CertyIQ

CertyIQ

Premium exam material

Get certification quickly with the CertyIQ Premium exam material.
Everything you need to prepare, learn & pass your certification exam easily. Lifetime free updates
First attempt guaranteed success.

https://www.CertyIQ.com

About CertyIQ

We here at CertyIQ eventually got enough of the industry's greedy exam paid for. Our team of IT professionals comes with years of experience in the IT industry Prior to training CertIQ we worked in test areas where we observed the horrors of the paywall exam preparation system.

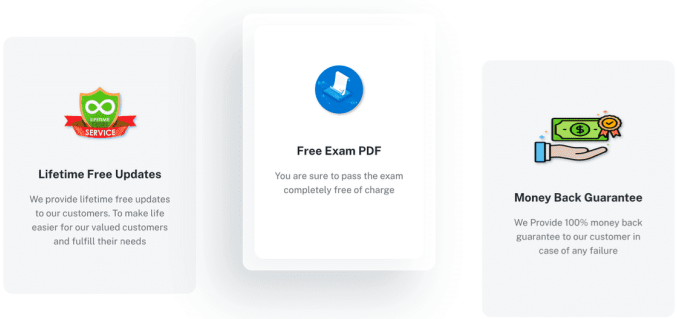
The misuse of the preparation system has left our team disillusioned. And for that reason, we decided it was time to make a difference. We had to make In this way, CertyIQ was created to provide quality materials without stealing from everyday people who are trying to make a living.

Doubt Support

We have developed a very scalable solution using which we are able to solve 400+ doubts every single day with an average rating of 4.8 out of 5.

https://www.certyiq.com

Mail us on - certyiqofficial@gmail.com



John

October 19, 2022

★★★★★

Thanks you so much for your help. I scored 972 in my exam today. More than 90% were from your PDFs!

Dana

September 04, 2022

★★★★★

Thanks a lot for this updated AZ-900 Q&A. I just passed my exam and got 974, I followed both of your Az-900 videos and the 6 PDF, the PDFs are very much valid, all answers are correct. Could you please create a similar video/PDF for DP900, your content/PDF's is really awesome. The team did a really good job. Thank You 😊.

Ahamed Shibly

2 months ago

★★★★★

Customer support is realy fast and helpful, I just finished my exam and this video along with the 6 PDF helped me pass! Definitely recommend getting the PDFs. Thank you!

October 22, 2022

★★★★★

Passed my exam today with 891 marks. Out of 52 questions, 51 were from certyiq PDFs including Contoso case study.
Thank You certyiq team!

Henry Rome

2 months ago

★★★★★

These questions are real and 100 % valid. Thank you so much for your efforts, also your 4 PDFs are awesome, I passed the DP900 exam on 1 Sept. With 968 marks. Thanks a lot, buddy!

Esmaria

2 months ago

★★★★★

Simple easy to understand explanations. To anyone out there wanting to write AZ900, I highly recommend 6 PDF's.Thank you so much, appreciate all your hard work in having such great content. Passed my exam Today - 3 September with 942 score.

Google

(Professional Data Engineer)

Professional Data Engineer on Google Cloud Platform

Total: **283 Questions**

Link: https://certyiq.com/papers?provider=google&exam=professional-data-engineer

Question: 1 CertyIQ

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

A. Threading
B. Serialization
C. Dropout Methods
D. Dimensionality Reduction

Answer: C

Explanation:

Bad performance of a model is either due to lack of relationship between dependent and independent variables used, or just overfit due to having used too many features and/or bad features.

A: Threading parallelisation can reduce training time, but if the selected featuers are the same then the resulting performance won't have changed

B: Serialization is only changing data into byte streams. This won't be useful.

C: This can show which features are bad. E.g. if it is one feature causing bad performance, then the dropout method will show it, so you can remove it from the model and retrain it.

D: This would become clear if the model did not fit the training data well. But the question says that the model fits the training data well, so D is not the answer.

Reference:

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

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

Question: 2 CertyIQ

You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

A. Continuously retrain the model on just the new data.
B. Continuously retrain the model on a combination of existing data and the new data.
C. Train on the existing data while using the new data as your test set.

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

Answer: B

Explanation:

As new data can be with new features. Hence the new data can be split to both training and test data to retrain as well as with existing data. because we have to use a combination of old and new test data as well as training data.

Question: 3 CertyIQ

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

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

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

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

D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Answer: C

Explanation:

Based on Google documentation, self-join is an anti-pattern because this option provides the least amount of inconvenience over using pre-specified date ranges or one table per clinic while also increasing performance due to avoiding self-joins.

Reference:

https://cloud.google.com/bigquery/docs/best-practices-performance-patterns

Question: 4 CertyIQ

You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old. What should you do?

A. Disable caching by editing the report settings.

B. Disable caching in BigQuery by editing table details.

C. Refresh your browser tab showing the visualizations.

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

Answer: A

Explanation:

