Google.Professional-Data-Engineer.v2022-08-11.q226

Exam Code:	Professional-Data-Engineer
Exam Name:	Google Certified Professional Data Engineer Exam
Certification Provider:	Google
Free Question Number:	226
Version:	v2022-08-11
# of views:	2380
# of Questions views:	4871
https://www.dumpsfiles.com/files/Google/Professional-Data-Engineer/Google.Professional-Data-Engineer.v2022-08-11.q226	

NEW QUESTION: 1
You work for a global shipping company. You want to train a model on 40 TB of data to predict which ships in each geographic region are likely to cause delivery delays on any given day. The model will be based on multiple attributes collected from multiple sources. Telemetry data, including location in GeoJSON format, will be pulled from each ship and loaded every hour. You want to have a dashboard that shows how many and which ships are likely to cause delays within a region. You want to use a storage solution that has native functionality for prediction and geospatial processing. Which storage solution should you use?
A. Cloud Bigtable
B. Cloud Datastore
C. Cloud SQL for PostgreSQL
D. BigQuery
Answer: D (LEAVE A REPLY)

NEW QUESTION: 2
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

NEW QUESTION: 3
You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data warehouse-style analytics on the data in Google Cloud and expose the dataset as files for batch analysis tools in other cloud providers. What should you do?
A. Store the warm data as files in Cloud Storage, and store the active data in BigQuery. Keep this ratio as 80% warm and 20% active.
B. Store and process the entire dataset in BigQuery.
C. Store and process the entire dataset in Cloud Bigtable.
D. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.
Answer: (SHOW ANSWER)

NEW QUESTION: 4
How can you get a neural network to learn about relationships between categories in a categorical feature?
A. Create a multi-hot column
B. Create a one-hot column
C. Create a hash bucket
D. Create an embedding column
Answer: D (LEAVE A REPLY)
Explanation
There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead of having just one value, like a continuous feature, it has many values, or dimensions. This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn't encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.
Both of these problems can be solved by representing a categorical feature with an embedding column. The idea is that each category has a smaller vector with, let's say, 5 values in it. But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic features in a neural network. The difference is that each category has a set of weights (5 of them in this case).
You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.
Reference:

https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/a-wide-and-dee

NEW QUESTION: 5
Google Cloud Bigtable indexes a single value in each row. This value is called the _____.
A. primary key
B. unique key
C. row key
D. master key
Answer: (SHOW ANSWER)
Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key.
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION: 6
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.
Reference: https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

NEW QUESTION: 7
Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data. How should you set up the log data transfer into Google Cloud?
A. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

B. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
C. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.
D. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
Answer: (SHOW ANSWER)

NEW QUESTION: 8
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. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
B. Introduce resource hierarchy to leverage access control policy inheritance.
C. Use Cloud Deployment Manager to automate access provision.
D. 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.
E. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
Answer: A,C (LEAVE A REPLY)

NEW QUESTION: 9
Which of these are examples of a value in a sparse vector? (Select 2 answers.)
A. [0, 5, 0, 0, 0, 0]
B. [0, 0, 0, 1, 0, 0, 1]
C. [0, 1]
D. [1, 0, 0, 0, 0, 0, 0]
Answer: C,D (LEAVE A REPLY)
Categorical features in linear models are typically translated into a sparse vector in which each possible value has a corresponding index or id. For example, if there are only three possible eye colors you can represent 'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and 'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words).

[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1.
[0, 5, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain
0 s and 1s.
Reference:
https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION: 10
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 a blend of the shortest and longest intents to be representative of all intents
B. Automate the more complicated requests first because those require more of the agents' time
C. Automate the 10 intents that cover 70% of the requests so that live agents can handle more complicated requests
D. Automate intents in places where common words such as "payment" appear only once so the software isn't confused
Answer: C (LEAVE A REPLY)

NEW QUESTION: 11
You work for a manufacturing plant that batches application log files together into a single log file once a day at
2:00 AM. You have written a Google Cloud Dataflow job to process that log file. You need to make sure the log file in processed once per day as inexpensively as possible. What should you do?
A. Change the processing job to use Google Cloud Dataproc instead.
B. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.
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: 12
Your company is using WILDCARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:
Syntax error : Expected end of statement but got "-" at [4:11]
SELECT age
FROM

bigquery-public-data.noaa_gsod.gsod
WHERE
age != 99
AND_TABLE_SUFFIX = '1929'
ORDER BY
age DESC
Which table name will make the SQL statement work correctly?
A. bigquery-public-data.noaa_gsod.gsod*
B. 'bigquery-public-data.noaa_gsod.gsod*`
C. 'bigquery-public-data.noaa_gsod.gsod'
D. 'bigquery-public-data.noaa_gsod.gsod'*
Answer: (SHOW ANSWER)

NEW QUESTION: 13
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?
A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
B. Use Cloud Dataproc to run your transformations. Use the diagnosecommand to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs.
Configure the job to use non-default Compute Engine machine types when needed.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 14
You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes.
You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload.
What should you do?
A. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
B. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.
C. Export Bigtable dump to GCS and run your analytical job on top of the exported files.

D. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
Answer: (SHOW ANSWER)

NEW QUESTION: 15
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 without any additional adjustment to your code.
B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.
C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.
D. Stay on CPUs, and increase the size of the cluster you're training your model on.
Answer: (SHOW ANSWER)
Cloud TPUs are not suited to the following workloads: [...] Neural network workloads that contain custom TensorFlow operations written in C++. Specifically, custom operations in the body of the main training loop are not suitable for TPUs.

NEW QUESTION: 16
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: device_id
Column data: date, data_point
B. Rowkey: date#device_id
Column data: data_point
C. Rowkey: date#data_point
Column data: device_id
D. Rowkey: date
Column data: device_id,data_point
E. Rowkey: data_point
Column data: device_id,date
Answer: E (LEAVE A REPLY)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 17
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 Spanner
D. Cloud SQL
Answer: C (LEAVE A REPLY)

NEW QUESTION: 18

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt.
You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?
A. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.
B. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
C. Add a SideInput that returns a Boolean if the element is corrupt.
D. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 19
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)
Explanation
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.
Reference: https://cloud.google.com/bigquery/querying-data#query-caching

NEW QUESTION: 20
You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:
* Executing the transformations on a schedule
* Enabling non-developer analysts to modify transformations
* Providing a graphical tool for designing transformations
What should you do?
A. Load each month's CSV data into BigQuery, and write a SQL query to transform the data to a standard schema. Merge the transformed tables together with a SQL query
B. Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe.
Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

C. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformation. The Python code should be stored in a revision control system and modified as the incoming data's schema changes
D. Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis
Answer: D (LEAVE A REPLY)

NEW QUESTION: 21
An organization maintains a Google BigQuery dataset that contains tables with user-level data. They want
to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-
level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for
other projects is assigned to those projects. What should they do?
A. Create and share an authorized view that provides the aggregate results.
B. Create and share a new dataset and view that provides the aggregate results.
C. Create and share a new dataset and table that contains the aggregate results.
D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.
Answer: D (LEAVE A REPLY)
Explanation/Reference:
Reference: https://cloud.google.com/bigquery/docs/access-control

NEW QUESTION: 22
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: 23
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. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC
D. Deploy the model and measure the real-world AUC; it's always higher because of generalization
Answer: C (LEAVE A REPLY)

NEW QUESTION: 24
What are the minimum permissions needed for a service account used with Google Dataproc?
A. Execute to Google Cloud Storage; write to Google Cloud Logging
B. Write to Google Cloud Storage; read to Google Cloud Logging
C. Execute to Google Cloud Storage; execute to Google Cloud Logging
D. Read and write to Google Cloud Storage; write to Google Cloud Logging
Answer: D (LEAVE A REPLY)
Service accounts authenticate applications running on your virtual machine instances to other Google Cloud Platform services. For example, if you write an application that reads and writes files on Google Cloud Storage, it must first authenticate to the Google Cloud Storage API. At a minimum, service accounts used with Cloud Dataproc need permissions to read and write to Google Cloud Storage, and to write to Google Cloud Logging.
Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

NEW QUESTION: 25
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/Reference:

Reference https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery

NEW QUESTION: 26
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)

NEW QUESTION: 27
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 >#<sensorid>#<timestamp>.
B. Use a row key of the form <timestamp>#<sensorid>.
C. Use a row key of the form <timestamp>.
D. Use a row key of the form <sensorid>.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 28
Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?
A. Rewrite the job in Pig.
B. Rewrite the job in Apache Spark.
C. Increase the size of the Hadoop cluster.

D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.
Answer: B (LEAVE A REPLY)
Spark performs in-memory processing and faster, which results in optimization of job's processing time.

