Google.Professional-Data-Engineer.v2023-03-14.q177

Exam Code:	Professional-Data-Engineer
Exam Name:	Google Certified Professional Data Engineer Exam
Certification Provider:	Google
Free Question Number:	177
Version:	v2023-03-14
# of views:	1973
# of Questions views:	1770
https://www.dumpsdb.com/dumps/Google/Professional-Data-Engineer/Google.Professional-Data-Engineer.v2023-03-14.q177	

NEW QUESTION: 1
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time.
This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?
A. Implement clustering in BigQuery on the ingest date column.
B. Re-create the table using data partitioning on the package delivery date.
C. Implement clustering in BigQuery on the package-tracking ID column.
D. Tier older data onto Cloud Storage files, and leverage extended tables.
Answer: (SHOW ANSWER)

NEW QUESTION: 2
All Google Cloud Bigtable client requests go through a front-end server _____ they are sent to a Cloud Bigtable node.
A. before
B. after
C. only if
D. once
Answer: A (LEAVE A REPLY)
Explanation

In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node.
The nodes are organized into a Cloud Bigtable cluster, which belongs to a Cloud Bigtable instance, which is a container for the cluster. Each node in the cluster handles a subset of the requests to the cluster.
When additional nodes are added to a cluster, you can increase the number of simultaneous requests that the cluster can handle, as well as the maximum throughput for the entire cluster.
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION: 3
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 an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.
C. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences.
Deploy the models using Cloud Dataproc. Call the models from your application.
D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 4
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 session window with a gap time duration of 60 minutes.
B. Use a sliding time window with a duration of 60 minutes.
C. Use a fixed-time window with a duration of 60 minutes.

D. Use a global window with a time based trigger with a delay of 60 minutes.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 5
What are two methods that can be used to denormalize tables in BigQuery?
A. 1) Split table into multiple tables; 2) Use a partitioned table
B. 1) Join tables into one table; 2) Use nested repeated fields
C. 1) Use a partitioned table; 2) Join tables into one table
D. 1) Use nested repeated fields; 2) Use a partitioned table
Answer: B (LEAVE A REPLY)
The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are dealing with sales transactions, you would write each individual fact to a record, along with the accompanying dimensions such as order and customer information.
The other method for denormalizing data takes advantage of BigQuery's native support for nested and repeated structures in JSON or Avro input data. Expressing records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which would be represented as nested, repeated elements.

NEW QUESTION: 6
After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.
What should you do?
A. Select random samples from the tables using the RAND() function and compare the samples.
B. Select random samples from the tables using the HASH() function and compare the samples.
C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.
D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.
Answer: C (LEAVE A REPLY)
Full comparison with this option, rest are comparison on sample which doesn't ensure all the data will be ok.

NEW QUESTION: 7

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 the sharded tables into a single partitioned table
B. Create separate views to cover each month, and query from these views
C. Convert all daily log tables into date-partitioned tables
D. Enable query caching so you can cache data from previous months
Answer: C (LEAVE A REPLY)

NEW QUESTION: 8
For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?
A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.
Answer: (SHOW ANSWER)
It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,
If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.
It is recommended to avoid creating your Compute Engine instance in a different region from your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.

NEW QUESTION: 9
Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?
A. Cloud Dataflow
B. Cloud Dataprep
C. Cloud Composer
D. Cloud Dataproc
Answer: D (LEAVE A REPLY)

NEW QUESTION: 10
You are working on a sensitive project involving private user dat
A. Grant the consultant the Cloud Dataflow Developer role on the project.
B. Create a service account and allow the consultant to log on with it.
C. 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?
D. Create an anonymized sample of the data for the consultant to work with in a different project.
E. Grant the consultant the Viewer role on the project.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 11
Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of dat A.
Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?
A. 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.
B. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.
C. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.
D. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 12
You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline.
Which component will be used for the data processing operation?
A. PCollection
B. Transform
C. Pipeline
D. Sink API
Answer: B (LEAVE A REPLY)
In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.
Reference: https://cloud.google.com/dataflow/model/programming-model

NEW QUESTION: 13

What are two of the characteristics of using online prediction rather than batch prediction?
A. It is optimized to handle a high volume of data instances in a job and to run more complex models.
B. Predictions are returned in the response message.
C. Predictions are written to output files in a Cloud Storage location that you specify.
D. It is optimized to minimize the latency of serving predictions.
Answer: B,D (LEAVE A REPLY)
Online prediction
.Optimized to minimize the latency of serving predictions.
.Predictions returned in the response message.
Batch prediction
.Optimized to handle a high volume of instances in a job and to run more complex models.
.Predictions written to output files in a Cloud Storage location that you specify.

NEW QUESTION: 14
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 (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 15
Which action can a Cloud Dataproc Viewer perform?
A. Submit a job.
B. Create a cluster.
C. Delete a cluster.
D. List the jobs.
Answer: D (LEAVE A REPLY)
A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.
Reference:
https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations_summary

NEW QUESTION: 16

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data.
What should you do?
A. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.
B. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
D. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket.
Manually destroy the key previously used for encryption, and rotate the key once.
Answer: (SHOW ANSWER)

NEW QUESTION: 17
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the
"Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?
A. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google cp Cloud.

B. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.
D. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket.
Manually destroy the key previously used for encryption, and rotate the key once.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 18
You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)
A. Denormalize the data as must as possible.
B. Preserve the structure of the data as much as possible.
C. Use BigQuery UPDATE to further reduce the size of the dataset.
D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.
Answer: A,D (LEAVE A REPLY)
Denormalization will help in performance by reducing query time, update are not good with bigquery.

NEW QUESTION: 19
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. Manually start the Cloud Dataflow job each morning when you get into the office.
B. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.
C. Change the processing job to use Google Cloud Dataproc instead.
D. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 20
Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?
A. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.
B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' projects. Restrict access to the Cloud Storage bucket.
C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit logs. Restrict access to the project with the exported logs.
D. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 21
You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.
What should you do?
A. Use Cloud Dataflow with Beam to detect errors and perform transformations.
B. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
C. Use federated tables in BigQuery with queries to detect errors and perform transformations.
D. Use Cloud Dataprep with recipes to detect errors and perform transformations.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 22
Which SQL keyword can be used to reduce the number of columns processed by BigQuery?
A. BETWEEN
B. WHERE
C. SELECT
D. LIMIT
Answer: C (LEAVE A REPLY)
SELECT allows you to query specific columns rather than the whole table.
LIMIT, BETWEEN, and WHERE clauses will not reduce the number of columns processed by BigQuery.
Reference: https://cloud.google.com/bigquery/launch-checklist#architecture_design_and_development_checklist

NEW QUESTION: 23
How would you query specific partitions in a BigQuery table?

A. Use the DAY column in the WHERE clause
B. Use the EXTRACT(DAY) clause
C. Use the __PARTITIONTIME pseudo-column in the WHERE clause
D. Use DATE BETWEEN in the WHERE clause
Answer: C (LEAVE A REPLY)
Explanation
Partitioned tables include a pseudo column named _PARTITIONTIME that contains a date-based timestamp for data loaded into the table. To limit a query to particular partitions (such as Jan 1st and 2nd of 2017), use a clause similar to this:
WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02') Reference: https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

NEW QUESTION: 24
Which of these statements about exporting data from BigQuery is false?
A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.
B. The only supported export destination is Google Cloud Storage.
C. Data can only be exported in JSON or Avro format.
D. The only compression option available is GZIP.
Answer: C (LEAVE A REPLY)
Explanation
Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported.
Reference: https://cloud.google.com/bigquery/docs/exporting-data

NEW QUESTION: 25
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: (SHOW ANSWER)
Bigquery understands UTF-8 encoding anything other than that will result in data issues with schema.

