings

Question #26

Topic 1

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. Most Voted
- 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. Most Voted

Correct Answer: C*Community vote distribution*

D (82%) B (18%)

Question #27

Topic 1

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

- A. Eliminate features that are highly correlated to the output labels.
- B. Combine highly co-dependent features into one representative feature. Most Voted
- C. Instead of feeding in each feature individually, average their values in batches of 3.
- D. Remove the features that have null values for more than 50% of the training records.

Correct Answer: B*Community vote distribution*

B (85%) D (15%)

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. Most Voted
- 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.

Correct Answer: D

Community vote distribution

B (79%) D (21%)

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>. Most Voted

Correct Answer: D

Community vote distribution

D (100%)

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. Most Voted
- 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. Most Voted

Correct Answer: C

Community vote distribution

B (57%)

D (41%)

[◀ Previous Questions](#)

[Next Questions ▶](#)

 Custom View Settings

Question #31

Topic 1

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.

Correct Answer: D*Community vote distribution*

| | | |
|---------|---------|---------|
| A (53%) | B (32%) | D (15%) |
|---------|---------|---------|

Question #32

Topic 1

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.

Correct Answer: A*Community vote distribution*

| |
|----------|
| A (100%) |
|----------|

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. Most Voted
- C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages. Most Voted
- D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

Correct Answer: B

Community vote distribution

B (73%) C (18%) 9%

Flowlogistic Case Study -

Company Overview -

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

Company Background -

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

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

- Databases
 - 8 physical servers in 2 clusters
 - SQL Server " user data, inventory, static data
 - 3 physical servers
 - Cassandra " metadata, tracking messages
 - 10 Kafka servers " tracking message aggregation and batch insert
- Application servers " customer front end, middleware for order/customs
 - 60 virtual machines across 20 physical servers
 - Tomcat " Java services
 - Nginx " static content
 - Batch servers
- Storage appliances
 - iSCSI for virtual machine (VM) hosts
 - Fibre Channel storage area network (FC SAN) " SQL server storage
 - Network-attached storage (NAS) image storage, logs, backups
- 10 Apache Hadoop /Spark servers
- Core Data Lake
- Data analysis workloads
- 20 miscellaneous servers
 - Jenkins, monitoring, bastion hosts,

Business Requirements -

Build a reliable and reproducible environment with scaled parity of production.

- - Aggregate data in a centralized Data Lake for analysis
 - Use historical data to perform predictive analytics on future shipments
 - Accurately track every shipment worldwide using proprietary technology
 - Improve business agility and speed of innovation through rapid provisioning of new resources
 - Analyze and optimize architecture for performance in the cloud
 - Migrate fully to the cloud if all other requirements are met

Technical Requirements -

- Handle both streaming and batch data
- Migrate existing Hadoop workloads

- Ensure architecture is scalable and elastic to meet the changing demands of the company.
- Use managed services whenever possible
- Encrypt data flight and at rest
- Connect a VPN between the production data center and cloud environment

SEO Statement -

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

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

CTO Statement -

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

CFO Statement -

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

Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to

BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?

- A. Store the common data in BigQuery as partitioned tables.
- B. Store the common data in BigQuery and expose authorized views.
- C. Store the common data encoded as Avro in Google Cloud Storage. Most Voted
- D. Store the common data in the HDFS storage for a Google Cloud Dataproc cluster.

Correct Answer: B

Community vote distribution

| | | |
|---------|---------|-----|
| C (57%) | B (29%) | 14% |
|---------|---------|-----|

Flowlogistic Case Study -

Company Overview -

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

Company Background -

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

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- ⇒ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- ⇒ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

