# Question:

Passnexam.com

#### Question 1:

Your company built a TensorFlow neutral-network model with a large number of neurons and layers. The model fits well for the training data. However, when tested against new data, it performs poorly. What method can you employ to address this?

(Designing data processing systems)

1. Threading
2. Serialization
3. **Dropout Methods**
4. Dimensionality Reduction

Answer is Dropout Methods

There are various ways to prevent overfitting when dealing with DNNs. In this post, we’ll review these techniques and then apply them specifically to TensorFlow models:

- Early Stopping

- L1 and L2 Regularization

- Dropout

- Max-Norm Regularization

- Data Augmentation

D is not correct here because it is used for normal ml models whereas dropout methods is used for neural networks.

#### Question 2:

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?

(Designing data processing systems)

1. Use federated data sources, and check data in the SQL query.
2. Enable BigQuery monitoring in Google Stackdriver and create an alert.
3. Import the data into BigQuery using the gcloud CLI and set max\_bad\_records to 0.
4. **Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.**

**Answer is Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis.**

A. Use federated data sources, and check data in the SQL query. - WRONG (Because we are changing source itself, i.e. SQL, MySQL, PstgresSQL) instead of correcting the problem

B. Enable BigQuery monitoring in Google Stackdriver and create an alert. (WRONG - Because setting and creating an alert will not solve the corrupted data problem)

C. Import the data into BigQuery using the gcloud CLI and set max\_bad\_records to 0. (wrong - here we are saying set max\_bad\_records = 0 (i.e let's load all bad records into bi-query)

D. Run a Google Cloud Dataflow batch pipeline to import the data into BigQuery, and push errors to another dead-letter table for analysis. (CORRECT - Dataflow is used for this pupose only i.e transform the data and dead letter queue pupose is to write any invalid records - so that it can be analyzed later (rather than ignoring))

#### Question 3:

Your weather app queries a database every 15 minutes to get the current temperature. The frontend is powered by Google App Engine and server millions of users. How should you design the frontend to respond to a database failure?

(Designing data processing systems)

1. Issue a command to restart the database servers.
2. **Retry the query with exponential backoff, up to a cap of 15 minutes.**
3. Retry the query every second until it comes back online to minimize staleness of data.
4. Reduce the query frequency to once every hour until the database comes back online.

App engine create applications that use Cloud SQL database connections effectively. Below is what is written in google cloud documnetation.

If your application attempts to connect to the database and does not succeed, the database could be temporarily unavailable. In this case, sending too many simultaneous connection requests might waste additional database resources and increase the time needed to recover. Using exponential backoff prevents your application from sending an unresponsive number of connection requests when it can't connect to the database.

This retry only makes sense when first connecting, or when first grabbing a connection from the pool. If errors happen in the middle of a transaction, the application must do the retrying, and it must retry from the beginning of a transaction. So even if your pool is configured properly, the application might still see errors if connections are lost.

#### Question 4:

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

(Designing data processing systems)

1. Include ORDER BY DESK on timestamp column and LIMIT to 1.
2. Use GROUP BY on the unique ID column and timestamp column and SUM on the values.
3. Use the LAG window function with PARTITION by unique ID along with WHERE LAG IS NOT NULL.
4. **Use the ROW\_NUMBER window function with PARTITION by unique ID along with WHERE row equals 1.**

Row Number equals 1 with partitioning will ensure only one record is fetched per partition

#### Question 5:

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?

(Designing data processing systems)

1. Use a fixed-time window with a duration of 60 minutes.
2. Use a sliding time window with a duration of 60 minutes.
3. **Use a session window with a gap time duration of 60 minutes.**
4. Use a global window with a time based trigger with a delay of 60 minutes.

There are 3 windowing concepts in dataflow and each can be used for below use case

1) Fixed window

2) Sliding window and

3) Session window.

**Fixed window** = any aggregation use cases, any batch analysis of data, relatively simple use cases.

**Sliding window** = **Moving averages of data**

**Session window** = user session data, click data and real time gaming analysis.

The question here is about user session data and hence session window.

#### Question 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?

(Designing data processing systems)

1. Create a Google Cloud Dataflow job to process the data.
2. Create a Google Cloud Dataproc cluster that uses persistent disks for HDFS.
3. Create a Hadoop cluster on Google Compute Engine that uses persistent disks.
4. **Create a Cloud Dataproc cluster that uses the Google Cloud Storage connector.**
5. Create a Hadoop cluster on Google Compute Engine that uses Local SSD disks.

#### Question 7:

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?

(Designing data processing systems)

1. **Put the data into Google Cloud Storage.**
2. Use preemptible virtual machines (VMs) for the Cloud Dataproc cluster.
3. Tune the Cloud Dataproc cluster so that there is just enough disk for all data.
4. Migrate some of the cold data into Google Cloud Storage, and keep only the hot data in Persistent Disk.

Description: First rule of dataproc is to keep data in GCS

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

B is not correct because this will decrease the compute cost but not the storage cost.

C is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Cloud Storage.

D is not correct because while this will reduce cost somewhat, it will not be as cost effective as using Cloud Storage.

#### Question 8:

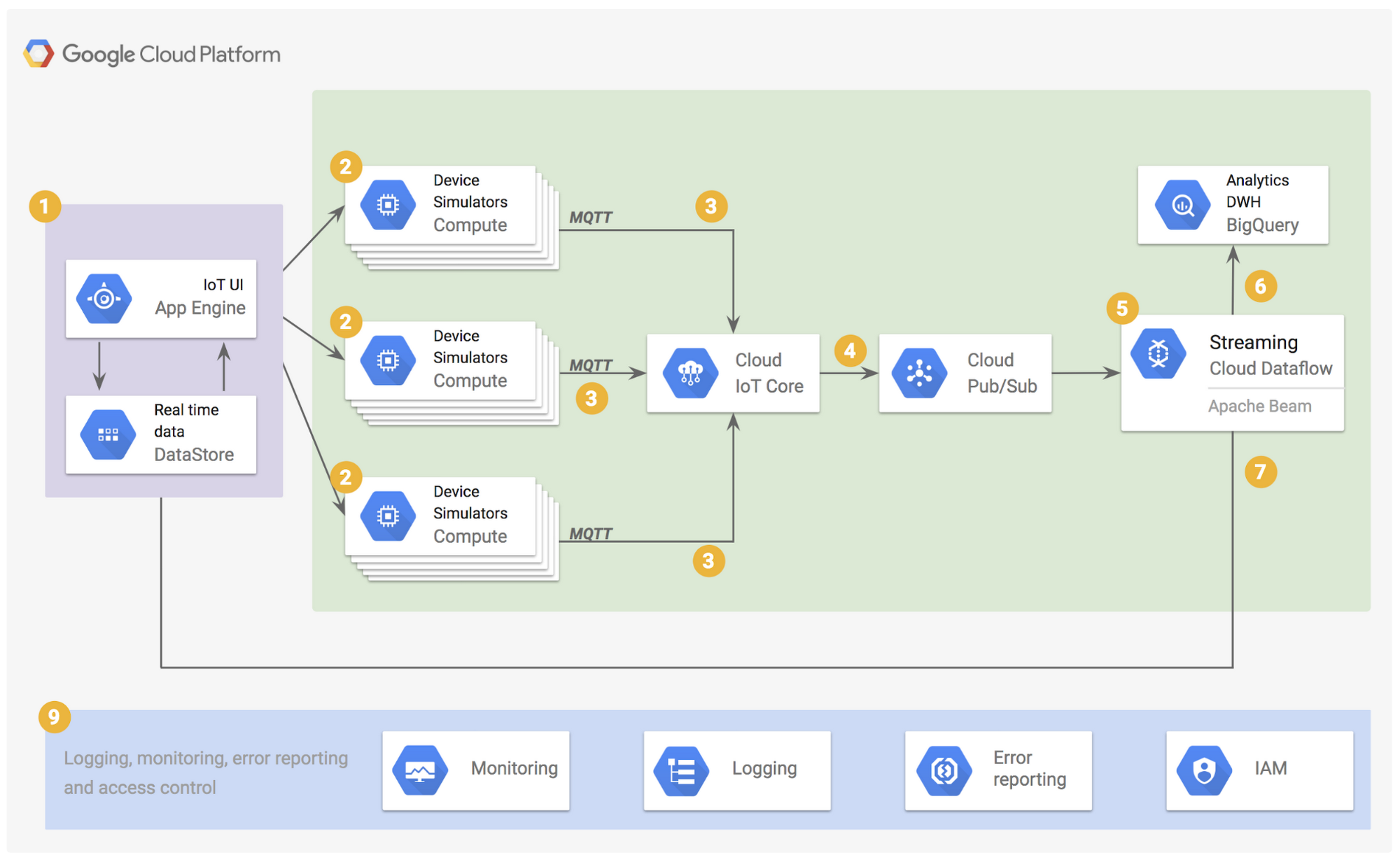
You are deploying 10,000 new Internet of Things devices to collect temperature data in your warehouses globally. You need to process, store and analyze these very large datasets in real time.

What should you do?

(Designing data processing systems)

1. Send the data to Google Cloud Datastore and then export to BigQuery.
2. **Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.**
3. Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
4. Export logs in batch to Google Cloud Storage and then spin up a Google Cloud SQL instance, import the data from Cloud Storage, and run an analysis as needed.

#### **End-to-end IoT solution on Google Cloud Platform**



The simulation, as explained above, starts from (1) App Engine that is a fully managed PaaS to implement scalable applications. In the App Engine frontend, the user triggers the generation of (simulated) devices and through APIs calls, the application will generate several instances on (2) Google Compute Engine, the IaaS solution to manage Virtual Machines (VMs). The following steps are executed at the startup of each VM:

* A public-and-private key pair is generated for each simulated device
* An instance of a (3) Java application that performs the following actions is launched for each simulated device:
  + registration of the device in *Cloud IoT Core*, a fully managed service designed to easily and securely connect, manage, and ingest data from globally dispersed devices
  + generation of a series of temperatures for a specified city
  + encapsulation of generated data into MQTT messages to make them available to Cloud IoT Core

Collected messages containing temperature values will then be published to a topic on (4) Cloud Pub/Sub, an enterprise message-oriented middleware. Here messages will be read in streaming mode by (5) Cloud Dataflow, a simplified stream and batch data processing solution, and then ingested into:

* (1) Cloud Datastore, a highly scalable, fully managed NoSQL database
* (6) BigQuery, a fast, highly scalable, cost-effective, and fully managed cloud data warehouse for analytics

Cloud **Datastore** will save data to be displayed directly into the UI of the App Engine application, while **BigQuery** will act as a **data warehouse** that will enable the execution of more in depth analysis.

All the logs generated by all components will then be ingested and monitored via (9) Stackdriver, a monitoring and management tool for services, containers, applications, and infrastructure. Permission and access will be managed via (9) Cloud IAM, a fine-grained access control and visibility tool for centrally managing cloud resources.

#### 

#### Question 9:

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?

(Designing data processing systems)

1. Grant the consultant the Viewer role on the project.
2. **Grant the consultant the Cloud Dataflow Developer role on the project.**
3. Create a service account and allow the consultant to log on with it.
4. Create an anonymized sample of the data for the consultant to work with in a different project.

The Dataflow developer role will not provide access to the underlying data.

#### Question 10:

#### Question 11:

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?

(Designing data processing systems)

1. Load the data every 30 minutes into a new partitioned table in BigQuery.
2. **Store and update the data in a regional Google Cloud Storage bucket and create a federated data source in BigQuery**
3. 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
4. Store the data in a file in a regional Google Cloud Storage bucket. Use Cloud Dataflow to query BigQuery and combine the data programmatically with the data stored in Google Cloud Storage.

regional storage is cheaper than BigQuery storage.

Use cases for external data sources include:

1. Loading and cleaning your data in one pass by querying the data from an external data source (a location external to BigQuery) and writing the cleaned result into BigQuery storage.
2. Having a small amount of frequently changing data that you join with other tables. As an external data source, the frequently changing data does not need to be reloaded every time it is updated.

#### Question 12:

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?

(Designing data processing systems)

1. BigQuery
2. **Cloud SQL**
3. Cloud Bigtable
4. Cloud Datastore

Answer is Cloud SQL

The database will be used to store all the transactional data of the product. what we need is the database only for store transactional data, not for analysis and ML. so the answer should be "the database that stores transactional data", which means, Cloud SQL. if you want to analyze or do ML you just specify Cloud SQL as a federated data source.

A: it's good for analysis but it costs too much to input/output data frequently.

C: BigTable is not good for transactional data.

D: okay datastore supports transactions, but it is weaker than RDB, and also, in this case, the data schema has already defined , you should use RDB.

#### Question 13: ??

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.)

(Designing data processing systems)

1. **Introduce data compression for each file to increase the rate file of file transfer.**
2. Contact your internet service provider (ISP) to increase your maximum bandwidth to at least 100 Mbps.
3. **Redesign the data ingestion process to use gsutil tool to send the CSV files to a storage bucket in parallel.**
4. **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.**
5. 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.

Answers are; AC or CD (not sure)

B - wrong (we need to provide solution without changing internet speed)

E - Bandwidth already low, so storage Transfer service will not help here.

#### Question 14:

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.)

(Designing data processing systems)

1. Redis
2. **HBase**
3. MySQL
4. **MongoDB**
5. **Cassandra**
6. HDFS with Hive

A. Redis - Redis is an in-memory non-relational key-value store. Redis is a great choice for **implementing a highly available in-memory cache to decrease data access latency, increase throughput,** and ease the load off your relational or NoSQL database and application. Since the question does not ask cache, A is discarded.

B. HBase - Meets reqs

**C. MySQL - they do not need ACID, so not needed.**

D. MongoDB - Meets reqs

E. Cassandra - Apache Cassandra is an open source NoSQL distributed database trusted by thousands of companies for scalability and high availability without compromising performance. Linear scalability and proven fault-tolerance on commodity hardware or cloud infrastructure make it the perfect platform for mission-critical data.

F. HDFS with Hive - Hive allows users to read, write, and manage petabytes of data using SQL. Hive is built on top of Apache Hadoop, which is an open-source framework used to efficiently store and process large datasets. As a result, Hive is closely integrated with Hadoop, and is designed to work quickly on petabytes of data. HIVE IS NOT A DATABSE.

#### Question 15:

#### 

#### Question 16:

Your globally distributed auction application allows users to bid on items. Occasionally, users place identical bids at nearly identical times, and different application servers process those bids. Each bid event contains the item, amount, user, and timestamp.

You want to collate those bid events into a single location in real time to determine which user bid first.

What should you do?

1. Create a file on a shared file and have the application servers write all bid events to that file. Process the file with Apache Hadoop to identify which user bid first.
2. Have each application server write the bid events to Cloud Pub/Sub as they occur. Push the events from Cloud Pub/Sub to a custom endpoint that writes the bid event information into Cloud SQL.
3. Set up a MySQL database for each application server to write bid events into. Periodically query each of those distributed MySQL databases and update a master MySQL database with bid event information.
4. Have each application server write the bid events to Google Cloud Pub/Sub as they occur. Use a pull subscription to pull the bid events using Google Cloud Dataflow. Give the bid for each item to the user in the bid event that is processed first.

#### Question 17:

Your organization has been collecting and analyzing data in Google BigQuery for 6 months. The majority of the data analyzed is placed in a time-partitioned table named events\_partitioned. To reduce the cost of queries, your organization created a view called events, which queries only the last 14 days of data. The view is described in legacy SQL. Next month, existing applications will be connecting to BigQuery to read the events data via an ODBC connection. You need to ensure the applications can connect.

Which two actions should you take? (Choose two.)

(Designing data processing systems)

1. Create a new view over events using standard SQL
2. Create a new partitioned table using a standard SQL query
3. **Create a new view over events\_partitioned using standard SQL**
4. **Create a service account for the ODBC connection to use for authentication**
5. Create a Google Cloud Identity and Access Management (Cloud IAM) role for the ODBC connection and shared "events"

A standard SQL query cannot reference a view defined using legacy SQL syntax. In order to connect through ODBC connection, we need to use standard SQL. So, we need to create a new view over events\_partitioned table using standard SQL which is C. Need service account to connect through ODBC which is option D.

#### 

#### Question 18:

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?

(Designing data processing systems)

1. Migrate the workload to Google Cloud Dataflow
2. **Use preemptible virtual machines (VMs) for the cluster**
3. Use a higher-memory node so that the job runs faster
4. Use SSDs on the worker nodes so that the job can run faster

Hadoop/Spark jobs are run on Dataproc, and the pre-emptible machines cost 80% less

#### Question 19:

Your infrastructure includes a set of YouTube channels. You have been tasked with creating a process for sending the YouTube channel data to Google Cloud for analysis. You want to design a solution that allows your world-wide marketing teams to perform ANSI SQL and other types of analysis on up-to-date YouTube channels log data.

How should you set up the log data transfer into Google Cloud?

(Designing data processing systems)

1. **Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.**
2. Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional bucket as a final destination.
3. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.
4. Use BigQuery Data Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.

Destination is GCS and having multi-regional so A is the best option available.

Even since BigQuery Data Transfer Service initially supports Google application sources like Google Ads, Campaign Manager, Google Ad Manager and YouTube but it does not support destination anything other than bq data set

#### Question 20:

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?

(Designing data processing systems)

1. Transform text files to compressed Avro using Cloud Dataflow. Use BigQuery for storage and query.
2. **Transform text files to compressed Avro using Cloud Dataflow. Use Cloud Storage and BigQuery permanent linked tables for query.**
3. Compress text files to gzip using the Grid Computing Tools. Use BigQuery for storage and query.
4. Compress text files to gzip using the Grid Computing Tools. Use Cloud Storage, and then import into Cloud Bigtable for query.

If you use gzip compression BigQuery cannot read the data in parallel….

The advantages of creating external tables are that they are fast to create so you skip the part of importing data and no additional monthly billing storage costs are accrued to your account since you only get charged for the data that is stored in the data lake, which is comparatively cheaper than storing it in BigQuery

A : Importing data into BigQuery may take more time compared to creating external tables on data. Additional storage costs by BigQuery is another issue which can be more expensive than Google Storage.

#### Question 21:

You are designing storage for 20 TB of text files as part of deploying a data pipeline on Google Cloud. Your input data is in CSV format. You want to minimize the cost of querying aggregate values for multiple users who will query the data in Cloud Storage with multiple engines.

Which storage service and schema design should you use?

(Designing data processing systems)

1. Use Cloud Bigtable for storage. Install the HBase shell on a Compute Engine instance to query the Cloud Bigtable data.
2. Use Cloud Bigtable for storage. Link as permanent tables in BigQuery for query.
3. **Use Cloud Storage for storage. Link as permanent tables in BigQuery for query.**
4. Use Cloud Storage for storage. Link as temporary tables in BigQuery for query.

#### Question 22:

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?

(Designing data processing systems)

1. Use Cloud SQL for storage. Add secondary indexes to support query patterns.
2. Use Cloud SQL for storage. Use Cloud Dataflow to transform data to support query patterns.
3. **Use Cloud Spanner for storage. Add secondary indexes to support query patterns.**
4. Use Cloud Spanner for storage. Use Cloud Dataflow to transform data to support query patterns.

Cloud Spanner allows transaction tables to scale horizontally and secondary indexes for range queries

#### Question 23:

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?

(Designing data processing systems)

1. **Cloud Bigtable**
2. Google BigQuery
3. Google Cloud Storage
4. Google Cloud Datastore

Bigtable is GCP’s managed wide-column database. It is also a good option for migrat-

ing on-premises **Hadoop HBase** databases to a managed database because Bigtable has

an HBase interface.

....

Cloud Bigtable is a wide-column NoSQL database used for high-volume databases that

require low millisecond (ms) latency. **Cloud Bigtable is used for IoT, time-series, finance,**

**and similar applications.**

#### Question 24:

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?

(Designing data processing systems)

1. **PigLatin using Pig**
2. HiveQL using Hive
3. Java using MapReduce
4. Python using MapReduce

Pig is a scripting language that can be used for checkpointing and splitting pipelines

#### Question 25:

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?

(Designing data processing systems)

1. **Install the OpenCensus Agent and create a custom metric collection application with a StackDriver exporter.**
2. Place the MariaDB instances in an Instance Group with a Health Check.
3. Install the StackDriver Logging Agent and configure fluentd in\_tail plugin to read MariaDB logs.
4. Install the StackDriver Agent and configure the MySQL plugin.

MariaDB needs costume metrics , and stackdriver built-in monitoring tools will not provide these metrics . Opencensus Agent will do this for you

#### Question 26:

You need to migrate a 2TB **relational database** to Google Cloud Platform. You do not have the resources to significantly refactor the application that uses this database and cost to operate is of primary concern.

Which service do you select for storing and serving your data?

(Designing data processing systems)

1. Cloud Spanner
2. Cloud Bigtable
3. Cloud Firestore
4. **Cloud SQL**

#### Question 27:

You are designing an Apache Beam pipeline to enrich data from **Cloud Pub/Sub** with static reference data from **BigQuery**. The reference data is small enough to fit in memory on a single worker. The pipeline should write enriched results to BigQuery for analysis.

Which job type and transforms should this pipeline use?

(Designing data processing systems)

1. Batch job, PubSubIO, side-inputs
2. Streaming job, PubSubIO, JdbcIO, side-outputs
3. **Streaming job, PubSubIO, BigQueryIO, side-inputs**
4. Streaming job, PubSubIO, BigQueryIO, side-outputs

You need pubsubIO and BigQueryIO for streaming data and writing enriched data back to BigQuery. side-inputs are a way to enrich the data

data will come from Pub/Sub, so it should be streaming, we'll need PubSubIO to be able to read from PubSub y BigQueryIO to be able to write to BigQuery, finally the side-inputs pattern let us enrich data

#### Question 28:

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.

