Google.Professional-Data-Engineer.v2026-09-10.q173

Exam Code:	Professional-Data-Engineer
Exam Name:	Google Certified Professional Data Engineer Exam
Certification Provider:	Google
Free Question Number:	173
Version:	v2026-09-10
# of views:	866
# of Questions views:	2219
https://www.dumpsfiles.com/files/Google/Professional-Data-Engineer/Google.Professional-Data-Engineer.v2026-09-10.q173	

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

60 virtual machines across 20 physical servers
- Tomcat - Java services
- Nginx - static content
- Batch servers
* Storage appliances
- iSCSI for virtual machine (VM) hosts
- Fibre Channel storage area network (FC SAN) - SQL server storage
- Network-attached storage (NAS) image storage, logs, backups
* 10 Apache Hadoop /Spark servers
- Core Data Lake
- Data analysis workloads
* 20 miscellaneous servers
- Jenkins, monitoring, bastion hosts,
Business Requirements
* Build a reliable and reproducible environment with scaled panty of production.
* Aggregate data in a centralized Data Lake for analysis
* Use historical data to perform predictive analytics on future shipments
* Accurately track every shipment worldwide using proprietary technology
* Improve business agility and speed of innovation through rapid provisioning of new resources
* Analyze and optimize architecture for performance in the cloud
* Migrate fully to the cloud if all other requirements are met
Technical Requirements
* Handle both streaming and batch data
* Migrate existing Hadoop workloads
* Ensure architecture is scalable and elastic to meet the changing demands of the company.
* Use managed services whenever possible
* Encrypt data flight and at rest
* Connect a VPN between the production data center and cloud environment SEO Statement We have grown so quickly that our inability to upgrade our infrastructure is really hampering further growth and efficiency. We are efficient at moving shipments around the world, but we are inefficient at moving data around.
We need to organize our information so we can more easily understand where our customers are and what they are shipping.
CTO Statement
IT has never been a priority for us, so as our data has grown, we have not invested enough in our technology. I have a good staff to manage IT, but they are so busy managing our infrastructure that I cannot get them to do the things that really matter, such as organizing our data, building the analytics, and figuring out how to implement the CFO' s tracking technology.
CFO Statement
Part of our competitive advantage is that we penalize ourselves for late shipments and deliveries.
Knowing where out shipments are at all times has a direct correlation to our bottom line and profitability. Additionally, I don't want to commit capital to building out a server environment.
Flowlogistic's management has determined that the current Apache Kafka servers cannot handle the data volume for their real-time inventory tracking system. You need to build a new system on Google Cloud Platform (GCP) that will feed the proprietary tracking

software. The system must be able to ingest data from a variety of global sources, process and query in real-time, and store the data reliably. Which combination of GCP products should you choose?

A. Cloud Load Balancing, Dataflow, and Cloud Storage

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

C. Dataflow, Cloud SQL, and Cloud Storage

D. Pub/Sub, Dataflow, and Local SSD

E. Pub/Sub, Dataflow, and Cloud Storage

Answer: (SHOW ANSWER)

NEW QUESTION: 2

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)

Explanation

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: 3

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 with the existing dataset.

B. Train your own image recognition model leveraging transfer learning techniques.

C. Use Cloud Vision AutoML, but reduce your dataset twice.

D. Use Cloud Vision API by providing custom labels as recognition hints.

Answer: (SHOW ANSWER)

NEW QUESTION: 4

You need to modernize your existing on-premises data strategy. Your organization currently uses.

* Apache Hadoop clusters for processing multiple large data sets, including on-premises Hadoop Distributed File System (HDFS) for data replication.

* Apache Airflow to orchestrate hundreds of ETL pipelines with thousands of job steps.

You need to set up a new architecture in Google Cloud that can handle your Hadoop workloads and requires minimal changes to your existing orchestration processes. What should you do?

A. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases Convert your ETL pipelines to Dataflow.
B. Use Bigtable for your large workloads, with connections to Cloud Storage to handle any HDFS use cases Orchestrate your pipelines with Cloud Composer.
C. Use Dataproc to migrate your Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Use Cloud Data Fusion to visually design and deploy your ETL pipelines.
D. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Orchestrate your pipelines with Cloud Composer..
Answer: D (LEAVE A REPLY)
Dataproc is a fully managed service that allows you to run Apache Hadoop and Spark workloads on Google Cloud. It is compatible with the open source ecosystem, so you can migrate your existing Hadoop clusters to Dataproc with minimal changes. Cloud Storage is a scalable, durable, and cost-effective object storage service that can replace HDFS for storing and accessing data. Cloud Storage offers interoperability with Hadoop through connectors, so you can use it as a data source or sink for your Dataproc jobs. Cloud Composer is a fully managed service that allows you to create, schedule, and monitor workflows using Apache Airflow. It is integrated with Google Cloud services, such as Dataproc, BigQuery, Dataflow, and Pub/Sub, so you can orchestrate your ETL pipelines across different platforms. Cloud Composer is compatible with your existing Airflow code, so you can migrate your existing orchestration processes to Cloud Composer with minimal changes.
The other options are not as suitable as Dataproc and Cloud Composer for this use case, because they either require more changes to your existing code, or do not meet your requirements. Dataflow is a fully managed service that allows you to create and run scalable data processing pipelines using Apache Beam. However, Dataflow is not compatible with your existing Hadoop code, so you would need to rewrite your ETL pipelines using Beam. Bigtable is a fully managed NoSQL database service that can handle large and complex data sets. However, Bigtable is not compatible with your existing Hadoop code, so you would need to rewrite your queries and applications using Bigtable APIs. Cloud Data Fusion is a fully managed service that allows you to visually design and deploy data integration pipelines using a graphical interface. However, Cloud Data Fusion is not compatible with your existing Airflow code, so you would need to recreate your orchestration processes using Cloud Data Fusion UI. Reference:
Dataproc overview
Cloud Storage connector for Hadoop
Cloud Composer overview

NEW QUESTION: 5
You are analyzing the price of a company's stock. Every 5 seconds, you need to compute a moving average of the past 30 seconds' worth of data. You are reading data from Pub/Sub and using DataFlow to conduct the analysis. How should you set up your windowed pipeline?
A. Use a sliding window with a duration of 5 seconds. Emit results by setting the following trigger: AfterProcessingTime.pastFirstElementInPane().plusDelayOf(Duration.standardSeconds(30))
B. Use a sliding window with a duration of 30 seconds and a period of 5 seconds. Emit results by setting the following trigger: AfterWatermark.pastEndOfWindow()
C. Use a fixed window with a duration of 30 seconds. Emit results by setting the following trigger: AfterWatermark.pastEndOfWindow().plusDelayOf(Duration.standardSeconds(5))
D. Use a fixed window with a duration of 5 seconds. Emit results by setting the following trigger: AfterProcessingTime.pastFirstElementInPane().plusDelayOf(Duration.standardSeconds(30))

Answer: C (LEAVE A REPLY)

NEW QUESTION: 6
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. 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. Reference:
Datastream overview
Creating a Datastream stream
Using Datastream with BigQuery

NEW QUESTION: 7
Your company has data assets across multiple Cloud Storage buckets and BigQuery datasets containing raw and processed data. The requirement is to establish a unified data governance framework that allows for centralized metadata discovery, data quality monitoring, and consistent security policy application across these various data stores without physically moving or duplicating the data. You need to implement a solution to achieve this federated governance. What should you do?
A. Deploy a centralized Cloud SQL database to store metadata extracted from BigQuery and Cloud Storage using custom scripts. Integrate the database with Looker Studio for data discovery and visualization.
Implement a custom policy engine using Cloud Run functions triggered by changes in IAM policies to enforce consistent security across projects.
B. Create a Looker Studio dashboard on BigQuery INFORMATION_SCHEMA views to visualize and monitor data quality.
Manage security using IAM policies at the project level, supplemented by BigQuery authorized views for granular access control.
C. Export metadata out of Dataplex Universal Catalog by running a metadata export job.
Implement Dataproc Metastore to manage table schemas and Apache Hive metastore for metadata discovery.
Manage security using a combination of BigQuery row-level security and Cloud Storage policies.
D. Use Dataplex to organize the BigQuery datasets and Cloud Storage buckets into lakes and zones.Use Dataplex for automated metadata discovery, centralized security policy management, data profiling, and data quality tasks.
Answer: D (LEAVE A REPLY)

Dataplex is Google Cloud's intelligent data fabric that enables organizations to centrally manage, monitor, and govern data across data lakes, data warehouses, and data marts. It is specifically designed to provide a unified interface for distributed data without moving the data.
* Unified Governance: Dataplex allows you to group distributed data (in GCS and BigQuery) into logical Lakes and Zones (e.g., Raw vs. Curated). This structure enables you to apply security policies once at the Lake or Zone level, which then propagates to all underlying assets.
* Metadata Discovery: Dataplex automatically scans and registers metadata from your buckets and datasets into the Search/Catalog interface, making it discoverable without manual scripts.
* Data Quality & Profiling: Dataplex includes built-in, fully managed features for Data Quality (declarative rules) and Data Profiling to monitor the health of your data directly within the governance framework.
* Correcting other options:
* A & B: These involve significant manual overhead, custom scripts, and fragmented tools. They do not provide a "unified data governance framework" that scales naturally across both GCS and BigQuery.
* C: Dataproc Metastore is primarily for Hive/Spark workloads and doesn't offer the comprehensive security and data quality governance required for a general-purpose federated framework across BigQuery and Cloud Storage.
Reference: Google Cloud Documentation on Dataplex:
"Dataplex is an intelligent data fabric that helps you unify distributed data and automate data management and governance across that data. With Dataplex, you can: Build a unified search and discovery experience across all your data. Centrally manage security and governance across data stored in Cloud Storage and BigQuery. Automate data quality and data profiling to ensure the reliability of your data assets." (Source: Dataplex Overview)
"Dataplex lets you organize your data into lakes and zones. A lake represents a logical data domain... A zone represents a sub-domain within a lake and is useful for categorizing data by its readiness (e.g., raw vs.
curated)." (Source: Dataplex terminology)

NEW QUESTION: 8
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. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
C. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
D. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.
Answer: (SHOW ANSWER)

NEW QUESTION: 9
You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK_STREAM. The column DTstores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRINGtype. Now, you want to compute web session durations of users who visit your site, and you want to change its data type

to the TIMESTAMP. You want to minimize the migration effort without making future queries computationally expensive. What should you do?

A. Delete the table CLICK_STREAM, and then re-create it such that the column DTis of the TIMESTAMPtype.
Reload the data.

B. Create a view CLICK_STREAM_V, where strings from the column DTare cast into TIMESTAMPvalues.
Reference the view CLICK_STREAM_Vinstead of the table CLICK_STREAMfrom now on.

C. Construct a query to return every row of the table CLICK_STREAM, while using the built-in function to cast strings from the column DTinto TIMESTAMPvalues. Run the query into a destination table NEW_CLICK_STREAM, in which the column TSis the TIMESTAMPtype. Reference the table NEW_CLICK_STREAMinstead of the table CLICK_STREAMfrom now on. In the future, new data is loaded into the table NEW_CLICK_STREAM.

D. Add two columns to the table CLICK STREAM: TSof the TIMESTAMPtype and IS_NEWof the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS_NEWto true. For future queries, reference the column TSinstead of the column DT, with the WHEREclause ensuring that the value of IS_NEWmust be true.

E. Add a column TSof the TIMESTAMPtype to the table CLICK_STREAM, and populate the numeric values from the column TSfor each row. Reference the column TSinstead of the column DTfrom now on.

Answer: D (LEAVE A REPLY)

NEW QUESTION: 10

You have a table that contains millions of rows of sales data, partitioned by date Various applications and users query this data many times a minute. The query requires aggregating values by using avg. max. and sum, and does not require joining to other tables. The required aggregations are only computed over the past year of data, though you need to retain full historical data in the base tables You want to ensure that the query results always include the latest data from the tables, while also reducing computation cost, maintenance overhead, and duration. What should you do?

A. Create a materialized view to aggregate the base table data Configure a partition expiration on the base table to retain only the last one year of partitions.

B. Create a materialized view to aggregate the base table data include a filter clause to specify the last one year of partitions.

C. Create a new table that aggregates the base table data include a filter clause to specify the last year of partitions. Set up a scheduled query to recreate the new table every hour.

D. Create a view to aggregate the base table data Include a filter clause to specify the last year of partitions.

Answer: C (LEAVE A REPLY)

A materialized view is a database object that contains the results of a query, which can be updated periodically. It can improve the performance and efficiency of queries that involve aggregations, joins, or filters. By creating a materialized view to aggregate the base table data and include a filter clause to specify the last one year of partitions, you can ensure that the query results always include the latest data from the tables, while also reducing computation cost, maintenance overhead, and duration. The materialized view will automatically refresh when the base table data changes, and will only use the partitions that match the filter clause. Option A is incorrect because it will delete the historical data from the base table, which is not desired. Option C is incorrect because it will create a redundant table that needs to be updated manually by a scheduled query, which is more complex and costly than using a materialized view. Option D is incorrect because a view does not store any data, but only references the base table data, which means it will not reduce the computation cost or duration of the query. Reference:
Materialized views, ML models in data warehouse - Google Cloud
Data Engineering with Google Cloud Platform - Packt Subscription

NEW QUESTION: 11
All Google Cloud Bigtable client requests go through a front-end server _____ they are sent to a Cloud Bigtable node.
A. before
B. after
C. only if
D. once
Answer: (SHOW ANSWER)
In a Cloud Bigtable architecture all client requests go through a front-end server before they are sent to a Cloud Bigtable node.
The nodes are organized into a Cloud Bigtable cluster, which belongs to a Cloud Bigtable instance, which is a container for the cluster. Each node in the cluster handles a subset of the requests to the cluster.
When additional nodes are added to a cluster, you can increase the number of simultaneous requests that the cluster can handle, as well as the maximum throughput for the entire cluster.
Reference:
https://cloud.google.com/bigtable/docs/overview

NEW QUESTION: 12
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 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. 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.
C. Set the BigQuery dataset to be multi-regional. In the event of an emergency, use a point-in-time snapshot to recover the data.
D. 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.
Answer: C (LEAVE A REPLY)
highly available = multi-regional:
https://cloud.google.com/bigquery/docs/locations
recovery strategy of this data that minimizes cost = point-in-time snapshot:
https://cloud.google.com/solutions/bigquery-data-warehouse#backup-and-recovery

NEW QUESTION: 13
You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat.
Here is some of the information you need to store:
The user profile: What the user likes and doesn't like to eat
The user account information: Name, address, preferred meal times
The order information: When orders are made, from where, to whom
The database will be used to store all the transactional data of the product. You want to optimize the data schem
a. Which Google Cloud Platform product should you use?
A. Cloud Bigtable
B. BigQuery

C. Cloud Datastore
D. Cloud SQL
Answer: B (LEAVE A REPLY)

NEW QUESTION: 14
Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow.
Numerous data logs are being are being generated during this step, and the team wants to analyze them.
Due to the dynamic nature of the campaign, the data is growing exponentially every hour. The data scientists have written the following code to read the data for a new key features in the logs.
BigQueryIO.Read
.named("ReadLogData")
.from("clouddataflow-readonly:samples.log_data")
You want to improve the performance of this data read. What should you do?
A. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
B. Specify the Tableobject in the code.
C. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.
D. Use .fromQuery operation to read specific fields from the table.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 15
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 Cloud Storage and BigQuery permanent linked tables for query.
B. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.
C. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
D. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
Answer: (SHOW ANSWER)

NEW QUESTION: 16
You have data located in BigQuery that is used to generate reports for your company. You have noticed some weekly executive report fields do not correspond to format according to company standards for example, report errors include different telephone formats and different country code identifiers. This is a frequent issue, so you need to create a recurring job to normalize the data. You want a quick solution that requires no coding What should you do?
A. Use Cloud Data Fusion and Wrangler to normalize the data, and set up a recurring job.
B. Use BigQuery and GoogleSQL to normalize the data, and schedule recurring quenes in BigQuery.
C. Create a Spark job and submit it to Dataproc Serverless.
D. Use Dataflow SQL to create a job that normalizes the data, and that after the first run of the job, schedule the pipeline to execute recurrently.
Answer: A (LEAVE A REPLY)

Cloud Data Fusion is a fully managed, cloud-native data integration service that allows you to build and manage data pipelines with a graphical interface. Wrangler is a feature of Cloud Data Fusion that enables you to interactively explore, clean, and transform data using a spreadsheet-like UI. You can use Wrangler to normalize the data in BigQuery by applying various directives, such as parsing, formatting, replacing, and validating data. You can also preview the results and export the wrangled data to BigQuery or other destinations. You can then set up a recurring job in Cloud Data Fusion to run the Wrangler pipeline on a schedule, such as weekly or daily. This way, you can create a quick and code-free solution to normalize the data for your reports. References:
* Cloud Data Fusion overview
* Wrangler overview
* Wrangle data from BigQuery
* [Scheduling pipelines]

NEW QUESTION: 17
You want to use a BigQuery table as a data sink. In which writing mode(s) can you use BigQuery as a sink?
A. Both batch and streaming
B. BigQuery cannot be used as a sink
C. Only batch
D. Only streaming
Answer: A (LEAVE A REPLY)
When you apply a BigQueryIO.Write transform in batch mode to write to a single table, Dataflow invokes a BigQuery load job. When you apply a BigQueryIO.Write transform in streaming mode or in batch mode using a function to specify the destination table, Dataflow uses BigQuery's streaming inserts Reference:
https://cloud.google.com/dataflow/model/bigquery-io

NEW QUESTION: 18
You are loading CSV files from Cloud Storage to BigQuery. The files have known data quality issues, including mismatched data types, such as STRINGS and INT64s in the same column, and inconsistent formatting of values such as phone numbers or addresses. You need to create the data pipeline to maintain data quality and perform the required cleansing and transformation. What should you do?
A. Use Data Fusion to transform the data before loading it into BigQuery.
B. Load the CSV files into a staging table with the desired schema, perform the transformations with SQL. and then write the results to the final destination table.
C. Create a table with the desired schema, toad the CSV files into the table, and perform the transformations in place using SQL.
D. Use Data Fusion to convert the CSV files lo a self-describing data formal, such as AVRO. before loading the data to BigOuery.

