Chapter 2

Designing Solutions to Meet Business Requirements

THE PROFESSIONAL CLOUD ARCHITECT CERTIFICATION EXAM OBJECTIVES COVERED IN THIS CHAPTER INCLUDE THE FOLLOWING:

- ✓ **1.1 Designing a solution infrastructure that meets business requirements**
- ✓ **1.2 Designing a solution infrastructure that meets technical requirements**
- ✓ **1.5 Envisioning future solution improvements**

One of the things that distinguishes a cloud architect from a cloud engineer is the architect's need to work with *business requirements*. Architects work with colleagues responsible for business strategy, planning, development, and operations. Before architects can begin to design solutions, they need to understand their organization's cloud vision and objectives that the business is trying to achieve. These objectives frame the possible solutions that are acceptable to the business.

When it comes to the Professional Cloud Architect exam, business requirements are pieces of information that will help you determine the most appropriate answer to some of the questions. It is important to remember that questions on the exam may have more than one answer that appears correct, but only one answer is the best solution to the problem posed. For example, a question may be asked about a small business application that needs to store data in a relational database, and the database will be accessed by only a few users in a single office. You *could* use the globally scalable relational database Cloud Spanner to solve this problem, but Cloud SQL would be a better fit to the requirements and cost less. If you come across a question that seems to have more than one answer, consider all the technical *and* business requirements. You may find that more than one option meets the technical or the business requirements, but only one option meets both.

This chapter reviews several key areas where business requirements are important to understand, including the following:

- Business use cases and product strategy
- Application design and cost considerations
- Systems integration and data management
- Compliance and regulations
- Security
- Success measures

Throughout the chapter, I will reference the case studies used in the Google Professional Cloud Architect exam.

Business Use Cases and Product Strategy

Business requirements may be high-level, broad objectives, or they may be tightly focused specifications of some aspect of a service. *High-level objectives* are tied to strategy, or plan,

to meet some vision and objective. These statements give us clues as to what the cloud solution will look like. In fact, we can often estimate technical requirements just from statements about business strategy and product offerings.

Let's look at the three case studies and see what kinds of information can be derived to help formulate a solution.

EHR Healthcare

The EHR Healthcare case study explicitly lists several business requirements, and from these statements, we can derive several facts about any solution.

- The company provides business-to-business services to insurance providers. The time it takes insurance providers to start using the system, known as *onboarding*, needs to be minimized.
- There is a mix of users, including medical offices, hospitals, and insurance providers. It is likely they all have different needs. For example, small medical offices may need more technical assistance when onboarding, while large insurance providers will likely have specialized data integration requirements.
- Medical records management services cannot have extended periods of downtime. These systems need to be available 99.9 percent of the time.
- Application performance is an issue. Latency needs to be reduced.
- Since the applications store and process protected health information such as medical history, maintaining data confidentiality is a top concern.
- The company is growing rapidly, and system administration costs cannot grow just because more infrastructure is added. System management practices should be designed to allow the organization to add infrastructure without needing to add staff to support it.
- The company wants to use its data to derive insights about industry trends.

This list of business requirements helps us start to understand or at least estimate some likely aspects of the technical requirements. Here are some examples of technical implications that should be considered based on the facts listed previously:

- Since the company is providing services to other businesses, customers will likely use public APIs.
- There are many legacy systems in the insurance industry, so there may be batch processing jobs as well.
- The need for availability calls for redundancy in infrastructure including compute, storage, and networking along with an architecture that prevents any single failure from making services unavailable.
- With a goal of deriving insights from data, the company will likely keep large amounts of data for extended periods of time. This coupled with the sensitivity of the data will require careful planning of access controls and data lifecycle management policies.

- Since the company serves customers in multiple nations and low latency is a goal, services and data will be served from multiple regions. For example, the EU's GDPR restricts the movement of records across national boundaries, which may have implications for region selection, storage strategy, and network topology.
- The adoption of managed services will likely lead to a decrease in infrastructure administration costs. AI and machine learning managed services will allow the company to start deriving insights from data faster than if they built ML models from scratch.

These are just possible requirements, but it helps to keep them in mind when analyzing a use case. This example shows how many possible requirements can be inferred from just a few statements about the business and product strategy. It also shows that architects need to draw on their knowledge of systems design to anticipate requirements that are not explicitly stated, such as the need to keep application response times low, which in turn may require replication of data across multiple regions.

Helicopter Racing League 

The Helicopter Racing League has several explicitly stated business requirements that fall into four categories: predictive analytics, increase viewership, operations, and increasing revenue.

The company wants to improve predictions during races, but they also want to expose their predictive models to business partners. Part of the business plan is to increase the type and amount of telemetry data collected.

Executives want to increase the number of concurrent viewers and reduce latency. This requirement will have a direct impact on technical requirements, especially related to scalability and geographic distribution of content and services.

Another business requirement is to minimize operational complexity and ensure compliance with regulations. This is a common requirement across the case studies.

The requirements specifically call for creating a merchandising revenue stream. It is not specifically detailed, but this may include branded merchandise such as clothing. This is a vague requirement and would require additional work to identify more specific details of the requirement.

Mountkirk Games Strategy 

Mountkirk Games is a company that makes online, session-based, multiplayer games for mobile platforms. The company is already using cloud computing. In the Mountkirk Games case study, we see an example of a product strategy that is focused on building on their cloud and application development experience to create a new product and improve the way they deliver services.

The business requirements include customer-facing requirements, such as supporting multiple gaming platforms and supporting multiple regions, which will improve latency and disaster recovery capabilities.

There is also a call for using managed services and pooled resources as well as support for dynamic scaling. We can see, given their use of Kubernetes, there will likely be an opportunity to use Google Kubernetes Engine as well as other managed services.

