Appendix
Answers to Review Questions

Chapter 1: Selecting Appropriate Storage Technologies

1. D. The correct answer is D. Stackdriver Logging is the best option because it is a managed service designed for storing logging data. Neither Option A nor B is as good a fit because the developer would have to design and maintain a relational data model and user interface to view and manage log data. Option C, Cloud Datastore, would not require a fixed data model, but it would still require the developer to create and maintain a user interface to manage log events.
2. B. The correct answer is B. Cloud Dataflow is a stream and batch processing service that is used for transforming data and processing streaming data. Option A, Cloud Dataproc, is a managed Hadoop and Spark service and not as well suited as Cloud Dataflow for the kind of stream processing specified. Option C, Cloud Dataprep, is an interactive tool for exploring and preparing data sets for analysis. Option D, Cloud SQL, is a relational database service, so it may be used to store data, but it is not a service specifically for ingesting and transforming data before writing to a database.
3. A. The correct answer is A, Cloud Storage, because the data in the files is treated as an atomic unit of data that is loaded into RStudio. Options B and C are incorrect because those are document databases and there is no requirement for storing the data in semi-structured format with support for fully indexed querying. Also, MongoDB is not a GCP service. Option D is incorrect because, although you could load CSV data into a Bigtable table, the volume of data is not sufficient to warrant using Bigtable.
4. C. The correct answer is C, BigQuery, which is a managed analytical database service that supports SQL and scales to petabyte volumes of data. Options A and B are incorrect because both are used for transaction processing applications, not analytics. Option D is incorrect because Bigtable does not support SQL.
5. C. The correct answer is C. Bigtable is the best storage service for IoT data, especially when a large number of devices will be sending data at short intervals. Option A is incorrect, because Cloud SQL is designed for transaction processing at a regional level. Option B is incorrect because Cloud Spanner is designed for transaction processing, and although it scales to global levels, it is not the best option for IoT data. Option D is incorrect because there is no need for indexed, semi-structured data.
6. C. The correct answer is C because the requirements call for a semi-structured schema. You will need to search players' possessions and not just look them up using a single key because of the requirement for facilitating trading. Option A is not correct. Transactional databases have fixed schemas, and this use case calls for a semi-structured schema. Option B is incorrect because it does not support indexed lookup, which is needed for searching. Option D is incorrect. Analytical databases are structured data stores.
7. A. The correct answer is A. Cloud Bigtable using the HBase API would minimize migration efforts, and since Bigtable is a managed service, it would help reduce operational costs. Option B is incorrect. Cloud Dataflow is a stream and batch processing service, not a database. Options C and D are incorrect. Relational databases are not likely to be appropriate choices for an HBase database, which is a wide-column NoSQL database, and trying to migrate from a wide-column to a relational database would incur unnecessary costs.
8. C. The correct answer is C because the FASTQ files are unstructured since their internal format is not used to organize storage structures. Also, 400 GB is large enough that it is not efficient to store them as objects in a database. Options A and B are incorrect because a NoSQL database is not needed for the given requirements. Similarly, there is no need to store the data in a structured database like Cloud Spanner, so Option D is incorrect.
9. D. The correct answer is D because the output is structured, will be queried with SQL, and will retrieve a large number of rows but few columns, making this a good use case for columnar storage, which BigQuery uses. Options A and B are not good options because neither

database supports SQL. Option C is incorrect because Cloud Storage is used for unstructured data and does not support querying the contents of objects.

10. B. The correct answer is B. Bigtable is a wide-column NoSQL database that supports semi-structured data and works well with datasets over 1 TB. Options A, D, and C are incorrect because they all are used for structured data. Option D is also incorrect because Cloud SQL does not currently scale to 40 TB in a single database.
11. B. The correct answer is B, write data to a Cloud Pub/Subtopic, which can scale automatically to existing workloads. The ingestion process can read data from the topic and data and then process it. Some data will likely accumulate early in every minute, but the ingestion process can catch up later in the minute after new data stops arriving. Option A is incorrect; Cloud Dataflow is a batch and stream processing service—it is not a message queue for buffering data. Option C is incorrect; Cloud SQL is not designed to scale for ingestion as needed in this example. Option D is incorrect; Cloud Dataprep is a tool for cleaning and preparing datasets for analysis.
12. A. The correct answer is A. This is a good use case for key-value databases because the value is looked up by key only and the value is a JSON structure. Option B is incorrect. Analytical databases are not a type of NoSQL database. Option C is not a good option because wide-column databases work well with larger databases, typically in the terabyte range. Option D is incorrect because the data is not modeled as nodes and links, such as a network model.
13. D. The correct answer is D. A document database could store the volume of data, and it provides for indexing on columns other than a single key. Options A and C do not support indexing on non-key attributes. Option B is incorrect because analytical is not a type of NoSQL database.
14. B. The correct answer is B; OLAP data models are designed to support drilling down and slicing and dicing. Option A is incorrect; OLTP models are designed to facilitate storing, searching, and retrieving individual records in a database. Option C is incorrect; OLAP databases often employ denormalization. Option D is incorrect; graph data models are used to model nodes and their relationships, such as those in social networks.
15. C. The correct answer is C. Cloud Spanner is the only globally scalable relational database for OLTP applications. Options A and B are incorrect because Cloud SQL will not meet the scaling requirements. Options B and D are incorrect because Cloud Datastore does not support OLTP models.

