Operational Excellence

1. 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.

- Evaluate external customer needs
- Evaluate internal customer needs
- Evaluate governance requirements
- Evaluate compliance requirements
- Evaluate threat landscape
- Evaluate tradeoffs
- Manage benefits and risks

2. 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.

- Resources have identified owners
- Processes and procedures have identified owners
- Operations activities have identified owners responsible for their performance
- Team members know what they are responsible for
- Mechanisms exist to identify responsibility and ownership
- Mechanisms exist to request additions, changes, and exceptions
- Responsibilities between teams are predefined or negotiated

3. 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.

- Executive Sponsorship
- Team members are empowered to take action when outcomes are at risk
- Escalation is encouraged
- Communications are timely, clear, and actionable
- Experimentation is encouraged
- Team members are enabled and encouraged to maintain and grow their skill sets
- Resource teams appropriately
- Diverse opinions are encouraged and sought within and across teams

4. 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.

- Implement application telemetry
- Implement and configure workload telemetry
- Implement user activity telemetry
- Implement dependency telemetry
- Implement transaction traceability

5. 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.

- Use version control
- Test and validate changes
- Use configuration management systems
- Use build and deployment management systems
- Perform patch management
- Share design standards
- Implement practices to improve code quality
- Use multiple environments
- Make frequent, small, reversible changes
- Fully automate integration and deployment

6. 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.

- Plan for unsuccessful changes
- Test and validate changes
- Use deployment management systems
- Test using limited deployments
- Deploy using parallel environments
- Deploy frequent, small, reversible changes
- Fully automate integration and deployment
- Automate testing and rollback

7. 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.

- Ensure personnel capability
- Ensure consistent review of operational readiness
- Use runbooks to perform procedures
- Use playbooks to investigate issues
- Make informed decisions to deploy systems and changes

8. 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.

- Identify key performance indicators
- Define workload metrics
- Collect and analyze workload metrics
- Establish workload metrics baselines
- Learn expected patterns of activity for workload
- Alert when workload outcomes are at risk

- Alert when workload anomalies are detected
- Validate the achievement of outcomes and the effectiveness of KPIs and metrics

9. 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.

- Identify key performance indicators
- Define operations metrics
- Collect and analyze operations metrics
- Establish operations metrics baselines
- Learn the expected patterns of activity for operations
- Alert when operations outcomes are at risk
- Alert when operations anomalies are detected
- Validate the achievement of outcomes and the effectiveness of KPIs and metrics

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

- Use processes for event, incident, and problem management
- Have a process per alert
- Prioritize operational events based on business impact
- Define escalation paths
- Enable push notifications
- Communicate status through dashboards
- Automate responses to events

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

- Have a process for continuous improvement
- Perform post-incident analysis
- Implement feedback loops
- Perform Knowledge Management
- Define drivers for improvement
- Validate insights
- Perform operations metrics reviews
- Document and share lessons learned
- Allocate time to make improvements

Performance Efficiency

1. 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.

- Understand the available services and resources
- Define a process for architectural choices
- Factor cost requirements into decisions
- Use policies or reference architectures
- Use guidance from your cloud provider or an appropriate partner
- Benchmark existing workloads
- Load test your workload

2. How do you select your compute solution?
The optimal compute solution for a workload varies based on application design, usage patterns, and configuration settings. Architectures can use different compute solutions for various components and enable different features to improve performance. Selecting the wrong compute solution for an architecture can lead to lower performance efficiency.

- Evaluate the available compute options
- Understand the available compute configuration options
- Collect compute-related metrics
- Determine the required configuration by right-sizing
- Use the available elasticity of resources
- Re-evaluate compute needs based on metrics

3. How do you select your storage solution?
The optimal storage solution for a system varies based on the kind of access method (block, file, or object), patterns of access (random or sequential), required throughput, frequency of access (online, offline, archival), frequency of update (WORM, dynamic), and availability and durability constraints. Well-architected systems use multiple storage solutions and enable different features to improve performance and use resources efficiently.

