Google.Professional-Data-Engineer.v2024-01-19.q177

Exam Code:	Professional-Data-Engineer
Exam Name:	Google Certified Professional Data Engineer Exam
Certification Provider:	Google
Free Question Number:	177
Version:	v2024-01-19
# of views:	1342
# of Questions views:	1770
https://www.freeqas.com/qa/Google/Professional-Data-Engineer/Google.Professional-Data-Engineer.v2024-01-19.q177.html	

NEW QUESTION: 1
For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?
A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.
Answer: C (LEAVE A REPLY)
Explanation
It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance, If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.
It is recommended to avoid creating your Compute Engine instance in a different region from your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.
Reference: https://cloud.google.com/bigtable/docs/creating-compute-instance

NEW QUESTION: 2
Case Study: 2 - MJTelco
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background

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

Solution Concept

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

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

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

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

Business Requirements

Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community. Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

Provide reliable and timely access to data for analysis from distributed research workers Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.

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

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

CEO Statement

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

CTO Statement

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

CFO Statement

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

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

NEW QUESTION: 3
A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with ingest-date partitioning.
Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?
A. Implement clustering in BigQuery on the ingest date column.
B. Tier older data onto Google Cloud Storage files and create a BigQuery table using GCS as an external data source.
C. Re-create the table using data partitioning on the package delivery date.
D. Implement clustering in BigQuery on the package-tracking ID column.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 4
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. Use an ETL tool to load the data from MySQL into Google BigQuery.
B. Connect an on-premises Apache Hadoop cluster to MySQL and perform ETL.
C. Mount the backups to Google Cloud SQL, and then process the data using Google Cloud Dataproc.
D. Add a node to the MySQL cluster and build an OLAP cube there.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 5
Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing on-premises applications every day. What should they do?
A. Execute gsutil rsync from the on-premises servers.
B. Write a job template in Cloud Dataproc to perform the data transfer.
C. Use Cloud Dataflow and write the data to Cloud Storage.
D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 6
Your company needs to upload their historic data to Cloud Storage. The security rules don't allow access from external IPs to their on-premises resources. After an initial upload, they will add new data from existing on-premises applications every day. What should they do?
A. Write a job template in Cloud Dataproc to perform the data transfer.
B. Execute gsutil rsyncfrom the on-premises servers.
C. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.
D. Use Cloud Dataflow and write the data to Cloud Storage.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 7
When creating a new Cloud Dataproc cluster with the projects.regions.clusters.create operation, these four values are required: project, region, name, and ____.
A. zone
B. node
C. label
D. type
Answer: A (LEAVE A REPLY)
At a minimum, you must specify four values when creating a new cluster with the projects.regions.clusters.create operation:
The project in which the cluster will be created
The region to use
The name of the cluster

The zone in which the cluster will be created
You can specify many more details beyond these minimum requirements. For example, you can also specify the number of workers, whether preemptible compute should be used, and the network settings.

NEW QUESTION: 8
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

NEW QUESTION: 9
Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file.
What is the most likely cause of this problem?
A. The CSV data loaded in BigQuery is not flagged as CSV.
B. The CSV data loaded in BigQuery is not using BigQuery's default encoding.
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: 10
You are designing a cloud-native historical data processing system to meet the following conditions:
* The data being analyzed is in CSV, Avro, and PDF formats and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.
* A streaming data pipeline stores new data daily.
* Peformance is not a factor in the solution.
* The solution design should maximize availability.
How should you design data storage for this solution?
A. Store the data in BigQuery. Access the data using the BigQuery Connector on Cloud Dataproc and Compute Engine.
B. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.
C. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.

D. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 11
Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?
A. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
B. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created
C. They have not assigned the timestamp, which causes the job to fail
D. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
Answer: (SHOW ANSWER)

NEW QUESTION: 12
Your company built a TensorFlow neutral-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly.
What method can you employ to address this?
A. Threading
B. Serialization
C. Dropout Methods
D. Dimensionality Reduction
Answer: C (LEAVE A REPLY)
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877

NEW QUESTION: 13
Flowlogistic Case Study
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology

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

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

* Databases
* 8 physical servers in 2 clusters
* SQL Server - user data, inventory, static data
* 3 physical servers
* Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert

* Application servers - customer front end, middleware for order/customs
* 60 virtual machines across 20 physical servers
* Tomcat - Java services
* Nginx - static content
* Batch servers

Storage appliances

* iSCSI for virtual machine (VM) hosts
* Fibre Channel storage area network (FC SAN) - SQL server storage
* Network-attached storage (NAS) image storage, logs, backups
* 10 Apache Hadoop /Spark servers
* Core Data Lake
* Data analysis workloads
* 20 miscellaneous servers
* Jenkins, monitoring, bastion hosts,

Business Requirements

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

Technical Requirements

* Handle both streaming and batch data

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

NEW QUESTION: 14
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. Partition 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. Cluster table data by create_date location_id and device_version
Answer: C (LEAVE A REPLY)

NEW QUESTION: 15
Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a Cassandra cluster on Google Compute Engine.

The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks. What should you do?

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

B. Grant the user access to Google Cloud Shell.

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

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

Answer: D (LEAVE A REPLY)

Datalab provides Jupyter for this kind of work.

NEW QUESTION: 16

Cloud Bigtable is Google's _____ Big Data database service.

A. Relational

B. mySQL

C. NoSQL

D. SQL Server

Answer: C (LEAVE A REPLY)

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: 17

Which of the following statements about Legacy SQL and Standard SQL is not true?

A. Standard SQL is the preferred query language for BigQuery.

B. If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.

C. One difference between the two query languages is how you specify fully-qualified table names (i.e. table names that include their associated project name).

D. You need to set a query language for each dataset and the default is Standard SQL.

Answer: D (LEAVE A REPLY)

You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.
Standard SQL has been the preferred query language since BigQuery 2.0 was released.
In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.
Due to the differences in syntax between the two query languages (such as with project-qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

NEW QUESTION: 18
Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?
A. Use K-means Clustering to detect faces in the pixels.
B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.
Answer: C (LEAVE A REPLY)
Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning. So deep is a strictly defined, technical term that means more than one hidden layer.
In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.
A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines. Feature engineering is difficult to perform on raw image data.
K-means Clustering is an unsupervised learning method used to categorize unlabeled data.
Reference: https://deeplearning4j.org/neuralnet-overview

NEW QUESTION: 19
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. Either the state or the city columns in the [myproject:mydataset.mytable]table have too many NULL values
C. Most rows in the [myproject:mydataset.mytable]table have the same value in the country column, causing data skew
D. The [myproject:mydataset.mytable] table has too many partitions
Answer: (SHOW ANSWER)

NEW QUESTION: 20
An organization maintains a Google BigQuery dataset that contains tables with user-level data. They want to expose aggregates of this data to other Google Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?
A. Create and share an authorized view that provides the aggregate results.
B. Create and share a new dataset and view that provides the aggregate results.
C. Create and share a new dataset and table that contains the aggregate results.
D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.
Answer: D (LEAVE A REPLY)
Explanation/Reference:
Reference: https://cloud.google.com/bigquery/docs/access-control

NEW QUESTION: 21
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 0s and 1s.

NEW QUESTION: 22
Which role must be assigned to a service account used by the virtual machines in a Dataproc cluster so they can execute jobs?
A. Dataproc Worker

B. Dataproc Viewer
C. Dataproc Runner
D. Dataproc Editor
Answer: A (LEAVE A REPLY)
Service accounts used with Cloud Dataproc must have Dataproc/Dataproc Worker role (or have all the permissions granted by Dataproc Worker role).
Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

NEW QUESTION: 23
Which SQL keyword can be used to reduce the number of columns processed by BigQuery?
A. BETWEEN
B. WHERE
C. SELECT
D. LIMIT
Answer: C (LEAVE A REPLY)
SELECT allows you to query specific columns rather than the whole table.
LIMIT, BETWEEN, and WHERE clauses will not reduce the number of columns processed by BigQuery.
Reference: https://cloud.google.com/bigquery/launch-checklist#architecture_design_and_development_checklist

NEW QUESTION: 24
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 SQ
Which pipeline should you use to meet these requirements?
A. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
B. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.
C. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.
D. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 25

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

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

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

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

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

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

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

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

NEW QUESTION: 26

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

A. details

B. value

C. null

D. id

Answer: B (LEAVE A REPLY)

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

Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-properties#formatting

NEW QUESTION: 27

Your team is working on a binary classification problem. You have trained a support vector machine (SVM) classifier with default parameters, and received an area under the Curve (AUC) of 0.87 on the validation set.

You want to increase the AUC of the model. What should you do?

A. Deploy the model and measure the real-world AUC; it's always higher because of generalization

B. Scale predictions you get out of the model (tune a scaling factor as a hyperparameter) in order to get the highest AUC

C. Perform hyperparameter tuning

D. Train a classifier with deep neural networks, because neural networks would always beat SVMs

Answer: (SHOW ANSWER)

NEW QUESTION: 28

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: (SHOW ANSWER)

Google suggests that we should provide access by following google hierarchy and groups for users with similar roles.