NEW QUESTION: 29
Your financial services company is moving to cloud technology and wants to store 50 TB of financial time-series data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?
A. Cloud Bigtable
B. Google BigQuery
C. Google Cloud Storage
D. Google Cloud Datastore
Answer: A (LEAVE A REPLY)
Explanation/Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series

NEW QUESTION: 30
Cloud Bigtable is Google's _____ Big Data database service.
A. Relational
B. mySQL
C. NoSQL
D. SQL Server
Answer: (SHOW ANSWER)
Cloud Bigtable is Google's NoSQL Big Data database service. It is the same database that Google uses for services, such as Search, Analytics, Maps, and Gmail. It is used for requirements that are low latency and high throughput including Internet of Things (IoT), user analytics, and financial data analysis.
Reference: https://cloud.google.com/bigtable/

NEW QUESTION: 31
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_releasedor all movies with tag=Comedyordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?
A. Manually configure the index in your index config as follows:

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

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

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

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

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

Answer: (SHOW ANSWER)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here: https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html (**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 32

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: (SHOW ANSWER)

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: 33
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.
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 he common data in the HDFS storage for a Google Cloud Dataproc cluster.
C. Store the common data in BigQuery as partitioned tables.
D. Store the common data in BigQuery and expose authorized views.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 34
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_releasedor all movies with tag=Comedyordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?
A. Manually configure the index in your index config as follows:

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

B. Set the following in your entity options: exclude_from_indexes = 'date_published'
C. Manually configure the index in your index config as follows:

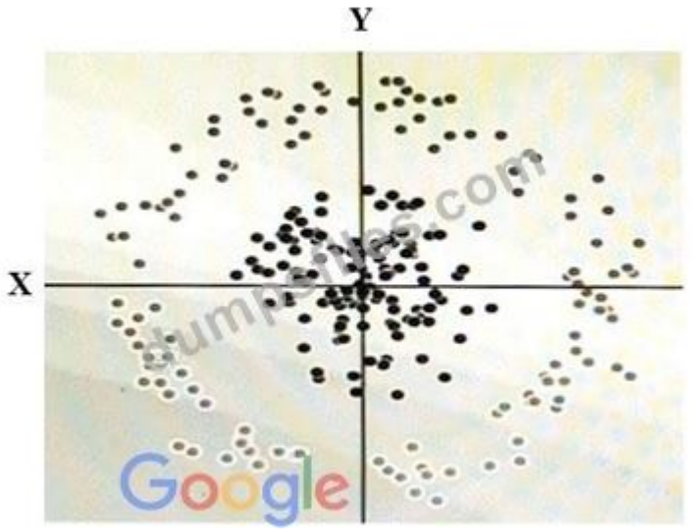
```
Indexes:
-kind: Movie
    Properties:
    -name: actors
    name: date_released
-kind: Movie
    Properties:
    -name: tags
    name: date_released
```

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

Answer: C (LEAVE A REPLY)

NEW QUESTION: 35

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+Y^2

B. X^2

C. cos(X)

D. Y^2

Answer: C (LEAVE A REPLY)

NEW QUESTION: 36

Which of the following are feature engineering techniques? (Select 2 answers)

A. Hidden feature layers

B. Feature prioritization
C. Crossed feature columns
D. Bucketization of a continuous feature
Answer: C,D (LEAVE A REPLY)
Selecting and crafting the right set of feature columns is key to learning an effective model.
Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.
Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.
Reference:
https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_mo del

NEW QUESTION: 37
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)
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: 38
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: B (LEAVE A REPLY)

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: 39
If a dataset contains rows with individual people and columns for year of birth, country, and income, how many of the columns are continuous and how many are categorical?
A. 1 continuous and 2 categorical
B. 3 categorical
C. 3 continuous
D. 2 continuous and 1 categorical
Answer: D (LEAVE A REPLY)
The columns can be grouped into two types--categorical and continuous columns:
A column is called categorical if its value can only be one of the categories in a finite set. For example, the native country of a person (U.S., India, Japan, etc.) or the education level (high school, college, etc.) are categorical columns.
A column is called continuous if its value can be any numerical value in a continuous range. For example, the capital gain of a person (e.g. $14,084) is a continuous column. Year of birth and income are continuous columns. Country is a categorical column. You could use bucketization to turn year of birth and/or income into categorical features, but the raw columns are continuous.
Reference: https://www.tensorflow.org/tutorials/wide#reading_the_census_data

NEW QUESTION: 40
Case Study 1 - Flowlogistic
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market.
Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it

because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

* Databases

8 physical servers in 2 clusters

- SQL Server - user data, inventory, static data

3 physical servers

- Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

- Tomcat - Java services
- Nginx - static content
- Batch servers

* Storage appliances

- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage
- Network-attached storage (NAS) image storage, logs, backups

* 10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

* 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements

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

Technical Requirements

* Handle both streaming and batch data
* Migrate existing Hadoop workloads
* Ensure architecture is scalable and elastic to meet the changing demands of the company.
* Use managed services whenever possible
* Encrypt data flight and at rest
* Connect a VPN between the production data center and cloud environment SEO Statement We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.
CTO Statement
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.
CFO Statement
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?
A. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
B. Cloud Pub/Sub, Cloud Dataflow, and Local SSD
C. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
D. Cloud Load Balancing, Cloud Dataflow, and Cloud Storage
Answer: A (LEAVE A REPLY)
Pub/Sub -> Dataflow for real time processing requirements.
https://codelabs.developers.google.com/codelabs/cpb104-pubsub/#0

NEW QUESTION: 41
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. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.
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. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.
Answer: C (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 42
Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?
A. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
B. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created
C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
D. They have not assigned the timestamp, which causes the job to fail
Answer: (SHOW ANSWER)

NEW QUESTION: 43
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 two.)
A. Use Cloud Deployment Manager to automate access provision.
B. Create distinct groups for various teams, and specify groups in Cloud IAM policies.

C. 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.
D. Introduce resource hierarchy to leverage access control policy inheritance.
E. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
Answer: A,B (LEAVE A REPLY)

NEW QUESTION: 44
An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?
A. Use federated data sources, and check data in the SQL query.
B. Enable BigQuery monitoring in Google Stackdriver and create an alert.
C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.
D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.
Answer: D (LEAVE A REPLY)
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: 45
You are analyzing the price of a company's stock. Every 5 seconds, you need to compute a moving average of the past 30 seconds' worth of data. You are reading data from Pub/Sub and using DataFlow to conduct the analysis. How should you set up your windowed pipeline?
A. Use a fixed window with a duration of 30 seconds. Emit results by setting the following trigger: AfterWatermark.pastEndOfWindow().plusDelayOf(Duration.standardSeconds(5))
B. Use a sliding window with a duration of 30 seconds and a period of 5 seconds. Emit results by setting the following trigger: AfterWatermark.pastEndOfWindow()
C. Use a sliding window with a duration of 5 seconds. Emit results by setting the following trigger: AfterProcessingTime.pastFirstElementInPane().plusDelayOf(Duration.standardSeconds(30))
D. Use a fixed window with a duration of 5 seconds. Emit results by setting the following trigger: AfterProcessingTime.pastFirstElementInPane().plusDelayOf(Duration.standardSeconds(30))
Answer: A (LEAVE A REPLY)

NEW QUESTION: 46
Case Study 1 - Flowlogistic
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market.
Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.
Solution Concept
Flowlogistic wants to implement two concepts using the cloud:

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

* Databases

8 physical servers in 2 clusters

- SQL Server - user data, inventory, static data

3 physical servers

- Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

- Tomcat - Java services
- Nginx - static content
- Batch servers

* Storage appliances

- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage
- Network-attached storage (NAS) image storage, logs, backups

* 10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

* 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements

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

Technical Requirements

* Handle both streaming and batch data
* Migrate existing Hadoop workloads
* Ensure architecture is scalable and elastic to meet the changing demands of the company.
* Use managed services whenever possible
* Encrypt data flight and at rest

* Connect a VPN between the production data center and cloud environment SEO Statement We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.
CTO Statement
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.
CFO Statement
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?
A. Store the common data in BigQuery as partitioned tables.
B. Store the common data in BigQuery and expose authorized views.
C. Store the common data encoded as Avro in Google Cloud Storage.
D. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.
Answer: B (LEAVE A REPLY)
DataProc can access data from Bigquery as well.

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 47
Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.

Which approach should you take?