Here again, we see that business requirements can help us frame and anticipate some likely technical requirements. There is not enough information in the business requirements to make definitive decisions about technical solutions, but they do allow us to start formulating possible solutions.

TerramEarth Strategy 

The TerramEarth business requirements include improving the ability to predict malfunctions in equipment, increasing the speed and reliability of development workflows, and enabling developers to create custom APIs more efficiently.

The business requirements do not explicitly call for using managed services, but given the emphasis on predictive analytics, it is likely that Vertex AI and other machine learning services, such as GPUs and TPUs, will be employed.

The business requirements also emphasize the importance of developer productivity, including remote workers.

Business requirements are a starting point for formulating a technical solution. Architects must apply their knowledge of systems design to map business requirements into possible technical requirements. After that, they can dig into explicit technical requirements to start to formulate technical solutions.

The key point of this section is that business requirements are not some kind of unnecessary filler in the case studies. They provide the broad context in which a solution will be developed. While they do not provide enough detail to identify solution components on their own, they do help us narrow the set of feasible technical solutions. Business requirements may help you rule out options to an exam question. For example, a data storage solution that distributes data across multiple regions by default may meet all technical requirements, but if a business requirement indicates the data to be stored must be located in a specific country, then the correct answer is the one that lets you limit where the data is stored.

Application Design and Cost Considerations 

In addition to specifying business and product strategy, business requirements may state things that you should consider in application design, such as a preference for managed services and the level of tolerance for disruptions in processing. Implicit in business requirements is the need to minimize costs while meeting business objectives.

One measure of costs is *total cost of ownership (TCO)*. TCO is the combination of all expenses related to maintaining a service, which can include the following:

- Software licensing costs
- Cloud computing costs, including infrastructure and managed services
- Cloud storage costs
- Data ingress and egress charges
- Cost of DevOps personnel to develop and maintain the service
- Cost of third-party services used in an application
- Charges against missed service-level agreements
- Network connectivity charges, such as those for a dedicated connection between an on-premises data center and Google Cloud

While you will want to minimize the TCO, you should be careful not to try to minimize the cost of each component separately. For example, you may be able to reduce the cost of DevOps personnel to develop and maintain a service if you increase your spending on managed services. Also, it is generally a good practice to find a feasible technical solution to a problem before trying to optimize that solution for costs.

Some of the ways to reduce costs while meeting application design requirements include managed services, using preemptible virtual machines, and data lifecycle management. Google also offers sustained uses discounts and reserved VMs, which can help reduce costs.

Managed Services 

Managed services are Google Cloud Platform services that do not require users to perform common configuration and maintenance operations. For example, Cloud SQL is a managed relational database service providing MySQL, SQL Server, and PostgreSQL databases. Database administration tasks, such as backing up data and patching operating systems, are performed by Google and not by customers using the service. Managed services are good options in the following cases:

- Users do not need low-level control over the resources providing the service, such as choosing the operating system to run in a VM.
- Managed services provide a functionality that would be difficult or expensive to implement, such as developing a machine vision application.
- There is little competitive advantage to performing resource management tasks. For example, the competitive advantage that may come from using Apache Spark for analytics stems from the algorithms and analysis methodologies, not from the administration of the Spark cluster.

Architects do not necessarily need to know the details of how managed services work. This is one of their advantages. Architects do not need to develop an expertise in natural language processing to use the Natural Language AI, but they do need to understand what kinds of functions managed services provide. See Table 2.1 for brief descriptions of some of the Google Cloud Platform managed services.

TABLE 2.1 Examples of Google Cloud Platform managed services

Service Name	Service Type	Description
AutoML Tables	AI and machine learning	Machine learning models for structured data
Recommendations AI	AI and machine learning	Personalized recommendations
Natural Language AI	AI and machine learning	Entity recognition, sentiment analysis, and language identification
Cloud Translation	AI and machine learning	Translate between languages
Cloud Vision	AI and machine learning	Understand contents of images
Dialogflow Essentials	AI and machine learning	Development suite for voice and text conversational apps
BigQuery	Analytics	Data warehousing and analytics
Cloud Datalab	Analytics	Interactive data analysis tool based on Jupyter Notebooks
Dataproc	Analytics	Managed Hadoop and Spark service
Cloud Data Fusion	Data management	Data integration and ETL tool
Data Catalog	Data management	Metadata management service
Dataflow	Data management	Stream and batch processing
Cloud Spanner	Database	Global relational database
Cloud SQL	Database	Regional relational database
Cloud Deployment Manager	Development	Infrastructure-as-code service
Cloud Pub/Sub	Messaging	Messaging service
Cloud Composer	Orchestration	Data workflow orchestration service
Bigtable	Storage	Wide column, NoSQL database

(Continues)

TABLE 2.1 Examples of Google Cloud Platform managed services *(continued)*

Service Name	Service Type	Description
Cloud Data Transfer	Storage	Bulk data transfer service
Cloud Memorystore	Storage	Managed cache service using Redis or memcached
Cloud Storage	Storage	Managed object storage service

This table is provided to show the breadth of Google Cloud Platform managed services. The services offered change over time. Managed services may be generally available, or they can be in beta. For a list of current services, see the Google Platform Services Summary at `cloud.google.com/terms/services`.

Reduced Levels of Services

One way to reduce costs is to accept lower levels of service in exchange for lower costs. In GCP there are a number of opportunities to reduce cost in exchange for lower levels of service, including the following:

- Using preemptible virtual machines instead of standard virtual machines
- Using standard tier networking instead of Premium tier
- Using Pub/Sub Lite instead of Pub/Sub
- Using Durable Reduced Availability Storage

These examples are representative of the kinds of trade-offs we make when opting for a lower-cost service.

