Top of Form

Question 1:

**Skipped**

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?

* 

**B. Grant the consultant the Cloud Dataflow Developer role on the project.**

**(Correct)**

* 

**A. Grant the consultant the Viewer role on the project.**

* 

**D. Create an anonymized sample of the data for the consultant to work with in a different project.**

* 

**C. Create a service account and allow the consultant to log on with it.**

**Explanation**

Yes it is B. From https://cloud.google.com/dataflow/docs/concepts/access-control "The developer who creates and examines jobs needs the roles/dataflow.admin role. For more sophisticated permissions management, the developer interacting with the Dataflow job needs the roles/dataflow.developer role. They need the roles/storage.objectAdmin or a related role to stage the required files. For debugging and quota checking, they need the project roles/compute.viewer role. Absent other role assignments, this role lets the developer create and cancel Dataflow jobs, but not interact with the individual VMs or access other Cloud services. The worker service account needs the roles/dataflow.worker role to process data for the Dataflow service. To access job data, the service account needs other roles such as roles/storage.objectAdmin."

Bottom of Form

Top of Form

Question 2:

**Skipped**

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?

* 

**B. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.**

**(Correct)**

* 

**A. Send the data to Google Cloud Datastore and then export to BigQuery.**

* 

**C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.**

* 

**D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.**

**Explanation**

B as the need to ingest it, transform and store the Cloud Pub/Sub, Cloud Dataflow, BigQuery is ideal stack to handle the IoT data. https://cloud.google.com/solutions/mobile/mobile-gaming-analysis-telemetry Option A is wrong as the Datastore is not an ideal ingestion service. Option C is wrong as Cloud Storage is not an ideal ingestion service and Dataproc is not a data warehousing solution. Option D is wrong as Cloud SQL is not a data warehousing solution.

Bottom of Form

Top of Form

Question 3:

**Skipped**

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?

* 

**B. Combine highly co-dependent features into one representative feature.**

**(Correct)**

* 

**D. Remove the features that have null values for more than 50% of the training records.**

* 

**A. Eliminate features that are highly correlated to the output labels.**

* 

**C. Instead of feeding in each feature individually, average their values in batches of 3.**

**Explanation**

Vote for 'B' combining features to createte a new feature is a step of "Feature construction" or decomposing or splitting features to create new features. Ideally, PCA should be apply if we want to reduce the dimension. Removing those columns / features - where Data is miss > 50% (may improve the speed) - but will decrease the accuracy as well. So instead of dropping features where we have missing data, we need to impute something

Bottom of Form

Top of Form

Question 4:

**Skipped**

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 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?**

* 

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

* 

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

* 

**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 Clod Pub/Sub.**

**(Correct)**

**Explanation**

Answer: B A. There is no indication that the application can do this. Moreover, due to networking problems, it is possible that Pub/Sub doesn't receive messages in order. This will analysis difficult. B. This makes sure that you have access to publishing timestamp which provides you with the correct ordering of messages. C. If timestamps are already messed up, BigQuery will get wrong results anyways. D. The timestamp we are interested in is when the data was produced by the publisher, not when it was received by Pub/Sub.

Bottom of Form

Top of Form

Question 5:

**Skipped**

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?

* 

**D. Maintain a database table to store the hash value and other metadata for 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.**

* 

**A. Assign global unique identifiers (GUID) to each data entry.**

**(Correct)**

**Explanation**

The best answer is "A". Answer "D" is not as efficient or error-proof due to two reasons 1. You need to calculate hash at sender as well as at receiver end to do the comparison. Waste of computing power. 2. Even if we discount the computing power, we should note that the system is sending inventory information. Two messages sent at different can denote same inventory level (and thus have same hash). Adding sender time stamp to hash will defeat the purpose of using hash as now retried messages will have different timestamp and a different hash. if timestamp is used as message creation timestamp than that can also be used as a UUID.

Bottom of Form

Top of Form

Question 6:

**Skipped**

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.)

