AWS Solutions Architect

Knowledge required to become an AWS Solutions Architect.

- Understanding Amazon Web Services (AWS)
  - What is Cloud Computing?
  - Why is the Cloud so Popular?
  - The Five Pillars of a Well-Architected Framework

Understanding Amazon Web Services (AWS)

An introduction to what AWS is and how it use used in modern Business Platforms

What is Cloud Computing?

Cloud computing is essentially a group of computers (servers) and other compute resources managed by a third party (AWS, Microsoft, Google) in a data center somewhere. One important characteristic of the leading cloud providers is the ability to quickly and frictionlessly provision resources.

The cloud computing model is one that offers computing services such as compute, storage, databases, networking, software, machine learning, and analytics over the internet and on demand.

Cloud Computing has quickly gained popularity for its ability to easily provision and manage resources. Before cloud computing, it was required that businesses setup, provision, and manage their own data centers. You can use AWS to run a virtual server for a couple dollars a months, but for the same hardware to be provisioned and managed privately can cost significantly more.

So how does this all work? Cloud computing uses a process called virtualization. Virtualization is the process of running multiple virtual instances on top of a physical computer system using an abstract layer sitting on top of actual hardware. A hypervisor is a computing layer that enables multiple operating systems to execute in the same physical compute resource. These operating systems running on top of these hypervisors are VMs – a component that can emulate a complete computing environment using only software but as if it was running on bare metal. Hypervisors, also known as Virtual Machine Monitors (VMMs), manage these VMs while running side by side. A hypervisor creates a logical separation between VMs, and it provides each of them with a slice of the available compute, memory, and storage resources.

What is Cloud Computing?

Cloud computing is essentially a group of computers (servers) and other compute resources managed by a third party (AWS, Microsoft, Google) in a data center somewhere.

What is Virtualization?

Cloud computing is essentially a group of computers (servers) and other compute resources managed by a third party (AWS, Microsoft, Google) in a data center somewhere.

Why is the Cloud so Popular?

There are multiple reasons why the cloud market is growing so fast. Some of them are listed here: Elasticity, Security, Availability, Faster hardware cycles, System administration staff, Faster time to market. These factors are crucial when trying to develop and maintain a modern tech stack and platform.

Elasticity may be one of the most important reasons for the cloud's popularity. More formally defined, elasticity is the ability of a computing environment to adapt to changes in workload by automatically provisioning or shutting down computing resources to match the capacity needed by the current workload.

One of the worst arguments for not migrating to the cloud are concerns over security. You probably have a better chance of getting into the Pentagon without a badge than getting into an Amazon data center. Amazon employees with access to the building must authenticate themselves four times to step on the data center floor.

In a cloud environment, spinning up new resources can take just a few minutes. So, we can configure minimal environments knowing that additional resources are a click away. These environments are also just as easily shut down and allow you to be flexible with your experimentation.

Another benefit to being on a cloud provider is the physical infrastructure around the compute resources. To increase resilience, data centers have discrete Uninterruptable Power Supplies (UPSes) and onsite backup generators.

It is also easy to upgrade to newer better compute resources, whenever AWS offers new and more powerful processor types, using them is as simple as stopping an instance, changing the processor type, and starting the instance again. In many cases, AWS may keep the price the same even when better and faster processors and technology become available.

The Five Pillars of a Well-Architected Framework

If there is one must-read white paper from AWS, it is the paper titled AWS Well-Architected Framework, which spells out the five pillars of a well-architected framework. The full paper can be found here.

> A **component** is the code, configuration, and AWS Resources that together deliver against a requirement. A component is often the unit of technical ownership, and is decoupled from other components.

> The term **workload** is used to identify a set of components that together deliver business value. A workload is usually the level of detail that business and technology leaders communicate about.

> We think about **architecture** as being how components work together in a workload. How components communicate and interact is often the focus of architecture diagrams.

> **Milestones** mark key changes in your architecture as it evolves throughout the product lifecycle (design, implementation, testing, go live, and in production).

> Within an organization the **technology portfolio** is the collection of workloads that are required for the business to operate.

General Design Principles

In essence the paper tries to provide guidance on how to build a well architected framework, while each pillar has different principles, there are a few principles that are overarching:

- **Stop guessing your capacity needs:** If you make a poor capacity decision when de ploying a workload, you might end up sitting on expensive idle resources or dealing with the performance implications of limited capacity. With cloud computing, these problems

can go away. You can use as much or as little capacity as you need, and scale up and down automatically.