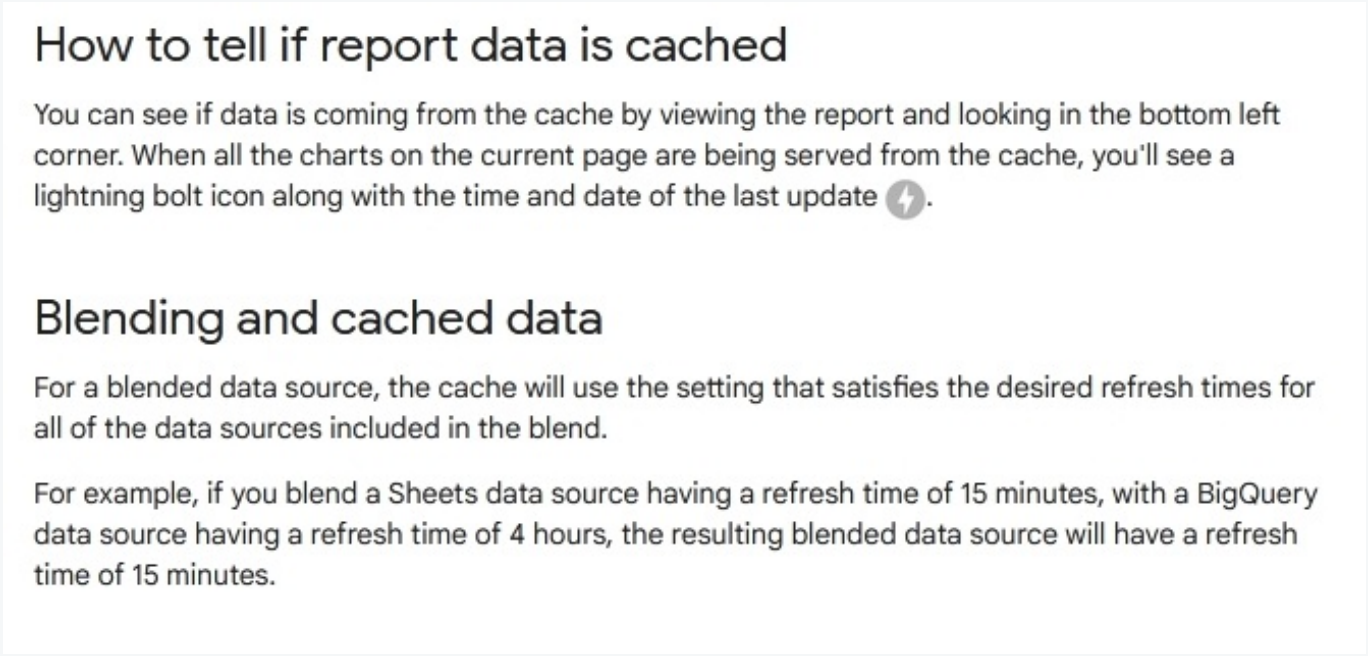
Disable caching by editing the report settings.

A cache is a temporary data storage system. Fetching cached data can be much faster than fetching it directly from the underlying data set, and helps reduce the number of queries sent, minimizing costs for paid data access.

Reference:

https://support.google.com/datastudio/answer/7020039?hl=en#zippy=%2Cin-this-article

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



How to tell if report data is cached

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

Blending and cached data

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

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

Question: 5 CertyIQ

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

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

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

C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.

D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

Answer: D

Explanation:

Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.

By running a Cloud Dataflow pipeline to import the data, you can perform data validation, cleaning and transformation before it gets loaded into BigQuery. Dataflow allows you to handle corrupted or incorrectly formatted rows by pushing them to another dead-letter table for analysis. This way, you can ensure that only clean and correctly formatted data is loaded into BigQuery for analysis.

Question: 6 CertyIQ

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

A. Issue a command to restart the database servers.

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

C. Retry the query every second until it comes back online to minimize staleness of data.

D. Reduce the query frequency to once every hour until the database comes back online.

Answer: B

Explanation:

App engine create applications that use Cloud SQL database connections effectively. Below is what is written in google cloud documentation.

If your application attempts to connect to the database and does not succeed, the database could be temporarily unavailable. In this case, sending too many simultaneous connection requests might waste additional database resources and increase the time needed to recover. Using exponential backoff prevents your application from sending an unresponsive number of connection requests when it can't connect to the database.

This retry only makes sense when first connecting, or when first grabbing a connection from the pool. If errors happen in the middle of a transaction, the application must do the retrying, and it must retry from the beginning of a transaction. So even if your pool is configured properly, the application might still see errors if connections are lost.

Reference:

https://cloud.google.com/sql/docs/mysql/manage-connections

Question: 7 CertyIQ

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

A. Linear regression

B. Logistic classification

C. Recurrent neural network

D. Feedforward neural network

Answer: A

Explanation:

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

Question: 8 CertyIQ

You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?

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

B. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.

C. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.

D. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.

Answer: D

Explanation:

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

Question: 9 CertyIQ

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

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

Which table name will make the SQL statement work correctly?

A. 'bigquery-public-data.noaa_gsod.gsod'

B. bigquery-public-data.noaa_gsod.gsod*

C. 'bigquery-public-data.noaa_gsod.gsod'*

D. 'bigquery-public-data.noaa_gsod.gsod*`

Answer: D

Explanation:

Reference:
https://cloud.google.com/bigquery/docs/wildcard-tables
" target="_blank" style="word-break: break-all;">

Filtering selected tables using _TABLE_SUFFIX

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

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

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

Question: 10 CertyIQ

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

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

Answer: BDF

Explanation:

bigquery.tables.create Create new tables.

bigquery.tables.delete Delete tables.

bigquery.tables.export Export table data out of BigQuery.

bigquery.tables.get Get table metadata.

To get table data, you need bigquery.tables.getData.

bigquery.tables.getData Get table data. This permission is required for querying table data.

To get table metadata, you need bigquery.tables.get.

bigquery.tables.list List tables and metadata on tables.

bigquery.tables.setCategory Set policy tags in table schema.

bigquery.tables.update

Update table metadata.

To update table data, you need bigquery.tables.updateData.

bigquery.tables.updateData

Update table data.

To update table metadata, you need bigquery.tables.update.

Question: 11 CertyIQ

You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:
- No interaction by the user on the site for 1 hour
Has added more than $30 worth of products to the basket
- Has not completed a transaction
You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?

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

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

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

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

Answer: C

Explanation:

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

Question: 12 CertyIQ

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data.
Which three steps should you take? (Choose three.)

A. Load data into different partitions.

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

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

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

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

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

Answer: BDF

Explanation:

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

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

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

By loading each client's data into a separate dataset, you ensure that each client's data is isolated from the data of other clients. Restricting access to each client's dataset to only approved users, as specified in D,

further enhances data security by ensuring that only authorized users can access the data. By using appropriate IAM roles for each client's users, as specified in F, you can grant different levels of access to different clients and their users, ensuring that each client has only the level of access required for their specific needs.

Question: 13 CertyIQ

You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.
Which Google database service should you use?

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

Answer: D

Explanation:

As user base grows, write transaction grows since we are dealing with POS (that not the place for reading but writing). In order to accommodate more writes in transactional flavor which can be horizontally scalable DATASTORE should be preferred.