NEW QUESTION: 29
Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?
A. Threading
B. Serialization
C. Dropout Methods
D. Dimensionality Reduction
Answer: C (LEAVE A REPLY)
Explanation
Reference
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505

NEW QUESTION: 30
You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? Choose 2 answers.
A. Total outstanding messages exceed the 10-MB maximum.
B. Publisher throughput quota is too small.
C. Error handling in the subscriber code is not handling run-time errors properly.
D. The subscriber code cannot keep up with the messages.
E. The subscriber code does not acknowledge the messages that it pulls.
Answer: C,D (LEAVE A REPLY)

NEW QUESTION: 31
You are building a new application that you need to collect data from in a scalable way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:
* Decoupling producer from consumer
* Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely
* Near real-time SQL query
* Maintain at least 2 years of historical data, which will be queried with SQL Which pipeline should you use to meet these requirements?
A. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.
B. Create an application that 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 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.
D. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 32
Which role must be assigned to a service account used by the virtual machines in a Dataproc cluster so they can execute jobs?
A. Dataproc Worker
B. Dataproc Viewer
C. Dataproc Runner
D. Dataproc Editor
Answer: A (LEAVE A REPLY)
Explanation
Service accounts used with Cloud Dataproc must have Dataproc/Dataproc Worker role (or have all the permissions granted by Dataproc Worker role).
Reference: https://cloud.google.com/dataproc/docs/concepts/service-accounts#important_notes

NEW QUESTION: 33
You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?
A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code
D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code
Answer: A (LEAVE A REPLY)

References:

NEW QUESTION: 34
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 Dataproc, Cloud Dataflow, BigQuery
B. Cloud Dataflow, Cloud SQL, Cloud Spanner
C. Cloud Pub/Sub, Cloud Storage, BigQuery
D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL
Answer: B (LEAVE A REPLY)

NEW QUESTION: 35
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. BigQuery and Cloud Bigtable
B. Cloud Datastore and Cloud Bigtable
C. BigQuery and Cloud Storage
D. Cloud Bigtable and Cloud SQL
Answer: A (LEAVE A REPLY)

NEW QUESTION: 36
What are all of the BigQuery operations that Google charges for?
A. Storage, queries, and streaming inserts
B. Storage, queries, and loading data from a file
C. Storage, queries, and exporting data
D. Queries and streaming inserts

Answer: (SHOW ANSWER)
Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations.

NEW QUESTION: 37
You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time.
Which two methods can accomplish this? Choose 2 answers.
A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
B. Use managed exportm, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
C. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
D. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.
E. Write an application that uses Cloud Datastore client libraries to read all the entities. Format the exported data into a JSON file. Apply compression before storing the data in Cloud Source Repositories.
Answer: C,E (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 38
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?
A. Create a table in BigQuery, and append the new samples for CPU and memory to the table
B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second
C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second
D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.
Answer: (SHOW ANSWER)
https://cloud.google.com/bigtable/docs/schema-design-time-series

NEW QUESTION: 39
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)
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.

NEW QUESTION: 40
Your chemical company needs to manually check documentation for customer order. You use a pull subscription in Pub/Sub so that sales agents get details from the order. You must ensure that you do not process orders twice with different sales agents and that you do not add more complexity to this workflow. What should you do?
A. Create a transactional database that monitors the pending messages.
B. Create a new Pub/Sub push subscription to monitor the orders processed in the agent's system.
C. Use Pub/Sub exactly-once delivery in your pull subscription.
D. Use a Deduphcate PTransform in Dataflow before sending the messages to the sales agents.
Answer: (SHOW ANSWER)
Pub/Sub exactly-once delivery is a feature that guarantees that subscriptions do not receive duplicate deliveries of messages based on a Pub/Sub-defined unique message ID. This feature is only supported by the pull subscription type, which is what you are using in this scenario. By enabling exactly-once delivery, you can ensure that each order is processed only once by a sales agent, and that no order is lost or duplicated. This also simplifies your workflow, as you do not need to create a separate database or subscription to monitor the pending or processed messages. References:
* Exactly-once delivery | Cloud Pub/Sub Documentation
* Cloud Pub/Sub Exactly-once Delivery feature is now Generally Available (GA)

NEW QUESTION: 41
Suppose you have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error.
SELECT person FROM `project1.example.table1` WHERE city = "London"
How would you correct the error?
A. Add ", UNNEST(person)" before the WHERE clause.
B. Change "person" to "person.city".

C. Change "person" to "city.person".
D. Add ", UNNEST(city)" before the WHERE clause.
Answer: (SHOW ANSWER)
To access the person.city column, you need to "UNNEST(person)" and JOIN it to table1 using a comma.
Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql#nested_repeated_results

NEW QUESTION: 42
You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?
A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.
C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.
Answer: D (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 43
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 Functions
B. Cloud Dataflow
C. Cloud Composer
D. Cloud Scheduler
Answer: C (LEAVE A REPLY)

NEW QUESTION: 44
You have a data stored in BigQuery. The data in the BigQuery dataset must be highly available. You need to define a storage, backup, and recovery strategy of this data that minimizes cost. How should you configure the BigQuery table?
A. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.
B. Set the BigQuery dataset to be regional. In the event of an emergency, use a point-in-time snapshot to recover the data.

C. Set the BigQuery dataset to be multi-regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.
D. Set the BigQuery dataset to be regional. Create a scheduled query to make copies of the data to tables suffixed with the time of the backup. In the event of an emergency, use the backup copy of the table.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 45
An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application. They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?
A. BigQuery
B. Cloud SQL
C. Cloud BigTable
D. Cloud Datastore
Answer: C (LEAVE A REPLY)

NEW QUESTION: 46
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 Speech-to-Text API
B. Cloud AutoML Natural Language
C. Cloud Natural Language API
D. Dialogflow Enterprise Edition
Answer: B (LEAVE A REPLY)

NEW QUESTION: 47
You store and analyze your relational data in BigQuery on Google Cloud with all data that resides in US regions. You also have a variety of object stores across Microsoft Azure and Amazon Web Services

(AWS), also in US regions. You want to query all your data in BigQuery daily with as little movement of data as possible. What should you do?
A. Load files from AWS and Azure to Cloud Storage with Cloud Shell gautil rsync arguments.
B. Create a Dataflow pipeline to ingest files from Azure and AWS to BigQuery.
C. Use the BigQuery Omni functionality and BigLake tables to query files in Azure and AWS.
D. Use BigQuery Data Transfer Service to load files from Azure and AWS into BigQuery.
Answer: (SHOW ANSWER)
BigQuery Omni is a multi-cloud analytics solution that lets you use the BigQuery interface to analyze data stored in other public clouds, such as AWS and Azure, without moving or copying the data. BigLake tables are a type of external table that let you query structured data in external data stores with access delegation. By using BigQuery Omni and BigLake tables, you can query data in AWS and Azure object stores directly from BigQuery, with minimal data movement and consistent performance. References:
* 1: Introduction to BigLake tables
* 2: Deep dive on how BigLake accelerates query performance
* 3: BigQuery Omni and BigLake (Analytics Data Federation on GCP)

NEW QUESTION: 48
You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?
A. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.
B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataproc job.
C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 'none' using a Cloud Dataprep job.
D. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 49
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: 50
The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster ____.
A. application node
B. conditional node
C. master node
D. worker node
Answer: C (LEAVE A REPLY)
The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster master node. The cluster master-host-name is the name of your Cloud Dataproc cluster followed by an -m suffix-for example, if your cluster is named "my- cluster", the master-host-name would be "my-cluster-m".
Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web- interfaces#interfaces

NEW QUESTION: 51
The Development and External teams nave the project viewer Identity and Access Management (1AM) role m a folder named Visualization. You want the Development Team to be able to read data from both Cloud Storage and BigQuery, but the External Team should only be able to read data from BigQuery. What should you do?

A. Create a VPC Service Controls perimeter containing both protects and Cloud Storage as a restricted API. Add the Development Team users to the perimeter's Access Level
B. Create a VPC Service Controls perimeter containing both protects and BigQuery as a restricted API Add the External Team users to the perimeter s Access Level

C. Create Virtual Private Cloud (VPC) firewall rules on the acme-raw-data protect that deny all Ingress traffic from the External Team CIDR range
D. Remove Cloud Storage IAM permissions to the External Team on the acme-raw-data project
Answer: B (LEAVE A REPLY)

NEW QUESTION: 52
You are developing a software application using Google's Dataflow SDK, and want to use conditional, for loops and other complex programming structures to create a branching pipeline. Which component will be used for the data processing operation?
A. PCollection
B. Transform
C. Pipeline
D. Sink API
Answer: (SHOW ANSWER)
Explanation
In Google Cloud, the Dataflow SDK provides a transform component. It is responsible for the data processing operation. You can use conditional, for loops, and other complex programming structure to create a branching pipeline.
Reference: https://cloud.google.com/dataflow/model/programming-model

NEW QUESTION: 53
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 service account with the ability to read the batch files and to write to BigQuery
B. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery
C. Grant the Project Owner role to a service account, and run the job with it
D. Restrict the Google Cloud Storage bucket so only you can see the files
Answer: A (LEAVE A REPLY)

NEW QUESTION: 54
Which of these is not a supported method of putting data into a partitioned table?
A. If you have existing data in a separate file for each day, then create a partitioned table and upload each file into the appropriate partition.
B. Run a query to get the records for a specific day from an existing table and for the destination table, specify a partitioned table ending with the day in the format "$YYYYMMDD".
C. Create a partitioned table and stream new records to it every day.

