Top of Form

Question 1:

**Skipped**

You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for  
BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account.  
What should you do?

* 

**B. Create an additional project to overcome the 2K on-demand per-project quota.**

* 

**C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.**

**(Correct)**

* 

**D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.**

* 

**A. Convert your batch BQ queries into interactive BQ queries.**

**Explanation**

Vote for 'C'. A - Interactive queries will not solve the issue (as it also compare available slots) B - Creating additional projects (not alllowed) D - Increase the slots (not possible) - max 2000 in on-demand

Bottom of Form

Top of Form

Question 2:

**Skipped**

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

* 

**B. Use the LIMIT keyword to reduce the number of rows returned.**

* 

**D. Use the bq query - -maximum\_bytes\_billed flag to restrict the number of bytes billed.**

* 

**C. Recreate the table with a partitioning column and clustering column.**

**(Correct)**

* 

**A. Create a separate table for each ID.**

**Explanation**

should be C: https://cloud.google.com/bigquery/docs/best-practices-costs

Bottom of Form

Top of Form

Question 3:

**Skipped**

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

* 

**D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.**

**(Correct)**

* 

**A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.**

* 

**C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.**

**(Correct)**

* 

**E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.**

* 

**B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.**

**Explanation**

Answer is C & D. C –> Adding more nodes to a cluster (not replication) can improve the write performance https://cloud.google.com/bigtable/docs/performance D –> since Google recommends adding nodes when storage utilization is > 70% https://cloud.google.com/bigtable/docs/modifying-instance#nodes

Bottom of Form

Top of Form

Question 4:

**Skipped**

You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?

* 

**A. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.**

* 

**D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.**

* 

**C. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.**

**(Correct)**

* 

**B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.**

**Explanation**

answer is C. A: limitation of 1500 dml B: Partitioning by item not supported C: Streaming for near real time D: Batch job will not be near real time. Near real time threshold is different for different use case. but a typical near real time is 5-10 secs. Batch job will invalidate this.

Bottom of Form

Top of Form

Question 5:

**Skipped**

You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?

* 

**B. Cloud Natural Language API**

* 

**C. Dialogflow Enterprise Edition**

* 

**D. Cloud AutoML Natural Language**

* 

**A. Cloud Speech-to-Text API**

**(Correct)**

**Explanation**

Option A - Cloud Speech-to-Text API. The question is just asking to " interpret customer voice commands" .. it does not mention anything related to sentiment analysis so NLP is not required. DialogFlow is more of a chat bot services typically suited for a "Service Desk" kind of setup - where clients will call a centralized helpdesk and automation is achieved through Chat bot services like - google Dialog flow

Bottom of Form

Top of Form

Question 6:

**Skipped**

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

* 

**C. Remove the bias from the data and collect applications that have been declined loans.**

* 

**B. Train a linear regression to predict a credit default risk score.**

**(Correct)**

* 

**D. Match loan applicants with their social profiles to enable feature engineering.**

* 

**A. Increase the size of the dataset by collecting additional data.**

**Explanation**

A is incorrect as you need to work with the data you have available C is an optimisation not a solution D is ethically incorrect and invasion to privacy, there could be several legal implications with this B although oversimplified but is a workable solution

Bottom of Form

Top of Form

Question 7:

**Skipped**

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

* 

**C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.**

* 

**A. Use Cloud Dataflow with Beam to detect errors and perform transformations.**

* 

**B. Use Cloud Dataprep with recipes to detect errors and perform transformations.**

**(Correct)**

* 

**D. Use federated tables in BigQuery with queries to detect errors and perform transformations.**

**Explanation**

Yes B. Honest speaking, sometime I thought the answers being posted here were intentionally to mislead people whose do not have proper knowledge on the subject, but just memorizing answers to pass the exam.

Bottom of Form

Top of Form

Question 8:

**Skipped**

You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.  
What should you do?