* 

**C. Put each client's BigQuery dataset into a different table.**

* 

**E. Only allow a service account to access the datasets.**

**(Correct)**

* 

**B. Load data into a different dataset for each client.**

**(Correct)**

* 

**F. Use the appropriate identity and access management (IAM) roles for each client's users.**

* 

**D. Restrict a client's dataset to approved users.**

**(Correct)**

* 

**A. Load data into different partitions.**

**Explanation**

According to me the answer is B, D AND E. When an identity calls a Google Cloud API, BigQuery requires that the identity has the appropriate permissions to use the resource. You can grant permissions by granting roles to a user, a group, or a service account. The question is about securing the data so that clients cannot see each other's data. A service account is a Google Account that is associated with your Google Cloud project

Bottom of Form

Top of Form

Question 7:

**Skipped**

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?

* 

**D. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.**

**(Correct)**

* 

**B. Grant the user access to Google Cloud Shell.**

* 

**C. Host a visualization tool on a VM on Google Compute Engine.**

* 

**A. Run a local version of Jupiter on the laptop.**

**Explanation**

D as Cloud Datalab provides a powerful interactive, scalable tool on Google Cloud with the ability to analyze, visualize data. https://cloud.google.com/datalab/ Cloud Datalab is a powerful interactive tool created to explore, analyze, transform and visualize data and build machine learning models on Google Cloud Platform. It runs on Google Compute Engine and connects to multiple cloud sendees easily so you can focus on your data science tasks.

Bottom of Form

Top of Form

Question 8:

**Skipped**

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?

* 

**D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.**

* 

**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.**

* 

**A. Update the current pipeline and use the drain flag.**

**(Correct)**

**Explanation**

Correct Option : A Explanation:-This option is correct as the key requirement is not to lose the data, the Dataflow pipeline can be stopped using the Drain option. Drain options would cause Dataflow to stop any new processing, but would also allow the existing processing to complete

Bottom of Form

Top of Form

Question 9:

**Skipped**

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.)

* 

**B. There are roughly equal occurrences of both normal and mutated samples in the database.**

* 

**D. You expect future mutations to have similar features to the mutated samples in the database.**

**(Correct)**

* 

**A. There are very few occurrences of mutations relative to normal samples.**

* 

**E. You already have labels for which samples are mutated and which are normal in the database.**

* 

**C. You expect future mutations to have different features from the mutated samples in the database.**

**Explanation**

I think that AD makes more sense. D is the explanation you gave. In the rest, A makes more sense, in any anomaly detection algorithm it is assumed a priori that you have much more "normal" samples than mutated ones, so that you can model normal patterns and detect patterns that are "off" that normal pattern. For that you will always need the no. of normal samples to be much bigger than the no. of mutated samples.

Bottom of Form

Top of Form

Question 10:

**Skipped**

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?

* 

**D. Clear your browser history for the past hour then reload the tab showing the virtualizations.**

* 

**C. Refresh your browser tab showing the visualizations.**

* 

**B. Disable caching in BigQuery by editing table details.**

* 

**A. Disable caching by editing the report settings.**

**(Correct)**

**Explanation**

A as Data Studio caches data for performance and as the latest data is not shown, the caching can be disabled to fetch the latest data. Refer GCP documentation - Data Studio Caching: https://support.google.com/datastudio/answer/7020039?hl=en Option B is wrong as BigQuery does not cache the data. Options C & D are wrong this would not allow fetching of latest data

Bottom of Form

Top of Form

Question 11:

**Skipped**

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?

* 

**C. "˜bigquery-public-data.noaa\_gsod.gsod'\***

* 

**D. "˜bigquery-public-data.noaa\_gsod.gsod\*`**

**(Correct)**

* 

**B. bigquery-public-data.noaa\_gsod.gsod\***

* 

**A. "˜bigquery-public-data.noaa\_gsod.gsod"˜**

**Explanation**

correct answer is D. The syntax is FROM `bigquery-public-data.noaa\_gsod.gsod\*` As most of the people suggested, the option D is not mentioned correctly in the questions.

Bottom of Form

Top of Form

Question 12:

**Skipped**

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?