NEW QUESTION: 26

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. Refresh your browser tab showing the visualizations.
B. Disable caching by editing the report settings.
C. Disable caching in BigQuery by editing table details.
D. Clear your browser history for the past hour then reload the tab showing the virtualizations.
Answer: (SHOW ANSWER)

NEW QUESTION: 27
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: D (LEAVE A REPLY)
Explanation/Reference: https://cloud.google.com/solutions/data-lifecycle-cloud-platform

NEW QUESTION: 28
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 a service account and grant dataset access to that account. Use the service account's private key to access the dataset
B. Create groups for your users and give those groups access to the dataset
C. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
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: (SHOW ANSWER)

NEW QUESTION: 29
Does Dataflow process batch data pipelines or streaming data pipelines?
A. Only Batch Data Pipelines

B. Both Batch and Streaming Data Pipelines
C. Only Streaming Data Pipelines
D. None of the above
Answer: B (LEAVE A REPLY)
Explanation
Dataflow is a unified processing model, and can execute both streaming and batch data pipelines
Reference: https://cloud.google.com/dataflow/

NEW QUESTION: 30
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 (LEAVE A REPLY)
A watermark is a threshold that indicates when Dataflow expects all of the data in a window to have arrived. If new data arrives with a timestamp that's in the window but older than the watermark, the data is considered late data.

NEW QUESTION: 31
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 (LEAVE A REPLY)
Explanation/Reference:
Reference: https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-
tensorflow-30505541d877

NEW QUESTION: 32
Which of these sources can you not load data into BigQuery from?
A. File upload
B. Google Drive
C. Google Cloud Storage
D. Google Cloud SQL
Answer: D (LEAVE A REPLY)
Explanation
You can load data into BigQuery from a file upload, Google Cloud Storage, Google Drive, or Google Cloud Bigtable. It is not possible to load data into BigQuery directly from Google Cloud SQL. One way to get data from Cloud SQL to BigQuery would be to export data from Cloud SQL to Cloud Storage and then load it from there.
Reference: https://cloud.google.com/bigquery/loading-data

NEW QUESTION: 33
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 nonkey 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: D (LEAVE A REPLY)
Reference: https://cloud.google.com/solutions/data-lifecycle-cloud-platform

NEW QUESTION: 34
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 (LEAVE A REPLY)
Reference:
https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understandyour-mobile-app

NEW QUESTION: 35
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 (LEAVE A REPLY)

NEW QUESTION: 36
You want to use a BigQuery table as a data sink. In which writing mode(s) can you use BigQuery as a sink?
A. Both batch and streaming
B. BigQuery cannot be used as a sink
C. Only batch
D. Only streaming
Answer: A (LEAVE A REPLY)
When you apply a BigQueryIO.Write transform in batch mode to write to a single table, Dataflow invokes a BigQuery load job. When you apply a BigQueryIO.Write transform in streaming mode or in batch mode using a function to specify the destination table, Dataflow uses BigQuery's streaming inserts

NEW QUESTION: 37

You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second. What should you do?
A. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a fixed time window of 1 hour.
Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
B. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
C. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuery. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour. If that number falls below 4000, send an alert.
D. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtable. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hour. If that number falls below 4000, send an alert.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 38
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: (SHOW ANSWER)
Explanation

NEW QUESTION: 39
You need to choose a database for a new project that has the following requirements:
* Fully managed
* Able to automatically scale up
* Transactionally consistent
* Able to scale up to 6 TB

* Able to be queried using SQL
Which database do you choose?
A. Cloud Bigtable
B. Cloud Datastore
C. Cloud SQL
D. Cloud Spanner
Answer: (SHOW ANSWER)

NEW QUESTION: 40
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. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
B. Convert the streaming insert code to batch load for individual messages.
C. Re-write the application to load accumulated data every 2 minutes.
D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 41
You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud. What should you do?
A. Use Cloud TPUs after implementing GPU kernel support for your customs ops.
B. Stay on CPUs, and increase the size of the cluster you're training your model on.
C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.
D. Use Cloud TPUs without any additional adjustment to your code.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 42
The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster ____.
A. application node
B. conditional node
C. master node

D. worker node

Answer: C (LEAVE A REPLY)

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix-for example, if your cluster is named "my- cluster", the master-host-name would be "my-cluster-m".

Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web- interfaces#interfaces

NEW QUESTION: 43

You have uploaded 5 years of log data to Cloud Storage A user reported that some data points in the log data are outside of their expected ranges, which indicates errors You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

A. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in
Cloud Storage

B. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage

C. Create a Compute Engine instance and create a new copy of the data in Cloud Storage Skip the rows with errors

D. Import the data from Cloud Storage into BigQuery Create a new BigQuery table, and skip the rows with errors.

Answer: (SHOW ANSWER)

NEW QUESTION: 44

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: C (LEAVE A REPLY)

Explanation/Reference: https://cloud.google.com/solutions/business-intelligence/

NEW QUESTION: 45

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. MySQL
C. HDFS with Hive
D. HBase
E. MongoDB
F. Cassandra
Answer: (SHOW ANSWER)

NEW QUESTION: 46
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 maximum number of workers
B. The disk size per worker
C. The zone
D. The number of workers
Answer: (SHOW ANSWER)

NEW QUESTION: 47
You are designing a cloud-native historical data processing system to meet the following conditions:
* The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.
* A streaming data pipeline stores new data daily.
* Peformance is not a factor in the solution.
* The solution design should maximize availability.
How should you design data storage for this solution?
A. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.
B. Store the data in BigQuery. Access the data using the BigQuery Connector on Cloud Dataproc and Compute Engine.
C. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
D. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 48
If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?
A. Unsupervised learning

B. Regressor
C. Classifier
D. Clustering estimator
Answer: (SHOW ANSWER)
Explanation
Regression is the supervised learning task for modeling and predicting continuous, numeric variables.
Examples include predicting real-estate prices, stock price movements, or student test scores.
Classification is the supervised learning task for modeling and predicting categorical variables.
Examples include predicting employee churn, email spam, financial fraud, or student letter grades.
Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset. Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.
Reference: https://elitedatascience.com/machine-learning-algorithms

NEW QUESTION: 49
The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster
____.
A. application node
B. conditional node
C. master node
D. worker node
Answer: C (LEAVE A REPLY)
Explanation
The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix-for example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m".
Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces

NEW QUESTION: 50
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. Train on the existing data while using the new data as your test set.
B. Continuously retrain the model on a combination of existing data and the new data.
C. Train on the new data while using the existing data as your test set.
D. Continuously retrain the model on just the new data.

Answer: C (LEAVE A REPLY)

NEW QUESTION: 51
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 encoded as Avro in Google Cloud Storage.
B. Store the common data in BigQuery as partitioned tables.
C. Store the common data in BigQuery and expose authorized views.
D. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.

Answer: C (LEAVE A REPLY)

NEW QUESTION: 52
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: C (LEAVE A REPLY)
Topic 2, 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.