Answer: A (LEAVE A REPLY)

Data Fusion's advantages:

Visual interface: Offers a user-friendly interface for designing data pipelines without extensive coding, making it accessible to a wider range of users.

Built-in transformations: Includes a wide range of pre-built transformations to handle common data quality issues, such as:

Data type conversions

Data cleansing (e.g., removing invalid characters, correcting formatting) Data validation (e.g., checking for missing values, enforcing constraints) Data enrichment (e.g., adding derived fields, joining with other datasets) Custom transformations: Allows for custom transformations using SQL or Java code for more complex cleaning tasks.

Scalability: Can handle large datasets efficiently, making it suitable for processing CSV files with potential data quality issues.

Integration with BigQuery: Integrates seamlessly with BigQuery, allowing for direct loading of transformed data.

Topic 2, MJTelco Case Study

Company Overview

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

Company Background

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

Solution Concept

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

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

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

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

Business Requirements

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

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

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

Technical Requirements

Ensure secure and efficient transport and storage of telemetry data

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

Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately 100m records/day Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis. Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.

NEW QUESTION: 19
You are designing a real-time system for a ride hailing app that identifies areas with high demand for rides to effectively reroute available drivers to meet the demand. The system ingests data from multiple sources to Pub/Sub. processes the data, and stores the results for visualization and analysis in real-time dashboards. The data sources include driver location updates every 5 seconds and app-based booking events from riders. The data processing involves real-time aggregation of supply and demand data for the last 30 seconds, every 2 seconds, and storing the results in a low-latency system for visualization. What should you do?
A. Group the data by using a tumbling window in a Dataflow pipeline, and write the aggregated data to Memorystore
B. Group the data by using a hopping window in a Dataflow pipeline, and write the aggregated data to Memorystore
C. Group the data by using a session window in a Dataflow pipeline, and write the aggregated data to BigQuery.
D. Group the data by using a hopping window in a Dataflow pipeline, and write the aggregated data to BigQuery.
Answer: B (LEAVE A REPLY)
A hopping window is a type of sliding window that advances by a fixed period of time, producing overlapping windows. This is suitable for the scenario where the system needs to aggregate data for the last 30 seconds, every 2 seconds, and provide real-time updates. A Dataflow pipeline can implement the hopping window logic using Apache Beam, and process both streaming and batch data sources. Memorystore is a low-latency, in-memory data store that can serve the aggregated data to the visualization layer. BigQuery is not a good choice for this scenario, as it is not optimized for low-latency queries and frequent updates.

NEW QUESTION: 20
You have designed an Apache Beam processing pipeline that reads from a Pub/Sub topic. The topic has a message retention duration of one day, and writes to a Cloud Storage bucket. You need to select a bucket location and processing strategy to prevent data loss in case of a regional outage with an RPO of 15 minutes. What should you do?
A. 1. Use a dual-region Cloud Storage bucket with turbo replication enabled.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.

B. 1. Use a regional Cloud Storage bucket.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by one day to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region and write in a bucket in the same region.
C. 1. Use a multi-regional Cloud Storage bucket.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 60 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.
D. 1. Use a dual-region Cloud Storage bucket.
2. Monitor Dataflow metrics with Cloud Monitoring to determine when an outage occurs.
3. Seek the subscription back in time by 15 minutes to recover the acknowledged messages.
4. Start the Dataflow job in a secondary region.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 21
You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?
A. PigLatin using Pig
B. HiveQL using Hive
C. Java using MapReduce
D. Python using MapReduce
Answer: A (LEAVE A REPLY)
Pig is scripting language which can be used for checkpointing and splitting pipelines.

NEW QUESTION: 22
Case Study: 3,
MJTelco Case Study
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost. Their management and operations teams are situated all around the globe creating many-to-many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:

Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments ?development/test, staging, and production ?
to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements
Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community. Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.
Provide reliable and timely access to data for analysis from distributed research workers Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
Technical Requirements
Ensure secure and efficient transport and storage of telemetry data Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately
100m records/day
Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis.
Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
Google Cloud Dataflow pipeline is now ready to start receiving data from the 50,000 installations. You want to allow Cloud Dataflow to scale its compute power up as required. Which Cloud Dataflow pipeline configuration setting should you update?
A. The zone
B. The number of workers
C. The maximum number of workers
D. The disk size per worker
Answer: A (LEAVE A REPLY)

NEW QUESTION: 23
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: (SHOW ANSWER)
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: 24
If you want to create a machine learning model that predicts the price of a particular stock based on its recent price history, what type of estimator should you use?
A. Unsupervised learning
B. Regressor
C. Classifier
D. Clustering estimator
Answer: B (LEAVE A REPLY)
Regression is the supervised learning task for modeling and predicting continuous, numeric variables.
Examples include predicting real-estate prices, stock price movements, or student test scores.
Classification is the supervised learning task for modeling and predicting categorical variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades. Clustering is an unsupervised learning task for finding natural groupings of observations (i.e. clusters) based on the inherent structure within your dataset. Examples include customer segmentation, grouping similar items in e-commerce, and social network analysis.
Reference: https://elitedatascience.com/machine-learning-algorithms

NEW QUESTION: 25
Your company produces 20,000 files every hour. Each data file is formatted as a comma separated values (CSV) file that is less than 4 KB. All files must be ingested on Google Cloud Platform before they can be processed. Your company site has a 200 ms latency to Google Cloud, and your Internet connection bandwidth is limited as 50 Mbps. You currently deploy a secure FTP (SFTP) server on a virtual machine in Google Compute Engine as the data ingestion point. A local SFTP client runs on a dedicated machine to transmit the CSV files as is. The goal is to make reports with data from the previous day available to the executives by 10:00 a.m. each day. This design is barely able to keep up with the current volume, even though the bandwidth utilization is rather low.
You are told that due to seasonality, your company expects the number of files to double for the next three months. Which two actions should you take? (Choose two.)
A. Create an S3-compatible storage endpoint in your network, and use Google Cloud Storage Transfer Service to transfer on-premises data to the designated storage bucket.
B. Introduce data compression for each file to increase the rate file of file transfer.
C. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
D. Assemble 1,000 files into a tape archive (TAR) file. Transmit the TAR files instead, and disassemble the CSV files in the cloud upon receiving them.

E. Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.
Answer: A,E (LEAVE A REPLY)

NEW QUESTION: 26
To give a user read permission for only the first three columns of a table, which access control method would you use?
A. Primitive role
B. Predefined role
C. Authorized view
D. It's not possible to give access to only the first three columns of a table.
Answer: (SHOW ANSWER)
An authorized view allows you to share query results with particular users and groups without giving them read access to the underlying tables. Authorized views can only be created in a dataset that does not contain the tables queried by the view.
When you create an authorized view, you use the view's SQL query to restrict access to only the rows and columns you want the users to see.

NEW QUESTION: 27
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. Establish a Cloud Interconnect between all remote data centers and Google.
C. Have the data acquisition devices publish data to Cloud Pub/Sub.
D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 28
You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output. Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?
A. Cancel
B. Drain
C. Stop
D. Finish
Answer: B (LEAVE A REPLY)
Using the Drain option to stop your job tells the Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources, but the Dataflow service will preserve any existing resources (such as worker instances) to finish processing and writing any buffered data in your pipeline.

NEW QUESTION: 29
If a dataset contains rows with individual people and columns for year of birth, country, and income, how many of the columns are continuous and how many are categorical?

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

NEW QUESTION: 30
You want to archive data in Cloud Storage. Because some data is very sensitive, you want to use the
"Trust No One" (TNO) approach to encrypt your data to prevent the cloud provider staff from decrypting your data. What should you do?
A. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in a different project that only the security team can access.
B. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key. Use gsutil cp to upload each encrypted file to the Cloud Storage bucket.
Manually destroy the key previously used for encryption, and rotate the key once.
C. Specify customer-supplied encryption key (CSEK) in the .boto configuration file. Use gsutil cp to upload each archival file to the Cloud Storage bucket. Save the CSEK in Cloud Memorystore as permanent storage of the secret.
D. Use gcloud kms keys create to create a symmetric key. Then use gcloud kms encrypt to encrypt each archival file with the key and unique additional authenticated data (AAD). Use gsutil to upload each encrypted file to the Cloud Storage bucket, and keep the AAD outside of Google cp Cloud.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 31
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: 32
You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of- Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about
100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.
You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)
A. Redis
B. HBase
C. MySQL
D. MongoDB
E. Cassandra
F. HDFS with Hive
Answer: B,D,F (LEAVE A REPLY)
Explanation/Reference:

NEW QUESTION: 33
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: 34
Which of these is NOT a way to customize the software on Dataproc cluster instances?
A. Set initialization actions
B. Modify configuration files using cluster properties
C. Configure the cluster using Cloud Deployment Manager

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

NEW QUESTION: 35
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: 36
Case Study 2 - MJTelco
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost. Their management and operations teams are situated all around the globe creating many-to- many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
* Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
* Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.

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

Business Requirements

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

Technical Requirements

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

CEO Statement

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

CTO Statement

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

CFO Statement

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

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

A. Rowkey: date#device_id
Column data: data_point

B. Rowkey: date
Column data: device_id, data_point

C. Rowkey: device_id
Column data: date, data_point

D. Rowkey: data_point
Column data: device_id, date

E. Rowkey: date#data_point
Column data: device_id

Answer: A (LEAVE A REPLY)

The most common query is for all the data for a given device for a given day", rowkey should have info for both device and date.

NEW QUESTION: 37
Which of the following are feature engineering techniques? (Select 2 answers)
A. Hidden feature layers
B. Feature prioritization
C. Crossed feature columns
D. Bucketization of a continuous feature
Answer: C,D (LEAVE A REPLY)
Explanation
Selecting and crafting the right set of feature columns is key to learning an effective model.
Bucketization is a process of dividing the entire range of a continuous feature into a set of consecutive bins/buckets, and then converting the original numerical feature into a bucket ID (as a categorical feature) depending on which bucket that value falls into.
Using each base feature column separately may not be enough to explain the data. To learn the differences between different feature combinations, we can add crossed feature columns to the model.
Reference:
https://www.tensorflow.org/tutorials/wide#selecting_and_engineering_features_for_the_model

NEW QUESTION: 38
You operate a database that stores stock trades and an application that retrieves average stock price for a given company over an adjustable window of time. The data is stored in Cloud Bigtable where the datetime of the stock trade is the beginning of the row key. Your application has thousands of concurrent users, and you notice that performance is starting to degrade as more stocks are added. What should you do to improve the performance of your application?
A. Change the row key syntax in your Cloud Bigtable table to begin with the stock symbol.
B. Change the row key syntax in your Cloud Bigtable table to begin with a random number per second.
C. Change the data pipeline to use BigQuery for storing stock trades, and update your application.
D. Use Cloud Dataflow to write summary of each day's stock trades to an Avro file on Cloud Storage.
Update your application to read from Cloud Storage and Cloud Bigtable to compute the responses.
Answer: B (LEAVE A REPLY)
Timestamp at starting of rowkey causes bottleneck issues.

NEW QUESTION: 39
Your company is running their first dynamic campaign, serving different offers by analyzing real-time data during the holiday season. The data scientists are collecting terabytes of data that rapidly grows every hour during their 30-day campaign. They are using Google Cloud Dataflow to preprocess the data and collect the feature (signals) data that is needed for the machine learning model in Google Cloud Bigtable.
The team is observing suboptimal performance with reads and writes of their initial load of 10 TB of data.
They want to improve this performance while minimizing cost. What should they do?
A. Redefine the schema by evenly distributing reads and writes across the row space of the table.
B. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.
C. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.
D. The performance issue should be resolved over time as the site of the BigDate cluster is increased.

Answer: (SHOW ANSWER)

NEW QUESTION: 40
Cloud Bigtable is Google's _____ Big Data database service.
A. Relational
B. mySQL
C. NoSQL
D. SQL Server
Answer: C (LEAVE A REPLY)
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: 41
Your company is currently setting up data pipelines for their campaign. For all the Google Cloud Pub/Sub streaming data, one of the important business requirements is to be able to periodically identify the inputs and their timings during their campaign. Engineers have decided to use windowing and transformation in Google Cloud Dataflow for this purpose. However, when testing this feature, they find that the Cloud Dataflow job fails for the all streaming insert. What is the most likely cause of this problem?
A. They have not set the triggers to accommodate the data coming in late, which causes the job to fail
B. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created
C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created
D. They have not assigned the timestamp, which causes the job to fail
Answer: C (LEAVE A REPLY)

NEW QUESTION: 42
The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?
A. Workers
B. Masters, workers, and parameter servers
C. Workers and parameter servers
D. Parameter servers
Answer: C (LEAVE A REPLY)
The CUSTOM tier is not a set tier, but rather enables you to use your own cluster specification.
When you use this tier, set values to configure your processing cluster according to these guidelines:
You must set TrainingInput.masterType to specify the type of machine to use for your master node. You may set TrainingInput.workerCount to specify the number of workers to use. You may set TrainingInput.parameterServerCount to specify the number of parameter servers to use. You can specify the type of machine for the master node, but you can't specify more than one master node.
Reference:
https://cloud.google.com/ml-engine/docs/training-overview#job_configuration_parameters

NEW QUESTION: 43
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. Remove Cloud Storage IAM permissions to the External Team on the acme-raw-data project
B. 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
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. 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
Answer: D (LEAVE A REPLY)

NEW QUESTION: 44
Your company has data assets across multiple Cloud Storage buckets and BigQuery datasets containing raw and processed data. The requirement is to establish a unified data governance framework that allows for centralized metadata discovery, data quality monitoring, and consistent security policy application across these various data stores without physically moving or duplicating the data. You need to implement a solution to achieve this federated governance.
What should you do?
A. 1. Export metadata out of Dataplex Universal Catalog by running a metadata export job.
2. Implement Dataproc Metastore to manage table schemas and Apache Hive metastore for metadata discovery.
3. Manage security using a combination of BigQuery row-level security and Cloud Storage policies.
B. 1. Use Dataplex to organize the BigQuery datasets and Cloud Storage buckets into lakes and zones.
2. Use Dataplex for automated metadata discovery, centralized security policy management, data profiling, and data quality tasks.
C. 1. Deploy a centralized Cloud SQL database to store metadata extracted from BigQuery and Cloud Storage using custom scripts.
2. Integrate the database with Looker Studio for data discovery and visualization.
3. Implement a custom policy engine using Cloud Run functions triggered by changes in IAM policies to enforce consistent security across projects.
D. 1. Create a Looker Studio dashboard on BigQuery INFORMATION_SCHEMAviews to visualize and monitor data quality.
2. Manage security using IAM policies at the project level, supplemented by BigQuery authorized views for granular access control.
Answer: (SHOW ANSWER)
Dataplex provides a federated data governance layer that logically organizes BigQuery datasets and Cloud Storage buckets into lakes and zones without moving data. It enables centralized metadata discovery through automatic cataloging, supports data profiling and data quality monitoring, and enforces consistent security policies across heterogeneous data stores, meeting all requirements for unified governance.

NEW QUESTION: 45
You have a requirement to insert minute-resolution data from 50,000 sensors into a BigQuery table. You expect significant growth in data volume and need the data to be available within 1 minute of ingestion for real-time analysis of aggregated trends. What should you do?