* 

**A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.**

**(Correct)**

* 

**D. Install the StackDriver Agent and configure the MySQL plugin.**

* 

**B. Place the MariaDB instances in an Instance Group with a Health Check.**

* 

**C. Install the StackDriver Logging Agent and configure fluentd in\_tail plugin to read MariaDB logs.**

**Explanation**

Answer : A MariaDB needs costume metrics , and stackdriver built-in monitoring tools will not provide these metrics . Opencensus Agent will do this for you For more info , refer to : https://cloud.google.com/monitoring/custom-metrics/open-census

Bottom of Form

Top of Form

Question 9:

**Skipped**

A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

* 

**D. Re-create the table using data partitioning on the package delivery date.**

* 

**B. Implement clustering in BigQuery on the package-tracking ID column.**

**(Correct)**

* 

**C. Tier older data onto Cloud Storage files, and leverage extended tables.**

* 

**A. Implement clustering in BigQuery on the ingest date column.**

**Explanation**

I think it is B, because clustering only on Date column doesnt make any sense. within single partition every record has same date , how it will sort the data , like it does for partition. So A does not make sense , it will be B only.

Bottom of Form

Top of Form

Question 10:

**Skipped**

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

* 

**A. Add a SideInput that returns a Boolean if the element is corrupt.**

* 

**B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.**

**(Correct)**

* 

**D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.**

* 

**C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.**

**Explanation**

B - seems to be better option since we need to filter out, question does not specify that we do need to store it into different Pcollection. https://beam.apache.org/documentation/transforms/python/overview/ ParDo is general purpose whereas partition splits the elements into do different pcollections. https://beam.apache.org/documentation/transforms/python/elementwise/partition/

Bottom of Form

Top of Form

Question 11:

**Skipped**

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

* 

**D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.**

* 

**B. Use Cloud Dataflow and write the data to Cloud Storage.**

* 

**A. Execute gsutil rsync from the on-premises servers.**

**(Correct)**

* 

**C. Write a job template in Cloud Dataproc to perform the data transfer.**

**Explanation**

should be A, dataflow is on cloud is external; "don't allow access from external IPs to their on-premises resources" so no dataflow.

Bottom of Form

Top of Form

Question 12:

**Skipped**

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

* 

**C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.**

**(Correct)**

* 

**D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.**

* 

**B. Select random samples from the tables using the HASH() function and compare the samples.**

* 

**A. Select random samples from the tables using the RAND() function and compare the samples.**

**Explanation**

Answer: C Description: Full comparison with this option, rest are comparison on sample which doesnot ensure all the data will be ok

Bottom of Form

Top of Form

Question 13:

**Skipped**

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems. What should you do?

* 

**B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.**

* 

**A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.**

**(Correct)**

* 

**C. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.**

* 

**D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.**

**Explanation**

I think it should be "A". In the question , it is mentioned, "You want to QUICKLY build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems". authorized view prevent accessing data for analysis.

Bottom of Form

Top of Form

Question 14:

**Skipped**

You are designing a data processing pipeline. The pipeline must be able to scale automatically as load increases. Messages must be processed at least once and must be ordered within windows of 1 hour. How should you design the solution?

* 

**B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.**

* 

**A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.**

* 

**D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.**

**(Correct)**

* 

**C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.**

**Explanation**

Indeed the correct answer is Option D. Again, not sure why Exam topic answer is deliberately chosen for a wrong answer, for such simple question.

Bottom of Form

Top of Form

Question 15:

**Skipped**

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis.  
Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)

* 

**B. Preserve the structure of the data as much as possible.**

* 

**C. Use BigQuery UPDATE to further reduce the size of the dataset.**

* 

**D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.**

* 

**E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.**

**(Correct)**

* 

**A. Denormalize the data as must as possible.**

**(Correct)**

**Explanation**

Correct Answer: AE A – Denormalisation helps improve performance. B, C - Not helping to address the problem. D – Append will increase the db size and cost involved for storage and also for large number of records to scan for queries by data science team which is costlier. E - Addresses the problem of maximising the usability of the data science team and the data. They can anayse the data exported to cloud storage instead of reading from bigquery which is expensive and impact performance considerably.