Social media posts can images/videos which cannot be stored in bigquery

#### Question 29:

You want to automate execution of a multi-step data pipeline running on Google Cloud. The pipeline includes **Cloud Dataproc** and **Cloud Dataflow** jobs that have multiple dependencies on each other. You want to use managed services where possible, and the pipeline will run every day.

Which tool should you use?

(Designing data processing systems)

1. cron
2. **Cloud Composer**
3. Cloud Scheduler
4. Workflow Templates on Cloud Dataproc

Cloud Composer is an Apache Airflow managed service, it serves well when orchestrating **interdependent pipelines**, and Cloud Scheduler is just a managed Cron service.

#### Question 30:

You work for a shipping company that uses handheld scanners to read shipping labels. Your company has strict data privacy standards that require scanners to only transmit recipients' personally identifiable information (PII) to analytics systems, which violates user privacy rules. You want to quickly build a scalable solution using cloud-native managed services to prevent exposure of PII to the analytics systems.

What should you do?

A. Create an authorized view in BigQuery to restrict access to tables with sensitive data.

B. Install a third-party data validation tool on Compute Engine virtual machines to check the incoming data for sensitive information.

C. Use Stackdriver logging to analyze the data passed through the total pipeline to identify transactions that may contain sensitive information.

**D. Build a Cloud Function that reads the topics and makes a call to the Cloud Data Loss Prevention API. Use the tagging and confidence levels to either pass or quarantine the data in a bucket for review.**

Protection of sensitive data, like personally identifiable information (PII), is critical to your business. Deploy de-identification in migrations, data workloads, and real-time data collection and processing.

#### Question 31:

You are a retailer that wants to integrate your online sales capabilities with different in-home assistants, such as Google Home. You need to interpret customer voice commands and issue an order to the backend systems.

Which solutions should you choose?

(Designing data processing systems)

1. Cloud Speech-to-Text API
2. Cloud Natural Language API
3. **Dialogflow Enterprise Edition**
4. Cloud AutoML Natural Language

since we need to recognize both voice and intent

Dialogflow can interpret the commands (intents) and integrates other applications e.g. backend systems.

#### Question 32:

You are designing a data processing pipeline. The pipeline must be able to **scale automatically as load increases.** Messages must be processed at least once and must be ordered within windows of 1 hour. How should you design the solution?

A. Use Apache Kafka for message ingestion and use Cloud Dataproc for streaming analysis.

B. Use Apache Kafka for message ingestion and use Cloud Dataflow for streaming analysis.

C. Use Cloud Pub/Sub for message ingestion and Cloud Dataproc for streaming analysis.

**D. Use Cloud Pub/Sub for message ingestion and Cloud Dataflow for streaming analysis.**

rule of thumb: If you see Kafka and Pub/Sub, always go with Pub/Sub in Google exam

Careful doing that: I got a question where you had to choose between Kafka and Pub/Sub... and the solution required to be able to replay all messages without time limit. So no Pub/Sub there.

This being a Google cert does not mean that they always force Google solutions.

**Dataflow has autoscaling feature and pubsub is best solution**

#### Question 33:

You need to set access to **BigQuery** for different departments within your company. Your solution should comply with the following requirements:

* Each department should have access only to their data.
* Each department will have one or more leads who need to be able to create and update tables and provide them to their team.
* Each department has data analysts who need to be able to query but not modify data.

How should you set access to the data in BigQuery?

A. Create a dataset for each department. Assign the department leads the role of OWNER, and assign the data analysts the role of WRITER on their dataset.

**B. Create a dataset for each department. Assign the department leads the role of WRITER, and assign the data analysts the role of READER on their dataset.**

C. Create a table for each department. Assign the department leads the role of Owner, and assign the data analysts the role of Editor on the project the table is in.

D. Create a table for each department. Assign the department leads the role of Editor, and assign the data analysts the role of Viewer on the project the table is in.

The permissions are required at dataset levels hence READER, WRITER & OWNER which are the primitive roles for dataset to be used.

I also choose B for two reasons. One is that we want access at department level. In C & D, it is at project level. That means, one lead of one department will have all permissions for different department if all tables are in same project.

#### Question 34: ??

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.

Option A; Cheap storage and it is a supported meathod

https://cloud.google.com/datastore/docs/export-import-entities

Option B; Data exported from one Datastore mode database can be imported into another Datastore mode database, even one in another project.

https://cloud.google.com/datastore/docs/export-import-entities

#### Question 35:

You are designing a cloud-native historical data processing system to meet the following conditions:

✑ The data being analyzed is **in CSV, Avro, and PDF formats** and will be accessed by multiple analysis tools including Cloud Dataproc, BigQuery, and Compute Engine.

✑ A **streaming data pipeline** stores new data daily.

**✑ Peformance is not a factor in the solution.**

✑ The solution design should **maximize availability.**

How should you design data storage for this solution?

A. Create a Cloud Dataproc cluster with high availability. Store the data in HDFS, and peform analysis as needed.

B. Store the data in BigQuery. Access the data using the BigQuery Connector on Cloud Dataproc and Compute Engine.

C. Store the data in a regional Cloud Storage bucket. Access the bucket directly using Cloud Dataproc, BigQuery, and Compute Engine.

**D. Store the data in a multi-regional Cloud Storage bucket. Access the data directly using Cloud Dataproc, BigQuery, and Compute Engine.**

D (since pdf cannot be stored in BigQuery, and also questions asks for availability)

Multi-region increases high availability and pdf can be stored in gcs

Question 36:

Your United States-based company has created an application for assessing and responding to user actions. The primary table's data volume grows by **250,000 records per second**. Many third parties use your application's APIs to build the functionality into their own frontend applications. Your application's APIs should comply with the following requirements:

**✑ Single global endpoint**

**✑ ANSI SQL support**

**✑ Consistent access to the most up-to-date data**

What should you do?

A. Implement BigQuery with no region selected for storage or processing.

**B. Implement Cloud Spanner with the leader in North America and read-only replicas in Asia and Europe.**

C. Implement Cloud SQL for PostgreSQL with the master in Norht America and read replicas in Asia and Europe.

D. Implement Cloud Bigtable with the primary cluster in North America and secondary clusters in Asia and Europe.

B: Cloud Spanner is the first scalable, enterprise-grade, globally-distributed, and strongly consistent database service built for the cloud specifically to combine the benefits of relational database structure with non-relational horizontal scale.

**Cloud Spanner** is a fully managed, mission-critical, relational database service that offers transactional consistency at global scale, schemas, SQL (ANSI 2011 with extensions), and automatic, synchronous replication for high availability.

Cloud Spanner has three types of replicas: read-write replicas, read-only replicas, and witness replicas. Bigquery cannot support 250K data ingestion/second , as ANSI SQL support is required , no other options left except Spanner.

A - BigQuery with NO Region ? (Looks wrong)

B - Spanner (SQL support and Scalable and have replicas ) - Looks correct

C - SQL (can't store so many records) (wrong)

D - Bigtable - NO SQL (wrong)

#### Question 37:

You are building an application to share financial market data with consumers, who will receive data feeds. Data is collected from the markets in real time.

Consumers will receive the data in the following ways:

✑ Real-time event stream

✑ ANSI SQL access to real-time stream and historical data

✑ Batch historical exports

Which solution should you use?

A. Cloud Dataflow, Cloud SQL, Cloud Spanner

**B. Cloud Pub/Sub, Cloud Storage, BigQuery**

C. Cloud Dataproc, Cloud Dataflow, BigQuery

D. Cloud Pub/Sub, Cloud Dataproc, Cloud SQL

**・Cloud Storage** for storing exported logs in batch mode.

**・Pub/Sub** for streaming exported logs in streaming mode.

**・Dataflow** for processing log data.

**・BigQuery** for storing processing output and supporting rich queries on that output.

✑ Real-time event stream -> Pub/Sub

✑ ANSI SQL access to real-time stream and historical data -> BigQuery

✑ Batch historical exports -> Cloud Storage

#### Question 38:

You are building a new data pipeline to share data between two different types of applications: jobs generators and job runners. Your solution must **scale to accommodate increases in usage** and must accommodate the addition of new applications without negatively affecting the performance of existing ones. What should you do?

A. Create an API using App Engine to receive and send messages to the applications

**B. Use a Cloud Pub/Sub topic to publish jobs, and use subscriptions to execute them**

C. Create a table on Cloud SQL, and insert and delete rows with the job information

D. Create a table on Cloud Spanner, and insert and delete rows with the job information

Pub/sub will be used to streaming data between application

Pub/sub is used to transmit data in real time and scale automatically

#### Question 39:

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

Vote for A

A - Correct , Transfer Appliance for moving offline data, large data sets, or data from a source with limited bandwidth

https://cloud.google.com/storage-transfer/docs/overview

B - Eliminated (Not recommended for large storage). recommended for < 1TB

C - Its ONLINE, but we have bandwidth issue - So eliminated.

D - Eliminated (Not recommended for large storage). recommended for < 1TB

Huge amount of data with log network bandwidth, Transfer applicate is best for moving data over 100TB

#### Question 40:

You receive data files in CSV format monthly from a third party. You need to cleanse this data, but every third month the schema of the files changes. Your requirements for implementing these transformations include:

✑ Executing the transformations on a **schedule**

**✑ Enabling non-developer analysts to modify transformations**

✑ Providing a **graphical tool** for designing transformations

What should you do?

**A. Use Cloud Dataprep to build and maintain the transformation recipes, and execute them on a scheduled basis**

B. Load each month's CSV data into BigQuery, and write a SQL query to transform the data to a standard schema. Merge the transformed tables together with a SQL query

C. Help the analysts write a Cloud Dataflow pipeline in Python to perform the transformation. The Python code should be stored in a revision control system and modified as the incoming data's schema changes

D. Use Apache Spark on Cloud Dataproc to infer the schema of the CSV file before creating a Dataframe. Then implement the transformations in Spark SQL before writing the data out to Cloud Storage and loading into BigQuery

Dataprep is used by non developers

#### Question 41:

You work for a shipping company that has distribution centers where packages move on delivery lines to route them properly. The company wants to add cameras to the delivery lines to **detect and track any visual damage to the packages in transit.** You need to create a way to automate the detection of damaged packages and flag them for human review in real time while the packages are in transit. Which solution should you choose?

A. Use BigQuery machine learning to be able to train the model at scale, so you can analyze the packages in batches.

**B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.**

C. Use the Cloud Vision API to detect for damage, and raise an alert through Cloud Functions. Integrate the package tracking applications with this function.

D. Use TensorFlow to create a model that is trained on your corpus of images. Create a Python notebook in Cloud Datalab that uses this model so you can analyze for damaged packages.

Answer is B. Train an AutoML model on your corpus of images, and build an API around that model to integrate with the package tracking applications.

AutoML is used to train model and do damage detection.

Auto Vision is used is a pre trained model used to detect **objects** in images.

**AutoML Vision enables you to train machine learning models to classify your images according to your own defined labels.**

* Train models from labeled images and evaluate their performance.
* Leverage a human labeling service for datasets with unlabeled images.
* Register trained models for serving through the **AutoML API.**

#### Question 42:

You operate an IoT pipeline built around Apache Kafka that normally receives around 5000 messages per second. You want to use Google Cloud Platform to create an alert as soon as the **moving average** over 1 hour drops below 4000 messages per second. What should you do?

**A. Consume the stream of data in Cloud Dataflow using Kafka IO. Set a sliding time window of 1 hour every 5 minutes. Compute the average when the window closes, and send an alert if the average is less than 4000 messages.**

B. 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.

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 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.

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 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.

Kafka IO and Dataflow is a valid option for interconnect (needless where Kafka is located - On Prem/Google Cloud/Other cloud)

**Sliding Window will help to calculate average.**

#### Question 43:

Your company is selecting a system to centralize data ingestion and delivery. You are considering messaging and data integration systems to address the requirements. The key requirements are:

✑ The ability to seek to a particular offset in a topic, **possibly back to the start of all data ever captured**

✑ Support for **publish/subscribe** semantics on hundreds of topics

✑ Retain per-key ordering

Which system should you choose?

**A. Apache Kafka**

B. Cloud Storage

C. Cloud Pub/Sub

D. Firebase Cloud Messaging

Pub sub can retain message only for 7 days maximum

#### Question 44:

You are planning to migrate your **current on-premises Apache Hadoop deployment to the cloud.** You need to ensure that the deployment is as **fault-tolerant and cost-effective** as possible for **long-running batch jobs.** You want to use a managed service. What should you do?

**A. Deploy a Cloud Dataproc cluster. Use a standard persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://**

B. Deploy a Cloud Dataproc cluster. Use an SSD persistent disk and 50% preemptible workers. Store data in Cloud Storage, and change references in scripts from hdfs:// to gs://

C. Install Hadoop and Spark on a 10-node Compute Engine instance group with standard instances. Install the Cloud Storage connector, and store the data in Cloud Storage. Change references in scripts from hdfs:// to gs://

D. Install Hadoop and Spark on a 10-node Compute Engine instance group with preemptible instances. Store data in HDFS. Change references in scripts from hdfs:// to gs://

Ask for cost effective so persistent disk are HDD which are cheaper in comparison to SSD.

Cloud Dataproc for Managed Cloud native application and HDD for cost-effective solution.

#### Question 45:

You need to choose a database for a new project that has the following requirements:

- Fully managed

- Able to automatically scale up

- Transactionally consistent

- Able to scale up to 6 TB

- Able to be queried using SQL

Which database do you choose?

(Designing data processing systems)

**A. Cloud SQL**

B. Cloud Bigtable

C. Cloud Spanner

D. Cloud Datastore

Answer is **Cloud SQL**

**It asks for scaling up which can be done in cloud sql,** horizontal scaling is not possible in cloud sql

Automatic storage increase

If you enable this setting, Cloud SQL checks your available storage every 30 seconds. If the available storage falls below a threshold size, Cloud SQL automatically adds additional storage capacity. If the available storage repeatedly falls below the threshold size, Cloud SQL continues to add storage until it reaches the maximum of 30 TB.

Answer could be : A

Both Cloud SQL and Cloud Spanner supports SQL.

With Cloud SQL, you can go up to 10 TB of storage which also satisfies the other requirement.

Consistency - Of course, with Cloud SQL, you have single master and read replicas. So technically the data will be consistent across all instances so to speak.

The reason why I didn't choose Spanner is - There is no requirement for HA DR, multi region, secondary indexes, etc.. So I choose "A"

#### Question 46:

What are two of the benefits of using denormalized data structures in BigQuery?

A. Reduces the amount of data processed, reduces the amount of storage required

**B. Increases query speed, makes queries simpler**

C. Reduces the amount of storage required, increases query speed

D. Reduces the amount of data processed, increases query speed

Cannot be A or C because:

"Denormalized schemas aren't storage-optimal, but BigQuery's low cost of storage addresses concerns about storage inefficiency."

Cannot be D because the amount of data processed is the same.

As for why is it "simpler", I don't see it directly stated but it is hinted at: "Expressing records by using nested and repeated fields simplifies data load using JSON or Avro files." and "Expressing records using nested and repeated structures can provide a more natural representation of the underlying data."

Denormalized data can never reduce data processed nor the amount of storage required.

#### Question 47:

Which of the following are examples of hyperparameters? (Select 2 answers.)

**A. Number of hidden layers**

**B. Number of nodes in each hidden layer**

C. Biases

D. Weights

#### Question 48:

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**

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.**

#### Question 49:

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

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.

#### Question 50:

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

Description: Drain is used to stop pipeline without the data loss

**Important information on draining a job**

Draining a job is not supported for **batch pipelines.**

Your pipeline continues to incur the cost of maintaining any associated Google Cloud resources until all processing and writing is finished.

If your streaming pipeline code includes a looping timer, the job cannot be drained.

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.

**Important information on canceling a job**

Canceling a job immediately halts the processing of the pipeline.

You might lose in-flight data when you cancel a job. In-flight data refers to data that is already read but is still being processed by the pipeline.

Data that's written from the pipeline to an output sink before you cancelled the job might still be accessible on your output sink.

If data loss is not a concern, canceling your job ensures that the Google Cloud resources that are associated with your job are shut down as soon as possible.

#### Question 51:

Which of the following statements is NOT true regarding Bigtable access roles?

A. Using IAM roles, you cannot give a user access to only one table in a project, rather than all tables in a project.

**B. To give a user access to only one table in a project, grant the user the Bigtable Editor role for that table.**

C. You can configure access control only at the project level.

D. To give a user access to only one table in a project, you must configure access through your application.

**Bigtable** uses Identity and Access Management (IAM) for access control.

For Bigtable, you can configure access control at the **project, instance, and table levels.** Here are some examples of using access control at the project level:

Allow a user to read from, but not write to, any table within the project.

Allow a user to read from and write to any table within the project, but not manage instances.

Allow a user to read from and write to any table within the project, and manage instances.

#### Question 52:

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

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.

#### Question 53:

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

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.

Cloud Bigtable 架构中，所有客户端请求在发送到 Cloud Bigtable 节点之前都经过前端服务器。

节点被组织成一个 Cloud Bigtable 集群，该集群属于一个 Cloud Bigtable 实例，该实例是集群的容器。集群中的每个节点都处理集群请求的一个子集。

将其他节点添加到集群时，您可以增加集群可以处理的同时请求数，以及整个集群的最大吞吐量。

#### Question 54:

In order to securely transfer web traffic data from your computer's web browser to the Cloud Dataproc cluster you should use a(n) \_\_\_\_\_.

1. VPN connection
2. Special browser
3. **SSH tunnel**
4. FTP connection

为了安全地将网络流量数据从计算机的网络浏览器传输到 Cloud Dataproc 集群，您应该使用 a(n) \_\_\_\_\_。

要连接到 Web 界面，建议使用 SSH 隧道来创建与主节点的安全连接。

#### Question 55:

The YARN ResourceManager and the HDFS NameNode interfaces are available on a Cloud Dataproc cluster \_\_\_\_.

1. application node
2. conditional node
3. **master node**
4. worker node

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 suffixfor example, if your cluster is named "my-cluster", the master-host-name would be "my-cluster-m".

#### Question 56:

Cloud Dataproc charges you only for what you really use with \_\_\_\_\_ billing.

**minute-by-minute**

One of the advantages of Cloud Dataproc is its low cost. Dataproc charges for what you really use with minute-by-minute billing and a low, ten-minute-minimum billing period.

#### Question 57:

Scaling a Cloud Dataproc cluster typically involves \_\_\_\_.

(Designing data processing systems)

1. **increasing or decreasing the number of worker nodes**
2. increasing or decreasing the number of master nodes
3. moving memory to run more applications on a single node
4. deleting applications from unused nodes periodically

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

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

2) decrease the number of workers to save money

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

#### Question 58:

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

(Designing data processing systems)

**value**

file\_prefix1:property1=**value1**,file\_prefix2:property2=value2,...

#### Question 59:

Which action can a Cloud Dataproc Viewer perform?

(Designing data processing systems)

Submit a job.

Create a cluster.

Delete a cluster.

**List the jobs.**

A Cloud Dataproc Viewer is limited in its actions based on its role. A viewer can only list clusters, get cluster details, list jobs, get job details, list operations, and get operation details.

### Dataproc roles

[Dataproc IAM roles](https://cloud.google.com/iam/docs/understanding-roles#dataproc-roles) are a bundle of one or more [permissions](https://cloud.google.com/dataproc/docs/concepts/iam/iam#permissions). You grant roles to users or groups to allow them to perform actions on the Dataproc resources in your project. For example, the Dataproc Viewer role contains the dataproc.\*.get and dataproc.\*.list permissions, which allow a user to get and list Dataproc clusters, jobs, and operations in a project.

#### 

#### Question 60:

Cloud Dataproc is a managed Apache Hadoop and Apache \_\_\_\_\_ service.

Spark

Cloud Dataproc is a managed **Apache Spark** and **Apache Hadoop** service that lets you use open source data tools for batch processing, querying, streaming, and machine learning.

#### Question 61:

When using Cloud Dataproc clusters, you can access the YARN web interface by configuring a browser to connect through a \_\_\_\_ proxy.

A. HTTPS

B. VPN

C. **SOCKS**

D. HTTP

When using Cloud Dataproc clusters, configure your browser to use the SOCKS proxy. The SOCKS proxy routes data intended for the Cloud Dataproc cluster through an SSH tunnel.

#### 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.**

no data storage as it is only for compute, preemptible workers can be taken away anytime so we cannot have only preemptible workers

### Secondary workers - preemptible and non-preemptible VMs

In addition to using standard Compute Engine VMs as Dataproc workers (called "primary" workers), **Dataproc** clusters can use "secondary" workers.

The following characteristics apply to all secondary workers in a Dataproc cluster:

**Processing only—Secondary workers do not store data.** They only function as processing nodes. Therefore, you can use secondary workers to scale compute without scaling storage.

**No secondary-worker-only clusters—Your cluster must have primary workers.** If you create a cluster and you do not specify the number of primary workers, Dataproc adds two primary workers to the cluster.

**Machine type—Secondary workers use the machine type of the cluster's primary workers.** For example, if you create a cluster with primary workers that use n1-standard-4 machine types, all secondary workers added to the cluster will also use n1-standard-4 machines.

**Persistent disk size—As a default, secondary workers are created with the smaller of 100GB or the primary worker boot disk size.** T

#### Question 63:

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

### Create a Dataproc cluster

You can create a new Dataproc cluster with the **CreateCluster API.**

You must specify the following values when creating a cluster:

* The **project** in which the cluster will be created
* The **name of the cluster**
* The **region** to use. If you specify the global region (the tutorial code uses a --global\_region flag to select the global region), you must also specify a **zone** (see zone\_uri). If you specify a non-global region and leave the zone\_uri field empty, Dataproc Auto Zone Placement will select a zone for your cluster.

#### Question 64:

Which role must be assigned to a service account used by the virtual machines in a Dataproc cluster so they can execute jobs?

**A. Dataproc Worker**

B. Dataproc Viewer

C. Dataproc Runner

D. Dataproc Editor

### Dataproc service accounts

The following service accounts are granted permissions required to perform Dataproc actions in the project where your cluster is located.