A. Use bq loadto load a batch of sensor data every 60 seconds.
B. Use a Cloud Dataflow pipeline to stream data into the BigQuery table.
C. Use the INSERT statement to insert a batch of data every 60 seconds.
D. Use the MERGE statement to apply updates in batch every 60 seconds.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 46
Your team has created several BigQuery curated datasets containing anonymized industry benchmark data.
You want to make these datasets easily discoverable and accessible for querying by external partner companies within their own Google Cloud projects. You need a secure and scalable solution. What should you do?
A. Grant the roles/bigquery.dataViewer IAM role to the partner group email addresses on the datasets.
B. Publish the datasets as listings within BigQuery sharing (Analytics Hub).
C. Create authorized views for each dataset and grant access to each partner.
D. Export the datasets to partner-specific Cloud Storage buckets.
Answer: (SHOW ANSWER)
Analytics Hub (part of BigQuery sharing) is the enterprise-grade, Google-recommended way to share data across organizational boundaries.
* Discoverability: Analytics Hub provides a central portal where partners can search for and "subscribe" to datasets. Traditional IAM grants (A) do not provide a discovery interface.
* Scalability: You can publish a single "listing" and allow many partners to subscribe to it. Each subscriber gets a "linked dataset" in their project.
* Security & Governance: The publisher retains control. You can see who has subscribed and even enforce "data egress" controls to prevent partners from copying or exporting the data while still allowing them to query it.
* Correcting other options:
* A: This works for a few users but lacks the metadata, searchability, and formal "subscription" lifecycle of Analytics Hub.
* C: Authorized views are for internal security within a project/org and don't solve the cross-org discovery problem.
* D: Exporting data creates data silos, increases storage costs, and loses the real-time query benefits of BigQuery.
Reference: Google Cloud Documentation on Analytics Hub:
"Analytics Hub is a data exchange platform that lets you share data and insights at scale across organizational boundaries... Listings let you share data without replicating the shared data... Subscribers discover data through the exchange and create a linked dataset in their own project to query the data." (Source: Introduction to BigQuery sharing)

NEW QUESTION: 47

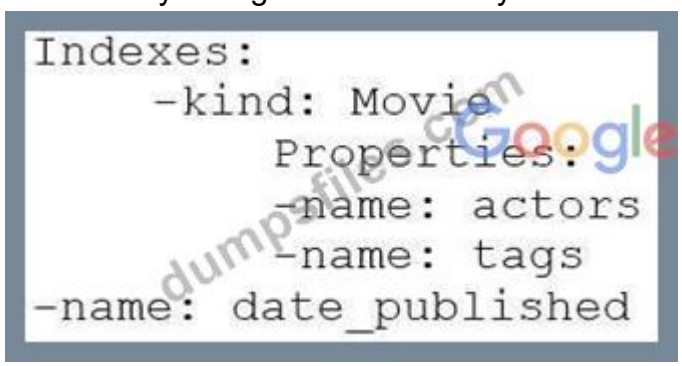
Which of the following is NOT true about Dataflow pipelines?
A. Dataflow pipelines are tied to Dataflow, and cannot be run on any other runner
B. Dataflow pipelines can consume data from other Google Cloud services
C. Dataflow pipelines can be programmed in Java
D. Dataflow pipelines use a unified programming model, so can work both with streaming and batch data sources
Answer: A (LEAVE A REPLY)
Explanation
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: 48
You work for a financial institution that lets customers register online. As new customers register, their user data is sent to Pub/Sub before being ingested into BigQuery. For security reasons, you decide to redact your customers' Government issued Identification Number while allowing customer service representatives to view the original values when necessary. What should you do?
A. Use BigQuery's built-in AEAD encryption to encrypt the SSN column. Save the keys to a new table that is only viewable by permissioned users.
B. Use BigQuery column-level security. Set the table permissions so that only members of the Customer Service user group can see the SSN column.
C. Before loading the data into BigQuery, use Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic hash.
D. Before loading the data into BigQuery, use Cloud Data Loss Prevention (DLP) to replace input values with a cryptographic format-preserving encryption token.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 49
You are deploying a new storage system for your mobile application, which is a media streaming service. You decide the best fit is Google Cloud Datastore. You have entities with multiple properties, some of which can take on multiple values. For example, in the entity 'Movie'the property 'actors'and the property 'tags' have multiple values but the property 'date released' does not. A typical query would ask for all movies with actor=<actorname>ordered by date_releasedor all movies with tag=Comedyordered by date_released.
How should you avoid a combinatorial explosion in the number of indexes?
A. Manually configure the index in your index config as follows:

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

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



Answer: A (LEAVE A REPLY)

NEW QUESTION: 50
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.
is rolling out their real-time inventory tracking system. The tracking devices will all send package-tracking messages, which will now go to a single Google Cloud Pub/Sub topic instead of the Apache Kafka cluster.
A subscriber application will then process the messages for real-time reporting and store them in Google BigQuery for historical analysis. You want to ensure the package data can be analyzed over time.
Which approach should you take?
A. Attach the timestamp and Package ID on the outbound message from each publisher device as they are sent to Clod Pub/Sub.
B. Use the NOW () function in BigQuery to record the event's time.
C. Use the automatically generated timestamp from Cloud Pub/Sub to order the data.
D. Attach the timestamp on each message in the Cloud Pub/Sub subscriber application as they are received.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 51
You are using Cloud Bigtable to persist and serve stock market data for each of the major indices. To serve the trading application, you need to access only the most recent stock prices that are streaming in How should you design your row key and tables to ensure that you can access the data with the most simple query?
A. Create one unique table for all of the indices, and then use the index and timestamp as the row key design

B. For each index, have a separate table and use a reverse timestamp as the row key design
C. For each index, have a separate table and use a timestamp as the row key design
D. Create one unique table for all of the indices, and then use a reverse timestamp as the row key design.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 52
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: A (LEAVE A REPLY)
Google charges for storage, queries, and streaming inserts. Loading data from a file and exporting data are free operations.

NEW QUESTION: 53
You are building a streaming Dataflow pipeline that ingests noise level data from hundreds of sensors placed near construction sites across a city. The sensors measure noise level every ten seconds, and send that data to the pipeline when levels reach above 70 dBA. You need to detect the average noise level from a sensor when data is received for a duration of more than 30 minutes, but the window ends when no data has been received for 15 minutes What should you do?
A. Use session windows with a 30-mmute gap duration.
B. Use tumbling windows with a 15-mmute window and a fifteen-minute. withAllowedLateness operator.
C. Use session windows with a 15-minute gap duration.
D. Use hopping windows with a 15-mmute window, and a thirty-minute period.
Answer: C (LEAVE A REPLY)
The key requirements for the windowing strategy are:
A window groups data for a specific sensor.
A window should contain data spanningat least30 minutes ("duration of more than 30 minutes" implies activity for this period).
A window for a sensorendswhen no data has been received from that sensor for 15 minutes (this is a gap).
This scenario perfectly describessession windows.
Session Windows:Session windows group elements (per key, e.g., per sensor ID) that arrive within a certain
"gap duration" of each other. A new session starts if data for a key arrives after the gap duration has passed since the last data point for that key.
In this case, if data stops arriving for a sensor for 15 minutes, the current session for that sensor closes. This matches "the window ends when no data has been received for 15 minutes." The "duration of more than 30 minutes" requirement is a condition you would applyafterthe session window closes. You'd calculate the duration of the data within the closed session window and only compute the average if that session's duration (span of event times within it) exceeds 30 minutes. Session windows themselves don't have a fixed duration; their duration is determined by data activity and the gap.
Let's analyze why other options are less suitable:
A (Hopping windows with a 15-minute window, and a thirty-minute period):Hopping windows have a fixed size and a fixed period. They create overlapping windows. This doesn't align with the dynamic nature of sessions ending based on inactivity. A 30-minute period with a 15-minute window means windows like [0:00-

0:15], [0:15-0:30], [0:30-0:45]. If activity is continuous, a 30-minute activity span would be covered, but the window closing is not based on a 15-minute gap of inactivity.
B (Tumbling windows with a 15-minute window and a fifteen-minute .withAllowedLateness operator):
Tumbling windows are fixed-size, non-overlapping windows. .withAllowedLateness deals with late data arriving for a window that has already passed its end time, not with defining the window based on activity gaps.
C (Session windows with a 30-minute gap duration):This would mean a session ends only if there's a 30- minute gap of inactivity. The requirement is a 15-minute gap.
Therefore, session windows with a 15-minute gap duration (Option D) correctly model the requirement for windows to close after 15 minutes of inactivity from a sensor. The subsequent filtering for sessions lasting more than 30 minutes is a downstream operation.
Reference:
Apache Beam Programming Guide > Windowing > Windowing functions > Session windows. "Session windowing assigns elements to windows that represent sessions of activity. A session window starts when the first element arrives for a key. If another element arrives for that key within the specified gap duration, that element is included in the existing session window. If an element arrives after the gap duration, a new session window starts for that element... Session windows are useful for data that is irregularly distributed with respect to time, such as user activity data." This directly matches the sensor data behavior: data arrives when noise is high, and a period of no data for 15 minutes should close the analysis window for that sensor.

NEW QUESTION: 54
You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?
A. Convert the streaming insert code to batch load for individual messages.
B. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
C. Re-write the application to load accumulated data every 2 minutes.
D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 55
You want to analyze hundreds of thousands of social media posts daily at the lowest cost and with the fewest steps.
You have the following requirements:
* You will batch-load the posts once per day and run them through the Cloud Natural Language API.
* You will extract topics and sentiment from the posts.
* You must store the raw posts for archiving and reprocessing.
* You will create dashboards to be shared with people both inside and outside your organization.
You need to store both the data extracted from the API to perform analysis as well as the raw social media posts for historical archiving. What should you do?
A. Store the social media posts and the data extracted from the API in BigQuery.
B. Store the social media posts and the data extracted from the API in Cloud SQL.
C. Store the raw social media posts in Cloud Storage, and write the data extracted from the API into BigQuery.
D. Feed to social media posts into the API directly from the source, and write the extracted data from the API into BigQuery.

Answer: D (LEAVE A REPLY)
Explanation

NEW QUESTION: 56
Given the record streams MJTelco is interested in ingesting per day, they are concerned about the cost of Google BigQuery increasing. MJTelco asks you to provide a design solution. They require a single large data table called tracking_table. Additionally, they want to minimize the cost of daily queries while performing fine-grained analysis of each day's events. They also want to use streaming ingestion. What should you do?
A. Create a table called tracking_table and include a DATE column.
B. Create a partitioned table called tracking_table and include a TIMESTAMP column.
C. Create sharded tables for each day following the pattern tracking_table_YYYYMMDD.
D. Create a table called tracking_table with a TIMESTAMP column to represent the day.
Answer: B (LEAVE A REPLY)
Explanation:

NEW QUESTION: 57
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)
Reference https://support.google.com/datastudio/answer/7020039?hl=en

NEW QUESTION: 58
You are migrating your data warehouse to BigQuery. You have migrated all of your data into tables in a dataset. Multiple users from your organization will be using the data. They should only see certain tables based on their team membership. How should you set user permissions?
A. Create SQL views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the SQL views
B. Assign the users/groups data viewer access at the table level for each table
C. Create authorized views for each team in the same dataset in which the data resides, and assign the users/groups data viewer access to the authorized views
D. Create authorized views for each team in datasets created for each team. Assign the authorized views data viewer access to the dataset in which the data resides. Assign the users/groups data viewer access to the datasets in which the authorized views reside
Answer: B (LEAVE A REPLY)

NEW QUESTION: 59

You are administering shared BigQuery datasets that contain views used by multiple teams in your organization. The marketing team is concerned about the variability of their monthly BigQuery analytics spend using the on-demand billing model. You need to help the marketing team establish a consistent BigQuery analytics spend each month. What should you do?
A. Create a BigQuery Standard pay-as-you go reservation with a baseline of 0 slots and autoscaling set to 500 for the marketing team, and bill them back accordingly.
B. Create a BigQuery Enterprise reservation with a baseline of 250 slots and autoscaling set to 500 for the marketing team, and bill them back accordingly.
C. Establish a BigQuery quota for the marketing team, and limit the maximum number of bytes scanned each day.
D. Create a BigQuery reservation with a baseline of 500 slots with no autoscaling for the marketing team, and bill them back accordingly.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 60
Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow.
Numerous data logs are being are being generated during this step, and the team wants to analyze them. Due to the dynamic nature of the campaign, the data is growing exponentially every hour.
The data scientists have written the following code to read the data for a new key features in the logs.
BigQueryIO.Read
.named("ReadLogData")
.from("clouddataflow-readonly:samples.log_data")
You want to improve the performance of this data read. What should you do?
A. Specify the TableReference object in the code.
B. Use .fromQuery operation to read specific fields from the table.
C. Use of both the Google BigQuery TableSchema and TableFieldSchema classes.
D. Call a transform that returns TableRow objects, where each element in the PCollexction represents a single row in the table.
Answer: (SHOW ANSWER)

NEW QUESTION: 61
Your company's customer_order table in BigOuery stores the order history for 10 million customers, with a table size of 10 PB. You need to create a dashboard for the support team to view the order history. The dashboard has two filters, countryname and username. Both are string data types in the BigQuery table. When a filter is applied, the dashboard fetches the order history from the table and displays the query results. However, the dashboard is slow to show the results when applying the filters to the following query:

```
SELECT date, order, status FROM customer_order
WHERE country = '<country_name>' AND username = '<username>'
```

How should you redesign the BigQuery table to support faster access?
A. Cluster the table by country field, and partition by username field.
B. Partition the table by country and username fields.
C. Cluster the table by country and username fields
D. Partition the table by _PARTITIONTIME.

Answer: C (LEAVE A REPLY)
To improve the performance of querying a large BigQuery table with filters on countryname and username, clustering the table by these fields is the most effective approach. Here's why option C is the best choice:
Clustering in BigQuery:
Clustering organizes data based on the values in specified columns. This can significantly improve query performance by reducing the amount of data scanned during query execution.
Clustering by countryname and username means that data is physically sorted and stored together based on these fields, allowing BigQuery to quickly locate and read only the relevant data for queries using these filters.
Filter Efficiency:
With the table clustered by countryname and username, queries that filter on these columns can benefit from efficient data retrieval, reducing the amount of data processed and speeding up query execution.
This directly addresses the performance issue of the dashboard queries that apply filters on these fields.
Steps to Implement:
Redesign the Table:
Create a new table with clustering on countryname and username:
CREATE TABLE project.dataset.new_table
CLUSTER BY countryname, username AS
SELECT * FROM project.dataset.customer_order;
Migrate Data:
Transfer the existing data from the original table to the new clustered table.
Update Queries:
Modify the dashboard queries to reference the new clustered table.
Reference:
BigQuery Clustering Documentation
Optimizing Query Performance

NEW QUESTION: 62
Which of these rules apply when you add preemptible workers to a Dataproc cluster (select 2 answers)?
A. Preemptible workers cannot use persistent disk.
B. Preemptible workers cannot store data.
C. If a preemptible worker is reclaimed, then a replacement worker must be added manually.
D. A Dataproc cluster cannot have only preemptible workers.
Answer: B,D (LEAVE A REPLY)

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

NEW QUESTION: 63
You are working on a sensitive project involving private user data. You have set up a project on Google Cloud Platform to house your work internally. An external consultant is going to assist with coding a complex transformation in a Google Cloud Dataflow pipeline for your project. How should you maintain users' privacy?
A. Grant the consultant the Viewer role on the project.
B. Grant the consultant the Cloud Dataflow Developer role on the project.
C. Create a service account and allow the consultant to log on with it.
D. Create an anonymized sample of the data for the consultant to work with in a different project.
Answer: (SHOW ANSWER)
A service account is a special type of Google account intended to represent a non-human user that needs to authenticate and be authorized to access data in Google APIs.
https://cloud.google.com/iam/docs/understanding-service-accounts

NEW QUESTION: 64
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: 65
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.
Reference:
https://cloud.google.com/dataproc/docs/tutorials/python-library-example#create_a_new_cloud_dataproc_cluste

NEW QUESTION: 66
You need to create a new transaction table in Cloud Spanner that stores product sales data. You are deciding what to use as a primary key. From a performance perspective, which strategy should you choose?
A. A random universally unique identifier number (version 4 UUID)
B. The original order identification number from the sales system, which is a monotonically increasing integer
C. The current epoch time
D. A concatenation of the product name and the current epoch time
Answer: A (LEAVE A REPLY)

NEW QUESTION: 67
You've migrated a Hadoop job from an on-prem cluster to dataproc and GCS. Your Spark job is a complicated analytical workload that consists of many shuffing operations and initial data are parquet files (on average 200-400 MB size each). You see some degradation in performance after the migration to Dataproc, so you'd like to optimize for it. You need to keep in mind that your organization is very cost-sensitive, so you'd like to continue using Dataproc on preemptibles (with 2 non-preemptible workers only) for this workload.
What should you do?
A. Increase the size of your parquet files to ensure them to be 1 GB minimum.
B. Switch 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: C (LEAVE A REPLY)
To optimize the performance of a complex Spark job on Dataproc that heavily relies on shuffling operations, and given the cost constraints of using preemptible VMs, switching from HDDs to SSDs and using HDFS as an intermediate storage layer can significantly improve performance. Here's why option C is the best choice:
Performance of SSDs:
SSDs provide much faster read and write speeds compared to HDDs, which is crucial for performance-intensive operations like shuffling in Spark jobs.
Using SSDs can reduce I/O bottlenecks during the shuffle phase of your Spark job, improving overall job performance.
Intermediate Storage with HDFS:

Copying data from Google Cloud Storage (GCS) to HDFS for intermediate storage can reduce latency compared to reading directly from GCS.
HDFS provides better locality and faster data access within the Dataproc cluster, which can significantly improve the efficiency of shuffling and other I/O operations.
Cost Considerations:
Although SSDs are more expensive than HDDs, the performance improvement for shuffle-heavy workloads can justify the cost, especially if the improved performance reduces the overall runtime and thereby the cost of using preemptible VMs.
Using preemptible VMs with SSDs for this workload balances the cost and performance trade-offs effectively.

