Google.Professional-Data-Engineer.v2023-04-24.q185

Exam Code:	Professional-Data-Engineer
Exam Name:	Google Certified Professional Data Engineer Exam
Certification Provider:	Google
Free Question Number:	185
Version:	v2023-04-24
# of views:	341
# of Questions views:	3200
https://www.dumpsfiles.com/files/Google/Professional-Data-Engineer/Google.Professional-Data-Engineer.v2023-04-24.q185	

NEW QUESTION: 1
Your neural network model is taking days to train. You want to increase the training speed. What can you do?
A. Subsample your training dataset.
B. Increase the number of input features to your model.
C. Increase the number of layers in your neural network.
D. Subsample your test dataset.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 2
Which Java SDK class can you use to run your Dataflow programs locally?
A. LocalRunner
B. DirectPipelineRunner
C. MachineRunner
D. LocalPipelineRunner
Answer: (SHOW ANSWER)
DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization.
Useful for small local execution and tests
Reference: https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRunner

NEW QUESTION: 3
You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?

A. PCollection
B. Transform
C. Pipeline
D. Sink API
Answer: (SHOW ANSWER)
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: 4
You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Userstable consisting of a FirstNamefield and a LastNamefield. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullNamefield consisting of the value of the FirstNamefield concatenated with a space, followed by the value of the LastNamefield for each employee. How can you make that data available while minimizing cost?
A. Create a Google Cloud Dataflow job that queries BigQuery for the entire Userstable, concatenates the FirstNamevalue and LastNamevalue for each user, and loads the proper values for FirstName, LastName, and FullNameinto a new table in BigQuery.
B. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastNameand FullName. Run a BigQuery load job to load the new CSV file into BigQuery.
C. Add a new column called FullNameto the Users table. Run an UPDATEstatement that updates the FullNamecolumn for each user with the concatenation of the FirstNameand LastNamevalues.
D. Create a view in BigQuery that concatenates the FirstNameand LastNamefield values to produce the FullName.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 5
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: (SHOW ANSWER)

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: 6

All Google Cloud Bigtable client requests go through a front-end server _____ they are sent to a Cloud Bigtable node.

A. before

B. after

C. only if

D. once

Answer: A (LEAVE A REPLY)

Explanation

In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node.

The nodes are organized into a Cloud Bigtable cluster, which belongs to a Cloud Bigtable instance, which is a container for the cluster. Each node in the cluster handles a subset of the requests to the cluster.

When additional nodes are added to a cluster, you can increase the number of simultaneous requests that the cluster can handle, as well as the maximum throughput for the entire cluster.

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

NEW QUESTION: 7

The marketing team at your organization provides regular updates of a segment of your customer dataset.

The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?

A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.

B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.

C. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.

D. Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.
Answer: D (LEAVE A REPLY)
https://cloud.google.com/blog/products/gcp/performing-large-scale-mutations-in-bigquery

NEW QUESTION: 8
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 Cloud Dataproc cluster that uses the Google Cloud Storage connector.
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 Hadoop cluster on Google Compute Engine that uses Local SSD disks.
E. Create a Google Cloud Dataflow job to process the data.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 9
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 Spanner for storage. Add secondary indexes to support query patterns.
C. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.
D. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
Answer: (SHOW ANSWER)

NEW QUESTION: 10
Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by 250,000 records per second. Many third parties use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:
* Single global endpoint
* ANSI SQL support
* Consistent access to the most up-to-date data
What should you do?

A. Implement Cloud SQL for PostgreSQL with the master in Norht America and read replicas in Asia and Europe.
B. Implement BigQuery with no region selected for storage or processing.
C. Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.
D. Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 11
You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data.
Which two actions should you take? (Choose two.)
A. Configure your Cloud Dataflow pipeline to use local execution
B. Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions
C. Increase the number of nodes in the Cloud Bigtable cluster
D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable
E. Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable
Answer: (SHOW ANSWER)
A - Local execution is useful for testing and debugging purposes, especially if your pipeline can use smaller in-memory datasets.
B- https://cloud.google.com/dataflow/docs/guides/specifying-exec-params C- increases both read and write performance D- Flatten merges multiple PCollection objects into a single logical PCollection.
E- Consider using CoGroupByKey if you have multiple data sets that provide information about related things .

NEW QUESTION: 12
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: 13
You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-
Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about
100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and
durability (ACID). However, high availability and low latency are required.
You need to analyze the data by querying against individual fields. Which three databases meet your
requirements? (Choose three.)
A. Redis
B. HBase
C. HDFS with Hive
D. Cassandra
E. MySQL
F. MongoDB
Answer: B,C,F (LEAVE A REPLY)

NEW QUESTION: 14
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 fron PubSub and write to GCS.
B. Deploy a Kafka cluster on GCE VM Instances with the PubSub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
C. Deploy 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.
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 fron PubSub and write to GCS.

Answer: (SHOW ANSWER)

NEW QUESTION: 15
What are two of the characteristics of using online prediction rather than batch prediction?
A. It is optimized to handle a high volume of data instances in a job and to run more complex models.
B. Predictions are returned in the response message.
C. Predictions are written to output files in a Cloud Storage location that you specify.
D. It is optimized to minimize the latency of serving predictions.
Answer: B,D (LEAVE A REPLY)
Explanation
Online prediction
Optimized to minimize the latency of serving predictions.
Predictions returned in the response message.
Batch prediction
Optimized to handle a high volume of instances in a job and to run more complex models.
Predictions written to output files in a Cloud Storage location that you specify.
Reference:
https://cloud.google.com/ml-engine/docs/prediction-overview#online_prediction_versus_batch_prediction

NEW QUESTION: 16
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 Bigtable
B. Cloud Datastore
C. BigQuery
D. Cloud SQL
Answer: D (LEAVE A REPLY)

NEW QUESTION: 17
You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?
A. Instead of feeding in each feature individually, average their values in batches of 3.
B. Remove the features that have null values for more than 50% of the training records.
C. Combine highly co-dependent features into one representative feature.
D. Eliminate features that are highly correlated to the output labels.
Answer: (SHOW ANSWER)

NEW QUESTION: 18
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: (SHOW ANSWER)
Explanation/Reference:
Reference: https://cloud.google.com/bigquery/docs/access-control

NEW QUESTION: 19
You are designing a pipeline that publishes application events to a Pub/Sub topic. Although message ordering is not important, you need to be able to aggregate events across disjoint hourly intervals before loading the results to BigQuery for analysis. What technology should you use to process and load this data to BigQuery while ensuring that it will scale with large volumes of events?
A. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.
B. Schedule a Cloud Function to run hourly, pulling all available messages from the Pub/Sub topic and performing the necessary aggregations.
C. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub/Sub topic and performing the necessary aggregations.