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

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

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

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

Answer: (SHOW ANSWER)

Topic 3, 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: 48
You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? (Choose two.)
A. Publisher throughput quota is too small.
B. Total outstanding messages exceed the 10-MB maximum.
C. Error handling in the subscriber code is not handling run-time errors properly.
D. The subscriber code cannot keep up with the messages.
E. The subscriber code does not acknowledge the messages that it pulls.
Answer: (SHOW ANSWER)
C, E: By not acknowleding the pulled message, this result in it be putted back in Cloud Pub/Sub, meaning the messages accumulate instead of being consumed and removed from Pub/Sub. The

same thing can happen ig the subscriber maintains the lease on the message it receives in case of an error. This reduces the overall rate of processing because messages get stuck on the first subscriber. Also, errors in Cloud Function do not show up in Stackdriver Log Viewer if they are not correctly handled.
A: No problem with publisher rate as the observed result is a higher number of messages and not a lower number.
B: if messages exceed the 10MB maximum, they cannot be published.
D: Cloud Functions automatically scales so they should be able to keep up.

NEW QUESTION: 49
Cloud Bigtable is Google's _____ Big Data database service.
A. Relational
B. mySQL
C. NoSQL
D. SQL Server
Answer: (SHOW ANSWER)
Explanation
Cloud Bigtable is Google's NoSQL Big Data database service. It is the same database that Google uses for services, such as Search, Analytics, Maps, and Gmail.
It is used for requirements that are low latency and high throughput including Internet of Things (IoT), user analytics, and financial data analysis.
Reference: https://cloud.google.com/bigtable/

NEW QUESTION: 50
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)
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: 51
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. Deploy the model and measure the real-world AUC; it's always higher because of generalization
B. Train a classifier with deep neural networks, because neural networks would always beat SVMs
C. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC
D. Perform hyperparameter tuning
Answer: (SHOW ANSWER)

NEW QUESTION: 52
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. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
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. 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.
Answer: C,E (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 53
Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by

10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.
You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (Choose two.)
A. Introduce data compression for each file to increase the rate file of file transfer.
B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.
D. Assemble 1,000 files into a tape archive (TAR) file. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.
E. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.
Answer: C,E (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 54
You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?
A. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
B. Update the current pipeline and use the drain flag.
C. Update the current pipeline and provide the transform mapping JSON object.
D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 55
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: 56
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: 57
You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.
Which Google database service should you use?
A. Cloud SQL
B. BigQuery
C. Cloud Bigtable
D. Cloud Datastore
Answer: D (LEAVE A REPLY)
https://cloud.google.com/datastore/docs/concepts/overview

NEW QUESTION: 58
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 using BigQuery's default encoding.
B. The CSV data has not gone through an ETL phase before loading into BigQuery.
C. The CSV data loaded in BigQuery is not flagged as CSV.
D. The CSV data has invalid rows that were skipped on import.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 59
You are deploying MariaDB SQL databases on GCE VM Instances and need to configure monitoring and alerting. You want to collect metrics including network connections, disk IO and replication status from MariaDB with minimal development effort and use StackDriver for dashboards and alerts.
What should you do?

A. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.
B. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.
C. Place the MariaDB instances in an Instance Group with a Health Check.
D. Install the StackDriver Agent and configure the MySQL plugin.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 60
You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the data. They should only see certain tables based on their team membership. How should you set user permissions?
A. Assign the users/groups data viewer access at the table level for each table
B. Create authorized views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the authorized views
C. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views
D. Create authorized views for each team in datasets created for each team. Assign the authorized views data viewer access to the dataset in which the data resides. Assign the users/groups data viewer access to the datasets in which the authorized views reside
Answer: B (LEAVE A REPLY)

NEW QUESTION: 61
Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?
A. Use K-means Clustering to detect faces in the pixels.
B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.
Answer: C (LEAVE A REPLY)
Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning. So deep is a strictly defined, technical term that means more than one hidden layer.
In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.

A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.
Feature engineering is difficult to perform on raw image data.
K-means Clustering is an unsupervised learning method used to categorize unlabeled data.
Reference: https://deeplearning4j.org/neuralnet-overview

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 62
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 try... catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.
B. 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.
C. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
D. Add a try... catch block to your DoFn that transforms the data, extract erroneous rows from logs.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 63
Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?
A. Use K-means Clustering to detect faces in the pixels.
B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.

Answer: (SHOW ANSWER)
Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning. So deep is a strictly defined, technical term that means more than one hidden layer.
In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.
A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines. Feature engineering is difficult to perform on raw image data.
K-means Clustering is an unsupervised learning method used to categorize unlabeled data.
Reference: https://deeplearning4j.org/neuralnet-overview

NEW QUESTION: 64
Suppose you have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error.
SELECT person FROM
`project1.example.table1` WHERE city = "London" How would you correct the error?
A. Add ", UNNEST(person)" before the WHERE clause.
B. Change "person" to "person.city".
C. Change "person" to "city.person".
D. Add ", UNNEST(city)" before the WHERE clause.
Answer: A (LEAVE A REPLY)
To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma.
Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_results

NEW QUESTION: 65
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. Store each data entry as the primary key in a separate database and apply an index.
B. Assign global unique identifiers (GUID) to each data entry.

C. Compute the hash value of each data entry, and compare it with all historical data.
D. Maintain a database table to store the hash value and other metadata for each data entry.
Answer: (SHOW ANSWER)

NEW QUESTION: 66
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: 67
Cloud Bigtable is Google's _____ Big Data database service.
A. Relational
B. mySQL
C. NoSQL
D. SQL Server
Answer: (SHOW ANSWER)
Cloud Bigtable is Google's NoSQL Big Data database service. It is the same database that Google uses for services, such as Search, Analytics, Maps, and Gmail.
It is used for requirements that are low latency and high throughput including Internet of Things (IoT), user analytics, and financial data analysis.
Reference: https://cloud.google.com/bigtable/

NEW QUESTION: 68
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 a table called tracking_table with a TIMESTAMP column to represent the day.
B. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
C. Create a table called tracking_table and include a DATE column.
D. Create a partitioned table called tracking_table and include a TIMESTAMP column.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 69
Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?
A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.
B. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
C. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.

D. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 70
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: 71
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 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. 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.
C. 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.
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: 72
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. Reinforcement learning to predict the location of a transaction.
D. Clustering to divide the transactions into N categories based on feature similarity.
E. Unsupervised learning to predict the location of a transaction.
F. Supervised learning to predict the location of a transaction.
Answer: (SHOW ANSWER)

NEW QUESTION: 73
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: C (LEAVE A REPLY)
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: 74
Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?
A. Use K-means Clustering to detect faces in the pixels.
B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.
Answer: C (LEAVE A REPLY)
Explanation
Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning.
So deep is a strictly defined, technical term that means more than one hidden layer.
In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.
A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.
Feature engineering is difficult to perform on raw image data.
K-means Clustering is an unsupervised learning method used to categorize unlabeled data.
Reference: https://deeplearning4j.org/neuralnet-overview

NEW QUESTION: 75
You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:
* Executing the transformations on a schedule
* Enabling non-developer analysts to modify transformations

* Providing a graphical tool for designing transformations
What should you do?
A. Load each month's CSV data into BigQuery, and write a SQL query to transform the data to a standard schema. Merge the transformed tables together with a SQL query
B. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformation. The Python code should be stored in a revision control system and modified as the incoming data's schema changes
C. Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe.
Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery
D. Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis
Answer: C (LEAVE A REPLY)

NEW QUESTION: 76
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

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:

https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 77
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 Transfer Appliance to copy the data to Cloud Storage
B. Use gsutil cp -Jto compress the content being uploaded to Cloud Storage
C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
D. 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
Answer: A (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 78
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?
A. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second
B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second
C. Create a table in BigQuery, and append the new samples for CPU and memory to the table
D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 79
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. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.

B. Add capacity (memory and disk space) to the database server by the order of 200.
C. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.
D. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 80
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 Google Stackdriver Audit Logs to review data access.
C. Use Stackdriver Monitoring to see the usage of BigQuery query slots.
D. Get the identity and access management IIAM) policy of each table
Answer: B (LEAVE A REPLY)

NEW QUESTION: 81
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. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to 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. Create an Authorized View with the provided query. Share the dataset that contains the view with 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: B (LEAVE A REPLY)

NEW QUESTION: 82
You need to compose visualizations for operations teams with the following requirements:
Which approach meets the requirements?

A. 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.
B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
C. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
D. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 83
You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?
A. Create a cron schedule in Cloud Dataprep.
B. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.
C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.
D. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 84
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: A (LEAVE A REPLY)
By creating an authorized view one assures that the data is current and avoids taking more storage space (and cost) in order to share a dataset. B and D are not cost optimal and C does not guarantee that the data is kept updated.

NEW QUESTION: 85

You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

A. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.

