Answer Sheet

Q1)

You are designing storage for event data as part of building a data pipeline on Google Cloud. Your input data is in CSV format.

You want to minimize the cost of querying individual values over time windows.

Which storage service and schema design should you use?

✅ Use Cloud Bigtable for storage. Design tall and narrow tables, and use a new row for each single event version.
Explanation:-This option is correct as its an event data (time series) and need to be restricted to individual values over time windows, it is best to use Bigtable with tall and narrow tables.
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.
As an optimization, you can use s

● Use Cloud Bigtable for storage. Design short and wide tables, and use a new column for each single event version.
Explanation:-This option is incorrect as short and wide tables and are ideal for storing time series data.

● Use Cloud Storage for storage. Join the raw file data with a BigQuery log table.
Explanation:-This option is incorrect as you do not need to use GCS/BQ for this scenario.

● Use Cloud Storage for storage. Write a Cloud Dataprep job to split the data into partitioned tables.
Explanation:-This option is incorrect as you do not need to use GCS/BQ for this scenario.

Q2)

You are building a data pipeline on Google Cloud. You need to prepare source data for a machine-learning model.

This involves quickly deduplicating rows from three input tables and also removing outliers from data columns where you do not know the data distribution.

What should you do?

● Write an Apache Spark job with a series of steps for Cloud Dataflow. The first step will examine the source data, and the second and third steps step will perform data transformations.
Explanation:-This option is incorrect as we should not use Apache Spark and Cloud Dataflow or Cloud Dataproc for this scenario.

● Write an Apache Spark job with a series of steps for Cloud Dataproc. The first step will examine the source data, and the second and third steps step will perform data transformations
Explanation:-This option is incorrect as we should not use Apache Spark and Cloud Dataflow or Cloud Dataproc for this scenario.

● Use Cloud Dataprep to preview the data distributions in sample source data table columns. Write a recipe to transform the data and add it to the Cloud Dataprep job.
Explanation:-This option is incorrect as you can simply use the suggested transformations instead of writing custom recipe in Cloud Dataprep.

✅ Use Cloud Dataprep to preview the data distributions in sample source data table columns. Click on each column name, click on each appropriate suggested transformation, and then click 'Add' to add each transformation to the Cloud Dataprep job.
Explanation:-This is the correct option as the requirements is to prepare/clean source data, use Cloud Dataprep suggested transformations to quickly build a transformation job.
Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code. Cloud Dataprep automatica

---

Q3)

You are setting up Cloud Dataproc to perform some data transformations using Apache Spark jobs.

The data will be used for a new set of non-critical experiments in your marketing group.

You want to set up a cluster that can transform a large amount of data in the most cost-effective way. What should you do?

● Set up a cluster in High Availability mode with high-memory machine types. Add 10 additional local SSDs.
Explanation:-This option is incorrect as this scenario does not call for High Availability mode because it handles non-critical experiments.

● Set up a cluster in High Availability mode with default machine types. Add 10 additional Preemptible worker nodes.
Explanation:-This option is incorrect as this scenario does not call for High Availability mode because it handles non-critical experiments.

✅ Set up a cluster in Standard mode with high-memory machine types. Add 10 additional Preemptible worker nodes.
Explanation:-This option is correct as Dataproc is a managed service which handles Spark and Hadoop jobs and Spark and high-memory machines only need the Standard mode. Also, using Preemptible nodes provides cost-efficiency as this is not mission-critical.
Note: Preemptible instances can be used to lower your Compute Engine costs for Cloud Dataproc clusters, but do not change the way you are billed for the Cloud Dataproc premium.

● Set up a cluster in Standard mode with the default machine types. Add 10 additional local SSDs.
Explanation:-This option is incorrect as local SSDs would cost more; instead, use Preemptible nodes to meet your objective of delivering a cost-effective solution.

Q4)

You want to display aggregate view counts for your YouTube channel data in Data Studio.

You want to see the video tiles and view counts summarized over the last 30 days.

You also want to segment the data by the Country Code using the fewest possible steps. What should you do?

● Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter.
Explanation:-This option is correct as you cannot produce a summarized report that meets your business requirements using the options listed.

✅ Set up a YouTube data source for your channel data for Data Studio. Set Views as the metric and set Video Title and Country Code as report dimensions.
Explanation:-This option is correct as there is no need to export; you can use the existing YouTube data source. Country Code is a dimension because it's a string and should be displayed as such, that is, showing all countries, instead of filtering.
⚫ Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title as a report dimension. Set Country Code as a filter.
Explanation:-This option is correct as you do not need to export data from YouTube to Cloud Storage; you can simply use the existing YouTube data source.
⚫ Export your YouTube views to Cloud Storage. Set up a Cloud Storage data source for Data Studio. Set Views as the metric and set Video Title and Country Code as report dimensions.
Explanation:-This option is correct as you do not need to export data from YouTube to Cloud Storage; you can simply use the existing YouTube data source.

Q5)

Your company wants to try out the cloud with low risk. They want to archive approximately 100 TB of their log data to the cloud and test the analytics features available to them there, while also retaining that data as a long-term disaster recovery backup.

Which two steps should they take? (Choose two answers)

✅ Load logs into Google BigQuery.
Explanation:-This option is correct as Google Cloud Storage can provide long term archival option and BigQuery provides analytics capabilities.
⚫ Load logs into Google Cloud SQL.
Explanation:-This option is incorrect as Cloud SQL is relational database and does not support the capacity required as well as not suitable for long term archival storage.
⚫ Import logs into Google Stackdriver.
Explanation:-This option is incorrect as Stackdriver is a monitoring, logging, alerting and debugging tool. It is not ideal for long term retention of data and does not provide analytics capabilities.
⚫ Insert logs into Google Cloud Bigtable.
Explanation:-This option is incorrect as Bigtable is a NoSQL solution and can be used for analytics. However it is ideal for data with low latency access and is expensive.
✅ Upload log files into Google Cloud Storage.
Explanation:-This option is correct as Google Cloud Storage can provide long term archival option and BigQuery provides analytics capabilities.