* 

**D. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.**

* 

**B. Use an ETL tool to load the data from MySQL into Google BigQuery.**

**(Correct)**

* 

**C. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.**

* 

**A. Add a node to the MySQL cluster and build an OLAP cube there.**

**Explanation**

I think it is B and not D: 1) There are no info regarding date freshness required for analytics. So nightly backup might be not enough as a source because it will only provide info one tie a day. 2) Dataproc is recommended as easiest way for migration of hadoop processes. SO no reason to use Dataproc for designing a new analytics processes. 3) The solution is really very limited if you will extend it in the future and add new data sources or create new aggregate tables. Where they should be created? 4) There is no info on which version is on prem MySQL database (I am not an expert in MySql) but I can imagine there might be compartibility issue for backup / restore between different releases

Bottom of Form

Top of Form

Question 13:

**Skipped**

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

* 

**B. Retry the query with exponential backoff, up to a cap of 15 minutes.**

**(Correct)**

* 

**A. Issue a command to restart the database servers.**

* 

**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.**

**Explanation**

Correct answer is B. App engine create applications that use Cloud SQL database connections effectively. Below is what is written in google cloud documnetation. If your application attempts to connect to the database and does not succeed, the database could be temporarily unavailable. In this case, sending too many simultaneous connection requests might waste additional database resources and increase the time needed to recover. Using exponential backoff prevents your application from sending an unresponsive number of connection requests when it can't connect to the database. This retry only makes sense when first connecting, or when first grabbing a connection from the pool. If errors happen in the middle of a transaction, the application must do the retrying, and it must retry from the beginning of a transaction. So even if your pool is configured properly, the application might still see errors if connections are lost. reference link is https://cloud.google.com/sql/docs/mysql/manage-connections

Bottom of Form

Top of Form

Question 14:

**Skipped**

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?

* 

**C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.**

* 

**B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.**

* 

**A. Put the data into Google Cloud Storage.**

**(Correct)**

* 

**D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.**

**Explanation**

Ans: A B: Wrong eVM wont solve the problem of larger storage prices. C: May be, but nothing mentioned in terms of what to tune in the question, also this is like-for-like migration so tuning may not be part of the migration. D: Again, this is like-for-like so need to define what is hot data and which is cold data, also persistent disk costlier than cloud storage.

Bottom of Form

Top of Form

Question 15:

**Skipped**

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?

* 

**B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.**

**(Correct)**

* 

**C. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.**

* 

**D. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.**

* 

**A. Check the dashboard application to see if it is not displaying correctly.**

**Explanation**

B as the issue can be debugged by running a fixed dataset and checking the output. Refer GCP documentation - Dataflow logging: https://cloud.google.com/dataflow/docs/guides/logging A is wrong as the Dashboard uses data provided by Dataflow, the input source for Dashboard seems to be the issue C is wrong as Monitoring will not help find missing messages in Cloud Pub/Sub D is wrong as Dataflow cannot be configured as Push endpoint with Cloud Pub/Sub.

Bottom of Form

Top of Form

Question 16:

**Skipped**

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?**

* 

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

* 

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

* 

**A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage**

**(Correct)**

* 

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

**Explanation**

A. B is incorrect, because local SSD wouldn't satisfy the needs. C is incorrect, because one of the requirements is 'Global', Cloud SQL is well suited for regional applications. Cloud Spanner is a better suit in that regard. D is incorrect, because Load Balancer is for web traffic, not messages.

Bottom of Form

Top of Form

Question 17:

**Skipped**

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.**

* 

**C. Store the common data encoded as Avro in Google Cloud Storage.**

**(Correct)**

* 

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

* 

**B. Store the common data in BigQuery and expose authorized views.**

**Explanation**

think the answer is C. Though data proc has connectors for cloud storage, Bigtable, and BigQuery, using a big query connector is a little more work compared to cloud storage and big table. The best thing while moving apache Hadoop and spark jobs to data proc is using cloud storage instead of HDFS. Your data in ORC, Parquet, Avro, or any other format will be used by different clusters or jobs, and you need data persistence if the cluster terminates. We can just replace HDFS:// with gs://. Even big query can read data in avro format. The best way to store data common to both workloads is cloud storage. Hence I am going with C.

Bottom of Form

Top of Form

Question 18:

**Skipped**

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?