D. Use ORDER BY to put a table's rows into chronological order and then change the table's type to "Partitioned".

Answer: D (LEAVE A REPLY)

You cannot change an existing table into a partitioned table. You must create a partitioned table from scratch. Then you can either stream data into it every day and the data will automatically be put in the right partition, or you can load data into a specific partition by using "$YYYYMMDD" at the end of the table name.

Reference: https://cloud.google.com/bigquery/docs/partitioned-tables

NEW QUESTION: 55

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 StackDriver Agent and configure the MySQL plugin.

B. Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.

C. Place the MariaDB instances in an Instance Group with a Health Check.

D. Install the StackDriver Logging Agent and configure fluentd in_tail plugin to read MariaDB logs.

Answer: (SHOW ANSWER)

NEW QUESTION: 56

Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?

A. An hourly watermark

B. An event time trigger

C. The with Allowed Lateness method

D. A processing time trigger

Answer: D (LEAVE A REPLY)

When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.

Processing time triggers. These triggers operate on the processing time ?the time when the data element is processed at any given stage in the pipeline. Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam's default trigger is event time-based.

Reference: https://beam.apache.org/documentation/programming-guide/#triggers

NEW QUESTION: 57

Scaling a Cloud Dataproc cluster typically involves ____.

A. increasing or decreasing the number of worker nodes

B. increasing or decreasing the number of master nodes

C. moving memory to run more applications on a single node

D. deleting applications from unused nodes periodically

Answer: A (LEAVE A REPLY)

Explanation

After creating a Cloud Dataproc cluster, you can scale the cluster by increasing or decreasing the number of worker nodes in the cluster at any time, even when jobs are running on the cluster. Cloud Dataproc clusters are typically scaled to:

1) increase the number of workers to make a job run faster

2) decrease the number of workers to save money

3) increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage

Reference: https://cloud.google.com/dataproc/docs/concepts/scaling-clusters

NEW QUESTION: 58

Which of these statements about exporting data from BigQuery is false?

A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.

B. The only supported export destination is Google Cloud Storage.

C. Data can only be exported in JSON or Avro format.

D. The only compression option available is GZIP.

Answer: (SHOW ANSWER)

Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported.

Reference: https://cloud.google.com/bigquery/docs/exporting-data

NEW QUESTION: 59

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

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

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

Answer: (SHOW ANSWER)

NEW QUESTION: 60

MJTelco Case Study

Company Overview

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

Company Background

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

Solution Concept

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

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

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

Business Requirements

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

Technical Requirements

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

CEO Statement

Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)
A. Ensure all the tables are included in global dataset.
B. Adjust the settings for each table to allow a related region-based security group view access.
C. Adjust the settings for each dataset to allow a related region-based security group view access.
D. Adjust the settings for each view to allow a related region-based security group view access.
E. Ensure each table is included in a dataset for a region.
Answer: (SHOW ANSWER)

NEW QUESTION: 61
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 an initialization action to execute the jobs
C. Create a Directed Acyclic Graph in Cloud Composer
D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster
Answer: A (LEAVE A REPLY)
Explanation/Reference: https://cloud.google.com/dataproc/docs/concepts/workflows/using-workflows

NEW QUESTION: 62
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 Dataproc for training existing Spark ML models, but start reading data directly from BigQuery
B. Rewrite your models on TensorFlow, and start using Cloud ML Engine
C. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery
D. Use Cloud ML Engine for training existing Spark ML models
Answer: D (LEAVE A REPLY)

NEW QUESTION: 63
Flowlogistic wants to use Google BigQuery as their primary analysis system, but they still have Apache Hadoop and Spark workloads that they cannot move to BigQuery. Flowlogistic does not know how to store the data that is common to both workloads. What should they do?
A. Store the common data encoded as Avro in Google Cloud Storage.
B. Store the common data in BigQuery as partitioned tables.
C. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.
D. Store the common data in BigQuery and expose authorized views.
Answer: (SHOW ANSWER)

NEW QUESTION: 64
You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes Cloud Dataproc and Cloud Dataflow jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day. Which tool should you use?
A. Cloud Scheduler
B. Cloud Composer
C. cron
D. Workflow Templates on Cloud Dataproc
Answer: D (LEAVE A REPLY)

NEW QUESTION: 65

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

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

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

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

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

Answer: D (LEAVE A REPLY)

Explanation

https://cloud.google.com/bigquery/docs/reference/standard-sql/analytic-function-concepts

NEW QUESTION: 66

Flowlogistic Case Study

Company Overview

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

Company Background

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

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

Databases

- 8 physical servers in 2 clusters
- SQL Server - user data, inventory, static data
- 3 physical servers
- Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert
Application servers - customer front end, middleware for order/customs
60 virtual machines across 20 physical servers
- Tomcat - Java services
- Nginx - static content
- Batch servers
Storage appliances
- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage
- Network-attached storage (NAS) image storage, logs, backups
10 Apache Hadoop /Spark servers
- Core Data Lake
- Data analysis workloads
20 miscellaneous servers
- Jenkins, monitoring, bastion hosts,
Business Requirements
Build a reliable and reproducible environment with scaled panty of production.
Aggregate data in a centralized Data Lake for analysis
Use historical data to perform predictive analytics on future shipments
Accurately track every shipment worldwide using proprietary technology
Improve business agility and speed of innovation through rapid provisioning of new resources
Analyze and optimize architecture for performance in the cloud
Migrate fully to the cloud if all other requirements are met
Technical Requirements
Handle both streaming and batch data
Migrate existing Hadoop workloads
Ensure architecture is scalable and elastic to meet the changing demands of the company.
Use managed services whenever possible
Encrypt data flight and at rest
Connect a VPN between the production data center and cloud environment
SEO Statement
We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.

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

NEW QUESTION: 67
Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data. Which three machine learning applications can you use?
(Choose three.)
A. Reinforcement learning to predict the location of a transaction.
B. Clustering to divide the transactions into N categories based on feature similarity.
C. Unsupervised learning to predict the location of a transaction.
D. Supervised learning to determine which transactions are most likely to be fraudulent.
E. Supervised learning to predict the location of a transaction.
F. Unsupervised learning to determine which transactions are most likely to be fraudulent.
Answer: A,B,F (LEAVE A REPLY)

NEW QUESTION: 68
By default, which of the following windowing behavior does Dataflow apply to unbounded data sets?
A. Windows at every 100 MB of data
B. Single, Global Window
C. Windows at every 1 minute
D. Windows at every 10 minutes
Answer: B (LEAVE A REPLY)

Dataflow's default windowing behavior is to assign all elements of a PCollection to a single, global window, even for unbounded PCollections Reference: https://cloud.google.com/dataflow/model/pcollection

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

NEW QUESTION: 70
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 SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD
D. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
Answer: A (LEAVE A REPLY)

NEW QUESTION: 71
What are two methods that can be used to denormalize tables in BigQuery?
A. 1) Split table into multiple tables; 2) Use a partitioned table
B. 1) Join tables into one table; 2) Use nested repeated fields
C. 1) Use a partitioned table; 2) Join tables into one table
D. 1) Use nested repeated fields; 2) Use a partitioned table
Answer: B (LEAVE A REPLY)
The conventional method of denormalizing data involves simply writing a fact, along with all its dimensions, into a flat table structure. For example, if you are dealing with sales transactions, you would write each

individual fact to a record, along with the accompanying dimensions such as order and customer information.
The other method for denormalizing data takes advantage of BigQuery's native support for nested and repeated structures in JSON or Avro input data. Expressing records using nested and repeated structures can provide a more natural representation of the underlying data. In the case of the sales order, the outer part of a JSON structure would contain the order and customer information, and the inner part of the structure would contain the individual line items of the order, which would be represented as nested, repeated elements.

NEW QUESTION: 72
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. Deploy small Kafka clusters in your data centers to buffer events.
B. Have the data acquisition devices publish data to Cloud Pub/Sub.
C. Establish a Cloud Interconnect between all remote data centers and Google.
D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.
Answer: (SHOW ANSWER)
Pubsub is global service with high message delivery capacity.