NEW QUESTION: 53
Which of these sources can you not load data into BigQuery from?
A. File upload
B. Google Drive
C. Google Cloud Storage
D. Google Cloud SQL
Answer: (SHOW ANSWER)
You can load data into BigQuery from a file upload, Google Cloud Storage, Google Drive, or Google Cloud Bigtable. It is not possible to load data into BigQuery directly from Google Cloud

SQL. One way to get data from Cloud SQL to BigQuery would be to export data from Cloud SQL to Cloud Storage and then load it from there.
Reference: https://cloud.google.com/bigquery/loading-data

NEW QUESTION: 54
Cloud Dataproc charges you only for what you really use with _____ billing.
A. month-by-month
B. minute-by-minute
C. week-by-week
D. hour-by-hour
Answer: B (LEAVE A REPLY)
One of the advantages of Cloud Dataproc is its low cost. Dataproc charges for what you really use with minute-by-minute billing and a low, ten-minute-minimum billing period.

NEW QUESTION: 55
The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?
A. Workers
B. Masters, workers, and parameter servers
C. Workers and parameter servers
D. Parameter servers
Answer: C (LEAVE A REPLY)
The CUSTOM tier is not a set tier, but rather enables you to use your own cluster specification. When you use this tier, set values to configure your processing cluster according to these guidelines:
You must set TrainingInput.masterType to specify the type of machine to use for your master node.
You may set TrainingInput.workerCount to specify the number of workers to use.
You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use.
You can specify the type of machine for the master node, but you can't specify more than one master node.
Reference: https://cloud.google.com/ml-engine/docs/training-overview#job_configuration_parameters

NEW QUESTION: 56
An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and cat the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?
A. Write an application that makes a queue in a NoSQL database

B. Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.
C. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.
D. Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic
Answer: A (LEAVE A REPLY)

NEW QUESTION: 57
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. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
B. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.
C. Redefine the schema by evenly distributing reads and writes across the row space of the table.
D. The performance issue should be resolved over time as the site of the BigDate cluster is increased.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 58
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. Dimensionality Reduction
B. Dropout Methods
C. Serialization
D. Threading
Answer: B (LEAVE A REPLY)

NEW QUESTION: 59
An aerospace company uses a proprietary data format to store its night dat
a. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming as few resources as possible.
What should you do?

A. Use a standard Dataflow pipeline to store the raw data m BigQuery and then transform the format later when the data is used
B. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format
C. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format
D. Write a she script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source
Answer: C (LEAVE A REPLY)

NEW QUESTION: 60
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
Scale and harden their PoC to support significantly more data flows generated when they ramp to more
- than 50,000 installations.
Refine their machine-learning cycles to verify and improve the dynamic models they use to control
- topology definition.
MJTelco will also use three separate operating environments - development/test, staging, and production
- to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements

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

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement

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

CTO Statement

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

CFO Statement

The project is too large for us to maintain the hardware and software required for the data and analysis.

Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

You need to compose visualization for operations teams with the following requirements:

- Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)
- The report must not be more than 3 hours delayed from live data.
- The actionable report should only show suboptimal links.
-

Most suboptimal links should be sorted to the top.
- Suboptimal links can be grouped and filtered by regional geography.
- User response time to load the report must be <5 seconds.
- You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?

A. Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.
B. Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API.
C. Export the data to a spreadsheet, compose a series of charts and tables, one for each possible combination of criteria, and spread them across multiple tabs.
D. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 61
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 SQL, and Cloud Storage
B. Cloud Dataflow, Cloud SQL, and Cloud Storage
C. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
D. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
E. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage
Answer: (SHOW ANSWER)

NEW QUESTION: 62
Government regulations in the banking industry mandate the protection of client's personally identifiable information (PII). Your company requires PII to be access controlled encrypted and compliant with major data protection standards In addition to using Cloud Data Loss Prevention (Cloud DIP) you want to follow Google-recommended practices and use service accounts to control access to PII. What should you do?
A. Use Cloud Storage to comply with major data protection standards. Use one service account shared by all users

B. Use Cloud Storage to comply with major data protection standards. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group
C. Use one service account to access a Cloud SQL database and use separate service accounts for each human user
D. Assign the required identity and Access Management (IAM) roles to every employee, and create a single service account to access protect resources
Answer: B (LEAVE A REPLY)

NEW QUESTION: 63
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?
A. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
B. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
C. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
D. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
Answer: (SHOW ANSWER)

NEW QUESTION: 64
Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)
A. The wide model is used for memorization, while the deep model is used for generalization.
B. A good use for the wide and deep model is a recommender system.
C. The wide model is used for generalization, while the deep model is used for memorization.
D. A good use for the wide and deep model is a small-scale linear regression problem.
Answer: A,B (LEAVE A REPLY)
Explanation
Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.

Reference: https://research.googleblog.com/2016/06/wide-deep-learning-better-together-with.html

NEW QUESTION: 65

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 Load Balancing, Cloud Dataflow, and Cloud Storage

B. Cloud Dataflow, Cloud SQL, and Cloud Storage

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

D. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
E. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
Answer: E (LEAVE A REPLY)

NEW QUESTION: 66
You are working on a sensitive project involving private user dat
a. 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 Cloud Dataflow Developer role on the project.
B. Create a service account and allow the consultant to log on with it.
C. Grant the consultant the Viewer role on the project.
D. Create an anonymized sample of the data for the consultant to work with in a different project.
Answer: (SHOW ANSWER)

NEW QUESTION: 67
Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?
A. Field promotion
B. Randomization
C. Salting
D. Hashing
Answer: (SHOW ANSWER)
By default, prefer field promotion. Field promotion avoids hotspotting in almost all cases, and it tends to make it easier to design a row key that facilitates queries.
Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotting

NEW QUESTION: 68
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 nonkey columns. What should you do?
A. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
B. Use Cloud Spanner 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 SQL for storage. Add secondary indexes to support query patterns.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 69
Which of these is not a supported method of putting data into a partitioned table?

A. If you have existing data in a separate file for each day, then create a partitioned table and upload each file into the appropriate partition.
B. Run a query to get the records for a specific day from an existing table and for the destination table, specify a partitioned table ending with the day in the format "$YYYYMMDD".
C. Create a partitioned table and stream new records to it every day.
D. Use ORDER BY to put a table's rows into chronological order and then change the table's type to "Partitioned".
Answer: D (LEAVE A REPLY)
You cannot change an existing table into a partitioned table. You must create a partitioned table from scratch. Then you can either stream data into it every day and the data will automatically be put in the right partition, or you can load data into a specific partition by using "$YYYYMMDD" at the end of the table name.
Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

NEW QUESTION: 70
When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?
A. 500 TB
B. 1 GB
C. 1 TB
D. 500 GB
Answer: (SHOW ANSWER)
Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions. It is not a good solution for less than 1 TB of data.