- ⇒ Databases
 - 8 physical servers in 2 clusters
 - SQL Server " user data, inventory, static data
 - 3 physical servers
 - Cassandra " metadata, tracking messages
 - 10 Kafka servers " tracking message aggregation and batch insert
 - ⇒ Application servers " customer front end, middleware for order/customs
 - 60 virtual machines across 20 physical servers
 - Tomcat " Java services
 - Nginx " static content
 - Batch servers
 - ⇒ Storage appliances
 - iSCSI for virtual machine (VM) hosts
 - Fibre Channel storage area network (FC SAN) " SQL server storage
 - Network-attached storage (NAS) image storage, logs, backups
 - ⇒ 10 Apache Hadoop /Spark servers
 - Core Data Lake
 - Data analysis workloads
 - ⇒ 20 miscellaneous servers
 - Jenkins, monitoring, bastion hosts,

Business Requirements -

- ⇒ Build a reliable and reproducible environment with scaled parity of production.
- ⇒ Aggregate data in a centralized Data Lake for analysis
- ⇒ Use historical data to perform predictive analytics on future shipments
- ⇒ Accurately track every shipment worldwide using proprietary technology
- ⇒ Improve business agility and speed of innovation through rapid provisioning of new resources
- ⇒ Analyze and optimize architecture for performance in the cloud
- ⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

- ⇒ Handle both streaming and batch data
- ⇒ Migrate existing Hadoop workloads
- ⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.

- Use managed services whenever possible
- Encrypt data flight and at rest
- Connect a VPN between the production data center and cloud environment

SEO Statement -

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

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

CTO Statement -

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

CFO Statement -

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

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

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

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

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

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

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

Correct Answer: C

Community vote distribution

A (94%) 6%

[◀ Previous Questions](#)

[Next Questions ➔](#)



- Expert Verified, Online, **Free**.

Custom View Settings

Flowlogistic Case Study -

Company Overview -

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

Company Background -

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

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

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

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

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

⇒ Databases

8 physical servers in 2 clusters

- SQL Server " user data, inventory, static data

3 physical servers

- Cassandra " metadata, tracking messages

10 Kafka servers " tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

- Tomcat " Java services

- Nginx " static content

- Batch servers

⇒ Storage appliances

- iSCSI for virtual machine (VM) hosts

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

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

⇒ 10 Apache Hadoop /Spark servers

- Core Data Lake

- Data analysis workloads

⇒ 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements -

⇒ Build a reliable and reproducible environment with scaled parity of production.

⇒ Aggregate data in a centralized Data Lake for analysis

⇒ Use historical data to perform predictive analytics on future shipments

⇒ Accurately track every shipment worldwide using proprietary technology

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

⇒ Analyze and optimize architecture for performance in the cloud

⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

Handle both streaming and batch data

-

- Migrate existing Hadoop workloads
- Ensure architecture is scalable and elastic to meet the changing demands of the company.
- Use managed services whenever possible
- Encrypt data flight and at rest
- Connect a VPN between the production data center and cloud environment

SEO Statement -

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

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

CTO Statement -

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

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment. Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?

- A. Export the data into a Google Sheet for virtualization.
- B. Create an additional table with only the necessary columns. Most Voted
- C. Create a view on the table to present to the virtualization tool. Most Voted
- D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

Correct Answer: C

Community vote distribution

C (60%)

B (40%)

Flowlogistic Case Study -

Company Overview -

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

Company Background -

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

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- ⇒ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- ⇒ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

- ⇒ Databases
 - 8 physical servers in 2 clusters
 - SQL Server " user data, inventory, static data
 - 3 physical servers
 - Cassandra " metadata, tracking messages
 - 10 Kafka servers " tracking message aggregation and batch insert
 - ⇒ Application servers " customer front end, middleware for order/customs
- 60 virtual machines across 20 physical servers
 - Tomcat " Java services
 - Nginx " static content
 - Batch servers
 - ⇒ Storage appliances
 - iSCSI for virtual machine (VM) hosts
 - Fibre Channel storage area network (FC SAN) " SQL server storage
 - Network-attached storage (NAS) image storage, logs, backups
 - ⇒ 10 Apache Hadoop /Spark servers
 - Core Data Lake
 - Data analysis workloads
 - ⇒ 20 miscellaneous servers
 - Jenkins, monitoring, bastion hosts,

Business Requirements -

- ⇒ Build a reliable and reproducible environment with scaled parity of production.
- ⇒ Aggregate data in a centralized Data Lake for analysis
- ⇒ Use historical data to perform predictive analytics on future shipments
- ⇒ Accurately track every shipment worldwide using proprietary technology
- ⇒ Improve business agility and speed of innovation through rapid provisioning of new resources
- ⇒ Analyze and optimize architecture for performance in the cloud
- ⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

- ⇒ Handle both streaming and batch data
- ⇒ Migrate existing Hadoop workloads
- ⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.