Question: 14 CertyIQ

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

A. There are very few occurrences of mutations relative to normal samples.
B. There are roughly equal occurrences of both normal and mutated samples in the database.
C. You expect future mutations to have different features from the mutated samples in the database.
D. You expect future mutations to have similar features to the mutated samples in the database.
E. You already have labels for which samples are mutated and which are normal in the database.

Answer: AC

Explanation:

Anomaly detection unsupervised learning

The objective of Unsupervised Anomaly Detection is to detect previously unseen rare objects or events without any prior knowledge about these. The only information available is that the percentage of anomalies in the dataset is small, usually less than 1%.

Reference:

https://paperswithcode.com/task/unsupervised-anomaly-detection#:~:text=The%20objective%20of%20Unsupervised%20Anomaly,%2C%20usually%20less%20than%201%

Question: 15 CertyIQ

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

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

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

C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.

D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.

Answer: D

Explanation:

The data is first comes to buffer and then written to Storage. If we are running queries in buffer we will face above mentioned issues. If we wait for the bigquery to write the data to storage then we won't face the issue. So We need to wait till it's written tio storage

Question: 16 CertyIQ

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

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

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

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

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

Answer: A

Explanation:

Description: First we need to know who is accessing what then we can create suitable policies. Stackdriver is used to track access logs for Bigquery,

Question: 17 CertyIQ

Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible. They also want to be able to persist data beyond the life of the cluster. What should you do?

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

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

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

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

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

Answer: D

Explanation:

Dataproc is used to migrate Hadoop and Spark jobs on GCP. Dataproc with GCS connected through Google Cloud Storage connector helps store data after the life of the cluster. When the job is high I/O intensive, then we need to create a small persistent disk.

Question: 18 CertyIQ

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

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

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

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

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

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

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

Answer: BCD

Explanation:

B - Not labelled as Fraud or not. So Unsupervised.

C - Clustering can be done based on location, amount etc.

D - Location is already given. So labelled. Hence supervised.

Fraud is not a feature, so unsupervised, location is given so supervised, Clustering can be done looking at the done with same features

BCD makes more sense to me. Its for sure not unsupervised, since locations are in the data already. Reinforcement also doesn't fit, as there no AI and no interactions with data from the observer.

Question: 19 CertyIQ

Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for- like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

A. Put the data into Google Cloud Storage.

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

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

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

Answer: A

Explanation:

First rule of dataproc is to keep data in GCS

Question: 20 CertyIQ

You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

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

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

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

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

Answer: D

Explanation:

The custom endpoint is not acknowledging the message, that is the reason for Pub/Sub to send the message again and again. Not acknowledging a message makes Pub/Sub to think it has not been received, so it sends duplicate messages.

Question: 21 CertyIQ

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

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

B. Compute the hash value of each data entry, and compare it with all historical data.

C. Store each data entry as the primary key in a separate database and apply an index.

D. Maintain a database table to store the hash value and other metadata for each data entry.

Answer: A

Explanation:

Inventory data can often be naturally duplicate. Assigning a unique GUID at sender's end is ensuring that we can track a unique record reliably at the recieving end and if there are issues which causes same field to be sent twice, we can easily dedup using the GUID with lesser hassle.

Answer "D" is not as efficient or error-proof due to two reasons

1. You need to calculate hash at sender as well as at receiver end to do the comparison. Waste of computing power.

2. Even if we discount the computing power, we should note that the system is sending inventory information.

Two messages sent at different can denote same inventory level (and thus have same hash). Adding sender time stamp to hash will defeat the purpose of using hash as now retried messages will have different timestamp and a different hash.

if timestamp is used as message creation timestamp than that can also be used as a UUID.

Question: 22 CertyIQ

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a
Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks.
What should you do?

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

B. Grant the user access to Google Cloud Shell.

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

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

Answer: D

Explanation:

Datalab provides Jupyter for this kind of work

Cloud Datalab -> AI Notebooks -> Vertex AI Workbench

Question: 23 CertyIQ

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

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

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

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

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

Answer: B

Explanation:

Pubsub for realtime, Dataflow for pipeline, Bigquery for analytics.You can use cloud data flow for both batch and streaming pipelines. Pub sub will be used to stream data into cloud data flow.

Question: 24 CertyIQ

You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table

CLICK_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. Now, you want to compute web session durations of users who visit your site, and you want to change its data type to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

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

B. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each row. Reference the column TS instead of the column DT from now on.

C. Create a view CLICK_STREAM_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK_STREAM_V instead of the table CLICK_STREAM from now on.

D. Add two columns to the table CLICK STREAM: TS of the TIMESTAMP type and IS_NEW of the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEW to true. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS_NEW must be true.

E. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.

Answer: E

Explanation:

more simple and reasonable. Also recommended if not concerned about cost but simplicity.

Reference:

https://cloud.google.com/bigquery/docs/manually-changing-schemas#changing_a_columns_data_type

Question: 25 CertyIQ

You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.

B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.

C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.

D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

Answer: D

Explanation:

Using the Stack driver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.

A and B are wrong since don't notify anything to the monitoring tool.

C has no filter on what will be notified. We want only some tables.

Question: 26 CertyIQ

You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?

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

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

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

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

Answer: B

Explanation:

As external consultant just going to assist with coding, it means he is not going to test pipeline himself most likely internal developer will perform this task(as project has private data) thus consultant does not need data access. B seems most appropriate option here as it will only allow consultant to verify logic or flow of the pipeline.

Question: 27 CertyIQ

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

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

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

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

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

Answer: B

Explanation:

Data that is co-dependent is high corelated is some kind of reduldant information in some cases. If the features x1, x2 and x3 are x2 = x1 + 1 and x3 = 2*x1, for example, x2 and x3 are reduldant because can be explained with x1 feature, so can be excluded of the the model. Other option is to group this features. There is a lot of ways to resolve, but the main ideia is to use data engineer in co-depedent features to reduce the number of features in the modelnull values can have many meanings and need different approach to handle, otherwise it causes inaccurate model, so not D

Question: 28 CertyIQ

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

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

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

A. Specify the TableReference object in the code.

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

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

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

Answer: B

Explanation:

Use .from Query operation to read specific fields from the table. Big Query IO. read. from() directly reads the whole table from Big Query. This function exports the whole table to temporary files in Google Cloud Storage, where it will later be read from. This requires almost no computation, as it only performs an export job, and later Dataflow reads from GCS (not from Big Query).