Chapter 2: Building and Operationalizing Storage Systems

1. B. The correct answer is B, Cloud SQL Proxy. Cloud SQL Proxy provides secure access to Second Generation instances without you having to create allow lists or configure SSL. The proxy manages authentication and automatically encrypts data. Option A is incorrect because TLS is the successor to SSL. It can be used to encrypt traffic, but it would require the DBA to manage certificates, so it is not as good an answer as Option B. Option C is incorrect; using an IP address does not ensure encryption of data. Option D is incorrect; there is no such thing as an auto-encryption feature in Cloud SQL.
2. B. The correct answer is B. Maintenance could be occurring. Maintenance on read replicas is not restricted to the maintenance window of the primary instance or to other windows, so it can occur anytime. That would make the read replica unavailable. Option A is incorrect because a database administrator would have to promote a read replica, and the problem stated that there is no pattern detected and DBAs were not performing database operations. Option C is incorrect; backups are not performed on read replicas. Option D is incorrect; Cloud SQL instances do not fail over to a read replica.
3. B. The correct answer is B, 65%. Options A and C are not recommended levels for any Cloud Spanner configuration. Option D, 45%, is the recommend CPU utilization for a multi-regional Cloud Spanner instance.
4. A. The correct answer is A. Since each node can store 2 TB, it will require at least 10 nodes. Options B and D are incorrect because they are more nodes than needed. Answer C is incorrect; five is not sufficient for storing 20 TB of data.
5. B. The correct answer is B. The database is multitenant, so each tenant, or customer, will query only its own data, so all that data should be in close proximity. Using customer ID first accomplishes this. Next, the sensor ID is globally unique, so data would be distributed evenly across database storage segments when sorting based on sensor ID. Since this is time-series data, virtually all data arriving at the same time will have timestamps around the same time. Using a timestamp early in the key could create hotspots. Using sensor ID first would avoid hotspots but would require more scans to retrieve customer data because multiple customers' data would be stored in each data block.
6. D. The correct answer is D. Description can be represented in a string. Health consists of three properties that are accessed together, so they can be grouped into an entity. Possessions need a recursive representation since a possession can include sets of other possessions. Options A and B are incorrect; character state requires multiple properties, and so it should not be represented in a single string. Option B is also incorrect, because possessions are complex objects and should not be represented in strings. Option C is incorrect; description is an atomic property and does not need to be modeled as an entity.

7. D. The correct answer is D—no entities are returned. The query requires a composite index, but the question stated that no additional indexes were created. All other answers are wrong because querying by property other than a key will only return entities found in an index.
8. A. The correct answer is A. Setting shorter TTLs will make keys eligible for eviction sooner, and a different eviction policy may lead to more evictions. For example, switching from an eviction policy that evicts only keys with a TTL to a policy that can evict any key may reduce memory use.
9. A. The correct answer is A. Regional storage is sufficient for serving users in the same geographic location and costs less than multi-regional storage. Option B is incorrect because it does not minimize cost, and there is no need for multi-regional storage. Options C and D are incorrect because Nearline and Coldline are less expensive only for infrequently accessed data.
10. C. The correct answer is C. A data retention policy will ensure that files are not deleted from a storage bucket until they reach a specified age. Options A and B are incorrect because files can be deleted from Coldline or multi-regional data unless a data retention policy is in place. Option D is incorrect because a lifecycle policy will change the storage type on an object but not prevent it from being deleted.
11. B. The correct answer is B, installing the Stackdriver Monitoring agent. This will collect application-level metrics and send them to Stackdriver for alerting and charting. Option A is incorrect because Stackdriver Logging does not collect metrics, but you would install the Stackdriver Logging agent if you also wanted to collect database logs. Option C is incorrect; Stackdriver Debug is for analyzing a running program. Option D is incorrect; by default, you will get only instance metrics and audit logs.
12. C. The correct answer is C. The queries are likely scanning more data than needed. Partitioning the table will enable BigQuery to scan only data within a partition, and clustering will improve the way column data is stored. Option A is incorrect because BigQuery organizes data according to table configuration parameters, and there is no indication that queries need to order results. Option B is incorrect; Standard SQL dialect has more SQL features but none of those are used. Also, it is unlikely that the query execution plan would be more efficient with Standard SQL. Option D is incorrect; it would actually require more data to be scanned and fetched because BigQuery uses a columnar data storage model.
13. A. The correct answer is A. You probably did not include the pseudo-column `_PARTITIONTIME` in the `WHERE` clause to limit the amount of data scanned. Options B and D are incorrect; the format of the file from which data is loaded does not affect the amount of data scanned. Option C is incorrect; the distinction between active and long-term data impacts only the cost of storage, not the execution of a query.
14. D. The correct answer is D. Stackdriver Monitoring collects metrics, and the slot metrics are the ones that show resource utilization related to queries. Options A and B are incorrect; logging does not collect the metrics that are needed. Option C is incorrect because CPU utilization is not a metric associated with a serverless service like BigQuery.
15. A. The correct answer is A. Since the files have sequential names, they may be loading in lexicographic order, and this can create hotspots. Option B is incorrect; the volume of data, not the format, will determine upload speed. Option C is incorrect; there should be no noticeable difference between the command-line SDK and the REST API. Option D is incorrect; writing to multi-regional storage would not make the uploads faster.

Chapter 3: Designing Data Pipelines

1. C. The correct answer is C, Cloud Pub/Sub, which is a scalable, managed messaging queue that is typically used for ingesting high-volume streaming data. Option A is incorrect; Cloud Storage does not support streaming inserts, but Cloud Pub/Sub is designed to scale for high-volume writes and has other features useful for stream processing, such as acknowledging and processing a message. Option B is incorrect; Cloud SQL is not designed to support high volumes of low-latency writes like the kind needed in IoT applications. Option D is incorrect; although BigQuery has streaming inserts, the database is designed for analytic operations.
2. B. The correct answer is B. The transformation stage is where business logic and filters are applied. Option A is incorrect; ingestion is when data is brought into the GCP environment. Option C is incorrect—that data should be processed, and problematic data removed before storing the data. Answer D is incorrect; by the analysis stage, data should be fully transformed and available for analysis.
3. B. The correct answer is B. Cloud Dataflow supports Python and is a serverless platform. Option A is incorrect because, although it supports Python, you have to create and configure clusters. Option C is incorrect; Cloud Spanner is a horizontally scalable global relational database. Option D is incorrect; Cloud Dataprep is an interactive tool for preparing data for analysis.
4. C. The correct answer is C; the new code disabled message acknowledgments. That caused Cloud Pub/Sub to consider the message outstanding for up to the duration of the acknowledgment wait time and then resend the message. Options A and B are incorrect; changing the subscription or topic IDs would cause problems but not the kind described. Option D is incorrect because the type of subscription does not influence whether messages are delivered multiple times.
5. A. The correct answer is A; messages may be delivered multiple times and therefore processed multiple times. If the logic were not idempotent, it could leave the application in an incorrect state, such as that which could

occur if you counted the same message multiple times. Options B and C are incorrect; the order of delivery does not require idempotent operations. Option D is incorrect; the time between messages is not a factor in requiring logic to be idempotent.

