Top of Form

Question 1:

**Skipped**

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

* 

**C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created**

* 

**D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created**

**(Correct)**

* 

**B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail**

* 

**A. They have not assigned the timestamp, which causes the job to fail**

**Explanation**

Answer: D Description: Caution: Beam’s default windowing behavior is to assign all elements of a PCollection to a single, global window and discard late data, even for unbounded PCollections. Before you use a grouping transform such as GroupByKey on an unbounded PCollection, you must do at least one of the following: —->>>>>>Set a non-global windowing function. See Setting your PCollection’s windowing function. Set a non-default trigger. This allows the global window to emit results under other conditions, since the default windowing behavior (waiting for all data to arrive) will never occur. —->>>>If you don’t set a non-global windowing function or a non-default trigger for your unbounded PCollection and subsequently use a grouping transform such as GroupByKey or Combine, your pipeline will generate an error upon construction and your job will fail. So it looks like D

Bottom of Form

Top of Form

Question 2:

**Skipped**

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

* 

**B. HiveQL using Hive**

* 

**D. Python using MapReduce**

* 

**C. Java using MapReduce**

* 

**A. PigLatin using Pig**

**(Correct)**

**Explanation**

Answer: A Description: Pig is scripting language which can be used for checkpointing and splitting pipelines

Bottom of Form

Top of Form

Question 3:

**Skipped**

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

* 

**C. Increase the size of the Hadoop cluster.**

* 

**B. Rewrite the job in Apache Spark.**

**(Correct)**

* 

**D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.**

* 

**A. Rewrite the job in Pig.**

**Explanation**

B: The objective is to not increase the cost at the sametime do the analyitics required. Mapreduce jobs are not efficient and fast as spark so it will avoid failing the jobs.

Bottom of Form

Top of Form

Question 4:

**Skipped**

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

* 

**C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset**

**(Correct)**

* 

**A. Create groups for your users and give those groups access to the dataset**

* 

**D. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset**

* 

**B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request**

**Explanation**

C: It says "do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset", then C is the best choice.

Bottom of Form

Top of Form

Question 5:

**Skipped**

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

* 

**C. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.**

* 

**B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.**

**(Correct)**

* 

**D. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.**

* 

**A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.**

**Explanation**

Answer: B Description: KMS stored on cloud helps in creating, rotating and destroying keys as needed and also it can be used to encrypt compute engines

Bottom of Form

Top of Form

Question 6:

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

* 

**B. Create a partitioned table called tracking\_table and include a TIMESTAMP column.**

**(Correct)**

* 

**C. Create sharded tables for each day following the pattern tracking\_table\_YYYYMMDD.**

* 

**A. Create a table called tracking\_table and include a DATE column.**

* 

**D. Create a table called tracking\_table with a TIMESTAMP column to represent the day.**

**Explanation**

https://cloud.google.com/bigquery/docs/partitioned-tables#dt\_partition\_shard - Supports B

Bottom of Form

Top of Form

Question 7:

**Skipped**

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

* 

**C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.**

**(Correct)**

* 

**A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.**

* 

**B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.**

* 

**D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.**

**Explanation**

Answer: C Description: Spanner allows transaction tables to scale horizontally and secondary indexes for range queries

Bottom of Form

Top of Form

Question 8:

**Skipped**

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

* 

**D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.**

* 

**A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.**

* 

**C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.**

* 

**B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.**

**(Correct)**

**Explanation**

B. My reasoning is that the phrase "Give the bid for each item to the user in the bid event that is processed first." in D is invalid. A bid should be given to a user who made it first based on the value of timestamp.

Bottom of Form

Top of Form

Question 9:

**Skipped**

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

* 

**A. BigQuery**

* 

**C. Cloud BigTable**

* 

**D. Cloud Datastore**

* 

**B. Cloud SQL**

**(Correct)**

**Explanation**

B - Cloud SQL should be the only correct answer. Required solution needs to support transactions as well as analysis through a BI tool. Firestore/Datastore does not support SQL syntax typically needed to do analysis done by a BI tool. BigQuery is not suitable for transactional use case. BigTable does not support SQL.