NEW QUESTION: 73
You want to rebuild your batch pipeline for structured data on Google Cloud You are using PySpark to conduct data transformations at scale, but your pipelines are taking over twelve hours to run To expedite development and pipeline run time, you want to use a serverless tool and SQL syntax You have already moved your raw data into Cloud Storage How should you build the pipeline on Google Cloud while meeting speed and processing requirements?
A. Ingest your data into Cloud SQL, convert your PySpark commands into SparkSQL queries to transform the data, and then use federated queries from BigQuery for machine learning.
B. Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipeline on Dataproc to write the data into BigQuery
C. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery
D. Ingest your data into BigQuery from Cloud Storage, convert your PySpark commands into BigQuery SQL queries to transform the data, and then write the transformations to a new table
Answer: B (LEAVE A REPLY)

NEW QUESTION: 74
After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want

to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.
What should you do?
A. Select random samples from the tables using the HASH() function and compare the samples.
B. Select random samples from the tables using the RAND() function and compare the samples.
C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.
D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 75
You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?
A. Use Transfer Appliance to copy the data to Cloud Storage
B. Use gsutil cp J to compress the content being uploaded to Cloud Storage
C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic
Answer: (SHOW ANSWER)
Huge amount of data with log network bandwidth, Transfer applicate is best for moving data over 100TB.

NEW QUESTION: 76
You create an important report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. You notice that visualizations are not showing data that is less than 1 hour old. What should you do?
A. Disable caching by editing the report settings.
B. Disable caching in BigQuery by editing table details.
C. Refresh your browser tab showing the visualizations.
D. Clear your browser history for the past hour then reload the tab showing the virtualizations.
Answer: A (LEAVE A REPLY)
https://support.google.com/datastudio/answer/7020039?hl=en

NEW QUESTION: 77
You are designing storage for very large text files for a data pipeline on Google Cloud. You want to support ANSI SQL queries. You also want to support compression and parallel load from the input locations using Google recommended practices. What should you do?
A. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
B. Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.
C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
D. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.
Answer: (SHOW ANSWER)
Avro is compressed format and dataflow for parallel pipeline and bigquery for storage.

NEW QUESTION: 78
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: (SHOW ANSWER)
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: 79
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 Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
C. Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.
D. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 80
Which Cloud Dataflow / Beam feature should you use to aggregate data in an unbounded data source every hour based on the time when the data entered the pipeline?
A. An hourly watermark
B. An event time trigger
C. The with Allowed Lateness method
D. A processing time trigger
Answer: D (LEAVE A REPLY)
When collecting and grouping data into windows, Beam uses triggers to determine when to emit the aggregated results of each window.
Processing time triggers. These triggers operate on the processing time - the time when the data element is processed at any given stage in the pipeline.
Event time triggers. These triggers operate on the event time, as indicated by the timestamp on each data element. Beam's default trigger is event time-based.
Reference: https://beam.apache.org/documentation/programming-guide/#triggers

NEW QUESTION: 81
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 dat
a. 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: B,C (LEAVE A REPLY)
References:

NEW QUESTION: 82
A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with latency under 100 milliseconds. You use the following query to generate predictions: SELECT predicted_label, user_id FROM ML.PREDICT (MODEL 'dataset.model', table user_features). How should you create the ML pipeline?
A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.

B. Create a Cloud Dataflow pipeline using BigQueryIO to read predictions for all users from the query. Write the results to Cloud Bigtable using BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.
C. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.
D. Create a Cloud Dataflow pipeline using BigQueryIO to read results from the query. Grant the Dataflow Worker role to the application service account.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 83
You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?
A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to `none' using a Cloud Dataproc job.
B. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to `none' using a Cloud Dataprep job.
C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.
D. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.
Answer: (SHOW ANSWER)

NEW QUESTION: 84
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, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
B. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.
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: 85
Your neural network model is taking days to train. You want to increase the training speed. What can you do?
A. Subsample your test dataset.

B. Subsample your training dataset.
C. Increase the number of input features to your model.
D. Increase the number of layers in your neural network.
Answer: D (LEAVE A REPLY)
Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

NEW QUESTION: 86
Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)
A. Disable writes to certain tables.
B. Restrict access to tables by role.
C. Ensure that the data is encrypted at all times.
D. Restrict BigQuery API access to approved users.
E. Segregate data across multiple tables or databases.
F. Use Google Stackdriver Audit Logging to determine policy violations.
Answer: B,D,F (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 87
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost. Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

MJTelco will also use three separate operating environments - development/test, staging, and production - to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements
* Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
* Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
* Provide reliable and timely access to data for analysis from distributed research workers
* Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
Technical Requirements
* Ensure secure and efficient transport and storage of telemetry data
* Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
* Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
* Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high- value problems instead of problems with our data pipelines.
You create a new report for your large team in Google Data Studio 360. The report uses Google BigQuery as its data source. It is company policy to ensure employees can view only the data associated with their region, so you create and populate a table for each region. You need to enforce the regional access policy to the data.
Which two actions should you take? (Choose two.)
A. Ensure each table is included in a dataset for a region.
B. Adjust the settings for each table to allow a related region-based security group view access.
C. Adjust the settings for each dataset to allow a related region-based security group view access.
D. Adjust the settings for each view to allow a related region-based security group view access.
E. Ensure all the tables are included in global dataset.
Answer: A,D (LEAVE A REPLY)

NEW QUESTION: 88
Which action can a Cloud Dataproc Viewer perform?
A. Submit a job.
B. Create a cluster.
C. Delete a cluster.
D. List the jobs.
Answer: D (LEAVE A REPLY)
Explanation
A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.
Reference:
https://cloud.google.com/dataproc/docs/concepts/iam#iam_roles_and_cloud_dataproc_operations_summary

NEW QUESTION: 89
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?
A. Create a table in BigQuery, and append the new samples for CPU and memory to the table
B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second
C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second
D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.
Answer: C (LEAVE A REPLY)
A tall and narrow table has a small number of events per row, which could be just one event, whereas a short and wide table has a large number of events per row. As explained in a moment, tall and narrow tables are best suited for time-series data. For time series, you should generally use tall and narrow tables. This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum (see Rows can be big but are not infinite).
https://cloud.google.com/bigtable/docs/schema-design-time-series#patterns_for_row_key_design

NEW QUESTION: 90
When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through a ____ proxy.
A. HTTPS
B. VPN

C. SOCKS
D. HTTP
Answer: C (LEAVE A REPLY)
When using Cloud Dataproc clusters, configure your browser to use the SOCKS proxy. The SOCKS proxy routes data intended for the Cloud Dataproc cluster through an SSH tunnel.
Reference: https://cloud.google.com/dataproc/docs/concepts/cluster-web- interfaces#interfaces

NEW QUESTION: 91
If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?
A. Do not use a production instance.
B. Run your test for at least 10 minutes.
C. Before you test, run a heavy pre-test for several minutes.
D. Use at least 300 GB of data.
Answer: (SHOW ANSWER)
If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test:
Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.
Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use 100 GB of data per node.
Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.
Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.

NEW QUESTION: 92
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. Materialize the dimensional data in views
B. Partition the data by transaction date
C. Denormalize the data
D. Shard the data by customer ID
Answer: (SHOW ANSWER)

NEW QUESTION: 93
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost. Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments - development/test, staging, and production - to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements
* Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community.
* Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
* Provide reliable and timely access to data for analysis from distributed research workers
* Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
Technical Requirements
* Ensure secure and efficient transport and storage of telemetry data
* Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
* Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day

* Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high- value problems instead of problems with our data pipelines.
MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?
A. Rowkey: date
Column data: device_id,data_point
B. Rowkey: date#device_id
Column data: data_point
C. Rowkey: device_id
Column data: date, data_point
D. Rowkey: data_point
Column data: device_id,date
E. Rowkey: date#data_point
Column data: device_id
Answer: D (LEAVE A REPLY)

NEW QUESTION: 94
Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?
A. The CSV data loaded in BigQuery is not using BigQuery's default encoding.
B. The CSV data has invalid rows that were skipped on import.
C. The CSV data loaded in BigQuery is not flagged as CSV.
D. The CSV data has not gone through an ETL phase before loading into BigQuery.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 95
You work for a manufacturing company that sources up to 750 different components, each from a different supplier. You've collected a labeled dataset that has on average 1000 examples for each unique component.
Your team wants to implement an app to help warehouse workers recognize incoming components based on a photo of the component. You want to implement the first working version of this app (as Proof-Of-Concept) within a few working days. What should you do?
A. Use Cloud Vision AutoML, but reduce your dataset twice.
B. Train your own image recognition model leveraging transfer learning techniques.
C. Use Cloud Vision API by providing custom labels as recognition hints.
D. Use Cloud Vision AutoML with the existing dataset.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 96
Flowlogistic Case Study
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.
Solution Concept
Flowlogistic wants to implement two concepts using the cloud:
- Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads
- Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.
Existing Technical Environment
Flowlogistic architecture resides in a single data center:
Databases
- 8 physical servers in 2 clusters
- SQL Server - user data, inventory, static data
3 physical servers

- Cassandra - metadata, tracking messages
10 Kafka servers - tracking message aggregation and batch insert
Application servers - customer front end, middleware for order/customs
▪
60 virtual machines across 20 physical servers
- Tomcat - Java services
- Nginx - static content
- Batch servers
Storage appliances
▪
- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage
- Network-attached storage (NAS) image storage, logs, backups
10 Apache Hadoop /Spark servers
▪
- Core Data Lake
- Data analysis workloads
20 miscellaneous servers
▪
- Jenkins, monitoring, bastion hosts,
Business Requirements
Build a reliable and reproducible environment with scaled panty of production.
▪
Aggregate data in a centralized Data Lake for analysis
▪
Use historical data to perform predictive analytics on future shipments
▪
Accurately track every shipment worldwide using proprietary technology
▪
Improve business agility and speed of innovation through rapid provisioning of new resources
▪
Analyze and optimize architecture for performance in the cloud
▪
Migrate fully to the cloud if all other requirements are met
▪
Technical Requirements
Handle both streaming and batch data
▪
Migrate existing Hadoop workloads
▪
Ensure architecture is scalable and elastic to meet the changing demands of the company.
▪
Use managed services whenever possible
▪
Encrypt data flight and at rest
▪
Connect a VPN between the production data center and cloud environment
▪
SEO Statement
We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.