6. A. The correct answer is A; a sliding window would have the data for the past four minutes. Option B is incorrect because tumbling windows do not overlap, and the requirement calls for using the last four messages so the window must slide. Options C and D are not actually names of window types.
7. A. The correct answer is A; you should use CloudPubSubConnector and Kafka Connect. The connector is developed and maintained by the Cloud Pub/Sub team for this purpose. Option B is incorrect since this is a less direct and efficient method. Option C requires maintaining a service. Option D is incorrect because there is no such service.
8. C. The correct answer is C. Use Cloud Dataflow templates to specify the pattern and provide parameters for users to customize the template. Option A is incorrect since this would require users to customize the code in the script. Options B and D are incorrect because Cloud Dataproc should not be used for this requirement. Also, Option D is incorrect because there are no Cloud Dataproc templates.
9. D. The correct answer is D. You should create an ephemeral cluster for each job and delete the cluster after the job completes. Option A is incorrect because that is a more complicated configuration. Option B is incorrect because it keeps the cluster running instead of shutting down after jobs complete. Option C is incorrect because it keeps the clusters running after the jobs complete.
10. A. The correct answer is A, Cloud Composer, which is designed to support workflow orchestration. Options B and C are incorrect because they are both implementations of the Apache Beam model that is used for executing stream and batch processing program. Option D is incorrect; Cloud Dataproc is a managed Hadoop and Spark service.
11. B. The correct answer is B. The data could be stored in Cloud Bigtable, which provides consistent, scalable performance. Option A is incorrect because Cloud Storage is an object storage system, not a database. Option C is incorrect, since Cloud Datastore is a document-style NoSQL database and is not suitable for a data warehouse. Option D is incorrect; Cloud Dataflow is not a database.
12. D. The correct answer is D. With change data capture, each change is a source system captured and recorded in a data store. Options A, B, and C all capture the state of source systems at a point in time and do not capture changes between those times.

Chapter 4: Designing a Data Processing Solution

1. B. The correct answer is B. IoT sensors can write data to a Cloud Pub/Sub topic. When a message is written, it can trigger a Cloud Function that runs the associated code. Cloud Functions can execute the Python validation check, and if the validation check fails, the message is removed from the queue. Option A is incorrect; Cloud Storage is not a for streaming ingestion. Option C is incorrect because BigQuery is an analytical database that could be used in later stages but not during ingest. Answer D is incorrect because Cloud Storage is not a suitable choice for high-volume streaming ingestion, and BigQuery is not suitable for storing data during ingestion.
2. A. The answer is A. This scenario calls for full control over the choice of the operating system, and the application is moving from a physical server so that it is not containerized. Compute Engine can run the application in a VM configured with Ubuntu 14.04 and the additional packages. Option B is incorrect because the application is not containerized (although it may be modified to be containerized). Option C is incorrect because the application cannot run in one of the language-specific runtimes of App Engine Standard. Option D is incorrect because the Cloud Functions product runs code in response to events and does not support long-running applications.
3. B. The correct answer is B, Kubernetes Engine, because the application will be designed using containerized microservices that should be run in a way that minimizes DevOps overhead. Option A is incorrect because Compute Engine would require more DevOps work to manage your own Kubernetes Cluster or configure managed instance groups to run different containers needed for each microservice. Options C and D are incorrect because App Engine Standard and Cloud Functions do not run containers.
4. C. The correct answer is C. Mean time between failure is used for measuring reliability. Options A and B are incorrect because they are related to utilization and efficiency but unrelated to reliability. Option D is incorrect, since mean time to recovery is used as a metric for restoring service after an outage. Mean time to recovery is important and would likely be included in negotiations, but it is not used as a measure of reliability.
5. C. The correct answer is C. A global load balancer is needed to distribute workload across multiple regions. Options A and B are incorrect because there is no indication in the requirements that object storage or a message queue is required. Option D is incorrect because there is no indication that a hybrid cloud is needed that would necessitate the use of a VPN or direct connect option.
6. A. The correct answer is A. The purpose of this queue is to list rooms on the platform so that as long each message is processed at least once, the room will appear in the listing. Options B and D are incorrect because processing does not have to be exactly once because listing a room is an idempotent operation. For example, adding a listing of the same room

twice does not change the listing since duplicate listing messages are dropped by the application. Option C is incorrect because no ordering is implied in the requirements.

7. C. The correct answer is C. Machines of different types have different failure characteristics and therefore will have their own models. Option A is incorrect; randomly distributing messages will mix metrics from different types of machines. Option B is incorrect because identifiers in close proximity are not necessarily from machines of the same type. Option D is incorrect; routing based on timestamp will mix metrics from different machine types.
8. B. The correct answer is B. The description of independent services, using SOAP, and deployed on virtual machines fits the definition of an SOA architecture. Answer A is incorrect; since there are multiple components, it is not a monolithic architecture. Option C could be a possibility, but it is not the best fit since the application uses SOAP and is deployed on VMs. Option D is incorrect because the application does not use a serverless deployment.
9. C. The correct answer is C. Microservices would allow each function to be deployed independently in its own container. Option A is incorrect; a monolithic architecture would make the update problems worse. Option B is incorrect, because hub-and-spoke is a message broker pattern. Option D is incorrect; pipelines are abstractions for thinking about workflows—they are not a type of architecture.
10. A. The correct answer is A. An assessment should be done first. Options B, C, and D are all parts of a data warehouse migration plan but come after the assessment phase.
11. A. The correct answer is A. Data sources, the data model, and ETL scripts would all be included. Options B and D are incorrect; technical requirements do not include information about business sponsors and their roles. Option C is incorrect because more than data sources should be included.
12. C. The correct answer is C. The company is incurring an opportunity cost because if they had migrated to a modern cloud-based data warehouse, the team would have had opportunities to develop new reports. Options A and B are incorrect; although they are kinds of expenses, they require expenditure of funds to be either a capital or an operating cost. Option D is not a type of cost.
13. C. The correct answer is C. Denormalization reduces the number of joins required and nested, and repeated fields can be used to store related data in a single row. Option A is incorrect; BigQuery does use Colossus, but that does not change the number of joins. Option B is incorrect; BigQuery does use columnar storage, but that does not affect the number of joins. Option D is incorrect; federated storage allows BigQuery to access data stored outside of BigQuery, but it does not change the need for joins.
14. D. The correct answer is D. Prioritizing low-risk use cases will allow the team to make progress on migrating while minimizing the impact if something goes wrong. Options A, B, and C are incorrect because they do not give priority to minimizing risk; other factors are prioritized in each case.
15. B. The correct answer is B. The set of tasks to verify a correct data warehouse migration include verifying schemas, data loads, transformations, and queries, among other things. Option A is incorrect because more is required than just verifying schemas and data loads. Options C and D are incorrect; the backlog of feature requests is important but not relevant to verifying the migration.