* **Dataproc VM service account:** VMs in a Dataproc cluster use this service account for Dataproc data plane operations, such reading and writing data from and to Cloud Storage and BigQuery.
* The **Dataproc Worker role** provides the VM service account with the m**inimum permissions** necessary to operate with Dataproc. Additional roles are necessary to grant permissions to read and write data to Google Cloud resources, such as BigQuery.
* **Dataproc Service Agent service account**

#### Question 65:

What are the minimum permissions needed for a service account used with Google Dataproc?

A. Execute to Google Cloud Storage; write to Google Cloud Logging

B. Write to Google Cloud Storage; read to Google Cloud Logging

C. Execute to Google Cloud Storage; execute to Google Cloud Logging

**D. Read and write to Google Cloud Storage; write to Google Cloud Logging**

Service 帐户将在您的虚拟机实例上运行的应用程序验证到其他 Google Cloud Platform 服务。例如，如果您编写一个在 Google Cloud Storage 上读写文件的应用程序，它必须首先通过 Google Cloud Storage API 进行身份验证。与 Cloud Dataproc 一起使用的服务帐号至少需要读取和写入 Google Cloud Storage 以及写入 Google Cloud Logging 的权限。

#### Question 66:

Which of the following job types are supported by Cloud Dataproc (select 3 answers)?

1. **Hive**
2. **Pig**
3. YARN
4. **Spark**

### **Dataproc** Frequently asked questions:

How does Dataproc work?

**Dataproc** is a managed framework that runs on the Google Cloud Platform and ties together several popular tools for processing data, including Apache Hadoop, Spark, Hive, and Pig. Dataproc has a set of control and integration mechanisms that coordinate the lifecycle, management, and coordination of clusters. **Dataproc is integrated with the** **YARN application manager** **to make managing and using your clusters easier.**

What type of jobs can I run?

**Dataproc** provides **out-of-the box** and **end-to-end support** for many of the most popular job types, including **Spark, Spark SQL, PySpark, MapReduce, Hive, and Pig jobs.**

What **Cluster Manager** does Dataproc use with Spark?

Dataproc runs **Spark on YARN.**

Can I run a persistent cluster?

Once started, **Dataproc clusters continue to run until shut down.** You can run a Dataproc cluster for as long as you need.

如何在集群上提交作业？

有多种方法可以在 Dataproc 集群上提交作业。最简单的方法是使用 Google Cloud Console 上的 Dataproc提交作业 页面或 gcloud CLI gcloud dataproc jobs submit 命令。如需以编程方式提交作业，请参阅 Dataproc API 参考。

我可以一次运行多个工作吗？

是的，您可以在 Dataproc 集群上一次运行多个作业。Cloud Dataproc 利用资源管理器 (YARN) 和特定于应用程序的配置（例如使用 Spark 进行扩展）来优化集群上资源的使用。作业性能将随着集群大小和活动作业的数量而扩展。

我可以取消集群上的作业吗？

确实。可以通过Google Cloud Console 网络界面或命令行取消作业。Dataproc 根据请求利用 YARN 应用程序取消来停止作业。

#### Question 67:

By default, which of the following windowing behavior does Dataflow apply to **unbounded** data sets?

A. Windows at every 100 MB of data

**B. Single, Global Window**

C. Windows at every 1 minute

D. Windows at every 10 minutes

Dataflow's default windowing behavior is to assign all elements of a PCollection to a single, global window, even for unbounded PCollections

It just ask for what is the default window that will apply if we don't specify window.

the default window for dataflow is global window.

#### Question 68:

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

#### Question 69:

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

1. **dataflow.worker**
2. dataflow.compute
3. dataflow.developer
4. dataflow.viewer

### Access control with IAM:

**Dataflow Worker**

(roles/dataflow.worker)

Provides the permissions necessary for a Compute Engine service account to execute work units for a Dataflow pipeline.

Lowest-level resources where you can grant this role:

* Project

**Dataflow Viewer**

(roles/dataflow.viewer)

Provides read-only access to all Dataflow-related resources.

Lowest-level resources where you can grant this role:

* Project

**Dataflow Developer**

(roles/dataflow.developer)

Provides the permissions necessary to execute and manipulate Dataflow jobs.

Lowest-level resources where you can grant this role:

* Project

**Dataflow Admin**

(roles/dataflow.admin)

Minimal role for creating and managing dataflow jobs.

#### Question 70:

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

A. PCollection

**B. Transform**

C. Pipeline

D. Sink API

### Apache Beam Concepts

This section contains summaries of fundamental concepts. On the **Apache Beam** website, the Apache Beam Programming Guide walks you through the basic concepts of **building pipelines using the Apache Beam SDKs.**

Basic concepts:

**Pipelines**

A pipeline encapsulates the entire series of computations involved in reading input data, transforming that data, and writing output data. The input source and output sink can be the same or of different types, allowing you to convert data from one format to another. Apache Beam programs start by constructing a Pipeline object, and then using that object as the basis for creating the pipeline's datasets. **Each pipeline represents a single, repeatable job.**

**PCollection**

A PCollection represents a potentially distributed, multi-element dataset that acts as the pipeline's data. Apache Beam **transforms** use PCollection objects as inputs and outputs for each step in your pipeline. A PCollection can hold a dataset of a fixed size or an unbounded dataset from a continuously updating data source.

**Transforms**

A transform represents a processing operation that transforms data. A transform takes one or more PCollections as input, performs an operation that you specify on each element in that collection, and produces one or more PCollections as output. A transform can perform nearly any kind of processing operation, including performing mathematical computations on data, converting data from one format to another, grouping data together, reading and writing data, filtering data to output only the elements you want, or combining data elements into single values.

转换表示转换数据的处理操作。转换将一个或多个PCollections 作为输入，对集合中的每个元素执行您指定的操作，并生成一个或多个 PCollections 作为输出。转换可以执行几乎任何类型的处理操作，包括对数据执行数学计算、将数据从一种格式转换为另一种格式、将数据组合在一起、读取和写入数据、过滤数据以仅输出您想要的元素，或将数据元素组合成单一值。

**ParDo**

ParDo is the core parallel processing operation in the Apache Beam SDKs, invoking a user-specified function on each of the elements of the input PCollection. ParDo collects the zero or more output elements into an output PCollection. The ParDo transform processes elements independently and possibly in parallel.

**Pipeline I/O**

Apache Beam I/O connectors let you read data into your pipeline and write output data from your pipeline. An I/O connector consists of a source and a sink. All Apache Beam sources and sinks are transforms that let your pipeline work with data from several different data storage formats. You can also write a custom I/O connector.

**Aggregation**

Aggregation is the process of computing some value from multiple input elements. The primary computational pattern for aggregation in Apache Beam is to group all elements with a common key and window. Then, it combines each group of elements using an associative and commutative operation.

**User-defined functions (UDFs)**

Some operations within Apache Beam allow executing user-defined code as a way of configuring the transform. For ParDo, user-defined code specifies the operation to apply to every element, and for Combine, it specifies how values should be combined. A pipeline might contain UDFs written in a different language than the language of your runner. A pipeline might also contain UDFs written in multiple languages.

**Runner**

Runners are the software that accepts a pipeline and executes it. Most runners are translators or adapters to massively parallel big-data processing systems. Other runners exist for local testing and debugging.

**Source**

A transform that reads from an external storage system. A pipeline typically reads input data from a source. The source has a type, which may be different from the sink type, so you can change the format of data as it moves through the pipeline.

**Sink**

A transform that writes to an external data storage system, like a file or a database.

#### Question 72:

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

1. Is payment TRANSACTION -- DB should able to perform full blown transaction (updating inventory, sales info etc, though not specified) , not just ATOMIC which DataStore provides

2. Its point-of-sale application, not ONLINE STORE where HIGH number of concurrent users ordering stuff.

3. User Base could grow exponentially - again more users does mot mean concurrent users and more processing power. Its only about storage.

4. Do not want to Manage infrastructure scaling. - Cloud SQL can scale in terms of storage.

5. **CloudStore is poor selection for OLTP application**

- Each property is index - so higher latency

1. 付款交易 - DB 应该能够执行完整的交易（更新库存、销售信息等，虽然没有指定），不仅仅是 DataStore 提供的 ATOMIC 2. 它的销售点应用程序，而不是在线商店，其中有大量并发用户订购东西。3. 用户群可能呈指数级增长——更多的用户并不意味着并发用户和更多的处理能力。它只是关于存储。4.不想管理基础设施扩展。- Cloud SQL 可以在存储方面进行扩展。5. CloudStore 对 OLTP 应用程序的选择很差 - 每个属性都是索引 - 所以延迟更高

#### 

#### Question 73:

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.

A cache is a temporary data storage system. **Fetching cached data can be much faster than fetching it directly from the underlying data set**, and helps reduce the number of queries sent, minimizing costs for paid data access.

#### Question 74:

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

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

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

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

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

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

#### Question 75:

You have spent a few days loading data from comma-separated values (CSV) files into the Google BigQuery table CLICK\_STREAM. The column DT stores the epoch time of click events. For convenience, you chose a simple schema where every field is treated as the STRING type. 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 DT is of the TIMESTAMP type. Reload the data.

B. Add a column TS of the TIMESTAMP type to the table CLICK\_STREAM, and populate the numeric values from the column TS for each row. Reference the column TS instead of the column DT from now on.

C. Create a view CLICK\_STREAM\_V, where strings from the column DT are cast into TIMESTAMP values. Reference the view CLICK\_STREAM\_V instead of the table CLICK\_STREAM from now on.

D. Add two columns to the table CLICK STREAM: TS of the TIMESTAMP type and IS\_NEW of the BOOLEAN type. Reload all data in append mode. For each appended row, set the value of IS\_NEW to true. For future queries, reference the column TS instead of the column DT, with the WHERE clause ensuring that the value of IS\_NEW must be true.

**E. Construct a query to return every row of the table CLICK\_STREAM, while using the built-in function to cast strings from the column DT into TIMESTAMP values. Run the query into a destination table NEW\_CLICK\_STREAM, in which the column TS is the TIMESTAMP type. Reference the table NEW\_CLICK\_STREAM instead of the table CLICK\_STREAM from now on. In the future, new data is loaded into the table NEW\_CLICK\_STREAM.**

**There are two ways to manually change a column's data type:**

**Using a SQL query: choose this option if you are more concerned about simplicity and ease of use, and you are less concerned about costs.**

**Recreating the table: choose this option if you are more concerned about costs, and you are less concerned about simplicity and ease of use.**

#### Question 76:

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?

1. **Update the current pipeline and use the drain flag.**
2. Update the current pipeline and provide the transform mapping JSON object.
3. Create a new pipeline that has the same Cloud Pub/Sub subscription and cancel the old pipeline.
4. Create a new pipeline that has a new Cloud Pub/Sub subscription and cancel the old pipeline.

Question mentions not to lose any data. All other options may lead to some data loss. If you want to prevent data loss as you bring down your streaming pipelines. 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.

#### Question 77:

Your software uses a simple JSON format for all messages. These messages are published to Google Cloud Pub/Sub, then processed with Google Cloud Dataflow to create a real-time dashboard for the CFO. During testing, you notice that some messages are missing in the dashboard.

You check the logs, and all messages are being published to Cloud Pub/Sub successfully.

What should you do next?

1. Check the dashboard application to see if it is not displaying correctly.
2. **Run a fixed dataset through the Cloud Dataflow pipeline and analyze the output.**
3. Use Google Stackdriver Monitoring on Cloud Pub/Sub to find the missing messages.
4. Switch Cloud Dataflow to pull messages from Cloud Pub/Sub instead of Cloud Pub/Sub pushing messages to Cloud Dataflow.

#### Question 78:

You work for a large fast food restaurant chain with over 400,000 employees. You store employee information in Google BigQuery in a Users table consisting of a FirstName field and a LastName field. A member of IT is building an application and asks you to modify the schema and data in BigQuery so the application can query a FullName field consisting of the value of the FirstName field concatenated with a space, followed by the value of the LastName field for each employee.

How can you make that data available while minimizing cost?

(Building and operationalizing data processing systems)

1. Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
2. **Add a new column called FullName to the Users table. Run an UPDATE statement that updates the FullName column for each user with the concatenation of the FirstName and LastName values.**
3. Create a Google Cloud Dataflow job that queries BigQuery for the entire Users table, concatenates the FirstName value and LastName value for each user, and loads the proper values for FirstName, LastName, and FullName into a new table in BigQuery.
4. Use BigQuery to export the data for the table to a CSV file. Create a Google Cloud Dataproc job to process the CSV file and output a new CSV file containing the proper values for FirstName, LastName and FullName. Run a BigQuery load job to load the new CSV file into BigQuery.

<https://www.examtopics.com/discussions/google/view/16819-exam-professional-data-engineer-topic-1-question-43/>

#### Question 79:

You have enabled the free integration between Firebase Analytics and Google BigQuery. Firebase now automatically creates a new table daily in BigQuery in the format app\_events\_YYYYMMDD. You want to query all of the tables for the past 30 days in **legacy** SQL.

What should you do?

1. **Use the TABLE\_DATE\_RANGE function**
2. Use the WHERE\_PARTITIONTIME pseudo column
3. Use WHERE date BETWEEN YYYY-MM-DD AND YYYY-MM-DD
4. Use SELECT IF.(date >= YYYY-MM-DD AND date <= YYYY-MM-DD

Description: Legacy sql uses table date range whereas standard sql uses table\_sufix for wildcard

Table wildcard functions:

TABLE\_DATE\_RANGE() Queries multiple daily tables that span a date range.

TABLE\_DATE\_RANGE\_STRICT() Queries multiple daily tables that span a date range, with

no missing dates.

TABLE\_QUERY() Queries tables whose names match a specified predicate.

#### Question 80:

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 assigned the timestamp, which causes the job to fail

B. They have not set the triggers to accommodate the data coming in late, which causes the job to fail

C. They have not applied a global windowing function, which causes the job to fail when the pipeline is created

**D. They have not applied a non-global windowing function, which causes the job to fail when the pipeline is created**

### Windowing:

Beam’s default **windowing** behavior is to assign all elements of a PCollection to a single, global window and discard late data, even for unbounded PCollections. Before you use a grouping transform such as GroupByKey on an unbounded PCollection, you must do at least one of the following:

* Set a **non-global windowing function.** See Setting your PCollection’s windowing function.
* Set a **non-default trigger.** This allows the global window to emit results under other conditions, since the default windowing behavior (waiting for all data to arrive) will never occur.

#### Question 81:

You architect a system to analyze seismic data. Your extract, transform, and load (ETL) process runs as a series of MapReduce jobs on an Apache Hadoop cluster. The ETL process takes days to process a data set because some steps are computationally expensive. Then you discover that a sensor calibration step has been omitted. How should you change your ETL process to carry out sensor calibration systematically in the future?

A. Modify the transformMapReduce jobs to apply sensor calibration before they do anything else.

**B. Introduce a new MapReduce job to apply sensor calibration to raw data, and ensure all other MapReduce jobs are chained after this.**

C. Add sensor calibration data to the output of the ETL process, and document that all users need to apply sensor calibration themselves.

D. Develop an algorithm through simulation to predict variance of data output from the last MapReduce job based on calibration factors, and apply the correction to all data.

Two reasons, it is a cleaner approach with single job to handle the calibration before the data is used in the pipeline. Second, doing this step in later stages can be complex and maintenance of those jobs in the future will become challenging.

#### Question 82:

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

A. BigQuery

**B. Cloud SQL**

C. Cloud BigTable

D. Cloud Datastore

Cloud SQL should be the only correct answer. Required solution needs to support transactions as well as analysis through a BI tool. Firestore/Datastore does not support SQL syntax typically needed to do analysis done by a BI tool. BigQuery is not suitable for transactional use case. BigTable does not support SQL.

#### Question 83:

You launched a new gaming app almost three years ago. You have been uploading log files from the previous day to a separate Google BigQuery table with the table name format LOGS\_yyyymmdd. You have been using table wildcard functions to generate daily and monthly reports for all time ranges. Recently, you discovered that some queries that cover long date ranges are exceeding the limit of 1,000 tables and failing. How can you resolve this issue?

A. Convert all daily log tables into date-partitioned tables

**B. Convert the sharded tables into a single partitioned table**

C. Enable query caching so you can cache data from previous months

D. Create separate views to cover each month, and query from these views

Answer is Convert the sharded tables into a single partitioned table

Google says that when you have multiple wildcard tables, best option is to shard it into single partitioned table. Time and cost efficient

### Partitioning versus sharding

Table sharding is the practice of storing data in multiple tables, using a naming prefix such as [PREFIX]\_YYYYMMDD.

Partitioning is recommended over table sharding, because partitioned tables perform better. With sharded tables, BigQuery must maintain a copy of the schema and metadata for each table. BigQuery might also need to verify permissions for each queried table. This practice also adds to query overhead and affects query performance.

If you previously created date-sharded tables, you can convert them into an ingestion-time partitioned table. For more information, see Convert date-sharded tables into ingestion-time partitioned tables.

#### Question 84:

You are integrating one of your internal IT applications and Google BigQuery, so users can query BigQuery from the application's interface. You do not want individual users to authenticate to BigQuery and you do not want to give them access to the dataset. You need to securely access BigQuery from your IT application. What should you do?

A. Create groups for your users and give those groups access to the dataset

B. Integrate with a single sign-on (SSO) platform, and pass each user's credentials along with the query request

**C. Create a service account and grant dataset access to that account. Use the service account's private key to access the dataset**

D. Create a dummy user and grant dataset access to that user. Store the username and password for that user in a file on the files system, and use those credentials to access the BigQuery dataset

Service Account are best option when granting access from tools/appllications

#### Question 85:

You are selecting services to write and transform JSON messages from Cloud Pub/Sub to BigQuery for a data pipeline on Google Cloud. You want to **minimize service costs.** You also want to monitor and accommodate input data volume that will vary in size with minimal manual intervention.

A. Use Cloud Dataproc to run your transformations. Monitor CPU utilization for the cluster. Resize the number of worker nodes in your cluster via the command line.

B. Use Cloud Dataproc to run your transformations. Use the diagnose command to generate an operational output archive. Locate the bottleneck and adjust cluster resources.

**C. Use Cloud Dataflow to run your transformations. Monitor the job system lag with Stackdriver. Use the default autoscaling setting for worker instances.**

D. Use Cloud Dataflow to run your transformations. Monitor the total execution time for a sampling of jobs. Configure the job to use non-default Compute Engine machine types when needed.

Dataproc does not seem to be a good solution in this case as it always require a manual intervention to adjust resources.

**Autoscaling** with dataflow will automatically handle changing data volumes with no manual intervention, and monitoring through Stackdriver can be used to spot bottleneck. Total execution time is not good there as it does not provide a precise view on potential bottleneck.

**Dataflow** provides a cost-effective solution to perform transformations on the streaming data, with autoscaling provides scaling without any intervention. System lag with Stackdriver provides monitoring for the streaming data. With autoscaling enabled, the Cloud Dataflow service automatically chooses the appropriate number of worker instances required to run your job.

#### Question 86:

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. Encrypted on Cloud Storage with user-supplied encryption keys. A separate decryption key will be given to each authorized user.

B. In a BigQuery dataset that is viewable only by authorized personnel, with the Data Access log used to provide the auditability.

C. In Cloud SQL, with separate database user names to each user. The Cloud SQL Admin activity logs will be used to provide the auditability.

D. 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.

Keywords here are

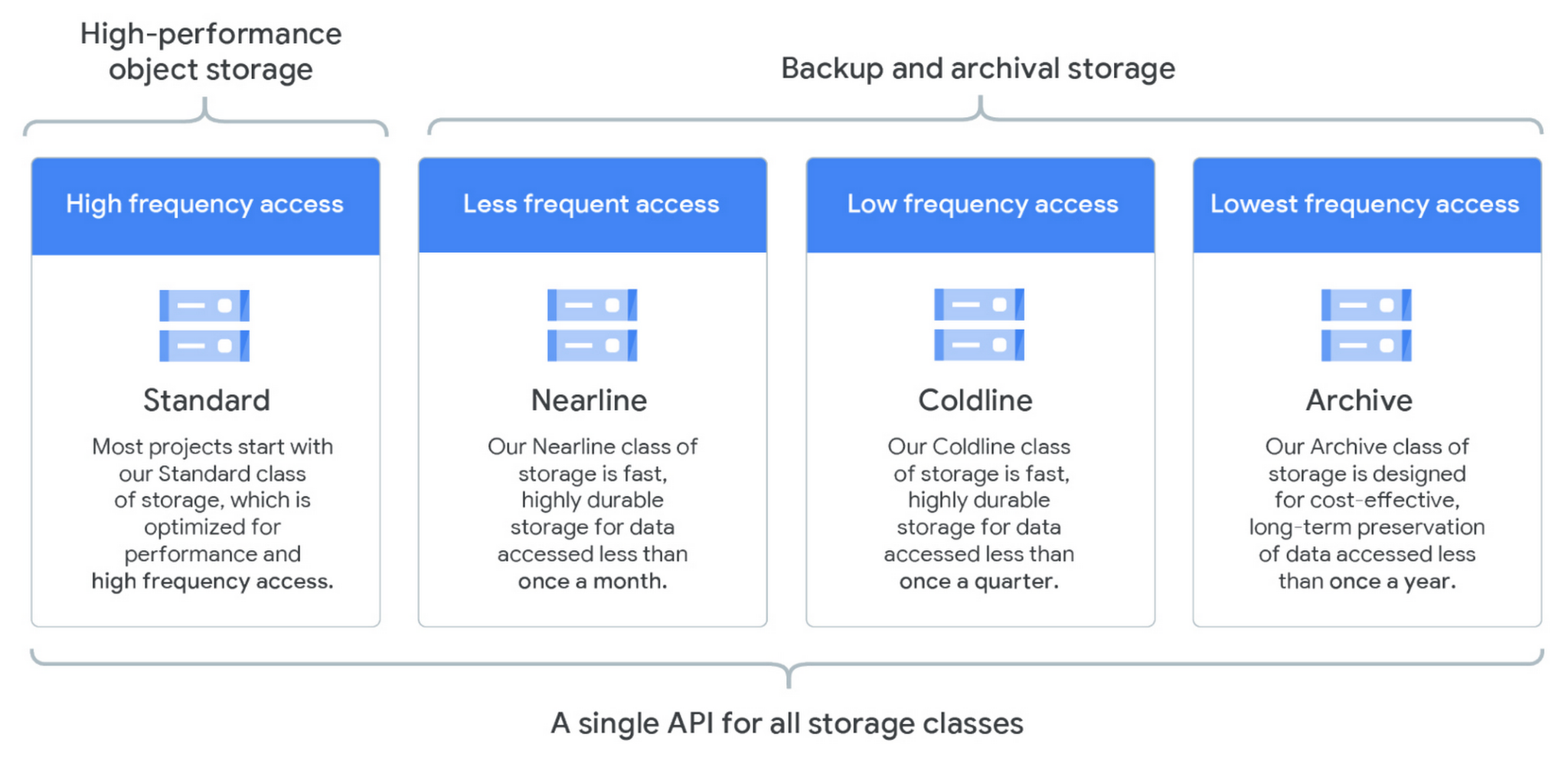
1. "Archived": Immutable and hence, BQ and Cloud SQL are ruled out

2. "Auditable": Means track any changes done.