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

NEW QUESTION: 97
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: B (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 98
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: C (LEAVE A REPLY)
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: 99
Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects. What should you do?
A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned_bytes
B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/ allocated_for_project
C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric
D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric
Answer: B (LEAVE A REPLY)
https://cloud.google.com/bigquery/docs/monitoring

NEW QUESTION: 100
Which of these statements about exporting data from BigQuery is false?
A. To export more than 1 GB of data, you need to put a wildcard in the destination filename.
B. The only supported export destination is Google Cloud Storage.
C. Data can only be exported in JSON or Avro format.
D. The only compression option available is GZIP.
Answer: (SHOW ANSWER)
Explanation
Data can be exported in CSV, JSON, or Avro format. If you are exporting nested or repeated data, then CSV format is not supported.
Reference: https://cloud.google.com/bigquery/docs/exporting-data

NEW QUESTION: 101
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. MySQL
B. MongoDB
C. Redis

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

NEW QUESTION: 102
You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?
A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name
C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code
D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code
Answer: D (LEAVE A REPLY)
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#changing_windowing

NEW QUESTION: 103
You need to migrate a Redis database from an on-premises data center to a Memorystore for Redis instance.
You want to follow Google-recommended practices and perform the migration for minimal cost. time, and effort. What should you do?
A. Make a secondary instance of the Redis database on a Compute Engine instance, and then perform a live cutover.
B. Write a shell script to migrate the Redis data, and create a new Memorystore for Redis instance.
C. Create a Dataflow job to road the Redis database from the on-premises data center. and write the data to a Memorystore for Redis instance
D. Make an RDB backup of the Redis database, use the gsutil utility to copy the RDB file into a Cloud Storage bucket, and then import the RDB tile into the Memorystore for Redis instance.
Answer: D (LEAVE A REPLY)
The import and export feature uses the native RDB snapshot feature of Redis to import data into or export data out of a Memorystore for Redis instance. The use of the native RDB format prevents lock-in and makes it very easy to move data within Google Cloud or outside of Google Cloud. Import and export uses Cloud Storage buckets to store RDB files. Reference:
https://cloud.google.com/memorystore/docs/redis/import-export-overview

NEW QUESTION: 104

You are developing a new deep teaming model that predicts a customer's likelihood to buy on your ecommerce site. Alter running an evaluation of the model against both the original training data and new test data, you find that your model is overfitting the data. You want to improve the accuracy of the model when predicting new data. What should you do?

A. Increase the size of the training dataset, and increase the number of input features.

B. Increase the size of the training dataset, and decrease the number of input features.

C. Reduce the size of the training dataset, and increase the number of input features.

D. Reduce the size of the training dataset, and decrease the number of input features.

Answer: B (LEAVE A REPLY)

https://machinelearningmastery.com/impact-of-dataset-size-on-deep-learning-model-skill-and-performance-estim

NEW QUESTION: 105

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)

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: 106

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)

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: 107
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: 108
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.

NEW QUESTION: 109
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost.

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

Solution Concept

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

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

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

Business Requirements

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

Technical Requirements

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

CEO Statement

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

CTO Statement

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

CFO Statement

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

Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
MJTelco's Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?
A. The disk size per worker
B. The number of workers
C. The maximum number of workers
D. The zone
Answer: D (LEAVE A REPLY)

NEW QUESTION: 110
You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis. Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)
A. Denormalize the data as must as possible.
B. Preserve the structure of the data as much as possible.
C. Use BigQuery UPDATE to further reduce the size of the dataset.
D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.
E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.
Answer: A,D (LEAVE A REPLY)
Denormalization will help in performance by reducing query time, update are not good with bigquery.

NEW QUESTION: 111
You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?
A. Use Transfer Appliance to copy the data to Cloud Storage
B. Use gsutil cp -Jto compress the content being uploaded to Cloud Storage
C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
D. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/ sec so it does not interfere with the production traffic
Answer: A (LEAVE A REPLY)
Explanation

NEW QUESTION: 112
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: 113
Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?
A. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.
B. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
C. The performance issue should be resolved over time as the site of the BigDate cluster is increased.
D. Redefine the schema by evenly distributing reads and writes across the row space of the table.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 114
An aerospace company uses a proprietary data format to store its night dat
a. You need to connect this new data source to BigQuery and stream the data into BigQuery. You want to efficiency import the data into BigQuery where consuming as few resources as possible. What should you do?
A. Use a standard Dataflow pipeline to store the raw data in BigQuery and then transform the format later when the data is used.
B. Write a shell script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source

C. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format
D. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format
Answer: C (LEAVE A REPLY)

NEW QUESTION: 115
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. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.
B. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.
C. Change the processing job to use Google Cloud Dataproc instead.
D. Manually start the Cloud Dataflow job each morning when you get into the office.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 116
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 to TFRecords formats (appr. 200MB per file) instead of parquet files.
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 from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.
Answer: D (LEAVE A REPLY)
In order to increase performance switch to SSD which will be costly, so to tackle this increase the boot disk size, bootsize is worker node cache size 100 Gb.

NEW QUESTION: 117
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: A,D (LEAVE A REPLY)

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: 118

Which of the following is NOT true about Dataflow pipelines?

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

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

C. Dataflow pipelines can be programmed in Java

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

Answer: (SHOW ANSWER)

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

NEW QUESTION: 119

You work for a large ecommerce company. You are using Pub/Sub to ingest the clickstream data to Google Cloud for analytics. You observe that when a new subscriber connects to an existing topic to analyze data, they are unable to subscribe to older data for an upcoming yearly sale event in two months, you need a solution that, once implemented, will enable any new subscriber to read the last 30 days of data. What should you do?

A. Create a new topic, and publish the last 30 days of data each time a new subscriber connects to an existing topic.

B. Set the topic retention policy to 30 days.

C. Set the subscriber retention policy to 30 days.

D. Ask the source system to re-push the data to Pub/Sub, and subscribe to it.

Answer: (SHOW ANSWER)

By setting the topic retention policy to 30 days, you can ensure that any new subscriber can access the messages that were published to the topic within the last 30 days1. This feature allows you to replay previously acknowledged messages or initialize new subscribers with historical data2. You can configure the topic retention policy by using the Cloud Console, the gcloud command-line tool, or the Pub/Sub API1. Option A is not efficient, as it requires creating a new topic and duplicating the data for each new subscriber, which would increase the storage costs and complexity. Option C is not effective, as it only affects the unacknowledged messages in a subscription, and does not allow new subscribers to access older data3. Option D is not feasible, as it depends on the source system's ability and willingness to re-push the data, and it may cause data duplication or inconsistency. References:

* 1: Create a topic | Cloud Pub/Sub Documentation | Google Cloud
* 2: Replay and purge messages with seek | Cloud Pub/Sub Documentation | Google Cloud
* 3: When is a PubSub Subscription considered to be inactive?

NEW QUESTION: 120
You are collecting IoT sensor data from millions of devices across the world and storing the data in BigQuery.
Your access pattern is based on recent data tittered by location_id and device_version with the following query:

You want to optimize your queries for cost and performance. How should you structure your data?
A. Cluster table data by create_date, partition by location 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 tocation_id and device_version
D. Partition table data by create_date, location_id and device_version
Answer: B (LEAVE A REPLY)

NEW QUESTION: 121
Which of the following is NOT one of the three main types of triggers that Dataflow supports?
A. Trigger based on element size in bytes
B. Trigger that is a combination of other triggers
C. Trigger based on element count
D. Trigger based on time
Answer: A (LEAVE A REPLY)
There are three major kinds of triggers that Dataflow supports: 1. Time-based triggers 2. Data-driven triggers.
You can set a trigger to emit results from a window when that window has received a certain number of data elements. 3. Composite triggers. These triggers combine multiple time-based or data-driven triggers in some logical way Reference: https://cloud.google.com/dataflow/model/triggers

NEW QUESTION: 122
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: B (LEAVE A REPLY)
Explanation
Reference https://cloud.google.com/bigquery/docs/gis-data

NEW QUESTION: 123
Cloud Dataproc charges you only for what you really use with _____ billing.
A. month-by-month
B. minute-by-minute
C. week-by-week
D. hour-by-hour
Answer: B (LEAVE A REPLY)
Explanation
One of the advantages of Cloud Dataproc is its low cost. Dataproc charges for what you really use with minute-by-minute billing and a low, ten-minute-minimum billing period.
Reference: https://cloud.google.com/dataproc/docs/concepts/overview

NEW QUESTION: 124
You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to minimize service costs. You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention. What should you do?
A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.
B. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.
C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.
D. Use Cloud Dataproc to run your transformations. Use the diagnosecommand to generate an operational output archive. Locate the bottleneck and adjust cluster resources.