Bottom of Form

Top of Form

Question 16:

**Skipped**

You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this?  
(Choose two.)

* 

**B. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.**

* 

**D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.**

* 

**C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.**

**(Correct)**

* 

**A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.**

**(Correct)**

* 

**E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.**

**Explanation**

AC https://cloud.google.com/datastore/docs/export-import-entities C: To import only a subset of entities or to import data into BigQuery, you must specify an entity filter in your export. B: Not correct since you want to store in a different environment than Datastore. Tho this statment is true: Data exported from one Datastore mode database can be imported into another Datastore mode database, even one in another project. A is correct Billing and pricing for managed exports and imports in Datastore Output files stored in Cloud Storage count towards your Cloud Storage data storage costs. Steps to Export all the entities 1. Go to the Datastore Entities Export page in the Google Cloud Console. 2. Go to the Datastore Export page 2. Set the Namespace field to All Namespaces, and set the Kind field to All Kinds. 3. Below Destination, enter the name of your "Cloud Storage bucket". 4. Click Export.

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'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?**

* 

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

* 

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

* 

**E. Cloud Dataflow, Cloud SQL, and Cloud Storage**

* 

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

**(Correct)**

* 

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

**Explanation**

Answer C: Look the 3 requirement in the question "ingest data from a variety of global sources, process and query in real-time, and store the data reliably" Ingest data from global sources: Pub-Sub Process and Query in realtime: Cloud SQL Store reliably: Cloud storage I can understand Databflow is required in case you need to analyze and transform data but question does not refer it.

Bottom of Form

Top of Form

Question 18:

**Skipped**

You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud  
Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

* 

**A. An alert based on a decrease of subscription/num\_undelivered\_messages for the source and a rate of change increase of instance/storage/ used\_bytes for the destination**

* 

**D. An alert based on an increase of instance/storage/used\_bytes for the source and a rate of change decrease of subscription/ num\_undelivered\_messages for the destination**

* 

**B. An alert based on an increase of subscription/num\_undelivered\_messages for the source and a rate of change decrease of instance/storage/ used\_bytes for the destination**

**(Correct)**

* 

**C. An alert based on a decrease of instance/storage/used\_bytes for the source and a rate of change increase of subscription/ num\_undelivered\_messages for the destination**

**Explanation**

The answer is B. subscription/num\_undelivered\_messages: the number of messages that subscribers haven't processed https://cloud.google.com/pubsub/docs/monitoring#monitoring\_forwarded\_undeliverable\_messages

Bottom of Form

Top of Form

Question 19:

**Skipped**

You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally.  
Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your  
Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

* 

**C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.**

**(Correct)**

* 

**D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.**

* 

**A. Edge TPUs as sensor devices for storing and transmitting the messages.**

* 

**B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.**

**Explanation**

C is correct: the main trick come from A, and response is that TPU only use when we have a deployed machine learning model that we don't have now.

Bottom of Form

Top of Form

Question 20:

**Skipped**

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

* 

**C. Create a feature cross of latitude and longitude, bucketize at the minute level and use L1 regularization during optimization.**

**(Correct)**

* 

**B. Create a numeric column from a feature cross of latitude and longitude.**

* 

**A. Provide latitude and longitude as input vectors to your neural net.**

* 

**D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.**

**Explanation**

Answer C seem to be the correct one, rather than it's not mentionned that the model overfits. Source : https://developers.google.com/machine-learning/crash-course/regularization-for-sparsity/l1-regularization

Bottom of Form

Top of Form

Question 21:

**Skipped**

You are designing a cloud-native historical data processing system to meet the following conditions:  
✑ The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute  
Engine.  
✑ A streaming data pipeline stores new data daily.  
✑ Peformance is not a factor in the solution.  
✑ The solution design should maximize availability.  
How should you design data storage for this solution?