Only D can provide the audibility piece!

I will go with D

Here’s an overview of how Archive fits into Cloud Storage classes:



#### Question 87:

After migrating ETL jobs to run on BigQuery, you need to verify that the output of the migrated jobs is the same as the output of the original. You've loaded a table containing the output of the original job and want to compare the contents with output from the migrated job to show that they are identical. The tables do not contain a primary key column that would enable you to join them together for comparison.

What should you do?

A. Select random samples from the tables using the RAND() function and compare the samples.

B. Select random samples from the tables using the HASH() function and compare the samples.

**C. Use a Dataproc cluster and the BigQuery Hadoop connector to read the data from each table and calculate a hash from non-timestamp columns of the table after sorting. Compare the hashes of each table.**

D. Create stratified random samples using the OVER() function and compare equivalent samples from each table.

Description:

options A B and D only will determine that it “might” be identical since is only a sample. HASH() can be helpful when doing bulk comparisons, but you still have to compare field by field to get the final answer.

#### Question 88:

You are a head of BI at a large enterprise company with multiple business units that each have different priorities and budgets. You use on-demand pricing for

**BigQuery** with a quota of 2K concurrent on-demand slots per project. Users at your organization sometimes don't get slots to execute their query and you need to correct this. You'd like to avoid introducing new projects to your account.

What should you do?

A. Convert your batch BQ queries into interactive BQ queries.

B. Create an additional project to overcome the 2K on-demand per-project quota.

**C. Switch to flat-rate pricing and establish a hierarchical priority model for your projects.**

D. Increase the amount of concurrent slots per project at the Quotas page at the Cloud Console.

Description:

If you use **BigQuery** with **on-demand pricing**, your project is allocated up to 2000 slots distributed among your project’s queries. The only lever available to you is to select interactive vs. batch query type. By default, BigQuery runs interactive queries. As soon as an interactive query is issued, it competes for slots with all other interactive queries that are concurrently running in other on-demand projects. Batch queries are queued and executed as soon as idle resources are available

You might simply prefer a fixed monthly bill, or encounter workloads that are extremely sensitive to query latency, and thus have predictability and control requirements that cannot be met by the on-demand service. For such situations, you can use the **flat-rate service.** In this model, a certain number of slots are dedicated to your project(s), and you can establish a hierarchical priority model amongst the projects. **The flat-rate model is especially suitable for large enterprises with multiple business units and workloads with varying priorities and budgets.**

#### Question 89:

You have an **Apache Kafka** cluster on-prem with topics containing web application logs. You need to **replicate** the data to Google Cloud for analysis in BigQuery and Cloud Storage. The preferred replication method is **mirroring** to **avoid deployment of Kafka Connect plugins.**

What should you do?

**A. Deploy a Kafka cluster on GCE VM Instances. Configure your on-prem cluster to mirror your topics to the cluster running in GCE. Use a Dataproc cluster or Dataflow job to read from Kafka and write to GCS.**

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

C. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Source connector. Use a Dataflow job to read from PubSub and write to GCS.

D. Deploy the PubSub Kafka connector to your on-prem Kafka cluster and configure PubSub as a Sink connector. Use a Dataflow job to read from PubSub and write to GCS.

Description:

### Kafka mirroring (MirrorMaker):

Kafka's mirroring feature makes it possible to maintain a replica of an existing Kafka cluster.

The solution specifically mentions mirroring and minimizing the use of Kafka Connect plugin.

#### Question 90:

Your team is responsible for developing and maintaining ETLs in your company. One of your Dataflow jobs is failing because of some errors in the input data, and you need to improve reliability of the pipeline (incl. being able to reprocess all failing data).

What should you do?

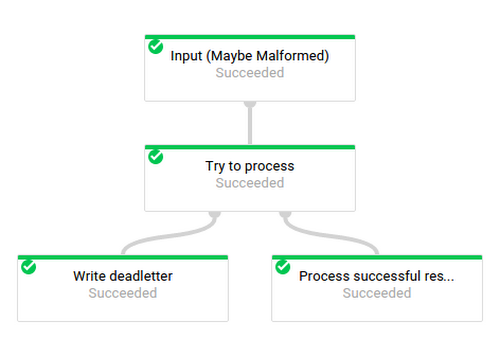
(Building and operationalizing data processing systems)

1. Add a filtering step to skip these types of errors in the future, extract erroneous rows from logs.
2. Add a try... catch block to your DoFn that transforms the data, extract erroneous rows from logs.
3. Add a try... catch block to your DoFn that transforms the data, write erroneous rows to PubSub directly from the DoFn.
4. **Add a try... catch block to your DoFn that transforms the data, use a sideOutput to create a PCollection that can be stored to PubSub later.**

Description:

If the failure is within the processing code of a DoFn, one way to handle this is to catch the exception, log an error, and then drop the input. The rest of the elements in the pipeline will be processed successfully, so progress can be made as normal. But just logging the elements isn’t ideal because it doesn’t provide an easy way to see these malformed inputs and reprocess them later.

**A better way to solve this would be to have a dead letter file where all of the failing inputs are written for later analysis and reprocessing. We can use a side output in Dataflow to accomplish this goal. For example:**



#### Question 91:

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?

(Building and operationalizing data processing systems)

1. Export Bigtable dump to GCS and run your analytical job on top of the exported files.
2. 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.
3. **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.**
4. Increase the size of your existing cluster twice and execute your analytics workload on your new resized cluster.

Description:

If you need a different consistency guarantee, Bigtable can also provide r**ead-your-writes consistency** when replication is enabled, which ensures that an application will never read data that is older than its most recent writes. **To gain read-your-writes consistency for a group of applications,** each application in the group must use an app profile that is configured for **single-cluster routing,** and all of the app profiles must route requests to the same cluster. You can use the instance's additional clusters at the same time for other purposes.

For some replication use cases, Bigtable can also provide strong consistency, which ensures that all of your applications see your data in the same state. To gain strong consistency, you use the single-cluster routing app-profile configuration for read-your-writes consistency that is described above, but you must not use the instance's additional clusters unless you need to fail over to a different cluster. Review the examples of replication settings to see if this is possible for your use case.

### Routing policies

Every app profile has a routing policy that controls which clusters handle incoming requests from your applications. Options for routing policies include the following:

* Single-cluster routing: Sends all requests to a single cluster that you specify.
* Multi-cluster routing to any cluster: Sends requests to the nearest available cluster in the instance.
* Cluster group routing: Sends requests to the nearest available cluster within a cluster group that you specify in the app profile settings.

#### Question 92:

You store historic data in Cloud Storage. You need to perform analytics on the historic data. You want to use a solution to detect invalid data entries and perform data transformations that will **not require programming or knowledge of SQL.**

What should you do?

A. Use Cloud Dataflow with Beam to detect errors and perform transformations.

**B. Use Cloud Dataprep with recipes to detect errors and perform transformations.**

C. Use Cloud Dataproc with a Hadoop job to detect errors and perform transformations.

D. Use federated tables in BigQuery with queries to detect errors and perform transformations.

Description:

Easy and powerful data preparation - Dataprep

With each gesture in the UI, Dataprep automatically suggests and predicts your next ideal data transformation. Once you’ve defined your sequence of transformations, Dataprep uses Dataflow or BigQuery under the hood, enabling you to process structured or unstructured datasets of any size with the **ease of clicks, not code.**

#### Question 93:

Your company needs to upload their historic data to Cloud Storage. **The security rules don't allow access from external IPs to their on-premises resources.** After an initial upload, they will add new data from existing on-premises applications every day. What should they do?

**A. Execute gsutil rsync from the on-premises servers.**

B. Use Cloud Dataflow and write the data to Cloud Storage.

C. Write a job template in Cloud Dataproc to perform the data transfer.

D. Install an FTP server on a Compute Engine VM to receive the files and move them to Cloud Storage.

Description:

Dataflow is on cloud is external; "don't allow access from external IPs to their on-premises resources" so no dataflow.

#### Question 94:

You need to copy millions of sensitive patient records from a relational database to BigQuery. The total size of the database is **10 TB.** You need to design a solution that is secure and time-efficient. What should you do?

A. Export the records from the database as an Avro file. Upload the file to GCS using gsutil, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

**B. Export the records from the database as an Avro file. Copy the file onto a Transfer Appliance and send it to Google, and then load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.**

C. Export the records from the database into a CSV file. Create a public URL for the CSV file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the CSV file into BigQuery using the BigQuery web UI in the GCP Console.

D. Export the records from the database as an Avro file. Create a public URL for the Avro file, and then use Storage Transfer Service to move the file to Cloud Storage. Load the Avro file into BigQuery using the BigQuery web UI in the GCP Console.

Description:

Without knowing the bandwidth, it is not possible to determine whether the upload can be completed within 7 days, as recommended by Google. So the safest and most performant way is to use Transfer Appliance.

### Deciding among Google's transfer options

Choosing a transfer option depends on your use case, as the following table shows.

|  |  |  |
| --- | --- | --- |
| **Where you're moving data from** | **Scenario** | **Suggested products** |
| Another cloud provider (for example, Amazon Web Services or Microsoft Azure) to Google Cloud | — | [Storage Transfer Service](https://cloud.google.com/storage-transfer/docs/overview) |
| Cloud Storage to Cloud Storage (two different buckets) | — | [Storage Transfer Service](https://cloud.google.com/storage-transfer/docs/overview) |
| Your private data center to Google Cloud | Enough bandwidth to meet your project deadline  for less than 1 TB of data | [gsutil](https://cloud.google.com/storage/docs/gsutil) |
| Your private data center to Google Cloud | Enough bandwidth to meet your project deadline  for more than 1 TB of data | [Storage Transfer Service](https://cloud.google.com/storage-transfer/docs/overview) for on-premises data |
| Your private data center to Google Cloud | Not enough bandwidth to meet your project deadline | [Transfer Appliance](https://cloud.google.com/transfer-appliance) |

#### Question 95:

You used Cloud Dataprep to create a recipe on a sample of data in a BigQuery table. You want to reuse this recipe on a daily upload of data with the same schema, after the load job with variable execution time completes. What should you do?

A. Create a cron schedule in Cloud Dataprep.

B. Create an App Engine cron job to schedule the execution of the Cloud Dataprep job.

C. Export the recipe as a Cloud Dataprep template, and create a job in Cloud Scheduler.

**D. Export the Cloud Dataprep job as a Cloud Dataflow template, and incorporate it into a Cloud Composer job.**

Description:

We’re happy to announce the latest release of Cloud Dataprep, which exposes orchestration APIs so you can integrate Cloud Dataprep within your schedulers or other orchestration solutions like Cloud Composer. This means you can expand your automation beyond Cloud Dataflow templates through direct integration in other tools to create repeatable data pipelines for your analytics and AI/ML initiatives—saving time and adding reliability. In addition, this API lets you use dynamic inputs and outputs through Cloud Dataprep variables or parameters—not available using Cloud Dataflow templates. As a result, you can re-use a single Cloud Dataprep flow to execute on a range of input/output values that are evaluated at runtime.

#### Question 96:

You have developed three data processing jobs.

1. One executes a Cloud Dataflow pipeline that transforms data uploaded to Cloud Storage and writes results to BigQuery.
2. The second ingests data from on-premises servers and uploads it to Cloud Storage.
3. The third is a Cloud Dataflow pipeline that gets information from third-party data providers and uploads the information to Cloud Storage.

You need to be able to schedule and monitor the execution of these three workflows and manually execute them when needed. What should you do?

**A. Create a Direct Acyclic Graph in Cloud Composer to schedule and monitor the jobs.**

B. Use Stackdriver Monitoring and set up an alert with a Webhook notification to trigger the jobs.

C. Develop an App Engine application to schedule and request the status of the jobs using GCP API calls.

D. Set up cron jobs in a Compute Engine instance to schedule and monitor the pipelines using GCP API calls.

Description:

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers.

Cloud composer is used to schedule the interdependent jobs

#### Question 97:

You have Cloud Functions written in Node.js that pull messages from Cloud Pub/Sub and send the data to BigQuery. You observe that the message processing rate on the Pub/Sub topic is orders of magnitude higher than anticipated, but there is no error logged in Stackdriver Log Viewer. What are the two most likely causes of this problem? (Choose two.)

A. Publisher throughput quota is too small.

B. Total outstanding messages exceed the 10-MB maximum.

**C. Error handling in the subscriber code is not handling run-time errors properly.**

D. The subscriber code cannot keep up with the messages.

**E. The subscriber code does not acknowledge the messages that it pulls.**

Description:

By not acknowleding the pulled message, this result in it be putted back in Cloud Pub/Sub, meaning the messages accumulate instead of being consumed and removed from Pub/Sub. The same thing can happen ig the subscriber maintains the lease on the message it receives in case of an error. This reduces the overall rate of processing because messages get stuck on the first subscriber. Also, errors in Cloud Function do not show up in Stackdriver Log Viewer if they are not correctly handled.

A: No problem with publisher rate as the observed result is a higher number of messages and not a lower number.

B: if messages exceed the 10MB maximum, they cannot be published.

D: Cloud Functions automatically scales so they should be able to keep up.

#### Question 98:

Your company has a hybrid cloud initiative. You have a complex data pipeline that moves data between cloud provider services and leverages services from each of the cloud providers. Which cloud-native service should you use to orchestrate the entire pipeline?

A. Cloud Dataflow

**B. Cloud Composer**

C. Cloud Dataprep

D. Cloud Dataproc

Description:

Cloud Composer is a fully managed workflow orchestration service that empowers you to author, schedule, and monitor pipelines that span across clouds and on-premises data centers.

Cloud Composer can help create workflows that connect data, processing, and services across clouds, giving you a unified data environment.

Built on the popular Apache Airflow open source project and operated using the Python programming language, Cloud Composer is free from lock-in and easy to use.

**Cloud Composer gives you the ability to connect your pipeline through a single orchestration tool whether your workflow Eves on-premises, in multiple clouds, or fully within GCP.** The ability to author, schedule, and monitor your workflows in a unified manner means you can break down the silos in your environment and focus less on infrastructure.

#### Question 99:

You use a dataset in BigQuery for analysis. You want to provide third-party companies with access to the same dataset. You need to keep the costs of data sharing **low** and ensure that the data is **current**. Which solution should you choose?

**A. Create an authorized view on the BigQuery table to control data access, and provide third-party companies with access to that view.**

B. Use Cloud Scheduler to export the data on a regular basis to Cloud Storage, and provide third-party companies with access to the bucket.

C. Create a separate dataset in BigQuery that contains the relevant data to share, and provide third-party companies with access to the new dataset.

D. Create a Cloud Dataflow job that reads the data in frequent time intervals, and writes it to the relevant BigQuery dataset or Cloud Storage bucket for third-party companies to use.

Description:

By creating an authorized view one assures that the data is current and avoids taking more storage space (and cost) in order to share a dataset.

BigQuery is a petabyte-scale analytics data warehouse that you can use to run SQL queries over vast amounts of data in near realtime.

Giving a view access to a dataset is also known as creating an authorized view in BigQuery’. An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query.

When you create the view, it must be created in a dataset separate from the source data queried by the view. Because you can assign access controls only at the dataset level, if the view is created in the same dataset as the source data, your data analysts would have access to both the view and the data.

#### Question 100:

A shipping company has live package-tracking data that is sent to an Apache Kafka stream in real time. This is then loaded into BigQuery. Analysts in your company want to query the tracking data in BigQuery to analyze geospatial trends in the lifecycle of a package. The table was originally created with **ingest-date partitioning.** Over time, the query processing time has increased. You need to implement a change that would improve query performance in BigQuery. What should you do?

A. Implement clustering in BigQuery on the ingest date column.

**B. Implement clustering in BigQuery on the package-tracking ID column.**

C. Tier older data onto Cloud Storage files, and leverage extended tables.

D. Re-create the table using data partitioning on the package delivery date.

Description:

In general, there are two typical usage patterns for clustering within a data warehouse:

Clustering on columns that have a very **high number of distinct values,** like **userId** or **transactionId**.

Clustering on multiple columns that are frequently used together. When clustering by multiple columns, the order of columns you specify is important. The order of the specified columns determines the sort order of the data. You can filter by any prefix of the clustering columns and get the benefits of clustering, like regionId, shopId and productId together; or regionId and shopId; or just regionId.

### BigQuery

Introduction to clustered tables

Both partitioning and clustering can improve performance and reduce query cost.

Use clustering under the following circumstances:

* You don't need strict cost guarantees before running the query.
* You need more **granularity** than partitioning alone allows. To get clustering benefits in addition to partitioning benefits, you can use the same column for both partitioning and clustering.
* Your queries commonly use filters or aggregation against multiple particular columns.
* The cardinality of the number of values in a column or group of columns is large.

Use partitioning under the following circumstances:

* You want to know query costs before a query runs. Partition pruning is done before the query runs, so you can get the query cost after partitioning pruning through a dry run. Cluster pruning is done when the query runs, so the cost is known only after the query finishes.
* You need partition-level management. For example, you want to set a partition expiration time, load data to a specific partition, or delete partitions.
* You want to specify how the data is partitioned and what data is in each partition. For example, you want to define time granularity or define the ranges used to partition the table for integer range partitioning.

Prefer clustering over partitioning under the following circumstances:

* Partitioning results in a small amount of data per partition (approximately less than 1 GB).
* Partitioning results in a large number of partitions beyond the limits on partitioned tables.
* Partitioning results in your mutation operations modifying most partitions in the table frequently (for example, every few minutes).

You can also combine partitioning with clustering. Data is first partitioned and then data in each partition is clustered by the clustering columns.

#### Question 101:

You are operating a Cloud Dataflow streaming pipeline. The pipeline aggregates events from a Cloud Pub/Sub subscription source, within a window, and sinks the resulting aggregation to a Cloud Storage bucket. The source has consistent throughput. You want to monitor an alert on behavior of the pipeline with Cloud Stackdriver to ensure that it is processing data. Which Stackdriver alerts should you create?

A. An alert based on a decrease of subscription/num\_undelivered\_messages for the source and a rate of change increase of instance/storage/ used\_bytes for the destination

**B. An alert based on an increase of subscription/num\_undelivered\_messages for the source and a rate of change decrease of instance/storage/ used\_bytes for the destination**

C. An alert based on a decrease of instance/storage/used\_bytes for the source and a rate of change increase of subscription/ num\_undelivered\_messages for the destination

D. An alert based on an increase of instance/storage/used\_bytes for the source and a rate of change decrease of subscription/ num\_undelivered\_messages for the destination

Description:

You would want to get alerted only if Pipeline fails & not if it is running fine. I think Option [B] is correct, because in event of Pipeline failure :

1) subscription/ num\_undelivered\_messages would pile up at a constant rate as the source has consistent throughput

2) instance/storage/ used\_bytes will get closer to zero. Hence need to monitor it's rate of change

#### Question 102:

You currently have a single on-premises Kafka cluster in a data center in the us-east region that is responsible for ingesting messages from IoT devices **globally**.

Because large parts of globe have poor internet connectivity, messages sometimes batch at the edge, come in all at once, and cause a spike in load on your

Kafka cluster. This is becoming difficult to manage and prohibitively expensive. What is the Google-recommended cloud native architecture for this scenario?

A. Edge TPUs as sensor devices for storing and transmitting the messages.

B. Cloud Dataflow connected to the Kafka cluster to scale the processing of incoming messages.

**C. An IoT gateway connected to Cloud Pub/Sub, with Cloud Dataflow to read and process the messages from Cloud Pub/Sub.**

D. A Kafka cluster virtualized on Compute Engine in us-east with Cloud Load Balancing to connect to the devices around the world.

Description:

Pubsub is global and dataflow can scale workers

Alterative to Kafka in google cloud native service is Pub/Sub and Dataflow punched with Pub/Sub is the google recommended option

#### Question 103:

You have a petabyte of analytics data and need to design a storage and processing platform for it. You must be able to perform data **warehouse-style** analytics on the data in Google Cloud and expose the dataset as files for **batch analysis** tools in **other cloud providers.** What should you do?

A. Store and process the entire dataset in BigQuery.

B. Store and process the entire dataset in Cloud Bigtable.

**C. Store the full dataset in BigQuery, and store a compressed copy of the data in a Cloud Storage bucket.**

D. Store the warm data as files in Cloud Storage, and store the active data in BigQuery. Keep this ratio as 80% warm and 20% active.