Bottom of Form

Top of Form

Question 10:

**Skipped**

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

* 

**D. Create a service account for the ODBC connection to use for authentication**

**(Correct)**

* 

**E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"**

* 

**B. Create a new partitioned table using a standard SQL query**

* 

**A. Create a new view over events using standard SQL**

* 

**C. Create a new view over events\_partitioned using standard SQL**

**(Correct)**

**Explanation**

Answer: C, D Description: Cannot create standard sql view on top of legacy sql, service account to connect cloud

Bottom of Form

Top of Form

Question 11:

**Skipped**

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

* 

**C. Use a higher-memory node so that the job runs faster**

* 

**A. Migrate the workload to Google Cloud Dataflow**

* 

**D. Use SSDs on the worker nodes so that the job can run faster**

* 

**B. Use pre-emptible virtual machines (VMs) for the cluster**

**(Correct)**

**Explanation**

Answer should be B. The workload in itself is already completing in very less time i.e. in 30 minutes so no optimization is required to reduce the same. However, we have to make sure that our cluster runs only for 30 mins in a week and hence we choose pre-emtible VMs.

Bottom of Form

Top of Form

Question 12:

**Skipped**

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

* 

**C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore**

* 

**B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery**

**(Correct)**

* 

**A. Load the data every 30 minutes into a new partitioned table in BigQuery.**

* 

**D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.**

**Explanation**

B is correct because regional storage is cheaper than BigQuery storage. A is not correct because it is not the cheapest way of accomplishing this task. C, D are not correct because it is not the least expensive option. Using Dataflow to query BigQuery adds unnecessary cost to the deployment and will cost more than using BigQuery natively.

Bottom of Form

Top of Form

Question 13:

**Skipped**

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

* 

**B. Create and share a new dataset and view that provides the aggregate results.**

* 

**D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.**

* 

**A. Create and share an authorized view that provides the aggregate results.**

**(Correct)**

* 

**C. Create and share a new dataset and table that contains the aggregate results.**

**Explanation**

Answer: A Description: Authorized view is used to protect the underlying data, authorized views are created in different dataset with restricted access

Bottom of Form

Top of Form

Question 14:

**Skipped**

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

* 

**C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.**

**(Correct)**

* 

**B. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.**

* 

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

* 

**D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.**

**Explanation**

answer C: BigQuery can access data in external sources, known as federated sources. Instead of first loading data into BigQuery, you can create a reference to an external source. External sources can be Cloud Bigtable, Cloud Storage, and Google Drive. When accessing external data, you can create either permanent or temporary external tables. Permanent tables are those that are created in a dataset and linked to an external source. Dataset-level access controls can be applied to these tables. When you are using a temporary table, a table is created in a special dataset and will be available for approxi- mately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse. "Dan Sullivan" Book

Bottom of Form

Top of Form

Question 15:

**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 visualization for operations teams with the following requirements:  
✑ Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)  
✑ The report must not be more than 3 hours delayed from live data.  
✑ The actionable report should only show suboptimal links.  
✑ Most suboptimal links should be sorted to the top.  
✑ Suboptimal links can be grouped and filtered by regional geography.  
✑ User response time to load the report must be <5 seconds.  
You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?

* 

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

* 

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

**(Correct)**

* 

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

* 

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

**Explanation**

Answer D: Data in SQL so querying becomes easier on any pattern. create mutiple charts, graphs to fulfill your requirements.

Bottom of Form

Top of Form

Question 16:

**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 is building a custom interface to share data. They have these requirements:  
1. They need to do aggregations over their petabyte-scale datasets.  
2. They need to scan specific time range rows with a very fast response time (milliseconds).  
Which combination of Google Cloud Platform products should you recommend?

* 

**D. BigQuery and Cloud Storage**

* 

**A. Cloud Datastore and Cloud Bigtable**

* 

**B. Cloud Bigtable and Cloud SQL**

* 

**C. BigQuery and Cloud Bigtable**

**(Correct)**

**Explanation**

C: They need to do aggregations over their petabyte-scale datasets: Bigquery They need to scan specific time range rows with a very fast response time (milliseconds): Bigtable

Bottom of Form

Top of Form

Question 17:

**Skipped**

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

* 

**A. Change the processing job to use Google Cloud Dataproc instead.**

* 

**D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.**

* 

**B. Manually start the Cloud Dataflow job each morning when you get into the office.**

* 

**C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.**

**(Correct)**

**Explanation**

Answer is C. https://cloud.google.com/appengine/docs/flexible/nodejs/scheduling-jobs-with-cron-yaml

Bottom of Form

Top of Form

Question 18:

**Skipped**

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

* 

**C. Use a service account with the ability to read the batch files and to write to BigQuery**

**(Correct)**

* 

**B. Grant the Project Owner role to a service account, and run the job with it**

* 

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

* 

**A. Restrict the Google Cloud Storage bucket so only you can see the files**

**Explanation**

Correct Answer : C Explanation:-This option is correct as the best practice is to use a service account with least privilege. Practices - Service Accounts A service account is a special type of Google account intended to represent a non-human user that needs to authenticate and be authorized to access data in Google APIs. Typically, service accounts are used in scenarios such as - • Running workloads on virtual machines (VMs).