- Use managed services whenever possible
- Encrypt data flight and at rest
- Connect a VPN between the production data center and cloud environment

SEO Statement -

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

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

CTO Statement -

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

CFO Statement -

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

Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single

Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in

Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.

Which approach should you take?

A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.

B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Cloud Pub/Sub. Most Voted

C. Use the NOW () function in BigQuery to record the event's time.

D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.

Correct Answer: B

Community vote distribution

B (67%)

D (33%)

MJTelco Case Study -**Company Overview -**

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- ⇒ Provide reliable and timely access to data for analysis from distributed research workers
- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

- ⇒ Ensure secure and efficient transport and storage of telemetry data
- ⇒ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
- ⇒ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
- ⇒ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

A. The zone

- B. The number of workers
- C. The disk size per worker
- D. The maximum number of workers Most Voted

Correct Answer: A

Community vote distribution

D (100%)

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⦿ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⦿ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⦿ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
 - ⦿ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
 - ⦿ Provide reliable and timely access to data for analysis from distributed research workers
- Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
-

Technical Requirements -

- ⦿ Ensure secure and efficient transport and storage of telemetry data
- ⦿ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
- ⦿ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
- ⦿ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You need to compose visualizations for operations teams with the following requirements:

- ⦿ The report must include telemetry data from all 50,000 installations for the most recent 6 weeks (sampling once every minute).
- ⦿ The report must not be more than 3 hours delayed from live data.
- ⦿ The actionable report should only show suboptimal links.

- ↳ Most suboptimal links should be sorted to the top.
- ↳ Suboptimal links can be grouped and filtered by regional geography.
- ↳ User response time to load the report must be <5 seconds.

Which approach meets the requirements?

- Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
- Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
- Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
- Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table. Most Voted

Correct Answer: C

Community vote distribution

D (75%)

C (25%)

MJTelco Case Study -**Company Overview -**

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.

- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

Provide reliable and timely access to data for analysis from distributed research workers

▪

- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

- ⇒ Ensure secure and efficient transport and storage of telemetry data

- ⇒ Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

- ⇒ Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

- ⇒ Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.

Which two actions should you take? (Choose two.)

- A. Ensure all the tables are included in global dataset.
- B. Ensure each table is included in a dataset for a region. Most Voted Most Voted
- C. Adjust the settings for each table to allow a related region-based security group view access. Most Voted
- D. Adjust the settings for each view to allow a related region-based security group view access.
- E. Adjust the settings for each dataset to allow a related region-based security group view access. Most Voted

Correct Answer: *BD*

Community vote distribution

BE (53%) BC (33%) 8%

[◀ Previous Questions](#)

[Next Questions ➔](#)



- Expert Verified, Online, **Free**.

Custom View Settings

MJTelco Case Study -

Company Overview -

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- Provide reliable and timely access to data for analysis from distributed research workers
- Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

- Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
- Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
- Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day.

Which schema should you use?

- A. Rowkey: date#device_id Column data: data_point Most Voted
- B. Rowkey: date Column data: device_id, data_point
- C. Rowkey: device_id Column data: date, data_point Most Voted
- D. Rowkey: data_point Column data: device_id, date
- E. Rowkey: date#data_point Column data: device_id

Correct Answer: D

Community vote distribution

| | | |
|---------|---------|-----|
| A (55%) | C (35%) | 10% |
|---------|---------|-----|

Question #42

Topic 1

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. Most Voted
- C. Increase the size of the Hadoop cluster.
- D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Correct Answer: A

Community vote distribution

| |
|----------|
| B (100%) |
|----------|

Question #43

Topic 1

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. Most Voted
- 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. Most Voted
- 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.