Q6)

A company wants to transfer petabyte scale of data to Google Cloud for their analytics, however are constrained on their internet connectivity?

Which GCP service can help them transfer the data quickly?

⚫ Transfer appliance and Dataprep to decrypt the data
Explanation:-This option is incorrect as Dataprep does not help is decrypting the data.
⚫ Google Transfer service using multiple VPN connections
Explanation:-This option is incorrect is wrong as Google Transfer Service does not support importing data from on-premises data center. It only supports online imports.
⚫ gustil with multiple VPN connections
Explanation:-This option is incorrect as the data is huge transferring it with gsutil would take a long time.
✅ Transfer appliance and rehydrator to decrypt the data
Explanation:-This option is correct as the data is huge it should be transferred using Transfer Appliance and use a Rehydrator to decrypt the data. Once you capture your data onto the Google Transfer Appliance, ship the appliance to the Google upload facility for rehydration. Data rehydration is the process by which you fully reconstitute the files so you can access and use the transferred data.
To rehydrate data, the data is first copied from the Transfer Appliance to your Cloud Storage staging bu

---

Q7)

A company has lot of data sources from multiple systems used for reporting.

Over a period of time, a lot data is missing and you are asked to perform anomaly detection.

How would you design the system?

✅ Load in Cloud Storage and use Dataprep with Data Studio
Explanation:-This option is correct as Dataprep provides data cleaning and automatically identifies anomalies in the data. It can integrated with Cloud Storage and BigQuery. Such that -
1. Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code.
2. Cloud Dataprep is an integrated partner service operated by Trifacta and based on their industry-leading data preparation solution. Google works clo sely with Trifacta to provide a seamless user experience that removes the need for up-front software installation, separate licensing costs, or ongoing operational overhead. Cloud Dataprep is fully managed and scales on-demand to meet your growing data preparation needs so you can stay focused on analysis. Refer: https://cloud.google.com/dataprep
⚫ Use Dataprep with Data Studio
Explanation:-This option is incorrect as Dataprep would not be able interact directly with local system.
⚫ Load in Cloud Storage and use Dataflow with Data Studio
Explanation:-This option is incorrect as Cloud Dataflow is a fully-managed service for transforming and enriching data in stream (real time) and batch (historical) modes with equal reliability and expressiveness -- no more complex workarounds or compromises needed. It does not provide anomaly detection.
⚫ Use Dataflow with Data Studio
Explanation:-This option is incorrect as Cloud Dataflow is a fully-managed service for transforming and enriching data in stream (real time) and

batch (historical) modes with equal reliability and expressiveness -- no more complex workarounds or compromises needed. It does not provide anomaly detection.

Q8)

Your company plans to migrate a multi-petabyte data set to the cloud. The data set must be available 24hrs a day.

Your business analysts have experience only with using a SQL interface.

How should you store the data to optimize it for ease of analysis?

- Stream data into Google Cloud Datastore.

Explanation:-This option is incorrect as Datastore does not provide a SQL interface and is a NoSQL solution.

- Put flat files into Google Cloud Storage.

Explanation:-This option is incorrect as Cloud Storage does not provide SQL interface.

- Insert data into Google Cloud SQL.

Explanation:-This option is incorrect as Cloud SQL cannot support multi-petabyte data. Storage limit for Cloud SQL is 10TB

- ✅ Load data into Google BigQuery.

Explanation:-This option is correct as BigQuery is the only of these Google products that supports an SQL interface and a high enough SLA (99.9%) to make it readily available.

Q9)

Your company hosts its data into multiple Cloud SQL databases. You need to export your Cloud SQL tables into BigQuery for analysis.]

How can the data be exported?

- Convert your Cloud SQL data to JSON format, then import directly into BigQuery

Explanation:-This option is incorrect as they are not supported options.

- ✅ Export your Cloud SQL data to Cloud Storage, then import into BigQuery

Explanation:-This option is correct as BigQuery does not provide direct load from Cloud SQL. The data needs to be loaded through Cloud Storage. There are many situations where you can query data without loading it. For all other situations, you must first load your data into BigQuery before you can run queries.

- Import data to BigQuery directly from Cloud SQL.

Explanation:-This option is incorrect as they are not supported options.

- Use the BigQuery export function in Cloud SQL to manage exporting data into BigQuery.

Explanation:-This option is incorrect as they are not supported options.

Q10)

Your BigQuery table needs to be accessed by team members who are not proficient in technology.

You want to simplify the columns they need to query to avoid confusion.

How can you do this while preserving all of the data in your table?

- Train your team members on how to query larger tables.

Explanation:-This option is incorrect as it is not a feasible solution.

- ✅ Create a query that uses the reduced number of columns they will access. Save this query as a view in a different dataset. Give your team members access to the new dataset and instruct them to query against the saved view instead of the main table.

Explanation:-This option is correct as the best way to limit and expose number of columns and access is to create a View. With BigQuery, the access can only be controlled on Datasets and Views, but not on tables.

- Apply column filtering to your table, and restrict the unfiltered view to yourself and those who need access to the full table.

Explanation:-This option is incorrect as column filtering cannot be applied to Table and it can be done through Views.