Bottom of Form

Top of Form

Question 19:

**Skipped**

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

* 

**B. The CSV data has invalid rows that were skipped on import.**

* 

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

* 

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

* 

**C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.**

**(Correct)**

**Explanation**

C is correct because this is the only situation that would cause successful import. A is not correct because if another data format other than CSV was selected then the data would not import successfully. B is not correct because the data was fully imported meaning no rows were skipped. D is not correct because whether the data has been previously transformed will not affect whether the source file will match the BigQuery table. https://cloud.google.com/bigquery/docs/loading-data#loading\_encoded\_data

Bottom of Form

Top of Form

Question 20:

**Skipped**

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

* 

**C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to "˜none' using a Cloud Dataprep job.**

* 

**A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to "˜none' using a Cloud Dataproc job.**

* 

**B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.**

**(Correct)**

* 

**D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 0 using a custom script.**

**Explanation**

Answer: B Description: Dataprep is best suited for this kind of activity, as we need to just do a small function of modulating the data and dataprep is UI based

Bottom of Form

Top of Form

Question 21:

**Skipped**

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

* 

**B. Subsample your training dataset.**

**(Correct)**

* 

**D. Increase the number of layers in your neural network.**

* 

**C. Increase the number of input features to your model.**

* 

**A. Subsample your test dataset.**

**Explanation**

B. A is incorrect because it's modifying TEST data. Both C and D will increase, not decrease the training time. One of the easiest questions here in my opinion. I wonder if questions like this will be on the actual exam.

Bottom of Form

Top of Form

Question 22:

**Skipped**

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

* 

**A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.**

* 

**C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.**

**(Correct)**

* 

**D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.**

* 

**B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.**

**Explanation**

Correct Answer: C Explanation:-This option is correct as Dataflow, provides a cost-effective solution to perform transformations on the streaming data, with autoscaling provides scaling without any intervention. System lag with Stackdriver provides monitoring for the streaming data. With autoscaling enabled, the Cloud Dataflow service automatically chooses the appropriate number of worker instances required to run your job.

Bottom of Form

Top of Form

Question 23:

**Skipped**

You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:  
✑ The user profile: What the user likes and doesn't like to eat  
✑ The user account information: Name, address, preferred meal times  
✑ The order information: When orders are made, from where, to whom  
The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

* 

**A. BigQuery**

**(Correct)**

* 

**C. Cloud Bigtable**

* 

**B. Cloud SQL**

* 

**D. Cloud Datastore**

**Explanation**

A: It says you need to store transactional data. You can optimize the schema in BigQuery. Then, as you want to do ML, you are going to do data explor, transf to build the analytical table. That analytical table could be input to a BQML model. Cloud SQL could work, but con GCP I think BigQuery is the right answer.

Bottom of Form

Top of Form

Question 24:

**Skipped**

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

* 

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

**(Correct)**

* 

**C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.**

* 

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

* 

**A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.**

**Explanation**

B. The question is focused on designing storage for very large files, with support for compression, ANSI SQL queries, and parallel loading from the input locations. This can be met using GCS for storage and Bigquery permanent tables with external data source in GCS.