B. Create a table in BigQuery, and append the new samples for CPU and memory to the table

C. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second

D. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second

Answer: D (LEAVE A REPLY)

NEW QUESTION: 86

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

C. Rowkey: device_id
Column data: date, data_point
D. Rowkey: data_point
Column data: device_id,date
E. Rowkey: date#data_point
Column data: device_id
Answer: A (LEAVE A REPLY)
The most common query is for all the data for a given device for a given day", rowkey should have info for both device and date.

NEW QUESTION: 87
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 cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google 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 a different project that only the security team can access.
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 and rotate the key once.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 88
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. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.
C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.
D. Edge TPUs as sensor devices for storing and transmitting the messages.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 89
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 a table called tracking_table and include a DATE column.
B. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
C. Create a table called tracking_table with a TIMESTAMP column to represent the day.
D. Create a partitioned table called tracking_table and include a TIMESTAMP column.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 90
You are planning to use Google's Dataflow SDK to analyze customer data such as displayed below. Your project requirement is to extract only the customer name from the data source and then write to an output PCollection.
Tom,555 X street
Tim,553 Y street
Sam, 111 Z street
Which operation is best suited for the above data processing requirement?
A. ParDo
B. Sink API
C. Source API
D. Data extraction
Answer: A (LEAVE A REPLY)
In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.
Reference: https://cloud.google.com/dataflow/model/par-do

NEW QUESTION: 91
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. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
B. 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.
C. Create an authorized view in BigQuery to restrict access to tables with sensitive data.
D. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.
Answer: (SHOW ANSWER)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 92
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 zone
C. The number of workers
D. The disk size per worker
Answer: B (LEAVE A REPLY)

NEW QUESTION: 93
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?
A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
B. Use Cloud Dataproc to run your transformations. Use the diagnosecommand to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs.
Configure the job to use non-default Compute Engine machine types when needed.
Answer: B (LEAVE A REPLY)
Explanation

NEW QUESTION: 94
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: (SHOW ANSWER)
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 Reference: https://cloud.google.com/dataflow/model/bigquery-io

NEW QUESTION: 95

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:

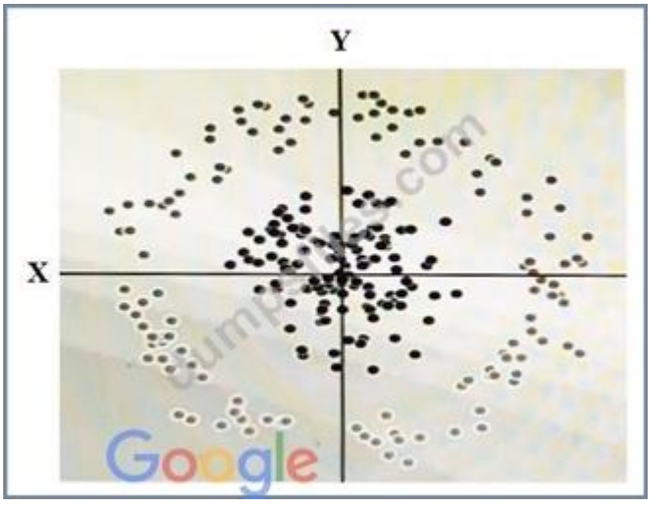
https://cloud.google.com/dataflow/model/triggers

NEW QUESTION: 96

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. cos(X)

C. X^2+Y^2

D. Y^2

Answer: (SHOW ANSWER)

NEW QUESTION: 97

How can you get a neural network to learn about relationships between categories in a categorical feature?

A. Create a multi-hot column
B. Create a one-hot column
C. Create a hash bucket
D. Create an embedding column
Answer: D (LEAVE A REPLY)
There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead of having just one value, like a continuous feature, it has many values, or dimensions. This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn't encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.
Both of these problems can be solved by representing a categorical feature with an embedding column.
The idea is that each category has a smaller vector with, let's say, 5 values in it. But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic features in a neural network. The difference is that each category has a set of weights (5 of them in this case).
You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.
Reference: https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine- course/a-wide-and-deep-model.html

NEW QUESTION: 98
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: A (LEAVE A REPLY)
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: 99
By default, which of the following windowing behavior does Dataflow apply to unbounded data sets?
A. Windows at every 100 MB of data
B. Single, Global Window
C. Windows at every 1 minute

D. Windows at every 10 minutes
Answer: B (LEAVE A REPLY)
Dataflow's default windowing behavior is to assign all elements of a PCollection to a single, global window, even for unbounded PCollections Reference:
https://cloud.google.com/dataflow/model/pcollection

NEW QUESTION: 100
Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?
A. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
B. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created
C. They have not assigned the timestamp, which causes the job to fail
D. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
Answer: (SHOW ANSWER)

NEW QUESTION: 101
You work for a financial institution that lets customers register online. As new customers register, their user data is sent to Pub/Sub before being ingested into BigQuery. For security reasons, you decide to redact your customers' Government issued Identification Number while allowing customer service representatives to view the original values when necessary. What should you do?
A. Before loading the data into BigQuery, use Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic format-preserving encryption token.
B. Before loading the data into BigQuery, use Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic hash.
C. Use BigQuery column-level security. Set the table permissions so that only members of the Customer Service user group can see the SSN column.
D. Use BigQuery's built-in AEAD encryption to encrypt the SSN column. Save the keys to a new table that is only viewable by permissioned users.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 102

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 a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.
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 Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 103
Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO.
During testing, you notice that some messages are missing in the dashboard. You check the logs, and all messages are being published to Cloud Pub/Sub successfully. What should you do next?
A. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
B. Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.
C. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.
D. Check the dashboard application to see if it is not displaying correctly.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 104
Which Google Cloud Platform service is an alternative to Hadoop with Hive?
A. Cloud Dataflow
B. Cloud Bigtable
C. BigQuery
D. Cloud Datastore
Answer: C (LEAVE A REPLY)
Explanation

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.
Google BigQuery is an enterprise data warehouse.
Reference: https://en.wikipedia.org/wiki/Apache_Hive

NEW QUESTION: 105
If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?
A. Do not use a production instance.
B. Run your test for at least 10 minutes.
C. Before you test, run a heavy pre-test for several minutes.
D. Use at least 300 GB of data.
Answer: A (LEAVE A REPLY)
If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test:
Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.
Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data.
However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use 100 GB of data per node.
Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.
Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.
Reference: https://cloud.google.com/bigtable/docs/performance

NEW QUESTION: 106
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. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
B. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.
C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
D. Put the data into Google Cloud Storage.
Answer: (SHOW ANSWER)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here: https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html (**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 107
You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?
A. Grant the consultant the Cloud Dataflow Developer role on the project.
B. Create a service account and allow the consultant to log on with it.
C. Create an anonymized sample of the data for the consultant to work with in a different project.
D. Grant the consultant the Viewer role on the project.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 108
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 try... catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.
B. Add a try... catch block to your DoFn that transforms the data, extract erroneous rows from logs.
C. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
D. Add a try... catch block to your sideOutput to create a PCollection that can be stored to PubSub later.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 109
You are designing the database schema for a machine learning-based food ordering service that will
predict what users want to eat. Here is some of the information you need to store:
The user profile: What the user likes and doesn't like to eat
- The user account information: Name, address, preferred meal times
-

The order information: When orders are made, from where, to whom
▪
The database will be used to store all the transactional data of the product. You want to optimize the data
schema. Which Google Cloud Platform product should you use?
A. Cloud Datastore
B. Cloud SQL
C. Cloud Bigtable
D. BigQuery
Answer: (SHOW ANSWER)

NEW QUESTION: 110
You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?
A. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster
B. Create a Directed Acyclic Graph in Cloud Composer
C. Create an initialization action to execute the jobs
D. Create a Cloud Dataproc Workflow Template
Answer: B (LEAVE A REPLY)

NEW QUESTION: 111
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 WHERE_PARTITIONTIME pseudo column
B. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD
C. Use the TABLE_DATE_RANGE function
D. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
Answer: C (LEAVE A REPLY)

NEW QUESTION: 112
You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt.
You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?
A. Add a SideInput that returns a Boolean if the element is corrupt.
B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.

C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.
D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.
Answer: B (LEAVE A REPLY)
Explanation

NEW QUESTION: 113
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: device_idColumn data: date, data_point
B. Rowkey: date#device_idColumn data: data_point
C. Rowkey: data_pointColumn data: device_id, date
D. Rowkey: date#data_pointColumn data: device_id
E. Rowkey: dateColumn data: device_id, data_point
Answer: C (LEAVE A REPLY)

NEW QUESTION: 114
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. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.
B. Add capacity (memory and disk space) to the database server by the order of 200.
C. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.
D. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 115
You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?
A. Create an initialization action to execute the jobs

B. Create a Directed Acyclic Graph in Cloud Composer
C. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster
D. Create a Cloud Dataproc Workflow Template
Answer: D (LEAVE A REPLY)

NEW QUESTION: 116
Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?
A. Export the data into a Google Sheet for virtualization.
B. Create an additional table with only the necessary columns.
C. Create a view on the table to present to the virtualization tool.
D. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.
Answer: (SHOW ANSWER)
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: 117
Which of the following is NOT a valid use case to select HDD (hard disk drives) as the storage for Google Cloud Bigtable?
A. You expect to store at least 10 TB of data.