- Create a copy of your table in a different dataset, and remove the unneeded columns from the copy. Have your team members run queries against this copy.

Explanation:-This option is incorrect as it is not an ideal solution, as it results in duplication of data. Also, deletion of Columns is not supported.

Q11)

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?

- Cloud SQL

Explanation:-This option is incorrect as Cloud SQL would need infrastructure scaling. Although storage can be automatically scaled (upto a limit), instance type needs to be changed as per the load manually.

- BigQuery

Explanation:-This option is incorrect as BigQuery is an data warehousing option.

- Cloud Bigtable

Explanation:-This option is incorrect as Bigtable is not a relational database but an NoSQL option.

- ✅ Cloud Datastore

Explanation:-This option is correct as the payment transactions would need a transactional data service Datastore can support the same. Also it is fully managed with NoOps required.

Q12)

Your infrastructure includes two 100-TB enterprise file servers. You need to perform a one-way, one-time migration of this data to the Google Cloud securely.

Only users in Germany will access this data. You want to create the most cost-effective solution. What should you do?

- ✅ Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.
- Use Transfer Appliance to transfer the offsite backup files to a Cloud Storage Multi-Regional bucket as a final destination.
- Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Regional storage bucket as a final destination.
- Use Storage Transfer Service to transfer the offsite backup files to a Cloud Storage Multi-Regional storage bucket as a final destination.

Q13)

Your company is planning to migrate their analytics data into BigQuery. There is a need to handle both batch and streaming data.

You are assigned the role to determine the costs that would be incurred for different operations.

What are all of the BigQuery operations that Google charges for?

- ✅ Storage, queries, and streaming inserts.
- Storage, queries, and loading data from a file.
- Storage, queries, and exporting data.
- Queries and streaming inserts.

Q14)

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?

- Create a Google Cloud Dataflow job to process the data.

Explanation:-This option is incorrect as Dataflow is not suited to execute Hadoop jobs.

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

Explanation:-This option is incorrect as HDFS is associated with the Cluster. If the cluster is terminated, the data would be lost.

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

Explanation:-This option is incorrect as Cluster on Compute Engine would increase infrastructure management and persistent disks would not provide scalability.

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

Explanation:-This option is correct as the requirement is to reuse Hadoop jobs with minimizing the infrastructure management with the ability to store data in a durable external storage, Dataproc with Cloud Storage would be an ideal solution.
Cloud Dataproc is a fast, easy-to-use, low-cost and fully managed service that lets you run the Apache Spark and Apache Hadoop ecosystem on Google Cloud Platform. Cloud Dataproc provisions big or small clusters rapidly, supports many popular job types, and is integ

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

Explanation:-This option is incorrect as Cluster on Compute Engine would increase infrastructure management and Local SSDs would not provide data durability.

Q15)

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?

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

Explanation:-This option is incorrect as the person column needs to be UNNEST for the nested city field to be used directly in the WHERE clause. Also, note this is standard SQL query by the reference of the table.
#standardSQL SELECT page.title FROM `bigquery-public-data.samples.github_nested`, UNNEST(payload.pages) AS page WHERE page

- Change "person" to "person.city".

Explanation:-This option is incorrect.

- Change "person" to "city.person".

Explanation:-This option is incorrect.

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

Explanation:-This option is incorrect.

Q16)

You have a Dataflow job that you want to cancel. It is a streaming IoT pipeline, and you want to ensure that any data that is in-flight is processed and written to the output with no data loss.

Which of the following commands can you use on the Dataflow monitoring console to stop the pipeline job?

- Cancel

Explanation:-This option is incorrect as Cancel does not handle in-flight messages and it might result in data loss.

- ✅ Drain

Explanation:-This option is correct as Drain command helps Dataflow process and complete in-flight messages and stops accepting any new ones. If you need to stop a running Cloud Dataflow job, you can do so by issuing a command using either the Cloud Dataflow Monitoring Interface or the Cloud Dataflow Command-line Interface. There are two possible commands you can issue to stop your job: Cancel and Drain.

Note: The Drain command is supported for streaming pipelines only.
Using the Drain option
Stop
Explanation:-This option is incorrect as Stop and Pause option do not exist.
Pause
Explanation:-This option is incorrect as Stop and Pause option do not exist.

Q17)

You currently have a Bigtable instance you've been using for development running a development instance type, using HDD's for storage. You are ready to upgrade your development instance to a production instance for increased performance. You also want to upgrade your storage to SSD's as you need maximum performance for your instance.

What should you do?

Build a Dataflow pipeline or Dataproc job to copy the data to the new cluster with SSD storage type.
Explanation:-This option is correct as the storage for the cluster cannot be updated. You need to define the new cluster and copy or import the data to it.
Switching between SSD and HDD storage
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 use a Dataflow or Hadoop MapReduce job to copy the data from one instance to another. Keep in mind that migrating an entire instance takes time, and you might need to add nodes to your Cloud Bigtable clusters before you migrate your instance. Refer: https://cloud.google.com/bigtable/docs/choosing-ssd-hdd
Upgrade your development instance to a production instance, and switch your storage type from HDD to SSD.
Explanation:-This option is incorrect as storage type cannot be changed.
Run parallel instances where one instance is using HDD and the other is using SSD.
Explanation:-This option is incorrect as it would have two clusters running at the same time with same data, thereby increasing cost.
Use the Bigtable instance sync tool in order to automatically synchronize two different instances, with one having the new storage configuration.
Explanation:-This option is incorrect as it would have two clusters running at the same time with same data, thereby increasing cost.

Q18)

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?