NEW QUESTION: 68
You need to modernize your existing on-premises data strategy. Your organization currently uses:
- Apache Hadoop clusters for processing multiple large data sets,
including on-premises Hadoop Distributed File System (HDFS) for data
replication.
- Apache Airflow to orchestrate hundreds of ETL pipelines with
thousands of job steps.
You need to set up a new architecture in Google Cloud that can handle your Hadoop workloads and requires minimal changes to your existing orchestration processes. What should you do?
A. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Convert your ETL pipelines to Dataflow.
B. Use Bigtable for your large workloads, with connections to Cloud Storage to handle any HDFS use cases. Orchestrate your pipelines with Cloud Composer.
C. Use Dataproc to migrate your Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Use Cloud Data Fusion to visually design and deploy your ETL pipelines.
D. Use Dataproc to migrate Hadoop clusters to Google Cloud, and Cloud Storage to handle any HDFS use cases. Orchestrate your pipelines with Cloud Composer.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 69
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)
Reference https://medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using- tensorflow-30505541d877

NEW QUESTION: 70
You are migrating a table to BigQuery and are deeding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID and the city and

state in which the store is located You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state city and individual store. You want to model this table to minimize query time and cost. What should you do?

A. Partition by transaction tome cluster by store ID first, then city, then stale

B. Top-level cluster by stale first, then city then store

C. Top-level cluster by store ID first, then city then state.

D. Partition by transaction time; cluster by state first, then city then store ID

Answer: (SHOW ANSWER)

NEW QUESTION: 71

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: C (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. Reference: 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: 72
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: 73
You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?
A. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster
B. Create a Cloud Dataproc Workflow Template
C. Create a Directed Acyclic Graph in Cloud Composer
D. Create an initialization action to execute the jobs
Answer: C (LEAVE A REPLY)

NEW QUESTION: 74
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. Grant the user access to Google Cloud Shell.
B. Run a local version of Jupiter on the laptop.
C. Deploy Google Cloud Datalab to a virtual machine (VM) on Google Compute Engine.
D. Host a visualization tool on a VM on Google Compute Engine.
Answer: (SHOW ANSWER)

NEW QUESTION: 75
You have a data analyst team member who needs to analyze data by using BigQuery. The data analyst wants to create a data pipeline that would load 200 CSV files with an average size of 15MB from a Cloud Storage bucket into BigQuery daily. The data needs to be ingested and transformed before being accessed in BigQuery for analysis. You need to recommend a fully managed, no-code solution for the data analyst. What should you do?
A. Create a Cloud Run function and schedule it to run daily using Cloud Scheduler to load the data into BigQuery.
B. Use the BigQuery Data Transfer Service to load files from Cloud Storage to BigQuery, create a BigQuery job which transforms the data using BigQuery SQL and schedule it to run daily.
C. Build a custom Apache Beam pipeline and run it on Dataflow to load the file from Cloud Storage to BigQuery and schedule it to run daily using Cloud Composer.
D. Create a pipeline by using BigQuery pipelines and schedule it to load the data into BigQuery daily.
Answer: B (LEAVE A REPLY)
The requirements are for a daily scheduled load, ingest, and transformation, and specifically a fully managed, no-code solution.
* Ingest (Load): The BigQuery Data Transfer Service (DTS) is the fully managed, serverless, and no- code solution for batch loading files (including CSV from Cloud Storage) into BigQuery on a schedule. This is the "ingest" part.
* Transform: After loading the raw data into a staging table using DTS, the transformation can be done using BigQuery SQL. This transformation query can then be automated using a Scheduled Query in BigQuery, which is also a fully managed and no-code feature that runs on a schedule.
* Fully Managed & No-Code: Both DTS for Cloud Storage and Scheduled Queries are native BigQuery features that are fully managed and configured through the console without requiring code, directly meeting the constraints.
* Correcting other options:
* A (Cloud Run + Script): Cloud Run requires writing a custom Python script, which violates the no-code requirement.
* C (Dataflow + Apache Beam + Cloud Composer): This is a powerful, highly scalable ETL solution, but it requires writing custom code (Apache Beam) and requires setting up and managing a workflow orchestrator (Cloud Composer/Airflow), which violates both the fully managed (Dataflow is serverless, but the code/pipeline itself is custom and needs maintenance) and no-code requirements.
* D (BigQuery pipelines): "BigQuery pipelines" is not a distinct, official product name in the Google Cloud documentation that fulfills a no-code scheduled ETL. The closest product is the combination of DTS and Scheduled Queries, as described in option B.
Reference: Google Cloud Documentation on BigQuery Data Transfer Service and Scheduled Queries:
"The BigQuery Data Transfer Service automates data movement into BigQuery on a scheduled, managed basis... The BigQuery Data Transfer Service supports loading data from Cloud Storage in one of the following formats: Comma-separated values (CSV)..." (Source: What is BigQuery Data Transfer Service? and Introduction to Cloud Storage transfers)
"A scheduled query is a query that BigQuery automatically runs at regular intervals. When you configure a scheduled query, you specify the GoogleSQL SELECT statement to run, the destination table for the query results, and the frequency of the query." (Source: Scheduling queries) This combination delivers a fully managed, no-code ELT (Extract-Load-Transform) pipeline.

NEW QUESTION: 76
You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?
A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name
B. Update the Cloud Dataflow pipeline inflight by passing the --updateoption with the --jobNameset to a new unique job name

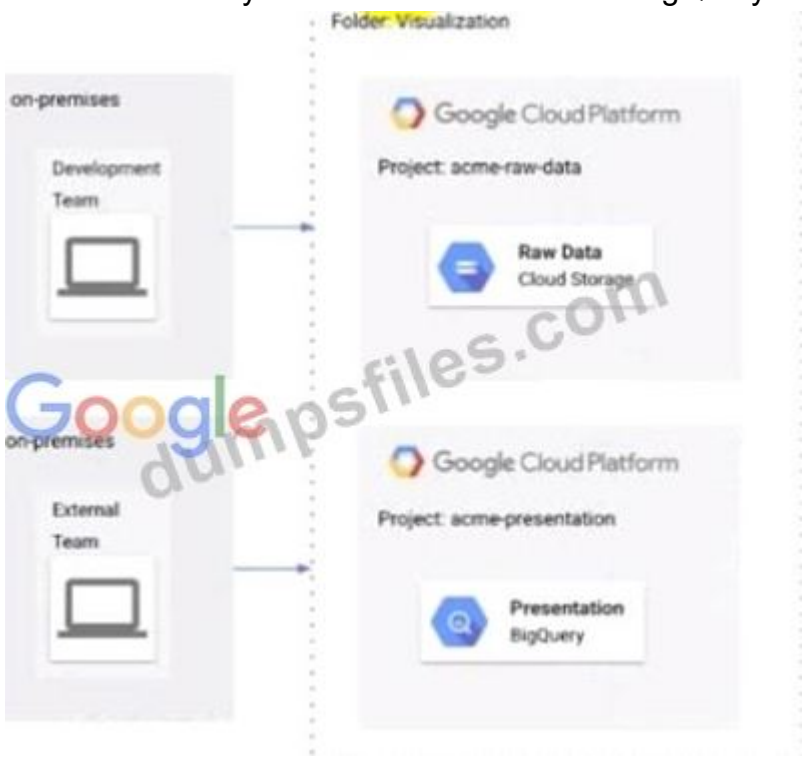
C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code
D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code
Answer: A (LEAVE A REPLY)
Explanation/Reference: https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline

NEW QUESTION: 77
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: C (LEAVE A REPLY)
Use L1 regularization when you need to assign greater importance to more influential features. It shrinks less important feature to 0.
L2 regularization performs better when all input features influence the output & all with the weights are of equal size.

NEW QUESTION: 78
You need to look at BigQuery data from a specific table multiple times a day. The underlying table you are querying is several petabytes in size, but you want to filter your data and provide simple aggregations to downstream users. You want to run queries faster and get up-to-date insights quicker. What should you do?
A. Run a scheduled query to pull the necessary data at specific intervals daily.
B. Create a materialized view based off of the query being run.
C. Use a cached query to accelerate time to results.
D. Limit the query columns being pulled in the final result.
Answer: B (LEAVE A REPLY)
Materialized views are precomputed views that periodically cache the results of a query for increased performance and efficiency. BigQuery leverages precomputed results from materialized views and whenever possible reads only changes from the base tables to compute up-to-date results. Materialized views can significantly improve the performance of workloads that have the characteristic of common and repeated queries. Materialized views can also optimize queries with high computation cost and small dataset results, such as filtering and aggregating large tables. Materialized views are refreshed automatically when the base tables change, so they

always return fresh data. Materialized views can also be used by the BigQuery optimizer to process queries to the base tables, if any part of the query can be resolved by querying the materialized view. References:
* Introduction to materialized views
* Create materialized views
* BigQuery Materialized View Simplified: Steps to Create and 3 Best Practices
* Materialized view in Bigquery

NEW QUESTION: 79
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 BigQuery as a restricted API Add the External Team users to the perimeter s Access Level
B. Create Virtual Private Cloud (VPC) firewall rules on the acme-raw-data protect that deny all Ingress traffic from the External Team CIDR range
C. Remove Cloud Storage IAM permissions to the External Team on the acme-raw-data project
D. 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
Answer: A (LEAVE A REPLY)

NEW QUESTION: 80
You need to deploy additional dependencies to all of a Cloud Dataproc cluster at startup using an existing initialization action. Company security policies require that Cloud Dataproc nodes do not have access to the Internet so public initialization actions cannot fetch resources. What should you do?
A. Deploy the Cloud SQL Proxy on the Cloud Dataproc master

B. Use an SSH tunnel to give the Cloud Dataproc cluster access to the Internet
C. Copy all dependencies to a Cloud Storage bucket within your VPC security perimeter
D. Use Resource Manager to add the service account used by the Cloud Dataproc cluster to the Network User role
Answer: C (LEAVE A REPLY)
https://cloud.google.com/dataproc/docs/concepts/configuring-clusters/init-actions

NEW QUESTION: 81
An external customer provides you with a daily dump of data from their database. The data flows into Google Cloud Storage GCS as comma-separated values (CSV) files. You want to analyze this data in Google BigQuery, but the data could have rows that are formatted incorrectly or corrupted. How should you build this pipeline?
A. Enable BigQuery monitoring in Google Stackdriver and create an alert.
B. Use federated data sources, and check data in the SQL query.
C. Import the data into BigQuery using the gcloud CLI and set max_bad_records to 0.
D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 82
You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on
Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed.
What should you do?
A. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
B. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
C. Create a dedicated service account, and use encryption at rest to reference your data stored in your
Compute Engine cluster instances as part of your API service calls.
D. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 83
You are building a model to make clothing recommendations. You know a user's fashion pis likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?
A. Continuously retrain the model on just the new data.
B. Continuously retrain the model on a combination of existing data and the new data.
C. Train on the existing data while using the new data as your test set.
D. Train on the new data while using the existing data as your test set.
Answer: (SHOW ANSWER)
We have to use a combination of old and new test data as well as training data.

NEW QUESTION: 84
You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in-time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment. You want to archive these snapshots for a long time. Which two methods can accomplish this? (Choose two.)
A. Write an application that uses Cloud Datastore client libraries to read all the entities. Treat each entity as a BigQuery table row via BigQuery streaming insert. Assign an export timestamp for each export, and attach it as an extra column for each row. Make sure that the BigQuery table is partitioned using the export timestamp column.
B. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
C. 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.
D. Use managed export, and then import to Cloud Datastore in a separate project under a unique namespace reserved for that export.
E. Use managed export, and then import the data into a BigQuery table created just for that export, and delete temporary export files.
Answer: (SHOW ANSWER)

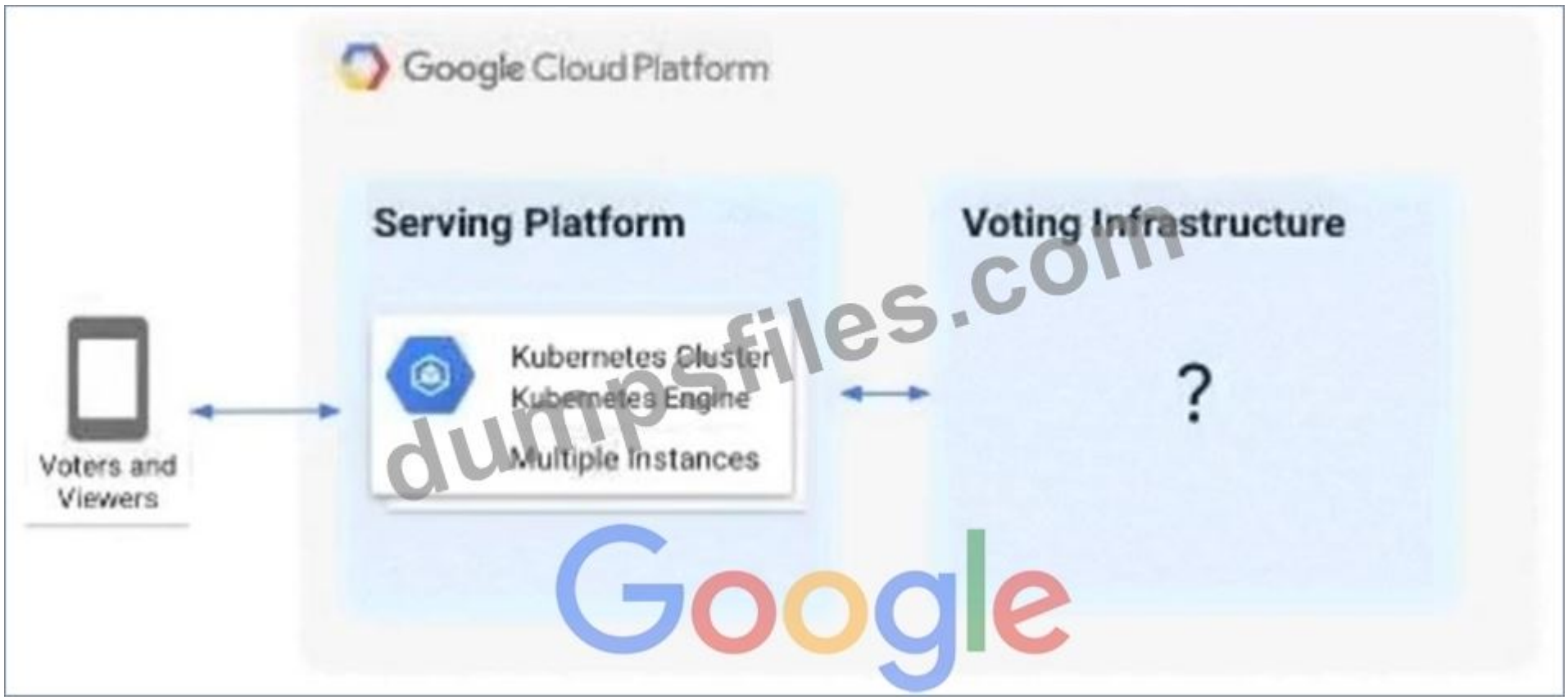
NEW QUESTION: 85
Your team is building a data lake platform on Google Cloud. As a part of the data foundation design, you are planning to store all the raw data in Cloud Storage. You are expecting to ingest approximately 25 GB of data a day and your billing department is worried about the increasing cost of storing old data. The current business requirements are:
- The old data can be deleted anytime.
- There is no predefined access pattern of the old data.
- The old data should be available instantly when accessed.
- There should not be any charges for data retrieval.
What should you do to optimize for cost?
A. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to coldline, 90 days to nearline, and 365 days to archive storage class. Delete old data as needed.
B. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 90 days to coldline, and 365 days to archive storage class. Delete old data as needed.
C. Create the bucket with the Autoclass storage class feature.
D. Create an Object Lifecycle Management policy to modify the storage class for data older than 30 days to nearline, 45 days to coldline, and 60 days to archive storage class. Delete old data as needed.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 86
You are migrating your data warehouse to Google Cloud and decommissioning your on-premises data center. Because this is a priority for your company, you know that bandwidth will be made available for the initial data load to the cloud. The files being transferred are not large in number, but each file is 90 GB Additionally, you want your transactional systems to continually update the warehouse on Google Cloud in real time. What tools should you use to migrate the data and ensure that it continues to write to your warehouse?
A. Storage Transfer Service for the migration, Pub/Sub and Cloud Data Fusion for the real-time updates

B. BigQuery Data Transfer Service for the migration, Pub/Sub and Dataproc for the real-time updates
C. gsutil for the migration; Pub/Sub and Dataflow for the real-time updates
D. gsutil for both the migration and the real-time updates
Answer: C (LEAVE A REPLY)
Use Gsutil when there is enough bandwidth to meet your project deadline for less than 1 TB of data.
Storage Transfer Service is for much larger volumes for migration. Moreover, Cloud Data Fusion and Dataproc are not ideal for real-time updates. BigQuery Data Transfer Service does not support all on-prem sources.
Reference:
https://cloud.google.com/architecture/migration-to-google-cloud-transferring-your-large-datasets#gsutil_for_smaller_transfers_of_on-premises_data

NEW QUESTION: 87
You created an analytics environment on Google Cloud so that your data scientist team can explore data without impacting the on-premises Apache Hadoop solution. The data in the on-premises Hadoop Distributed File System (HDFS) cluster is in Optimized Row Columnar (ORC) formatted files with multiple columns of Hive partitioning. The data scientist team needs to be able to explore the data in a similar way as they used the on-premises HDFS cluster with SQL on the Hive query engine. You need to choose the most cost-effective storage and processing solution. What should you do?
A. Copy the ORC files on Cloud Storage, then deploy a Dataproc cluster for the data scientist team.
B. Import the ORC files lo Bigtable tables for the data scientist team.
C. Import the ORC files to BigOuery tables for the data scientist team.
D. Copy the ORC files on Cloud Storage, then create external BigQuery tables for the data scientist team.
Answer: (SHOW ANSWER)

NEW QUESTION: 88
A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3 minute period. You are in charge of the Voting restructure* and must ensure that the platform can handle the load and Hal all votes are processed. You must display partial results write voting is open. After voting doses you need to count the votes exactly once white optimizing cost. What should you do?