D. Create a streaming Dataflow job that reads continually from the Pub/Sub topic and performs aggregations using tumbling windows.
Answer: (SHOW ANSWER)
Explanation/Reference:

NEW QUESTION: 20
Which of the following is not possible using primitive roles?
A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.
B. Give UserA owner access and UserB editor access for all datasets in a project.
C. Give a user access to view all datasets in a project, but not run queries on them.
D. Give GroupA owner access and GroupB editor access for all datasets in a project.
Answer: C (LEAVE A REPLY)
Primitive roles can be used to give owner, editor, or viewer access to a user or group, but they can't be used to separate data access permissions from job-running permissions.
Reference: https://cloud.google.com/bigquery/docs/access-control#primitive_iam_roles

NEW QUESTION: 21
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. Denormalize the data as must as possible.
B. 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.
C. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
D. Use BigQuery UPDATE to further reduce the size of the dataset.
E. Preserve the structure of the data as much as possible.
Answer: A,B (LEAVE A REPLY)

NEW QUESTION: 22
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: (SHOW ANSWER)
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: 23
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. 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.
B. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.
D. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 24
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: 25
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: data_pointColumn data: device_id, date
B. Rowkey: date#data_pointColumn data: device_id
C. Rowkey: date#device_idColumn data: data_point
D. Rowkey: dateColumn data: device_id, data_point
E. Rowkey: device_idColumn data: date, data_point
Answer: A (LEAVE A REPLY)

NEW QUESTION: 26
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 Cloud Dataproc Workflow Template

B. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster
C. Create a Directed Acyclic Graph in Cloud Composer
D. Create an initialization action to execute the jobs
Answer: (SHOW ANSWER)

NEW QUESTION: 27
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: B (LEAVE A REPLY)
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: 28
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. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
B. Preserve the structure of the data as much as possible.
C. Denormalize the data as must as possible.
D. Use BigQuery UPDATE to further reduce the size of the dataset.
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)

NEW QUESTION: 29

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 Stackdriver Monitoring to see the usage of BigQuery query slots.

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

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

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

Answer: A (LEAVE A REPLY)

NEW QUESTION: 30

You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?

A. Specify customer-supplied encryption key (CSEK) in the .botoconfiguration file. Use gsutil cpto upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.

B. Use gcloud kms keys createto create a symmetric key. Then use gcloud kms encryptto encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.

C. Specify customer-supplied encryption key (CSEK) in the .botoconfiguration file. Use gsutil cpto upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.

D. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encryptto encrypt each archival file with the key. Use gsutil cpto upload each encrypted file to the Cloud Storage bucket.

Manually destroy the key previously used for encryption, and rotate the key once.

Answer: (SHOW ANSWER)

NEW QUESTION: 31

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]

NEW QUESTION: 32
Cloud Dataproc is a managed Apache Hadoop and Apache _____ service.
A. Blaze
B. Spark
C. Fire
D. Ignite
Answer: B (LEAVE A REPLY)
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.

NEW QUESTION: 33
You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for long-running batch jobs. You want to use a managed service. What should you do?
A. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances. Store data in HDFS. Change references in scripts from hdfs:// to gs://
B. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from hdfs:// to gs://
C. Deploy a Cloud Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://
D. Deploy a Cloud Dataproc cluster. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

Answer: D (LEAVE A REPLY)

NEW QUESTION: 34
You have uploaded 5 years of log data to Cloud Storage A user reported that some data points in the log data are outside of their expected ranges, which indicates errors You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons What should you do?
A. Create a Compute Engine instance and create a new copy of the data in Cloud Storage Skip the rows with errors
B. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage
C. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage
D. Import the data from Cloud Storage into BigQuery Create a new BigQuery table, and skip the rows with errors.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 35
You need to migrate a 2TB relational database to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.
Which service do you select for storing and serving your data?
A. Cloud Spanner
B. Cloud SQL
C. Cloud Bigtable
D. Cloud Firestore
Answer: B (LEAVE A REPLY)

NEW QUESTION: 36
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 Cloud Dataproc Workflow Template
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 Directed Acyclic Graph in Cloud Composer
Answer: B (LEAVE A REPLY)

NEW QUESTION: 37
You want to optimize your queries for cost and performance. How should you structure your data?
A. Cluster table data by create_date partition by locationed and device_version
B. Cluster table data by create_date location_id and device_version
C. Partition table data by create_date cluster table data by location_Id and device_version
D. Partition table data by create_date, location_id and device_version
Answer: (SHOW ANSWER)

NEW QUESTION: 38
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data.
What should you do?
A. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
B. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil cp to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google Cloud.
C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.
D. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket.
Manually destroy the key previously used for encryption, and rotate the key once and rotate the key once.
Answer: (SHOW ANSWER)

NEW QUESTION: 39
You need to choose a database for a new project that has the following requirements:
* Fully managed
* Able to automatically scale up
* Transactionally consistent
* Able to scale up to 6 TB
* Able to be queried using SQL
Which database do you choose?
A. Cloud Bigtable
B. Cloud SQL

C. Cloud Datastore
D. Cloud Spanner
Answer: (SHOW ANSWER)

NEW QUESTION: 40
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. Y^2
C. cos(X)
D. X^2+Y^2
Answer: D (LEAVE A REPLY)

NEW QUESTION: 41
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. Place the MariaDB instances in an Instance Group with a Health Check.
C. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.
D. Install the StackDriver Agent and configure the MySQL plugin.
Answer: (SHOW ANSWER)
The GitHub repository named google-fluentd-catch-all-config which includes the configuration files for the Logging agent for ingesting the logs from various third-party software packages.

NEW QUESTION: 42
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world.
The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments - development/test, staging, and production - to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements
* Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
* Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
* Provide reliable and timely access to data for analysis from distributed research workers
* Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
Technical Requirements
Ensure secure and efficient transport and storage of telemetry data
Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement

Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high- value problems instead of problems with our data pipelines.
MJTelco is building a custom interface to share data. They have these requirements:
1. They need to do aggregations over their petabyte-scale datasets.
2. They need to scan specific time range rows with a very fast response time (milliseconds).
Which combination of Google Cloud Platform products should you recommend?
A. Cloud Datastore and Cloud Bigtable
B. BigQuery and Cloud Bigtable
C. Cloud Bigtable and Cloud SQL
D. BigQuery and Cloud Storage
Answer: B (LEAVE A REPLY)

NEW QUESTION: 43
The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster ____.
A. application node
B. conditional node
C. master node
D. worker node
Answer: C (LEAVE A REPLY)
The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix-for example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m".