Preemptible Virtual Machines

One way to minimize computing costs is to use *preemptible virtual machines*, which are short-lived VMs that cost about 80 percent less than their nonpreemptible counterparts. Here are some things to keep in mind about preemptible VMs when considering business and technical requirements:

- Preemptible VMs may be shut down at any time by Google. They will be shut down after running for 24 hours.
- GCP will signal a VM before shutting down. The VM will have 30 seconds to stop processes gracefully.

- Preemptible VMs can have local SSDs and GPUs if additional storage and compute resources are needed.
- If you want to replace a preempted VM automatically, you can create a managed instance group for preemptible VMs. If resources are available, the managed instance group will replace the preempted VM.
- Preemptible VMs can be used with some managed services, such as Cloud Dataproc, to reduce the overall cost of the service.

Preemptible VMs are well suited for batch jobs or other workloads that can be disrupted and restarted. They are not suitable for running services that need high availability, such as a database or user-facing service, like a web server.

Preemptible VMs are also not suitable for applications that manage state in memory or on the local SSD. Preemptible VMs cannot be configured to live migrate; when they are shut down, locally persisted data and data in memory are lost. If you want to use preemptible VMs with stateful applications, consider using Cloud Memorystore, a managed Redis or memcached service for caching data, or a database to store state.

Google Cloud offers Spot VMs, which are similar to preemptible VMs but do not necessarily shut down within the first 24 hours of running. At the time of writing, Spot VMs are in Pre-GA, may have limited support, and may not be feature complete. Throughout this book, we will use the term *Preemptible VMs* to refer to low-cost instances that may be provisioned as traditional Preemptible VMs or as Spot VMs.

Standard vs. Premium Tier Networking

Google Cloud offers two levels or tiers of networking, Standard and Premium.

Standard Tier networking is the lower-performance option. Performance and availability are typically like other cloud providers that depend on the public internet. There are no global SLAs, and Cloud Load Balancing is limited to regional load balancing.

With Premium Tier networking, you experience high performance and reliability, low latency, and a global SLA. You can also use global load balancers that can distribute load across regions. Network traffic is carried on Google's network, not the public internet.

Pub/Sub Lite vs. Pub/Sub

Pub/Sub is a highly scalable messaging service in Google Cloud. Pub/Sub Lite is also horizontally scalable but costs less and provides lower levels of service than Pub/Sub.

Pub/Sub provides for per-message parallelism, automatic scaling, and global routing. Service endpoints are global and regional.

Pub/Sub Lite offers lower availability and durability than Pub/Sub. Messages replication is limited to a single zone unlike Pub/Sub, which provides multizone replication in a single region. Pub/Sub Lite service endpoints are regional, not global. The Lite service also requires users of the service to manage resource capacity.

Pub/Sub Lite can cost up to 85 percent less than Cloud Pub/Sub, but Google recommends that Pub/Sub as the default choice for a managed messaging service.

Durable Reduced Availability Storage

Durable Reduced Availability (DRA) Storage is like Standard Storage but with lower performance and availability. DRA Storage has a 99 percent availability, while Standard Storage has greater than 99.99 percent availability in dual-regions and multiregions and 99.99 percent in regions.

The documentation for Durable Reduced Ability Storage, Multi-Regional Storage, and Regional Storage states that "[u]nless you already are using one of these additional classes, you should use Standard Storage instead." (Source: `cloud.google.com/storage/docs/storage-classes`)

Data Lifecycle Management

When assessing the application design and cost implications of business requirements, consider how to manage storage costs. Storage options lie on a spectrum from short-term to archival storage.

- *Memorystore* is a cache, and data should be kept in cache only if it is likely to be used by an application in the very near future. The purpose of this storage is to reduce the latency of accessing data.

  Caches are not durable. Data stored in Memorystore can disappear at any time. Only data that can be retrieved from another source or regenerated should be stored in a cache.
- Databases, like *CloudSQL* and *Firestore*, store data that needs to be persistently stored and readily accessed by an application or user. Data should be stored in the database when it could possibly be queried or updated. When data is no longer required to be queried or updated, it can be exported and stored in object storage.
- In the case of *time-series databases*, data may be aggregated by larger time spans as time goes on. For example, an application may collect performance metrics every minute. After three days, there is no need to query to the minute level of detail, and data can be aggregated to the hour level. After one month, data can be aggregated to the day level. This incremental aggregation will save space and improve response times for queries that span large time ranges.
- *Object storage* is often used for unstructured data and backups. Standard Storage class should be used for frequently accessed data. If data is accessed at most once a month, then Nearline storage can be used. When data is not likely to be accessed more than once in 90 days, then Coldline storage should be used. Archive storage is appropriate for objects that are not accessed more than once per year.

Consider how to take advantage of Cloud Storage's lifecycle management features, which allow you to specify actions to perform on objects when specific events occur. The two actions supported are deleting an object or changing its storage class. Standard Class storage objects can be migrated to either Nearline, Coldline, or Archive storage. Nearline storage can migrate to Coldline storage or Archive storage. Coldline storage can be migrated to Archive storage. DRA Storage can be transitioned to the other storage classes.

Lifecycle conditions can be based on the following:

- The age of an object
- When it was created, including CreatedBefore and CustomTimeBefore conditions
- Days since a custom time metadata field on an object
- The object's storage class
- The number of versions of an object as well as the number of days since the object became noncurrent
- Whether or not the object is "live" (an object in nonversions bucketed is "live"; archived objects are not live)
- Storage class

You can monitor data lifecycle management either by using Cloud Storage usage logs or by enabling Pub/Sub notifications for Cloud Storage buckets. The latter will send a message to a Pub/Sub topic when an action occurs.

Systems Integration and Data Management

Business requirements can give information that is useful for identifying dependencies between systems and how data will flow through those systems.

Systems Integration Business Requirements

One of an architect's responsibilities is to ensure that systems work together. Business requirements will not specify technical details about how applications should function together, but they will state what needs to happen to data or what functions need to be available to users.