Correct Answer: C

Community vote distribution

| | | | |
|---------|---------|---------|----|
| A (44%) | B (30%) | C (19%) | 7% |
|---------|---------|---------|----|

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

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

Indexes:

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

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

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

Indexes:

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

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

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

Correct Answer: A

Community vote distribution

A (79%)

D (21%)

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 is 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. Most Voted
- D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Correct Answer: C

Community vote distribution

C (100%)

 Custom View Settings

Question #46

Topic 1

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

What should you do?

- A. Load the data every 30 minutes into a new partitioned table in BigQuery.
- B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery Most Voted
- C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore
- D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Correct Answer: B*Community vote distribution*

| | | |
|---------|-----|----|
| B (78%) | 13% | 9% |
|---------|-----|----|

Question #47

Topic 1

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

- The user profile: What the user likes and doesn't like to eat
- The user account information: Name, address, preferred meal times
- The order information: When orders are made, from where, to whom

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

- A. BigQuery Most Voted
- B. Cloud SQL
- C. Cloud Bigtable
- D. Cloud Datastore Most Voted

Correct Answer: A*Community vote distribution*

| | | |
|---------|---------|---------|
| A (43%) | B (37%) | D (20%) |
|---------|---------|---------|

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

Correct Answer: C

Community vote distribution

C (100%)

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

You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (Choose two.)

- A. Introduce data compression for each file to increase the rate file of file transfer.
- B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
- C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel. Most Voted
- D. Assemble 1,000 files into a tape archive (TAR) file. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them. Most Voted
- E. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.

Correct Answer: CE

Community vote distribution

CD (82%)

Other

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 Most Voted

C. MySQL

D. MongoDB Most Voted

E. Cassandra Most Voted

F. HDFS with Hive

Correct Answer: BDF

Community vote distribution

BDE (94%) 6%

◀ Previous Questions

Next Questions ▶

 Custom View Settings

Question #51

Topic 1

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 Most Voted
- B. Reduce the number of training examples
- C. Use a smaller set of features Most Voted
- D. Use a larger set of features
- E. Increase the regularization parameters Most Voted
- F. Decrease the regularization parameters

Correct Answer: ADF*Community vote distribution*

ACE (93%) 7%

Question #52

Topic 1

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

Correct Answer: B*Community vote distribution*

C (71%) B (29%)

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

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

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



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

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

Correct Answer: A

Community vote distribution

D (87%) 13%

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. Most Voted
- 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. Most Voted

Correct Answer: C

Community vote distribution

B (51%) D (47%)

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

- A. Create a new view over events using standard SQL
- B. Create a new partitioned table using a standard SQL query
- C. Create a new view over events_partitioned using standard SQL Most Voted
- D. Create a service account for the ODBC connection to use for authentication Most Voted
- E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection to shared `events`

Correct Answer: AE

Community vote distribution

CD (53%) DE (24%) CE (24%)

◀ Previous Questions

Next Questions ▶

 Custom View Settings

Question #56

Topic 1

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)

Correct Answer: A

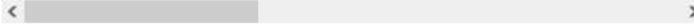
Reference:

<https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understand-your-mobile-app?hl=am>

Building complex queries

What if we want to run a query across both platforms of our app over a specific date range? Since Firebase Analytics data is split into tables for each day, we can do this using BigQuery's **TABLE_DATE_RANGE** function. This query returns a count of the cities users are coming from over a one week period:

```
01  SELECT
02    user_dim.geo_info.city,
03    COUNT(user_dim.geo_info.city) as city_count
04  FROM
05  TABLE_DATE_RANGE([firebase-analytics-sample-data:alpha], DATE_SUB(CURRENT_DATE(), INTERVAL 7 DAY), CURRENT_DATE())
06  TABLE_DATE_RANGE([firebase-analytics-sample-data:beta], DATE_SUB(CURRENT_DATE(), INTERVAL 7 DAY), CURRENT_DATE())
07  GROUP BY
08    user_dim.geo_info.city
09  ORDER BY
10    city_count DESC
```

*Community vote distribution*

A (100%)