NEW QUESTION: 44
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 a partitioned table called tracking_table and include a TIMESTAMP column.
C. Create a table called tracking_table with a TIMESTAMP column to represent the day.
D. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 45
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: 46

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

A. PigLatin using Pig

B. Python using MapReduce

C. Java using MapReduce

D. HiveQL using Hive

Answer: B (LEAVE A REPLY)

NEW QUESTION: 47

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

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

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

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

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

Answer: C (LEAVE A REPLY)

Explanation/Reference:

NEW QUESTION: 48

You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity 'Movie' the property 'actors' and the property 'tags' have multiple values but the property 'date released' does not. A typical query would ask for all movies with actor=<actorname> ordered by date_released or all movies with tag=Comedy ordered by date_released. How should you avoid a combinatorial explosion in the number of indexes?
A. Set the following in your entity options: exclude_from_indexes = 'date_published'
B. Manually configure the index in your index config as follows:
C. Set the following in your entity options: exclude_from_indexes = 'actors, tags'
D. Manually configure the index in your index config as follows:
Answer: (SHOW ANSWER)

NEW QUESTION: 49
You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app_events_YYYYMMDD. You want to query all of the tables for the past 30 days in legacy SQL. What should you do?
A. Use the TABLE_DATE_RANGE function
B. Use the WHERE_PARTITIONTIME pseudo column
C. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
D. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD
Answer: A (LEAVE A REPLY)
Reference:
https://cloud.google.com/blog/products/gcp/using-bigquery-and-firebase-analytics-to-understandyour-mobile-app

NEW QUESTION: 50
You work for a mid-sized enterprise that needs to move its operational system transaction data from an on- premises database to GCP. The database is about 20 TB in size. Which database should you choose?
A. Cloud SQL
B. Cloud Bigtable
C. Cloud Spanner
D. Cloud Datastore
Answer: A (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 51
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)
Explanation
Partitioned tables include a pseudo column named _PARTITIONTIME that contains a date-based timestamp for data loaded into the table. To limit a query to particular partitions (such as Jan 1st and 2nd of 2017), use a clause similar to this:
WHERE _PARTITIONTIME BETWEEN TIMESTAMP('2017-01-01') AND TIMESTAMP('2017-01-02') Reference: https://cloud.google.com/bigquery/docs/partitioned-tables#the_partitiontime_pseudo_column

NEW QUESTION: 52
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning. Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?
A. Implement clustering in BigQuery on the package-tracking ID column.
B. Implement clustering in BigQuery on the ingest date column.
C. Re-create the table using data partitioning on the package delivery date.
D. Tier older data onto Cloud Storage files, and leverage extended tables.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 53
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. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.
B. Specify the TableReference object in the code.
C. Use .fromQuery operation to read specific fields from the table.
D. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.

Answer: (SHOW ANSWER)

NEW QUESTION: 54
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data.
What should you do?
A. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.
B. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. 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.
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 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.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 55
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: (SHOW ANSWER)
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: 56
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 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. Adjust the settings for each view to allow a related region-based security group view access.
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 all the tables are included in global dataset.
E. Ensure each table is included in a dataset for a region.
Answer: (SHOW ANSWER)

NEW QUESTION: 57
Which of these numbers are adjusted by a neural network as it learns from a training dataset (select 2 answers)?
A. Weights
B. Biases
C. Continuous features
D. Input values
Answer: (SHOW ANSWER)
Explanation

A neural network is a simple mechanism that's implemented with basic math. The only difference between the traditional programming model and a neural network is that you let the computer determine the parameters (weights and bias) by learning from training datasets.
Reference:
https://cloud.google.com/blog/big-data/2016/07/understanding-neural-networks-with-tensorflow-playground

NEW QUESTION: 58
A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3 minute period. You are in charge of the Voting restructure* and must ensure that the platform can handle the load and Hal all votes are processed. You must display partial results write voting is open. After voting doses you need to count the votes exactly once white optimizing cost. What should you do?
A. Create a Memorystore instance with a high availability (HA) configuration
B. Write votes to a Pub Sub tope and have Cloud Functions subscribe to it and write voles to BigQuery
C. Write votes to a Pub/Sub tope and toad into both Bigtable and BigQuery via a Dataflow pipeline Query Bigtable for real-time results and BigQuery for later analysis Shutdown the Bigtable instance when voting concludes
Answer: C (LEAVE A REPLY)
D Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas

NEW QUESTION: 59
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than
50,000 installations.
* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

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

Business Requirements

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

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data

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

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

CEO Statement

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

CTO Statement

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

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

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

CFO Statement

The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high- value problems instead of problems with our data pipelines.
You need to compose visualization for operations teams with the following requirements:
* Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once every minute)
* The report must not be more than 3 hours delayed from live data.
* The actionable report should only show suboptimal links.
* Most suboptimal links should be sorted to the top.
* Suboptimal links can be grouped and filtered by regional geography.
* User response time to load the report must be <5 seconds.
You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month. What should you do?
A. 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.
B. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.
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 series of charts and tables, one for each possible combination of criteria.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 60
You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)
A. There are very few occurrences of mutations relative to normal samples.
B. There are roughly equal occurrences of both normal and mutated samples in the database.
C. You expect future mutations to have different features from the mutated samples in the database.
D. You expect future mutations to have similar features to the mutated samples in the database.

E. You already have labels for which samples are mutated and which are normal in the database.

Answer: (SHOW ANSWER)

Explanation

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set.

https://en.wikipedia.org/wiki/Anomaly_detection

NEW QUESTION: 61

You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in How should you design your row key and tables to ensure that you can access the data with the most simple query?

A. Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.

B. Create one unique table for all of the indices, and then use the index and timestamp as the row key design

C. For each index, have a separate table and use a reverse timestamp as the row key design

D. For each index, have a separate table and use a timestamp as the row key design

Answer: B (LEAVE A REPLY)

NEW QUESTION: 62

Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?

A. Preemptible workers cannot use persistent disk.

B. Preemptible workers cannot store data.

C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.

D. A Dataproc cluster cannot have only preemptible workers.

Answer: B,D (LEAVE A REPLY)

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

NEW QUESTION: 63
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: 64
You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings.
Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?
A. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
B. Re-write the application to load accumulated data every 2 minutes.

C. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.
D. Convert the streaming insert code to batch load for individual messages.
Answer: (SHOW ANSWER)

NEW QUESTION: 65
Which of the following is not possible using primitive roles?
A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.
B. Give UserA owner access and UserB editor access for all datasets in a project.
C. Give a user access to view all datasets in a project, but not run queries on them.
D. Give GroupA owner access and GroupB editor access for all datasets in a project.
Answer: C (LEAVE A REPLY)
Explanation
Primitive roles can be used to give owner, editor, or viewer access to a user or group, but they can't be used to separate data access permissions from job-running permissions.
Reference: https://cloud.google.com/bigquery/docs/access-control#primitive_iam_roles

NEW QUESTION: 66
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 with the PubSub Kafka connector configured as a Sink connector. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.
C. Deploy 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.
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: 67
You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing operations and initial data are parquet files (on average 200-
400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is

very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.
What should you do?
A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
B. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.
C. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
D. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 68
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: (SHOW ANSWER)
https://cloud.google.com/blog/products/databases/getting-started-with-time-series-trend-predictions-using- gcp

NEW QUESTION: 69
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: (SHOW ANSWER)
ference: https://cloud.google.com/solutions/business-intelligence/

NEW QUESTION: 70

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: (SHOW ANSWER)

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: 71

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 Natural Language API

B. Cloud Speech-to-Text API

C. Cloud AutoML Natural Language

D. Dialogflow Enterprise Edition

Answer: D (LEAVE A REPLY)

NEW QUESTION: 72

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

What should you do?

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

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

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

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

Answer: (SHOW ANSWER)

NEW QUESTION: 73
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.

NEW QUESTION: 74
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 preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.
B. Increase the cluster size with more non-preemptible workers.
C. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.
D. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 75
You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure.
What should you do?
A. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
B. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
C. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

D. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 76
You are building new real-time data warehouse for your company and will use Google BigQuery streaming inserts. There is no guarantee that data will only be sent in once but you do have a unique ID for each row of data and an event timestamp. You want to ensure that duplicates are not included while interactively querying data. Which query type should you use?
A. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
B. Include ORDER BY DESK on timestamp column and LIMIT to 1.
C. Use the ROW_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.
D. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 77
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 is building a custom interface to share data. They have these requirements:
They need to do aggregations over their petabyte-scale datasets. They need to scan specific time range rows with a very fast response time (milliseconds). Which combination of Google Cloud Platform products should you recommend?
A. Cloud Datastore and Cloud Bigtable
B. Cloud Bigtable and Cloud SQL
C. BigQuery and Cloud Storage
D. BigQuery and Cloud Bigtable
Answer: D (LEAVE A REPLY)

NEW QUESTION: 78
When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied errors. What could be the reason for that?
A. Your gcloud does not have access to the BigQuery resources
B. BigQuery cannot be accessed from local machines
C. You are missing gcloud on your machine
D. Pipelines cannot be run locally
Answer: (SHOW ANSWER)
When reading from a Dataflow source or writing to a Dataflow sink using DirectPipelineRunner, the Cloud Platform account that you configured with the gcloud executable will need access to the corresponding source/sink

NEW QUESTION: 79
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: 80
Your company is implementing a data warehouse using BigQuery, and you have been tasked with designing the data model You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?
A. Shard the data by customer ID
B. Materialize the dimensional data in views
C. Partition the data by transaction date
D. Denormalize the data
Answer: (SHOW ANSWER)

NEW QUESTION: 81
The _____ for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.
A. Cloud Dataflow connector
B. DataFlow SDK
C. BiqQuery API
D. BigQuery Data Transfer Service
Answer: A (LEAVE A REPLY)
Explanation
The Cloud Dataflow connector for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline. You can use the connector for both batch and streaming operations.
Reference: https://cloud.google.com/bigtable/docs/dataflow-hbase

NEW QUESTION: 82
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.

NEW QUESTION: 83
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 --updateoption with the --jobNameset 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: A (LEAVE A REPLY)
Explanation/Reference: https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

NEW QUESTION: 84
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?
A. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
B. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
C. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 85
Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?
A. A sequential numeric ID
B. A timestamp followed by a stock symbol
C. A non-sequential numeric ID
D. A stock symbol followed by a timestamp
Answer: A,B (LEAVE A REPLY)
using a timestamp as the first element of a row key can cause a variety of problems.
In brief, when a row key for a time series includes a timestamp, all of your writes will target a single node; fill that node; and then move onto the next node in the cluster, resulting in hotspotting.
Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, since new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes. [https://cloud.google.com/bigtable/docs/schema-design]
Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotting

NEW QUESTION: 86
Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine.
The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks. What should you do?
A. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.
B. Host a visualization tool on a VM on Google Compute Engine.
C. Grant the user access to Google Cloud Shell.
D. Run a local version of Jupiter on the laptop.
Answer: (SHOW ANSWER)

NEW QUESTION: 87
You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.
You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)

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

NEW QUESTION: 88
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?
A. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
B. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table.
Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
C. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
D. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
Answer: (SHOW ANSWER)

NEW QUESTION: 89
Which of the following is not possible using primitive roles?
A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.
B. Give UserA owner access and UserB editor access for all datasets in a project.
C. Give a user access to view all datasets in a project, but not run queries on them.
D. Give GroupA owner access and GroupB editor access for all datasets in a project.
Answer: (SHOW ANSWER)
Primitive roles can be used to give owner, editor, or viewer access to a user or group, but they can't be used to separate data access permissions from job-running permissions.

NEW QUESTION: 90

Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?
A. Issue a command to restart the database servers.
B. Retry the query with exponential backoff, up to a cap of 15 minutes.
C. Retry the query every second until it comes back online to minimize staleness of data.
D. Reduce the query frequency to once every hour until the database comes back online.
Answer: B (LEAVE A REPLY)
https://cloud.google.com/sql/docs/mysql/manage-connections#backoff

NEW QUESTION: 91
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 Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"
C. Create a service account for the ODBC connection to use for authentication
D. Create a new view over events using standard SQL
E. Create a new partitioned table using a standard SQL query
Answer: A,C (LEAVE A REPLY)

NEW QUESTION: 92
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 Spanner 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. Use Cloud Dataflow to transform data to support query patterns.
D. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 93
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. 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
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. Assign the users/groups data viewer access at the table level for each table
Answer: B (LEAVE A REPLY)

NEW QUESTION: 94
You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing operations and initial data are parquet files (on average
200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.
What should you do?
A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
B. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.
C. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.
D. Switch from HDDs to SSDs, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
Answer: (SHOW ANSWER)

NEW QUESTION: 95

MJTelco Case Study

Company Overview

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

Company Background

Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.

Solution Concept

MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

Scale and harden their PoC to support significantly more data flows generated when they ramp to more
than 50,000 installations.

Refine their machine-learning cycles to verify and improve the dynamic models they use to control
topology definition.

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

Business Requirements

Scale up their production environment with minimal cost, instantiating resources when and where
needed in an unpredictable, distributed telecom user community.

Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

Provide reliable and timely access to data for analysis from distributed research workers

Maintain isolated environments that support rapid iteration of their machine-learning models without
affecting their customers.