Let's review examples of systems integration considerations in the case studies. These are representative examples of system integration considerations; it is not an exhaustive list.

EHR Healthcare Systems Integration

The EHR Healthcare Systems case study notes that there are several legacy file and API-based integrations with insurance providers that will be replaced over the next several years.

The existing systems will not be migrated to the cloud. This is an example of a rip-and-replace migration strategy.

Even though the existing systems will not be migrated, new cloud-native systems will be developed. As an architect working on that project, you would consider several challenges, including the following:

- Understanding the volume and types of data exchanged
- Deciding how to authenticate service requests
- Encrypting data at rest and in transit
- Managing encryption keys
- Decoupling services to accommodate spikes in service demand
- Designing ingestion and data pipelines
- Monitoring and logging for service performance as well as security
- Using multiregion storage and compute resources for high availability while operating within any regulations that put constraints on where data may be stored and processed

In addition to these technical design issues, the architect and business sponsors will need to determine how to retire existing on-premises systems while bringing the new systems online without disrupting services.

Helicopter Racing League

The Helicopter Racing League is highly focused on improving predictive analytics and integrating their findings with the viewer platform. Consider two types of analytics described in the case study: (1) viewer consumption patterns and engagement and (2) race predictions.

To understand viewer consumption patterns and engagement, the company will need to collect details about viewer behaviors during races. This will likely require ingestion systems that can scale to large volumes of data distributed over a wide geographic area. The ingestion system will likely feed a streaming analysis data pipeline (Cloud Dataflow would be a good option for this service), and the results of the initial analysis as well as telemetry data may be stored for further analysis.

In fact, the data may be stored in two different systems for further analysis. BigQuery is optimized for scanning large volumes of data and would make it a good choice for analyzing data that spans a race or multiple races and entails hundreds of terabytes of data. Bigtable provides low-latency writes and is highly performant for key-based lookups and small scans, such as time-series data for a single viewer over the past 10 minutes.

Mountkirk Games Systems Integration

Let's consider how datastores and microservices architectures can influence systems integration.

Online games, like those produced by Mountkirk Games, use more than one type of datastore. Player data, such as the player's in-game possessions and characteristics, could be stored in a document database like Cloud Datastore, while the time-series data could be

stored in Bigtable, and billing information may be kept in a transaction processing relational database. Architects should consider how data will be kept complete and consistent across datastores. For example, if a player purchases a game item, then the application needs to ensure that the item is added to the player's possession record in the player database and that the charge is authorized by the payment system. If the payment is not authorized, the possession should not be available to the player.

Mountkirk Games uses a microservices architecture. Microservices are single services that implement one single function of an application, such as storing player data or recording player actions. An aggregation of microservices implements an application. Microservices make their functions accessible through application programming interfaces (APIs). Depending on security requirements, services may require that calls to their API functions are authenticated. High-risk services, such as a payment service, may require more security controls than other services. Cloud Endpoints may help to manage APIs and help secure and monitor calls to microservices.

TerramEarth Systems Integration

From the description of the current system, we can see that on-board applications communicate with a centralized data collection system. Some of the data is collected in batches when vehicles return to base, and some is collected as it is generated.

As part of planning for envisioning future needs, an architect working with TerramEarth should consider how to support vehicles sending more data directly, eventually retiring the batch data load process in favor of having real-time or near-real-time data uploads for all vehicles. This would require planning an ingest pipeline that could receive data reliably and perform any preprocessing necessary.

Services that store and analyze the data will need to scale to support millions of vehicles transmitting data. To accomplish this, consider using a Cloud Pub/Sub queue, which allows decoupling and buffering so that data is not lost if the ingestion services cannot keep up with the rate that new data is received.

TerramEarth will likely want to share the growing inventory of data with dealers. This will require integrating TerramEarth and dealer systems. The architect should gather additional details to understand how best to share data with dealers. For example, an architect may ask, "What features of the data are significant to dealers?" Also, architects should consider how dealers would access the data. Dealer applications could query TerramEarth APIs for data, or TerramEarth could send data directly to dealer data warehouses or other reporting platforms.

Each case study has examples of systems integration requirements. When considering these requirements, keep in mind the structure and volume of data exchanged, the frequency of data exchange, the need for authentication, the reliability of each system, how to prevent data loss in the event of a problem with one of the services, and how to protect services from intentional or unintentional bursts in API requests that could overwhelm the receiving services.

Data Management Business Requirements

In addition to using business requirements to understand which systems need to work together, architects can use those requirements to understand data management business requirements. At a minimum, data management considerations include the following:

- How much data will be collected and stored?
- How long will it be stored?
- What processing will be applied to the data?
- Who will have access to the data?

How Much Data Is Stored?

One of the first questions asked about data management is "How much data are we expecting, and at what rate will we receive it?" Knowing the expected volumes of data will help plan for storage.

If data is being stored in a managed service like Cloud Storage, then Google will manage the provisioning of storage space, but you should still understand the volume of data so that you can accurately estimate storage costs and work within storage limits.

It is important to plan for adequate storage capacity. Those responsible for managing storage will also need to know the rate at which new data arrives and existing data is removed from the system. This will give the growth rate in storage capacity.

How Long Is Data Stored?

It is also important to understand how long data will be stored in various storage systems. Data may first arrive in a Cloud Pub/Sub queue, but it is immediately processed by a Cloud Function that removes the data from the queue, transforms the data, and writes it to another storage system. In this case, the data is usually in the queue for a short period of time. If the ingestion process is down, data may accumulate in the Cloud Pub/Sub topic. Since Cloud Pub/Sub is a managed service, the DevOps team would not have to allocate additional storage to the queue if there is a backup of data. GCP will take care of that. They will, however, have to consider how long the data should be retained in the queue. For example, there may be little or no business value in data that is more than seven days old. In that case, the team should configure the Cloud Pub/Sub queue with a seven-day retention period.