Rewrite the job in Pig.
Explanation:-This option is incorrect as Pig is wrapper and would initiate Map Reduce jobs
Rewrite the job in Apache Spark.
Explanation:-This option is correct as Spark can improve the performance as it performs lazy in-memory execution.
Spark is important because it does part of its pipeline processing in memory rather than copying from disk. For some applications, this makes Spark extremely fast. With a Spark pipeline, you have two different kinds of operations, transforms and actions. Spark builds its pipeline used an abstraction called a directed graph.
Increase the size of the Hadoop cluster
Explanation:-This option is incorrect as it would increase the cost.
Decrease the size of the Hadoop cluster but also rewrite the job in Hive.
Explanation:-This option is incorrect as Hive is wrapper and would initiate Map Reduce jobs. Also, reducing the size would reduce performance.

Q19)

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?

Create a view in BigQuery that concatenates the FirstName and LastName field values to produce the FullName.
Explanation:-This option is correct as it is better to create materialized tables instead of views as the query would be executed everytime. Refer BigQuery Best Practices
Best practice: If possible, materialize your query results in stages.
If you create a large, multi-stage query, each time you run it, BigQuery reads all the data that is required by the query. You are billed for all the data that is read each time the query is run.
Instead, break your query into stages where each stage materi
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.
Explanation:-This option is incorrect as DML are limited by quotas. Maximum number of combined UPDATE, DELETE, and MERGE statements per day per table — 200
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.
Explanation:-This option is correct as the best option is to create a new table with the updated columns. Dataflow provides a serverless NoOps option to convert data.
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.

Explanation:-This option is incorrect as Dataproc would need provisioning of servers and writing scripts.

Q20)

A company's BigQuery data is currently stored in external CSV files in Cloud Storage.

As the data has increased over the period of time, the query performance has dropped.

What steps can help improve the query performance maintaining the cost-effectiveness?

✅ Import the data into BigQuery for better performance.
Explanation:-This option is correct as the performance issue is because the data is stored in a non-optimal format in an external storage medium. Query performance for external data sources may not be as high as querying data in a native BigQuery table. If query speed is a priority, load the data into BigQuery instead of setting up an external data source. The performance of a query that includes an external data source depends on the external storage type. For example, querying data stored in Cloud Sto

⚪ Request more slots for greater capacity to improve performance.
Explanation:-This option is incorrect as there is feature to request more slots.

⚪ Divide the data into partitions based on date.
Explanation:-This option is incorrect as partitioning of data at source would not improve query time for all use cases.

⚪ Time to move to Cloud Bigtable; it is faster in all cases.
Explanation:-This option is incorrect as Bigtable is more ideal for NoSQL data type and can get very expensive

---

Q21)

A client is using Cloud SQL database to serve infrequently changing lookup tables that host data used by applications.

The applications will not modify the tables. As they expand into other geographic regions they want to ensure good performance.

What do you recommend?

⚪ Migrate to Cloud Spanner
Explanation:-This option is incorrect as Cloud Spanner is suitable for read/write operations, as the requirement is mainly for read, read replicas would be best fit.

✅ Read replicas
Explanation:-This option is correct as read replica will increase the availability of the service and can be located closer to the users in the new geographies. Cloud SQL provides the ability to replicate a master instance to one or more read replicas. A read replica is a copy of the master that reflects changes to the master instance in almost real time.

⚪ Instance high availability configuration
Explanation:-This option is incorrect as high availability is for failover and not for performance.

⚪ Migrate to Cloud Storage
Explanation:-This option is incorrect as Cloud Storage is not ideal storage for relational data.

Q22)

A company wants to connect cloud applications to an Oracle database in its data center.

Requirements are a maximum of 9 Gbps of data and a Service Level Agreement (SLA) of 99%.

Which option best suits the requirements?

⚪ Implement a high-throughput Cloud VPN connection
Explanation:-This option is incorrect as Cloud VPN is over the internet through IPSec VPN at a low cost for your data bandwidth needs up to 3.0 Gbps.

⚪ Cloud Router with VPN
Explanation:-This option is incorrect as Cloud Router helps only in dynamic routing.

⚪ Dedicated Interconnect
Explanation:-This option is incorrect as Dedicated Interconnect is suitable for High bandwidth connections with a minimum of 10 Gbps. Traffic flows directly between networks, not through the public Internet.

✅ Partner Interconnect
Explanation:-This option is correct as Partner Interconnect is useful for data up to 10 Gbps and is offered by ISPs with SLAs.
Flexible capacity options with a minimum of 50 Mbps. More points of connectivity through one of our supported service providers. Traffic between networks flows through a service provider, not through the public Internet.
Google provides an SLA for the connection between Google and service provider. Whether an end-to-end SLA for the connection is offered, depends on your ser

---

Q23)

An organization wishes to enable real time analytics on user interactions on their web application.

They estimate that there will be 1000 interactions per second and wishes to use services, which are ops free.

Which combination of services can be used in this case?

⚪ App Engine, Dataproc, DataStudio
Explanation:-This option is incorrect as Dataflow and Dataproc would need processing and storage.

⚪ Compute Engine, BigQuery Streaming Inserts, DataStudio
Explanation:-This option is incorrect as Compute Engine would not be Ops free.

✅ App Engine, BigQuery Streaming Inserts, DataStudio
Explanation:-This option is correct as the focus is more on NoOps, the App Engine can be used to capture and insert the data into BigQuery using streaming inserts. The data can then be analyzed and visualized using DataStudio.

⚪ App Engine, Dataflow, DataStudio

Explanation:-This option is incorrect as Dataflow and Dataproc would need processing and storage.