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

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

Correct Answer: C

Community vote distribution

D (77%) A (15%) 8%

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

- A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
- B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this. Most Voted
- C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
- D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Correct Answer: A

Community vote distribution

B (100%)

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

A. BigQuery Most Voted

B. Cloud SQL Most Voted

C. Cloud BigTable

D. Cloud Datastore

Correct Answer: *B*

Reference:

<https://cloud.google.com/sql/>

Cloud SQL

Fully managed relational database service for MySQL, PostgreSQL, and SQL Server. Run the same relational databases you know with their rich extension collections, configuration flags and developer ecosystem, but without the hassle of self management.

[Try Cloud SQL free](#)

[Contact sales](#)

- ✓ Reduce maintenance cost with fully managed [MySQL](#), [PostgreSQL](#) and [SQL Server](#) databases
- ✓ Ensure business continuity with reliable and secure services backed by 24/7 SRE team
- ✓ Automate database provisioning, storage capacity management, and other time-consuming tasks
- ✓ Database observability made easy for developers with Cloud SQL Insights
- ✓ Easy integration with existing apps and Google Cloud services like GKE and BigQuery

Community vote distribution

B (68%)

A (30%)

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 Most Voted
- 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

Correct Answer: A

Community vote distribution

B (94%) 6%

[◀ Previous Questions](#)

[Next Questions ▶](#)

 Custom View Settings

Question #61

Topic 1

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 Most Voted
- 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

Correct Answer: A*Community vote distribution*

B (100%)

Question #62

Topic 1

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?

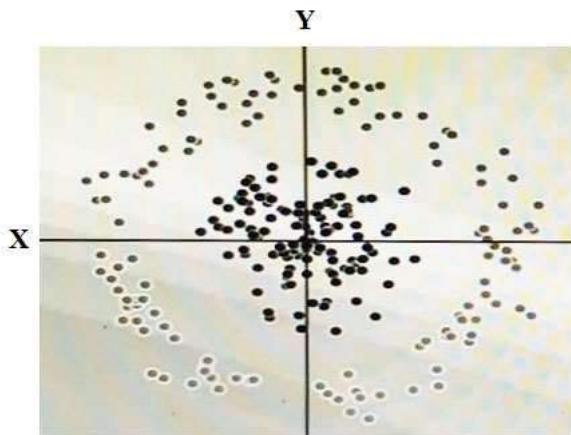
- A. Set a single global window to capture all the data.
- B. Set sliding windows to capture all the lagged data.
- C. Use watermarks and timestamps to capture the lagged data. Most Voted
- D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for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

Correct Answer: B*Community vote distribution*

C (87%)

7%

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



- A. $X^2 + Y^2$ Most Voted
- B. X^2
- C. Y^2
- D. $\cos(X)$

Correct Answer: D

Community vote distribution

| | |
|---------|---------|
| A (76%) | B (24%) |
|---------|---------|

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 Most Voted
- 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

Correct Answer: C

Community vote distribution

| |
|----------|
| C (100%) |
|----------|

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. Most Voted
- C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataprep job.
- D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 0 using a custom script.

Correct Answer: C

Community vote distribution

B (100%)

[◀ Previous Questions](#)

[Next Questions ➔](#)

 Custom View Settings

Question #66

Topic 1

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. Most Voted
- 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.

Correct Answer: C*Community vote distribution*

B (100%)

Question #67

Topic 1

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.

Correct Answer: C*Community vote distribution*

C (100%)

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. Most Voted
- 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.

Correct Answer: B

Community vote distribution

C (96%) 4%

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. Most Voted
- 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.

Correct Answer: B

Community vote distribution

A (50%) C (38%) 12%

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

Correct Answer: D

Community vote distribution

B (73%) A (27%)

[◀ Previous Questions](#)

[Next Questions ➔](#)

 Custom View Settings

Question #71

Topic 1

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

- A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.
- B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.
- C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.
- D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Correct Answer: A*Community vote distribution*

A (100%)

Question #72

Topic 1

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

Correct Answer: A*Community vote distribution*

C (100%)

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

Correct Answer: C

Community vote distribution

C (100%)

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

Which product should they use to store the data?