If data is stored in Cloud Storage, you can take advantage of the service's lifecycle policies to delete or change the storage class of data as needed. For example, if the data is rarely accessed after 30 days, data can be stored in Nearline while Coldline storage is a good option for data accessed not more than once in 90 days. If it is accessed once a year or less often, then Archive storage is an appropriate choice.

When data is stored in a database, you will have to develop procedures for removing data when it is no longer needed. In this case, the data could be backed up to Cloud Storage for archiving, or it could be deleted without keeping a copy. You should consider the trade-offs

between the possible benefits of having data available and the cost of storing it. For instance, machine learning models can take advantage of large volumes of data, so it may be advisable to archive data even if you cannot anticipate a use for the data at this time.

What Processing Is Applied to the Data?

There are several things to consider about how data is processed. These include the following:

- Distance between the location of stored data and services that will process the data
- Volume of data that is moved from storage to processing services
- Acceptable latency when reading and writing the data
- Stream or batch processing
- In the case of stream processing, how long to wait for late arriving data

The distance between the location of data and where it is processed is an important consideration. This affects both the time it takes to read and write the data as well as, in some cases, the network costs for transmitting the data. If data will be accessed from multiple geographic regions, consider using multiregional storage when using Cloud Storage. If you are storing data in a relational database, consider replicating data to a read-only replica located in a region closer to where the data will be read. If there is a single write instance of the database, then this approach will not improve the time to ingest or update the data.

Understand if the data will be processed in batches or as a stream. Batch processes tend to tolerate longer latencies, so moving data between regions may not create problems for meeting business requirements around how fast the data needs to be processed. If data is now being processed in batch but it will be processed as a stream in the future, consider using Cloud Dataflow, Google's managed Apache Beam service. Apache Beam provides a unified processing model for batch and stream processing.

When working with stream processing, you should consider how you will deal with late arriving and missing data. It is a common practice in stream processing to assume that no data older than a specified time will arrive. For example, if a process collects telemetry data from sensors every minute, a process may wait up to five minutes for data. If the data has not arrived by then, the process assumes that it will never arrive.

A common architecture pattern is to consume data asynchronously by having data producers write data to a Cloud Pub/Sub topic and then having data consumers read from that topic. This helps prevent data loss when consumers cannot keep up with producers. Asynchronous data consumption also enables higher degrees of parallelism, which promotes scalability.

Business requirements help shape the context for systems integration and data management. They also impose constraints on acceptable solutions. In the context of the exam, business requirements may infer technical requirements that can help you identify the correct answer to a question.

Compliance and Regulation

Businesses and organizations may be subject to regulations. For example, it is likely that Mountkirk Games accepts payment using credit cards and so is subject to financial services regulations governing payment cards. Part of analyzing business requirements is to understand which, if any, regulations require compliance. Regulations have different requirements, depending on their objectives. Some widely recognized regulations include the following:

- *Health Insurance Portability and Accountability Act (HIPAA)* addresses privacy security of medical information in the United States.
- *General Data Protection Regulation (GDPR)* defines privacy protections for people in and citizens of the European Union.
- The *Sarbanes-Oxley (SOX) Act* regulates business reporting of publicly traded companies to ensure the accuracy and reliability of financial statements to mitigate the risk of corporate fraud. This is a U.S. federal regulation.
- *Children's Online Privacy Protection Act (COPPA)* is a U.S. law that regulates websites that collect personal information to protect children under the age of 13.
- *Payment Card Industry Data Security Standard (PCI DSS)* is an industry security standard that applies to businesses that accept payment cards. The regulation specifies security controls that must be in place to protect cardholders' data.

It is important to understand what regulations apply to the systems you design. Some regulations apply because the business or organization developing cloud applications operates in a particular jurisdiction. HIPAA applies to healthcare providers with patients and clients in the United States. Companies that operate in the state of California in the United States may also subject to the California Consumer Privacy Act. If a business operates in North America but has customers in Europe, it may be subject to GDPR.

Some regulations apply by virtue of the industry in which the business or organization operates. HIPAA governs healthcare providers and others with access to protected health information. Banks in the United States are subject to the Financial Services Modernization Act, also known as the Gram-Leach-Bliley Act (GLBA), specifying privacy protections for consumers' nonpublic financial information.

Privacy Regulations

Regulations placed on data are often designed to ensure privacy and protect the integrity of data. A large class of regulations govern privacy. HIPAA, GLBA, GDPR, and a host of national laws are designed to limit how personal data is used and to provide individuals with some level of control over their information. More than 40 countries, the European Union, and Singapore have privacy regulations. (See `www.privacypolicies.com/blog/privacy-law-by-country` for a list of countries and links to additional information.) Industry regulations, like PCI DSS, also include protections for keeping data confidential.

From an architect's perspective, privacy regulations require that we plan on ways to protect data through its entire lifecycle. This begins when data is collected, for example, when a patient enters medical information into a doctor's scheduling application. Protected data should be encrypted before transmitting it to cloud applications and databases. Data should also be encrypted when stored. This is sometimes called encrypting *data in transit/motion* and *data at rest*.

Access controls should be in place to ensure that only authenticated and authorized people and service accounts can access protected data. In some cases, applications may need to log changes to data. In those cases, logs must be tamperproof.

Networks and servers should be protected with firewalls and other measures to limit access to servers that process protected data. With Google Cloud, architects and developers can take advantage of the Cloud Identity-Aware Proxy to verify a user's identity in the context of a request to a service and determine whether that operation should be allowed.

Security best practices should be used as well. This includes following the *principle of least privilege*, so users and service accounts have only the permissions that are required to carry out their responsibilities. Also practice *defense in depth*. That principle assumes any security control may be compromised, so systems and data should be protected with multiple different types of controls.