Q24)

A startup plans to use a data processing platform, which supports both batch and streaming applications.

They would prefer to have a hands-off/serverless data processing platform to start with.

Which GCP service is suited for them?

- Dataproc

Explanation:-This option is incorrect as Google Cloud Dataproc is a fast, easy to use, managed Spark and Hadoop service for distributed data processing. It is not serverless and more suited for batch processing.

- Dataprep

Explanation:-This option is incorrect as Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis. It does not help process batch and streaming data.

- ✅ Dataflow

Explanation:-This option is correct as Dataflow helps design data processing pipelines and is a fully managed service for strongly consistent, parallel data-processing pipelines. It provides an SDK for Java with composable primitives for building data-processing pipelines for batch or continuous processing. This service manages the life cycle of Google Compute Engine resources of the processing pipeline(s). It also provides a monitoring user interface for understanding pipeline health.

Cloud Dataflo

- BigQuery

Explanation:- This option is incorrect as BigQuery is an analytics data warehousing solution.

Q25)

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.

How should you design the system in Google Cloud?

- Send the data to Google Cloud Datastore and then export to BigQuery.
- ✅ Send the data to Google Cloud Pub/Sub, stream Cloud Pub/Sub to Google Cloud Dataflow, and store the data in Google BigQuery.
- Send the data to Cloud Storage and then spin up an Apache Hadoop cluster as needed in Google Cloud Dataproc whenever analysis is required.
- 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.

Q26)

Your company is in a highly regulated industry. You have 2 groups of analysts, who perform the initial analysis and sanitization of the data.

You now need to provide analyst three secure access to these BigQuery query results, but not the underlying tables or datasets.

How would you share the data?

- Export the query results to a public Cloud Storage bucket.

Explanation:-This option is incorrect as a public Cloud Storage bucket is accessible to all.

- ✅ Create a BigQuery Authorized View and assign a project-level user role to analyst three.

Explanation:-This option is correct as you need to copy or store the query results in a separate dataset and provide authorization to view and/or use that dataset. The other solutions are not secure. 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 a

- Assign the bigquery.resultsonly.viewer role to analyst three.

Explanation:-This option is incorrect as there is no resultsonly viewer role.

- Create a BigQuery Authorized View and assign an organizational level role to analyst three.

Explanation:-This option is incorrect as an Organizational role would provide access to the underlying data as well.

Q27)

Your company is making the move to Google Cloud and has chosen to use a managed database service to reduce overhead.

Your existing database is used for a product catalog that provides real-time inventory tracking for a retailer. Your database is 500 GB in size.

The data is semi-structured and does not need full atomicity. You are looking for a truly no-ops/serverless solution.

What storage option should you choose?

- ✅ Cloud Datastore
- Cloud Bigtable
- Cloud SQL
- BigQuery

Q28) Which of these numbers are adjusted by a neural network as it learns from a training dataset? (Choose two)

- Continuous features

- Input values
- ✅ Weights
- ✅ Biases

Q29)

A user wishes to generate reports on petabyte scale data using a Business Intelligence (BI) tools.

Which storage option provides integration with BI tools and supports OLAP workloads up to petabyte-scale?

- Bigtable
- Cloud Datastore
- Cloud Storage
- ✅ BigQuery

Q30)

Your company is planning to migrate their historical dataset into BigQuery.

This data would be exposed to the data scientists for perform analysis using BigQuery ML.

The data scientists would like to know which ML models does the BigQuery ML support. What would be your answer? (Choose 2)

- K Means

Explanation:-This option is Incorrect.

- Principal Component Analysis

Explanation:-This option is Incorrect.

- ✅ Multiclass logistic regression for Classification

Explanation:-This option is correct as BigQuery ML supports Linear regression, Binary Logistic regression and Multiclass logistic regression.
BigQuery ML currently supports the following types of models - 1. Linear regression — These models can be used for predicting a numerical value.
2. Binary logistic regression — These models can be used for predicting one of two classes (such as identifying whether an email is spam).
3. Multiclass logistic regression for classification — These models can be used to predict multiple possible values such as whether an input is "low-value," "medium-value," or "high-value." Labels can have up to 50 unique values. In BigQuery ML, multiclass logistic regression training uses a multinomial classifier with a cross-entropy loss function.
Refer: https://cloud.google.com/bigquery-ml/docs/introduction

- Random Forest

Explanation:-This option is Incorrect.

- ✅ Linear Regression

Explanation:-This option is correct as BigQuery ML supports Linear regression, Binary Logistic regression and Multiclass logistic regression.
BigQuery ML currently supports the following types of models - 1. Linear regression — These models can be used for predicting a numerical value.
2. Binary logistic regression — These models can be used for predicting one of two classes (such as identifying whether an email is spam).
3. Multiclass logistic regression for classification — These models can be used to predict multiple possible values such as whether an input is "low-value," "medium-value," or "high-value." Labels can have up to 50 unique values. In BigQuery ML, multiclass logistic regression training uses a multinomial classifier with a cross-entropy loss function.
Refer: https://cloud.google.com/bigquery-ml/docs/introduction

---

Q31)

Your company wants to develop an REST based application for text analysis to identify entities and label by types such as person, organization, location, events, products, and media from within a text.

You need to do a quick Proof of Concept (PoC) to implement and demo the same. How would you design your application?

- Create and Train a model using Tensorflow and Develop an REST based wrapper over it

Explanation:-This option is incorrect as they do not provide quick results.

- Create and Train a model using BigQuery ML and Develop an REST based wrapper over it

Explanation:-This option is incorrect as they do not provide quick results.