* 

**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.**

**(Correct)**

* 

**B. Enable BigQuery monitoring in Google Stackdriver and create an alert.**

* 

**A. Use federated data sources, and check data in the SQL query.**

**Explanation**

The answer is D. An ETL pipeline will be implemented for this scenario. Check out handling invalid inputs in cloud data flow https://cloud.google.com/blog/products/gcp/handling-invalid-inputs-in-dataflow ParDos . . . and don’ts: handling invalid inputs in Dataflow using Side Outputs as a “Dead Letter” file

Bottom of Form

Top of Form

Question 19:

**Skipped**

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?

* 

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

**(Correct)**

* 

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

* 

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

* 

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

**Explanation**

D as the Cloud Pub/Sub will deliver duplicate messages only if there has been no acknowledgement from the subscriber. Refer GCP documentation - Cloud Pub/Sub FAQs - Duplicates: https://cloud.google.com/pubsub/docs/faq#duplicates Why are there too many duplicate messages? Cloud Pub/Sub guarantees at-least-once message delivery, which means that occasional duplicates are to be expected. However, a high rate of duplicates may indicate that the client is not acknowledging messages within the configured ack\_deadline\_seconds, and Cloud Pub/Sub is retrying the message delivery.

Bottom of Form

Top of Form

Question 20:

**Skipped**

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?

* 

**B. Logistic classification**

* 

**A. Linear regression**

**(Correct)**

* 

**C. Recurrent neural network**

* 

**D. Feedforward neural network**

**Explanation**

Correct answer is A. A tip here to decide when a liner regression should be used or logistics regression needs to be used. If you are forecasting that is the values in the column that you are predicting is numeric, it is always liner regression. If you are classifying, that is buy or no buy, yes or no, you will be using logistics regression.

Bottom of Form

Top of Form

Question 21:

**Skipped**

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?

* 

**B. Get the identity and access management IIAM) policy of each table**

* 

**D. Use the Google Cloud Billing API to see what account the warehouse is being billed to.**

* 

**A. Use Google Stackdriver Audit Logs to review data access.**

**(Correct)**

* 

**C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.**

**Explanation**

A is correct because this is the best way to get granular access to data showing which users are accessing which data. B is not correct because we already know that all users already have access to all data, so this information is unlikely to be useful. It will also not show what users have done, just what they can do. C is not correct because slot usage will not inform security policy. D is not correct because a billing account is typically shared among many people and will only show the amount of data queried and stored https://cloud.google.com/bigquery/docs/reference/auditlogs/#mapping-audit-entries-to-log-streams https://cloud.google.com/bigquery/docs/monitoring#slots-available

Bottom of Form

Top of Form

Question 22:

**Skipped**

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.  
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?

* 

**B. The number of workers**

* 

**C. The disk size per worker**

* 

**D. The maximum number of workers**

**(Correct)**

* 

**A. The zone**

**Explanation**

The correct answer is D. Please look for the details in below https://cloud.google.com/dataflow/docs/guides/specifying-exec-params We need to specify and set execution parameters for cloud data flow . Also, to enable autoscaling, set the following execution parameters when you start your pipeline: --autoscaling\_algorithm=THROUGHPUT\_BASED --max\_num\_workers=N The objective of autoscaling streaming pipelines is to minimize backlog while maximizing worker utilization and throughput, and quickly react to spikes in load. By enabling autoscaling, you don't have to choose between provisioning for peak load and fresh results. Workers are added as CPU utilization and backlog increase and are removed as these metrics come down. This way, you’re paying only for what you need, and the job is processed as efficiently as possible.

Bottom of Form

Top of Form

Question 23:

**Skipped**

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.**

**(Correct)**

* 

**D. Train on the new data while using the existing data as your test set.**

* 

**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.**

**Explanation**

I would go with B (but A is also possible If the market is fast changing) https://datascience.stackexchange.com/questions/12761/should-a-model-be-re-trained-if-new-observations-are-available: Suppose that your model attempts to predict customers' behavior, e.g. how likely is a customer to purchase your product given an offer tailored for him. Clearly, the market changes over time, customers' preferences change, and your competitors adjust. You should adjust as well, so you need to retrain periodically. In such a case I would recommend to add new data, but also omit old data that is not relevant anymore. If the market is fast changing, you should even consider retraining periodically based on new data only. https://docs.aws.amazon.com/machine-learning/latest/dg/retraining-models-on-new-data.html: It is a good practice to continuously monitor the incoming data and retrain your model on newer data if you find that the data distribution has deviated significantly from the original training data distribution