Chapter 5: Building and Operationalizing Processing Infrastructure

1. C. The correct answer is C. A managed instance group will provision instances as required to meet the load and stay within the bounds set for the number of instances. Option A is incorrect; Cloud Functions are for event-driven processing, not continually monitoring metrics. Option B is incorrect because it is not the most efficient way to scale instances. Option D is incorrect, since the requirements call for Compute Engine instances, not a Hadoop/Spark cluster.
2. B. The correct answer is B. Autohealing uses a health check function to determine whether an application is functioning correctly, and if not, the instance is replaced. Option A is incorrect; autoscaling adds or removes instances based on instance metrics. Option C is incorrect; redundancy is a feature of instance groups, but it is not the mechanism that replaces poorly performing nodes. Option D is incorrect; eventual consistency describes a model for storing writes in a way that they will eventually be visible to all queries.
3. C. The correct answer is C, defining an instance template using the `gcloud compute instance-templates create` command. Options A and B are incorrect, since there is no need to create each instance individually. Option D is incorrect. `cbt` is the command-line utility for working with Cloud Bigtable.
4. A. The correct answer is A; Kubernetes uses pods as the smallest deployable unit. Options B and C are incorrect because deployments and replicas are Kubernetes abstractions, but they are not used as the mechanism for logically encapsulating containers. Option D is incorrect, since pods and containers are not synonymous.
5. A. The correct answer is A; the `kubectl scale deployment` command specifying the desired number of replicas is the correct command. Option B is incorrect, since this would set the number of replicas to 2. Options C and D are incorrect; there is no `gcloud containers scale deployment` command.

6. C. The correct answer is C, using two clusters with one dedicated for receiving write operations and the other responsible for batch processing. Options A and B are incorrect because you do not specify the operating system used in Bigtable. Option D is incorrect; a cluster cannot have multiple node pools in the same zone.
7. D. The correct answer is D; the number of master nodes cannot be changed. Options A and B are incorrect; there is no `--num-masters` or `--add-masters` parameter in the `gcloud dataproc clusters update` command. Option C is incorrect; `cbt` is the command-line utility for working with Cloud Bigtable.
8. A. The correct answer is A; the total CPU utilization by the deployment is used as the basis for making scaling decisions. Option B is incorrect; some CPUs in the cluster may be used by other deployments. Options C and D are incorrect because the decision is based on overall utilization, not any individual pod.
9. A. The correct answer is A; `app.yaml` is the configuration file used to specify the runtime. Option B is incorrect; `queue.yaml` is used to configure task queues. Option C is incorrect; `dispatch.yaml` is used to override routing rules. Option D is incorrect; `cron.yaml` is used to schedule tasks.
10. B. The correct answer is B. The `--max-instances` parameter limits the number of concurrently executing function instances. Option A is incorrect; `--limit` is not a parameter used with function deployments. Option C is incorrect; labels are not used to control configuration of functions. Option D is incorrect; language-specific parameters are not used to configure Cloud Functions.
11. D. The correct answer is D; data access logs have a 30-day data retention period. Options A, B, and C are incorrect; they all have 400-day retention periods.
12. C. The correct answer is C; Stackdriver Monitoring collects performance metrics. Option A is incorrect; Stackdriver Debugger is used to inspect the state of running code. Option B is incorrect; Stackdriver Logging is used to collect semi-structured data about events. Option D is incorrect; Stackdriver Trace is used to collect information about the time required to execute functions in a call stack.
13. B. The correct answer is B; Stackdriver Logging is used to collect semi-structured data about events. Option A is incorrect; Stackdriver Debugger is used to inspect the state of running code. Option C is incorrect; Stackdriver Monitoring is used to collect performance metrics. Option D is incorrect; Stackdriver Trace is used to collect information about the time required to execute functions in a call stack.
14. D. The correct answer is D; Stackdriver Trace is used to collect information about the time required to execute functions in a call stack. Option A is incorrect; Stackdriver Debugger is used to inspect the state of running code. Option B is incorrect; Stackdriver Logging is used to collect semi-structured data about events. Option C is incorrect; Stackdriver Monitoring is used to collect performance metrics.
15. B. The correct answer is B; `maxNumWorkers` specifies the maximum number of instances that can be run for a Cloud Dataflow pipeline. Option A is incorrect; `numWorkers` is the initial number of workers. Option C is incorrect; `streaming` specifies whether streaming mode is enabled. Option D is incorrect; it is not an actual parameter.

Chapter 6: Designing for Security and Compliance

1. C. The correct answer is C. This is an appropriate use case for primitive roles because there are few users working in a development environment, not production, and working with data that does not contain sensitive information. In this case, there is no need for fine-grained access controls. Options A and B are incorrect because they would require more administration, and fine-grained access controls are not needed. Option D is incorrect; access control lists are used with Cloud Storage resources and should be used only when roles are insufficient.
2. A. The correct answer is A; the `iam.roles.create` permission is needed to create custom roles. Option B is incorrect; it is not an actual permission. Options C and D are incorrect; they are examples of fictitious roles, not permissions.
3. C. The correct answer is C. A service account associated with the application should have the `roles/storage.objectCreator` assigned to it. Options A and B are incorrect; those are identities associated with actual users. Option D is incorrect; access control lists can be assigned to a bucket, but roles are assigned to identities.
4. B. The correct answer is B. Policy A applies to all departments, so it should be assigned at the organizational level. Policies B, C, D, and E are department specific and apply to all projects, so they can be inherited by projects when they are assigned to the departments folder. Option A is incorrect; policy A belongs at the organizational level, and each of the other policies should apply only to one department's folder. Option C is incorrect; the policies should not be assigned to individual projects. Option D is incorrect because policy A belongs at the organization level, and policies B, C, D and E belong at the folder level.
5. A. The correct answer is A. Since the application needs to read the contents of only the object, the `roles/storage.objectViewer` role is sufficient. Options B grants more permissions than needed. Option C would not allow the application to read the object. Option D has more permissions than needed.
6. B. The correct answer is B. The `roles/BigQuery.jobUser` role allows users to run jobs, including queries. Option A is incorrect because that would grant more permissions than needed. Option C is incorrect; it

would allow access to table and dataset metadata. Option D is incorrect; there is no such role.