NEW QUESTION: 71
You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time.
Which two methods can accomplish this? (Choose two.)
A. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.
B. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
C. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp

for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.
E. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
Answer: (SHOW ANSWER)

NEW QUESTION: 72
You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?
A. Cloud Dataflow
B. Cloud Functions
C. Cloud Scheduler
D. Cloud Composer
Answer: C (LEAVE A REPLY)

NEW QUESTION: 73
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 a combination of existing data and the new data.
B. Train on the existing data while using the new data as your test set.
C. Train on the new data while using the existing data as your test set.
D. Continuously retrain the model on just the new data.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 74
You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query - -dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries.
What should you do?
A. Create a separate table for each ID.
B. Use the LIMIT keyword to reduce the number of rows returned.
C. Recreate the table with a partitioning column and clustering column.
D. Use the bq query - -maximum_bytes_billedflag to restrict the number of bytes billed.
Answer: B (LEAVE A REPLY)
Explanation

NEW QUESTION: 75
You need to create a new transaction table in Cloud Spanner that stores product sales data. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?
A. The current epoch time
B. A concatenation of the product name and the current epoch time
C. A random universally unique identifier number (version 4 UUID)
D. The original order identification number from the sales system, which is a monotonically increasing integer
Answer: C (LEAVE A REPLY)
https://cloud.google.com/spanner/docs/schema-and-data-model#choosing_a_primary_key

NEW QUESTION: 76
Which of the following is NOT one of the three main types of triggers that Dataflow supports?
A. Trigger based on element size in bytes
B. Trigger that is a combination of other triggers
C. Trigger based on element count
D. Trigger based on time
Answer: (SHOW ANSWER)
Explanation
There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers.
You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way Reference:
https://cloud.google.com/dataflow/model/triggers

NEW QUESTION: 77
Which software libraries are supported by Cloud Machine Learning Engine?
A. Theano and TensorFlow

B. Theano and Torch
C. TensorFlow
D. TensorFlow and Torch
Answer: (SHOW ANSWER)
Explanation
Cloud ML Engine mainly does two things:
Enables you to train machine learning models at scale by running TensorFlow training applications in the cloud.
Hosts those trained models for you in the cloud so that you can use them to get predictions about new data.
Reference: https://cloud.google.com/ml-engine/docs/technical-overview#what_it_does

NEW QUESTION: 78
Which of these statements about exporting data from BigQuery is false?
A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.
B. The only supported export destination is Google Cloud Storage.
C. Data can only be exported in JSON or Avro format.
D. The only compression option available is GZIP.
Answer: C (LEAVE A REPLY)
Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported.

NEW QUESTION: 79
You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)
A. Change the zone of your Cloud Dataflow pipeline to run in us-central1
B. Increase the number of max workers
C. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery
D. Use a larger instance type for your Cloud Dataflow workers
E. Create a temporary table in Cloud Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery
Answer: B,D (LEAVE A REPLY)

NEW QUESTION: 80

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

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

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

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

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

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

Answer: (SHOW ANSWER)

NEW QUESTION: 81

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 (LEAVE A REPLY)

Until or unless the message is not acknowledged within defined ack window period for every message, we will get duplicate (number of retries to send message can be defined).
https://cloud.google.com/pubsub/docs/troubleshooting#dupes

NEW QUESTION: 82

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 (LEAVE A REPLY)
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.

NEW QUESTION: 83
You need to create a new transaction table in Cloud Spanner that stores product sales data. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?
A. The current epoch time
B. A concatenation of the product name and the current epoch time
C. A random universally unique identifier number (version 4 UUID)
D. The original order identification number from the sales system, which is a monotonically increasing integer
Answer: C (LEAVE A REPLY)
Explanation/Reference: https://www.uuidgenerator.net/version4

NEW QUESTION: 84
What is the general recommendation when designing your row keys for a Cloud Bigtable schema?
A. Include multiple time series values within the row key
B. Keep the row keep as an 8 bit integer
C. Keep your row key reasonably short
D. Keep your row key as long as the field permits
Answer: C (LEAVE A REPLY)
Explanation
A general guide is to, keep your row keys reasonably short. Long row keys take up additional memory and storage and increase the time it takes to get responses from the Cloud Bigtable server.
Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

NEW QUESTION: 85
What are two of the benefits of using denormalized data structures in BigQuery?
A. Reduces the amount of data processed, reduces the amount of storage required
B. Increases query speed, makes queries simpler
C. Reduces the amount of storage required, increases query speed
D. Reduces the amount of data processed, increases query speed
Answer: B (LEAVE A REPLY)

Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINs on large tables, but with a denormalized data structure, you don't have to use JOINs, since all of the data has been combined into one table. Denormalization also makes queries simpler because you do not have to use JOIN clauses. Denormalization increases the amount of data processed and the amount of storage required because it creates redundant data.
https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

NEW QUESTION: 86
You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?
A. Cloud Composer
B. Workflow Templates on Cloud Dataproc
C. Cloud Scheduler
D. cron
Answer: B (LEAVE A REPLY)

NEW QUESTION: 87
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. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer

Service to transfer on-premices data to the designated storage bucket.
C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in
parallel.
D. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
E. 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.
Answer: B,C (LEAVE A REPLY)

NEW QUESTION: 88
You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline.
Which component will be used for the data processing operation?
A. PCollection
B. Transform
C. Pipeline
D. Sink API
Answer: (SHOW ANSWER)
Explanation
In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.
Reference: https://cloud.google.com/dataflow/model/programming-model

NEW QUESTION: 89
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. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery
B. 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
C. Load the data every 30 minutes into a new partitioned table in BigQuery.
D. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

Answer: C (LEAVE A REPLY)

NEW QUESTION: 90
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. Grant the Project Owner role to a service account, and run the job with it
B. Restrict the Google Cloud Storage bucket so only you can see the files
C. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files
and write to BigQuery
D. Use a service account with the ability to read the batch files and to write to BigQuery
Answer: A (LEAVE A REPLY)

NEW QUESTION: 91
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: B,C,D (LEAVE A REPLY)
Fraud is not a feature, so unsupervised, location is given so supervised, Clustering can be done looking at the done with same features.

NEW QUESTION: 92
Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set. You want to increase the AUC of the model. What should you do?
A. Perform hyperparameter tuning
B. Train a classifier with deep neural networks, because neural networks would always beat SVMs
C. Deploy the model and measure the real-world AUC; it's always higher because of generalization
D. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC
Answer: A (LEAVE A REPLY)
https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568

NEW QUESTION: 93
You are designing a cloud-native historical data processing system to meet the following conditions:
* The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.
* A streaming data pipeline stores new data daily.
* Peformance is not a factor in the solution.
* The solution design should maximize availability.
How should you design data storage for this solution?
A. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
B. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.
C. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.
D. Store the data in BigQuery. Access the data using the BigQuery Connector on Cloud Dataproc and Compute Engine.
Answer: (SHOW ANSWER)

NEW QUESTION: 94
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 Hadoop cluster on Google Compute Engine that uses Local SSD disks.