Bottom of Form

Top of Form

Question 24:

**Skipped**

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**

**(Correct)**

**Explanation**

D as the payment transactions would need a transactional data service Datastore can support the same. Also, it is fully managed with NoOps required. A is wrong as Cloud SQL would need infrastructure scaling. Although storage can be automatically scaled (up to a limit), instance type needs to be changed as per the load manually. B is wrong as BigQuery is a data warehousing option. C is wrong as Bigtable is not a relational database but a NoSQL option.

Bottom of Form

Top of Form

Question 25:

**Skipped**

Your company built a TensorFlow neutral-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?

* 

**B. Serialization**

* 

**D. Dimensionality Reduction**

* 

**A. Threading**

* 

**C. Dropout Methods**

**(Correct)**

**Explanation**

Vote for 'C' https://keras.io/api/layers/regularization\_layers/dropout/ Option A - wrong (as Threading is to make training faster) Option B - wrong (Serialization) used while saving the model Option D - Wrong (Dimensionality Reduction) - This is though core parameter, but in question it's mention model works okay on Training Data..So dimension is not the issue here

Bottom of Form

Top of Form

Question 26:

**Skipped**

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.)

* 

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

* 

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

**(Correct)**

* 

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

* 

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

* 

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

**(Correct)**

* 

**D. Supervised learning to predict the location of a transaction.**

**(Correct)**

**Explanation**

Answer: B, C, D Description: Fraud is not a feature, so unsupervised, location is given so supervised, Clustering can be done looking at the done with same features

Bottom of Form

Top of Form

Question 27:

**Skipped**

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?

* 

**B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.**

* 

**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.**

* 

**C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.**

**(Correct)**

* 

**A. Add capacity (memory and disk space) to the database server by the order of 200.**

**Explanation**

C is correct because this option provides the least amount of inconvenience over using pre-specified date ranges or one table per clinic while also increasing performance due to avoiding self-joins. A is not correct because adding additional compute resources is not a recommended way to resolve database schema problems. B is not correct because this will reduce the functionality of the database and make running reports more difficult. D is not correct because this will likely increase the number of tables so much that it will be more difficult to generate reports vs. the correct option. https://cloud.google.com/bigquery/docs/best-practices-performance-patterns https://cloud.google.com/bigquery/docs/reference/standard-sql/query-syntax#explicit-alias-visibility

Bottom of Form

Top of Form

Question 28:

**Skipped**

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>.**

* 

**C. Use a row key of the form <timestamp>#<sensorid>.**

* 

**B. Use a row key of the form <sensorid>.**

* 

**D. Use a row key of the form >#<sensorid>#<timestamp>.**

**(Correct)**

**Explanation**

A is not correct because this will cause most writes to be pushed to a single node (known as hotspotting) B is not correct because this will not allow for multiple readings from the same sensor as new readings will overwrite old ones. C is not correct because this will cause most writes to be pushed to a single node (known as hotspotting) D is correct because it will allow for retrieval of data based on both sensor id and timestamp but without causing hotspotting.

Bottom of Form

Top of Form

Question 29:

**Skipped**

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.)

* 

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

**(Correct)**

* 

**E. Adjust the settings for each dataset to allow a related region-based security group view access.**

* 

**A. Ensure all the tables are included in global dataset.**

* 

**B. Ensure each table is included in a dataset for a region.**

**(Correct)**

* 

**D. Adjust the settings for each view to allow a related region-based security group view access.**

**Explanation**

B is definitely correct. But now that we have table level access control in big query, will C also be a right answer? But because question says table is populated for each region, dataset is also created region specific considering which E is correct.

Bottom of Form

Top of Form

Question 30:

**Skipped**

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.**

**(Correct)**

**Explanation**

D as the key requirement is to have notification on a particular table. It can be achieved using advanced log filter to filter only the table logs and create a project sink to Cloud Pub/Sub for notification. Refer GCP documentation - Advanced Logs Filters: https://cloud.google.com/logging/docs/view/advanced-queries A is wrong as advanced filter will help in filtering. However, there is no notification sends. B is wrong as it would send all the logs and BigQuery does not provide notifications. C is wrong as it would send all the logs.