Big Query IO .read .from Query() executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs).

Question: 29 CertyIQ

Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

A. Use a row key of the form <timestamp>.

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

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

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

Answer: D

Explanation:

Best practices of bigtable states that rowkey should not be only timestamp or have timestamp at starting. It's better to have sensorid and timestamp as rowkey

Question: 30 CertyIQ

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

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

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

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

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

Answer: B

Explanation:

Bigquery is most suitable for analytical purposes and the question is asking about 'minimal impact' on current DB

A is correct, unless you are not a google partner and you want to spend money and time on infra.

C and D are also correct, if you are a Hadoop master and you still want to be on a local environment for C and for both answers you are just solving the ETL part.

B is the correct answer since you are performing the ETL and using a specialized analytic tool (BigQuery) for which is the main issue of this question (perform analytics without having an impact on the operations).

Question: 31 CertyIQ

You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?

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

B. Update the current pipeline and provide the transform mapping JSON object.

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

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

Answer: A

Explanation:

This option is correct as the key requirement is not to lose

the data, the Dataflow pipeline can be stopped using the Drain option.

Drain options would cause Dataflow to stop any new processing, but would

also allow the existing processing to complete

Question: 32 CertyIQ

Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?

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

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

C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.

D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

Answer: A

Explanation:

Cloud Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Cloud Bigtable distribute the workload across all of the nodes in your cluster. If reads and writes cannot be spread across all of your Cloud Bigtable nodes, performance will suffer.

If you find that you're reading and writing only a small number of rows, you might need to redesign your schema so that reads and writes are more evenly distributed.

Question: 33 CertyIQ

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud
Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?

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

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

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

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

Answer: B

Explanation:

Stack driver monitoring is for performance, not logging of missing data.

Question: 34 CertyIQ

Flowlogistic Case Study -

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

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

Solution Concept -
Flowlogistic wants to implement two concepts using the cloud:
✑ Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their

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

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

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

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

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

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

CFO Statement -
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to
BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?

A. Store the common data in BigQuery as partitioned tables.

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

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

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

Answer: C

Explanation:

avro data can be accessed by spark as well

Question: 35 CertyIQ

Flowlogistic Case Study -

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

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

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

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

Business Requirements -
✑ Build a reliable and reproducible environment with scaled panty of production.

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

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

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

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

CFO Statement -
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system.
You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

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

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

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

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

Answer: A

Explanation:

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

as explained by JayZeeLee :

B is incorrect, because local SSD wouldn't satisfy the needs.

C is incorrect, because one of the requirements is 'Global', Cloud SQL is well suited for regional applications. Cloud Spanner is a better suit in that regard.

D is incorrect, because Load Balancer is for web traffic, not messages.

Question: 36 CertyIQ

Flowlogistic Case Study -

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

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

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

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

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

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

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

shipping.

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

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

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

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

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

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

Answer: C

Explanation:

A logical view can be created with only the required columns which is required for visualization. B is not the right option as you will create a table and make it static. What happens when the original data is updated. This new table will not have the latest data and hence view is the best possible option here.

Question: 37 CertyIQ

Flowlogistic Case Study -

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

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

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

Existing Technical Environment -
Flowlogistic architecture resides in a single data center:
✑ Databases
8 physical servers in 2 clusters

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

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

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

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

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

CFO Statement -
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single
Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in
Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?

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

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

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

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

Answer: B

Explanation:

A. There is no indication that the application can do this. Moreover, due to networking problems, it is possible that Pub/Sub doesn't receive messages in order. This will analysis difficult.

B. This makes sure that you have access to publishing timestamp which provides you with the correct ordering of messages.

C. If timestamps are already messed up, BigQuery will get wrong results anyways.

D. The timestamp we are interested in is when the data was produced by the publisher, not when it was received by Pub/Sub.

Question: 38 CertyIQ

MJTelco Case Study -

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

Company Background -
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
- Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
- Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

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

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

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

CEO Statement -
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed

data pipelines to meet our reliability and capacity commitments.

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

CFO Statement -
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?

A. The zone
B. The number of workers
C. The disk size per worker
D. The maximum number of workers

Answer: D

Explanation:

A: The zone has nothing to do with scaling computer power.

B: The key word here is, "Scale its compute power up AS REQUIRED", with this answer, the number of workers would immediately scale the computer power.

C: we need to scale compute power, not storage

D: is the correct answer, changing the Number of Maximum workers will allow Dataflow to add up to that number of workers if required.

Reference:

https://cloud.google.com/dataflow/docs/reference/pipeline-options#resource_utilization

Question: 39 CertyIQ

MJTelco Case Study -

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

Company Background -
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
✑ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
✑ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments `" development/test, staging, and production `" to meet the needs of running experiments, deploying new features, and serving production customers.

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

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

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

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

CFO Statement -
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
You need to compose visualizations for operations teams with the following requirements:
✑ The report must include telemetry data from all 50,000 installations for the most resent 6 weeks (sampling once every minute).
✑ The report must not be more than 3 hours delayed from live data.
✑ The actionable report should only show suboptimal links.
✑ Most suboptimal links should be sorted to the top.
✑ Suboptimal links can be grouped and filtered by regional geography.
✑ User response time to load the report must be <5 seconds.
Which approach meets the requirements?

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

B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.

C. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.

D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.

Answer: D

Explanation:

1. DataStudio and BQ are the simpliest way to do it
2. They also can activate BI Engine feature to improve the response time.

Question: 40 CertyIQ

MJTelco Case Study -

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

Company Background -
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept -
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
✑ Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
✑ Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments `" development/test, staging, and production `" to meet the needs of running experiments, deploying new features, and serving production customers.

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

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

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

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

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

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

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

Answer: BE

Explanation:

Even if now BQ offers table level access control, since the number of tables can be expected to be high, controlling access at the dataset level would ease operations. That is why I would still go for E instead of C.

Question: 41 CertyIQ

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?