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

Correct Answer: A

Reference:

<https://cloud.google.com/bigtable/docs/schema-design-time-series>

The basic design patterns for storing time-series data in Bigtable are as follows:

- Rows are time buckets
 - New columns for new events
 - New cells for new events
- Rows represent single timestamps
 - Serialized column data
 - Unserialized column data

Community vote distribution

A (79%)

C (21%)

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

- A. Create and share an authorized view that provides the aggregate results. Most Voted
- B. Create and share a new dataset and view that provides the aggregate results. Most Voted
- C. Create and share a new dataset and table that contains the aggregate results.
- D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Correct Answer: A

Reference:

<https://cloud.google.com/bigquery/docs/share-access-views>

BigQuery is a petabyte-scale analytics data warehouse that you can use to run SQL queries over vast amounts of data in near real time.

Giving a view access to a dataset is also known as creating an [authorized view](#) in BigQuery. An authorized view lets you share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query. In this tutorial, you create an authorized view.

Community vote distribution

A (51%)

B (49%)

◀ Previous Questions

Next Questions ➔

 Custom View Settings

Question #76

Topic 1

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

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

Correct Answer: B*Community vote distribution*

B (56%)

D (42%)

Question #77

Topic 1

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

- A. Subsample your test dataset.
- B. Subsample your training dataset. 
- C. Increase the number of input features to your model.
- D. Increase the number of layers in your neural network.**

Correct Answer: D

Reference:

<https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d>

Neural networks are machine learning algorithms that provide state of the accuracy on many use cases. But, a lot of times the accuracy of the network we are building might not be satisfactory or might not take us to the top positions on the leaderboard in data science competitions. Therefore, we are always looking for better ways to improve the performance of our models. There are many techniques available that could help us achieve that. Follow along to get to know them and to build your own accurate neural network.

Community vote distribution

B (94%)

6%

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

Correct Answer: D

Community vote distribution

A (80%) C (20%)

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

Correct Answer: C

Community vote distribution

C (63%) A (38%)

MJTelco Case Study -**Company Overview -**

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- ⇒ Provide reliable and timely access to data for analysis from distributed research workers
- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

MJTelco is building a custom interface to share data. They have these requirements:

1. They need to do aggregations over their petabyte-scale datasets.
2. They need to scan specific time range rows with a very fast response time (milliseconds).

Which combination of Google Cloud Platform products should you recommend?

- A. Cloud Datastore and Cloud Bigtable
- B. Cloud Bigtable and Cloud SQL
- C. BigQuery and Cloud Bigtable
- D. BigQuery and Cloud Storage

Correct Answer: C

Community vote distribution

C (100%)

[◀ Previous Questions](#)

[Next Questions ▶](#)



- Expert Verified, Online, **Free**.

Custom View Settings

MJTelco Case Study -**Company Overview -**

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⦿ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⦿ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⦿ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- ⦿ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- ⦿ Provide reliable and timely access to data for analysis from distributed research workers
- ⦿ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You need to compose visualization for operations teams with the following requirements:

- ⦿ Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)
- ⦿ The report must not be more than 3 hours delayed from live data.
- ⦿ The actionable report should only show suboptimal links.

- Most suboptimal links should be sorted to the top.
- Suboptimal links can be grouped and filtered by regional geography.

- User response time to load the report must be <5 seconds.

You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?

A. Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.

B. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.

Most Voted

C. Export the data to a spreadsheet, compose a series of charts and tables, one for each possible combination of criteria, and spread them across multiple tabs.

D. Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API. **Most Voted**

Correct Answer: B

Community vote distribution

D (65%)

B (35%)

MJTelco Case Study -**Company Overview -**

MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.

Company Background -

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and providers in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

- ⇒ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- ⇒ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments "development/test, staging, and production" to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements -

- ⇒ Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
- ⇒ Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
- ⇒ Provide reliable and timely access to data for analysis from distributed research workers
- ⇒ Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements -

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement -

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.

CTO Statement -

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement -

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?