7. D. Option D is correct. You do not need to configure any settings to have data encrypted at rest in GCP. Options B, C, and D are all incorrect because no configuration is required.
8. A. The correct answer is A: AES256 or AES128 encryption. Option B is incorrect, but it is a strong encryption algorithm and could be used to encrypt data. Option C is incorrect; DES is a weak encryption algorithm that is easily broken by today's methods. Option D is incorrect; Blowfish is strong encryption algorithm designed as a replacement for DES and other weak encryption algorithms.
9. B. The correct answer is B; the data encryption key is encrypted using a key encryption key. Option A is incorrect; there are no hidden locations on disk that are inaccessible from a hardware perspective. Option C is incorrect; keys are not stored in a relational database. Option D is incorrect; an elliptic curve encryption algorithm is not used.
10. C. The correct answer is C. The risk analysis job assesses the likelihood that redacted data can be re-identified. Option A and Option B are incorrect. The results are not measures of counts or percent of times that data is redacted. Option D is incorrect. The result is not a list of InfoType patterns detected.
11. B. The correct answer is B. You should prioritize the order of scanning, starting with the most at-risk data. Option A is incorrect; identifying InfoTypes to use comes later. Option C is incorrect; a risk analysis is done after inspection. Option D is incorrect; that is not the recommended first step.
12. C. The correct answer is C; COPPA is a regulation that governs the collection of data from children under the age of 13. Option A is incorrect; HIPAA is a healthcare regulation. Option B is incorrect; GDPR is a European Union privacy regulation. Option D is incorrect; FedRAMP applies to cloud providers supplying services to U.S. federal agencies.

Chapter 7: Designing Databases for Reliability, Scalability, and Availability

1. C. The correct answer is C. Hotspots occur when a row-key design does not adequately distribute read/write load for a given query pattern. Option A is incorrect; Cloud Bigtable does not have secondary indexes. Option B is incorrect; Cloud Bigtable does not use partition keys. Options D is incorrect; the hotspots are not caused by failure to use a read replica, although using a read replica may reduce the load on the nodes with hotspots.
2. A. The correct answer is A. This table should be designed as a tall and narrow one with a single dataset in each row. Option B is incorrect because starting a row-key with the date and hour will lead to hotspotting. Option C is incorrect, since changing the start time will not change the parts of the design that make querying by ranges more difficult than they need to be. Option D is incorrect; the structure of rows should be changed from wide to narrow.
3. B. The correct answer is B. The only way to achieve strong consistency in Cloud Bigtable is by having all reads routed from a single cluster and using the other replicas only for failover. Option A is incorrect; Cloud Bigtable does not have secondary indexes. Option C is incorrect; moving tablets does not impact read consistency. Option D is incorrect; a poor row-key design can impact performance but not consistency.
4. C. The correct answer is C; the `STRUCT` data type is used to store ordered type fields, and this is the closest to a document structure. Option A is incorrect; all elements of an array are of the same data type, but items in a document may consist of different data types. Option B is incorrect because although a document could be represented in a string, it does not provide field-level access to data like a `STRUCT` does. Option D is incorrect; JSON is not a valid data type in Cloud Spanner.
5. B. The correct answer is B. Since the problematic queries involved joins of hierarchically related tables, interleaving the data of the tables could improve join performance. Option A is incorrect; Cloud Bigtable, not Cloud Spanner, uses replicated clusters. Option C is incorrect; the `STORING` clause is used to create indexes that can answer queries using just the index, and that would not address the join performance problem. Option D is incorrect; an execution plan might help you understand the cause of a performance problem, but it would not on its own improve query performance.
6. A. The correct answer is A. By using a hash of the natural key, you will avoid hotspotting and keeping the natural key data in the table will make it available to users. Option B is incorrect because Cloud Spanner automatically creates splits based on load, and if the database performance is adversely affected, then splitting is no longer sufficient to address the problem. Option C is incorrect; interleaved tables reduce the number of I/O operations performed when retrieving related data. Option D is incorrect; adding more secondary indexes will not change the hotspotting pattern of the primary index.
7. C. The correct answer is C. The likely cause is that you are using a UUID generator that uses time or other sequential values at the start of the key. Options A and B are incorrect because secondary indexes are not related to primary key hotspots. Option D is incorrect; UUIDs have a fixed, constant length, so the size of the string used to store them should not need to be increased.
8. E. The correct answer is E. There should be a small number of tall and narrow tables. Option A is incorrect because it does not include one table for equities and one for bonds. Option B is incorrect; this would lead to

many small tables rather than fewer large tables. Option C is incorrect because it is missing the use of tall and narrow tables. Option D is incorrect because it includes Option B and not Option C.

9. D. The correct answer is D. The enterprise is building a data lake for large volumes of data that is not yet organized for analytics. Options A and B are incorrect because there is no information about how data will be queried or what data would be included. Option C is incorrect because Cloud Spanner is a global, horizontally scalable relational database designed for transaction processing.
10. A. The correct answer is A. Since data is being written to partitions, and the data modeler did not specify a timestamp or integer column as the partition key, it must be partitioned based on ingestion time. Options B and C are incorrect because they require a column to be specified that has a partition value. Option D is incorrect; clustered tables are not a form of partitioning.
11. B. The correct answer is B. FLOAT64 is not a supported cluster column type. Answer A is incorrect; the table is not external because data is loaded into BigQuery. Option C is incorrect; there is no such thing as a FLOAT64 partition type. Option D is incorrect; clustering keys do not need to be integers or timestamps—they can be data, Bool, geography, INT64, numeric, string, or timestamp.
12. C. The correct answer is C; comma-separated values, Avro, and newline-delimited JSON are supported. Options A and B are both missing at least one supported format. Option D is incorrect because the Parquet format is not supported in Google Drive, although it is supported in Cloud Storage.