- **Test systems at production scale:** In the cloud, you can create a production-scale test environment on demand, complete your testing, and then decommission the resources. Because you only pay for the test environment when it's running, you can simulate your live environment for a fraction of the cost of testing on premises.
- **Automate to make architectural experimentation easier:** Automation allows you to create and replicate your workloads at low cost and avoid the expense of manual effort. You can track changes to your automation, audit the impact, and revert to previous parameters when necessary.
- **Allow for evolutionary architectures:** Allow for evolutionary architectures. In a tra ditional environment, architectural decisions are often implemented as static, onetime events, with a few major versions of a system during its lifetime. As a business and its context continue to evolve, these initial decisions might hinder the system's ability to deliver changing business requirements. In the cloud, the capability to automate and test on demand lowers the risk of impact from design changes. This allows systems to evolve over time so that businesses can take advantage of innovations as a standard practice.
- **Drive architectures using data:** In the cloud, you can collect data on how your ar chitectural choices affect the behavior of your workload. This lets you make factbased decisions on how to improve your workload. Your cloud infrastructure is code, so you can use that data to inform your architecture choices and improvements over time.
- **Improve through game days:** Test how your architecture and processes perform by regularly scheduling game days to simulate events in production. This will help you understand where improvements can be made and can help develop organizational experience in dealing with events.

This table is a highlight of what a well-architected framework should encompass.

Name	**Description**
Operational Excellence	The ability to support development and run workloads effectively, gain insight into their operations, and to continuously improve supporting processes and procedures to deliver business value.
Security	The security pillar describes how to take advantage of cloud technologies to protect data, systems, and assets in a way that can improve your security posture.
Reliability	The reliability pillar encompasses the ability of a workload to perform its intended function correctly and consistently when it’s expected to. This includes the ability to operate and test the workload through its total lifecycle. This paper provides in-depth, best practice guidance for implementing reliable workloads on AWS.

Performance Efficiency	The ability to use computing resources efficiently to meet system requirements, and to maintain that efficiency as demand changes and technologies evolve.
Cost Optimization	The ability to run systems to deliver business value at the lowest price point.
(and part of Performance Efficiency) Sustainability	The ability to continually improve sustainability impacts by reducing energy consumption and increasing efficiency across all components of a workload by maximizing the benefits from the provisioned resources and minimizing the total resources required.

First Pillar - Operational Excellence

Operational Excellence will be the definitive factor in the success of any well-architected framework. It sets the standard for the rest of the architecture and constantly working towards and maintain operational excellence will assure that your framework is well-architected.

- Perform operations (infrastructure) as code.
- Make frequent, small, reversible changes.
- Refine operations procedures frequently.
- Anticipate failure.
- Learn from all operational failures.

There are four best practice areas for operational excellence in the cloud:

Organization

Your teams need to have a shared understanding of your entire workload, their role in it, and shared business goals to set the priorities that will enable business success. Well-defined priorities will maximize the benefits of your efforts. Evaluate internal and external customer needs involving key stakeholders, including business, development, and operations teams, to determine where to focus efforts. Evaluating cus tomer needs will ensure that you have a thorough understanding of the support that is required to achieve business outcomes. Ensure that you are aware of guidelines or obligations defined by your organizational governance and external factors, such as regulatory compliance requirements and industry standards, that may mandate or emphasize specific focus. Validate that you have mechanisms to identify changes to internal governance and external compliance requirements. If no requirements are identified, ensure that you have applied due diligence to this determination. Review your priorities regularly so that they can be updated as needs change.

How do you determine what your priorities are?

Everyone needs to understand their part in enabling business success. Have shared goals in order to set priorities for resources. This will maximize the benefits of your efforts.

How do you structure your organization to support your business outcomes?

Your teams must understand their part in achieving business outcomes. Teams need to understand their roles in the success of other teams, the role of other teams in their success, and have shared goals. Understanding responsibility, ownership, how decisions are made, and who has authority to make decisions will help focus efforts and maximize the benefits from your teams.

How does your organizational culture support your business outcomes?

Provide support for your team members so that they can be more effective in taking action and supporting your business outcome.

Prepare

To prepare for operational excellence, you have to understand your workloads and their expected behaviors. You will then be able design them to provide insight to their status and build the procedures to support them.

Design your workload so that it provides the information necessary for you to understand its internal state (for example, metrics, logs, events, and traces) across all components in support of observability and investigating issues. Iterate to develop the telemetry necessary to monitor the health of your workload, identify when outcomes are at risk, and enable effective responses. When instrumenting your workload, capture a broad set of information to enable situational awareness (for example, changes in state, user activity, privilege access, utilization counters), knowing that you can use filters to select the most useful information over time.