Data Integrity Regulations

Data integrity regulations are designed to protect against fraud. SOX, for example, requires regulated businesses to have controls on systems that prevent tampering with financial data. In addition, businesses need to be able to demonstrate that they have these controls in place. This can be done with application logs and reports from security systems, such as vulnerability scanners or anti-malware applications.

Depending on regulations, applications that collect or process sensitive data may need to use message digests and digital signing to protect data with tamper-proof protections.

Many of the controls used to protect privacy, such as encryption and blocking mechanisms, like firewalls, are also useful for protecting data integrity.

In addition to stated business requirements, it is a good practice to review compliance and regulations with the business owners of the systems that you design. You will often find that the security protections required by regulations overlap with the controls that are used in general to secure information systems and data.

Security

Information security, also known as infosec and cybersecurity, is a broad topic. In this section, you will focus on understanding high-level security requirements based on business requirements. Chapter 7, "Designing for Security and Legal Compliance," will go into more detail on cloud security measures.

Business requirements for security tend to fall into three areas: confidentiality, integrity, and availability.

Confidentiality

Confidentiality is about limiting access to data. Only users and service accounts with legitimate business needs should have access to data. Even if regulations do not require keeping some data confidential, it is a good practice to protect confidentiality. Using HTTPS instead of HTTP and encrypting data at rest should be standard practice. Fortunately, for GCP users, Google Cloud provides encryption at rest by default.

When we use default encryption, Google manages the encryption keys. This requires the least work from customers and DevOps teams. If there is a business requirement that the customer and not Google manage the keys, you can design for customer-managed encryption keys using Cloud KMS, or you can use customer-supplied encryption keys. In the former case, keys are kept in the cloud. When using customer-supplied keys, they are stored outside of GCP's key management infrastructure.

Protecting servers and networks is also part of ensuring confidentiality. When collecting business requirements, look for requirements for additional measures, for example, if a particular hardened operating system must be used. This can limit your choice of computing services. Also determine what kind of authentication is required. Will multifactor authentication be needed? Start thinking about roles and permissions. Will custom IAM roles be required? Determine what kinds and level of audit logging are required.

Integrity

Protecting *data integrity* is a goal of some of the regulations discussed earlier, but it is a general security requirement in any business application. The basic principle is that only people or service accounts with legitimate business needs should be able to change data and then only for legitimate business purposes.

Access controls are a primary tool for protecting data integrity. Google Cloud Platform has defined many roles to grant permissions easily according to common business roles. For example, App Engine has roles for administrators, code viewers, deployers, and others. This allows security administrators to assign fine-grained roles to users and service accounts while still maintaining least privileges.

Server and network security measures also contribute to protecting data integrity.

When collecting and analyzing business requirements, seek to understand the roles that are needed to carry out business operations and which business roles or positions will be assigned those roles. Pay particular attention to who is allowed to view and update data, and use separate roles for users who have read-only access.

Availability

Availability is a bit different from confidentiality and integrity. Here the goal is to ensure that users have access to a system. Malicious activities, such as distributed denial-of-service (DDoS) attacks, malware infection, and encrypting data without authorization (ransomware attacks), can degrade availability.

During the requirements-gathering phase of a project, consider any unusual availability requirements. With respect to security, the primary focus is on preventing malicious acts. From a reliability perspective, availability is about ensuring redundant systems and failover mechanisms to ensure that services continue to operate despite component failures.

Security should be discussed when collecting business requirements. At this stage, it is more important to understand what the business expects in terms of confidentiality, integrity, and availability. We get into technical and implementation details after first understanding the business requirements.

Success Measures

Businesses and other organizations are moving to the cloud because of its value. Businesses can more efficiently develop, deploy, and run applications, especially when they are designed in ways that take advantage of the cloud. Decision-makers typically want to measure the value of their projects. This enables them to allocate resources to the more beneficial projects while avoiding others that may not prove worthwhile. Two common ways to measure progress and success are key performance indicators and return on investment.

Key Performance Indicators

KPIs are a measurable value of some aspect of the business or operations that indicates how well the organization is achieving its objectives. A sales department may have total value of sales in the last week as a KPI, while a DevOps team might use CPU utilization as a KPI of efficient use of compute resources.

Project KPIs

Project managers may use KPIs to measure the progress of a cloud migration project. KPIs in that case may include a volume of data migrated to the cloud and no longer stored on-premises, the number of test cases run each day in the cloud instead of on-premises, or the number of workload hours running in the cloud instead of on-premises.

You can see from these examples that KPIs can be highly specific and tailored to a particular kind of project. Often, you will have to define how you will measure a KPI. For example, a workload hour may be defined based on the wall clock time and the number of CPUs dedicated to a workload.

The definition of a KPI should allow for an obvious way to measure the indicator. The details of the definition should be stated early in the project to help team members understand the business objectives and how they will be measured.

Operations KPI

Line-of-business managers may use KPIs to measure how well operations are running. These KPIs are closely aligned with business objectives. A retailer may use total sales revenue, while a telecommunications company may monitor reduction in customer churn, in other words, customers taking their business to a competitor. A financial institution that makes loans might use the number of applications reviewed as a measure of how well the business is running.

For architects, it is important to know how the business will measure the success of a project or operation. KPIs help us understand what is most important to the business and what drives decision-makers to invest in a project or line of business.

Return on Investment

ROI is a way of measuring the monetary value of an investment. ROI is expressed as a percentage, and it is based on the value of some aspect of the business after an investment when compared to its value before the investment. The return, or increase or loss, after an investment divided by the cost of the investment is the ROI. The formula for ROI is as follows:

$$\text{ROI} = \left[\left(\text{value of investment} - \text{cost of investment}\right) / \text{cost of investment}\right] * 100$$

The value of investment is measured for a fixed period of time, such as 1 year or 3 years. For example, if a company invests $100,000 in new equipment and this investment generates a value of $145,000 over 3 years, then the ROI is 45 percent over 3 years.