Chapter 8: Understanding Data Operations for Flexibility and Portability

1. B. The correct answer is B. Cloud Catalog is designed to help data consumers understand what data is available, what it means, and how it can be used. Option A is incorrect; Cloud Composer is a managed workflow service. Option C is incorrect; Cloud Dataprep is used to prepare data for analysis and machine learning. Option D is incorrect; Data Studio is used for reporting and visualizing data.
2. A. The correct answer is A. Cloud Composer is a managed workflow service based on Apache Airflow. Option B is incorrect; Data Catalog is a metadata management service. Option C is incorrect; Cloud Dataprep is used to prepare data for analysis and machine learning. Option D is incorrect; Data Studio is used for reporting and visualizing data.
3. C. The correct answer is C. Cloud Dataprep is used to prepare data for analysis such as this, as well as machine learning. Option A is incorrect; Cloud Composer is a managed workflow service based on Apache Airflow. Option B is incorrect; Cloud Catalog is a metadata management service. Option D is incorrect; Data Studio is used for reporting and visualizing data.
4. D. The correct answer is D. Data Studio is the GCP tool to use for reporting and visualizing data. Option A is incorrect; Cloud Composer is a managed workflow service based on Apache Airflow. Option B is incorrect; Data Catalog is a metadata management service. Option C is incorrect; Cloud Dataprep is used to prepare data for analysis and machine learning.
5. B. The correct answer is B. CSV and JSON are the only formats supported for exporting compressed data from Cloud Dataprep. Option A is incorrect because it does not include JSON. Options C and D are incorrect because they include AVRO.
6. B. The correct answer is B. Extracted data sources work with a static snapshot of a dataset, which gives better performance than live connection data sources. Option A is incorrect because extracted connections are faster than live connections. Option C is incorrect because there is no compound connection. Option D is incorrect because blended connections are designed to query data from up to five data sources.
7. A. The correct answer is A. Live connections will update data in reports automatically. Option B is incorrect; you would need to update the extracted dataset manually in order to refresh data. Option C is incorrect; there is no such thing as a compound connection. Option D is incorrect because it includes extracted connections.
8. B. The correct answer is B. `conda install` or `pip install` can be run from within a Jupyter Notebook. Option A is incorrect; they do not have to go outside Jupyter Notebook to install the library. Option C is incorrect; the Linux package manager is not used to install Python libraries. Option D is incorrect; Python libraries are shared using Python-specific package managers, so users do not have to work with source code directly.
9. C. The correct answer is C. Cloud Composer is a workflow orchestration service that can perform all the tasks mentioned. Option A is incorrect; Cloud Dataprep is a service for preparing data for analysis and machine learning. Option B is incorrect; Cloud Dataproc is a managed Hadoop/Spark service. Option D is incorrect; Data Studio is a reporting and visualization tool.
10. B. The correct answer is B. Cloud Datalab is a managed Jupyter Notebook service that supports interactive analysis and ad hoc programming. Option A is incorrect; Cloud Dataproc is a managed Hadoop/Spark service. Option C is incorrect; Cloud Composer is a workflow orchestration service based on Apache Airflow. Option D is incorrect; Data Studio is a reporting and visualization tool.

Chapter 9: Deploying Machine Learning Pipelines

1. A. The correct answer is A. Cloud Storage is an object storage system that makes no assumptions about the internal structure of objects. Option B is incorrect; Cloud Spanner is a globally scalable relational database and provides for highly structured data schemas. Option C is incorrect; Cloud Dataprep is a tool for preparing data for analysis and machine learning but not for storage. Option D is incorrect; Cloud Pub/Sub is used for streaming ingestion, not batch ingestion.
2. D. The correct answer is D. Cloud Pub/Sub is designed for this kind of streaming ingestion, and it can scale to meet the expected growth in the number of sensors. Option A is incorrect; Cloud Storage is used for batch ingestion, not streaming ingestion. Option B is incorrect; the data will eventually be stored in Cloud Bigtable, but it should be written to a Cloud Pub/Sub topic that can buffer the data prior to having the data consumed by a Cloud Dataflow service. Option C is incorrect; BigQuery Streaming Insert should be used only when the streaming data is being stored in BigQuery.
3. B. The correct answer is B. The data is available in a data lake, so there is no need to ingest the data. Thus, the next step should be to understand the distribution and quality of data using Cloud Dataprep. Option A is incorrect; the data has already been ingested into a data lake. Options C and D are incorrect; the data should not be transformed until it has been evaluated.
4. B. The correct answer is B. Validation is used to assess the quality of model predictions when tuning hyperparameters. Option A is incorrect; training data is used to learn parameters of the model, not hyperparameters. Option C is incorrect; test data is used to measure the quality of a model after hyperparameters have been tuned. Option D is incorrect; there is no such thing as hyperparameter data.
5. A. The correct answer is A. The first thing to do is to explore the data to understand any quality issues and to perform feature engineering. Feature engineering can reduce the amount of time needed to train a model and improve performance. Option B is incorrect; visualizations of model evaluation metrics will not help with either the time to build or the quality of the model. Option C is incorrect; cross-validation is useful for evaluation, not for reducing the time to build a model. Option D is incorrect; it is too early to tune hyperparameters, and feature engineering should occur before that.
6. B. The correct answer is B. This is an example of underfitting. If the model performs poorly across multiple algorithms and when evaluating using the same data that was used to train the model, then that is underfitting and it is likely caused by too little training data. Option A is incorrect; the model is not overfitting the training data because, if that were the case, the accuracy would be high when evaluated with the training data. Option C is incorrect; the opposite is the likely cause of underfitting. Option D is incorrect; tenfold cross-validation for evaluation is a reasonable technique for evaluating the quality of a machine learning model.
7. C. The correct answer is C. The model is overfitting the training data, so adding a penalty factor using L1 and L2 regularization will reduce overfitting. Option A is incorrect; using confusion matrices will not change the model. Option B is incorrect; regularization is done during training, not evaluating. Option D is incorrect because it is not the best answer. Tuning hyperparameters may improve evaluation metrics but not as effectively as applying regularization.
8. B. The correct answer is B. There has likely been a change in sales patterns since the model was trained, and the model should be retrained with data that more closely reflects the actual distribution of sales data today. Option A is incorrect; if a model were overfitting, it would perform poorly initially, as well as several months later. Option C is incorrect; performance metrics, such as CPU utilization, will not help diagnose a quality of recommendation problem. Option D is incorrect; if the model were underfitting, it would perform poorly initially, as well as several months later.
9. A. The correct answer is A. The problem calls for analyzing images, not videos, and the task is identifying objects so one of the Object Detection services should be used. Since the data is used for analysis and long-term decision making, detection does not need to be performed at the edge. Option B is incorrect because analysis does not require a real-time result to make a decision at the edge. Options C and D are incorrect because this application uses images, not videos.
10. C. The correct answer is C. AutoML Tables is designed to build machine learning models using structured data. It also automates common tasks such as feature engineering. Options A and B are incorrect because they require machine learning knowledge to use. Option D is incorrect; AutoML Natural Language is used to classify texts and other natural language artifacts, not structured data.
11. C. The correct answer is C. BigQuery ML provides access to machine learning algorithms from within SQL, and there is no need to move the data from BigQuery. Also, BigQuery builds models faster than AutoML, so BigQuery ML best fits the requirements. Options A and B are incorrect because they require some machine learning experience to use. Option D is incorrect because AutoML Tables can take an hour or more to build, so C is a better option.
12. B. The correct answer is B. Spark MLib includes association rules for frequent pattern mining. Option A is incorrect; Cloud Dataflow is a stream and batch processing service. Options C and D are incorrect; BigQuery ML and AutoML Tables do not include algorithms for frequent pattern mining.

