Google Professional Cloud Data Engineer Actual Exam

1. We highly recommend you should take Google Professional Data Engineer Actual Exam Version because it include actual exam questions and highlighted answers are collected and verified in our exam. It will help you pass exam in easier way.

A.

Answer(s): B is correct because of the requirement to support occasionally (schema) changing JSON files and aggregate ANSI SQL queries: you need to use BigQuery, and it is quickest to use ‘Automatically detect’ for schema changes.

---

2. You use a Hadoop cluster both for serving analytics and for processing and transforming data. The data is currently stored on HDFS in Parquet format. The data processing jobs run for 6 hours each night. Analytics users can access the system 24 hours a day. Phase 1 is to quickly migrate the entire Hadoop environment without a major re-architecture. Phase 2 will include migrating to BigQuery for analytics and to Cloud Dataflow for data processing. You want to make the future migration to BigQuery and Cloud Dataflow easier by following Google-recommended practices and managed services. What should you do?

A. A. Lift and shift Hadoop/HDFS to Cloud Dataproc.

B. B. Lift and shift Hadoop/HDFS to Compute Engine.

C. C. Create a single Cloud Dataproc cluster to support both analytics and data processing, and point it at a Cloud Storage bucket that contains the Parquet files that were previously stored on HDFS.

D. D. Create separate Cloud Dataproc clusters to support analytics and data processing, and point both at the same Cloud Storage bucket that contains the Parquet files that were previously stored on HDFS.

Answer(s): D Is correct because it leverages a managed service (Cloud Dataproc), the data is stored on GCS in Parquet format which can easily be loaded into BigQuery in the future and the Cloud Dataproc clusters are job specific.

---

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

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

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

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

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

Answer(s): D is correct because it will just pick out a single row for each set of duplicates.

---

4. https://cloud.google.com/storage/
https://cloud.google.com/storage/transfer/
4. You are designing a streaming pipeline for ingesting player interaction data for a mobile game. You want the pipeline to handle out-of-order data delayed up to 15 minutes on a per-player basis and exponential growth in global users. What should you do?

A. A. Design a Cloud Dataflow streaming pipeline with session windowing and a minimum gap duration of 15 minutes. Use “individual player” as the key. Use Cloud Pub/Sub as a message bus for ingestion.

B. B. Design a Cloud Dataflow streaming pipeline with session windowing and a minimum gap duration of 15 minutes. Use “individual player” as the key. Use Apache Kafka as a message bus for ingestion.

C. C. Design a Cloud Dataflow streaming pipeline with a single global window of 15 minutes. Use Cloud Pub/Sub as a message bus for ingestion.

D. D. Design a Cloud Dataflow streaming pipeline with a single global window of 15 minutes. Use Apache Kafka as a message bus for ingestion.

Answer(s): A is correct because the question requires delay be handled on a per-player basis and session windowing will do that. PubSub handles the need to scale exponentially with traffic coming from around the globe.

---

5. https://beam.apache.org/documentation/programming-guide/#windowing
https://cloud.google.com/pubsub/architecture
5. Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

A. A. The CSV data loaded in BigQuery is not flagged as CSV.

B. B. The CSV data had invalid rows that were skipped on import.

C. C. The CSV data loaded in BigQuery is not using BigQuery's default encoding.

D. D. The CSV data has not gone through an ETL phase before loading into BigQuery.

Answer(s): C is correct because this is the only situation that would cause successful import.

---

6. https://cloud.google.com/bigquery/docs/loading-data#loading_encoded_data
6. 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. A. Create a Google Cloud Dataflow job to process the data.

B. B. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.

C. C. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.

D. D. Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.

E. E. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

Answer(s): D is correct because it uses managed services, and also allows for the data to persist on GCS beyond the life of the cluster.

---

7. You work for an economic consulting firm that helps companies identify economic trends as they happen. As part of your analysis, you use Google BigQuery to correlate customer data with the average prices of the 100 most common goods sold, including bread, gasoline, milk, and

others. The average prices of these goods are updated every 30 minutes. You want to make sure this data stays up to date so you can combine it with other data in BigQuery as cheaply as possible. What should you do?
A. Load the data every 30 minutes into a new partitioned table in BigQuery.
B. Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery.
C. Store the data in Google Cloud Datastore. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Cloud Datastore.
D. Store the data in a file in a regional Google Cloud Storage bucket. Use Google Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

A. Reveal

Answer(s): B is correct because regional storage is cheaper than BigQuery storage.

---

8. You have 250,000 devices which produce a JSON device status event every 10 seconds. You want to capture this event data for outlier time series analysis. What should you do?

A. A. Ship the data into BigQuery. Develop a custom application that uses the BigQuery API to query the dataset and displays device outlier data based on your business requirements.

B. B. Ship the data into BigQuery. Use the BigQuery console to query the dataset and display device outlier data based on your business requirements.

C. C. Ship the data into Cloud Bigtable. Use the Cloud Bigtable cbt tool to display device outlier data based on your business requirements.