B. You will mostly run batch workloads with scans and writes, rather than frequently executing random reads of a small number of rows.
C. You need to integrate with Google BigQuery.
D. You will not use the data to back a user-facing or latency-sensitive application.
Answer: C (LEAVE A REPLY)
Explanation
For example, if you plan to store extensive historical data for a large number of remote-sensing devices and then use the data to generate daily reports, the cost savings for HDD storage may justify the performance tradeoff. On the other hand, if you plan to use the data to display a real-time dashboard, it probably would not make sense to use HDD storage-reads would be much more frequent in this case, and reads are much slower with HDD storage.
Reference: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd

NEW QUESTION: 118
Which of the following statements about Legacy SQL and Standard SQL is not true?
A. Standard SQL is the preferred query language for BigQuery.
B. If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
C. One difference between the two query languages is how you specify fully-qualified table names (i.e. table names that include their associated project name).
D. You need to set a query language for each dataset and the default is Standard SQL.
Answer: D (LEAVE A REPLY)
You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.
Standard SQL has been the preferred query language since BigQuery 2.0 was released.
In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.
Due to the differences in syntax between the two query languages (such as with project- qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

NEW QUESTION: 119
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)
Explanation/Reference: https://cloud.google.com/solutions/data-lifecycle-cloud-platform

NEW QUESTION: 120
Your financial services company is moving to cloud technology and wants to store 50 TB of financial time-series data in the cloud. This data is updated frequently and new data will be streaming in all the time.
Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?
A. Google BigQuery
B. Google Cloud Datastore
C. Google Cloud Storage
D. Cloud Bigtable
Answer: D (LEAVE A REPLY)

NEW QUESTION: 121
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. Logistic classification
B. Recurrent neural network
C. Linear regression
D. Feedforward neural network
Answer: C (LEAVE A REPLY)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 122
You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table

using an insert job, but you do not want to receive notifications for other tables. What should you do?
A. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.
B. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.
C. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.
D. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.
Answer: (SHOW ANSWER)

NEW QUESTION: 123
You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.
You have the following requirements:
* You will batch-load the posts once per day and run them through the Cloud Natural Language API.
* You will extract topics and sentiment from the posts.
* You must store the raw posts for archiving and reprocessing.
* You will create dashboards to be shared with people both inside and outside your organization.
You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?
A. Store the social media posts and the data extracted from the API in BigQuery.
B. Store the social media posts and the data extracted from the API in Cloud SQL.
C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.
D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.
Answer: (SHOW ANSWER)
Explanation

NEW QUESTION: 124
Case Study 1 - Flowlogistic
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market.
Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology

for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

* Databases

8 physical servers in 2 clusters

- SQL Server - user data, inventory, static data

3 physical servers

- Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

- Tomcat - Java services
- Nginx - static content
- Batch servers

* Storage appliances

- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage
- Network-attached storage (NAS) image storage, logs, backups

* 10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

* 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements

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

Technical Requirements
* Handle both streaming and batch data
* Migrate existing Hadoop workloads
* Ensure architecture is scalable and elastic to meet the changing demands of the company.
* Use managed services whenever possible
* Encrypt data flight and at rest
* Connect a VPN between the production data center and cloud environment SEO Statement We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.
CTO Statement
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.
CFO Statement
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?
A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
B. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.
C. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.
D. Use the NOW () function in BigQuery to record the event's time.
Answer: (SHOW ANSWER)

NEW QUESTION: 125
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. Tier older data onto Cloud Storage files, and leverage extended tables.
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. Implement clustering in BigQuery on the ingest date column.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 126
You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of you Cloud Bigtable cluster. Which two actions can you take to accomplish this? (Choose two.)
A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.
E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.
Answer: C,D (LEAVE A REPLY)
C -> Adding more nodes to a cluster (not replication) can improve the write performance
https://cloud.google.com/bigtable/docs/performance
D -> since Google recommends adding nodes when storage utilization is > 70%
https://cloud.google.com/bigtable/docs/modifying-instance#nodes

NEW QUESTION: 127
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. Use .fromQueryoperation to read specific fields from the table.
B. Specify the TableReferenceobject in the code.

C. Use of both the Google BigQuery TableSchemaand TableFieldSchemaclasses.
D. Call a transform that returns TableRowobjects, where each element in the PCollectionrepresents a single row in the table.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 128
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 the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector. Use a Dataflow job to read from PubSub and write to GCS.
B. Deploy a Kafka cluster on GCE VM Instances. 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.
C. 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.
D. 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.
Answer: (SHOW ANSWER)

NEW QUESTION: 129
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: (SHOW ANSWER)
Denormalization will help in performance by reducing query time, update are not good with bigquery.

NEW QUESTION: 130
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: 131
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: 132
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: 133
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 your network bandwidth from Compute Engine to Cloud Storage.
B. Increase the size of the Google Persistent Disk on your server.
C. Increase the CPU size on your server.
D. Increase your network bandwidth from your datacenter to GCP.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 134
You work on a regression problem in a natural language processing domain, and you have 100M labeled exmaples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?
A. Increase the share of the test sample in the train-test split.

B. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.
C. Try out regularization techniques (e.g., dropout of batch normalization) to avoid overfitting.
D. Try to collect more data and increase the size of your dataset.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 135
You need to choose a database to store time series CPU and memory usage for millions of computers.
You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?
A. Create a table in BigQuery, and append the new samples for CPU and memory to the table
B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second
C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second
D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.
Answer: C (LEAVE A REPLY)
https://cloud.google.com/bigtable/docs/schema-design-time-series

NEW QUESTION: 136
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?
A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
C. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.
D. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
Answer: (SHOW ANSWER)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 137
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)
A. Ensure all the tables are included in global dataset.
B. Adjust the settings for each dataset to allow a related region-based security group view access.
C. Adjust the settings for each table to allow a related region-based security group view access.
D. Ensure each table is included in a dataset for a region.
E. Adjust the settings for each view to allow a related region-based security group view access.
Answer: D,E (LEAVE A REPLY)

NEW QUESTION: 138
Which of these are examples of a value in a sparse vector? (Select 2 answers.)
A. [0, 5, 0, 0, 0, 0]
B. [0, 0, 0, 1, 0, 0, 1]
C. [0, 1]
D. [1, 0, 0, 0, 0, 0, 0]
Answer: (SHOW ANSWER)
Categorical features in linear models are typically translated into a sparse vector in which each possible value has a corresponding index or id. For example, if there are only three possible eye colors you can represent 'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and
'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words).
[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single
1.
[0, 5, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s.

Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION: 139
Cloud Dataproc is a managed Apache Hadoop and Apache _____ service.
A. Blaze
B. Spark
C. Fire
D. Ignite
Answer: B (LEAVE A REPLY)
Explanation
Cloud Dataproc is a managed Apache Spark and Apache Hadoop service that lets you use open source data tools for batch processing, querying, streaming, and machine learning.
Reference: https://cloud.google.com/dataproc/docs/

NEW QUESTION: 140
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. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.
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. 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.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 141
In order to securely transfer web traffic data from your computer's web browser to the Cloud Dataproc cluster you should use a(n) _____.
A. VPN connection
B. Special browser
C. SSH tunnel
D. FTP connection
Answer: C (LEAVE A REPLY)
Explanation
To connect to the web interfaces, it is recommended to use an SSH tunnel to create a secure connection to the master node.
Reference:

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

NEW QUESTION: 142
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 dat
A. Unsupervised learning to determine which transactions are most likely to be fraudulent.
B. Which three machine learning applications can you use? (Choose three.)
C. Reinforcement learning to predict the location of a transaction.
D. Supervised learning to predict the location of a transaction.
E. Clustering to divide the transactions into N categories based on feature similarity.
F. Unsupervised learning to predict the location of a transaction.
G. Supervised learning to determine which transactions are most likely to be fraudulent.
Answer: A,D,G (LEAVE A REPLY)

NEW QUESTION: 143
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: 144
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 2 answers.
A. Preserve the structure of the data as much as possible.
B. Use BigQuery UPDATE to further reduce the size of the dataset.
C. 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.
D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.