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data
Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows
each.
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately
100m records/day
Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems
both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis.
Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region.
You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)
A. Adjust the settings for each dataset to allow a related region-based security group view access.
B. Ensure each table is included in a dataset for a region.
C. Ensure all the tables are included in global dataset.
D. Adjust the settings for each view to allow a related region-based security group view access.
E. Adjust the settings for each table to allow a related region-based security group view access.
Answer: B,D (LEAVE A REPLY)

NEW QUESTION: 96
Which of these sources can you not load data into BigQuery from?
A. File upload
B. Google Drive
C. Google Cloud Storage
D. Google Cloud SQL
Answer: (SHOW ANSWER)
You can load data into BigQuery from a file upload, Google Cloud Storage, Google Drive, or Google Cloud Bigtable. It is not possible to load data into BigQuery directly from Google Cloud SQL. One way to get data from Cloud SQL to BigQuery would be to export data from Cloud SQL to Cloud Storage and then load it from there.

NEW QUESTION: 97
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 ParDo transform in Cloud Dataflow to discard corrupt elements.
B. Add a SideInput that returns a Boolean if the element is corrupt.
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: (SHOW ANSWER)

NEW QUESTION: 98
Cloud Dataproc is a managed Apache Hadoop and Apache _____ service.
A. Blaze
B. Spark
C. Fire
D. Ignite
Answer: (SHOW ANSWER)
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: 99
Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?
A. Preemptible workers cannot use persistent disk.

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

NEW QUESTION: 100
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. Storage Transfer Service for the migration, Pub/Sub and Cloud Data Fusion for the real-time updates
B. gsutil for the migration; Pub/Sub and Dataflow for the real-time updates
C. gsutil for both the migration and the real-time updates
D. BigQuery Data Transfer Service lor the migration, Pub/Sub and Dataproc for the real-time updates
Answer: B (LEAVE A REPLY)

NEW QUESTION: 101
You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products of features of the platform. What should you do?
A. Export the information to Cloud Stackdriver, and set up an Alerting policy

B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver
C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs
D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs
Answer: (SHOW ANSWER)
Monitoring does not only provide you with access to Dataflow-related metrics, but also lets you to create alerting policies and dashboards so you can chart time series of metrics and choose to be notified when these metrics reach specified values.

NEW QUESTION: 102
To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?
A. gcloud ml-engine local train
B. gcloud ml-engine jobs submit training
C. gcloud ml-engine jobs submit training local
D. You can't run a TensorFlow program on your own computer using Cloud ML Engine .
Answer: A (LEAVE A REPLY)
gcloud ml-engine local train - run a Cloud ML Engine training job locally This command runs the specified module in an environment similar to that of a live Cloud ML Engine Training Job.
This is especially useful in the case of testing distributed models, as it allows you to validate that you are properly interacting with the Cloud ML Engine cluster configuration.
Reference: https://cloud.google.com/sdk/gcloud/reference/ml-engine/local/train

NEW QUESTION: 103
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 a new Cloud Pub/Sub subscription and cancel the old pipeline.
B. Update the current pipeline and use the drain flag.
C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
D. Update the current pipeline and provide the transform mapping JSON object.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 104

You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?
A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
B. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
C. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.
D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 105
You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application.
What should you do?
A. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset
B. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset
C. Create groups for your users and give those groups access to the dataset
D. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request
Answer: (SHOW ANSWER)

NEW QUESTION: 106
You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available.
How should you use this data to train the model?
A. Continuously retrain the model on just the new data.
B. Continuously retrain the model on a combination of existing data and the new data.
C. Train on the existing data while using the new data as your test set.
D. Train on the new data while using the existing data as your test set.
Answer: C (LEAVE A REPLY)
Explanation
https://cloud.google.com/automl-tables/docs/prepare

NEW QUESTION: 107
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 data pipeline to use BigQuery for storing stock trades, and update your application.
B. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
C. 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.
D. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 108
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 loaded in BigQuery is not flagged as CSV.
C. The CSV data has invalid rows that were skipped on import.
D. The CSV data has not gone through an ETL phase before loading into BigQuery.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 109

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. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
B. Use Cloud Deployment Manager to automate access provision.
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. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
Answer: (SHOW ANSWER)

NEW QUESTION: 110
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.

NEW QUESTION: 111
You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines. Which storage service and schema design should you use?
A. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
B. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.
C. Use Cloud Storage 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: 112
Your company receives both batch- and stream-based event dat
a. 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. Use watermarks and timestamps to capture the lagged data.
B. Set sliding windows to capture all the lagged data.
C. Set a single global window to capture all the data.
D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 113
You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?
A. The message body for the sensor event is too large.
B. Your custom endpoint has an out-of-date SSL certificate.
C. The Cloud Pub/Sub topic has too many messages published to it.
D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.
Answer: (SHOW ANSWER)

NEW QUESTION: 114
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 the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.
C. Store and process the entire dataset in BigQuery.
D. Store and process the entire dataset in Cloud Bigtable.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 115
Flowlogistic Case Study
Company Overview

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

Company Background

The company started as a regional trucking company, and then expanded into other logistics market.

Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

Use their proprietary technology in a real-time inventory-tracking system that indicates the location of
their loads

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

Databases

8 physical servers in 2 clusters

- SQL Server - user data, inventory, static data

3 physical servers

- Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

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

Storage appliances

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

10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

20 miscellaneous servers
- Jenkins, monitoring, bastion hosts,

Business Requirements

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

Aggregate data in a centralized Data Lake for analysis

Use historical data to perform predictive analytics on future shipments

Accurately track every shipment worldwide using proprietary technology

Improve business agility and speed of innovation through rapid provisioning of new resources

Analyze and optimize architecture for performance in the cloud

Migrate fully to the cloud if all other requirements are met

Technical Requirements

Handle both streaming and batch data

Migrate existing Hadoop workloads

Ensure architecture is scalable and elastic to meet the changing demands of the company.

Use managed services whenever possible

Encrypt data flight and at rest

Connect a VPN between the production data center and cloud environment

SEO Statement

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

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

CTO Statement

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

CFO Statement

Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability.

Additionally, I don't want to commit capital to building out a server environment.

Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the

messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?
A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
B. Use the NOW () function in BigQuery to record the event's time.
C. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.
D. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 116
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 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.
B. 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.
C. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.
D. Build 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.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 117
You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second. What should you do?
A. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a 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.

B. 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.
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: (SHOW ANSWER)

NEW QUESTION: 118
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: (SHOW ANSWER)
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: 119
You work for a manufacturing plant that batches application log files together into a single log file once a
day at 2:00 AM. You have written a Google Cloud Dataflow job to process that log file. You need to make
sure the log file in processed once per day as inexpensively as possible. What should you do?