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

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

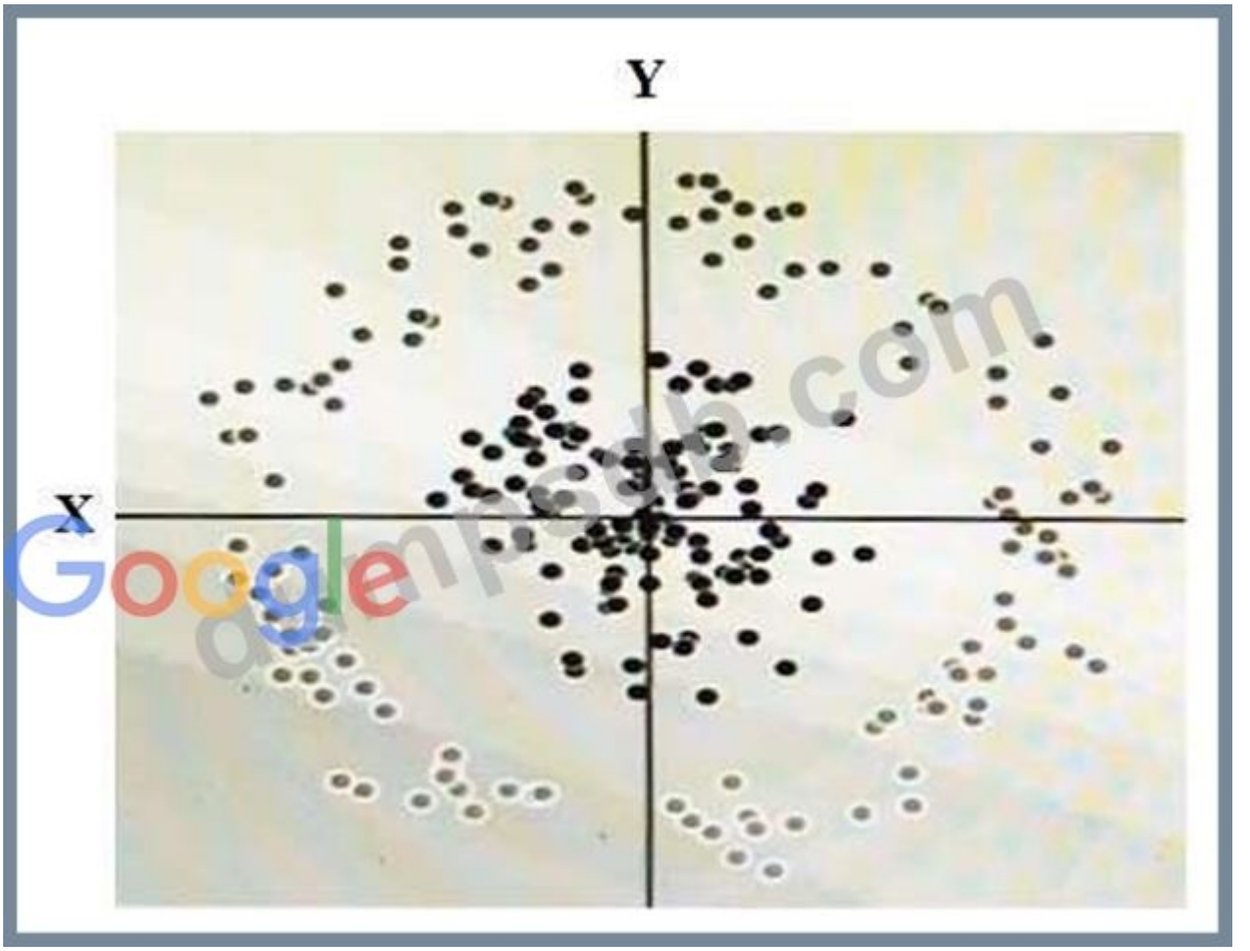
D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

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

Answer: D (LEAVE A REPLY)

NEW QUESTION: 95

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. X^2

B. X^2+Y^2

C. cos(X)

D. Y^2

Answer: (SHOW ANSWER)

NEW QUESTION: 96

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. MongoDB
B. HDFS with Hive
C. Redis
D. Cassandra
E. HBase
F. MySQL
Answer: A,B,E (LEAVE A REPLY)

NEW QUESTION: 97
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: A,D,F (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 98
Which of the following are examples of hyperparameters? (Select 2 answers.)
A. Number of hidden layers
B. Number of nodes in each hidden layer
C. Biases
D. Weights
Answer: A,B (LEAVE A REPLY)
If model parameters are variables that get adjusted by training with existing data, your hyperparameters are the variables about the training process itself. For example, part of setting up a deep neural network is deciding how many "hidden" layers of nodes to use between the input layer and the output layer, as well as how many nodes each layer should use. These variables are not directly related to the training data at all.
They are configuration variables. Another difference is that parameters change during a training job, while the hyperparameters are usually constant during a job.

Weights and biases are variables that get adjusted during the training process, so they are not hyperparameters.
Reference: https://cloud.google.com/ml-engine/docs/hyperparameter-tuning-overview

NEW QUESTION: 99
Case Study: 2 - MJTelco
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:
Which approach meets the requirements?
A. 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.
B. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
C. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 100
You architect a system to analyze seismic dat
a. 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. 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.
B. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
C. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
D. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
Answer: (SHOW ANSWER)

NEW QUESTION: 101
Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries 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. Google Cloud Datastore
B. Google Cloud Storage
C. Cloud Bigtable
D. Google BigQuery
Answer: C (LEAVE A REPLY)

NEW QUESTION: 102
When a Cloud Bigtable node fails, ____ is lost.
A. all data
B. no data
C. the last transaction
D. the time dimension
Answer: B (LEAVE A REPLY)
Explanation
A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.
Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:
Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud Bigtable simply updates the pointers for each node.
Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.

When a Cloud Bigtable node fails, no data is lost
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION: 103
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. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.
C. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
D. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
Answer: (SHOW ANSWER)

NEW QUESTION: 104
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: (SHOW ANSWER)
Using Hash values we can remove duplicate values from a database. Hashvalues will be same for duplicate data and thus can be easily rejected.

NEW QUESTION: 105
Which of the following are examples of hyperparameters? (Select 2 answers.)
A. Number of hidden layers
B. Number of nodes in each hidden layer
C. Biases
D. Weights
Answer: A,B (LEAVE A REPLY)
If model parameters are variables that get adjusted by training with existing data, your hyperparameters are the variables about the training process itself. For example, part of setting up a deep neural network is deciding how many "hidden" layers of nodes to use between the input layer and the output layer, as well as how many nodes each layer should use. These variables are not directly related to the training data at all. They are configuration variables.

Another difference is that parameters change during a training job, while the hyperparameters are usually constant during a job.
Weights and biases are variables that get adjusted during the training process, so they are not hyperparameters.
Reference: https://cloud.google.com/ml-engine/docs/hyperparameter-tuning-overview

NEW QUESTION: 106
Which of these operations can you perform from the BigQuery Web UI?
A. Upload a file in SQL format.
B. Load data with nested and repeated fields.
C. Upload a 20 MB file.
D. Upload multiple files using a wildcard.
Answer: B (LEAVE A REPLY)
Explanation
You can load data with nested and repeated fields using the Web UI.
You cannot use the Web UI to:
- Upload a file greater than 10 MB in size
- Upload multiple files at the same time
- Upload a file in SQL format
All three of the above operations can be performed using the "bq" command.
Reference: https://cloud.google.com/bigquery/loading-data

NEW QUESTION: 107
When a Cloud Bigtable node fails, ____ is lost.
A. all data
B. no data
C. the last transaction
D. the time dimension
Answer: (SHOW ANSWER)
A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format.

Each tablet is associated with a specific Cloud Bigtable node. Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:
Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud Bigtable simply updates the pointers for each node. Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.
When a Cloud Bigtable node fails, no data is lost
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION: 108
You are updating the code for a subscriber to a Put/Sub feed. You are concerned that upon deployment the subscriber may erroneously acknowledge messages, leading to message loss. You subscriber is not set up to retain acknowledged messages. What should you do to ensure that you can recover from errors after deployment?
A. Set up the Pub/Sub emulator on your local machine Validate the behavior of your new subscriber togs before deploying it to production
B. Enable dead-lettering on the Pub/Sub topic to capture messages that aren't successful acknowledged if an error occurs after deployment, re-deliver any messages captured by the dead-letter queue
C. Use Cloud Build for your deployment if an error occurs after deployment, use a Seek operation to locate a tmestamp logged by Cloud Build at the start of the deployment
D. Create a Pub/Sub snapshot before deploying new subscriber code. Use a Seek operation to re-deliver messages that became available after the snapshot was created
Answer: D (LEAVE A REPLY)

NEW QUESTION: 109
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 of both the Google BigQuery TableSchema and TableFieldSchema classes.
C. Use .fromQuery operation to read specific fields from the table.

D. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.
Answer: (SHOW ANSWER)

NEW QUESTION: 110
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. Use the diagnosecommand to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
B. 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.
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 Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 111
You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices globally. Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?
A. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
B. Edge TPUs as sensor devices for storing and transmitting the messages.
C. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.
D. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
Answer: (SHOW ANSWER)

NEW QUESTION: 112
You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for BigQuery with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account.
What should you do?

A. Convert your batch BQ queries into interactive BQ queries.
B. Create an additional project to overcome the 2K on-demand per-project quota.
C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.
D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.
Answer: C (LEAVE A REPLY)
Explanation
Explanation/Reference:
Reference https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery

NEW QUESTION: 113
You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing low and ensure that the data is current. Which solution should you choose?
A. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.
B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.
C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.
D. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.
Answer: B (LEAVE A REPLY)
Explanation

NEW QUESTION: 114
You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:
* Decoupling producer from consumer
* Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
* Near real-time SQL query
* Maintain at least 2 years of historical data, which will be queried with SQL Which pipeline should you use to meet these requirements?
A. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.
B. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
C. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.

D. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 115
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 dat
a. 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 service account for the ODBC connection to use for authentication
D. Create a new view over events_partitioned using standard SQL
E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"
Answer: A,E (LEAVE A REPLY)

NEW QUESTION: 116
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: (SHOW ANSWER)
Spanner allows transaction tables to scale horizontally and secondary indexes for range queries.

NEW QUESTION: 117
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?
A. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
C. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table.
Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
D. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
Answer: (SHOW ANSWER)

NEW QUESTION: 118
You are migrating a table to BigQuery and are deeding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and state in which the store is located You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state city and individual store. You want to model this table to minimize query time and cost. What should you do?
A. Top-level cluster by store ID first, then city then state.
B. Partition by transaction time; cluster by state first, then city then store ID
C. Top-level cluster by stale first, then city then store
D. Partition by transaction tome cluster by store ID first, then city, then stale
Answer: (SHOW ANSWER)

NEW QUESTION: 119
What are two of the characteristics of using online prediction rather than batch prediction?
A. It is optimized to handle a high volume of data instances in a job and to run more complex models.
B. Predictions are returned in the response message.
C. Predictions are written to output files in a Cloud Storage location that you specify.
D. It is optimized to minimize the latency of serving predictions.
Answer: B,D (LEAVE A REPLY)
Explanation
Online prediction
Optimized to minimize the latency of serving predictions.
Predictions returned in the response message.
Batch prediction
Optimized to handle a high volume of instances in a job and to run more complex models.
Predictions written to output files in a Cloud Storage location that you specify.
Reference:
https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

NEW QUESTION: 120
As your organization expands its usage of GCP, many teams have started to create their own projects.
Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects.
Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies.
Which two steps should you take? Choose 2 answers.
A. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.
B. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
C. Introduce resource hierarchy to leverage access control policy inheritance.
D. Use Cloud Deployment Manager to automate access provision.
E. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
Answer: D,E (LEAVE A REPLY)

NEW QUESTION: 121
You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?
A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.
D. Use Cloud Dataflow to write summary of each day's stock trades to an Avro file on Cloud Storage.
Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 122

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

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

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

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

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

Answer: D (LEAVE A REPLY)

NEW QUESTION: 123
You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure.
What should you do?
A. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
B. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.
C. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.
D. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 124
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 (LEAVE A REPLY)
https://cloud.google.com/compute/docs/disks/customer-managed-encryption

NEW QUESTION: 125
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 sliding time window with a duration of 60 minutes.
B. Use a fixed-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: (SHOW ANSWER)

NEW QUESTION: 126
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: A (LEAVE A REPLY)

NEW QUESTION: 127
Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?
A. Preemptible workers cannot use persistent disk.
B. Preemptible workers cannot store data.
C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.
D. A Dataproc cluster cannot have only preemptible workers.
Answer: (SHOW ANSWER)

The following rules will apply when you use preemptible workers with a Cloud Dataproc cluster:
. Processing only--Since preemptibles can be reclaimed at any time, preemptible workers do not store data. Preemptibles added to a Cloud Dataproc cluster only function as processing nodes. . No preemptible-only clusters--To ensure clusters do not lose all workers, Cloud Dataproc cannot create preemptible-only clusters.
. Persistent disk size--As a default, all preemptible workers are created with the smaller of 100GB or the primary worker boot disk size. This disk space is used for local caching of data and is not available through HDFS.
The managed group automatically re-adds workers lost due to reclamation as capacity permits.
Reference: https://cloud.google.com/dataproc/docs/concepts/preemptible-vms

NEW QUESTION: 128
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 Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
B. Send the data to Google Cloud Datastore and then export to BigQuery.
C. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
D. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.
Answer: (SHOW ANSWER)

NEW QUESTION: 129
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: data_point
Column data: device_id,date
C. Rowkey: date#data_point
Column data: device_id
D. Rowkey: device_id
Column data: date, data_point
E. Rowkey: date
Column data: device_id,data_point
Answer: B (LEAVE A REPLY)

NEW QUESTION: 130
You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?
A. Dialogflow Enterprise Edition
B. Cloud Speech-to-Text API
C. Cloud Natural Language API
D. Cloud AutoML Natural Language
Answer: A (LEAVE A REPLY)

NEW QUESTION: 131
You have an Apache Kafka cluster on-prem with topics containing web application logs. You need to replicate the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is mirroring to avoid deployment of Kafka Connect plugins.
What should you do?
A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
B. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.

C. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connector. Use a Dataflow job to read from PubSub and write to GCS.
D. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector. Use a Dataflow job to read from PubSub and write to GCS.
Answer: A (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 132
You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?
A. Use gsutil cp -J to compress the content being uploaded to Cloud Storage
B. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic
C. Use Transfer Appliance to copy the data to Cloud Storage
D. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
Answer: C (LEAVE A REPLY)

NEW QUESTION: 133
Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?
A. A sequential numeric ID
B. A timestamp followed by a stock symbol
C. A non-sequential numeric ID
D. A stock symbol followed by a timestamp
Answer: A,B (LEAVE A REPLY)
using a timestamp as the first element of a row key can cause a variety of problems.
In brief, when a row key for a time series includes a timestamp, all of your writes will target a single node; fill that node; and then move onto the next node in the cluster, resulting in hotspotting.
Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, since new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes. [https://cloud.google.com/bigtable/docs/schema-design] Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotting

NEW QUESTION: 134
You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU. Currently, your pipeline is located in

europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

A. Increase the number of max workers

B. Use a larger instance type for your Cloud Dataflow workers

C. Change the zone of your Cloud Dataflow pipeline to run in us-central1

D. Create a temporary table in Cloud Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery

E. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

Answer: (SHOW ANSWER)

Explanation/Reference:

NEW QUESTION: 135

Dataproc clusters contain many configuration files. To update these files, you will need to use the --properties option. The format for the option is: file_prefix:property=_____.

A. details

B. value

C. null

D. id

Answer: B (LEAVE A REPLY)

To make updating files and properties easy, the --properties command uses a special format to specify the configuration file and the property and value within the file that should be updated. The formatting is as follows: file_prefix:property=value.