- A. Create a table called tracking_table and include a DATE column.

B. Create a partitioned table called tracking_table and include a TIMESTAMP column.

C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.

D. Create a table called tracking_table with a TIMESTAMP column to represent the day.

Correct Answer: B

Community vote distribution

B (100%)

Flowlogistic Case Study -

Company Overview -

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

Company Background -

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

Solution Concept -

Flowlogistic wants to implement two concepts using the cloud:

- ⇒ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- ⇒ Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.

Existing Technical Environment -

Flowlogistic architecture resides in a single data center:

- ⇒ Databases
 - 8 physical servers in 2 clusters
 - SQL Server " user data, inventory, static data
 - 3 physical servers
 - Cassandra " metadata, tracking messages
 - 10 Kafka servers " tracking message aggregation and batch insert
- ⇒ Application servers " customer front end, middleware for order/customs
 - 60 virtual machines across 20 physical servers
 - Tomcat " Java services
 - Nginx " static content
 - Batch servers
- ⇒ Storage appliances
 - iSCSI for virtual machine (VM) hosts
 - Fibre Channel storage area network (FC SAN) " SQL server storage
- Network-attached storage (NAS) image storage, logs, backups
 - ⇒ 10 Apache Hadoop /Spark servers
- Core Data Lake
 - Data analysis workloads
 - ⇒ 20 miscellaneous servers
 - Jenkins, monitoring, bastion hosts,

Business Requirements -

- ⇒ Build a reliable and reproducible environment with scaled parity of production.
- ⇒ Aggregate data in a centralized Data Lake for analysis
- ⇒ Use historical data to perform predictive analytics on future shipments
- ⇒ Accurately track every shipment worldwide using proprietary technology
- ⇒ Improve business agility and speed of innovation through rapid provisioning of new resources
- ⇒ Analyze and optimize architecture for performance in the cloud
- ⇒ Migrate fully to the cloud if all other requirements are met

Technical Requirements -

- ⇒ Handle both streaming and batch data
- ⇒ Migrate existing Hadoop workloads
- ⇒ Ensure architecture is scalable and elastic to meet the changing demands of the company.

- Use managed services whenever possible
- Encrypt data flight and at rest
- Connect a VPN between the production data center and cloud environment

SEO Statement -

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

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

CTO Statement -

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

CFO Statement -

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment. Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.

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

- A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage Most Voted
- B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
- C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
- D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage
- E. Cloud Dataflow, Cloud SQL, and Cloud Storage

Correct Answer: C

Community vote distribution

A (100%)

Question #84

Topic 1

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

- A. Select random samples from the tables using the RAND() function and compare the samples.
- B. Select random samples from the tables using the HASH() function and compare the samples.
- C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table. Most Voted
- D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Correct Answer: B

Community vote distribution

C (86%)

14%

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

Correct Answer: C

Reference:

<https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery>

You might simply prefer a fixed monthly bill, or encounter workloads that are extremely sensitive to query latency, and thus have predictability and control requirements that cannot be met by the on-demand service. For such situations, you can use the [flat-rate service](#). In this model, a certain number of slots are dedicated to your project(s), and you can establish a hierarchical priority model amongst the projects. The flat-rate model is especially suitable for large enterprises with multiple business units and workloads with varying priorities and budgets. For instance, the arrangement illustrated below gives priority to queries that are issued from the "Dashboarding" project over the queries from the other two projects. But even with prioritization, slots won't be wasted. If the prioritized "Dashboarding" project does not use all its dedicated slots, they'll be distributed among the remaining projects. Even data stored in the "Data Science" project can be queried from the "Dashboarding" project with a higher priority than when it's queried from within the "Data Science" project itself.

Community vote distribution

C (100%)

[◀ Previous Questions](#)

[Next Questions ➔](#)

 Custom View Settings

Question #86

Topic 1

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

- A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS. Most Voted
- B. Deploy a Kafka cluster on GCE VM Instances with the Pub/Sub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
- C. Deploy the Pub/Sub Kafka connector to your on-prem Kafka cluster and configure Pub/Sub as a Source connector. Use a Dataflow job to read from Pub/Sub and write to GCS.
- D. Deploy the Pub/Sub Kafka connector to your on-prem Kafka cluster and configure Pub/Sub as a Sink connector. Use a Dataflow job to read from Pub/Sub and write to GCS.