Description:

BigQuery for analytics processing and Cloud Storage for exposing the data as files

#### Question 104:

You use BigQuery as your centralized analytics platform. New data is loaded every day, and an ETL pipeline modifies the original data and prepares it for the final users. This ETL pipeline is regularly modified and can generate errors, but sometimes the errors are detected only after 2 weeks. You need to provide a method to recover from these errors, and your backups should be optimized for storage costs. How should you organize your data in BigQuery and store your backups?

A. Organize your data in a single table, export, and compress and store the BigQuery data in Cloud Storage.

**B. Organize your data in separate tables for each month, and export, compress, and store the data in Cloud Storage.**

C. Organize your data in separate tables for each month, and duplicate your data on a separate dataset in BigQuery.

D. Organize your data in separate tables for each month, and use snapshot decorators to restore the table to a time prior to the corruption.

Description:

Store your data in different tables for specific time periods. This method ensures that you will need to restore only a subset of data to a new table, rather than a whole dataset.

#### Question 106:

You are building a new application that you need to collect data from in a **scalable** way. Data arrives continuously from the application throughout the day, and you expect to generate approximately 150 GB of JSON data per day by the end of the year. Your requirements are:

✑ Decoupling producer from consumer

✑ Space and cost-efficient storage of the raw ingested data, which is to be stored indefinitely

✑ Near real-time SQL query

✑ Maintain at least 2 years of historical data, which will be queried with SQL

Which pipeline should you use to meet these requirements?

A. Create an application that provides an API. Write a tool to poll the API and write data to Cloud Storage as gzipped JSON files.

B. Create an application that writes to a Cloud SQL database to store the data. Set up periodic exports of the database to write to Cloud Storage and load into BigQuery.

C. Create an application that publishes events to Cloud Pub/Sub, and create Spark jobs on Cloud Dataproc to convert the JSON data to Avro format, stored on HDFS on Persistent Disk.

**D. Create an application that publishes events to Cloud Pub/Sub, and create a Cloud Dataflow pipeline that transforms the JSON event payloads to Avro, writing the data to Cloud Storage and BigQuery.**

Description:

Real-time made real easy:

Adopt simple ingestion for complex events

Ingest and analyze hundreds of m\illions of events per second from applications or devices virtually anywhere on the globe with Pub/Sub. Or directly stream millions of events per second into your data warehouse for SQL-based analysis with **BigQuery's streaming API.**

Unify stream and batch processing without lock-in

**Unify streaming and batch data analysis with equal ease and build cohesive data pipelines with Dataflow.** Dataflow ensures exactly-once processing, making your streaming pipelines more reliable and consistent for mission-critical applications. Data engineers can reuse code through Dataflow’s open source SDK, Apache Beam, which provides pipeline portability for hybrid or multi-cloud environments.

#### Question 107:

You have several Spark jobs that run on a Cloud Dataproc cluster on a schedule. Some of the jobs run in sequence, and some of the jobs run concurrently. You need to automate this process. What should you do?

A. Create a Cloud Dataproc Workflow Template

B. Create an initialization action to execute the jobs

**C. Create a Directed Acyclic Graph in Cloud Composer**

D. Create a Bash script that uses the Cloud SDK to create a cluster, execute jobs, and then tear down the cluster

Description:

After further consideration C. Workflows will not be able to handle sequential jobs as it has only time based triggers.

1) Create a Dataproc workflow template that runs a Spark PI job

2) Create an Apache Airflow DAG that Cloud Composer will use to start the workflow at a specific time.

#### Question 108:

Data Analysts in your company have the **Cloud IAM Owner** role assigned to them in their projects to allow them to work with multiple GCP products in their projects. Your organization requires that all BigQuery data access logs be retained for **6 months**. You need to ensure that only audit personnel in your company can access the data access logs for all projects. What should you do?

A. Enable data access logs in each Data Analyst's project. Restrict access to Stackdriver Logging via Cloud IAM roles.

B. Export the data access logs via a project-level export sink to a Cloud Storage bucket in the Data Analysts' projects. Restrict access to the Cloud Storage bucket.

C. Export the data access logs via a project-level export sink to a Cloud Storage bucket in a newly created projects for audit logs. Restrict access to the project with the exported logs.

**D. Export the data access logs via an aggregated export sink to a Cloud Storage bucket in a newly created project for audit logs. Restrict access to the project that contains the exported logs.**

Description:

### Configure aggregated sinks

**Sinks** control how Cloud Logging routes logs. Using sinks, you can route some or all of your logs to supported destinations.

You can also create aggregated sinks to specify that your sink combine and route log entries from the Google Cloud resources contained by an organization or folder. For instructions, see **Configure aggregated sinks.**

**Aggregated log sink** will create a single sink for all projects, the destination can be a google cloud storage, pub/sub topic, bigquery table or a cloud logging bucket. without aggregated sink this will be required to be done for each project individually which will be cumbersome.

Obviously, the auditor needs to check all projects accessed by data analyst which is not project-level, a higher level like folder or organization level, this can only be done via the aggregate sink.

#### Question 109:

You are operating a streaming Cloud Dataflow pipeline. Your engineers have a new version of the pipeline with a different windowing algorithm and triggering strategy. You want to update the running pipeline with the new version. You want to ensure that no data is lost during the update. What should you do?

A. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to the existing job name

B. Update the Cloud Dataflow pipeline inflight by passing the --update option with the --jobName set to a new unique job name

C. Stop the Cloud Dataflow pipeline with the Cancel option. Create a new Cloud Dataflow job with the updated code

**D. Stop the Cloud Dataflow pipeline with the Drain option. Create a new Cloud Dataflow job with the updated code**

Description:

#### Question 110:

You want to migrate an on-premises Hadoop system to Cloud Dataproc. Hive is the primary tool in use, and the data format is Optimized Row Columnar (ORC).

All ORC files have been successfully copied to a Cloud Storage bucket. **You need to replicate some data to the cluster's local Hadoop Distributed File System (HDFS) to maximize performance.** What are two ways to start using Hive in Cloud Dataproc? (Choose two.)

A. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to HDFS. Mount the Hive tables locally.

**B. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to any node of the Dataproc cluster. Mount the Hive tables locally.**

**C. Run the gsutil utility to transfer all ORC files from the Cloud Storage bucket to the master node of the Dataproc cluster. Then run the Hadoop utility to copy them do HDFS. Mount the Hive tables from HDFS.**

D. Leverage Cloud Storage connector for Hadoop to mount the ORC files as external Hive tables. Replicate external Hive tables to the native ones.

E. Load the ORC files into BigQuery. Leverage BigQuery connector for Hadoop to mount the BigQuery tables as external Hive tables. Replicate external Hive tables to the native ones.

Description:

HDFS lies on datanode, data on masternode needs to be copied on datanode

B for managed hive table option, C for external hive table

#### Question 111:

You are implementing several batch jobs that must be executed on a schedule. These jobs have many interdependent steps that must be executed in a specific order. Portions of the jobs involve executing shell scripts, running Hadoop jobs, and running queries in BigQuery. The jobs are expected to run for many minutes up to several hours. If the steps fail, they must be retried a fixed number of times. Which service should you use to manage the execution of these jobs?

A. Cloud Scheduler

B. Cloud Dataflow

C. Cloud Functions

**D. Cloud Composer**

Description:

the main point is that Cloud Composer should be used when there is inter-dependencies between the job, e.g. we need the output of a job to start another whenever the first finished, and use dependencies coming from first job.

All interdependent tasks need to be run through cloud composer whereas small/adhoc tasks need to be run via scheduler.

Cloud Scheduler is a fully managed enterprise-grade cron job scheduler.

#### Question 112:

You want to build a managed Hadoop system as your data lake. The data transformation process is composed of a series of Hadoop jobs executed in sequence.

To accomplish the design of separating storage from compute, you decided to use the Cloud Storage connector to store all input data, output data, and intermediary data. However, you noticed that one Hadoop job runs very slowly with Cloud Dataproc, when compared with the on-premises bare-metal Hadoop environment (8-core nodes with 100-GB RAM). Analysis shows that this particular Hadoop job is disk I/O intensive. You want to resolve the issue. What should you do?

A. Allocate sufficient memory to the Hadoop cluster, so that the intermediary data of that particular Hadoop job can be held in memory

**B. Allocate sufficient persistent disk space to the Hadoop cluster, and store the intermediate data of that particular Hadoop job on native HDFS**

C. Allocate more CPU cores of the virtual machine instances of the Hadoop cluster so that the networking bandwidth for each instance can scale up

D. Allocate additional network interface card (NIC), and configure link aggregation in the operating system to use the combined throughput when working with Cloud Storage

Description:

Local HDFS storage is a good option if:

* Your jobs require a lot of metadata operations—for example, you have thousands of partitions and directories, and each file size is relatively small.
* You modify the HDFS data frequently or you rename directories. (Cloud Storage objects are immutable, so renaming a directory is an expensive operation because it consists of copying all objects to a new key and deleting them afterwards.)
* You heavily use the append operation on HDFS files.
* You have workloads that involve heavy I/O. For example, you have a lot of partitioned writes, such as the following:
* spark.read().write.partitionBy(...).parquet("gs://")
* You have I/O workloads that are especially sensitive to latency. For example, you require single-digit millisecond latency per storage operation.

#### Question 113:

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

Description:

If you create a Dataproc cluster with internal IP addresses only, attempts to access the Internet in an initialization action will fail unless you have configured routes to direct the traffic through a NAT or a VPN gateway. Without access to the Internet, you can enable Private Google Access, and place job dependencies in Cloud Storage; cluster nodes can download the dependencies from Cloud Storage from internal IPs.

#### Question 114:

You work for a mid-sized enterprise that needs to move its operational system transaction data from an on-premises database to GCP. The database is about 20

TB in size. Which database should you choose?

**A. Cloud SQL**

B. Cloud Bigtable

C. Cloud Spanner

D. Cloud Datastore

Description:

Cloud SQL storage limits

* Dedicated core: Up to 64 TB.
* Shared core: Up to 3 TB

#### Question 115:

You have data pipelines running on BigQuery, Cloud Dataflow, and Cloud Dataproc. You need to perform health checks and monitor their behavior, and then notify the team managing the pipelines if they fail. You also need to be able to work across multiple projects. Your preference is to use managed products of features of the platform. What should you do?

**A. Export the information to Cloud Stackdriver, and set up an Alerting policy**

B. Run a Virtual Machine in Compute Engine with Airflow, and export the information to Stackdriver

C. Export the logs to BigQuery, and set up App Engine to read that information and send emails if you find a failure in the logs

D. Develop an App Engine application to consume logs using GCP API calls, and send emails if you find a failure in the logs

Description:

Monitoring does not only provide you with access to Dataflow-related metrics, but also lets you to create alerting policies and dashboards so you can chart time series of metrics and choose to be notified when these metrics reach specified values.

#### Question 116:

Suppose you have a table that includes a nested column called "city" inside a column called "person", but when you try to submit the following query in BigQuery, it gives you an error.

SELECT person FROM `project1.example.table1` WHERE city = "London"

How would you correct the error?

A. Add ", UNNEST(person)" before the WHERE clause.

**B. Change "person" to "person.city".**

C. Change "person" to "city.person".

D. Add ", UNNEST(city)" before the WHERE clause.

Description:

The qestion is about nest nor repeated. Nested doesn't need unnest, while repeated do.

#### Question 117:

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

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

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

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

D. The only compression option available is GZIP.

Description:

When you export data from BigQuery, note the following:

* You cannot export table data to a local file, to Google Sheets, or to Google Drive. The only supported export location is Cloud Storage. For information on saving query results, see Downloading and saving query results.
* You can export up to 1 GB of table data to a single file. If you are exporting more than 1 GB of data, use a wildcard to export the data into multiple files. When you export data to multiple files, the size of the files will vary.
* You cannot export nested and repeated data in CSV format. Nested and repeated data is supported for Avro and JSON exports.
* When you export data in JSON format, INT64 (integer) data types are encoded as JSON strings to preserve 64-bit precision when the data is read by other systems.
* You cannot export data from multiple tables in a single export job.
* You cannot choose a compression type other than GZIP when you export data using the Cloud Console or the classic BigQuery web UI.

#### Question 118:

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

Description:

#### Question 119:

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.**

Description:

A. By default, a query's results are not cached. (False)

B. BigQuery caches query results for 48 hours. (False - 24 hours)

C. Query results are cached even if you specify a destination table. False

When a destination table is specified in the job configuration, the Cloud Console, the bq command-line tool, or the API, the query results are not cached.

https://cloud.google.com/bigquery/docs/cached-results#cache-exceptions

D. There is no charge for a query that retrieves its results from cache. (True)

#### Question 120:

Which of these sources can you not load data into BigQuery from?

A. File upload

B. Google Drive

C. Google Cloud Storage

**D. Google Cloud SQL**

Description:

Only Cloud sql is not available to be used for loading data directly.

On console we can see that we can use file, bigtable, drive and GS.

#### Question 121:

How would you query specific partitions in a BigQuery table?

A. Use the DAY column in the WHERE clause

B. Use the EXTRACT(DAY) clause

**C. Use the \_\_PARTITIONTIME pseudo-column in the WHERE clause**

D. Use DATE BETWEEN in the WHERE clause

Description:

Logical partition column is queried using \_partitiontime

Ingestion time partitioning

When you create a table partitioned by ingestion time, BigQuery automatically assigns rows to partitions based on the time when BigQuery ingests the data. You can choose hourly, daily, monthly, or yearly granularity for the partitions. Partitions boundaries are based on UTC time.

An ingestion-time partitioned table has a pseudocolumn named \_PARTITIONTIME. The value of this column is the ingestion time for each row, truncated to the partition boundary (such as hourly or daily). For example, suppose that you create an ingestion-time partitioned table with hourly partitioning and send data at the following times:

#### Question 122:

Which SQL keyword can be used to reduce the number of columns processed by BigQuery?

A. BETWEEN

B. WHERE

**C. SELECT**

D. LIMIT

Description:

SELECT is for few columns among all columns,

LIMIT is to have a limited number of rows as result but the table is completely scanned

SELECT 是针对所有列中的少数列，LIMIT 是结果的行数有限，但表已被完全扫描

#### Question 123:

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.

Description:

#### Question 124:

What are two methods that can be used to denormalize tables in BigQuery?

A. 1) Split table into multiple tables; 2) Use a partitioned table

**B. 1) Join tables into one table; 2) Use nested repeated fields**

C. 1) Use a partitioned table; 2) Join tables into one table

D. 1) Use nested repeated fields; 2) Use a partitioned table

Description:

Denormalizing data involves grouping together several tables linked by references, into a single table, statically performing the appropriate join operations.

**GCP recommends to denormalize a dimension table larger than 10 GB,**

unless you get tangible evidence that the costs of data manipulation and UPDATE and DELETE operations outweigh the benefits of optimal queries.

Denormalization says join tables to create one table and nested repeated fields to make query run faster

#### Question 125:

Which of these is not a supported method of putting data into a partitioned table?

A. If you have existing data in a separate file for each day, then create a partitioned table and upload each file into the appropriate partition.

B. Run a query to get the records for a specific day from an existing table and for the destination table, specify a partitioned table ending with the day in the format "$YYYYMMDD".

C. Create a partitioned table and stream new records to it every day.

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

Description:

#### Question 126:

Which of these numbers are adjusted by a neural network as it learns from a training dataset (select 2 answers)?

(Building and operationalizing data processing systems)

1. **Weights**
2. **Biases**
3. Continuous features
4. Input values

Description:

The two are adjust to create a perfect model

#### Question 127:

Which TensorFlow function can you use to configure a categorical column if you don't know all of the possible values for that column?

A. categorical\_column\_with\_vocabulary\_list

**B. categorical\_column\_with\_hash\_bucket**

C. categorical\_column\_with\_unknown\_values

D. sparse\_column\_with\_keys

Description:

**vocabulary list** is used when column values are known(incremental values are added as categorical columns starting from 0)

**hash\_bucket** is used when you don’t know the values

#### Question 128: Amazon?

A company ingests a large set of clickstream data in nested JSON format from different sources and stores it in Amazon S3. Data analysts need to analyze this data in combination with data stored in an Amazon Redshift cluster. Data analysts want to build a cost-effective and automated solution for this need.

Which solution meets these requirements?

(Building and operationalizing data processing systems)

1. Use Apache Spark SQL on Amazon EMR to convert the clickstream data to a tabular format. Use the Amazon Redshift COPY command to load the data into the Amazon Redshift cluster.
2. Use AWS Lambda to convert the data to a tabular format and write it to Amazon S3. Use the Amazon Redshift COPY command to load the data into the Amazon Redshift cluster.
3. **Use the Relationalize class in an AWS Glue ETL job to transform the data and write the data back to Amazon S3. Use Amazon Redshift Spectrum to create external tables and join with the internal tables.**
4. Use the Amazon Redshift COPY command to move the clickstream data directly into new tables in the Amazon Redshift cluster.

Description:

#### Question 129:

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

A. An hourly watermark

B. An event time trigger

C. The with Allowed Lateness method

**D. A processing time trigger**

Description:

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

**Processing time triggers.** These triggers operate on the processing time the time when the data element is processed at any given stage in the pipeline.

**Event time triggers.** These triggers operate on the event time, as indicated by the timestamp on each data element. Beams default trigger is event time-based.

#### Question 130:

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

Description:

In Google Cloud dataflow SDK, you can use the ParDo to extract only a customer name of each element in your PCollection.

come on.. this would never be in the exame

#### Question 131:

Does Dataflow process batch data pipelines or streaming data pipelines?

(Building and operationalizing data processing systems)

1. Only Batch Data Pipelines
2. **Both Batch and Streaming Data Pipelines**
3. Only Streaming Data Pipelines
4. None of the above

Description:

#### Question 132:

The \_\_\_\_\_\_\_\_\_ for Cloud Bigtable makes it possible to use Cloud Bigtable in a Cloud Dataflow pipeline.

**A. Cloud Dataflow connector**

B. DataFlow SDK

C. BiqQuery API

D. BigQuery Data Transfer Service

Description:

#### Question 133:

The Dataflow SDKs have been recently transitioned into which Apache service?

A. Apache Spark

B. Apache Hadoop

C. Apache Kafka

**D. Apache Beam**

Description:

#### Question 134:

Which Java SDK class can you use to run your Dataflow programs locally?

(Building and operationalizing data processing systems)

LocalRunner

**DirectPipelineRunner**

MachineRunner

LocalPipelineRunner

Description:

DirectPipelineRunner allows you to execute operations in the pipeline directly, without any optimization. Useful for small local execution and tests

#### Question 135:

Which of the following is NOT one of the three main types of triggers that Dataflow supports?

(Building and operationalizing data processing systems)

1. **Trigger based on element size in bytes**
2. Trigger that is a combination of other triggers
3. Trigger based on element count
4. Trigger based on time

Description:

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

#### Question 136:

What Dataflow concept determines when a Window's contents should be output based on certain criteria being met?

A. Sessions

B. OutputCriteria

C. Windows

**D. Triggers**

Description:

Triggers control when the elements for a specific key and window are output. As elements arrive, they are put into one or more windows by a Window transform and its associated WindowFn, and then passed to the associated Trigger to determine if the Windows contents should be output.

#### Question 137:

When running a pipeline that has a BigQuery source, on your local machine, you continue to get permission denied errors. What could be the reason for that?

(Building and operationalizing data processing systems)

1. **Your gcloud does not have access to the BigQuery resources**
2. BigQuery cannot be accessed from local machines
3. You are missing gcloud on your machine
4. Pipelines cannot be run locally

Description:

#### Question 138:

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

A. Continuously retrain the model on just the new data.

**B. Continuously retrain the model on a combination of existing data and the new data.**

C. Train on the existing data while using the new data as your test set.

D. Train on the new data while using the existing data as your test set.

Description:

#### Question 139:

You are creating a model to predict housing prices. Due to budget constraints, you must run it on a single resource-constrained virtual machine. Which learning algorithm should you use?

**A. Linear regression**

B. Logistic classification

C. Recurrent neural network

D. Feedforward neural network

Description:

A as Supervised learning using Regression can help build a model to predict house prices.

Option B is wrong as Classification would not help to solve the problem.

Options C & D are wrong as they would need more resources.

#### Question 140:

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.**

Description:

#### Question 141:

Business owners at your company have given you a database of bank transactions. Each row contains the user ID, transaction type, transaction location, and transaction amount. They ask you to investigate what type of machine learning can be applied to the data.

Which three machine learning applications can you use? (Choose three.)

(Operationalizing machine learning models)

1. Supervised learning to determine which transactions are most likely to be fraudulent.
2. **Unsupervised learning to determine which transactions are most likely to be fraudulent.**
3. **Clustering to divide the transactions into N categories based on feature similarity.**
4. **Supervised learning to predict the location of a transaction.**
5. Reinforcement learning to predict the location of a transaction.
6. F. Unsupervised learning to predict the location of a transaction.

Description:

#### Question 142:

Your company has hired a new data scientist who wants to perform complicated analyses across very large datasets stored in Google Cloud Storage and in a

Cassandra cluster on Google Compute Engine. The scientist primarily wants to create labelled data sets for machine learning projects, along with some visualization tasks. She reports that her laptop is not powerful enough to perform her tasks and it is slowing her down. You want to help her perform her tasks.

What should you do?

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

B. Grant the user access to Google Cloud Shell.

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

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

Description:

Datalab provides Jupyter for this kind of work

#### Question 143:

You are building a model to predict whether or not it will rain on a given day. You have thousands of input features and want to see if you can improve training speed by removing some features while having a minimum effect on model accuracy. What can you do?

A. Eliminate features that are highly correlated to the output labels.

**B. Combine highly co-dependent features into one representative feature.**

C. Instead of feeding in each feature individually, average their values in batches of 3.

D. Remove the features that have null values for more than 50% of the training records.

Description:

A: correlated to output means that feature can contribute a lot to the model. so not a good idea.

C: you need to run with almost same number, but you will iterate twice, once for averaging and second time to feed the averaged value.

D: removing features even if it 50% nulls is not good idea, unless you prove that it is not at all correlated to output. But this is nowhere so can remove.