A. Write votes to a Pub/Sub tope and toad into both Bigtable and BigQuery via a Dataflow pipeline Query Bigtable for real-time results and BigQuery for later analysis Shutdown the Bigtable instance when voting concludes

B. Write votes to a Pub Sub tope and have Cloud Functions subscribe to it and write voles to BigQuery

C. Create a Memorystore instance with a high availability (HA) configuration

D. Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas

Answer: A (LEAVE A REPLY)

NEW QUESTION: 89

You are collecting loT 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:

```
SELECT
    MAX(temperature)
FROM
    acme_iot_data.sensors
WHERE
    create_date > DATE_SUB(CURRENT_DATE(), INTERVAL 7 day)
    AND location_id = "SW1W9TQ"
    AND device_version = "202007r3"
```


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

A. Partition table data by create_date cluster table data by tocation_id and device_version
B. Cluster table data by create_date location_id and device_version
C. Partition table data by create_date, location_id and device_version
D. Cluster table data by create_date, partition by location and device_version
Answer: A (LEAVE A REPLY)

NEW QUESTION: 90
Your financial services company is moving to cloud technology and wants to store 50 TB of financial time-series data in the cloud. This data is updated frequently and new data will be streaming in all the time.
Your company also wants to move their existing Apache Hadoop jobs to the cloud to get insights into this data. Which product should they use to store the data?
A. Google BigQuery
B. Google Cloud Datastore
C. Google Cloud Storage
D. Cloud Bigtable
Answer: D (LEAVE A REPLY)

NEW QUESTION: 91
You are planning to use Google's Dataflow SDK to analyze customer data such as displayed below. Your project requirement is to extract only the customer name from the data source and then write to an output PCollection.
Tom,555 X street
Tim,553 Y street
Sam, 111 Z street
Which operation is best suited for the above data processing requirement?
A. ParDo
B. Sink API
C. Source API
D. Data extraction
Answer: A (LEAVE A REPLY)
In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.

NEW QUESTION: 92

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)
Dataflow is a unified processing model, and can execute both streaming and batch data pipelines Reference: https://cloud.google.com/dataflow/

NEW QUESTION: 93
You are defining the data governance strategy for a new BigQuery table with medical and financial data. You want a scalable solution that ensures the clinical researchers can access patient medical data without financial information, while allowing the accounting team to access only financial data with minimal patient identifiers. What should you do?
A. Implement column-level security policies in BigQuery tables with IAM permissions.
B. Create separate tables for personally identifiable information (PII), financial data, and anonymized medical data. Use IAM permissions to control access to each table.
C. Implement row-level security policies in BigQuery tables with IAM permissions.
D. Create separate datasets with authorized views exposing only approved data.
Answer: (SHOW ANSWER)
Authorized views allow you to expose only the approved columns or rows from a BigQuery table to specific groups, enabling scalable, fine-grained access control for different teams without duplicating the underlying data.

NEW QUESTION: 94
You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings. Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?
A. Re-write the application to load accumulated data every 2 minutes.
B. Convert the streaming insert code to batch load for individual messages.
C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.
Answer: (SHOW ANSWER)
The data is first comes to buffer and then written to Storage. If we are running queries in buffer we will face above mentioned issues. If we wait for the bigquery to write the data to storage then we won't face the issue.
So We need to wait till it's written tio storage

NEW QUESTION: 95
You recently deployed several data processing jobs into your Cloud Composer 2 environment. You notice that some tasks are failing in Apache Airflow. On the monitoring dashboard, you see an increase in the total workers' memory usage, and there were worker pod evictions. You need to resolve these errors. What should you do?
Choose 2 answers

A. Increase the directed acyclic graph (DAG) file parsing interval.
B. Increase the memory available to the Airflow workers.
C. Increase the maximum number of workers and reduce worker concurrency.
D. Increase the memory available to the Airflow triggerer.
E. Increase the Cloud Composer 2 environment size from medium to large.
Answer: B,C (LEAVE A REPLY)
To resolve issues related to increased memory usage and worker pod evictions in your Cloud Composer 2 environment, the following steps are recommended:
Increase Memory Available to Airflow Workers:
By increasing the memory allocated to Airflow workers, you can handle more memory-intensive tasks, reducing the likelihood of pod evictions due to memory limits.
Increase Maximum Number of Workers and Reduce Worker Concurrency:
Increasing the number of workers allows the workload to be distributed across more pods, preventing any single pod from becoming overwhelmed.
Reducing worker concurrency limits the number of tasks that each worker can handle simultaneously, thereby lowering the memory consumption per worker.
Steps to Implement:
Increase Worker Memory:
Modify the configuration settings in Cloud Composer to allocate more memory to Airflow workers. This can be done through the environment configuration settings.
Adjust Worker and Concurrency Settings:
Increase the maximum number of workers in the Cloud Composer environment settings.
Reduce the concurrency setting for Airflow workers to ensure that each worker handles fewer tasks at a time, thus consuming less memory per worker.
Reference:
Cloud Composer Worker Configuration
Scaling Airflow Workers

NEW QUESTION: 96
You are migrating a table to BigQuery and are deciding on the data model. Your table stores information related to purchases made across several store locations and includes information like the time of the transaction, items purchased, the store ID, and the city and state in which the store is located. You frequently query this table to see how many of each item were sold over the past 30 days and to look at purchasing trends by state, city, and individual store. How would you model this table for the best query performance?
A. Partition by transaction time; cluster by state first, then city, then store ID.
B. Partition by transaction time; cluster by store ID first, then city, then state.
C. Top-level cluster by state first, then city, then store ID.
D. Top-level cluster by store ID first, then city, then state.
Answer: A (LEAVE A REPLY)
https://cloud.google.com/bigquery/docs/querying-clustered-tables

NEW QUESTION: 97

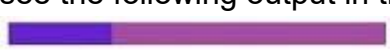
Which of the following is NOT one of the three main types of triggers that Dataflow supports?
A. Trigger based on element size in bytes
B. Trigger that is a combination of other triggers
C. Trigger based on element count
D. Trigger based on time
Answer: A (LEAVE A REPLY)
There are three major kinds of 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: 98
You are creating a data model in BigQuery that will hold retail transaction dat a. Your two largest tables, sales_transation_header and sales_transation_line. have a tightly coupled immutable relationship. These tables are rarely modified after load and are frequently joined when queried. You need to model the sales_transation_header and sales_transation_line tables to improve the performance of data analytics queries. What should you do?
A. Create a sal es_transaction table that Stores the sales_tran3action_header and sales_transaction_line data as a JSON data type.
B. Create a sale3_transaction table that holds the sales_transaction_header information as rows and the sales_transaction_line rows as nested and repeated fields.
C. Create a sale_transaction table that holds the sales_transaction_header and sales_transaction_line information as rows, duplicating the sales_transaction_header data for each line.
D. Create separate sales_transation_header and sales_transation_line tables and. when querying, specify the sales transition line first in the WHERE clause.
Answer: B (LEAVE A REPLY)
BigQuery supports nested and repeated fields, which are complex data types that can represent hierarchical and one-to-many relationships within a single table. By using nested and repeated fields, you can denormalize your data model and reduce the number of joins required for your queries. This can improve the performance and efficiency of your data analytics queries, as joins can be expensive and require shuffling data across nodes. Nested and repeated fields also preserve the data integrity and avoid data duplication. In this scenario, the sales_transaction_header and sales_transaction_line tables have a tightly coupled immutable relationship, meaning that each header row corresponds to one or more line rows, and the data is rarely modified after load. Therefore, it makes sense to create a single sales_transaction table that holds the sales_transaction_header information as rows and the sales_transaction_line rows as nested and repeated fields. This way, you can query the sales transaction data without joining two tables, and use dot notation or array functions to access the nested and repeated fields. For example, the sales_transaction table could have the following schema:
Table
Field name
Type
Mode
id
INTEGER
NULLABLE

order_time
TIMESTAMP
NULLABLE
customer_id
INTEGER
NULLABLE
line_items
RECORD
REPEATED
line_items.sku
STRING
NULLABLE
line_items.quantity
INTEGER
NULLABLE
line_items.price
FLOAT
NULLABLE
To query the total amount of each order, you could use the following SQL statement:
SQL

```
SELECT id, SUM(line_items.quantity * line_items.price) AS total_amount
FROM sales_transaction
GROUP BY id;
```

AI-generated code. Review and use carefully. More info on FAQ.
Reference:
Use nested and repeated fields
BigQuery explained: Working with joins, nested & repeated data
Arrays in BigQuery - How to improve query performance and optimise storage

NEW QUESTION: 99
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. Either the state or the city columns in the [myproject:mydataset.mytable] table have too many NULL values
B. Most rows in the [myproject:mydataset.mytable] table have the same value in the country column, causing data skew
C. The [myproject:mydataset.mytable] table has too many partitions
D. Users are running too many concurrent queries in the system
Answer: (SHOW ANSWER)

NEW QUESTION: 100
You are creating the CI'CD cycle for the code of the directed acyclic graphs (DAGs) running in Cloud Composer. Your team has two Cloud Composer instances: one instance for development and another instance for production. Your team is using a Git repository to maintain and develop the code of the DAGs. You want to deploy the DAGs automatically to Cloud Composer when a certain tag is pushed to the Git repository. What should you do?
A. 1 Use Cloud Build to copy the code of the DAG to the Cloud Storage bucket of the development instance for DAG testing.
2. If the tests pass, use Cloud Build to build a container with the code of the DAG and the KubernetesPodOperator to deploy the container to the Google Kubernetes Engine (GKE) cluster of the production instance.
B. 1 Use Cloud Build to copy the code of the DAG to the Cloud Storage bucket of the development instance for DAG testing.
2. If the tests pass, use Cloud Build to copy the code to the bucket of the production instance.
C. 1 Use Cloud Build to build a container with the code of the DAG and the KubernetesPodOperator to deploy the code to the Google Kubernetes Engine (GKE) cluster of the development instance for testing.
2. If the tests pass, use the KubernetesPodOperator to deploy the container to the GKE cluster of the production instance.
D. 1. Use Cloud Build to build a container and the Kubemetes Pod Operator to deploy the code of the DAG to the Google Kubernetes Engine (GKE) cluster of the development instance for testing.
2. If the tests pass, copy the code to the Cloud Storage bucket of the production instance.
Answer: (SHOW ANSWER)

NEW QUESTION: 101
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: A (LEAVE A REPLY)
Huge amount of data with log network bandwidth, Transfer applicate is best for moving data over
100TB.

NEW QUESTION: 102
You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?
A. Update the current pipeline and use the drain flag.
B. Update the current pipeline and provide the transform mapping JSON object.
C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.
Answer: A (LEAVE A REPLY)
The key requirement is not to lose the data, the Dataflow pipeline can be stopped using the Drain option.

Drain options would cause Dataflow to stop any new processing, but would also allow the existing processing to complete.

NEW QUESTION: 103
You are building a data pipeline on Google Cloud. You need to prepare data using a casual method for a machine-learning process. You want to support a logistic regression model. You also need to monitor and adjust for null values, which must remain real-valued and cannot be removed. What should you do?
A. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to `none' using a Cloud Dataproc job.
B. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to 0 using a Cloud Dataprep job.
C. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.
D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to `none' using a Cloud Dataprep job.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 104
Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use Hadoop jobs they have already created and minimize the management of the cluster as much as possible.
They also want to be able to persist data beyond the life of the cluster. What should you do?
A. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.
B. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
C. Create a Google Cloud Dataflow job to process the data.
D. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
E. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 105
The Dataflow SDKs have been recently transitioned into which Apache service?
A. Apache Spark
B. Apache Hadoop
C. Apache Kafka
D. Apache Beam
Answer: D (LEAVE A REPLY)
Explanation
Dataflow SDKs are being transitioned to Apache Beam, as per the latest Google directive Reference:
https://cloud.google.com/dataflow/docs/

NEW QUESTION: 106
You are designing a stateful data processing pipeline that reads data from a Cloud Storage bucket and writes transformed data to a BigQuery table. The pipeline must be highly available and resilient to zonal failures within the us-central1 region. You need to configure a Dataflow pipeline ensuring minimal disruption during a zonal outage. What should you do?
A. Deploy the Dataflow job to a single zone within us-central1 and configure it to use a regional persistent disk to store its state.
B. Launch the Dataflow job with the --region us-central1 parameter.
C. Deploy the Dataflow job to a single zone within us-central1 and use a multi-regional Cloud Storage bucket to store its state.

D. Launch the Dataflow job with the --zone us-central1-a parameter.
Answer: B (LEAVE A REPLY)
To ensure high availability and resilience against zonal outages in Dataflow, you must leverage Regional Endpoints.
* Regional Placement: By using the --region flag (e.g., us-central1) without specifying a specific --zone, Dataflow automatically places workers in different zones within that region. If a specific zone in us- central1 fails, the Dataflow service backend can automatically restart or migrate workers to a healthy zone within the same region.
* State Management: Dataflow handles state (shuffling and persistent state) natively. By operating at the regional level, the control plane can maintain the job's continuity even if a zone goes offline.
* Correcting other options:
* A & D: Specifying a single zone (either by the --zone parameter or by manual deployment to one zone) creates a single point of failure. If that zone goes down, the pipeline stops entirely.
* C: While multi-regional GCS is highly available, it doesn't solve the compute resilience issue. If the Dataflow workers are in a single zone and that zone fails, the processing stops regardless of where the data is stored.
Reference: Google Cloud Documentation on Dataflow Regional Endpoints:
"By default, the Dataflow service runs your job in a regional endpoint. Using a regional endpoint provides high availability and geographic redundancy... If you specify a region but not a zone, Dataflow automatically selects a zone within that region to run your workers. This helps ensure that your job is resilient to zonal failures." (Source: Dataflow Regional Endpoints)
"In the event of a zonal outage, the Dataflow service attempts to schedule workers in another zone within the same region." (Source: Dataflow high availability)

NEW QUESTION: 107
A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3 minute period. You are in charge of the Voting restructure* and must ensure that the platform can handle the load and Hal all votes are processed. You must display partial results write voting is open.
After voting doses you need to count the votes exactly once white optimizing cost. What should you do?
A. Create a Memorystore instance with a high availability (HA) configuration
B. Write votes to a Pub/Sub tope and toad into both Bigtable and BigQuery via a Dataflow pipeline Query Bigtable for real-time results and BigQuery for later analysis Shutdown the Bigtable instance when voting concludesD Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas
C. Write votes to a Pub Sub tope and have Cloud Functions subscribe to it and write voles to BigQuery
Answer: B (LEAVE A REPLY)

NEW QUESTION: 108

Your organization is modernizing their IT services and migrating to Google Cloud. You need to organize the data that will be stored in Cloud Storage and BigQuery. You need to enable a data mesh approach to share the data between sales, product design, and marketing departments.
What should you do?
A. 1. Create multiple projects for storage of the data for each of your departments' applications.
2. Enable each department to create Cloud Storage buckets and BigQuery datasets.
3. In Dataplex, map each department to a data lake and the Cloud Storage buckets, and map the BigQuery datasets to zones.
4. Enable each department to own and share the data of their data lakes.
B. 1. Create a project for storage of the data for your organization.
2. Create a central Cloud Storage bucket with three folders to store the files for each department.
3. Create a central BigQuery dataset with tables prefixed with the department name.
4. Give viewer rights for the storage project for the users of your departments.
C. 1. Create multiple projects for storage of the data for each of your departments' applications.
2. Enable each department to create Cloud Storage buckets and BigQuery datasets.
3. Publish the data that each department shared in Analytics Hub.
4. Enable all departments to discover and subscribe to the data they need in Analytics Hub.
D. 1. Create a project for storage of the data for each of your departments.
2. Enable each department to create Cloud Storage buckets and BigQuery datasets.
3. Create user groups for authorized readers for each bucket and dataset.
4. Enable the IT team to administer the user groups to add or remove users as the departments' request.
Answer: A (LEAVE A REPLY)

NEW QUESTION: 109
As your organization expands its usage of GCP, many teams have started to create their own projects.
Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects.
Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way.
You want to simplify access control management by minimizing the number of policies.
Which two steps should you take? (Choose two.)
A. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.
B. Create distinct groups for various teams, and specify groups in Cloud IAM policies.
C. Introduce resource hierarchy to leverage access control policy inheritance.
D. Use Cloud Deployment Manager to automate access provision.
E. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.
Answer: B,D (LEAVE A REPLY)

NEW QUESTION: 110
You are building a model to make clothing recommendations. You know a user's fashion preference is likely to change over time, so you build a data pipeline to stream new data back to the model as it becomes available. How should you use this data to train the model?