How do you design your workload so that you can understand its state?

Design your workload so that it provides the information necessary across all components (for example, metrics, logs, and traces) for you to understand its internal state. This enables you to provide effective responses when appropriate.

How do you reduce defects, ease remediation, and improve flow into production?

Adopt approaches that improve flow of changes into production, that enable refactoring, fast feedback on quality, and bug fixing. These accelerate beneficial changes entering production, limit issues deployed, and enable rapid identification and remediation of issues introduced through deployment activities.

How do you mitigate deployment risks?

Adopt approaches that provide fast feedback on quality and enable rapid recovery from changes that do not have desired outcomes. Using these practices mitigates the impact of issues introduced through the deployment of changes.

How do you know that you are ready to support a workload?

Evaluate the operational readiness of your workload, processes and procedures, and personnel to understand the operational risks related to your workload.

Operate

Successful operation of a workload is measured by the achievement of business and customer outcomes. Define expected outcomes, determine how success will be measured, and identify metrics that will be used in those calculations to determine if your workload and operations are successful. Operational health includes both the health of the workload and the health and success of the operations activities performed in support of the workload (for example, deployment and incident response). Establish metrics baselines for improvement, investigation, and intervention, collect and analyze your metrics, and then validate your understanding of operations success and how it changes over time. Use collected metrics to determine if you are satisfying customer and business needs, and identify areas for improvement.

How do you understand the health of your workload?

Define, capture, and analyze workload metrics to gain visibility to workload events so that you can take appropriate action.

How do you understand the health of your operations?

Define, capture, and analyze operations metrics to gain visibility to operations events so that you can take appropriate action.

How do you manage workload and operations events?
Prepare and validate procedures for responding to events to minimize their disruption to your workload.

Evolve

You must learn, share, and continuously improve to sustain operational excellence. Dedicate work cycles to making continuous incremental improvements. Perform post incident analysis of all customer impacting events. Identify the contributing factors and preventative action to limit or prevent recurrence. Communicate contributing factors with affected communities as appropriate. Regularly evaluate and prioritize opportunities for improvement (for example, feature requests, issue remediation, and compliance requirements), including both the workload and operations procedures. Include feedback loops within your procedures to rapidly identify areas for improvement and capture learnings from the execution of operations.

How do you evolve operations?
Dedicate time and resources for continuous incremental improvement to evolve the effectiveness and efficiency of your operations.

Second Pillar - Security

To enable system security and to guard against nefarious actors and vulnerabilities, AWS recommends these architectural principles:

- Always enable traceability.
- Apply security at all levels.
- Implement the principle of least privilege.
- Secure the system at all levels: application, data, operating system, and hardware. Automate security best practices.

There are six best practice areas for security in the cloud:

Security

To operate your workload securely, you must apply overarching best practices to every area of security. Take requirements and processes that you have defined in operational excellence at an organizational and workload level, and apply them to all areas.

How do you securely operate your workload?

To operate your workload securely, you must apply overarching best practices to every area of security. Take requirements and processes that you have defined in operational excellence at an organizational and workload level, and apply them to all areas. Staying up to date with AWS and industry recommendations and threat intelligence helps you evolve your threat model and control objectives. Automating security processes, testing, and validation allow you to scale your security operations.

Identity and Access Management

Identity and access management are key parts of an information security program, ensuring that only authorized and authenticated users and components are able to access your resources, and only in a manner that you intend. For example, you should define principals (that is, accounts, users, roles, and services that can perform actions in your account), build out policies aligned with these principals, and implement strong credential management. These privilege-management elements form the core of authentication and authorization.

How do you manage identities for people and machines?

There are two types of identities you need to manage when approaching operating secure AWS workloads. Understanding the type of identity you need to manage and grant access helps you ensure the right identities have access to the right resources under the right conditions. Human Identities: Your administrators, developers, operators, and end users require an identity to access your AWS environments and applications. These are members of your organization, or external users with whom you collaborate, and who interact with your AWS resources via a web browser, client application, or interactive command-line tools. Machine Identities: Your service applications, operational tools, and workloads require an identity to make requests to AWS services - for example, to read data. These identities include machines running in your AWS environment such as Amazon EC2 instances or AWS Lambda functions. You may also manage machine identities for external parties who need access. Additionally, you may also have machines outside of AWS that need access to your AWS environment.