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

Answer: D

Explanation:

rowkey be Device_Id+Date(reverse)

Question: 42 CertyIQ

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

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

Answer: B

Explanation:

SPARK > hadoop, pig, hive

Spark performs in-memory processing and faster, which results in optimization of job's processing time

Question: 43 CertyIQ

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

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

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

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

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

Answer: A

Explanation:

Answer will be A because when you create View it does not store extra space and its a logical representation, for rest of the option you need to write large code and extra processing for dataflow/dataproc

Question: 44 CertyIQ

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

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

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

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

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

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

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

Answer: A

Explanation:

From Google cloud documentation

The rows of an index table are sorted first by ancestor and then by property values, in the order specified in the index definition. The perfect index for a query, which allows the query to be executed most efficiently, is defined on the following properties, in order:

Properties used in equality filters

Property used in an inequality filter (of which there can be no more than one)

Properties used in sort orders

Properties used in projections (that are not already included in sort orders)

Question: 45 CertyIQ

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.

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.

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

Answer: C

Explanation:

A: Dataproc is a managed Apache Spark and Apache Hadoop service, makes no sense to use it

B: This might sound as the cheapest, but is highly error prone, besides, anyone in charge of this has a salary and I doubt is a low one.

C: This is the easiest/fastest/cheapest way to trigger job runs, you can even set retry attempts.

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

D: Setting this would be much more expensive than the cron-job

Question: 46 CertyIQ

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

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

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

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

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

Answer: B

Explanation:

As per google docs on BigQuery:

Use cases for external data sources include:

Loading and cleaning your data in one pass by querying the data from an external data source (a location external to BigQuery) and writing the cleaned result into BigQuery storage.

Having a small amount of frequently changing data that you join with other tables. As an external data source, the frequently changing data does not need to be reloaded every time it is updated.

Question: 47 CertyIQ

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

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

A. BigQuery

B. Cloud SQL

C. Cloud Bigtable

D. Cloud Datastore

Answer: A

Explanation:

You want to optimize the data schema + Machine Learning --> Bigquery. So A

Question: 48 CertyIQ

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

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

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

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

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

Answer: C

Explanation:

If you don't specify an encoding, or if you specify UTF-8 encoding when the CSV file is not UTF-8 encoded, BigQuery attempts to convert the data to UTF-8. Generally, your data will be loaded successfully, but it may not match byte-for-byte what you expect."

Reference:

https://cloud.google.com/bigquery/docs/loading-data-cloud-storage-csv#details_of_loading_csv_data

Question: 49 CertyIQ

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

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

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

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

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.

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.

Answer: CD

Explanation:

A: size is small enough that compressing each file will not help (indeed, it may even add overhead).

B: bandwidth is not a problem, no need to increase.

C: Parallel uploading the files with -m will increase speed in general.

D: many individual small files are a problem, since each file adds overhead to the processing and upload to GCS, and the upload sped of GCS is proportional to the size. If we pack all the small files in a bigger single TAR, it will improve the overall performance.

E: Storage Transfer Service is intended to move 100s of TB, and requires a 300Mbps connection as minimum (the doc even states that if your connection is less than 300Mbps is better to use gsutil).

Reference:

https://cloud.google.com/storage-transfer/docs/on-prem-overview#requirements

https://jbrojbrojbro.medium.com/parallel-uploads-for-smaller-files-387ff86afc74

Question: 50 CertyIQ

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

A. Redis
B. HBase
C. MySQL
D. MongoDB
E. Cassandra
F. HDFS with Hive

Answer: BDE

Explanation:

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.

Question: 51 CertyIQ

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

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

Answer: ACE

Explanation:

The tools to prevent overfitting: less variables, regularization, early ending on the training...

Overfitting means that the classifier knows too well the data and fails to generalize. We should use a smaller number of features to help the classifier generalize, and more examples so that it can have more variety.

The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set.

- Adding more training data will increase the complexity of the training set and help with the variance problem.

- Reducing the feature set will ameliorate the overfitting and help with the variance problem.

- Increasing the regularization parameter will reduce overfitting and help with the variance problem.

Reference:

https://github.com/mGalarnyk/datasciencecoursera/blob/master/Stanford_Machine_Learning/Week6/AdviceQuiz.r

Question: 52 CertyIQ

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

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

Answer: C

Explanation:

The answer is C because Service Account is the best way to access the BigQuery API if your application can run jobs associated with service credentials rather than an end-user's credentials, such as a batch processing pipeline. https://cloud.google.com/bigquery/docs/authentication

Data owners cant create jobs or queries. -> B out We need service Account -> D out Access only granting me does not solve the problem -> A out The answer is C. (Minimum rights to perform the job)

Question: 53 CertyIQ

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

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

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

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

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

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

Answer: D

Explanation:

D Image says that average(dark) and maximum(light) have difference in few times, this it is a skew The color indicators show the relative timings for all steps across all stages. For example, the COMPUTE step of Stage 00 shows a bar whose shaded fraction is 21/30 since 30ms is the maximum time spent in a single step of any stage. The parallel input information shows that each stage required only a single worker, so there's no variance between average and slowest timings.

Reference:

https://cloud.google.com/bigquery/query-plan

https://cloud.google.com/bigquery/docs/best-practices-performance-patterns

Question: 54 CertyIQ

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

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

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

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

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

Answer: B

Explanation:

A and C is out since it does not give the real-time feature. The battle is between B and D.

If we think about the business scenario we need to give the bid to who published it first. D is given to the who processed first. Technically it can be implemented bcz DataFlow is only support pull subscription with cloud pub-sub.

In answer B, events are pushed to the endpoint. Explicitly they haven't mentioned that it pushes to the "Cloud Data Flow". It may be a custom API endpoint. Some people have a misunderstanding about this point. There is no clue about that. It may be a REST API endpoint or application.

The event is consist of the timestamp field itself. So by inserting it into the DB we can find out who is the winner.

Based on that we can Mark answer as B.

Question: 55 CertyIQ

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

A. Create a new view over events using standard SQL

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

C. Create a new view over events_partitioned using standard SQL

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

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

Answer: CD

Explanation:

C = A standard SQL query cannot reference a view defined using legacy SQL syntax.