Chapter 10: Choosing Training and Serving Infrastructure

1. C. The correct answer is C. The requirements call for high-precision arithmetic and parallelization, so that indicates using a GPU. There is a small amount of data, and you want to work with it interactively, so a single machine with a GPU will suffice. Options A and B are incorrect because TPUs do not support high-precision arithmetic. Also, Option A requires more resources than needed for a small dataset. Option D is incorrect because this is an interactive workload, so there is no need for the high availability provided by a managed instance group and there are more resources allocated than needed for this workload.
2. D. The correct answer is D. The TPU strategy meets all of the requirements of synchronous training on TPUs. The other strategies all apply to GPUs and/or CPUs and therefore do not meet the requirements.
3. C. The correct answer is C. This is an example of a latency problem that might be resolved by serving the model closer to where the data is generated. Options A and B are incorrect because overfitting and underfitting are problems with model training not serving. Option D is incorrect; there is no indication that the volume of data processed is a problem.
4. B. The correct answer is B. The raw data does not need to be stored and with limited bandwidth; it is best to minimize the amount of data transmitted, so sending just the average is correct. Also, because the network connection is sometimes unreliable, you should use Cloud Dataflow to implement stream processing logic, such as handling late-arriving data and inserting default values for missing data. Option A is incorrect because it sends too much data. Option C is incorrect because it sends too much data and stores it directly in BigQuery without preprocessing for missing values and other business logic. Option D is incorrect because it stores data directly in BigQuery without preprocessing for missing values and other business logic.
5. C. The correct answer is C. Repeaters are used in networks to boost signal strength. There is no indication that this is needed, and in any case, that is a network implementation choice and not a comparable part of the IoT architecture of the other components. Options A, B, and D are all part of the standard IoT architecture.
6. C. The correct answer is C. Cloud Pub/Sub is a globally scalable messaging queue that can ingest large volumes of data and buffer it for other services. Option A is incorrect; Cloud Storage is not used for streaming high-volume data into GCP; it is used for batch uploads. Option B is incorrect; BigQuery streaming inserts are not used for ingestion in the reference model. Option D is incorrect; Cloud Bigtable is not used for ingestion.
7. C. The correct answer is C. Edge TPU is designed for inferencing on edge devices. Since the model is used to help autonomous vehicles improve their ability to track objects in adverse weather conditions, low latency is essential. Options A and B are incorrect because they serve the model from a central service rather than at the edge. Option D is incorrect; GPUs are used when training models, not when using them for inference.
8. D. The correct answer is D. Cloud Dataflow is a stream and batch processing service based on Apache Beam. Option A is incorrect; Cloud Storage is not used for stream processing—it is an object storage service. Option B is incorrect; BigQuery streaming inserts are for storing data in BigQuery partitioned tables. Option C is incorrect; Cloud Pub/Sub is used for ingestion, not stream processing.
9. B. The correct answer is B. This is a typical use case for TPUs because the model is built on TensorFlow using only basic operations and no custom operations, so TPUs are an option. The long training time on CPUs indicate that this is a good option for TPUs. Option A is incorrect; this would only cut the training time in half, assuming a linear speedup. Option C is incorrect because only CPUs are used and not TPUs. Option D is incorrect; App Engine is used for scalable web applications and not used for training models.
10. A. The correct answer is A. A Fortran library optimized for highly parallel, high-precision arithmetic that runs on GPUs would be a good option for training this model. Option B is incorrect; TPUs do not support high-precision arithmetic. Option C is not the best choice because it would not be as performant as a GPU-based solution, but it could be a backup option in case the Fortran library did not function on the GPU. Option D is incorrect; Cloud Functions do not run Fortran code and are not suitable to running workloads such as this.

Chapter 11: Measuring, Monitoring, and Troubleshooting Machine Learning Models

1. B. The correct answer is B. Sales price is a continuous value, so a regression algorithm would be used to predict it. Option A is incorrect; classifiers are used to predict discrete categories. Option C is incorrect; decision trees are used for classification. Option D is incorrect; reinforcement learning is a type of machine learning that learns from the environment.
2. A. The correct answer is A. The question is asking about building a binary classifier, so a logistic regression would work. Option B is incorrect; K-means clustering is an unsupervised learning algorithm, and this is a supervised learning problem. Options C and D are incorrect because simple linear regression and multiple linear regression are used for predicting continuous values.
3. C. The correct answer is C. The data being sent is time-series data, and they are trying to detect anomalies. Option A is incorrect; this problem does not call for partitioning the dataset. Option B is incorrect because K-

means is a clustering algorithm, and this problem does not call for partitioning a dataset. Option D is incorrect; there is no learning from the environment.