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

NEW QUESTION: 120
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. Increase the share of the test sample in the train-test split.
C. Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.
D. Try to collect more data and increase the size of your dataset.
Answer: (SHOW ANSWER)

NEW QUESTION: 121
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: 122
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. Create an anonymized sample of the data for the consultant to work with in a different project.
B. Create a service account and allow the consultant to log on with it.
C. Grant the consultant the Viewer role on the project.
D. Grant the consultant the Cloud Dataflow Developer role on the project.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 123
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: (SHOW ANSWER)
Explanation

NEW QUESTION: 124

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 assigned the timestamp, which causes the job to fail
C. They have not applied a non-global windowing function, which causes the job to fail when the pipeline
is created
D. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
Answer: A (LEAVE A REPLY)

NEW QUESTION: 125
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. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
B. Use Cloud Deployment Manager to automate access provision.
C. Introduce resource hierarchy to leverage access control policy inheritance.
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. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
Answer: B,E (LEAVE A REPLY)

NEW QUESTION: 126

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

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

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

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

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

Answer: (SHOW ANSWER)

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

NEW QUESTION: 127

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: 128

Which of the following IAM roles does your Compute Engine account require to be able to run pipeline jobs?

A. dataflow.worker

B. dataflow.compute

C. dataflow.developer

D. dataflow.viewer

Answer: A (LEAVE A REPLY)

Explanation

The dataflow.worker role provides the permissions necessary for a Compute Engine service account to execute work units for a Dataflow pipeline Reference: https://cloud.google.com/dataflow/access-control

NEW QUESTION: 129

You are using Google BigQuery as your data warehouse. Your users report that the following simple query is running very slowly, no matter when they run the query:
SELECT country, state, city FROM [myproject:mydataset.mytable] GROUP BY country You check the query plan for the query and see the following output in the Read section of Stage:1:
What is the most likely cause of the delay for this query?
A. Users are running too many concurrent queries in the system
B. Most rows in the [myproject:mydataset.mytable]table have the same value in the country column, causing data skew
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: A (LEAVE A REPLY)

NEW QUESTION: 130
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 dat
a. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? Choose 2 answers.
A. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
B. 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.
C. Preserve the structure of the data as much as possible.
D. Use BigQuery UPDATE to further reduce the size of the dataset.
E. Denormalize the data as must as possible.
Answer: (SHOW ANSWER)

NEW QUESTION: 131
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'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 Load Balancing, Cloud Dataflow, and Cloud Storage
D. Cloud Pub/Sub, Cloud SQL, and Cloud Storage

Answer: (SHOW ANSWER)

NEW QUESTION: 132
You're training a model to predict housing prices based on an available dataset with real estate properties.
Your plan is to train a fully connected neural net, and you've discovered that the dataset contains latitude and longitude of the property. Real estate professionals have told you that the location of the property is highly influential on price, so you'd like to engineer a feature that incorporates this physical dependency.
What should you do?
A. Provide latitude and longitude as input vectors to your neural net.
B. Create a numeric column from a feature cross of latitude and longitude.
C. Create a feature cross of latitude and longitude, bucketize at the minute level and use L1 regularization during optimization.
D. Create a feature cross of latitude and longitude, bucketize it at the minute level and use L2 regularization during optimization.
Answer: (SHOW ANSWER)
Explanation
Reference https://cloud.google.com/bigquery/docs/gis-data

NEW QUESTION: 133
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: (SHOW ANSWER)
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: 134
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)
Explanation
Regression is the supervised learning task for modeling and predicting continuous, numeric variables.
Examples include predicting real-estate prices, stock price movements, or student test scores.
Classification is the supervised learning task for modeling and predicting categorical variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades.
Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset. Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.
Reference: https://elitedatascience.com/machine-learning-algorithms

NEW QUESTION: 135
Your company is implementing a data warehouse using BigQuery and you have been tasked with designing the data model You move your on-premises sales data warehouse with a star data schema to BigQuery but notice performance issues when querying the data of the past 30 days Based on Google's recommended practices, what should you do to speed up the query without increasing storage costs?
A. Partition the data by transaction date
B. Materialize the dimensional data in views
C. Shard the data by customer ID
D. Denormalize the data
Answer: A (LEAVE A REPLY)

NEW QUESTION: 136
You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?
A. Deploy the Cloud SQL Proxy on the Cloud Dataproc master
B. Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet
C. Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter
D. Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role
Answer: C (LEAVE A REPLY)
https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions

NEW QUESTION: 137
You are planning to migrate your current on-premises Apache Hadoop deployment to the cloud. You need to ensure that the deployment is as fault-tolerant and cost-effective as possible for long-running batch jobs. You want to use a managed service. What should you do?
A. Deploy a Cloud Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://
B. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances.
Store data in HDFS. Change references in scripts from hdfs:// to gs://
C. Deploy a Cloud Dataproc cluster. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://
D. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from hdfs:// to gs://
Answer: (SHOW ANSWER)

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)
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: 139
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: 140
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. Create an anonymized sample of the data for the consultant to work with in a different project.
B. Create a service account and allow the consultant to log on with it.
C. Grant the consultant the Viewer role on the project.
D. Grant the consultant the Cloud Dataflow Developer role on the project.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 141
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: D (LEAVE A REPLY)
When query results are retrieved from a cached results table, you are not charged for the query.
BigQuery caches query results for 24 hours, not 48 hours.
Query results are not cached if you specify a destination table.
A query's results are always cached except under certain conditions, such as if you specify a destination table.

NEW QUESTION: 142
You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:
- No interaction by the user on the site for 1 hour
- Has added more than $30 worth of products to the basket
- Has not completed a transaction
You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?
A. Use a fixed-time window with a duration of 60 minutes.
B. Use a sliding time window with a duration of 60 minutes.
C. Use a session window with a gap time duration of 60 minutes.
D. Use a global window with a time based trigger with a delay of 60 minutes.
Answer: C (LEAVE A REPLY)
It will send a message per user after that user is inactive for 60 minutes. Session window works well for capturing a session per user basis.

NEW QUESTION: 143
You work for an advertising company, and you've developed a Spark ML model to predict click-through rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?
A. Use Cloud ML Engine for training existing Spark ML models
B. Rewrite your models on TensorFlow, and start using Cloud ML Engine
C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery
D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery
Answer: (SHOW ANSWER)
https://cloud.google.com/dataproc/docs/tutorials/bigquery-sparkml

NEW QUESTION: 144

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is 10 TB. You need to design a solution that is secure and time-efficient. What should you do?
A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
C. Export the records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.
D. Export the records from the database as an Avro file. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.
Answer: (SHOW ANSWER)
Google recommends that enterprises use Transfer Appliance in cases where it would take them over a week to upload data to the cloud via the internet, or when an enterprise needs to migrate over 60 TB of data.