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

NEW QUESTION: 111
You need to create a SQL pipeline. The pipeline runs an aggregate SQL transformation on a BigQuery table every two hours and appends the result to another existing BigQuery table. You need to configure the pipeline to retry if errors occur. You want the pipeline to send an email notification after three consecutive failures. What should you do?
A. Create a BigQuery scheduled query to run the SQL transformation with schedule options that repeats every two hours, and enable notification to Pub/Sub topic. Use Pub/Sub and Cloud Functions to send an email after three failed executions.
B. Use the BigQueryInsertJobOperator in Cloud Composer, set the retry parameter to three, and set the email_on_failure parameter to true.
C. Create a BigQuery scheduled query to run the SQL transformation with schedule options that repeats every two hours, and enable email notifications.
D. Use the BigQueryUpsertTableOperator in Cloud Composer, set the retry parameter to three, and set the email_on_failure parameter to true.
Answer: (SHOW ANSWER)

NEW QUESTION: 112
How can you get a neural network to learn about relationships between categories in a categorical feature?
A. Create a multi-hot column
B. Create a one-hot column
C. Create a hash bucket
D. Create an embedding column
Answer: (SHOW ANSWER)
There are two problems with one-hot encoding. First, it has high dimensionality, meaning that instead of having just one value, like a continuous feature, it has many values, or dimensions.
This makes computation more time-consuming, especially if a feature has a very large number of categories. The second problem is that it doesn't encode any relationships between the categories. They are completely independent from each other, so the network has no way of knowing which ones are similar to each other.
Both of these problems can be solved by representing a categorical feature with an embedding column. The idea is that each category has a smaller vector with, let's say, 5 values in it. But unlike a one-hot vector, the values are not usually 0. The values are weights, similar to the weights that are used for basic features in a neural network. The difference is that each category has a set of weights (5 of them in this case).
You can think of each value in the embedding vector as a feature of the category. So, if two categories are very similar to each other, then their embedding vectors should be very similar too.
Reference:
https://cloudacademy.com/google/introduction-to-google-cloud-machine-learning-engine-course/a-wide-and-deep-model.html

NEW QUESTION: 113
Your Cloud Storage data lake has raw, processed, and historical data in different buckets. Data older than two years is rarely accessed, and all data must be retained for no longer than seven years. You are concerned about rising storage costs. How should you control costs for the historical data bucket?
A. Write a script on a Compute Engine instance, triggered daily by Cloud Scheduler, to scan all objects and delete any older than seven years.
B. Configure an Object Lifecycle Management rule to transition objects older than two years to the Archive storage class and eventually delete them after seven years.
C. Enable the Autoclass feature on your Cloud Storage buckets and select Opt-in to object transitions to Coldline and Archive storage classes.
D. Replicate the buckets to a different region with lower storage costs and configure an Object Lifecycle Management rule to delete objects after seven years.
Answer: (SHOW ANSWER)
Object Lifecycle Management rules allow you to automatically transition infrequently accessed objects to lower-cost storage classes, such as Archive, after a defined age and to delete them after the seven-year retention limit. This native, policy-driven approach minimizes storage costs over time while ensuring compliance with data retention requirements without manual intervention.

NEW QUESTION: 114
You are choosing a NoSQL database to handle telemetry data submitted from millions of Internet-of-Things (IoT) devices. The volume of data is growing at 100 TB per year, and each data entry has about 100 attributes. The data processing pipeline does not require atomicity, consistency, isolation, and durability (ACID). However, high availability and low latency are required.
You need to analyze the data by querying against individual fields. Which three databases meet your requirements? (Choose three.)
A. Redis
B. HBase
C. MySQL
D. MongoDB
E. Cassandra
F. HDFS with Hive
Answer: (SHOW ANSWER)
Explanation:

NEW QUESTION: 115
You are preparing data to serve a sales demand prediction model. The training data undergoes several pre-processing steps, including scaling numerical features and one-hot encoding categorical features. The model is deployed on Vertex AI Endpoints. You need to prevent training-serving skew and ensure accurate predictions in production. You want a solution that is easy to implement. What should you do?
A. Implement a custom handler within the Vertex AI Endpoint to automatically perform data transformations before the model makes a prediction.
B. Replicate the exact same pre-processing logic in the inference pipeline that was used during model training.
C. Store the raw, unprocessed data in a separate Cloud Storage bucket exclusively for serving.
D. Ensure the serving data is a smaller, random sample of the training data.

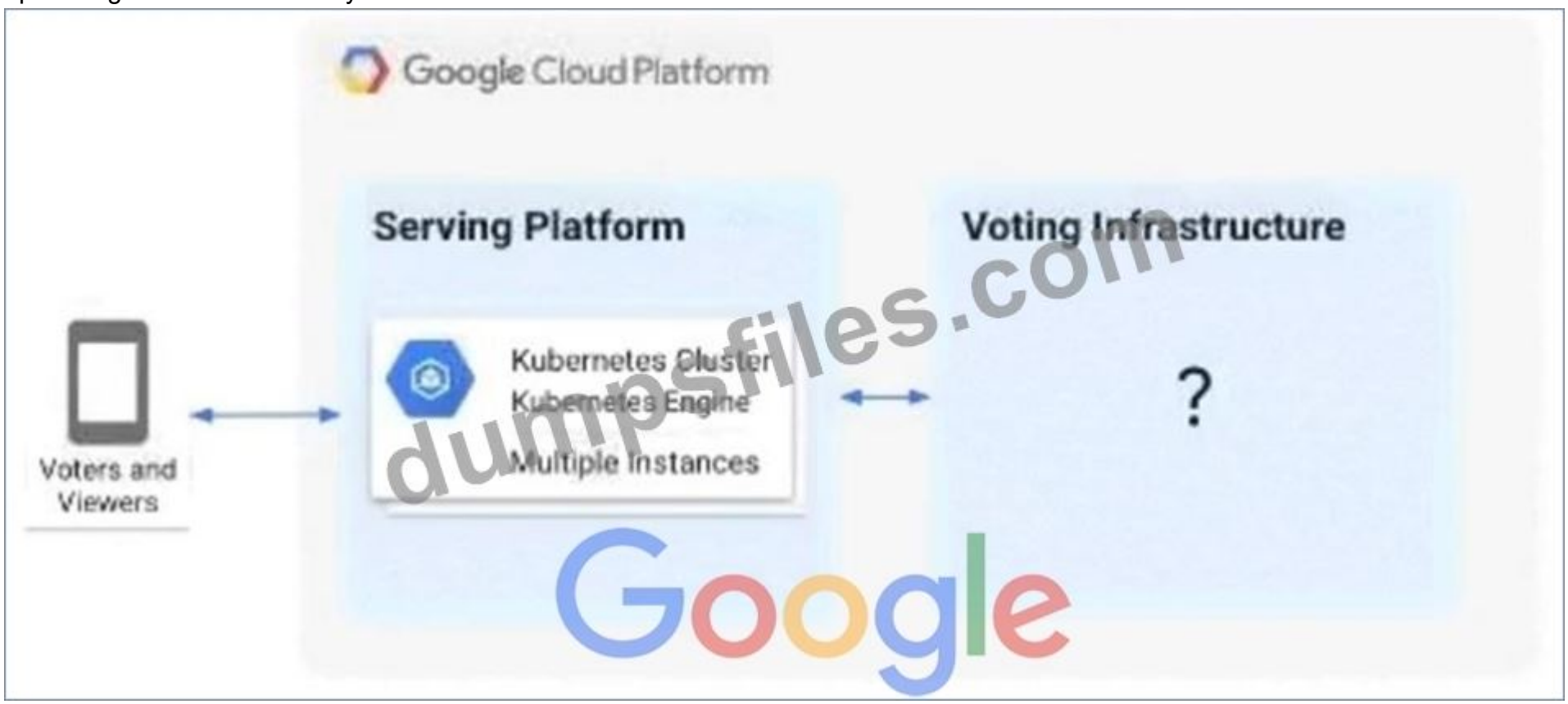
Answer: B (LEAVE A REPLY)
Replicating the exact pre-processing logic from training in the inference pipeline ensures consistency between training and serving, preventing training-serving skew and maintaining prediction accuracy without adding unnecessary complexity.

NEW QUESTION: 116
If you're running a performance test that depends upon Cloud Bigtable, all the choices except one below are recommended steps. Which is NOT a recommended step to follow?
A. Do not use a production instance.
B. Run your test for at least 10 minutes.
C. Before you test, run a heavy pre-test for several minutes.
D. Use at least 300 GB of data.
Answer: A (LEAVE A REPLY)
Explanation
If you're running a performance test that depends upon Cloud Bigtable, be sure to follow these steps as you plan and execute your test:
Use a production instance. A development instance will not give you an accurate sense of how a production instance performs under load.
Use at least 300 GB of data. Cloud Bigtable performs best with 1 TB or more of data. However, 300 GB of data is enough to provide reasonable results in a performance test on a 3-node cluster. On larger clusters, use
100 GB of data per node.
Before you test, run a heavy pre-test for several minutes. This step gives Cloud Bigtable a chance to balance data across your nodes based on the access patterns it observes.
Run your test for at least 10 minutes. This step lets Cloud Bigtable further optimize your data, and it helps ensure that you will test reads from disk as well as cached reads from memory.
Reference: https://cloud.google.com/bigtable/docs/performance

NEW QUESTION: 117
You want to store your team's shared tables in a single dataset to make data easily accessible to various analysts. You want to make this data readable but unmodifiable by analysts. At the same time, you want to provide the analysts with individual workspaces in the same project, where they can create and store tables for their own use, without the tables being accessible by other analysts. What should you do?
A. Give analysts the BigQuery Data Viewer role at the project level Create one other dataset, and give the analysts the BigQuery Data Editor role on that dataset.
B. Give analysts the BigQuery Data Viewer role at the project level Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the project level.
C. Give analysts the BigQuery Data Viewer role on the shared dataset. Create a dataset for each analyst, and give each analyst the BigQuery Data Editor role at the dataset level for their assigned dataset
D. Give analysts the BigQuery Data Viewer role on the shared dataset Create one other dataset and give the analysts the BigQuery Data Editor role on that dataset.
Answer: C (LEAVE A REPLY)

The BigQuery Data Viewer role allows users to read data and metadata from tables and views, but not to modify or delete them. By giving analysts this role on the shared dataset, you can ensure that they can access the data for analysis, but not change it. The BigQuery Data Editor role allows users to create, update, and delete tables and views, as well as read and write data. By giving analysts this role at the dataset level for their assigned dataset, you can provide them with individual workspaces where they can store their own tables and views, without affecting the shared dataset or other analysts' datasets. This way, you can achieve both data protection and data isolation for your team. Reference:
BigQuery IAM roles and permissions
Basic roles and permissions

NEW QUESTION: 118
A live TV show asks viewers to cast votes using their mobile phones. The event generates a large volume of data during a 3-minute period. You are in charge of the "Voting infrastructure" and must ensure that the platform can handle the load and that all votes are processed. You must display partial results while voting is open. After voting closes, you need to count the votes exactly once while optimizing cost. What should you do?



A. Write votes to a Pub/Sub topic and have Cloud Functions subscribe to it and write votes to BigQuery.
B. Write votes to a Pub/Sub topic and load into both Bigtable and BigQuery via a Dataflow pipeline.Query Bigtable for real-time results and BigQuery for later analysis. Shut down the Bigtable instance when voting concludes.
C. Create a Memorystore instance with a high availability (HA) configuration.
D. Create a Cloud SQL for PostgreSQL database with high availability (HA) configuration and multiple read replicas.
Answer: (SHOW ANSWER)

NEW QUESTION: 119
What is the general recommendation when designing your row keys for a Cloud Bigtable schema?

A. Include multiple time series values within the row key
B. Keep the row keep as an 8 bit integer
C. Keep your row key reasonably short
D. Keep your row key as long as the field permits
Answer: (SHOW ANSWER)
A general guide is to, keep your row keys reasonably short. Long row keys take up additional memory and storage and increase the time it takes to get responses from the Cloud Bigtable server.

NEW QUESTION: 120
You are designing a basket abandonment system for an ecommerce company. The system will send a message to a user based on these rules:
* No interaction by the user on the site for 1 hour
* Has added more than $30 worth of products to the basket
* Has not completed a transaction
You use Google Cloud Dataflow to process the data and decide if a message should be sent. How should you design the pipeline?
A. Use a sliding time window with a duration of 60 minutes.
B. Use a global window with a time based trigger with a delay of 60 minutes.
C. Use a session window with a gap time duration of 60 minutes.
D. Use a fixed-time window with a duration of 60 minutes.
Answer: (SHOW ANSWER)

NEW QUESTION: 121
You are building an ELT solution in BigQuery by using Dataform. You need to perform uniqueness and null value checks on your final tables. What should you do to efficiently integrate these checks into your pipeline?
A. Create Dataplex data quality tasks.
B. Build Dataform assertions into your code.
C. Write a Spark-based stored procedure.
D. Build BigQuery user-defined functions (UDFs).
Answer: B (LEAVE A REPLY)

NEW QUESTION: 122
When a Cloud Bigtable node fails, ____ is lost.
A. all data

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

NEW QUESTION: 123
An aerospace company uses a proprietary data format to store its night data. 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. Use Apache Hive to write a Dataproc job that streams the data into BigQuery in CSV format
C. Write a shell script that triggers a Cloud Function that performs periodic ETL batch jobs on the new data source
D. Use an Apache Beam custom connector to write a Dataflow pipeline that streams the data into BigQuery in Avro format
Answer: D (LEAVE A REPLY)

NEW QUESTION: 124
Your company's data platform ingests CSV file dumps of booking and user profile data from upstream sources into Cloud Storage. The data analyst team wants to join these datasets on the email field available in both the datasets to perform analysis. However, personally identifiable information (PII) should not be accessible to the analysts. You need to de-identify the email field in both the datasets before loading them into BigQuery for analysts. What should you do?
A. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud Data Loss Prevention (Cloud DLP) with masking as the de-identification transformations type.
2. Load the booking and user profile data into a BigQuery table.
B. 1. Create a pipeline to de-identify the email field by using recordTransformations in Cloud DLP with format-preserving encryption with FFX as the de-identification transformation type.
2. Load the booking and user profile data into a BigQuery table.
C. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.
2. Create a policy tag with the email mask as the data masking rule.
3. Assign the policy to the email field in both tables. A
4. Assign the Identity and Access Management bigquerydatapolicy.maskedReader role for the BigQuery tables to the analysts.
D. 1. Load the CSV files from Cloud Storage into a BigQuery table, and enable dynamic data masking.

2. Create a policy tag with the default masking value as the data masking rule.
3. Assign the policy to the email field in both tables.
4. Assign the Identity and Access Management bigquerydatapolicy.maskedReader role for the BigQuery tables to the analysts
Answer: B (LEAVE A REPLY)
Cloud DLP is a service that helps you discover, classify, and protect your sensitive data. It supports various de-identification techniques, such as masking, redaction, tokenization, and encryption. Format-preserving encryption (FPE) with FFX is a technique that encrypts sensitive data while preserving its original format and length. This allows you to join the encrypted data on the same field without revealing the actual values. FPE with FFX also supports partial encryption, which means you can encrypt only a portion of the data, such as the domain name of an email address. By using Cloud DLP to de-identify the email field with FPE with FFX, you can ensure that the analysts can join the booking and user profile data on the email field without accessing the PII. You can create a pipeline to de-identify the email field by usingrecordTransformations in Cloud DLP, which allows you to specify the fields and the de-identification transformations to apply to them.
You can then load the de-identified data into a BigQuery table for analysis. References:
* De-identify sensitive data | Cloud Data Loss Prevention Documentation
* Format-preserving encryption with FFX | Cloud Data Loss Prevention Documentation
* De-identify and re-identify data with the Cloud DLP API
* De-identify data in a pipeline

NEW QUESTION: 125
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: A (LEAVE A REPLY)
Huge amount of data with log network bandwidth, Transfer applicate is best for moving data over 100TB.

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

NEW QUESTION: 127
You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?
A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
B. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
C. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.
D. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 128
You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?
A. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster
B. Create a Directed Acyclic Graph in Cloud Composer
C. Create an initialization action to execute the jobs
D. Create a Cloud Dataproc Workflow Template
Answer: D (LEAVE A REPLY)

NEW QUESTION: 129
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: 130
You have a Standard Tier Memorystore for Redis instance deployed in a production environment.
You need to simulate a Redis instance failover in the most accurate disaster recovery situation, and ensure that the failover has no impact on production data. What should you do?
A. Create a Standard Tier Memorystore for Redis instance in a development environment. Initiate a manual failover by using the force-data-loss data protection mode.
B. Increase one replica to Redis instance in production environment. Initiate a manual failover by using the force-data-loss data protection mode.
C. Create a Standard Tier Memorystore for Redis instance in the development environment. Initiate a manual failover by using the limited-data-loss data protection mode.
D. Initiate a manual failover by using the limited-data-loss data protection mode to the Memorystore for Redis instance in the production environment.
Answer: (SHOW ANSWER)

NEW QUESTION: 131
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 hosted on a Compute Engine instance that makes a push subscription to the Pub/ Sub topic
B. Write an application that makes a queue in a NoSQL database
C. Use Cloud Composer to subscribe to a Pub/Sub tope and can the Python API.
D. Use a Pub/Sub push subscription to trigger a Cloud Function to pass the data to tie Python API.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 132
You need to store and analyze social media postings in Google BigQuery at a rate of 10,000 messages per minute in near real-time. Initially, design the application to use streaming inserts for individual postings.
Your application also performs data aggregations right after the streaming inserts. You discover that the queries after streaming inserts do not exhibit strong consistency, and reports from the queries might miss in-flight data. How can you adjust your application design?
A. Re-write the application to load accumulated data every 2 minutes.
B. Convert the streaming insert code to batch load for individual messages.
C. Load the original message to Google Cloud SQL, and export the table every hour to BigQuery via streaming inserts.
D. Estimate the average latency for data availability after streaming inserts, and always run queries after waiting twice as long.
Answer: (SHOW ANSWER)

NEW QUESTION: 133
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. Convert your PySpark commands into SparkSQL queries to transform the data; and then run your pipelineb on Dataproc to write the data into BigQuery
B. 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.
C. 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
D. Use Apache Beam Python SDK to build the transformation pipelines, and write the data into BigQuery
Answer: C (LEAVE A REPLY)
Dataproc is diffrent than Dataproc Serveless. This question is talking about dataproc.
By the way dp serverless support both pyspark and sparkSql no need of conversion.