Answer: (SHOW ANSWER)

NEW QUESTION: 125
Which of the following statements is NOT true regarding Bigtable access roles?
A. Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.
B. To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.
C. You can configure access control only at the project level.
D. To give a user access to only one table in a project, you must configure access through your application.
Answer: B (LEAVE A REPLY)
For Cloud Bigtable, you can configure access control at the project level. For example, you can grant the ability to:
Read from, but not write to, any table within the project.
Read from and write to any table within the project, but not manage instances.
Read from and write to any table within the project, and manage instances.
Reference: https://cloud.google.com/bigtable/docs/access-control

NEW QUESTION: 126
Case Study: 1 - Flowlogistic
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market. Because they have not updated their infrastructure, managing and tracking orders and shipments has become a bottleneck. To improve operations, Flowlogistic developed proprietary technology for tracking shipments in real time at the parcel level. However, they are unable to deploy it because their technology stack, based on Apache Kafka, cannot support the processing volume. In addition, Flowlogistic wants to further analyze their orders and shipments to determine how best to deploy their resources.
Solution Concept
Flowlogistic wants to implement two concepts using the cloud:
Use their proprietary technology in a real-time inventory-tracking system that indicates the location of their loads Perform analytics on all their orders and shipment logs, which contain both structured and unstructured data, to determine how best to deploy resources, which markets to expand info. They also want to use predictive analytics to learn earlier when a shipment will be delayed.
Existing Technical Environment
Flowlogistic architecture resides in a single data center:
Databases
8 physical servers in 2 clusters
SQL Server - user data, inventory, static data

3 physical servers
Cassandra - metadata, tracking messages
10 Kafka servers - tracking message aggregation and batch insert
Application servers - customer front end, middleware for order/customs 60 virtual machines across 20 physical servers Tomcat - Java services Nginx - static content Batch servers Storage appliances iSCSI for virtual machine (VM) hosts Fibre Channel storage area network (FC SAN) ?SQL server storage Network-attached storage (NAS) image storage, logs, backups Apache Hadoop /Spark servers Core Data Lake Data analysis workloads
20 miscellaneous servers
Jenkins, monitoring, bastion hosts,
Business Requirements
Build a reliable and reproducible environment with scaled panty of production. Aggregate data in a centralized Data Lake for analysis Use historical data to perform predictive analytics on future shipments Accurately track every shipment worldwide using proprietary technology Improve business agility and speed of innovation through rapid provisioning of new resources Analyze and optimize architecture for performance in the cloud Migrate fully to the cloud if all other requirements are met Technical Requirements Handle both streaming and batch data Migrate existing Hadoop workloads Ensure architecture is scalable and elastic to meet the changing demands of the company.
Use managed services whenever possible
Encrypt data flight and at rest
Connect a VPN between the production data center and cloud environment SEO Statement We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.
CTO Statement
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.
CFO Statement
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries. Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic's CEO wants to gain rapid insight into their customer base so his sales team can be better informed in the field. This team is not very technical, so they've purchased a visualization tool to simplify the creation of BigQuery reports. However, they've been overwhelmed by all the data in the table, and are spending a lot of money on queries trying to find the data they need. You want to solve their problem in the most cost-effective way. What should you do?
A. Create an additional table with only the necessary columns.

B. Create a view on the table to present to the virtualization tool.
C. Create identity and access management (IAM) roles on the appropriate columns, so only they appear in a query.
D. Export the data into a Google Sheet for virtualization.
Answer: (SHOW ANSWER)

NEW QUESTION: 127
Which of the following statements about Legacy SQL and Standard SQL is not true?
A. Standard SQL is the preferred query language for BigQuery.
B. If you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
C. One difference between the two query languages is how you specify fully-qualified table names (i.e. table names that include their associated project name).
D. You need to set a query language for each dataset and the default is Standard SQL.
Answer: (SHOW ANSWER)
You do not set a query language for each dataset. It is set each time you run a query and the default query language is Legacy SQL.
Standard SQL has been the preferred query language since BigQuery 2.0 was released.
In legacy SQL, to query a table with a project-qualified name, you use a colon, :, as a separator. In standard SQL, you use a period, ., instead.
Due to the differences in syntax between the two query languages (such as with project- qualified table names), if you write a query in Legacy SQL, it might generate an error if you try to run it with Standard SQL.
Reference:
https://cloud.google.com/bigquery/docs/reference/standard-sql/migrating-from-legacy-sql

NEW QUESTION: 128
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 Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.
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 a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 129
Suppose you have a dataset of images that are each labeled as to whether or not they contain a human face. To create a neural network that recognizes human faces in images using this labeled dataset, what approach would likely be the most effective?
A. Use K-means Clustering to detect faces in the pixels.
B. Use feature engineering to add features for eyes, noses, and mouths to the input data.
C. Use deep learning by creating a neural network with multiple hidden layers to automatically detect features of faces.
D. Build a neural network with an input layer of pixels, a hidden layer, and an output layer with two categories.
Answer: (SHOW ANSWER)
Traditional machine learning relies on shallow nets, composed of one input and one output layer, and at most one hidden layer in between. More than three layers (including input and output) qualifies as "deep" learning.
So deep is a strictly defined, technical term that means more than one hidden layer.
In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer.
A neural network with only one hidden layer would be unable to automatically recognize high-level features of faces, such as eyes, because it wouldn't be able to "build" these features using previous hidden layers that detect low-level features, such as lines.
Feature engineering is difficult to perform on raw image data.
K-means Clustering is an unsupervised learning method used to categorize unlabeled data.
Reference: https://deeplearning4j.org/neuralnet-overview

NEW QUESTION: 130
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.
Reference: https://cloud.google.com/bigquery/querying-data#query-caching

NEW QUESTION: 131
Your new customer has requested daily reports that show their net consumption of Google Cloud compute resources and who used the resources. You need to quickly and efficiently generate these daily reports. What should you do?
A. Do daily exports of Cloud Logging data to BigQuery. Create views filtering by project, log type, resource, and user.
B. Filter data in Cloud Logging by project, resource, and user; then export the data in CSV format.
C. Filter data in Cloud Logging by project, log type, resource, and user, then import the data into BigQuery.
D. Export Cloud Logging data to Cloud Storage in CSV format. Cleanse the data using Dataprep, filtering by project, resource, and user.
Answer: B (LEAVE A REPLY)
https://cloud.google.com/logging/docs/view/logs-explorer-interface?cloudshell=true

NEW QUESTION: 132
For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?
A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.
B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.
C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.
D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.
Answer: C (LEAVE A REPLY)
It is recommended to create your Compute Engine instance in the same zone as your Cloud Bigtable instance for the best possible performance,
If it's not possible to create a instance in the same zone, you should create your instance in another zone within the same region. For example, if your Cloud Bigtable instance is located in us-central1-b, you could create your instance in us-central1-f. This change may result in several milliseconds of additional latency for each Cloud Bigtable request.
It is recommended to avoid creating your Compute Engine instance in a different region from your Cloud Bigtable instance, which can add hundreds of milliseconds of latency to each Cloud Bigtable request.

NEW QUESTION: 133
Case Study 1 - Flowlogistic
Company Overview
Flowlogistic is a leading logistics and supply chain provider. They help businesses throughout the world manage their resources and transport them to their final destination. The company has grown rapidly, expanding their offerings to include rail, truck, aircraft, and oceanic shipping.
Company Background
The company started as a regional trucking company, and then expanded into other logistics market.