4. A. The correct answer is A. The perceptron algorithm is used to train a binary classifier based on artificial neurons. Options B and C are incorrect; those algorithms can build classifiers but not ones based on an artificial neuron. Option D is incorrect; linear regression is used to predict a continuous value.
5. C. The correct answer is C. A baseline model is the simplest model, and it is used as a reference point. Options A and B are incorrect; this is a regression problem, not a classification problem. Option D is incorrect because the problem does not call for partitioning the data.
6. A. The correct answer is A. Bucketing maps continuous values to ranges, for example, from 0.00 to 10.00. Each bucket is a discrete value of the derived feature. Option B is incorrect because the problem does not call for reducing dimensions. Option C is incorrect because principal component analysis is a type of dimension reduction. Option D is incorrect; gradient descent is a technique used when training a model.
7. C. The correct answer is C. L2 regularization calculates a penalty based on the sum-of-the-squares of the weights. L1 regularization should be used when you want less relevant features to have weights close to zero. Option A is incorrect; gradient descent does not lessen the impact of a feature. Option B is incorrect; a large number of epochs will not reduce the impact of outliers, and it may actually lead to overfitting. Option D is incorrect; backpropagation is a technique used to train neural networks.
8. B. The correct answer is B. The model overfits the training data, and dropout is a regularization method for neural networks. Options A and C are incorrect and could actually make the overfitting problem worse. Option D is incorrect; ReLU is a type of activation function, not a regularization method.
9. A. The correct answer is A. The dataset is unbalanced, and undersampling legitimate products will lead to a more balanced dataset. Option B is incorrect; dropout is a regularization technique used with neural networks. Option C is incorrect; L1 regularization reduces the risk of overfitting. Option D is incorrect; AUC, or area under the curve, is a way to evaluate the quality of a model.
10. C. The correct answer is C. This is an example of reporting bias because the dataset did not reflect the population. Options A and B are incorrect; applying L1 regularization or dropout would not cause a model to perform well with training and test data but not with more representative data. Option A is incorrect; there is no indication that a human is involved in the decision making.

Chapter 12: Leveraging Prebuilt Models as a Service

1. A. The correct answer is A. The Cloud Video Intelligence API can identify objects and track objects across frames. Option B is incorrect because it cannot track objects across frames. Option C is incorrect because, although streaming traffic data is a form of time-series data, it does not support object recognition or object tracking. Option D is incorrect; Cloud Dataflow is a batch and stream processing service and may be used for its stream processing capabilities, but it does not have object identification or object tracking capabilities.
2. B. The correct answer is B. The Cloud Vision API supports explicit content identification, also known as Safe Search. Option A is incorrect since there is no requirement to support video on the site. Option C is incorrect; the site does need to analyze time-series data, which is what Cloud Inference API is used for. Option D is incorrect; Cloud Dataprep is used to prepare data for analysis and machine learning.
3. B. The correct answer is B. The Cloud Vision API supports up to 2,000 images per batch. Option A is incorrect because if the wrong function were called, none of the operations would succeed. Option C is incorrect since all buckets have the same access controls and some operations succeed. Option D is incorrect; images are loaded from Cloud Storage buckets.
4. D. The correct answer is D. Intents categorize a speaker's intention for a single statement, such as asking for a recommendation. Option A is incorrect; entities are nouns extracted from dialogue. Option B is incorrect; fulfillments are used to connect a service to an integration. Option C is incorrect; integrations are applications that process end-user interactions, such as deciding what to recommend.
5. C. The correct answer is C. Both changing the device specification to wearable and using WaveNet-quality voice will improve the output. Options A and B are both partially correct, but not a completely correct answer. Option D is incorrect; Base64 is an encoding for binary data, not text. Option E is incorrect because it includes Option C.
6. B. The correct answer is B. Google recommends a minimum sampling rate of 16,000 Hz. Option A is incorrect; WaveNet is used for speech synthesis, not speech to text. Option C is incorrect; SSML is also for text to speech. Option D is incorrect because it includes Option A.
7. C. The correct answer is C. The gRPC API is only available with the advanced version of the Translation API. Option A is incorrect; WaveNet is a speech synthesis option, not a translation option. Option B is incorrect; the basic version does not provide a gRPC API. Option D is incorrect because Option A is included in the choices.
8. A. The correct answer is A. The first two steps are to import libraries and to create a translation client data structure. Option B is incorrect because the translation client can't be created when importing the libraries first. Options C and D are incorrect because there is no need to pass a

parameter into the API with the operation when there is a specific function call for translating.

9. A. The correct answer is A. The goal is to identify people, which are one kind of entity, so entity extraction is the correct functionality. Options B and C are incorrect because there is no requirement to understand the sentiment of the communications. Option D is incorrect because syntactic analysis does not help with identifying individuals.
10. A. The correct answer is A. Click-through rate (CTR) is the default optimization, and it maximizes the likelihood that the user engages the recommendation. Option B is incorrect; revenue per order is only available with the "frequently bought together" recommendation type. Option C is incorrect; conversation rate optimizes for the likelihood that the user purchases the recommended product. Option D is incorrect; total revenue is a metric for measuring performance, not an optimization objective.
11. C. The correct answer is C. Recommended for you predicts the next product with which the customer is likely to engage. Option A is incorrect; it provides a list of products that the customer is likely to purchase. Option B is incorrect; it provides a list of products often purchased together. Option D is incorrect; the recently viewed recommendation type provides a list of recently viewed items.
12. A. The correct answer is A. The Cloud Inference API is designed for this kind of time-series analysis and anomaly detection. Option B is incorrect; AutoML Tables is for working with structured, tabular data. Option C is incorrect; this is not a vision problem. Option D is incorrect; there is no such thing as the Cloud Anomaly Detection API.