NEW QUESTION: 134
Case Study: 2 - MJTelco
Company Overview
MJTelco is a startup that plans to build networks in rapidly growing, underserved markets around the world. The company has patents for innovative optical communications hardware. Based on these patents, they can create many reliable, high-speed backbone links with inexpensive hardware.
Company Background
Founded by experienced telecom executives, MJTelco uses technologies originally developed to overcome communications challenges in space. Fundamental to their operation, they need to create a distributed data infrastructure that drives real-time analysis and incorporates machine learning to continuously optimize their topologies. Because their hardware is inexpensive, they plan to overdeploy the network allowing them to account for the impact of dynamic regional politics on location availability and cost. Their management and operations teams are situated all around the globe creating many-to- many relationship between data consumers and provides in their system. After careful consideration, they decided public cloud is the perfect environment to support their needs.
Solution Concept
MJTelco is running a successful proof-of-concept (PoC) project in its labs. They have two primary needs:
Scale and harden their PoC to support significantly more data flows generated when they ramp to more than 50,000 installations.
Refine their machine-learning cycles to verify and improve the dynamic models they use to control topology definition.
MJTelco will also use three separate operating environments ?development/test, staging, and production ?
to meet the needs of running experiments, deploying new features, and serving production customers.
Business Requirements
Scale up their production environment with minimal cost, instantiating resources when and where needed in an unpredictable, distributed telecom user community. Ensure security of their proprietary data to protect their leading-edge machine learning and analysis.

Provide reliable and timely access to data for analysis from distributed research workers Maintain isolated environments that support rapid iteration of their machine-learning models without affecting their customers.
Technical Requirements
Ensure secure and efficient transport and storage of telemetry data Rapidly scale instances to support between 10,000 and 100,000 data providers with multiple flows each.
Allow analysis and presentation against data tables tracking up to 2 years of data storing approximately
100m records/day
Support rapid iteration of monitoring infrastructure focused on awareness of data pipeline problems both in telemetry flows and in production learning cycles.
CEO Statement
Our business model relies on our patents, analytics and dynamic machine learning. Our inexpensive hardware is organized to be highly reliable, which gives us cost advantages. We need to quickly stabilize our large distributed data pipelines to meet our reliability and capacity commitments.
CTO Statement
Our public cloud services must operate as advertised. We need resources that scale and keep our data secure. We also need environments in which our data scientists can carefully study and quickly adapt our models. Because we rely on automation to process our data, we also need our development and test environments to work as we iterate.
CFO Statement
The project is too large for us to maintain the hardware and software required for the data and analysis.
Also, we cannot afford to staff an operations team to monitor so many data feeds, so we will rely on automation and infrastructure. Google Cloud's machine learning will allow our quantitative researchers to work on our high-value problems instead of problems with our data pipelines.
MJTelco needs you to create a schema in Google Bigtable that will allow for the historical analysis of the last 2 years of records. Each record that comes in is sent every 15 minutes, and contains a unique identifier of the device and a data record. The most common query is for all the data for a given device for a given day. Which schema should you use?
A. Rowkey: date#data_pointColumn data: device_id
B. Rowkey: device_idColumn data: date, data_point
C. Rowkey: dateColumn data: device_id, data_point
D. Rowkey: date#device_idColumn data: data_point
E. Rowkey: data_pointColumn data: device_id, date
Answer: E (LEAVE A REPLY)

NEW QUESTION: 135
You need to create a near real-time inventory dashboard that reads the main inventory tables in your BigQuery data warehouse. Historical inventory data is stored as inventory balances by item and location. You have several thousand updates to inventory every hour. You want to maximize performance of the dashboard and ensure that the data is accurate. What should you do?
A. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.
C. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.

D. Use the BigQuery bulk loader to batch load inventory changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 136
Your company maintains a hybrid deployment with GCP, where analytics are performed on your anonymized customer data. The data are imported to Cloud Storage from your data center through parallel uploads to a data transfer server running on GCP. Management informs you that the daily transfers take too long and have asked you to fix the problem. You want to maximize transfer speeds. Which action should you take?
A. Increase the size of the Google Persistent Disk on your server.
B. Increase the CPU size on your server.
C. Increase your network bandwidth from your datacenter to GCP.
D. Increase your network bandwidth from Compute Engine to Cloud Storage.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 137
You are responsible for writing your company's ETL pipelines to run on an Apache Hadoop cluster. The pipeline will require some checkpointing and splitting pipelines. Which method should you use to write the pipelines?
A. HiveQL using Hive
B. PigLatin using Pig
C. Python using MapReduce
D. Java using MapReduce
Answer: C (LEAVE A REPLY)

NEW QUESTION: 138
Your neural network model is taking days to train. You want to increase the training speed. What can you do?
A. Subsample your test dataset.
B. Subsample your training dataset.
C. Increase the number of input features to your model.
D. Increase the number of layers in your neural network.
Answer: D (LEAVE A REPLY)
Explanation/Reference: https://towardsdatascience.com/how-to-increase-the-accuracy-of-a-neural-network-9f5d1c6f407d

NEW QUESTION: 139
You want to encrypt the customer data stored in BigQuery. You need to implement for-user crypto-deletion on data stored in your tables. You want to adopt native features in Google Cloud to avoid custom solutions. What should you do?
A. Create a customer-managed encryption key (CMEK) in Cloud KMS. Associate the key to the table while creating the table.
B. Create a customer-managed encryption key (CMEK) in Cloud KMS. Use the key to encrypt data before storing in BigQuery.
C. Implement Authenticated Encryption with Associated Data (AEAD) BigQuery functions while storing your data in BigQuery.
D. Encrypt your data during ingestion by using a cryptographic library supported by your ETL pipeline.
Answer: A (LEAVE A REPLY)
To implement for-user crypto-deletion and ensure that customer data stored in BigQuery is encrypted, using native Google Cloud features, the best approach is to use Customer-Managed Encryption Keys (CMEK) with Cloud Key Management Service (KMS).
Here's why:
Customer-Managed Encryption Keys (CMEK):
CMEK allows you to manage your own encryption keys using Cloud KMS. These keys provide additional control over data access and encryption management.
Associating a CMEK with a BigQuery table ensures that data is encrypted with a key you manage.
For-User Crypto-Deletion:
For-user crypto-deletion can be achieved by disabling or destroying the CMEK. Once the key is disabled or destroyed, the data encrypted with that key cannot be decrypted, effectively rendering it unreadable.
Native Integration:
Using CMEK with BigQuery is a native feature, avoiding the need for custom encryption solutions. This simplifies the management and implementation of encryption and decryption processes.
Steps to Implement:
Create a CMEK in Cloud KMS:
Set up a new customer-managed encryption key in Cloud KMS.
Associate the CMEK with BigQuery Tables:
When creating a new table in BigQuery, specify the CMEK to be used for encryption.
This can be done through the BigQuery console, CLI, or API.
Reference:
BigQuery and CMEK
Cloud KMS Documentation
Encrypting Data in BigQuery

NEW QUESTION: 140
Government regulations in the banking industry mandate the protection of client's personally identifiable information (PII). Your company requires PII to be access controlled encrypted and compliant with major data protection standards In addition to using Cloud Data Loss Prevention (Cloud DIP) you want to follow Google-recommended practices and use service accounts to control access to PII. What should you do?
A. Use Cloud Storage to comply with major data protection standards. Use multiple service accounts attached to IAM groups to grant the appropriate access to each group

B. Assign the required identity and Access Management (IAM) roles to every employee, and create a single service account to access protect resources
C. Use Cloud Storage to comply with major data protection standards. Use one service account shared by all users
D. Use one service account to access a Cloud SQL database and use separate service accounts for each human user
Answer: (SHOW ANSWER)

NEW QUESTION: 141
Government regulations in your industry mandate that you have to maintain an auditable record of access to certain types of datA. Assuming that all expiring logs will be archived correctly, where should you store data that is subject to that mandate?
A. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.
B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.
C. In a bucket on Cloud Storage that is accessible only by an AppEngine service that collects user information and logs the access before providing a link to the bucket.
D. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 142
You set up a streaming data insert into a Redis cluster via a Kafka cluster. Both clusters are running on Compute Engine instances. You need to encrypt data at rest with encryption keys that you can create, rotate, and destroy as needed. What should you do?
A. Create a dedicated service account, and use encryption at rest to reference your data stored in your Compute Engine cluster instances as part of your API service calls.
B. Create encryption keys in Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
C. Create encryption keys locally. Upload your encryption keys to Cloud Key Management Service. Use those keys to encrypt your data in all of the Compute Engine cluster instances.
D. Create encryption keys in Cloud Key Management Service. Reference those keys in your API service calls when accessing the data in your Compute Engine cluster instances.
Answer: B (LEAVE A REPLY)
https://cloud.google.com/compute/docs/disks/customer-managed-encryption

NEW QUESTION: 143
Your analytics team wants to build a simple statistical model to determine which customers are most likely to work with your company again, based on a few different metrics. They want to run the model on Apache Spark, using data housed in Google Cloud Storage, and you have recommended using Google Cloud Dataproc to execute this job. Testing has shown that this workload can run in approximately 30 minutes on a 15-node cluster, outputting the results into Google BigQuery. The plan is to run this workload weekly.
How should you optimize the cluster for cost?
A. Use SSDs on the worker nodes so that the job can run faster
B. Use a higher-memory node so that the job runs faster
C. Migrate the workload to Google Cloud Dataflow

D. Use pre-emptible virtual machines (VMs) for the cluster
Answer: (SHOW ANSWER)

NEW QUESTION: 144
You have Google Cloud Dataflow streaming pipeline running with a Google Cloud Pub/Sub subscription as the source. You need to make an update to the code that will make the new Cloud Dataflow pipeline incompatible with the current version. You do not want to lose any data when making this update. What should you do?
A. Update the current pipeline and use the drain flag.
B. Update the current pipeline and provide the transform mapping JSON object.
C. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
D. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.
Answer: B (LEAVE A REPLY)
If any transform names in your pipeline have changed, you must supply a transform mapping and pass it using the --transformNameMapping option.
https://cloud.google.com/dataflow/docs/guides/updating-a-pipeline#preventing_compatibility_breaks

NEW QUESTION: 145
Which Google Cloud Platform service is an alternative to Hadoop with Hive?
A. BigQuery
B. Cloud Dataflow
C. Cloud Datastore
D. Cloud Bigtable
Answer: A (LEAVE A REPLY)

NEW QUESTION: 146
You have a job that you want to cancel. It is a streaming pipeline, and you want to ensure that any data that is in-flight is processed and written to the output. Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?
A. Cancel
B. Drain
C. Stop
D. Finish
Answer: B (LEAVE A REPLY)
Using the Drain option to stop your job tells the Dataflow service to finish your job in its current state. Your job will immediately stop ingesting new data from input sources, but the Dataflow
service will preserve any existing resources (such as worker instances) to finish processing and writing any buffered data in your pipeline.

NEW QUESTION: 147
When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?
A. 500 TB
B. 1 GB

C. 1 TB
D. 500 GB
Answer: C (LEAVE A REPLY)
Cloud Bigtable is not a relational database. It does not support SQL queries, joins, or multi-row transactions. It is not a good solution for less than 1 TB of data.

NEW QUESTION: 148
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: 149
You decided to use Cloud Datastore to ingest vehicle telemetry data in real time. You want to build a storage system that will account for the long-term data growth, while keeping the costs low. You also want to create snapshots of the data periodically, so that you can make a point-in- time (PIT) recovery, or clone a copy of the data for Cloud Datastore in a different environment.
You want to archive these snapshots for a long time. Which two methods can accomplish this?
(Choose two.)
A. Use managed export, and store the data in a Cloud Storage bucket using Nearline or Coldline class.
B. Use managed export, 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: A,B (LEAVE A REPLY)
https://cloud.google.com/datastore/docs/export-import-entities

NEW QUESTION: 150
You are developing an Apache Beam pipeline to extract data from a Cloud SQL instance by using JdbcIO.
You have two projects running in Google Cloud. The pipeline will be deployed and executed on Dataflow in Project A. The Cloud SQL instance is running jn Project B and does not have a public IP address. After deploying the pipeline, you noticed that the pipeline

failed to extract data from the Cloud SQL instance due to connection failure. You verified that VPC Service Controls and shared VPC are not in use in these projects.
You want to resolve this error while ensuring that the data does not go through the public internet. What should you do?
A. Set up VPC Network Peering between Project A and Project B. Add a firewall rule to allow the peered subnet range to access all instances on the network.
B. Turn off the external IP addresses on the Dataflow worker. Enable Cloud NAT in Project A.
C. Set up VPC Network Peering between Project A and Project B. Create a Compute Engine instance without external IP address in Project B on the peered subnet to serve as a proxy server to the Cloud SQL database.
D. Add the external IP addresses of the Dataflow worker as authorized networks in the Cloud SOL instance.
Answer: C (LEAVE A REPLY)
* Option A is incorrect because VPC Network Peering alone does not enable connectivity to Cloud SQL instances with private IP addresses. You also need to configure private services access and allocate an IP address range for the service producer network1.
* Option B is incorrect because Cloud NAT does not support Cloud SQL instances with private IP addresses. Cloud NAT only provides outbound connectivity for resources that do not have public IP addresses, such as VMs, GKE clusters, and serverless instances2.
* Option C is correct because it allows you to use a Compute Engine instance as a proxy server to connect to the Cloud SQL database over the peered network. The proxy server does not need an external IP address because it can communicate with the Dataflow workers and the Cloud SQL instance using internal IP addresses. You need to install the Cloud SQL Auth proxy on the proxy server and configure it to use a service account that has the Cloud SQL Client role.
* Option D is incorrect because it requires you to assign public IP addresses to the Dataflow workers, which exposes the data to the public internet. This violates the requirement of ensuring that the data does not go through the public internet. Moreover, adding authorized networks does not work for Cloud SQL instances with private IP addresses.

NEW QUESTION: 151
You're using Bigtable for a real-time application, and you have a heavy load that is a mix of read and writes.
You've recently identified an additional use case and need to perform hourly an analytical job to calculate certain statistics across the whole database. You need to ensure both the reliability of your production application as well as the analytical workload.
What should you do?
A. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
B. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.
C. Add a second cluster to an existing instance with a multi-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
D. Add a second cluster to an existing instance with a single-cluster routing, use live-traffic app profile for your regular workload and batch-analytics profile for the analytics workload.
Answer: C (LEAVE A REPLY)

NEW QUESTION: 152
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.
Reference: https://www.tensorflow.org/tutorials/linear#feature_columns_and_transformations

NEW QUESTION: 153
You are designing the database schema for a machine learning-based food ordering service that will predict what users want to eat. Here is some of the information you need to store:
The user profile: What the user likes and doesn't like to eat
- The user account information: Name, address, preferred meal times
- The order information: When orders are made, from where, to whom
- The database will be used to store all the transactional data of the product. You want to optimize the data schema. Which Google Cloud Platform product should you use?

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

NEW QUESTION: 154
Your company receives both batch- and stream-based event dat
a. You want to process the data using Google Cloud Dataflow over a predictable time period. However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?
A. Set a single global window to capture all the data.
B. Use watermarks and timestamps to capture the lagged data.

C. Set sliding windows to capture all the lagged data.
D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.
Answer: (SHOW ANSWER)

NEW QUESTION: 155
Your company's customer_order table in BigQuery stores the order history for 10 million customers, with a table size of 10 PB. You need to create a dashboard for the support team to view the order history. The dashboard has two filters, country_name and username. Both are string data types in the BigQuery table. When a filter is applied, the dashboard fetches the order history from the table and displays the query results. However, the dashboard is slow to show the results when applying the filters to the following query:

```
SELECT date, order, status FROM customer_order
WHERE country = '<country_name>' AND username = '<username>'
```

How should you redesign the BigQuery table to support faster access?
A. Cluster the table by country field, and partition by username field.
B. Cluster the table by country and username fields.
C. Partition the table by country and username fields.
D. Partition the table by _PARTITIONTIME.
Answer: B (LEAVE A REPLY)

NEW QUESTION: 156
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 Bigtable and Cloud SQL
C. BigQuery and Cloud Storage
D. Cloud Datastore and Cloud Bigtable
Answer: (SHOW ANSWER)

NEW QUESTION: 157
You are administering a BigQuery on-demand environment. Your business intelligence tool is submitting hundreds of queries each day that aggregate a large (50 TB) sales history fact table at the day and month levels. These queries have a slow response time and are exceeding cost expectations. You need to decrease response time, lower query costs, and minimize maintenance. What should you do?
A. Build materialized views on top of the sales table to aggregate data at the day and month level.
B. Build authorized views on top of the sales table to aggregate data at the day and month level.