#### Question 144:

Your company is performing data preprocessing for a learning algorithm in Google Cloud Dataflow. Numerous data logs 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 PCollection represents a single row in the table.

Description:

BigQueryIO.read.from() directly reads the whole table from BigQuery.

This function exports the whole table to temporary files in Google Cloud Storage, where it will later be read from.

This requires almost no computation, as it only performs an export job, and later Dataflow reads from GCS (not from BigQuery).

BigQueryIO.read.fromQuery() executes a query and then reads the results received after the query execution. Therefore, this function is more time-consuming, given that it requires that a query is first executed (which will incur in the corresponding economic and computational costs).

#### Question 145:

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?

1. Use a row key of the form <timestamp>.
2. Use a row key of the form .
3. Use a row key of the form #.
4. **Use a row key of the form >#<sensorid>#<timestamp>.**

Description:

Best practices of bigtable states that rowkey should not be only timestamp or have timestamp at starting. It’s better to have sensorid and timestamp as rowkey.

#### Question 146:

You are training a spam classifier. You notice that you are overfitting the training data.

Which three actions can you take to resolve this problem? (Choose three.)

(Operationalizing machine learning models)

**Get more training examples**

Reduce the number of training examples

**Use a smaller set of features**

Use a larger set of features

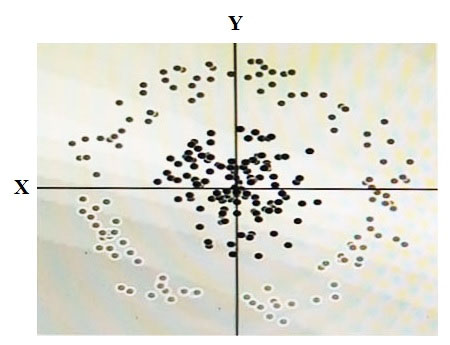
**Increase the regularization parameters**

Decrease the regularization parameters

Description:

#### Question 147:

You have some data, which is shown in the graphic below. The two dimensions are X and Y, and the shade of each dot represents what class it is. You want to classify this data accurately using a linear algorithm. To do this you need to add a synthetic feature. What should the value of that feature be?

​​

**A. X^2+Y^2**

B. X^2

C. Y^2

D. cos(X)

Description:

#### Question 148:

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 "˜none' using a Cloud Dataprep job.

D. Use Cloud Dataflow to find null values in sample source data. Convert all nulls to 0 using a custom script.

Description:

Dataprep is the tool. A or B.

Since they need to have a real-valued cannot be null N/A or empty, have to be “0”, so it has to be B.

#### Question 149:

You are developing an application that uses a **recommendation engine** on Google Cloud. Your solution should display new videos to customers based on past views. Your solution needs to generate labels for the entities in videos that the customer has viewed. Your design must be able to provide very fast filtering suggestions based on data from other customer preferences on several TB of data. What should you do?

A. Build and train a complex classification model with Spark MLlib to generate labels and filter the results. Deploy the models using Cloud Dataproc. Call the model from your application.

B. Build and train a classification model with Spark MLlib to generate labels. Build and train a second classification model with Spark MLlib to filter results to match customer preferences. Deploy the models using Cloud Dataproc. Call the models from your application.

**C. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud Bigtable, and filter the predicted labels to match the user's viewing history to generate preferences.**

D. Build an application that calls the Cloud Video Intelligence API to generate labels. Store data in Cloud SQL, and join and filter the predicted labels to match the user's viewing history to generate preferences.

Description:

1. Rather than building a new model - it is better to use Google provide APIs, here - Google Video Intelligence. So option A and B rules out

2. Between SQL and Bigtable - Bigtable is the better option as Bigtable support row-key filtering. Joining the filters is not required.

#### Question 150:

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

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

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. 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.

Description:

Entity analysis -> Identify entities within documents receipts, invoices, and contracts and label them by types such as date, person, contact information, organization, location, events, products, and media.

Sentiment analysis -> Understand the overall opinion, feeling, or attitude sentiment expressed in a block of text.

-- Avoid Custom models

#### Question 151:

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.

Description:

L1 where some features are more influential, L2 when everyone contributes

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.

#### Question 152:

You work for a bank. You have a labelled dataset that contains information on already granted loan application and whether these applications have been defaulted. You have been asked to train a model to predict default rates for credit applicants.

What should you do?

A. Increase the size of the dataset by collecting additional data.

**B. Train a linear regression to predict a credit default risk score.**

C. Remove the bias from the data and collect applications that have been declined loans.

D. Match loan applicants with their social profiles to enable feature engineering.

Description:

#### Question 153:

You are creating a new pipeline in Google Cloud to stream IoT data from Cloud Pub/Sub through Cloud Dataflow to BigQuery. While previewing the data, you notice that roughly 2% of the data appears to be corrupt. You need to modify the Cloud Dataflow pipeline to filter out this corrupt data. What should you do?

A. Add a SideInput that returns a Boolean if the element is corrupt.

**B. Add a ParDo transform in Cloud Dataflow to discard corrupt elements.**

C. Add a Partition transform in Cloud Dataflow to separate valid data from corrupt data.

D. Add a GroupByKey transform in Cloud Dataflow to group all of the valid data together and discard the rest.

Description:

ParDo is a Beam **transform** for generic parallel processing. ParDo is useful for common data processing operations, including:

a. **Filtering** a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection, or discard it.

b. **Formatting** or type-converting each element in a data set.

c. **Extracting** parts of each element in a data set.

d. **Performing** computations on each element in a data set.

A does not help

C Partition is a Beam transform for PCollection objects that store the same data type. Partition splits a single PCollection into a fixed number of smaller collections. Again, does not help

D GroupByKey is a Beam transform for processing collections of key/value pairs. GroupByKey is a good way to aggregate data that has something in common

#### Question 154:

You need to create a data pipeline that copies time-series transaction data so that it can be queried from within BigQuery by your data science team for analysis.

Every hour, thousands of transactions are updated with a new status. The size of the intitial dataset is 1.5 PB, and it will grow by 3 TB per day. The data is heavily structured, and your data science team will build machine learning models based on this data. You want to maximize performance and usability for your data science team. Which two strategies should you adopt? (Choose two.)

**A. Denormalize the data as much as possible.**

B. Preserve the structure of the data as much as possible.

C. Use BigQuery UPDATE to further reduce the size of the dataset.

**D. Develop a data pipeline where status updates are appended to BigQuery instead of updated.**

E. Copy a daily snapshot of transaction data to Cloud Storage and store it as an Avro file. Use BigQuery's support for external data sources to query.

Description:

A: Denormalization increases query speed for tables with billions of rows because BigQuery's performance degrades when doing JOINs on large tables, but with a denormalized data structure, you don't have to use JOINs, since all of the data has been combined into one table.

Denormalization also makes queries simpler because you do not have to use JOIN clauses.

https://cloud.google.com/solutions/bigquery-data-warehouse#denormalizing\_data

Denormalization will help in performance by reducing query time, update is not good with big query.

And append has better performance than Update.

#### Question 155:

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

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

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

Description:

POC is a small-scale experiments

#### Question 156:

You are working on a niche product in the image recognition domain. Your team has developed a model that is dominated by custom C++ TensorFlow ops your team has implemented. These ops are used inside your main training loop and are performing bulky matrix multiplications. It currently takes up to several days to train a model. You want to decrease this time significantly and keep the cost low by using an accelerator on Google Cloud.

A. Use Cloud TPUs without any additional adjustment to your code.

B. Use Cloud TPUs after implementing GPU kernel support for your customs ops.

C. Use Cloud GPUs after implementing GPU kernel support for your customs ops.

**D. Stay on CPUs, and increase the size of the cluster you're training your model on.**

Description:

<https://www.examtopics.com/discussions/google/view/17236-exam-professional-data-engineer-topic-1-question-127/>

CPUs:

• Quick prototyping that requires maximum flexibility

• Simple models that do not take long to train

• Small models with small effective batch sizes

• Models that are dominated by custom TensorFlow operations written in C++

• Models that are limited by available I/O or the networking bandwidth of the host system

#### Question 157:

You work on a regression problem in a natural language processing domain, and you have 100M labeled exmaples in your dataset. You have randomly shuffled your data and split your dataset into train and test samples (in a 90/10 ratio). After you trained the neural network and evaluated your model on a test set, you discover that the root-mean-squared error (RMSE) of your model is twice as high on the train set as on the test set.

How should you improve the performance of your model?

(Operationalizing machine learning models)

1. Increase the share of the test sample in the train-test split.
2. Try to collect more data and increase the size of your dataset.
3. Try out regularization techniques (e.g., dropout of batch normalization) to avoid overfitting.
4. **Increase the complexity of your model by, e.g., introducing an additional layer or increase sizing the size of vocabularies or n-grams used.**

Description:

"Training error is small and test error is big" is an indication of overfitting.

#### Question 158:

A data scientist has created a BigQuery ML model and asks you to create an ML pipeline to serve predictions. You have a REST API application with the requirement to serve predictions for an individual user ID with **latency under 100 milliseconds.** You use the following query to generate predictions: SELECT predicted\_label, user\_id FROM ML.PREDICT (MODEL "˜dataset.model', table user\_features). How should you create the ML pipeline?

A. Add a WHERE clause to the query, and grant the BigQuery Data Viewer role to the application service account.

B. Create an Authorized View with the provided query. Share the dataset that contains the view with the application service account.

C. Create a Cloud Dataflow pipeline using BigQueryIO to read results from the query. Grant the Dataflow Worker role to the application service account.

**D. Create a Cloud Dataflow pipeline using BigQueryIO to read predictions for all users from the query. Write the results to Cloud Bigtable using BigtableIO. Grant the Bigtable Reader role to the application service account so that the application can read predictions for individual users from Cloud Bigtable.**

Description:

The key reason for pick D is the 100ms requirement. Bigtable provides lowest latency

#### Question 159:

You work for an advertising company, and you've developed a Spark ML model to predict click-through rates at advertisement blocks. You've been developing everything at your on-premises data center, and now your company is migrating to Google Cloud. Your data center will be closing soon, so a rapid lift-and-shift migration is necessary. However, the data you've been using will be migrated to migrated to BigQuery. You periodically retrain your Spark ML models, so you need to migrate existing training pipelines to Google Cloud. What should you do?

A. Use Cloud ML Engine for training existing Spark ML models

B. Rewrite your models on TensorFlow, and start using Cloud ML Engine

**C. Use Cloud Dataproc for training existing Spark ML models, but start reading data directly from BigQuery**

D. Spin up a Spark cluster on Compute Engine, and train Spark ML models on the data exported from BigQuery

Description:

Only C make sense for rapid Lift and Shift

A Cloud Dataproc cluster has the Spark components, including Spark ML, installed.

#### Question 160:

You work for a global shipping company. You want to train a model on 40 TB of data to predict which ships in each geographic region are likely to cause delivery delays on any given day. The model will be based on multiple attributes collected from multiple sources. Telemetry data, including location in GeoJSON format, will be pulled from each ship and loaded every hour.

You want to have a dashboard that shows how many and which ships are likely to cause delays within a region. You want to use a storage solution that has native functionality for prediction and geospatial processing.

Which storage solution should you use?

(Operationalizing machine learning models)

**A. BigQuery**

B. Cloud Bigtable

C. Cloud Datastore

D. Cloud SQL for PostgreSQL

Description:

Geospatial and ML functionality is with bigquery

GeoJson + Native functionality for prediction -> BigQuery

#### Question 161:

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

**A. Perform hyperparameter tuning**

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

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

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

Description:

AUC is scale-invariant. It measures how well predictions are ranked, rather than their absolute values.

Preprocessing/scaling should be done with input features, instead of predictions (output)

#### 

#### Question 162:

You need to choose a database to store time series CPU and memory usage for millions of computers. You need to store this data in one-second interval samples. Analysts will be performing real-time, ad hoc analytics against the database. You want to avoid being charged for every query executed and ensure that the schema design will allow for future growth of the dataset. Which database and data model should you choose?

A. Create a table in BigQuery, and append the new samples for CPU and memory to the table

B. Create a wide table in BigQuery, create a column for the sample value at each second, and update the row with the interval for each second

**C. Create a narrow table in Cloud Bigtable with a row key that combines the Computer Engine computer identifier with the sample time at each second**

D. Create a wide table in Cloud Bigtable with a row key that combines the computer identifier with the sample time at each minute, and combine the values for each second as column data.

Description:

A tall and narrow table has a small number of events per row, which could be just one event, whereas a short and wide table has a large number of events per row. As explained in a moment, tall and narrow tables are best suited for time-series data.

For time series, you should generally use tall and narrow tables. This is for two reasons: Storing one event per row makes it easier to run queries against your data. Storing many events per row makes it more likely that the total row size will exceed the recommended maximum (see Rows can be big but are not infinite).

Bigtable is best suited to the following scenarios: time-series data (e.g. CPU and memory usage over time for multiple servers), financial data (e.g. transaction histories, stock prices, and currency exchange rates), and IoT (Internet of Things) use cases.

#### 

#### Question 163:

Which of these operations can you perform from the BigQuery Web UI?

(Operationalizing machine learning models)

1. Upload a file in SQL format.
2. **Load data with nested and repeated fields.**
3. Upload a 20 MB file.
4. Upload multiple files using a wildcard.

Description:

Wildcards and comma-separated lists are not supported when you load files from a local data source. Files must be loaded individually.

When using the Cloud Console, files loaded from a local data source cannot exceed 10 MB. For larger files, load the file from Cloud Storage

#### Question 164:

Why do you need to split a machine learning dataset into training data and test data?

(Operationalizing machine learning models)

1. So you can try two different sets of features
2. **To make sure your model is generalized for more than just the training data**
3. To allow you to create unit tests in your code
4. So you can use one dataset for a wide model and one for a deep model

Description:

#### Question 165:

The CUSTOM tier for Cloud Machine Learning Engine allows you to specify the number of which types of cluster nodes?

(Operationalizing machine learning models)

1. Workers
2. Masters, workers, and parameter servers
3. **Workers and parameter servers**
4. Parameter servers

Description:

#### Question 166:

Which software libraries are supported by Cloud Machine Learning Engine?

(Operationalizing machine learning models)

1. Theano and TensorFlow
2. Theano and Torch
3. **TensorFlow**
4. TensorFlow and Torch

Description:

#### Question 167:

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.

Description:

Description: Wide model is used for memorization and deep model is used for generalization to make model think like human, both needs to be used to create a recommender system like search.

#### Question 168:

To run a TensorFlow training job on your own computer using Cloud Machine Learning Engine, what would your command start with?

(Operationalizing machine learning models)

1. **gcloud ml-engine local train**
2. gcloud ml-engine jobs submit training
3. gcloud ml-engine jobs submit training local
4. You can't run a TensorFlow program on your own computer using Cloud ML Engine .

Description:

#### Question 169:

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?

(Operationalizing machine learning models)

1. Unsupervised learning
2. **Regressor**
3. Classifier
4. Clustering estimator

Description:

#### Question 170:

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.

Description:

#### Question 171:

What are two of the characteristics of using online prediction rather than batch prediction?

(Operationalizing machine learning models)

1. It is optimized to handle a high volume of data instances in a job and to run more complex models.
2. **Predictions are returned in the response message.**
3. Predictions are written to output files in a Cloud Storage location that you specify.
4. **It is optimized to minimize the latency of serving predictions.**

Description:

#### Question 172:

Which of these are examples of a value in a sparse vector? (Select 2 answers.)

(Operationalizing machine learning models)

[0, 5, 0, 0, 0, 0]

[0, 0, 0, 1, 0, 0, 1]

**[0, 1]**

**[1, 0, 0, 0, 0, 0, 0]**

Description:

#### Question 173:

How can you get a neural network to learn about relationships between categories in a categorical feature?

(Operationalizing machine learning models)

1. Create a multi-hot column
2. Create a one-hot column
3. Create a hash bucket
4. **Create an embedding column**

Description:

#### Question 174:

Which Google Cloud Platform service is an alternative to Hadoop with Hive?

(Operationalizing machine learning models)

1. Cloud Dataflow
2. Cloud Bigtable
3. **BigQuery**
4. Cloud Datastore

Description:

Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data summarization, query, and analysis.

Google BigQuery is an enterprise data warehouse.

#### Question 175:

You want to use a database of information about tissue samples to classify future tissue samples as either normal or mutated. You are evaluating an unsupervised anomaly detection method for classifying the tissue samples. Which two characteristic support this method? (Choose two.)

(Operationalizing machine learning models)

1. You already have labels for which samples are mutated and which are normal in the database.
2. **You expect future mutations to have similar features to the mutated samples in the database.**
3. You expect future mutations to have different features from the mutated samples in the database.
4. There are roughly equal occurrences of both normal and mutated samples in the database.
5. **There are very few occurrences of mutations relative to normal samples.**

Description:

#### Question 176:

You designed a database for patient records as a pilot project to cover a few hundred patients in three clinics. Your design used a single database table to represent all patients and their visits, and you used self-joins to generate reports. The server resource utilization was at 50%. Since then, the scope of the project has expanded. The database must now store 100 times more patient records. You can no longer run the reports, because they either take too long or they encounter errors with insufficient compute resources. How should you adjust the database design?

A. Add capacity (memory and disk space) to the database server by the order of 200.

B. Shard the tables into smaller ones based on date ranges, and only generate reports with prespecified date ranges.

**C. Normalize the master patient-record table into the patient table and the visits table, and create other necessary tables to avoid self-join.**

D. Partition the table into smaller tables, with one for each clinic. Run queries against the smaller table pairs, and use unions for consolidated reports.

Description:

C is correct because this option provides the least amount of inconvenience over using pre-specified date ranges or one table per clinic while also increasing performance due to avoiding self-joins.

based on Google documentation, self-join is an anti-pattern

#### 

#### Question 177:

Your company is using WILDCARD tables to query data across multiple tables with similar names. The SQL statement is currently failing with the following error:

# Syntax error : Expected end of statement but got "-" at [4:11] SELECT age FROM bigquery-public-data.noaa\_gsod.gsod WHERE age != 99 AND\_TABLE\_SUFFIX = "˜1929' ORDER BY age DESC Which table name will make the SQL statement work correctly?

(Ensuring solution quality)

"˜bigquery-public-data.noaa\_gsod.gsod"˜

bigquery-public-data.noaa\_gsod.gsod\*

"˜bigquery-public-data.noaa\_gsod.gsod'\*

**`bigquery-public-data.noaa\_gsod.gsod\*`**

Description:

it follows the correct wildcard syntax of enclosing the table name in backticks and including the \* wildcard character.

#### Question 178:

Your company is in a highly regulated industry. One of your requirements is to ensure individual users have access only to the minimum amount of information required to do their jobs. You want to enforce this requirement with Google BigQuery. Which three approaches can you take? (Choose three.)

A. Disable writes to certain tables.

**B. Restrict access to tables by role.**

C. Ensure that the data is encrypted at all times.

**D. Restrict BigQuery API access to approved users.**

E. Segregate data across multiple tables or databases.

**F. Use Google Stackdriver Audit Logging to determine policy violations**

Description:

Access control on table level is now possible in BigQuery

A. Disable writes to certain tables. ---> Read is still available(not minimal access)

C. Ensure that the data is encrypted at all times. ---> Data is encrypted by default.

E. Segregate data across multiple tables or databases. ---> Normalization is of no help here.

#### Question 179:

Your company handles data processing for a number of different clients. Each client prefers to use their own suite of analytics tools, with some allowing direct query access via Google BigQuery. You need to secure the data so that clients cannot see each other's data. You want to ensure appropriate access to the data.

Which three steps should you take? (Choose three.)

A. Load data into different partitions.

**B. Load data into a different dataset for each client.**

C. Put each client's BigQuery dataset into a different table.

**D. Restrict a client's dataset to approved users.**

E. Only allow a service account to access the datasets.

**F. Use the appropriate identity and access management (IAM) roles for each client's users.**

Description:

#### Question 180:

You want to use Google Stackdriver Logging to monitor Google BigQuery usage. You need an instant notification to be sent to your monitoring tool when new data is appended to a certain table using an insert job, but you do not want to receive notifications for other tables. What should you do?

A. Make a call to the Stackdriver API to list all logs, and apply an advanced filter.

B. In the Stackdriver logging admin interface, and enable a log sink export to BigQuery.

C. In the Stackdriver logging admin interface, enable a log sink export to Google Cloud Pub/Sub, and subscribe to the topic from your monitoring tool.

**D. Using the Stackdriver API, create a project sink with advanced log filter to export to Pub/Sub, and subscribe to the topic from your monitoring tool.**

Description:

D as the key requirement is to have notification on a particular table. It can be achieved using advanced log filter to filter only the table logs and create a project sink to Cloud Pub/Sub for notification.

Refer GCP documentation - Advanced Logs Filters: https://cloud.google.com/logging/docs/view/advanced-queries

A is wrong as advanced filter will help in filtering. However, there is no notification sends.

B is wrong as it would send all the logs and BigQuery does not provide notifications.

C is wrong as it would send all the logs.

#### Question 181:

Your company's customer and order databases are often under heavy load. This makes performing analytics against them difficult without harming operations.

The databases are in a MySQL cluster, with nightly backups taken using mysqldump. You want to perform analytics with minimal impact on operations. What should you do?

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

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

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

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

Description:

It is a GOOGLE exam. The answer won't be on-premise or OLAP cubes even if it is the easiest. The answer is B

#### Question 182:

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. The performance issue should be resolved over time as the site of the BigDate cluster is increased.

C. Redesign the schema to use a single row key to identify values that need to be updated frequently in the cluster.

D. Redesign the schema to use row keys based on numeric IDs that increase sequentially per user viewing the offers.

Description:

If you find that you're reading and writing only a small number of rows, you might need to redesign your schema so that reads and writes are more evenly distributed.