How do you manage permissions for people and machines?

Manage permissions to control access to people and machine identities that require access to AWS and your workload. Permissions control who can access what, and under what conditions.

Detection

You can use detective controls to identify a potential security threat or incident. They are an essential part of governance frameworks and can be used to support a quality process, a legal or

compliance obligation, and for threat identification and response efforts. There are different types of detective controls. For example, conducting an inventory of assets and their detailed attributes promotes more effective decision making (and lifecycle controls) to help establish operational baselines. You can also use internal auditing, an examination of controls related to information systems, to ensure that practices meet policies and requirements and that you have set the correct automated alerting notifications based on defined conditions. These controls are important reactive factors that can help your organization identify and understand the scope of anomalous activity.

How do you detect and investigate security events?

Capture and analyze events from logs and metrics to gain visibility. Take action on security events and potential threats to help secure your workload.

Infrastructure Protection

Infrastructure protection encompasses control methodologies, such as defense in depth, necessary to meet best practices and organizational or regulatory obligations. Use of these methodologies is critical for successful, ongoing operations in either the cloud or on-premises.

How do you protect your network resources?

Any workload that has some form of network connectivity, whether it’s the internet or a private network, requires multiple layers of defense to help protect from external and internal network-based threats.

How do you protect your compute resources?

Compute resources in your workload require multiple layers of defense to help protect from external and internal threats. Compute resources include EC2 instances, containers, AWS Lambda functions, database services, IoT devices, and more.

Data Protection

Before architecting any system, foundational practices that influence security should be in place. For example, data classification provides a way to categorize organizational data based on levels of sensitivity, and encryption protects data by way of rendering it unintelligible to unauthorized access. These tools and techniques are important because they support objectives such as preventing financial loss or complying with regulatory obligations.

How do you classify your data?

Classification provides a way to categorize data, based on criticality and sensitivity in order to help you determine appropriate protection and retention controls.

How do you protect your data at rest?

Protect your data at rest by implementing multiple controls, to reduce the risk of unauthorized access or mishandling.

How do you protect your data in transit?

Protect your data in transit by implementing multiple controls to reduce the risk of unauthorized access or loss.

Incident Response

Even with extremely mature preventive and detective controls, your organization should still put processes in place to respond to and mitigate the potential impact of security incidents. The architecture of your workload strongly affects the ability of your teams to operate effectively during an incident, to isolate or contain systems, and to restore operations to a known good state. Putting in place the tools and access ahead of a security incident, then routinely practicing incident response through game days, will help you ensure that your architecture can accommodate timely investigation and recovery.

How do you anticipate, respond to, and recover from incidents?

Preparation is critical to timely and effective investigation, response to, and recovery from security incidents to help minimize disruption to your organization.

Third Pillar - Reliability

Reliability is one of, if not the most important pillars. *Can we rely on this framework?* For example, at any given time, there are at least six copies of any object stored in Amazon S3, meaning it has only a 0.00001% chance of having a data loss. The well-architected framework paper recommends these design principles to enhance reliability:

- Continuously test backup and recovery processes.
- Design systems so that they can automatically recover from a single component failure.
- Leverage horizontal scalability whenever possible to enhance overall system availability.

- Use automation to provision and shut down resources depending on traffic and usage to minimize resource bottlenecks.
- Manage change with automation.

There are four best practice areas for reliability in the cloud:

Foundations

Foundational requirements are those whose scope extends beyond a single workload or project. Before architecting any system, foundational requirements that influence reliability should be in place. For example, you must have sufficient network bandwidth to your data center.

How do you manage service quotas and constraints?

For cloud-based workload architectures, there are service quotas (which are also referred to as service limits). These quotas exist to prevent accidentally provisioning more resources than you need and to limit request rates on API operations so as to protect services from abuse. There are also resource constraints, for example, the rate that you can push bits down a fiber-optic cable, or the amount of storage on a physical disk.

How do you plan your network topology?

Workloads often exist in multiple environments. These include multiple cloud environments (both publicly accessible and private) and possibly your existing data center infrastructure. Plans must include network considerations such as intra- and inter-system connectivity, public IP address management, private IP address management, and domain name resolution.

Workload Architecture

A reliable workload starts with upfront design decisions for both software and infrastructure. Your architecture choices will impact your workload behavior across all five Well-Architected pillars. For reliability, there are specific patterns you must follow.

How do you design your workload service architecture?

Build highly scalable and reliable workloads using a service-oriented architecture (SOA) or a microservices architecture. Service-oriented architecture (SOA) is the practice of making soft ware components reusable via service interfaces. Microservices architecture goes further to make components smaller and simpler.