Bottom of Form

Top of Form

Question 31:

**Skipped**

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.)

* 

**C. Ensure that the data is encrypted at all times.**

* 

**D. Restrict BigQuery API access to approved users.**

**(Correct)**

* 

**A. Disable writes to certain tables.**

* 

**E. Segregate data across multiple tables or databases.**

* 

**F. Use Google Stackdriver Audit Logging to determine policy violations.**

**(Correct)**

* 

**B. Restrict access to tables by role.**

**(Correct)**

**Explanation**

Correct: BDF bigquery.tables.create Create new tables. bigquery.tables.delete Delete tables. bigquery.tables.export Export table data out of BigQuery. bigquery.tables.get Get table metadata. To get table data, you need bigquery.tables.getData. bigquery.tables.getData Get table data. This permission is required for querying table data. To get table metadata, you need bigquery.tables.get. bigquery.tables.list List tables and metadata on tables. bigquery.tables.setCategory Set policy tags in table schema. bigquery.tables.update Update table metadata. To update table data, you need bigquery.tables.updateData. bigquery.tables.updateData Update table data. To update table metadata, you need bigquery.tables.update.

Bottom of Form

Top of Form

Question 32:

**Skipped**

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?

* 

**B. Convert the streaming insert code to batch load for individual messages.**

**(Correct)**

* 

**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.**

* 

**A. Re-write the application to load accumulated data every 2 minutes.**

**Explanation**

B. Streams data into BigQuery one record at a time without needing to run a load job: https://cloud.google.com/bigquery/docs/reference/rest/v2/tabledata/insertAll Instead of using a job to load data into BigQuery, you can choose to stream your data into BigQuery one record at a time by using the tabledata.insertAll method. This approach enables querying data without the delay of running a load job: https://cloud.google.com/bigquery/streaming-data-into-bigquery The BigQuery Storage Write API is a unified data-ingestion API for BigQuery. It combines the functionality of streaming ingestion and batch loading into a single high-performance API. You can use the Storage Write API to stream records into BigQuery that become available for query as they are written, or to batch process an arbitrarily large number of records and commit them in a single atomic operation. Committed mode. Records are available for reading immediately as you write them to the stream. Use this mode for streaming workloads that need minimal read latency. https://cloud.google.com/bigquery/docs/write-api

Bottom of Form

Top of Form

Question 33:

**Skipped**

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?

* 

**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.**

* 

**A. Include ORDER BY DESK on timestamp column and LIMIT to 1.**

* 

**D. Use the ROW\_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.**

**(Correct)**

**Explanation**

Answer: D Description: Row Number equals 1 with partitioning will ensure only one record is fetched per partition

Bottom of Form

Top of Form

Question 34:

**Skipped**

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?

* 

**C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.**

* 

**A. Specify the TableReference object in the code.**

* 

**D. Call a transform that returns TableRow objects, where each element in the PCollection represents a single row in the table**

* 

**B. Use .fromQuery operation to read specific fields from the table.**

**(Correct)**

**Explanation**

B BigQueryIO.read.from() directly reads the whole table from BigQuery. This function exports the whole table to temporary files in Google Cloud Storage, where it will later be read from. This requires almost no computation, as it only performs an export job, and later Dataflow reads from GCS (not from BigQuery). BigQueryIO.read.fromQuery() executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs).

Bottom of Form

Top of Form

Question 35:

**Skipped**

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?

* 

**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.**

* 

**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.**

* 

**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)**

* 

**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.**

* 

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

**Explanation**

E as the column type cannot be changed and the column needs to casting loaded into a new table using either SQL Query or import/export. Refer GCP documentation - BigQuery Changing Schema: https://cloud.google.com/bigquery/docs/manually-changing-schemas#changing\_a\_columns\_data\_type A is wrong as with this approach all the data would be lost and needs to be reloaded B is wrong as numeric values cannot be used directly and would need casting. C is wrong as view is not materialized views, so the future queries would always be taxed as the casting would be done always.