D = For the ODBC drivers is needed a service account which will get a standard Bigquery role

Question: 56 CertyIQ

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

A. Use the TABLE_DATE_RANGE function

B. Use the WHERE_PARTITIONTIME pseudo column

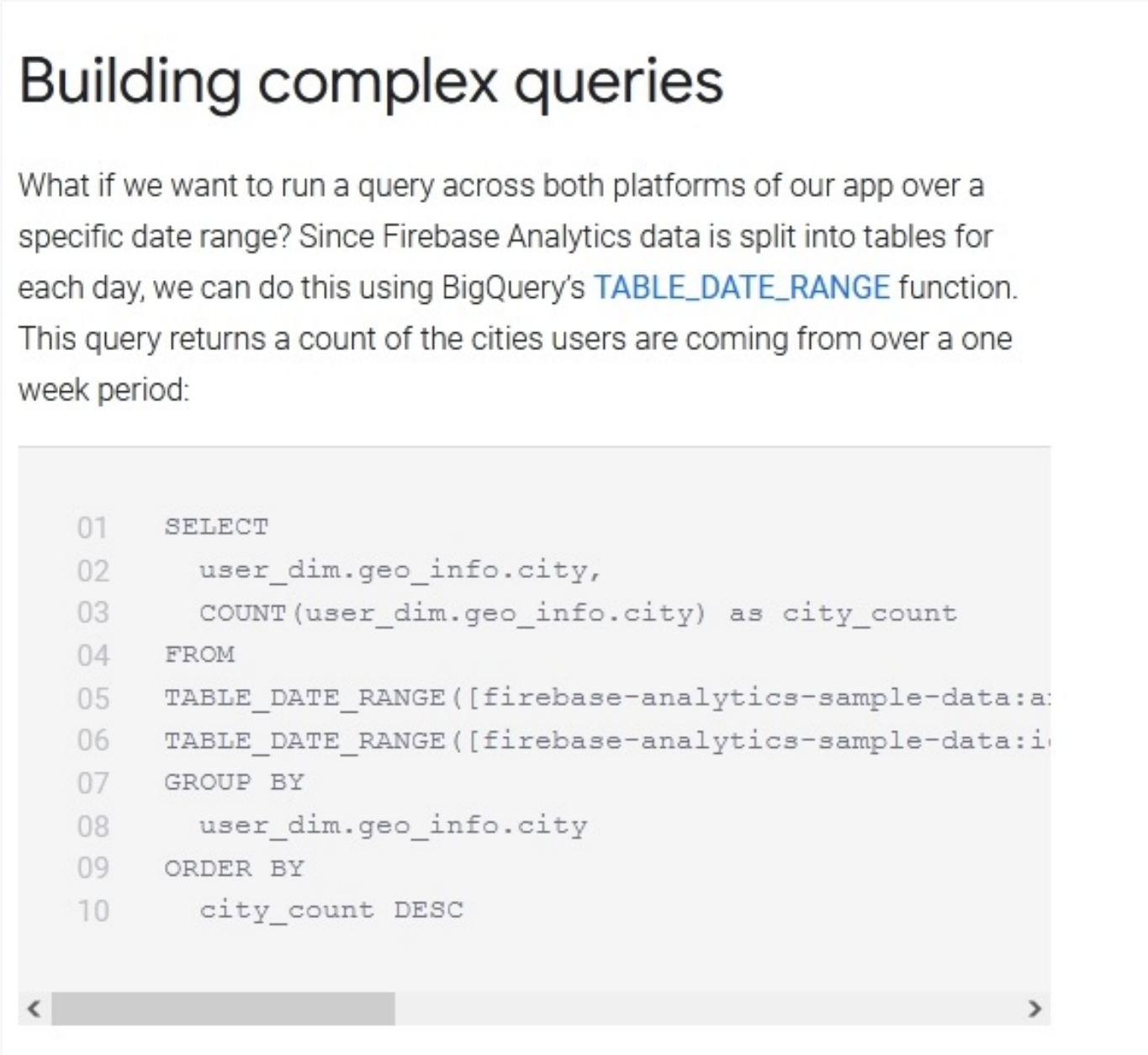
C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD

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

Answer: A

Explanation:

Reference:
https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understand-your-mobile-app?hl=am
" target="_blank" style="word-break: break-all;">



Question: 57 CertyIQ

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

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

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

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

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

Answer: D

Explanation:

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

Question: 58 CertyIQ

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

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

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

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

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

Answer: B

Explanation:

Two reasons, it is a cleaner approach with single job to handle the calibration before the data is used in the pipeline. Second, doing this step in later stages can be complex and maintenance of those jobs in the future will become challenging.