- Understand storage characteristics and requirements
- Evaluate available configuration options
- Make decisions based on access patterns and metrics

4. How do you select your database solution?
The optimal database solution for a system varies based on requirements for availability, consistency, partition tolerance, latency, durability, scalability, and query capability. Many systems use different database solutions for various subsystems and enable different features to improve performance. Selecting the wrong database solution and features for a system can lead to lower performance efficiency.

- Understand data characteristics
- Evaluate the available options
- Collect and record database performance metrics
- Choose data storage based on access patterns
- Optimize data storage based on access patterns and metrics

5. How do you configure your networking solution?
The optimal network solution for a workload varies based on latency, throughput requirements, jitter, and bandwidth. Physical constraints, such as user or on-premises resources, determine location options. These constraints can be offset with edge locations or resource placement.

- Understand how networking impacts performance
- Evaluate available networking features
- Choose appropriately sized dedicated connectivity or VPN for hybrid workloads
- Leverage load-balancing and encryption offloading
- Choose network protocols to improve performance
- Choose your workload's location based on network requirements
- Optimize network configuration based on metrics

6. 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.

- Stay up-to-date on new resources and services
- Define a process to improve workload performance
- Evolve workload performance over time

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

- Record performance-related metrics
- Analyze metrics when events or incidents occur
- Establish Key Performance Indicators (KPIs) to measure workload performance
- Use monitoring to generate alarm-based notifications
- Review metrics at regular intervals
- Monitor and alarm proactively

8. 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.

- Understand the areas where performance is most critical
- Learn about design patterns and services
- Identify how tradeoffs impact customers and efficiency
- Measure the impact of performance improvements
- Use various performance-related strategies

Security

1. 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.

- Separate workloads using accounts
- Secure AWS account
- Identify and validate control objectives
- Keep up to date with security threats
- Keep up to date with security recommendations
- Automate testing and validation of security controls in pipelines
- Identify and prioritize risks using a threat model
- Evaluate and implement new security services and features regularly

2. 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.

- Use strong sign-in mechanisms
- Use temporary credentials
- Store and use secrets securely
- Rely on a centralized identity provider
- Audit and rotate credentials periodically
- Leverage user groups and attributes

3. 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.

- Define access requirements
- Grant least privilege access
- Establish emergency access process
- Reduce permissions continuously
- Define permission guardrails for your organization
- Manage access based on life cycle
- Analyze public and cross account access
- Share resources securely

4. 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.

- Configure service and application logging
- Analyze logs, findings, and metrics centrally
- Automate response to events
- Implement actionable security events

5. 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.

- Create network layers
- Control traffic at all layers
- Automate network protection
- Implement inspection and protection

6. How do you protect your compute resources?
Compute resources in your workload require multiple layers of defence to help protect from external and internal threats. Compute resources include EC2 instances, containers, AWS Lambda functions, database services, IoT devices, and more.

- Perform vulnerability management
- Reduce attack surface
- Implement managed services
- Automate compute protection
- Enable people to perform actions at a distance
- Validate software integrity

7. 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.

- Identify the data within your workload
- Define data protection controls
- Automate identification and classification
- Define data lifecycle management

8. 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.

- Implement secure key management
- Enforce encryption at rest
- Automate data at rest protection
- Enforce access control
- Use mechanisms to keep people away from data

9. 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.

- Implement secure key and certificate management
- Enforce encryption in transit
- Automate detection of unintended data access
- Authenticate network communications

10. 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.

- Identify key personnel and external resources
- Develop incident management plans
- Prepare forensic capabilities
- Automate containment capability
- Pre-provision access
- Pre-deploy tools
- Run game days

Cost Optimization

1. 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.