- ✅ Use Cloud Natural Language API and Develop an REST based wrapper over it

Explanation:-This option is correct as the solution needs to developed quickly, the Cloud Natural Language API can be used to perform text analysis.
Cloud Natural Language API reveals the structure and meaning of text by offering powerful machine learning models in an easy-to-use REST API. And with AutoML Natural Language Beta you can build and train ML models easily, without extensive ML expertise. You can use Natural Language to extract information about people, places, events, and much more mentioned

- Use Cloud Vision API and Develop an REST based wrapper over it

Explanation:-This option is incorrect as Cloud Vision is for image analysis and not text analysis.

Q32)

Your company wants to transcribe the conversations between the manufacturing employees at real time.

The conversations are recorded using old radio systems in the 8000Hz frequency. They are in English with a short duration of 35-40 secs.

You need to design the system inline with Google recommended best practice. How would you design the application?

- ✅ Use Cloud Speech-to-Text API in synchronous mode

Explanation:-This option is correct as Speech-to-Text can be used to convert short duration audio in synchronous calls. As well as it is recommended not to re-sample the data, if it is coming at a lower sampling rate from the source.
Lower sampling rates may reduce accuracy. However, avoid re-sampling. For example, in telephony the native rate is commonly 8000 Hz, which is the rate that should be sent to the service.

Synchronous speech recognition returns the recognized text for short audio (less t
- Use Cloud Speech-to-Text API in asynchronous mode

Explanation:-This option is incorrect.
- Re-sample the audio using 16000Hz frequency and Use Cloud Speech-to-Text API in synchronous mode

Explanation:-This option is incorrect.
- Re-sample the audio using 16000Hz frequency and Use Cloud Speech-to-Text API in asynchronous mode

Explanation:-This option is incorrect.

Q33)

You have lot of Spark jobs. Some jobs need to run independently while others can run parallelly.

There is also inter-dependency between the jobs and the dependent jobs should not be triggered unless the previous ones are completed.

How do you orchestrate the pipelines?

- Cloud Dataproc

Explanation:-This option is incorrect as Google Cloud Dataproc is a fast, easy to use, managed Spark and Hadoop service for distributed data processing. It does not help easy orchestration.
- Cloud Scheduler

Explanation:-This option is incorrect as Cloud Scheduler is a fully managed enterprise-grade cron job scheduler. It is not an orchestration tool.
- Schedule jobs on a single Compute Engine using Cron.

Explanation:-This option is incorrect as it does not help orchestrate the dependency between jobs, but merely schedule them.
- ✅ Cloud Composer

Explanation:-This option is correct as Cloud Composer can help create workflows that connect data, processing, and services across clouds, giving you a unified data environment.
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. Built on the popular Apache Airflow open source project and operated using the Python programming language, Cloud Composer is free from lock-in and ea

---

Q34)

Your company is planning to host its analytics data in BigQuery. You are required to control access to the dataset with least privilege meeting the following guidelines

Each team has multiple Team Leaders, who should have the ability to create, delete tables, but not delete dataset.

Each team has Data Analysts, who should be able to query data, but not modify itHow would you design the access control?

- Grant Team leader group - OWNER and Data Analyst - WRITER
- Grant Team leader group - OWNER and Data Analyst - READER
- ✅ Grant Team leader group - WRITER and Data Analyst - READER
- Grant Team leader group - READER and Data Analyst - WRITER

Q35)

Your company wants to develop a system to measure the feedback of their products from the reviews posted by people on various Social media platforms. The reviews are mainly text based.

You need to do a quick Proof of Concept (PoC) to implement and demo the same. How would you design your application?

- Create and Train a sentiment analysis model using Tensorflow

Explanation:-This option is incorrect as building and training a senetiment analysis model using Tensorflow would take time and effort.
- Use Cloud Speech-to-Text API for sentiment analysis

Explanation:-This option is incorrect as Speech-to-Text API is for audio to text conversion.
- ✅ Use Cloud Natural Language API for sentiment analysis

Explanation:-This option is correct as Natural Language processing provides pre-model to perform sentiment analysis.
You can use Cloud Natural Language to extract information about people, places, events, and much more mentioned in text documents, news articles, or blog posts. You can use it to understand sentiment about your product on social media or parse intent from customer conversations happening in a call center or a messaging app. You can analyze text uploaded in your request or integrate with y
- Use Cloud Vision API for sentiment analysis

Explanation:-This option is incorrect as Cloud Vision is for image analysis.

Q36)

Your company receives a lot of financial data in CSV files. The files need to be processed, cleaned and transformed before they are made available for analytics.

The schema of the data also changes every third month. The Data analysts should be able to perform the tasks

1. No prior knowledge of any language with no coding

2. Provided a GUI tool to build and modify the schema

What solution best fits the need?

- Use Dataflow code and provide Data Analysts the access to the code. Store the schema externally to be easily modified.

Explanation:-This option is incorrect as they would need programming knowledge.
- ✅ Use Dataprep with transformation recipes.

Explanation:-This option is correct as Dataprep can be used to handle schema changes by Data Analysts without any programming knowledge, but through an easy to use GUI.
Cloud Dataprep by Trifacta is an intelligent data service for visually exploring, cleaning, and preparing structured and unstructured data for analysis.

Cloud Dataprep is serverless and works at any scale. There is no infrastructure to deploy or manage. Easy data preparation with clicks and no code. Visually explore and interact wit

- Use Dataproc spark and provide Data Analysts the access to the code. Store the schema externally to be easily modified.