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

Solution Concept

Flowlogistic wants to implement two concepts using the cloud:

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

Existing Technical Environment

Flowlogistic architecture resides in a single data center:

* Databases

8 physical servers in 2 clusters

- SQL Server - user data, inventory, static data

3 physical servers

- Cassandra - metadata, tracking messages

10 Kafka servers - tracking message aggregation and batch insert

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

60 virtual machines across 20 physical servers

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

* Storage appliances

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

* 10 Apache Hadoop /Spark servers

- Core Data Lake
- Data analysis workloads

* 20 miscellaneous servers

- Jenkins, monitoring, bastion hosts,

Business Requirements

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

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

NEW QUESTION: 134
An online retailer has built their current application on Google App Engine. A new initiative at the company mandates that they extend their application to allow their customers to transact directly via the application. They need to manage their shopping transactions and analyze combined data from multiple datasets using a business intelligence (BI) tool. They want to use only a single database for this purpose. Which Google Cloud database should they choose?
A. BigQuery
B. Cloud SQL
C. Cloud BigTable

D. Cloud Datastore
Answer: C (LEAVE A REPLY)
Reference: https://cloud.google.com/solutions/business-intelligence/

NEW QUESTION: 135
Does Dataflow process batch data pipelines or streaming data pipelines?
A. Only Batch Data Pipelines
B. Both Batch and Streaming Data Pipelines
C. Only Streaming Data Pipelines
D. None of the above
Answer: B (LEAVE A REPLY)
Explanation
Dataflow is a unified processing model, and can execute both streaming and batch data pipelines
Reference: https://cloud.google.com/dataflow/

NEW QUESTION: 136
You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?
A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
B. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
C. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.
D. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
Answer: (SHOW ANSWER)

NEW QUESTION: 137

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

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

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

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

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

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

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

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

NEW QUESTION: 138
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)
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: 139
You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will not require programming or knowledge of SQL.
What should you do?
A. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.
B. Use Cloud Dataprep with recipes to detect errors and perform transformations.
C. Use Cloud Dataflow with Beam to detect errors and perform transformations.
D. Use federated tables in BigQuery with queries to detect errors and perform transformations.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 140
Which Java SDK class can you use to run your Dataflow programs locally?
A. LocalRunner
B. DirectPipelineRunner
C. MachineRunner
D. LocalPipelineRunner
Answer: B (LEAVE A REPLY)
DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization. Useful for small local execution and tests

NEW QUESTION: 141
You architect a system to analyze seismic dat
a. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?
A. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.
B. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.
C. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.
D. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 142
You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc

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

NEW QUESTION: 143
Your company built a TensorFlow neural-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly.
What method can you employ to address this?
A. Threading
B. Serialization
C. Dropout Methods
D. Dimensionality Reduction
Answer: (SHOW ANSWER)
https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877

NEW QUESTION: 144
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 and expose authorized views.
B. Store he common data in the HDFS storage for a Google Cloud Dataproc cluster.
C. Store the common data encoded as Avro in Google Cloud Storage.
D. Store the common data in BigQuery as partitioned tables.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 145
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: 146
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 Cloud Pub/Sub topic has too many messages published to it.
B. The message body for the sensor event is too large.
C. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.

D. Your custom endpoint has an out-of-date SSL certificate.
Answer: (SHOW ANSWER)

NEW QUESTION: 147
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: 148
Case Study: 2 - MJTelco
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost. Their management and operations teams are situated all around the globe creating many-to- many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

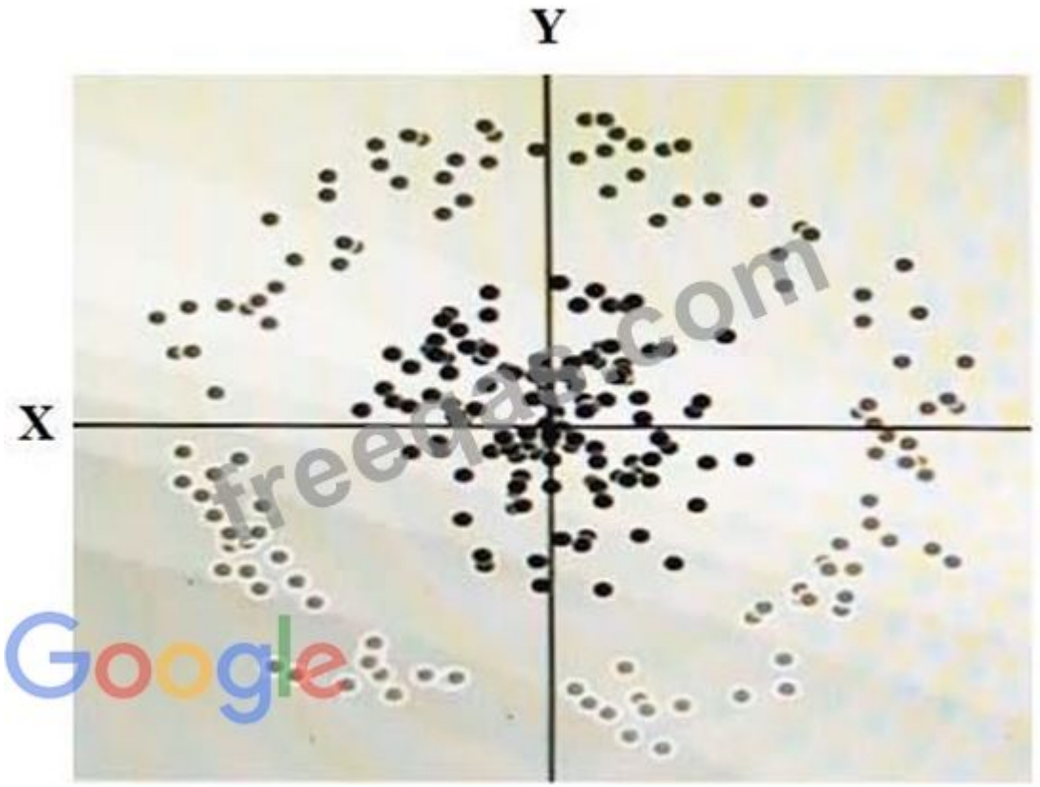
Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments ?development/test, staging, and production ? to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements
Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community. Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
Provide reliable and timely access to data for analysis from distributed research workers Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
Technical Requirements
Ensure secure and efficient transport and storage of telemetry data Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day
Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
You need to compose visualizations for operations teams with the following requirements:
Which approach meets the requirements?
A. Load the data into Google Cloud Datastore tables, write a Google App Engine Application that queries all rows, applies a function to derive the metric, and then renders results in a table using the Google charts and visualization API.
B. Load the data into Google BigQuery tables, write Google Apps Script that queries the data, calculates the metric, and shows only suboptimal rows in a table in Google Sheets.

C. Load the data into Google Sheets, use formulas to calculate a metric, and use filters/sorting to show only suboptimal links in a table.
D. Load the data into Google BigQuery tables, write a Google Data Studio 360 report that connects to your data, calculates a metric, and then uses a filter expression to show only suboptimal rows in a table.
Answer: (SHOW ANSWER)

NEW QUESTION: 149
Your company is using WHILECARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:
Syntax error : Expected end of statement but got "-" at [4:11] SELECT age FROM bigquery-public-data.noaa_gsod.gsod WHERE age != 99 AND_TABLE_SUFFIX = `1929' ORDER BY age DESC Which table name will make the SQL statement work correctly?
A. `bigquery-public-data.noaa_gsod.gsod*`
B. `bigquery-public-data.noaa_gsod.gsod`
C. bigquery-public-data.noaa_gsod.gsod*
D. `bigquery-public-data.noaa_gsod.gsod'*
Answer: C (LEAVE A REPLY)

NEW QUESTION: 150
Google Cloud Bigtable indexes a single value in each row. This value is called the _____.
A. primary key
B. unique key
C. row key
D. master key
Answer: C (LEAVE A REPLY)
Cloud Bigtable is a sparsely populated table that can scale to billions of rows and thousands of columns, allowing you to store terabytes or even petabytes of data. A single value in each row is indexed; this value is known as the row key.
Reference: https://cloud.google.com/bigtable/docs/overview

NEW QUESTION: 151
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: C (LEAVE A REPLY)

NEW QUESTION: 152
Your financial services company is moving to cloud technology and wants to store 50 TB of financial timeseries data in the cloud. This data is updated frequently and new data will be streaming in all the time. Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data.
Which product should they use to store the data?
A. Cloud Bigtable
B. Google BigQuery

C. Google Cloud Storage
D. Google Cloud Datastore
Answer: (SHOW ANSWER)
Reference: https://cloud.google.com/bigtable/docs/schema-design-time-series

NEW QUESTION: 153
Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs are being are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour.
The data scientists have written the following code to read the data for a new key features in the logs.
BigQueryIO.Read
.named("ReadLogData")
.from("clouddataflow-readonly:samples.log_data")
You want to improve the performance of this data read. What should you do?
A. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
B. Use .fromQuery operation to read specific fields from the table.
C. Specify the TableReference object in the code.
D. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 154
Your organization has two Google Cloud projects, project A and project B. In project A, you have a Pub/Sub topic that receives data from confidential sources. Only the resources in project A should be able to access the data in that topic. You want to ensure that project B and any future project cannot access data in the project A topic. What should you do?
A. Configure VPC Service Controls in the organization with a perimeter around the VPC of project A.
B. Add firewall rules in project A so only traffic from the VPC in project A is permitted.
C. Configure VPC Service Controls in the organization with a perimeter around project A.
D. Use Identity and Access Management conditions to ensure that only users and service accounts in project A can access resources in project.
Answer: D (LEAVE A REPLY)
Identity and Access Management (IAM) is the recommended way to control access to Pub/Sub resources, such as topics and subscriptions. IAM allows you to grant roles and permissions to users and service accounts at the project level or the individual resource level. You can also use IAM conditions to specify additional attributes for granting or denying access, such as time, date, or origin. By using IAM conditions, you can ensure that only the resources in project A can access the data in the project A topic, regardless of the network configuration or the VPC Service Controls. You can also prevent project B and any future project from accessing the data in the project A topic by not granting them any roles or permissions on the topic.