- Establish a cost optimization function
- Establish a partnership between finance and technology
- Establish cloud budgets and forecasts
- Implement cost awareness in your organizational processes
- Report and notify on cost optimization
- Monitor cost proactively
- Keep up to date with new service releases

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

- Develop policies based on your organization requirements
- Implement goals and targets
- Implement an account structure
- Implement groups and roles
- Implement cost controls
- Track project lifecycle

3. 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.

- Configure detailed information sources
- Identify cost attribution categories
- Establish organization metrics
- Configure billing and cost management tools
- Add organization information to cost and usage
- Allocate costs based on workload metrics

4. 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.

- Track resources over their life time
- Implement a decommissioning process
- Decommission resources
- Decommission resources automatically

5. 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.

- Identify organization requirements for cost
- Analyze all components of this workload
- Perform a thorough analysis of each component
- Select software with cost effective licensing
- Select components of this workload to optimize cost in line with organization priorities
- Perform cost analysis for different usage over time

6. 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.

- Perform cost modelling
- Select resource type, size, and number based on data
- Select resource type, size, and number automatically based on metrics

7. How do you use pricing models to reduce cost?
Use the pricing model that is most appropriate for your resources to minimize expense.

- Perform pricing model analysis
- Implement regions based on cost
- Select third party agreements with cost efficient terms
- Implement pricing models for all components of this workload
- Perform pricing model analysis at the master account level

8. 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.

- Perform data transfer modeling
- Select components to optimize data transfer cost
- Implement services to reduce data transfer costs

9. 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).

- Perform an analysis on the workload demand
- Implement a buffer or throttle to manage demand
- Supply resources dynamically

10. How do you evaluate new services?
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.

- Develop a workload review process
- Review and analyze this workload regularly

Reliability

1. 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.

- Aware of service quotas and constraints
- Manage service quotas across accounts and regions
- Accommodate fixed service quotas and constraints through architecture
- Monitor and manage quotas
- Automate quota management
- Ensure that a sufficient gap exists between the current quotas and the maximum usage to accommodate failover

2. 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.

- Use highly available network connectivity for your workload public endpoints
- Provision redundant connectivity between private networks in the cloud and on-premises environments
- Ensure IP subnet allocation accounts for expansion and availability
- Prefer hub-and-spoke topologies over many-to-many mesh
- Enforce non-overlapping private IP address ranges in all private address spaces where they are connected

3. 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 software components reusable via service interfaces. Microservices architecture goes further to make components smaller and simpler.

- Choose how to segment your workload
- Build services focused on specific business domains and functionality
- Provide service contracts per API

4. 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 negatively impact other components or the workload. These best practices prevent failures and improve mean time between failures (MTBF).

- Identify which kind of distributed system is required
- Implement loosely coupled dependencies
- Do constant work
- Make all responses idempotent

5. 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).

- Implement graceful degradation to transform applicable hard dependencies into soft dependencies
- Throttle requests
- Control and limit retry calls
- Fail fast and limit queues
- Set client timeouts
- Make services stateless where possible
- Implement emergency levers

6. 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.

- Monitor all components for the workload (Generation)
- Define and calculate metrics (Aggregation)
- Send notifications (Real-time processing and alarming)
- Automate responses (Real-time processing and alarming)
- Analytics
- Conduct reviews regularly
- Monitor end-to-end tracing of requests through your system

7. 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.

- Use automation when obtaining or scaling resources
- Obtain resources upon detection of impairment to a workload
- Obtain resources upon detection that more resources are needed for a workload
- Load test your workload

8. 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.

- Use runbooks for standard activities such as deployment
- Integrate functional testing as part of your deployment
- Integrate resiliency testing as part of your deployment
- Deploy using immutable infrastructure
- Deploy changes with automation

9. 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).

- Identify and back up all data that needs to be backed up, or reproduce the data from sources
- Secure and encrypt backups
- Perform data backup automatically
- Perform periodic recovery of the data to verify backup integrity and processes

10. 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.

- Deploy the workload to multiple locations
- Select the appropriate locations for your multi-location deployment

- Automate recovery for components constrained to a single location
- Use bulkhead architectures to limit scope of impact

11. 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.