The table's schema is not designed correctly. To get good performance from Cloud Bigtable, it's essential to design a schema that makes it possible to distribute reads and writes evenly across each table

#### Question 183:

Your company has recently grown rapidly and now ingesting data at a significantly higher rate than it was previously. You manage the daily batch MapReduce analytics jobs in Apache Hadoop. However, the recent increase in data has meant the batch jobs are falling behind. You were asked to recommend ways the development team could increase the responsiveness of the analytics without increasing costs. What should you recommend they do?

A. Rewrite the job in Pig.

**B. Rewrite the job in Apache Spark.**

C. Increase the size of the Hadoop cluster.

D. Decrease the size of the Hadoop cluster but also rewrite the job in Hive.

Description:

Spark performs in-memory processing and faster, which results in optimization of job’s processing time

The objective is to not increase the cost at the sametime do the analyitics required. Mapreduce jobs are not efficient and fast as spark so it will avoid failing the jobs.

#### Question 184:

You work for a manufacturing plant that batches application log files together into a single log file once a day at 2:00 AM. You have written a Google Cloud

Dataflow job to process that log file. You need to make sure the log file in processed once per day as inexpensively as possible. What should you do?

A. Change the processing job to use Google Cloud Dataproc instead.

B. Manually start the Cloud Dataflow job each morning when you get into the office.

**C. Create a cron job with Google App Engine Cron Service to run the Cloud Dataflow job.**

D. Configure the Cloud Dataflow job as a streaming job so that it processes the log data immediately.

Description:

Scheduler for adhoc jobs – 3 jobs free and $0.10 per job

#### Question 185:

Your company is loading comma-separated values (CSV) files into Google BigQuery. The data is fully imported successfully; however, the imported data is not matching byte-to-byte to the source file. What is the most likely cause of this problem?

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

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

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

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

Description:

C is correct because this is the only situation that would cause successful import.

A is not correct because if another data format other than CSV was selected then the data would not import successfully.

B is not correct because the data was fully imported meaning no rows were skipped.

D is not correct because whether the data has been previously transformed will not affect whether the source file will match the BigQuery table.

#### Question 186:

You are implementing security best practices on your data pipeline. Currently, you are manually executing jobs as the Project Owner. You want to automate these jobs by taking nightly batch files containing non-public information from Google Cloud Storage, processing them with a Spark Scala job on a Google Cloud

Dataproc cluster, and depositing the results into Google BigQuery.

How should you securely run this workload?

A. Restrict the Google Cloud Storage bucket so only you can see the files

B. Grant the Project Owner role to a service account, and run the job with it

**C. Use a service account with the ability to read the batch files and to write to BigQuery**

D. Use a user account with the Project Viewer role on the Cloud Dataproc cluster to read the batch files and write to BigQuery

Description:

#### Question 187:

Your company receives both batch- and stream-based event data. You want to process the data using Google Cloud Dataflow over a predictable time period.

However, you realize that in some instances data can arrive late or out of order. How should you design your Cloud Dataflow pipeline to handle data that is late or out of order?

A. Set a single global window to capture all the data.

B. Set sliding windows to capture all the lagged data.

**C. Use watermarks and timestamps to capture the lagged data.**

D. Ensure every datasource type (stream or batch) has a timestamp, and use the timestamps to define the logic for lagged data.

Description:

"Watermark in implementation is a monotonically increasing timestamp. When Beam/Dataflow see a record with an event timestamp that is earlier than the watermark, the record is treated as late data."

#### Question 188:

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.

Description:

A makes no sense, you need to use your own keys.

You don’t create keys locally and upload them, you should import it to make it work..using the kms public key…not just “uploading” it. C is also out.

IT’s between B and D

Cloud KMS is a cloud-hosted key management service that lets you manage cryptographic keys for your cloud services the same way you do on-premises, You can generate, use, rotate, and destroy cryptographic keys from there.

Since you want to encrypt data at rest, is B, you don’t use them for any API calls.

#### Question 189:

An organization maintains a Google BigQuery dataset that contains tables with user-level data. They want to expose aggregates of this data to other Google

Cloud projects, while still controlling access to the user-level data. Additionally, they need to minimize their overall storage cost and ensure the analysis cost for other projects is assigned to those projects. What should they do?

**A. Create and share an authorized view that provides the aggregate results.**

B. Create and share a new dataset and view that provides the aggregate results.

C. Create and share a new dataset and table that contains the aggregate results.

D. Create dataViewer Identity and Access Management (IAM) roles on the dataset to enable sharing.

Description:

B will not minimize storage cost, neither will C

I dont think is D: dataviewer role will allow this:

– Read the dataset’s metadata and to list tables in the dataset.

– Read data and metadata from the dataset’s tables.

This is not what we want. They’ll see our users data!

A looks good: An authorized view allows you to share query results with particular users and groups without giving them access to the underlying tables. and also enforcing row-level access, which is what we want.

#### Question 190:

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.

Description:

Description: Subsampling is the method to increase the training speed

#### Question 191:

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 CPU size on your server.

B. Increase the size of the Google Persistent Disk on your server.

**C. Increase your network bandwidth from your datacenter to GCP.**

D. Increase your network bandwidth from Compute Engine to Cloud Storage.

Description:

#### Question 192:

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.

Description:

#### Question 193:

You have a data pipeline that writes data to Cloud Bigtable using well-designed row keys. You want to monitor your pipeline to determine when to increase the size of you Cloud Bigtable cluster. Which two actions can you take to accomplish this? (Choose two.)

A. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Read pressure index is above 100.

B. Review Key Visualizer metrics. Increase the size of the Cloud Bigtable cluster when the Write pressure index is above 100.

**C. Monitor the latency of write operations. Increase the size of the Cloud Bigtable cluster when there is a sustained increase in write latency.**

**D. Monitor storage utilization. Increase the size of the Cloud Bigtable cluster when utilization increases above 70% of max capacity.**

E. Monitor latency of read operations. Increase the size of the Cloud Bigtable cluster of read operations take longer than 100 ms.

Description:

D: In general, do not use more than 70% of the hard limit on total storage, so you have room to add more data. If you do not plan to add significant amounts of data to your instance, you can use up to 100% of the hard limit

C: If this value is frequently at 100%, you might experience increased latency. Add nodes to the cluster to reduce the disk load percentage.

The key visualizer metrics options, suggest other things other than increase the cluster size.

#### 

#### Question 194:

You have a query that filters a BigQuery table using a WHERE clause on timestamp and ID columns. By using bq query "" -dry\_run you learn that the query triggers a full scan of the table, even though the filter on timestamp and ID select a tiny fraction of the overall data. You want to reduce the amount of data scanned by BigQuery with minimal changes to existing SQL queries. What should you do?

A. Create a separate table for each ID.

B. Use the LIMIT keyword to reduce the number of rows returned.

**C. Recreate the table with a partitioning column and clustering column.**

D. Use the bq query - -maximum\_bytes\_billed flag to restrict the number of bytes billed.

Description:

#### Question 195:

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 load to 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.

Description:

#### Question 196:

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. Leverage BigQuery UPDATE statements to update the inventory balances as they are changing.**

B. Partition the inventory balance table by item to reduce the amount of data scanned with each inventory update.

C. Use the BigQuery streaming the stream changes into a daily inventory movement table. Calculate balances in a view that joins it to the historical inventory balance table. Update the inventory balance table nightly.

D. 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.

Description:

There is no dml limitation anymore, so A is also OK

#### Question 197:

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.

Description:

highly available = multi-regional:

recovery strategy of this data that minimizes cost = point-in-time snapshot:

#### Question 198:

You are managing a Cloud Dataproc cluster. You need to make a job run faster while minimizing costs, without losing work in progress on your clusters. What should you do?

A. Increase the cluster size with more non-preemptible workers.

B. Increase the cluster size with preemptible worker nodes, and configure them to forcefully decommission.

C. Increase the cluster size with preemptible worker nodes, and use Cloud Stackdriver to trigger a script to preserve work.

**D. Increase the cluster size with preemptible worker nodes, and configure them to use graceful decommissioning.**

Description:

Graceful decommissioning will ensure that the data is processed by worker before it is removed by Yarn

After creating a Dataproc cluster, you can adjust ("scale") the cluster by increasing or decreasing the number of primary or secondary worker nodes in the cluster. You can scale a Dataproc cluster at any time, even when jobs are running on the cluster.

Use Dataproc Autoscaling. Instead of manually scaling clusters, enable Autoscaling to have Dataproc set the "right" number of workers for your workloads.

Why scale a Dataproc cluster?

* to increase the number of workers to make a job run faster
* to decrease the number of workers to save money (see Graceful Decommissioning as an option to use when downsizing a cluster to avoid losing work in progress).
* to increase the number of nodes to expand available Hadoop Distributed Filesystem (HDFS) storage

#### Question 199:

You have historical data covering the last three years in BigQuery and a data pipeline that delivers new data to BigQuery daily. You have noticed that when the

Data Science team runs a query filtered on a date column and limited to 30""90 days of data, the query scans the entire table. You also noticed that your bill is increasing more quickly than you expected. You want to resolve the issue as cost-effectively as possible while maintaining the ability to conduct SQL queries.

What should you do?

**A. Re-create the tables using DDL. Partition the tables by a column containing a TIMESTAMP or DATE Type.**

B. Recommend that the Data Science team export the table to a CSV file on Cloud Storage and use Cloud Datalab to explore the data by reading the files directly.

C. Modify your pipeline to maintain the last 30""90 days of data in one table and the longer history in a different table to minimize full table scans over the entire history.

D. Write an Apache Beam pipeline that creates a BigQuery table per day. Recommend that the Data Science team use wildcards on the table name suffixes to select the data they need.

Description:

With partitions the performance will improve for selecting 30-90 days data. Also the storage cost will reduce as the old partitions (not updated in last 90 days) will qualify for Long-Term storage rates.

#### Question 200:

You operate a logistics company, and you want to improve event delivery reliability for vehicle-based sensors. You operate small data centers around the world to capture these events, but leased lines that provide connectivity from your event collection infrastructure to your event processing infrastructure are unreliable, with unpredictable latency. You want to address this issue in the **most cost-effective way.** What should you do?

A. Deploy small Kafka clusters in your data centers to buffer events.

**B. Have the data acquisition devices publish data to Cloud Pub/Sub.**

C. Establish a Cloud Interconnect between all remote data centers and Google.

D. Write a Cloud Dataflow pipeline that aggregates all data in session windows.

Description:

The most cost effective (cheapest) way is to use PubSub. It can handle messages with high latency.

#### Question 201:

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.

Description:

#### Question 202:

As your organization expands its usage of GCP, many teams have started to create their own projects. Projects are further multiplied to accommodate different stages of deployments and target audiences. Each project requires unique access control configurations. The central IT team needs to have access to all projects.

Furthermore, data from Cloud Storage buckets and BigQuery datasets must be shared for use in other projects in an ad hoc way. You want to simplify access control management by minimizing the number of policies. Which two steps should you take? (Choose two.)

A. Use Cloud Deployment Manager to automate access provision.

**B. Introduce resource hierarchy to leverage access control policy inheritance.**

**C. Create distinct groups for various teams, and specify groups in Cloud IAM policies.**

D. Only use service accounts when sharing data for Cloud Storage buckets and BigQuery datasets.

E. For each Cloud Storage bucket or BigQuery dataset, decide which projects need access. Find all the active members who have access to these projects, and create a Cloud IAM policy to grant access to all these users.

Description:

B & C

Google Cloud resources are organized hierarchically, where the organization node is the root node in the hierarchy, the projects are the children of the organization, and the other resources are descendents of projects.

You can set Cloud Identity and Access Management (Cloud IAM) policies at different levels of the resource hierarchy. Resources inherit the policies of the parent resource. The effective policy for a resource is the union of the policy set at that resource and the policy inherited from its parent.

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

#### Question 203:

You are running a pipeline in Cloud Dataflow that receives messages from a Cloud Pub/Sub topic and writes the results to a BigQuery dataset in the EU.

Currently, your pipeline is located in europe-west4 and has a maximum of 3 workers, instance type n1-standard-1. You notice that during peak periods, your pipeline is struggling to process records in a timely fashion, when all 3 workers are at maximum CPU utilization. Which two actions can you take to increase performance of your pipeline? (Choose two.)

**A. Increase the number of max workers**

**B. Use a larger instance type for your Cloud Dataflow workers**

C. Change the zone of your Cloud Dataflow pipeline to run in us-central1

D. Create a temporary table in Cloud Bigtable that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Bigtable to BigQuery

E. Create a temporary table in Cloud Spanner that will act as a buffer for new data. Create a new step in your pipeline to write to this table first, and then create a new pipeline to write from Cloud Spanner to BigQuery

Description:A & B

instance n1-standard-1 is low configuration and hence need to be larger configuration, definitely B should be one of the option.

Increase max workers will increase parallelism and hence will be able to process faster given larger CPU size and multi core processor instance type is chosen. Option A can be a better step.

#### Question 204:

You have a data pipeline with a Cloud Dataflow job that aggregates and writes time series metrics to Cloud Bigtable. This data feeds a dashboard used by thousands of users across the organization. You need to support additional concurrent users and reduce the amount of time required to write the data. Which two actions should you take? (Choose two.)

A. Configure your Cloud Dataflow pipeline to use local execution

**B. Increase the maximum number of Cloud Dataflow workers by setting maxNumWorkers in PipelineOptions**

**C. Increase the number of nodes in the Cloud Bigtable cluster**

D. Modify your Cloud Dataflow pipeline to use the Flatten transform before writing to Cloud Bigtable

E. Modify your Cloud Dataflow pipeline to use the CoGroupByKey transform before writing to Cloud Bigtable

Description:

The maximum number of Compute Engine instances to be made available to your pipeline during execution. Note that this can be higher than the initial number of workers (specified by num\_workers to allow your job to scale up, automatically or otherwise.

Adding nodes to the original cluster: You can add 3 nodes to the cluster, for a total of 6 nodes. The write throughput for the instance doubles, but the instance's data is available in only one zone:

#### Question 205:

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. The current epoch time

B. A concatenation of the product name and the current epoch time

**C. A random universally unique identifier number (version 4 UUID)**

D. The original order identification number from the sales system, which is a monotonically increasing integer

Description:

Choosing a primary key

* [Hash the key](https://cloud.google.com/spanner/docs/schema-design#fix_hash_the_key) and store it in a column. Use the hash column (or the hash column and the unique key columns together) as the primary key.
* [Swap the order](https://cloud.google.com/spanner/docs/schema-design#fix_swap_key_order) of the columns in the primary key.
* [Use a Universally Unique Identifier (UUID)](https://cloud.google.com/spanner/docs/schema-design#uuid_primary_key). Version 4 [UUID](https://tools.ietf.org/html/rfc4122) is recommended, because it uses random values in the high-order bits. Don't use a UUID algorithm (such as version 1 UUID) that stores the timestamp in the high order bits.
* [Bit-reverse](https://cloud.google.com/spanner/docs/schema-design#bit_reverse_primary_key) sequential values.

#### Question 206:

Each analytics team in your organization is running BigQuery jobs in their own projects. You want to enable each team to monitor slot usage within their projects.

What should you do?

A. Create a Stackdriver Monitoring dashboard based on the BigQuery metric query/scanned\_bytes

**B. Create a Stackdriver Monitoring dashboard based on the BigQuery metric slots/allocated\_for\_project**

C. Create a log export for each project, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

D. Create an aggregated log export at the organization level, capture the BigQuery job execution logs, create a custom metric based on the totalSlotMs, and create a Stackdriver Monitoring dashboard based on the custom metric

Description:

Viewing project and reservation slot usage in Stackdriver Monitoring

Information is available from the "Slots Allocated" metric in Stackdriver Monitoring. This metric information includes a per-reservation and per-job breakdown of slot usage. The information can also be visualized by using the custom charts metric explorer.

#### Question 207:

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. Assign the users/groups data viewer access at the table level for each table**

B. 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

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

Description:

#### Question 208:

You plan to deploy Cloud SQL using MySQL. You need to ensure high availability in the event of a zone failure. What should you do?

**A. Create a Cloud SQL instance in one zone, and create a failover replica in another zone within the same region.**

B. Create a Cloud SQL instance in one zone, and create a read replica in another zone within the same region.

C. Create a Cloud SQL instance in one zone, and configure an external read replica in a zone in a different region.

D. Create a Cloud SQL instance in a region, and configure automatic backup to a Cloud Storage bucket in the same region.

Description:

The HA (High Availability) configuration, sometimes called a cluster, provides data redundancy. A Cloud SQL instance configured for HA is also called a regional instance and is located in a primary and secondary zone within the configured region. Within a regional instance, the configuration is made up of a primary instance and a standby instance

#### Question 209:

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

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

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. 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.**

Description:

#### Question 210:

Which of the following is **not** possible using primitive roles?

A. Give a user viewer access to BigQuery and owner access to Google Compute Engine instances.

B. Give UserA owner access and UserB editor access for all datasets in a project.

**C. Give a user access to view all datasets in a project, but not run queries on them.**

D. Give GroupA owner access and GroupB editor access for all datasets in a project.

Description:

#### Question 211:

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.**

Description:

#### Question 212:

Which methods can be used to reduce the number of rows processed by BigQuery?

(Ensuring solution quality)

1. **Splitting tables into multiple tables; putting data in partitions**
2. Splitting tables into multiple tables; putting data in partitions; using the LIMIT clause
3. Putting data in partitions; using the LIMIT clause
4. Splitting tables into multiple tables; using the LIMIT clause

Description: not c: LIMIT would still perform a full scan. It will limit the output of the query, not the data scanned

#### Question 213:

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?

(Ensuring solution quality)

1 continuous and 2 categorical

3 categorical

3 continuous

**2 continuous and 1 categorical**

Description:

Year can be any value, income can be any value, so continuous, and country is categorical as values are finite

#### Question 214:

What is the recommended action to do in order to switch between SSD and HDD storage for your Google Cloud Bigtable instance?

A. create a third instance and sync the data from the two storage types via batch jobs

**B. export the data from the existing instance and import the data into a new instance**

C. run parallel instances where one is HDD and the other is SDD

D. the selection is final and you must resume using the same storage type

Description:

When you create a Cloud Bigtable instance and cluster, your choice of SSD or HDD storage for the cluster is permanent. You cannot use the Google Cloud

Platform Console to change the type of storage that is used for the cluster.

If you need to convert an existing HDD cluster to SSD, or vice-versa, you can export the data from the existing instance and import the data into a new instance.

Alternatively, you can write -

a Cloud Dataflow or Hadoop MapReduce job that copies the data from one instance to another.

#### Question 215:

What is the HBase Shell for Cloud Bigtable?

(Ensuring solution quality)

1. The HBase shell is a GUI based interface that performs administrative tasks, such as creating and deleting tables.
2. **The HBase shell is a command-line tool that performs administrative tasks, such as creating and deleting tables.**
3. The HBase shell is a hypervisor based shell that performs administrative tasks, such as creating and deleting new virtualized instances.
4. The HBase shell is a command-line tool that performs only user account management functions to grant access to Cloud Bigtable instances.

Description:

#### Question 216:

Google Cloud Bigtable indexes a single value in each row. This value is called the \_\_\_\_\_\_\_.

(Ensuring solution quality)

primary key

unique key

**row key**

master key

Description:

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.**

#### Question 217:

Cloud Bigtable is a recommended option for storing very large amounts of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_?

(Ensuring solution quality)

multi-keyed data with very high latency

multi-keyed data with very low latency

**single-keyed data with very low latency**

single-keyed data with very high latency

Description:

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. Cloud Bigtable is ideal for storing very large amounts of single-keyed data with very low latency. It supports high read and write throughput at low latency, and it is an ideal data source for MapReduce operations.

#### Question 218:

When you store data in Cloud Bigtable, what is the recommended minimum amount of stored data?

(Ensuring solution quality)

500 TB

1 GB

**1 TB**

500 GB

Description:

#### Question 219:

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.**

Description:

#### Question 220:

Which is the preferred method to use to avoid hotspotting in time series data in Bigtable?

**Field promotion**

Randomization

Salting

Hashing

Description:

By default, prefer field promotion. Field promotion avoids hotspotting in almost all cases, and it tends to make it easier to design a row key that facilitates queries.

#### Question 221:

Which is not a valid reason for poor Cloud Bigtable performance?

(Ensuring solution quality)

The workload isn't appropriate for Cloud Bigtable.

The table's schema is not designed correctly.

**The Cloud Bigtable cluster has too many nodes.**

There are issues with the network connection.

Description:

The Cloud Bigtable cluster doesn't have enough nodes. If your Cloud Bigtable cluster is overloaded, adding more nodes can improve performance. Use the monitoring tools to check whether the cluster is overloaded.

#### Question 222:

When a Cloud Bigtable node fails, \_\_\_\_ is lost.

(Ensuring solution quality)

all data

**no data**

the last transaction

the time dimension

Description:

#### Question 223:

Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?

**A. A sequential numeric ID**

**B. A timestamp followed by a stock symbol**

C. A non-sequential numeric ID

D. A stock symbol followed by a timestamp

Description:

1. Row keys that start with a timestamp. This will cause sequential writes to be pushed onto a single node, creating a hotspot. If you put a timestamp in a row key, you need to precede it with a high-cardinality value like a user ID to avoid hotspotting.

2.Sequential numeric IDs. Suppose your system assigns a numeric ID to each of your application's users. You might be tempted to use the user's numeric ID as the row key for your table. However, because new users are more likely to be active users, this approach is likely to push most of your traffic to a small number of nodes.

#### Question 224:

Which row keys are likely to cause a disproportionate number of reads and/or writes on a particular node in a Bigtable cluster (select 2 answers)?

(Ensuring solution quality)

A sequential numeric ID

A timestamp followed by a stock symbol

A non-sequential numeric ID

A stock symbol followed by a timestamp

Description:

#### Question 225:

For the best possible performance, what is the recommended zone for your Compute Engine instance and Cloud Bigtable instance?