Question: 59 CertyIQ

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

B. Cloud SQL

C. Cloud BigTable

D. Cloud Datastore

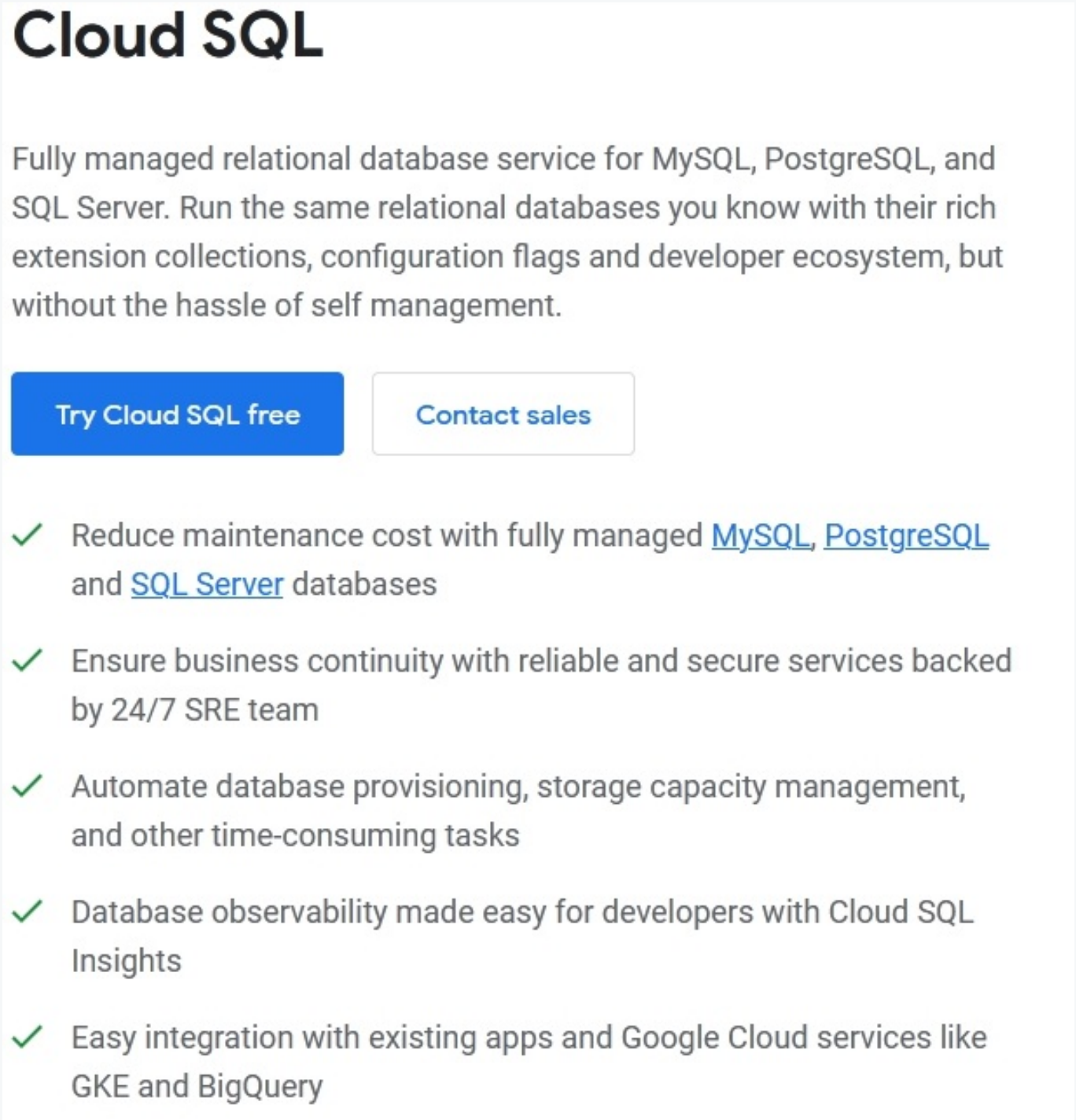
Answer: B

Explanation:

Big query is not suitable for transactional use case, and Cloud SQL supports transactions as well as analysis through a BI tool.

Reference:

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



Question: 60 CertyIQ

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

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

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

Answer: B

Explanation:

Convert MANY sharded tables into a single ONE (partitioned) table

C'mon, how much time are you going to take to partition every single table you have? second point and the most important, you have a table for every SINGLE DAY "LOGS_YYYYMMDD" partitioning every table will end on scanning all the records of each table when you query them by date ranges using the wildcards, there will be no difference on time-partitioning each table versus consuming them as described.

Question: 61 CertyIQ

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

A. Migrate the workload to Google Cloud Dataflow

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

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

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

Answer: B

Explanation:

Preemptible workers are the default secondary worker type. They are reclaimed and removed from the cluster if they are required by Google Cloud for other tasks. Although the potential removal of preemptible workers can affect job stability, you may decide to use preemptible instances to lower per-hour compute costs for non-critical data processing or to create very large clusters at a lower total cost

Reference:

https://cloud.google.com/dataproc/docs/concepts/compute/secondary-vms

Question: 62 CertyIQ

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

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

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

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

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

Answer: C

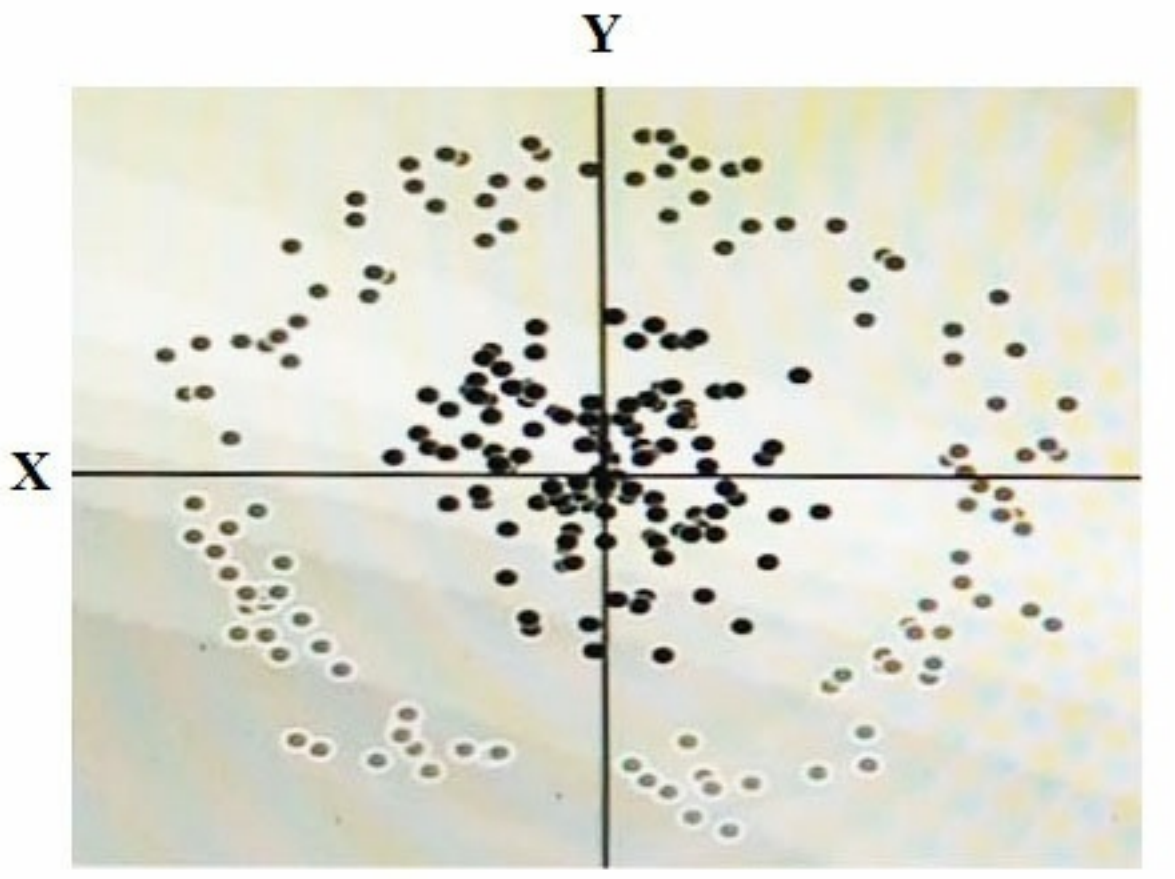
Explanation:

"Watermark in implementation is a monotonically increasing timestamp. When Beam/Dataflow see a record with an event timestamp that is earlier than the watermark, the record is treated as late data."

A is a direct No, if data don't have timestamp, we'll only have the procesing time and not the "event time". B is not either, sliding windows are not for this. Hopping|sliding windowing is useful for taking running averages of data, but not to process late data. D looks correct but has one concept missing, the watermark to know if the process time is ok with the event time or not. I'm not 100% sure is incorrect. If, since we have a "predictable time period", might be this will do. I mean, if our dashboard is shown after the last input data has arrived (single global window), this should be ok. We'd have a "perfect watermark". Anyway we'd need triggering .

Question: 63 CertyIQ

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



A. X2+Y2

B. X2

C. Y2

D. cos(X)

Answer: A

Explanation:

For fitting a linear classifier when the data is in a circle use A.

Question: 64 CertyIQ

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

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

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

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

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

Answer: C

Explanation:

Service Account are best option when granting access from tools/appllications

Question: 65 CertyIQ

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

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

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

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

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

Answer: B

Explanation:

real-valued can not be null N/A or empty, have to be “0”, so it has to be B.

Dataprep suites this, so none of dataflow options qualify as answer. Then 0 can be real-value than a "~none'.

Question: 66 CertyIQ

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

needed. What should you do?

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

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

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

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

Answer: B

Explanation:

A makes no sense, you need to use your own keys. You don't create keys locally and upload them, you should import it to make it work..using the kms public key...not just "uploading" it. C is also out. IT's between B and D Cloud KMS is a cloud-hosted key management service that lets you manage cryptographic keys for your cloud services the same way you do on-premises, You can generate, use, rotate, and destroy cryptographic keys from there. Since you want to encrypt data at rest, is B, you don't use them for any API calls.

Reference:

https://cloud.google.com/compute/docs/disks/customer-managed-encryption

https://cloud.google.com/security/encryption-at-rest/

Question: 67 CertyIQ

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

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

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

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

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

Answer: C

Explanation:

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

Question: 68 CertyIQ

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data

pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?

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

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

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

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

Answer: C

Explanation:

best suitable for the purpose with autoscaling and google recommended transform engine between pubsub n bq

C only as referred by MaxNRG

C. Dataproc does not seem to be a good solution in this case as it always require a manual intervention to adjust resources. Autoscaling with dataflow will automatically handle changing data volumes with no manual intervention, and monitoring through Stackdriver can be used to spot bottleneck. Total execution time is not good there as it does not provide a precise view on potential bottleneck.

Question: 69 CertyIQ

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

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

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

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

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

Answer: A

Explanation:

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

Question: 70 CertyIQ

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

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

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

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.

Answer: B

Explanation:

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.

Question: 71 CertyIQ

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

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

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

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

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

Answer: A

Explanation:

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

Question: 72 CertyIQ

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

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

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

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

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

Answer: C

Explanation:

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 approximately 24 hours. Temporary tables are useful for one-time operations, such as loading data into a data warehouse.

"Dan Sullivan" Book

Question: 73 CertyIQ

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

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

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

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

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

Answer: C

Explanation:

Spanner allows transaction tables to scale horizontally and secondary indexes for range queries

Question: 74 CertyIQ

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

A. Cloud Bigtable

B. Google BigQuery

C. Google Cloud Storage

D. Google Cloud Datastore

Answer: A

Explanation:

Reference:
https://cloud.google.com/bigtable/docs/schema-design-time-series
" target="_blank" style="word-break: break-all;">

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

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

Question: 75 CertyIQ

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

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

Answer: A

Explanation:

Reference:
https://cloud.google.com/bigquery/docs/share-access-views
" target="_blank" style="word-break: break-all;">

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

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

Question: 76 CertyIQ

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

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

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

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

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

Answer: D

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!

Question: 77 CertyIQ

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

A. Subsample your test dataset.

B. Subsample your training dataset.

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

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

Answer: B

Explanation:

It is B. D would improve the accuracy, not speed.

Question: 78 CertyIQ

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

A. PigLatin using Pig

B. HiveQL using Hive

C. Java using MapReduce

D. Python using MapReduce

Answer: A

Explanation:

Pig is scripting language which can be used for checkpointing and splitting pipelines

Question: 79 CertyIQ

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

A. Increase the CPU size on your server.

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

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

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

Answer: C

Explanation:

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.

Question: 80 CertyIQ

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?

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

Answer: C

Explanation:

Bigquery and Big table =PB storage capacity

Bigtable=to read scan rows Big query select row to read

Thank you

Thank you for being so interested in the premium exam material.
I'm glad to hear that you found it informative and helpful.

But Wait

I wanted to let you know that there is more content available in the full version.
The full paper contains additional sections and information that you may find helpful, and I encourage you to download it to get a more comprehensive and detailed view of all the subject matter.

Download Full Version Now

Future is Secured

100% Pass Guarantee

24/7 Customer Support

Mail us - certyiqofficial@gmail.com

Free Updates

Lifetime Free Updates!

Total: **283 Questions**

Link: https://certyiq.com/papers?provider=google&exam=professional-data-engineer