NEW QUESTION: 145
You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:
* Decoupling producer from consumer
* Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
* Near real-time SQL query
* Maintain at least 2 years of historical data, which will be queried with SQL Which pipeline should you use to meet these requirements?
A. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.
B. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
C. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
D. Create an application that 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: (SHOW ANSWER)

NEW QUESTION: 146

To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?
A. gcloud ml-engine local train
B. gcloud ml-engine jobs submit training
C. gcloud ml-engine jobs submit training local
D. You can't run a TensorFlow program on your own computer using Cloud ML Engine .
Answer: A (LEAVE A REPLY)
gcloud ml-engine local train - run a Cloud ML Engine training job locally
This command runs the specified module in an environment similar to that of a live Cloud ML Engine Training Job.
This is especially useful in the case of testing distributed models, as it allows you to validate that you are properly interacting with the Cloud ML Engine cluster configuration.

NEW QUESTION: 147
You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors.
You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the most cost-effective way. What should you do?
A. Write a Cloud Dataflow pipeline that aggregates all data in session windows.
B. Establish a Cloud Interconnect between all remote data centers and Google.
C. Deploy small Kafka clusters in your data centers to buffer events.
D. Have the data acquisition devices publish data to Cloud Pub/Sub.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 148
Flowlogistic is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster. A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?
A. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
B. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.
C. Use the NOW () function in BigQuery to record the event's time.
D. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.
Answer: B (LEAVE A REPLY)
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: 149
Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp. You want to collate those bid events into a single location in real time to determine which user bid first. What should you do?
A. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.
B. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
C. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
D. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull
Answer: C (LEAVE A REPLY)
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.

NEW QUESTION: 150
You have spent a few days loading data from comma-separated values (CSV) files into the Google
BigQuery table CLICK_STREAM. The column DTstores the epoch time of click events. For convenience,
you chose a simple schema where every field is treated as the STRINGtype. 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. Add a column TSof the TIMESTAMPtype to the table CLICK_STREAM, and populate the numeric
values from the column TSfor each row. Reference the column TSinstead of the column DTfrom now
on.
B. Delete the table CLICK_STREAM, and then re-create it such that the column DTis of the TIMESTAMP
type. Reload the data.
C. Add two columns to the table CLICK STREAM: TSof the TIMESTAMPtype and IS_NEWof the
BOOLEANtype. Reload all data in append mode. For each appended row, set the value of IS_NEWto
true. For future queries, reference the column TSinstead of the column DT, with the WHEREclause
ensuring that the value of IS_NEWmust be true.
D. Create a view CLICK_STREAM_V, where strings from the column DTare cast into TIMESTAMPvalues.
Reference the view CLICK_STREAM_Vinstead of the table CLICK_STREAMfrom 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 DTinto TIMESTAMPvalues. Run the query into a destination table
NEW_CLICK_STREAM, in which the column TSis the TIMESTAMPtype. Reference the table
NEW_CLICK_STREAMinstead of the table CLICK_STREAMfrom now on. In the future, new data is
loaded into the table NEW_CLICK_STREAM.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 151
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 bq query - -maximum_bytes_billed flag to restrict the number of bytes billed.
B. Use the LIMIT keyword to reduce the number of rows returned.
C. Create a separate table for each ID.
D. Recreate the table with a partitioning column and clustering column.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 152
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: (SHOW ANSWER)
Explanation/Reference:
Reference https://cloud.google.com/blog/products/gcp/busting-12-myths-about-bigquery

NEW QUESTION: 153

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

NEW QUESTION: 154
When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?
A. 500 TB
B. 1 GB
C. 1 TB
D. 500 GB
Answer: C (LEAVE A REPLY)
Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions. It is not a good solution for less than 1 TB of data.

NEW QUESTION: 155
Your company is streaming real-time sensor data from their factory floor into Bigtable and they have
noticed extremely poor performance. How should the row key be redesigned to improve Bigtable
performance on queries that populate real-time dashboards?
A. Use a row key of the form <timestamp>#<sensorid>.
B. Use a row key of the form >#<sensorid>#<timestamp>.
C. Use a row key of the form <timestamp>.
D. Use a row key of the form <sensorid>.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 156
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: (SHOW ANSWER)
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.

NEW QUESTION: 157
Flowlogistic Case Study
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market.
Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.
Solution Concept
Flowlogistic wants to implement two concepts using the cloud:
Use their proprietary technology in a real-time inventory-tracking system that indicates the location of
their loads

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

Databases

8 physical servers in 2 clusters

- SQL Server - user data, inventory, static data

3 physical servers

- Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

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

Storage appliances

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

10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements

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

Aggregate data in a centralized Data Lake for analysis

Use historical data to perform predictive analytics on future shipments

Accurately track every shipment worldwide using proprietary technology

Improve business agility and speed of innovation through rapid provisioning of new resources

Analyze and optimize architecture for performance in the cloud

Migrate fully to the cloud if all other requirements are met

Technical Requirements

Handle both streaming and batch data

Migrate existing Hadoop workloads

Ensure architecture is scalable and elastic to meet the changing demands of the company.

Use managed services whenever possible

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

NEW QUESTION: 158
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. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer
Service to transfer on-premices data to the designated storage bucket.
D. Introduce data compression for each file to increase the rate file of file transfer.
E. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in
parallel.
Answer: C,E (LEAVE A REPLY)

NEW QUESTION: 159
The marketing team at your organization provides regular updates of a segment of your customer dataset. The marketing team has given you a CSV with 1 million records that must be updated in BigQuery. When you use the UPDATE statement in BigQuery, you receive a quotaExceeded error. What should you do?
A. Reduce the number of records updated each day to stay within the BigQuery UPDATE DML statement limit.
B. Increase the BigQuery UPDATE DML statement limit in the Quota management section of the Google Cloud Platform Console.
C. Import the new records from the CSV file into a new BigQuery table. Create a BigQuery job that merges the new records with the existing records and writes the results to a new BigQuery table.
D. Split the source CSV file into smaller CSV files in Cloud Storage to reduce the number of BigQuery UPDATE DML statements per BigQuery job.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 160
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 an Authorized View with the provided query. Share the dataset that contains the view with the application service account.
C. Create a Cloud Dataflow pipeline using BigQueryIOto read results from the query. Grant the Dataflow Worker role to the application service account.
D. 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.
Answer: (SHOW ANSWER)

NEW QUESTION: 161
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: (SHOW ANSWER)
Explanation
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: 162
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: 163
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)
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/Trigger

NEW QUESTION: 164
You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application.
What should you do?
A. Create groups for your users and give those groups access to the dataset
B. 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. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query
request
D. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset
Answer: D (LEAVE A REPLY)

NEW QUESTION: 165
You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking

nightly batch files containing non- public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud Dataproc cluster, and depositing the results into Google BigQuery.
How should you securely run this workload?
A. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery
B. Use a service account with the ability to read the batch files and to write to BigQuery
C. Restrict the Google Cloud Storage bucket so only you can see the files
D. Grant the Project Owner role to a service account, and run the job with it
Answer: D (LEAVE A REPLY)

NEW QUESTION: 166
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.
Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
Scale and harden their PoC to support significantly more data flows generated when they ramp to more
than 50,000 installations.
Refine their machine-learning cycles to verify and improve the dynamic models they use to control
topology definition.
MJTelco will also use three separate operating environments - development/test, staging, and production
- to meet the needs of running experiments, deploying new features, and serving production customers.