- Monitor all components of the workload to detect failures
- Fail over to healthy resources
- Automate healing on all layers
- Rely on the data plane and not the control plane during recovery
- Use static stability to prevent bimodal behaviour
- Send notifications when events impact availability

12. 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.

- Use playbooks to investigate failures
- Perform post-incident analysis
- Test functional requirements
- Test scaling and performance requirements
- Test resiliency using chaos engineering
- Conduct game days regularly

13. 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 your workload. Set these based on business needs. Implement a strategy to meet these objectives, considering locations and function of workload resources and data. The probability of disruption and cost of recovery are also key factors that help to inform the business value of providing disaster recovery for a workload.

- Define recovery objectives for downtime and data loss
- Use defined recovery strategies to meet the recovery objectives
- Test disaster recovery implementation to validate the implementation
- Manage configuration drift at the DR site or Region
- Automate recovery

Sustainability

1. How do you select Regions to support your sustainability goals?
Choose Regions where you will implement your workloads based on both your business requirements and sustainability goals.

- Choose Regions near Amazon renewable energy projects and Regions where the grid has a published carbon intensity that is lower than other locations (or Regions).

2. How do you take advantage of user behavior patterns to support your sustainability goals?
The way users consume your workloads and other resources can help you identify improvements to meet sustainability goals. Scale infrastructure to continually match user load and ensure that only the minimum resources required to support users are deployed. Align service levels to customer needs. Position resources to limit the network required for users to consume them. Remove existing, unused assets. Identify created assets that are unused and stop generating them. Provide your team members with devices that support their needs with minimized sustainability impact.

- Scale infrastructure with user load
- Align SLAs with sustainability goals
- Stop the creation and maintenance of unused assets
- Optimize geographic placement of workloads for user locations
- Optimize team member resources for activities performed

3. How do you take advantage of software and architecture patterns to support your sustainability goals?
Implement patterns for performing load smoothing and maintaining consistent high utilization of deployed resources to minimize the resources consumed. Components might become idle from lack of use because of changes in user behavior over time. Revise patterns and architecture to consolidate under-utilized components to increase overall utilization. Retire components that are no longer required. Understand the performance of your workload components, and optimize the components that consume the most resources. Be aware of the devices your customers use to access your services, and implement patterns to minimize the need for device upgrades.

- Optimize software and architecture for asynchronous and scheduled jobs
- Remove or refactor workload components with low or no use
- Optimize areas of code that consume the most time or resources
- Optimize impact on customer devices and equipment
- Use software patterns and architectures that best support data access and storage patterns

4. How do you take advantage of data access and usage patterns to support your sustainability goals?
Implement data management practices to reduce the provisioned storage required to support your workload, and the resources required to use it. Understand your data, and use storage technologies and configurations that best support the business value of the data and how it's used. Lifecycle data to more efficient, less performant storage when requirements decrease, and delete data that's no longer required.

- Implement a data classification policy
- Use technologies that support data access and storage patterns
- Use lifecycle policies to delete unnecessary data
- Minimize over-provisioning in block storage
- Remove unneeded or redundant data

- Use shared file systems or object storage to access common data
- Minimize data movement across networks
- Back up data only when difficult to recreate

5. How do your hardware management and usage practices support your sustainability goals?
Look for opportunities to reduce workload sustainability impacts by making changes to your hardware management practices. Minimize the amount of hardware needed to provision and deploy, and select the most efficient hardware for your individual workload.

- Use the minimum amount of hardware to meet your needs
- Use instance types with the least impact
- Use managed services
- Optimize your use of GPUs

6. How do your development and deployment processes support your sustainability goals?
Look for opportunities to reduce your sustainability impact by making changes to your development, test, and deployment practices

- Adopt methods that can rapidly introduce sustainability improvements
- Keep your workload up to date
- Increase utilization of build environments
- Use managed device farms for testing