E. Denormalize the data as must as possible.
Answer: C,D (LEAVE A REPLY)

NEW QUESTION: 145
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world.
The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments - development/test, staging, and production - to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements
* Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
* Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
* Provide reliable and timely access to data for analysis from distributed research workers
* Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
Technical Requirements
* Ensure secure and efficient transport and storage of telemetry data

* Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
* Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately
100m records/day
* Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high- value problems instead of problems with our data pipelines.
You need to compose visualizations for operations teams with the following requirements:
* The report must include telemetry data from all 50,000 installations for the most resent 6 weeks (sampling once every minute).
* The report must not be more than 3 hours delayed from live data.
* The actionable report should only show suboptimal links.
* Most suboptimal links should be sorted to the top.
* Suboptimal links can be grouped and filtered by regional geography.
* User response time to load the report must be <5 seconds.
Which approach meets the requirements?
A. Load the data into Google 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.
B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.
C. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.

D. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 146
Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?
A. Store the common data in BigQuery and expose authorized views.
B. Store the common data encoded as Avro in Google Cloud Storage.
C. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.
D. Store the common data in BigQuery as partitioned tables.
Answer: (SHOW ANSWER)

NEW QUESTION: 147
You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity 'Movie' the property 'actors' and the property 'tags' have multiple values but the property 'date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

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

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

Google dumpsfiles.com

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

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

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

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

A. Option C
B. Option A
C. Option B.
D. Option D
Answer: (SHOW ANSWER)

NEW QUESTION: 148
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 bq loadto load a batch of sensor data 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 the MERGE statement to apply updates in batch every 60 seconds.
Answer: C (LEAVE A REPLY)
Explanation

NEW QUESTION: 149
Which of the following are feature engineering techniques? (Select 2 answers)
A. Hidden feature layers
B. Feature prioritization
C. Crossed feature columns
D. Bucketization of a continuous feature
Answer: C,D (LEAVE A REPLY)
Explanation
Selecting and crafting the right set of feature columns is key to learning an effective model.
Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.
Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.
Reference:
https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

NEW QUESTION: 150
Case Study: 1 - Flowlogistic
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.
Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?
A. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.
B. Create a view on the table to present to the virtualization tool.
C. Create an additional table with only the necessary columns.
D. Export the data into a Google Sheet for virtualization.
Answer: (SHOW ANSWER)

NEW QUESTION: 151
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. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
B. 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
C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset
D. Create groups for your users and give those groups access to the dataset
Answer: C (LEAVE A REPLY)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 152
You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.

You have the following requirements:
* You will batch-load the posts once per day and run them through the Cloud Natural Language API.
* You will extract topics and sentiment from the posts.
* You must store the raw posts for archiving and reprocessing.
* You will create dashboards to be shared with people both inside and outside your organization.
You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?
A. Store the social media posts and the data extracted from the API in BigQuery.
B. Store the social media posts and the data extracted from the API in Cloud SQL.
C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.
D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.
Answer: C (LEAVE A REPLY)
Social media posts can images/videos which cannot be stored in bigquery/

NEW QUESTION: 153
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. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.
C. Create an authorized view in BigQuery to restrict access to tables with sensitive data.
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: 154
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. 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.
B. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
C. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.
D. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.
E. Introduce data compression for each file to increase the rate file of file transfer.
Answer: C,D (LEAVE A REPLY)

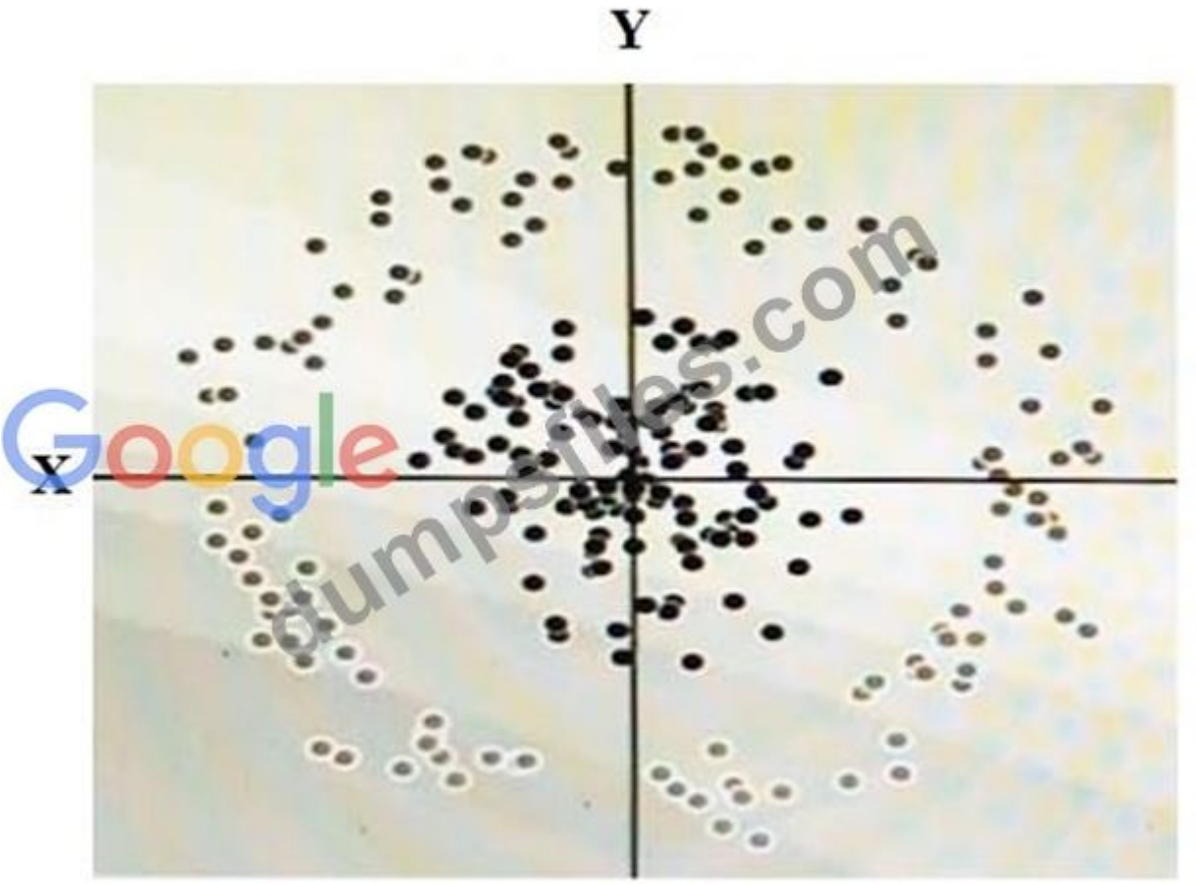
NEW QUESTION: 155
You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?
A. Increase the cluster size with more non-preemptible workers.
B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.
D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.
Answer: D (LEAVE A REPLY)
Explanation
Reference https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/flex

NEW QUESTION: 156
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. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.

B. Specify the TableReference object in the code.
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: D (LEAVE A REPLY)

NEW QUESTION: 157
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,E (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 158
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.



Google
dumpsites.com

To do this you need to add a synthetic feature. What should the value of that feature be?

A. Y^2

B. cos(X)

C. X^2

D. X^2+Y^2

Answer: B (LEAVE A REPLY)

NEW QUESTION: 159

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)

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: 160
You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?
A. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.
B. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.
C. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
D. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
Answer: (SHOW ANSWER)

NEW QUESTION: 161
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)
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: 162
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: (SHOW ANSWER)

NEW QUESTION: 163
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)
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: 164
What are two methods that can be used to denormalize tables in BigQuery?
A. 1) Use nested repeated fields; 2) Use a partitioned table
B. 1) Split table into multiple tables; 2) Use a partitioned table
C. 1) Join tables into one table; 2) Use nested repeated fields
D. 1) Use a partitioned table; 2) Join tables into one table
Answer: (SHOW ANSWER)

NEW QUESTION: 165
You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:
SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country You check the query plan for the query and see the following output in the Read section of Stage:1:

What is the most likely cause of the delay for this query?
A. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew

B. Users are running too many concurrent queries in the system
C. The [myproject:mydataset.mytable] table has too many partitions
D. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
Answer: B (LEAVE A REPLY)

NEW QUESTION: 166
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. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
B. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.
C. Specify the Tableobject in the code.
D. Use .fromQuery operation to read specific fields from the table.
Answer: B (LEAVE A REPLY)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here:
https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html
(**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 167
Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel
uploads to a data transfer server running on GCP. Management informs you that the daily transfers take

too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action
should you take?
A. Increase the CPU size on your server.
B. Increase the size of the Google Persistent Disk on your server.
C. Increase your network bandwidth from your datacenter to GCP.
D. Increase your network bandwidth from Compute Engine to Cloud Storage.
Answer: (SHOW ANSWER)
Explanation/Reference:

NEW QUESTION: 168
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.
Reference:
https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing_data

NEW QUESTION: 169
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_releasedor all movies with tag=Comedyordered 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. Manually configure the index in your index config as follows:

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

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

Answer: (SHOW ANSWER)

NEW QUESTION: 170

Which of the following is NOT true about Dataflow pipelines?