Bottom of Form

Top of Form

Question 25:

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

* 

**E. Rowkey: date#data\_point Column data: device\_id**

* 

**C. Rowkey: device\_id Column data: date, data\_point**

* 

**D. Rowkey: data\_point Column data: device\_id, date**

* 

**B. Rowkey: date Column data: device\_id, data\_point**

* 

**A. Rowkey: date#device\_id Column data: data\_point**

**(Correct)**

**Explanation**

A because, as per guidelines, rowkey should have less granular data first followed by next granular data.. In question, its for a given day, so date is the least granular. Date comes first followed by device id. This followed by the data gives all columns

Bottom of Form

Top of Form

Question 26:

**Skipped**

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

* 

**B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.**

* 

**D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.**

* 

**C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.**

* 

**A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.**

**(Correct)**

**Explanation**

Answer: A Description: My take on this is for sensor calibration you just need to update the transform function, rather than creating a whole new mapreduce job and storing/passing the values to next job

Bottom of Form

Top of Form

Question 27:

**Skipped**

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

* 

**C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.**

**(Correct)**

* 

**E. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.**

* 

**B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.**

* 

**A. Introduce data compression for each file to increase the rate file of file transfer.**

* 

**D. Assemble 1,000 files into a tape archive (TAR) file. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.**

**(Correct)**

**Explanation**

E cannot be: Transfer Service is recommended for 300mbps or faster https://cloud.google.com/storage-transfer/docs/on-prem-overview Bandwidth is not an issue, so B is not an answer Cloud Storage loading gets better throughput the larger the files are. Therefore making them smaller with compression does not seem a solution. -m option to do parallel work is recommended. Therefore A is not and C is an answer. https://medium.com/@duhroach/optimizing-google-cloud-storage-small-file-upload-performance-ad26530201dc That leaves D as the other option. It is true you cannot user tar directly with gsutil, but you can load the tar file to Cloud Storage, move the file to a Compute Engine instance with Linux, use tar to split files and copy them back to Cloud Storage. Batching many files in a larger tar will improve Cloud Storage throughput. So, given the alternatives, I think answer is CD

Bottom of Form

Top of Form

Question 28:

**Skipped**

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

* 

**B. HBase**

**(Correct)**

* 

**D. MongoDB**

**(Correct)**

* 

**C. MySQL**

* 

**A. Redis**

* 

**E. Cassandra**

**(Correct)**

* 

**F. HDFS with Hive**

**Explanation**

Answer is BDE - A. Redis - Redis is an in-memory non-relational key-value store. Redis is a great choice for implementing a highly available in-memory cache to decrease data access latency, increase throughput, and ease the load off your relational or NoSQL database and application. Since the question does not ask cache, A is discarded. B. HBase - Meets reqs C. MySQL - they do not need ACID, so not needed. D. MongoDB - Meets reqs E. Cassandra - Apache Cassandra is an open source NoSQL distributed database trusted by thousands of companies for scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data. F. HDFS with Hive - Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data. HIVE IS NOT A DATABSE.

Bottom of Form

Top of Form

Question 29:

**Skipped**

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

* 

**A. Cloud Bigtable**

**(Correct)**

* 

**D. Google Cloud Datastore**

* 

**C. Google Cloud Storage**

* 

**B. Google BigQuery**

**Explanation**

A: Bigtable is GCP’s managed wide-column database. It is also a good option for migrat- ing on-premises Hadoop HBase databases to a managed database because Bigtable has an HBase interface. .... Cloud Bigtable is a wide-column NoSQL database used for high-volume databases that require low millisecond (ms) latency. Cloud Bigtable is used for IoT, time-series, finance, and similar applications. "Dan Sullivan" Official Google Cloud Certified Professional Data Engineer Study Guide-Sybex book

Bottom of Form

Top of Form

Question 30:

**Skipped**

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

* 

**B. Increase the size of the Google Persistent Disk on your server.**

* 

**A. Increase the CPU size on your server.**

* 

**C. Increase your network bandwidth from your datacenter to GCP.**

**(Correct)**

* 

**D. Increase your network bandwidth from Compute Engine to Cloud Storage.**

**Explanation**

Answer: C Description : Speed of data transfer depends on Bandwidth Few things in computing highlight the hardware limitations of networks as transferring large amounts of data. Typically you can transfer 1 GB in eight seconds over a 1 Gbps network. If you scale that up to a huge dataset (for example, 100 TB), the transfer time is 12 days. Transferring huge datasets can test the limits of your infrastructure and potentially cause problems for your business.

