EXAM PROFESSIONAL DATA ENGINEER TOPIC 1 QUESTION 206 DISCUSSION

Actual exam question from Google's Professional Data Engineer
Question #: 206
Topic #: 1
[All Professional Data Engineer Questions]

You need ads data to serve AI models and historical data for analytics. Longtail and outlier data points need to be identified. You want to cleanse the data in near-real time before running it through AI models. What should you do?

A. Use Cloud Storage as a data warehouse, shell scripts for processing, and BigQuery to create views for desired datasets.

B. Use Dataflow to identify longtail and outlier data points programmatically, with BigQuery as a sink.

C. Use BigQuery to ingest, prepare, and then analyze the data, and then run queries to create views.

D. Use Cloud Composer to identify longtail and outlier data points, and then output a usable dataset to BigQuery.

Show Suggested Answer

by e70ea9e at *Dec. 30, 2023, 9:24 a.m.*

Comments

Type your comment...

Submit

Y___ash 7 months, 3 weeks ago

Selected Answer: B

Dataflow for Real-Time Processing: Dataflow allows you to process data in near-real time, making it well-suited for identifying longtail and outlier data points as they occur. You can use Dataflow to implement custom data cleansing and outlier detection

longtail and outlier data points as they occur. You can use Dataflow to implement custom data cleansing and outlier detection algorithms that operate on streaming data.

BigQuery as a Sink: Using BigQuery as a sink allows you to store the cleaned and processed data efficiently for further analysis or use in AI models. Dataflow can write the cleaned data to BigQuery tables, enabling seamless integration with downstream processes.

upvoted 2 times

JyoGCP 8 months, 3 weeks ago

Selected Answer: B

B. Use Dataflow to identify longtail and outlier data points programmatically, with BigQuery as a sink.

upvoted 1 times

datapassionate 9 months, 3 weeks ago

Selected Answer: B

B. Use Dataflow to identify longtail and outlier data points programmatically, with BigQuery as a sink.

upvoted 1 times

Matt_108 9 months, 3 weeks ago

Selected Answer: B

B: Dataflow, solves exactly the use case described

upvoted 1 times

MaxNRG 10 months ago

Selected Answer: B

B is the best option for cleansing the ads data in near real-time before running it through AI models.
The key reasons are:
• Dataflow allows for stream processing of data in near real-time. This allows you to identify and cleanse longtail and outlier data points as the data is streamed in.
• Dataflow has built-in capabilities for detecting and handling outliers and anomalies in streaming data. This makes it well-suited for programmatically identifying longtail and outlier data points.
• Using BigQuery as the output sink allows the cleansed data to be immediatley available for analysis and serving to AI models. BigQuery can act as a serving layer for the models.
• Options A, C, and D either don't provide real-time processing (A and C) or don't easily integrate with BigQuery for analysis and serving (D).

upvoted 2 times

MaxNRG 10 months ago

So B is the best architecture here to meet the needs of near real-time cleansing, identification of longtail/outlier data points, and integration with BigQuery for serving AI models.

upvoted 1 times

raaad 10 months, 1 week ago

Selected Answer: B

- Dataflow is a fully managed service for stream and batch data processing and is well-suited for real-time data processing tasks like identifying longtail and outlier data points.
- Using BigQuery as a sink allows to efficiently store the cleansed and processed data for further analysis and serving it to AI models.

upvoted 1 times

e70ea9e 10 months, 1 week ago

Selected Answer: B

Real-time Data Processing: Dataflow excels at handling large-scale, streaming data with low latency, enabling near-real-time cleansing.
Scalability: Easily scales to handle growing data volumes and processing needs.
Programmatic Data Cleaning: Allows you to write custom logic in Apache Beam for identifying longtail and outlier data points accurately and efficiently.
Integration with BigQuery: Seamless integration with BigQuery enables you to store cleansed data for AI model training and historical analytics.
Cost-Effective: Dataflow's pay-as-you-go model optimizes costs for real-time data processing.

upvoted 1 times

Platform

- Home
- Examtopics PRO
- All Exams
- Training Courses