D. D. Ship the data into Cloud Bigtable. Install and use the HBase shell for Cloud Bigtable to query the table for device outlier data based on your business requirements.

Answer(s): C is correct because the data type, volume, and query pattern best fits BigTable capabilities and also Google best practices as linked below.

---

9. https://cloud.google.com/bigtable/docs/go/cbt-overview
https://cloud.google.com/bigtable/docs/installing-hbase-shell
9. You are selecting a messaging service for log messages that must include final result message ordering as part of building a data pipeline on Google Cloud. You want to stream input for 5 days and be able to query the current status. You will be storing the data in a searchable repository. How should you set up the input messages?

A. A. Use Cloud Pub/Sub for input. Attach a timestamp to every message in the publisher.

B. B. Use Cloud Pub/Sub for input. Attach a unique identifier to every message in the publisher.

C. C. Use Apache Kafka on Compute Engine for input. Attach a timestamp to every message in the publisher.

D. D. Use Apache Kafka on Compute Engine for input. Attach a unique identifier to every message in the publisher.

Answer(s): A is correct because of recommended Google practices; see the links below.

---

10. https://cloud.google.com/pubsub/docs/ordering
http://www.jesse-anderson.com/2016/07/apache-kafka-and-google-cloud-pubsub/
10. You want to publish system metrics to Google Cloud from a large number of on-prem hypervisors and VMs for analysis and creation of dashboards. You have an existing custom monitoring agent deployed to all the hypervisors and your on-prem metrics system is unable to handle the load. You want to design a system that can collect and store metrics at scale. You don't want to manage your own time series database. Metrics from all agents should be written to the same table but agents must not have permission to modify or read data written by other agents.What should you do?

A. A. Modify the monitoring agent to publish protobuf messages to Cloud PubSub. Use a Dataproc cluster or Dataflow job to consume messages from Pubsub and write to BigTable.

B. B. Modify the monitoring agent to write protobuf messages directly to BigTable.

C. C. Modify the monitoring agent to write protobuf messages to HBase deployed on GCE VM Instances

D. D. Modify the monitoring agent to write protobuf messages to Cloud Pubsub. Use a Dataproc cluster or Dataflow job to consume messages from Pubsub and write to Cassandra deployed on GCE VM Instances.

Answer(s): A Is correct because Bigtable can store and analyze time series data, and the solution is using managed services which is what the requirements are calling for.

---

11. You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud. You intend to use ANSI SQL to run queries for your analysts. How should you transform the input data?

A. A. Use BigQuery for storage. Use Cloud Dataflow to run the transformations.

B. B. Use BigQuery for storage. Use Cloud Dataproc to run the transformations.

C. C. Use Cloud Storage for storage. Use Cloud Dataflow to run the transformations.

D. D. Use Cloud Storage for storage. Use Cloud Dataproc to run the transformations.

Answer(s): B is correct because of the requirement to use custom Spark transforms; use Cloud Dataproc. ANSI SQL queries require the use of BigQuery.

---

12. https://stackoverflow.com/questions/46436794/what-is-the-difference-between-google-cloud-dataflow-and-google-cloud-dataproc
12. You are designing a relational data repository on Google Cloud to grow as needed. The data will be transactionally consistent and added from any location in the world. You want to monitor and adjust node count for input traffic, which can spike unpredictably. What should you do?

A. A. Use Cloud Spanner for storage. Monitor storage usage and increase node count if more than 70% utilized.

B. B. Use Cloud Spanner for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

C. C. Use Cloud Bigtable for storage. Monitor data stored and increase node count if more than 70% utilized.

D. D. Use Cloud Bigtable for storage. Monitor CPU utilization and increase node count if more than 70% utilized for your time span.

Answer(s): B is correct because of the requirement to globally scalable transactions—use Cloud Spanner. CPU utilization is the recommended metric for scaling, per Google best practices, linked below.

---

13. https://cloud.google.com/spanner/docs/monitoring
https://cloud.google.com/bigtable/docs/monitoring-instance
13. You have a Spark application that writes data to Cloud Storage in Parquet format. You scheduled the application to run daily using DataProcSparkOperator and Apache Airflow DAG by Cloud Composer. You want to add tasks to the DAG to make the data available to BigQuery

users. You want to maximize query speed and configure partitioning and clustering on the table. What should you do?

A. A. Use “BashOperator” to call “bq insert”.

B. B. Use “BashOperator” to call “bq cp” with the “–append” flag.

C. C. Use “GoogleCloudStorageToBigQueryOperator” with “schema_object” pointing to a schema JSON in Cloud Storage and “source_format” set to “PARQUET”.

D. D. Use “BigQueryCreateExternalTableOperator” with “schema_object” pointing to a schema JSON in Cloud Storage and “source_format” set to “PARQUET”.

Answer(s): C is correct because it loads the data and sets partitioning and clustering.