Correct Answer: A*Community vote distribution*

A (100%)

Question #87

Topic 1

You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffling operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload. What should you do?

- A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
- B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.
- C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
- D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size. Most Voted

Correct Answer: D*Community vote distribution*

D (61%)

A (30%)

9%

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 Pub/Sub PubSub directly from the DoFn. Most Voted**
- 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 Pub/Sub later. Most Voted

Correct Answer: C

Community vote distribution

D (73%)

C (27%)

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

What should you do?

- A. Provide latitude and longitude as input vectors to your neural net.
- B. Create a numeric column from a feature cross of latitude and longitude.**
- C. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L1 regularization during optimization. Most Voted
- D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.

Correct Answer: B

Reference:

<https://cloud.google.com/bigquery/docs/gis-data>

Loading geospatial data

Single points on Earth can be described by just a longitude, latitude pair. For example, you can load a CSV file that contains longitude and latitude values and then use the `ST_GEOPOINT` function to convert them into `GEOGRAPHY` values.

For more complex geographies, you can load the following geospatial data formats into a `GEOGRAPHY` column:

- Well-known text (WKT)
- Well-known binary (WKB)
- GeoJSON

Community vote distribution

C (88%)

6%

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. Most Voted

Correct Answer: C

Community vote distribution

D (82%) A (18%)

[◀ Previous Questions](#)

[Next Questions ➔](#)

 Custom View Settings

Question #91

Topic 1

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

What should you do?

- A. Increase the size of the dataset by collecting additional data.
- B. Train a linear regression to predict a credit default risk score. Most Voted
- C. Remove the bias from the data and collect applications that have been declined loans.
- D. Match loan applicants with their social profiles to enable feature engineering. Most Voted

Correct Answer: B*Community vote distribution*

| | | |
|---------|---------|---------|
| B (54%) | D (25%) | C (21%) |
|---------|---------|---------|

Question #92

Topic 1

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 Most Voted

Correct Answer: D*Community vote distribution*

| |
|----------|
| D (100%) |
|----------|

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

What should you do?

- A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
- B. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload. Most Voted
- C. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload. Most Voted
- D. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

Correct Answer: B

Community vote distribution

C (65%) B (35%)

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

- A. Batch job, PubSubIO, side-inputs
- B. Streaming job, PubSubIO, JdbcIO, side-outputs
- C. Streaming job, PubSubIO, BigQueryIO, side-inputs Most Voted
- D. Streaming job, PubSubIO, BigQueryIO, side-outputs

Correct Answer: C

Community vote distribution

C (94%) 6%

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

- A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
- B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
- C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
Most Voted
- D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
Most Voted
- E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

Correct Answer: AC

Community vote distribution

CD (86%) 14%

◀ Previous Questions

Next Questions ▶

 Custom View Settings

Question #96

Topic 1

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.

You have the following requirements:

- You will batch-load the posts once per day and run them through the Cloud Natural Language API.
- You will extract topics and sentiment from the posts.
- You must store the raw posts for archiving and reprocessing.
- You will create dashboards to be shared with people both inside and outside your organization.

You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

- A. Store the social media posts and the data extracted from the API in BigQuery.
- B. Store the social media posts and the data extracted from the API in Cloud SQL.
- C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.** Most Voted
- D. Feed the social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Correct Answer: C*Community vote distribution*

C (100%)

Question #97

Topic 1

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

What should you do?

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

Correct Answer: A*Community vote distribution*

B (94%)

6%

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

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

Correct Answer: A

Community vote distribution

A (100%)

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

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

Correct Answer: C

Community vote distribution

C (100%)

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

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

Correct Answer: C

Community vote distribution

B (100%)

[← Previous Questions](#)

[Next Questions →](#)

25 votes.