* 

**B. Store the data in BigQuery. Access the data using the BigQuery Connector on Cloud Dataproc and Compute Engine.**

* 

**D. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.**

**(Correct)**

* 

**A. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.**

* 

**C. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.**

**Explanation**

Answer: D Description: Multi-region increases high availability and pdf can be stored in gcs

Bottom of Form

Top of Form

Question 22:

**Skipped**

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?

* 

**A. cron**

* 

**D. Workflow Templates on Cloud Dataproc**

* 

**C. Cloud Scheduler**

* 

**B. Cloud Composer**

**(Correct)**

**Explanation**

Should be: B Cloud Composer is an Apache Airflow managed service, it serves well when orchestrating interdependent pipelines, and Cloud Scheduler is just a managed Cron service. Link: https://stackoverflow.com/questions/59841146/cloud-composer-vs-cloud-scheduler

Bottom of Form

Top of Form

Question 23:

**Skipped**

You need to set access to BigQuery for different departments within your company. Your solution should comply with the following requirements:  
✑ Each department should have access only to their data.  
Each department will have one or more leads who need to be able to create and update tables and provide them to their team.

****

**✑ Each department has data analysts who need to be able to query but not modify data.  
How should you set access to the data in BigQuery?**

* 

**B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.**

**(Correct)**

* 

**D. Create a table for each department. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.**

* 

**C. Create a table for each department. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.**

* 

**A. Create a dataset for each department. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.**

**Explanation**

Correct answer is B D not correct because it is granting project level editor/viewer access which is against the principle of least privilege and not the best practice.

Bottom of Form

Top of Form

Question 24:

**Skipped**

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?

* 

**B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.**

**(Correct)**

* 

**C. Export the records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.**

* 

**D. Export the records from the database as an Avro file. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.**

* 

**A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.**

**Explanation**

You are transferring sensitive patient information, so C & D are ruled out. Choice comes down to A & B. Here it gets tricky. How to choose Transfer Appliance: (https://cloud.google.com/transfer-appliance/docs/2.0/overview) Without knowing the bandwidth, it is not possible to determine whether the upload can be completed within 7 days, as recommended by Google. So the safest and most performant way is to use Transfer Appliance. Therefore my choice is B.

Bottom of Form

Top of Form

Question 25:

**Skipped**

You have developed three data processing jobs. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to  
BigQuery. The second ingests data from on-premises servers and uploads it to Cloud Storage. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage. You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

* 

**B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.**

* 

**D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.**

**(Correct)**

* 

**C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.**

* 

**A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.**

**Explanation**

Three workflows are independent. They need to be executed & monitored manually. Cloud composer will add no value. We dont need to connect three workflows in a single workflow. Option D is a simple and efficient solution. Not all solutions require use of Google managed services. Answer is "D"

Bottom of Form

Top of Form

Question 26:

**Skipped**

You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?

* 

**B. Have the data acquisition devices publish data to Cloud Pub/Sub.**

* 

**C. Establish a Cloud Interconnect between all remote data centers and Google.**

**(Correct)**

* 

**A. Deploy small Kafka clusters in your data centers to buffer events.**

* 

**D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.**

**Explanation**

C. This is a tricky one. The issue here is the unreliable connection between data collection and data processing infrastructure, and to resolve it in a cost-effective manner. However, it also mentions that the company is using leased lines. I think replacing the leased lines with Cloud InterConnect would solve the problem, and hopefully not be an added expense. https://cloud.google.com/interconnect/docs/concepts/overview

Bottom of Form

Top of Form

Question 27:

**Skipped**

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

* 

**B. Streaming job, PubSubIO, JdbcIO, side-outputs**

* 

**A. Batch job, PubSubIO, side-inputs**

* 

**D. Streaming job, PubSubIO, BigQueryIO, side-outputs**

* 

**C. Streaming job, PubSubIO, BigQueryIO, side-inputs**

**(Correct)**

**Explanation**

Answer is C, You need pubsubIO and BigQueryIO for streaming data and writing enriched data back to BigQuery. side-inputs are a way to enrich the data https://cloud.google.com/architecture/e-commerce/patterns/slow-updating-side-inputs

Bottom of Form

Top of Form

Question 28:

**Skipped**

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

* 

**D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.**

* 

**C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.**

* 

**B. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.**

* 

**A. Increase the size of your parquet files to ensure them to be 1 GB minimum.**

**(Correct)**

**Explanation**

Should be A: https://stackoverflow.com/questions/42918663/is-it-better-to-have-one-large-parquet-file-or-lots-of-smaller-parquet-files https://www.dremio.com/tuning-parquet/ C & D will improve performance but need to pay more $$

Bottom of Form

Top of Form

Question 29:

**Skipped**

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

* 

**C. Use the INSERT statement to insert a batch of data every 60 seconds.**

* 

**A. Use bq load to load a batch of sensor data every 60 seconds.**

* 

**D. Use the MERGE statement to apply updates in batch every 60 seconds.**

* 

**B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.**

**(Correct)**

**Explanation**

Option [B] is correct. 'C' is incorrect because there's a different method -'tabledata.insertAll' for streaming inserts

Bottom of Form

Top of Form

Question 30:

**Skipped**

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. Which solution should you choose?

* 

**A. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.**

**(Correct)**

* 

**B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.**

* 

**C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.**

* 

**D. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.**

**Explanation**

A: By creating an authorized view one assures that the data is current and avoids taking more storage space (and cost) in order to share a dataset. B and D are not cost optimal and C does not guarantee that the data is kept updated

Bottom of Form

Top of Form

Question 31:

**Skipped**

You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?

* 

**A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.**

**(Correct)**

* 

**B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.**

* 

**C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.**

* 

**D. Use Cloud Dataflow to write summary of each day's stock trades to an Avro file on Cloud Storage. Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.**

**Explanation**

Option A. Below document explains Having EXCHANGE and SYMBOL in the leading positions in the row key will naturally distribute activity. https://cloud.google.com/bigtable/docs/schema-design-time-series

Bottom of Form

Top of Form

Question 32:

**Skipped**

You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

* 

**C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.**

* 

**B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.**

* 

**D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.**

**(Correct)**

* 

**A. Create a cron schedule in Cloud Dataprep.**

**Explanation**

Answer: D Description: Dataprep can be run on Dataflow using template and cloud composer will create dependency on previous job

Bottom of Form

Top of Form

Question 33:

**Skipped**

You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the  
Data Science team runs a query filtered on a date column and limited to 30""90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries.  
What should you do?

* 

**A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type.**

**(Correct)**

* 

**C. Modify your pipeline to maintain the last 30""90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.**

* 

**B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.**

* 

**D. Write an Apache Beam pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.**

**Explanation**

The D solution is obviously discarded. The request NOT require ONLY LAST 30-90 days, so the C solution is not the right solution. In addition to this, the request ask to keep the possibility to made queries, so B is wrost. Is not mandatory make the queries while you make the modify so the right answer is A

Bottom of Form

Top of Form

Question 34:

**Skipped**

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

* 

**D. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector. Use a Dataflow job to read from PubSub and write to GCS.**

* 

**B. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.**

* 

**C. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connector. Use a Dataflow job to read from PubSub and write to GCS.**

* 

**A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.**

**(Correct)**

**Explanation**

A. https://cwiki.apache.org/confluence/pages/viewpage.action?pageId=27846330 The solution specifically mentions mirroring and minimizing the use of Kafka Connect plugin. D would be the more Google Cloud-native way of implementing the same, but the requirement is better met by A.