---

14. https://cloud.google.com/bigquery/docs/loading-data
https://airflow.incubator.apache.org/integration.html#bigquerycreateemptytableoperator
https://cloud.google.com/bigquery/docs/reference/bq-cli-reference
https://cloud.google.com/bigquery/docs/bq-command-line-tool
https://airflow.incubator.apache.org/integration.html#googlecloudstoragetobigqueryoperator
https://cloud.google.com/bigquery/external-data-sources
14. You have a website that tracks page visits for each user and then creates a Cloud Pub/Sub message with the session ID and URL of the page. You want to create a Cloud Dataflow pipeline that sums the total number of pages visited by each user and writes the result to BigQuery. User sessions timeout after 30 minutes. Which type of Cloud Dataflow window should you choose?

A. A. A single global window

B. B. Fixed-time windows with a duration of 30 minutes

C. C. Session-based windows with a gap duration of 30 minutes

D. D. Sliding-time windows with a duration of 30 minutes and a new window every 5 minute

Answer(s): C is correct because it continues to sum user page visits during their browsing session and completes at the same time as the session timeout.

---

15. https://cloud.google.com/dataflow/docs/resources
15. You are designing a basket abandonment system for an ecommerce company. The system

will send a message to a user based on these rules: a). No interaction by the user on the site for 1 hour b). Has added more than $30 worth of products to the basket c). 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. A. Use a fixed-time window with a duration of 60 minutes.

B. B. Use a sliding time window with a duration of 60 minutes.

C. C. Use a session window with a gap time duration of 60 minutes.

D. D. Use a global window with a time based trigger with a delay of 60 minutes.

Answer(s): C is correct because it will send a message per user after that user is inactive for 60 minutes.

---

16. https://beam.apache.org/documentation/programming-guide/#windowing
16. You need to stream time-series data in Avro format, and then write this to both BigQuery and Cloud Bigtable simultaneously using Cloud Dataflow. You want to achieve minimal end-to-end latency. Your business requirements state this needs to be completed as quickly as possible. What should you do?

A. A. Create a pipeline and use ParDo transform.

B. B. Create a pipeline that groups the data into a PCollection and uses the Combine transform.

C. C. Create a pipeline that groups data using a PCollection and then uses Cloud Bigtable and BigQueryIO transforms.

D. D. Create a pipeline that groups data using a PCollection, and then use Avro I/O transform to write to Cloud Storage. After the data is written, load the data from Cloud Storage into BigQuery and Cloud Bigtable.

Answer(s): C is correct because this is the right set of transformations that accepts and writes to the required data stores.

---

17. https://cloud.google.com/blog/products/gcp/guide-to-common-cloud-dataflow-use-case-patterns-part-1
17. Your company’s on-premises Apache Hadoop servers are approaching end-of-life, and IT has

decided to migrate the cluster to Google Cloud Dataproc. A like-for-like migration of the cluster would require 50 TB of Google Persistent Disk per node. The CIO is concerned about the cost of using that much block storage. You want to minimize the storage cost of the migration. What should you do?

A. A. Put the data into Google Cloud Storage.

B. B. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.

C. C. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.

D. D. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

Answer(s): A is correct because Google recommends using Google Cloud Storage instead of HDFS as it is much more cost effective especially when jobs aren't running.

---

18. https://cloud.google.com/dataproc/docs/concepts/connectors/cloud-storage
18. You are designing storage for two relational tables that are part of a 10-TB database on Google Cloud. You want to support transactions that scale horizontally. You also want to optimize data for range queries on non-key columns. What should you do?

A. A. Use Cloud SQL for storage. Add secondary indexes to support query patterns.

B. B. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.

C. C. Use Cloud Spanner for storage. Add secondary indexes to support query patterns.

D. D. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Answer(s): C is correct because Cloud Spanner scales horizontally, and you can create secondary indexes for the range queries that are required.

---

19. Your company is streaming real-time sensor data from their factory floor into Bigtable and they have noticed extremely poor performance. How should the row key be redesigned to improve Bigtable performance on queries that populate real-time dashboards?

A. A. Use a row key of the form .

B. B. Use a row key of the form .

C. C. Use a row key of the form #.

D. D. Use a row key of the form #.

Answer(s): D is correct because it will allow for retrieval of data based on both sensor id and timestamp but without causing hotspotting.

---

20. https://cloud.google.com/bigtable/docs/schema-design
https://cloud.google.com/bigtable/docs/schema-design-time-series#ensure_that_your_row_key_avoids_hotspotting
20. You are developing an application on Google Cloud that will automatically generate subject labels for users' blog posts. You are under competitive pressure to add this feature quickly, and you have no additional developer resources. No one on your team has experience with machine learning. What should you do?

A. A. Call the Cloud Natural Language API from your application. Process the generated Entity Analysis as labels.

B. B. Call the Cloud Natural Language API from your application. Process the generated Sentiment Analysis as labels.

C. C. Build and train a text classification model using TensorFlow. Deploy the model using Cloud Machine Learning Engine. Call the model from your application and process the results as labels.

D. D. Build and train a text classification model using TensorFlow. Deploy the model using a Kubernetes Engine cluster. Call the model from your application and process the results as labels.

Answer(s): A is correct because it provides a managed service and a fully trained model, and the user is pulling the entities, which is the right label.