C. Enable BI Engine and add your sales table as a preferred table.
D. Create a scheduled query to build sales day and sales month aggregate tables on an hourly basis.
Answer: (SHOW ANSWER)
To improve response times and reduce costs for frequent queries aggregating a large sales history fact table, materialized views are a highly effective solution. Here's why option A is the best choice:
Materialized Views:
Materialized views store the results of a query physically and update them periodically, offering faster query responses for frequently accessed data.
They are designed to improve performance for repetitive and expensive aggregation queries by precomputing the results.
Efficiency and Cost Reduction:
By building materialized views at the day and month level, you significantly reduce the computation required for each query, leading to faster response times and lower query costs.
Materialized views also reduce the need for on-demand query execution, which can be costly when dealing with large datasets.
Minimized Maintenance:
Materialized views in BigQuery are managed automatically, with updates handled by the system, reducing the maintenance burden on your team.
Steps to Implement:
Identify Aggregation Queries:
Analyze the existing queries to identify common aggregation patterns at the day and month levels.
Create Materialized Views:
Create materialized views in BigQuery for the identified aggregation patterns. For example CREATE MATERIALIZED VIEW project.dataset.sales_daily_summary AS SELECT DATE(transaction_time) AS day, SUM(amount) AS total_sales FROM project.dataset.sales GROUP BY day; CREATE MATERIALIZED VIEW project.dataset.sales_monthly_summary AS SELECT EXTRACT(YEAR FROM transaction_time) AS year, EXTRACT(MONTH FROM transaction_time) AS month, SUM(amount) AS total_sales FROM project.dataset.sales GROUP BY year, month; Query Using Materialized Views:
Update existing queries to use the materialized views instead of directly querying the base table.
Reference:
BigQuery Materialized Views
Optimizing Query Performance

NEW QUESTION: 158
You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the moving average over 1 hour drops below 4000 messages per second. What should you do?
A. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a fixed time window of 1 hour. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
B. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.
C. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to BigQuery. Use Cloud Scheduler to run a script every five minutes that counts the number of rows created in BigQuery in the last hour. If that number falls below

4000, send an alert.
D. Use Kafka Connect to link your Kafka message queue to Cloud Pub/Sub. Use a Cloud Dataflow template to write your messages from Cloud Pub/Sub to Cloud Bigtable. Use Cloud Scheduler to run a script every hour that counts the number of rows created in Cloud Bigtable in the last hour. If that number falls below
4000, send an alert.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 159
You need to look at BigQuery data from a specific table multiple times a day. The underlying table you are querying is several petabytes in size, but you want to filter your data and provide simple aggregations to downstream users. You want to run queries faster and get up-to-date insights quicker. What should you do?
A. Create a materialized view based off of the query being run.
B. Limit the query columns being pulled in the final result.
C. Use a cached query to accelerate time to results.
D. Run a scheduled query to pull the necessary data at specific intervals daily.
Answer: (SHOW ANSWER)

NEW QUESTION: 160
You work for a large ecommerce company. You store your customer's order data in Bigtable. You have a garbage collection policy set to delete the data after 30 days and the number of versions is set to 1. When the data analysts run a query to report total customer spending, the analysts sometimes see customer data that is older than 30 days. You need to ensure that the analysts do not see customer data older than 30 days while minimizing cost and overhead. What should you do?
A. Schedule a job daily to scan the data in the table and delete data older than 30 days.
B. Set the expiring values of the column families to 30 days and set the number of versions to 2.
C. Set the expiring values of the column families to 29 days and keep the number of versions to 1.
D. Use a timestamp range filter in the query to fetch the customer's data for a specific range.
Answer: D (LEAVE A REPLY)

NEW QUESTION: 161
You have an upstream process that writes data to Cloud Storage. This data is then read by an Apache Spark job that runs on Dataproc. These jobs are run in the us-central1 region, but the data could be stored anywhere in the United States. You need to have a recovery process in place in case of a catastrophic single region failure. You need an approach with a maximum of 15 minutes of data loss (RPO=15 mins). You want to ensure that there is minimal latency when reading the data. What should you do?
A. 1. Create a dual-region Cloud Storage bucket in the us-central1 and us-south1 regions.
2. Enable turbo replication.
3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in the us-south1 region.
4. In case of a regional failure, redeploy your Dataproc duster to the us-south1 region and continue reading from the same bucket.
B. 1. Create a dual-region Cloud Storage bucket in the us-central1 and us-south1 regions.
2. Enable turbo replication.
3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in the same region.
4. In case of a regional failure, redeploy the Dataproc clusters to the us-south1 region and read from the same bucket.

C. 1. Create a Cloud Storage bucket in the US multi-region.
2. Run the Dataproc cluster in a zone in the ua-central1 region, reading data from the US multi-region bucket.
3. In case of a regional failure, redeploy the Dataproc cluster to the us-central2 region and continue reading from the same bucket.
D. 1. Create two regional Cloud Storage buckets, one in the us-central1 region and one in the us-south1 region.
2. Have the upstream process write data to the us-central1 bucket. Use the Storage Transfer Service to copy data hourly from the us-central1 bucket to the us-south1 bucket.
3. Run the Dataproc cluster in a zone in the us-central1 region, reading from the bucket in that region.
4. In case of regional failure, redeploy your Dataproc clusters to the us-south1 region and read from the bucket in that region instead.
Answer: B (LEAVE A REPLY)
To ensure data recovery with minimal data loss and low latency in case of a single region failure, the best approach is to use a dual-region bucket with turbo replication. Here's why option B is the best choice:
* Dual-Region Bucket:
* A dual-region bucket provides geo-redundancy by replicating data across two regions, ensuring high availability and resilience against regional failures.
* The chosen regions (us-central1 and us-south1) provide geographic diversity within the United States.
* Turbo Replication:
* Turbo replication ensures that data is replicated between the two regions within 15 minutes, meeting the Recovery Point Objective (RPO) of 15 minutes.
* This minimizes data loss in case of a regional failure.
* Running Dataproc Cluster:
* Running the Dataproc cluster in the same region as the primary data storage (us-central1) ensures minimal latency for normal operations.
* In case of a regional failure, redeploying the Dataproc cluster to the secondary region (us-south1) ensures continuity with minimal data loss.
Steps to Implement:
* Create a Dual-Region Bucket:
* Set up a dual-region bucket in the Google Cloud Console, selecting us-central1 and us-south1 regions.
* Enable turbo replication to ensure rapid data replication between the regions.
* Deploy Dataproc Cluster:
* Deploy the Dataproc cluster in the us-central1 region to read data from the bucket located in the same region for optimal performance.
* Set Up Failover Plan:
* Plan for redeployment of the Dataproc cluster to the us-south1 region in case of a failure in the us- central1 region.
* Ensure that the failover process is well-documented and tested to minimize downtime and data loss.
Reference Links:
* Google Cloud Storage Dual-Region
* Turbo Replication in Google Cloud Storage
* Dataproc Documentation

NEW QUESTION: 162

Your company is planning to migrate a large on-premises data warehouse to BigQuery. The data is currently stored in a proprietary, vendor-specific format. You need to perform a batch migration of this data to BigQuery. What should you do?

A. Use the bq command-line tool to load the data directly from the on-premises data warehouse.

B. Use the BigQuery Data Transfer Service.

C. Export the data to CSV files, upload the files to Cloud Storage, then load the files into BigQuery.

D. Use Datastream to replicate the data in real time.

Answer: (SHOW ANSWER)

Comprehensive and Detailed Explanation:

The challenge here is dealing with a "proprietary, vendor-specific format" for a one-time batch migration.

Option C is the correct answer because it represents the most universal and reliable pattern for migration from any source system. By first exporting the data into a standard, interoperable format like CSV (or preferably, a self-describing format like Avro or Parquet), you decouple the process from the proprietary source. These standard files can then be easily uploaded to Cloud Storage (the recommended staging area for BigQuery loads) and loaded into BigQuery in a highly performant and parallelized manner.

Option A is incorrect because the bq command-line tool cannot connect directly to an on-premises data warehouse to pull data. It loads data from files or streams.

Option B is incorrect because the BigQuery Data Transfer Service (DTS) has connectors for specific, common data sources (like Teradata, Redshift, S3). It is unlikely to have a connector for a generic
"proprietary, vendor-specific format."

Option D is incorrect because Datastream is a Change Data Capture (CDC) service designed for real-time replication of databases, not for a large-scale, one-time batch migration of a data warehouse.

Reference (Google Cloud Documentation Concepts):Google Cloud's "Data warehouse migration to BigQuery" guide outlines several migration strategies. For batch data transfer, the recommended path is Extract, Transfer, Load (ETL). This involves extracting data from the source into files (in formats like CSV, Avro, Parquet), transferring those files to Cloud Storage, and then loading them into BigQuery. This approach is recommended for its reliability and compatibility with any source system

NEW QUESTION: 163

Which of the following statements about the Wide & Deep Learning model are true? (Select 2 answers.)

A. The wide model is used for memorization, while the deep model is used for generalization.

B. A good use for the wide and deep model is a recommender system.

C. The wide model is used for generalization, while the deep model is used for memorization.

D. A good use for the wide and deep model is a small-scale linear regression problem.

Answer: A,B (LEAVE A REPLY)

Can we teach computers to learn like humans do, by combining the power of memorization and generalization? It's not an easy question to answer, but by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), one can combine the strengths of both to bring us one step closer. At Google, we call it Wide & Deep Learning. It's useful for generic large-scale regression and classification problems with sparse inputs (categorical features with a large number of possible feature values), such as recommender systems, search, and ranking problems.

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

NEW QUESTION: 164

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.

NEW QUESTION: 165
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)
Explanation
DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization.
Useful for small local execution and tests
Reference:
https://cloud.google.com/dataflow/java-sdk/JavaDoc/com/google/cloud/dataflow/sdk/runners/DirectPipelineRun

NEW QUESTION: 166
Your company is migrating their 30-node Apache Hadoop cluster to the cloud. They want to re-use
Hadoop jobs they have already created and minimize the management of the cluster as much as possible.
They also want to be able to persist data beyond the life of the cluster. What should you do?
A. Create a Google Cloud Dataflow job to process the data.
B. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.
C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.
E. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.

Answer: A (LEAVE A REPLY)

NEW QUESTION: 167
You are planning to use Google's Dataflow SDK to analyze customer data such as displayed below. Your project requirement is to extract only the customer name from the data source and then write to an output PCollection.
Tom,555 X street
Tim,553 Y street
Sam, 111 Z street
Which operation is best suited for the above data processing requirement?
A. ParDo
B. Sink API
C. Source API
D. Data extraction
Answer: A (LEAVE A REPLY)
In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.

NEW QUESTION: 168
You are designing the architecture to process your data from Cloud Storage to BigQuery by using Dataflow.
The network team provided you with the Shared VPC network and subnetwork to be used by your pipelines.
You need to enable the deployment of the pipeline on the Shared VPC network. What should you do?
A. Assign the compute. networkUser role to the Dataflow service agent.
B. Assign the compute.networkUser role to the service account that executes the Dataflow pipeline.
C. Assign the dataflow, admin role to the Dataflow service agent.
D. Assign the dataflow, admin role to the service account that executes the Dataflow pipeline.
Answer: (SHOW ANSWER)
To use a Shared VPC network for a Dataflow pipeline, you need to specify the subnetwork parameter with the full URL of the subnetwork, and grant the service account that executes the pipeline the compute.networkUser role in the host project. This role allows the service account to use the subnetworks in the Shared VPC network. The Dataflow service agent does not need this role, as it only creates and manages the resources for the pipeline, but does not execute it. The dataflow.admin role is not related to the network access, but to the permissions to create and delete Dataflow jobs and resources. References:
* Specify a network and subnetwork | Cloud Dataflow | Google Cloud
* How to config dataflow Pipeline to use a Shared VPC?

NEW QUESTION: 169
You are building a report-only data warehouse where the data is streamed into BigQuery via the streaming API. Following Google's best practices, you have both a staging and a production table for the data. How should you design your data loading to ensure that there is only one master dataset without affecting performance on either the ingestion or reporting pieces?
A. Have a staging table that is an append-only model, and then update the production table every three hours with the changes written to staging
B. Have a staging table that is an append-only model, and then update the production table every ninety minutes with the changes written to staging
C. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every three hours
D. Have a staging table that moves the staged data over to the production table and deletes the contents of the staging table every thirty minutes
Answer: C (LEAVE A REPLY)
Following common extract, transform, load (ETL) best practices, we used a staging table and a separate production table so that we could load data into the staging table without impacting users of the data. The design we created based on ETL best practices called for first deleting all the records from the staging table, loading the staging table, and then replacing the production table with the contents.
When using the streaming API, the BigQuery streaming buffer remains active for about 30 to 60 minutes or more after use, which means that you can't delete or change data during that time.
Since we used the streaming API, we scheduled the load every three hours to balance getting data into BigQuery quickly and being able to subsequently delete the data from the staging table during the load process.
Building a script with BigQuery on the back end.
Reference:
https://cloud.google.com/blog/products/data-analytics/moving-a-publishing-workflow-to-bigquery-for-new-data-insights

NEW QUESTION: 170
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 0 using a Cloud Dataprep job.
B. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to using a custom script.
C. Use Cloud Dataprep to find null values in sample source data. Convert all nulls to `none' using a Cloud Dataproc job.
D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to `none' using a Cloud Dataprep job.
Answer: (SHOW ANSWER)

NEW QUESTION: 171
You need to move 2 PB of historical data from an on-premises storage appliance to Cloud Storage within six months, and your outbound network capacity is constrained to 20 Mb/sec. How should you migrate this data to Cloud Storage?
A. Use gsutil cp -J to compress the content being uploaded to Cloud Storage
B. Use trickle or ionice along with gsutil cp to limit the amount of bandwidth gsutil utilizes to less than 20 Mb/sec so it does not interfere with the production traffic

C. Create a private URL for the historical data, and then use Storage Transfer Service to copy the data to Cloud Storage
D. Use Transfer Appliance to copy the data to Cloud Storage
Answer: D (LEAVE A REPLY)

NEW QUESTION: 172
You are creating a data model in BigQuery that will hold retail transaction data. Your two largest tables, sales_transation_header and sales_transation_line. have a tightly coupled immutable relationship. These tables are rarely modified after load and are frequently joined when queried. You need to model the sales_transation_header and sales_transation_line tables to improve the performance of data analytics queries.
What should you do?
A. Create a sal es_transaction table that Stores the sales_tran3action_header and sales_transaction_line data as a JSON data type.
B. Create a sales_transaction table that holds the sales_transaction_header information as rows and thesales_transaction_line rows as nested and repeated fields.
C. Create a sale_transaction table that holds the sales_transaction_header and sales_transaction_line information as rows, duplicating the sales_transaction_header data for each line.
D. Create separate sales_transation_header and sales_transation_line tables and. when querying, specify the sales transition line first in the WHERE clause.
Answer: B (LEAVE A REPLY)
BigQuery supports nested and repeated fields, which are complex data types that can represent hierarchical and one-to-many relationships within a single table. By using nested and repeated fields, you can denormalize your data model and reduce the number of joins required for your queries. This can improve the performance and efficiency of your data analytics queries, as joins can be expensive and require shuffling data across nodes. Nested and repeated fields also preserve the data integrity and avoid data duplication. In this scenario, the sales_transaction_header and sales_transaction_line tables have a tightly coupled immutable relationship, meaning that each header row corresponds to one or more line rows, and the data is rarely modified after load. Therefore, it makes sense to create a single sales_transaction table that holds the sales_transaction_header information as rows and the sales_transaction_line rows as nested and repeated fields. This way, you can query the sales transaction data without joining two tables, and use dot notation or array functions to access the nested and repeated fields. For example, the sales_transaction table could have the following schema:
Table
Field name
Type
Mode
id
INTEGER
NULLABLE
order_time
TIMESTAMP
NULLABLE
customer_id
INTEGER
NULLABLE

line_items
RECORD
REPEATED
line_items.sku
STRING
NULLABLE
line_items.quantity
INTEGER
NULLABLE
line_items.price
FLOAT
NULLABLE
To query the total amount of each order, you could use the following SQL statement:
SQL
SELECTid,SUM(line_items.quantity*line_items.price)AStotal_amount
FROMsales_transaction
GROUPBYid;
AI-generated code. Review and use carefully. More info on FAQ.
References:
Use nested and repeated fields
BigQuery explained: Working with joins, nested & repeated data
Arrays in BigQuery - How to improve query performance and optimise storage

NEW QUESTION: 173
You want to process payment transactions in a point-of-sale application that will run on Google Cloud Platform. Your user base could grow exponentially, but you do not want to manage infrastructure scaling.
Which Google database service should you use?
A. Cloud SQL
B. BigQuery
C. Cloud Bigtable
D. Cloud Datastore
Answer: (SHOW ANSWER)
https://cloud.google.com/datastore/docs/concepts/overview