Bottom of Form

Top of Form

Question 36:

**Skipped**

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?

* 

**D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.**

**(Correct)**

* 

**C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.**

* 

**E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.**

* 

**B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.**

* 

**A. Create a Google Cloud Dataflow job to process the data.**

**Explanation**

D is correct because it uses managed services, and also allows for the data to persist on GCS beyond the life of the cluster. A is not correct because the goal is to re-use their Hadoop jobs and MapReduce and/or Spark jobs cannot simply be moved to Dataflow. B is not correct because the goal is to persist the data beyond the life of the ephemeral clusters, and if HDFS is used as the primary attached storage mechanism, it will also disappear at the end of the cluster’s life. C is not correct because the goal is to use managed services as much as possible, and this is the opposite. E is not correct because the goal is to use managed services as much as possible, and this is the opposite.

Bottom of Form

Top of Form

Question 37:

**Skipped**

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?

* 

**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.**

* 

**B. The performance issue should be resolved over time as the site of the BigDate cluster is increased.**

* 

**A. Redefine the schema by evenly distributing reads and writes across the row space of the table.**

**(Correct)**

**Explanation**

A as the schema needs to be redesigned to distribute the reads and writes evenly across each table. Refer GCP documentation - Bigtable Performance: https://cloud.google.com/bigtable/docs/performance The table's schema is not designed correctly. To get good performance from Cloud Bigtable, it's essential to design a schema that makes it possible to distribute reads and writes evenly across each table. See Designing Your Schema for more information. https://cloud.google.com/bigtable/docs/schema-design Option B is wrong as increasing the size of cluster would increase the cost. Option C is wrong as single row key for frequently updated identifiers reduces performance Option D is wrong as sequential IDs would degrade the performance. A safer approach is to use a reversed version of the user's numeric ID, which spreads traffic more evenly across all of the nodes for your Cloud Bigtable table.

Bottom of Form

Top of Form

Question 38:

**Skipped**

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 visualizations for operations teams with the following requirements:  
✑ The report must include telemetry data from all 50,000 installations for the most resent 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?

* 

**A. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.**

* 

**B. 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.**

* 

**D. 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.**

**(Correct)**

* 

**C. 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.**

**Explanation**

D. A and B are incorrect, because Google Sheets are not the best fit to handle large amount of data. C may work, but it requires building an application which equates to more work. D is more efficient, therefore a better option.

Bottom of Form

Top of Form

Question 39:

**Skipped**

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?

* 

**C. Use a session window with a gap time duration of 60 minutes.**

**(Correct)**

* 

**B. Use a sliding time window with a duration of 60 minutes.**

* 

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

* 

**D. Use a global window with a time based trigger with a delay of 60 minutes.**

**Explanation**

The correct answer is C. There are 3 windowing concepts in dataflow and each can be used for below use case 1) Fixed window 2) Sliding window and 3) Session window. Fixed window = any aggregation use cases, any batch analysis of data, relatively simple use cases. Sliding window = Moving averages of data Session window = user session data, click data and real time gaming analysis. The question here is about user session data and hence session window.

Bottom of Form

Top of Form

Question 40:

**Skipped**

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?

* 

**D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.**

* 

**C. Create a view on the table to present to the virtualization tool.**

* 

**B. Create an additional table with only the necessary columns.**

**(Correct)**

* 

**A. Export the data into a Google Sheet for virtualization.**

**Explanation**

Answer: B A. Data is too big to be virtualized using Google Sheets. B. Creating the new table won't be able to keep up with the new data inserts. C. This will help to select appropriate columns as well as will be able to deal with new data inserts. D. You cannot restrict access to selected columns using IAM. Views should be used to do that.

Bottom of Form

**Continue**

Retake test

**Course content**

Practice tests

* Lecture completed

Start

Practice Test 1: PEACTICE TEST 1

* Lecture incomplete

Start

Practice Test 2: PRACTICE TEST 2

* Lecture incomplete

Start

Practice Test 3: PRACTICE TEST 3

Overview

Q&AQuestions and answers

Notes

Announcements

Reviews

Learning tools