NEW QUESTION: 136

Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer dat

a. 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 size of the Google Persistent Disk on your server.

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

C. Increase the CPU size on your server.

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

Answer: (SHOW ANSWER)

NEW QUESTION: 137
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 eventsdata 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_partitioned using standard SQL
B. Create a new partitioned table using a standard SQL query
C. Create a service account for the ODBC connection to use for authentication
D. Create a new view over events using standard SQL
E. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection
and shared "events"
Answer: D,E (LEAVE A REPLY)

NEW QUESTION: 138
You work for an advertising company, and you've developed a Spark ML model to predict click-through rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?
A. Use Cloud ML Engine for training existing Spark ML models
B. Rewrite your models on TensorFlow, and start using Cloud ML Engine
C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery

D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery
Answer: C (LEAVE A REPLY)
https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml

NEW QUESTION: 139
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 a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset
B. Create groups for your users and give those groups access to the dataset
C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset
D. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
Answer: (SHOW ANSWER)

NEW QUESTION: 140
A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions.
You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features). How should you create the ML pipeline?
A. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.
B. Create a Cloud Dataflow pipeline using BigQueryIOto read predictions for all users from the query. Write the results to Cloud Bigtable using BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.
C. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.
D. Create a Cloud Dataflow pipeline using BigQueryIOto read results from the query. Grant the Dataflow Worker role to the application service account.
Answer: (SHOW ANSWER)

NEW QUESTION: 141

You are collecting IoT sensor data from millions of devices across the world and storing the data in BigQuery. Your access pattern is based on recent data tittered by location_id and device_version with the following query:

You want to optimize your queries for cost and performance. How should you structure your data?

A. Partition table data by create_date cluster table data by tocation_id and device_version

B. Partition table data by create_date, location_id and device_version

C. Cluster table data by create_date, partition by location and device_version

D. Cluster table data by create_date location_id and device_version

Answer: (SHOW ANSWER)

NEW QUESTION: 142

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

What should you do?

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

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

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

D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API.

Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 143

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 your network bandwidth from Compute Engine to Cloud Storage.
C. Increase your network bandwidth from your datacenter to GCP.
D. Increase the size of the Google Persistent Disk on your server.
Answer: (SHOW ANSWER)

NEW QUESTION: 144
You work for a large financial institution that is planning to use Dialogflow to create a chatbot for the company's mobile app You have reviewed old chat logs and lagged each conversation for intent based on each customer's stated intention for contacting customer service About 70% of customer requests are simple requests that are solved within 10 intents The remaining 30% of inquiries require much longer, more complicated requests Which intents should you automate first?
A. Automate the 10 intents that cover 70% of the requests so that live agents can handle more complicated requests
B. Automate a blend of the shortest and longest intents to be representative of all intents
C. Automate the more complicated requests first because those require more of the agents' time
D. Automate intents in places where common words such as "payment" appear only once so the software isn't confused
Answer: A (LEAVE A REPLY)

NEW QUESTION: 145
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 Local SSD

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

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

D. Cloud Dataflow, Cloud SQL, and Cloud Storage

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

Answer: (SHOW ANSWER)

NEW QUESTION: 146

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 (LEAVE A REPLY)
Explanation
Reference
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505
Topic 1, 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
* 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.

NEW QUESTION: 147

You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?
A. Use the MERGE statement to apply updates in batch every 60 seconds.
B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
C. Use the INSERT statement to insert a batch of data every 60 seconds.
D. Use bq load to load a batch of sensor data every 60 seconds.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 148
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: (SHOW ANSWER)
Forecasting and Liner regression is used for predicting housing price.

NEW QUESTION: 149
You need to choose a database for a new project that has the following requirements:
* Fully managed
* Able to automatically scale up
* Transactionally consistent
* Able to scale up to 6 TB
* Able to be queried using SQL
Which database do you choose?
A. Cloud SQL
B. Cloud Bigtable
C. Cloud Spanner
D. Cloud Datastore
Answer: C (LEAVE A REPLY)
https://cloud.google.com/products/databases

NEW QUESTION: 150
Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?
A. Cloud Dataflow
B. Cloud Composer
C. Cloud Dataprep

D. Cloud Dataproc
Answer: B (LEAVE A REPLY)
Cloud Composer uses airflow which is open source and can help to orchestrate jobs.

NEW QUESTION: 151
You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:
* Decoupling producer from consumer
* Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
* Near real-time SQL query
* Maintain at least 2 years of historical data, which will be queried with SQL Which pipeline should you use to meet these requirements?
A. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
B. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
C. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 152
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. MongoDB
B. HDFS with Hive
C. Cassandra
D. MySQL
E. HBase
F. Redis
Answer: A,B,E (LEAVE A REPLY)

NEW QUESTION: 153
Flowlogistic Case Study
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market.
Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.
Solution Concept
Flowlogistic wants to implement two concepts using the cloud:
Use their proprietary technology in a real-time inventory-tracking system that indicates the location of
- their loads
Perform analytics on all their orders and shipment logs, which contain both structured and unstructured
- data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.
Existing Technical Environment
Flowlogistic architecture resides in a single data center:
Databases
- 8 physical servers in 2 clusters
- SQL Server - user data, inventory, static data
- 3 physical servers
- Cassandra - metadata, tracking messages
10 Kafka servers - tracking message aggregation and batch insert

Application servers - customer front end, middleware for order/customs

- 60 virtual machines across 20 physical servers
- Tomcat - Java services
- Nginx - static content
- Batch servers

Storage appliances

- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage

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

10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements

Build a reliable and reproducible environment with scaled panty of production.

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

Technical Requirements

Handle both streaming and batch data

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

SEO Statement

We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

We need to organize our information so we can more easily understand where our customers are and what they are shipping.

CTO Statement

IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.
CFO Statement
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability.
Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?
A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage
E. Cloud Dataflow, Cloud SQL, and Cloud Storage
Answer: C (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 154
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. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.
B. Change the processing job to use Google Cloud Dataproc instead.
C. Manually start the Cloud Dataflow job each morning when you get into the office.
D. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 155
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 (LEAVE A REPLY)
Explanation
https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

NEW QUESTION: 156
When a Cloud Bigtable node fails, ____ is lost.
A. all data
B. no data
C. the last transaction
D. the time dimension
Answer: (SHOW ANSWER)
A Cloud Bigtable table is sharded into blocks of contiguous rows, called tablets, to help balance the workload of queries. Tablets are stored on Colossus, Google's file system, in SSTable format. Each tablet is associated with a specific Cloud Bigtable node.
Data is never stored in Cloud Bigtable nodes themselves; each node has pointers to a set of tablets that are stored on Colossus. As a result:
Rebalancing tablets from one node to another is very fast, because the actual data is not copied. Cloud Bigtable simply updates the pointers for each node.
Recovery from the failure of a Cloud Bigtable node is very fast, because only metadata needs to be migrated to the replacement node.
When a Cloud Bigtable node fails, no data is lost

NEW QUESTION: 157
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 (LEAVE A REPLY)
The recommendation requires filtering based on several TB of data, therefore BigTable is the recommended option vs Cloud SQL which is limited to 10TB.