How do you design interactions in a distributed system to prevent failures?

Distributed systems rely on communications networks to interconnect components, such as servers or services. Your workload must operate reliably despite data loss or latency in these networks. Components of the distributed system must operate in a way that does not neg atively impact other components or the workload. These best practices prevent failures and improve mean time between failures (MTBF).

How do you design interactions in a distributed system to mitigate or withstand failures?

Distributed systems rely on communications networks to interconnect components (such as servers or services). Your workload must operate reliably despite data loss or latency over these networks. Components of the distributed system must operate in a way that does not negatively impact other components or the workload. These best practices enable workloads to withstand stresses or failures, more quickly recover from them, and mitigate the impact of such impairments. The result is improved mean time to recovery (MTTR).

Change Management

Changes to your workload or its environment must be anticipated and accommodated to achieve reliable operation of the workload. Changes include those imposed on your workload, such as spikes in demand, as well as those from within, such as feature deployments and security patches.

How do you monitor workload resources?

Logs and metrics are powerful tools to gain insight into the health of your workload. You can configure your workload to monitor logs and metrics and send notifications when thresholds are crossed or significant events occur. Monitoring enables your workload to recognize when low-performance thresholds are crossed or failures occur, so it can recover automatically in response.

How do you design your workload to adapt to changes in demand?

A scalable workload provides elasticity to add or remove resources automatically so that they closely match the current demand at any given point in time.

How do you implement change?

Controlled changes are necessary to deploy new functionality, and to ensure that the workloads and the operating environment are running known software and can be patched or replaced in a predictable manner. If these changes are uncontrolled, then it makes it difficult to predict the effect of these changes, or to address issues that arise because of them.

Failure Management

In any system of reasonable complexity, it is expected that failures will occur. Reliability requires that your workload be aware of failures as they occur and take action to avoid impact on availability. Workloads must be able to both withstand failures and automatically repair issues.

How do you back up data?

Back up data, applications, and configuration to meet your requirements for recovery time objectives (RTO) and recovery point objectives (RPO).

How do you use fault isolation to protect your workload?

Fault isolated boundaries limit the effect of a failure within a workload to a limited number of components. Components outside of the boundary are unaffected by the failure. Using multiple fault isolated boundaries, you can limit the impact on your workload.

How do you design your workload to withstand component failures?

Workloads with a requirement for high availability and low mean time to recovery (MTTR) must be architected for resiliency.

How do you test reliability?

After you have designed your workload to be resilient to the stresses of production, testing is the only way to ensure that it will operate as designed, and deliver the resiliency you expect.

How do you plan for disaster recovery (DR)?

Having backups and redundant workload components in place is the start of your DR strategy. RTO and RPO are your objectives for restoration of availability. Set these based on business needs. Implement a strategy to meet these objectives, considering locations and function of

workload resources and data.

Fourth Pillar - Performance Efficiency

When it comes to performance efficiency, the recommended design best practices are as follows: Democratize advanced technologies. Take advantage of AWS's global infrastructure to deploy your application globally with minimal cost and to provide low latency. Leverage serverless architectures wherever possible. Deploy multiple configurations to see which one delivers better performance.

- Democratize advanced technologies
- Go global in minutes
- Use serverless architectures
- Experiment more often

There are four best practice areas for performance efficiency in the cloud:

Selection

The optimal solution for a particular workload varies, and solutions often combine multiple approaches. Well-architected workloads use multiple solutions and enable different features to improve performance.

How do you select the best performing architecture?
Often, multiple approaches are required for optimal performance across a workload. Well-architected systems use multiple solutions and features to improve performance.

Review

Cloud technologies are rapidly evolving and you must ensure that workload components are using the latest technologies and approaches to continually improve performance. You must continually evaluate and consider changes to your workload components to ensure you are meeting its performance and cost objectives. New technologies, such as machine learning and artificial intelligence (AI), can allow you to re-imagine customer experiences and innovate across all of your business workloads.

How do you evolve your workload to take advantage of new releases?

When architecting workloads, there are finite options that you can choose from. However, over time, new technologies and approaches become available that could improve the performance of your workload.

Monitoring

After you implement your workload, you must monitor its performance so that you can remediate any issues before they impact your customers. Monitoring metrics should be used to raise alarms when thresholds are breached.

How do you monitor your resources to ensure they are performing?
System performance can degrade over time. Monitor system performance to identify degra dation and remediate internal or external factors, such as the operating system or application load.