Bottom of Form

Top of Form

Question 35:

**Skipped**

You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?

* 

**A. Set the BigQuery dataset to be regional. In the event of an emergency, use a point-in-time snapshot to recover the data.**

* 

**C. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.**

**(Correct)**

* 

**B. Set the BigQuery dataset to be regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.**

* 

**D. Set the BigQuery dataset to be multi-regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.**

**Explanation**

Answer: C Description: In multiregional, data is not lost and recovery time is ms. Regional, zone failure results in data loss

Bottom of Form

Top of Form

Question 36:

**Skipped**

You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.  
You have the following requirements:  
✑ You will batch-load the posts once per day and run them through the Cloud Natural Language API.  
✑ You will extract topics and sentiment from the posts.  
✑ You must store the raw posts for archiving and reprocessing.  
✑ You will create dashboards to be shared with people both inside and outside your organization.  
You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?

* 

**B. Store the social media posts and the data extracted from the API in Cloud SQL.**

* 

**D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.**

* 

**C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.**

**(Correct)**

* 

**A. Store the social media posts and the data extracted from the API in BigQuery.**

**Explanation**

Answer should be C, because they ask you to save a copy of the raw posts for archival, which may not be possible if you directly feed the posts to the API.

Bottom of Form

Top of Form

Question 37:

**Skipped**

You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? (Choose two.)

* 

**E. The subscriber code does not acknowledge the messages that it pulls.**

**(Correct)**

* 

**B. Total outstanding messages exceed the 10-MB maximum.**

* 

**D. The subscriber code cannot keep up with the messages.**

* 

**C. Error handling in the subscriber code is not handling run-time errors properly.**

**(Correct)**

* 

**A. Publisher throughput quota is too small.**

**Explanation**

Answer: C, E Description: C, E: By not acknowleding the pulled message, this result in it be putted back in Cloud Pub/Sub, meaning the messages accumulate instead of being consumed and removed from Pub/Sub. The same thing can happen ig the subscriber maintains the lease on the message it receives in case of an error. This reduces the overall rate of processing because messages get stuck on the first subscriber. Also, errors in Cloud Function do not show up in Stackdriver Log Viewer if they are not correctly handled. A: No problem with publisher rate as the observed result is a higher number of messages and not a lower number. B: if messages exceed the 10MB maximum, they cannot be published. D: Cloud Functions automatically scales so they should be able to keep up.

Bottom of Form

Top of Form

Question 38:

**Skipped**

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

* 

**B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.**

* 

**C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.**

* 

**D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.**

**(Correct)**

* 

**A. Increase the cluster size with more non-preemptible workers.**

**Explanation**

Answer: D Description: Graceful decommissioning will ensure that the data is processed by worker before it is removed by Yarn

Bottom of Form

Top of Form

Question 39:

**Skipped**

You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.  
Which service do you select for storing and serving your data?

* 

**B. Cloud Bigtable**

* 

**D. Cloud SQL**

**(Correct)**

* 

**A. Cloud Spanner**

* 

**C. Cloud Firestore**

**Explanation**

D Cloud SQL supports MySQL 5.6 or 5.7, and provides up to 624 GB of RAM and 30 TB of data storage, with the option to automatically increase the storage size as needed.

Bottom of Form

Top of Form

Question 40:

**Skipped**

Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

* 

**A. Cloud Dataflow**

* 

**D. Cloud Dataproc**

* 

**C. Cloud Dataprep**

* 

**B. Cloud Composer**

**(Correct)**

**Explanation**

B: Hybrid and multi-cloud Ease your transition to the cloud or maintain a hybrid data environment by orchestrating workflows that cross between on-premises and the public cloud. Create workflows that connect data, processing, and services across clouds to give you a unified data environment. https://cloud.google.com/composer#section-2

Bottom of Form