NEW QUESTION: 158
When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through a ____ proxy.
A. HTTPS
B. VPN
C. SOCKS
D. HTTP
Answer: C (LEAVE A REPLY)
Explanation
When using Cloud Dataproc clusters, configure your browser to use the SOCKS proxy. The SOCKS proxy routes data intended for the Cloud Dataproc cluster through an SSH tunnel.
Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web-interfaces#interfaces

NEW QUESTION: 159
Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:
* The ability to seek to a particular offset in a topic, possibly back to the start of all data ever captured
* Support for publish/subscribe semantics on hundreds of topics
* Retain per-key ordering
Which system should you choose?
A. Apache Kafka
B. Cloud Storage
C. Cloud Pub/Sub
D. Firebase Cloud Messaging
Answer: A (LEAVE A REPLY)
These are the functionalities which are currently lagging/not-available with Pub/Sub.

NEW QUESTION: 160
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. Your custom endpoint has an out-of-date SSL certificate.
B. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.
C. The message body for the sensor event is too large.
D. The Cloud Pub/Sub topic has too many messages published to it.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 161
When you design a Google Cloud Bigtable schema it is recommended that you _____.
A. Avoid schema designs that are based on NoSQL concepts
B. Create schema designs that are based on a relational database design
C. Avoid schema designs that require atomicity across rows
D. Create schema designs that require atomicity across rows
Answer: (SHOW ANSWER)
All operations are atomic at the row level. For example, if you update two rows in a table, it's possible that one row will be updated successfully and the other update will fail. Avoid schema designs that require atomicity across rows.
Reference: https://cloud.google.com/bigtable/docs/schema-design#row-keys

NEW QUESTION: 162
Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data).
What should you do?
A. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
B. Add a try... catch block to your DoFn that transforms the data, extract erroneous rows from logs.
C. Add a try... catch block to your DoFn that transforms the data, use a sideOutput to create a PCollection that can be stored to PubSub later.
D. Add a try... catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.
Answer: (SHOW ANSWER)

NEW QUESTION: 163
Which of these is NOT a way to customize the software on Dataproc cluster instances?

A. Set initialization actions
B. Modify configuration files using cluster properties
C. Configure the cluster using Cloud Deployment Manager
D. Log into the master node and make changes from there
Answer: (SHOW ANSWER)
You can access the master node of the cluster by clicking the SSH button next to it in the Cloud Console.
You can easily use the --properties option of the dataproc command in the Google Cloud SDK to modify many common configuration files when creating a cluster. When creating a Cloud Dataproc cluster, you can specify initialization actions in executables and/or scripts that Cloud Dataproc will run on all nodes in your Cloud Dataproc cluster immediately after the cluster is set up. [https://cloud.google.com/dataproc/ docs/concepts/configuring-clusters/init-actions]
Reference: https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/cluster-properties

NEW QUESTION: 164
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than
50,000 installations.
* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

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

Business Requirements

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

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data

Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.

CEO Statement

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

CTO Statement

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure.

We also need environments in which our data scientists can carefully study and quickly adapt our models.

Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.

CFO Statement

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

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

A. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
B. Create a partitioned table called tracking_table and include a TIMESTAMP column.
C. Create a table called tracking_table with a TIMESTAMP column to represent the day.
D. Create a table called tracking_table and include a DATE column.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 165
You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time.
Which two methods can accomplish this? Choose 2 answers.
A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
B. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.
C. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
D. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
E. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.
Answer: B,D (LEAVE A REPLY)

NEW QUESTION: 166
You want to rebuild your batch pipeline for structured data on Google Cloud You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax You have already moved your raw data into Cloud Storage How should you build the pipeline on Google Cloud while meeting speed and processing requirements?
A. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery
B. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table
C. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated queries from BigQuery for machine learning.

D. Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipeline on Dataproc to write the data into BigQuery
Answer: (SHOW ANSWER)

NEW QUESTION: 167
Which of these statements about BigQuery caching is true?
A. By default, a query's results are not cached.
B. BigQuery caches query results for 48 hours.
C. Query results are cached even if you specify a destination table.
D. There is no charge for a query that retrieves its results from cache.
Answer: (SHOW ANSWER)
When query results are retrieved from a cached results table, you are not charged for the query.
BigQuery caches query results for 24 hours, not 48 hours.
Query results are not cached if you specify a destination table.
A query's results are always cached except under certain conditions, such as if you specify a destination table.

NEW QUESTION: 168
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: D (LEAVE A REPLY)
Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

NEW QUESTION: 169
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. Unsupervised learning to predict the location of a transaction.
B. Supervised learning to determine which transactions are most likely to be fraudulent.
C. Unsupervised learning to determine which transactions are most likely to be fraudulent.
D. Clustering to divide the transactions into N categories based on feature similarity.
E. Reinforcement learning to predict the location of a transaction.
F. Supervised learning to predict the location of a transaction.
Answer: C,D,E (LEAVE A REPLY)

NEW QUESTION: 170
MJTelco is building a custom interface to share dat
a. They have these requirements:
They need to do aggregations over their petabyte-scale datasets.
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. BigQuery and Cloud Bigtable
B. Cloud Bigtable and Cloud SQL
C. Cloud Datastore and Cloud Bigtable
D. BigQuery and Cloud Storage
Answer: (SHOW ANSWER)

NEW QUESTION: 171
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 the Google Cloud Billing API to see what account the warehouse is being billed to.
B. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
C. Use Google Stackdriver Audit Logs to review data access.
D. Get the identity and access management IIAM) policy of each table
Answer: (SHOW ANSWER)

NEW QUESTION: 172
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. 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.
E. 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.
Answer: E (LEAVE A REPLY)

NEW QUESTION: 173
You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time. Consumers will receive the data in the following ways:
Real-time event stream
ANSI SQL access to real-time stream and historical data
Batch historical exports
Which solution should you use?
A. Cloud Pub/Sub, Cloud Storage, BigQuery
B. Cloud Dataproc, Cloud Dataflow, BigQuery
C. Cloud Dataflow, Cloud SQL, Cloud Spanner
D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL
Answer: C (LEAVE A REPLY)

NEW QUESTION: 174
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: C (LEAVE A REPLY)
Explanation

NEW QUESTION: 175
You work for a mid-sized enterprise that needs to move its operational system transaction data from an on- premises database to GCP. The database is about 20 TB in size. Which database should you choose?
A. Cloud SQL
B. Cloud Bigtable
C. Cloud Spanner
D. Cloud Datastore
Answer: A (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 176
To give a user read permission for only the first three columns of a table, which access control method would you use?
A. Primitive role
B. Predefined role
C. Authorized view
D. It's not possible to give access to only the first three columns of a table.
Answer: C (LEAVE A REPLY)
An authorized view allows you to share query results with particular users and groups without giving them read access to the underlying tables. Authorized views can only be created in a dataset that does not contain the tables queried by the view.
When you create an authorized view, you use the view's SQL query to restrict access to only the rows and columns you want the users to see.
Reference: https://cloud.google.com/bigquery/docs/views#authorized-views

NEW QUESTION: 177
Which TensorFlow function can you use to configure a categorical column if you don't know all of the possible values for that column?
A. categorical_column_with_vocabulary_list
B. categorical_column_with_hash_bucket

C. categorical_column_with_unknown_values
D. sparse_column_with_keys
Answer: B (LEAVE A REPLY)
If you know the set of all possible feature values of a column and there are only a few of them, you can use categorical_column_with_vocabulary_list. Each key in the list will get assigned an auto-incremental ID starting from 0.
What if we don't know the set of possible values in advance? Not a problem. We can use categorical_column_with_hash_bucket instead. What will happen is that each possible value in the feature column occupation will be hashed to an integer ID as we encounter them in training.
Reference: https://www.tensorflow.org/tutorials/wide