In cloud migration projects, the investment includes the cost of cloud services, employee and contractor costs, and any third-party service costs. The value of the investment can include the expenses saved by not replacing old equipment or purchasing new equipment, savings due to reduced power consumption in a data center, and new revenue generated by applications and services that scale up in the cloud but were constrained when run on-premises.

Success measures such as KPIs and ROI are a formal way of specifying what the organization values with respect to a project or line of business. As an architect, you should know which success measures are being used so that you can understand how the business measures the value of the systems that you design.

Summary

The first stages of a cloud project should begin with understanding the business use cases and product strategy. This information sets the context for later work on the technical requirements analysis.

One part of business requirements analysis includes application design and cost considerations. Application design considerations include assessing the possible use of managed services and lower classes of service that cost less than standard services. Data lifecycle management is also a factor in application design.

In addition to business drivers, consider regulations that may apply to your projects. Many regulations are designed to protect individuals' privacy or to ensure the integrity of data to prevent fraud. Compliance with regulations may require additional security controls or application features that otherwise would not be implemented.

Security business requirements can be framed around three objectives: protecting confidentiality, preserving the integrity of data, and ensuring the availability of services, especially with respect to malicious acts that could disrupt services. There may be ancillary security requirements as well. For example, if a company sends a digital purchase order, the recipient should be expected to send a signed acknowledgment proving that they received it. This kind of nonrepudiation requirement may occur with external business processes.

Business and other organizations will often monitor the progress of projects and the efficiency and profitability of lines of business using success measures, such as KPIs and ROI.

Exam Essentials

Study the case studies: EHR Healthcare, Helicopter Racing League, Mountkirk Games, and TerramEarth. You will have access to the case studies during the exam, but you can save time if you are already familiar with the details of each. Also, think through the implications of the business requirements to understand how they constrain technical solution options.

Understand business terms like total cost of ownership (TCO), key performance indicators (KPIs), and return on investment (ROI). You will almost certainly not have to calculate any of these measures, but you should understand what they measure and why they are used by business executives and analysts.

Learn about Google Cloud Platform managed services and for what purposes they are used. These services can be used instead of deploying and managing applications and managing servers, storage, networking, and so forth. If a business requirement includes using managed services or reducing the workload on a DevOps team, you may be able to use one or more of these services to solve problems presented on the exam.

Understand the elements of data management. This includes how much data will be collected and stored, how long it will be stored, what processing will be applied to it, and who will have access to it.

Understand how compliance with regulations can introduce additional business requirements. Businesses and organizations may be subject to one or more regulations. Part of analyzing business requirements is to understand which, if any, regulations require compliance.

Understand the three main aspects of security with respect to business requirements. They are confidentiality, integrity, and availability. Confidentiality focuses on limiting access to data to those who have a legitimate business need for it. Integrity ensures that data is not tampered with or corrupted in ways that could enable fraud or other misuse. Availability addresses the need to prevent malicious acts from disrupting access to services.

Know why decision-makers use success measures. Two well-known success measures are KPIs and ROI. KPIs are measures that are specific to a business operation. ROI is a general measure based on the cost of an investment versus the additional benefit realized because of that investment.

Review Questions

1. In the TerramEarth case study, the volume of data and compute load will be most affected by what characteristics of the TerramEarth systems?
   - **A.** The number of dealers and customers
   - **B.** The number of vehicles, the number of sensors on vehicles, network connectivity, and the types of data collected
   - **C.** The type of storage used
   - **D.** Compliance with regulations

2. You have received complaints from customers about long wait times while loading application pages in their browsers, especially pages with several images. Your director has tasked you with reducing latency when accessing and transmitting data to a client device outside the cloud. Which of the following would you use? (Choose two.)
   - **A.** Multiregional storage
   - **B.** Coldline storage
   - **C.** CDN
   - **D.** Cloud Pub/Sub
   - **E.** Cloud Dataflow

3. Mountkirk Games will analyze game players' usage patterns. This will require collecting time-series data including game state. What database would be a good option for doing this?
   - **A.** BigQuery
   - **B.** Bigtable
   - **C.** Cloud Spanner
   - **D.** Cloud Storage

4. You have been hired to consult with a new data warehouse team. They are struggling to meet schedules because they repeatedly find problems with data quality and must write preprocessing scripts to clean the data. What managed service would you recommend for addressing these problems?
   - **A.** Cloud Dataflow
   - **B.** Cloud Dataproc
   - **C.** Cloud Dataprep
   - **D.** Cloud Datastore

5. You have deployed an application that receives data from sensors on manufacturing equipment. Sometimes more data arrives than can be processed by the current set of Compute Engine instances. Business managers do not want to run additional VMs. What changes could you make to ensure that data is not lost because it cannot be processed as it is sent from the equipment? Assume that business managers want the lowest-cost solution.

   A. Write data to local SSDs on the Compute Engine VMs.

   B. Write data to Cloud Memorystore and have the application read data from the cache.

   C. Write data from the equipment to a Cloud Pub/Sub queue and have the application read data from the queue.

   D. Tune the application to run faster.

6. Your company uses Apache Spark for data science applications. Your manager has asked you to investigate running Spark in the cloud. Your manager's goal is to lower the overall cost of running and managing Spark. What would you recommend?

   A. Run Apache Spark in Compute Engine.

   B. Use Cloud Dataproc with preemptible virtual machines.

   C. Use Cloud Data Fusion.

   D. Use Cloud Memorystore with Apache Spark running in Compute Engine.

7. You are working with a U.S. hospital to extract data from a legacy electronic health record (EHR) system. The hospital has offered to provide business requirements, but there is little information about regulations in the documented business requirements. What regulations would you look to for more guidance on complying with relevant regulations?