Explanation:-This option is incorrect as they would need programming knowledge.

- Use DataLab with transformation recipes.

Explanation:-This option is incorrect as they would need programming knowledge.

Q37)

Your company has assigned fixed number for slots to each project for BigQuery.

Each project wants to monitor the number of available slots. How can the monitoring be configured?

- Monitor the BigQuery Slots Used metric

Explanation:-This option is incorrect.

- Monitor the BigQuery Slots Pending metric

Explanation:-This option is incorrect.

- Monitor the BigQuery Slots Allocated metric

Explanation:-This option is incorrect.

- ✅ Monitor the BigQuery Slots Available metric

Explanation:-This option is correct as BigQuery provides 2 metrics for Slots. Slots Allocated to the project and Slots Available for the project. BigQuery Slots available slots Total number of slots available to the project. If the project shares a reservation of slots with other projects the slots being used by the other projects is not depicted.

Q38)

A company has migrated their Hadoop cluster to the cloud and is now using Cloud Dataproc with the same settings and same methods as in the data center.

What would you advise them to do to make better use of the cloud environment?

- Upgrade to the latest version of HDFS. Change the settings in Hadoop components to optimize for the different kinds of work in the mix.

Explanation:-This option is incorrect.

- Find more jobs to run so the cluster utilizations will cost-justify the expense.

Explanation:-This option is incorrect.

- ✅ Store persistent data off-cluster. Start a cluster for one kind of work then shut it down when it is not processing data.

Explanation:-This option is correct as Storing persistent data off the cluster allows the cluster to be shut down when not processing data. And it allows separate clusters to be started per job or per kind of work, so tuning is less important.

1. Direct data access – Store your data in Cloud Storage and access it directly, with no need to transfer it into HDFS first.
2. HDFS compatibility – You can easily access your data in Cloud Storage using the gs:// prefix instead of hdfs://.
3. Interoperabi

- Migrate from Cloud Dataproc to an open source Hadoop Cluster hosted on Compute Engine, because this is the only way to get all the Hadoop customizations needed for efficiency.

Explanation:-This option is incorrect.

Q39)

Your company processes high volumes of IoT data that are time-stamped. The total data volume can be several petabytes.

The data needs to be written and changed at a high speed. You want to use the most performant storage option for your data.

Which product should you use?

- Cloud Datastore

Explanation:-This option is incorrect as Cloud Datastore is not the most performant product for frequent writes or timestamp-based queries.

- Cloud Storage

Explanation:- This option is incorrect as Cloud Storage is designed for object storage not for this type of data ingestion and collection.

- ✅ Cloud Bigtable

Explanation:-This option is correct as Cloud Bigtable is the most performant storage option to work with IoT and time series data. Google Cloud Bigtable is a fast, fully managed, highly-scalable NoSQL database service. It is designed for the collection and retention of data from 1TB to hundreds of PB.

- BigQuery

Explanation:- This option is incorrect as BigQuery is more of an a scalable, fully managed enterprise data warehousing solution and not ideal fast changing data.

Q40)

You are asked to design next generation of smart helmet for accident detection and reporting system. Each helmet will push 10kb of biometric data In JSON format every 1 second to a collection platform that will process and use trained machine learning model to predict and detect if an accident happens and send notification.

Management has tasked you to architect the platform ensuring the following requirements are met:

· Provide the ability for real-time analytics of the inbound biometric data

· Ensure ingestion and processing of the biometric data is highly durable. Elastic and parallel

· The results of the analytic processing should be persisted for data mining to improve the accident detection ML model in the future.

Which architecture outlined below win meet the initial requirements for the platform?

- Utilize Cloud Storage to collect the inbound sensor data, analyze data with Dataproc and save the results to BigQuery.

Explanation:-This option is incorrect as Cloud Storage is not an ideal ingestion service for real time high frequency data. Also Dataproc is a fast, easy-to-use, fully-managed cloud service for running Apache Spark and Apache Hadoop clusters in a simpler, more cost-efficient way.

✅ Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to BigQuery.

Explanation:-This option is correct as Cloud Pub/Sub provides elastic and scalable ingestion, Dataflow provides processing and BigQuery analytics.

Google Cloud Pub/Sub provides a globally durable message ingestion service. By creating topics for streams or channels, you can enable different components of your application to subscribe to specific streams of data without needing to construct subscriber-specific channels on each device. Cloud Pub/Sub also natively connects to other Cloud Platform services,

⚪ Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Cloud SQL.

Explanation:-This option is incorrect as Cloud SQL is a relational database and not suited for analytics data storage.

⚪ Utilize Cloud Pub/Sub to collect the inbound sensor data, analyze the data with Dataflow and save the results to Bigtable.

Explanation:-This option is incorrect as Bigtable is not ideal for long term analytics data storage.

Q41)

You are designing storage for CSV files and using an I/O-intensive custom Apache Spark transform as part of deploying a data pipeline on Google Cloud.

You are using ANSI SQL to run queries for your analysts. You want to support complex aggregate queries and reuse existing code.

How should you transform the input data?

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

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

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

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

Q42)

You have migrated your Hadoop jobs with external dependencies on a Dataproc cluster.

As a security requirement, the cluster has been setup using internal IP addresses only and does not have a direct Internet connectivity.

How can the cluster be configured to allow the installation of the dependencies?

⚪ Setup a SSH tunnel to Internet and route outbound requests through it.

Explanation:-These options are incorrect as they would not allow secure outbound connection.

✅ Store the external dependencies in Cloud Storage and modify the initialization scripts

Explanation:-This option is correct as the Dataproc cluster is configured with internal IP addresses only, the dependencies can be stored in Cloud Storage so that they can be accessed using internal IPs.