A. Dataflow pipelines are tied to Dataflow, and cannot be run on any other runner

B. Dataflow pipelines can consume data from other Google Cloud services

C. Dataflow pipelines can be programmed in Java

D. Dataflow pipelines use a unified programming model, so can work both with streaming and batch data sources

Answer: (SHOW ANSWER)

Dataflow pipelines can also run on alternate runtimes like Spark and Flink, as they are built using the Apache Beam SDKs Reference: https://cloud.google.com/dataflow/

NEW QUESTION: 171

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 a random number per second.
B. 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.
C. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
D. Change the data pipeline to use BigQuery for storing stock trades, and update your application.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 172
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)
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: 173
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 Hadoop cluster on Google Compute Engine that uses persistent disks.
C. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
D. Create a Google Cloud Dataflow job to process the data.
E. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 174
What are all of the BigQuery operations that Google charges for?
A. Storage, queries, and streaming inserts

B. Storage, queries, and loading data from a file
C. Storage, queries, and exporting data
D. Queries and streaming inserts
Answer: (SHOW ANSWER)
Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations.
Reference: https://cloud.google.com/bigquery/pricing

NEW QUESTION: 175
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_RANGEfunction
B. Use the WHERE_PARTITIONTIMEpseudo column
C. Use WHEREdate 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)
Explanation/Reference: https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understand- your-mobile-app?hl=am

NEW QUESTION: 176
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)
Explanation/Reference:
Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

NEW QUESTION: 177
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: 178

Case Study: 3,

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.
Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?
A. The zone
B. The maximum number of workers
C. The disk size per worker
D. The number of workers
Answer: A (LEAVE A REPLY)

NEW QUESTION: 179

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. 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.
B. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
C. Send the data to Google Cloud Datastore and then export to BigQuery.
D. Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
Answer: (SHOW ANSWER)

NEW QUESTION: 180
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 is not acknowledging messages within the acknowledgement deadline.
B. Your custom endpoint has an out-of-date SSL certificate.
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: (SHOW ANSWER)

NEW QUESTION: 181
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 Transfer Appliance to copy the data to Cloud Storage
B. Use gsutil cp -Jto compress the content being uploaded to Cloud Storage
C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
D. 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
Answer: A (LEAVE A REPLY)
Explanation

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here: https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html (**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 182
The Dataflow SDKs have been recently transitioned into which Apache service?
A. Apache Spark
B. Apache Hadoop
C. Apache Kafka
D. Apache Beam
Answer: (SHOW ANSWER)
Explanation
Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive
Reference: https://cloud.google.com/dataflow/docs/

NEW QUESTION: 183
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?
A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.
C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs.
Configure the job to use non-default Compute Engine machine types when needed.
Answer: C (LEAVE A REPLY)
Dataflow is good with autoscaling and stackdriver to monitor CPU and Storage.

NEW QUESTION: 184

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_releasedor all movies with tag=Comedyordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

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

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

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

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

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

A. Option A
B. Option D
C. Option C
D. Option B.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 185
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 a table called tracking_table and include a DATE column.
B. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
C. Create a partitioned table called tracking_table and include a TIMESTAMP column.
D. Create a table called tracking_table with a TIMESTAMP column to represent the day.
Answer: (SHOW ANSWER)

NEW QUESTION: 186
You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/ Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket.
The source has consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?
A. An alert based on a decrease of subscription/num_undelivered_messages for the source and a rate of change increase of instance/storage/used_bytes for the destination
B. An alert based on an increase of subscription/num_undelivered_messages for the source and a rate of change decrease of instance/storage/used_bytes for the destination
C. An alert based on a decrease of instance/storage/used_bytes for the source and a rate of change increase of subscription/num_undelivered_messages for the destination
D. An alert based on an increase of instance/storage/used_bytes for the source and a rate of change decrease of subscription/num_undelivered_messages for the destination
Answer: B (LEAVE A REPLY)

Increase in number of undelivered messages shows that the messages are not getting subscribed.

NEW QUESTION: 187
Your company currently runs a large on-premises cluster using Spark, Hive, and HDFS in a colocation facility.
The cluster is designed to accommodate peak usage on the system; however, many jobs are batch in nature, and usage of the cluster fluctuates quite dramatically. Your company is eager to move to the cloud to reduce the overhead associated with on-premises infrastructure and maintenance and to benefit from the cost savings.
They are also hoping to modernize their existing infrastructure to use more serverless offerings in order to take advantage of the cloud. Because of the timing of their contract renewal with the colocation facility, they have only 2 months for their initial migration. How would you recommend they approach their upcoming migration strategy so they can maximize their cost savings in the cloud while still executing the migration in time?
A. Modernize the Spark workload for Dataflow and the Hive workload for BigQuery.
B. Migrate the Spark workload to Dataproc plus HDFS, and modernize the Hive workload for BigQuery.
C. Migrate the workloads to Dataproc plus Cloud Storage; modernize later.
D. Migrate the workloads to Dataproc plus HDFS; modernize later.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 188
Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was
previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the
recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways
the development team could increase the responsiveness of the analytics without increasing costs. What
should you recommend they do?
A. Rewrite the job in Pig.
B. Increase the size of the Hadoop cluster.
C. Rewrite the job in Apache Spark.
D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 189
Which of these are examples of a value in a sparse vector? (Select 2 answers.)
A. [0, 5, 0, 0, 0, 0]

B. [0, 0, 0, 1, 0, 0, 1]
C. [0, 1]
D. [1, 0, 0, 0, 0, 0, 0]
Answer: C,D (LEAVE A REPLY)
Explanation
Categorical features in linear models are typically translated into a sparse vector in which each possible value has a corresponding index or id. For example, if there are only three possible eye colors you can represent
'eye_color' as a length 3 vector: 'brown' would become [1, 0, 0], 'blue' would become [0, 1, 0] and 'green' would become [0, 0, 1]. These vectors are called "sparse" because they may be very long, with many zeros, when the set of possible values is very large (such as all English words).
[0, 0, 0, 1, 0, 0, 1] is not a sparse vector because it has two 1s in it. A sparse vector contains only a single 1.
[0, 5, 0, 0, 0, 0] is not a sparse vector because it has a 5 in it. Sparse vectors only contain 0s and 1s.
Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION: 190
You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?
A. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.
B. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.
C. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.
D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.
Answer: (SHOW ANSWER)

NEW QUESTION: 191
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 Tableobject in the code.
B. Use .fromQuery operation to read specific fields from the table.
C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
D. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.
Answer: B (LEAVE A REPLY)
BigQueryIO.read.from() directly reads the whole table from BigQuery. This function exports the whole table to temporary files in Google Cloud Storage, where it will later be read from. This requires almost no computation, as it only performs an export job, and later Dataflow reads from GCS (not from BigQuery).
BigQueryIO.read.fromQuery() executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs).

NEW QUESTION: 192
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. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.
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. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 193
Why do you need to split a machine learning dataset into training data and test data?
A. So you can try two different sets of features
B. To make sure your model is generalized for more than just the training data

C. To allow you to create unit tests in your code
D. So you can use one dataset for a wide model and one for a deep model
Answer: B (LEAVE A REPLY)
The flaw with evaluating a predictive model on training data is that it does not inform you on how well the model has generalized to new unseen data. A model that is selected for its accuracy on the training dataset rather than its accuracy on an unseen test dataset is very likely to have lower accuracy on an unseen test dataset. The reason is that the model is not as generalized. It has specialized to the structure in the training dataset. This is called overfitting.
Reference: https://machinelearningmastery.com/a-simple-intuition-for-overfitting/

NEW QUESTION: 194
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

Column data: device_id, data_point

B. Rowkey: data_point

Column data: device_id, date

C. Rowkey: date#device_id

Column data: data_point

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

NEW QUESTION: 195
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. Cloud AutoML Natural Language
B. Dialogflow Enterprise Edition
C. Cloud Speech-to-Text API
D. Cloud Natural Language API
Answer: (SHOW ANSWER)

NEW QUESTION: 196
Which of the following is not true about Dataflow pipelines?
A. Pipelines are a set of operations
B. Pipelines represent a data processing job
C. Pipelines represent a directed graph of steps
D. Pipelines can share data between instances
Answer: D (LEAVE A REPLY)
Explanation
The data and transforms in a pipeline are unique to, and owned by, that pipeline. While your program can create multiple pipelines, pipelines cannot share data or transforms Reference: https://cloud.google.com/dataflow/model/pipelines

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here: https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html (**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 197

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

B. Use the bq query - -maximum_bytes_billed flag to restrict the number of bytes billed.

C. Create a separate table for each ID.

D. Recreate the table with a partitioning column and clustering column.

Answer: (SHOW ANSWER)

NEW QUESTION: 198

If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?