Bottom of Form

Top of Form

Question 31:

**Skipped**

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

* 

**B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.**

* 

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

* 

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

**(Correct)**

* 

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

**Explanation**

Correct Answer: A Destination is GCS and having multi-regional so A is the best option available. Even since BigQuery Data Transfer Service initially supports Google application sources like Google Ads, Campaign Manager, Google Ad Manager and YouTube but it does not support destination anything other than bq data set

Bottom of Form

Top of Form

Question 32:

**Skipped**

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

* 

**A. Set a single global window to capture all the data.**

* 

**C. Use watermarks and timestamps to capture the lagged data.**

**(Correct)**

* 

**B. Set sliding windows to capture all the lagged data.**

* 

**D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.**

**Explanation**

Answer: C Description: A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

Bottom of Form

Top of Form

Question 33:

**Skipped**

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

* 

**A. Use the TABLE\_DATE\_RANGE function**

**(Correct)**

* 

**B. Use the WHERE\_PARTITIONTIME pseudo column**

* 

**C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD**

* 

**D. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD**

**Explanation**

Answer: A Description: Legacy sql uses table date range whereas standard sql uses table\_sufix for wildcard

Bottom of Form

Top of Form

Question 34:

**Skipped**

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

* 

**B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.**

* 

**D. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.**

**(Correct)**

* 

**C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.**

* 

**A. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.**

**Explanation**

Keywords here are 1. "Archived": Immutable and hence, BQ and Cloud SQL are ruled out 2. "Auditable": Means track any changes done. Only D can provide the audibility piece! I will go with D

Bottom of Form

Top of Form

Question 35:

**Skipped**

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

* 

**B. Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.**

**(Correct)**

* 

**D. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.**

* 

**C. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.**

* 

**A. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.**

**Explanation**

Correct: B BigQuery has no quota on the DML statements. (Search Google - does bigquery have quota for update). Why not C: This is a one time activity and SQL is the easiest way to program it. DataFlow is way overkill for this. You will need to find an engineer who can develop DataFlow pipelines. Whereas, SQL is so much more widely known and easier. One of the great features about BigQuery is its SQL interface. Even for BigQueryML services.

Bottom of Form

Top of Form

Question 36:

**Skipped**

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

* 

**B. Reduce the number of training examples**

* 

**A. Get more training examples**

**(Correct)**

* 

**D. Use a larger set of features**

* 

**C. Use a smaller set of features**

**(Correct)**

* 

**E. Increase the regularization parameters**

**(Correct)**

* 

**F. Decrease the regularization parameters**

**Explanation**

ACE: Reason: Overfitting results in noisy model and the predictions wont be with good accuracy. Feed more apt training samples for a better model. More number of features implies more dimensionality which would also result in poor performance.

Bottom of Form

Top of Form

Question 37:

**Skipped**

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

* 

**B. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.**

* 

**D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.**

* 

**C. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.**

**(Correct)**

* 

**A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc. Call the model from your application.**

**Explanation**

Answer: C A & B - Need to build your own model, so discarded as options C D can do the job here using Cloud Video Intelligence API. BigTable is better option. So C is correct

Bottom of Form

Top of Form

Question 38:

**Skipped**

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

* 

**A. Convert all daily log tables into date-partitioned tables**

* 

**B. Convert the sharded tables into a single partitioned table**

**(Correct)**

* 

**D. Create separate views to cover each month, and query from these views**

* 

**C. Enable query caching so you can cache data from previous months**

**Explanation**

Answer: B Description: Google says that when you have multiple wildcard tables, best option is to shard it into single partitioned table. Time and cost efficient

Bottom of Form

Top of Form

Question 39:

**Skipped**

You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:  
SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country  
You check the query plan for the query and see the following output in the Read section of Stage:1:

****

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

* 

**C. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values**

* 

**D. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew**

* 

**A. Users are running too many concurrent queries in the system**

**(Correct)**

* 

**B. The [myproject:mydataset.mytable] table has too many partitions**

**Explanation**

If you read this https://medium.com/slalom-build/using-bigquery-execution-plans-to-improve-query-performance-af141b0cc33d C can't be right because the skewness happen when the column you use for grouping contains lots of null values, here C mentions columns that aren't part of the grouping clause. D, that's not how data get skewed, it gets skewed due to null values. A is the only answer here.

Bottom of Form

Top of Form

Question 40:

**Skipped**

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

* 

**A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.**

**(Correct)**

* 

**B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.**

* 

**D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.**

* 

**C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.**

**Explanation**

Correct Answer : A Entity analysis -> Identify entities within documents receipts, invoices, and contracts and label them by types such as date, person, contact information, organization, location, events, products, and media. Sentiment analysis -> Understand the overall opinion, feeling, or attitude sentiment expressed in a block of text. -- Avoid Custom models

Bottom of Form