Business Requirements
Scale up their production environment with minimal cost, instantiating resources when and where
needed in an unpredictable, distributed telecom user community.
Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
Provide reliable and timely access to data for analysis from distributed research workers
Maintain isolated environments that support rapid iteration of their machine-learning models without
affecting their customers.
Technical Requirements
Ensure secure and efficient transport and storage of telemetry data
Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately
100m records/day
Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis.
Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
You need to compose visualization for operations teams with the following requirements:
Telemetry must include data from all 50,000 installations for the most recent 6 weeks (sampling once
every minute)
The report must not be more than 3 hours delayed from live data.

The actionable report should only show suboptimal links.
Most suboptimal links should be sorted to the top.
Suboptimal links can be grouped and filtered by regional geography.
User response time to load the report must be <5 seconds.
You create a data source to store the last 6 weeks of data, and create visualizations that allow viewers to see multiple date ranges, distinct geographic regions, and unique installation types. You always show the latest data without any changes to your visualizations. You want to avoid creating and updating new visualizations each month.
What should you do?
A. Look through the current data and compose a series of charts and tables, one for each possible combination of criteria.
B. Load the data into relational database tables, write a Google App Engine application that queries all rows, summarizes the data across each criteria, and then renders results using the Google Charts and visualization API.
C. Look through the current data and compose a small set of generalized charts and tables bound to criteria filters that allow value selection.
D. 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.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 167
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: D (LEAVE A REPLY)
Explanation

NEW QUESTION: 168
You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?
A. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
B. Re-write the application to load accumulated data every 2 minutes.
C. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.
D. Convert the streaming insert code to batch load for individual messages.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 169
You are designing a pipeline that publishes application events to a Pub/Sub topic. You need to aggregate events across hourly intervals before loading the results to BigQuery for analysis. Your solution must be scalable so it can process and load large volumes of events to BigQuery. What should you do?
A. Create a Cloud Function to perform the necessary data processing that executes using the Pub/Sub trigger every time a new message is published to the topic.
B. Schedule a Cloud Function to run hourly, pulling all avertable messages from the Pub/Sub topic and performing the necessary aggregations
C. Create a streaming Dataflow job to continually read from the Pub/Sub topic and perform the necessary aggregations using tumbling windows
D. Schedule a batch Dataflow job to run hourly, pulling all available messages from the Pub-Sub topic and performing the necessary aggregations
Answer: C (LEAVE A REPLY)

NEW QUESTION: 170
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_model

NEW QUESTION: 171
Your weather app queries a database every 15 minutes to get the current temperature. The frontend is
powered by Google App Engine and server millions of users. How should you design the frontend to
respond to a database failure?
A. Issue a command to restart the database servers.
B. Retry the query every second until it comes back online to minimize staleness of data.
C. Reduce the query frequency to once every hour until the database comes back online.
D. Retry the query with exponential backoff, up to a cap of 15 minutes.
Answer: (SHOW ANSWER)

NEW QUESTION: 172
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)

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

NEW QUESTION: 173
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: 174
You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?
A. Cloud Dataflow
B. Cloud Scheduler
C. Cloud Composer
D. Cloud Functions
Answer: B (LEAVE A REPLY)

NEW QUESTION: 175
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: 176
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. cos(X)
C. Y^2
D. X^2
Answer: B (LEAVE A REPLY)

NEW QUESTION: 177
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: (SHOW ANSWER)
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.

NEW QUESTION: 178
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: (SHOW ANSWER)
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: 179
The Dataflow SDKs have been recently transitioned into which Apache service?
A. Apache Spark
B. Apache Hadoop
C. Apache Kafka
D. Apache Beam
Answer: D (LEAVE A REPLY)
Explanation
Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive
Reference: https://cloud.google.com/dataflow/docs/

NEW QUESTION: 180
Your company's on-premises Apache Hadoop servers are approaching end-of-life, and IT has decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage.
You want to minimize the storage cost of the migration. What should you do?
A. Put the data into Google Cloud Storage.
B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.

D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.
Answer: (SHOW ANSWER)
Explanation/Reference:
Reference: https://cloud.google.com/dataproc/

NEW QUESTION: 181
Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the eventsdata via an ODBC connection. You need to ensure the applications can connect. Which two actions should you take? (Choose two.)
A. Create a new partitioned table using a standard SQL query
B. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"
C. Create a new view over events using standard SQL
D. Create a new view over events_partitioned using standard SQL
E. Create a service account for the ODBC connection to use for authentication
Answer: B,C (LEAVE A REPLY)

NEW QUESTION: 182
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. Introduce resource hierarchy to leverage access control policy inheritance.
C. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
E. 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.
Answer: B,C (LEAVE A REPLY)
Google suggests that we should provide access by following google hierarchy and groups for users with similar roles.

NEW QUESTION: 183
You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems. Which solutions should you choose?
A. Dialogflow Enterprise Edition
B. Cloud Speech-to-Text API
C. Cloud Natural Language API
D. Cloud AutoML Natural Language
Answer: D (LEAVE A REPLY)

NEW QUESTION: 184
Which methods can be used to reduce the number of rows processed by BigQuery?
A. Splitting tables into multiple tables; putting data in partitions
B. Splitting tables into multiple tables; putting data in partitions; using the LIMIT clause
C. Putting data in partitions; using the LIMIT clause
D. Splitting tables into multiple tables; using the LIMIT clause
Answer: A (LEAVE A REPLY)
If you split a table into multiple tables (such as one table for each day), then you can limit your query to the data in specific tables (such as for particular days). A better method is to use a partitioned table, as long as your data can be separated by the day.
If you use the LIMIT clause, BigQuery will still process the entire table.
Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

NEW QUESTION: 185
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