Tradeoffs

When you architect solutions, think about tradeoffs to ensure an optimal approach. Depending on your situation, you could trade consistency, durability, and space for time or latency, to deliver higher performance.

How do you use tradeoffs to improve performance?
When architecting solutions, determining tradeoffs enables you to select an optimal approach. Often you can improve performance by trading consistency, durability, and space for time and latency.

Fifth Pillar - Cost Optimization

To enhance cost optimization, these principles are suggested: Use a consumption model. Leverage economies of scale whenever possible. Reduce expenses by limiting the use of company-owned data centers. Constantly analyze and account for infrastructure expenses. Whenever possible, use AWS - managed services instead of services that you need to manage yourself. This should lower your administration expenses.

- Implement Cloud Financial Management
- Adopt a consumption model
- Measure overall efficiency

- Stop spending money on undifferentiated heavy lifting
- Analyze and attribute expenditure

There are five best practice areas for cost optimization in the cloud:

Practice Cloud Financial Management

With the adoption of cloud, technology teams innovate faster due to shortened approval, procurement, and infrastructure deployment cycles. A new approach to financial management in the cloud is required to realize business value and financial success. This approach is Cloud Financial Management, and builds capability across your organization by implementing organizational wide knowledge building, programs, resources, and processes.

How do you implement cloud financial management?
Implementing Cloud Financial Management enables organizations to realize business value and financial success as they optimize their cost and usage and scale on AWS.

Expenditure and usage awareness

The increased flexibility and agility that the cloud enables encourages innovation and fast-paced development and deployment. It eliminates the manual processes and time associated with provisioning on-premises infrastructure, including identifying hardware specifications, negotiating price quotations, managing purchase orders, scheduling shipments, and then deploying the resources. However, the ease of use and virtually unlimited on-demand capacity requires a new way of thinking about expenditures.

How do you govern usage?
Establish policies and mechanisms to ensure that appropriate costs are incurred while objec tives are achieved. By employing a checks-and-balances approach, you can innovate without overspending.

How do you monitor usage and cost?
Establish policies and procedures to monitor and appropriately allocate your costs. This allows you to measure and improve the cost efficiency of this workload.

How do you decommission resources?

Implement change control and resource management from project inception to end-of-life. This ensures you shut down or terminate unused resources to reduce waste.

Cost-effective resources

Using the appropriate instances and resources for your workload is key to cost savings. For example, a reporting process might take five hours to run on a smaller server but one hour to run on a larger server that is twice as expensive. Both servers give you the same outcome, but the smaller server incurs more cost over time.

How do you evaluate cost when you select services?

Amazon EC2, Amazon EBS, and Amazon S3 are building-block AWS services. Managed services, such as Amazon RDS and Amazon DynamoDB, are higher level, or application level, AWS services. By selecting the appropriate building blocks and managed services, you can optimize this workload for cost. For example, using managed services, you can reduce or remove much of your administrative and operational overhead, freeing you to work on applications and business-related activities.

How do you meet cost targets when you select resource type, size and number?

Ensure that you choose the appropriate resource size and number of resources for the task at hand. You minimize waste by selecting the most cost effective type, size, and number.

How do you use pricing models to reduce cost?

Use the pricing model that is most appropriate for your resources to minimize expense.

How do you plan for data transfer charges?

Ensure that you plan and monitor data transfer charges so that you can make architectural decisions to minimize costs. A small yet effective architectural change can drastically reduce your operational costs over time.

Manage demand and supply resources

When you move to the cloud, you pay only for what you need. You can supply resources to match the workload demand at the time they’re needed, this eliminates the need for costly and wasteful over provisioning. You can also modify the demand, using a throttle, buffer, or queue to smooth the demand and serve it with less resources resulting in a lower cost, or process it at a later time with a batch service.

How do you manage demand, and supply resources?

For a workload that has balanced spend and performance, ensure that everything you pay for is used and avoid significantly underutilizing instances. A skewed utilization metric in either direction has an adverse impact on your organization, in either operational costs (degraded performance due to over-utilization), or wasted AWS expenditures (due to over-provisioning).

Optimize over time

As AWS releases new services and features, it's a best practice to review your existing architectural decisions to ensure they continue to be the most cost effective. As your requirements change, be aggressive in decommissioning resources, entire services, and systems that you no longer require.

How do you evaluate new services?

As AWS releases new services and features, it's a best practice to review your existing archi tectural decisions to ensure they continue to be the most cost effective.