# Professional Data Engineer.38q

Number: Google Data Engineer
Passing Score: 800
Time Limit: 120 min



**Google Certified Professional - Data Engineer**

**Exam A**

**QUESTION 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.
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:** D
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 2**
Your company is using `WILDCARD` tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

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

Which table name will make the SQL statement work correctly?

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

**Correct Answer:** D
**Section: (none)**
**Explanation**

**Explanation/Reference:**
Reference: https://cloud.google.com/bigquery/docs/wildcard-tables

**QUESTION 3**
Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)



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



**Correct Answer:** BDF
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 4**
Your startup has never implemented a formal security policy. Currently, everyone in the company has access to the datasets stored in Google BigQuery. Teams have freedom to use the service as they see fit, and they have not documented their use cases. You have been asked to secure the data warehouse. You need to discover what everyone is doing. What should you do first?

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

**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 5**
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:** A
**Section: (none)**
**Explanation**



**Explanation/Reference:**

**QUESTION 6**
Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use?
(Choose three.)

A. Supervised learning to determine which transactions are most likely to be fraudulent.
B. Unsupervised learning to determine which transactions are most likely to be fraudulent.
C. Clustering to divide the transactions into N categories based on feature similarity.
D. Supervised learning to predict the location of a transaction.
E. Reinforcement learning to predict the location of a transaction.
F. Unsupervised learning to predict the location of a transaction.

**Correct Answer:** BCE

**Section: (none)**
**Explanation**
**Explanation/Reference:**

**QUESTION 7**
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.

**Correct Answer:** B
**Section: (none)**
**Explanation**

**Explanation/Reference:**



**QUESTION 8**
Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiency?

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

**Correct Answer:** D
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 9**
You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table `CLICK_STREAM`. The column `DT` stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the `STRING` type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the `TIMESTAMP`. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

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



**Correct Answer:** D
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 10**
You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

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

**Correct Answer:** B

**Section: (none)**
**Explanation**

**Explanation/Reference:**
**QUESTION 11**
Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour. The data scientists have written the following code to read the data for a new key features in the logs.

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

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

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



**Correct Answer:** D
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 12**
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>`.

**Correct Answer:** A

**Section: (none)**
**Explanation**

**Explanation/Reference:**
**QUESTION 13**
Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations. The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

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

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**



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

**Technical Requirements**

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

**SEO Statement**

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.

**CTO Statement**

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

**CFO Statement**

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

Flowlogistic 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.
D. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.

**Correct Answer:** B
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 15**

**Flowlogistic Case Study**

**Company Overview**

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

**Company Background**

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

**Solution Concept**

Flowlogistic wants to implement two concepts using the cloud:

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



**Existing Technical Environment**

Flowlogistic architecture resides in a single data center:

- Databases
  8 physical servers in 2 clusters
  - SQL Server – user data, inventory, static data
  3 physical servers
  - Cassandra – metadata, tracking messages

10 Kafka servers – tracking message aggregation and batch insert

- Application servers – customer front end, middleware for order/customs
  60 virtual machines across 20 physical servers
  - Tomcat – Java services
  - Nginx – static content
  - Batch servers
- Storage appliances
  - iSCSI for virtual machine (VM) hosts
  - Fibre Channel storage area network (FC SAN) – SQL server storage

- Network-attached storage (NAS) image storage, logs, backups
- 10 Apache Hadoop /Spark servers
  - Core Data Lake
  - Data analysis workloads
- 20 miscellaneous servers
  - Jenkins, monitoring, bastion hosts,

**Business Requirements**

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

**Technical Requirements**



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

**SEO Statement**

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.

**CTO Statement**

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

**CFO Statement**

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

Flowlogistic's 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.
C. Create a view on the table to present to the virtualization tool.
D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 16**
**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 provides 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.
C. Adjust the settings for each table to allow a related region-based security group view access.
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.

**Correct Answer:** BD
**Section: (none)**
**Explanation**

**Explanation/Reference:**
**QUESTION 17**
Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?



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

**Correct Answer:** A

**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 18**
You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a `Users` table consisting of a `FirstName` field and a `LastName` field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a `FullName` field consisting of the value of the `FirstName` field concatenated with a space, followed by the value of the `LastName` field for each employee. How can you make that data available while minimizing cost?

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

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 19**
You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity `'Movie'` the property `'actors'` and the property `'tags'` have multiple values but the property `'date released'` does not. A typical query would ask for all movies with `actor=<actorname>` ordered by `date_released` or all movies with `tag=Comedy` ordered by `date_released`. How should you avoid a combinatorial explosion in the number of indexes? A. Manually configure the index in your index config as follows:

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

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

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



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

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

**Correct Answer:** A
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 20**
You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:
- The user profile: What the user likes and doesn't like to eat
- The user account information: Name, address, preferred meal times
- The order information: When orders are made, from where, to whom

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

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

**Correct Answer:** A
**Section: (none)**
**Explanation**
**Explanation/Reference:**



**QUESTION 21**
You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

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

**Correct Answer:** ADF
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 22**
You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery. How should you securely run this workload?

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

**Correct Answer:** B
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 23**
Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?

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

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 24**
Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

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

**Correct Answer:** B
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 25**
You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application. What should you do?

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

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 26**
You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?

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

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 27**
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?



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

**Correct Answer:** B
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 28**
Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?

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

**Correct Answer:** B
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 29**
You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?

A. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.
D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.



**Correct Answer:** A
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 30**
You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on non-key columns. What should you do?

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

**Correct Answer:** D
**Section: (none)**
**Explanation**

**Explanation/Reference:**
Reference: https://cloud.google.com/solutions/data-lifecycle-cloud-platform

**QUESTION 31**
Your financial services company is moving to cloud technology and wants to store 50 TB of financial time-series data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?

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

**Correct Answer:** A
**Section: (none)**
**Explanation**



**Explanation/Reference:**
Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series

**QUESTION 32**
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
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 33**
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
**Section: (none)**
**Explanation**

**Explanation/Reference:**
Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

**QUESTION 34**
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
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 35**
Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?

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

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 36**
**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 provides 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.
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.

**Correct Answer:** B
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 37**
**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 provides 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
**Section: (none)**
**Explanation**

**Explanation/Reference:**

**QUESTION 38**
**Flowlogistic Case Study**

**Company Overview**

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

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

**Solution Concept**

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

**Existing Technical Environment**

Flowlogistic architecture resides in a single data center:

- Databases
  - 8 physical servers in 2 clusters
  - SQL Server – user data, inventory, static data

- 3 physical servers
- Cassandra – metadata, tracking messages

10 Kafka servers – tracking message aggregation and batch insert

- Application servers – customer front end, middleware for order/customs
  - 60 virtual machines across 20 physical servers
  - Tomcat – Java services
  - Nginx – static content
  - Batch servers
- Storage appliances
  - iSCSI for virtual machine (VM) hosts
  - Fibre Channel storage area network (FC SAN) – SQL server storage

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

- 10 Apache Hadoop /Spark servers
  - Core Data Lake
  - Data analysis workloads
- 20 miscellaneous servers
  - Jenkins, monitoring, bastion hosts,



**Business Requirements**

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

**Technical Requirements**

- Handle both streaming and batch data ▪

Migrate existing Hadoop workloads

- Ensure architecture is scalable and elastic to meet the changing demands of the company.
- Use managed services whenever possible
- Encrypt data flight and at rest

Connect a VPN between the production data center and cloud environment

**SEO Statement**
We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.

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

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

Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?



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

**Correct Answer:** C
**Section: (none)**
**Explanation**

**Explanation/Reference:**