Option A is not a good solution, as VPC Service Controls are designed to prevent data exfiltration from Google Cloud resources to the public internet, not to control access between Google Cloud projects. VPC Service Controls create a perimeter around the resources of one or more projects, and restrict the communication with resources outside the perimeter. However, VPC Service Controls do not apply to Pub/Sub, as Pub/Sub is not associated with any specific IP address or VPC network. Therefore, configuring VPC Service Controls with a perimeter around the VPC of project A would not prevent project B or any future project from accessing the data in the project A topic, if they have the necessary IAM roles and permissions.
Option B is not a good solution, as firewall rules are used to control the ingress and egress traffic to and from the VPC network of a project. Firewall rules do not apply to Pub/Sub, as Pub/Sub is not associated with any specific IP address or VPC network. Therefore, adding firewall rules in project A to only permit traffic from the VPC in project A would not prevent project B or any future project from accessing the data in the project A topic, if they have the necessary IAM roles and permissions.
Option C is not a good solution, as VPC Service Controls are designed to prevent data exfiltration from Google Cloud resources to the public internet, not to control access between Google Cloud projects. VPC Service Controls create a perimeter around the resources of one or more projects, and restrict the communication with resources outside the perimeter. However, VPC Service Controls do not apply to Pub/Sub, as Pub/Sub is not associated with any specific IP address or VPC network. Therefore, configuring VPC Service Controls with a perimeter around project A would not prevent project B or any future project from accessing the data in the project A topic, if they have the necessary IAM roles and permissions. References: Access control with IAM | Cloud Pub/Sub Documentation | Google Cloud, [Using IAM Conditions | Cloud IAM Documentation | Google Cloud], [VPC Service Controls overview | Google Cloud], [Using VPC Service Controls | Google Cloud], [Pub/Sub tier capabilities | Memorystore for Redis | Google Cloud].

NEW QUESTION: 155
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. 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://
C. 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://
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: 156

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: 157
Scaling a Cloud Dataproc cluster typically involves ____.
A. increasing or decreasing the number of worker nodes
B. increasing or decreasing the number of master nodes
C. moving memory to run more applications on a single node
D. deleting applications from unused nodes periodically
Answer: (SHOW ANSWER)
After creating a Cloud Dataproc cluster, you can scale the cluster by increasing or decreasing the number of worker nodes in the cluster at any time, even when jobs are running on the cluster. Cloud Dataproc clusters are typically scaled to:
1) increase the number of workers to make a job run faster
2) decrease the number of workers to save money
3) increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage

NEW QUESTION: 158
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 Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"
B. Create a new view over events_partitioned using standard SQL
C. Create a new partitioned table using a standard SQL query
D. Create a new view over events using standard SQL
E. Create a service account for the ODBC connection to use for authentication
Answer: B,E (LEAVE A REPLY)

NEW QUESTION: 159
You are training a spam classifier. You notice that you are overfitting the training data. Which three actions can you take to resolve this problem? (Choose three.)

A. Get more training examples
B. Reduce the number of training examples
C. Use a smaller set of features
D. Use a larger set of features
E. Increase the regularization parameters
F. Decrease the regularization parameters
Answer: A,D,F (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 160
You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query - -dry_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?
A. Create a separate table for each ID.
B. Use the LIMIT keyword to reduce the number of rows returned.
C. Recreate the table with a partitioning column and clustering column.
D. Use the bq query - -maximum_bytes_billedflag to restrict the number of bytes billed.
Answer: B (LEAVE A REPLY)
Explanation

NEW QUESTION: 161
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: 162

You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants.
What should you do?
A. Match loan applicants with their social profiles to enable feature engineering.
B. Remove the bias from the data and collect applications that have been declined loans.
C. Train a linear regression to predict a credit default risk score.
D. Increase the size of the dataset by collecting additional data.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 163
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the "Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?
A. Use gcloud kms keys 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.
B. 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.
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. 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.
Answer: (SHOW ANSWER)

NEW QUESTION: 164
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: A,B (LEAVE A REPLY)
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.

NEW QUESTION: 165

An online brokerage company requires a high volume trade processing architecture. You need to create a secure queuing system that triggers jobs. The jobs will run in Google Cloud and cat the company's Python API to execute trades. You need to efficiently implement a solution. What should you do?
A. Write an application that makes a queue in a NoSQL database
B. Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.
C. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.
D. Write an application hosted on a Compute Engine instance that makes a push subscription to the Pub/Sub topic
Answer: A (LEAVE A REPLY)

NEW QUESTION: 166
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: 167
You have uploaded 5 years of log data to Cloud Storage A user reported that some data points in the log data are outside of their expected ranges, which indicates errors You need to address this issue and be able to run the process again in the future while keeping the original data for compliance reasons. What should you do?

A. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to a new dataset in Cloud Storage
B. Create a Cloud Dataflow workflow that reads the data from Cloud Storage, checks for values outside the expected range, sets the value to an appropriate default, and writes the updated records to the same dataset in Cloud Storage
C. Import the data from Cloud Storage into BigQuery Create a new BigQuery table, and skip the rows with errors.
D. Create a Compute Engine instance and create a new copy of the data in Cloud Storage Skip the rows with errors
Answer: B (LEAVE A REPLY)

NEW QUESTION: 168
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 failover replica in another zone within the same region.
B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.
C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.
D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.
Answer: (SHOW ANSWER)
https://cloud.google.com/sql/docs/mysql/high-availability

NEW QUESTION: 169
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 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.
B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
C. 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.
D. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 170
Data Analysts in your company have the Cloud IAM Owner role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for 6 months. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?
A. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.
B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' projects. Restrict access to the Cloud Storage bucket.
C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit logs. Restrict access to the project with the exported logs.
D. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.
Answer: D (LEAVE A REPLY)
https://cloud.google.com/iam/docs/roles-audit-logging#scenario_external_auditors

NEW QUESTION: 171
You have a data processing application that runs on Google Kubernetes Engine (GKE). Containers need to be launched with their latest available configurations from a container registry. Your GKE nodes need to have GPUs. local SSDs, and 8 Gbps bandwidth. You want to efficiently provision the data processing infrastructure and manage the deployment process. What should you do?
A. Use Compute Engi.no startup scriots to pull container Images, and use gloud commands to provision the infrastructure.
B. Use GKE to autoscale containers, and use gloud commands to provision the infrastructure.
C. Use Cloud Build to schedule a job using Terraform build to provision the infrastructure and launch with the most current container images.
D. Use Dataflow to provision the data pipeline, and use Cloud Scheduler to run the job.
Answer: C (LEAVE A REPLY)
https://cloud.google.com/architecture/managing-infrastructure-as-code

NEW QUESTION: 172
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. Have the data acquisition devices publish data to Cloud Pub/Sub.
B. Deploy small Kafka clusters in your data centers to buffer events.
C. Establish a Cloud Interconnect between all remote data centers and Google.

D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.
Answer: (SHOW ANSWER)

NEW QUESTION: 173
Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable. The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data. They want to improve this performance while minimizing cost. What should they do?
A. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
B. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.
C. The performance issue should be resolved over time as the site of the BigDate cluster is increased.
D. Redefine the schema by evenly distributing reads and writes across the row space of the table.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 174
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. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above
70% of max capacity.
B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.
C. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.
D. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.
E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.
Answer: (SHOW ANSWER)

NEW QUESTION: 175
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. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.

B. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.
C. Put the data into Google Cloud Storage.
D. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
Answer: (SHOW ANSWER)

NEW QUESTION: 176
You have an Oracle database deployed in a VM as part of a Virtual Private Cloud (VPC) network. You want to replicate and continuously synchronize 50 tables to BigQuery. You want to minimize the need to manage infrastructure. What should you do?
A. Create a Datastream service from Oracle to BigQuery, use a private connectivity configuration to the same VPC network, and a connection profile to BigQuery.
B. Create a Pub/Sub subscription to write to BigQuery directly Deploy the Debezium Oracle connector to capture changes in the Oracle database, and sink to the Pub/Sub topic.
C. Deploy Apache Kafka in the same VPC network, use Kafka Connect Oracle Change Data Capture (CDC), and Dataflow to stream the Kafka topic to BigQuery.
D O Deploy Apache Kafka in the same VPC network, use Kafka Connect Oracle change data capture (CDC), and the Kafka Connect Google BigQuery Sink Connector.
Answer: A (LEAVE A REPLY)
Datastream is a serverless, scalable, and reliable service that enables you to stream data changes from Oracle and MySQL databases to Google Cloud services such as BigQuery, Cloud SQL, Google Cloud Storage, and Cloud Pub/Sub. Datastream captures and streams database changes using change data capture (CDC) technology. Datastream supports private connectivity to the source and destination systems using VPC networks. Datastream also provides a connection profile to BigQuery, which simplifies the configuration and management of the data replication. References:
* Datastream overview
* Creating a Datastream stream
* Using Datastream with BigQuery

NEW QUESTION: 177
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, copy initial data from GCS to HDFS, run the Spark job and copy results back to GCS.
C. Switch to TFRecords formats (appr. 200MB per file) instead of parquet files.
D. Switch from HDDs to SSDs, override the preemptible VMs configuration to increase the boot disk size.

Answer: (SHOW ANSWER)