   A. GDPR

   B. SOX

   C. HIPAA

   D. PCI DSS

8. What security control can be used to help detect changes to data?

   A. Firewall rules

   B. Message digests

   C. Authentication

   D. Authorization

9. Your company has a data classification scheme for categorizing data as secret, sensitive, private, and public. There are no confidentiality requirements for public data. All other data must be encrypted at rest. Secret data must be encrypted with keys that the company controls. Sensitive and private data can be encrypted with keys managed by a third party. Data will be stored in GCP. What would you recommend to meet these requirements while minimizing cost and administrative overhead?

   A. Use Cloud KMS to manage keys for all data.

   B. Use Cloud KMS for secret data and Google default encryption for other data.

   C. Use Google default encryption for all data.

   D. Use a custom encryption algorithm for all data.

10. You manage a service with several databases. The queries to the relational database are increasing in latency. Reducing the amount of data in tables will improve performance and reduce latency. The application administrator has determined that approximately 60 percent of the data in the database is more than 90 days old and has never been queried and does not need to be in the database. You are required to keep the data for five years in case it is requested by auditors. What would you propose to decrease query latency without increasing costs—or at least keeping any cost increases to a minimum?

    A. Horizontally scale the relational database.

    B. Vertically scale the relational database.

    C. Export data more than 90 days old, store it in Cloud Storage Archive class storage, and delete that data from the relational database.

    D. Export data more than 90 days old, store it in Cloud Storage multiregional class storage, and delete that data from the relational database.

11. Your company is running several custom applications that were written by developers who are no longer with the company. The applications frequently fail. The DevOps team is paged more for these applications than any others. You propose replacing those applications with several managed services in GCP. A manager who is reviewing your cost estimates for using managed services in GCP notes that the cost of the managed services will be more than what they pay for internal servers. What would you recommend as the next step for the manager?

    A. Nothing. The manager is correct—the costs are higher. You should reconsider your recommendation.

    B. Suggest that the manager calculate total cost of ownership, which includes the cost to support the applications as well as infrastructure costs.

    C. Recommend running the custom applications in Compute Engine to lower costs.

    D. Recommend rewriting the applications to improve reliability.

12. A director at an online gaming startup has asked for your recommendation on how to measure the success of the migration to GCP. The director is particularly interested in customer satisfaction. What KPIs would you recommend?

    A. Average revenue per customer per month

    B. Average time played per customer per week

    C. Average time played per customer per year

    D. Average revenue per customer per year

13. Mountkirk Games is implementing a player analytics system. You have been asked to document requirements for a stream processing system that will ingest and preprocess data before writing it to the database. The preprocessing system will collect data about each player for one minute and then write a summary of statistics about that database. The project manager has provided the list of statistics to calculate and a rule for calculating values for missing data. What other business requirements would you ask of the project manager?

    A. How long to store the data in the database?

    B. What roles and permissions should be in place to control read access to data in the database?

    C. How long to wait for late-arriving data?

    D. A list of managed services that can be used in this project.

14. A new data warehouse project is about to start. The data warehouse will collect data from 14 different sources initially, but this will likely grow over the next 6 to 12 months. What managed GCP service would you recommend for managing metadata about the data warehouse sources?
    - **A.** Data Catalog
    - **B.** Cloud Dataprep
    - **C.** Cloud Dataproc
    - **D.** BigQuery

15. You are consulting for a multinational company that is moving its inventory system to GCP. The company wants to use a managed database service, and it requires SQL and strong consistency. The database should be able to scale to global levels. What service would you recommend?
    - **A.** Bigtable
    - **B.** Cloud Spanner
    - **C.** Cloud Datastore
    - **D.** BigQuery

16. TerramEarth has interviewed dealers to better understand their needs regarding data. Dealers would like to have access to the latest data available, and they would like to minimize the amount of data they have to store in their databases and object storage systems. How would you recommend that TerramEarth provide data to their dealers?
    - **A.** Extract dealer data to a CSV file once per night during off-business hours and upload it to a Cloud Storage bucket accessible to the dealer.
    - **B.** Create an API that dealers can use to retrieve specific pieces of data on an as-needed basis.
    - **C.** Create a database dump using the database export tool so that dealers can use the database import tool to load the data into their databases.
    - **D.** Create a user account on the database for each dealer and have them log into the database to run their own queries.

17. Your company has large volumes of unstructured data stored on several network-attached storage systems. The maintenance costs are increasing, and management would like to consider alternatives. What GCP storage system would you recommend?
    - **A.** Cloud SQL
    - **B.** Cloud Storage
    - **C.** Cloud Datastore
    - **D.** Bigtable

18. A customer-facing application is built using a microservices architecture. One of the services does not scale as fast as the service that sends it data. This causes the sending service to wait while the other service processes the data. You would like to change the integration to use asynchronous instead of synchronous calls. What is one way to do this?

A. Create a Cloud Pub/Sub topic, have the sending service write data to the topic, and have the receiving service read from the topic.

B. Create a Cloud Storage bucket, have the sending service write data to the topic, and have the receiving service read from the topic.

C. Have the sending service write data to local drives, and have the receiving service read from those drives.

D. Create a Bigtable database, have the sending service write data to the topic, and have the receiving service read from the topic.

19. A key initiative at TerramEarth is to use the data that is collected from vehicles to predict when equipment will break down. What managed services would you recommend TerramEarth to consider?

A. Bigtable

B. Cloud Dataflow

C. Cloud AutoML

D. Cloud Spanner

20. A team of data scientists is more proficient with statistics than with coding extraction, transformation and loading pipelines. The data scientists would like to use a managed service specifically designed for ETL. What GCP service would you recommend?

A. Cloud Data Fusion

B. Cloud BigQuery

C. Cloud Data Catalog

D. Cloud Pub/Sub