A. Have the Compute Engine instance in the furthest zone from the Cloud Bigtable instance.

B. Have both the Compute Engine instance and the Cloud Bigtable instance to be in different zones.

**C. Have both the Compute Engine instance and the Cloud Bigtable instance to be in the same zone.**

D. Have the Cloud Bigtable instance to be in the same zone as all of the consumers of your data.

Description:

#### Question 226:

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

Description:

#### Question 227:

You work for a car manufacturer and have set up a data pipeline using Google Cloud Pub/Sub to capture anomalous sensor events. You are using a push subscription in Cloud Pub/Sub that calls a custom HTTPS endpoint that you have created to take action of these anomalous events as they occur. Your custom

HTTPS endpoint keeps getting an inordinate amount of duplicate messages. What is the most likely cause of these duplicate messages?

A. The message body for the sensor event is too large.

B. Your custom endpoint has an out-of-date SSL certificate.

C. The Cloud Pub/Sub topic has too many messages published to it.

**D. Your custom endpoint is not acknowledging messages within the acknowledgement deadline.**

Description:

The custom endpoint is not acknowledging the message, that is the reason for Pub/Sub to send the message again and again.

#### Question 228:

Your company uses a proprietary system to send inventory data every 6 hours to a data ingestion service in the cloud. Transmitted data includes a payload of several fields and the timestamp of the transmission. If there are any concerns about a transmission, the system re-transmits the data. How should you deduplicate the data most efficiency?

**A. Assign global unique identifiers (GUID) to each data entry.**

B. Compute the hash value of each data entry, and compare it with all historical data.

C. Store each data entry as the primary key in a separate database and apply an index.

D. Maintain a database table to store the hash value and other metadata for each data entry.

Description:

# 

# BigQuery

## Audit logs

Audit logs help you answer "who did what, where, and when?" within your Google Cloud resources with the same level of transparency as in on-premises environments. Enabling audit logs helps your security, auditing, and compliance entities monitor Google Cloud data and systems for possible vulnerabilities or external data misuse.

Data Access audit logs-- except for BigQuery

Data Access audit logs-- are disabled by default because audit logs can be quite large.

Stackdriver is used to track access logs for Bigquery.

## Manually change table schemas

**Change a column's name**

**Change a column's data type**

**Delete columns from table schema**

There are two ways to manually rename a column, change a column's data type:

* Using a SQL query: choose this option if you are more concerned about simplicity and ease of use, and you are less concerned about costs.
* Recreating the table: choose this option if you are more concerned about costs, and you are less concerned about simplicity and ease of use.

**Change column mode**

Currently, the only supported modification you can make to a column's mode is changing it from REQUIRED to NULLABLE. Changing a column's mode from REQUIRED to NULLABLE is also called column relaxation.

You can manually change a column's mode by [exporting](https://cloud.google.com/bigquery/docs/exporting-data) your table data to Cloud Storage, and then [loading](https://cloud.google.com/bigquery/docs/loading-data) the data into a new table with a schema definition that specifies the correct mode for the column. You can also use the load job to overwrite the existing table.

## Dataflow Developer

(roles/dataflow.developer)

Provides the permissions necessary to execute and manipulate Dataflow jobs.

Lowest-level resources where you can grant this role:

* Project

## [Accessing external (federated) data sources with BigQuery’s data access layer](https://cloud.google.com/blog/products/gcp/accessing-external-federated-data-sources-with-bigquerys-data-access-layer)

There are cases in data lake architectures where external federated data sources may make sense. Traditionally we find that data lakes are comprised of raw, un/semi-structured, and schema on read.

Corner cases for external data sources:

* Avoiding duplicate data in BigQuery storage
* Queries that do not have strong performance requirements
* **Small amount of frequently changing data to join with other tables in BigQuery**
* Irregular, non scheduled, non shared queries

**From BigQuery to Google Cloud Storage**

There are two ways to use BigQuery to query files in a Cloud Storage bucket in the same project; by querying a temporary or permanent table.

**Permanent table—**You create a table in a BigQuery dataset that is linked to your external data source. This allows you to use BigQuery dataset-level IAM roles to share the table with others who may have access to the underlying external data source. **Use permanent tables when you need to share the table with others.**

**Temporary table—**You submit a command that includes a query and creates a non-permanent table linked to the external data source. With this approach you do not create a table in one of your BigQuery datasets, so make sure to give consideration towards sharing the query or table. Consider using a temporary table for one-time, ad-hoc queries, or for one time extract, transform, or load (ETL) workflows

## 

## Introduction to external data sources

An external data source is a data source that you can query directly from BigQuery, even though the data is not stored in BigQuery storage.

BigQuery supports the following external data sources:

* Bigtable
* Cloud Spanner
* Cloud SQL
* Cloud Storage
* Drive

Use cases for external data sources include:

For ELT (extract-load-transform) workloads, loading and cleaning your data in one pass and writing the cleaned result into BigQuery storage, by using a CREATE TABLE ... AS SELECT query.

Joining BigQuery tables with frequently changing data from an external data source. By querying the external data source directly, you don't need to reload the data into BigQuery storage every time it changes.

BigQuery has two different mechanisms for querying external data:

**External tables**

An external table is a table that acts like a standard BigQuery table. The table metadata, including the table schema, is stored in BigQuery storage, but the data itself resides in the external source.

External tables can be temporary or permanent. A permanent external table is contained inside a dataset, and you manage it in the same way that you manage a standard BigQuery table. For example, you can view the table properties, set access controls, and so forth. You can query the table and join it with other tables.

You can use external tables with the following data sources:

* Bigtable
* Cloud Storage
* Drive

**Federated queries**

A federated query is a way to send a query statement to an external database and get the result back as a temporary table. Federated queries use the BigQuery Connection API to establish a connection with the external database. In your standard SQL query, you use the EXTERNAL\_QUERY function to send a query statement to the external database, using that database's SQL dialect. The results are converted to BigQuery standard SQL data types.

You can use federated queries with the following external databases:

* Cloud Spanner
* Cloud SQL

## Create an authorized view

When would you use ELT?

* Extract data from files in Cloud Storage into BigQuery.
* Transform the data on the fly using BigQuery views, or store into new tables.

A problem like data being out of range can be solved in BigQuery without an

intermediate transformation step. Invalid data can be filtered out using a BigQuery

view, and everyone can access the view rather than the raw data.

[BigQuery](https://cloud.google.com/bigquery/docs) is a petabyte-scale analytics data warehouse that you can use to run SQL queries over vast amounts of data in near real time.

Giving a view access to a dataset is also known as creating an [authorized view](https://cloud.google.com/bigquery/docs/authorized-views) in BigQuery. An authorized view lets you share query results with particular users and groups without giving them access to the underlying tables. You can also use the view's SQL query to restrict the columns (fields) the users are able to query. In this tutorial, you create an authorized view.

## 

## Loading Avro data from Cloud Storage

**Advantages of Avro**

Avro is the preferred format for loading data into BigQuery. Loading Avro files has the following advantages over CSV and JSON (newline delimited):

**The Avro binary format:**

* Is faster to load. The data can be read in **parallel**, even if the data blocks are **compressed**.
* Doesn't require typing or serialization.
* Is easier to parse because there are no encoding issues found in other formats such as ASCII.

When you load Avro files into BigQuery, the table schema is automatically retrieved from the self-describing source data.

## loading data into BigQuery

There are several ways to ingest data into BigQuery:

* Batch load a set of data records.
* Stream individual records or batches of records.
* Use queries to generate new data and append or overwrite the results to a table.
* Use a third-party application or service.

### Batch loading

Options for batch loading in BigQuery include the following:

* **Load jobs.** Load data from Cloud Storage or from a local file by creating a [load job](https://cloud.google.com/bigquery/docs/batch-loading-data). The records can be in Avro, CSV, JSON, ORC, or Parquet format.
* **BigQuery Data Transfer Service.** Use [BigQuery Data Transfer Service](https://cloud.google.com/bigquery-transfer/docs/transfer-service-overview) to automate loading data from Google Software as a Service (SaaS) apps or from third-party applications and services.
* **BigQuery Storage Write API.** The Storage Write API lets you batch-process an arbitrarily large number of records and commit them in a single atomic operation. If the commit operation fails, you can safely retry the operation. Unlike BigQuery load jobs, the Storage Write API does not require staging the data to intermediate storage such as Cloud Storage.
* **Other managed services.** Use other managed services to export data from an external data store and import it into BigQuery. For example, you can load data from Firestore exports.

### Streaming

With streaming, you continually send smaller batches of data in real time, so the data is available for querying as it arrives. Options for streaming in BigQuery include the following:

* Storage Write API. The Storage Write API supports high-throughput streaming ingestion with exactly-once delivery semantics.
* Dataflow. Use [Dataflow](https://cloud.google.com/dataflow/docs) with the Apache Beam SDK to set up a streaming pipeline that writes to BigQuery.
* BigQuery Connector for SAP. The BigQuery Connector for SAP enables near real time replication of SAP data directly into BigQuery. For more information, see the [BigQuery Connector for SAP planning guide](https://cloud.google.com/solutions/sap/docs/bq-connector-for-sap-planning).

# Cloud Logging

[Cloud Logging](https://cloud.google.com/logging) is a service that allows you to store, search, monitor, and alert on log data and events from your Google Cloud Platform (GCP) infrastructure services and your applications. You can view and analyze log data in real time via Logs Viewer, command line or Cloud SDK.

Using a [Logging sink](https://cloud.google.com/logging/docs/export), you can build an event-driven system to detect and respond to log events in real time. Cloud Logging can help you to build this event-driven architecture through its [integration](https://cloud.google.com/logging/docs/export/using_exported_logs#pubsub-overview) with [Cloud Pub/Sub](https://cloud.google.com/pubsub) and a serverless computing service such as [Cloud Functions](https://cloud.google.com/functions) or [Cloud Run](https://cloud.google.com/run).

**Pub/Sub topic** – In Pub/Sub, you can create a topic to which to direct the log sink and use the Pub/Sub message to trigger a cloud function.

**Cloud Function** – In Cloud Functions, you can create logic to evaluate the received logs based on your business requirements.

# Dataproc

Dataproc is used to migrate Hadoop and Spark jobs on GCP. Dataproc with GCS connected through Google Cloud Storage connector helps store data after the life of the cluster. When the job is high I/O intensive, then we need to create a small persistent disk.

i) re-use hadoop jobs with minimum cluster management - Cloud Dataproc

ii) Persist data beyond the life of cluster - use Cloud Storage.

# Cloud Pub/Sub

Dealing with duplicates and forcing retries:

When you do not acknowledge a message before its [acknowledgement deadline](https://cloud.google.com/pubsub/docs/subscriber#delivery) has expired, Pub/Sub resends the message. As a result, Pub/Sub can send duplicate messages. Use Cloud Monitoring to [monitor](https://cloud.google.com/pubsub/docs/monitoring) acknowledge operations with the expired response code to detect this condition. To get this data, select the [subscription/expired\_ack\_deadlines\_count](https://cloud.google.com/monitoring/api/metrics_gcp#pubsub/subscription/expired_ack_deadlines_count) metric.

# Cloud Datalab

Use Cloud Datalab to easily explore, visualize, analyze, and transform data using familiar languages, such as Python and SQL, interactively.

Use Cloud Datalab to easily explore, visualize, analyze, and transform data using familiar languages, such as Python and SQL, interactively. Pre-installed Jupyter introductory, sample, and tutorial notebooks, show you how to:

* Access, analyze, monitor, and visualize data
* Use notebooks with Python, TensorFlow Machine Learning, and Google Analytics, Google BigQuery, and Google Charts APIs

# Cloud Bigtable

## schema design

The following general concepts apply to Bigtable schema design:

* Bigtable is a key/value store, not a relational store. It does not support joins, and transactions are supported only within a single row.
* Each table has only one index, the row key. There are no secondary indices. Each row key must be unique.
* Rows are sorted lexicographically by row key, from the lowest to the highest byte string. Row keys are sorted in big-endian byte order (sometimes called network byte order), the binary equivalent of alphabetical order.
* Column families are not stored in any specific order.
* Columns are grouped by column family and sorted in lexicographic order within the column family. For example, in a column family called SysMonitor with column qualifiers of ProcessName, User, %CPU, ID, Memory, DiskRead, and Priority, Bigtable stores the columns in this order:

Graphical user interface, application, table

Description automatically generated with medium confidence

* The intersection of a row and column can contain multiple timestamped cells. Each cell contains a unique, timestamped version of the data for that row and column.
* All operations are atomic at the row level. This means that an operation affects either an entire row or none of the row.
* Ideally, both reads and writes should be distributed evenly across the row space of a table.
* Bigtable tables are sparse. A column doesn't take up any space in a row that doesn't use the column.

### Row keys

In many cases, you should design row keys that start with a common value and end with a granular value.

Row keys to avoid:

* Row keys that start with a timestamp, don't use a timestamp by itself or at the beginning of a row key.
* Row keys that cause related data to not be grouped together.
* Sequential numeric IDs.
* Frequently updated identifiers.
* Hashed values.
* Values expressed as raw bytes

## Troubleshoot performance issues in Bigtable:

Make sure you're reading and writing many different rows in your table. Bigtable performs best when reads and writes are evenly distributed throughout your table, which helps Bigtable distribute the workload across all of the nodes in your cluster. If reads and writes cannot be spread across all of your Bigtable nodes, performance will suffer.  
If you find that you're reading and writing only a small number of rows, you might need to [redesign your schema](https://cloud.google.com/bigtable/docs/schema-design) so that reads and writes are more evenly distributed.

# Apache Spark

Apache Spark is faster than Hadoop/Pig/MapReduce

# 

# Datastore

When the same property is repeated multiple times, Firestore in Datastore mode can detect exploding indexes and suggest an alternative index. However, in all other circumstances (such as the query defined in this example), a Datastore mode database will generate an exploding index. In this case, you can circumvent the exploding index by manually configuring an index in your index configuration file:

Graphical user interface

Description automatically generated with low confidence

# The App Engine Cron Service

The App Engine Cron Service allows you to configure regularly scheduled tasks that operate at defined times or regular intervals. These tasks are commonly known as *cron jobs*. These cron jobs are automatically triggered by the App Engine Cron Service. For instance, you might use this to send out a report email on a daily basis, to update some cached data every 10 minutes, or to update some summary information once an hour.

Free applications can have up to 20 scheduled tasks. Paid applications can have up to 250 scheduled tasks.

# Cloud Storage Transfer Service

Should you use gsutil or Storage Transfer Service?

The [gsutil](https://cloud.google.com/storage/docs/gsutil) command-line tool also enables you to transfer data between Cloud Storage and other locations. While you can use gsutil to transfer from another cloud storage provider to Cloud Storage, Storage Transfer Service is recommended for this use case. You can also use gsutil to transfer between Cloud Storage buckets. This can be helpful for moving data between [bucket locations](https://cloud.google.com/storage/docs/locations).

Follow these rules of thumb when deciding whether to use gsutil or Storage Transfer Service:

Graphical user interface, application

Description automatically generated

# 

# Preemptible Instance

Preemptible VM instances are available **at much lower price**—a 60-91% discount—compared to the price of standard VMs. H

Preemptible instance limitations:

* Compute Engine always stops preemptible instances after they run for 24 hours.
* Preemptible instances are finite Compute Engine resources, so they might not always be available.
* Preemptible instances can't live migrate to a regular VM instance, or be set to automatically restart when there is a maintenance event.

# Cloud Spanner

Cloud Spanner is a fully managed, mission-critical, relational database service that offers transactional consistency at global scale, automatic, synchronous replication for high availability, and support for two SQL dialects: Google Standard SQL (ANSI 2011 with extensions) and PostgreSQL.

## Secondary indexes

* Adding a secondary index on a column makes it more efficient to look up data in that column.
* secondary indexes can also help Cloud Spanner execute scans more efficiently, enabling index scans rather than full table scans.

# OpenCensus

**Create custom metrics with OpenCensus:**

Although Cloud Monitoring provides an API that supports defining and collecting custom metrics, it is a low-level, proprietary API. **OpenCensus provides an API that follows the style of the language community, along with an exporter that sends your metric data to Cloud Monitoring through the Monitoring API for you.**

OpenCensus also has **good support for application tracing;** see **OpenCensus Tracing** for a general overview. Cloud Trace recommends using OpenCensus for trace instrumentation. To **collect both metric and trace data** from your services, you can use a single distribution of libraries.

# Cloud SQL

Cloud SQL instances support MySQL 8.0, 5.7 (default), and 5.6, and provide up to 624 GB of RAM and 64 TB of data storage, with the option to automatically increase the storage size, as needed.

# Using an appropriate side-input pattern for data enrichment

In streaming analytics applications, it is common to enrich data with additional information that might be useful for further analysis. For example, if you have the storeId for a transaction, you might want to add information about the store location. You would typically add this additional information by taking an element and "denormalizing" it by bringing in information from a lookup table.

We recommend using the Apache Beam Side input patterns to handle slow updating side inputs.

# Using Apache Beam schemas to work with structured data

We recommend using **Apache Beam** schemas to make **processing structured data easier.**

Converting your objects to Rows lets you produce very clean Java code, which makes your **directed acyclic graph (DAG)** building exercise easier. You can also reference object properties as fields within the analytics statements that you create, instead of having to call methods.

# Cloud Data Loss Prevention (Cloud DLP)

There are two main types of transformations performed by Cloud DLP:

* recordTransformations
* infoTypeTransformations

Both **recordTransformations** and **infoTypeTransformations** methods can de-identify and encrypt sensitive information in your data. For example, you can transform the values in the US\_SOCIAL\_SECURITY\_NUMBER column to be unidentifiable or use tokenization to obscure it while keeping referential integrity.

The **infoTypeTransformations** method enables you to inspect for sensitive data and transform the finding. For example, if you have unstructured or free-text data, the infoTypeTransformations method can help you identify a SSN inside of a sentence and encrypt the SSN value while leaving the rest of the text intact. You can also define custom infoTypes methods.

The **recordTransformations** method enables you to apply a transformation configuration per field when using structured or tabular data. With the recordTransformations method, you can apply the same transformation across every value in that field such as hashing or tokenizing every value in a column with SSN column as the field or header name.

With the recordTransformations method , you can also mix in the infoTypeTransformations method that only apply to the values in the specified fields. For example, you can use an infoTypeTransformations method inside of a recordTransformations method for the field named comments to redact any findings for US\_SOCIAL\_SECURITY\_NUMBER that are found inside the text in the field.

In increasing order of complexity, the de-identification processes are as follows:

Redaction: Remove the sensitive content with no replacement of content.

Masking: Replace the sensitive content with fixed characters.

Encryption: Replace sensitive content with encrypted strings, possibly reversibly.

**Working with delimited data:**

Often, data consists of records delimited by a selected character, with fixed types in each column, like a **CSV file**. **For this class of data, you can apply de-identification transformations (recordTransformations) directly, without inspecting the data.** For example, you can expect a column labeled SSN to contain only SSN data. You don't need to inspect the data to know that the infoType detector is US\_SOCIAL\_SECURITY\_NUMBER. However, free-form columns labeled Additional Details can contain sensitive information, but the infoType class is unknown beforehand. **For a free-form column, you need to inspect the infoTypes detector (infoTypeTransformations) before applying de-identification transformations.** Cloud DLP allows both of these transformation types to co-exist in a single de-identification template. Cloud DLP includes more than 100 built-in infoTypes detectors. You can also create custom types or modify built-in infoTypes detectors to find sensitive data that is unique to your organization.

# Dialogflow Enterprise

Here’s a little more on what **Dialogflow** offers:

* **Conversational interaction powered by machine learning**: Dialogflow uses natural language processing to build conversational experiences faster and iterate more quickly. Provide a few examples of what a user might say and Dialogflow will build a unique model that can learn what actions to trigger and what data to extract so it provides the most relevant and precise responses to your users.
* **Build once and deploy everywhere**: Use Dialogflow to build a conversational app and deploy it on your website, your app or 32 different platforms, including the Google Assistant and other popular messaging services. Dialogflow also supports multiple languages and multilingual experiences so you can reach users around the world.
* **Advanced fulfillment options:** Fulfillment defines the corresponding action in response to whatever a user says, such as processing an order for a pizza or triggering the right answer to your user's question. Dialogflow allows you to connect to any webhook for fulfillment whether it's hosted in the public cloud or on-premises. Dialogflow’s integrated code editor allows you to code, test and implement these actions directly within Dialogflow's console.
* **Voice control with speech recognition**: Starting today, Dialogflow enables your conversational app to respond to voice commands or voice conversations. It's available within a single API call, combining speech recognition with natural language understanding.

The **enterprise edition** expands on all the benefits of Dialogflow, offering greater flexibility and support to meet the needs of large-scale businesses.

# Cloud Dataflow

**Streaming analytics now simpler, more cost-effective in Cloud Dataflow:**

# Google Cloud transfer options:

**Transfer Appliance** for moving offline data, large data sets, or data from a source with limited bandwidth

**BigQuery Data Transfer Service** to move data from SaaS applications to BigQuery.

**Transfer service for on-premises data** to move data from your on-premises machines to Cloud Storage.

# AutoML Vision Object Detection documentation:

**AutoML Vision Object Detection** enables developers to train custom machine learning models that are capable of detecting individual **objects** in a given image along with its bounding box and label.

The AutoML Vision Object Detection release includes the following features:

**Object localization** - Detects multiple objects in an image and provides information about the object and where the object was found in the image.

**API/UI** - **Provides an API** and custom user interface for importing your dataset from a Google Cloud Storage hosted CSV file and training images, for adding and removing annotations from imported images, for training and reviewing model evaluation metrics, and for using your model with online prediction.

## ParDo

ParDo is a Beam transform for generic parallel processing. ParDo is useful for common data processing operations, including:

* + Filtering a data set. You can use ParDo to consider each element in a PCollection and either output that element to a new collection, or discard it.
  + Formatting or type-converting each element in a data set.
  + Extracting parts of each element in a data set.
  + Performing computations on each element in a data set.