A. Do not use a production instance.

B. Run your test for at least 10 minutes.

C. Before you test, run a heavy pre-test for several minutes.

D. Use at least 300 GB of data.

Answer: (SHOW ANSWER)

If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test:

Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.

Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use 100 GB of data per node.

Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes. Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.

Reference: https://cloud.google.com/bigtable/docs/performance

NEW QUESTION: 199

You work on a regression problem in a natural language processing domain, and you have 100M labeled exmaples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set. How should you improve the performance of your model?

A. Try out regularization techniques (e.g., dropout of batch normalization) to avoid overfitting.

B. Try to collect more data and increase the size of your dataset.

C. Increase the share of the test sample in the train-test split.
D. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 200
You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?
A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code
D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code
Answer: D (LEAVE A REPLY)
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing

NEW QUESTION: 201
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 global window with a time based trigger with a delay of 60 minutes.
C. Use a fixed-time window with a duration of 60 minutes.
D. Use a sliding time window with a duration of 60 minutes.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 202
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)
Explanation

NEW QUESTION: 203
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. 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.
B. 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.
C. Delete the table CLICK_STREAM, and then re-create it such that the column DT is of the TIMESTAMP type. Reload the data.
D. Add a column TS of the TIMESTAMP type to the table CLICK_STREAM, and populate the numeric values from the column TS for each row. Reference the column TS instead of the column DT from now on.
E. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW_CLICK_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW_CLICK_STREAM instead of the table CLICK_STREAM from now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 204
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 number of workers
B. The disk size per worker
C. The zone
D. The maximum number of workers
Answer: C (LEAVE A REPLY)

NEW QUESTION: 205
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: C (LEAVE A REPLY)
Explanation
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.
Reference: https://cloud.google.com/bigtable/docs/creating-compute-instance

NEW QUESTION: 206
What Dataflow concept determines when a Window's contents should be output based on certain criteria being met?
A. Sessions
B. OutputCriteria
C. Windows
D. Triggers
Answer: D (LEAVE A REPLY)
Explanation
Triggers control when the elements for a specific key and window are output. As elements arrive, they are put into one or more windows by a Window transform and its associated WindowFn, and then passed to the associated Trigger to determine if the Windows contents should be output.
Reference:

https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/transforms/windowing/Trig

NEW QUESTION: 207
What is the HBase Shell for Cloud Bigtable?
A. The HBase shell is a GUI based interface that performs administrative tasks, such as creating and deleting tables.
B. The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables.
C. The HBase shell is a hypervisor based shell that performs administrative tasks, such as creating and deleting new virtualized instances.
D. The HBase shell is a command-line tool that performs only user account management functions to grant access to Cloud Bigtable instances.
Answer: B (LEAVE A REPLY)
Explanation
The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables. The Cloud Bigtable HBase client for Java makes it possible to use the HBase shell to connect to Cloud Bigtable.
Reference: https://cloud.google.com/bigtable/docs/installing-hbase-shell

NEW QUESTION: 208
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)
Explanation
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: 209
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 has invalid rows that were skipped on import.
B. The CSV data loaded in BigQuery is not flagged as CSV.
C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.
D. The CSV data has not gone through an ETL phase before loading into BigQuery.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 210
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: 211
You are deploying a new storage system for your mobile application, which is a media streaming service.
You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity `Movie' the property `actors' and the property
`tags' have multiple values but the property `date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?

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

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

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

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

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

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

A. Option A
B. Option C
C. Option D
D. Option B.
Answer: A (LEAVE A REPLY)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here: https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html (**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

NEW QUESTION: 212
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 Load Balancing, Cloud Dataflow, and Cloud Storage
C. Cloud Pub/Sub, Cloud Dataflow, and Cloud Storage
D. Cloud Pub/Sub, Cloud SQL, and Cloud Storage
Answer: D (LEAVE A REPLY)

NEW QUESTION: 213
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 Hadoop cluster on Google Compute Engine that uses persistent disks.
C. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
D. Create a Google Cloud Dataflow job to process the data.
E. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
Answer: E (LEAVE A REPLY)

NEW QUESTION: 214
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

Scale and harden their PoC to support significantly more data flows generated when they ramp to more
- than 50,000 installations.
Refine their machine-learning cycles to verify and improve the dynamic models they use to control
- topology definition.
MJTelco will also use three separate operating environments - development/test, staging, and production
- to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements

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

Technical Requirements

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

CEO Statement

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

CTO Statement

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and

quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis.
Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?
A. Rowkey: date#device_id
Column data: data_point
B. Rowkey: date
Column data: device_id,data_point
C. Rowkey: data_point
Column data: device_id,date
D. Rowkey: date#data_point
Column data: device_id
E. Rowkey: device_id
Column data: date, data_point
Answer: C (LEAVE A REPLY)

NEW QUESTION: 215
Which Google Cloud Platform service is an alternative to Hadoop with Hive?
A. Cloud Dataflow
B. Cloud Bigtable
C. BigQuery
D. Cloud Datastore
Answer: C (LEAVE A REPLY)
Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.
Google BigQuery is an enterprise data warehouse.
Reference: https://en.wikipedia.org/wiki/Apache_Hive

NEW QUESTION: 216
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)
Legacy sql uses table date range whereas standard sql uses table_sufix for wildcard.

NEW QUESTION: 217
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. 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.
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. 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.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 218
You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output. Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?
A. Cancel
B. Drain
C. Stop
D. Finish
Answer: (SHOW ANSWER)
Explanation
Using the Drain option to stop your job tells the Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources, but the Dataflow

service will preserve any existing resources (such as worker instances) to finish processing and writing any buffered data in your pipeline.
Reference: https://cloud.google.com/dataflow/pipelines/stopping-a-pipeline

NEW QUESTION: 219
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: (SHOW ANSWER)
Explanation
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.
Reference: https://cloud.google.com/dataproc/docs/concepts/overview

NEW QUESTION: 220
Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)
A. Create a new view over events_partitioned using standard SQL
B. Create a new view over events using standard SQL
C. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"
D. Create a service account for the ODBC connection to use for authentication
E. Create a new partitioned table using a standard SQL query
Answer: (SHOW ANSWER)

NEW QUESTION: 221
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: 222
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. Re-create the table using data partitioning on the package delivery date.
B. Implement clustering in BigQuery on the ingest date column.
C. Implement clustering in BigQuery on the package-tracking ID column.
D. Tier older data onto Cloud Storage files, and leverage extended tables.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 223
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)
ference: https://cloud.google.com/solutions/business-intelligence/

NEW QUESTION: 224

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

NEW QUESTION: 225
You are building a teal-lime prediction engine that streams files, which may contain Pll (personal identifiable information) data, into Cloud Storage and eventually into BigQuery You want to ensure that the sensitive data is masked but still maintains referential Integrity, because names and emails are often used as join keys How should you use the Cloud Data Loss Prevention API (DLP API) to ensure that the Pll data is not accessible by unauthorized individuals?
A. Redact all Pll data, and store a version of the unredacted data in a locked-down bucket
B. Scan every table in BigQuery, and mask the data it finds that has Pll
C. Create a pseudonym by replacing Pll data with a cryptographic format-preserving token
D. Create a pseudonym by replacing the Pll data with cryptogenic tokens, and store the non-tokenized data in a locked-down button.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 226
You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time What tools should you use to migrate the data and ensure that it continues to write to your warehouse?
A. gsutil for both the migration and the real-time updates
B. Storage Transfer Service for the migration, Pub/Sub and Cloud Data Fusion for the real-time updates
C. BigQuery Data Transfer Service lor the migration, Pub/Sub and Dataproc for the real-time updates
D. gsutil for the migration; Pub/Sub and Dataflow for the real-time updates
Answer: (SHOW ANSWER)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here: https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html (**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)

Valid Professional-Data-Engineer Dumps shared by TrainingDump.com for Helping Passing Professional-Data-Engineer Exam! TrainingDump.com now offer the **newest Professional-Data-Engineer exam dumps**, the TrainingDump.com Professional-Data-Engineer exam **questions have been updated** and **answers have been corrected** get the **newest** TrainingDump.com Professional-Data-Engineer dumps with Test Engine here: https://www.trainingdump.com/Google/Professional-Data-Engineer-practice-exam-dumps.html (**403** Q&As Dumps, **40%OFF** **Special Discount:** **Exam-Tests**)