If you create a Cloud 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

⚪ Setup a SOCKS proxy and route outbound requests through it.

Explanation:-These options are incorrect as they would not allow secure outbound connection.

⚪ Setup the Dataproc master node is public subnet to be able to download external dependencies

Explanation:-These options are incorrect as they would not allow secure outbound connection.

Q43)

Your company hosts a 2PB on-premises Hadoop cluster with sensitive data.

They want to plan the migration of the cluster to Google Cloud as part of phase 1 activity before the jobs are moved.

Current network speed between the colocation and cloud is 10Gbps. What is the efficient way to transfer the data?

✅ Use Transfer appliance to transfer the data to Cloud Storage

⚪ Expose the data as a public URL and Storage Transfer Service to transfer it

⚪ Use gsutil command to transfer the data to Cloud Storage

⚪ Use hadoop distcp command to copy the data between cluster

Q44)

Your company is planning to migrate its data first to Google Cloud Storage.

You need to keep the contents of this bucket in sync with a new Google Cloud Storage bucket to support a backup storage destination.

What is the best method to achieve this?

⚪ Once per week, use a gsutil cp command to copy over newly modified files.

Explanation:-This option is incorrect as copy can be used to copy, however there needs to be more handling to keep it in sync.

✅ Use gsutil rsync commands to keep both locations in sync.

Explanation:-This option is correct as the data transfer is between on-premises and Google Cloud, the gsutil rsync can be used to keep the source and destination in sync.

gsutil rsync command makes the contents under dst_url the same as the contents under src_url, by copying any missing files/objects (or those whose data has changed), and (if the -d option is specified) deleting any extra files/objects. src_url must specify a directory, bucket, or bucket subdirectory.

⚪ Use Storage Transfer Service to keep both the source and destination in sync.

Explanation:-This option is incorrect as the data is not available in an online location.

Use gsutil -m cp to keep both locations in sync.
Explanation:-This option is incorrect as copy can be used to copy, however there needs to be more handling to keep it in sync.

Q45)

You have a real time data processing pipeline running in Dataflow.

As a part of changed requirement you need to update the windowing and triggering strategy for the pipeline.

You want to update the pipeline without any loss of in-flight messages. What is the best way to deploy the changes?

Stop with pipeline using the drain option and use new Dataflow pipeline
Explanation:-This option is incorrect as with Drain option the windows and triggers would closed immediately. When you issue the Drain command, Cloud Dataflow immediately closes any in-process windows and fires all triggers. The system does not wait for any outstanding time-based windows to finish. For example, if your pipeline is ten minutes into a two-hour window when you issue the Drain command, Cloud Dataflow won't wait for the remainder of the window to finish. It will close the window immediately with
Stop with pipeline using the cancel option and use new Dataflow pipeline
Explanation:-This option is incorrect as Cancel immediately halts processing, you may lose any "in-flight" data.
✅ Pass the --update option with --jobname parameter to the same name as the job you want to update
Explanation:-This option is correct as Dataflow allows updates to the existing pipeline in case of compatible changes while saving the intermediate state data.

When you update a job on the Cloud Dataflow service, you replace the existing job with a new job that runs your updated pipeline code. The Cloud Dataflow service retains the job name, but runs the replacement job with an updated jobId.

The replacement job preserves any intermediate state data from the prior job, as well as any buffer
Pass the --update option with --jobname parameter to the new job name you want to update
Explanation:-This option is incorrect as the job name should be the same.

Q46)

Your company is working on real time click stream analysis.

They want to implement a feature to capture user click during a session and aggregate the count for that session. Session timeout is 30 mins.

How would you design the data processing?

Use Dataflow and fixed windowing of 30 minutes
Explanation:-This option is incorrect as Fixed and Global windowing would not work.
✅ Use Dataflow and Session windowing with gap duration of 30 minutes
Explanation:-This option is correct as Dataflow would help in performing real time analytics and data count aggregation over a window. Session windows to track the session for the aggregate click count by the user.
A session window function defines windows that contain elements that are within a certain gap duration of another element. Session windowing applies on a per-key basis and is useful for data that is irregularly distributed with respect to time. For example, a data stream representing user mou
Use Dataflow and Global window with gap duration of 30 minutes
Explanation:-This option is incorrect as Fixed and Global windowing would not work.
Use Dataproc and store the data in BigQuery and aggregate the same
Explanation:-This option is incorrect as Dataproc and BigQuery would not provide real time analytics.

Q47)

Your company is building a package tracking application to track the complete lifecycle of the package.

The data is stored in a BigQuery time partitioned table.

Over the period of time the data in the table and grown manifold and Data Scientists are complaining of slowness in their package tracking queries.

How can the table be modified to improve the performance and maintaining cost effectiveness?

Import the table data to Bigtable
Explanation:-This option is incorrect as Bigtable would not be a cost effective option.
Change the partitioned table column from time to date
Explanation:-This option is incorrect as changing the partitioning from time to date would be impact queries on packages.
✅ Update to table to perform clustering on package id
Explanation:-This option is correct as Clustering the data on the package Id can greatly improve the performance.

Clustering can improve the performance of certain types of queries such as queries that use filter clauses and queries that aggregate data. When data is written to a clustered table by a query job or a load job, BigQuery sorts the data using the values in the clustering columns. These values are used to organize the data into multiple blocks in BigQuery storage. When you submit a query
Ask the Data Scientists to use LIMIT parameter on the queries
Explanation:-This option is incorrect as LIMIT parameter does limit the amount of data queried.
